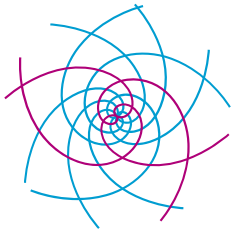
WWU
MÜNSTER

Mutual Information based Parameter Extraction for Spreading Cell Colonies

Doctoral Thesis

Ramona Sasse

- 2021 -



MM
Mathematics
Münster
Cluster of Excellence

living.knowledge

Fach: Mathematik

Mutual Information based Parameter Extraction for Spreading Cell Colonies

Inauguraldissertation

zur Erlangung des Doktorgrades der Naturwissenschaften

– Dr. rer. nat. –

im Fachbereich Mathematik und Informatik

der Mathematisch-Naturwissenschaftlichen Fakultät

der Westfälischen Wilhelms-Universität Münster

eingereicht von

Ramona Sasse

aus Coesfeld

– 2021 –

Dekan:	Prof. Dr. Xiaoyi Jiang
Erster Gutachter:	Prof. Dr. Benedikt Wirth Westfälische Wilhelms-Universität Münster
Zweiter Gutachter:	Prof. Dr. Martin Burger Friedrich-Alexander-Universität Erlangen-Nürnberg
Tag der mündlichen Prüfung:	23.07.2021
Tag der Promotion:	23.07.2021

To Christin and Stephan, who always believed in me,

and to my parents for their love and support throughout my life!

Zusammenfassung

Bis heute ist die Entwicklung und das Wachstum von Tumorzellkolonien noch nicht in Gänze erforscht. Insbesondere wenn es zu speziellen Ausbreitungsphänomenen kommt, die sich dadurch auszeichnen, dass innerhalb der Kolonie eine Zellmasse mit einem auffallend anderem Erscheinungsbild entsteht, sind Wissenschaftler daran interessiert, dieses Auftreten und Anwachsen besser zu verstehen. Das Ableiten von Ausbreitungsmerkmalen für die gesamte Zellkolonie als auch Wachstumseigenschaften für diese innere Teilkolonie sind wichtige Forschungsfragen in der Grundlagenforschung im biomedizinischen Umfeld. Von besonderem Interesse ist auch der Einfluss von Medikamenten auf dieses dynamische Zellverhalten. Basierend auf einer Kooperation mit Wissenschaftlern des Pharmakonzerns AstraZeneca UK Ltd. erforschen wir das Ausbreitungsphänomen dieser Zellkolonien von einem mathematischen Standpunkt aus und verwenden dazu Mikroskopiebilder, die eben diesen Wachstumsprozess über die Zeit abbilden. In dem Kontext dieser Arbeit untersuchen wir das Ausbreitungsverhalten von Zellpopulationen basierend auf einer initialen Krebszelle eines Lungentumors. In den betrachteten Phasenkontrastbildern beobachten wir ebenfalls zwei Teilkolonien mit markanten Texturunterschieden in den Mikroskopiebildern. Deswegen sind wir besonders interessiert an der Untersuchung von Ausbreitungsverhalten dieser zwei Teilkolonien.

Mit Hilfe von Modellierung und mathematischer Optimierung zielen wir darauf ab, Ausbreitungseigenschaften herauszufiltern, die diesen Prozess beschreiben. Dazu bedienen wir uns des Konzeptes der *Transinformation* (engl.: *mutual information*), die den gegenseitigen Informationsgehalt von zwei Zufallsgrößen beschreibt (*frei übersetzt nach [20]*). In unserem Fall konzentrieren wir uns auf die Transinformation zwischen zwei Bildern. Dazu leiten wir einerseits sogenannte *Merkmalsbilder* (engl.: *feature images*) basierend auf lokalen Textureigenschaften der Phasenkontrastbilder her. Zum anderen generieren wir mit Hilfe eines mathematischen Ausbreitungsmodells *Klassifizierungsbilder* (engl.: *classification images*). Mit diesen teilen wir ein Bild in unterschiedliche Teilbereiche, sogenannte "Klassen" ein. Das Modell basierend auf bestimmten Ausbreitungseigenschaften beschreibt zum Beispiel, ob in einem Punkt Zellen zu erwarten sind oder nicht. Auf dieser Grundlage können wir den Bildbereich in Hintergrundbereiche *ohne* Zellen und Vordergrundbereiche *mit* Zellen einteilen. Indem wir die Transinformation zwischen den Merkmalsbildern und den Klassifizierungsbildern maximieren, können wir Ausbreitungseigenschaften ableiten, die direkt mit dem Klassifizierungsbild oder vielmehr mit dem Ausbreitungsmodell verknüpft sind. Im Optimum beschreibt die Ausbreitung im Klassifizierungsbild dann möglichst genau die Ausbreitung in den Mikroskopiedaten. Mit diesen Ausbreitungseigenschaften möchten wir Mathematiker einen Beitrag leisten, um das Verständnis für das Wachstum und die Ausbreitung von Tumorzellpopulationen zu schärfen, und so schließlich weitere Grundlagen für die Krebsforschung und Arzneimittelentwicklung schaffen.

Für die Merkmalsbilder nutzen wir lokale Textureigenschaften, die wir direkt aus den Mikroskopiebildern berechnen können. In diesem Zusammenhang befassen wir uns kurz mit der Segmentierung von Zellkolonien, welche die weitere mathematische Beschreibung des Ausbreitungsprozesses motiviert. Dazu führen wir anschließend ein Modell basierend auf Partiellen Differentialgleichungen (PDG) ein. Mit diesem Modell können wir nicht nur die Ausbreitung *einer* Zellkolonie beschreiben, sondern sind zusätzlich in der Lage, zwei Untergruppen innerhalb der Kolonie mit verschiedenen zellulären Erscheinungsbildern zu identifizieren und zu charakterisieren. Schließlich reduzieren wir das Modell

zu einer vereinfachten Version. Dieses vereinfachte Modell erfüllt weiterhin wichtige Kerneigenschaften des PDG-Modells und beschreibt zwei sich konzentrisch, d.h. kreisförmig mit gemeinsamen Mittelpunkt, ausbildende Koloniefronten, die hintereinander durch das Gebiet propagieren.
Mit dem Optimierungsproblem zur Maximierung der Transinformation berücksichtigen wir schließlich zugleich die Texturinformationen aus den Mikroskopiebildern anhand der Merkmalsbilder gemeinsam mit den Modelleigenschaften inhärent in den Klassifizierungsbildern. In diesem Sinne ist ein Großteil der vorliegenden Arbeit auf eine genaue mathematische Analyse des zugrundeliegenden Optimierungsproblems ausgerichtet. Tatsächlich führen wir ein Minimierungsproblem ein, um die *negative* Transinformation zu optimieren. Dazu beschäftigen wir uns mit Konvergenzaussagen bezüglich verschiedener Diskretisierungen sowie der Konvergenz von Minimierern.
Ein weiteres zentrales Thema dieser Dissertation ist die numerische Lösung des Optimierungsproblems. Als Konzeptnachweis betrachten wir ein vereinfachtes Beispielproblem, in dem wir uns künstlich erzeugter, simulierter Klassifizierungs- und Merkmalsbilder bedienen. Danach werden beispielhaft zwei Zeitreihen mit sich ausbreitenden Zellkolonien des AstraZeneca-Datensatzes verarbeitet.
Zum Abschluss der Arbeit diskutieren wir das Potential, spezielle Herausforderungen und mögliche Erschwernisse im Zusammenhang mit dem gewählten Ansatz. Zudem wird ein kurzer Ausblick auf Zukunftsstudien gegeben, die unter anderem mögliche Modellverbesserungen oder die Verwendung von weiterem Vorwissen berücksichtigen.
Insgesamt sehen wir besonderes Potential in unserem Ansatz, der einerseits die Unterscheidung zweier Teilkolonien erlaubt und andererseits mathematische Optimierung für die Modellierung von Ausbreitungsprozessen einer Zellkolonie mit Hilfe von Bilddaten verknüpft und somit dedizierte Aussagen zum Ausbreitungsverhalten ermöglicht. Dies ist ein besonderer Vorteil gegenüber Klassifizierungsmethoden, die lediglich die Bestimmung von verschiedenen zellulären Strukturen und Erscheinungen zum Beispiel unterstützen.

Abstract

Until today, the development of a growing cell colony due to spatial cell spreading and cell division based on a single initial cancer cell is not yet fully understood. Especially when it comes to certain spreading phenomena, where the colony not only grows but researchers also observe an inner bulk developing within the colony with a significantly different appearance, we aim to improve the understanding of colony growth. The influence of medical treatments and drug treatments on the dynamic behavior of cells is of special interest in this context. The extraction of spreading characteristics for the total colony as well as growth properties for this inner subcolony are important tasks in biomedical basic research and of particular importance with respect to the development of future cancer treatments. Based on a joint project with scientists from the pharmaceutical company AstraZeneca UK Ltd., we tackle the spreading phenomenon of cell colonies from a mathematical perspective in this thesis. We investigate the growth of cell colonies started from single cells which were originally derived from lung tumor tissue. In the inspected phase contrast images, we observe two subcolonies with strikingly different texture appearances in the microscopy images. For this reason, we are particularly interested in the investigation of the spreading of *two* subcolonies.

By means of model fitting via a *mutual information* (MI) based optimization approach, we aim for the extraction of special spreading properties that correlate to the colony growth process observable in microscopy data. The MI describes the mutual amount of information between two random variables [20]. In our case, we are interested in the mutual information between two images. To be more precise, we derive *feature images* based on extracted texture features from the phase contrast images which facilitate the localization of a spreading cell colony. Next, we generate *classification images* based on a mathematical spreading model. With this we identify different parts within an image and *classify* them according to certain characteristics. For example, we identify empty background regions *without* cells and foreground regions *containing* cells on the basis of a given model that identifies whether a specified location contains cells or not. The model itself is directly connected with certain spreading properties. By maximizing the mutual information between these features and classification images, we derive *spreading properties* that are closely linked to the classification images and, consequently, to the underlying growth model. With the derived spreading properties, we facilitate the analysis of spreading cancer cells investigated in the drug development process.

For the feature images, we exploit basic local texture features directly extracted from the microscopy images. We briefly touch upon segmenting the cell colonies which motivates further the investigation of the spreading cell colonies from a mathematical point of view. For this purpose, we introduce a mathematical model based on partial differential equations (PDEs). We emphasize that instead of merely identifying a colony's area, we focus on a model that enables differentiating even between *two* subpopulations of cellular appearances *within* a colony. Then, we present a simplified spreading approach which still preserves the fundamental properties of the previous PDE model and can capture two consecutive concentric spreading colony fronts.

With the derived MI-based optimization problem, we consider the texture features from the original imaging data and the mathematical spreading model jointly. One central topic of the present thesis deals with a thorough analysis of this optimization problem. To this end, we introduce a minimization problem to optimize the *negative* MI. This way, we focus on various related convergence statements

including the convergence of discretizations and the convergence of minimizers.
A second major topic is the numerical solution of our optimization problem. As a proof of concept, we first consider a simplified toy problem consisting of simulated classification and feature images. Secondly, we test the optimization approach on two exemplary time series of the AstraZeneca data set capturing the growth process of two colonies. We conclude this work with a discussion of the applied approach by considering its potential, related challenges and possible obstacles. Finally, we give an outlook on conceivable future studies including model improvements and using more prior knowledge in the optimization approach.
We highlight that the power of our approach lies within the differentiation of two subcolonies on the one hand. On the other hand, we see a particular potential in using mathematical optimization techniques for fitting cell spreading models to imaging data to facilitate the extraction of novel colony growth properties. This is a major advantage of our approach over classification methods which are solely used to distinguish different cellular appearances for example.

Keywords: mathematical optimization, mutual information, parameter extraction, spreading model, PDE model for cell colonies, model fitting, convergence of discretizations, segmentation, image analysis, texture features in microscopy images, cell colony growth, single cell cloning

Acknowledgments

Contents

1

Introduction

The focus of this thesis is the investigation of spreading cell colonies from a mathematical perspective. We deal with a data set of microscopy images capturing the growth process of cancerous cell colonies. The research project was initiated during a three month secondment to the industrial host AstraZeneca in Cambridge in the summer of 2018. Cancer research is a highly relevant field in the pharmaceutical industry and scientists not only at AstraZeneca are aiming for a better understanding of developing cancer cell populations. For example, in [75] the authors observed a significant increase in research and publications related to pediatric cancer over a time span of 10 years from 2007 to 2016. In our setting, biologists at AstraZeneca investigate the growing and spreading process of cell colonies descending from one initial cancer cell which originates from a lung tumor. In the "Global cancer statistics 2020" by Sung et al., it is stated that "with an estimated 2.2 million new cancer cases and 1.8 million deaths, lung cancer is the second most commonly diagnosed cancer and the leading cause of cancer death in 2020" which stresses depressingly the high topicality of research for this type of cancer [74]. To facilitate the understanding of the growth process for such a cancer cell population and its interactions with potential drug treatments, we are tackling the research question from a mathematical point of view.

The collaboration with AstraZeneca is supported by the international program on Nonlocal Methods for Arbitrary Data Sources (NoMADS). International researchers from computer vision and applied mathematics cooperate in this multidisciplinary project with companies fostering a better understanding of the current challenges in industrial applications and aiming for higher applicability of nonlocal methods while building up a fundamental mathematical background of the new principles at the same time [58]. In the joint project with AstraZeneca, we as applied mathematicians incorporate computer-aided cell colony classification based on *non-local* texture information of the microscopy images, which highlights that this project fits perfectly into the NoMADS context. The texture characteristics extracted from the grayscale phase contrast images are essential to identify cells and recognize the spreading colony in comparison to the domain's background regions. In Figure 1.1, we visualize such a spreading colony in sketches for three different time points. In the first frame, we observe only one initial cell. The colony starts to develop and spread in the following time frames due to cell proliferation, i.e., cell division, and cell migration.

The idea to investigate spreading cell colonies from a mathematical perspective comes from current tasks that researchers at AstraZeneca are facing when dealing with phase contrast images capturing the development process. In their research for new medical treatments, they are interested in not only finding empty wells, i.e., identifying wells without a growing colony, but also in the effect of

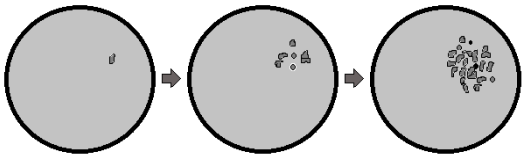


Figure 1.1: The development of a growing cell population sketched for three example time points.

different chemical perturbations on the colony growth process. It would be also interesting to detect wells where a colony starts from not only one but even more than one initial cell. However, this is an open challenge for future studies and we focus here on the colony growth process assuming in general only *one* initial cell as the origin of a the colony. By combining mathematical modeling, optimization techniques and image processing, we develop a novel approach to analyze the spreading cell populations. Spreading properties based on a mathematical model are derived directly from the phase contrast image data with a newly implemented software tool explicitly for this project. This task-specific spreading information has the potential to facilitate the investigation process and reveal new insights on cell population spreading in the future. By this, mathematicians and biologists in the cooperation aim for improving the time-consuming and cumbersome task of analyzing the data manually and, ultimately, accelerating the drug development process to fight spreading cancer cells.

In the course of this doctoral thesis, we introduce the pharmaceutical background of investigations for developing cell colonies, focus on image processing for the given microscopy data, present possible mathematical models for spreading cell populations and finally derive an optimization problem to extract spreading information for the captured cell populations directly from the imaging data. We start in Chapter 2 with an introduction to the application's background in the pharmaceutical setting. In this context, we give details on the image acquisition process and the experiment's design.
In Chapter 3, we also further discuss the imaging data with respect to particular texture properties. While describing the initial image analysis, we motivate the investigation of the spreading colonies in the data from a mathematics-informatics viewpoint touching common tasks in computer vision. In this chapter, we also highlight remarkable texture changes within cell populations that facilitate the localization of a growing cell population. In Figure 1.2, we present one example microscopy image in which we emphasize the different texture regions with colored frames. In this well, i.e., in this tiny shell-like experimental domain, we observe a growing cell population. For visualization aspects we use here an enhanced contrast within the well's domain to highlight the cell colony and facilitate its detection by eye. While the patch framed blue shows the colony's texture appearance close to the leading edge where individual cells are distinguishable, the red one is located near the center of the colony where single cell detection is rather challenging and complicated. As a reference patch, we present the texture present in the well's background in the green frame. For this sake, we introduce in Section 3.3 *feature images* being based on certain texture characteristics. For this purpose, we apply basic measures derived from the field of descriptive statistics: We extract the minimal and maximal gray value and calculate the interquartile range (cf. [65]) in a local neighborhood. These feature images are particularly important for the later optimization problem as we use them to couple the real data with *classification images* based on a given mathematical spreading model.
In Chapter 4, we introduce two possible modeling approaches that capture a growth process of a cell

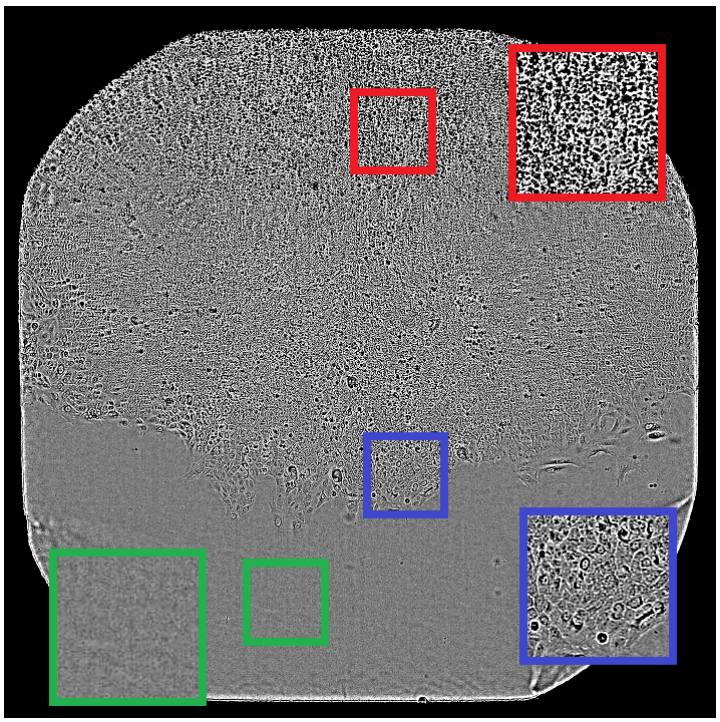


Figure 1.2: An example of a growing colony with highlighted different texture appearances for the domain's background (green), the colony near the leading front (blue) and a patch close to the heart of the colony (red).

colony and correlate to certain spreading characteristics, which we refer to as *spreading properties* in the following. On the one hand, we present in Section 4.1 a spreading model based on partial differential equations (PDEs) which is motivated by the Fisher-Kolmogoroff equation, the Lotka-Volterra system and the SIR[1]-model [55, 56] as well as relates in some sense to an interaction model of two different cell types presented in [34]. On the other hand, we introduce a simplified model based on concentric spreading colony fronts of constant spreading velocity for two subpopulations in Section 4.2. We stress that we are aiming for a model that can distinguish between two different subcolonies. The two subcolonies are supposed to capture the development of the two different texture regions, i.e., to identify populations consisting of cells near the colony's front versus cells close to the colony's center part. Based on the mathematical model, we generate a *classification image* to reveal different subregions within a cell population and to identify background areas in the image domain. In Figure 1.3, we sketch classification images that reflect a concentric spreading colony consisting of two subcolonies (light vs. dark brown) at two example time points.
Having the classification images and the feature images at hand, we concentrate on optimization based model fitting in Chapter 5 to extract spreading information linked to the applied model directly from the texture features. We introduce the concept of mutual information (MI) in Section 5.1 which describes "the amount of information about one random variable function contained in another random function" [20]. In our context, our optimization problem is based on maximizing the mutual information between the classification images and the feature images. To be more precise, we aim for spreading properties related to optimized classification images for which the mutual information

[1] S=susceptible, I=infective, R=recovered people

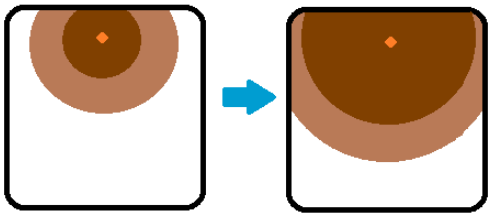


Figure 1.3: A sketch of classification images for two time points revealing a concentric spreading colony which consists of two subcolonies highlighted in light and dark brown regions.

between these and the feature images is maximal. For a more figurative interpretation of this approach, we state that the "optimized" spreading phenomenon depicted in the classification images matches the observed colony spreading in the feature images best when the mutual information between classification and feature images is maximal. Our optimization problem is inspired by mutual information based image registration often applied in the context of images in the medical field [61].
In Section 5.2, we focus on our MI-maximization problem. As we consider a gradient-based numerical solution, we focus on gradient terms of the mutual information. One important statement in this section focuses on derivative terms for histogram measures which are required for the gradient of mutual information (cf. Theorem 5.40). When considering the classification image and the feature image jointly, we calculate joint histograms which are essential to compute the mutual information between both images.
For the numerical approach, we need to consider diverse discretization stages. In Section 5.3, various intermediate convergence results are given that are of certain importance when relating the optimization problem in a *discretized* setting to an originally *continuous* one. For example, we concentrate on the convergence of images defined on discrete pixel grids when considering increasing resolutions. As a second example, we mention discretized histogram measures that are based on discrete histogram bins. For these histogram measures related to the classification and feature images, we prove convergence to the pushforward of the Lebesgue measure of the related spaces with respect to the image mappings. The whole sections serves as a preparation for the profound analysis of the optimization problem in Section 5.4. The main statements of this section are the *existence* and *convergence* of minimizers (cf. Theorems 5.97 and 5.98).
In the final Section 5.5 of this main chapter, we perform numerical evaluations. As a proof of concept, we test our approach of deriving spreading information based on MI-optimization for classification and feature images on an artificial toy example. In a second step, we apply the developed approach to real data of AstraZeneca. For two example time series capturing spreading cell colonies, we numerically solve the optimization problem and fit the concentric spreading model to the texture data extracted from the original microscopy images. In both cases, the numerical solver converges to a minimizer which validates our model.
In our work, we focus on a thorough analytical assessment of the considered optimization approach and present a deeper numerical analysis of the given problem. To this end, we validate numerically the convergence statements for decreasing discretization scales of pixel widths and binning sizes by means of our toy example. Additionally, we evaluate matching texture regions based on their optimized classification values for the example colonies of the AstraZeneca data.

The power of our MI-based model fitting approach lies in the expectation that we are able to determine similar texture regions occurring in different wells' time series by assigning them to similar classification values through the optimization and, thereby, achieve a comparability across the different wells. For example, we aim for a numerical solution to identify background areas vs. a colony's region across all recorded time-lapse phase contrast images based on their nature of texture. We stress that we are not only applying a texture classification approach here. In our approach, we incorporate prior knowledge on an expected spreading behavior of a developing cell colony. To be more precise, we include a specific spreading model into our approach instead of only focusing on a texture classification. With this approach, we are indeed able to extract *spreading properties* for a growing colony captured in the imaging data.

Moreover, we emphasize that our main objective is fitting a model that captures not only *one* spreading colony but which even identifies *two* different subcolonies. We observe that in some parts of a growing cell population single cell detection would be possible — either manually or with a sophisticated segmentation approach. In other parts of a cell population, especially in the inner bulk of a growing cell colony, a striking texture change impedes and even prevents single cell detection. In the considered models, we include not only *one* colony but also the possibility of a *second* colony emerging within the *first* one. The significantly different texture near the center part could correlate to cell debris due to sick or dead cells. Another valid interpretation is that the cells are closely packed in this region and already lie on top of each other. As they start moving out of focus, the imaging device cannot capture their outline accurately enough anymore. A profound analysis of the biological cause of this second strikingly different texture appearances is beyond the scope of this thesis. Therefore, we use figurative names for the two subpopulations throughout this thesis dividing the colony in regions of *normal* cellular appearances versus regions of *abnormal* cellular appearances.

As it turns out with the numerical tests on the real data, the identification of this "second" colony is more challenging than expected. For the two example wells, we observe that our approach rather identifies the *transition* area between background and cell populations as the "first" subcolony, and the main population as the "second" subcolony. In this context, we see great potential in model improvements or incorporating additional prior information on the inner bulk of cells we expected as the "second" colony region. Hence, we discuss the findings of this thesis in the final Chapter 6 and give an outlook on future studies. In this sense, we reflect on the applied approach and consider further improvements on different levels.

2

Application background in the pharmaceutical field

AstraZeneca UK Ltd. is a global pharmaceutical company developing medicines and drugs. Next to cancer research the company is investigating treatments for patients suffering from cardiovascular and metabolic diseases or respiratory, inflammations and autoimmune diseases [5].
In the BioPharmaceuticals R&D unit at AstraZeneca in Cambridge, scientists focus on early stage, preclinical research. They identify and develop new drugs, build mechanistic data packages of existing drugs and interrogate biological systems to help identify new drug targets. In this research unit, the biologists investigate the reaction and behavior of cells when they undergo various treatments. This basic research is essential in the pipeline of developing new or improving existing drugs. In the oncology division, the scientists focus on treatments and drugs related to cancerous diseases. They use microscopy and immunohistochemistry alongside a variety of other non-imaging based destructive techniques to investigate influences on the cell development or the cells' life cycle by highlighting specific structures as for example certain proteins within the cells. In this context, one can identify how a cell responds to a particular drug treatment, indicate if a specific gene was activated or knocked out in an experiment with the help of biomarkers, using fluorescently labeled antibodies to visualize them within the cell context. The biomarkers themselves are measureable parameters used as meaningful indicators to compare different experiments [37].
To account for company secrecy regulations and comply with confidentiality rules, we do not give any further details on the implemented biomarkers in the joint project we are dealing with in this thesis. Still, we state some basic information on the biological experiment and introduce the resulting imaging data in the following sections.

2.1 Image recording and experimental setting

The time-lapse phase contrast image data set we are focusing on in this project captures the development of cell populations over time. We are dealing with a large data set consisting of five multi-well plates with 16×24 wells, i.e., 384 wells, on each plate and imaged at eight discrete time points. One well is a tiny experimental domain with an approximately square surface of width 3.7 mm and height 11.7 mm which is usually filled with a total volume of 0.01 – 0.1 ml. When setting the initial time point

as 0 and referencing the next time points in relation to the first initial time point, the following seven images are recorded approximately two hours, 17 hours, five days and 15 hours, seven days and 17 hours, 11 days and 15 hours, 14 days and 15 hours and 18 days and 15 hours after the initial time point. In Table 2.1, the exact time stamps are given with the date in the format [yyyy-mm-dd] in the first column and for each plate in the following columns the time of day in the format [hh:mm:ss]. While the dates are non-equidistantly distributed over the whole time interval of almost 20 days, the time steps between the recordings per day are more evenly distributed. Since the recording is happening in a high-throughput microscopy system where the plates are selected and entered for imaging in an automatic way, the time steps between the images of consecutive plates are approximately five to 10 minutes apart.

day	plate 1	plate 2	plate 3	plate 4	plate 5
1: 2018-08-08	16:58:10	17:10:54	17:27:58	17:38:07	17:51:26
2: 2018-08-08	19:04:01	19:16:31	19:21:07	19:25:41	19:30:20
3: 2018-08-09	10:08:06	10:15:07	10:19:44	10:24:23	10:29:07
4: 2018-08-14	07:41:37	07:49:10	07:53:44	07:58:18	08:02:56
5: 2018-08-16	10:24:54	10:34:12	10:38:45	10:43:26	10:48:05
6: 2018-08-20	08:37:50	08:44:18	08:48:52	08:53:27	08:58:06
7: 2018-08-23	08:08:21	08:18:07	08:22:48	08:27:29	08:32:14
8: 2018-08-27	08:13:57	08:19:21	08:24:00	08:28:39	08:33:22

Table 2.1: Discrete time stamps of recordings per well plate in the format [yyyy-mm-dd] indicating the date in the first column and the time of day in the format [hh:mm:ss] in the following columns per plate represented.

The wells on a plate are numbered in an alpha-numerical way. Each row is identified by a letter, starting with A at the top row and ending with P in the 16^{th} row, while the columns are numbered from 1 to 24. In Figure 2.1, an example well plate is shown. With a close look, the column and row identifiers are observable. Moreover, the hand which holds the plate serves as a reference system highlighting the total size of the well plate.
The biologists use for imaging "a high contrast imager designed for single cell imaging, identification and clonal outgrowth characterization", the Cell Metric®, from Solentim [46]. It captures for each well the whole well's domain in high resolution and at pre-defined time points. For more technical aspects on the machine, we refer to its datasheet [47]. For each well, an image of size 1548×1548 pixels is recorded on this high throughput machine using phase contrast imaging. No more than 15 of the 24 columns per well plate are used because of the experimental design. In each of the wells only one initial cell is entered in liquid media. Here, Dulbecco's Modified Eagle's Medium (Sigma) supplemented with 10% Foetal Calf Serum (Gibco) and 1% GlutaMax (ThermoFisher) is used. For more information on the components in the media, we refer to a website of an example supplier [78]. Loading single cells into the plate is a slow process, so the biologists load the cells in certain batches to limit the amount of time for the cells being out of a temperature and CO_2 controlled incubator. If biologists use the full plate, it will be likely that the cells entered first are influenced or even suffering from being exposed for a longer preparation period. For this reason, only the first 15 columns will

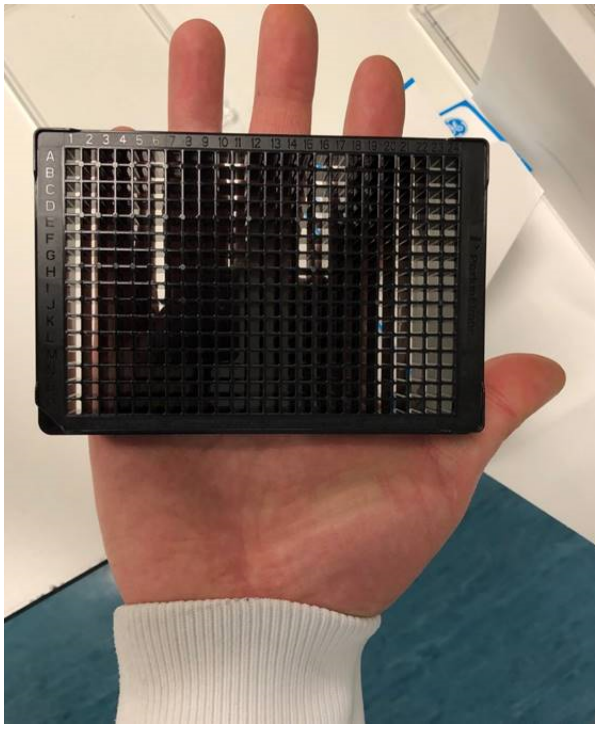


Figure 2.1: An example of a well plate shown on a hand serving as a reference size [81].

be taken into consideration whereas the other columns are left blank.
The cells used in this experiment were originally extracted from a lung cancer tumor. They have since then been in continuous in vitro cell culture for many years. Although all initial cells come from the same tumor tissue, one cannot assume that they are all genetically identical. Tumour cells themselves are always different from each other [27, 73]. This so called cell heterogeneity makes the development of treatments especially challenging with different mutations present across different cells, and subtle alterations in expression pathways and activity.
In our data set, the experimental setup, e.g., the liquid media or the genetic treatment of the initial cell, in three consecutive columns is prepared with the same biomarker. This is achieved by preparing the cells firstly on a transfection plate where a sample of cells is exposed and manipulated by one specific strategy corresponding to a distinct biomarker. Next, single cells are selected and sorted column-wise into the target well plates. This means that no more than 3×16 initial cells receive the same genetic manipulations. As the cells are not counted before the preparation stage, it can happen that not all three columns are completely filled with an initial cell in every well. Moreover, it can happen that no colony formation is observed because the initial cell does not survive the preparation stage or it dies soon after it is put into the target well. Those empty wells will be neglected in the later analysis. The replicates of one treatment are essential to get meaningful statistical output. Furthermore, those replicates are important to account for statistical outliers because of the differences and unique nature of tumor cells themselves. In other words, only a single particularly significant occurrence of a cell colony per three-column well group for one biomarker cannot be generalized to be true for all cells expressing this biomarker. Taking this into consideration is important to account for the individuality and diversity of cells within one tumor [27, 73].
During the time-lapse imaging, the initial cell either dies (*apoptotic cell*) or starts to replicate (*mitotic*

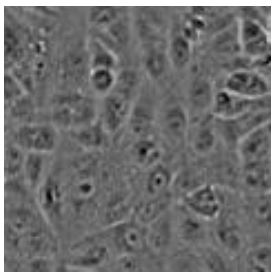
cell). When the new cells perform mitosis recurrently, a whole colony starts to grow. This is the starting point of our joint work between applied mathematicians and AstraZeneca biologists. In the next subsection, we motivate further this interdisciplinary collaboration.

2.2 Motivation for cell colony investigation

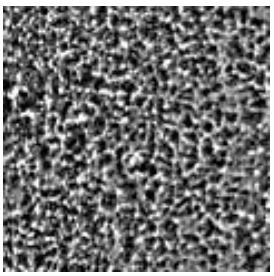
Cancer research is indeed a highly relevant field of research since it "is a major public health problem worldwide" [69]. For example "[it] is the second leading cause of death in the United States" as stated in the Cancer statistics, 2020 by Siegel et al. [69]. This stresses the severeness of the disease with potentially fatal consequences. Nevertheless, the behavior and the spreading of tumor cells is not yet overall well understood until today which highlights the high topicality of this research. To develop new drugs and better treatments, it is essential to analyze the effects on the cell level in an early stage research. For this reason, we develop a novel approach to analyze the growth process of a cell colony with a software solution to extract spreading properties. We emphasize that the main interest is not single cell analysis but rather measuring effects on the colony level. With this interdisciplinary project between AstraZeneca scientists and applied mathematicians, we want to make our contribution to foster a better understanding of those cells' spreading behavior and the formation process of such a cell colony.
In the given experimental setting, it is crucial to estimate and compare the effect of the various treatments and manipulations on the colony spreading behavior. As we are dealing with cancer cells here, the aim is to identify biomarkers which correlate to wells in which we observe a slowing down phenomenon in the colony spreading or ultimately stopping the spreading process. Even the observation of an increasing spreading speed related to certain treatments can reveal new insights on the mechanisms of the cell colony. Consequently, we are in general not interested in wells where no colony forms at all but rather in those where we can extract new information from the growth process of a cell population to gain more insights about the spreading of cancer cells by considering their exposure to different treatments.

At the beginning of this NoMADS cooperation on cell colony development, the standard procedure was that AstraZeneca biologists investigated the phase contrast images manually and chose relevant wells and their time series by visual inspections. This manual detection and investigation of growing colonies is influenced by subjective assessments and difficult to replicate. An automated way to process the data set, filtering out relevant wells with growing colonies, would already be an improvement over the time-consuming and cumbersome manual assessments of growing cell populations.
In our project, we aim for a software solution using image analysis, mathematical modeling and optimization to support AstraZeneca biologists in the future in their evaluation process. Our main goal is to extract new spreading properties related to the growing process of a cell population. We want to capture the *area* covered by the cell population as well as *temporal information* on the development of the cell colony. We stress that we are indeed interested in an approximation of the colony area and do not focus on single cell detection or single cell tracking.

To gain a better understanding of the temporal development of a cell population, we use mathematical modeling to explore properties like spreading speeds and spreading directions, cf. Chapter 4. At the same time we measure the growth of the colony by detecting the area covered by cells. Moreover, we want to separate different parts within the colony based on varying texture appearances. One of this subcolony regions represents cells with "normal" appearance for which we assume that they can perform mitosis to contribute to even more colony growth. The second subcolony attracts attention due to a remarkable change in texture in the images. For this colony area different interpretations are valid. The first interpretation is that this could reveal areas of apoptotic cells and cell debris where no more colony growth is to be expected. On the other hand, the striking texture changes could also reflect regions of the colony where cells are very closely packed, lying on top of each other and, thereby, bringing forward this distinct texture where no single cell segmentation is possible anymore for the given image quality. As the exact biological background is not yet fully understood, we call these regions with striking texture "abnormal" colony parts.
In Figure 2.2, we show an example patch for a colony region considered to be "normal" (cf. Figure 2.2a) and another cropped image representing an area within an "abnormal" colony region (cf. Figure 2.2b). While in the first image patch single cells are perceivable, the cryptic texture in the second image conceals the present cell population significantly. In Section 3.1, we introduce the given image data more thoroughly and go into more details considering the appearing texture changes.



(a) A patch showing cells in a"normal" colony region.



(b) A patch representing an "abnormal" colony area.

Figure 2.2: Example image patches for the two different subcolonies.

In Figure 2.3 a sketch is used to visualize a growing cell population in a circular well domain over time. This draft reflects the initial cell at the beginning and a growing accumulation of cells over time. In the last well, single cells are no longer distinguishable in the center part of the cell population; representing a similar texture change we observe in the real data set.

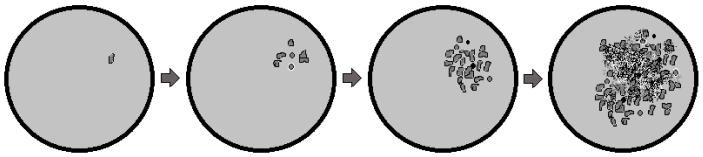


Figure 2.3: A sketch of a growing cell colony.

A closer evaluation of this second subcolony with the significant texture change requires additional imaging. With a fluorescence tag or a stain, individual cellular compartments such as the nucleus or

the cytoplasm could be labeled and made visible in a new color channel. Actually, the "color image" is also a grayscale image which we consider to reflect a certain color channel. For example, we visualize a *green fluorescent protein* (abbreviation: *GFP*) [80] in a green color channel and a red fluorescent protein (abbreviation: *RFP*) in a red color channel whereas a DAPI staining is usually shown in a blue channel. The choice of a specific color channel is based on specific filter and presentation settings. The naming and the usual color channel representation are based on the emission intensity and spectra depending on the emission and excitation wavelength. Without going deeper into the details of fluorescence microscopy, we refer the interested reader to [30] for a more thorough introduction. To stress the importance of fluorescent proteins in cell analysis, we point out that Osamu Shimomura, Martin Chalfie and Roger Y. Tsien were awarded with the Nobel Prize in Chemistry "for the discovery and development of the green fluorescent protein, GFP" in 2008 [57].
As the investigated cells already have a histone-red-fluorescent protein tag, the preferred way is to record a time course consisting directly of phase contrast images and the red color channel. This allows to observe the development of labeled nuclei over time. If the color data allows nuclei segmentation, one can derive the exact number of cells within the colony for each time point. In contrast to this, a DNA marker based on a DAPI staining, would support the estimation of total cells in the culture only for one specific time point as the staining is applied to dead cells after fixing them and permeabilising the cells' membranes. Consequently, this staining impedes an estimation of the total cell number over a certain time course. For more information on the DAPI staining, we refer to the article by Kapuscinski [39]. Instead of the DAPI end point staining, one could also consider live-cell-nuclei stains based on Hoechst fluorescent dyes [13]. This is an alternative for live cell staining. However, the Hoechst staining requires a dye to intercalate with the cells' DNA and is toxic to cells in the long term, i.e., it makes the cells ultimately sick. In this respect, we do not suggest this staining for time course analysis either.
In color channel images, one could aim for single cell counting by counting detected cell nuclei after segmenting them. With a specific number of cells counted in one image, one could estimate the cell density within the colony based on the previously segmented colony area. For now, it is unclear if segmenting individual nuclei is achievable — especially in the second colony area where no single cells are perceivable. Without having the color channel accessible, we cannot predict to get reliable nuclei segmentation results. However, if the image quality allows nucleus counting after segmentation, this will facilitate the investigation of the hypothesis of increasing cell densities in areas of those remarkable texture changes.
The deeper analysis of what exactly is happening in the second subcolony with the striking texture is out of the scope of this thesis. Here, we aim for a model capturing two different colony subregions. Further on, we will only differentiate between colony parts of the first subcolony and the second subcolony. In the course of this thesis, we call the first one where single cells are still visible and identifiable **normal** colony area. The second part of the colony where the new texture in the images does not reveal single cell boundaries is described as the **abnormal** subcolony in the further course.

In the next chapter on preliminary image processing, we focus on the given data set by highlighting the special texture observable in the images. Even more, we introduce texture features which we will use in the later course to distinguish the different colony regions.

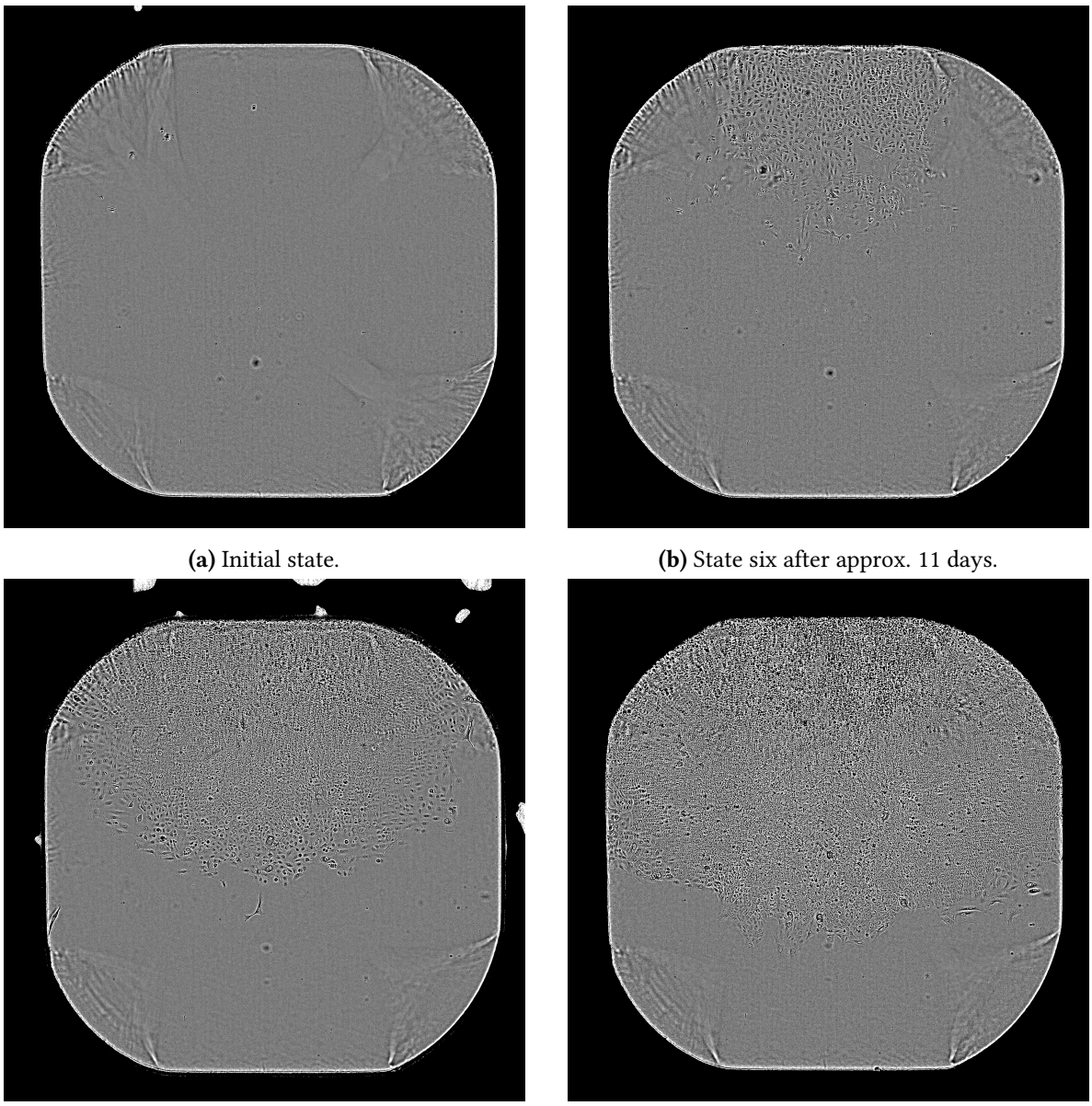
3

Preliminary image processing & feature images

In this chapter, we investigate the present image data in more details and focus on texture properties. Based on these, we generate feature images which capture the growth process of developing cell colonies.
To get a first idea of the spreading process of the cell colonies, we apply classical image processing approaches. With the help of segmentation and registration methods, we are able to calculate an estimate for the colony area and to track the moving colony fronts over time (cf. Section 3.2). However, before we delve into the details, we present in Section 3.1 an example of a growing colony to illustrate the spreading process and also the image quality we are dealing with. We highlight the remarkable texture within a cell population observable in the images and already introduce the relevant texture features we use in the later course of this work when approaching the colony development from a mathematical perspective. For this purpose, we already present the mathematical context of the feature images and theoretical concepts related to noise effects influencing the feature data in Section 3.3, more precisely in Section 3.3.1. In our setting, we exploit local texture properties to calculate feature images. They are introduced in Section 3.3.2 in particular. In Section 3.3.3, we conclude with alternative texture descriptors which could potentially be used to derive feature images as well.

3.1 Introduction to the image data set

In this section, we focus on the given image data and present an example of a growing colony observed in well B4 on plate 1 at selected time frames. In the further course of this subsection, we highlight texture changes when zooming in to small subdomains in the imaging field of view.
The growing colonies are captured in time-lapse phase contrast images and for visualization aspects, we apply a limited color range in the preview of the selected time points in Figure 3.1. By this, we achieve a contrast enhancement within the well's domain and the colony spreading is better perceivable by eye. In the first state in Figure 3.1a, it is very difficult even for experienced biologists to locate the initial cell. Due to the imaging quality, single cell detection is not straightforward and in this example there are several spots in the left upper quadrant of the well which either could be the initial cell or only mark some artificial structures. Anyway, there are even more spots visible which



(a) Initial state.

(b) State six after approx. 11 days.

(c) State seven after approx. fourteen days.

(d) State eight after approx. eighteen days.

Figure 3.1: Colony development in well B4 of plate 1 (limited color range for contrast enhancement).

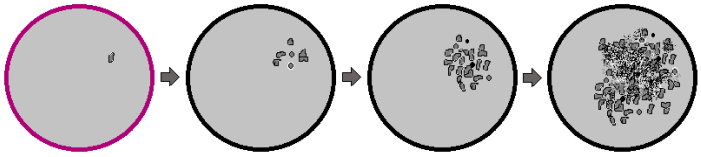
are slightly out of focus and, consequently, appear a bit blurry (cf. right lower quadrant or middle of lower half of the well). Since we are interested in the total colony domain and not in detecting single cells, those drawbacks because of the image quality are no obstacles to detect global colony estimates in the further course.

In the given images, we are facing even more artifact structures than only out of focus objects. We can observe reflections and wave-like structures close to the well's boundary. Those might arise from movements in the liquid media when positioning the well for imaging and the reflections might occur due to lightning effects on the well's boundaries during the imaging process. In some images there are also other reflections present outside of the well, cf. Figure 3.1a and Figure 3.1c. The strategy we are following to deal with those artefacts is to crop them out of the images by first segmenting the well's domain and then focusing only on the well's domain itself. By using a slightly shrinked

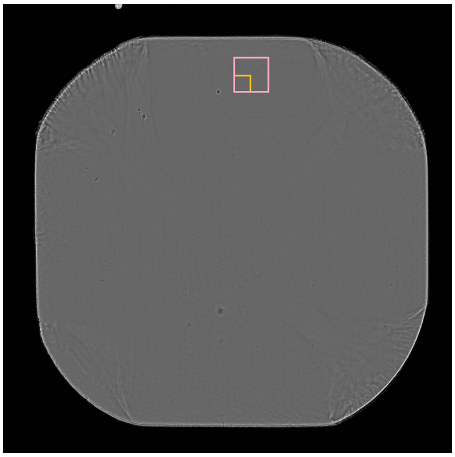
version of the well's domain as a cropping frame, we also get rid of the brighter shining effects close to the well's boundaries.

To be more precise, we only determine for one reference well its domain by applying Chan Vese segmentation [14]. This results in a reference well domain which we use to register all other wells' time frames first and then cropping them to a smaller field of view within the wells' areas with a shrinked version of this reference domain. The registration step is necessary when using one reference well domain since the wells are not positioned completely identical during the imaging process so that slight shifting effects are observable when comparing several images. We apply a rigid registration approach, i.e., only allowing "rigid" transformations as rotations or translations, via the MATLAB function *imregtform* [53].

To conclude the short excursion on artefacts in the images, it is also important to mention that in rare cases we observed bubbles in the media or out of focus structures revealing particles or small hairs on top of the wells' lids, which might influence the colony segmentation later on. Since we are aiming for a global software solution for the colony detection task, we do not focus on removing artefacts present in only a few images of the data set. We rather accept some mis-segmentations due to rare artifact structures in the images.

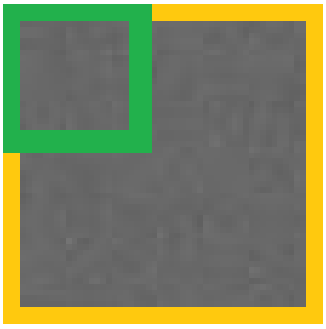


(a) Colony sketch marking the initial state.

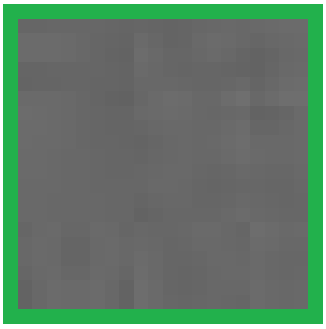


(b) Well B4 on plate 1 at the initial state.

(c) Zoom in to 120 × 120 pixels.



(d) Zoom in to 50 × 50 pixels.

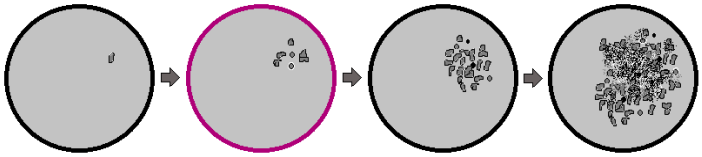


(e) Zoom in to 20 × 20 pixels.

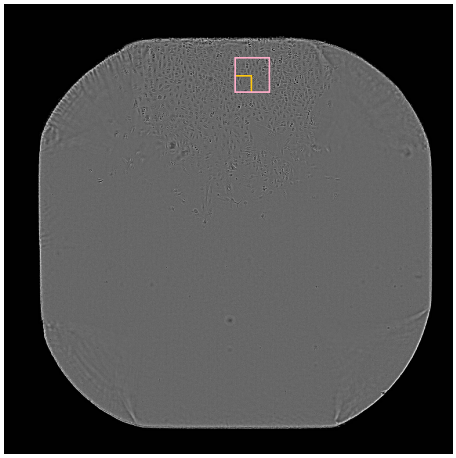
Figure 3.2: Small zoomed in square domains in initial time stamp of well B4 of plate 1.

In Figures 3.2 to 3.5, we focus on the growing colony at the selected time points individually. After a sketch of the current development state, we repeat again the microscopy images with small artificial

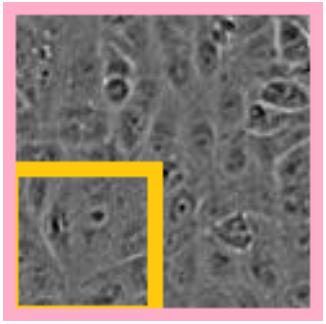
squares in pink and yellow added. We zoom into these small windows and add another close-up within the yellow frame highlighted with a green frame. In contrast to the images shown in Figure 3.1, we do not apply a limit color range here. Instead we use the original color scale which impedes the manual colony detection. However, we use the small square domains to zoom in to those small regions of interest. By this, we facilitate the inspection of texture changes based on the original color scale in those subdomains over time while the colony is developing.
Coming back to the *initial state* when the image of well B4 in Figure 3.1a is recorded, we want to stress that there is not yet a colony developed and the little squares in pink and yellow focus on a background region of the well. In Figure 3.2, we zoom into these little squares for the first time point of the time series. We observe that the grayscale values do not differ a lot in this background patch although it is not one constant grayscale value. But compared to the other time points presented in Figures 3.3 to 3.5, this background region is a smooth gray tone without a huge range of grayscale values.



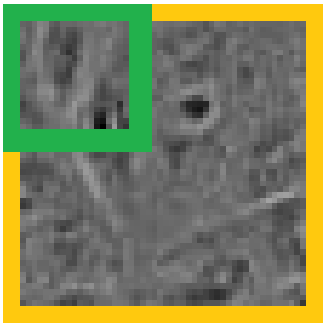
(a) Colony sketch marking the sixth state.



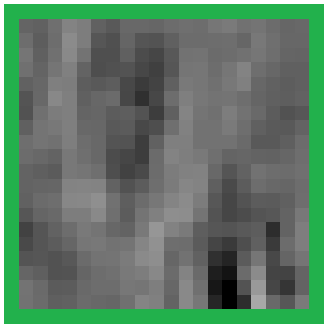
(b) Well B4 on plate 1 at the sixth state.



(c) Zoom in to 120 × 120 pixels.



(d) Zoom in to 50 × 50 pixels.

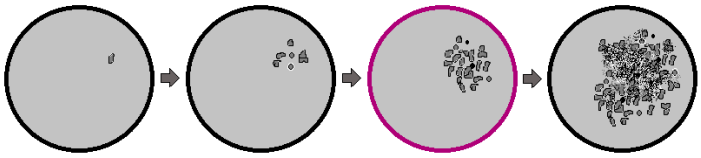


(e) Zoom in to 20 × 20 pixels.

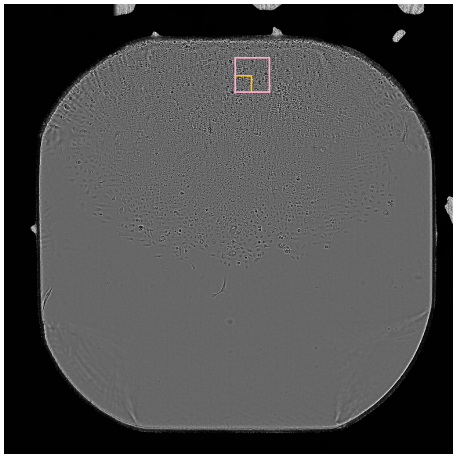
Figure 3.3: Small zoomed in square domains of well B4 of plate 1 after approx. 11 days (sixth state).

In the time frames 6, 7 and 8 depicted in Figures 3.1b to 3.1d, the small square windows focus on regions where cells are located and where the cell colony emerges. We focus on small zoomed in windows in Figures 3.3 to 3.5. While in Figure 3.3c single cell features are perceivable with brighter structures marking cell boundaries or dark spots highlighting nuclei, still not every single cell is distinctively identifiable. Moreover, we notice that cell shapes differ. While some cells are more elongated or stretched, cells that are about to perform mitosis are rounding up and showing the typical

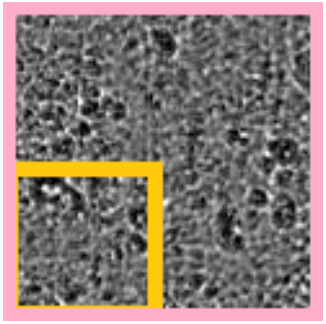
bright halo effect around them [32]. It is worth mentioning that in this early stage of an emerging cell colony, single cells can be identified and the colony region is not yet overcrowded.
In the closest zoomed in version in Figure 3.3d, we already observe a higher variance in the gray values for this sixth time frame of the image series of well B4, compared to the previously described initial state. Due to brighter pixels located close to cell boundaries and darker nuclei pixels this gray value distribution is significantly different from the smooth gray tones in a background patch shown in the previous Figure 3.2d.
For the following recorded time frame number 7, we observe a more crowded population in Figure 3.4. In the pink zooming window in Figure 3.4c, it looks like the focused well area is densely occupied by cells. Still the brighter pixels are hints for cell boundaries. Compared to the previous state in Figure 3.3c, the single cells look more squeezed as there is not enough space for them to stretch out. Even more, we want to highlight that we still observe darker spots pointing to cell nuclei. Still, we already see more blurry grayish regions where cell boundaries are no longer detectable by visual inspection. We point out that those areas resemble background regions on a very fine scale and might impede differing between background, i.e., empty regions, and colony regions where cell boundaries are blurred.



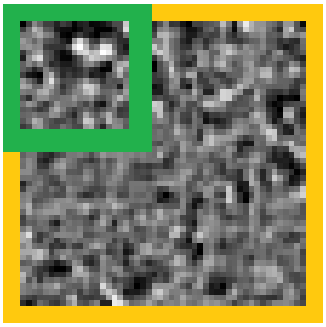
(a) Colony sketch marking the seventh state.



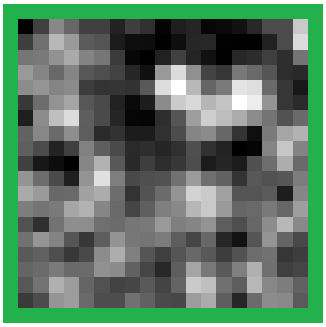
(b) Well B4 on plate 1 at the seventh state.



(c) Zoom in to 120 × 120 pixels.



(d) Zoom in to 50 × 50 pixels.

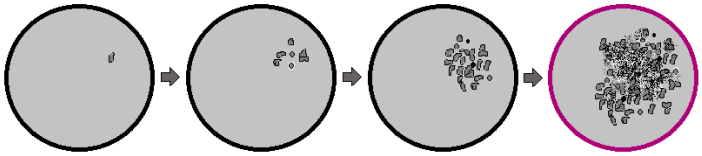


(e) Zoom in to 20 × 20 pixels.

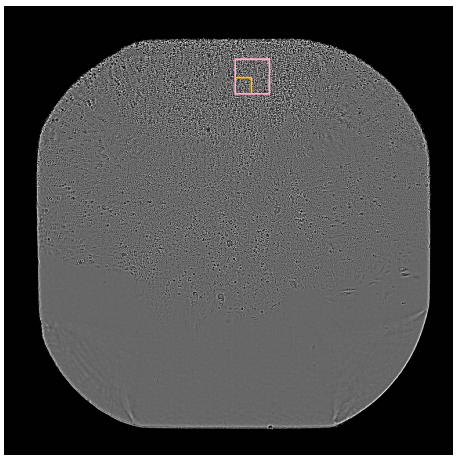
Figure 3.4: Small zoomed in square domains of well B4 of plate 1 after approx. fourteen days (seventh state).

For the final state number 8 presented in Figure 3.5, we stress that the variance of gray values is even larger than compared to the previous state in Figure 3.4. We observe more dark pixels and also more

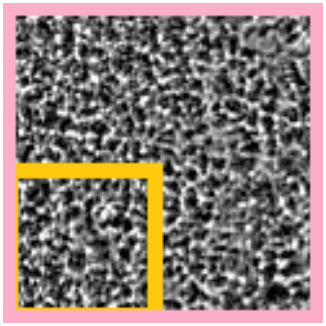
bright regions. It is not straightforward to detect individual cells here. Especially, the very dark and very bright pixels are striking. The dark ones could either be related to apoptotic cells or cell debris or they could reveal in combination with the very bright pixels cells lying on top of each other. If the cell colony is very crowded such that cells start to lie on top of each other, some cells will move out of focus. Consequently, dark and bright pixels could possibly reveal shadowing effects due to those out of focus objects as well.



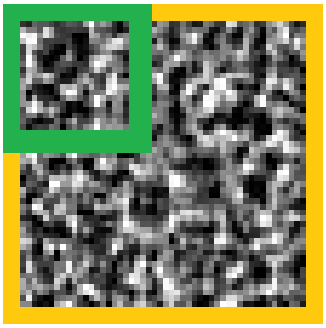
(a) Colony sketch marking the final eighth state.



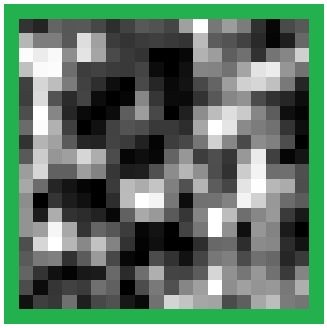
(b) Well B4 on plate 1 at the eighth state.



(c) Zoom in to 120 × 120 pixels.



(d) Zoom in to 50 × 50 pixels.



(e) Zoom in to 20 × 20 pixels.

Figure 3.5: Small zoomed in square domains of well B4 of plate 1 after approx. eighteen days (final state).

Once more, we want to stress that the biological interpretation of this texture is not intuitively straightforward. To get an idea what exactly is happening in those dense regions, additional imaging is required. For example, another color channel marking single cells to allow counting them after segmentation of this new signal would reveal new information that helps to identify the biological meaning of those areas. In the scope of this thesis we do not go deeper into details of this analysis. We rather concentrate on two different colony areas in the further course — one for the region where single cells could be visible (“normal” cell colony) and the second one where we observe the significant texture change such that single cells are no longer distinguishable (“abnormal” cell colony).

In the presented small patches we can observe the increasing variance of grayscale values over time. In Figure 3.6, we show histograms of the gray values present in the depicted 50×50 pixels patches for the time points 1, 6, 7 and 8 again. Those small subpatches are extracted from similar

colony regions as the ones given before. The histograms highlight the increasing variance of occurring grayscale values in those small square subdomains.

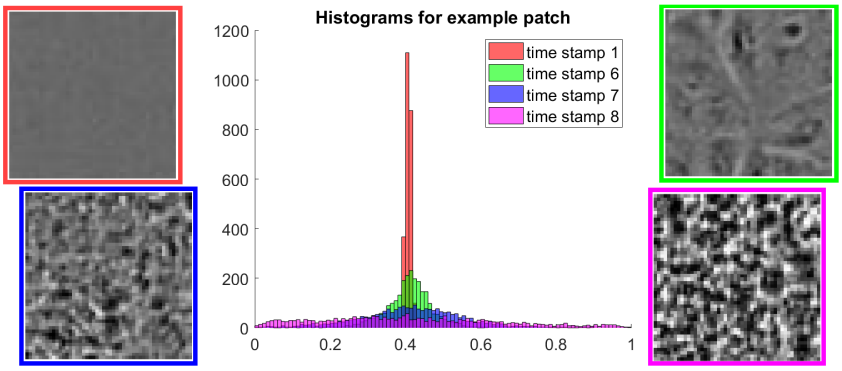


Figure 3.6: Gray value distributions for small subpatches within well B4 at time frame 1, 6, 7 and 8 in one histogram.

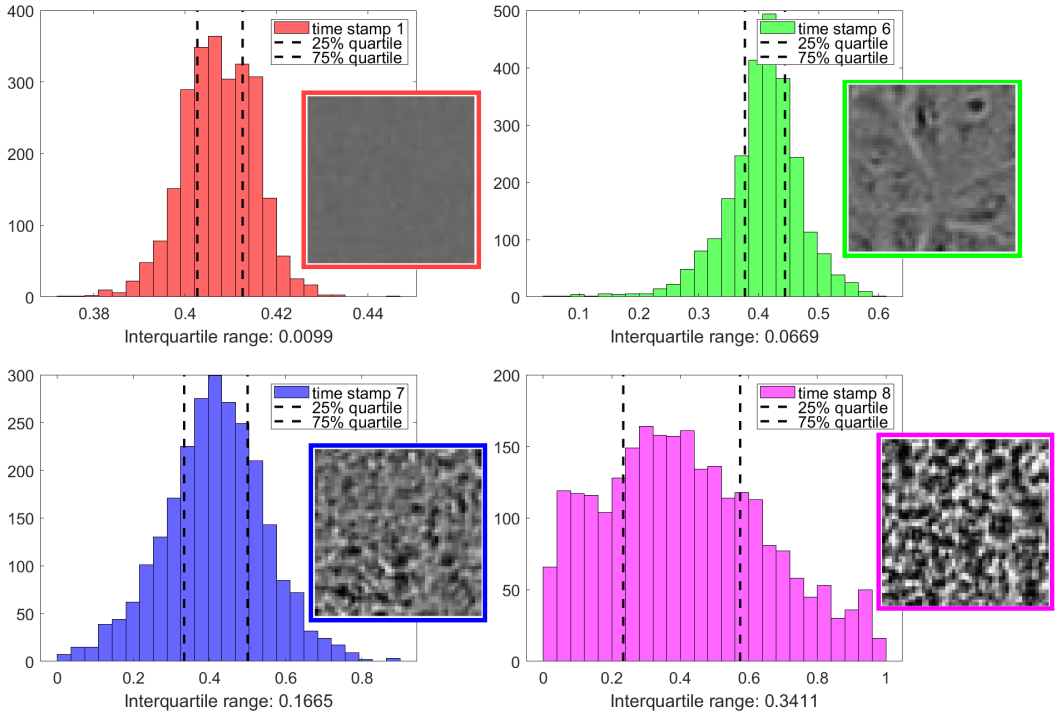


Figure 3.7: Gray value distributions for small subpatches within well B4 at time frame 1, 6, 7 and 8 in separate histograms.

Instead of overlaying the histograms for the different subpatches, we present in Figure 3.7 the histograms separately. Here, the interquartile ranges of the histograms are marked with dashed lines. We observe that the limits of the x-axis for the different histograms are changing significantly and this is true for the interquartile ranges as well. In the background regions depicted in the first patch with the red frame, we get an interquartile range of smaller than 0.01 whereas this property increases when entering a colony region (green frame) up to a value around 0.07. In the more crowded patches of cell colonies at frame 7 and 8, we get interquartile ranges of about 0.17 and even 0.34. This re-

flects the increasing presence of very dark and very bright pixels whereas in the first patches the grayscale values are concentrated closely around a value of 0.41 which relates to the smooth grayish background.

In the final Section 3.3 of this chapter those gray values play an even bigger role when we describe the selected feature set that we use for the upcoming colony analysis. Before we focus more on those features, we start with a short detour: With the help of segmentation results of the colony area, we motivate in Section 3.2 the further spreading analysis.

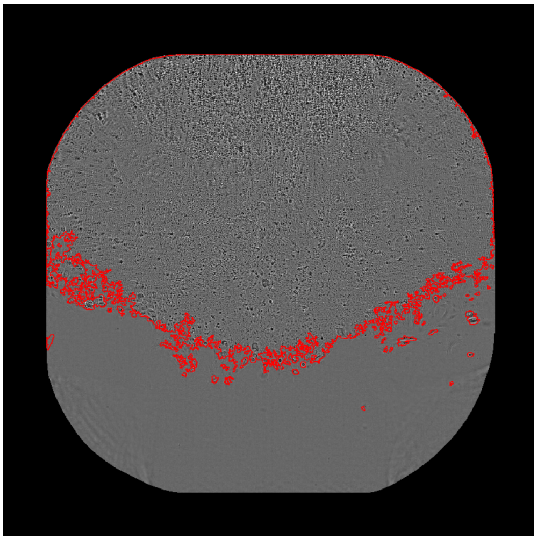
3.2 Initial colony segmentation and moving fronts

One common task in image processing when dealing with microscopy data is *segmentation*. In the process of image segmentation, the image gets split into several distinct parts — or *segments* — consisting of neighbouring pixel groups which share a common property. When dealing with grayscale images dividing the domain into foreground and background, for example, is one classical segmentation task and can be accomplished by grayscale thresholding. Each pixel gets assigned to a certain pixel group depending on its gray value and if it is below or above the selected threshold value. Another well known approach is k-means clustering which is used to generate a segmentation based on the gray or color values in an image. Based on the value of each pixel, the image is partitioned into k different segments which not necessary need to be connected. With the k-means approach, each pixel is assigned to that segment with the nearest mean value of the occurring gray values in that partition. In other words, we can say each pixel is *labeled* with one of k labels to identify the cluster it belongs to. In this light, we stress that the segmentation task can also be interpreted as a classification task since every pixel is labeled — or *classified*. All pixels with the same label are then considered to belong to the same object or similar structures. Labeling can also be achieved based on a different property. A famous example is edge detection where lines separating different objects are identified based on their gradient values.
Without going into further details on possible segmentation approaches, we refer the interested reader to the work of Kristian Bredies and Dirk Lorenz. In their book on Mathematical Image Processing, various techniques for image analysis are introduced [11]. Furthermore, we recommend the book on Pattern Classification of Duda et al. [19] for more information on segmentation and clustering approaches as for example the k-means algorithm.

An intuitive way to get an idea of the spreading colony and its moving front is based on segmenting the colony area. By applying a combination of spatio-temporal gradients, thresholding and smoothing with morphological operations, an estimate for the colony area can be identified. Without describing the implemented task-specific segmentation approach in full detail, we just briefly motivate the different substeps. Although the grayscale images are of low contrast, cell boundaries are observable in the initial states. As shown in Section 3.1, cell boundaries are prominent due to a brighter halo effect or a darker shadow effect compared to the background or the cell's body. Both effects result in spatial gradient information. The same holds true for more occupied colony regions at later time

stamps because of the significant texture changes — although single cell boundaries are rarely visible themselves.

Adding temporal gradient information based on frame differences, the growing colony region is highlighted. By thresholding those gradients, we get a rough estimate of the colony area. With the help of morphological operations (*dilation* and *erosion*, *image opening* and *closing*, cf. [17]), we fill gaps and smooth the colony area. As we are not interested in single cell tracking, it is only a slight trade off that we miss some single cells in this approach. In total we can assume that we detect about 90% of the colony area with this approach. In Figure 3.8 the segmented colony area for well B4 of the first plate at the final state after approximately 18 days is shown exemplarily. We observe in Figure 3.8a that indeed not all single cells are segmented, especially at the right lower colony border we miss a few cells. We refer to Figure 3.1d for a contrast enhanced preview of this colony state. We point out that the segmentation map in Figure 3.8a still serves well as an initial global estimate for the colony region and its binary map is shown in Figure 3.8b.



(a) The boundary of a cell colony (red) on cropped original data.

(b) A binary mask of the colony area with the well's border in purple.

Figure 3.8: Segmentation result for the cell colony in well B4 of plate 1 at the final state.

Based on those binary maps for the colony regions, we track the moving colony fronts. By applying again morphological operations, we fill up the front line and remove some detached objects (or artifacts) in the mask that are not linked to the main colony area and too small. With those smeared front lines we generate a map displaying the growing colony region, cf. Figure 3.9. The colored front regions show two main aspects. On the one hand they represent the gain of colony area between two consecutive time stamps. On the other hand, the time stamp when the new front line is observed in the microscopy data is depicted in the color of the region area. This map already reveals in a very condensed way the colony growth over time for biological assessments. To facilitate the evaluation process, those moving wave fronts are plotted in a grid structure resembling the original plate map (cf. Figure 2.1) as shown in Figure 3.10. This representation helps to identify at a single glance in which wells growing colonies are present. Moreover, it allows a comparison between different biomarkers: As always three neighboring columns relate to the same biomarker, one can observe that for the

biomarker of columns seven to nine most colonies are present on the first plate. With the color-coded time stamp information, the growing colonies are also comparable with respect to their initial growth starting time point. We remark that the second and third columns are left blank in this experiment due to too few initial cells for this biomarker. Those empty occurences were already described in Section 2.1. Columns 13 to 15 are only filled on the final plate to avoid starting a sixth plate for only one biomarker group. For this reasons, the last three columns are empty as well. Finally, we highlight that well B4 shown in a magnified version in Figure 3.9 is present in Figure 3.10 in the second row and forth column according to the alpha-numerical numbering of the wells per plate.

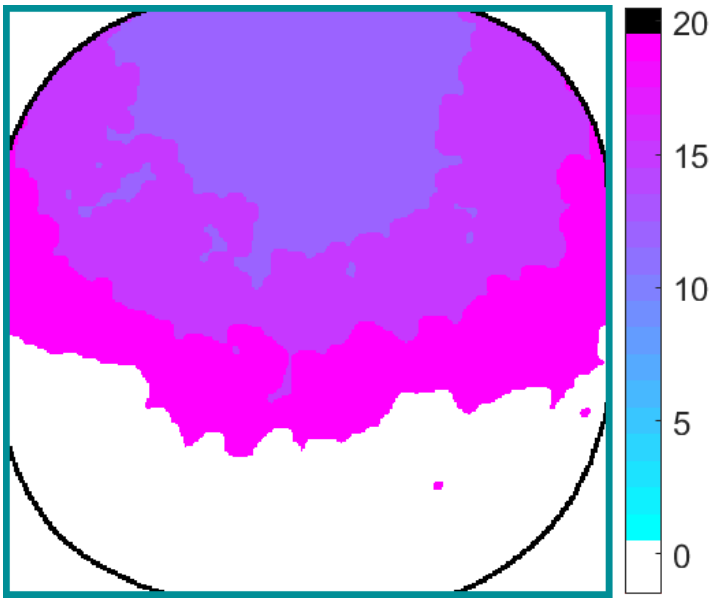


Figure 3.9: Map of growing cell colony in well B4 of the first plate with color-coded observation time stamps (in days) for the moving front lines.

The colony growth observable in those maps arouses our interest in the spreading process of cell colonies. Even more, the question comes up if we can derive spreading properties from the data without performing a colony segmentation first. In the following chapter, we use mathematical modeling to describe the spreading process.

In the further course of this thesis, we aim for an approach considering a spreading model and the given data at the same time to extract properties describing the colony growth directly from the image data via numerical optimization techniques. Therefore, we derive feature images from the microscopy data which will be of particular importance in the later optimization and which are introduced in the next subsection in more detail.

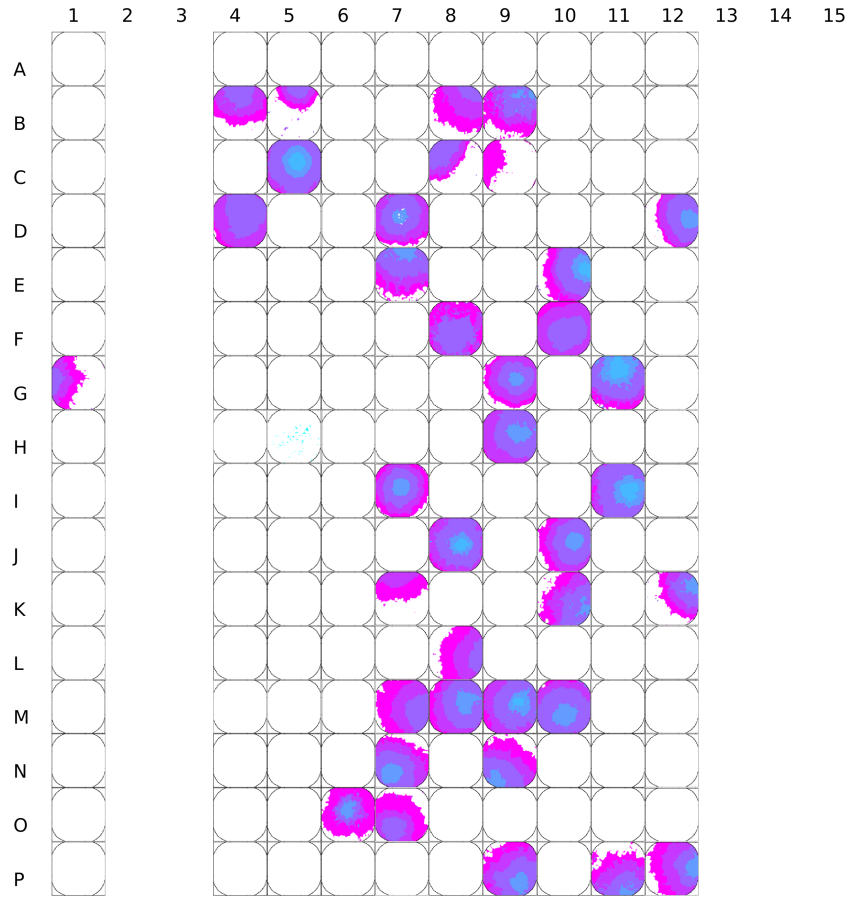


Figure 3.10: Plate map of moving colony front lines for the first plate.

3.3 Feature images for colony analysis

Based on the given microscopy data capturing the development of growing cell colonies, we introduce feature images. The feature images are implemented in the optimization process later to automatically derive spreading properties correlated with a mathematical model based on the original data. In this way, we want to extract spreading information on the growing cell population observable in the microscopy images.
The main goal of the first part of this section is deriving estimates for the probabilities when we consider the feature images to be corrupted by Gaussian noise. After that we show concrete examples for feature images based on local texture information and expand on alternative feature descriptors.

3.3.1 Theoretical background of feature images

In this section, we focus on the general setting and the theoretical background of the used feature images. They are based on the given imaging data and, additionally, we consider them to be corrupted

by Gaussian noise. We give a series of definitions and propositions related to the applied concepts. The main aim of this section is to develop probability measures related to the disturbed feature images plus certain estimates that we make use of in the later course of this work when dealing with existence and convergence results related to the optimization problem (cf. Section 5.4).
We start with a basic definition of a feature image I_1 in an abstract setting after introducing the spatio-temporal domain it is living on. We complement this definition with a uniform probability distribution that we relate with this space in the further course.

Definition 3.1 (Semi-discrete spatio-temporal domain Ω_T with a uniform probability distribution)
Based on a rectangular domain $\Omega = [0, L] \times [0, W]$ in $\mathbb{R}^2$ and a spatio-temporal cylinder $\Omega \times [0, T]$ with $T > 0$, we define the semi-discrete spatio-temporal domain by

$$\Omega_T = \Omega \times \{t_1, \dots, t_{n_T}\}$$

where we consider discrete time points $\{t_1, \dots, t_{n_T}\} \subset [0, T]$, i.e., we already assume a certain discretization in the temporal domain. By L and W, we denote the length and width of the rectangular spatial domain while the end time point of the time interval is denoted by T. The total number of discrete time points is denoted by n_T.
We consider a uniform probability distribution related to the spatio-temporal domain Ω_T defined as

$$P_{\Omega_T}(A) = \sum_{i=1}^{n_T} \int\limits_{A \cap (\Omega \times t_i)} \frac{1}{\kappa(\Omega_T)} \, \mathrm{d}\kappa(\boldsymbol{x}, t_i) = \frac{1}{n_T \, \kappa(\Omega)} \sum_{i=1}^{n_T} \kappa(A \cap (\Omega \times t_i))$$

for $A \in \mathcal{B}(\Omega_T)$. The multi-dimensional Lebesgue measure related to $\Omega \subset \mathbb{R}^2$ is denoted by κ here.

We remark that we use

$$\kappa(\Omega_T) = n_T \kappa(\Omega) \tag{3.1}$$

without specifying a separate measure for the semi-discrete spatio-temporal domain. In this sense, we use

$$\int\limits_{\Omega_T} f(\boldsymbol{x}, t) \; \mathrm{d}(\boldsymbol{x}, t) = \sum_{i=1}^{n_T} \int\limits_{\Omega} f(\boldsymbol{x}, t_i) \; \mathrm{d}\boldsymbol{x}$$

for the integration of a function $f : \Omega_T \to \mathbb{R}$.
After defining our spatio-temporal domain Ω_T, we introduce the feature image and the related probability measure.

Definition 3.2 (Feature image I_1)
The feature image I_1 is given as a function living on the image domain and mapping to a feature space $\mathcal{F} = \mathbb{R}^n$

$$I_1 : \Omega_T \to \mathcal{F}. \tag{3.2}$$

The probability that I_1 maps values from Ω_T to a measurable $F \in \mathcal{B}(\mathcal{F})$ is given as the pushforward measure with

$$P_{\mathcal{F}}(F) := I_{1\#}P_{\Omega_T}(F) = P_{\Omega_T}\left(I_1^{-1}(F)\right)$$

with P_{Ω_T} being the probability measure related to the measurable space $(\Omega_T, \mathcal{B}(\Omega_T))$ (cf. Definition 3.1).

This first definition of the feature image I_1 lives in a very general setting. At this point, we only say that the feature image maps into *some* space $\mathcal{F} = \mathbb{R}^n$ without specifying the $n \in \mathbb{N}$ here while we consider a three dimensional feature space later on. The general approach allows us to formulate the optimization problem later based on an arbitrary feature image I_1. Of course, although we grant some flexibility here, we stress that the feature image still needs to correlate to the original data.
Before giving two examples for feature images, we expand briefly on the nature of the original microscopy image data. We assume that the original images consist of only one channel, consequently mapping to $\mathbb{R}$, and are piecewise continuous in a perfect setting. Due to inaccuracies in the image acquisition process, we observe random intensity oscillations in the microscopy images. These disturbances of the original images are known as "noise". There is a whole field in image processing dealing with denoising approaches to approximate the underlying original images by smoothing out noise effects while still preserving meaningful structures in the image as for example edges [11, 12]. Without analyzing those noise effects prominent in the original data in more detail, we assume that the feature images will also be influenced by noise effects.
Moreover, we assume that the noise effect is independent of the location within the image domain and the disturbances in one location of the feature image are independent to the disturbances in another location. Additionally, we suppose that we are facing an additive noise model with zero mean in every location on the image domain. Lastly, we also presume that the noise level is signal independent, i.e., the corruption is independent from the underlying feature values. One commonly used approach to model such noise effects is to apply additive Gaussian noise with zero mean [12]. We summarize this effect of Gaussian noise in an definition for our *noise image*.

Definition 3.3 (Noise image I_N)
Let $\left(\Omega_0, \mathcal{E}, P^0\right)$ be a probability space. We denote the Gaussian noise disturbing the feature images with the real-valued multi-dimensional random variable

$$I_N : \Omega_0 \times \Omega_T \to \mathcal{F} = \mathbb{R}^n.$$

We also call this random variable *noise image* and consider the observation space $\mathbb{R}^n$ to be equipped with the related Borel σ–algebra $\mathcal{B}(\mathbb{R}^n)$. For every location $(\boldsymbol{x}, t) \in \Omega_T$, we consider the features of the noise image to be identically distributed such that

$$P^0\left(I_N(\cdot, (\boldsymbol{x}, t)) \in A\right) = \mathcal{N}\left(0, \sigma^2 \mathbf{1}_{n\times n}\right)(A) \qquad \forall A \in \mathcal{B}(\mathbb{R}^n)$$

holds for an arbitrary $\sigma > 0$. We take into consideration that the features are independent of each other with the multivariate Gauß distribution $\mathcal{N}\left(0, \sigma^2 \mathbf{1}_{n\times n}\right)$ for which the covariance matrix is

concentrated along the main diagonal. Here, the identity matrix in $\mathbb{R}^{n\times n}$ is denoted with $\mathbf{1}_{n\times n}$. Moreover, we assume that the family of random variables $(I_N(\cdot,(\boldsymbol{x},t)))_{(\boldsymbol{x},t)\in\Omega_T}$ are stochastically independent to take into account that the disturbances in one location do not correlate to another location.

The pushforward of the probability measure P^0 with respect to I_N coincides with the Gaussian normal distribution. In preparation of later proofs, we already express the effect of the additive Gaussian noise with zero mean and standard deviation of σ^2 on the related image histograms or, more precisely, on the related probability density functions.

Definition 3.4 (Probabilities related to the noise image)
For the probabilities related to the noise image given by

$$I_N(\cdot,(\boldsymbol{x},t)) \sim \mathcal{N}\left(0,\sigma^2\mathbf{1}_{n\times n}\right) \qquad \forall\,(\boldsymbol{x},t)\in\Omega_T$$

we introduce an abbreviated form with

$$\begin{aligned} P_N(F) &= \mathcal{N}\left(0,\sigma^2\mathbf{1}_{n\times n}\right)(F) \\ &= I_N(\cdot,(\boldsymbol{x},t))_{\#}P^0(F) \end{aligned}$$

for all $F\in\mathcal{B}(\boldsymbol{\mathcal{F}})$. The probability measure of the Gaussian normal distribution is absolutely continuous with respect to the Lebesgue measure and we denote the corresponding probability density function by p_N with

$$\begin{aligned} p_N &: \boldsymbol{\mathcal{F}} \to \mathbb{R}_+ \\ P_N(F) &= \int_F p_N(f)\ \mathrm{d}f \quad \forall F\in\mathcal{B}(\boldsymbol{\mathcal{F}}) \\ P_N(\boldsymbol{\mathcal{F}}) &= \int_{\boldsymbol{\mathcal{F}}} p_N(f)\ \mathrm{d}f = 1. \end{aligned}$$

Given this noise image, we formulate now our *disturbed feature image* I_1^{d} which consists basically of the original feature image I_1 with additive gaussian noise.

Definition 3.5 (Disturbed feature image I_1^{d})
With the above definitions, we consider I_1 to be the feature image and I_N to be the noise image. The *disturbed feature image* is then defined as

$$I_1^{\mathrm{d}} : \Omega_0\times\Omega_T\to\boldsymbol{\mathcal{F}}, \qquad I_1^{\mathrm{d}}(\omega,(\boldsymbol{x},t)) = I_1(\boldsymbol{x},t) + I_N(\omega,(\boldsymbol{x},t)). \tag{3.3}$$

The disturbed feature image is a real-valued multi-dimensional random variable as well.

To prepare the derivation of probabilities related to the disturbed feature image I_1, we show in the next step that the feature image I_1 and the noise image I_N are stochastically independent.

Proposition 3.6 (Stochastic independence of the feature image and the noise image)
We consider the probability space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T), P^*)$ with $P^* := P^0 \otimes P_{\Omega_T}$. Let I_1 be the feature image defined in Definition 3.2 and I_N be the noise image introduced in Definition 3.5.
For the extended feature image

$$\bar{I}_1 : \Omega_0 \times \Omega_T \to \boldsymbol{\mathcal{F}}, \qquad \bar{I}_1(\omega, (\boldsymbol{x}, t)) = I_1(\boldsymbol{x}, t)$$

it holds that it is stochastically independent to the noise image I_N.

Proof. We consider the distribution of the joined mapping $(\bar{I}_1, I_N) : \Omega_0 \times \Omega_T \to \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{F}}$. In this context, we examine $\bar{I}_1$, I_N and $(\bar{I}_1, I_N)$ as random variables and use the Theorem 2.21 in [40]. To show stochastic independence, we need to prove that

$$(\bar{I}_1, I_N)_{\#} P^*(E) = (\bar{I}_{1\#} P^*) \otimes (I_{N\#} P^*)(E)$$

for all measurable $E \subset \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{F}}$ holds.
We show the statement for $E := F_1 \times F_2$ with arbitrary $F_1, F_2 \in \mathcal{B}(\boldsymbol{\mathcal{F}})$ and split the proof in following three separate statements.

1. $(\bar{I}_1, I_N)_{\#} P^*(F_1 \times F_2) = P_{\boldsymbol{\mathcal{F}}}(F_1) P_N(F_2)$
2. $\bar{I}_{1\#} P^*(F_1) = P_{\boldsymbol{\mathcal{F}}}(F_1)$
3. $I_{N\#} P^*(F_2) = P_N(F_2)$

Statement 1:
To begin with, we state that for the set $Z := \bar{I}_1^{-1}(F_1) \cap I_N^{-1}(F_2) \subset \Omega_0 \times \Omega_T$ the following equivalence

$$(\omega, (\boldsymbol{x}, t)) \in Z \Leftrightarrow \bar{I}_1(\omega, (\boldsymbol{x}, t)) = I_1(\boldsymbol{x}, t) \in F_1 \wedge I_N(\omega, (\boldsymbol{x}, t)) \in F_2$$

holds and this implies

$$\mathbf{1}_Z(\omega, (\boldsymbol{x}, t)) = \mathbf{1}_{F_1}(I_1(\boldsymbol{x}, t)) \cdot \mathbf{1}_{F_2}(I_N(\omega, (\boldsymbol{x}, t))). \tag{3.4}$$

With this at hand, we derive

$$\begin{aligned}
(\bar{I}_1, I_N)_{\#} P^*(F_1 \times F_2) &= P^*\left(\bar{I}_1^{-1}(F_1) \cap I_N^{-1}(F_2)\right) \\
&= P^*(Z) \\
&= \int_{\Omega_0 \times \Omega_T} \mathbf{1}_Z(\omega, (\boldsymbol{x}, t)) \underbrace{\mathrm{d}P^*(\omega, (\boldsymbol{x}, t))}_{\mathrm{d}\left(P^0 \otimes P_{\Omega_T}\right)(\omega, (\boldsymbol{x}, t))} \\
&\overset{\text{Equation (3.4)}}{=} \int_{\Omega_0} \int_{\Omega_T} \mathbf{1}_{F_1}(I_1(\boldsymbol{x}, t)) \cdot \mathbf{1}_{F_2}(N(\omega, (\boldsymbol{x}, t))) \, \mathrm{d}P_{\Omega_T}(\boldsymbol{x}, t) \, \mathrm{d}P^0(\omega) \\
&\overset{\substack{\text{Fubini's theorem} \\ \text{thm. 2.16 in [40]}}}{=} \int_{\Omega_T} \mathbf{1}_{F_1}(I_1(\boldsymbol{x}, t)) \underbrace{\int_{\Omega_0} \mathbf{1}_{F_2}(N(\omega, (\boldsymbol{x}, t))) \, \mathrm{d}P^0(\omega)}_{P_N(F_2)} \mathrm{d}P_{\Omega_T}(\boldsymbol{x}, t) \\
&= \int_{\Omega_T} \mathbf{1}_{F_1}(I_1(\boldsymbol{x}, t)) \, \mathrm{d}P_{\Omega_T}(\boldsymbol{x}, t) \, P_N(F_2) \\
&= P_{\boldsymbol{\mathcal{F}}}(F_1) P_N(F_2).
\end{aligned}$$

✓

Statement 2:
We exploit the fact that $\bar{I}_1$ is defined to be "constant" for any $\omega \in \Omega_0$ and derive

$$\begin{aligned}
\bar{I}_{1\#}P^*\left(F_1\right) &= P^*\left(\left\{\left(\omega,\left(\boldsymbol{x},t\right)\right) \in \Omega_0 \times \Omega_T \middle| I_1\left(\boldsymbol{x},t\right) \in F_1\right\}\right) \\
&= P^*\left(\Omega_0 \times I_1^{-1}\left(F_1\right)\right) \\
&= P^0 \otimes P_{\Omega_T}\left(\Omega_0 \times I_1^{-1}\left(F_1\right)\right) \\
&= \underbrace{P^0\left(\Omega_0\right)}_{=1} P_{\Omega_T}\left(I_1^{-1}\left(F_1\right)\right) \\
&\overset{\text{Definition 3.2}}{=} P_{\boldsymbol{\mathcal{F}}}\left(F_1\right).
\end{aligned}$$

✓

Statement 3:
Similarly, we derive

$$\begin{aligned}
I_{N\#}P^*\left(F_2\right) &= P^*\left(\left\{\left(\omega,\left(\boldsymbol{x},t\right)\right) \in \Omega_0 \times \Omega_T \middle| I_N\left(\omega,\left(\boldsymbol{x},t\right)\right) \in F_2\right\}\right) \\
&\overset{\substack{\text{Fubini's theorem} \\ \text{thm. 2.16 in [40]}}}{=} \int_{\Omega_T} \underbrace{\int_{\Omega_0} \mathbf{1}_{F_2}\left(I_N\left(\omega,\left(\boldsymbol{x},t\right)\right)\right) \, \mathrm{d}P^0\left(\omega\right)}_{=I_{N\#}P^0(F_2)=P_N(F_2)} \, \mathrm{d}P_{\Omega_T}\left(\boldsymbol{x},t\right) \\
&= P_N\left(F_2\right) \underbrace{\int_{\Omega_T} 1 \, \mathrm{d}P_{\Omega_T}\left(\boldsymbol{x},t\right)}_{=1} \\
&= P_N\left(F_2\right).
\end{aligned}$$

✓

By making use of the three statements, the assertion

$$\left(\bar{I}_1, I_N\right)_{\#} P^*\left(E\right) = \left(\bar{I}_{1\#}P^*\right) \otimes \left(I_{N\#}P^*\right)\left(E\right)$$

follows directly for all $E = F_1 \times F_2$ for arbitrary measurable $F_1, F_2 \in \boldsymbol{\mathcal{F}}$. □

One could also derive the stochastic independence based on the fact that I_1 is actually deterministic and, consequently, stochastic independent to any other stochastic event. With the stochastic independence of the feature image and the noise image at hand, we delve into a statement on the probability measure and probability density function for the disturbed feature image.

Proposition 3.7 (Probability density functions of the disturbed feature image I_1^{d})
Given a feature image I_1 and a noise image I_N as stated in the previous Definitions 3.2 and 3.5, we use $P_{\boldsymbol{\mathcal{F}}}$ and P_N as the related probability measures living on $\boldsymbol{\mathcal{F}}$. Let $P_{\boldsymbol{\mathcal{F}}}^{\mathrm{d}}$ be the probability measure

related to the mapping of the disturbed feature image, i.e., we consider it as the pushforward measure

$$P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(F) := I^{\mathrm{d}}_{1\,\#}P^*(F) = P^*\left(\left(I^{\mathrm{d}}_1\right)^{-1}(F)\right)$$

with the product measure $P^* = P^0 \otimes P_{\Omega_T}$ related to the measurable space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T))$ again and with $F \in \mathcal{B}(\boldsymbol{\mathcal{F}})$ arbitrary.
Then it holds for the probability measure of the disturbed feature image

$$P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(F) = \int_{\boldsymbol{\mathcal{F}}} P_N(F - y)\ \mathrm{d}P_{\boldsymbol{\mathcal{F}}}(y).$$

The related probability density function of the disturbed feature image is given as the convolution of the measure $P_{\boldsymbol{\mathcal{F}}}$ with the Gaussian kernel p_N resulting in

$$p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f) = \int_{\boldsymbol{\mathcal{F}}} p_N(f - y)\ \mathrm{d}P_{\boldsymbol{\mathcal{F}}}(y)$$

for $f \in \boldsymbol{\mathcal{F}}$ and by integration with respect to the measure $P_{\boldsymbol{\mathcal{F}}}$.

Proof. The statement follows directly with the stochastic independence of the extended feature image $\bar{I}_1$ and the noise image I_N considered as mappings from $\Omega_0 \times \Omega_T$ to $\boldsymbol{\mathcal{F}}$ and with the probability space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T), P^*)$ where P^* denotes again the product measure $P^0 \otimes P_{\Omega_T}$. According to the Definition 2.32 in [40], it holds that the probability measure for the sum of two independent random variables is given as the convolution of their individual probability measures. Consequently, we derive for I^{d}_1 defined in Equation (3.3)

$$P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}} = P_N * P_{\boldsymbol{\mathcal{F}}} \tag{3.5}$$

by making use of the previous statement on the stochastic independence of $\bar{I}_1$ and I_N in Proposition 3.6. Let $F \in \mathcal{B}(\boldsymbol{\mathcal{F}})$ be measurable and arbitrary. Then it holds with Equation (3.5) and the convolution of measures (cf. Definition 14.17 in [40])

$$\begin{aligned}
P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(F) &= (P_N * P_{\boldsymbol{\mathcal{F}}})(F)\,(P_N \otimes P_{\boldsymbol{\mathcal{F}}})(\{(f_1, f_2) \in \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{F}} \mid f_1 + f_2 \in F\}) \\
&= \int_{\boldsymbol{\mathcal{F}}} \int_{\boldsymbol{\mathcal{F}}} \mathbf{1}_F(f_1 + f_2)\ \mathrm{d}P_N(f_1)\ \mathrm{d}P_{\boldsymbol{\mathcal{F}}}(f_2) \\
&= \int_{\boldsymbol{\mathcal{F}}} \int_{\boldsymbol{\mathcal{F}}} \mathbf{1}_{F-f_2}(f_1)\ \mathrm{d}P_N(f_1)\ \mathrm{d}P_{\boldsymbol{\mathcal{F}}}(f_2) \\
&= \int_{\boldsymbol{\mathcal{F}}} \int_{F-f_2} 1\ \mathrm{d}P_N(f_1)\ \mathrm{d}P_{\boldsymbol{\mathcal{F}}}(f_2) = \int_{\boldsymbol{\mathcal{F}}} P_N(F - f_2)\ \mathrm{d}P_{\boldsymbol{\mathcal{F}}}(f_2).
\end{aligned}$$

By making use of P_N being absolutely continuous to the Lebesgue measure on $\boldsymbol{\mathcal{F}}$ and with applying the corresponding probability density function p_N, we can extend this to

$$\begin{aligned}
P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(F) &= \int_{\boldsymbol{\mathcal{F}}} P_N(F - f_2)\ \mathrm{d}P_{\boldsymbol{\mathcal{F}}}(f_2) \\
&= \int_{\boldsymbol{\mathcal{F}}} \int_F p_N(f - f_2)\ \mathrm{d}f\ \mathrm{d}P_{\boldsymbol{\mathcal{F}}}(f_2) \\
&\overset{\text{Fubini's theorem}}{=} \int_F \int_{\boldsymbol{\mathcal{F}}} p_N(f - f_2)\ \mathrm{d}P_{\boldsymbol{\mathcal{F}}}(f_2)\ \mathrm{d}f
\end{aligned}$$

with $\mathrm{d}f$ marking the integration with respect to the Lebesgue measure on $\mathcal{F}$. Finally, we receive with this transformation the probability density function with respect to the Lebesgue measure on $\mathcal{F}$ and related to the disturbed feature image and living on $\mathcal{F}$ to be given by

$$p_{\mathcal{F}}^{\mathrm{d}}(f) = \int_{\mathcal{F}} p_N(f-y)\ \mathrm{d}P_{\mathcal{F}}(y).$$

□

The used concept of absolute continuity of measures is introduced more thoroughly in a later section and we refer the reader especially to Definition 5.5.
Before we continue with concrete examples of feature images, we comment on the effect of the Gaussian noise when it comes to joint probability density distributions. Therefore, we start by introducing a joint probability measure when considering feature images together with a second "image modality", i.e., in this context we include classification images as a second data source regarding joint probabilities.

Definition 3.8 (Probability measures on classification and joint spaces)
We assume that we have a second function I_2 for a classification image, mapping from the spatio-temporal domain Ω_T to the classification space $\mathcal{C} = \mathbb{R}$. Let the corresponding probability measure be $P_{\mathcal{C}}$ describing the probability that I_2 maps values from Ω_T to a measurable $C \in \mathcal{B}(\mathcal{C})$, i.e.,

$$P_{\mathcal{C}}(C) = P_{\Omega_T}\left(I_2^{-1}(C)\right)$$

with P_{Ω_T} being the probability measure related to the measurable space $(\Omega_T, \mathcal{B}(\Omega_T))$ (cf. Definition 3.1) again. We assume that $P_{\mathcal{C}}$ is absolutely continuous to the Lebesgue measure and the related probability density function is given by $p_{\mathcal{C}} : \mathcal{C} \to \mathbb{R}_+$ for which it holds that

$$P_{\mathcal{C}}(C) = \int_C p_{\mathcal{C}}(c)\ \mathrm{d}c \quad \forall C \in \mathcal{B}(\mathcal{C})$$
$$P_{\mathcal{C}}(\mathcal{C}) = \int_{\mathcal{C}} p_{\mathcal{C}}(c)\ \mathrm{d}c = 1.$$

We introduce the joint probability measure $P_{\mathcal{F}\times\mathcal{C}}$ living on $\mathcal{F}\times\mathcal{C}$ with

$$P_{\mathcal{F}\times\mathcal{C}}(E) := (I_1, I_2)_{\#} P_{\Omega_T}(E) = P_{\Omega_T}\left(\{(\boldsymbol{x}, t) \in \Omega_T \mid (I_1(\boldsymbol{x}, t), I_2(\boldsymbol{x}, t)) \in E\}\right)$$

for arbitrary $E \in \mathcal{B}(\mathcal{F}\times\mathcal{C})$ and for $E = F \times C$ with $F \in \mathcal{B}(\mathcal{F})$ and $C \in \mathcal{B}(\mathcal{C})$, we can further derive

$$P_{\mathcal{F}\times\mathcal{C}}(E) = P_{\Omega_T}\left(\left\{I_1^{-1}(F) \cap I_2^{-1}(C)\right\}\right).$$

The joint probability measure corresponds to the probability of different feature and classification combinations when regarding the pushforward of P_{Ω_T} with respect to the joint mapping of the feature image and the classification image $(I_1, I_2) : \Omega_T \to \mathcal{F}\times\mathcal{C}$.

To establish a relation between the joint probability measure $P_{\mathcal{F}\times\mathcal{C}}$ and the individual measure on the classification space $\mathcal{C}$, we apply the disintegration of measures. Our next theorem is oriented on the disintegration theorem from [4] stated in Theorem 5.3.1 or, more precisely, on the remark

following this theorem dealing with the disintegration for product spaces. We adapt the statement to our setting concerning the product space $\mathcal{F} \times \mathcal{C}$ and cite the concept without focusing on its proof. As we will use the concept of disintegration in the further course when deriving derivative terms of histograms in Section 5.2.4, cf. Theorem 5.29 especially, again but in a slightly different form, we label the following theorem as the first version.

Theorem 3.9 (Adapted Disintegration Theorem — Version 1)
We consider the probability space $(\mathcal{F} \times \mathcal{C}, \mathcal{B}(\mathcal{F} \times \mathcal{C}), P_{\mathcal{F}\times\mathcal{C}})$ and the measurable spaces $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$, $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$. Let the natural projection onto the classification domain $\mathcal{C}$ be

$$\pi_{\mathcal{C}} : \mathcal{F} \times \mathcal{C} \to \mathcal{C}$$

and let

$$\mu := \pi_{\mathcal{C}\#} P_{\mathcal{F}\times\mathcal{C}} = P_{\mathcal{F}\times\mathcal{C}} \circ \pi_{\mathcal{C}}^{-1}$$

be a probability measure for the measurable space $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$. Then each fiber $\pi_{\mathcal{C}}^{-1}(c) = \mathcal{F} \times \{c\}$ can canonically be identified with $\mathcal{F}$ for any $c \in \mathcal{C}$. Moreover, there exists a μ-almost everywhere uniquely determined Borel family of probability measures $\boldsymbol{\nu} = \{\nu_c\}_{c\in\mathcal{C}}$ related to the measurable space $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$ such that

$$P_{\mathcal{F}\times\mathcal{C}}(A) = \int_{\mathcal{C}} \nu_c\left(A \cap \pi_{\mathcal{C}}^{-1}(c)\right) \, \mathrm{d}\mu(c) = \int_{\mathcal{C}} \nu_c\left(\left\{\left(f, c'\right) \in A \,|\, c' = c\right\}\right) \, \mathrm{d}\mu(c)$$

holds for any measurable set $A \in \mathcal{B}(\mathcal{F} \times \mathcal{C})$. Furthermore, it holds

$$\int_{\mathcal{F}\times\mathcal{C}} g(f,c) \, \mathrm{d}P_{\mathcal{F}\times\mathcal{C}}(f,c) = \int_{\mathcal{C}} \int_{\pi_{\mathcal{C}}^{-1}(c)=\mathcal{F}} g(f,c) \, \mathrm{d}\nu_c(f) \, \mathrm{d}\mu(c)$$

for every Borel map $g : \mathcal{F} \times \mathcal{C} \to [0, +\infty]$.
We use the notation $P_{\mathcal{F}\times\mathcal{C}} = (\mu, \boldsymbol{\nu})$ for a disintegration along $\mathcal{C}$ of the probability measure.

For the remainder of this section, we consider one fixed family of probability measures $\boldsymbol{\nu} = \{\nu_c\}_{c\in\mathcal{C}}$ for the disintegration of the joint measure $P_{\mathcal{F}\times\mathcal{C}}$ with respect to its projection on the classification space $\mathcal{C}$, i.e., the measure $\mu = \pi_{\mathcal{C}\#} P_{\mathcal{F}\times\mathcal{C}}$. With this disintegration of the probability measure on the joint space, we can establish a direct relation between the joint measure $P_{\mathcal{F}\times\mathcal{C}}$ and the measure $P_{\mathcal{C}}$ related to the individual space $\mathcal{C}$.

Proposition 3.10
We consider the probability measure $P_{\mathcal{C}}$ related to the classification space $\mathcal{C}$ and its probability density function $p_{\mathcal{C}}$ as defined in Definition 3.8. Let $(\mu, \boldsymbol{\nu})$ be the disintegration along $\mathcal{C}$ of $P_{\mathcal{F}\times\mathcal{C}}$ related to the projection map $\pi_{\mathcal{C}}$ as formulated in Theorem 3.9. Then it holds that

$$P_{\mathcal{C}}(C) = \pi_{\mathcal{C}\#} P_{\mathcal{F}\times\mathcal{C}}(C) \tag{3.6}$$

for any $C \in \mathcal{B}(\mathcal{C})$ and $p_{\mathcal{C}}$ is also the probability density function to $\mu = \pi_{\mathcal{C}\#} P_{\mathcal{F}\times\mathcal{C}}$ when integrating with respect to the Lebesgue measure on $\mathcal{C}$:

$$\mu(C) = \int_C p_{\mathcal{C}}(c)\ \mathrm{d}c \tag{3.7}$$

Proof. Let $C \in \mathcal{B}(\mathcal{C})$ be arbitrary. We derive with the joint measure $P_{\mathcal{F}\times\mathcal{C}}$ in Definition 3.8

$$\begin{aligned}
\pi_{\mathcal{C}\#} P_{\mathcal{F}\times\mathcal{C}}(C) &= \pi_{\mathcal{C}\#}\left((I_1, I_2)_{\#} P_{\Omega_T}\right)(C) \\
&= (I_1, I_2)_{\#} P_{\Omega_T}\underbrace{\left(\pi_{\mathcal{C}}^{-1}(C)\right)}_{=\mathcal{F}\times C} \\
&= P_{\Omega_T}\left(I_1^{-1}(\mathcal{F}) \cap I_2^{-1}(C)\right) \\
&= P_{\Omega_T}\left(\Omega_T \cap I_2^{-1}(C)\right) \\
&= P_{\Omega_T}\left(I_2^{-1}(C)\right) = P_{\mathcal{C}}(C)
\end{aligned}$$

to prove the equality of the projected joint measure with the previously defined probability measure related to the classification space $\mathcal{C}$. With this equality the statement on the probability density function $p_{\mathcal{C}}$ in Equation (3.7) follows directly. □

Building up on Proposition 3.7, we focus now on the joint probability when considering a disturbed feature image I_1^{d} affected by Gaussian noise.

Definition 3.11 (Joint probability measure for disturbed features)
We introduce the joint probability measure $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ on $\mathcal{F}\times\mathcal{C}$ related to feature images corrupted by Gaussian noise and the classification images to be given by

$$P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(E) := \left(I_1^{\mathrm{d}}, I_2\right)_{\#} P^*(E) = P^*\left(\left\{(\omega, (\boldsymbol{x}, t)) \in \Omega_0 \times \Omega_T \mid \left(I_1^{\mathrm{d}}(\omega, (\boldsymbol{x}, t)), I_2(\boldsymbol{x}, t)\right) \in E\right\}\right)$$

for arbitrary $E \in \mathcal{B}(\mathcal{F}\times\mathcal{C})$ and with $P^* = P^0 \otimes P_{\Omega_T}$.

Since the Gaussian noise is only affecting the probabilities of related disturbed features, it is an intuitive assumption to consider the disturbed joint probability resulting from convolution of the original probability measure $P_{\mathcal{F}\times\mathcal{C}}$ with the probability measure of the related noise model along the subspace $\mathcal{F}$. In the next proposition, we focus on this hypothesis.

Proposition 3.12
Let $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$, $P_{\mathcal{F}\times\mathcal{C}}$ and P_N be the introduced probability measures in Definitions 3.4, 3.8 and 3.11. Considering the convolution of measures, we state that

$$P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} = P_N *_{\mathcal{F}} P_{\mathcal{F}\times\mathcal{C}}$$

holds with $*_{\mathcal{F}}$ denoting the convolution only along the subspace $\mathcal{F}$. This joint probability measure considering the noise effects on the feature images $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ is absolutely continuous with respect to the Lebesgue measure on $\mathcal{F}\times\mathcal{C}$. Based on the disintegration of $P_{\mathcal{F}\times\mathcal{C}} = (P_{\mathcal{C}}, \nu)$ along the

classification space $\mathcal{C}$ (cf. Theorem 3.9 and proposition 3.10), we can derive the related probability density function to be given by

$$p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c) = (p_N * \nu_c)(f)\, p_{\mathcal{C}}(c)\,.$$

Proof. To show the equality

$$P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} = P_N *_{\mathcal{F}} P_{\mathcal{F}\times\mathcal{C}},$$

we consider the convolution of measures (cf. e.g. Definition 14.17 in [40]) which acts in this case only on values related to the space $\mathcal{F}$, i.e., it holds

$$(P_N *_{\mathcal{F}} P_{\mathcal{F}\times\mathcal{C}})(E) = P^*\left(\{(\omega,(\boldsymbol{x},t)) \in \Omega_0\times\Omega_T \mid (I_N(\omega,(\boldsymbol{x},t)) + I_1(\boldsymbol{x},t), I_2(\boldsymbol{x},t)) \in E\}\right)$$

for an arbitrary $E \in \mathcal{B}(\mathcal{F}\times\mathcal{C})$. With this follows the statement directly with the disturbed feature image I^{d}_1 in Definition 3.5 and $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ in Definition 3.11.
With the Gaussian noise disturbing the feature images, we get that $P^{\mathrm{d}}_{\mathcal{F}}$ is absolutely continuous with respect to the Lebesgue measure on the feature space $\mathcal{F}$. Moreover, the probability measure $P_{\mathcal{C}}$ is considered to be absolutely continuous with respect to the Lebesgue measure on the classification space $\mathcal{C}$ and the corresponding probability density function is given by $p_{\mathcal{C}}$. Based on this and by applying the disintegration theorem (cf. Theorem 3.9), we derive the joint probability density function related to the joint measure $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ and living on the joint space $\mathcal{F}\times\mathcal{C}$. For the disintegration of the joint measure, we use $P_{\mathcal{F}\times\mathcal{C}} = (P_{\mathcal{C}}, \boldsymbol{\nu})$. We consider $A \subset \mathcal{F}\times\mathcal{C}$ to be measurable and derive

$$\begin{aligned}
P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(A) &= (P_N *_{\mathcal{F}} P_{\mathcal{F}\times\mathcal{C}})(A)\\
&= \int_{\mathcal{F}}\int_{\mathcal{F}\times\mathcal{C}} p_N(f-y)\,\mathbf{1}_A(f,c)\;\mathrm{d}P_{\mathcal{F}\times\mathcal{C}}(y,c)\;\mathrm{d}f\\
&= \int_{\mathcal{F}}\int_{\mathcal{C}}\int_{\mathcal{F}} p_N(f-y)\,\mathbf{1}_A(f,c)\;\mathrm{d}\nu_c(y)\;\mathrm{d}P_{\mathcal{C}}(c)\;\mathrm{d}f\\
&= \int_{\mathcal{F}}\int_{\mathcal{C}} \underbrace{\int_{\mathcal{F}} p_N(f-y)\;\mathrm{d}\nu_c(y)}_{=(p_N*\nu_c)(f)}\,\mathbf{1}_A(f,c)\;\mathrm{d}P_{\mathcal{C}}(c)\;\mathrm{d}f\\
&= \int_{\mathcal{F}}\int_{\mathcal{C}} \mathbf{1}_A(f,c)\,(p_N*\nu_c)(f)\,\underbrace{\mathrm{d}P_{\mathcal{C}}(c)}_{=p_{\mathcal{C}}(c)\,\mathrm{d}c}\,\mathrm{d}f\\
&= \int_{\mathcal{F}}\int_{\mathcal{C}} \mathbf{1}_A(f,c)\,(p_N*\nu_c)(f)\,p_{\mathcal{C}}(c)\;\mathrm{d}c\,\mathrm{d}f\\
&= \int_{\mathcal{F}\times\mathcal{C}} \mathbf{1}_A(f,c)\,p_{\mathcal{C}}(c)\,(p_N*\nu_c)(f)\;\mathrm{d}(f,c)\\
&= \int_{A} p_{\mathcal{C}}(c)\,(p_N*\nu_c)(f)\;\mathrm{d}(f,c)\,.
\end{aligned}$$

This proves that $p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} : \mathcal{F}\times\mathcal{C} \to \mathbb{R}_+$ defined by $p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c) := (p_N*\nu_c)(f)\,p_{\mathcal{C}}(c)$ is indeed the probability density function of the probability measure $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ when considering integration with respect to the Lebesgue measure on $\mathcal{F}\times\mathcal{C}$. Moreover, the existence of the probability density

functions shows the absolute continuity of $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ with respect to the Lebesgue measure on the joint space $\mathcal{F}\times\mathcal{C}$. □

We conclude the excursion on probability measures and related density functions when considering noisy features with a final estimation, that will be crucial in later proofs.

Proposition 3.13
We consider the probability measure $P_{\mathcal{C}}$ on the classification space $\mathcal{C}$ and the probability measure $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ on the joint space $\mathcal{F}\times\mathcal{C}$ when including the effect of noisy features. For the related density functions with respect to the Lebesgue measures the following inequality holds:

$$p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c) \leq \|p_N\|_\infty p_{\mathcal{C}}(c) \quad \text{for all } (f,c)\in\mathcal{F}\times\mathcal{C}.$$

Proof. For arbitrary $(f,c)\in\mathcal{F}\times\mathcal{C}$ we obtain

$$\begin{aligned} p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c) &= (p_N * \nu_c)(f)\, p_{\mathcal{C}}(c) \\ &= \int_{\mathcal{F}} \underbrace{p_N(f-y)}_{\leq\|p_N\|_\infty} \mathrm{d}\nu_c(y)\, p_{\mathcal{C}}(c) \\ &\leq \|p_N\|_\infty \underbrace{\int_{\mathcal{F}} \mathrm{d}\nu_c(y)}_{=1} p_{\mathcal{C}}(c) \\ &= \|p_N\|_\infty p_{\mathcal{C}}(c). \end{aligned}$$

We exploit here in the final step that ν_c for all $c\in\mathcal{C}$ is a probability measure on $\mathcal{F}$. Moreover, we use the fact that for the noise image I_N which is $\mathcal{N}\left(0,\sigma^2\mathbf{1}_{n\times n}\right)$-randomly distributed for a standard deviation $\sigma>0$, we know that its probability density function p_N is bounded in the infinity-norm with

$$\|p_N\|_\infty < \infty.$$

□

Due to the Gaussian noise model, the values of the disturbed feature image I_1^{d} can be distributed over the whole feature space such that

$$I_1^{\mathrm{d}}(\omega_0,\Omega_T) = \mathcal{F}$$

for arbitrary $\omega_0\in\Omega_0$ holds. We assume that the occurring features in the original feature image I_1 are bounded, i.e., the related probability measure has a bounded support. This is a valid assumption when considering that the features are extracted based on gray values that are themselves naturally bounded.
When thinking about the related probability distributions, we see that the application of a Gaussian filter which is itself not compactly supported, results in a probability distribution living on the whole feature space such that

$$\operatorname{supp}\left(p^{\mathrm{d}}_{\mathcal{F}}\right) = \mathcal{F}$$

holds. Because of the convolution with the Gaussian function p_N, it holds that $p^{\mathrm{d}}_{\mathcal{F}}$ converges towards zero without ever reaching 0 for features "converging to infinity" themselves in the sense that for example their Euclidean norm converges to infinity. To facilitate the later analysis, we introduce the following subset of the feature space to achieve a lower bound for the probability density distribution related to the disturbed feature image I^{d}_1.

Definition 3.14 (Reduced feature space $\mathcal{F}'$)
Let $\delta > 0$, $\delta \ll 1$, be arbitrary but fixed. Then we define the reduced features space as

$$\mathcal{F}' := \left\{ f \in \mathcal{F} \,|\, p^{\mathrm{d}}_{\mathcal{F}}(f) > \delta \right\}.$$

For simplifications in later context we assume that the n-dimensional reduced space is given as the Cartesian product $\mathcal{F}' = \mathcal{F}'_1 \times \cdots \times \mathcal{F}'_n$. Additionally, we point out an interesting and important interpretation of the noisy feature images in the following remark for later applications.

Remark 3.15. When considering the noise effect on our feature images, it is always depending on an $\omega_0 \in \Omega_0$. Indeed we regard $I^{\mathrm{d}}_1 : \Omega_0 \times \Omega_T \to \mathcal{F}$ as a random variable considering the joint measure space $(\Omega_0 \times \Omega_T, \mathcal{E} \times \mathcal{B}(\Omega_T), P^*)$. As we are including here the whole domain Ω_0 instead of just focusing on one arbitrary but fixed $\omega_0 \in \Omega_0$, we remark that this corresponds to the interpretation that our feature image for a certain time point t is not only affected by *one* noisy image but rater by a infinite number of noise images (considering all $\omega_0 \in \Omega_0$).

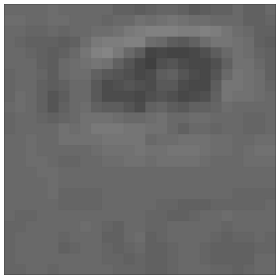
As a result of this subsection, we derived an upper bound for the joint probability density function in Proposition 3.13 and introduced the reduced feature space in the last definition. Both concepts are strongly related to probabilities affected by the noisy features and are crucial for the later analysis of our main optimization problem in Section 5.4.

After introducing the feature image and its noisy counterpart in a general setting, we focus next on different feature extraction approaches. We start with features derived from local basic texture information.

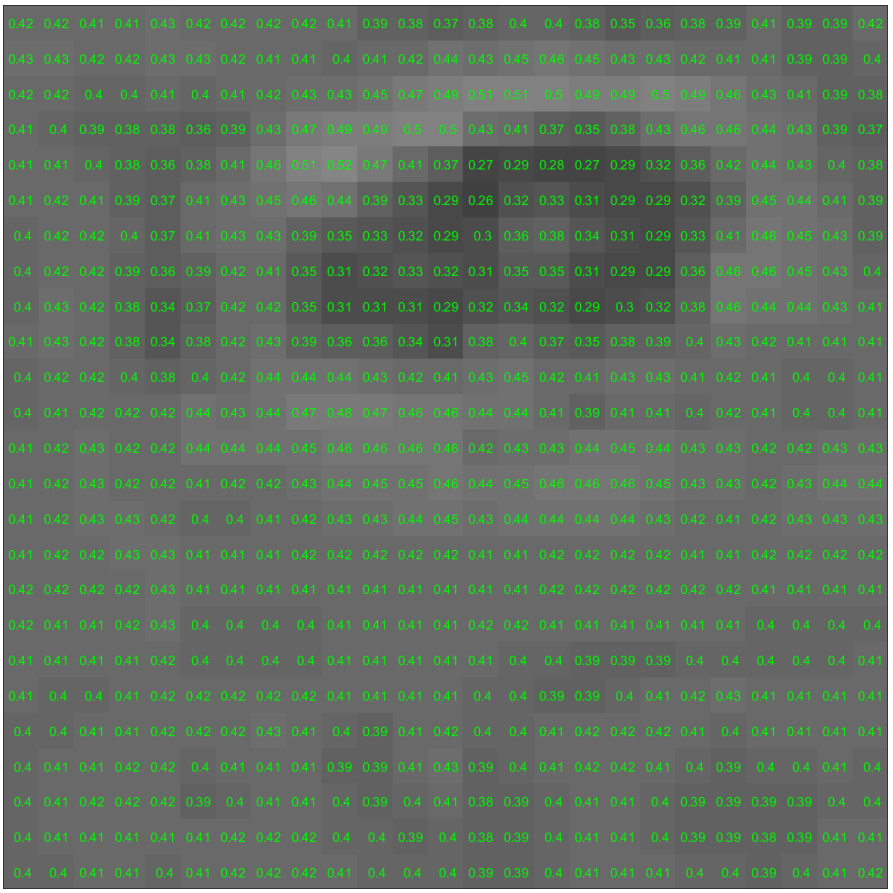
3.3.2 Features based on local texture information

We are aiming for an optimization approach which facilitates the model fitting based on features extracted from the microscopy images. In this section, we describe some simple texture features which we will use in the later numerical tests to validate the optimization approach. With minimum and maximum gray values of small subregions in the images and also with local interquartile ranges, we use properties often used in descriptive statistics to extract basic texture characteristics. Furthermore, those features describe texture properties which are also perceivable for the human eye and so it is more straightforward to link them manually with specific texture regions which are expected to reveal particular areas of a cell colony.

The texture features we are focusing on are based on the given gray values in the pixels of the microscopy image. Inspired by the appearance of a cell with significant cell boundaries, we derive the following properties. In Figure 3.11a there is an individual cell located in the subpatch of a microscopy image revealing the underlying pixel grid. Based on the grayscale values, we derive local maxima and local minima of the gray values, cf. Figure 3.11b. We observe that the cell is surrounded by a brighter halo effect while the inside of the cell is noticeable due to the darker pixels compared to the background area. When considering the grayscale values to be in a range between 0 and 1, we consider the brighter pixels to have a larger gray value while the darker ones have a smaller gray value. In the extreme cases, a black pixel corresponds to 0 whereas white represents a pixel of gray value 1. Based on this definition, we apply local dilation and erosion operations with disk filter kernels to extract local minima and local maxima. The radii of the disk kernels differ slightly. While we use a larger radius for the local maxima of 10 pixels, we apply a smaller one of 5 pixels for the local minima. The idea here is to ensure that the filter for local maxima is large enough to avoid dark spots *within* a cell after applying the dilation kernel. Likewise, we allow a smaller kernel for the erosion filter to focus on local minima because we want to enforce that the filter stresses the dark pixel *inside* a cell without increasing the cell's outline a lot.



(a) Image patch of the single cell revealing the pixel structure.



(b) Gray scale values of each pixel on the subpatch with the single cell.

Figure 3.11: A single cell in a subpatch of a microscopy image.

We illustrate the extraction of local minima in Figure 3.12 and Figure 3.13. In Figure 3.12, we mark disk shaped regions with a red, yellow and light blue circle in which we focus on the local minima.

The center pixel of each circle is marked with a small rectangle and the pixel with the smallest gray value is marked by a small hexagon in the same color as the corresponding circle region. Figure 3.13 shows the resulting image after the erosion process. The local minimum of each circle is assigned to the corresponding center pixel marked again with a small rectangle.
For the given example subpatch with a single cell, we illustrate the result of the erosion and dilation

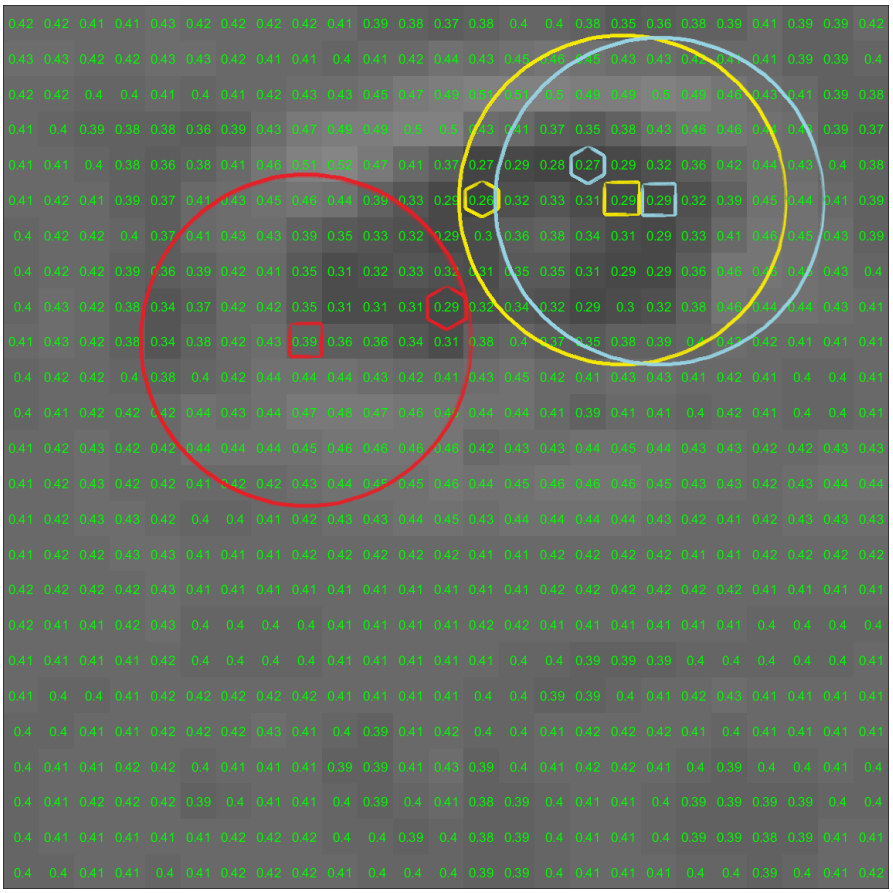


Figure 3.12: Gray scale values of each pixel on a sub patch with a single cell with circular regions identifying local neighborhoods for the erosion operation.

processes for the extraction of local minima and maxima in Figure 3.14. Here, we clearly observe that the extraction of local maxima is based on a greater filter kernel such that the total cell area is stressed by a bright spot in the local maxima image without returning any darker holes inside, cf. Figure 3.14c. This also results in a slightly larger "cell area" in the local maxima image compared to the cell's region in the local minima image. Next to the local maxima and local minima gray values, we also focus on the distribution of grayscale values in a small neighborhood. Therefore, we use for the third texture feature the interquartile range (cf. [65]) of occurring gray values in a small square neighborhood (25×25 pixels). We already introduced the interquartile ranges for small image patches in Section 3.1 and illustrated this measure in Figures 3.6 and 3.7.
Finally, we exemplify those basic texture features for small subpatches in Figure 3.15. For four example image patches at significant time points, we plot the extracted texture features into the three dimensional feature space, i.e., each dot in the middle three-dimensional feature plot represents a pixel of the subimage framed in the same color. The patch for the first time point with the red frame

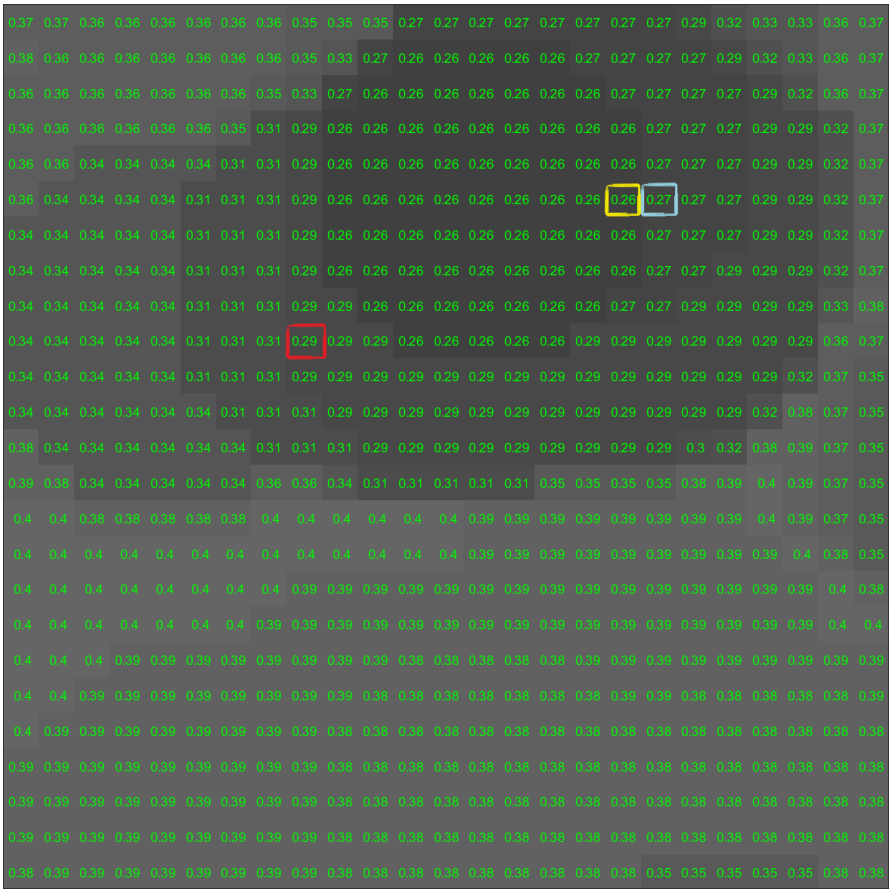


Figure 3.13: Resulting gray values after eroding the image with a disk shaped kernel.

represents an empty field of view only showing the background of a well. In green boundaries, we display a patch with separated, individual cells at the sixth time stamp. For the last two time points (time stamp 7 and 8) the patches in a blue and purple frame show significant changes in the texture. The range of grayscale values is particularly higher here and no single cells are observable anymore. In the three dimensional graph the first axis represents

$$1 - \text{local minima}.$$

We use this inverted version of the local minima to enforce an increasing course of extracted features when moving forward in time. While for the first background patch we have a smooth gray tone, we get increasing values for the local maxima and, respectively, decreasing values for local minima for patches of later time points. All in all, we observe that for the given time points we get an increasing course in all three plotted dimensions when proceeding in time. The extracted features accumulate accordingly approximately on the related diagonal plane in the three dimensional plot.
Based on the nature of the gray values normalized between 0 and 1[1], we conclude that the extracted features are certainly living in

$$[0,1]^3 \subset \mathcal{F} = \mathbb{R}^3. \tag{3.8}$$

[1] In other contexts a normalization between 0 and 255 is also commonly used.

(a) Original gray values of subpatch with single cell.

(b) Extracted local minima after erosion operation.

(c) Extracted local maxima after dilation operation.

Figure 3.14: Local texture features after erosion and dilation in comparison to original gray values.

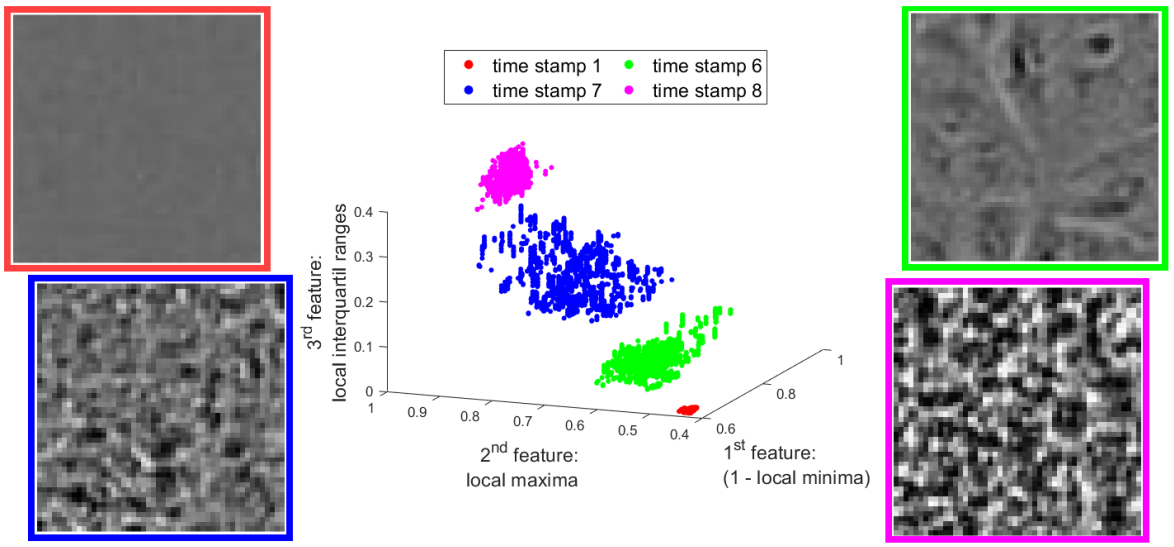
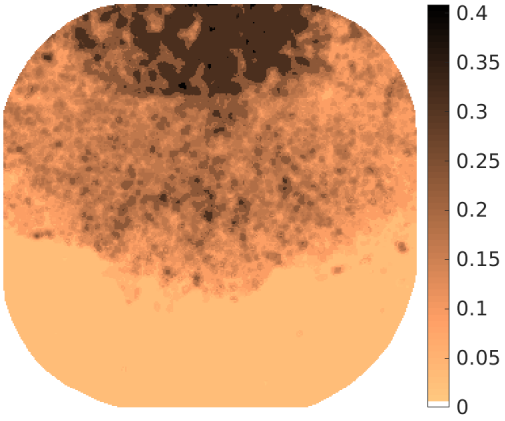
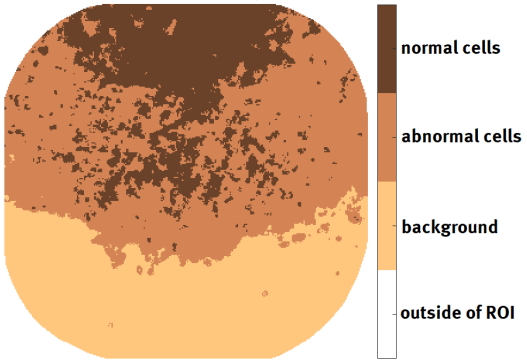


Figure 3.15: Extracted texture information in the 3d feature space for four example subpatches for significant occurrences in the microscopy data.

With those three texture features we aim for an optimization approach to fit a given spreading model to our data. Given those basic texture features, we could already aim for a first classification of the colony area based on thresholding in the feature space. As an example, we present in Figure 3.16 the third texture feature representing the interquartile ranges for an example well at the final time point. Additionally, we show the thresholded version classifying the well area into three different segments: The background area, a region with "normal" cells and the area of "abnormal" cells. Certainly, this needs a further validation with biologists and possibly additional imaging of the cell colony.
Considering segmentation results for the total colony area based on those thresholded features for example or as given in Figure 3.8, we can observe the spreading process of the cell colony in the real data by focusing on the moving colony fronts based on the segmented colony area. In Figure 3.9 we already presented such moving colony fronts of a cell population in an example well over time. The time points are reflected in the varying colors of the moving front domains such that the earlier time stamps correspond to blue wave fronts while the later ones are shown in purple and even pink with a legend of time points given in *days*. Again, we stress that the segmentation results depend on manually selected threshold values there.



(a) Smoothed interquartile ranges revealing different colony sections.



(b) Classification of different regions in field of view based on thresholded interquartile ranges.

Figure 3.16: The third feature representing local interquartile ranges of gray values reveals different areas within a cell colony.

As we want to reduce the pre-processing steps and use feature data close to the original data, we apply the model fitting approach in the further course on the basic texture features presented in this section. With this we avoid the necessity of hand-crafted threshold values for the segmentation approach, plus allow model fitting for two subcolonies. We remind that we are not only focusing on the spreading of the total cell population which is also represented in the segmentation masks, but rather aim to derive spreading properties for a subcolony of "normal" cells and another one for the "abnormal" cells. With a mutual information based optimization approach, we intend to derive spreading properties based on the texture information which also reveal the different colony areas. We focus more on possible spreading models and on the optimization approach in the next chapters. Before, we briefly comment on other extraction methods to derive feature images that could be used in the optimization process.

3.3.3 Alternative feature images

In this section we want to elaborate on three other feature extraction approaches that could be used to incorporate the data into the optimization model. Briefly, we discuss their benefits and drawbacks individually and conclude with some further literature reference to extract texture properties from imaging data.

In our optimization problem the feature images are our tool to link the applied spreading model with the given data. One straightforward approach to incorporate the data is to use the given microscopy images as the actual feature image. This would accelerate the pre-processing significantly as we do not need a specific method to *extract* the features from the data. Then again, it is important to take into account that the microscopy images might be affected by noise. Therefore, it could be beneficial to apply some pre-processing steps to smooth noise effects. This pre-processing is still

faster than extracting several different texture features. In comparison with the introduced local texture features in Section 3.3.2, this smoothing can be implemented more efficiently by applying a filtering method via convolutional operations. A famous approach for noise reduction is for example to apply a Gaussian filter [11]. Convolutional filters are in this context superior to the texture feature extraction in Section 3.3.2 because the implementation is computationally expensive for those *non-linear* filters determining local minima, local maxima and local interquartile ranges.

Improving the used texture features is also a step forward in another direction. The features extracted in Section 3.3.2 are simple texture features and easy to grasp. Still, they are also sensitive to differences in the imaging process. For example, variances in the exposure or lighting of the imaging set up can have a great impact on the local minima and local maxima features. So in general, it would be better to rule out such sensitivities by focusing more on the variances of the gray values in small neighborhoods. Indeed, the local interquartile ranges are already a step in the right direction. Calculating the variance or standard deviation of local grayscale values can contribute to identify differences in varying colony regions more significantly. Even more or in addition to this feature, one could also calculate local entropy values based on small neighborhood regions. Entropy measures the dispersion of probabilities for particular events, here for grayscale values, in certain neighborhood regions [61] and it can be used to describe texture characteristics of a total image or in small local sub patches as in out context [31]. We will focus on the different interpretations for entropy in the further course of this thesis when we introduce the concept of mutual information in Section 5.1. For now, we stress that local entropy values are well designed to grasp local variances of texture information and, consequently, to differentiate varying texture regions within our imaging field of view in the microscopy data.

In the past few years, neural networks are gaining in importance in the field of image processing [68]. Especially, deep neural networks (abbreviation: DNN) have become increasingly important in the last years. A deep neural network is a special case of a neural network which consists of multiple hidden layers. The name of a neural network is derived from the similarity of its structure with the constitution of the human brain: Neurons are connected with other neurons, receive input information from preceding neurons and send output information to other neighboring neurons. In the artificial network, this is incorporated by different layers of neurons. Based on the received input from preceding layers, neurons in the current layer can send output information to the neurons in the next level. By this, we simulate an information flow in which the neurons "contribute" to extract specific insights of the given input data. The input information, here, e.g., an image, is passed through to derive a particular output, such as a classification of the image in the example case.
Neurons of neighboring layers are connected via weighting terms and contribute this way to an information flow to the neurons in the next layer. The information of a neuron depends on the weighted sum of the pieces of information handed through by the connected neurons of the previous layer. Following a certain sequence of operations by calculating this weighted sum and then applying an activation function, the information for the neuron in the current layer is calculated. This layer structure of connected neurons resembles the connection of biological neurons via synapses.
With increasing computational power this machine learning technique started outperforming classi-

cal approaches for image processing tasks. A broad range of applications from image classification and pattern recognition to medical image analysis and the analysis of financial time series is significant for this artificial intelligence approach. For a more detailed introduction to the underlying concepts of deep learning and neural networks, we refer the interested reader to [70]. In [1], the author additionally discusses a wide range of applications, e.g., "image captioning [and] image classification". Without going deeper into the different types or application systems, we focus now on how this modern and powerful approach can facilitate our process of fitting a model to the imaging data.
A convolutional neural network (abbreviation: CNN) as for example the U-Net in [64] can be used to extract feature maps using multiple convolutional layers combined with downsampling. In a CNN some layers are convolutional layers, i.e., instead of just calculating a weighted sum, the new information is calculated by applying convolutional operations on the previous layers followed by the activation function again. In the case of the U-Net structure, pooling operations are used to reduce the current size of the current feature map in the first half of the net. The second one includes up-sampling strategies to accomplish the original image size again. This results in the famous "shape" of the U-Net resembling the letter *U* when sketching the different layers and moving down for down-sampling layers and up in the second half for the up-sampling layers. This approach allows the extraction of features at a more sophisticated level. Especially, the extracted features are not straightforwardly perceivable for the human eye when multiple convolutional layers are combined. This is a major contrast to the simple texture features described in Section 3.3.2. We expect that the more advanced features derived with a neural net contain much more information than the simple texture features and, consequently, should have a greater impact on a good model fitting when applying the optimization on the model and the features simultaneously. For this reason, we recommend to investigate the optimization based on neural net features in the future. In this context, we do not take a deeper look into this approach. Instead of focusing on an appropriate network structure to extract features, we will use the simple texture features described in the previous section for our numerical tests.

Before we proceed with the further course of this thesis, we complete this section on texture information with some more references on properties referred to for texture analysis from the literature. In [9] Bharati et al. investigate different texture-based approaches for quality assessment of steel sheets as a direct application in an industrial field. As a statistical measure they use gray level co-occurrence matrices and also focus on multi-step approaches incorporating for example two dimensional Fourier Transform. In any case, they suggest to use methods that incorporate "spatial information". In 2004, they recommended the wavelet texture analysis. We consider that more advanced features based on wavelets can also improve our model fitting.
In the publication of Sharma et al. from 2001, the authors compared different texture descriptors on the basis of the publicly available database "Meastex" [67]. In their tests the co-occurrence matrices and another approach called "Law's method" performed best. However, it is important to state that they did not include an approach based on wavelets in their analysis. Nevertheless, we want to highlight one statement which we consider to be still important 20 years later: They pointed out that the best performance was achieved when they combined "features from all five methods" [67]. With this in mind, we conclude that it is advisable to always consider more than one sole texture

descriptor to generate our feature images.
Next, we refer to a more recent "Study of statistical methods for texture analysis" [63]. The authors compare among others gray level co-occurrence matrices, local binary patterns, autocorrelation functions and histogram patterns for texture classification tasks in different contexts. As it is stated, for different tasks different approaches are more frequently used and for "biomedical analysis" the gray level co-occurence matrix is considered to be "a powerful descriptor of texture" [63]. Moreover, they conclude it to be "the most efficient method to extract texture features for classification [with] a discrimination purpose" [63] which indicates that it is well applicable in our context to differentiate texture of two subcolonies of a cell population. The authors also include more modern approaches which also consider concepts of machine learning as for example the "local spiking pattern" which is based on "a 2-dimensional neural network" [18]. They also evaluate the "SRITCSD method" which uses a support vector machine for classification and "applies the [singular value decomposition] to enhance image textures of an image [and then] extracts the texture features in the DWT domain of the SVD version of the image" [15]. For more details on the compared approaches and the results for some benchmark datasets, we refer the reader to the complete study of Ramola et al. [63].
In this short literature review, we presented different methods used for texture classification and commented on their applicability based on the cited references. We stress that in our context, we are not aiming at texture classification as the main task since our overall objective is to find a model fitting approach to extract spreading information from microscopy data. Still, we point out that the mentioned methods might be superior to our basic texture descriptors used to derive feature images for our upcoming analysis and, consequently, could contribute to fit a spreading model more accurately.

After introducing the feature images that we use to incorporate the original data in the optimization process, we present modeling approaches that capture the development of a cell colony next. Inspired by the moving colony fronts observed in the data, cf. Figure 3.9, we focus on modeling approaches that correspond to similar spreading behaviors. In the next chapter, we introduce a system of partial differential equations that favors related moving wave fronts and also present a simplified version that supports similar circular spreading phenomena.

4

Mathematical modeling for spreading cell colonies

Mathematical modeling is used in a wide range of applications. In the automotive industry for example, it is used for crash test simulations. Material models are implemented to simulate crash test situations and, consequently, support the development of robust and save vehicles for the future. In [76], Miranville and Teman introduce mathematical models for continuum mechanics, for example. Also in financial mathematics and insurance calculations mathematical modeling is crucial to estimate future price developments. For these applications, we refer the interested reader to for example the textbooks dealing with financial models by Lamberton and Lapeyre or by Øksendal (cf. [42, 59]). In [41], Kratka points out the necessity of modeling for risk assessment in the context of insurances. She highlights that for newly arising risks which can affect the world globally a transition is needed away from models "strictly based on historical data" to models that also work with few data. Such global risks are for example "energetic risks", or risks related with changing climate conditions and even consequences of global pandemics as for example the on-going COVID-19 crisis. With this in mind, we want to continue with various models related to spreading phenomena. Such model gained recently huge interest again because of the currently spreading SARS-CoV-2 virus.

A famous spreading model is based on the Fisher-Kolmogoroff equation. This is a special version of a partial differential equation that is often used to capture spatial spreading phenomena which exhibit propagation wave structures [55, 56]. The original model was designed to capture the spreading of a gene within a population while considering spatial interaction in a diffusion term and logistic growth in a reaction term considering the reproduction rate [55].

In chapter 11 in [56], Murray introduces a growth model for brain tumors. The presented model describes the spatial and time dependent tumor cell concentration based on a reaction-diffusion partial differential equation (PDE). While the reaction term models a growth term, the diffusion term captures the spatial mobility of cells. In this context, they apply first a spatially homogeneous brain tissue modeled by a constant diffusion coefficient. In a second step, they account for spatial heterogeneities within the tissue by considering a space-dependent diffusion function to differentiate for example between gray and white matter in the brain.

As we are more interested in modeling *two* different cell concentrations possibly occurring in a cancer cell colony as observed in the given microscopy data, we orientate our model more on a compartment model similar to the ones introduced for epidemic spreading in chapter 13 in [56]. Those spreading models can capture the interaction of different (human) population groups as for example during an

epidemic. While this SIR-model[1] was used in the literature to understand the spread of the Black Death, the same model was used recently to model global and local spreading of the COVID-19 epidemic [48, 56].

Furthermore, we mention the Lotka-Volterra system which is often used to mathematically model interacting species. A popular example is the predator-prey interaction between foxes and rabbits (cf. section 3.1 in [55]), where the population growth of both species is unconditionally linked to the other's. While the fox population can only grow when enough prey (*rabbits*) is present, their dominance consequently results in a lower reproducibility of the rabbit population. Contrary, if fewer rabbits are present, the foxes' population will go down because of the missing prey, and then, consequently, the rabbit population can grow again due to the absence of the dominant predator.

Considering different aspects of the Fisher-Kolmogoroff equation, Lotka-Volterra systems and the SIR model from above again, we focus in Section 4.1 on a model based on two different "population" groups to analyze the spreading of a cell colony. In our setting, we aim to capture the spreading of "normal" cells and "abnormal" cells with similar traveling wavefront speeds. Those two groups are related to the different texture regions we observe in our microscopy data (cf. Section 3.3). While we consider some basic spreading properties for the two different subpopulations in our PDE model, we want to refer to other models introduced in the literature and which focus explicitly on cell spreading. Gerisch and Chaplain present in [29] for example a continuum model for cancer cell invasion of tissue and consider an *in vivo* setting which is contrary to our *in vitro* experimental data of AstraZeneca. Still, we see aspects which could be important to improve our basic PDE model. For example, they consider cell-cell adhesion effects. Their terms for cell migration and cell proliferation could also be helpful to define diffusion and reaction terms. Nevertheless, we do not recommend to use their model in our context. While they consider two more concentrations concerning the extracellular matrix and one related to a specific enzyme next to the cancer cell concentration, we aim for a model capturing *two* different cell concentrations which both relate to the total cell concentration. Of course, we could additionally consider another concentration representing the wells' liquid media with its ingredients and nutrients for the cells. However, this is not the main focus of our work and goes beyond the scope of this thesis.

In [34], the authors use a system of PDEs to model the interaction and concurrent spreading of cancer cells and other native cell types. They investigate the spreading of fibroblast cells and melanoma cells in a first experiment separately before evaluating the spreading phenomena in a co-culture. Based on the individual experiments, they derive diffusivity coefficients and proliferation rates for each of the used cell tpyes with the help of cell density histograms, estimating the leading front of a spreading culture with ImageJ's Sobel filter and the population's area. Based on the spreading parameters for the individual cultures, they numerically simulate the spreading in the co-culture and compare it to the experiments. We stress that the main difference to our approach is that we aim for numerically deriving the spreading parameters for our two cell types directly from the image data. Moreover, we do not study the competition between two different cell types but are interested in a transition model that captures the change of *normal* to *abnormal* cell appearances. Still, we point out another interesting consideration they implement in their approach. Based on the assumption of radial symmetric spreading, they define the cell densities depending on the time point and only the

[1] S=susceptible, I=infective, R=recovered people

radial position. If only one cell type is present, the system is simplified to a single Fisher-Kolmogoroff equation in radial coordinates.
In [79], Treloar et al. investiagte the sensitivity of spreading properties like the diffusivity and the proliferation rates depending on special assay geometries. In the previously cited paper [34], the authors used circular barrier assays and also the wells recorded in our experiments have a certain shape. As Treloar et al. found out that the *in vitro* assay shapes indeed may influence the spreading characteristics and our well geometry is not perfectly circular shaped anyway, we conclude our approach with a PDE system based on two dimensional spatial coordinates to be more appropriate for our investigations than the system of PDEs introduced in [34] and depending on a radial coordinate.
In [35], the authors use simulations based on different models for preliminary tumor growth and metastatic spreading and compare it with the spreading observed in mice experiments. For the spreading phenomenon they use a PDE model to test the different approaches to model growth and spreading effects. While they rather recommend three different growth models, we use in our model a logistic growth term for simplification aspects. We refer the interested reader to the original literature for the three suggested models of Bertalanffy, Gomperts and West referenced by Hartung et al. in [35]. The approach of Bertalanffy was initially described in [8]. The Gompertz model can be checked out in the book of Wheldon on "Mathematical models in cancer research" [83]. For further information on the model of West, we refer to [33] and [82].

After this review on spreading models for cancer cells in the literature, we establish a system of partial differential equations to model and approximate the colony development in our imaging data in the following Section 4.1. Based on this, we further introduce a reduced approach based on concentric circles in Section 4.2. In the final Section 4.3, we briefly elaborate on different ways to fit a model to given data.

4.1 A PDE model for colony spreading

We are aiming for an optimization approach that we can use to derive spreading parameters describing the colony growth of a tumor cell colony directly from the given imaging data. In this section we develop a system of partial differential equations to model a spreading phenomenon that satisfies certain assumptions. We start with elaborating on various conditions that we want to incorporate into our model in the following list of assumptions.

Assumption 4.1
Considering the biological background of the given experiment focusing on the development of a growing cell colony, we state certain assumptions and requirements that a model of partial differential equations should meet.

1. We consider two different cell groups for normal and abnormal cells. The associated cell concentrations are denoted by c_n and c_a, respectively, and are living on the spatio-temporal

domain denoted by $\Omega_T = \Omega \times [0, T]$. We use $\left[\frac{1}{(\text{space})^2}\right]$ as the associated unit of these cell concentrations.

2. There exists a maximal number of cells that can be in one location. More precisely, we assume an upper bound for the total cell concentration consisting of both cell groups. This maximal cell concentration limit is the carrying capacity denoted by K in the unit of $\left[\frac{1}{(\text{space})^2}\right]$.

3. Cells within the normal group are regarded as mobile cells, i.e., we assume that these cells can migrate, which is reflected in a diffusion term for the normal cell concentration c_n. The diffusion is assumed to be space and time independent and, consequently, it is reflected in a diffusion constant denoted by D here. Its unit is given as $\left[\frac{(\text{space})^2}{\text{time}}\right]$.

4. The cell proliferation, i.e., the reproduction of cells is incorporated in a logistic growth term for the normal cells which includes the relative total cell population. In this relative total concentration, we consider normal and abnormal cells added together and compared to the limiting carrying capacity. The reproductivity is reflected in the reproduction rate or proliferation rate denoted by r and of unit $\left[\frac{1}{\text{time}}\right]$.

5. Depending on the relative total cell concentration consisting of both the normal and the abnormal cell populations, we implement a reaction term that captures the transition from normal cells to abnormal cells. The rate for this transition is labeled with m and reflects a certain "mortality" of the normal cells. The unit of this rate is given as $\left[\frac{1}{\text{time}}\right]$ in line with the previous proliferation rate.

6. The abnormal cells do not proliferate and the only growth term for this cell concentration is derived from the state transition term of the normal cells. The abnormal cell population develops from corrupted or "dying" normal cells.

7. We consider the abnormal cells to be a kind of immobile bulk of cells. They do not move around by themselves. The only movement they could perform arises from displacements of neighboring and migrating normal cells.

In this list we refer to certain characteristics that we relate to the cell spreading phenomena we are facing in the microscopy images.. Moreover, some properties are typical for well-known population models that were already mentioned in the introductory text of this chapter. Based on these arguments we define our PDE model as follows:

Definition 4.2 (System of PDEs for cell colony development)
By the following system of partial differential equations, we define the growth phenomenon for a cell colony fulfilling the requirements and assumptions in Assumption 4.1. It models the interaction between two different cell types for normal cells (c_n) and abnormal cells c_a:

$$c_{n,t} = rc_n\left(1 - \frac{c_n + c_a}{K}\right) - mc_n\frac{c_n + c_a}{K} + Dc_{n,xx}$$
$$c_{a,t} = mc_n\left(\frac{c_n + c_a}{K}\right).$$

Both concentrations live on the spatio-temporal domain Ω_T (cf. Definition 3.1), are non-negative and are bounded from above by the carrying capacity K such that

$$c_{n,t},\ c_{a,t} : \Omega_T \to [0, K]$$

holds. The temporal derivative of the given concentration is denoted by $c_{\cdot,t}$ and the second partial derivatives in the spatial dimension are given as $c_{\cdot,xx}$.

The equation for the concentration of normal cells incorporates a logistic growth term ($rc_n\left(1 - \frac{c_n+c_a}{K}\right)$) and also a diffusion term ($Dc_{n,xx}$) to account for the migration process. When we skip the last missing term, namely the reaction term modeling the state transition from normal to abnormal cells, the equation for the normal cells resembles the famous Fisher-Kolmogoroff equation (cf. equation (11.17) in [55]). It is a well known result that the Fisher-Kolmogoroff equation exhibits a traveling wave solution in the one dimensional case with a constant wave speed.
In two dimensions, we assume a constant traveling wave speed in the limiting case for this system as well if we consider an unbounded, infinite domain. Without proving this rigorously, we present a possible line of arguments to show this assumption. We point out that far away from the colony's origin the curvature of the contour lines decreases as sketched in Figure 4.1. If we focus on a very small piece of the wave front at such a location very far away from the origin, e.g., in the red frame, the traveling wave approaches a two dimensional wave front which is almost constant in one direction. With green arrows, we indicate the spreading directions. In the red frame the wavefront moves locally constantly, i.e., with a constant wave front profile in the vertical direction, which is highlighted with the two parallel spreading arrows. If the wavefront converged to be locally constant in one direction, we could derive from the one dimensional case that the wave speed converged locally to a constant value in this limiting case, too.
Of course, our domain is far from being unbounded. Moreover, Markham et al. show that there exists indeed a transient spreading velocity before it converges to a constant one [50]. In their case, they even focus only on a one dimensional example. Although they do not use a system of PDEs but different strategies to approximate cell occupancies on a one dimensional grid, we still consider that in our context the spreading velocity is transient as well before converging to constancy. Still, we do not want to derive this transient velocities more thoroughly and, instead, assume a constant spreading speed for simplification aspects.
Coming back to our setting, we are facing two main differences compared to the simpler 1D Fisher-Kolmogoroff model. Firstly, we are interested in concentrations living on a two dimensional spatial

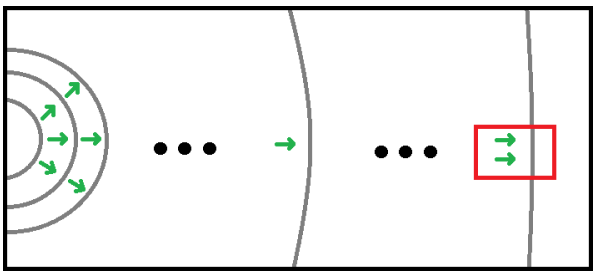


Figure 4.1: Sketch of a traveling wave front in two dimension in an unbounded domain.

domain instead of a one dimensional one. Despite the previous argumentation, this two dimensional domain is considered to be bounded since it resembles the well domain in the experiments. Secondly, we are not only concentrating on *one* cell concentration, but investigate an interaction model between two different cell species. While for the historic Fisher-Komogoroff model a traveling wave solution exists which joins the two stable state solutions of no concentration and a totally occupied domain, we consider in our context three stable state solutions. One stable state is assumed to match an empty domain again, the second stable state considers the domain to be occupied by normal cells and the last one reflects the domain inhabited completely by abnormal cells.

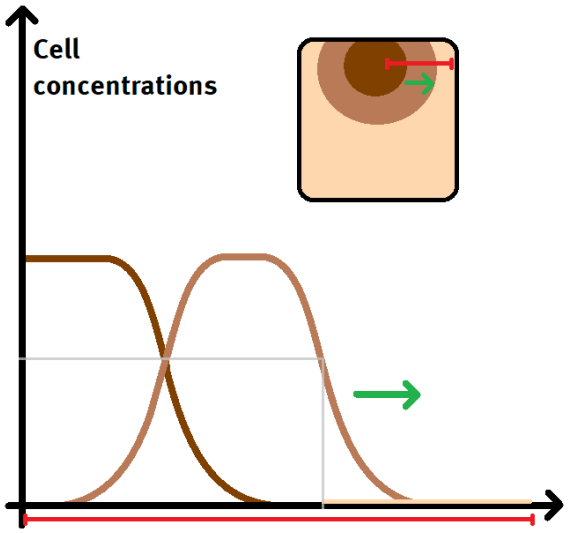


Figure 4.2: Sketch of cell concentrations and moving colony fronts for a one dimensional cut in a two-dimensional domain.

In Figure 4.2, we sketch such a traveling wavefront when considering a one dimensional cut. This cut is marked in red in the upper right corner of the two dimensional colony preview and also below the x-axis of the plot. Here, the light brown color corresponds to the normal cells whereas the dark brown area marks the region with abnormal cells. The background and consequently empty area is shown in beige. We present in the two dimensional preview a kind of "classification image" that divides the image domain into different areas which are associated with the background region, the normal cell population or abnormals cells. Instead of including transition zones between the subcolonies or subclasses, we only mark the "dominant" groups, i.e., if the cell concentrations exceed a certain threshold we classify the region to belong to the cell popular with maximum concentration

in that location. This threshold is marked with a horizontal gray line in the cross-section plot. If both cell concentrations are below this threshold value, we identify the corresponding region as background area. In the cross-section plot, the corresponding part is on the right of the vertical gray line and marked with the beige curve *on* the x-axis corresponding to *no cells* and, consequently, a "concentration" value equaling 0.
In light brown we observe a traveling wave front symbolizing a ring area in two dimensions for the normal cells. The traveling direction is marked with a green arrow parallel to the red cutting line in both plots. The second wave front represents the moving front of the abnormal cell colony. Here, we include the "interaction" phenomenon that a rising abnormal cell concentration corresponds directly with the falling concentration of normal cells since they are changing their state. Lastly, we point out that the plateau regions for both wave curves correspond to the carrying capacity and this value cannot be exceeded.
Based on the literature, we stress that such wave-like patterns as shown in Figure 4.2 are possible also in two dimensions and for interaction models (cf. chapter 1 in [56]). A thorough stability analysis of the given model is complex and not the focus of this thesis. Instead of performing a deeper analysis of solutions for the given model, we perform a model reduction next. The simplified model should capture the basic spreading properties we relate with the above system of PDEs, namely a constant traveling wave speed and two wavefronts where the second one for abnormal cells follows the first one for normal cells.
The main interest of this thesis is deriving properties for a spreading model based on maximizing mutual information between model characteristics and the imaging data. Therefore, as a proof of concept, it is ideal to focus on a simplified model first. For more advanced future studies, we suggest to look into integrating a PDE model into the optimization problem to extract properties like reproduction rate, transition rate, diffusion coefficient and a carrying capacity directly and foster the understanding of the underlying biological phenomenon for the spreading cell colonies even more.

4.2 Concentric circles for colony spreading

Instead of applying the PDE spreading model introduced in the Section 4.1, we present a simplified model where we use an approximation by applying circular spreading. In Figure 4.3, a simplified

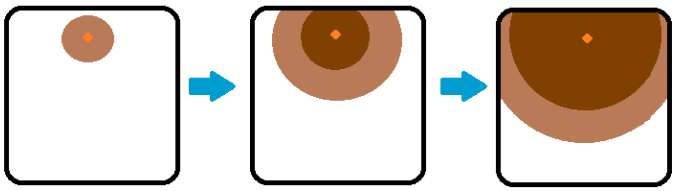


Figure 4.3: Sketch of simplified concentric spreading over time of a cell colony in a well domain.

concentric spreading phenomenon is sketched for a cell colony in a well domain. Consisting of two different cell groups marked in light and dark brown, two colony areas are observable. Both cell groups start their spreading process initially at the same origin marked with an orange diamond. We assume to have a constant spreading speed v. This is a valid assumption, remembering that the PDE

model results in a constant traveling wave speed in the one dimensional case and is expected to show a similar behavior in two dimensions when considering an infinite domain. Of course, we are never facing unbounded domains in the real world experimental settings. However, considering a constant wave speed serves here as one important property for the reduced model.
We motivate this simplification by stating that in mathematical modeling we often need to make some simplifications to develop a model that both captures the real world scenario appropriately and results in a solvable model we can approximate with numerical methods. While solving the PDE model is also numerically possible, we prefer here the simplified model as this is more straightforward to implement in a numerical solution. Overall, we are focusing in this thesis mainly on the concept of extracting spreading properties based on maximizing the mutual information between the original data, or to be more precise the texture data, and an applicable spreading model. For a proof of concept, the simplified model serves well to check the convergence in a numerical solution of the related optimization problem (cf. Section 5.5.1).
Based on the model for colony spreading, we want to generate a classification image I_2 consisting of a labeling according to

$$I_2 \approx \begin{cases} 0 : \text{outside of the cell colony,} \\ 1 : \text{within the normal cell regions,} \\ 2 : \text{within the abnormal cell regions,} \end{cases} \tag{4.1}$$

that lives in our spatio-temporal domain Ω_T defined in Definition 3.1. We note that similar to the real data set we assume the time domain to be discretized resulting in n_T discrete time points

$$T^{\Delta} = \{t_1, \cdots, t_{n_T}\} \text{ with } t_1 = 0,\ t_{n_T} = T \tag{4.2}$$

and temporal time steps

$$\Delta t_i = t_{i+1} - t_i, \quad i = 1, \cdots, n_T - 1.$$

The superscript h in the set of time points T^{Δ} is used to indicate the *discretization* and to distinguish it from the end time point denoted with T. In the given data set, we are dealing with $n_T = 8$ discrete time points which are not equidistantly distributed in the time interval. Consequently, the temporal time steps Δt_i are varying for $i = 1, \cdots, n_T - 1$. In the upcoming definition, we collect the the spreading properties and define their underlying space.

Definition 4.3 (Parameter space $\boldsymbol{P}$)
We define the parameter space

$$\boldsymbol{P} = [0, L] \times [0, W] \times \left\{ \left(\hat{t}_{0,n}, \hat{t}_{0,a} \right) \in [0, T]^2 \,|\, \hat{t}_{0,n} \leq \hat{t}_{0,a} \right\} \times [0, v_{\max}]. \tag{4.3}$$

The spreading properties $\boldsymbol{p} = (\boldsymbol{x}_0, t_{0,n}, t_{0,a}, v) \in \boldsymbol{P}$ capture the variables defining the spreading process with $\boldsymbol{x}_0$ denoting the origin of the cell colony, $t_{0,n}$ and $t_{0,a}$ denoting the starting time points for the two different wave fronts for the population region consisting of normal cells (subscript n) and the abnormal colony division(subscript a) and the constant traveling wave speed given by v.

We point out, that the parameter space is designed in such a way that the following assumptions hold:

(i) The colony's origin is within the spatial domain: $\boldsymbol{x}_0 \in \Omega$;

(ii) The colony fronts can only start in the valid time interval: $\hat{t}_{0,n}, \hat{t}_{0,a} \in [0, T]$;

(iii) The colony front for the *abnormal* cells cannot start before the front for the *normal* cells starts to propagate;

(iv) The traveling wavefront speed is limited by a maximal speed of $v_{\max} = \frac{\max\{L,W\}}{\max_i \Delta t_i}$;

(v) For physical reasoning the wave speed cannot be negative.

All those assumptions ensure that a circular spreading colony is always observable within the given spatio-temporal domain Ω_T. Moreover, the restriction for the spreading speed guarantees that a new colony cannot grow that fast that we cannot capture its development with the given temporal discretization, i.e., we follow the assumption that if a colony starts to grow in one time frame, it cannot grow so fast that in the consecutive time frame the total well area is covered by cells already. In the given data set, the two initial time frames are recorded very close to each other — the second one serving as a control imaging to make sure that the set up is correct. The biologists select the different time stamps for recording based on their knowledge that the cells in focus would not grow populations that fast.

The classification image I_2 is calculated with the following circle equation based on a given parameter setting $\boldsymbol{p}$,

$$\begin{aligned} &k_j : \boldsymbol{P} \times \Omega_T \to [0,1], \\ &k_j(\boldsymbol{p}, \boldsymbol{x}, t) = v(t - t_{0,j}) - \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \qquad j = n, a. \end{aligned} \tag{4.4}$$

combined with a smoothed Heaviside function. We remark that $\|\cdot\|_2$ is used for the Euclidean norm. Based on this circle equation, we define that for both colony regions of normal and abnormal cells the circle equations for $j = n, a$ returns for a fixed $\boldsymbol{p} \in \boldsymbol{P}$

$$k_j(\boldsymbol{p}, \boldsymbol{x}, t) \begin{cases} > 0 : \text{if } (\boldsymbol{x}, t) \text{ is in the subcolony;} \\ < 0 : \text{if } (\boldsymbol{x}, t) \text{ is outside of the subcolony;} \\ = 0 : \text{if } (\boldsymbol{x}, t) \text{ is on the subcolony's moving front.} \end{cases}$$

The discontinuous Heaviside function

$$\begin{aligned} &H : \mathbb{R} \to \{0, 1\} \\ &H(x) = \begin{cases} 0 : x < 0 \\ 1 : x \geq 0 \end{cases} \end{aligned}$$

is approximated by the continuous function

$$\begin{aligned} \sigma &: \mathbb{R} \to [0,1] \\ \sigma(x) &= \frac{1}{1+\exp\left(-\frac{x}{\varepsilon_0}\right)}. \end{aligned} \tag{4.5}$$

If $\varepsilon_0 \to 0$ holds, we achieve a better approximation of the jump function. In Figure 4.4, a plot of the continuous function σ for $\varepsilon_0 = 1, 0.5, 0.25$ visualizes this convergence to the Heaviside function for ε_0 converging to 0.

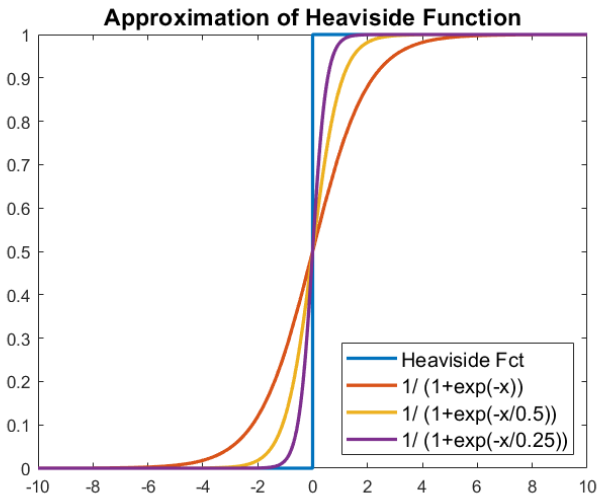


Figure 4.4: Approximation of the Heaviside Function.

We define the classification image I_2 as a continuous function for any $t \in \{t_1, \dots, t_{n_T}\}$ with the help of the circle equations in Equation (4.4) and the smooth approximation of the Heaviside function in Equation (4.5).

Definition 4.4 (Classification image I_2)
For a fixed model parameter $\varepsilon_0 > 0$, the classification image I_2 is given as a continuous function in $\boldsymbol{p} \in \boldsymbol{P}$ and $\boldsymbol{x} \in \Omega$ by

$$I_2 : \boldsymbol{P} \times \Omega_T \to \boldsymbol{\mathcal{C}}, \tag{4.6}$$

$$I_2(\boldsymbol{p}, \boldsymbol{x}, t) = \frac{1}{1+\exp\left(-\frac{1}{\varepsilon_0} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)} + \frac{1}{1+\exp\left(-\frac{1}{\varepsilon_0} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)} \tag{4.7}$$

by using the circle equation from Equation (4.4) and with the classification space given as $\boldsymbol{\mathcal{C}} = \mathbb{R}$.

With this given model and the smooth approximation of the Heaviside function in the each summand, we ensure that the classification image I_2 actually maps to the $[0,2]$ and, consequently, we can define a reduced classification space as follows:

Definition 4.5 (Reduced classification space $\mathcal{C}'$)
We define the reduced classification space $\mathcal{C}' \subset [0,2]$ as the image of the mapping I_2 which depends on the model parameter $\varepsilon_0 > 0$, i.e. the steepness of the approximated Heaviside functions.

Depending on the given parameter set $\boldsymbol{p} \in \boldsymbol{P}$ the function I_2 returns for each discrete time point $t \in \{t_1, \dots, t_{n_T}\}$ a smooth labeling for the different colony regions such that Equation (4.1) holds. *Smooth* means here that the mapping is continuous and continuously differentiable on the space $\boldsymbol{P} \times \Omega$ for all $t \in \{t_1, \dots, t_{n_T}\}$ which holds true due to the construction of I_2. Moreover, we point out that depending on ε_0 the widths of the transition regions between the different subcolonies and the background regions vary. For larger ε_0, the transition areas are more smeared out compared to small ε_0. For $\varepsilon_0 \to 0$, the continuous classification image approaches discontinuous jumps between the different classification areas as claimed in Equation (4.1) more closely.
Without going into detail on the choice of ε_0 here, we stress that we use it as a pre-set and inherent model parameter. As a natural extension, we could include the selection of an appropriate parameter ε_0 for scaling transition regions into our model. To allow transition areas of different widths between background and normal cell colony compared to normal cells versus abnormal cells, it could be beneficial to use two different ε_0 for the respective areas, i.e., in the different summands for calculating I_2 via Equation (4.6). This can be especially favorable as we cannot assume a priori that both regions are of similar width for sure. However, we will restrict ourselves to a pre-selected ε_0 for our analysis in Chapter 5 and leave improved and automatically selected transition parameters for future studies.

For later references in sections on optimization and discretization effects in Chapter 5, we elaborate here on the partial derivatives for the classification image with respect to the parameter set $\boldsymbol{p}$. As a composition of different subfunctions we can denote the gradient of the classification image by using the gradients of the circle equations $\nabla_{\boldsymbol{p}} k_n$ and $\nabla_{\boldsymbol{p}} k_a$ and derive

$$\begin{aligned}
&\nabla_{\boldsymbol{p}} I_2(\boldsymbol{p}, \boldsymbol{x}, t) \\
=&\frac{-\exp\left(-\frac{1}{\varepsilon_0} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)\left(-\frac{1}{\varepsilon_0}\nabla_{\boldsymbol{p}} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)}{\left(1+\exp\left(-\frac{1}{\varepsilon_0} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)\right)^2} + \frac{-\exp\left(-\frac{1}{\varepsilon_0} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)\left(-\frac{1}{\varepsilon_0}\nabla_{\boldsymbol{p}} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)}{\left(1+\exp\left(-\frac{1}{\varepsilon_0} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)\right)^2} \\
=&\frac{1}{\varepsilon_0}\left(\frac{\exp\left(-\frac{1}{\varepsilon_0} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)\left(\nabla_{\boldsymbol{p}} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)}{\left(1+\exp\left(-\frac{1}{\varepsilon_0} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)\right)^2} + \frac{\exp\left(-\frac{1}{\varepsilon_0} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)\left(\nabla_{\boldsymbol{p}} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)}{\left(1+\exp\left(-\frac{1}{\varepsilon_0} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)\right)^2}\right).
\end{aligned} \tag{4.8}$$

The gradient of the classification image is of importance later when focusing on optimization used to fit our colony model to the given data. Based on partial derivatives, we derive the gradients for the circle equation for the normal cells' front k_n as

$$\nabla_{\boldsymbol{p}} k_n(\boldsymbol{p}, \boldsymbol{x}, t) = \begin{pmatrix} \frac{\partial k_n(\boldsymbol{p},\boldsymbol{x},t)}{\partial x_{0,1}} \\ \frac{\partial k_n(\boldsymbol{p},\boldsymbol{x},t)}{\partial x_{0,2}} \\ \frac{\partial k_n(\boldsymbol{p},\boldsymbol{x},t)}{\partial t_{0,n}} \\ \frac{\partial k_n(\boldsymbol{p},\boldsymbol{x},t)}{\partial t_{0,a}} \\ \frac{\partial k_n(\boldsymbol{p},\boldsymbol{x},t)}{\partial v} \end{pmatrix} = \begin{pmatrix} -\left(\|\boldsymbol{x}-\boldsymbol{x_0}\|_2\right)^{-1}(x_{0,1}-x_1) \\ -\left(\|\boldsymbol{x}-\boldsymbol{x_0}\|_2\right)^{-1}(x_{0,2}-x_2) \\ -v \\ 0 \\ t-t_{0,n} \end{pmatrix} \tag{4.9}$$

and for the abnormal colony's front k_a as

$$\nabla_{\boldsymbol{p}} k_a\left(\boldsymbol{p}, \boldsymbol{x}, t\right) = \begin{pmatrix} \frac{\partial k_a(\boldsymbol{p},\boldsymbol{x},t)}{\partial x_{0,1}} \\ \frac{\partial k_a(\boldsymbol{p},\boldsymbol{x},t)}{\partial x_{0,2}} \\ \frac{\partial k_a(\boldsymbol{p},\boldsymbol{x},t)}{\partial t_{0,n}} \\ \frac{\partial k_a(\boldsymbol{p},\boldsymbol{x},t)}{\partial t_{0,a}} \\ \frac{\partial k_a(\boldsymbol{p},\boldsymbol{x},t)}{\partial v} \end{pmatrix} = \begin{pmatrix} -\left(\|\boldsymbol{x}-\boldsymbol{x}_0\|_2\right)^{-1}\left(x_{0,1}-x_1\right) \\ -\left(\|\boldsymbol{x}-\boldsymbol{x}_0\|_2\right)^{-1}\left(x_{0,2}-x_2\right) \\ 0 \\ -v \\ t-t_{0,a} \end{pmatrix}. \tag{4.10}$$

After having introduced two models for cell colony development — one based on a system of partial differential equations and a reduced version of it preserving main properties of the PDE model, we want to bring the model and the data into one context next. For this purpose, we use again the extracted feature image I_1^{d} introduced in Section 3.3. Before we dive into our approach based on model fitting via Mutual Information Maximization in Chapter 5, we first reflect on other approaches used to fit a mathematical model to given data in the next subsection.

4.3 Review of model fitting approaches

When we want to fit a mathematical model to real world data, one commonly used approach is to apply manual estimations and iterative calibration steps. In [34], the authors estimate diffusion and reaction coefficients in the used PDE model to capture cell spreading by introducing certain assumptions for simplifications. In the end, the authors estimate the occurring diffusivity of the cells and their proliferation rates by investigating subregions in which they consider the assumptions to be true.

Xun et al. also deal with parameter estimation for PDEs in [84] "to model complex dynamic systems in applied sciences such as biology" in their article from 2013. With "a parameter cascading method and a Bayesian approach" they aim to reduce the numerically load which was due to multiple evaluations of the PDE system under various "candidate parameter" estimates [84]. For a more detailed introduction to parameter estimation, we refer the reader to chapter 6 in [65].

Instead of deriving parameters based on "empirical observations", Long et al. suggest a "data-driven" approach in [45]. Based on the recently emerging neural networks, they introduce their "PDE-Net" which can *learn* the underlying PDEs of complex dynamic systems. Based on their approach, the authors are able to predict the dynamics for a "relatively long time".

With "PESTO", Stapor et al. provide a "Parameter EStimation TOolbox" that is applicable for many different applications, e.g., for fitting models used in computational biology [71]. The MATLAB toolbox can be flexibly used for parameter estimation of PDEs and ordinary differential equations (abbreviations: ODEs). Moreover, the authors stress that they incorporated modern optimization concepts such as multi-start techniques and "automated starting point selection" [71].

Instead of estimating the underlying (PDE) model with the aforementioned approaches, we suggest another multi-step concept. Firstly, we consider to have the real data in form of feature images at hand. The idea is now to simplify the multidimensional feature-data to some classification images, determining for example normal colony regions, abnormal regions and background regions. For this

puprose, one could imagine a thresholding approach only acting on one of the features, e.g., on the third feature based on local interquartile ranges which could reveal already different occurrences in the imaging domain. Based on the thresholded values one can consequently identify different "classes" to generate a classification image. By considering that this image consists only of integer values for the "main classes" encoding background area, normal cell colony area and abnormal cell colony area we can also interpret it as a segmentation mask. We refer to Figure 3.16b) which depicts an example of such a classification image derived from the thresholded interquartile ranges.
Secondly, we consider a classification image based on the contemplated model similar to the classification image introduced in Section 4.2. Of course, this is depending on specific model parameters which we want to optimize in the end again. Then, we aim for minimizing the *distance* between these two versions of the classification images by optimizing the underlying parameter settings. In this sense, we of course need a distance measure.
In [38], Jiang et al. review various criteria for evaluating and comparing different segmentation results. In our case, our model-based classification images can easily be transferred to segmentation masks revealing the three main classes for no cells, normal cells and abnormal cells with values in the classification space close to $0, 1$ or 2, respectively. This can be achieved by applying thresholding. The investigated measures in [38] are derived from the context of comparing clustering results. The idea originates from the interpretation of the segmentation task "as one of data clustering" [38]. In this paper, the authors present different distance measures of clusterings related to "counting pairs", "set matching" and "Information-theoretic distances of clusterings". For the latter category, they deal with the concept of mutual information and present a specific version to compare clusterings. They also cite the normalized version of mutual information for clusterings from [72]. We point out this MI-related measure explicitly since we investigate the concept of mutual information more closely for our setting of model fitting in Chapter 5. Coming back to [38], the authors state in their conclusion that the measures "may be biased in certain situations". Therefore, they suggest to define a combination of measures from the different categories aiming for bias reduction. Instead of discussing the various measures more thoroughly here, we refer the reader to the original paper. While we propose above to use thresholding applied to the feature images based on the local interquartile ranges to derive classification images based on the feature data, we focus on another approach to generate these classification images next.
An alternative approach to derive such classification images from the original data or from the feature images is based on graph theory. We consider the concept of "point cloud sparsification via Cut-Pursuit" (cf. chapter 8 in [25] or [77]). Via efficient graph cut algorithms, the authors produce a sparsification of a given point cloud by following a "coarse-to-fine optimization scheme". For more details on the implemented cutting strategy for graph edges, we refer to the introduction of "Cut Pursuit" by Landrieu and Obozinski [43]. In this sense, we point out that the application of "Cut Pursuit" is not limited to point clouds but can be used for any data for which we can find a graph representation. As an example, we refer to images living on pixel grids for which can generate a graph representation by introducing nodes for each pixel and edge connections between neighboring pixels. In our context, we could generate a finite weighted graph for the point cloud consisting of the three-dimensional feature vectors of the pixels in our image domain. In this sense, we also interpret each pixel as a node in the graph and connect these with other pixels by defining certain relational

measures. In close cooperation with Fjedor Gaede, we tested this approach for one example time frame of a growing colony by applying a software framework he developed for his contributions in [25, 77]. The generated edge connections for the graph are based on *local* and *non-local* information concerning the pixels in the image domain. The idea is to include information on *neighboring* pixels in the image domain (*local information*) and also favor feature vectors which are *close* to each other in the feature space but not necessarily in the image domain (*non-local information*). Based on this information, we generate edges and their weighting terms. The final result of our test case for the feature data of the final time point for well B4 on plate 1 is shown in Figure 4.5. In this case, we sparsified a point cloud consisting of 351×351 feature vectors for a downsampled image domain with cropped out lower and right blank areas outside of the well. The returned state of the graph cut approach reveals 288 final nodes in the sparsified graph. In Figure 4.5, the pixels which are represented by the same node in the reduced graph are colored identically. We observe that the classification based on the final node numbers indeed classifies similar texture regions to the same color (cf. Figure 3.1d) while also including spatial, local information. This results in a "smoothed" mask with very few sparse points which seem to be "mis-classified" (cf. darker blue spots within other colored regions). This could be improved by adjusting certain regularization parameters in the optimization functional or by changing the weighting functions used to incorporate the local and non-local information. However, we do not want to go too deep into the applied concepts and techniques of this approach. For further information, we refer the interested reader to the cited works [25, 43, 77]. Eventually, we would assume to get appropriate classification images consisting only of three different classes if we were sparsifying the point cloud further — and actually quite extremely to get only three final nodes in the representative graph. Alternatively, one could use the graph based approach for data reduction and apply a subsequent technique to generate the segmentation into three main classes. We refer to [28] for further information on this graph reduction to a "super-pixel graph" for simplifying and accelerating following partitioning tasks such as for our classification problem.

As the result of our graph based feature point cloud sparsification shown in the image domain in Figure 4.5 resembles a result for a semantic segmentation problem, we want to refer additionally to [44]. In this article, the authors propose a semantic segmentation approach based on deep neural nets when considering point cloud data. In our context, we are facing a point cloud in the feature space consisting of the feature vectors for the pixels of our (possibly downsampled) domain whereas Landrieu and Simonovsky inspect LiDAR data [44]. We stress that approaches like this could also help to get representative segmentation masks — or classification images — for our feature data.

As we already stated above, determining good segmentation masks is not enough for model fitting. We suggested to use two versions of the classifications images — one depending on the feature data and one depending on the model and the inherent parameters — and then calculating the distance between the images with a pre-set measure. Then, we could apply an optimization approach to minimize this *error*, i.e., the distance, between the predicted classification image based on the model and the other one derived from the feature images. This approach consequently consists of several different substeps.

To improve this, we consider in the next chapter an optimization approach working directly with the given feature data and the derived classification images in the previous Section 4.2. Inspired by mutual information based image registration approaches as presented in [61], we elaborate in the

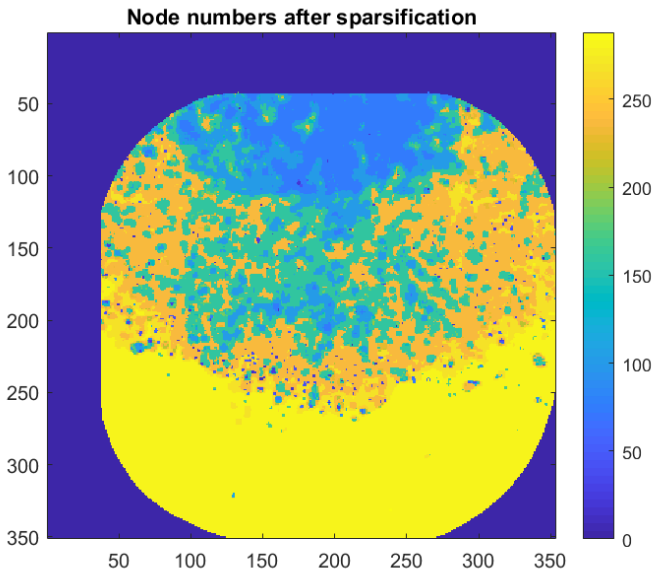


Figure 4.5: Final node numbers after point cloud sparsification with the Cut Pursuit algorithm for the final time point of well B4 on plate 1 with a growing cell colony.

up-coming Chapter 5 on how to implement and exploit this concept in our setting. In this sense, we pursue our ultimate goal of finding good model parameters for classification images based on the reduced model to approximate the cell colony spreading observed in microscopy data and incorporated in the feature images.

5

Mutual information based model fitting

The concept of mutual information is used here to "overlay" different kinds of information and "align" them in an optimal way such that the information gain considering both inputs jointly is maximal. As this description of mutual information is a bit figurative, we start with an illustration based on the registration problem for medical images in Section 5.1. Hereafter, we will translate the concept of mutual information back to our setting and derive the main optimization problem we are dealing with in this thesis in Section 5.2. In our context, we want to match classification images based on a spreading model with information extracted from microscopy images capturing growing cell colonies. For this purpose, we focus in subsections on histogram measures, derivative terms related to theses histograms and the gradient for discretized mutual information.
As we are using recorded imaging data and aim for a numerical solution of this optimization approach, we need to consider various discretization steps which are the main topic of Section 5.3 together with related convergence results. Before we concentrate on the numerical analysis, we focus on a convergence analysis for this discretized problem and the expected, continuous model in Section 5.4. This is crucial to make sure that our numerical results truly relates to an optimum in the continuous setting. For the numerical tests in Section 5.5, we introduce a toy example before applying the optimization problem on data from AstraZeneca.

5.1 Mutual information - a brief introduction

We start with a brief motivation of image registration for medical images based on [11], before defining mutual information (abbrev. MI) explicitly. After that we come back to the cell colony development and transfer the application of mutual information into our setting.
Let us consider two different kinds of medical imaging modalities, for example images recorded with computed tomography (CT) and images based on magnetic resonance imaging (MRI). Without going into detail, it is known that both imaging modalities have their advantages - and also their drawbacks. With computed tomography, we can capture solid structures like bones within the human body. On the contrary, soft tissues cannot be visualized adequately as x-rays pass through easily. On the other hand magnetic resonance tomography cannot capture bones as well as a CT, whereas blood vessels and soft tissues are distinctively shown in MRI data. For certain medical assessments, diagnosis or surgery planning it is important to get as much information about the body parts in focus as possible. For this purpose, image registration is one of the most important tasks in medical image analysis. As

the positioning and orientation of the body is not necessarily completely identical when applying two different imaging techniques, it is important to get an adjusted view later on in which both images are overlaid and deformed in such a way that the same body parts are depicted in the same image area.
Another example when registration is essential is when tissues or structures are imaged repeatedly with some time delay, e.g., before and after a surgery. In this case, we cannot assume that the body is positioned identically either in the imaging device - and also the imaged part might appear slightly different due to healing processes or courses of disease. To account for those misalignments, a registered version of the given images can be computationally derived.
One approach for image registration is based on the concept of mutual information. Based on a measure for information, we are aiming for *reducing* "the uncertainty [of a] possible outcome", *increasing* "the information gain" given the input data and a *reduction* of the "dispersion of [the present] probability distribution" [61] when maximizing the mutual information. We add some more content to this rather formal description for mutual information and start with a definition of the Shannon entropy as a measure of information (cf. [61, 66]).

Definition 5.1 (Shannon entropy)
Given the events $e_1, \dots, e_m$ occurring with probabilities $p_1, \dots, p_m$, we define the *Shannon entropy* as

$$H = \sum_{i=1}^{m} p_i \log\left(\frac{1}{p_i}\right) = -\sum_{i=1}^{m} p_i \log(p_i).$$

This measure directly weights the information content by the probability for each event individually. Moreover, this information measure accounts for the different probabilities with the logarithmic term. When an event has a very low probability, i.e., it is a very rare event, the information gain is high, in particular it is higher than for an event with a very high probability. It is expected that an event with a very high probability does not contribute to new information a lot since the negative logarithm takes on higher values for probabilities closer to zero. Additionally, the logarithmic term accounts for an event which is definitely occurring, i.e., its probability is 1, then this event does not contribute to any new information gain. Consistently, its share in the Shannon entropy is 0. Of course, we only consider probabilities smaller than or equal to 1 here.
Based on the Shannon entropy as a measure for information, we now define the mutual information for two different "event series" A, B based on the definition given in [61].

Definition 5.2 (Mutual information — discrete events)
Let A, B be two different "event series" with probabilities $p(a)$ and $p(b)$ for $a \in A$ and $b \in B$. Then the mutual information MI is defined as

$$\mathrm{MI}(A, B) = \sum_{\substack{a \in A \\ b \in B}} p(a, b) \log\left(\frac{p(a, b)}{p(a)\, p(b)}\right).$$

Transferring this definition again to the imaging context, we can interpret A and B as two images with gray values a and b, respectively. The probabilities $p(a)$ and $p(b)$ correspond to the probabilities of

the given gray values, i.e., for image A living on a discrete pixel grid $p(a)$ equals the ratio between the number of pixels with gray value a and the total number of pixels in this image. In this context, the joint probability $p(a,b)$ corresponds to the probability of pixels in image A taking the gray value a while the same pixels take the value b in image B. We remark that this definition of mutual information is related to the Kullback-Leibler distance [61] which is defined as

$$\sum_{i=1}^{m} p(i) \log\left(\frac{p(i)}{q(i)}\right)$$

for two probability distributions p, q. Applying the Kullback-Leibler distance between the joint probability distribution $p(a,b)$ and the joint distribution in case of independence, $p(a)\,p(b)$, we obtain again the definition of mutual information given in Definition 5.2.
When maximizing the mutual information, we gain the most information and reduce the uncertainty. Coming back to our example of the two different imaging modalities, CT and MRI, we could apply a deformation function on one of the images, e.g., on the MRI data. Maximizing then the mutual information of the original CT image and the deformed MRI image, we can optimize the deformation parameters to achieve a desirable alignment of prominent structures in both images.
This is one way of registering images. We refer the interested reader to the literature to read more about different registration techniques. A brief overview of registration approaches derived from variational methods can be found in [11] for example.
In the next section, we focus on the definition of an optimization problem based on mutual information for our investigation of cell colony spreading. By maximizing the mutual information between feature and classification images we aim for deriving spreading properties which both capture the colony growth phenomenon in the feature image and can be used as model parameters to generate the classification images.

5.2 Optimization problem for cell colony spreading

In the previous section, we introduced the concept of maximizing the mutual information between two images derived from different medical imaging modalities to achieve an adequate alignment. We now transfer this approach to our cell colony context.
Throughout this section, we consider the multi-channel feature image $I_1^{\mathrm{d}} : \Omega_0 \times \Omega_T \to \mathcal{F}$ influenced by Gaussian noise and introduced in Section 3.3 (cf. Definition 3.5). To prepare the later analysis we already limit the domain of the considered features to $\mathcal{F}'$ focusing only on features which are occurring with a certain probability greater than a small limit bounded away from zero (cf. Definition 3.14). Such a feature image mapping could for example consider resetting the occurring features of very low probabilities by a default replacement value or by the mean value of expected feature values in a small neighborhood around an inadequate feature. To avoid an overload of notations, we do not introduce another feature image definition which indeed only maps to $\mathcal{F}'$. We merely neglect features occurring with too low probabilities and replace their probabilities by 0. In this sense, we accept a certain inaccuracy here and proceed with the reduced space $\mathcal{F}'$ instead of $\mathcal{F}$ in the further course.

Besides from the feature data, we use the classification image $I_2 : \boldsymbol{P} \times \Omega_T \to \boldsymbol{\mathcal{C}}$ from Section 4.2 (cf. Definition 4.4) for the further course. By definition the classification image I_2 depends continuously on $\boldsymbol{p} \in \boldsymbol{P}$.
Based on both image types, we introduce our MI-based optimization problem in the following Section 5.2.2. Before we introduce the mutual information concerning probabilities related to our feature and classification images, we start with an excursion on the Radon-Nikodym theorem in Section 5.2.1. With this, we facilitate the further analysis and derivation of the mutual information not only based on *probability measures* but also considering related *probability density functions*.
Since we want to solve the maximization of mutual information with a gradient based optimization solver, we need to derive gradient terms. This is the main focus of Section 5.2.4. To prepare the derivative terms and the histogram definitions in that section, we first concentrate on the measure-theoretical setting (cf. Section 5.2.3). We conclude Section 5.2 with translating the histograms, the mutual information and its gradient terms into a discrete setting (cf. Section 5.2.5).

5.2.1 The Radon-Nikodym theorem - a small excursion

We adapt the Radon-Nikodym theorem to our context, here. We skip its proof, but add some interesting properties and results related to the Radon-Nikodym derivative. For the main proof and more related concepts to the Radon-Nikodym theorem, we refer the interested reader to [54].
We start with the following definition of a density function for a measure (cf. Definition 7.1 in [54]). Although we choose here the measurable space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T))$ for our space-time domain, the definition is not bound to this specific space and can be transferred to any measure space.

Definition 5.3
Let $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T), \lambda_1)$ be a measure space and $f : \Omega_0 \times \Omega_T \to [0, \infty]$ a nonnegative measurable function. The following

$$\lambda_2 : \mathcal{E} \otimes \mathcal{B}(\Omega_T) \to [0, \infty), \quad A \mapsto \int_A f \, \mathrm{d}\lambda_1$$

defines a measure with density f with respect to λ_1 and this is also abbreviated by

$$\lambda_2 = f\lambda_1.$$

It can be shown that $\lambda_2 = f\lambda_1$ is indeed a measure. We will skip this proof here and leave it to the interested readers to prove it themselves or check out details in the referenced literature.
Additionally, we state the following theorem cited from 7.3 in [54] to use it later in the proof for the histogram derivatives.

Theorem 5.4
Let $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T), \lambda_1)$ be a measure space. Let f, g be two nonnegative, measurable functions and $\lambda_2 := f\lambda_1$. Then it holds that g is λ_2–integrable if and only if $g \cdot f$ is λ_1–integrable. In this case the following identity holds:

$$\int g \, \mathrm{d}\lambda_2 = \int g \cdot f \, \mathrm{d}\lambda_1.$$

For further preparations of the famous Radon-Nikodym theorem, we clarify the notion of *absolute continuity* in the context of measures (cf. Definitions 7.9 and 7.24 in [54], Definition 3.5 in [12], Definition 7.30 in [40]).

Definition 5.5 (Absolute continuity of measures)
Let λ_1, λ_2 be measures on the measurable space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T))$. The measure λ_2 is *absolutely continuous* with respect to λ_1, if $\lambda_2(A) = 0$ holds for every $A \in \mathcal{E} \otimes \mathcal{B}(\Omega_T)$ with $\lambda_1(A) = 0$. This is written as $\lambda_2 \ll \lambda_1$.

We now state the Radon-Nikodym theorem based on the formulations in theorems 7.13 in [54], 7.34 in [40] and Theorem 3.6 in [12].

Theorem 5.6 (Radon-Nikodym theorem for finite measures)
Let λ_1, λ_2 be two finite measures of the measurable space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T))$ with $\lambda_2 \ll \lambda_1$. Then there exists a measurable function $f : \Omega_0 \times \Omega_T \to [0, \infty)$ with

$$\lambda_2(A) = \int_A f \, \mathrm{d}\lambda_1$$

for every $A \in \mathcal{E} \otimes \mathcal{B}(\Omega_T)$, i.e., λ_2 has the density f with respect to λ_1.
The function f is also referred to as the *Radon-Nikodym derivative* of λ_2 with respect to λ_1, typically denoted by $f = \frac{\mathrm{d}\lambda_2}{\mathrm{d}\lambda_1}$.

Proof. We refer the interested reader to the proofs stated in [54] for Theorem 7.13 and in [40] for Theorem 7.34. □

We conclude with a remark on the used notation associated with the Radon-Nikodym derivative.

Remark 5.7. We want to draw the reader's attention to the Radon-Nikodym derivative given as $f = \frac{\mathrm{d}\lambda_2}{\mathrm{d}\lambda_1}$. Depending on the literature source the Radon-Nikodym derivative is also sometimes denoted with $f = \frac{\lambda_2}{\lambda_1}$. We stress that both expressions are considered to describe the same concept and the " d" is only a notational difference.

To give a broader overview and to emphasize the benefits related to the theory around the Radon-Nikodym derivative, we conclude with the following lemma connecting the Radon-Nikodym derivative with the Lebesgue measures.

Lemma 5.8 (Radon-Nikodym derivative replacement by density functions)
Let λ_1, λ_2 be finite (probability) measures on the measurable (probability) space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T))$ and λ_3 be another measure on the measurable space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T))$. If $\lambda_1 \ll \lambda_3$ and $\lambda_2 \ll \lambda_3$ hold, then there exist (probability) density functions $f_1, f_2 : \Omega_0 \times \Omega_T \to [0, \infty)$ for λ_1, λ_2 with respect to the measure λ_3, i.e.,

$$\lambda_1(A) = \int_A f_1 \, \mathrm{d}\lambda_3$$
$$\lambda_2(A) = \int_A f_2 \, \mathrm{d}\lambda_3$$

for any $A \in \mathcal{E} \otimes \mathcal{B}(\Omega_T)$.
Moreover, if $\lambda_1 \ll \lambda_2$ holds, the function $\frac{f_1}{f_2}$ is well-defined and equals the Radon-Nikodym derivative $\frac{\mathrm{d}\lambda_1}{\mathrm{d}\lambda_2}$.

Proof. The existence of the (probability) density functions f_1, f_2 in the first part of this lemma is a direct consequence from Theorem 5.6. Considering the Radon-Nikodym derivatives $f_1 = \frac{\mathrm{d}\lambda_1}{\mathrm{d}\lambda_3}$ and $f_2 = \frac{\mathrm{d}\lambda_2}{\mathrm{d}\lambda_3}$ from the first part of the lemma, we can derive the statement of the second part consequently by the following transformations:

$$\frac{\mathrm{d}\lambda_1}{\mathrm{d}\lambda_2} = \frac{\frac{\mathrm{d}\lambda_1}{\mathrm{d}\lambda_3}}{\frac{\mathrm{d}\lambda_2}{\mathrm{d}\lambda_3}} = \frac{f_1}{f_2}.$$

This is well-defined, i.e., we avoid dividing by 0 because of the absolute continuity of the related measures. □

We remark that the statement in the above lemma is true for any measure λ_3 on the measure space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T))$ for which it holds that the other two measures are both absolutely continuous with respect to the third measure. We use this lemma in the further course when considering the third measure λ_3 to be the Lebesgue measure on the measure space.
Finally, we cite another interesting property related to the Radon-Nikodym derivatives which is also known as the "first chain rule" (cf. the first rule in 7.20 in [54]).

Lemma 5.9 (Chain rule with Radon-Nikodym derivative)
Let λ_2, λ_1 be finite measures on the measure space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T))$ and $\lambda_1 \ll \lambda_2$. For an integrable function f and any measurable set $A \in \mathcal{E} \otimes \mathcal{B}(\Omega_T)$, it holds that

$$\int_A f \, \mathrm{d}\lambda_1 = \int_A f \frac{\mathrm{d}\lambda_1}{\mathrm{d}\lambda_2} \, \mathrm{d}\lambda_2.$$

Proof. We refer to the proof of the first statement in 7.20 in [54]. □

After this excursion on the theory related to Radon-Nikodym derivatives, we focus next on calculating the mutual information based on probabilities related to our feature and classification images.

5.2.2 Introduction to the MI optimization problem

In the preceding Section 5.1 of this chapter, we motivated the application of mutual information for the registration of medical images and stated the definition of mutual information in Definition 5.2 very roughly. In particular, the example was based on discrete images, i.e., images living on a pixel grid. We now take a step back and consider non-discretized images living on the true time-space domain Ω_T without considering the discretization of Ω based on a pixel grid. This is the starting point for our continuous definition of mutual information. We stress here that both the domain $\Omega_0 \times \Omega_T$ and the codomain $\mathcal{F}$ of I_1^{d} and $\mathcal{C}$ of I_2, respectively, are considered to be continuous. To be more precise, the spaces Ω_0, Ω, $\mathcal{F}$ and $\mathcal{C}$ are not simplified or discretized in any way so far whereas we consider the temporal dimension of Ω_T to be discretized by pre-selected time points $\{t_1, \dots, t_{n_T}\} \subset [0, T]$ (cf. Definition 3.1). However, instead of focusing on the whole feature space $\mathcal{F}$, we remind the reader that we are focusing only on $\mathcal{F}'$ which contains only features occurring with a certain probability greater than a small limit bounded away from zero (cf. Definition 3.14). Furthermore, we remind the reader that I_1^{d} is considered to be a random variable for which the Gaussian additive noise is related to the subdomain Ω_0 as introduced in Definition 3.5.
We begin with recapitulating the probability density functions in the feature and classification spaces (cf. Proposition 3.7, Definition 3.8 and Proposition 3.12).

Definition 5.10 (Probability density functions and probability measures)
We define the probability density functions

$$\begin{aligned} p_{\mathcal{F}}^{\mathrm{d}} : \mathcal{F}' \to \mathbb{R}_+ \quad &\text{with } P_{\mathcal{F}}^{\mathrm{d}}\left(\mathcal{F}'\right) = \int_{\mathcal{F}'} p_{\mathcal{F}}^{\mathrm{d}}(f)\ \mathrm{d}f \leq 1 \\ p_{\mathcal{C}} : \mathcal{C} \to \mathbb{R}_+ \quad &\text{with } P_{\mathcal{C}}(\mathcal{C}) = \int_{\mathcal{C}} p_{\mathcal{C}}(c)\ \mathrm{d}c = 1 \\ p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}} : \mathcal{F}' \times \mathcal{C} \to \mathbb{R}_+ \quad &\text{with } P_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}}\left(\mathcal{F}' \times \mathcal{C}\right) = \int_{\mathcal{F}'\times\mathcal{C}} p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}}(f,c)\ \mathrm{d}(f,c) \leq 1 \end{aligned}$$

capturing the probability of occurring values $f \in \mathcal{F}'$, $c \in \mathcal{C}$ and $(f,c) \in \mathcal{F}' \times \mathcal{C}$, respectively, given the images I_1^{d} and I_2 living on $\Omega_0 \times \Omega_T$ and $\boldsymbol{P} \times \Omega_T$. The related probability measures can be calculated by integration of the probability density functions with respect to the Lebesgue measures of the related spaces. It holds

$$\begin{aligned} P_{\mathcal{F}}^{\mathrm{d}}(A_{\mathcal{F}'}) &= \int_{A_{\mathcal{F}'}} p_{\mathcal{F}}^{\mathrm{d}}(f)\ \mathrm{d}f \\ P_{\mathcal{C}}(A_{\mathcal{C}}) &= \int_{A_{\mathcal{C}}} p_{\mathcal{C}}(c)\ \mathrm{d}c \\ P_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}}(A_{\mathcal{F}'\times\mathcal{C}}) &= \int_{A_{\mathcal{F}'\times\mathcal{C}}} p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}}(f,c)\ \mathrm{d}(f,c) \end{aligned}$$

for arbitrary measurable subsets $A_{\mathcal{F}'} \subset \mathcal{F}'$, $A_{\mathcal{C}} \subset \mathcal{C}$ and $A_{\mathcal{F}'\times\mathcal{C}} \subset \mathcal{F}' \times \mathcal{C}$.

We stress here that the above inequalities in the integral terms are a direct consequence of cutting off certain features with very low probability in the definition of the reduced feature space $\mathcal{F}'$, cf. Definition 3.14.

Remark 5.11. We remark that the individual probability distributions can be calculated via integration over the other space of the joint probability as follows:

$$p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f) = \int_{\boldsymbol{\mathcal{C}}} p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f,c)\ \mathrm{d}c$$
$$p_{\boldsymbol{\mathcal{C}}}(c) = \int_{\boldsymbol{\mathcal{F}}} p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f,c)\ \mathrm{d}f$$

Under the assumption that our classification image is continuous and even continuously differentiable on $\boldsymbol{P}\times\Omega$ for any $t\in\{t_1,\dots,t_{n_T}\}$, we can conclude that the probability density function $p_{\boldsymbol{\mathcal{C}}}$ of a classification image corresponding to a fixed $\boldsymbol{p}\in\boldsymbol{P}$ and any time point $t\in\{t_1,\dots,t_{n_T}\}$ is continuous on its range $I_2(\boldsymbol{p},\Omega,t)$, too. Even more, the joint probability $p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ is continuous in the direction of $\boldsymbol{\mathcal{C}}$. Since the classification image I_2 depends on the parameter setting $\boldsymbol{p}$, we remark that $p_{\boldsymbol{\mathcal{C}}}$ and $p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ depend on $\boldsymbol{p}$, too, because of their dependence on I_2, respectively.
With those probabilities at hand, we define now the mutual information in the continuous setting oriented on Theorem 1 in [20] and Remark 4.3 in [60].

Definition 5.12 (Mutual information — continuous setting)
Let the probability measures $P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}$, $P_{\boldsymbol{\mathcal{C}}}$ and the joint probability measure $P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ be given and describing the probabilities related to images I_1^{d} and I_2 living on $\Omega_0\times\Omega_T$ and Ω_T. We assume that $P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \ll P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\otimes P_{\boldsymbol{\mathcal{C}}}$. The mutual information between I_1^{d} and I_2 is then defined by

$$\mathrm{MI}\left(P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\right) := \int_{\boldsymbol{\mathcal{F}}'\times\boldsymbol{\mathcal{C}}} \log\left(\frac{\mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}}{\mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\otimes P_{\boldsymbol{\mathcal{C}}}}\right)\mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}. \tag{5.1}$$

We stress that it is crucial that the joint probability measure $P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ is absolutely continuous with respect to the product measure $P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\otimes P_{\boldsymbol{\mathcal{C}}}$ to define the above calculation of mutual information based on probability measures.
We already derived the transition from probability measures to probability density functions in the previous course of this thesis, especially in Section 3.3. With this at hand, we motivate the following proposition on calculating the mutual information based on given probability density functions.

Proposition 5.13 (Mutual information based on PDFs)
Let the probability measures $P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}$, $P_{\boldsymbol{\mathcal{C}}}$ and the joint probability measure $P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ be given and describing the probabilities related to images I_1^{d} and I_2 living on $\Omega_0\times\Omega_T$ and Ω_T. When considering the related probability density functions $p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}$, $p_{\boldsymbol{\mathcal{C}}}$ and $p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$, the mutual information can be calculated by the following integration with respect to the Lebesgue measure on the joint space:

$$\mathrm{MI}\left(p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\right) := \int_{\boldsymbol{\mathcal{F}}'\times\boldsymbol{\mathcal{C}}} p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f,c)\log\left(\frac{p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f,c)}{p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f)\,p_{\boldsymbol{\mathcal{C}}}(c)}\right)\mathrm{d}(f,c). \tag{5.2}$$

Before we focus on the proof of this statement, we remark that we denote the underlying probabilities in the parenthesis on the left hand side for which we calculate the mutual information. The definitions

of the MI in Definition 5.12 and Proposition 5.13 could be read as MI(*"probabilities"*). For notational simplicity, we do not introduce two different symbols for the MI based on *probability measures* and the MI based on related *probability density functions* when considering integration with respect to the Lebesgue measure.

Proof. In Equation (5.1) the quotient in the logarithmic term is a Radon-Nikodym derivative (cf. [20]). Since, $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ and also $P_{\mathcal{F}} \otimes P_{\mathcal{C}}$ are absolutely continuous with respect to the Lebesgue measure on the joint space $\mathcal{F}\times\mathcal{C}$ (cf. Propositions 3.7 and 3.12 and Definition 3.8), we can replace the Radon-Nikodym derivative by the quotient of the related probability density functions (cf. Lemma 5.8)

$$\frac{\mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}}{\mathrm{d}P^{\mathrm{d}}_{\mathcal{F}} \otimes P_{\mathcal{C}}} = \frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}}{p^{\mathrm{d}}_{\mathcal{F}} \cdot p_{\mathcal{C}}}.$$

Additionally, we exploit Theorem 5.6 and Lemma 5.9 to change the integration measure implementing the relevant density function $p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ in the integrand term. □

The Radon-Nikodym derivative, the cited statements and the concept of absolute continuity of measures were already the main topic in Section 5.2.1.
To grasp the inherent dependencies hidden in the above definition of the mutual information, we expand the dependencies as follows by stating that the probabilities depend on the corresponding given images:

$$\mathrm{MI}\left(p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\right) = \mathrm{MI}\left(p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\left(I_1^{\mathrm{d}}, I_2\right)\right) = \mathrm{MI}\left(p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\left(I_1^{\mathrm{d}}, I_2\left(\boldsymbol{p}\right)\right)\right). \tag{5.3}$$

We clarify the well-definedness of the integral in the mutual information by approaching possible difficulties next.

Lemma 5.14
Due to the given noisy feature data we can infer that the joint probability measure $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ is absolutely continuous with respect to the product measure $P^{\mathrm{d}}_{\mathcal{F}} \otimes P_{\mathcal{C}}$, i.e., for all measurable $A \subset \mathcal{F}' \times \mathcal{C}$ with $P^{\mathrm{d}}_{\mathcal{F}} \otimes P_{\mathcal{C}}(A) = 0$ it holds that $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(A) = 0$.
Equivalently, for all measurable $A \subset \mathcal{F}' \times \mathcal{C}$ with $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(A) \neq 0$ it holds that $P^{\mathrm{d}}_{\mathcal{F}} \otimes P_{\mathcal{C}}(A) \neq 0$.

Proof. We prove the second statement of the lemma by applying the disintegration theorem stated in Theorem 3.9. We consider $A \subset \mathcal{F}' \times \mathcal{C}$ measurable with $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(A) \neq 0$. With

$$P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(A) = \int_A p_{\mathcal{C}}(c)\,(p_N * \nu_c)(f)\ \mathrm{d}(f,c)$$

and the non-negativity of the probability density function $p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c) = p_{\mathcal{C}}(c)\,(p_N * \nu_c)(f)$, it follows that a subset

$$A' = \left\{(f,c) \in A \subset \mathcal{F}' \times \mathcal{C} \,|\, p_{\mathcal{C}}(c) \neq 0 \wedge (p_N * \nu_c)(f) \neq 0\right\}$$

with positive Lebesgue measure exists. With this it follows directly that

$$P^{\mathrm{d}}_{\mathcal{F}} \otimes P_{\mathcal{C}}(A) \geq P^{\mathrm{d}}_{\mathcal{F}} \otimes P_{\mathcal{C}}\left(A'\right) = \int_{A'} p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c)\ \mathrm{d}(f,c) > 0$$

because A' is not a Lebesgue null set, $p_{\mathcal{C}}(c) > 0$ for all $(f, c) \in A'$ and $p^{\mathrm{d}}_{\mathcal{F}} > 0$ due to the considered Gaussian noise. We recall that $p^{\mathrm{d}}_{\mathcal{F}}$ is calculated by convolution with the Gaussian density which is in turn positive everywhere (cf. Proposition 3.7). For the limited space $\mathcal{F}'$, we even have that $p^{\mathrm{d}}_{\mathcal{F}}$ is bounded away from 0 (cf. Definition 3.14). This proves the absolute continuity of $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ with respect to the product measure $P^{\mathrm{d}}_{\mathcal{F}} \otimes P_{\mathcal{C}}$. □

In the following remark, we use this lemma to show the well-definedness of the integral term for the mutual information.

Remark 5.15. We want to point out a few essential properties related to the integral definition for mutual information. We recall that the integral for the MI is given by

$$\int_{\mathcal{F}'\times\mathcal{C}} \log\left(\frac{\mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}}{\mathrm{d}P^{\mathrm{d}}_{\mathcal{F}} \otimes P_{\mathcal{C}}}\right) \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$$

when considering the probability measures (cf. Equation (5.1)).

- In the logarithmic term it is problematic if the numerator approaches zero as the logarithm is not defined in zero. However, as this is only happening if $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(A)$ approaches zero, too, it is not a difficulty for the MI calculation based on probability measures as we are integrating with respect to the measure $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ anyway and, consequently, we could also consider the integration over the subset of $\mathcal{F}' \times \mathcal{C}$ where $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ is indeed positive.
- Another issue is when the denominator is zero in the logarithmic term. However, due to the aforementioned Lemma 5.14, we can ensure that the denominator can only be zero if the numerator is zero, too. With the previous remark it is clear that this is not problematic as we could consider integrating over the subset of $\mathcal{F}' \times \mathcal{C}$ where $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ is indeed positive.
- Finally, we stress that this interpretation is also valid in our setting where we limit the feature space to the space where the probabilities are greater than a small value bounded away from zero (cf. Definition 3.14).

In this sense, the calculation of MI based on the integration over the set $S \subset \mathcal{F}' \times \mathcal{C}$ where $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} > 0$ holds remedies the issues:

$$\int_S \log\left(\frac{\mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}}{\mathrm{d}P^{\mathrm{d}}_{\mathcal{F}} \otimes P_{\mathcal{C}}}\right) \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}.$$

Analogously, we can consider the integrand to be well-defined

$$p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c) \log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c)}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c)}\right)$$

when calculating the MI with the help of probability density functions (cf. Equation (5.2)). In this setting, we either way consider the integration only over the support of $p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ or set

$$p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c)\log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c)}{p^{\mathrm{d}}_{\mathcal{F}}(f)\,p_{\mathcal{C}}(c)}\right) = \begin{cases} p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c)\log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c)}{p^{\mathrm{d}}_{\mathcal{F}}(f)p_{\mathcal{C}}(c)}\right) & \text{for } (f,c)\in\mathcal{F}'\times\mathcal{C} \\ & \text{with } p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c)\neq 0 \\ 0 & \text{else.} \end{cases}$$

With the help of $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} \ll P^{\mathrm{d}}_{\mathcal{F}}\otimes P_{\mathcal{C}}$, we pointed out that the mutual information is indeed well-defined.

We aim for a best fitting parameter setting $\boldsymbol{p}\in\boldsymbol{P}$ such that the corresponding classification image I_2 based on the concentric spreading model captures well the spreading colony observable in the microscopy data and highlighted even more in the feature images I^{d}_1. For this we define the following optimization problem bearing in mind the previously described MI calculation and also, importantly, the dependence of the classification image I_2 on the parameter setting $\boldsymbol{p}$.

Definition 5.16 (MI optimization problem)
For a given multi-channel feature image $I^{\mathrm{d}}_1:\Omega_0\times\Omega_T\to\mathcal{F}'$ and a classification image $I_2:\boldsymbol{P}\times\Omega_T\to\mathcal{C}$ depending on a parameter setting $\boldsymbol{p}\in\boldsymbol{P}$, we have the related probability density functions

$$p^{\mathrm{d}}_{\mathcal{F}} = p^{\mathrm{d}}_{\mathcal{F}}\left(I^{\mathrm{d}}_1\right), \qquad p_{\mathcal{C}} = p_{\mathcal{C}}\left(I_2(\boldsymbol{p})\right), \qquad p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} = p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\left(I^{\mathrm{d}}_1, I_2(\boldsymbol{p})\right).$$

Based on those probabilities we define the *maximization of mutual information* as

$$\operatorname*{argmax}_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{MI}\left(p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\right). \tag{MAX MI}$$

Since the identity

$$\operatorname*{argmax}_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{MI}\left(p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\right) = \operatorname*{argmin}_{\boldsymbol{p}\in\boldsymbol{P}} -\mathrm{MI}\left(p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\right).$$

holds, we can transfer the maximization of mutual information to the subsequent minimization problem. This gives us the main optimization problem we focus on throughout the course of this work.

Definition 5.17 (Main minimization problem)
For a given multi-channel feature image $I^{\mathrm{d}}_1:\Omega_0\times\Omega_T\to\mathcal{F}'$ and a classification image $I_2:\boldsymbol{P}\times\Omega_T\to\mathcal{C}$ depending on a parameter setting $\boldsymbol{p}\in\boldsymbol{P}$, we have the related probability density functions

$$p^{\mathrm{d}}_{\mathcal{F}} = p^{\mathrm{d}}_{\mathcal{F}}\left(I^{\mathrm{d}}_1\right), \qquad p_{\mathcal{C}} = p_{\mathcal{C}}\left(I_2(\boldsymbol{p})\right), \qquad p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} = p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\left(I^{\mathrm{d}}_1, I_2(\boldsymbol{p})\right).$$

With the functional

$$\mathrm{F}:\boldsymbol{P}\to\mathbb{R}, \qquad \mathrm{F}(\boldsymbol{p}) := -\mathrm{MI}\left(p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\right)$$

where the joint probability depends on the given parameter set $\boldsymbol{p}$, we introduce the *main minimization problem*

$$\underset{\boldsymbol{p}\in P}{\operatorname{argmin}} \mathrm{F}(\boldsymbol{p}). \tag{MIN -MI}$$

We translate it directly to a minimization problem. This is more convenient as most of the optimization solvers are implemented for minimization problems by default and we will apply a minimization solver in our numerical tests as well (cf. Section 5.5). Moreover, we are considering a gradient-based optimization solver and, therefore, we advance with the derivation of the gradient term for mutual information in the further course (Section 5.2.4) . To facilitate the understanding of our line of arguments for the gradient derivation, we first introduce histogram definitions in the same chapter after a short excursion on the measure-theoretical setting in the next Section 5.2.3.

5.2.3 Measure-theoretical setting

When approaching the histogram definition from a measure-theoretical viewpoint, we need to clarify the basic setting we are using. We begin with a collection of some basic notations we are using throughout the further course of Chapter 5.

Notation 5.18
We recapitulate the probability measures introduced in Section 3.3.1. Additionally, we introduce some measure spaces which are essential for the later analysis as well.

1. Measures related to the Gaussian noise and the spatio-temporal domain Ω_T:

   - The measure space for our space-time domain Ω_T is given by $(\Omega_T, \mathcal{B}(\Omega_T), \kappa)$. The Lebesgue measure is denoted with κ and $\mathcal{B}(\Omega_T)$ is the Borel $\sigma-$algebra containing all measurable subsets of our semi-discrete spatio-temporal domain Ω_T (cf. Definition 3.1).

   - The probability space to model the noise effects corrupting the feature image is denoted with $\left(\Omega_0, \mathcal{E}, P^0\right)$. The probability measure is denoted with P^0 and the pushforward of P^0 with respect to the noise image I_N coincides with a Gaussian normal distribution (cf. Definitions 3.3 and 3.4).

   - We denote the measure space combining the spatio-temporal domain Ω_T with the noise space Ω_0 with $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T), \lambda)$ where we use the product measure $\lambda = P^0 \otimes \kappa$.

   - The probability space related to Ω_T is given as $(\Omega_T, \mathcal{B}(\Omega_T), P_{\Omega_T})$ with the uniform probability measure P_{Ω_T} and the Borel $\sigma-$algebra of Ω_T (cf. Definition 3.1).

   - The probability space combining the spatio-temporal domain Ω_T with the noise space Ω_0 is denoted with $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T), P^*)$ where we use $P^* = P^0 \otimes P_{\Omega_T}$ and the product $\sigma-$algebra.

2. Measures related to the classification image I_2 and the feature image I_1:

   - The measure space for the classification space $\mathcal{C}$ is given by $(\mathcal{C}, \mathcal{B}(\mathcal{C}), \kappa)$. The one dimensional Lebesgue measure is denoted with κ and $\mathcal{B}(\mathcal{C})$ is the Borel σ–algebra containing all measurable subsets of $\mathcal{C}$.

   - The measure space for the reduced classification space $\mathcal{C}'$ (cf. Definition 4.5) is given by $(\mathcal{C}', \mathcal{B}(\mathcal{C}'), \kappa)$. The one dimensional Lebesgue measure is denoted by κ and $\mathcal{B}(\mathcal{C}')$ is the Borel σ–algebra containing all measurable subsets of $\mathcal{C}'$.

   - The probability space related to $\mathcal{C}$ is given as $(\mathcal{C}, \mathcal{B}(\mathcal{C}), P_{\mathcal{C}})$ with the probability measure $P_{\mathcal{C}} = {I_2}_{\#} P_{\Omega_T}$ as the pushforward of P_{Ω_T} with respect to the classification image I_2 and the Borel σ–algebra of the classification space $\mathcal{C}$. While the measure $P_{\mathcal{C}}$ was already introduced in Definition 3.8 for a general classification image, we now consider I_2 as defined in Definition 4.4.

   - The measure space for the feature space $\mathcal{F}$ is given by $(\mathcal{F}, \mathcal{B}(\mathcal{F}), \kappa)$. The Lebesgue measure for the feature space is denoted by κ and $\mathcal{B}(\mathcal{F})$ is the Borel σ–algebra containing all measurable subsets of $\mathcal{F}$.

   - The measure space for the reduced feature space $\mathcal{F}'$ (cf. Definition 3.14) is given by $(\mathcal{F}', \mathcal{B}(\mathcal{F}'), \kappa)$. The Lebesgue measure for the reduced feature space is denoted with κ and $\mathcal{B}(\mathcal{F}')$ is the Borel σ–algebra containing all measurable subsets of $\mathcal{F}'$.

   - The probability space related to $\mathcal{F}$ is given as $(\mathcal{F}, \mathcal{B}(\mathcal{F}), P_{\mathcal{F}})$ with the probability measure $P_{\mathcal{F}} = {I_1}_{\#} P_{\Omega_T}$ as the pushforward of P_{Ω_T} with respect to the feature image I_1 and the Borel σ–algebra of the classification space $\mathcal{F}$ (cf. Definition 3.2).

   - The measure space for our joint feature-classification space $\mathcal{F} \times \mathcal{C}$ is given by $(\mathcal{F} \times \mathcal{C}, \mathcal{B}(\mathcal{F} \times \mathcal{C}), \kappa)$. The Lebesgue measure is here denoted with κ and $\mathcal{B}(\mathcal{F} \times \mathcal{C})$ is the Borel σ–algebra containing all measurable subsets of $\mathcal{F} \times \mathcal{C}$.

   - The probability space related to $\mathcal{F} \times \mathcal{C}$ is given as $(\mathcal{F} \times \mathcal{C}, \mathcal{B}(\mathcal{F} \times \mathcal{C}), P_{\mathcal{F} \times \mathcal{C}})$ with the probability measure $P_{\mathcal{F} \times \mathcal{C}} = (I_1, I_2)_{\#} P_{\Omega_T}$ as the pushforward of P_{Ω_T} with respect to the joint image (I_1, I_2) and the Borel σ–algebra of the joint space $\mathcal{F} \times \mathcal{C}$ (cf. Definition 3.8).

3. Measures related to the feature image corrupted by Gaussian noise I_1^{d}:

   - We denote with $P_{\mathcal{F}}^{\mathrm{d}} = {I_1^{\mathrm{d}}}_{\#} P^*$ the pushforward of P^* with respect to the disturbed feature image I_1^{d} which maps from $\Omega_0 \times \Omega_T$ to $\mathcal{F}$ (cf. Proposition 3.7).

   - We denote with $P_{\mathcal{F} \times \mathcal{C}}^{\mathrm{d}} = \left(I_1^{\mathrm{d}}, I_2\right)_{\#} P^*$ the pushforward of P^* with respect to the joint image $\left(I_1^{\mathrm{d}}, I_2\right)$ which maps from $\Omega_0 \times \Omega_T$ to $\mathcal{F} \times \mathcal{C}$ (cf. Proposition 3.12).

Remark 5.19. We remark that we denote the Lebesgue measure related to the different spaces Ω_T, $\mathcal{C}$, $\mathcal{F}$ and $\mathcal{F} \times \mathcal{C}$ with the same symbol κ to reduce the notational complexity. Instead of noting down

the related dimension in the superscript or subscript of the measure, we recall that for Ω_T we use the two dimensional Lebesgue measure corresponding to the measurable space $(\Omega, \mathcal{B}(\Omega))$ for any discrete time point $t \in \{t_1, \dots, t_{n_T}\}$ (cf. Definition 3.1 and Equation (3.1)). In case of the classification space $\boldsymbol{\mathcal{C}}$, we use the one dimensional Lebesgue measure. For the general feature space $\boldsymbol{\mathcal{F}}$, we use the n-dimensional Lebesgue measure and for our specific setting of a three dimensional feature space, we use the three dimensional Lebesgue measure. Consequently, for the joint space $\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$ we use the $n+1$ dimensional or $3+1$, respectively, dimensional Lebesgue measure.

Remark 5.20. Without proving it explicitly, we state that for the pushforward of P_{Ω_T} with respect to the classification image I_2 as defined in Definition 4.4, i.e., $P_{\boldsymbol{\mathcal{C}}} = I_{2\#}P_{\Omega_T}$ it holds that $P_{\boldsymbol{\mathcal{C}}} \ll \kappa$. Consequently, there exists a probability density function $p_{\boldsymbol{\mathcal{C}}}$ such that the statements in Definition 3.8 are still true and the subsequent statements in that Section 3.3.1 concerning the classification space $\boldsymbol{\mathcal{C}}$ or the joint space $\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$ are also still valid.

Based on the mappings for the feature images (cf. Definition 3.2) and the classification images (cf. Definition 4.4), we define the joint image mapping.

Definition 5.21 (Joint mapping)
We define the joint mapping for the feature image and the classification image by

$$\begin{aligned}&\boldsymbol{I} : \boldsymbol{P} \times \Omega_0 \times \Omega_T \to \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}},\\ &\boldsymbol{I}(\boldsymbol{p}, \omega, (\boldsymbol{x}, t)) = \left(I_1^{\mathrm{d}}, I_2\right)(\boldsymbol{p}, \omega, (\boldsymbol{x}, t)) = \left(I_1^{\mathrm{d}}(\omega, (\boldsymbol{x}, t)), I_2(\boldsymbol{p}, \boldsymbol{x}, t)\right).\end{aligned}$$

To approach the histogram definition from a measure-theoretical viewpoint, we start by recapitulating explicitly the definition of the *pushforward measure* (cf. Definition 2.16 in [54] and Definition 1.98 in [40]) with respect to the joint mapping $\boldsymbol{I} = \left(I_1^{\mathrm{d}}, I_2\right)$.

Definition 5.22 (Pushforward measure)
For a fixed parameter setting $\boldsymbol{p} \in \boldsymbol{P}$, we focus on the $(\mathcal{E} \otimes \mathcal{B}(\Omega_T), \mathcal{B}(\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}))$-measurable mapping of the product space $\Omega_0 \times \Omega_T$ into the joint image spaces $\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$:

$$\begin{aligned}&\boldsymbol{I} : \Omega_0 \times \Omega_T \to \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}},\\ &\boldsymbol{I}(\boldsymbol{p}, \omega, (\boldsymbol{x}, t)) = \left(I_1^{\mathrm{d}}, I_2\right)(\boldsymbol{p}, \omega, (\boldsymbol{x}, t)) = \left(I_1^{\mathrm{d}}(\omega, (\boldsymbol{x}, t)), I_2(\boldsymbol{p}, \boldsymbol{x}, t)\right).\end{aligned}$$

We define the *pushforward measure* of λ with respect to $\boldsymbol{I}$

$$\lambda^{\boldsymbol{I}} : \mathcal{B}(\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}) \to [0, \infty], \qquad A' \mapsto \lambda\left(\boldsymbol{I}^{-1}\left(A'\right)\right) \in [0, \infty]$$

for any $A' \in \mathcal{B}(\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}})$. This is a measure on $(\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}, \mathcal{B}(\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}))$.

Remark 5.23. Considering $A' \in \mathcal{B}(\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}})$, we complement the definition of the pushforward measure with the following equivalent notations which are also commonly used in the literature:

$$\lambda^{\boldsymbol{I}}\left(A'\right) = \left(\boldsymbol{I}_{\#}\lambda\right)\left(A'\right) = \left(\lambda \circ \boldsymbol{I}^{-1}\right)\left(A'\right) = \lambda\left(\boldsymbol{I}^{-1}\left(A'\right)\right).$$

We point out that this pushforward measure does not coincide with $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} = \left(I_1^{\mathrm{d}}, I_2\right)_{\#} P^*$ since $\lambda = P^0 \otimes \kappa$ holds while $P^* = P^0 \otimes P_{\Omega_T}$ holds. We stress that κ is the the Lebesgue measure on Ω_T whereas P_{Ω_T} is the uniform probability measure on Ω_T. Consequently, it holds that

$$P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} = \left(I_1^{\mathrm{d}}, I_2\right)_{\#} P^* = \frac{1}{\kappa\left(\Omega_T\right)} \left(I_1^{\mathrm{d}}, I_2\right)_{\#} \lambda = \frac{1}{n_T\, \kappa\left(\Omega\right)} \left(I_1^{\mathrm{d}}, I_2\right)_{\#} \lambda, \tag{5.4}$$

i.e we can transform $\boldsymbol{I}_{\#}\lambda$ to $\boldsymbol{I}_{\#}P^* = P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ via a normalization step.
Next, we note down the transformation theorem related to the pushforward measure (cf. the transformation formula 3.1 in [21], Theorem 3.20 in [54] or Theorem 4.10 in [40]).

Theorem 5.24
Let $\boldsymbol{g} : \mathcal{F} \times \mathcal{C} \to \bar{\mathbb{R}}$ be a $\mathcal{B}\left(\mathcal{F} \times \mathcal{C}\right)$-measurable, non-negative function and λ a measure on the measurable space $\left(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}\left(\Omega_T\right)\right)$. The mapping $\boldsymbol{I}$ is given as previously defined in Definition 5.22 for a fixed parameter setting $\boldsymbol{p} \in \boldsymbol{P}$. For a measurable subset $A' \subset \mathcal{F} \times \mathcal{C}$ the identity

$$\int\limits_{\boldsymbol{I}^{-1}(A')} \boldsymbol{g} \circ \boldsymbol{I}\left(\omega, (\boldsymbol{x}, t)\right)\, d\lambda\left(\boldsymbol{x}\right) = \int\limits_{A'} \boldsymbol{g}\, d\left(\lambda \circ \boldsymbol{I}^{-1}\right)(f, c)$$

holds. A measurable function $\boldsymbol{g} : \mathcal{F} \times \mathcal{C} \to \bar{\mathbb{R}}$ is $\boldsymbol{I}_{\#}\lambda$-integrable if and only if $g \circ \boldsymbol{I}$ is λ-integrable.

After having introduced the measure-theoretical setting, we can focus next on the definition of a histogram based on the pushforward measure.

5.2.4 Histogram definition and partial derivatives

For the optimization problem, we need to evaluate and analyze the mutual information based on a probability distribution. This probability distribution is related to a histogram. To be more precise, a probability measure can be derived from a histogram measure by applying a normalization. For an example for such a normalization step, we refer to Equation (5.4) because we define the histogram based on the pushforward measure next.
In addition to this histogram measure, we are also interested in partial derivatives for it (cf. Equation (5.17)). To ensure differentiability, we introduce a smoothing step via convolution with a smooth mollifier. First, we define the histogram with the previously introduced pushforward measure.

Definition 5.25 (Histogram definition — part 1)
In general, we define the histogram $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}$ in the joint feature-classification space $\mathcal{F} \times \mathcal{C}$ for a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ by

$$\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}\left(A'\right) = \lambda\left(\left\{(\omega, (\boldsymbol{x}, t)) \in \Omega_0 \times \Omega_T \,:\, \boldsymbol{I}\left(\boldsymbol{p}, \omega, (\boldsymbol{x}, t)\right) \in A'\right\}\right) \in [0, \infty]$$

for $A' \in \mathcal{B}\left(\mathcal{F} \times \mathcal{C}\right)$, the Borel σ–algebra of $\mathcal{F} \times \mathcal{C}$, and with the measure λ of the measurable space $\left(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}\left(\Omega_T\right)\right)$.

Remark 5.26. We remark that the histogram for $A' \in \mathcal{B}(\mathcal{F} \times \mathcal{C})$ can also be written as

$$\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}\left(A'\right) = \lambda\left(\boldsymbol{I}^{-1}\left(A'\right)\right) = \left(\lambda \circ \boldsymbol{I}^{-1}\right)\left(A'\right) = \left(\boldsymbol{I}_{\#}\lambda\right)\left(A'\right)$$

by making use of the pushforward measure defined in Definition 5.22 and its equivalent notations as given in Remark 5.23.

Similarly, we set the histograms on the individual spaces $\boldsymbol{H}_{\mathcal{F}}$ and $\boldsymbol{H}_{\mathcal{C}}$.

Definition 5.27 (Histogram definition — part 2)
The histograms on the separate spaces when considering a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ can be defined analogously via the pushforward measure as

$$\begin{aligned}\boldsymbol{H}_{\mathcal{F}}\left(A'_F\right) &= \left(I_{1\,\#}^{\mathrm{d}}\lambda\right)\left(A'_F\right),\\ \boldsymbol{H}_{\mathcal{C}}\left(A'_C\right) &= \left(I_{2\#}\kappa\right)\left(A'_C\right)\end{aligned}$$

for $A'_F \in \mathcal{B}(\mathcal{F})'$ and $A'_C \in \mathcal{B}(\mathcal{C})'$ and with the pushforward measures of λ and κ related to the corresponding spaces, i.e., $\Omega_0 \times \Omega_T$ for the pushforward of the disturbed feature image I_1^{d} and Ω_T for the pushforward of the classification image I_2.

In the previous definitions, we have defined the histogram measures for a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$. In the next remark, we focus on the dependence of the histogram measures on this parameter set $\boldsymbol{p}$.

Remark 5.28 (Histograms as mappings). The histogram measures $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}$ and $\boldsymbol{H}_{\mathcal{C}}$ defined in Definitions 5.25 and 5.27 depend on a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ because of the dependence of the underlying classification image I_2 on $\boldsymbol{p}$ (cf. Definition 4.4). While we do not specify this dependence in the symbols for the histogram measure, we emphasize that we can understand the histograms $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}} = \boldsymbol{I}_{\#}\lambda$ and $\boldsymbol{H}_{\mathcal{C}} = I_{2\#}\kappa$ as mappings in the sense that for any fixed $A' \in \mathcal{B}(\mathcal{F} \times \mathcal{C})$ and any fixed $C \in \mathcal{B}(\mathcal{C})$

$$\begin{aligned}&\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}} : \boldsymbol{P} \to [0,\infty], && \boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}(\boldsymbol{p}) = \boldsymbol{I}(\boldsymbol{p},\cdot)_{\#}\lambda\left(A'\right),\\ &\boldsymbol{H}_{\mathcal{C}} : \boldsymbol{P} \to [0,\infty], && \boldsymbol{H}_{\mathcal{C}}(\boldsymbol{p}) = I_2(\boldsymbol{p},\cdot)_{\#}\kappa(C).\end{aligned}$$

We can understand the histograms as mappings from the parameter space $\boldsymbol{P}$ to measures on $(\mathcal{F}\times\mathcal{C}, \mathcal{B}(\mathcal{F}\times\mathcal{C}))$ denoted by $\mathcal{M}(\mathcal{B}(\mathcal{F}\times\mathcal{C}))$, or to measures on $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$ denoted by $\mathcal{M}(\mathcal{B}(\mathcal{C}))$, respectively. Without introducing a separate notation for these histogram considered as mappings in the sense that

$$\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}} : \boldsymbol{P} \to \mathcal{M}(\mathcal{B}(\mathcal{F}\times\mathcal{C})), \qquad \boldsymbol{H}_{\mathcal{C}} : \boldsymbol{P} \to \mathcal{M}(\mathcal{B}(\mathcal{C})),$$

we stress that we need to keep the dependence of the histograms on the parameter setting in mind.

It is crucial to bear this dependence of the histograms on the parameter setting in mind when thinking about derivatives of the histograms with respect to the parameter setting. We emphasize that we are interested in these derivatives since we want to find the optimal parameter setting to maximize the mutual information between the feature images and the classification images by

applying a gradient-based numerical solver.
To facilitate the derivation of gradient terms for the joint histogram, we introduce a smoothing step. We smooth the measure by convolving the histogram with a smooth mollifier which is living only in the classification domain $\mathcal{C}$ because the parameter setting is also influencing the histograms only through the classification values. To achieve this, we first state a version of the disintegration theorem adapted to this context. After that we are able to focus on the derivatives of the histogram, finally.
The next theorem is based on the disintegration theorem given in [4] (cf. Theorem 5.3.1) and in particular on the remark after Theorem 5.3.1 in [4] dealing with the special case of disintegration for product spaces. We directly adapt the theorem's statement to our notation and used spaces in this section compared to the first version of this theorem stated in Theorem 3.9.
We are mainly interested in the disintegration of our histogram measure $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}} = \boldsymbol{I}_{\#}\lambda$. As the cited theorem is stated for probability measures, we first focus on the disintegration of the probability measure $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} = \boldsymbol{I}_{\#}P^{*}$ and then extend this in a second step to the disintegration of $\boldsymbol{I}_{\#}\lambda$ by considering the normalization stated in Equation (5.4).

Theorem 5.29 (Adapted Disintegration Theorem — Version 2)
We consider the probability space $\left(\mathcal{F}\times\mathcal{C}, \mathcal{B}\left(\mathcal{F}\times\mathcal{C}\right), P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\right)$ and the measurable spaces $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$, $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$. Let the natural projection onto the classification domain $\mathcal{F}$ be

$$\pi_{\mathcal{F}} : \mathcal{F}\times\mathcal{C} \to \mathcal{F}$$

and we have with

$$P^{\mathrm{d}}_{\mathcal{F}} = \pi_{\mathcal{F}\#}\left(P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\right) = \left(P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\right)\circ\pi^{-1}_{\mathcal{F}}.$$

a probability measure for the measurable space $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$. Then each fiber $\pi^{-1}_{\mathcal{F}}(f) = \{f\}\times\mathcal{C}$ can canonically be identified with $\mathcal{C}$ for any $f \in \mathcal{F}$. Moreover, there exists a $P^{\mathrm{d}}_{\mathcal{F}}$-almost everywhere uniquely determined Borel family of probability measures $\boldsymbol{P}_{\mathcal{C}} = \left\{P_{\mathcal{C}f}\right\}_{f\in\mathcal{F}}$ on the measurable space $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$ such that

$$P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\left(A'\right) = \int_{\mathcal{F}} P_{\mathcal{C}f}\left(A'\cap\pi^{-1}_{\mathcal{F}}(f)\right)\ \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}}(f) = \int_{\mathcal{F}} P_{\mathcal{C}f}\left(\left\{\left(f', c\right)\in A' \,|\, f' = f\right\}\right)\ \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}}(f)$$

holds for any measurable set $A' \in \mathcal{B}\left(\mathcal{F}\times\mathcal{C}\right)$. Furthermore, it holds that

$$\int_{\mathcal{F}\times\mathcal{C}} g(f,c)\ \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c) = \int_{\mathcal{F}}\int_{\pi^{-1}_{\mathcal{F}}(f)=\mathcal{C}} g(f,c)\ \mathrm{d}P_{\mathcal{C}f}(c)\ \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}}(f) \tag{5.5}$$

for every Borel map $g : \mathcal{F}\times\mathcal{C} \to [0, +\infty]$.
We use the notation $P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}} = \left(\boldsymbol{P}_{\mathcal{C}}, P^{\mathrm{d}}_{\mathcal{F}}\right)$ for a disintegration along $\mathcal{F}$ of the probability measure.

Without proving this theorem, we introduce a direct consequence in the following remark in which we focus on the disintegration of $\boldsymbol{I}_{\#}\lambda$.

Remark 5.30. We consider the same setting as in Theorem 5.29 and Notation 5.18.
We recall that

$$P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} = \boldsymbol{I}_{\#}P^{*} = P^{*}\circ\boldsymbol{I}^{-1} = \frac{1}{\kappa\left(\Omega_T\right)}\lambda\circ\boldsymbol{I}^{-1}$$

holds (cf. Equation (5.4)). In this spirit, the probability measure P^{*} is the normalization of λ by considering a uniform probability distribution on Ω_T. We specify that we can also disintegrate $\boldsymbol{I}_{\#}\lambda$ by reverting this normalization, i.e. we could scale $P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}$ or the family $\boldsymbol{P}_{\boldsymbol{\mathcal{C}}}$ by $\kappa\left(\Omega_T\right) = n_T\kappa\left(\Omega\right)$ (cf. Equation (3.1)). In the following, we use $\boldsymbol{\mu} = \kappa\left(\Omega_T\right)\boldsymbol{P}_{\boldsymbol{\mathcal{C}}}$, i.e., $\boldsymbol{\mu} = \left\{\mu_f\right\}_{f\in\boldsymbol{\mathcal{F}}}$ with $\mu_f = \kappa\left(\Omega_T\right)P_{\boldsymbol{\mathcal{C}}f}$ and use $\left(\boldsymbol{\mu}, P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\right)$ as the disintegration of the *unnormalized* measure $\boldsymbol{I}_{\#}\lambda$.

In the further course, we consider one fixed family of probability measures $\boldsymbol{\mu} = \left\{\mu_f\right\}_{f\in\boldsymbol{\mathcal{F}}}$ for the disintegration of the pushforward measure $\boldsymbol{I}_{\#}\lambda$ with respect to its normalized projection on the feature space $\boldsymbol{\mathcal{F}}$, i.e., the measure $P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}} = \frac{1}{\kappa(\Omega_T)}\pi_{\boldsymbol{\mathcal{F}}\#}\left(\boldsymbol{I}_{\#}\lambda\right) = \pi_{\boldsymbol{\mathcal{F}}\#}\left(\boldsymbol{I}_{\#}P^{*}\right) = \pi_{\boldsymbol{\mathcal{F}}\#}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$.

Remark 5.31. A direct consequence of the disintegration in Theorem 5.29 and the previous remark on the disintegration $\left(\boldsymbol{\mu}, P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\right)$ of the measure $\boldsymbol{I}_{\#}\lambda$ along $\boldsymbol{\mathcal{F}}$ is that similar to Equation (5.5) it holds

$$\int_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} g\left(f,c\right)\ \mathrm{d}\boldsymbol{I}_{\#}\lambda\left(f,c\right) = \int_{\boldsymbol{\mathcal{F}}}\left(\int_{\boldsymbol{\mathcal{C}}} g\left(f,c\right)\ \mathrm{d}\mu_f\left(c\right)\right)\mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right).$$

for any Borel map $g : \boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}} \to \left[0,+\infty\right]$.

We use this disintegration to smooth the histogram $\boldsymbol{H}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} = \boldsymbol{I}_{\#}\lambda$. To prepare the smoothing step, we cite the convolution of measures first for the sake of completeness. The following lemma is based on the Proposition 5.4.1 in [6].

Lemma 5.32
Let λ_1, λ_2 be two finite measures on the measurable space $\left(\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}, \mathcal{B}\left(\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}\right)\right)$. For any measurable set $A \subset \boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}$, the convolution of the measures is defined as

$$\left(\lambda_1 * \lambda_2\right)\left(A\right) = \int_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\int_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \mathbf{1}_A\left(\left(f,c\right)+\left(f',c'\right)\right)\ \mathrm{d}\lambda_1\left(f,c\right)\ \mathrm{d}\lambda_2\left(f',c'\right). \tag{5.6}$$

Then holds that $\left(\lambda_1 * \lambda_2\right)$ is a measure as well on $\left(\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}, \mathcal{B}\left(\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}\right)\right)$.

Proof. We refer to the proof stated in [6] for Proposition 5.4.1. □

An equivalent formulation of Equation (5.6) is given by

$$\left(\lambda_1 * \lambda_2\right)\left(A\right) = \int_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \lambda_2\left(A-\left(f,c\right)\right)\ \mathrm{d}\lambda_1\left(f,c\right).$$

While we defined the convolution here for the joint feature classification space, it is also well defined for measures on other measurable spaces, e.g., for the measurable space $\left(\boldsymbol{\mathcal{C}}, \mathcal{B}\left(\boldsymbol{\mathcal{C}}\right)\right)$. To smooth our histogram measure, we apply convolution along the classification space $\boldsymbol{\mathcal{C}}$.

Definition 5.33 (Smoothing histogram by convolution)

1. Let $\eta \in C^\infty(\mathcal{C})$ be a smooth, non-negative function with $\int_{\mathcal{C}} \eta \, \mathrm{d}c = 1$ and which is compactly supported on $B(0,1)$. Moreover, we consider again the measure $\boldsymbol{I}_\# \lambda$ on $\mathcal{F} \times \mathcal{C}$ with its disintegration $\left(\boldsymbol{\mu}, P^{\mathrm{d}}_{\mathcal{F}}\right)$ along $\mathcal{F}$. We consider $\hat{\eta}$ as the measure on $\mathcal{C}$ with the probability density η with respect to the Lebesgue measure κ on the one dimensional classification space $\mathcal{C}$. Then we define the convolution operation along the subspace $\mathcal{C}$ of $\mathcal{F} \times \mathcal{C}$ by considering

$$\hat{\eta} *_{\mathcal{C}} \boldsymbol{I}_\# \lambda := \hat{\eta} *_{\mathcal{C}} \left(\boldsymbol{\mu}, P^{\mathrm{d}}_{\mathcal{F}}\right) = \left(\left\{\hat{\eta} * \mu_f\right\}_{f \in \mathcal{F}}, P^{\mathrm{d}}_{\mathcal{F}}\right)$$

   with $\hat{\eta} * \mu_f$ the convolution of measures on the measurable space $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$.

2. Let $\eta_{\varepsilon_1} := \frac{1}{\varepsilon_1} \eta\left(\frac{\cdot}{\varepsilon_1}\right)$ be a smooth mollification kernel with support $B(0, \varepsilon_1)$, $\int_{\mathcal{C}} \eta_{\varepsilon_1} \, \mathrm{d}c = 1$ and $\lim_{\varepsilon_1 \to 0} \eta_{\varepsilon_1}(0) = \delta(0)$, i.e. converging to the dirac delta function for vanishing ε_1. We define the smoothing operation along the $\mathcal{C}$–axis for the histogram measure $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}} = \boldsymbol{I}_\# \lambda$ by

$$\boldsymbol{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}} := \hat{\eta}_{\varepsilon_1} *_{\mathcal{C}} \boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}.$$

   when considering $\hat{\eta}_{\varepsilon_1}$ to be the measure with the density η_{ε_1} with respect to the Lebesgue measure.

Remark 5.34. We stress that $\boldsymbol{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}$ is still a measure on the measurable space $(\mathcal{F} \times \mathcal{C}, \mathcal{B}(\mathcal{F} \times \mathcal{C}))$. The convolution with a smooth mollifier in the classification domain $\mathcal{C}$ results in a smoothed version of the histogram, i.e., in the classification direction we achieve a differentiable representation in the sense that the histogram has a differentiable density with respect to the Lebesgue measure in the $\mathcal{C}$-direction. The derivatives are important for the approximation of the derivative of our mutual information term. In the following, we regard $\boldsymbol{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}$ as a smoothed measure living on the measurable space $(\mathcal{F} \times \mathcal{C}, \mathcal{B}(\mathcal{F} \times \mathcal{C}))$.

This smoothing step becomes crucial in the further course when we consider different discretization aspects to solve our optimization problem numerically. Underlying pixel grids of the feature images and classification images as well as binning effects in the feature and classification spaces can impede the calculation of gradient terms because of possibly occurring delta peaks or piecewise constant histogram measures. To smooth out such discontinuities, we introduce the smoothing by convolution with a mollifier.

Example 5.35. An example for a smooth mollifier as described in Definition 5.33 is the bump function

$$\eta : \mathbb{R} \mapsto \mathbb{R}, \qquad \eta(x) = \begin{cases} \frac{1}{C_{\text{norm.}}} \exp\left(\frac{-1}{1-x^2}\right), & |x| < 1 \\ 0, & |x| \geq 1. \end{cases}$$

We choose the normalization constant $C_{\text{norm.}} \approx 0.4440$ such that the mollifer's property

$$\int_{\mathbb{R}} \eta(x) \, dx = 1$$

holds. Based on this bump function, we construct the mollifier in the classification space with

$$\eta_{\varepsilon_1} : \boldsymbol{\mathcal{C}} \mapsto \mathbb{R}, \qquad \eta_{\varepsilon_1}(c) := \frac{1}{\varepsilon_1}\eta\left(\frac{c}{\varepsilon_1}\right).$$

Consequently, the compact support of η_{ε_1} is the interval $[-\varepsilon_1, \varepsilon_1]$.

Notation 5.36
The mollifier η introduced in Example 5.35 is also called the *standard mollifier* (Definitions in C.4. in [22]). In the following we use $\eta_{\varepsilon_1} = \frac{1}{\varepsilon_1}\eta\left(\frac{\cdot}{\varepsilon_1}\right)$ for the smoothing effect along the classification space.

Remark 5.37. We carry out explicitly the smoothing step for the histogram measure $\boldsymbol{H}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} = \boldsymbol{I}_{\#}\lambda$ using its disintegration $\left(\boldsymbol{\mu}, P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\right)$ as introduced in Remark 5.30. We focus on an arbitrary measurable subset $A' \subset \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$ and make use of the convolution of measures. The transformation goes as follows:

$$\begin{aligned}
\boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(A') &= \hat{\eta}_{\varepsilon_1} *_{\mathcal{C}} \boldsymbol{H}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(A') \\
&= \int_{\boldsymbol{\mathcal{C}}} \boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(A' - (\mathbf{0}, c)\right) \mathrm{d}\hat{\eta}_{\varepsilon_1}(c) \\
&= \int_{\boldsymbol{\mathcal{C}}} \eta_{\varepsilon_1}(c)\, \boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(A' - (\mathbf{0}, c)\right) \mathrm{d}c \\
&\overset{\text{disintegration}}{=} \int_{\boldsymbol{\mathcal{C}}} \eta_{\varepsilon_1}(c) \int_{\boldsymbol{\mathcal{F}}} \mu_f\left(\left\{\left(f', c'\right) \in A' - (\mathbf{0}, c) \mid f' = f\right\}\right) \mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f)\, \mathrm{d}c \\
&\overset{\text{Fubini}}{=} \int_{\boldsymbol{\mathcal{F}}} \int_{\boldsymbol{\mathcal{C}}} \eta_{\varepsilon_1}(c)\, \mu_f\left(\left\{\left(f', c'\right) \in A' - (\mathbf{0}, c) \mid f' = f\right\}\right) \mathrm{d}c\, \mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f) \\
&= \int_{\boldsymbol{\mathcal{F}}} \left(\hat{\eta}_{\varepsilon_1} * \mu_f\right)\left(\left\{\left(f', c'\right) \in A' \mid f' = f\right\}\right) \mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f).
\end{aligned}$$

This highlights that $\left(\left\{\hat{\eta}_{\varepsilon_1} * \mu_f\right\}_{f\in\boldsymbol{\mathcal{F}}}, P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\right)$ disintegrates the smoothed histogram $\boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ along the feature space $\boldsymbol{\mathcal{F}}$.

Remark 5.38. As another remark we state the integral term after executing the smoothing convolution and the disintegration for $\boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ when considering $A' = \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$:

$$\boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}) = \int_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \mathrm{d}\boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f, c) = \int_{\boldsymbol{\mathcal{F}}} \int_{\boldsymbol{\mathcal{C}}} \int_{\boldsymbol{\mathcal{C}}} \eta_{\varepsilon_1}\left(\hat{c} - c'\right) \mathrm{d}\mu_f\left(c'\right) \mathrm{d}\hat{c}\, \mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f).$$

We point out that this directly follows from the transformations in Remark 5.37 when including $A' = \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$ and the substitution of c by $\hat{c} - c'$.

In this sense, we remark that we can explicitly write down the density function of the convolution measure.

Remark 5.39. Considering the disintegration $\left(\left\{\hat{\eta}_{\varepsilon_1} * \mu_f\right\}_{f\in\boldsymbol{\mathcal{F}}}, P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\right)$ of the smoothed histogram $\boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ along the space $\boldsymbol{\mathcal{F}}$, we can derive the density function of the convolved measure $\hat{\eta}_{\varepsilon_1} * \mu_f$ for an arbitrary but fixed $f \in \boldsymbol{\mathcal{F}}$ as follows. For a set $C \in \mathcal{B}(\boldsymbol{\mathcal{C}})$ we derive

$$\begin{aligned}
\left(\hat{\eta}_{\varepsilon_1} * \mu_f\right)(C) &= \int_{\boldsymbol{\mathcal{C}}}\int_{\boldsymbol{\mathcal{C}}} \mathbf{1}_C\left(c+c'\right)\,\mathrm{d}\hat{\eta}_{\varepsilon_1}(c)\,\mathrm{d}\mu_f\left(c'\right)\\
&= \int_{\boldsymbol{\mathcal{C}}}\int_{\boldsymbol{\mathcal{C}}} \eta_{\varepsilon_1}(c)\,\mathbf{1}_C\left(c+c'\right)\,\mathrm{d}c\,\mathrm{d}\mu_f\left(c'\right)\\
&\overset{\substack{\text{subst.:}\\ c=\hat{c}-c'}}{=} \int_{\boldsymbol{\mathcal{C}}}\int_{\boldsymbol{\mathcal{C}}} \eta_{\varepsilon_1}\left(\hat{c}-c'\right)\mathbf{1}_C(\hat{c})\,\mathrm{d}\hat{c}\,\mathrm{d}\mu_f\left(c'\right)\\
&\overset{\text{Fubini}}{=} \int_{\boldsymbol{\mathcal{C}}}\int_{\boldsymbol{\mathcal{C}}} \eta_{\varepsilon_1}\left(\hat{c}-c'\right)\mathbf{1}_C(\hat{c})\,\mathrm{d}\mu_f\left(c'\right)\,\mathrm{d}\hat{c} = \int_{C}\int_{\boldsymbol{\mathcal{C}}} \eta_{\varepsilon_1}\left(\hat{c}-c'\right)\,\mathrm{d}\mu_f\left(c'\right)\,\mathrm{d}\hat{c}
\end{aligned}$$

which shows that indeed $\int_{\boldsymbol{\mathcal{C}}} \eta_{\varepsilon_1}(\cdot - c')\,\mathrm{d}\mu_f(c')$ can be considered as the density function with respect to the Lebesgue measure of the convolved measure for an arbitrary but fixed $f \in \boldsymbol{\mathcal{F}}$. We stress that the integration domain in the substitution step is not changing since we consider $\boldsymbol{\mathcal{C}} = \mathbb{R}$.

To solve the optimization problem of maximizing the mutual information of classification images and features images numerically, we are aiming for a gradient-based optimization technique. The goal of maximizing the MI is to find an optimal parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ which corresponds to a classification result that captures best the colony's spreading given in the data and represented in the feature images. Before we derive the gradient of our MI term, we deal with derivatives for the smoothed histogram measure.

The dependence of the histogram on the parameter set is given by the classification image I_2. We refer to Remark 5.28 to recapitulate the dependence of the histogram measures on the parameter setting $\boldsymbol{p} \in \boldsymbol{P}$. As discussed above, we introduced a smoothing step by convolution with a smooth mollifier to ensure differentiability along the classification axis. So now, we are interested in the partial derivatives of this smoothed histogram, $\frac{\partial \boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}}{\partial p_i}$ for each parameter $p_i \in \boldsymbol{p}$. In Section 4.2 we introduced the classification image I_2 in Definition 4.4. The partial derivatives of I_2 with respect to the parameters $p_i \in \boldsymbol{p}$ are also stated in that section as well, cf. Equation (4.8). We use $\nabla_{\boldsymbol{p}} I_2$, exploit properties related to the Radon-Nikodym theorem (cf. Section 5.2.1) and make use of the disintegration theorem (Theorem 5.40) to show the following statement on the directional derivative of the smoothed histogram:

Theorem 5.40
Let $\boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ be the smoothed histogram from Definition 5.33 and the derivative of the mollification kernel be given by η'_{ε_1}. For any parameter $p_i \in \boldsymbol{p}$, the directional derivative is given by

$$\frac{\partial \boldsymbol{H}^{\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}}{\partial p_i} = -\hat{\eta}'_{\varepsilon_1} *_{\boldsymbol{\mathcal{C}}} \boldsymbol{I}_{\#}\left(\frac{\partial I_2}{\partial p_i}\lambda\right) \tag{5.7}$$

with $\hat{\eta}'_{\varepsilon_1} = \eta'_{\varepsilon_1}\kappa$, i.e., the measure $\hat{\eta}'_{\varepsilon_1}$ has the density η'_{ε_1} with respect to the Lebesgue measure κ on $(\boldsymbol{\mathcal{C}}, \mathcal{B}(\boldsymbol{\mathcal{C}}))$.

Proof. Let $g \in C^0(\mathcal{F} \times \mathcal{C})$ be an arbitrary, continuous map and measurable. In the following we denote with the dual pair $\langle g, \sigma \rangle$ the integration of g with respect to a measure σ on $(\mathcal{F} \times \mathcal{C}, \mathcal{B}(\mathcal{F} \times \mathcal{C}))$, e.g., for the histogram measure $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}} = \boldsymbol{I}_{\#}\lambda$, we get with its disintegration $\left(\boldsymbol{\mu}, P^{\mathrm{d}}_{\mathcal{F}}\right)$ and $\boldsymbol{\mu} = \left\{\mu_f\right\}_{f\in\mathcal{F}}$

$$\langle g, \boldsymbol{I}_{\#}\lambda\rangle = \int_{\mathcal{F}\times\mathcal{C}} g(f,c)\ \mathrm{d}\boldsymbol{I}_{\#}\lambda(f,c) \overset{\text{Remark 5.30}}{=} \int_{\mathcal{F}}\int_{\mathcal{C}} g(f,c)\ \mathrm{d}\mu_f(c)\ \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}}(f)$$

and, analogously, for the smoothed histogram measure $\boldsymbol{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}$ and its disintegration $\left(\left\{\hat{\eta}_{\varepsilon_1} * \mu_f\right\}_{f\in\mathcal{F}}, P^{\mathrm{d}}_{\mathcal{F}}\right)$

$$\begin{aligned}
\langle g, \boldsymbol{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}\rangle &= \int_{\mathcal{F}\times\mathcal{C}} g(f,c)\ \mathrm{d}\boldsymbol{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}(f,c)\\
&\overset{\text{Definition 5.33}}{=} \int_{\mathcal{F}}\int_{\mathcal{C}} g(f,c)\ \mathrm{d}\left(\hat{\eta}_{\varepsilon_1} * \mu_f\right)(c)\ \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}}(f)\\
&\overset{\text{Remark 5.39}}{=} \int_{\mathcal{F}}\int_{\mathcal{C}} g(f,c)\int_{\mathcal{C}} \eta_{\varepsilon_1}(c-\hat{c})\ \mathrm{d}\mu_f(\hat{c})\ \mathrm{d}c\ \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}}(f).
\end{aligned}$$

We consider this as the starting point for the following transformations. Let p_i be an arbitrary parameter in our parameter set $\boldsymbol{p}$. Now, we focus on the partial derivative by using the dual pair:

$$\begin{aligned}
\langle g, \frac{\partial}{\partial p_i}\boldsymbol{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}\rangle &= \frac{\partial}{\partial p_i}\langle g, \boldsymbol{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}\rangle\\
&\overset{\text{Remark 5.39}}{=} \frac{\partial}{\partial p_i}\int_{\mathcal{F}}\int_{\mathcal{C}}\int_{\mathcal{C}} g(f,c)\,\eta_{\varepsilon_1}(c-\hat{c})\ \mathrm{d}\mu_f(\hat{c})\ \mathrm{d}c\ \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}}(f)\\
&\overset{\text{Fubini}}{=} \frac{\partial}{\partial p_i}\int_{\mathcal{C}}\int_{\mathcal{F}}\int_{\mathcal{C}} g(f,c)\,\eta_{\varepsilon_1}(c-\hat{c})\ \mathrm{d}\mu_f(\hat{c})\ \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}}(f)\ \mathrm{d}c\\
&\overset{\substack{\text{disintegration,}\\ \text{cf. Remark 5.31}}}{=} \frac{\partial}{\partial p_i}\int_{\mathcal{C}}\int_{\mathcal{F}\times\mathcal{C}} g(f,c)\,\eta_{\varepsilon_1}(c-\hat{c})\ \mathrm{d}\boldsymbol{I}_{\#}\lambda(\hat{c},f)\ \mathrm{d}c\\
&\overset{\text{Theorem 5.24}}{=} \frac{\partial}{\partial p_i}\int_{\mathcal{C}}\int_{\Omega_0\times\Omega_T} g\left(I_1^{\mathrm{d}}(\omega,(\boldsymbol{x},t)),c\right)\eta_{\varepsilon_1}(c-I_2(\boldsymbol{p},\boldsymbol{x},t))\ \mathrm{d}\lambda(\omega,(\boldsymbol{x},t))\ \mathrm{d}c\\
&\overset{\substack{\text{re-order differentiation}\\ \text{and integration (*)}}}{=} -\int_{\mathcal{C}}\int_{\Omega_0\times\Omega_T} g\left(I_1^{\mathrm{d}}(\omega,(\boldsymbol{x},t)),c\right)\eta'_{\varepsilon_1}(c-I_2(\boldsymbol{p},\boldsymbol{x},t))\underbrace{\frac{\partial I_2(\boldsymbol{p},\boldsymbol{x},t)}{\partial p_i}}_{\substack{\text{with}\\ f:=\frac{\partial I_2(\boldsymbol{p},\boldsymbol{x},t)}{\partial p_i},\ \mathrm{d}\lambda_1:=\mathrm{d}\lambda(\omega,(\boldsymbol{x},t))\\ \mathrm{d}\lambda_2:=\mathrm{d}\left(\frac{\partial I_2(\boldsymbol{p},\cdot)}{\partial p_i}\lambda\right)(\omega,(\boldsymbol{x},t))\\ \text{in Theorem 5.4}}}\ \mathrm{d}\lambda(\omega,(\boldsymbol{x},t))\ \mathrm{d}c\\
&\overset{\text{Theorem 5.4}}{=} -\int_{\mathcal{C}}\int_{\Omega_0\times\Omega_T} g\left(I_1^{\mathrm{d}}(\omega,(\boldsymbol{x},t)),c\right)\eta'_{\varepsilon_1}(c-I_2(\boldsymbol{p},\boldsymbol{x},t))\ \mathrm{d}\left(\frac{\partial I_2(\boldsymbol{p},\cdot)}{\partial p_i}\lambda\right)(\omega,(\boldsymbol{x},t))\ \mathrm{d}c\\
&\overset{\text{Theorem 5.24}}{=} -\int_{\mathcal{C}}\int_{\mathcal{F}\times\mathcal{C}} g(f,c)\,\eta'_{\varepsilon_1}(c-\hat{c})\ \mathrm{d}\boldsymbol{I}_{\#}\left(\frac{\partial I_2(\boldsymbol{p},\cdot)}{\partial p_i}\lambda\right)(\hat{c},f)\ \mathrm{d}c\\
&\overset{\text{Fubini}}{=} -\int_{\mathcal{F}\times\mathcal{C}}\int_{\mathcal{C}} g(f,c)\,\eta'_{\varepsilon_1}(c-\hat{c})\ \mathrm{d}\boldsymbol{I}_{\#}\left(\frac{\partial I_2(\boldsymbol{p},\cdot)}{\partial p_i}\lambda\right)(\hat{c},f)\ \mathrm{d}c = \langle g, \mathcal{K}^{\varepsilon_1}\rangle
\end{aligned}$$

with

$$\mathcal{K}^{\varepsilon_1} := -\hat{\eta}'_{\varepsilon_1} *_{\mathcal{C}} \boldsymbol{I}_{\#}\left(\frac{\partial I_2}{\partial p_i}\lambda\right)$$

by considering a disintegration $(\hat{\boldsymbol{\mu}}, \hat{\nu})$ along the space $\boldsymbol{\mathcal{F}}$ of the pushforward of the measure $\frac{\partial I_2}{\partial p_i}\lambda$ with respect to the joint mapping $\boldsymbol{I}$. In this sense, $\hat{\nu}$ denotes a probability measure on the measurable space $(\boldsymbol{\mathcal{F}}, \mathcal{B}(\boldsymbol{\mathcal{F}}))$ and $\hat{\boldsymbol{\mu}} = \left\{\hat{\mu}_f\right\}_{f\in\boldsymbol{\mathcal{F}}}$ a family of appropriately scaled measures on $(\boldsymbol{\mathcal{C}}, \mathcal{B}(\boldsymbol{\mathcal{C}}))$.
For the sake of completeness, we remark that it is indeed valid to change the order of integration and differentiation in (*) as the integrand is a continuous function and the integration variables do not depend on $\boldsymbol{p}$. □

The partial derivative of the histogram with respect to a parameter $p_i \in \boldsymbol{p}$ is an essential ingredient, as we will see in the next section. Since we are aiming for a gradient-based numerical solver for the MI optimization problem, we will focus next on the gradient of MI.

5.2.5 Discretized histograms, discrete MI and its gradient

To solve the optimization problem numerically, an intuitive approach is applying gradient-based optimization techniques as for example gradient descent. So, this section is mainly dedicated to deriving the gradient of MI. Before we dive right into the details, we focus on different aspects related to discrete histogram definitions first to introduce a definition of mutual information in a discrete context.

Discrete histograms and discrete MI

We begin with a discretized histogram that we will use to define MI in a discretized setting. Firstly, it is a valid assumption to use a discrete histogram because in numerics we are unconditionally bound to use a finite number of entries to approximate continuous values and also to perform calculations with a discrete histogram. So it is an essential consideration when aiming for a computer aided solution. Secondly, we are facing a limited amount of data to generate our histograms from as well, since both the classification and the features images are living on a discrete pixel grid in the actual application setting anyway. This is already an outlook to the consequent Section 5.3 where we will motivate the use of various discretization aspects more thoroughly. Still we want to clarify the discretization of the feature and classifaction spaces now to prepare the definition of a discrete histogram.
We state that in this section, we focus on the reduced feature space $\boldsymbol{\mathcal{F}}'$ again omitting features of very low probabilities as defined in Definition 3.14 and we use $\boldsymbol{\mathcal{C}}'$ defined in Definition 4.5 and, equivalently, given as

$$\boldsymbol{\mathcal{C}}' := \operatorname{supp}(p_{\boldsymbol{\mathcal{C}}}) \subset \boldsymbol{\mathcal{C}}.$$

We stress here that it is a valid assumption that the support of the probability density function $p_{\boldsymbol{\mathcal{C}}}$ is bounded since the classification image I_2 as defined in Definition 4.4 is only mapping to a subset of $[0, 2]$.

Definition 5.41 (Discretization of $\mathcal{F}'$ and $\mathcal{C}'$)
We consider a discrete setting in our joint feature-classification space $\mathcal{F}' \times \mathcal{C}'$ with pairwise disjoint discrete bins $\mathcal{B}_{\mathcal{F},i}$ in $\mathcal{F}'$ and $\mathcal{B}_{\mathcal{C},j}$ in $\mathcal{C}'$, respectively, with $i = 1, \dots, N_{\mathcal{F}}$ and $j = 1, \dots, N_{\mathcal{C}}$ resulting in finite discretizations for $\mathcal{F}'$ and $\mathcal{C}'$ with

$$\bigcup_{i=1}^{N_{\mathcal{F}}} \mathcal{B}_{\mathcal{F},i} = \mathcal{F}', \qquad \bigcup_{j=1}^{N_{\mathcal{C}}} \mathcal{B}_{\mathcal{C},j} = \mathcal{C}'.$$

When using equidistant binning widths Δc and Δf, we can calculate the numbers of necessary bins

$$N_{\mathcal{C}} = \frac{|\mathcal{C}'|}{\Delta c}, \qquad N_{\mathcal{F}} = \prod_{i=1}^{n} \frac{|\mathcal{F}'_i|}{\Delta f}$$

when assuming the n-dimensional reduced feature space to be given as the Cartesian product $\mathcal{F}' = \mathcal{F}'_1 \times \cdots \times \mathcal{F}'_n$.

In case of our three dimensional feature space based on the texture information introduced in Section 3.3, we can calculate the number of feature bins as follows

$$N_{\mathcal{F}} = \frac{|\mathcal{F}'_1|}{\Delta f} \cdot \frac{|\mathcal{F}'_2|}{\Delta f} \cdot \frac{|\mathcal{F}'_3|}{\Delta f} \tag{5.8}$$

which we will use in the further course. In case of a classification range between 0 and 2 we can approximate the number of bins in the classification domain with

$$N_{\mathcal{C}} = \frac{|\mathcal{C}'|}{\Delta c} \leq \frac{2}{\Delta c}. \tag{5.9}$$

Without loss of generality, we consider only bin width Δc and Δf such that $N_{\mathcal{C}}$ and $N_{\mathcal{F}}$ are indeed integer values.
The discrete histograms can be defined as follows when applying the just introduced binning strategy.

Definition 5.42 (Discrete histograms — three versions)
We begin with the discrete histograms defined as arrays and mark them with bars. Each entry is given by

$$\begin{aligned}
\overline{\boldsymbol{H}}_{\mathcal{F}}(i) &= \left(I_{1\,\#}^{\mathrm{d}}\lambda\right)(\mathcal{B}_{\mathcal{F},i}), && i = 1, \dots, N_{\mathcal{F}} \\
\overline{\boldsymbol{H}}_{\mathcal{C}}(j) &= \left(I_2(\boldsymbol{p}, \cdot)_{\#}\kappa\right)(\mathcal{B}_{\mathcal{C},j}), && j = 1, \dots, N_{\mathcal{C}} \\
\overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j) &= \left(\boldsymbol{I}(\boldsymbol{p}, \cdot)_{\#}\lambda\right)(\mathcal{B}_{\mathcal{F},i} \times \mathcal{B}_{\mathcal{C},j}), && i = 1, \dots, N_{\mathcal{F}},\ j = 1, \dots, N_{\mathcal{C}}
\end{aligned}$$

In these arrays, we collect in each entry for the corresponding bin the pushforward of λ with respect to I_1^{d} and $\boldsymbol{I}$ and, respectively, the pushforward of κ with respect to I_2, i.e. the measure of the preimages that get mapped into the bin corresponding to that entry. Based on these arrays, we define histogram measures $\hat{\boldsymbol{H}}_{\mathcal{F}}$, $\hat{\boldsymbol{H}}_{\mathcal{C}}$ and $\hat{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}$ that are non-zero on the $\mathcal{F}'$, $\mathcal{C}'$ and $\mathcal{F}' \times \mathcal{C}'$. For

each measurable $C \subset \mathcal{C}'$ and $F \subset \mathcal{F}'$, we implement the following relation between the array-like histograms and the new histogram measures:

$$\hat{\boldsymbol{H}}_{\mathcal{F}}(F) = \sum_{i=1}^{N_{\mathcal{F}}} \frac{|F \cap \mathcal{B}_{\mathcal{F},i}|}{|\mathcal{B}_{\mathcal{F},i}|} \overline{\boldsymbol{H}}_{\mathcal{F}}(i) = \frac{1}{\Delta f^3} \sum_{i=1}^{N_{\mathcal{F}}} |F \cap \mathcal{B}_{\mathcal{F},i}| \, \overline{\boldsymbol{H}}_{\mathcal{F}}(i),$$

$$\hat{\boldsymbol{H}}_{\mathcal{C}}(C) = \sum_{j=1}^{N_{\mathcal{C}}} \frac{|C \cap \mathcal{B}_{\mathcal{C},j}|}{|\mathcal{B}_{\mathcal{C},j}|} \overline{\boldsymbol{H}}_{\mathcal{C}}(j) = \frac{1}{\Delta c} \sum_{j=1}^{N_{\mathcal{C}}} |C \cap \mathcal{B}_{\mathcal{C},j}| \, \overline{\boldsymbol{H}}_{\mathcal{C}}(j),$$

$$\hat{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(F \times C) = \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \frac{|F \cap \mathcal{B}_{\mathcal{F},i}|}{|\mathcal{B}_{\mathcal{F},i}|} \frac{|C \cap \mathcal{B}_{\mathcal{C},j}|}{|\mathcal{B}_{\mathcal{C},j}|} \overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j) = \frac{1}{\Delta c \Delta f^3} \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} |F \cap \mathcal{B}_{\mathcal{F},i}| \, |C \cap \mathcal{B}_{\mathcal{C},j}| \, \overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j).$$

In this context $|\cdot|$ is used to indicate the Lebesgue measure. Additionally, we stress that they are labeled with the hat symbol whereas the original, continuous histogram measures $\boldsymbol{H}$ are bare bold letters lacking any symbol on top (cf. Definition 5.25).
Finally, we introduce piecewise constant histogram density functions with

$$\hat{h}_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}_+ \qquad \hat{h}_{\mathcal{F}}(f) = \frac{1}{\Delta f^3} \sum_{i=1}^{N_{\mathcal{F}}} \mathbf{1}_{\mathcal{B}_{\mathcal{F},i}}(f) \, \overline{\boldsymbol{H}}_{\mathcal{F}}(i)$$

$$\hat{h}_{\mathcal{C}} : \mathcal{C} \to \mathbb{R}_+ \qquad \hat{h}_{\mathcal{C}}(c) = \frac{1}{\Delta c} \sum_{j=1}^{N_{\mathcal{C}}} \mathbf{1}_{\mathcal{B}_{\mathcal{C},j}}(c) \, \overline{\boldsymbol{H}}_{\mathcal{C}}(j)$$

$$\hat{h}_{\mathcal{F}\times\mathcal{C}} : \mathcal{F} \times \mathcal{C} \to \mathbb{R}_+ \qquad \hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c) = \frac{1}{\Delta c \Delta f^3} \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \mathbf{1}_{\mathcal{B}_{\mathcal{F},i}}(f) \, \mathbf{1}_{\mathcal{B}_{\mathcal{C},j}}(c) \, \overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j).$$

On each bin $\mathcal{B}_{\mathcal{F},i}$, $\mathcal{B}_{\mathcal{C},j}$ or $\mathcal{B}_{\mathcal{F},i} \times \mathcal{B}_{\mathcal{C},j}$ for $i = 1, \dots, N_{\mathcal{F}}$ and $j = 1, \dots, N_{\mathcal{C}}$ those density functions are constant and outside of $\mathcal{F}', \mathcal{C}'$ or $\mathcal{F}' \times \mathcal{C}'$, respectively, they are 0.

We remark that the discrete histogram measures are based on a normalization step by the bin sizes Δf^3 and Δc to preserve the "total mass". This motivates the following lemma.

Lemma 5.43 (Mass conservation in discrete histograms)
For the array-like discrete histogram and the derived discrete histogram measures, the total mass in both histograms is identical. Consequently, the following identities hold:

$$\sum_{i=1}^{N_{\mathcal{F}}} \overline{\boldsymbol{H}}_{\mathcal{F}}(i) = \hat{\boldsymbol{H}}_{\mathcal{F}}\left(\mathcal{F}'\right),$$

$$\sum_{j=1}^{N_{\mathcal{C}}} \overline{\boldsymbol{H}}_{\mathcal{C}}(j) = \hat{\boldsymbol{H}}_{\mathcal{C}}\left(\mathcal{C}'\right),$$

$$\sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j) = \hat{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}\left(\mathcal{F}' \times \mathcal{C}'\right).$$

Furthermore, for the histograms related to the reduced feature space it holds that they approximate the total mass of the spatio-temporal domain Ω_T, i.e. the following inequalities

$$\hat{\boldsymbol{H}}_{\mathcal{F}}\left(\mathcal{F}'\right) \leq |\Omega_T|$$

$$\hat{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}\left(\mathcal{F}' \times \mathcal{C}'\right) \leq |\Omega_T|$$

hold and for the histogram related to the classification space the equality

$$\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{C}}}\left(\boldsymbol{\mathcal{C}}'\right) = |\Omega_T|$$

holds with the Lebesgue measure κ of the spatio-temporal domain given by $|\Omega_T| = n_T \kappa(\Omega) = n_T \cdot L \cdot W$.

Proof. We show the arguments exemplarily in the feature space $\boldsymbol{\mathcal{F}}'$. In the classification space $\boldsymbol{\mathcal{C}}'$ and in the joint space $\boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}'$ the statements follow similarly. We start on the right hand side with the discrete histogram measures and derive

$$\begin{aligned}
\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}\left(\boldsymbol{\mathcal{F}}'\right) &\overset{\text{Definition 5.41}}{=} \sum_{i=1}^{N_{\boldsymbol{\mathcal{F}}}} \hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}\left(\mathcal{B}_{\boldsymbol{\mathcal{F}},i}\right) \\
&\overset{\text{Definition 5.42}}{=} \sum_{i=1}^{N_{\boldsymbol{\mathcal{F}}}} \sum_{i_2=1}^{N_{\boldsymbol{\mathcal{F}}}} \frac{\left|\mathcal{B}_{\boldsymbol{\mathcal{F}},i} \cap \mathcal{B}_{\boldsymbol{\mathcal{F}},i_2}\right|}{\left|\mathcal{B}_{\boldsymbol{\mathcal{F}},i_2}\right|} \overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}\left(i_2\right) \\
&\overset{\text{bins pw. disjoint}}{=} \sum_{i=1}^{N_{\boldsymbol{\mathcal{F}}}} \frac{\left|\mathcal{B}_{\boldsymbol{\mathcal{F}},i} \cap \mathcal{B}_{\boldsymbol{\mathcal{F}},i}\right|}{\left|\mathcal{B}_{\boldsymbol{\mathcal{F}},i}\right|} \overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}\left(i\right) = \sum_{i=1}^{N_{\boldsymbol{\mathcal{F}}}} \overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}\left(i\right)
\end{aligned}$$

which shows the statement of mass conservation for the feature space histograms.
For the second statement in the setting of the feature space $\boldsymbol{\mathcal{F}}'$, we derive

$$\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}\left(\boldsymbol{\mathcal{F}}'\right) = \sum_{i=1}^{N_{\boldsymbol{\mathcal{F}}}} \overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}\left(i\right) \overset{\text{Definition 5.42}}{=} \sum_{i=1}^{N_{\boldsymbol{\mathcal{F}}}} \left(I_{1\,\#}^{\mathrm{d}}\lambda\right)\left(\mathcal{B}_{\boldsymbol{\mathcal{F}},i}\right) \overset{\text{additivity}}{=} \left(I_{1\,\#}^{\mathrm{d}}\lambda\right)\left(\boldsymbol{\mathcal{F}}'\right)$$

by exploiting that the bins are considered to be pairwise disjoint as well as that their union results in the total space $\boldsymbol{\mathcal{F}}'$ again. Similarly, one can derive

$$\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(\boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}'\right) = \left(\boldsymbol{I}_{\#}\lambda\right)\left(\boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}'\right).$$

With this in mind, we continue with the following estimations

$$\begin{aligned}
\left(I_{1\,\#}^{\mathrm{d}}\lambda\right)\left(\boldsymbol{\mathcal{F}}'\right) = \left(I_{1\,\#}^{\mathrm{d}}\left(P^0 \otimes \kappa\right)\right)\left(\boldsymbol{\mathcal{F}}'\right) &\leq \left(I_{1\,\#}^{\mathrm{d}}\left(P^0 \otimes \kappa\right)\right)\left(\boldsymbol{\mathcal{F}}\right) \\
&= \int_{\Omega_0 \times \Omega_T} \underbrace{\mathbf{1}_{\boldsymbol{\mathcal{F}}}\left(I_1^{\mathrm{d}}\left(\omega,(\boldsymbol{x},t)\right)\right)}_{=1} \mathrm{d}\left(P^0 \otimes \kappa\right)\left(\omega,(\boldsymbol{x},t)\right) \\
&\overset{\text{Fubini's theorem}}{=} \underbrace{\int_{\Omega_0} 1 \,\mathrm{d}P^0\left(\omega\right)}_{=1} \int_{\Omega_T} 1 \,\mathrm{d}\kappa = \kappa\left(\Omega_T\right)
\end{aligned}$$

and, analogously, it holds that

$$\begin{aligned}
\left(\boldsymbol{I}_{\#}\lambda\right)\left(\boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}'\right) &= \left(\boldsymbol{I}_{\#}\left(P^0 \otimes \kappa\right)\right)\left(\boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}'\right) \\
&\leq \left(\boldsymbol{I}_{\#}\left(P^0 \otimes \kappa\right)\right)\left(\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}\right) \\
&= P^0\left(\Omega_0\right) \cdot \kappa\left(\Omega_T\right) = \kappa\left(\Omega_T\right).
\end{aligned}$$

Here, we exploit the fact that the pushforwards with respect to the mappings I_1^{d} and $\boldsymbol{I}$ considering the total spaces $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$ equal the total mass of the original spatio-temporal domain Ω_T. For the histogram related only to the classification space $\boldsymbol{\mathcal{C}}'$, we even get the equality

$$\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{C}}}\left(\boldsymbol{\mathcal{C}}'\right) = \left(I_{2\#}\kappa\right)\left(\boldsymbol{\mathcal{C}}'\right) = \kappa\left(\Omega_T\right)$$

because $\boldsymbol{\mathcal{C}}'$ contains the support of I_2. By recalling that $\kappa(\Omega_T) = n_T\,\kappa(\Omega) = n_T \cdot L \cdot W$ holds, we have shown the complete statement. □

Remark 5.44 (Missed mass due to neglected lower probability features). We only approximate the total mass of the spatio-temporal domain Ω_T with the histogram measures $\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}(\boldsymbol{\mathcal{F}}')$ and $\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(\boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}')$ since we are neglecting features of very low probabilities in the reduced feature space $\boldsymbol{\mathcal{F}}'$. Consequently, we miss certain areas in the spatio-temporal domain which are indeed mapped to these features of very low probabilities. We avoid introducing a new, adjusted definition for the disturbed feature image I_1^{d} to replace the feature values of very low probabilities by others which are lying within $\boldsymbol{\mathcal{F}}'$. Instead, we assume that due to the very low probabilities of the missed features in $\boldsymbol{\mathcal{F}}\backslash\boldsymbol{\mathcal{F}}'$, we also miss only very few "mass" of the spatio-temporal domain Ω_T and we infer that we approximate the total mass well. In this sense, we use the approximations

$$\left(I_{1\,\#}^{\mathrm{d}}\lambda\right)\left(\boldsymbol{\mathcal{F}}'\right) \approx |\Omega_T|\,, \qquad\qquad \left(\boldsymbol{I}_{\#}\lambda\right)\left(\boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}'\right) \approx |\Omega_T| \tag{5.10}$$

in the further course.

Next, we introduce the following relation between the piecewise constant histogram density functions and the discrete histogram measures in the next lemma.

Lemma 5.45
The density functions $\hat{h}_{\boldsymbol{\mathcal{F}}}$, $\hat{h}_{\boldsymbol{\mathcal{C}}}$ and $\hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ defined in Definition 5.42 are density functions with respect to the Lebesgue measure for the discrete histogram measures $\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}$, $\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{C}}}$ and $\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ such that it holds

$$\begin{aligned}
\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}(F) &= \int_F \hat{h}_{\boldsymbol{\mathcal{F}}}(f)\ \mathrm{d}f,\\
\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{C}}}(C) &= \int_C \hat{h}_{\boldsymbol{\mathcal{C}}}(c)\ \mathrm{d}c,\\
\hat{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(F\times C) &= \int_{F\times C} \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f,c)\ \mathrm{d}(f,c)
\end{aligned}$$

for arbitrary measurable subsets $C \subset \boldsymbol{\mathcal{C}}'$ and $F \subset \boldsymbol{\mathcal{F}}'$.

Proof. We show the identity for the statement corresponding to the feature space $\mathcal{F}'$. Let F be an arbitrary measurable subset of $\mathcal{F}'$. We start with the identity on the right hand side and perform the following transformations:

$$\begin{aligned}
\int_F \hat{h}_{\mathcal{F}}(f)\ \mathrm{d}f &\overset{\text{cf. Definition 5.42}}{=} \int_F \frac{1}{\Delta f^3} \sum_{i=1}^{N_{\mathcal{F}}} \mathbf{1}_{\mathcal{B}_{\mathcal{F},i}}(f)\, \overline{\boldsymbol{H}}_{\mathcal{F}}(i)\ \mathrm{d}f \\
&\frac{1}{\Delta f^3} \sum_{i=1}^{N_{\mathcal{F}}} \overline{\boldsymbol{H}}_{\mathcal{F}}(i) \int_F \mathbf{1}_{\mathcal{B}_{\mathcal{F},i}}(f)\ \mathrm{d}f \\
&\frac{1}{\Delta f^3} \sum_{i=1}^{N_{\mathcal{F}}} \overline{\boldsymbol{H}}_{\mathcal{F}}(i) \int_{F \cap \mathcal{B}_{\mathcal{F},i}} \mathrm{d}f \\
&= \frac{1}{\Delta f^3} \sum_{i=1}^{N_{\mathcal{F}}} \overline{\boldsymbol{H}}_{\mathcal{F}}(i)\, |F \cap \mathcal{B}_{\mathcal{F},i}| \overset{\text{cf. Definition 5.42}}{=} \hat{\boldsymbol{H}}_{\mathcal{F}}(F).
\end{aligned}$$

The remaining two statements for the spaces $\mathcal{C}'$ and $\mathcal{F}' \times \mathcal{C}'$ can be shown similarly. □

In the later Section 5.3.4 on discretization aspects, we will add a convergence result for the discretized histograms, more precisely for their histogram density functions when applying binning widths Δf, Δc converging to 0. In the current section, we focus next on the relation between the histograms for "one" space and the histogram on the joint space.

Lemma 5.46 (Relation between individual and joint histogram)
The relation between the individual histograms for the continuous histogram measures and for the discretized array-valued histograms is given as follows:

1. Based on the joint histogram measure $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}$, it holds for the histograms $\boldsymbol{H}_{\mathcal{F}}$ and $\boldsymbol{H}_{\mathcal{C}}$ and for arbitrary measurable subsets $A'_F \subset \mathcal{F}'$ and $A'_C \subset \mathcal{C}'$ that

$$\begin{aligned}
\boldsymbol{H}_{\mathcal{F}}\left(A'_F\right) &= \int_{A'_F \times \mathcal{C}'} 1\ \mathrm{d}\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}(f,c), \\
\boldsymbol{H}_{\mathcal{C}}\left(A'_C\right) &= \int_{\mathcal{F}' \times A'_C} 1\ \mathrm{d}\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}(f,c),
\end{aligned}$$

   i.e., the individual histograms are calculated by applying integration along the other space.

2. Based on the discretized joint histogram $\overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}$ as an array, it holds for the individual, array-valued histograms that

$$\overline{\boldsymbol{H}}_{\mathcal{F}}(i) = \sum_{j=1}^{N_{\mathcal{C}}} \overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j), \qquad i = 1,\dots,N_{\mathcal{F}} \tag{5.11}$$

$$\overline{\boldsymbol{H}}_{\mathcal{C}}(j) = \sum_{i=1}^{N_{\mathcal{F}}} \overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j), \qquad j = 1,\dots,N_{\mathcal{C}}. \tag{5.12}$$

Proof. We start with the proof of the first part. Let A'_F be an arbitrary subset of $\boldsymbol{\mathcal{F}'}$. When we perform the following transformations

$$\begin{aligned}\int_{A'_F\times\boldsymbol{\mathcal{C}'}} 1\,\mathrm{d}\boldsymbol{H}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f,c) &= \int_{A'_F\times\boldsymbol{\mathcal{C}'}} 1\,\mathrm{d}\left(\lambda\circ\boldsymbol{I}^{-1}\right)(f,c)\\ &\overset{(*)}{=} \int_{\boldsymbol{I}^{-1}\left(A'_F\times\boldsymbol{\mathcal{C}'}\right)=\left(I_1^{\mathrm{d}}\right)^{-1}\left(A'_F\right)} 1\,\mathrm{d}\lambda\left(\omega,(\boldsymbol{x},t)\right)\\ &= \int_{A'_F} 1\,\mathrm{d}\left(\lambda\circ I_1^{\mathrm{d}}\right)(f,c) \overset{\text{Definition 5.27}}{=} \boldsymbol{H}_{\boldsymbol{\mathcal{F}}}\left(A'_F\right)\end{aligned}$$

and use in $(*)$ the following identity for the integration domain

$$\boldsymbol{I}^{-1}\left(A'_F\times\boldsymbol{\mathcal{C}'}\right)=\left(I_1^{\mathrm{d}},I_2\right)^{-1}\left(A'_F\times\boldsymbol{\mathcal{C}'}\right)=\left(I_1^{\mathrm{d}}\right)^{-1}\left(A'_F\right)\cap\left(\Omega_0\times\Omega_T\right)=\left(I_1^{\mathrm{d}}\right)^{-1}\left(A'_F\right),$$

we finally end up with the definition of one individual histogram.
To show the identity

$$\boldsymbol{H}_{\boldsymbol{\mathcal{C}}}\left(A'_C\right)=\int_{\boldsymbol{\mathcal{F}'}\times A'_C} 1\,\mathrm{d}\boldsymbol{H}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f,c)$$

we can apply similar transformations related to I_2 on $\boldsymbol{\mathcal{C}'}$ to obtain the definition for $\boldsymbol{H}_{\boldsymbol{\mathcal{C}}}$ stated in Definition 5.27.

✓

The second part can be derived in a similar way when considering summation over one-dimension of the two dimensional joint histogram-array $\overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ and thinking about one fixed index i or j as a "subdomain" of the corresponding joint space similar to the integration along one subdomain $\boldsymbol{\mathcal{F}'}$ or $\boldsymbol{\mathcal{C}'}$ above. We show the statement again for the histogram for the feature space, $\overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}$ while the statement for the histogram $\overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{C}}}$ can be derived analogously. For an arbitrary $i\in\{1,\ldots,N_{\boldsymbol{\mathcal{F}}}\}$, it holds that

$$\begin{aligned}\sum_{j=1}^{N_{\boldsymbol{\mathcal{C}}}}\overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(i,j) &= \sum_{j=1}^{N_{\boldsymbol{\mathcal{C}}}}\left(\boldsymbol{I}(\boldsymbol{p},\cdot)_{\#}\lambda\right)\left(\mathcal{B}_{\boldsymbol{\mathcal{F}},i}\times\mathcal{B}_{\boldsymbol{\mathcal{C}},j}\right)\\ &= \sum_{j=1}^{N_{\boldsymbol{\mathcal{C}}}}\int_{\mathcal{B}_{\boldsymbol{\mathcal{F}},i}\times\mathcal{B}_{\boldsymbol{\mathcal{C}},j}} 1\,\mathrm{d}\left(\boldsymbol{I}(\boldsymbol{p},\cdot)_{\#}\lambda\right)\\ &= \int_{\mathcal{B}_{\boldsymbol{\mathcal{F}},i}\times\boldsymbol{\mathcal{C}'}} 1\,\mathrm{d}\left(\boldsymbol{I}(\boldsymbol{p},\cdot)_{\#}\lambda\right)\\ &= \ldots = \overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}(i).\end{aligned}$$

We skipped in the end a few steps. However, the statement can be indeed proved like this by applying disintegration of the joint measure $\left(\boldsymbol{I}(\boldsymbol{p},\cdot)_{\#}\lambda\right)$ and by considering that the preimage of $\boldsymbol{\mathcal{C}'}$ for I_2 is again the whole spatio-temporal domain Ω_T which then allows to only focus on the preimage of $\mathcal{B}_{\boldsymbol{\mathcal{F}},i}$ under I_1^{d}.

✓

□

In the initial definition of MI in Definition 5.12, we used probability density functions related to the occurring classification values for a certain parameter settings $\boldsymbol{p} \in \boldsymbol{P}$ in the classification image I_2 as well as probability density function for the feature values occurring in the feature image I_1^{d}, which incorporate the original microscopy data. Moreover, we used the corresponding joint probability density function related to the joint mapping $\boldsymbol{I} = \left(I_1^{\mathrm{d}}, I_2\right)$. With these probability density functions (PDFs), we can calculate the mutual information as stated in Definition 5.12. Before we define a discretized version of mutual information, we establish the relation between the probability density functions and the discretized histograms.

Lemma 5.47 (Relation between probability density functions and histogram density functions) The (piecewise constant) probability density functions can be calculated based on the discretized (piecewise constant) histogram density functions $\hat{h}_{\mathcal{F}}$, $\hat{h}_{\mathcal{C}}$ and $\hat{h}_{\mathcal{F}\times\mathcal{C}}$ via a normalization step such that

$$\begin{aligned}
\hat{p}_{\mathcal{F}} &: \mathcal{F} \to \mathbb{R}_+ & \hat{p}_{\mathcal{F}}(f) &= \frac{1}{n_T \cdot L \cdot W} \hat{h}_{\mathcal{F}}(f) \\
\hat{p}_{\mathcal{C}} &: \mathcal{C} \to \mathbb{R}_+ & \hat{p}_{\mathcal{C}}(c) &= \frac{1}{n_T \cdot L \cdot W} \hat{h}_{\mathcal{C}}(c) \\
\hat{p}_{\mathcal{F}\times\mathcal{C}} &: \mathcal{F} \times \mathcal{C} \to \mathbb{R}_+ & \hat{p}_{\mathcal{F}\times\mathcal{C}}(f,c) &= \frac{1}{n_T \cdot L \cdot W} \hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c)
\end{aligned}$$

holds with the normalization constant equal to the total mass of the domain Ω_T

$$|\Omega_T| = n_T \cdot L \cdot W.$$

Proof. The functions $\hat{p}_{\mathcal{F}}$, $\hat{p}_{\mathcal{C}}$ and $\hat{p}_{\mathcal{F}\times\mathcal{C}}$ introduced with the histogram density functions are indeed probability density functions, since it holds that

$$\hat{p}_{\mathcal{F}}, \hat{p}_{\mathcal{C}}, \hat{p}_{\mathcal{F}\times\mathcal{C}} \geq 0$$

because of the non-negativity of the different histogram measures and the histogram density function introduced in Definition 5.42. Moreover, they are integrable similarly to the histogram density functions and with

$$\begin{aligned}
\int_{\mathcal{F}} \hat{p}_{\mathcal{F}}(f)\ \mathrm{d}f &= \int_{\mathcal{F}} \frac{1}{n_T \cdot L \cdot W} \hat{h}_{\mathcal{F}}(f)\ \mathrm{d}f \\
&\overset{\text{Definition 5.42}}{=} \frac{1}{n_T \cdot L \cdot W} \int_{\mathcal{F}} \frac{1}{\Delta f^3} \sum_{i=1}^{N_{\mathcal{F}}} \mathbf{1}_{\mathcal{B}_{\mathcal{F},i}}(f)\, \overline{\boldsymbol{H}}_{\mathcal{F}}(i)\ \mathrm{d}f \\
&= \frac{1}{n_T \cdot L \cdot W} \frac{1}{\Delta f^3} \sum_{i=1}^{N_{\mathcal{F}}} \overline{\boldsymbol{H}}_{\mathcal{F}}(i) \int_{\mathcal{B}_{\mathcal{F},i}} \mathrm{d}f \\
&= \frac{1}{n_T \cdot L \cdot W} \frac{1}{\Delta f^3} \overline{\boldsymbol{H}}_{\mathcal{F}}(i) \sum_{i=1}^{N_{\mathcal{F}}} \underbrace{\int_{\mathcal{B}_{\mathcal{F},i_2}} \mathrm{d}f}_{=\Delta f^3} \\
&= \frac{1}{n_T \cdot L \cdot W} \sum_{i=1}^{N_{\mathcal{F}}} \overline{\boldsymbol{H}}_{\mathcal{F}}(i) \overset{\text{cf. Lemma 5.43}}{\approx} 1.
\end{aligned}$$

we show that $\hat{p}_{\mathcal{F}}$ is also normalized. That $\hat{p}_{\mathcal{C}}$ and $\hat{p}_{\mathcal{F}\times\mathcal{C}}$ are normalized can equivalently be shown with the cited statements. We remark that for $\hat{p}_{\mathcal{F}}$ and $\hat{p}_{\mathcal{F}\times\mathcal{C}}$, we exploit the fact that the total mass is approximated by the given histograms well enough when considering to cut off only features of very low probabilities to get the underlying feature space $\mathcal{F}'$ (cf. Remark 5.44). □

Without going into the details of discretizing the spatio-temporal domain Ω_T, we still introduce the following relation. We focus on the discretization details of Ω_T later on in Section 5.3, especially in Definition 5.55.

Remark 5.48. As our spatio-temporal domain Ω_T is already semi-discretized because we only consider discrete time points $t \in \{t_1, \dots, t_{n_T}\}$ (cf. Definition 3.1), we only need to discretize further the spatial domain Ω. For this purpose, we apply an equidistant spatial width h. Consequently, the volume of one discrete pixel equals h^2. With this at hand, we add another relation for the normalization constant given as the measure of the spatio-temporal domain Ω_T

$$n_T \cdot L \cdot W = |\Omega_T| = \sum_{t_i=1}^{n_T} \int_{\Omega} 1 \, \mathrm{d}\boldsymbol{x} = n_T \sum_{i=1}^{n_{\tilde{p}}} \mathrm{h}^2 = n_T \cdot n_{\tilde{p}} \cdot \mathrm{h}^2 \tag{5.13}$$

with n_T denoting the number of discrete time points and, similarly, $n_{\tilde{p}}$ the number of spatial pixels in Ω with area h^2.
Since the discrete, array valued histograms $\overline{\boldsymbol{H}}_{\mathcal{F}}$, $\overline{\boldsymbol{H}}_{\mathcal{C}}$ and $\overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}$ accumulate the pushforward measures with respect to the mappings I_1^{d}, I_2 and $\boldsymbol{I}$, of one bin per array entry, we can approximate the histogram entries for a discrete pixel grid of Ω_T as follows: For each histogram entry, we count the pixel that get mapped into the related bin and multiply this amount by the volume of one discrete pixel h^2. With this in mind, it is clear that when summing up all histogram entries, we end up with the total volume of our domain again.

$$\sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j) = n_T \cdot n_{\tilde{p}} \cdot \mathrm{h}^2 \overset{\text{Equation (5.13)}}{=} n_T \cdot L \cdot W$$

For a discretization of Ω_T, we add another discrete histogram definition that is based on "counting" pixels of the discretized domain Ω_T that get mapped by our mappings I_1^{d}, I_2 and $\boldsymbol{I}$ into certain bins of the spaces $\mathcal{F}'$, $\mathcal{C}'$ and $\mathcal{F}' \times \mathcal{C}'$. In this sense, we consider piecewise constant image mappings $\hat{I}_1^{\mathrm{d,h}}$, $\hat{I}_2^{\mathrm{h}}$ and $\left(\hat{I}_1^{\mathrm{d,h}}, \hat{I}_2^{\mathrm{h}}\right)$ living on the discretized pixel grid. We skip exact definitions for the discretized variables here. However, we refer to Definitions 5.55, 5.56 and 5.63 in Section 5.3.1 and Section 5.3.2 for a precise introduction of the pixel grid and the discretized image mappings. In the these upcoming sections, we investigate the discretized feature and classification images in more details with respect to convergence properties.

Definition 5.49 (Discrete histogram — counting pixels of a discretized domain Ω_T)
We define discrete, array valued histograms when considering a discretized pixel grid for Ω for each time point $t \in \{t_1, \dots, t_{n_T}\}$ and with pixel width h. Each entry is given by

$$\begin{aligned}
\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}}(i) &:= \#\left\{\text{pixel } v \,|\, \hat{I}_1^{\mathrm{d,h}}(v) \in \mathcal{B}_{\mathcal{F},i}\right\}, && i = 1, \dots, N_{\mathcal{F}} \\
\overset{\circ}{\boldsymbol{H}}_{\mathcal{C}}(j) &:= \#\left\{\text{pixel } v \,|\, \hat{I}_2^{\mathrm{h}} \in \mathcal{B}_{\mathcal{C},j}\right\}, && j = 1, \dots, N_{\mathcal{C}} \\
\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j) &:= \#\left\{\text{pixel } v \,|\, \left(\hat{I}_1^{\mathrm{d,h}}, \hat{I}_2^{\mathrm{h}}\right)(v) \in (\mathcal{B}_{\mathcal{F},i} \otimes \mathcal{B}_{\mathcal{C},j})\right\}, && i = 1, \dots, N_{\mathcal{F}},\ j = 1, \dots, N_{\mathcal{C}}.
\end{aligned}$$

We mark the new histogram version with a circle superscript. Given the discretization sizes of the spatio-temporal domain Ω_T, we can calculate the total number of histogram entries via

$$N_{\text{hist. entries}} = \frac{n_T \cdot L \cdot W}{\mathrm{h}^2}. \tag{5.14}$$

Without performing the calculations explicitly, we state that it

$$N_{\text{hist. entries}} = \sum_{i=1}^{N_{\mathcal{F}}} \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}}(i) = \sum_{j=1}^{N_{\mathcal{C}}} \overset{\circ}{\boldsymbol{H}}_{\mathcal{C}}(j) = \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j). \tag{5.15}$$

holds. With the pixel counting histogram definition we can establish the following relation to the previously defined discrete histograms introduced in Definition 5.42.

Lemma 5.50 (Relation of array-valued histograms)
Let Ω_T be discretized such that for each time point $t \in \{t_1, \dots, t_{n_T}\}$ the spatial domain Ω is given as a pixel grid of pixel width h. Then the following identities hold

$$\begin{aligned}
\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}}(i) &= \frac{1}{\mathrm{h}^2}\overline{\boldsymbol{H}}_{\mathcal{F}}(i), && i = 1, \dots, N_{\mathcal{F}}, \\
\overset{\circ}{\boldsymbol{H}}_{\mathcal{C}}(j) &= \frac{1}{\mathrm{h}^2}\overline{\boldsymbol{H}}_{\mathcal{C}}(j), && j = 1, \dots, N_{\mathcal{C}}, \\
\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j) &= \frac{1}{\mathrm{h}^2}\overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j), && i = 1, \dots, N_{\mathcal{F}},\ j = 1, \dots, N_{\mathcal{C}}.
\end{aligned}$$

Proof. The statement is a direct consequence of the corresponding definitions and considering the explanations given in Remark 5.48. □

With this relations in mind, we can now define a discretized version of MI compared to Definition 5.12 by making use of the discrete histograms.

Definition 5.51 (Discretized mutual information)
Let Ω_T be discretized such that for each time point $t \in \{t_1, \dots, t_{n_T}\}$ the spatial domain Ω is given as a pixel grid of pixel width h. Given the discrete, array-valued histograms $\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}}$, $\overset{\circ}{\boldsymbol{H}}_{\mathcal{C}}$ and $\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}$

based on pixel counting and with a finite number of entries $N_{\text{hist. entries}}$, we introduce the mutual information via

$$\mathrm{MI}^{\Delta f,\Delta c}\left(\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}\right) = \frac{1}{N_{\text{hist. entries}}} \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j) \log\left(\frac{N_{\text{hist. entries}} \cdot \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}}(i) \cdot \overset{\circ}{\boldsymbol{H}}_{\mathcal{C}}(j)}\right). \tag{5.16}$$

This definition is based on the formulation of MI in [49] and is here adapted to our notations.
With the following proposition, we show that for discrete histogram arrays $\overline{\boldsymbol{H}}_{\mathcal{F}}$, $\overline{\boldsymbol{H}}_{\mathcal{C}}$ and $\overline{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}$, the discretized MI coincides with the continuous MI derived from the piecewise constant probability distributions related to the corresponding histogram measures.

Proposition 5.52 (Matching of discrete and continuous MI)
For the discrete histogram measures $\hat{\boldsymbol{H}}_{\mathcal{F}}$, $\hat{\boldsymbol{H}}_{\mathcal{C}}$ and $\hat{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}$ and their corresponding piecewise constant probability distributions $\hat{p}_{\mathcal{F}}$, $\hat{p}_{\mathcal{C}}$ and $\hat{p}_{\mathcal{F}\times\mathcal{C}}$ it holds that

$$\mathrm{MI}\left(\hat{p}_{\mathcal{F}\times\mathcal{C}}\right) = \mathrm{MI}^{\Delta f,\Delta c}\left(\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}\right),$$

when considering the related histograms based on pixel counting $\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}}$, $\overset{\circ}{\boldsymbol{H}}_{\mathcal{C}}$ and $\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}$.

Proof. The main ingredients for the proof are stated in Definition 5.42 and Lemma 5.47. We start on the left hand side with the MI calculation based on the given probability density functions related

to the discrete histograms and perform various transformations as stated below to achieve the MI definition in a discrete setting when using the histograms based on pixel counting.

$$
\begin{aligned}
\mathrm{MI}\left(\hat{p}_{\mathcal{F}\times\mathcal{C}}\right) &= \int_{\mathcal{F}\times\mathcal{C}} \hat{p}_{\mathcal{F}\times\mathcal{C}}(f,c)\log\left(\frac{\hat{p}_{\mathcal{F}\times\mathcal{C}}(f,c)}{\hat{p}_{\mathcal{F}}(f)\cdot\hat{p}_{\mathcal{C}}(c)}\right)\mathrm{d}(f,c)\\
&\overset{\text{cf. Definition 5.41}}{=} \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \int_{\mathcal{B}_{\mathcal{F},i}}\int_{\mathcal{B}_{\mathcal{C},j}} \hat{p}_{\mathcal{F}\times\mathcal{C}}(f,c)\log\left(\frac{\hat{p}_{\mathcal{F}\times\mathcal{C}}(f,c)}{\hat{p}_{\mathcal{F}}(f)\cdot\hat{p}_{\mathcal{C}}(c)}\right)\mathrm{d}(f,c)\\
&\overset{\text{cf. Lemma 5.47}}{=} \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \int_{\mathcal{B}_{\mathcal{F},i}}\int_{\mathcal{B}_{\mathcal{C},j}} \frac{1}{n_T\cdot L\cdot W}\hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c)\log\left(\frac{\frac{1}{n_T\cdot L\cdot W}\hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c)}{\frac{1}{n_T\cdot L\cdot W}\hat{h}_{\mathcal{F}}(f)\cdot\frac{1}{n_T\cdot L\cdot W}\hat{h}_{\mathcal{C}}(c)}\right)\mathrm{d}(f,c)\\
&\overset{\substack{\text{pcw. const. of }\hat{h}'s,\\ \text{cf. Definition 5.42}}}{=} \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \frac{1}{n_T\cdot L\cdot W}\frac{\overline{H}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\Delta f^3\Delta c}\log\left(n_T\cdot L\cdot W\cdot\frac{\frac{\overline{H}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\Delta f^3\Delta c}}{\frac{\overline{H}_{\mathcal{F}}(i)}{\Delta f^3}\cdot\frac{\overline{H}_{\mathcal{C}}(j)}{\Delta c}}\right)\underbrace{\int_{\mathcal{B}_{\mathcal{F},i}}\int_{\mathcal{B}_{\mathcal{C},j}}\mathrm{d}(f,c)}_{=\Delta f^3\Delta c}\\
&= \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \frac{1}{n_T\cdot L\cdot W}\overline{H}_{\mathcal{F}\times\mathcal{C}}(i,j)\log\left(n_T\cdot L\cdot W\cdot\frac{\overline{H}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\overline{H}_{\mathcal{F}}(i)\cdot\overline{H}_{\mathcal{C}}(j)}\right)\\
&\overset{\text{cf. Lemma 5.50}}{=} \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \frac{\mathrm{h}^2}{n_T\cdot L\cdot W}\mathring{H}_{\mathcal{F}\times\mathcal{C}}(i,j)\log\left(n_T\cdot L\cdot W\cdot\frac{\mathrm{h}^2\mathring{H}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\mathrm{h}^2\mathring{H}_{\mathcal{F}}(i)\cdot\mathrm{h}^2\mathring{H}_{\mathcal{C}}(j)}\right)\\
&= \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \frac{\mathrm{h}^2}{n_T\cdot L\cdot W}\mathring{H}_{\mathcal{F}\times\mathcal{C}}(i,j)\log\left(\frac{n_T\cdot L\cdot W}{\mathrm{h}^2}\cdot\frac{\mathring{H}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\mathring{H}_{\mathcal{F}}(i)\cdot\mathring{H}_{\mathcal{C}}(j)}\right)\\
&\overset{\text{cf. Equation (5.14)}}{=} \frac{1}{N_{\text{hist. entries}}}\sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \mathring{H}_{\mathcal{F}\times\mathcal{C}}(i,j)\log\left(\frac{N_{\text{hist. entries}}\cdot\mathring{H}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\mathring{H}_{\mathcal{F}}(i)\cdot\mathring{H}_{\mathcal{C}}(j)}\right)\\
&= \mathrm{MI}^{\Delta f,\Delta c}\left(\mathring{H}_{\mathcal{F}\times\mathcal{C}}\right).
\end{aligned}
$$

□

Having now a discrete version of the MI at hand, we derive in the next paragraph the gradient for this discrete version. These derivative terms are of importance later when focusing on a numerical solution of our main optimization problem given in Definition 5.17.

Gradient for discrete MI

The discretized histograms and also a discretized definition of MI are the starting point for us to derive the gradient term of the discretized MI to use this for the numerical analysis later on in Section 5.5. The numerical solution is based on a gradient-based optimization approach. Consequently, the gradient terms are indeed essential for the upcoming analysis. First, we cite the gradient of mutual information from [49] in an adapted version to our setting and notation in the following theorem.

Theorem 5.53 (Gradient of MI)
For the discrete formulation of mutual information introduced in Definition 5.51 the gradient is given as

$$\nabla \mathrm{MI}^{\Delta f,\Delta c}(\boldsymbol{p}) = \left\{ \frac{\partial \mathrm{MI}^{\Delta f,\Delta c}}{\partial p_k} \right\}$$

for a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ with the individual parameters p_k and with

$$\frac{\partial \mathrm{MI}^{\Delta f,\Delta c}}{\partial p_k} = \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \frac{\partial \mathrm{MI}^{\Delta f,\Delta c}}{\partial \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)} \cdot \frac{\partial \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\partial p_k} \tag{5.17}$$

$$= \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \frac{1}{N_{\text{hist. entries}}} \left(\log\left(\frac{N_{\text{hist. entries}} \cdot \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\overset{\circ}{\boldsymbol{H}}_{\mathcal{C}}(j) \cdot \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}}(i)} \right) - \mathrm{MI}^{\Delta f,\Delta c} \right) \cdot \frac{\partial \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\partial p_k}, \tag{5.18}$$

if the partial derivatives of the joint histogram $\frac{\partial \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\partial p_k}$ for the parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ exist.

In this sense, we can understand that "each gradient component $[k]$ is thus expressed as the sum over all histogram entries of the change in each entry when changing $[p_k]$, weighted by the influence of this change on [the MI]" as stated by Maes et al. in [49]. In the cited paper, the authors do not state the proof for the gradient of mutual information explicitly. We add the proof here for the sake of completeness.

Proof. To derive the given gradient term we focus on each individual partial derivative $\frac{\partial \mathrm{MI}^{\Delta f,\Delta c}}{\partial p_k}$ for one parameter p_k in our parameter setting $\boldsymbol{p} \in \boldsymbol{P}$. Furthermore, we consider that the partial derivatives $\frac{\partial \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\partial p_k}$ exists for all histogram entries and every parameter $p_k \in \boldsymbol{p}$ and that they measure the change in the histogram entry $\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)$ when changing the parameter p_k. This is a valid assumption because of our smoothing steps introduced in Definition 5.33 and the derived partial derivatives (cf. Theorem 5.40).
Let p_k be an arbitrary parameter in $\boldsymbol{p}$. Considering the chain rule it follows directly that

$$\frac{\partial \mathrm{MI}^{\Delta f,\Delta c}}{\partial p_k} = \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \frac{\partial \mathrm{MI}^{\Delta f,\Delta c}}{\partial \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)} \cdot \frac{\partial \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\partial p_k}.$$

Consequently, we need to derive now the derivative term

$$\frac{\partial \mathrm{MI}^{\Delta f,\Delta c}}{\partial \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)} = \frac{1}{N_{\text{hist. entries}}} \left(\log\left(\frac{N_{\text{hist. entries}} \cdot \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}(i,j)}{\overset{\circ}{\boldsymbol{H}}_{\mathcal{C}}(j) \cdot \overset{\circ}{\boldsymbol{H}}_{\mathcal{F}}(i)} \right) - \mathrm{MI}^{\Delta f,\Delta c} \right).$$

To facilitate the proof, we introduce the following abbreviated notation:

$$\begin{aligned} N &:= N_{\text{hist. entries}} \\ H_{i,j} &:= \overset{\circ}{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(i,j) \\ H_i &:= \overset{\circ}{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}(i) \\ H_j &:= \overset{\circ}{\boldsymbol{H}}_{\boldsymbol{\mathcal{C}}}(j) \\ \mathrm{MI} &:= \mathrm{MI}^{\Delta f,\Delta c}. \end{aligned}$$

Based on the one hand on a similar statement for the pixel counting histograms $\overset{\circ}{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}$ and $\overset{\circ}{\boldsymbol{H}}_{\boldsymbol{\mathcal{C}}}$ as the second statement in Lemma 5.46 for the array valued histograms $\overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{F}}}$ and $\overline{\boldsymbol{H}}_{\boldsymbol{\mathcal{C}}}$ and on the other hand on Equation (5.15), we use the following relations with the abbreviated notation

$$\begin{aligned} H_i &= \sum_j H_{i,j} \\ H_j &= \sum_i H_{i,j} \\ N &= \sum_{i,j} H_{i,j}. \end{aligned} \tag{5.19}$$

Next, we denote the histogram entry with $H_{\hat{i},\hat{j}}$ for which we are looking for its partial derivative of the mutual information. Considering the sums in Equation (5.19), we start by stating the following partial derivatives

$$\begin{aligned} \frac{\partial N}{\partial H_{\hat{i},\hat{j}}} &= 1 \\ \frac{\partial H_i}{\partial H_{\hat{i},\hat{j}}} &= \begin{cases} 1, & i = \hat{i} \\ 0, & i \neq \hat{i} \end{cases}, \\ \frac{\partial H_j}{\partial H_{\hat{i},\hat{j}}} &= \begin{cases} 1, & j = \hat{j} \\ 0, & j \neq \hat{j} \end{cases} \end{aligned} \tag{5.20}$$

which we need for the partial derivative of the mutual information. With these prerequisites we derive the partial derivative by using the product rule, the quotient rule and the chain rule for the discrete MI (cf. Definition 5.51):

$$\begin{aligned} \frac{\partial \mathrm{MI}}{\partial H_{\hat{i},\hat{j}}} &= -\frac{1}{N^2}\sum_{i,j} H_{i,j}\log\left(\frac{N\cdot H_{i,j}}{H_i\cdot H_j}\right) \\ &\quad + \frac{1}{N}\left(\log\left(\frac{N\cdot H_{\hat{i},\hat{j}}}{H_{\hat{i}}\cdot H_{\hat{j}}}\right) + \sum_{i,j} H_{i,j}\frac{H_i\cdot H_j}{N\cdot H_{i,j}}\frac{\partial}{\partial H_{\hat{i},\hat{j}}}\left(\frac{N\cdot H_{i,j}}{H_i\cdot H_j}\right)\right) \\ &= \frac{1}{N}\left(\log\left(\frac{N\cdot H_{\hat{i},\hat{j}}}{H_{\hat{i}}\cdot H_{\hat{j}}}\right) - \mathrm{MI}\right) + \underbrace{\frac{1}{N}\sum_{i,j}\frac{H_i\cdot H_j}{N}\frac{\partial}{\partial H_{\hat{i},\hat{j}}}\left(\frac{N\cdot H_{i,j}}{H_i\cdot H_j}\right)}_{=\spadesuit}. \end{aligned}$$

We continue by transforming the term on the right hand side marked with a ♠ by exploiting Equations (5.19) and (5.20):

$$
\begin{aligned}
\spadesuit = & \frac{1}{N} \sum_{i,j} \frac{H_i \cdot H_j}{N} \frac{\partial}{\partial H_{\hat{i},\hat{j}}} \left(\frac{N \cdot H_{i,j}}{H_i \cdot H_j} \right) \\
= & \frac{1}{N^2} \frac{\left(H_{\hat{i},\hat{j}} + N\right)\left(H_{\hat{i}} \cdot H_{\hat{j}}\right) - \left(N \cdot H_{\hat{i},\hat{j}}\right)\left(H_{\hat{j}} + H_{\hat{i}}\right)}{H_{\hat{i}} \cdot H_{\hat{j}}} && \text{(here: } i = \hat{i},\ j = \hat{j}\text{)} \\
& + \frac{1}{N^2} \sum_{i=\hat{i}, j \neq \hat{j}} \frac{H_{\hat{i},j}\left(H_{\hat{i}} \cdot H_j\right) - \left(N \cdot H_{\hat{i},j}\right) H_j}{H_{\hat{i}} \cdot H_j} && \text{(here: } i = \hat{i},\ j \neq \hat{j}\text{)} \\
& + \frac{1}{N^2} \sum_{i \neq \hat{i}, j = \hat{j}} \frac{H_{i,\hat{j}}\left(H_i \cdot H_{\hat{j}}\right) - \left(N \cdot H_{i,\hat{j}}\right) H_i}{H_i \cdot H_{\hat{j}}} && \text{(here: } i \neq \hat{i},\ j = \hat{j}\text{)} \\
& + \frac{1}{N^2} \sum_{i \neq \hat{i}, j \neq \hat{j}} \frac{H_{i,j}\left(H_i \cdot H_j\right)}{H_i \cdot H_j} && \text{(here: } i \neq \hat{i},\ j \neq \hat{j}\text{)} \\
= & \frac{1}{N^2} \left(\left(H_{\hat{i},\hat{j}} + N\right) - \frac{\left(N \cdot H_{\hat{i},j}\right)\left(H_{\hat{j}} + H_{\hat{i}}\right)}{H_{\hat{i}} \cdot H_{\hat{j}}} \right) && \text{(here: } i = \hat{i},\ j = \hat{j}\text{)} \\
& + \frac{1}{N^2} \sum_{i=\hat{i}, j \neq \hat{j}} H_{\hat{i},j} - \frac{N \cdot H_{\hat{i},j}}{H_{\hat{i}}} && \text{(here: } i = \hat{i},\ j \neq \hat{j}\text{)} \\
& + \frac{1}{N^2} \sum_{i \neq \hat{i}, j = \hat{j}} H_{i,\hat{j}} - \frac{\left(N \cdot H_{i,\hat{j}}\right)}{H_{\hat{j}}} && \text{(here: } i \neq \hat{i},\ j = \hat{j}\text{)} \\
& + \frac{1}{N^2} \sum_{i \neq \hat{i}, j \neq \hat{j}} H_{i,j} && \text{(here: } i \neq \hat{i},\ j \neq \hat{j}\text{)} \\
= & \frac{H_{\hat{i},\hat{j}} + N}{N^2} - \frac{1}{N} \frac{H_{\hat{i},\hat{j}}}{H_{\hat{j}}} - \frac{1}{N} \frac{H_{\hat{i},\hat{j}}}{H_{\hat{i}}} && \text{(previously: } i = \hat{i},\ j = \hat{j}\text{)} \\
& + \frac{1}{N^2} \left(H_{\hat{i}} - H_{\hat{i},\hat{j}}\right) - \frac{1}{N} \frac{H_{\hat{i}} - H_{\hat{i},\hat{j}}}{H_{\hat{i}}} && \text{(previously: } i = \hat{i},\ j \neq \hat{j}\text{)} \\
& + \frac{1}{N^2} \left(H_{\hat{j}} - H_{\hat{i},\hat{j}}\right) - \frac{1}{N} \frac{H_{\hat{j}} - H_{\hat{i},\hat{j}}}{H_{\hat{j}}} && \text{(previously: } i \neq \hat{i},\ j = \hat{j}\text{)} \\
& + \frac{1}{N^2} \left(N - H_{\hat{j}} - H_{\hat{i}} + H_{\hat{i},\hat{j}}\right) && \text{(previously: } i \neq \hat{i},\ j \neq \hat{j}\text{)} \\
= & \frac{1}{N^2} \left(H_{\hat{i},\hat{j}} + H_{\hat{i}} - H_{\hat{i},\hat{j}} + H_{\hat{j}} - H_{\hat{i},\hat{j}} - H_{\hat{j}} - H_{\hat{i}} + H_{\hat{i},\hat{j}}\right) \\
& + \frac{1}{N} \left(1 - \frac{H_{\hat{i},\hat{j}}}{H_{\hat{j}}} - \frac{H_{\hat{i},\hat{j}}}{H_{\hat{i}}} - 1 + \frac{H_{\hat{i},\hat{j}}}{H_{\hat{i}}} - 1 + \frac{H_{\hat{i},\hat{j}}}{H_{\hat{j}}} + 1\right) \\
= & \ 0
\end{aligned}
$$

When we plug this in for ♠, we receive

$$\frac{\partial \mathrm{MI}}{\partial H_{\hat{i},\hat{j}}} = \frac{1}{N}\left(\log\left(\frac{N \cdot H_{\hat{i},\hat{j}}}{H_{\hat{i}} \cdot H_{\hat{j}}}\right) - \mathrm{MI}\right)$$

and this proves the statement. □

After having derived the gradient terms for the Mutual Information that are to be used in numerical optimization later on, we need to deal with various discretization aspects before we focus on the actual numerical solution. When applying those discretizations, we need to essentially ensure that the minimizers of the discretized optimization problem coincide with or more precisely converge to minimizers of the true, continuous optimization problem. Therefore, the next two sections are dedicated to the various discretization steps and the detailed convergence analysis.

5.3 Discretizations for numerical approach

In the previous section, we defined the optimization problem to determine an optimal parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ in a continuous setting. For various reasons, we cannot solve the optimization problem in a continuous setting. In this section, we focus on the different discretization aspects and directly elaborate on convergence results connecting the discretized setting to a continuous one. To be more precise, we delve into the discretization of the image data based on discrete pixel grids and focus on histogram discretized by binning strategies. In this context, we show the convergence of the discretized histogram measure towards the pushforward measure or, more precisely, focus on the L^1-convergence of the associated histogram density functions. To complete this section on intermediate convergence results that we use later on in the convergence analysis of the MI optimization problem, we deal in the last subsection with the L^1–convergence of a sequence of probability density functions related to the various discretization steps as well as their pointwise convergence for a subsequence.

5.3.1 Feature images on a discrete pixel grid

In this first subsection, we inspect the feature images in a discretized setting. We introduced the multi-channel feature image I_1 in Definition 3.2. In this context, we interpret the feature image to be living on a semi-discrete spatio-temporal domain $\Omega_T = \Omega \times \{t_1, \dots, t_{n_T}\}$. However, we cannot expect to have a continuous spatial domain Ω. For once, the feature extraction is based on the given microscopy data which in turn is given as discrete pixel images for each time point $t \in \{t_1, \dots, t_{n_T}\}$. Additionally, we aim for a numerical solution of the optimization problem. So it is essential to discretize the spatial domain Ω to deal with it in numerical implementations. Due to limited computer memory and computation power, we cannot apply arbitrarily fine discretizations and, consequently, can approximate a problem which is originally given in a continuous setting only up to a certain discretization error.
We start with a recapitulation of our semi-discrete spatio-temporal domain from Definition 3.1. This

definition takes into consideration the setting we are facing on the application side where we have microscopy images for pre-selected time points at hand.

Definition 5.54 (Semi-discrete spatio-temporal domain)
For a spatio-temporal cylinder $\Omega \times [0, T]$, we define that a semi-discrete spatio-temporal domain to be given by

$$\Omega_T = \Omega \times \{t_1, \dots, t_{n_T}\}$$

where we consider concrete time frames for pre-selected discrete time points $\{t_1, \dots, t_{n_T}\} \subset [0, T]$, i.e., we already assume a certain discretization of the temporal axis. The corresponding time steps are denoted by $\Delta t_i = t_{i+1} - t_i$ for $i = 1, \dots, n_T - 1$ and the total number of time frames is denoted with n_T.

We supplement this with a norm for a function f living on $\Omega_T = \Omega \times \{t_1, \dots, t_{n_T}\}$ to be given by the sum of norms for each time point, e.g., for the L^1-norm we consider

$$\|f\|_{L^1(\Omega_T)} = \sum_{i=1}^{n_T} \|f(\cdot, t_i)\|_{L^1(\Omega)}. \tag{5.21}$$

In the same sense, we consider the integration of a function $f : \Omega_T \to \mathbb{R}$ living on our semi-discrete spatio-temporal domain to be given by

$$\int_{\Omega_T} f(\boldsymbol{x}, t) \; \mathrm{d}(\boldsymbol{x}, t) = \sum_{i=1}^{n_T} \int_{\Omega} f(\boldsymbol{x}, t_i) \; \mathrm{d}\boldsymbol{x}. \tag{5.22}$$

Based on this semi-discrete spatio-temporal domain, we elaborate on a spatio-temporal domain discretized spatially as well in the following definition.

Definition 5.55 (Discretized spatio-temporal domain)
For a semi-discrete spatio-temporal domain $\Omega_T = \Omega \times \{t_1, \dots, t_{n_T}\}$, we define the discretized domain based on the discrete pixel grid Ω^{h} with pixel width h and the given discrete time points $\{t_1, \dots, t_{n_T}\}$. This leads to an approximation of the spatio-temporal domain with

$$\Omega_T \approx \Omega^{\mathrm{h}} \times \{t_1, \dots, t_{n_T}\}.$$

We define

$$\Omega^{\mathrm{h}} = \bigcup_{i=1}^{n_{\tilde{p}}} \Omega_{\tilde{p}_i}$$

with $\Omega_{\tilde{p}_i}$ as the domain for pixel $\tilde{p}_i$. Without loss of generality, we define the pixel domains $\Omega_{\tilde{p}}$ to be open. The total number of pixels for our rectangular domain $\Omega = [0, L] \times [0, W]$ is given as $n_{\tilde{p}} = \frac{L}{h}\frac{W}{h}$. Without loss of generality, we assume h to be such that $n_{\tilde{p}}$ is integer-valued.
For notational simplicity, we write $\tilde{p} \in \Omega^{\mathrm{h}}$ to refer to the domain $\Omega_{\tilde{p}} \subset \Omega^{\mathrm{h}}$.

We use here the notation $\tilde{p}$ for a pixel to avoid confusion with probability density functions denoted with p. We now focus on the convergence of a discretized image I_1^{h} to the original feature image I_1

for $h \to 0$, i.e., for increasing resolution of the discretized image. The statements and convergence results work for multi-channel images analogously, but for notational reasons we deal here only with a one-channel feature image.

We start with the following definition for the discretized feature image.

Definition 5.56 (Discretized feature image I_1^{h})
In the discretized setting, the "feature image" $\hat{I}_1^{\mathrm{h}}$ for a time point $t \in \{t_1, \dots, t_{n_T}\}$ is given as an array consisting of the feature values for each pixel. For a pixel $\tilde{p} \in \Omega^{\mathrm{h}}$, we define the values per pixel to be given by the mean value of I_1 in this pixel at this time point t. We define the discretized image I_1^{h} as the piecewise constant continuation of $\hat{I}_1^{\mathrm{h}}$ living on Ω_T again. For any $t \in \{t_1, \dots, t_{n_T}\}$ and a pixel $\tilde{p} \in \Omega^{\mathrm{h}}$, it holds that

$$\hat{I}_1^{\mathrm{h}}(\tilde{p}, t) = \fint_{\Omega_{\tilde{p}}} I_1(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x} = I_1^{\mathrm{h}}(\boldsymbol{x}, t) \quad \forall \boldsymbol{x} \in \Omega_{\tilde{p}}. \tag{5.23}$$

We remark, that we use for notational simplicity a hat symbol ($\hat{\cdot}$) for discrete values accumulated in a list or array structure. Before we delve into the convergence statement, we cite the definition of the total variation from Definition 4.81 in [12] and also refer to the Defintion 3.4 in [3].

Definition 5.57 (Total variation)
Let $\Omega \subset \mathbb{R}^N$ be a Lipschitz domain and let $u \in L^1(\Omega)$ be a locally integrable function on Ω. Then we define the total variation of u in Ω by

$$\mathrm{TV}_\Omega(u) := \sup\left\{ \int_\Omega u(\boldsymbol{x}) \ \mathrm{div}(\varphi(\boldsymbol{x})) \ \mathrm{d}\boldsymbol{x} : \varphi \in C_c^\infty\left(\Omega, \mathbb{R}^N\right), \|\varphi\|_\infty \leq 1 \right\},$$

with $C_c^\infty\left(\Omega, \mathbb{R}^N\right)$ the space of test functions on Ω.

Additionally, we introduce the space of functions of bounded variation based on the same Definition 4.81 in [12]. For an equivalent formulation for this space of functions of bounded variation, we refer to Proposition 3.6 in [3].

Definition 5.58 (Space of bounded variation)
The space of functions of bounded variation in Ω is defined by

$$\mathrm{BV}(\Omega) := \left\{ u \in L^1(\Omega) : \mathrm{TV}_\Omega(u) < \infty \right\}.$$

For an interpretation of the space of bounded variation, we state exemplarily that functions in this space do not allow infinitely many oscillations. In this case, the total variation would not be finite. In contrast, functions with finite many discontinuous jumps are included in this space as their total variation is indeed finite. In our setting, we are dealing with microscopy images which we consider to be of bounded variation. In this sense, the microscopy images may contain sharp edges, i.e.,

discontinuous jumps. In line with this, we consider the feature image I_1 to be of bounded variation. It follows directly that for any $t \in \{t_1, \dots, t_{n_T}\}$

$$I_1(\cdot, t) \in \mathrm{BV}(\Omega)$$

and therewith

$$I_1(\cdot, t) \in \mathrm{BV}\left(\Omega_{\tilde{p}}\right) \quad \forall \tilde{p} \in \Omega^{\mathrm{h}}.$$

By definition this means that for arbitrary $t \in \{t_1, \dots, t_{n_T}\}$ there exists a constant $C_{\mathrm{BV},t} > 0$ such that

$$\mathrm{TV}_\Omega(I_1(\cdot, t)) < C_{\mathrm{BV},t}. \tag{5.24}$$

It follows that $C_{\mathrm{BV}} := \max\{C_{\mathrm{BV},t} \mid t \in \{t_1, \dots, t_{n_T}\}\} > 0$ exists such that for any $t \in \{t_1, \dots, t_{n_T}\}$ holds

$$\mathrm{TV}_\Omega(I_1(\cdot, t)) < C_{\mathrm{BV}}. \tag{5.25}$$

Focusing on a pixel domain $\tilde{p}$, there exists constants $C_{\mathrm{BV},\tilde{p}} > 0$ as well such that

$$\mathrm{TV}_{\Omega_{\tilde{p}}}(I_1(\cdot, t)) < C_{\mathrm{BV},\tilde{p}}$$

holds. It follows directly that $C_{\mathrm{BV},\tilde{p}} \leq C_{\mathrm{BV}}$ for all $\tilde{p} \in \Omega^{\mathrm{h}}$ holds.
To prepare the convergence proof, we cite the Poincaré-Wirtinger inequality in spaces of bounded variation (cf. Theorem 3.2 in [7]).

Theorem 5.59 (Poincaré-Wirtinger in equality in $\mathrm{BV}(\Omega)$)
Let $\Omega \subset \mathbb{R}^N$ be a Lipschitz open bounded set. Then there exists a constant C_Ω such that

$$\|u - \fint_\Omega u\|_{L^1(\Omega)} \leq C_\Omega \mathrm{TV}_\Omega(u) \quad \forall u \in \mathrm{BV}(\Omega).$$

Proof. We refer the interested reader to the proof of Theorem 3.2 in the article on "Poincaré-Wirtinger inequalities in bounded variation function spaces" by Bergounioux [7]. □

For the later proof we need a scaled version of the Poincaré-Wirtinger inequality. For that we use the following proposition.

Proposition 5.60 (Poincaré-Wirtinger inequality for scaled domains)
We consider the domain $\Omega_1 = (0,1)^2$ and an arbitrary function $u \in \mathrm{BV}(\Omega_1)$. Let C_{Ω_1} be the related constant for the Poincaré-Wirtinger inequality. For a scaled function $u_{\mathrm{h}} \in \mathrm{BV}(\Omega_{\mathrm{h}})$ with $u_{\mathrm{h}} = u\left(\frac{x}{\mathrm{h}}\right)$ living on the scaled domain $\Omega_{\mathrm{h}} := (0, \mathrm{h})^2$, it holds

$$\|u_{\mathrm{h}} - \fint_{\Omega_{\mathrm{h}}} u_{\mathrm{h}}\|_{L^1(\Omega_{\mathrm{h}})} \leq \mathrm{h} C_{\Omega_1} \mathrm{TV}_{\Omega_{\mathrm{h}}}(u_{\mathrm{h}}).$$

Before we delve into the proof of this statement, we stress that we use the notation Ω_{h} with the subscript h for the *scaled* domain whereas Ω^{h} with h in the superscript denotes the discretized pixel grid domain.

Proof. We start the proof with the transition from Ω_h to Ω_1 or vice versa which can be achieved via

$$\begin{aligned} f : \Omega_h \to \Omega_1, \quad f(\boldsymbol{x}) &= \frac{\boldsymbol{x}}{h} \quad \text{with} \quad |\det Df| = \frac{1}{h^2} \\ g : \Omega_1 \to \Omega_h, \quad g(\boldsymbol{x}) &= h\boldsymbol{x} \quad \text{with} \quad |\det Dg| = h^2. \end{aligned}$$

With Theorem 5.59 it follows that a constant $C_{\Omega_1} > 0$ exists such that for all $u \in \mathrm{BV}(\Omega_1)$ it holds

$$\| u - \fint_{\Omega_1} u \|_{L^1(\Omega_1)} \leq C_{\Omega_1} \mathrm{TV}_{\Omega_1}(u).$$

Considering now the scaled space Ω_h, the scaled function $u_h = u \circ f$ and similarly $u = u_h \circ g$, we can show that

$$\begin{aligned} \| u_h - \fint_{\Omega_h} u_h \|_{L^1(\Omega_h)} &= \int_{\Omega_h} | u_h(\boldsymbol{x}) - \fint_{\Omega_h} u_h(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} | \, \mathrm{d}\boldsymbol{x} \\ &\overset{\substack{\text{integration} \\ \text{by subst.}}}{=} h^2 \int_{\Omega_1} | \underbrace{u_h(h\boldsymbol{x})}_{=u(\boldsymbol{x})} - \underbrace{\fint_{\Omega_h} u_h(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}}_{=\fint_{\Omega_1} u(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}} | \, \mathrm{d}\boldsymbol{x} \\ &= h^2 \int_{\Omega_1} | u(\boldsymbol{x}) - \fint_{\Omega_1} u(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} | \, \mathrm{d}\boldsymbol{x} \\ &\overset{\substack{\text{Poincaré-Wirtinger} \\ \text{inequality,} \\ \text{cf. Theorem 5.59}}}{\leq} h^2 C_{\Omega_1} \mathrm{TV}_{\Omega_1}(u) \\ &\overset{(*)}{=} h^2 C_{\Omega_1} \frac{1}{h} \mathrm{TV}_{\Omega_h}(u_h) \\ &= h C_{\Omega_1} \mathrm{TV}_{\Omega_h}(u_h) \end{aligned}$$

holds. We apply in $(*)$ the following transformation

$$\begin{aligned} \mathrm{TV}_{\Omega_1}(u) &\overset{\text{Definition 5.57}}{=} \sup \left\{ \int_{\Omega_1} u(\boldsymbol{x}) \left(\varphi_{x_1}(\boldsymbol{x}) + \varphi_{x_2}(\boldsymbol{x}) \right) \, \mathrm{d}\boldsymbol{x} : \varphi \in C_c^\infty\left(\Omega_1, \mathbb{R}^N\right), \, \|\varphi\|_\infty \leq 1 \right\} \\ &\overset{\substack{\text{integration} \\ \text{by subst.}}}{=} \sup \left\{ \frac{1}{h^2} \int_{\Omega_h} u_h(\boldsymbol{x}) \, h \cdot \left(\varphi_{h,x_1}(\boldsymbol{x}) + \varphi_{h,x_2}(\boldsymbol{x}) \right) \, \mathrm{d}\boldsymbol{x} : \varphi_h \in C_c^\infty\left(\Omega_h, \mathbb{R}^N\right), \, \|\varphi_h\|_\infty \leq 1 \right\} \\ &= \frac{1}{h} \sup \left\{ \int_{\Omega_h} u_h(\boldsymbol{x}) \left(\varphi_{h,x_1}(\boldsymbol{x}) + \varphi_{h,x_2}(\boldsymbol{x}) \right) \, \mathrm{d}\boldsymbol{x} : \varphi_h \in C_c^\infty\left(\Omega_h, \mathbb{R}^N\right), \, \|\varphi_h\|_\infty \leq 1 \right\} \\ &= \frac{1}{h} \mathrm{TV}_{\Omega_h}(u_h). \end{aligned}$$

Here, it is important to consider not only the determinant when applying the substitution. It is also necessary to account for the transformation within the derivative terms. When $\varphi_h(\boldsymbol{x}) = \varphi\left(\frac{\boldsymbol{x}}{h}\right)$ holds, then for a partial derivative we get

$$\frac{\partial}{\partial x_i} \varphi_h(\boldsymbol{x}) = \varphi_{h,x_i}(\boldsymbol{x}) = \frac{1}{h} \varphi_{x_i}\left(\frac{\boldsymbol{x}}{h}\right).$$

□

We now prove our main result for the discretization in the spatial domain for the feature image.

Theorem 5.61
Let $I_1(\cdot, t) \in \mathrm{BV}(\Omega)$ for every $t \in \{t_1, \dots, t_{n_T}\}$. Then it holds that the piecewise constant feature image I_1^{h} derived from the discretized features living on a pixel grid converges in L^1 to the feature image I_1 if the pixel width h converges to 0. The following estimate holds

$$\|I_1 - I_1^{\mathrm{h}}\|_{L^1(\Omega_T)} \leq n_T \, C_{\Omega_1} \, \mathrm{h} \, C_{\mathrm{BV}}$$

with n_T the number of discrete time point sin our semi-discrete spatio-temporal domain Ω_T and the constants C_{Ω_1}, C_{BV} from Proposition 5.60 and Equation (5.25).

Proof. We start with the convergence in $L^1(\Omega)$ for any $t \in \{t_1, \dots, t_{n_T}\}$.

$$\begin{aligned}
\|I_1(\cdot, t) - I_1^{\mathrm{h}}(\cdot, t)\|_{L^1(\Omega)} &= \int_\Omega |I_1(\boldsymbol{x}, t) - I_1^{\mathrm{h}}(\boldsymbol{x}, t)| \, \mathrm{d}\boldsymbol{x} \\
&= \sum_{\tilde{p} \in \Omega^{\mathrm{h}}} \int_{\Omega_{\tilde{p}}} |I_1(\boldsymbol{x}, t) - I_1^{\mathrm{h}}(\boldsymbol{x}, t)| \, \mathrm{d}\boldsymbol{x} \\
&\overset{\text{Equation (5.23)}}{=} \sum_{\tilde{p} \in \Omega^{\mathrm{h}}} \int_{\Omega_{\tilde{p}}} |I_1(\boldsymbol{x}, t) - \fint_{\Omega_{\tilde{p}}} I_1(\boldsymbol{y}, t) \, \mathrm{d}\boldsymbol{y}| \, \mathrm{d}\boldsymbol{x} \\
&\overset{\text{Proposition 5.60}}{\leq} C_{\Omega_1} \mathrm{h} \sum_{\tilde{p} \in \Omega^{\mathrm{h}}} \mathrm{TV}_{\Omega_{\tilde{p}}}(I_1) \\
&\overset{(*)}{\leq} C_{\Omega_1} \, \mathrm{h} \, \mathrm{TV}_\Omega(I_1) \\
&\overset{\text{Equation (5.25)}}{\leq} C_{\Omega_1} \, \mathrm{h} \, C_{\mathrm{BV}}
\end{aligned}$$

with constants C_1, $C_{\mathrm{BV}} > 0$ the $L^1\left(\Omega^{\mathrm{h}}\right)$-convergence to 0 follows directly with $\mathrm{h} \to 0$. We note down that in $(*)$ we apply the following approximation where we use that the pixel domains are defined to be open (cf. Definition 5.55). This is important for the application of the Poincaré-Wirtinger inequality. Moreover, they are disjoint

$$\Omega_{\tilde{p}} \cap \Omega_{\tilde{q}} = \varnothing$$

and with $\varphi \in C_c^\infty\left(\Omega, \mathbb{R}^N\right)$ defined as

$$\varphi := \sum_{\tilde{p} \in \Omega^{\mathrm{h}}} \mathbf{1}_{\Omega_{\tilde{p}}} \varphi_{\Omega_{\tilde{p}}}$$

with the characteristic functions $\mathbf{1}_{\Omega_{\tilde{p}}}$ living on Ω^{h} combined with the test functions per pixel $\varphi_{\Omega_{\tilde{p}}} \in C_c^\infty\left(\Omega_{\tilde{p}}, \mathbb{R}^N\right)$ we have a continuous continuation on Ω and also a valid test function on Ω. With this construction in mind, it is possible to move the summation into the supremum for the total variation in $(*)$ and, consequently, get the TV-estimation on the whole domain Ω. By considering Equation (5.21), we conclude that

$$\|I_1 - I_1^{\mathrm{h}}\|_{L^1(\Omega_T)} = \sum_{i=1}^{n_T} \|I_1(\cdot, t_i) - I_1^{\mathrm{h}}(\cdot, t_i)\|_{L^1(\Omega)} \leq n_T \, C_{\Omega_1} \, \mathrm{h} \, C_{\mathrm{BV}}$$

holds which proves the L^1-convergence for our discretized spatio-temporal domain as $\mathrm{h} \to 0$. □

In the following lemma, we focus on the possibility of changing integration and limit as a direct consequence from the L^1-convergence.

Lemma 5.62
We consider a sequence of functions $(f_n)_{n\in\mathbb{N}}$ that converges in L^1 to f. Then the order of taking the limit of the sequence and integrating it can be changed such that

$$\lim_{n\to\infty} \int f_n = \int f$$

holds.

Proof. We divide the proof in two steps showing the lower inequality and greater inequality separately.

"$\leq$"

$$\begin{aligned}
\lim_{n\to\infty} \int f_n - \int f &= \lim_{n\to\infty} \int f_n - f \\
&\leq \lim_{n\to\infty} \left| \int f_n - f \right| \\
&\leq \lim_{n\to\infty} \int |f_n - f| \overset{L^1\text{conv.}}{=} 0 \\
\Rightarrow \lim_{n\to\infty} \int f_n &\leq \int f.
\end{aligned}$$

"$\geq$"

$$\begin{aligned}
\int f - \lim_{n\to\infty} \int f_n &= \lim_{n\to\infty} \int f - f_n \\
&\leq \lim_{n\to\infty} \left| \int f - f_n \right| \\
&\leq \lim_{n\to\infty} \int |f - f_n| \overset{L^1\text{conv.}}{=} 0 \\
\Rightarrow \lim_{n\to\infty} \int f_n &\geq \int f.
\end{aligned}$$

□

In later proofs, we exploit this property to change the order of integration and taking the limit when dealing with the discrete feature images converging in L^1.
Similar to the discretization of the feature images, we consider the classification image on a discretized pixel grid. Next, we show that this discretized image converges uniformly with respect to the spreading parameters $\boldsymbol{p}$ to the continuous classification image based on the concentric spreading model.

5.3.2 Classification images on a discrete pixel grid

We introduced the classification image I_2 in Definition 4.4 on a spatio-temporal domain Ω_T combined with the parameter space $\boldsymbol{P}$. As we aim for a numerical solution of the optimization problem, it is

essential to discretize the spatio-temporal domain Ω_T as shown in Definition 5.55. Similar to the previous section dealing with the L^1-convergence of the feature image, we show in this section the convergence of I_2^h against I_2 in L^1 for a pixel width $\mathrm{h} \to 0$.
Again, we use the discretized domain defined in Definition 5.55. For define the discretized classification image I_2^{h} as follows:

Definition 5.63 (Discretized classification image I_2^{h})
In the discretized setting, the "classification image" $\hat{I}_2^{\mathrm{h}}$ for a time point $t \in \{t_1, \dots, t_{n_T}\}$ and a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ is given as an array consisting of the classification values for each pixel $\tilde{p} \in \Omega^{\mathrm{h}}$. For each pixel, we define the values per pixel to be given by the value of I_2 in the center point of this pixel. We define the discretized image I_2^{h} as the piecewise constant continuation of $\hat{I}_2^{\mathrm{h}}$ living on Ω_T again. For any $t \in \{t_1, \dots, t_{n_T}\}$, a fixed parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ and a pixel $\tilde{p} \in \Omega^{\mathrm{h}}$, it holds that

$$\hat{I}_2^{\mathrm{h}}\left(\boldsymbol{p}, \tilde{p}, t\right) = I_2\left(\boldsymbol{p}, z_{\tilde{p}}, t\right) = I_2^{\mathrm{h}}\left(\boldsymbol{p}, \boldsymbol{x}, t\right) \quad \forall \boldsymbol{x} \in \Omega_{\tilde{p}}. \tag{5.26}$$

with $z_{\tilde{p}}$ denoting the center point of pixel $\Omega_{\tilde{p}}$.

As we are considering square pixel regions, we can substantiate for a pixel domain $\Omega_{\tilde{p}}$ and its center point $z_{\tilde{p}} \in \Omega_{\tilde{p}}$ that

$$\left\|\boldsymbol{x} - z_{\tilde{p}}\right\|_\infty < \mathrm{h} \qquad \forall \boldsymbol{x} \in \Omega_{\tilde{p}} \tag{5.27}$$

and the norm is even lower equal than $\frac{\mathrm{h}}{2}$. Now, we prove the convergence of this discrete classification image living on a pixel grid to the continuous classification image defined in Definition 4.4.

Theorem 5.64 (L^1-convergence of classification images)
The piecewise constant classification image I_2^{h} derived from the discretized classification values living on a pixel grid converges in L^1 to the smooth classification image I_2 if the pixel width h converges to 0.

Proof. By definition of the classification image I_2 in Definition 4.4, it is continuous on Ω_T. Let $t \in \{t_1, \dots, t_{n_T}\}$ and $\boldsymbol{p} \in \boldsymbol{P}$ be arbitrary but fixed. Then we can define a constant $C = C\left(\boldsymbol{p}, t\right)$, i.e., depending on the considered time point t and the spreading parameter setting $\boldsymbol{p}$, by setting

$$C := \max_{\hat{\boldsymbol{x}} \in \Omega} \left\|\nabla_{\boldsymbol{x}} I_2\left(\boldsymbol{p}, \hat{\boldsymbol{x}}, t\right)\right\|_1. \tag{5.28}$$

The gradient with respect to the spatial variable is denoted with $\nabla_{\boldsymbol{x}}$. This definition is valid since I_2 is continuously differentiable with respect to all variables (cf. Definition 4.4). Since a digitized classification image, i.e., an image only consisting of integer values representing certain classes, is here approximated smoothly with a Heaviside function to avoid jumps and create continuously differentiable transition regions, it is a valid conclusion that $C < \infty$ holds.
We remark that C depends on the scaling parameter ε_0 which determines the steepness of the Heaviside approximation, cf. Definition 4.4. However, this parameter is considered as a fixed model

parameter so that C can indeed be used as a finite constant for a fixed ε_0.
Let $\delta > 0$ be arbitrary. Then there exists an $\mathrm{h}' > 0$ with $\mathrm{h}' < \frac{\delta}{C \cdot L \cdot W}$, such that for all $\mathrm{h} \leq \mathrm{h}'$ holds

$$\begin{aligned}
\left\| I_2^{\mathrm{h}}(\boldsymbol{p}, \cdot, t) - I_2(\boldsymbol{p}, \cdot, t) \right\|_{L^1(\Omega)} &= \int_\Omega \left| I_2^{\mathrm{h}}(\boldsymbol{p}, \boldsymbol{x}, t) - I_2(\boldsymbol{p}, \boldsymbol{x}, t) \right| \mathrm{d}\boldsymbol{x} \\
&= \sum_{\tilde{p} \in \Omega^{\mathrm{h}}} \int_{\Omega_{\tilde{p}}} \left| I_2^{\mathrm{h}}(\boldsymbol{p}, \boldsymbol{x}, t) - I_2(\boldsymbol{p}, \boldsymbol{x}, t) \right| \mathrm{d}\boldsymbol{x} \\
&\overset{\text{Equation (5.26)}}{=} \sum_{\tilde{p} \in \Omega^{\mathrm{h}}} \int_{\Omega_{\tilde{p}}} \left| I_2\left(\boldsymbol{p}, \boldsymbol{z}_{\tilde{p}}, t\right) - I_2(\boldsymbol{p}, \boldsymbol{x}, t) \right| \mathrm{d}\boldsymbol{x} \\
&\overset{\substack{\text{mean value} \\ \text{theorem} \\ \boldsymbol{y} \in \overline{\boldsymbol{z}_{\tilde{p}} \boldsymbol{x}}}}{=} \sum_{\tilde{p} \in \Omega^{\mathrm{h}}} \int_{\Omega_{\tilde{p}}} \left| \nabla_{\boldsymbol{x}} I_2(\boldsymbol{p}, \boldsymbol{y}, t) \cdot \left(\boldsymbol{z}_{\tilde{p}} - \boldsymbol{x}\right) \right| \mathrm{d}\boldsymbol{x} \\
&\overset{\substack{\text{Cauchy} \\ \text{inequality}}}{\leq} \sum_{\tilde{p} \in \Omega^{\mathrm{h}}} \int_{\Omega_{\tilde{p}}} \underbrace{\left\| \nabla_{\boldsymbol{x}} I_2(\boldsymbol{p}, \boldsymbol{y}, t) \right\|_1}_{\substack{\leq C, \\ \text{cf. Equation (5.28)}}} \underbrace{\left\| \boldsymbol{z}_{\tilde{p}} - \boldsymbol{x} \right\|_\infty}_{\substack{< \mathrm{h}, \\ \text{cf. Equation (5.27)}}} \mathrm{d}\boldsymbol{x} \\
&< C \cdot \mathrm{h} \underbrace{\sum_{\tilde{p} \in \Omega^{\mathrm{h}}} \int_{\Omega_{\tilde{p}}} 1 \, \mathrm{d}\boldsymbol{x}}_{= L \cdot W} \\
&= C \cdot L \cdot W \cdot \mathrm{h} \leq C \cdot L \cdot W \cdot \mathrm{h}' < \delta.
\end{aligned} \tag{5.29}$$

For the sake of completeness, we state that we use $\overline{\boldsymbol{z}_{\tilde{p}} \boldsymbol{x}}$ as the notation for the line segment between $\boldsymbol{z}_{\tilde{p}}$ and $\boldsymbol{x}$. This shows the convergence of the discretized image I_2^{h} to the continuous image I_2 in $L^1(\Omega)$ for a fixed time point $t \in \{t_1, \ldots, t_{n_T}\}$.
To show the convergence of the classification images considering the whole semi-discrete spatio-temporal domain Ω_T, i.e., not limiting the analysis to one fixed time point $t \in \{t_1, \ldots, t_{n_T}\}$, we start with an estimate for the norm of the spatial gradient by defining the following constant $C^* = C^*(\boldsymbol{p}) > 0$ by

$$C^* = \max_{t \in \{t_1, \ldots, t_{n_T}\}} C(\boldsymbol{p}, t) = \max_{\substack{t \in \{t_1, \ldots, t_{n_T}\}, \\ \hat{\boldsymbol{x}} \in \Omega}} \left\| \nabla_{\boldsymbol{x}} I_2(\boldsymbol{p}, \hat{\boldsymbol{x}}, t) \right\|_1 \tag{5.30}$$

based on Equation (5.28). This constant C^* only depends on the parameter set. By considering Equation (5.21) and Equation (5.29) with $C = C(\boldsymbol{p}, t)$, we conclude that for arbitrary $\delta > 0$ there exists an $\mathrm{h}' > 0$ with $\mathrm{h}' < \frac{\delta}{C^* \cdot L \cdot W}$, such that for all $\mathrm{h} \leq \mathrm{h}'$ holds

$$\begin{aligned}
\left\| I_2^{\mathrm{h}} - I_2 \right\|_{L^1(\Omega_T)} &= \sum_{i=1}^{n_T} \left\| I_2^{\mathrm{h}}(\cdot, t_i) - I_2(\cdot, t_i) \right\|_{L^1(\Omega)} \\
&\leq \sum_{i=1}^{n_T} C(\boldsymbol{p}, t_i) \cdot L \cdot W \cdot \mathrm{h} \\
&\leq n_T \cdot C^* \cdot L \cdot W \cdot \mathrm{h} \\
&\leq C^* \cdot L \cdot W \cdot \mathrm{h}' < \delta
\end{aligned}$$

holds which proves the L^1-convergence for the classification images living on our discretized spatio-temporal domain as $\mathrm{h} \to 0$. □

With the previous theorem, we show the convergence of the discretized classification images to the continuous one when considering a vanishing grid size h. Since we choose an arbitrary but fixed parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ in the beginning of the proof, the considered constants C and C^* depend on the parameter setting. In the next proposition, we go a step further and derive a constant that is independent of our parameter setting to show that the discretized classification image I_2^{h} converges uniformly for $(\boldsymbol{p}, (\boldsymbol{x}, t)) \in \boldsymbol{P} \times \Omega_T$.

Proposition 5.65
The piecewise constant classification image I_2^{h} derived from the discretized classification values living on a pixel grid converges uniformly for parameter settings $\boldsymbol{p} \in \boldsymbol{P}$ and $(\boldsymbol{x}, t) \in \Omega_T$ to the smooth classification image I_2 if the pixel width h converges to 0.

Proof. To show this statement, we consider a similar starting position as in the proof of Theorem 5.64 and also follow a similar line of arguments. Based on the definition of I_2 in Definition 4.4 and its gradient terms defined in Equation (4.8), we derive a constant C which is independent of $(\boldsymbol{p}, (\boldsymbol{x}, t)) \in \boldsymbol{P} \times \Omega_T$ and that bounds the norm of the gradient. We recapitulate the spatial gradient from Equation (4.8)

$$\nabla_{\boldsymbol{x}} I_2(\boldsymbol{p}, \boldsymbol{x}, t) = \frac{1}{\varepsilon_0}\left(\frac{\exp\left(-\frac{1}{\varepsilon_0} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)\left(\nabla_{\boldsymbol{x}} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)}{\left(1+\exp\left(-\frac{1}{\varepsilon_0} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right)\right)^2} + \frac{\exp\left(-\frac{1}{\varepsilon_0} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)\left(\nabla_{\boldsymbol{x}} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)}{\left(1+\exp\left(-\frac{1}{\varepsilon_0} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right)\right)^2}\right)$$

with Equations (4.9) and (4.10)

$$\nabla_{\boldsymbol{x}} k_n(\boldsymbol{p}, \boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} k_a(\boldsymbol{p}, \boldsymbol{x}, t) = \frac{1}{\|\boldsymbol{x} - \boldsymbol{x}_0\|_2}\begin{pmatrix}(x_1 - x_{0,1})\\(x_2 - x_{0,2})\end{pmatrix}.$$

We approximate the norm of this as follows

$$\left\|\nabla_{\boldsymbol{x}} k_n(\boldsymbol{p}, \boldsymbol{x}, t)\right\|_1 = \left\|\nabla_{\boldsymbol{x}} k_a(\boldsymbol{p}, \boldsymbol{x}, t)\right\|_1 = \frac{\sqrt{(x_1 - x_{0,1})^2} + \sqrt{(x_2 - x_{0,2})^2}}{\sqrt{(x_1 - x_{0,1})^2 + (x_2 - x_{0,2})^2}} \leq \sqrt{2}. \tag{5.31}$$

This is true as we briefly show by applying the substitution $\alpha := x_1 - x_{0,1}$ and $\beta := x_2 - x_{0,2}$ and derive

$$\frac{\sqrt{\alpha^2} + \sqrt{\beta^2}}{\sqrt{\alpha^2 + \beta^2}} \leq \sqrt{2} \quad \Leftrightarrow \quad \alpha^2 + \beta^2 + 2\sqrt{\alpha^2}\sqrt{\beta^2} \leq 2\left(\alpha^2 + \beta^2\right) \quad \Leftrightarrow \quad 0 \leq \left(\sqrt{\alpha^2} - \sqrt{\beta^2}\right)^2$$

which holds for arbitrary $\alpha, \beta \in \mathbb{R}$ and, consequently, also for our substituted coordinates.
For the next approximations, we exploit the structure of the parameter space $\boldsymbol{P}$ as given in Equation (4.3) and the definition of our spatio-temporal domain Ω_T (cf. Definition 3.1). Considering the circle equations for the normal and abnormal colony fronts defined in Equation (4.4), we derive

$$\begin{aligned}\exp\left(-\frac{1}{\varepsilon_0} k_j(\boldsymbol{p}, \boldsymbol{x}, t)\right) &= \exp\left(-\frac{1}{\varepsilon_0}\left(v\left(t - t_{0,j}\right) - \|\boldsymbol{x} - \boldsymbol{x}_0\|_2\right)\right)\\ &\leq \exp\left(-\frac{1}{\varepsilon_0}\left(-v_{\max} T - \sqrt{L^2 + W^2}\right)\right) := C^{\dagger}\end{aligned} \tag{5.32}$$

which holds for both fronts, i.e., for $j = n, a$, and the constant $C^\dagger$ is positive. We conclude with the following approximation of the norm for the spatial gradient of the classification image for $(\boldsymbol{p}, \boldsymbol{x}, t) \in \boldsymbol{P} \times \Omega_T$ arbitrary:

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{x}} I_2\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right\|_1 &= \left\| \frac{1}{\varepsilon_0} \left(\frac{\exp\left(-\frac{1}{\varepsilon_0} k_n\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right) \nabla_{\boldsymbol{x}} k_n\left(\boldsymbol{p}, \boldsymbol{x}, t\right)}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_n\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right)\right)^2} + \frac{\exp\left(-\frac{1}{\varepsilon_0} k_a\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right) \nabla_{\boldsymbol{x}} k_a\left(\boldsymbol{p}, \boldsymbol{x}, t\right)}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_a\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right)\right)^2} \right) \right\|_1 \\
&\overset{\substack{\text{triangle}\\ \text{inequality}}}{\leq} \frac{1}{\varepsilon_0} \Bigg(\underbrace{\exp\left(-\frac{1}{\varepsilon_0} k_n\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right)}_{\substack{\leq C^\dagger, \text{ cf.}\\ \text{Equation (5.32)}}} \underbrace{\left(\frac{1}{1 + \exp\left(-\frac{1}{\varepsilon_0} k_n\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right)} \right)^2}_{\leq 1 \text{ by def., cf. Equation (4.5)}} \underbrace{\left\|\nabla_{\boldsymbol{x}} k_n\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right\|_1}_{\substack{< \sqrt{2}, \text{ cf.}\\ \text{Equation (5.31)}}} \\
&\qquad + \underbrace{\exp\left(-\frac{1}{\varepsilon_0} k_a\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right)}_{\substack{\leq C^\dagger, \text{ cf.}\\ \text{Equation (5.32)}}} \underbrace{\left(\frac{1}{1 + \exp\left(-\frac{1}{\varepsilon_0} k_a\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right)} \right)^2}_{\leq 1 \text{ by def., cf. Equation (4.5)}} \underbrace{\left\|\nabla_{\boldsymbol{x}} k_a\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right\|_1}_{\substack{< \sqrt{2}, \text{ cf.}\\ \text{Equation (5.31)}}} \Bigg) \\
&\leq \frac{2\sqrt{2}}{\varepsilon_0} C^\dagger =: C^*
\end{aligned}$$

With the the newly defined constant $C^* > 0$ at hand, we show the uniform convergence for I_2^{h} to I_2 for vanishing h. Let $(\boldsymbol{p}, (\boldsymbol{x}, t)) \in \boldsymbol{P} \times \Omega_T$ be arbitrary and let $\delta > 0$ be arbitrary, too. Then there exists an $\mathrm{h}' > 0$ with $\mathrm{h}' < \frac{\delta}{C^*}$ such that for all $\mathrm{h} \leq \mathrm{h}'$ holds

$$\begin{aligned}
\left|I_2^{\mathrm{h}}\left(\boldsymbol{p}, (\boldsymbol{x}, t)\right) - I_2\left(\boldsymbol{p}, (\boldsymbol{x}, t)\right)\right| &\overset{\text{Equation (5.26)}}{=} \left|I_2\left(\boldsymbol{p}, \boldsymbol{z}_{\tilde{p}}, t\right) - I_2\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right| \\
&\overset{\substack{\text{mean value}\\ \text{theorem}\\ \boldsymbol{y} \in \overline{\boldsymbol{z}_{\tilde{p}} \boldsymbol{x}}}}{=} \left|\nabla_{\boldsymbol{x}} I_2\left(\boldsymbol{p}, \boldsymbol{y}, t\right) \cdot \left(\boldsymbol{z}_{\tilde{p}} - \boldsymbol{x}\right)\right| \\
&\overset{\substack{\text{Cauchy}\\ \text{inequality}}}{\leq} \underbrace{\left\|\nabla_{\boldsymbol{x}} I_2\left(\boldsymbol{p}, \boldsymbol{y}, t\right)\right\|_1}_{\leq C^*} \underbrace{\left\|\boldsymbol{z}_{\tilde{p}} - \boldsymbol{x}\right\|_\infty}_{\substack{< h,\\ \text{cf. Equation (5.27)}}} < C^* \cdot h \leq C^* \cdot h' < \delta.
\end{aligned}$$

Since our constant C^* does not depend on any of the parameter $(\boldsymbol{p}, (\boldsymbol{x}, t)) \in \boldsymbol{P} \times \Omega_T$, this proves the uniform convergence of the discretized classification image to the continuous one when considering vanishing grid sizes. For the sake of completeness, we point that the center pixel point $\boldsymbol{z}_{\tilde{p}}$ for the pixel containing $\boldsymbol{x}$ depends naturally on the current grid size h. However, this does not impede our statement of uniform convergence as we approximate the gradient term for the classification image in a more general case by using our independent constant C^*. □

While we did not derive precise approximations for the constants C and C^* in Equations (5.28) and (5.30) in the proof of Theorem 5.64 when considering $\boldsymbol{p} \in \boldsymbol{P}$ and $t \in \{t_1, \ldots, t_{n_T}\}$ fixed, we remark that we could also apply the global estimate

$$\left\|\nabla_{\boldsymbol{x}} I_2\left(\boldsymbol{p}, \boldsymbol{x}, t\right)\right\|_1 \leq \frac{2\sqrt{2}}{\varepsilon_0} C^\dagger$$

as derived in the last proof for arbitrary $(\boldsymbol{p}, (\boldsymbol{x}, t)) \in \boldsymbol{P} \times \Omega_T$ (cf. Equation (5.32)).
In the next section, we focus on another statement for uniform convergence of the classification images with respect to the spreading parameter $\boldsymbol{p} \in \boldsymbol{P}$.

5.3.3 Uniform convergence of classification images

After having shown the convergence of I_2^h to I_2 for a pixel width h converging to 0, we add another convergence result related to the classification image. For later references, we show the uniform convergence of the classification image depending on the parameter setting $\boldsymbol{p} \in \boldsymbol{P}$. We recapitulate two statements on uniform convergence and uniform continuity which we make use of in the later proof of the main statement of this subsection.

Proposition 5.66 (Uniform convergence in compositions)
Let $(f_n)_{n\in\mathbb{N}}$ be a uniformly convergent sequence of functions $f_n : X \to \mathbb{R}$ with $f : X \to \mathbb{R}$ the limiting function and $F : \mathbb{R} \to \mathbb{R}$ a uniformly continuous function. Then it holds that $(F \circ f_n)_{n\in\mathbb{N}}$ is uniformly convergent.

Proof. Let $\varepsilon > 0$ be arbitrary. Because of the uniform continuity of F there exists a $\delta > 0$ such that for all $(y_1, y_2) \in \mathbb{R}^2$ with $|y_1 - y_2| < \delta$ it holds

$$|F(y_1) - F(y_2)| < \varepsilon.$$

Moreover, with the uniform convergence of $(f_n)_{n\in\mathbb{N}}$ it follows that an $N_0 \in \mathbb{N}$ exists such that for all $x \in X$ and all $N \geq N_0$ it holds

$$|f_n(x) - f(x)| < \delta.$$

For $\varepsilon \to 0$, the statement follows directly via

$$|(F \circ f_n)(x) - (F \circ f)(x)| = |F(f_n(x)) - F(f(x))| < \varepsilon.$$

□

Proposition 5.67 (Uniform convergence for addition)
Let $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ be uniformly convergent sequences of functions $f_n, g_n : X \to \mathbb{R}$ with $f, g : X \to \mathbb{R}$ the limiting functions. Then it holds that $(f_n + g_n)_{n\in\mathbb{N}}$ is uniformly convergent with the limiting function $(f + g)$.

Proof. Let $\varepsilon > 0$ be arbitrary. Because of the uniform convergence of $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ there exist $N_1, N_2 \in \mathbb{N}$ such that

$$|f_n(\boldsymbol{x}) - f(\boldsymbol{x})| < \frac{\varepsilon}{2}, \quad \forall N \geq N_1 \qquad\qquad |g_n(\boldsymbol{x}) - g(\boldsymbol{x})| < \frac{\varepsilon}{2}, \quad \forall N \geq N_2$$

hold. If we set $N_0 = \max\{N_1, N_2\}$ then it follows directly with the triangle inequality that for all $N \geq N_0$ it holds

$$|(f_n + g_n)(\boldsymbol{x}) - (f + g)(\boldsymbol{x})| \leq |f_n(\boldsymbol{x}) - f(\boldsymbol{x})| + |g_n(\boldsymbol{x}) - g(\boldsymbol{x})| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since ε was chosen arbitrarily, the statement follows for $\varepsilon \to 0$. □

To prepare the proof of I_2 being uniformly convergent in $(\boldsymbol{x}, t) \in \Omega_T$ for a converging parameter sequence, e.g., $\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon \to 0} \boldsymbol{p}$, we elaborate on the following two additional prerequisites.

Proposition 5.68
The smoothed Heaviside function defined in Equation (4.5) is uniformly continuous on $\mathbb{R}$.

Proof. We have the smoothed Heaviside approximation from Equation (4.5)

$$f : \mathbb{R} \to [0,1], \qquad f(x) = \frac{1}{1 + \exp\left(-\frac{x}{\varepsilon_0}\right)}$$

and its derivative given as

$$f' : \mathbb{R} \to [0,1], \qquad f'(x) = \frac{\exp\left(-\frac{x}{\varepsilon_0}\right)}{\varepsilon_0 \left(1 + \exp\left(-\frac{x}{\varepsilon_0}\right)\right)^2}$$

which attains its maximum in $x = 0$ with $f'(0) = \frac{1}{4\varepsilon_0}$. Let $\varepsilon > 0$ be arbitrary and ε_0 be again the inherent and fixed model parameter controlling the transition width or, more precisely, the steepness of the smoothed Heaviside approximation. Then it exists a $\delta > 0$ with $\delta < 4\varepsilon\varepsilon_0$ such that for all $x_1, x_2 \in \mathbb{R}$ with $|x_1 - x_2| < \delta$ holds

$$|f(x_1) - f(x_2)| \overset{\substack{\text{mean value theorem} \\ x_0 \in \overline{x_1 x_2}}}{=} \left|f'(x_0)(x_1 - x_2)\right| = \underbrace{\frac{\exp\left(-\frac{x_0}{\varepsilon_0}\right)}{\varepsilon_0 \left(1 + \exp\left(-\frac{x_0}{\varepsilon_0}\right)^2\right)}}_{\leq f'(0)} \underbrace{|x_1 - x_2|}_{<\delta} < \frac{1}{4\varepsilon_0}\delta < \varepsilon.$$

Since ε was chosen arbitrarily, this shows the uniform continuity of the smoothed Heaviside function for $\varepsilon \to 0$. □

Proposition 5.69
Let k_j denote again the circle equation describing the normal colony front ($j = n$) and the abnormal front ($j = a$) as defined in Equation (4.4). For $j = n, a$, we define a sequence of functions $\left((k_j)_\varepsilon\right)_{\varepsilon>0}$ by

$$(k_j)_\varepsilon := k_j(\boldsymbol{p}_\varepsilon, \cdot) : \Omega_T \to [0,1].$$

for a sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ in the parameter space $\boldsymbol{P}$. Let the sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ be converging to $\boldsymbol{p} \in \boldsymbol{P}$ for $\varepsilon \to 0$ in the sense that $\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2 < \varepsilon$ for all $\varepsilon > 0$ holds. Then it holds that the sequence of functions $\left((k_j)_\varepsilon\right)_{\varepsilon>0}$ is uniformly convergent with the limiting function

$$k_j := k_j(\boldsymbol{p}, \cdot) : \Omega_T \to [0,1].$$

Proof. Let $\boldsymbol{p}_\varepsilon = (\boldsymbol{x}_{0,\varepsilon}, t_{0,n,\varepsilon}, t_{0,a,\varepsilon}, v_\varepsilon)$ be converging to $\boldsymbol{p} = (\boldsymbol{x}_0, t_{0,n}, t_{0,a}, v)$ for $\varepsilon \to 0$ in the sense that $\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2 < \varepsilon$ for all $\varepsilon > 0$ holds. Consequently it also holds that

$$\left.\begin{aligned} &|v_\varepsilon - v| \\ &|t_{0,n,\varepsilon} - t_{0,n}| \\ &|t_{0,a,\varepsilon} - t_{0,a}| \\ &\|\boldsymbol{x}_{0,\varepsilon} - \boldsymbol{x}_0\|_2 \end{aligned}\right\} \leq \|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2 < \varepsilon.$$

Let $\delta > 0$ be arbitrary now. Because of the converging parameter set, we can choose $\varepsilon' > 0$ such that for all $\varepsilon < \varepsilon'$ even

$$\begin{aligned} |v_\varepsilon - v| &< \frac{\delta}{3T} \\ |v_\varepsilon t_{0,j,\varepsilon} - v t_{0,j}| &< \frac{\delta}{3} \\ \|\boldsymbol{x}_{0,\varepsilon} - \boldsymbol{x}_0\|_2 &< \frac{\delta}{3} \end{aligned}$$

holds. It follows now for arbitrary $(\boldsymbol{x}, t) \in \Omega_T$ and $\varepsilon < \varepsilon'$

$$\begin{aligned} &|k_j(\boldsymbol{p}_\varepsilon, \boldsymbol{x}, t) - k_j(\boldsymbol{p}, \boldsymbol{x}, t)| \\ &= |v_\varepsilon(t - t_{0,j,\varepsilon}) - \|\boldsymbol{x}_{0,\varepsilon} - \boldsymbol{x}\|_2 - v(t - t_{0,j}) + \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2| \\ &\overset{\text{triangle inequality}}{\leq} \underbrace{t|v_\varepsilon - v|}_{<T\frac{\delta}{3T}} + \underbrace{|v_\varepsilon t_{0,j,\varepsilon} - v t_{0,j}|}_{<\frac{\delta}{3}} + \underbrace{\big|\|\boldsymbol{x}_{0,\varepsilon} - \boldsymbol{x}\|_2 - \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2\big|}_{\substack{\text{reverse triangle inequality} \\ \leq \|\boldsymbol{x}_{0,\varepsilon} - \boldsymbol{x} - \boldsymbol{x}_0 + \boldsymbol{x}\|_2 \\ = \|\boldsymbol{x}_{0,\varepsilon} - \boldsymbol{x}_0\|_2 < \frac{\delta}{3}}} \\ &< T\frac{\delta}{3T} + \frac{\delta}{3} + \frac{\delta}{3} = \delta. \end{aligned}$$

Since we choose ε' independently from $(\boldsymbol{x}, t) \in \Omega_T$, this proves the uniformly convergence of $(k_j)_\varepsilon$ to k_j for $j = a, n$ for $\boldsymbol{p}_\varepsilon \overset{\varepsilon \to 0}{\longrightarrow} \boldsymbol{p}$ and with $\delta \to 0$. □

With these prerequisites at hand, we can now show the uniform convergence of I_2.

Theorem 5.70
We define a sequence of functions $((I_2)_\varepsilon)_{\varepsilon>0}$ by

$$(I_2)_\varepsilon := I_2(\boldsymbol{p}_\varepsilon, \cdot) : \Omega_T \to [0,1].$$

for a sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ in the parameter space $\boldsymbol{P}$ and with the classification image I_2 as defined in Definition 4.4. Let the sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ be converging to $\boldsymbol{p} \in \boldsymbol{P}$ for $\varepsilon \to 0$ in the sense that $\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2 < \varepsilon$ for all $\varepsilon > 0$ holds. Then it holds that

$$(I_2)_\varepsilon = I_2(\boldsymbol{p}_\varepsilon, (\cdot)) \to I_2(\boldsymbol{p}, (\cdot)) \quad \text{for } \varepsilon \to 0$$

with uniform convergence in the spatio-temporal domain Ω_T.

Proof. For a converging sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ to $\boldsymbol{p} \in \boldsymbol{P}$ for $\varepsilon \to 0$ we get with Propositions 5.66, 5.68 and 5.69 that each summand in the definition of the classification image in Definition 4.4, i.e.,

$$\frac{1}{1+\exp\left(-\frac{1}{\varepsilon_0} k_j\left(\boldsymbol{p}_\varepsilon, \boldsymbol{x}, t\right)\right)} \quad \text{for } j = n, a,$$

is uniformly convergent in the spatio-temporal domain Ω_T for $\varepsilon \to 0$.
With Proposition 5.67, we conclude that then the classification image combining both summands converges uniformly for $\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon \to 0} \boldsymbol{p}$. This proves the statement. □

We conclude an equivalent statement when considering the discretized classification images I_2^{h}.

Theorem 5.71
We define a sequence of functions $\left(\left(I_2^{\mathrm{h}}\right)_\varepsilon\right)_{\varepsilon>0}$ by

$$\left(I_2^{\mathrm{h}}\right)_\varepsilon := I_2^{\mathrm{h}}\left(\boldsymbol{p}_\varepsilon, \cdot\right) : \Omega_T \to [0,1].$$

for a sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ in the parameter space $\boldsymbol{P}$ and with the discretized classification image I_2^{h} living on a pixel grid as defined in Definition 5.63. Let the sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ be converging to $\boldsymbol{p} \in \boldsymbol{P}$ for $\varepsilon \to 0$ in the sense that $\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2 < \varepsilon$ for all $\varepsilon > 0$ holds. Then it holds that

$$\left(I_2^{\mathrm{h}}\right)_\varepsilon = I_2^{\mathrm{h}}\left(\boldsymbol{p}_\varepsilon, (\cdot)\right) \to I_2^{\mathrm{h}}\left(\boldsymbol{p}, (\cdot)\right) \quad \text{for } \varepsilon \to 0,$$

with uniform convergence in the spatio-temporal domain Ω_T.

Proof. The statement follows directly with Theorem 5.70. We consider $(\boldsymbol{x}, t) \in \Omega_T$ to be arbitrary. It holds that

$$\left|I_2^{\mathrm{h}}\left(\boldsymbol{p}_\varepsilon, (\boldsymbol{x}, t)\right) - I_2^{\mathrm{h}}\left(\boldsymbol{p}, (\boldsymbol{x}, t)\right)\right| = \left|I_2\left(\boldsymbol{p}_\varepsilon, \left(\boldsymbol{z}_{\tilde{p}}, t\right)\right) - I_2\left(\boldsymbol{p}, \left(\boldsymbol{z}_{\tilde{p}}, t\right)\right)\right|$$

with $\boldsymbol{z}_{\tilde{p}}$ the center point of the pixel containing $\boldsymbol{x}$. With this equality, we can trace the statement back to the one in Theorem 5.70 which states the uniform convergence in the spatio-temporal domain of $I_2\left(\boldsymbol{p}_\varepsilon, \cdot\right)$ to $I_2\left(\boldsymbol{p}, \cdot\right)$ for the parameter $\boldsymbol{p}_\varepsilon$ converging to $\boldsymbol{p}$ for $\varepsilon \to 0$. □

We conclude this section on the convergence of the classification images with the following theorem.

Theorem 5.72
We define a sequence of discrete classification images $\left(\left(I_2^{\mathrm{h}}\right)_\varepsilon\right)_{\varepsilon>0}$ by

$$\left(I_2^{\mathrm{h}}\right)_\varepsilon := I_2^{\mathrm{h}}\left(\boldsymbol{p}_\varepsilon, (\cdot)\right) : \Omega_T \to [0,1]$$

for a sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ in the parameter space $\boldsymbol{P}$ and a pixel width h based on Definition 5.63. Let the pixel width h be converging to 0 and let the sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ be converging to $\boldsymbol{p} \in \boldsymbol{P}$ for $\varepsilon \to 0$

in the sense that $\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2 < \varepsilon$ for all $\varepsilon > 0$ holds. Then it holds that the sequence of functions $\left(\left(I_2^{\mathrm{h}}\right)_\varepsilon\right)_{\varepsilon>0}$ converges in $L^1(\Omega_T)$, i.e.,

$$I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon, (\cdot)) \xrightarrow{L^1} I_2(\boldsymbol{p}, (\cdot)).$$

Proof. The statement is a direct consequence of Theorem 5.70 and proposition 5.65. We derive

$$\begin{aligned}
&\left\| I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon, \cdot) - I_2(\boldsymbol{p}, \cdot) \right\|_{L^1(\Omega_T)} = \int_{\Omega_T} \left| I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon, (\boldsymbol{x}, t)) - I_2(\boldsymbol{p}, (\boldsymbol{x}, t)) \right| \mathrm{d}(\boldsymbol{x}, t) \\
&= \int_{\Omega_T} \left| I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon, (\boldsymbol{x}, t)) - I_2(\boldsymbol{p}_\varepsilon, (\boldsymbol{x}, t)) + I_2(\boldsymbol{p}_\varepsilon, (\boldsymbol{x}, t)) - I_2(\boldsymbol{p}, (\boldsymbol{x}, t)) \right| \mathrm{d}(\boldsymbol{x}, t) \\
&\overset{\text{triangle inequality}}{\leq} \underbrace{\int_{\Omega_T} \left| I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon, (\boldsymbol{x}, t)) - I_2(\boldsymbol{p}_\varepsilon, (\boldsymbol{x}, t)) \right| \mathrm{d}(\boldsymbol{x}, t)}_{\longrightarrow 0 \text{ for } \mathrm{h}\to 0, \text{ cf. Proposition 5.65}} + \underbrace{\int_{\Omega_T} \left| I_2(\boldsymbol{p}_\varepsilon, (\boldsymbol{x}, t)) - I_2(\boldsymbol{p}, (\boldsymbol{x}, t)) \right| \mathrm{d}(\boldsymbol{x}, t)}_{R.S.}
\end{aligned}$$

and apply the Hölder inequality on the second summand

$$R.S. \leq \underbrace{\int_{\Omega_T} |1| \, \mathrm{d}(\boldsymbol{x}, t)}_{L \cdot W \cdot T < \infty} \underbrace{\left\| I_2(\boldsymbol{p}_\varepsilon, \cdot) - I_2(\boldsymbol{p}, \cdot) \right\|_\infty}_{\longrightarrow 0 \text{ for } \varepsilon\to 0, \text{ cf. Theorem 5.70}}.$$

This proves the convergence of $I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon, \cdot) \xrightarrow{L^1} I_2(\boldsymbol{p}, \cdot)$ for $\mathrm{h} \to 0$ and $\boldsymbol{p}_\varepsilon \to \boldsymbol{p}$ for $\varepsilon \to 0$. □

The statement in Theorem 5.71 could be used for an alternative proof of the last theorem when considering a different way of "adding 0", i.e., by adding and subtracting $I_2^{\mathrm{h}}(\boldsymbol{p}, \cdot)$ before the application of the triangle inequality. However, we do not focus in more details on such an alternative approach. With this we finalize the first convergence results dealing with the convergence for discrete feature and classification images and the uniform continuity of the classification image on the parameter setting. In the next section, we focus on the convergence of discrete histograms and their corresponding histogram density functions.

5.3.4 Convergence of histogram density functions

In this section, we deal with the convergence of the discretized histogram measure. More precisely, we are interested in the L^1-convergence of the corresponding histogram density functions. To this end, we start with a brief recapitulation of the main histogram versions and recall or introduce corresponding density functions. The histograms are related to our mappings for the feature images I_1^{d} (cf. Definition 3.5) and the classification images I_2 (cf. Definition 4.4) as well as the joint mapping of feature images and classification images $\boldsymbol{I}$ (cf. Definition 5.21). The histogram measures of our mappings indeed have densities with respect to the Lebesgue measures on the measurable spaces $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$, $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$ or $(\mathcal{F} \times \mathcal{C}, \mathcal{B}(\mathcal{F} \times \mathcal{C}))$, respectively. In Section 3.3.1, we focused on corresponding probability density functions for probability measures which coincide with the histogram measures after a normalization step. We point out that because of the noise effects on the feature images (cf. Propositions 3.7 and 3.12) and the way we generate our classification images based on a smoothed Heaviside

function (Definition 4.4), we can indeed assume that the histogram measure $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}$ has a density function $h_{\mathcal{F}\times\mathcal{C}}$ with respect to the Lebesgue measure on $(\mathcal{F}\times\mathcal{C}, \mathcal{B}(\mathcal{F}\times\mathcal{C}))$.

Definition 5.73 (Histograms part 1: recapitulation)
We differentiate between a continuous histogram and its smoothed version.

1. We have an ideal histogram (cf. Definition 5.25) without any discretization given as the pushforward measure $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}} = \boldsymbol{I}_{\#}\lambda$ and it holds

$$\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}\left(A'\right) = \int_{A'} 1 \, \mathrm{d}\left(\boldsymbol{I}_{\#}\lambda\right)(f,c) = \left(\boldsymbol{I}_{\#}\lambda\right)\left(A'\right)$$

   for an arbitrary $A' \in \mathcal{B}(\mathcal{F}\times\mathcal{C})$.

2. With a smooth mollification kernel $\eta_{\varepsilon_1} := \frac{1}{\varepsilon_1}\eta\left(\frac{\cdot}{\varepsilon_1}\right)$, we use

$$\boldsymbol{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}} := \eta_{\varepsilon_1} *_{\mathcal{C}} \boldsymbol{H}_{\mathcal{F}\times\mathcal{C}} \tag{5.33}$$

   as a smoothed histogram along the $\mathcal{C}$−axis for the original histogram measure $\boldsymbol{H}_{\mathcal{F}\times\mathcal{C}}$ (cf. Definition 5.33).

We continue with more definitions related to our histogram measures and which take the various discretization effects into consideration.

Remark 5.74 (Histograms part 2: recapitulation & extension). We recapitulate the definition of a discretized histogram and its density function when considering a binning of the feature and classification spaces. Moreover, we introduce the histogram measures related to classification and feature images living on a discretized pixel grid. We state another notation for a discretized histogram considering both the binning effects and the discretized image mappings. Finally, we take the smoothing effects into consideration which are based on the convolution with a smooth mollifier.

1. Considering discrete bins $\mathcal{B}_{\mathcal{F},i}, \mathcal{B}_{\mathcal{C},j}$ for $i \in \{1,\dots,N_{\mathcal{F}}\}$, $j \in \{1,\dots,N_{\mathcal{C}}\}$ of binning sizes Δc and Δf for the feature space $\mathcal{F}'$ and the classification space $\mathcal{C}'$ (cf. Definition 5.41), we denote with

$$\begin{aligned}\hat{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}\left(A'\right) &= \sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \frac{\left|A' \cap \left(\mathcal{B}_{\mathcal{F},i}\times\mathcal{B}_{\mathcal{C},j}\right)\right|}{\left|\mathcal{B}_{\mathcal{F},i}\right|\cdot\left|\mathcal{B}_{\mathcal{C},j}\right|}\left(\boldsymbol{I}(\boldsymbol{p},\cdot)_{\#}\lambda\right)\left(\mathcal{B}_{\mathcal{F},i}\times\mathcal{B}_{\mathcal{C},j}\right)\\ &= \frac{1}{\Delta c\Delta f^3}\sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}}\left|A' \cap \left(\mathcal{B}_{\mathcal{F},i}\times\mathcal{B}_{\mathcal{C},j}\right)\right|\left(\boldsymbol{I}(\boldsymbol{p},\cdot)_{\#}\lambda\right)\left(\mathcal{B}_{\mathcal{F},i}\times\mathcal{B}_{\mathcal{C},j}\right)\end{aligned}$$

   the discrete histogram for all $A' \in \mathcal{B}(\mathcal{F}'\times\mathcal{C}')$. We refer to Definition 5.42 for more details on the discrete histograms. We additionally recall from the same definition the histogram density function to be given by

$$\hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c) = \frac{1}{\Delta c\Delta f^3}\sum_{i,j=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \mathbf{1}_{\mathcal{B}_{\mathcal{F},i}}(f)\,\mathbf{1}_{\mathcal{B}_{\mathcal{C},j}}(c)\left(\boldsymbol{I}(\boldsymbol{p},\cdot)_{\#}\lambda\right)\left(\mathcal{B}_{\mathcal{F},i}\times\mathcal{B}_{\mathcal{C},j}\right)$$

   for all $(f,c) \in \mathcal{F}'\times\mathcal{C}'$ and $\hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c) = 0$ for all $(f,c) \in \mathcal{F}\times\mathcal{C}\backslash\mathcal{F}'\times\mathcal{C}'$.

2. We consider discretization effects in the image data and parameter settings, i.e., a parameter set $\boldsymbol{p}_\varepsilon$ of the sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ in $\boldsymbol{P}$ converging to $\boldsymbol{p}$ and a small width $h > 0$ of an underlying pixel grid for the discretized images. The histogram measure based on this discretized image data is given as

$$\boldsymbol{H}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} = \boldsymbol{I}^{\mathrm{h}}(\boldsymbol{p}_\varepsilon, \cdot)_{\#}\lambda.$$

   For notational issues, we do not add an ε in the superscript marking the approximation of the parameter settings. We rather consider the superscript h to represent both the pixel width and the ε parameter.

3. We denote a histogram measure that considers discrete binnings in the features and classification space as well as discretization effects in the underlying image mappings by $\hat{\boldsymbol{H}}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$. It is defined as

$$\hat{\boldsymbol{H}}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(A') = \frac{1}{\Delta c \Delta f^3} \sum_{i,j=1}^{N_{\boldsymbol{\mathcal{F}}}, N_{\boldsymbol{\mathcal{C}}}} \left|A' \cap (\mathcal{B}_{\boldsymbol{\mathcal{F}},i} \times \mathcal{B}_{\boldsymbol{\mathcal{C}},j})\right| \boldsymbol{H}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(\mathcal{B}_{\boldsymbol{\mathcal{F}},i} \times \mathcal{B}_{\boldsymbol{\mathcal{C}},j})$$

   for all $A' \in \mathcal{B}(\boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}')$. We denote the histogram density function by

$$\hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f, c) = \frac{1}{\Delta c \Delta f^3} \sum_{i,j=1}^{N_{\boldsymbol{\mathcal{F}}}, N_{\boldsymbol{\mathcal{C}}}} \mathbf{1}_{\mathcal{B}_{\boldsymbol{\mathcal{F}},i}}(f)\, \mathbf{1}_{\mathcal{B}_{\boldsymbol{\mathcal{C}},j}}(c) \left(\boldsymbol{I}^{\mathrm{h}}(\boldsymbol{p}_\varepsilon, \cdot)_{\#}\lambda\right)(\mathcal{B}_{\boldsymbol{\mathcal{F}},i} \times \mathcal{B}_{\boldsymbol{\mathcal{C}},j}) \tag{5.34}$$

   for all $(f, c) \in \boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}'$ and $\hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f, c) = 0$ for all $(f, c) \in \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}} \backslash \boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}'$.

4. We denote histogram measures related to smoothed histograms obtained from convolution along the classification axis with the mollifier η_{ε_1} from Example 5.35 and Notation 5.36 with an ε_1 in the superscript. For example, the smoothed discrete histogram measure influenced by approximation effects due to images living on discrete pixel grids and an approximated parameter setting $\boldsymbol{p}_\varepsilon$ is denoted by $\hat{\boldsymbol{H}}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$.

We remark that we only recapitulate the different histogram definitions for the joint feature and classification space $\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$. The histograms living on the individual spaces $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{C}}$ are derived similarly and can be recapitulated with the stated references.
In the definitions of the histogram measures living on the spaces $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$, we consider the push-forward measures of the random variables I^{d}_1 and $\boldsymbol{I}$ with respect to the measure λ (cf. Notation 5.18). We refer additionally to Lemma 5.47 to remind the reader of the derived piecewise constant density and probability density functions. The histogram density functions $\hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ and $\hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ are defined on the space $\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$. However, we stress that according to their definitions they are considered to be equal to zero on $\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}} \backslash \boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}}'$ (cf. Definition 5.42).
Before we delve into the convergence statement, we want to specify the notion of the histogram measures related to image mappings on discretized pixel grids more thoroughly when considering the feature images to be affected by Gaussian noise. For this purpose, we recapitulate and conclude in the following lemma the relations between the different histograms and feature mappings corresponding to the feature space $\boldsymbol{\mathcal{F}}$. Similar relations for the joint feature and classification mappings can be derived analogously.

Lemma 5.75
We consider the feature image I_1 (cf. Definition 3.2), the disturbed feature image I_1^{d} (cf. Definition 3.5) and the piecewise constant feature image I_1^{h} (cf. Definition 5.56). Let $I_1^{\mathrm{d,h}}$ be the discretized feature image affected by Gaussian noise, defined analogously to I_1^{d} by

$$I_1^{\mathrm{d,h}} : \Omega_0 \times \Omega_T \to \mathcal{F}, \qquad I_1^{\mathrm{d,h}}(\omega, (\boldsymbol{x}, t)) = I_1^{\mathrm{h}}(\boldsymbol{x}, t) + I_N(\omega, (\boldsymbol{x}, t)). \tag{5.35}$$

We consider the pushforward measures of κ with respect to the original feature image and discretized feature image to be given by given by $I_{1\#}\kappa$ and $I_{1\#}^{\mathrm{h}}\kappa$. With the measure $\lambda = P^0 \otimes \kappa$ on the measurable space $(\Omega_0 \times \Omega_T, \mathcal{E} \otimes \mathcal{B}(\Omega_T))$, it holds that

$$\begin{aligned} I_{1\#}^{\mathrm{d}}\lambda &= P_N * I_{1\#}\kappa, \\ I_{1\#}^{\mathrm{d,h}}\lambda &= P_N * I_{1\#}^{\mathrm{h}}\kappa. \end{aligned} \tag{5.36}$$

We skip the proof of this lemma because the statements can be proven similarly to Proposition 3.7 in Section 3.3.1. We stress that $*$ refers here to the convolution of measures of the measurable space $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$. In similar statements corresponding to the joint feature and classification space $\mathcal{F} \times \mathcal{C}$, one would consider the convolution along the space $\mathcal{F}$.

To focus now on the convergence statement for the histogram density functions, we use the following discretization aspects. We consider now the approximation of a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$ by the sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0} \subset \boldsymbol{P}$ with $\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon\to 0} \boldsymbol{p}$ in the sense that $\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2 < \varepsilon$ holds for all $\varepsilon > 0$. Moreover, we include discretizations effects due to image mappings living on pixel grids of width h and binning width Δf and Δc for histogram measures. Finally, we take into consideration the smoothing of the histogram along the classification space $\mathcal{C}$ by a mollifier scaled by $\varepsilon_1 > 0$ and introduced for deriving the histogram derivatives (cf. Section 5.2.4, especially Definition 5.33 and Theorem 5.40). We show that for a certain convergence order of the different discretization parameters the histogram density functions convergence. We state this in the upcoming theorem and based on various substeps dealing with intermediate convergence results we prove the statement. Since we consider for the main theorem and the lemmas focusing on intermediate results always the same setting, we start by defining this general setting first.

Definition 5.76 (Setting for L^1-convergence of histogram density functions)
Based on the following assumptions and considerations, we define the *general setting* for the upcoming convergence results in this section. Let $\hat{h}^{\mathrm{h},\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}$ be the piecewise constant histogram density function related to a discretized setting considering the smoothing mollification along $\mathcal{C}$, the binning of feature and classification spaces, the image mappings $I_1^{\mathrm{d,h}}$ and I_2^{h} based on a discrete pixel grid as well as the approximation of a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$. We assume that the various discretization parameters converge according to

$$h \to 0, \qquad \boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon\to 0} \boldsymbol{p}, \qquad \Delta c \to 0, \qquad \Delta f \to 0, \qquad \varepsilon_1 \to 0 \tag{5.37}$$

by preserving the convergence of the following relations

$$\frac{h}{\Delta c} \to 0, \qquad \frac{\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2}{\Delta c} \to 0, \qquad \frac{\Delta c}{\varepsilon_1} \to 0, \tag{5.38}$$

i.e., we enforce that h and $\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2$ converge faster to 0 than Δc and similarly Δc converges faster to 0 than the parameter ε_1 scaling the mollification kernel's width.

Based on this general setting, we state our main theorem for this section:

Theorem 5.77 (L^1-convergence of histogram density functions)
We consider the setting and convergence orders introduced in Definition 5.76. Then it holds that

$$\hat{h}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \xrightarrow{L^1} h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$$

with $h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ being the L^1-function describing the density of the histogram measure when considering the original setting without any discretization effects, i.e., the density function of the pushforward measure $\boldsymbol{I}_\#\lambda$.

We postpone the proof of the main result. For its preparation, we begin with intermediate results in the upcoming lemmas. To get an overview of the considered substeps, we list the intermediate convergence statements first:

1. In Lemma 5.78, we show the convergence $\hat{h}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \xrightarrow{L^1} \hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ for $\varepsilon_1 \to 0$.

2. In Lemma 5.79, we show the convergence $\hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \xrightarrow{L^1} h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ for $\Delta c, \Delta f \to 0$.

3. In Lemma 5.83, we show the convergence $\hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \xrightarrow{L^1} \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ for $\mathrm{h} \to 0$, $\frac{h}{\Delta c} \to 0$, $\boldsymbol{p}_\varepsilon \to \boldsymbol{p}$ and $\frac{\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2}{\Delta c} \to 0$ by considering the following convergences:

   a) In Lemma 5.80, we show the convergence $\sum_{m=1}^{N_{\boldsymbol{\mathcal{F}}}} \left| I_1^{\mathrm{d,h}}{}_\# \lambda - I_1^{\mathrm{d}}{}_\# \lambda \right| (\mathcal{B}_{\boldsymbol{\mathcal{F}},m}) \to 0$ for $\mathrm{h} \to 0$.

   b) In Lemma 5.81, we show the convergence $\sum_{n=1}^{N_{\boldsymbol{\mathcal{C}}}} \left| I_2^{\mathrm{h}}(\boldsymbol{p},\cdot)_\# \kappa - I_2(\boldsymbol{p},\cdot)_\# \kappa \right| (\mathcal{B}_{\boldsymbol{\mathcal{C}},n}) \to 0$ for $\mathrm{h} \to 0$ and $\frac{h}{\Delta c} \to 0$.

      We refer to Figure 5.1 for a motivation that the convergence order of $\Delta c \to 0$ and $\mathrm{h} \to 0$ given by $\frac{h}{\Delta c} \to 0$ is indeed crucial.

   c) In Lemma 5.82, we show the convergence $\sum_{n=1}^{N_{\boldsymbol{\mathcal{C}}}} |I_2(\boldsymbol{p}_\varepsilon,\cdot)_\# \kappa - I_2(\boldsymbol{p},\cdot)_\# \kappa| (\mathcal{B}_{\boldsymbol{\mathcal{C}},n}) \to 0$ for $\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon\to 0} \boldsymbol{p}$ and $\frac{\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2}{\Delta c} \to 0$.

The convergences in Items 1 to 3 are the main ingredients for the proof to show the convergence $\hat{h}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \xrightarrow{L^1} h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$ stated in our main Theorem 5.77.
Following now the substeps listed above, we start with the first intermediate convergence result in Item 1.

Lemma 5.78 (Mollification effect along $\mathcal{C}$)
We consider the setting introduced in Definition 5.76. Then it holds that $\hat{h}^{\mathrm{h},\varepsilon_1}_{\mathcal{F}\times\mathcal{C}} \xrightarrow{L^1} \hat{h}^{\mathrm{h}}_{\mathcal{F}\times\mathcal{C}}$ for the convergence of $\varepsilon_1 \to 0$.

Proof. We know that $\hat{h}_{\mathcal{F}\times\mathcal{C}}$ is a L^1-function because it holds that

$$\int_{\mathcal{F}'\times\mathcal{C}'} \left|\hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c)\right| \,\mathrm{d}(f,c) = \int_{\mathcal{F}'\times\mathcal{C}'} \hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c) \,\mathrm{d}(f,c) \leq \kappa(\Omega_T) < \infty.$$

We refer to Remark 5.44, to recall why the discretized histograms on the reduced feature and classification spaces only approximate the total mass $\kappa(\Omega_T)$.
The statement follows then directly with Theorem 2.14 (2) in [2] which states that for any $f \in L^p$ with $1 \leq p < \infty$ and a Dirac sequence such as our $(\eta_{\varepsilon_1})_{\varepsilon_1>0}$ it holds that

$$f * \eta_{\varepsilon_1} \to f \qquad \text{in } L^p \text{ for } \varepsilon_1 \to 0.$$

Consequently, the convergence holds in particular also for $p = 1$. □

Next, we focus on the convergence stated in Item 2.

Lemma 5.79 (Binning of spaces $\mathcal{F}'$ and $\mathcal{C}'$)
We consider the setting introduced in Definition 5.76. Then it holds that $\left\|\hat{h}_{\mathcal{F}\times\mathcal{C}} - h_{\mathcal{F}\times\mathcal{C}}\right\|_{L^1} \to 0$ for the convergences of $\Delta c, \Delta f \to 0$.

Proof. We know that $\hat{h}_{\mathcal{F}\times\mathcal{C}}$ and $h_{\mathcal{F}\times\mathcal{C}}$ are L^1-functions because it holds that

$$\int_{\mathcal{F}'\times\mathcal{C}'} \left|\hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c)\right| \,\mathrm{d}(f,c) = \int_{\mathcal{F}'\times\mathcal{C}'} \hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c) \,\mathrm{d}(f,c) \leq \kappa(\Omega_T) < \infty,$$
$$\int_{\mathcal{F}'\times\mathcal{C}'} \left|h_{\mathcal{F}\times\mathcal{C}}(f,c)\right| \,\mathrm{d}(f,c) = \int_{\mathcal{F}'\times\mathcal{C}'} h_{\mathcal{F}\times\mathcal{C}}(f,c) \,\mathrm{d}(f,c) \leq \kappa(\Omega_T) < \infty.$$

We refer to Remark 5.44, to recall why the discretized histograms on the reduced feature and classification spaces only approximate the total mass $\kappa(\Omega_T)$.
Furthermore, we recall that the piecewise constant histogram density function $\hat{h}_{\mathcal{F}\times\mathcal{C}}$ is constant on each bin combination $\mathcal{B}_{\mathcal{F},i} \times \mathcal{B}_{\mathcal{C},j}$ with $i \in \{1,\dots,N_{\mathcal{F}}\}$ and $j \in \{1,\dots,N_{\mathcal{C}}\}$ and coincides with the averaged pushforward measure of this bin (cf. Remark 5.74), i.e., for $(f,c) \in \mathcal{B}_{\mathcal{F},i} \times \mathcal{B}_{\mathcal{C},j}$ it holds that

$$\hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c) = \frac{1}{\Delta c \Delta f^3} \left(\boldsymbol{I}(\boldsymbol{p},\cdot)_{\#}\lambda\right)(\mathcal{B}_{\mathcal{F},i} \times \mathcal{B}_{\mathcal{C},j}) = \frac{1}{\Delta c \Delta f^3} \int_{\mathcal{B}_{\mathcal{F},i}\times\mathcal{B}_{\mathcal{C},j}} h_{\mathcal{F}\times\mathcal{C}}(f,c) \,\mathrm{d}(f,c).$$

We consider now $\delta > 0$ to be arbitrary. Since $C_c^\infty(\mathcal{F}\times\mathcal{C})$ lies dense in $L^1(\mathcal{F}\times\mathcal{C})$ (cf. Theorem 2.14 (3) in [2]), we can find a continuous function $g \in C_c^\infty(\mathcal{F}\times\mathcal{C})$ such that

$$\|h_{\mathcal{F}\times\mathcal{C}} - g\|_{L^1} < \delta \tag{5.39}$$

holds. Since g is continuous, we can conclude that for the same $\delta > 0$ there exists a $\delta' > 0$ such that

$$|g(f_1, c_1) - g(f_2, c_2)| < \delta$$

for all $(f_1, c_1), (f_2, c_2) \in \mathcal{F} \times \mathcal{C}$ with $\|(f_1, c_1) - (f_2, c_2)\|_2 < \delta'$. Since we consider $\Delta c, \Delta f \to 0$, we can assume Δc and Δf to be small enough such that $\Delta c + 3\Delta f < \delta'$ holds. Then it follows for $(f, c), (f', c') \in \mathcal{B}_{\mathcal{F},i} \times \mathcal{B}_{\mathcal{C},j}$ with $i \in \{1, \dots, N_{\mathcal{F}}\}$ and $j \in \{1, \dots, N_{\mathcal{C}}\}$ arbitrary that

$$\left\|(f, c) - \left(f', c'\right)\right\|_2 < \Delta c + 3\Delta f < \delta' \tag{5.40}$$

and, consequently,

$$\left|g(f, c) - g\left(f', c'\right)\right| < \delta \tag{5.41}$$

hold. Based on these binning widths Δc and Δf, we define a piecewise constant approximation of g by

$$\hat{g}(f, c) = \frac{1}{\Delta c \Delta f^3} \sum_{i,j=1}^{N_{\mathcal{F}}, N_{\mathcal{C}}} \mathbf{1}_{\mathcal{B}_{\mathcal{F},i}}(f)\, \mathbf{1}_{\mathcal{B}_{\mathcal{C},j}}(c) \int\limits_{\mathcal{B}_{\mathcal{F},i} \times \mathcal{B}_{\mathcal{C},j}} g(f, c)\ \mathrm{d}(f, c).$$

Consequently, for any (f, c) in an arbitrary bin combination $\mathcal{B}_{\mathcal{F},i} \times \mathcal{B}_{\mathcal{C},j}$ with $i \in \{1, \dots, N_{\mathcal{F}}\}$ and $j \in \{1, \dots, N_{\mathcal{C}}\}$ it holds that

$$\hat{g}(f, c) = \frac{1}{\Delta c \Delta f^3} \int\limits_{\mathcal{B}_{\mathcal{F},i} \times \mathcal{B}_{\mathcal{C},j}} g\left(f', c'\right)\ \mathrm{d}\left(f', c'\right).$$

Now, we can conclude that

$$\begin{aligned}
\|g - \hat{g}\|_{L^1} &= \int\limits_{\mathcal{F}' \times \mathcal{C}'} |g(f, c) - \hat{g}(f, c)|\ \mathrm{d}(f, c) \\
&= \sum_{m,n=1}^{N_{\mathcal{F}}, N_{\mathcal{C}}} \int\limits_{\mathcal{B}_{\mathcal{F},m} \times \mathcal{B}_{\mathcal{C},n}} \left| g(f, c) - \frac{1}{\Delta c \Delta f^3} \int\limits_{\mathcal{B}_{\mathcal{F},m} \times \mathcal{B}_{\mathcal{C},n}} g\left(f', c'\right)\ \mathrm{d}\left(f', c'\right) \right| \mathrm{d}(f, c) \\
&\leq \sum_{m,n=1}^{N_{\mathcal{F}}, N_{\mathcal{C}}} \int\limits_{\mathcal{B}_{\mathcal{F},m} \times \mathcal{B}_{\mathcal{C},n}} \frac{1}{\Delta c \Delta f^3} \int\limits_{\mathcal{B}_{\mathcal{F},m} \times \mathcal{B}_{\mathcal{C},n}} \underbrace{\left|g(f, c) - g\left(f', c'\right)\right|}_{<\delta} \mathrm{d}\left(f', c'\right)\ \mathrm{d}(f, c) \\
&< \delta \underbrace{\left|\mathcal{F}'\right| \left|\mathcal{C}'\right|}_{<\infty}
\end{aligned}$$

holds because of the continuity of g and Equations (5.40) and (5.41). We infer that there exists a finite constant $C_{\mathcal{F}', \mathcal{C}'} > 0$ such that $|\mathcal{F}'|\,|\mathcal{C}'| < C_{\mathcal{F}', \mathcal{C}'} < \infty$ and we conclude then that

$$\|g - \hat{g}\|_{L^1} < C_{\mathcal{F}', \mathcal{C}'} \delta \tag{5.42}$$

holds. Next, we approximate the L^1-norm of the difference between both piecewise constant functions $\hat{h}_{\mathcal{F}\times\mathcal{C}}$ and $\hat{g}$:

$$\begin{aligned}
&\left\|\hat{h}_{\mathcal{F}\times\mathcal{C}} - \hat{g}\right\|_{L^1} = \int\limits_{\mathcal{F}'\times\mathcal{C}'} \left|\hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c) - \hat{g}(f,c)\right| \,\mathrm{d}(f,c) \\
&= \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \int\limits_{\mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}} \Big| \underbrace{\hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c)}_{\substack{=\text{const. on}\\ \mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}}} - \underbrace{\hat{g}(f,c)}_{\substack{=\text{const. on}\\ \mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}}} \Big| \,\mathrm{d}(f,c) \\
&= \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \underbrace{\frac{1}{\Delta c \Delta f^3} \int\limits_{\mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}} \mathrm{d}(f,c)}_{=1} \left| \int\limits_{\mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}} h_{\mathcal{F}\times\mathcal{C}}(f',c') \,\mathrm{d}(f',c') - \int\limits_{\mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}} g(f',c') \,\mathrm{d}(f',c') \right| \\
&\leq \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \int\limits_{\mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}} \left|h_{\mathcal{F}\times\mathcal{C}}(f',c') - g(f',c')\right| \,\mathrm{d}(f',c') \\
&= \int\limits_{\mathcal{F}'\times\mathcal{C}'} \left|h_{\mathcal{F}\times\mathcal{C}}(f',c') - g(f',c')\right| \,\mathrm{d}(f',c') = \|h_{\mathcal{F}\times\mathcal{C}} - g\|_{L^1}
\end{aligned}$$

holds and with Equation (5.39) it follows directly that

$$\left\|\hat{h}_{\mathcal{F}\times\mathcal{C}} - \hat{g}\right\|_{L^1} \leq \|h_{\mathcal{F}\times\mathcal{C}} - g\|_{L^1} < \delta \tag{5.43}$$

holds. By applying the Minkowski inequality as well as exploiting Equation (5.39), Equation (5.42) and Equation (5.43), we finally derive that

$$\begin{aligned}
\left\|\hat{h}_{\mathcal{F}\times\mathcal{C}} - h_{\mathcal{F}\times\mathcal{C}}\right\|_{L^1} &= \left\|\hat{h}_{\mathcal{F}\times\mathcal{C}} - \hat{g} + \hat{g} - g + g - h_{\mathcal{F}\times\mathcal{C}}\right\|_{L^1} \\
&\leq \left\|\hat{h}_{\mathcal{F}\times\mathcal{C}} - \hat{g}\right\|_{L^1} + \|\hat{g} - g\|_{L^1} + \|g - h_{\mathcal{F}\times\mathcal{C}}\|_{L^1} \\
&< \left(2 + C_{\mathcal{F}',\mathcal{C}'}\right)\delta
\end{aligned}$$

holds. As we choose for any $\delta > 0$ our binning widths Δf and Δc small enough to account for the epsilon-delta definition of continuity for the function g (cf. Equations (5.40) and (5.41)), a vanishing δ implies directly Δc, $\Delta f \to 0$. Since $\delta > 0$ is arbitrary, the L^1-convergence result finally follows with $\delta \to 0$. □

With Figure 5.1 we motivate the next convergence statements which consider vanishing grid sizes Δf and Δc as well as the pixel width converging to 0. Based on the first sketch on the left which depicts a histogram measure of a piecewise constant image mapping (green) and its discretized version due to binning effects of width Δ (red), we consider two different scenarios. We focus on vanishing discretization sizes for the binning by $\Delta \to 0$ for, e.g., the classification or feature space, and for the pixel width with $h \to 0$. For a piecewise constant image mapping the corresponding pushforward measure consists of several Delta peaks (shown in green). When we consider then a binning of the horizontal axis, e.g., of the feature or classification space, we observe a histogram measure which has a piecewise constant histogram density function (shown in red). We highlight that a decreasing pixel width h corresponds to an increasing number of occurring feature and classification values in the images and, consequently, results in more Delta peaks in the histogram whereas a decreasing

bin width Δ results in a higher number of discrete bins in the histogram. Focusing now on the convergence order, we will expect the red histogram for the binned space to converge towards the green histogram consisting of Delta peaks (upper sketch on the right) if the binning width converges to 0 for a fixed pixel width h and even if the binning width converges faster to zero than the pixel width, i.e., $\frac{\Delta}{\mathrm{h}} \to 0$. The histogram measure consisting of Delta peaks is not absolutely continuous with respect to the Lebesgue measure and, consequently, it does not have a density function with respect to the Lebesgue measure. For this reason, we want to avoid that the binning width Δ converges faster than the pixel width h. In the other case, we will expect the discretized histogram with a piecewise constant density function to converge towards the original density function of the image mapping without a pixel grid discretization (gray graph in lower right plot) if the pixel width converges faster than the binning width, i.e., $\frac{\mathrm{h}}{\Delta} \to 0$. With this in mind, we focus on the next convergence results.

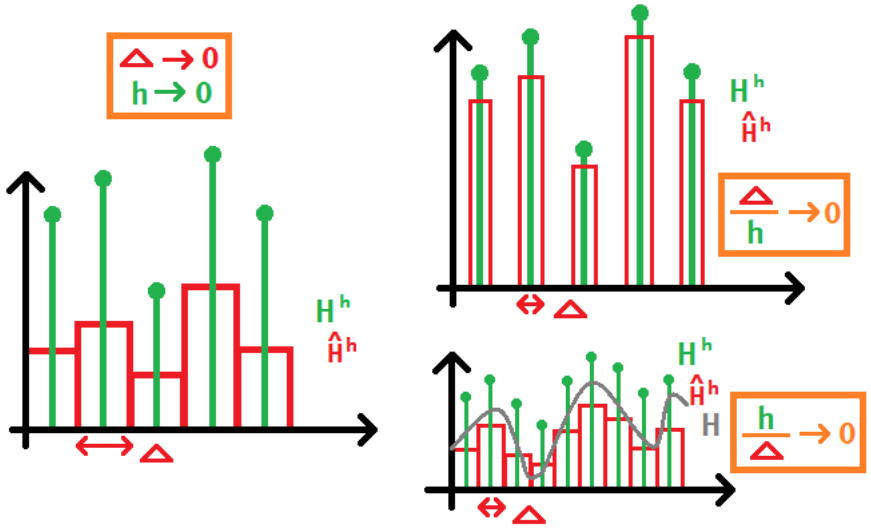


Figure 5.1: Sketches of a histogram measure corresponding to a piecewise constant image mapping living on a pixel grid of width h (green, Delta peaks) and a histogram measure based on a binning of width Δ with a piecewise constant density function (red) are shown for discretization sizes h and Δ converging to 0. Based on the left sketch, two different scenarios for the convergence order are presented. In the upper right sketch the binning width Δ converges faster than h, i.e. $\frac{\Delta}{\mathrm{h}} \to 0$ resulting in $\hat{H}^{\mathrm{h}} \to H^{\mathrm{h}}$, and in the lower right sketch the pixel width converges faster than Δ, i.e. $\frac{\mathrm{h}}{\Delta} \to 0$ allowing $\hat{H}^{\mathrm{h}} \to H$ with H being the original histogram measure without any discretization effects.

Lemma 5.80 (Discretized image mapping $I_1^{\mathrm{d,h}}$ on binned space $\mathcal{F}'$)
We consider the setting introduced in Definition 5.76. Then it holds for $\mathrm{h} \to 0$ that

$$\sum_{m=1}^{N_{\mathcal{F}}} \left| {I_1^{\mathrm{d,h}}}_{\#}\lambda - {I_1^{\mathrm{d}}}_{\#}\lambda \right| (\mathcal{B}_{\mathcal{F},m}) \to 0.$$

Proof. We focus first on an arbitrary bin $\mathcal{B}_{\mathcal{F},m}$ with $m \in \{1, \dots, N_{\mathcal{F}}\}$ and derive the following approximations. For the pushforward measure of $\lambda = P^0 \otimes \kappa$ with respect to the feature image, we recapitulate

that ${I_1^{\mathrm{d,h}}}_{\#}\lambda = P_N * {I_1^{\mathrm{h}}}_{\#}\kappa$ and ${I_1^{\mathrm{d}}}_{\#}\lambda = P_N * {I_{1\#}}\kappa$ hold (cf. Equation (5.36)) and recall that p_N denotes the probability density function of the probability measure P_N with respect to the Lebesgue measure (cf. Definition 3.4). We point out that p_N is the probability density function related to the Gaussian normal distribution for a fixed standard deviation σ^2 living on the feature space $\mathcal{F}$. Moreover, p_N is continuous and also continuously differentiable. Since its first partial derivatives are bounded, the probability density function p_N is Lipschitz continuous. We denote the corresponding Lipschitz constant with L and without calculating it explicitly, we state that naturally $0 < L < \infty$ holds. With this in mind, we derive

$$
\begin{aligned}
&\left|{I_1^{\mathrm{d,h}}}_{\#}\lambda\left(\mathcal{B}_{\mathcal{F},m}\right) - {I_1^{\mathrm{d}}}_{\#}\lambda\left(\mathcal{B}_{\mathcal{F},m}\right)\right| \\
&= \left| \int_{\mathcal{B}_{\mathcal{F},m}} \int_{\mathcal{F}} p_N\left(f-y\right)\ \mathrm{d}\left({I_1^{\mathrm{h}}}_{\#}\kappa\right)(y)\ \mathrm{d}f - \int_{\mathcal{B}_{\mathcal{F},m}} \int_{\mathcal{F}} p_N\left(f-y\right)\ \mathrm{d}\left(I_{1\#}\kappa\right)(y)\ \mathrm{d}f \right| \\
&= \left| \int_{\mathcal{B}_{\mathcal{F},m}} \int_{\Omega_T} p_N\left(f - I_1^{\mathrm{h}}\left(\boldsymbol{x},t\right)\right) - p_N\left(f - I_1\left(\boldsymbol{x},t\right)\right)\ \mathrm{d}\left(\boldsymbol{x},t\right)\ \mathrm{d}f \right| \\
&\leq \int_{\mathcal{B}_{\mathcal{F},m}} \int_{\Omega_T} \left| p_N\left(f - I_1^{\mathrm{h}}\left(\boldsymbol{x},t\right)\right) - p_N\left(f - I_1\left(\boldsymbol{x},t\right)\right)\right|\ \mathrm{d}\left(\boldsymbol{x},t\right)\ \mathrm{d}f \\
&\overset{\text{Equation (5.22)}}{=} \int_{\mathcal{B}_{\mathcal{F},m}} \sum_{i=1}^{n_T} \int_{\Omega} \left| p_N\left(f - I_1^{\mathrm{h}}\left(\boldsymbol{x},t_i\right)\right) - p_N\left(f - I_1\left(\boldsymbol{x},t_i\right)\right)\right|\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}f \\
&\overset{\text{Definition 5.55}}{=} \int_{\mathcal{B}_{\mathcal{F},m}} \sum_{i=1}^{n_T} \sum_{\tilde{p}\in\Omega^{\mathrm{h}}} \int_{\Omega_{\tilde{p}}} \left| p_N\left(f - I_1^{\mathrm{h}}\left(\boldsymbol{x},t_i\right)\right) - p_N\left(f - I_1\left(\boldsymbol{x},t_i\right)\right)\right|\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}f \\
&\overset{L\text{–continuity}}{\leq} \int_{\mathcal{B}_{\mathcal{F},m}} \sum_{i=1}^{n_T} \sum_{\tilde{p}\in\Omega^{\mathrm{h}}} \int_{\Omega_{\tilde{p}}} L \underbrace{\left|\left(f - I_1^{\mathrm{h}}\left(\boldsymbol{x},t_i\right)\right) - \left(f - I_1\left(\boldsymbol{x},t_i\right)\right)\right|}_{=\left|I_1^{\mathrm{h}}\left(\boldsymbol{x},t_i\right) - I_1\left(\boldsymbol{x},t_i\right)\right|}\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}f \\
&\overset{\substack{\text{Proposition 5.60}\\ \text{Theorem 5.61}}}{\leq} \int_{\mathcal{B}_{\mathcal{F},m}} n_T \cdot L \cdot C_{\Omega_1} \cdot h \cdot C_{\mathrm{BV}}\ \mathrm{d}f = \Delta f^3 \cdot n_T \cdot L \cdot C_{\Omega_1} \cdot C_{\mathrm{BV}} \cdot h.
\end{aligned}
$$

Consequently, there exists a finite constant $C_\star > 0$ such that $n_T \cdot L \cdot C_{\Omega_1} \cdot C_{\mathrm{BV}} < C_\star < \infty$ holds and we conclude that

$$
\left|{I_1^{\mathrm{d,h}}}_{\#}\lambda\left(\mathcal{B}_{\mathcal{F},m}\right) - {I_1^{\mathrm{d}}}_{\#}\lambda\left(\mathcal{B}_{\mathcal{F},m}\right)\right| < C_\star\, \Delta f^3\, h
$$

holds for an arbitrary bin $\mathcal{B}_{\mathcal{F},m}$ with $m \in \{1,\dots,N_{\mathcal{F}}\}$. With this at hand, we derive

$$
\begin{aligned}
\sum_{m=1}^{N_{\mathcal{F}}} \left|{I_1^{\mathrm{d,h}}}_{\#}\lambda\left(\mathcal{B}_{\mathcal{F},m}\right) - {I_1^{\mathrm{d}}}_{\#}\lambda\left(\mathcal{B}_{\mathcal{F},m}\right)\right| &< N_{\mathcal{F}}\, C_\star\, \Delta f^3\, h \\
&\overset{\text{Equation (5.8)}}{=} \frac{\left|\mathcal{F}_1'\right|}{\Delta f} \cdot \frac{\left|\mathcal{F}_2'\right|}{\Delta f} \cdot \frac{\left|\mathcal{F}_3'\right|}{\Delta f}\, C_\star\, \Delta f^3\, h = \underbrace{\left|\mathcal{F}_1'\right| \cdot \left|\mathcal{F}_2'\right| \cdot \left|\mathcal{F}_3'\right|}_{<\infty}\, C_\star\, h,
\end{aligned}
$$

i.e., the convergence follows for $\mathrm{h} \to 0$ and in fact even independently from the binning width Δf. □

When recalling the sketches of discretization effects influencing histogram measures as shown in Figure 5.1, we point out that in fact we proved the convergence

$$\sum_{m=1}^{N_{\mathcal{F}}} \left| {I_1^{\mathrm{d,h}}}_{\#}\lambda - {I_1^{\mathrm{d}}}_{\#}\lambda \right| (\mathcal{B}_{\mathcal{F},m}) \to 0$$

independently from the binning width Δf and do not require a specific order of convergence for $\mathrm{h} \to 0$ and $\Delta f \to 0$. We stress that this effect derives from the smoothing effect by the convolution with the Gaussian normal density function p_N to take noise effects into consideration. To be more precise, in our sketch we would not expect green Dirac Delta peaks for ${I_1^{\mathrm{d,h}}}_{\#}\lambda$ but instead smoothed Delta peaks which are in particular smoothed for any pixel width h.
When we focus on the pushforward of κ with respect to the classification image, we show with the next statement that in this context we indeed need to apply the convergence order $\frac{h}{\Delta c} \to 0$ in particular to ensure the following convergence for $I_2^{\mathrm{h}}(\boldsymbol{p},\cdot)_{\#}\kappa \to I_2(\boldsymbol{p},\cdot)_{\#}\kappa$ for any parameter $\boldsymbol{p} \in \boldsymbol{P}$:

Lemma 5.81 (Discretized classification mapping I_2^{h} on binned space $\mathcal{C}'$)
We consider the setting introduced in Definition 5.76. Then it holds with $\boldsymbol{p} \in \boldsymbol{P}$ arbitrary that

$$\sum_{n=1}^{N_{\mathcal{C}}} \left| I_2^{\mathrm{h}}(\boldsymbol{p},\cdot)_{\#}\kappa - I_2(\boldsymbol{p},\cdot)_{\#}\kappa \right| (\mathcal{B}_{\mathcal{C},n}) \to 0$$

for the convergences of $\mathrm{h} \to 0$ and $\frac{h}{\Delta c} \to 0$.

Proof. Let $\boldsymbol{p} \in \boldsymbol{P}$ and $t \in \{t_1, \dots, t_{n_T}\}$ be arbitrary but fixed for now. We focus first on an arbitrary bin $\mathcal{B}_{\mathcal{C},n}$ with $n \in \{1, \dots, N_{\mathcal{C}}\}$ and derive the following approximations. For the pushforward of κ with respect to the classification image in its discretized and its original version, we want to find an approximation for

$$\left| I_2^{\mathrm{h}}(\boldsymbol{p},\cdot,t)_{\#}\kappa - I_2(\boldsymbol{p},\cdot,t)_{\#}\kappa \right| (\mathcal{B}_{\mathcal{C},n}) .$$

In Figure 5.2, we show a sketch of one time frame for the classification image living on the discrete pixel grid (red) compared to the original classification image indicated by three smooth circular contour lines in the two dimensional domain (left sketch) and a radial cut focusing on the course of the classification images I_2 and I_2^{h} in radial direction (right plot). The cutting is illustrated in the two dimensional plot on the left by a purple cutting edge. In the cross-section plot on the right, we use horizontal green lines to mark the different bins of width Δc which are numbered on the vertical axis. Additionally, we indicate pixel borders on the horizontal axis by vertical red lines which are one pixel width h apart from each other. In Figure 5.3, we sketch the course of the classification images I_2 and I_2^{h} for a radial cut for two different settings. On the left side, we consider bins related to classification regions corresponding to small slopes in radial direction of the original mapping I_2 while in the right plot we also consider regions with a high slope in radial direction. Below the two plots, we mark the preimages of the different bins in brown under the original classification image I_2 and under the discretized classification image I_2^{h} in blue, i.e., we present $I_2^{-1}(\mathcal{B}_{\mathcal{C},n})$ and $I_2^{\mathrm{h}^{-1}}(\mathcal{B}_{\mathcal{C},n})$ with $n \in \{1, \dots, 5\}$.
Since the piecewise constant discrete image I_2^{h} is defined based on I_2 by assigning each pixel the value

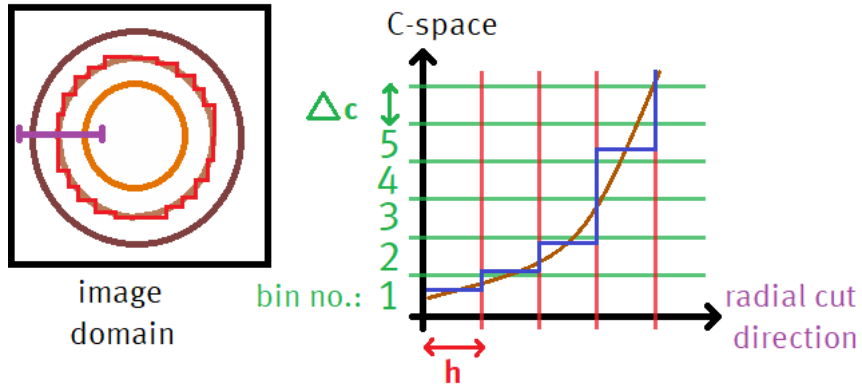


Figure 5.2: On the left, a sketch of three contour lines corresponding to three distinct classification values (brown circles) is presented in the image domain. The contour line of the middle circle is overlaid by a discrete approximation in red, i.e., by the corresponding circle on the discrete pixel grid. A cutting edge in radial direction is shown in purple and the related cross-section plot is shown on the right. The original, continuous classification values are shown in brown and in blue they are approximated by the discrete classification values on the pixel grid according to Definition 5.63. The grid width h is marked on the horizontal axis and the binning width Δc is depicted in green on the vertical axis.

of I_2 in its center point, it cannot happen that the preimage of a bin in $\mathcal{C}'$ is only empty under I_2 but not under I_2^{h}. With this in mind, it is sufficient to only consider bins in $\mathcal{C}'$ for which the preimage under I_2 is not empty because in the other case the preimage under I_2^{h} would be empty as well and, consequently, it would hold that

$$\left| I_2^{\mathrm{h}}(\boldsymbol{p}, \cdot, t)_{\#}\kappa - I_2(\boldsymbol{p}, \cdot, t)_{\#}\kappa \right| (\mathcal{B}_{\mathcal{C},n}) = 0.$$

We differ now between two cases. First, we consider $\mathcal{B}_{\mathcal{C},n}$ such that its preimages under both mappings I_2 and I_2^{h} are not empty. Depending on the slope of I_2 in radial direction and depending on the bin width, it can happen that the preimage of I_2^{h} is empty (cf. bin 3 and bin 4 in the right plot in Figure 5.3). We focus on this setting in the second case.
For the first case, we investigate the regions on the horizontal cut in the image domain which correspond to the preimage of bin 2 in the left plot and to the preimage of bins 1, 2 and 5 in the right plot in Figure 5.3. We observe that the preimages under I_2 and I_2^{h} overlap, i.e., they are not disjoint. However, they are not necessarily identical either. In fact, we can observe for each bin slight shifting effects when comparing the preimage under I_2 and I_2^{h}. Nevertheless, this mismatch is smaller than the pixel width h for both boundary regions in radial direction, i.e., the left and right boundaries of the preimages under I_2 and I_2^{h} for each of the bins focused on for this first case. Considering now again the two dimensional setting, we are facing ring-shaped preimages for the classification image I_2 and discretized ring-shapes for I_2^{h} with staircasing effects based on the underlying pixel grid. Before we dive into the approximation of the error between the preimages of I_2 and I_2^{h} for bin $\mathcal{B}_{\mathcal{C},n}$, we use the following estimate for the perimeter corresponding to the maximal possible radius for a circle to fit into our domain $\Omega = [0, L] \times [0, W]$:

$$2\pi R \leq \mathrm{Per}(\Omega) = 2(L + W) \tag{5.44}$$

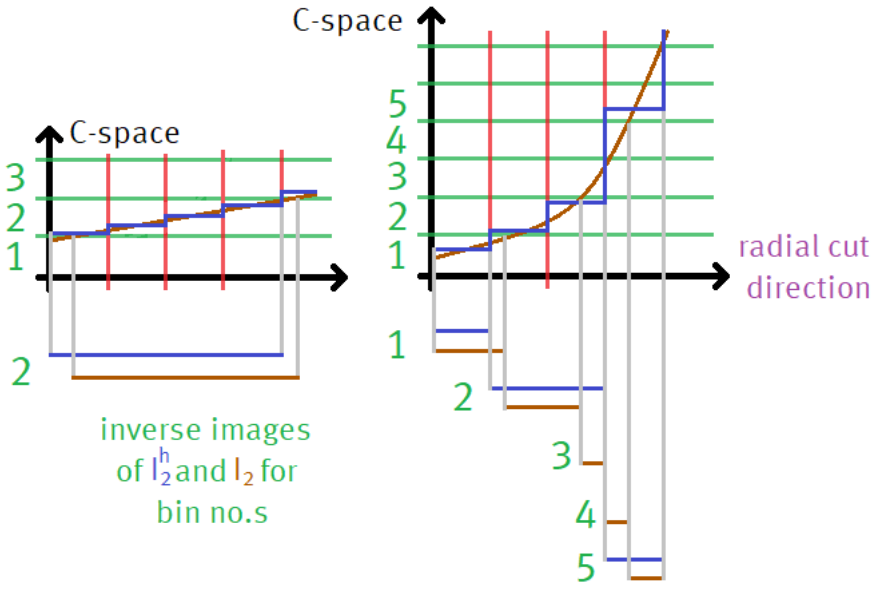


Figure 5.3: Based on the cross-section plot of a horizontal cut in the radial direction introduced in Figure 5.2, we present the regions related to the preimage of the classification mapping I_2 (brown) and its discrete approximation I_2^{h} (blue). On the left, we focus on a bin in the classification space for which the classification image rises only very slightly in the radial direction whereas on the right, we present also bins which correspond to high radial gradients of the classification mapping.

i.e., we use the perimeter of our domain as an upper bound for the circle's perimeter with a radius for which holds $R \leq \frac{1}{2} \min\{L, W\}$. The perimeter is here denoted by $\mathrm{Per}(\cdot)$. Based on the length of the perimeter of this maximal circle and by considering the width of the mismatched regions to be smaller than h on the inner and on the outer ring shapes of the perimeter of the preimages, we derive the following approximation:

$$\begin{aligned}\left|I_2^{\mathrm{h}}(\boldsymbol{p},\cdot,t)_{\#}\kappa - I_2(\boldsymbol{p},\cdot,t)_{\#}\kappa\right|(\mathcal{B}_{\mathcal{C},n}) &\leq \mathrm{h}\cdot\mathrm{Per}\left(I_2^{\mathrm{h}}(\boldsymbol{p},\cdot,t)^{-1}(\mathcal{B}_{\mathcal{C},n}) \cup I_2(\boldsymbol{p},\cdot,t)^{-1}(\mathcal{B}_{\mathcal{C},n})\right)\\ &\leq 2\cdot\mathrm{h}\cdot 2\,\pi\,R \leq 4\,(L+W)\,\mathrm{h}.\end{aligned}$$

Here, we apply an approximation of the perimeter of the union of the preimages of I_2 and I_2^{h} by twice the perimeter of a circle corresponding to the maximal possible radius R which we then in turn approximate by the upper bound given by the perimeter of the domain Ω (cf. Equation (5.44)). We use a quite rough approximation for the considered perimeter of the union of the preimages of I_2 and I_2^{h}. However, we stress that it is important for us to find an upper bound which is independent of the binning width Δc. Eventually, there exists a constant $C^{\dagger} > 0$ such that $4\,(L+W) < C^{\dagger}$ holds, i.e. $C^{\dagger}$ only depends on our spatial domain Ω. For later reference we conclude that

$$\left|I_2^{\mathrm{h}}(\boldsymbol{p},\cdot,t)_{\#}\kappa - I_2(\boldsymbol{p},\cdot,t)_{\#}\kappa\right|(\mathcal{B}_{\mathcal{C},n}) \leq C^{\dagger}\,\mathrm{h} \tag{5.45}$$

holds.

Next, we focus on the second case dealing with regions where the classification image rises significantly in radial direction such that the discretized classification images skips certain bins in the classification space $\mathcal{C}'$, i.e., we consider bins for which the preimage under I_2^{h} is empty while it is not

under I_2. Exemplarily, we refer to the bins 3 and 4 in the right plot of Figure 5.3 for which the preimage of I_2^{h} is empty. Here, the brown graph representing the classification image I_2 rises distinctively so that the discretized image I_2^{h} (piecewise constant blue graph) only maps to the neighboring bins 2 and 5 when considering the pixel width h and that I_2^{h} maps to the classification value of the center point in each pixel (cf. Definition 5.63).
We consider now $\mathcal{B}_{\mathcal{C},n}$ to be a bin for which the preimage of I_2^{h} is indeed empty. For the preimage of I_2 of this bin it holds consequently that the widths in radial direction is smaller than one pixel width h. With this in mind, we can derive the following approximation

$$\begin{aligned}\left|I_2^{\mathrm{h}}\left(\boldsymbol{p},\cdot,t\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot,t\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right) &= I_2\left(\boldsymbol{p},\cdot,t\right)_{\#}\kappa\left(\mathcal{B}_{\mathcal{C},n}\right)\\ &= \kappa\left(I_2\left(\boldsymbol{p},\cdot,t\right)^{-1}\left(\mathcal{B}_{\mathcal{C},n}\right)\right) \leq \mathrm{h}\,2\,\pi\,R \leq 2\left(L+W\right)\mathrm{h}\end{aligned}$$

by using the maximal possible radius to estimate the perimeter of the circular domain of the preimage and further using its upper bound given in Equation (5.44). Since $2\left(L+W\right) < C^{\dagger}$ holds, the following approximation is also valid for this second case:

$$\left|I_2^{\mathrm{h}}\left(\boldsymbol{p},\cdot,t\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot,t\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right) \leq C^{\dagger}\,\mathrm{h}. \tag{5.46}$$

As $t \in \{t_1,\ldots,t_{n_T}\}$ was chosen arbitrarily, we can deduce from Equations (5.45) and (5.46)

$$\left|I_2^{\mathrm{h}}\left(\boldsymbol{p},\cdot\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right) = \sum_{i=1}^{n_T}\left|I_2^{\mathrm{h}}\left(\boldsymbol{p},\cdot,t_i\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot,t_i\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right) \leq n_T\,C^{\dagger}\,\mathrm{h}.$$

Since the bin $\mathcal{B}_{\mathcal{C},n}$ was chosen arbitrarily as well we can derive further that

$$\sum_{n=1}^{N_{\mathcal{C}}}\left|I_2^{\mathrm{h}}\left(\boldsymbol{p},\cdot\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right) \leq N_{\mathcal{C}}\,n_T\,C^{\dagger}\,\mathrm{h} \overset{\text{Equation (5.9)}}{\leq} 2\,n_T\,C^{\dagger}\,\frac{\mathrm{h}}{\Delta c}$$

holds. With the convergence of $\frac{\mathrm{h}}{\Delta c} \to 0$ as stated in the assumptions of this lemma, we conclude that $\sum_{n=1}^{N_{\mathcal{C}}}\left|I_2^{\mathrm{h}}\left(\boldsymbol{p},\cdot\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right) \to 0$ holds. □

Lemma 5.82 (Approximated parameter setting for classification image I_2)
We consider the setting introduced in Definition 5.76. Then it holds that

$$\sum_{n=1}^{N_{\mathcal{C}}}\left|I_2\left(\boldsymbol{p}_{\varepsilon},\cdot\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right) \to 0 \qquad \text{for } \boldsymbol{p}_{\varepsilon} \xrightarrow{\varepsilon\to 0} \boldsymbol{p} \text{ and } \frac{\left\|\boldsymbol{p}_{\varepsilon}-\boldsymbol{p}\right\|_2}{\Delta c} \to 0.$$

Proof. Let $t \in \{t_1,\ldots,t_{n_T}\}$ be arbitrary but fixed for now. We focus first again on an arbitrary bin $\mathcal{B}_{\mathcal{C},n}$ with $n \in \{1,\ldots,N_{\mathcal{C}}\}$ and derive an approximation for $\left|I_2\left(\boldsymbol{p}_{\varepsilon},\cdot,t\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot,t\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right)$. Let the bin be given by $\mathcal{B}_{\mathcal{C},n} = \left[c_1,c_2\right]$ with $c_1 < c_2$ and $c_1, c_2 \in \mathcal{C}' \subset (0,2)$. We set $\mathcal{C}^* := (0,2)$.

Let $\left(\boldsymbol{p}_{\varepsilon}\right)_{\varepsilon>0}$ be a sequence of parameter settings converging to $\boldsymbol{p} \in P$ for $\varepsilon \to 0$ in the sense that $\left\|\boldsymbol{p}_{\varepsilon}-\boldsymbol{p}\right\|_2 < \varepsilon$ holds for all $\varepsilon > 0$. We recapitulate the individual parameter in the sets to be given by

$$\begin{aligned}\boldsymbol{p}_{\varepsilon} &= \left(\boldsymbol{x}_{0,\varepsilon}, t_{0,n,\varepsilon}, t_{0,a,\varepsilon}, v_{\varepsilon}\right),\\ \boldsymbol{p} &= \left(\boldsymbol{x}_0, t_{0,n}, t_{0,a}, v\right).\end{aligned}$$

For both parameter sets we consider a radial cut through their origins $\boldsymbol{x}_{0,\varepsilon}$ and $\boldsymbol{x}_0$. To find the radius of the contour line corresponding to the preimage of I_2 for a certain classification value, we make use of the implicit function theorem (cf. Theorem 2 in §8 of [23]). With this we aim for an implicit representation of the radius depending on the parameter setting $\boldsymbol{p}_\varepsilon$ or $\boldsymbol{p}$ and on the classification values c_1 and c_2.
We define the reduced parameter space

$$\boldsymbol{P}^* = \left\{ \left(\hat{t}_{0,n}, \hat{t}_{0,a} \right) \in (0,T)^2 \,|\, \hat{t}_{0,n} < \hat{t}_{0,a} \right\} \times (0, v_{\max})$$

for the starting time points and the velocities given in the reduced parameter sets $\boldsymbol{p}^* = (t_{0,n}, t_{0,a}, v)$ and $\boldsymbol{p}_\varepsilon^* = (t_{0,n,\varepsilon}, t_{0,a,\varepsilon}, v_\varepsilon)$. It follows directly that

$$\left\| \boldsymbol{p}_\varepsilon^* - \boldsymbol{p}^* \right\|_2 \leq \left\| \boldsymbol{p}_\varepsilon - \boldsymbol{p} \right\|_2 . \tag{5.47}$$

We ensure here that $\boldsymbol{P}^*$ is open in comparison to the original parameter space as defined in Definition 4.3 and we point out that we neglect the spatial origins $\boldsymbol{x}_0$ and $\boldsymbol{x}_{0,\varepsilon}$ in the reduced settings. Additionally, we define $\mathcal{R}^* := (0, R)$ with R denoting a maximal possible radius similar to the one used for the previous Lemma 5.81 (cf. Equation (5.44)). Here, we use an upper bound for the radii to be given by the diameter of the domain Ω, i.e., $R = \operatorname{diam}(\Omega) = \sqrt{L^2 + W^2}$. Oriented on the notation of the cited theorem in the book of Forster on Analysis 2 (cf. Theorem 2 in §8 of [23]), we define

$$F : \boldsymbol{P}^* \times \boldsymbol{\mathcal{C}}^* \times \mathcal{R}^* \to \mathbb{R} \qquad F\left(\boldsymbol{p}^*, c, r\right) = I_2^*\left(\boldsymbol{p}^*, r\right) - c$$

based on adjusted definitions for the classification image I_2^* and circular fronts k_j^* with $j = n, a$ which depend on the reduced parameter sets and the radius instead of a location $\boldsymbol{x}$ and its distance to the origin $\boldsymbol{x}_0$. To be more precise, we apply

$$\begin{aligned} &k_j^* : \boldsymbol{P}^* \times \mathcal{R}^* \to \mathbb{R}, \qquad k_j^*\left(\boldsymbol{p}^*, r\right) = v\left(t - t_{0,j}\right) - r, \qquad j = n, a \\ &I_2^* : \boldsymbol{P}^* \times \mathcal{R}^* \to \boldsymbol{\mathcal{C}}^*, \qquad I_2^*\left(\boldsymbol{p}^*, r\right) = \frac{1}{1 + \exp\left(-\frac{1}{\varepsilon_0} k_n^*\left(\boldsymbol{p}^*, r\right)\right)} + \frac{1}{1 + \exp\left(-\frac{1}{\varepsilon_0} k_a^*\left(\boldsymbol{p}^*, r\right)\right)}. \end{aligned} \tag{5.48}$$

which are closely related to the original definitions given in Equation (4.4) and in Definition 4.4. While those definitions introduced in Section 4.2 were depending on spatial coordinates and the current time point and also included the spatial colony's origin, we here focus on radial dependence and assume the fixed time point $t \in \{t_1, \ldots, t_{n_T}\}$. To be more precise, it is rather a reparameterization since originally the classification values were calculated based on the distance between a location $\boldsymbol{x}$ and the origin $\boldsymbol{x}_0$ which is now replaced by the radius.
Analogously to I_2 it holds that I_2^* is continuously differentiable with the gradient given by

$$\nabla_{(\boldsymbol{p}^*, r)} I_2^*\left(\boldsymbol{p}^*, r\right) = \frac{1}{\varepsilon_0} \left(\frac{\exp\left(-\frac{1}{\varepsilon_0} k_n^*\left(\boldsymbol{p}^*, r\right)\right) \left(\nabla_{(\boldsymbol{p}^*, r)} k_n^*\left(\boldsymbol{p}^*, r\right)\right)}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_n^*\left(\boldsymbol{p}^*, r\right)\right)\right)^2} + \frac{\exp\left(-\frac{1}{\varepsilon_0} k_a^*\left(\boldsymbol{p}^*, r\right)\right) \left(\nabla_{(\boldsymbol{p}^*, r)} k_a^*\left(\boldsymbol{p}^*, r\right)\right)}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_a^*\left(\boldsymbol{p}^*, r\right)\right)\right)^2} \right).$$

with the gradient of the adjusted circular fronts given by

$$\nabla_{(\boldsymbol{p}^*,r)} k_n^* (\boldsymbol{p}^*, r) = \begin{pmatrix} \frac{\partial k_n^*(\boldsymbol{p}^*,r)}{\partial t_{0,n}} \\ \frac{\partial k_n^*(\boldsymbol{p}^*,r)}{\partial t_{0,a}} \\ \frac{\partial k_n^*(\boldsymbol{p}^*,r)}{\partial v} \\ \frac{\partial k_n^*(\boldsymbol{p}^*,r)}{\partial r} \end{pmatrix} = \begin{pmatrix} -v \\ 0 \\ t - t_{0,n} \\ -1 \end{pmatrix}, \quad \nabla_{(\boldsymbol{p}^*,r)} k_a^* (\boldsymbol{p}^*, r) = \begin{pmatrix} \frac{\partial k_a^*(\boldsymbol{p}^*,r)}{\partial t_{0,n}} \\ \frac{\partial k_a^*(\boldsymbol{p}^*,r)}{\partial t_{0,a}} \\ \frac{\partial k_a^*(\boldsymbol{p}^*,r)}{\partial v} \\ \frac{\partial k_a^*(\boldsymbol{p}^*,r)}{\partial r} \end{pmatrix} = \begin{pmatrix} 0 \\ -v \\ t - t_{0,a} \\ -1 \end{pmatrix}. \tag{5.49}$$

We refer to Equations (4.8) to (4.10) for the derivative terms of the original classification image formulations and the corresponding circular fronts. As I_2^* is continuously differentiable for the inherent parameter, we conclude that F is also continuously differentiable. Based on the above derivative terms, we proceed with the partial derivatives of the function F which are given by

$$\begin{aligned} \nabla_{(\boldsymbol{p}^*,r)} F (\boldsymbol{p}^*, c, r) &= \nabla_{(\boldsymbol{p}^*,r)} I_2^* (\boldsymbol{p}^*, r) \\ \frac{\partial F}{\partial c} (\boldsymbol{p}^*, c, r) &= -1. \end{aligned}$$

In particular, it holds that

$$\begin{aligned} \frac{\partial F (\boldsymbol{p}^*, c, r)}{\partial r} &= \frac{\partial I_2^* (\boldsymbol{p}^*, r)}{\partial r} \\ &= \frac{1}{\varepsilon_0} \left(\frac{\exp\left(-\frac{1}{\varepsilon_0} k_n^* (\boldsymbol{p}^*, r)\right) \left(\frac{\partial k_a^*}{\partial r} (\boldsymbol{p}^*, r)\right)}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_n^* (\boldsymbol{p}^*, r)\right)\right)^2} + \frac{\exp\left(-\frac{1}{\varepsilon_0} k_a^* (\boldsymbol{p}^*, r)\right) \left(\frac{\partial k_a^*}{\partial r} (\boldsymbol{p}^*, r)\right)}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_a^* (\boldsymbol{p}^*, r)\right)\right)^2} \right) \\ &= -\frac{1}{\varepsilon_0} \left(\frac{\exp\left(-\frac{1}{\varepsilon_0} k_n^* (\boldsymbol{p}^*, r)\right)}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_n^* (\boldsymbol{p}^*, r)\right)\right)^2} + \frac{\exp\left(-\frac{1}{\varepsilon_0} k_a^* (\boldsymbol{p}^*, r)\right)}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_a^* (\boldsymbol{p}^*, r)\right)\right)^2} \right) < 0 \end{aligned}$$

since the numerator for both fraction is strictly positive and the fractions cannot equal 0. Consequently, $\frac{\partial F(\boldsymbol{p}^*,c,r)}{\partial r}$ is invertible or all $(\boldsymbol{p}^*, c, r) \in \boldsymbol{P}^* \times \boldsymbol{C}^* \times \mathcal{R}^*$. We can even show that $\left|\frac{\partial F(\boldsymbol{p}^*,c,r)}{\partial r}\right|$ is bounded away from zero such that the absolute value of its inverse is actually finite and does not converge to infinity. This can be seen in the following way. Similar to Equation (5.32), we can derive that

$$\begin{aligned} \exp\left(-\frac{1}{\varepsilon_0} k_j^* (\boldsymbol{p}^*, r)\right) &= \exp\left(-\frac{1}{\varepsilon_0} (v (t - t_{0,j}) - r)\right) \\ &\geq \exp\left(-\frac{1}{\varepsilon_0} (v_{\max} (T - 0) - 0)\right) \\ &= \exp\left(-\frac{1}{\varepsilon_0} v_{\max} T\right) := C^{\diamond}, \\ \exp\left(-\frac{1}{\varepsilon_0} k_j^* (\boldsymbol{p}^*, r)\right) &= \exp\left(-\frac{1}{\varepsilon_0} (v (t - t_{0,j}) - r)\right) \\ &\leq \exp\left(-\frac{1}{\varepsilon_0} (v_{\max} (0 - T) - R)\right) \\ &= \exp\left(-\frac{1}{\varepsilon_0} (-v_{\max} T - R)\right) := C^{\clubsuit} \end{aligned} \tag{5.50}$$

for $j = n, a$ hold. Consequently, we can derive the following lower bound for the absolute value of the partial derivative with respect to r

$$\begin{aligned}\left|\frac{\partial F(\boldsymbol{p}^*, c, r)}{\partial r}\right| &\geq \frac{1}{\varepsilon_0} C^{\diamondsuit} \left(\frac{1}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_n^*(\boldsymbol{p}^*, r)\right)\right)^2} + \frac{1}{\left(1 + \exp\left(-\frac{1}{\varepsilon_0} k_a^*(\boldsymbol{p}^*, r)\right)\right)^2} \right) \\ &\geq \frac{1}{\varepsilon_0} C^{\diamondsuit} \left(\frac{1}{\left(1 + C^{\clubsuit}\right)^2} + \frac{1}{\left(1 + C^{\clubsuit}\right)^2} \right) =: C^{\dagger} > 0\end{aligned}$$

and in same way an upper bound for its reciprocal is given by

$$\left| \frac{1}{\frac{\partial F(\boldsymbol{p}^*, c, r)}{\partial r}} \right| \leq \frac{1}{C^{\dagger}}. \tag{5.51}$$

Before we dive into the application of the implicit function theorem, we finish our preparations by deriving further upper bounds for the absolute values of the following partial derivatives:

$$\frac{\partial F}{\partial t_{0,n}}, \qquad \frac{\partial F}{\partial t_{0,a}}, \qquad \frac{\partial F}{\partial v}, \qquad \frac{\partial F}{\partial c}.$$

For this purpose, we derive with the same estimates introduced in Equation (5.50) and with approximations for the absolute values of the partial derivatives for the circle equations (cf. Equation (5.49)) the following upper bounds

$$\begin{aligned}\left|\frac{\partial F(\boldsymbol{p}^*, c, r)}{\partial t_{0,n}}\right| &= \frac{\partial I_2^*(\boldsymbol{p}^*, r)}{\partial t_{0,n}} \leq \frac{1}{\varepsilon_0} \frac{v_{\max} C^{\clubsuit}}{\left(1 + C^{\diamondsuit}\right)^2}, \\ \left|\frac{\partial F(\boldsymbol{p}^*, c, r)}{\partial t_{0,a}}\right| &= \frac{\partial I_2^*(\boldsymbol{p}^*, r)}{\partial t_{0,a}} \leq \frac{1}{\varepsilon_0} \frac{v_{\max} C^{\clubsuit}}{\left(1 + C^{\diamondsuit}\right)^2}, \\ \left|\frac{\partial F(\boldsymbol{p}^*, c, r)}{\partial v}\right| &= \frac{\partial I_2^*(\boldsymbol{p}^*, r)}{\partial v} \leq \frac{1}{\varepsilon_0} \frac{2TC^{\clubsuit}}{\left(1 + C^{\diamondsuit}\right)^2}, \\ \left|\frac{\partial F(\boldsymbol{p}^*, c, r)}{\partial c}\right| &= 1.\end{aligned} \tag{5.52}$$

Let $(\tilde{\boldsymbol{p}}^*, \tilde{c}, \tilde{r}) \in \boldsymbol{P}^* \times \boldsymbol{\mathcal{C}}^* \times \mathcal{R}^*$ be given such that $F(\tilde{\boldsymbol{p}}^*, \tilde{c}, \tilde{r}) = 0$ holds, i.e., that the value of I_2^* for the radius $\tilde{r}$ and the reduced parameter set $\tilde{\boldsymbol{p}}^*$ equals the classification value $\tilde{c}$. It holds that $\frac{\partial F}{\partial r}$ is also invertible in $(\tilde{\boldsymbol{p}}^*, \tilde{c}, \tilde{r})$. By applying the implicit function theorem, it follows that there exist open sets $\boldsymbol{P}_1^* \subset \boldsymbol{P}^*$ containing $\tilde{\boldsymbol{p}}^*$, $\boldsymbol{\mathcal{C}}_1^* \subset \boldsymbol{\mathcal{C}}^*$ containing $\tilde{c}$ and $\mathcal{R}_1^* \subset \mathcal{R}^*$ containing $\tilde{r}$, respectively, and a continuously differentiable function

$$\boldsymbol{r} : \boldsymbol{P}_1^* \times \boldsymbol{\mathcal{C}}_1^* \to \mathcal{R}_1^* \qquad \text{with } \boldsymbol{r}(\tilde{\boldsymbol{p}}^*, \tilde{c}) = \tilde{r} \tag{5.53}$$

such that for all $\boldsymbol{p}^* \in \boldsymbol{P}_1^*$, $c \in \boldsymbol{\mathcal{C}}_1^*$ and $r \in \mathcal{R}_1^*$

$$F(\boldsymbol{p}^*, c, r) = 0 \quad \Leftrightarrow \quad r = \boldsymbol{r}(\boldsymbol{p}^*, c)$$

holds. Moreover, the partial derivatives of $\boldsymbol{r}$ are given by

$$\begin{aligned}
\frac{\partial \boldsymbol{r}}{\partial t_{0,n}} &= -\left(\frac{\partial F}{\partial r}\right)^{-1} \frac{\partial F}{\partial t_{0,n}}, \\
\frac{\partial \boldsymbol{r}}{\partial t_{0,a}} &= -\left(\frac{\partial F}{\partial r}\right)^{-1} \frac{\partial F}{\partial t_{0,a}}, \\
\frac{\partial \boldsymbol{r}}{\partial v} &= -\left(\frac{\partial F}{\partial r}\right)^{-1} \frac{\partial F}{\partial v}, \\
\frac{\partial \boldsymbol{r}}{\partial c} &= -\left(\frac{\partial F}{\partial r}\right)^{-1} \frac{\partial F}{\partial c}
\end{aligned} \tag{5.54}$$

(cf. remark 4 after theorem 2 in [23]). When considering the upper bounds for the absolute values of the partial derivatives as stated in Equations (5.51) and (5.52), we easily derive bounds for the absolute values of the partial derivatives of $\boldsymbol{r}$ given in Equation (5.54).
Since the partial derivatives of $\boldsymbol{r}$ are bounded, we conclude that $\boldsymbol{r}$ is Lipschitz continuous in $\boldsymbol{p}^*$ and c. We denote a corresponding Lipschitz constant with L and without calculating it explicitly, we state that naturally $0 < L < \infty$ holds.

With this we have finally all prerequisites at hand to focus on the main statement of this lemma. More precisely, we focus now on an approximation of $|I_2(\boldsymbol{p}_\varepsilon, \cdot, t)_\# \kappa - I_2(\boldsymbol{p}, \cdot, t)_\# \kappa|(\mathcal{B}_{\mathcal{C},n})$ with t and $\mathcal{B}_{\mathcal{C},n} = [c_1, c_2]$ still arbitrary but fixed as stated in the introduction of this proof. When considering the classification value for a fixed parameter setting, we know that it is decreasing for increasing radius (cf. Equation (5.48)). Similarly, it also holds that $\boldsymbol{r}$ is also decreasing for increasing classification values for a fixed parameter setting. With Equation (5.54), we easily see that

$$\frac{\partial \boldsymbol{r}}{\partial c} = -\left(\frac{\partial F}{\partial r}\right)^{-1} \frac{\partial F}{\partial c} = \left(\frac{\partial I_2^*}{\partial r}\right)^{-1}$$

holds. Consequently, both partial derivatives have the same sign and therewith the same monotonicity. For illustration effects, we refer to the sketches in Figure 5.4. With this monotonicity in mind, it holds

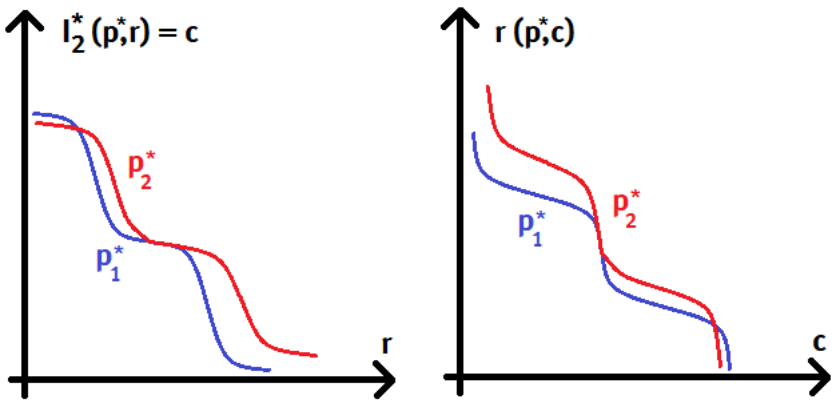


Figure 5.4: Sketches for the classification value depending on the radius r (left sketch) and, vice versa, the radius depending on the classification value c (right sketch). For two example parameter sets $\boldsymbol{p}_1^*$ (blue) and $\boldsymbol{p}_2^*$ (red), one can observe monotonously decreasing classification values for increasing radii (left plot) and monotonously decreasing radii for increasing classification values (right plot).

that

$$\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_1\right) > \boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_2\right), \qquad \boldsymbol{r}\left(\boldsymbol{p}^*, c_1\right) > \boldsymbol{r}\left(\boldsymbol{p}^*, c_2\right)$$

and, consequently, we can calculate the measure of the preimages for $\mathcal{B}_{\boldsymbol{C},n}$ by calculating the area of the corresponding rings in the image domain (cf. Figure 5.5). This leads to the following approxima-

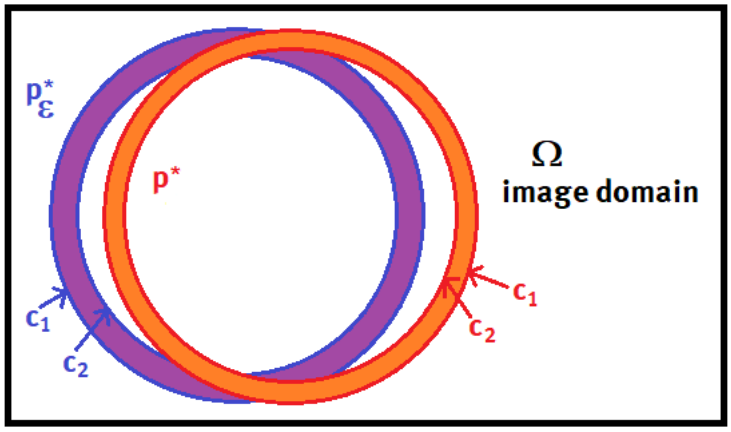


Figure 5.5: Sketches of ring-shaped preimages for two parameter sets $\boldsymbol{p}_\varepsilon^*$ (blue) and $\boldsymbol{p}^*$ (red) which are contained in our image domain Ω. The outer perimeters of the preimages of $\mathcal{B}_{\boldsymbol{C},n} = [c_1, c_2]$ under I_2^* correspond to the lower classification value c_1 and the inner ones correspond to the larger classification value c_2.

tion

$$\begin{aligned}
&\left|I_2^*\left(\boldsymbol{p}_\varepsilon^*,\cdot\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right) - I_2^*\left(\boldsymbol{p}^*,\cdot\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right)\right| \\
&= \pi\left|\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_1\right)^2 - \boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_2\right)^2 - \left(\boldsymbol{r}\left(\boldsymbol{p}^*, c_1\right)^2 - \boldsymbol{r}\left(\boldsymbol{p}^*, c_2\right)^2\right)\right| \\
&\overset{\text{triangle ineq.}}{\leq} \pi\left|\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_1\right)^2 - \boldsymbol{r}\left(\boldsymbol{p}^*, c_1\right)^2\right| + \pi\left|\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_2\right)^2 - \boldsymbol{r}\left(\boldsymbol{p}^*, c_2\right)^2\right| \\
&\overset{\substack{3^{\text{rd}}\text{ binomial}\\ \text{formula}}}{=} \pi\left|\left(\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_1\right) - \boldsymbol{r}\left(\boldsymbol{p}^*, c_1\right)\right)\underbrace{\left(\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_1\right) + \boldsymbol{r}\left(\boldsymbol{p}^*, c_1\right)\right)}_{\leq 2R}\right| \\
&\quad + \pi\left|\left(\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_2\right) - \boldsymbol{r}\left(\boldsymbol{p}^*, c_2\right)\right)\underbrace{\left(\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_2\right) + \boldsymbol{r}\left(\boldsymbol{p}^*, c_2\right)\right)}_{\leq 2R}\right| \\
&\leq 2\pi R\left(\left|\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_1\right) - \boldsymbol{r}\left(\boldsymbol{p}^*, c_1\right)\right| + \left|\boldsymbol{r}\left(\boldsymbol{p}_\varepsilon^*, c_2\right) - \boldsymbol{r}\left(\boldsymbol{p}^*, c_2\right)\right|\right) \overset{\text{L-continuity}}{\leq} 4\,\pi\,R\,L\,\left\|\boldsymbol{p}_\varepsilon^* - \boldsymbol{p}^*\right\|_2.
\end{aligned} \tag{5.55}$$

If the preimages of $\mathcal{B}_{\boldsymbol{C},n}$ under $I_2\left(\boldsymbol{p}_\varepsilon,\cdot,t\right)$ and $I_2\left(\boldsymbol{p},\cdot,t\right)$, i.e., under our original classification mappings considering $\boldsymbol{p}_\varepsilon$ and $\boldsymbol{p}$, are lying completely within our spatial domain Ω, it holds that

$$\begin{aligned}
I_2\left(\boldsymbol{p}_\varepsilon,\cdot,t\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right) &= I_2^*\left(\boldsymbol{p}_\varepsilon^*,\cdot\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right), \\
I_2\left(\boldsymbol{p},\cdot,t\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right) &= I_2^*\left(\boldsymbol{p}^*,\cdot\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right)
\end{aligned}$$

with κ denoting in each case the Lebesgue measure. In this case, the approximation

$$\begin{aligned}
\left|I_2\left(\boldsymbol{p}_\varepsilon,\cdot,t\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right) - I_2\left(\boldsymbol{p},\cdot,t\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right)\right| &= \left|I_2^*\left(\boldsymbol{p}_\varepsilon^*,\cdot\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right) - I_2^*\left(\boldsymbol{p}^*,\cdot\right)_{\#}\kappa\left(\mathcal{B}_{\boldsymbol{C},n}\right)\right| \\
&\leq 4\,\pi\,R\,L\,\left\|\boldsymbol{p}_\varepsilon^* - \boldsymbol{p}^*\right\|_2 \\
&\leq 4\,\pi\,R\,L\,\left\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\right\|_2.
\end{aligned} \tag{5.56}$$

follows directly with Equations (5.47) and (5.55).
Since $\mathcal{B}_{\mathcal{C},n}$ with $n \in \{1,\dots,N_{\mathcal{C}}\}$ and $t \in \{t_1,\dots,t_{n_T}\}$ were chosen arbitrarily, this finally leads to the following approximation:

$$\begin{aligned}\sum_{n=1}^{N_{\mathcal{C}}} \left|I_2\left(\boldsymbol{p}_\varepsilon,\cdot\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right) &= \sum_{n=1}^{N_{\mathcal{C}}}\sum_{i=1}^{n_T} \left|I_2\left(\boldsymbol{p}_\varepsilon,\cdot,t_i\right)_{\#}\kappa - I_2\left(\boldsymbol{p},\cdot,t_i\right)_{\#}\kappa\right|\left(\mathcal{B}_{\mathcal{C},n}\right)\\ &\overset{\text{Equation (5.56)}}{\leq} N_{\mathcal{C}}\, n_T\, 4\,\pi\, R\, L\, \left\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\right\|_2\\ &\overset{\text{Equation (5.9)}}{\leq} \frac{2}{\Delta c} n_T\, 4\,\pi\, R\, L\, \left\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\right\|_2\\ &= 8\, n_T\, \pi\, R\, L \frac{\left\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\right\|_2}{\Delta c}\end{aligned}$$

The convergence is a direct consequence now when considering the stated convergence of the discretization parameter to fulfill $\frac{\left\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\right\|_2}{\Delta c} \to 0$ and recalling that all other occurring parameter in the approximation, i.e., n_T, R and L, do not depend on our discretization parameter.

Having shown the convergence of the measures of the preimages for bins $\mathcal{B}_{\mathcal{C},n}$, $n \in \{1,\dots,N_{\mathcal{C}}\}$, given as complete ring-shaped domains in Ω, we infer that the convergence holds as well when only having subsets of the ring-shaped areas lying within our domain without deriving explicit approximations of the error. To see this, we assume that the preimages for an arbitrary bin $\mathcal{B}_{\mathcal{C},n}$ are only given as ring segments in Ω and apply then a domain continuation of Ω to Ω' such that the total ring-shaped preimages are lying within Ω'. For theses ring-shaped preimages on Ω', it holds again Equation (5.55). We conclude from the convergence of the measures of the preimages in Ω' for $\left\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\right\|_2 \to 0$ that also the measures of their subsets are converging when focusing again only on Ω.

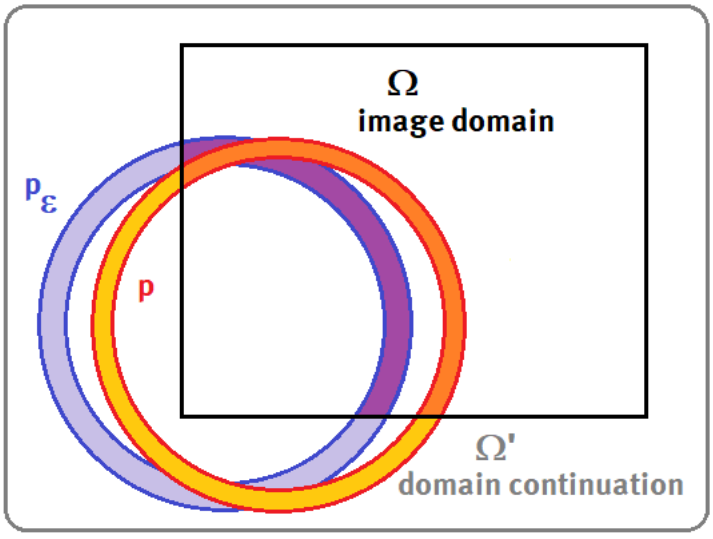


Figure 5.6: Sketches of ring-shaped preimages for two parameter sets $\boldsymbol{p}_\varepsilon$ and $\boldsymbol{p}$ which are only partially lying within the image domain Ω. The convergence of the measures for the preimages in the continuation of the domain Ω' follows for $\left\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\right\|_2 \to 0$ with Equation (5.56).

We illustrate this effect in Figure 5.6 with the preimage shown with blue boundaries for $\boldsymbol{p}_\varepsilon$ and with red boundaries for $\boldsymbol{p}$. The corresponding segments lying within the image domain Ω are marked in

dark purple and orange while the segments lying outside of Ω are shown in lighter colors (purple and yellow). With the approximation in Equation (5.55), it follows that the area of the total ring-shaped preimages converges. Moreover, we know that with $\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2 \to 0$ the convergences $p_{\varepsilon,i} \to p_i$ follows for all parameters in $\boldsymbol{p}_\varepsilon$ and $\boldsymbol{p}$. In this sense, the origin $\boldsymbol{x}_{0,\varepsilon}$ converges to $\boldsymbol{x}_0$ and also the starting time points and the velocity converge, i.e., $(t_{0,n,\varepsilon}, t_{0,a,\varepsilon}, v_\varepsilon) \to (t_{0,n}, t_{0,a}, v)$, for $\varepsilon \to 0$. This results in the blue ring converging towards the red one in the sketch, i.e., its center point converges to the one of the red one because of the convergence of $\boldsymbol{x}_{0,\varepsilon}$ and its width approaches the width of the red domain because of the convergences of $t_{0,n,\varepsilon}, t_{0,a,\varepsilon}, v_\varepsilon$. Consequently, also their areas of the segments lying *within* the image domain converge, finally. □

Lemma 5.83 (Discretized image mappings and approximated parameter settings in binned spaces) We consider the setting introduced in Definition 5.76. Then it holds that

$$\left\| \hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1} \to 0$$

for the convergences $\mathrm{h} \to 0$, $\frac{h}{\Delta c} \to 0$, $\boldsymbol{p}_\varepsilon \to \boldsymbol{p}$ and $\frac{\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2}{\Delta c} \to 0$.

Proof. We show $\left\| \hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1} \to 0$ for convergence of the discretization and approximation effects as stated in Definition 5.76. We consider now the discretization effects due to a discrete pixel grid as well as the binning widths for the piecewise constant histogram density functions. Recalling

the measures on $\mathcal{F} \times \mathcal{C}$, we remind the reader that $\lambda = P^0 \otimes \kappa$ holds (cf. Notation 5.18). With the following transformation, we split statement into intermediate substeps:

$$
\begin{aligned}
&\left\| \hat{h}^{\mathrm{h}}_{\mathcal{F}\times\mathcal{C}} - \hat{h}_{\mathcal{F}\times\mathcal{C}} \right\|_{L^1} = \int_{\mathcal{F}'\times\mathcal{C}'} \left| \hat{h}^{\mathrm{h}}_{\mathcal{F}\times\mathcal{C}}(f,c) - \hat{h}_{\mathcal{F}\times\mathcal{C}}(f,c) \right| \mathrm{d}(f,c) \\
&= \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \int_{\mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}} \frac{1}{\Delta c \Delta f^3} \left| \left[\boldsymbol{I}^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)_{\#}\lambda - \boldsymbol{I}(\boldsymbol{p}_\varepsilon,\cdot)_{\#}\lambda \right] (\mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}) \right| \mathrm{d}(f,c) \\
&= \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \left| \left[\boldsymbol{I}^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)_{\#}\lambda - \boldsymbol{I}(\boldsymbol{p}_\varepsilon,\cdot)_{\#}\lambda \right] (\mathcal{B}_{\mathcal{F},m}\times\mathcal{B}_{\mathcal{C},n}) \right| \\
&= \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \Bigg| \int_{\Omega_0\times\Omega_T} \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d,h}}\right)(\omega_0,(\boldsymbol{x},t)) \cdot \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)\right)(\boldsymbol{x},t) \\
&\qquad\qquad - \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d}}\right)(\omega_0,(\boldsymbol{x},t)) \cdot \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2(\boldsymbol{p},\cdot)\right)(\boldsymbol{x},t) \ \mathrm{d}\lambda(\omega_0,(\boldsymbol{x},t)) \Bigg| \\
&\leq \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \int_{\Omega_0\times\Omega_T} \Bigg| \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d,h}}\right)(\omega_0,(\boldsymbol{x},t)) \cdot \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)\right)(\boldsymbol{x},t) \\
&\qquad\qquad - \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d}}\right)(\omega_0,(\boldsymbol{x},t)) \cdot \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2(\boldsymbol{p},\cdot)\right)(\boldsymbol{x},t) \Bigg| \ \mathrm{d}\lambda(\omega_0,(\boldsymbol{x},t)) \\
&\overset{\text{adding 0}}{=} \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \int_{\Omega_0\times\Omega_T} \Bigg| \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d,h}}\right)(\omega_0,(\boldsymbol{x},t)) \cdot \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)\right)(\boldsymbol{x},t) \\
&\qquad\qquad - \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d}}\right)(\omega_0,(\boldsymbol{x},t)) \cdot \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)\right)(\boldsymbol{x},t) \\
&\qquad\qquad + \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d}}\right)(\omega_0,(\boldsymbol{x},t)) \cdot \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)\right)(\boldsymbol{x},t) \\
&\qquad\qquad - \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d}}\right)(\omega_0,(\boldsymbol{x},t)) \cdot \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2(\boldsymbol{p},\cdot)\right)(\boldsymbol{x},t) \Bigg| \ \mathrm{d}\lambda(\omega_0,(\boldsymbol{x},t)) \\
&\overset{\text{triangle ineq.}}{\leq} \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \int_{\Omega_0\times\Omega_T} \underbrace{\left|\left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)\right)(\boldsymbol{x},t)\right|}_{\leq 1} \\
&\qquad\qquad \cdot \left|\left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d,h}}\right)(\omega_0,(\boldsymbol{x},t)) - \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d}}\right)(\omega_0,(\boldsymbol{x},t))\right| \\
&\qquad\qquad + \underbrace{\left|\left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d}}\right)(\omega_0,(\boldsymbol{x},t))\right|}_{\leq 1} \\
&\qquad\qquad \cdot \left|\left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)\right)(\boldsymbol{x},t) - \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2(\boldsymbol{p},\cdot)\right)(\boldsymbol{x},t)\right| \ \mathrm{d}\lambda(\omega_0,(\boldsymbol{x},t)) \\
&\leq \sum_{m,n=1}^{N_{\mathcal{F}},N_{\mathcal{C}}} \int_{\Omega_0\times\Omega_T} \left|\left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d,h}}\right)(\omega_0,(\boldsymbol{x},t)) - \left(\mathbf{1}_{\mathcal{B}_{\mathcal{F},m}} \circ I_1^{\mathrm{d}}\right)(\omega_0,(\boldsymbol{x},t))\right| \ \mathrm{d}\lambda(\omega_0,(\boldsymbol{x},t)) \\
&\qquad\qquad + \underbrace{\int_{\Omega_0} \mathrm{d}P^0(\omega_0)}_{=1} \int_{\Omega_T} \left|\left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)\right)(\boldsymbol{x},t) - \left(\mathbf{1}_{\mathcal{B}_{\mathcal{C},n}} \circ I_2(\boldsymbol{p},\cdot)\right)(\boldsymbol{x},t)\right| \mathrm{d}\kappa(\boldsymbol{x},t) \\
&= \underbrace{\sum_{m=1}^{N_{\mathcal{F}}} \left| I_1^{\mathrm{d,h}}{}_{\#}\lambda - I_{1\,\#}^{\mathrm{d}}\lambda \right| (\mathcal{B}_{\mathcal{F},m})}_{\to 0, \text{ cf. Lemma 5.80}} \\
&\qquad + \underbrace{\sum_{n=1}^{N_{\mathcal{C}}} \left| I_2^{\mathrm{h}}(\boldsymbol{p}_\varepsilon,\cdot)_{\#}\kappa - I_2(\boldsymbol{p}_\varepsilon,\cdot)_{\#}\kappa \right| (\mathcal{B}_{\mathcal{C},n})}_{\to 0, \text{ cf. Lemma 5.81}} + \underbrace{\sum_{n=1}^{N_{\mathcal{C}}} \left| I_2(\boldsymbol{p}_\varepsilon,\cdot)_{\#}\kappa - I_2(\boldsymbol{p},\cdot)_{\#}\kappa \right| (\mathcal{B}_{\mathcal{C},n})}_{\to 0, \text{ cf. Lemma 5.82}}
\end{aligned}
$$

where we used the convergence for the first summand because of Lemma 5.80 for $\mathrm{h} \to 0$. The convergence of the second summand follows with Lemma 5.81 and $\mathrm{h} \to 0$ and $\frac{h}{\Delta c} \to 0$. The statement was shown for $\boldsymbol{p} \in \boldsymbol{P}$ arbitrary and, consequently, also holds for $\boldsymbol{p}_\varepsilon \in \boldsymbol{P}$. Last but not least the third summand converges with $\boldsymbol{p}_\varepsilon \to \boldsymbol{p}$ and $\frac{\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2}{\Delta c} \to 0$ as shown in Lemma 5.82. □

After having shown results for intermediate steps, we proceed with the convergence proof of our main Theorem 5.77.

Proof of Theorem 5.77. To show the convergence $\left\| \hat{h}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1} \to 0$ with the considered convergence orders of the different discretization quantities as stated in Definition 5.76, we split the statement into three main subconvergence results which in turn were already proven in the lemmas on intermediate convergence results above. With the Minkowski inequality (i.e. the triangle inequality for the L^1-norm), it holds that

$$\begin{aligned} \left\| \hat{h}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1} &= \left\| \hat{h}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - \hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} + \hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} + \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1} \\ &\leq \underbrace{\left\| \hat{h}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - \hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1}}_{(1)} + \underbrace{\left\| \hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1}}_{(3)} + \underbrace{\left\| \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1}}_{(2)}. \end{aligned} \tag{5.57}$$

We focus on the three summands individually:

(1) In Lemma 5.78, we proved that $\left\| \hat{h}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - \hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1} \to 0$ for $\varepsilon_1 \to 0$.

✓

(2) In Lemma 5.79, we focused on the convergence depending solely on the binning widths and proved that $\left\| \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1} \to 0$ for $\Delta c, \Delta f \to 0$.

✓

(3) In Lemma 5.83, we proved $\left\| \hat{h}^{\mathrm{h}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} - \hat{h}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \right\|_{L^1} \to 0$ for $\mathrm{h} \to 0$, $\frac{h}{\Delta c} \to 0$, $\boldsymbol{p}_\varepsilon \to \boldsymbol{p}$ and $\frac{\|\boldsymbol{p}_\varepsilon - \boldsymbol{p}\|_2}{\Delta c} \to 0$.

✓

Finally, this proves the convergence $\hat{h}^{\mathrm{h},\varepsilon_1}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}} \xrightarrow{L^1} h_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$. □

Remark 5.84. We point out that for convergence reasons, we do not necessarily need that the scaling parameter of the mollifier ε_1 converges slower than the parameter Δc for the binning width in the classification space. Indeed, we have shown for the first substep in the previous proof the L^1-convergence of the smoothed histogram for the mollification parameter $\varepsilon_1 \to 0$ (cf. Lemma 5.78). We stress that we still assume that the mollifier's support converges slower than the binning widths to ensure that the histogram binning does not cancel out the smoothing effect. By enforcing the stated convergence order in the Theorem 5.77, we make sure that this cannot happen which is crucial

for deriving gradient terms based on the smoothed histograms (cf. Section 5.2.4). Precisely, the requirement that Δc converges faster than the mollifier's support is a preparation for our numerical implementations where we will consider a discrete mollifier for which we need to ensure that its support is not smaller than the binning width Δc.
Moreover, we will consider for the numerical tests that the pixel width converges first, followed by the bin widths for the discrete histograms and that last but not least the mollification kernel converges. It is essential that we have far less discrete bins in the histograms compared to the number of total pixels and with the above argumentation, it is important that the support of the mollification kernel is greater than the binning width Δc.

Following a a similar line of arguments and intermediate convergence results, we can show the L^1-convergence of the histogram density functions for the separate spaces $\mathcal{F}$ and $\mathcal{C}$ stated in the next proposition. To recall the used measures, we refer to Notation 5.18.

Proposition 5.85 (L^1-convergence of histogram density functions)
Let $\hat{h}^{\mathrm{h}}_{\mathcal{F}}$ and $\hat{h}^{\mathrm{h},\varepsilon_1}_{\mathcal{C}}$ be the piecewise constant histogram density functions related to a discretized setting considering the smoothing mollification along $\mathcal{C}$, the binning of feature and classification spaces, the image mappings $I_1^{\mathrm{d,h}}$ and I_2^{h} based on a discrete pixel grid as well as the approximation of a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$. We consider the setting and convergence orders introduced in Definition 5.76 again. Then it holds that

$$\hat{h}^{\mathrm{h}}_{\mathcal{F}} \xrightarrow{L^1} h_{\mathcal{F}} \qquad \text{and} \qquad \hat{h}^{\mathrm{h},\varepsilon_1}_{\mathcal{C}} \xrightarrow{L^1} h_{\mathcal{C}}$$

with $h_{\mathcal{F}}$ and $h_{\mathcal{C}}$ being the L^1-functions describing the densities of the histogram measures with respect to the Lebesgue measure and when considering the original setting without any discretization effects, i.e., the histogram measures are given by the pushforward measures ${I_1^{\mathrm{d}}}_{\#}\lambda$ and $I_{2\#}\kappa$.

Proof. Following the same substeps as performed for the proof of Theorem 5.77 and with similar transformations, one can show the L^1-convergence for the histogram density functions for the separate spaces $\mathcal{C}$ and $\mathcal{F}$. □

With the L^1-convergence for the histogram density functions, we have paved the way for similar statements related to probability density functions. In the next section, we focus on their L^1-convergence and focus on possible pointwise convergence.

5.3.5 Convergence of probability density functions

In this section we investigate the probability density functions more thoroughly. Based on the classification and feature image data, we introduced them as normalized histogram density functions. These in turn are related to the pushforward of the measures λ and κ with respect to the corresponding image mappings I_1^{d}, I_2 and $\boldsymbol{I}$ (cf. Notation 5.18). Now, we focus on certain properties which are particularly important for the convergence proofs in Section 5.4.

To begin with, we recapitulate that the discretized probability density functions are step functions. In Lemma 5.47 we introduced the relation between the piecewise constant probability functions, namely $\hat{p}_{\mathcal{F}}$, $\hat{p}_{\mathcal{C}}$ and $\hat{p}_{\mathcal{F}\times\mathcal{C}}$, and the histogram density functions $\hat{h}_{\mathcal{F}}$, $\hat{h}_{\mathcal{C}}$ and $\hat{h}_{\mathcal{F}\times\mathcal{C}}$ related to the binning of the feature and classification spaces. While in that context, we considered the original mappings I_1^{d} for the noisy feature images and I_2 for the classification images, we can expand this to the image mappings $I_1^{\mathrm{d,h}}$ and I_2^{h} based on a discrete pixel grid. In this context, we can derive piecewise constant histogram density functions as well and receive via a normalization step similarly to Lemma 5.47 probability density functions that are also piecewise constant.
Based on the L^1-convergence of the histogram density functions shown in the previous section Section 5.3.4, we infer the L^1-convergence of the probability density functions.

Proposition 5.86 (L^1-convergence of PDFs)
Let $p_{\mathcal{F}}^{\mathrm{d},\varepsilon}$, $p_{\mathcal{C}}^{\varepsilon}$ and $p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d},\varepsilon}$ be the piecewise constant probability density functions related to a discretized setting with respect to smoothing mollification along $\mathcal{C}$, the binning of feature and classification spaces, the image mappings $I_1^{\mathrm{d,h}}$ and I_2^{h} based on a discrete pixel grid as well as the approximation of a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$. We consider the convergence orders of the discretization parameters as introduced in Definition 5.76. For the various discretization parameters, we introduce a condensed notation by representing all discretization parameters, precisely Δc, Δf, h, $\boldsymbol{p}_\varepsilon$ and ε_1, by a small ε in the superscript of the probability density functions. Then it holds that

$$p_{\mathcal{F}}^{\mathrm{d},\varepsilon} \xrightarrow{L^1} p_{\mathcal{F}}^{\mathrm{d}}, \qquad p_{\mathcal{C}}^{\varepsilon} \xrightarrow{L^1} p_{\mathcal{C}}, \qquad p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d},\varepsilon} \xrightarrow{L^1} p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}}$$

converge in L^1 for "$\varepsilon \to 0$", i.e., for fulfilling Equations (5.37) and (5.38) in Definition 5.76, and with $p_{\mathcal{F}}^{\mathrm{d}}$, $p_{\mathcal{C}}$ and $p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}}$ the L^1-functions describing the densities related to probability measures when considering the original setting with no smoothing effects along the $\mathcal{C}$-axis and neither discretizations in the image domain nor in classification and feature spaces.

Proof. The statement follows directly from the L^1-convergences of the histogram density functions shown in Theorem 5.77 and in Proposition 5.85 when considering a normalization step. By dividing by the Lebesgue measure of the spatio-temporal domain Ω_T, we can derive the probability density functions from the histogram density functions by (cf. Equation (5.4) and Lemma 5.47). □

With the L^1-convergence, we can deduce that the piecewise constant probability density functions converge almost everywhere pointwise, too. We state this in the following theorem.

Theorem 5.87 (Pointwise convergence of PDFs)
Let $\left(p_{\mathcal{F}}^{\mathrm{d},\varepsilon}\right)_{\varepsilon>0}$, $\left(p_{\mathcal{C}}^{\varepsilon}\right)_{\varepsilon>0}$ and $\left(p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d},\varepsilon}\right)_{\varepsilon>0}$ be our sequences of step functions related to a discretized setting with respect to smoothing mollification along $\mathcal{C}$, the binning of feature and classification spaces, the image mappings $I_1^{\mathrm{d,h}}$ and I_2^{h} based on a discrete pixel grid as well as the approximation of a parameter setting $\boldsymbol{p} \in \boldsymbol{P}$. We consider the convergence orders of the discretization parameters as introduced in Definition 5.76. The various discretization parameters Δc, Δf, h, $\boldsymbol{p}_\varepsilon$ and ε_1 are again represented by the small ε in the superscript of the probability density functions. Then it

holds that for each sequence of step functions there exists a converging subsequence denoted with $\tilde{\varepsilon}$ and representing subsequences for the various discretization parameters such that

$$\begin{aligned} p_{\mathcal{F}}^{\mathrm{d},\tilde{\varepsilon}} &\to p_{\mathcal{F}}^{\mathrm{d}} \quad \text{for almost all } f \in \mathcal{F} \\ p_{\mathcal{C}}^{\tilde{\varepsilon}} &\to p_{\mathcal{C}} \quad \text{for almost all } c \in \mathcal{C} \\ p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d},\tilde{\varepsilon}} &\to p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}} \quad \text{for almost all } (f,c) \in \mathcal{F}\times\mathcal{C} \end{aligned}$$

holds for "$\tilde{\varepsilon} \to 0$", i.e., for fulfilling the convergences in Equations (5.37) and (5.38) in Definition 5.76 for the according subsequences.

Proof. The statement follows directly with the L^1-convergence of the step functions proved in Proposition 5.86 and with lemma A1.11 in [2]. The cited lemma states that for a sequence converging in L^1, a subsequence exists which converges almost everywhere pointwise. □

With these statements, we can infer not only L^1-convergence but also pointwise convergence for a sequence of probability density functions after a transition to an adequate subsequence. Theorem 5.87 ensures that such subsequences exist for our discretized probability density functions when ensuring a certain convergence order of the discretized quantities.
In the following section we focus on the existence of a minimizer for our main optimization problem and on the convergence of minimizers for vanishing discretization scales. In this context, we will exploit the stated convergence results for the probability density functions. Without marking the extracted subsequences explicitly, we infer that we are already focusing on an adequate subsequence for which both the L^1 and the pointwise convergences hold.

5.4 Analysis of MI-based optimization

After introducing the concept of mutual information and deriving the optimization problem, we focus in this section on a thorough analysis of the given minimization problem. To solve the optimization numerically, we need to consider various discretization steps for which we prepared some intermediate convergence results in the previous section. In this context, it is important to ensure the convergence of minima of the discretized optimization problem to the true minimum of the corresponding problem in the continuous setting when considering vanishing discretization scales. Yet, it is essential to tackle the question of the existence of a possible minimizer for the optimization problem first.
We introduce the direct method, Γ-convergence and equi-coercivity serving as a toolkit to prove the existence and convergence statements that we focus on in this section. While these concepts are given in a general case, we transfer them to our setting and notation to facilitate a smooth transition to our specific context for maximizing the mutual information between the multi-channel feature image I_1^{d} influenced by Gaussian noise and the classification image I_2. The analytical statements proven in this section are the baseline for the numerical analysis of the optimization problem in the subsequent sections.

5.4.1 Prerequisites for existence and convergence proofs

The existence of a minimizer and also the convergence of minima are fundamental results in the calculus of variations. Based on the definitions and theorems given in Braides' book on "Γ-convergence for beginners" [10], we recapitulate the main statements we use in the progress of our analysis in this section. We adapt the statements to our notation to facilitate the later analysis of the main minimization problem given in Definition 5.17.
We start with two definitions on coercivity and lower semi-continuity which are essential for the direct method. The lower-continuity is introduced as given in Definition 1.2 in [10].

Definition 5.88 (Lower semi-continuity)
A functional $F : \boldsymbol{P} \to \mathbb{R} \cup \{-\infty, \infty\}$ is lower semi-continuous (*l.s.c*) in $\boldsymbol{p} \in \boldsymbol{P}$, if for every sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ converging to $\boldsymbol{p}$ it holds that

$$\mathrm{F}(\boldsymbol{p}) \leq \liminf_{\varepsilon \to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon).$$

If this holds for all $\boldsymbol{p} \in \boldsymbol{P}$, F is lower semi-continuous on $\boldsymbol{P}$.

The coercivity of a functional is cited from Braides' definition next (cf. defintion 1.19 in [10]).

Definition 5.89 (Coercivity condition)
A sequence of functions $(F_\varepsilon)_{\varepsilon>0}$ with $F_\varepsilon : \boldsymbol{P} \to \mathbb{R} \cup \{-\infty, \infty\}$ is called *equi-mildly coercive* if there exists a compact set $\boldsymbol{K} \subset \boldsymbol{P}$ with $\boldsymbol{K} \neq \varnothing$ such that

$$\inf_{\boldsymbol{p} \in \boldsymbol{P}} \mathrm{F}_\varepsilon(\boldsymbol{p}) = \inf_{\boldsymbol{p} \in \boldsymbol{K}} \mathrm{F}_\varepsilon(\boldsymbol{p}) \text{ for all } \varepsilon > 0.$$

Both concepts are crucial when proving existence of a minimizer for an optimization problem via the direct method. Following Braides course in the paragraph on "A maieutic approach to Γ−convergence. Direct methods" in the preface of [10], we describe the main steps of the famous approach to justify the existence of a minimum for a given functional.

Definition 5.90 (Direct method)
Given a functional $F : \boldsymbol{P} \to \mathbb{R}$ living on the metric space $\boldsymbol{P}$, it takes three main steps to derive the existence of a minimizer for the optimization problem

$$\min_{\boldsymbol{p} \in \boldsymbol{P}} \mathrm{F}(\boldsymbol{p}).$$

1. The *existence of a minimizing sequence* $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ ensures that the condition

$$\lim_{\varepsilon \to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon) = \inf_{\boldsymbol{p} \in \boldsymbol{P}} \mathrm{F}(\boldsymbol{p})$$

is met.

2. The *extraction of a converging subsequence* $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ such that $\boldsymbol{p}_\varepsilon \to \hat{\boldsymbol{p}}\,(\varepsilon \to 0)$ holds.

3. The *lower semi-continuity of the minimization functional* F (cf. Definition 5.88) results in

$$\mathrm{F}(\hat{\boldsymbol{p}}) \leq \liminf_{\varepsilon\to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon)$$

and, consequently, that the infimum is actually reached in $\hat{\boldsymbol{p}}$.

We remark that we denote the original minimization sequence and the extracted converging subsequence both with $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$. Following all three steps, we see that

$$\mathrm{F}(\hat{\boldsymbol{p}}) \leq \liminf_{\varepsilon\to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon) = \inf_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{F}(\boldsymbol{p}) \leq \mathrm{F}(\hat{\boldsymbol{p}})$$

and, consequently, equality holds. In this sense, one can show that minimizers for F exist and $\hat{\boldsymbol{p}}$ is indeed one candidate for a minimizer. We stress here that the direct method only proves the existence of minimizers and does not focus on the uniqueness of a possible minimizer.
To complement the three steps, we enumerate conditions and circumstances that facilitate the direct method following the argumentation of Braides in [10].

Remark 5.91. For the direct method certain assumptions on the problem setting contribute to perform the three main steps stated above efficiently.

- The existence of a minimizing sequence in step 1 follows directly if the given minimization functional is bounded from below such that $\mathrm{F}(\boldsymbol{p}) > -\infty$ for all $\boldsymbol{p} \in \boldsymbol{P}$ since the existence of the infimum results in the existence of a related minimization sequence.

- If the minimizing sequence lies in a compact set, the existence of a converging subsequence for step 2 follows directly with the property of compactness.

- If the functional F is coercive, the existence of a converging subsequence for step 2 is a direct consequence of the definition of coercivity, cf. Definition 5.89

In the proof of lower-semicontinuity for our minimization functional F, we apply the Lemma of Fatou in the next section. Based on the Theorem 4.6 in [54] and the lemma A1.20 in [2], we state Fatou's Lemma as a tool of choice to prove the later statements.

Theorem 5.92 (Fatou's Lemma)
Let $(\mathcal{F}\times\mathcal{C}, \mathcal{B}(\mathcal{F}\times\mathcal{C}), \kappa)$ be a measure space. For a sequence $(f_\varepsilon)_{\varepsilon>0}$ of nonnegative, measurable and integrable functions $f_\varepsilon : \mathcal{F}\times\mathcal{C} \to \mathbb{R}$, it holds

$$\int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} f_\varepsilon \,\mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} f_\varepsilon \,\mathrm{d}\kappa.$$

We state propositions which follow directly with Fatou's Lemma and which facilitate the proceeding analysis of existence and convergence of minimizers.

Proposition 5.93 (Consequences of Fatou's Lemma)
Let $(\mathcal{F} \times \mathcal{C}, \mathcal{B}(\mathcal{F} \times \mathcal{C}), \kappa)$ be a measure space and $(f_\varepsilon)_{\varepsilon>0}$ be a sequence of measurable and integrable functions $f_\varepsilon : \mathcal{F} \times \mathcal{C} \to \mathbb{R}$.

1. If $f_\varepsilon \geq C$ for all $\varepsilon > 0$, it holds
$$\int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} f_\varepsilon \, \mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} f_\varepsilon \, \mathrm{d}\kappa.$$

2. If $f_\varepsilon \geq g_\varepsilon$ for all $\varepsilon > 0$ with $(g_\varepsilon)_{\varepsilon>0}$ a sequence of L^1-functions $g_\varepsilon : \mathcal{F} \times \mathcal{C} \to \mathbb{R}$ which converges in L^1 as well as pointwise to a another L^1-function $g : \mathcal{F} \times \mathcal{C} \to \mathbb{R}$, it holds
$$\int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} f_\varepsilon \, \mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} f_\varepsilon \, \mathrm{d}\kappa.$$

3. If $f_\varepsilon \leq g_\varepsilon$ for all $\varepsilon > 0$ with $(g_\varepsilon)_{\varepsilon>0}$ a sequence of L^1-functions $g_\varepsilon : \mathcal{F} \times \mathcal{C} \to \mathbb{R}$ which converges in L^1 and pointwise to a another L^1-function $g : \mathcal{F} \times \mathcal{C} \to \mathbb{R}$, it holds
$$\int_{\mathcal{F}\times\mathcal{C}} \limsup_{\varepsilon\to 0} f_\varepsilon \, \mathrm{d}\kappa \geq \limsup_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} f_\varepsilon \, \mathrm{d}\kappa.$$

Proof. We prove the statements separately by applying Fatou's Lemma given in Theorem 5.92 on newly defined functions h_ε:

1. We define $h_\varepsilon := f_\varepsilon - C$ for all $\varepsilon > 0$. With $h_\varepsilon \geq 0$ for all $\varepsilon > 0$ and Fatou's Lemma, we get the following equivalent statements:
$$\begin{aligned}
& \int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} h_\varepsilon \, \mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} h_\varepsilon \, \mathrm{d}\kappa \\
\Leftrightarrow \quad & \int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} (f_\varepsilon - C) \, \mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} (f_\varepsilon - C) \, \mathrm{d}\kappa \\
\Leftrightarrow \quad & \int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} f_\varepsilon \, \mathrm{d}\kappa - \int_{\mathcal{F}\times\mathcal{C}} C \, \mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} f_\varepsilon \, \mathrm{d}\kappa - \int_{\mathcal{F}\times\mathcal{C}} C \, \mathrm{d}\kappa \\
\Leftrightarrow \quad & \int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} f_\varepsilon \, \mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} f_\varepsilon \, \mathrm{d}\kappa
\end{aligned}$$
This shows the first statement.

✓

2. For the second statement we define $h_\varepsilon := f_\varepsilon - g_\varepsilon$ for all $\varepsilon > 0$. With $h_\varepsilon \geq 0$ for all $\varepsilon > 0$, we get with Fatou's Lemma

$$\begin{aligned} & \int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} h_\varepsilon \, \mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} h_\varepsilon \, \mathrm{d}\kappa \\ \Leftrightarrow \quad & \int_{\mathcal{F}\times\mathcal{C}} \underbrace{\liminf_{\varepsilon\to 0} (f_\varepsilon - g_\varepsilon)}_{=:\spadesuit} \, \mathrm{d}\kappa \leq \underbrace{\liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} (f_\varepsilon - g_\varepsilon) \, \mathrm{d}\kappa}_{=:\diamondsuit}. \end{aligned}$$

For the terms denoted with $\spadesuit$ and $\diamondsuit$, we exploit the pointwise convergence of $(g_\varepsilon)_{\varepsilon>0}$ for the former one and the L^1-convergence for the latter one:

$$\spadesuit \geq \liminf_{\varepsilon\to 0} f_\varepsilon + \liminf_{\varepsilon\to 0} (-g_\varepsilon) = \liminf_{\varepsilon\to 0} f_\varepsilon - \limsup_{\varepsilon\to 0} g_\varepsilon = \liminf_{\varepsilon\to 0} f_\varepsilon - \lim_{\varepsilon\to 0} g_\varepsilon = \liminf_{\varepsilon\to 0} f_\varepsilon - g,$$

$$\begin{aligned} \diamondsuit &= \liminf_{\varepsilon\to 0} \left(\int f_\varepsilon \, \mathrm{d}\kappa + \int -g_\varepsilon \, \mathrm{d}\kappa \right) \\ &\leq \liminf_{\varepsilon\to 0} \left(\int f_\varepsilon \, \mathrm{d}\kappa \right) + \limsup_{\varepsilon\to 0} \left(\int -g_\varepsilon \, \mathrm{d}\kappa \right) \\ &= \liminf_{\varepsilon\to 0} \left(\int f_\varepsilon \, \mathrm{d}\kappa \right) - \liminf_{\varepsilon\to 0} \left(\int g_\varepsilon \, \mathrm{d}\kappa \right) \\ &= \liminf_{\varepsilon\to 0} \left(\int f_\varepsilon \, \mathrm{d}\kappa \right) - \lim_{\varepsilon\to 0} \left(\int g_\varepsilon \, \mathrm{d}\kappa \right) = \liminf_{\varepsilon\to 0} \left(\int f_\varepsilon \, \mathrm{d}\kappa \right) - \int g \end{aligned}$$

Plugging these in the initial inequality, we derive

$$\begin{aligned} & \int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} f_\varepsilon \, \mathrm{d}\kappa - \int_{\mathcal{F}\times\mathcal{C}} g \, \mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} f_\varepsilon \, \mathrm{d}\kappa - \int_{\mathcal{F}\times\mathcal{C}} g \, \mathrm{d}\kappa \\ \Leftrightarrow \quad & \int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} f_\varepsilon \, \mathrm{d}\kappa \leq \liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} f_\varepsilon \, \mathrm{d}\kappa. \end{aligned}$$

which leads finally to the claimed statement.
We point out that it is also possible to prove the first statement with the help of the second one by considering $g_\varepsilon \equiv C$, i.e. the constant function of value C.

✓

3. The third statement follows with the second one applied to $-f_\varepsilon \geq -g_\varepsilon$ for all $\varepsilon > 0$:

$$\begin{aligned} & \underbrace{\int_{\mathcal{F}\times\mathcal{C}} \liminf_{\varepsilon\to 0} -f_\varepsilon \, \mathrm{d}\kappa}_{=-\int \limsup_{\varepsilon\to 0} f_\varepsilon \, \mathrm{d}\kappa} \leq \underbrace{\liminf_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} -f_\varepsilon \, \mathrm{d}\kappa}_{=-\limsup_{\varepsilon\to 0} \int f_\varepsilon \, \mathrm{d}\kappa} \\ \Leftrightarrow \quad & \int_{\mathcal{F}\times\mathcal{C}} \limsup_{\varepsilon\to 0} f_\varepsilon \, \mathrm{d}\kappa \geq \limsup_{\varepsilon\to 0} \int_{\mathcal{F}\times\mathcal{C}} f_\varepsilon \, \mathrm{d}\kappa \end{aligned}$$

✓

□

Not only the existence of a minimizer is of interest for us. As we are dealing with a sequence of functionals approximating the original optimization problem given in Definition 5.17 due to various discretization aspects introduced in Section 5.3, we also need to prove another convergence statement. We need to show that minimizers of the approximating functionals converge to a minimizer of the original functional if the functionals converge themselves to the original optimization problem. Therefore, we need again the coercivity of the minimization functional, cf. Definition 5.89, as it is a necessary condition for the convergence of minima. For the second criterion for convergence of minima, we cite the Definition 1.5 in [10] to introduce the concept of Γ-convergence. The Γ-convergence is a well-known concept in the context of convergence for functionals.

Definition 5.94 (Γ-convergence)
A sequence of functions $(F_\varepsilon)_{\varepsilon>0}$ with $F_\varepsilon : \boldsymbol{P} \to \mathbb{R} \cup \{-\infty, \infty\}$ for all $\varepsilon > 0$ Γ-converges in $\boldsymbol{P}$ to $F : \boldsymbol{P} \to \mathbb{R} \cup \{-\infty, \infty\}$, if for all $\boldsymbol{p} \in \boldsymbol{P}$ we have

(i) (*lim inf inequality*) for every sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ with $\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon\to 0} \boldsymbol{p}$, it holds

$$\mathrm{F}(\boldsymbol{p}) \leq \liminf_{\varepsilon\to 0} \mathrm{F}_\varepsilon(\boldsymbol{p}_\varepsilon);$$

(ii) (*lim sup inequality*) there exists a sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ with $\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon\to 0} \boldsymbol{p}$ such that

$$\mathrm{F}(\boldsymbol{p}) \geq \limsup_{\varepsilon\to 0} \mathrm{F}_\varepsilon(\boldsymbol{p}_\varepsilon);$$

or alternatively

(ii)' (*existence of a recovery sequence*) there exists a sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ with $\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon\to 0} \boldsymbol{p}$ such that

$$\mathrm{F}(\boldsymbol{p}) = \lim_{\varepsilon\to 0} \mathrm{F}_\varepsilon(\boldsymbol{p}_\varepsilon).$$

F is called the Γ-limit of $(F_\varepsilon)_{\varepsilon>0}$ and also denoted with $F = \Gamma - \lim_{\varepsilon\to 0} \mathrm{F}_\varepsilon$.

We already introduced for the second condition for Γ-convergence, i.e., for the *lim sup inequality*, an alternative condition with the *recovery sequence* to prepare our proofs in the later course.
Finally, we have the ingredients at hand to introduce the theorem on the convergence of minima stated in Theorem 1.21 in [10].

Theorem 5.95 (Convergence of minima)
Let $(\boldsymbol{P}, d)$ be a metric space with a distance measure d and let $(F_\varepsilon)_{\varepsilon>0}$ be a sequence of functions with $\mathrm{F}_\varepsilon : \boldsymbol{P} \to \mathbb{R} \cup \{-\infty, \infty\}$ for all $\varepsilon > 0$. If for this sequence it holds that

(i) F_ε is mildly-coercive for every $\varepsilon > 0$,

(ii) $(F_\varepsilon)_{\varepsilon>0}$ Γ-converges to $F : \boldsymbol{P} \to \mathbb{R} \cup \{-\infty, \infty\}$,

then it holds

$$\min_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{F}(\boldsymbol{p}) = \liminf_{\varepsilon\to 0}\inf_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{F}_\varepsilon(\boldsymbol{p}).$$

Moreover, if $(\boldsymbol{p}_\varepsilon)$ is a pre-compact sequence $(\boldsymbol{p}_\varepsilon)$ such that

$$\lim_{\varepsilon\to 0} \mathrm{F}_\varepsilon(\boldsymbol{p}_\varepsilon) = \liminf_{\varepsilon\to 0}\inf_{\boldsymbol{P}} \mathrm{F}_\varepsilon(\boldsymbol{p}),$$

i.e., $(\boldsymbol{p}_\varepsilon)$ is a minimizing sequence of $(F_\varepsilon)_{\varepsilon>0}$, then $(\boldsymbol{p}_\varepsilon)$ converges (up to subsequences) to a minimizer of F.

Proof. Without going into details of the proof here, we refer the interested reader to the proof of Theorem 1.21 in [10]. □

For the sake of completeness, we recapitulate briefly a definition of a pre-compact sequence given in Definition 1.16 in [10].

Definition 5.96 (Pre-compactness)
A set $\boldsymbol{K} \subset \boldsymbol{P}$ is called *pre-compact* if its closure is compact. This means that all sequences $(\boldsymbol{x}_j)_j$ admit a converging subsequence. The limit of the subsequence may lay outside of $\boldsymbol{K}$ but on the closure of $\boldsymbol{K}$.

According to our main optimization problem introduced in Definition 5.17, we want to optimize the parameter setting $\boldsymbol{p}$ to minimize our optimization functional F or, equivalently, maximize the mutual information as stated in Definition 5.16. To incorporate the parameter setting $\boldsymbol{p}$ within the relevant terms, we remind the reader of the dependencies stated in Equation (5.3). As the classification values for I_2 depend continuously on $\boldsymbol{p}$, we will denote this dependency within the probability density functions including classification information by the notation "$c(\boldsymbol{p})$".
In the following course, we use the introduced concepts to show the existence of minimizers and the convergence of minimizers.

5.4.2 Existence of minimizers

In this section we focus on the proof of the existence of a minimizer for the given optimization problem introduced in Definition 5.17. The probabilities used in the statements depend on the feature and classification images and the classification images depend continuously on the parameter setting $\boldsymbol{p}$. Following the steps of the direct method presented in Definition 5.90, we want to prove the following statement.

Theorem 5.97 (Existence of a minimizer)
There exists at least one minimizing parameter set $\hat{\boldsymbol{p}} \in \boldsymbol{P}$ for the minimization problem

$$\min_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{F}(\boldsymbol{p}) = \min_{\boldsymbol{p}\in\boldsymbol{P}} -\mathrm{MI}(\boldsymbol{p})$$

introduced in Definition 5.17 and with the mutual information calculated with the probability density functions as stated in Proposition 5.13.

Proof. To show the existence of a minimizing parameter set $\hat{\boldsymbol{p}} \in \boldsymbol{P}$, we follow the approach of the direct method in Definition 5.90 and, accordingly, start with the existence of a minimizing sequence.

1. Existence of a minimizing sequence
As stated in Remark 5.91, we first show that the minimization functional is bounded from below. For this purpose, we choose an arbitrary parameter set $\boldsymbol{p} \in \boldsymbol{P}$ and conclude the existence of a lower bound with the following chain of inequalities:

$$\begin{aligned}
\mathrm{F}(\boldsymbol{p}) &= -\int_{\mathcal{F}'\times\mathcal{C}} \log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c(\boldsymbol{p}))}\right) \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c) \\
&\overset{\substack{\text{Jensen's inequality,}\\ \text{cf. Theorem 6.29 in [54]}}}{\geq} -\log\left(\int_{\mathcal{F}'\times\mathcal{C}} \frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c(\boldsymbol{p}))} \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f,c)\right) \\
&= -\log\left(\int_{\mathcal{F}'\times\mathcal{C}} p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))\, \frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c(\boldsymbol{p}))}\, \mathrm{d}(f,c)\right) \\
&\overset{-\log(x)\geq 1-x}{\geq} 1 - \int_{\mathcal{F}'\times\mathcal{C}} p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))\, \frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c(\boldsymbol{p}))}\, \mathrm{d}(f,c) \\
&\overset{\text{Proposition 3.13}}{\geq} 1 - \|p_N\|_\infty \int_{\mathcal{F}'\times\mathcal{C}} \frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))}{p^{\mathrm{d}}_{\mathcal{F}}(f)}\, \mathrm{d}(f,c) \\
&\overset{\text{Definition 3.14}}{>} 1 - \frac{\|p_N\|_\infty}{\delta} \int_{\mathcal{F}'\times\mathcal{C}} p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))\, \mathrm{d}(f,c) \\
&\overset{\text{Definition 5.10}}{\geq} 1 - \frac{\|p_N\|_\infty}{\delta} > -\infty.
\end{aligned}$$

In this context, we remind the reader of p_N being the probability density function related to the noise image (cf. Definition 3.4). This results in

$$\mathrm{F}(\boldsymbol{p}) \geq \inf_{\boldsymbol{p}'\in\boldsymbol{P}} \mathrm{F}(\boldsymbol{p}') > -\infty \quad \forall \boldsymbol{p} \in \boldsymbol{P}.$$

Consequently, the infimum of the optimization function exists and we set

$$M := \inf_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{F}(\boldsymbol{p})$$

with $M > -\infty$. The existence of a minimizing sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ in $\boldsymbol{P}$ with

$$M \leq \mathrm{F}(\boldsymbol{p}_\varepsilon) \leq M + \varepsilon \quad \forall \varepsilon > 0$$

follows directly. It holds that $\lim_{\varepsilon\to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon) = \inf_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{F}(\boldsymbol{p}) = M$.

✓

Still, we have not shown yet that the infimum is actually attained. Proving this is the underlying objective of the following steps.

2. Extraction of a converging subsequence
With the given model of concentric spreading wave fronts, we introduced the parameter space $\boldsymbol{P}$ as a compact set in Equation (4.3). From the compactness property of the parameter space $\boldsymbol{P}$, it follows directly that for any sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ in $\boldsymbol{P}$, we can extract a convergent subsequence. Without changing notations the convergent subsequence is now $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ with

$$\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon\to 0} \hat{\boldsymbol{p}} \in \boldsymbol{P}.$$

This subsequence is still a minimizing sequence such that

$$\begin{aligned} M \le \mathrm{F}(\boldsymbol{p}_\varepsilon) \le M + \varepsilon \quad \forall \varepsilon > 0, \\ \lim_{\varepsilon\to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon) = \inf_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{F}(\boldsymbol{p}) = M \end{aligned}$$

is still true.

✓

3. Lower semi-continuity of the minimization functional
In this step we need to prove that for a converging sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ with $\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon\to 0} \hat{\boldsymbol{p}} \in \boldsymbol{P}$, it holds that

$$\liminf_{\varepsilon\to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon) \ge \mathrm{F}(\hat{\boldsymbol{p}}).$$

We start with setting the integrand term as a function $h : \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}} \to \mathbb{R}$ and approximate it as follows

$$\begin{aligned} h_\varepsilon(f, c(\boldsymbol{p}_\varepsilon)) &:= -\log\left(\frac{p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f, c(\boldsymbol{p}_\varepsilon))}{p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f)\, p_{\boldsymbol{\mathcal{C}}}(c(\boldsymbol{p}_\varepsilon))}\right) \\ &\overset{-\log(x)\ge 1-x}{\ge} 1 - \frac{p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f, c(\boldsymbol{p}_\varepsilon))}{p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f)\, p_{\boldsymbol{\mathcal{C}}}(c(\boldsymbol{p}_\varepsilon))} \\ &\overset{\text{Proposition 3.13}}{\ge} 1 - \|p_N\|_\infty \frac{1}{p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}(f)} \\ &\overset{\text{Definition 3.14}}{>} 1 - \frac{\|p_N\|_\infty}{\delta} > -\infty. \end{aligned}$$

So there exists a constant $C \in \mathbb{R}$ such that $h_\varepsilon(f, c) \ge C$ for all $(f, c) \in \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$. We apply a version of the Lemma of Fatou, i.e, the first statement in Proposition 5.93 on h_ε and receive

$$\underbrace{\int_{\boldsymbol{\mathcal{F}}'\times\boldsymbol{\mathcal{C}}} \liminf_{\varepsilon\to 0} h_\varepsilon(f, c(\boldsymbol{p}_\varepsilon))\; \mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f, c)}_{L.S.} \le \underbrace{\liminf_{\varepsilon\to 0} \int_{\boldsymbol{\mathcal{F}}'\times\boldsymbol{\mathcal{C}}} h_\varepsilon(f, c(\boldsymbol{p}_\varepsilon))\; \mathrm{d}P^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}(f, c)}_{R.S.}.$$

For the right hand side, it holds per definition that

$$R.S. = \liminf_{\varepsilon\to 0} \int_{\mathcal{F}'\times\mathcal{C}} -\log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}_\varepsilon))}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c(\boldsymbol{p}_\varepsilon))}\right) \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c) = \liminf_{\varepsilon\to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon).$$

For the left hand side, we exploit the fact that the classification depends continuously on the given parameter setting and additionally that sthe probability density functions are pointwise convergent almost everywhere (cf. Theorem 5.87). Together with the continuity of the logarithm, this results in

$$\begin{aligned}
L.S. &= \int_{\mathcal{F}'\times\mathcal{C}} \liminf_{\varepsilon\to 0} -\log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}_\varepsilon))}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c(\boldsymbol{p}_\varepsilon))}\right) \mathrm{d}P^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c) \\
&= \int_{\mathcal{F}'\times\mathcal{C}} \liminf_{\varepsilon\to 0} -p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}_\varepsilon)) \log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}_\varepsilon))}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c(\boldsymbol{p}_\varepsilon))}\right) \mathrm{d}(f, c) \\
&= \int_{\mathcal{F}'\times\mathcal{C}} -p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\hat{\boldsymbol{p}})) \log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\hat{\boldsymbol{p}}))}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c(\hat{\boldsymbol{p}}))}\right) \mathrm{d}(f, c) \\
&= \mathrm{F}(\hat{\boldsymbol{p}}).
\end{aligned}$$

Plugging in the estimates for the left and right hand side terms, we see that $\mathrm{F}(\hat{\boldsymbol{p}}) \leq \liminf_{\varepsilon\to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon)$ holds, i.e., the optimization functional F is lower semi-continuous.

✓

In conclusion of all three steps for the direct method, we get that there exists a convergent minimizing sequence $(\boldsymbol{p}_\varepsilon)_{\varepsilon>0}$ with $\boldsymbol{p}_\varepsilon \xrightarrow{\varepsilon\to 0} \hat{\boldsymbol{p}}$ such that the following holds:

$$M \leq \mathrm{F}(\hat{\boldsymbol{p}}) \leq \liminf_{\varepsilon\to 0} \mathrm{F}(\boldsymbol{p}_\varepsilon) \leq \liminf_{\varepsilon\to 0} M + \varepsilon = \lim_{\varepsilon\to 0} M + \varepsilon = M.$$

This results in equality for $\varepsilon \to 0$ and it follows that the infimum is attained. This proves the existence of a minimizer for the given minimization functional F. □

After having shown that there exists at least one minimizer for the optimization functional, we focus in the next section on the convergence of minimizers when estimating the optimization functional with approximative functionals.

5.4.3 Convergence of minimizers

This section is devoted to a convergence result when considering a sequence of functionals approximating the main optimization functional (cf. Definition 5.17) and their minimizers. Based on Theorem 5.95, we show that this sequence of minimizers converges to a minimizer for our main problem.

Theorem 5.98 (Convergence of minimizers)
We approximate our minimization problem

$$\min_{\boldsymbol{p}\in P} \mathrm{F}(\boldsymbol{p}) = \min_{\boldsymbol{p}\in P} -\mathrm{MI}(\boldsymbol{p})$$

due to discretizations described in section Section 5.3 with

$$\min_{\boldsymbol{p}\in\boldsymbol{P}} \mathrm{F}_\varepsilon(\boldsymbol{p}) = \min_{\boldsymbol{p}\in\boldsymbol{P}} -\mathrm{MI}_\varepsilon(\boldsymbol{p}), \quad \varepsilon > 0.$$

More precisely, we include the binning widths for discrete histograms Δc in the classification space $\boldsymbol{\mathcal{C}}$ and Δf in the feature space $\boldsymbol{\mathcal{F}}$, the discretization based on discrete pixel grids for imaging data with the pixel width given as h, and eventually the parameter $\varepsilon_1 > 0$ scaling the mollification kernel to smooth the discretized histograms along the $\boldsymbol{\mathcal{C}}$-axis. We postulate that a sequence of minimizers $\hat{\boldsymbol{p}}_\varepsilon \in \boldsymbol{P}$ of the discretized optimizations functionals F_ε converges (up to subsequences) to a minimizer $\hat{\boldsymbol{p}} \in \boldsymbol{P}$ of the original optimization functional F for

$$h \to 0, \qquad \Delta c \to 0, \qquad \Delta f \to 0, \qquad \varepsilon_1 \to 0 \tag{5.58}$$

by preserving the convergence of the following relations

$$\frac{h}{\Delta c} \to 0, \qquad \frac{\|\hat{\boldsymbol{p}}_\varepsilon - \hat{\boldsymbol{p}}\|_2}{\Delta c} \to 0, \qquad \frac{\Delta c}{\varepsilon_1} \to 0, \tag{5.59}$$

i.e., we enforce that h and $\|\hat{\boldsymbol{p}}_\varepsilon - \hat{\boldsymbol{p}}\|_2$ converge faster to 0 than Δc and similarly Δc converges faster to 0 than the parameter ε_1 scaling the mollification kernel's width.

Proof. We use Theorem 5.95 to show the convergence of minima for the approximated functionals F_ε to a minimizer of the true optimization functional F for $\varepsilon \to 0$. For this, we need to prove the required properties of this theorem.

1. Coercivity of the functionals F_ε *for every* $\varepsilon > 0$
With the given model of concentric spreading wave fronts, we introduced the parameter space $\boldsymbol{P}$ as a compact set in Equation (4.3). From the compactness property of the parameter space $\boldsymbol{P}$ the coercivity follows directly (cf. Definition 5.89).

✓

2. $\Gamma-$*convergence:* $\mathrm{F}_\varepsilon \overset{\Gamma}{\to} \mathrm{F}$ for $\varepsilon \to 0$
Based on Definition 5.94, we prove the $\Gamma-$convergence of our minimization functional. To facilitate the notation in the following, we use

$$p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}, \quad p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}, \quad p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}$$

to incorporate the discretization parameter ε in the notation of the probability density functions in a discretized setting including a pixel width h, bin widths Δf and Δc for $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{C}}$, respectively, and also ε_1 as the scaling parameter of a mollification kernel for the discretized histograms. In this sense, ε represents again the different discretization quantities. This notation is in line with the L^1-convergence of the probability density functions stated in Proposition 5.86 and their pointwise convergence stated in Theorem 5.87. With this abbreviation in the labelling of the different probability

density functions, we replace the original densities $p^{\mathrm{d}}_{\mathcal{F}}$, $p_{\mathcal{C}}$ and $p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}$ in the discretized setting and in the corresponding discretized MI functional as follows

$$\mathrm{F}_{\varepsilon}(\boldsymbol{p}) = -\mathrm{MI}_{\varepsilon}(\boldsymbol{p}) = -\int_{\mathcal{F}'\times\mathcal{C}} p^{\mathrm{d},\varepsilon}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))\log\left(\frac{p^{\mathrm{d},\varepsilon}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))}{p^{\mathrm{d},\varepsilon}_{\mathcal{F}}(f)\, p^{\varepsilon}_{\mathcal{C}}(c(\boldsymbol{p}))}\right)\mathrm{d}(f,c).$$

We start with the first condition to prove Γ−convergence:
(i) lim inf inequality
Let $\boldsymbol{p}\in\boldsymbol{P}$ be an arbitrary parameter setting and $(\boldsymbol{p}_{\varepsilon})_{\varepsilon>0}\subset\boldsymbol{P}$ a sequence converging to $\boldsymbol{p}$ for $\varepsilon\to 0$ in the sense that $\|\boldsymbol{p}_{\varepsilon}-\boldsymbol{p}\|_2<\varepsilon$ for all $\varepsilon>0$.
Moreover, we state that also in the discretized setting we consider the feature images to be affected by Gaussian noise and infer

$$p^{\mathrm{d},\varepsilon}_{\mathcal{F}} > \delta$$

to be true for all $f\in\mathcal{F}'$ (cf. Definition 3.14). In this respect, we point out that we consider $p^{\mathrm{d},\varepsilon}_{\mathcal{F}}$ to be equal to 0 similarly to $p^{\mathrm{d}}_{\mathcal{F}}$ in $\mathcal{F}\backslash\mathcal{F}'$. Additionally, we use the following estimation

$$p^{\mathrm{d},\varepsilon}_{\mathcal{F}\times\mathcal{C}}(f,c)\le\|p_N\|_{\infty}p^{\varepsilon}_{\mathcal{C}}(c)$$

which holds true for all $(f,c)\in\mathcal{F}'\times\mathcal{C}$. This inequality for the probability density functions in the discretized case can be derived similarly to Proposition 3.13.
With these two estimations at hand, we can derive a lower bound for the integrand in the same way as done in the third part of the proof for Theorem 5.97 dealing with the existence of a minimizer. By applying again the Lemma of Fatou (cf. Proposition 5.93), we can derive

$$\begin{aligned}
\liminf_{\varepsilon\to 0}\ \mathrm{F}_{\varepsilon}(\boldsymbol{p}_{\varepsilon}) &= \liminf_{\varepsilon\to 0}\ -\int_{\mathcal{F}'\times\mathcal{C}} p^{\mathrm{d},\varepsilon}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}_{\varepsilon}))\log\left(\frac{p^{\mathrm{d},\varepsilon}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}_{\varepsilon}))}{p^{\mathrm{d},\varepsilon}_{\mathcal{F}}(f)\, p^{\varepsilon}_{\mathcal{C}}(c(\boldsymbol{p}_{\varepsilon}))}\right)\mathrm{d}(f,c)\\
&\overset{\text{Fatou's Lemma}}{\ge} -\int_{\mathcal{F}'\times\mathcal{C}} \liminf_{\varepsilon\to 0}\ p^{\mathrm{d},\varepsilon}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}_{\varepsilon}))\log\left(\frac{p^{\mathrm{d},\varepsilon}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}_{\varepsilon}))}{p^{\mathrm{d},\varepsilon}_{\mathcal{F}}(f)\, p^{\varepsilon}_{\mathcal{C}}(c(\boldsymbol{p}_{\varepsilon}))}\right)\mathrm{d}(f,c)\\
&\overset{\text{cf. Theorem 5.87}}{=} -\int_{\mathcal{F}'\times\mathcal{C}} p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))\log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}(f, c(\boldsymbol{p}))}{p^{\mathrm{d}}_{\mathcal{F}}(f)\, p_{\mathcal{C}}(c(\boldsymbol{p}))}\right)\mathrm{d}(f,c) = \mathrm{F}(\boldsymbol{p})
\end{aligned}$$

by respecting the required convergence orders in Equations (5.58) and (5.59) which are necessary for the application of Theorem 5.87. We refer the reader to Equation (5.3) to recapitulate the inherent dependencies of the probability density functions on the joint image mapping $\boldsymbol{I}$ and on the separate image mappings I^{d}_1, I_2, respectively. Analogously to the proof of lower semi-continuity for Theorem 5.97, we apply here again in the final steps that the classification depends continuously on the parameter setting (cf. Theorem 5.70) and exploit the pointwise convergence of probability density functions (cf. Theorem 5.87). Eventually, this shows the *lim inf condition.*

✓

(ii) lim sup inequality
We show the second condition for Γ−convergence with the help of of the following parameter sequence. Let $\boldsymbol{p}\in\boldsymbol{P}$ be arbitrary and $\boldsymbol{p}_{\varepsilon}=\boldsymbol{p}$ for all $\varepsilon>0$, i.e., a constant sequence which naturally

converges to $\boldsymbol{p}$ for $\varepsilon \to 0$.
We move the minus into the integral in our optimization functional and denote the integrand of the MI_ε–term then with $f_\varepsilon : \boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}} \to \mathbb{R}$. For f_ε we perform the following transformations

$$\begin{aligned}
f_\varepsilon\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right) &= -p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right) \log\left(\frac{p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right)}{p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right)}\right) \\
&\overset{\substack{\text{minus into}\\ \text{logarithm}}}{=} p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right) \log\left(\frac{p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right)}{p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right)}\right) \\
&\overset{\log(x)\le x-1}{\le} p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right) \left(\frac{p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right)}{p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right)} - 1\right) \\
&= p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) - \underbrace{p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}\times\boldsymbol{\mathcal{C}}}\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right)}_{\ge 0} \\
&\le p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) =: g_\varepsilon\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right).
\end{aligned}$$

For this upper bound function $g_\varepsilon : \boldsymbol{\mathcal{F}}' \times \boldsymbol{\mathcal{C}} \to \mathbb{R}_+$, we derive

$$g_\varepsilon\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right) \to g := p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right) p_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}\right)\right)$$

for $\varepsilon \to 0$ and for almost all $(f, c) \in \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$ with c depending on $\boldsymbol{p}_\varepsilon$ and $\boldsymbol{p}$, respectively, by applying the pointwise convergence almost everywhere for the probability density function (cf. Theorem 5.87). Moreover, we can conclude that $g_\varepsilon \to g$ in $L^1\left(\boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}\right)$ for $\varepsilon \to 0$ since

$$\begin{aligned}
&\int_{\boldsymbol{\mathcal{F}}'\times\boldsymbol{\mathcal{C}}} \left|g_\varepsilon\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right) - g\left(f, c\left(\boldsymbol{p}\right)\right)\right| \,\mathrm{d}\left(f, c\right) \\
&= \int_{\boldsymbol{\mathcal{F}}'\times\boldsymbol{\mathcal{C}}} \left|p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) - p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right) p_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}\right)\right)\right| \,\mathrm{d}\left(f, c\right) \\
&= \int_{\boldsymbol{\mathcal{F}}'\times\boldsymbol{\mathcal{C}}} \left|p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) - p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) + p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) - p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right) p_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}\right)\right)\right| \,\mathrm{d}\left(f, c\right) \\
&\le \int_{\boldsymbol{\mathcal{F}}'\times\boldsymbol{\mathcal{C}}} \left|p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) - p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right)\right| \,\mathrm{d}\left(f, c\right) \\
&\qquad + \int_{\boldsymbol{\mathcal{F}}'\times\boldsymbol{\mathcal{C}}} \left|p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right) p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) - p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right) p_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}\right)\right)\right| \,\mathrm{d}\left(f, c\right) \\
&= \underbrace{\int_{\boldsymbol{\mathcal{C}}} p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) \,\mathrm{d}c}_{=1} \underbrace{\int_{\boldsymbol{\mathcal{F}}} \left|p^{\mathrm{d},\varepsilon}_{\boldsymbol{\mathcal{F}}}\left(f\right) - p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right)\right| \,\mathrm{d}f}_{\to 0 \quad \text{for } \varepsilon\to 0} + \underbrace{\int_{\boldsymbol{\mathcal{F}}} p^{\mathrm{d}}_{\boldsymbol{\mathcal{F}}}\left(f\right) \,\mathrm{d}f}_{\le 1} \underbrace{\int_{\boldsymbol{\mathcal{C}}} \left|p^{\varepsilon}_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}_\varepsilon\right)\right) - p_{\boldsymbol{\mathcal{C}}}\left(c\left(\boldsymbol{p}\right)\right)\right| \,\mathrm{d}c}_{\to 0 \quad \text{for } \varepsilon\to 0} \\
&\to 0 \quad \text{for } \varepsilon \to 0
\end{aligned}$$

where we make use of the L^1–convergence of $p^{\mathrm{d},\varepsilon}_{\mathcal{F}} \to p^{\mathrm{d}}_{\mathcal{F}}$ and $p^{\varepsilon}_{\mathcal{C}} \to p_{\mathcal{C}}$ for $\varepsilon \to 0$.
Now, we can apply the Lemma of Fatou's consequence for an integrand bounded from above by an L^1–converging sequence of functions (cf. third statement in Proposition 5.93).

$$\begin{aligned}
\limsup_{\varepsilon\to 0} \ \mathrm{F}_\varepsilon\left(\boldsymbol{p}_\varepsilon\right) &= \limsup_{\varepsilon\to 0} \int_{\mathcal{F}'\times\mathcal{C}} f_\varepsilon\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right) \ \mathrm{d}\left(f, c\right) \\
&\overset{\text{cf. Proposition 5.93}}{\leq} \int_{\mathcal{F}'\times\mathcal{C}} \limsup_{\varepsilon\to 0} f_\varepsilon\left(f, c\left(\boldsymbol{p}_\varepsilon\right)\right) \ \mathrm{d}\left(f, c\right) \\
&\overset{\text{cf. Theorem 5.87}}{=} \int_{\mathcal{F}'\times\mathcal{C}} -p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\left(f, c\left(\boldsymbol{p}\right)\right) \log\left(\frac{p^{\mathrm{d}}_{\mathcal{F}\times\mathcal{C}}\left(f, c\left(\boldsymbol{p}\right)\right)}{p^{\mathrm{d}}_{\mathcal{F}}\left(f\right) p_{\mathcal{C}}\left(c\left(\boldsymbol{p}\right)\right)}\right) \mathrm{d}\left(f, c\right) = \mathrm{F}\left(\boldsymbol{p}\right)
\end{aligned}$$

where we made use of the pointwise convergence of the probability density functions almost everywhere and the continuity of the logarithm. With this the *lim sup condition* is proven.

✓

This shows with Theorem 5.95 the convergence of minima $\hat{\boldsymbol{p}}_\varepsilon \in \boldsymbol{P}$ for the discretized functionals F_ε to a minimizer $\hat{\boldsymbol{p}} \in \boldsymbol{P}$ for the original functional F for $\varepsilon \to 0$. □

We enforced a certain order of convergences for the various discretization steps to achieve an overall convergence result. This ordering was especially important to make use of the shown convergence results for the probability density functions in Section 5.3.5. In this context, we refer to Remark 5.84 for a comment on the specific convergence orders. In our numerical tests, the convergence orders are of particular importance in the further course. With the convergence statement at hand, we proceed in the next section to a numerical solution of our optimization problem.

5.5 Numerics of MI-based optimization

The main focus of this section is solving the optimization problem introduced in Definition 5.17 numerically. Therefore, it is crucial to consider the various discretization aspects focused on in Section 5.3 and resulting in a discretized optimization problem. As shown in Theorem 5.98, this discretized minimization problem Γ–converges to our initial problem.
In this section, we apply a software solution developed explicitly for this project. For our implementations, we used MATLAB R2018a. On the one hand, the software code performs pre-processing for the given microscopy data to extract the described features (cf. Section 3.3) from the microscopy data. We point out that we consider in this section the feature image not to be a random variable anymore but rather one specific realization which already contains some noise effects. On the other hand, we apply a gradient-based MATLAB solver for the numerical optimization approach. We describe this procedure in more detail in the upcoming Section 5.5.1. Before we elaborate on the numerics in particular, we introduce a toy example at the beginning of the next section. This artificial problem is the baseline for the following proof of concept.
In the second part of this section, we focus on the real data set provided by the pharmaceutical company AstraZeneca. In Section 5.5.2, we extract spreading information from two example wells' time

series with the help of the developed software solution. For a second proof of concept, we process both time series which capture growing cell populations simultaneously. With this we evaluate the power of the implemented MI-based optimization to map similar texture appearances from different wells and at varying time points to the same subcolonies by identifying regions consisting of similar and differing texture appearances.

5.5.1 Proof of Concept by means of a Toy Problem

In this section, we introduce an artificial example based on a simplified feature image. In combination with a set of classification images, we start by highlighting the effect of manipulations of the spreading parameters on the joint histogram which is a key element when calculating the MI. This the main focus of the first Section 5.5.1.1. In the following Section 5.5.1.2, we derive continuous representations of the corresponding probability density functions and compare them to discrete histograms for the given toy problem. Finally, we present histograms based on different discretization scales for the pixel widths, binning widths and mollifcation kernels. This and the solution of the minimization problem for the different discretization scales is the central theme of the last Section 5.5.1.3.

5.5.1.1 Introduction to the Toy Problem

In this section, we introduce our toy problem for which we calculate feature and classification images. We focus on the effect of different parameter disturbances in the image domain as well as on related joint histograms. These joint histograms are especially important when aiming for MI-based parameter estimations which we focus on in the later Section 5.5.1.3.

The toy problem is based on one-channel feature images generated in the same way as we model the classification images (cf. Section 4.2 and Definition 4.4 in particular). Those features are considered to be given data in the further course of this section whereas the classification images are based on model parameters that correspond to certain spreading properties. In the end, those spreading parameters are to be optimized in the numerical minimization approach.
We consider the unit square as the spatial domain Ω and the time interval $[0, 1]$. Similar to the setting for the real data given by AstraZeneca, we focus on discrete time stamps. In this case, we consider $t_1 = 0.5$ and $t_2 = 1$ to be the discrete time points. To facilitate the analysis of the toy example, we only focus on *one* developing colony front. A straightforward interpretation is, for example, the differentiation between *colony area* and *background region* when thinking about the application background when working on real microscopy data. We summarize the given information on the toy example in the next definition and directly introduce the known spreading parameters.

Definition 5.99 (Toy problem)
The toy problem consists of feature images living on the semi-discrete spatio-temporal domain $\Omega_T := [0,1]^2 \times \{t_1, t_2\}$ with $t_1 = 0.5$ and $t_2 = 1$. They are modeled via

$$I_1 : \Omega_T \to \mathcal{F}, \qquad I_1(\boldsymbol{p}, \boldsymbol{x}, t) = \frac{1}{1 + \exp\left(-\frac{1}{\varepsilon_0}\left(v(t - t_0) - \|\boldsymbol{x} - \boldsymbol{x}_0\|_2\right)\right)}$$

with the model parameter $\varepsilon_0 = 0.1$ and the fixed parameter set

$$\boldsymbol{p} = (\boldsymbol{x}_0, t_0, v) = \left(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, 0, 0.5\right).$$

The classification image is defined as

$$I_2 : \Omega_T \to \mathcal{C}, \qquad I_2(\hat{\boldsymbol{p}}, \boldsymbol{x}, t) = \frac{1}{1 + \exp\left(-\frac{1}{\varepsilon_0}\left(\hat{v}\left(t - \hat{t}_0\right) - \|\boldsymbol{x} - \hat{\boldsymbol{x}}_0\|_2\right)\right)}.$$

While we assume the same smoothing parameter $\varepsilon_0 = 0.1$ for the Heaviside mollification, we introduce a second set of spreading properties $\hat{\boldsymbol{p}} = \left(\hat{\boldsymbol{x}}_0, \hat{t}_0, \hat{v}\right)$. These parameters are to be approximated by solving

$$\underset{\hat{\boldsymbol{p}} \in P}{\operatorname{argmin}}\, \mathrm{F}_\varepsilon(\hat{\boldsymbol{p}}) = \underset{\hat{\boldsymbol{p}} \in P}{\operatorname{argmin}} - \mathrm{MI}_\varepsilon(\hat{\boldsymbol{p}}), \quad \varepsilon > 0.$$

In this context, the optimization functional describes the minimization of the negative mutual information concerning the given classification and feature images. Furthermore, the parameter $\varepsilon > 0$ represents the discretization effects which influence the pixel widths, binning widths and widths of the mollification kernel's support (cf. Theorem 5.98).

The values of the feature images as well as the classification images are getting mapped to the open interval $(0,1)$ because of the smooth approximation of the Heaviside function. We remark that we neglect the features to be affected by noise effects. Instead, we consider the above defined feature image for which we know the optimal spreading properties to be equal to $\boldsymbol{p}$. Still, in our optimization functional we include a smoothed version of the probabilities related to the classification image. We introduce this in Section 5.2.4 to derive derivative terms for the optimization functional. Instead of approximating the given original parameter $\boldsymbol{p}$ exactly, we expect the spreading properties $\hat{\boldsymbol{p}}$ to maximize the MI between the given feature image and the smoothed classification image related to $\hat{\boldsymbol{p}}$ and which corresponds to the smoothed probabilities in the classification space.

Before we delve into the analysis of our toy problem and include the mollification in the classification space, we start first by presenting the effect of various parameter combination in the image domain and on the corresponding joint histograms. We use manipulations of absolute value 0.15 for the different dimension to modify the classification parameters compared to the ground truth spreading parameters used to model the feature images (cf. Table 5.1). In Figure 5.7, the spreading phenomena for different parameter settings are shown with the first time stamp at $t = t_1 = 0.5$ in the first column and the second one at $t = t_2 = 1$ in the second column. We stress that in all subplots, the same color scaling holds such that the same color in all twelve plots represents the exact same value

in the interval $[0, 1]$. In the first row marked with "no disturbance", the image values are presented when assuming the exact parameters defined in the first row of Table 5.1 and also used to model I_1 as defined in Definition 5.99. These images serve as a kind of ground truth classification images capturing the exact same spreading phenomenon as given in the feature images for time points t_1 and t_2.
In the next two rows, spreading behaviors for varying initial spreading time points t_0 are presented. When assuming a later initial time point t_0 marking the kick off time point for colony growths, we observe that the spreading front is *behind* the colony front compared to the first row. This is reflected in smaller circles matching the color of larger circles in the first row for the ground truth spreading. In the third row, we observe the contrary behavior: For an earlier initial time point t_0 compared to the ground truth time point, we get larger circular contour lines of the same color compared to the first row.
A similar appearance is revealed when dealing with a larger spreading speed v. The circles of the same color are here larger than those in the first row. However, while we observe for the first time point at $t = t_1 = 0.5$ a similar spreading appearance as for the earlier t_0 in the third row, we do have a significant difference in the last time point at $t = t_2 = 1$. Here, we observe that the colony front indeed moves faster in comparison to the first three rows when comparing the corresponding diameter of the circular contour lines of the same classification value, i.e., the same color.
Row number five illustrates the direct contrary behavior visualizing a slower spreading speed v. We observe that the contour lines, i.e., the colony front moves slower.
In the last row, the spreading speed and the initial time point are matching the ground truth values whereas the colony's origin is slightly shifted in both spatial dimensions. Comparing first and last row, we highlight that the circular contour lines of the same color are of the same size, i.e., they have identical radii. However, they are shifted slightly towards the lower right corner in the last row. For all the previous parameter settings the origin was set to the mid point of our domain.

In Figure 5.8, one dimensional cuts in the image domain are used to plot the feature or classification values along this cutting line. While we only recapitulate the reference images for the first time point for the different parameter settings in the first two rows, we plot the image values along

classification parameters $\hat{\boldsymbol{p}}$			interpretation
$\hat{\boldsymbol{x}}_0$	$\hat{t}_0$	$\hat{v}$	
$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	0	0.5	no disturbance, equal to $\hat{\boldsymbol{p}}$
$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	0.15	0.5	later t_0
$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	−0.15	0.5	earlier t_0
$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	0	0.65	larger v
$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	0	0.45	smaller v
$\begin{pmatrix} 0.65 \\ 0.65 \end{pmatrix}$	0	0.5	shifted $\boldsymbol{x}_0$

Table 5.1: Used property settings for different parameter disturbances of magnitude 0.15 per dimension compared to the ground truth parameters applied to generate the feature images.

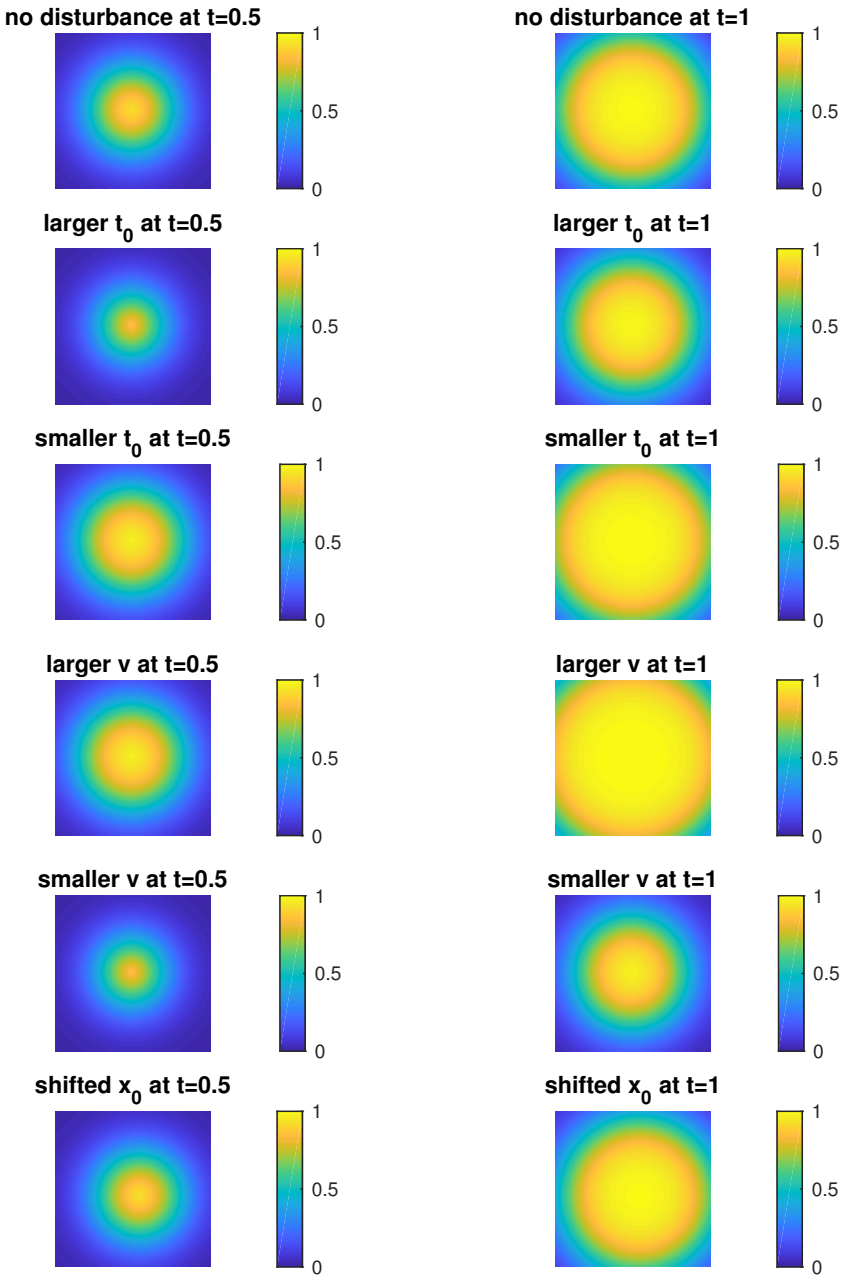


Figure 5.7: Spreading observations in the image domain at two time points for different spreading properties.

the one dimensional cuts for both time points in the last two rows of subplots in that figure. The one dimensional cut is a horizontal line in the middle of the domain and is marked for the example images in the first two rows. The colors of the cutting lines for the different parameter settings match the colors of the corresponding cross-section plots in the lower part.
One effect that becomes more prominent with these illustrations is that for our given parameter disturbances the classification images are identical for the first time point when comparing the effect of a *later* t_0 with a *smaller* v and an *earlier* t_0 with a *larger* v. We observe that the values on the cutting edge match which is depicted in the matching red and pink curves or the green and cyan curves, respectively.

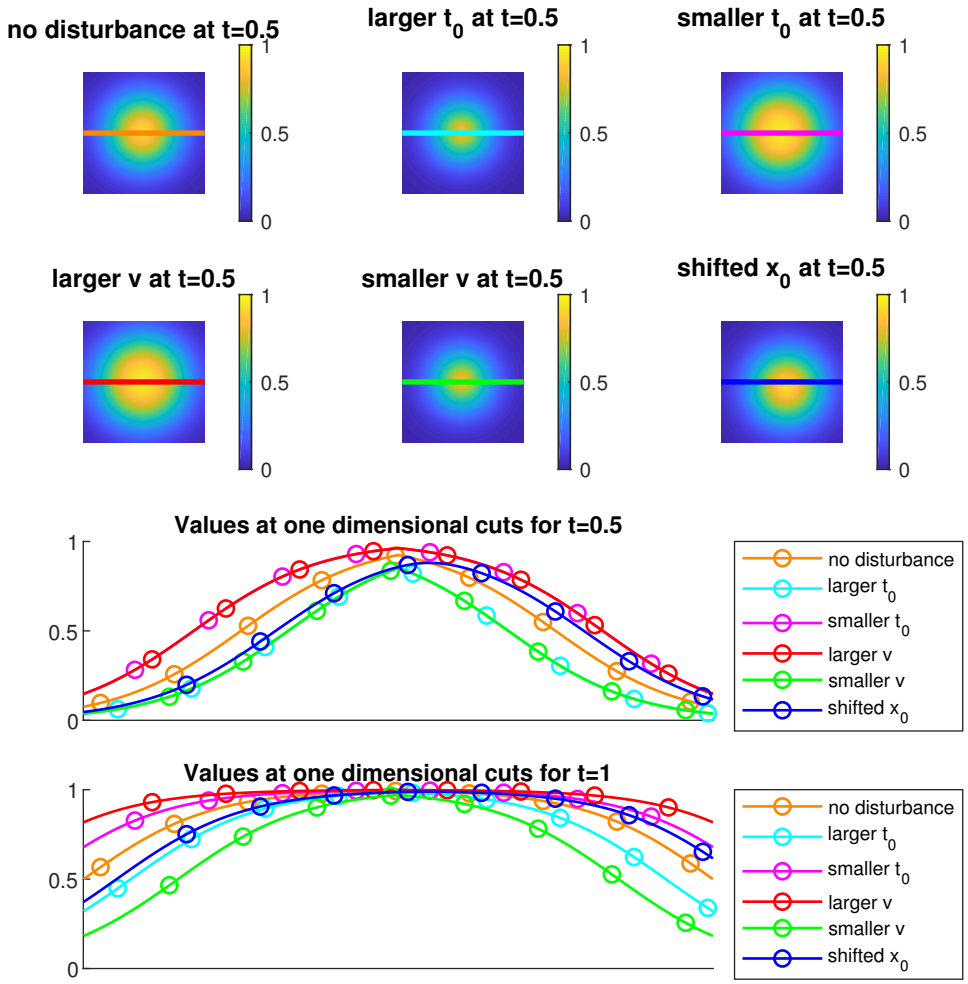


Figure 5.8: Spreading observations in the image domain at the first time point for different spreading properties with one dimensional cuts through the image domain indicating where we extract the one dimensional front. The one dimensional fronts are plotted for both time points.

By including both time stamps again, the development of the classification values along the one dimensional cuts highlight that the front lines for a larger spreading speed v (red) and an earlier initial time point t_0 (pink) correspond to moving fronts *in front of* the baseline in orange for the ground truth parameter settings. This is in line with the interpretations for the previous Figure 5.7. Again in Figure 5.8, we observe that for a smaller spreading speed v (green) and a later t_0 (cyan) the values are lower than for the orange ground truth data resulting in fronts moving *behind* the base front related to the undisturbed parameter settings.
The shifted origin $\boldsymbol{x}_0$ (blue) results in a shifted center position for the front lines. This effect is stressed by the fact that the maximum value for the blue one dimensional curve is shifted compared to the maximum values of all other curves for which the origin matches the ground truth origin.

After having shown the influence of the spreading parameters in the image domain, we move on to histogram visualizations. We investigate the effect of the parameter disturbances on the joint histogram and on its support in particular. In Figure 5.9, we show the joint histograms in the joint feature-classification spaces. The feature image I_1 is defined in Definition 5.99 and corresponds to

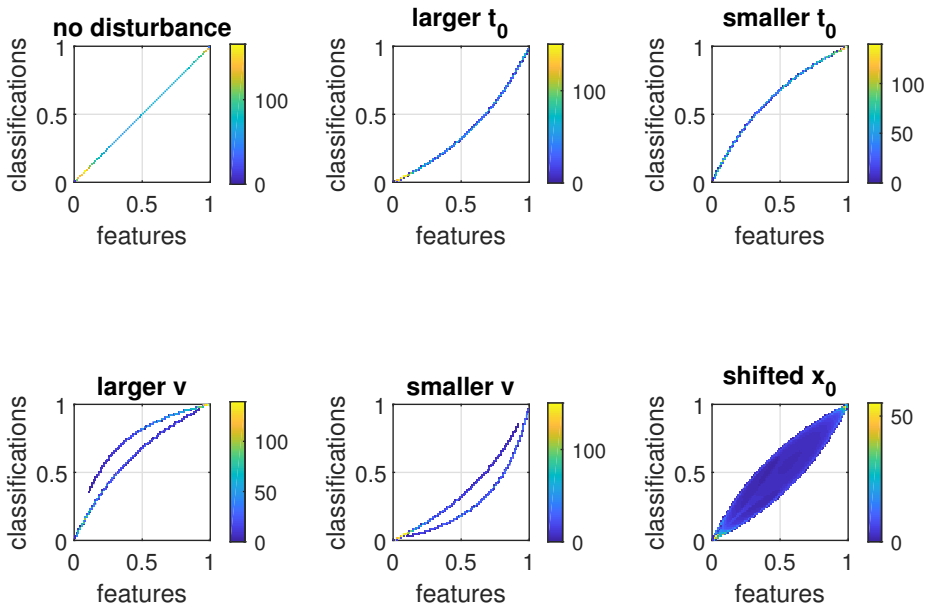


Figure 5.9: Joint histograms for variations of spreading parameters for the classification image. The different parameter disturbances affect the support of the joint histograms substantially.

the spreading behavior shown in the first row of Figure 5.7. For the classification images we consider the applied parameter disturbances used in the same figure. We stress that we also include the classification image to be matching *exactly* the feature image. This results in a histogram which is concentrated along the main diagonal in the joint space, cf. the first plot in Figure 5.9.
Based on variations in the initial time point t_0, the next two histograms in the first row of this figure are generated. Here, we observe the curvature of the histogram's support which reflects the slight mis-match of feature-classification combinations. For a later t_0 used to model the spreading initiated a bit later in the classification image compared to the spreading in the feature image, the overlaying contour lines of classification image and feature images reveal that the feature values are slightly larger than the classification values. This is reflected in the curve of the second subplot bended slightly towards the lower right corner. Since in both images a jump function between 0 and 1 is approximated, the support of the histogram converges again towards $(0,0)$ and $(1,1)$ in the joint space. A bending of the support towards the upper left corner as shown in the third subplot is revealed for the spreading phenomena related to an earlier t_0.
A similar *bending effect* is present for parameter disturbances related to variations of the spreading speed v. Additionally to this, we observe that we do not necessarily get an exact matching of one classification value with a feature value. Depending on the magnitude of the difference between the spreading speeds, we get some values which cannot be mapped uniquely. This appearance is present because contour lines of a specific value in the classification images can be related to contour lines of varying feature values in the various time frames. This is due to the differing spreading speed and prevents the histogram support to be related to a curve which could be parametrized by an injective mapping from the classification to the feature space or vice versa.
In the last subplot, we observe that the histogram is not concentrated along a curve in the joint space anymore but rather maps to whole "domain". This happens because of the shifted origin $\boldsymbol{x}_0$ resulting

in a classification value being mapped to a sequence or interval of feature values and vice versa. This effect occurs because contour lines for the classification and feature values do not match in their location in the image domain any more. We still observe a higher accumulation towards the limiting values in the corners $(0,0)$ and $(1,1)$.

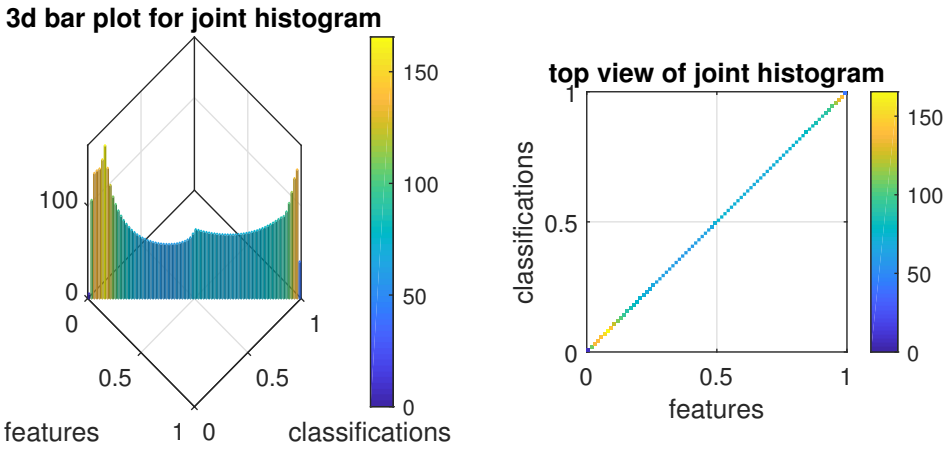


Figure 5.10: Joint histogram for "optimal" spreading parameters for the classification images.

In all of the previously described histograms, the probability of certain feature-classification combinations was resembled in the *height* of the related bars. As we presented all those histograms in a top view perspective, the actual height was encoded in the color scale, i.e., according to the colorbars next to the histograms one can derive the probability of certain combinations. To illustrate this, we show in Figure 5.10 the histogram again for the "optimal" spreading parameter settings and compare the three dimensional bar plot with the tile plot from a top view perspective. Here the color and, respectively, the values match in both plots and the distribution from higher accumulations near the corner $(0,0)$ (yellow) over to smaller values (greenish, bluish) in the middle part of the diagonal to higher accumulations again (orange, yellow) near the corner $(1,1)$ is visualized. As for the optimal parameter settings the joint histogram is concentrated around the main diagonal in the joint feature-classification space, the thin bars in the three dimensional plot are also located on this diagonal.
There are two more effects that we want to point out. First of all, we observe that the maximum value is attained not exactly at the corner $(0,0)$ but near the corner and the accumulation drops again towards the corner. This particular effect results from the design of our toy problem. We concentrate on the unit square and a circular spreading phenomenon with its origin in the center of our domain. This leads to highest accumulations for contour lines of maximal radius that are still lying *within* our domain. As contour lines with larger radius, i.e., with radius larger than 0.5, start moving out of our domain, the mass of the preimage for the related feature or classification values drops again. Consequently, we observe decreasing probabilities for those small feature and classification values which are only appearing in the corner regions.
The second effect is the particular kink near the mid point $(0.5, 0.5)$. This is due to our discontinuous representation of the temporal axis for our domain. Actually, we only focus on *two* discrete time points $t_1 = 0.5$ and $t_2 = 1$. For both of them, we archive continuous probability distributions as focused on next in Figure 5.11 depicting probability functions for the feature images. If we had included an

infinite number of time frames to discretize the temporal axis, i.e. considered a continuous representation of the temporal axis, we would have received a continuous probability distribution.

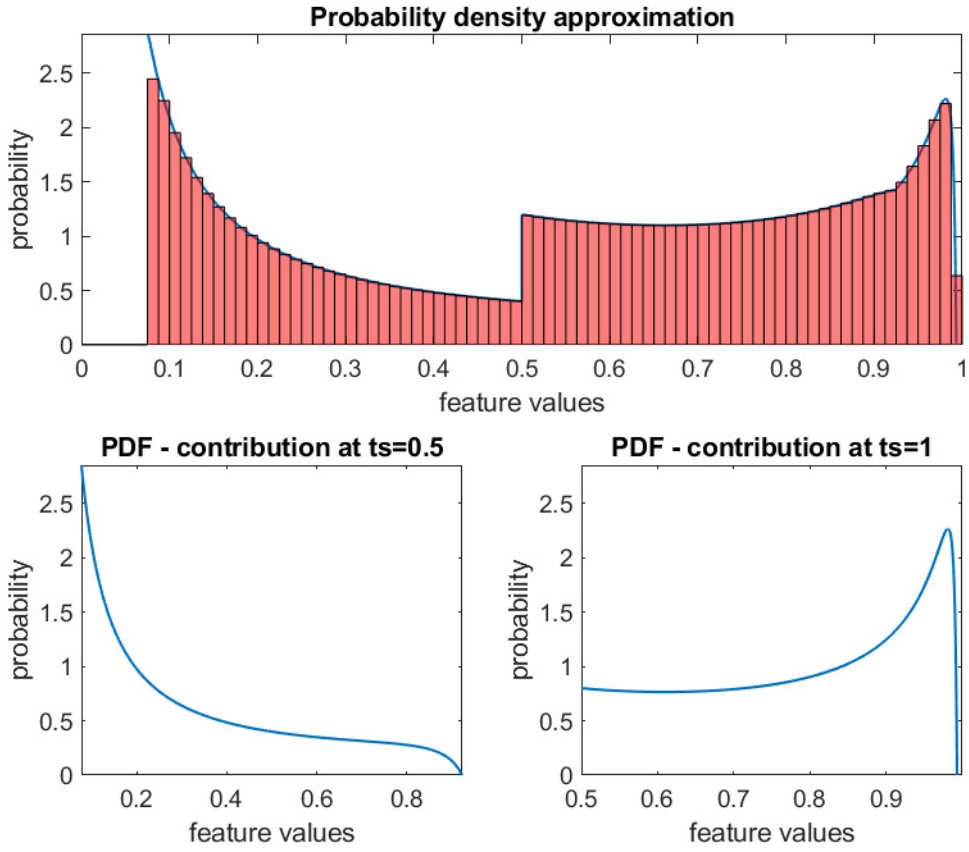


Figure 5.11: Continuous probability density function ($p_{\mathcal{F}}$) in the feature space compared to the discrete histogram and separated for two finite time stamps ($ts = t_1 = 0.5$ and $ts = t_2 = 1$).

With Figure 5.11, we focus on a continuous representation of the probability density function $p_{\mathcal{F}}$ compared to a discrete histogram. Furthermore, we present the individual probability density functions when focusing only on *one* time point in the second row of subplots. We delve into the exact derivation of the continuous representation for the probability density functions in the further course. Here, we use it to highlight that a discrete histogram approximates the smooth function shown in blue in the first subplot of Figure 5.11. In the second row, we plot the individual contributions of each time point to the general probability density function based on two time frames shown in the first row's plot. The aforementioned kink prominent in the probability density function including both time stamps is due to the fact that for the second time frame at $t = 1$ the feature values' range is approximately $[0.5, 1)$ while for the first one it is approximately $(0, 1)$. However, not only the varying range results in the kink structure. It is also influenced significantly by the steep increase for the smallest feature values per time point and especially for the second time point's values near 0.5.

In Figure 5.12, we stress the effect of parameter disturbances on the locations of the occurring kink in

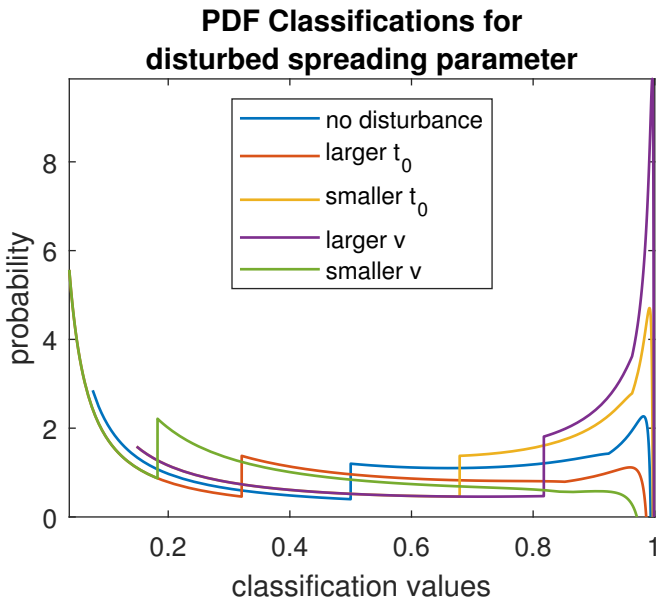


Figure 5.12: Effect of varying parameter settings on the continuous probability density function in the classification space $p_{\mathcal{C}}$.

the probability density function and also on accumulations towards the boundaries. As for a smaller spreading speed v or a later initial time point t_0, we observe smaller classification values in total when considering both time points. Particularly, the occurring values in the second time frame at $t = t_2 = 1$ are significantly smaller towards the boundary regions (cf. Figure 5.7). As larger circular contour lines result in higher probabilities, we observe that the kink structure is moving towards smaller values, i.e., towards the left on the classification axis. The contrary effect is observable for a larger spreading speed v or an earlier initial time point t_0. This results in a shifted kink towards the right on the classification axis and also in significantly higher accumulations for large classification values.

In the next section, we derive exact expressions for the probability density functions in the classification or feature space. We used those functions for the previous comparisons and illustration related to continuous probability density functions, i.e., in Figures 5.11 and 5.12. Next, we also derive the joint probability density function when considering the optimal, and consequently identical, parameter setting for feature and classification images.

5.5.1.2 Derivation of continuous probability density functions

Based on the previous section and the therein introduced toy problem (cf. Definition 5.99), we focus now on continuous representations for the probability density functions related to an adjusted version of our toy problem. For this purpose, we derive explicit formulations of the probability density functions in the separate spaces $\mathcal{C}$ and $\mathcal{F}$ as well as in the joint space $\mathcal{F} \times \mathcal{C}$.

We start with some simplifications for our toy problem and focus only on *one* parameter that might

be disturbed. To be more precise, instead of the unit square in $\mathbb{R}^2$ we use a circular domain with radius 0.5 and the origin at $(0.5, 0.5)$ as the spatial domain Ω. The center point matches exactly the origin $\boldsymbol{x}_0$ of the spreading phenomenon described with I_1. Moreover, we assume that the classification image is already modeled with correct $\boldsymbol{x}_0$ and v. So, we expect that only the initial time point differs compared to the t_0 in the parameter set used to model the captured feature images. We call this possibly disturbed initial time point for the classification images $\hat{t}_0$ again. Furthermore, we reduce the complexity further by considering only *one* temporal time point, e.g., $t = t_1 = 0.5$. This is summarized in the following definition for the adjusted toy problem.

Definition 5.100 (Adjusted toy problem)
We consider the toy problem introduced in Definition 5.99 adjusted such that it is living on a spherical domain $\Omega = B(\boldsymbol{x}_0, r_{\max})$ in $\mathbb{R}^2$ with the maximal radius $r_{\max} := 0.5$, evaluated at one temporal time point, e.g., $t = t_1 = 0.5$, and based on the parameter settings

$$\boldsymbol{p} = (\boldsymbol{x}_0, t_0, v) = \left(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, 0, 0.5\right)$$

for the feature image and

$$\hat{\boldsymbol{p}} = \left(\hat{\boldsymbol{x}}_0, \hat{t}_0, \hat{v}\right) = \left(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \hat{t}_0, 0.5\right)$$

for the classification image where already $\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0$ and $\hat{v} = v$ hold.

Before we focus on the related probability density functions, we first derive functions describing the radius associated with certain feature or classification values and their derivatives.

Proposition 5.101 (Radius functions in the classifaction and feature spaces)
Based on Definition 5.100, we can derive the radius functions

$$\begin{aligned} r_{\mathcal{F}} &: [f_{\min}, f_{\max}] \to [0, r_{\max}], & r_{\mathcal{F}}(f) &= \varepsilon_0 \log\left(\frac{1}{c} - 1\right) + v(t - t_0) \\ r_{\mathcal{C}} &: [c_{\min}, c_{\max}] \to [0, r_{\max}], & r_{\mathcal{C}}(c) &= \varepsilon_0 \log\left(\frac{1}{c} - 1\right) + v\left(t - \hat{t}_0\right) \end{aligned}$$

that determine for every occurring feature or classification value the radius of the corresponding contour circle on which the value is attained in the feature or, respectively, classification image. The limits $c_{\min}$, $c_{\max}$, $f_{\min}$ and $f_{\max}$ can be derived from the designed toy problem as well.

Proof. We use the functions introduced in Definition 5.99 for the feature and classification image

$$\begin{aligned} I_1(\boldsymbol{p}, \boldsymbol{x}, t) &= \frac{1}{1 + \exp\left(-\frac{1}{\varepsilon_0}\left(v(t - t_0) - \|\boldsymbol{x} - \boldsymbol{x}_0\|_2\right)\right)} \\ I_2(\hat{\boldsymbol{p}}, \boldsymbol{x}, t) &= \frac{1}{1 + \exp\left(-\frac{1}{\varepsilon_0}\left(\hat{v}\left(t - \hat{t}_0\right) - \|\boldsymbol{x} - \hat{\boldsymbol{x}}_0\|_2\right)\right)}. \end{aligned}$$

We define the radius to be $r := \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$ and re-define the feature and classification images depending on the radius instead of a concrete position $\boldsymbol{x}$ by asserting that the calculated image value for the

specific radius is attained on the contour line matching this radius or respectively in all $\boldsymbol{x} \in \Omega$ for which holds $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 = r$:

$$I_1(r,t) := \frac{1}{1+\exp\left(-\frac{1}{\varepsilon_0}\left(v(t-t_0)-r\right)\right)},$$
$$I_2(r,t) := \frac{1}{1+\exp\left(-\frac{1}{\varepsilon_0}\left(\hat{v}\left(t-\hat{t}_0\right)-r\right)\right)}.$$

Since t is a fixed time point in the reduced toy problem, we can simplify this further by introducing the function $\boldsymbol{f}$ and $\boldsymbol{c}$

$$\boldsymbol{f} : [0, r_{\max}] \to \boldsymbol{\mathcal{F}}, \qquad \boldsymbol{f}(r) := I_1(r,t),$$
$$\boldsymbol{c} : [0, r_{\max}] \to \boldsymbol{\mathcal{C}}, \qquad \boldsymbol{c}(r) := I_2(r,t)$$

and derive firstly the interval limits by exploiting that the minimal values are attained at the domain boundary and the maximal values are attained at the domain center, or rather the spreading phenomena's origins:

$$f_{\min} := \boldsymbol{f}(r_{\max}) = I_1(r_{\max}, t)$$
$$f_{\max} := \boldsymbol{f}(0) = I_1(0,t)$$
$$c_{\min} := \boldsymbol{c}(r_{\max}) = I_2(r_{\max}, t)$$
$$c_{\max} := \boldsymbol{c}(0) = I_2(0,t).$$

Now, we use the functions $\boldsymbol{f}$ and $\boldsymbol{c}$ to derive the radii functions. We set $\boldsymbol{f}(r) = f$ and perform the following equivalence transformations.

$$f = \frac{1}{1+\exp\left(-\frac{1}{\varepsilon_0}\left(v(t-t_0)-r\right)\right)}$$
$$\Leftrightarrow \quad 1/f - 1 = \exp\left(-\frac{1}{\varepsilon_0}\left(v(t-t_0)-r\right)\right)$$
$$\Leftrightarrow \quad r = \varepsilon_0 \log(1/f - 1) + v(t-t_0) =: \boldsymbol{r}_{\boldsymbol{\mathcal{F}}}(f).$$

Equivalently, we derive

$$\boldsymbol{r}_{\boldsymbol{\mathcal{C}}}(c) := \varepsilon_0 \log(1/c - 1) + v\left(t - \hat{t}_0\right)$$

which proves the statement. □

Next, we derive derivative forms for the newly introduced radius functions. They are of importance later on when focusing on the probability density functions.

Lemma 5.102 (Derivatives of the radius functions $r_{\mathcal{C}}$ and $r_{\mathcal{F}}$)
For the radius functions $r_{\mathcal{F}}$ and $r_{\mathcal{C}}$ from the previous Proposition 5.101, the derivatives are given as

$$r'_{\mathcal{F}}(f) = \frac{\varepsilon_0}{f^2 - f}, \qquad r'_{\mathcal{C}}(c) = \frac{\varepsilon_0}{c^2 - c}.$$

Proof. Based on the given definitions of the radius functions $r_{\mathcal{F}}$ and $r_{\mathcal{C}}$, we derive

$$r'_{\mathcal{F}}(f) = \frac{1}{\frac{1}{f} - 1}(-1)\frac{1}{f^2} = \frac{1}{f^2 - f}$$

and, equivalently,

$$r'_{\mathcal{C}}(c) = \frac{1}{c^2 - c}.$$

□

With this at hand, we prove the following statement on probability density functions in the classification space and in the feature space.

Proposition 5.103 (Probability density functions $p_{\mathcal{F}}$ and $p_{\mathcal{C}}$)
For the probability measures $P_{\mathcal{F}}$ and $P_{\mathcal{C}}$ related to the given feature and classification images defined in Definitions 5.99 and 5.100, it holds that they are absolutely continuous with respect to the Lebesgue measures on the corresponding space $\mathcal{F}$ or, respectively, $\mathcal{C}$ and there exist probability density functions such that

$$P_{\mathcal{F}}(F) = \int_F p_{\mathcal{F}}(f)\ \mathrm{d}f \qquad \forall F \subset \mathcal{F},$$
$$P_{\mathcal{C}}(C) = \int_C p_{\mathcal{F}}(c)\ \mathrm{d}c \qquad \forall C \subset \mathcal{C}$$

with the probability density functions given by

$$p_{\mathcal{F}}(f) = -8 r_{\mathcal{F}}(f)\, r'_{\mathcal{C}}(f),$$
$$p_{\mathcal{C}}(c) = -8 r_{\mathcal{C}}(c)\, r'_{\mathcal{C}}(c).$$

Proof. We recall that the Lebesgue measures for the measurable spaces $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$ and $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$ is denoted by κ as introduced in Notation 5.18. We assume that probability density functions $p_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}$ and $p_{\mathcal{C}} : \mathcal{C} \to \mathbb{R}$ exist such that

$$P_{\mathcal{F}}(F) = \int_F p_{\mathcal{F}}(f)\ \mathrm{d}\kappa(f)$$
$$P_{\mathcal{C}}(C) = \int_C p_{\mathcal{F}}(c)\ \mathrm{d}\kappa(c)$$

for arbitrary $F \subset \mathcal{F}$ and $C \subset \mathcal{C}$ hold. We identify the Lebesgue measures' notations with $\mathrm{d}\kappa(f) = \mathrm{d}f$ and $\mathrm{d}\kappa(c) = \mathrm{d}c$ to match the used notation in the proposition.

To derive the exact terms for the probability density functions, we focus on small example subset intervals $F = [f_1, f_2]$ and $C = [c_1, c_2]$ and state that

$$\begin{aligned} p_{\mathcal{F}}(f_2) &= \frac{\mathrm{d}}{\mathrm{d}f_2} P_{\mathcal{F}}([f_1, f_2]) \\ p_{\mathcal{C}}(c_2) &= \frac{\mathrm{d}}{\mathrm{d}c_2} P_{\mathcal{C}}([c_1, c_2]) \end{aligned} \tag{5.60}$$

hold. Without loss of generality, we infer that $f_1 < f_2$ and $c_1 < c_2$ hold.
For the given toy problem modeled with the above defined radius function, we denote that $\boldsymbol{r}_{\mathcal{C}}$ and $\boldsymbol{r}_{\mathcal{F}}$ are strictly monotone decreasing for increasing values c or f, respectively. Equivalently, it holds that $\boldsymbol{c}$ and $\boldsymbol{f}$ are strictly monotone decreasing for increasing radius values r. With this in mind, we determine the probability measures of the subsets to be given by

$$\begin{aligned} P_{\mathcal{C}}([c_1, c_2]) &= \frac{1}{|\Omega|}\pi\left(\boldsymbol{r}_{\mathcal{C}}(c_1)^2 - \boldsymbol{r}_{\mathcal{C}}(c_2)^2\right) \\ P_{\mathcal{F}}([f_1, f_2]) &= \frac{1}{|\Omega|}\pi\left(\boldsymbol{r}_{\mathcal{F}}(f_1)^2 - \boldsymbol{r}_{\mathcal{F}}(f_2)^2\right) \end{aligned}$$

as the probability measure equals the "ring area" defined by the two corresponding radii for the given values c_1, c_2 and f_1, f_2 compared to the area of the total domain. For the spherical domain Ω with radius $r_{\max} = 0.5$, it holds that

$$|\Omega| = \pi r_{\max}^2 = \frac{\pi}{4}.$$

Plugging these information into Equation (5.60), we receive

$$p_{\mathcal{C}}(c_2) = \frac{-2\pi}{|\Omega|}\boldsymbol{r}_{\mathcal{C}}(c_2) \cdot \boldsymbol{r}'_{\mathcal{C}}(c_2) = -8\boldsymbol{r}_{\mathcal{C}}(c_2) \cdot \boldsymbol{r}'_{\mathcal{C}}(c_2)$$

and, equivalently,

$$p_{\mathcal{F}}(f_2) = -8\boldsymbol{r}_{\mathcal{F}}(f_2) \cdot \boldsymbol{r}'_{\mathcal{F}}(f_2).$$

With this we have derived for the probability measures $P_{\mathcal{F}}$ and $P_{\mathcal{C}}$ probability density functions with respect to the Lebesgue measures on the related measurable spaces $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$ and $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$. The absolute continuity, i.e., $P_{\mathcal{F}} \ll \kappa$ and $P_{\mathcal{C}} \ll \kappa$ follows directly since for any Lebesgue null sets $F \subset \mathcal{F}$ and $C \subset \mathcal{C}$, it holds that

$$\begin{aligned} P_{\mathcal{F}}(F) &= \int_F p_{\mathcal{F}}(f)\ \mathrm{d}\kappa(f) = 0, \\ P_{\mathcal{C}}(C) &= \int_C p_{\mathcal{F}}(c)\ \mathrm{d}\kappa(c) = 0. \end{aligned}$$

This proves the statement of the proposition. □

To derive the probability density function in the joint space, we apply the concept of line integrals. We point out that by design of our example problem, we can assume that every classification value is matched to a unique feature value and vice versa. Therefore, it is a valid assumption that the support of the joint probability measure $P_{\mathcal{F}\times\mathcal{C}}$ is a curve in the joint space $\mathcal{F} \times \mathcal{C}$ and also the considered probability density function is living on that curve. For the proof of the joint probability density function, we first cite the concept of integration along a curve from [24].

Definition 5.104 (Integration along a curve)
Let $\gamma : [a, b] \to \mathbb{R}^n$ be a piecewise continuous curve. It holds that $\gamma([a, b])$ is a compact subset of $\mathbb{R}^n$. We consider $f : \gamma([a, b]) \to \mathbb{R}$ to be a continuous function.
Then the integration of f along the curve γ is described by

$$\int_\gamma f \, \mathrm{ds} := \int_a^b f(\gamma(t)) \, \|\gamma'(t)\|_2 \, \mathrm{d}t.$$

Since the joint probability measure is concentrated on a curve with Lebesgue measure zero in the joint space, it follows directly that the joint probability measure is not absolutely continuous to the Lebesue measure, i.e.,

$$P_{\mathcal{F}\times\mathcal{C}} \not\ll \kappa$$

holds with the Lebesgue measure κ on the measurable space $(\mathcal{F} \times \mathcal{C}, \mathcal{B}(\mathcal{F} \times \mathcal{C}))$ according to Notation 5.18. Still, it is possible to derive a closed form expression for the joint probability density function $p_{\mathcal{F}\times\mathcal{C}}$ when considering the integration along the $\mathcal{H}^1$ curve that the probability measure $P_{\mathcal{F}\times\mathcal{C}}$ is living on. To prepare the corresponding proof, we first derive a parameterization of the curve.

Proposition 5.105 (Parameterization along a curve)
Let the $\mathcal{H}^1$ curve on which the probability measure $P_{\mathcal{F}\times\mathcal{C}}$ lives be $S := \mathrm{supp}(P_{\mathcal{F}\times\mathcal{C}})$. The function $s : \mathcal{F} \to \mathcal{F} \times \mathcal{C}$ defined as

$$s(f) = (f, \tilde{c}(f))$$

with

$$\tilde{c} : [f_{\min}, f_{\max}] \to \mathcal{C}, \qquad \tilde{c}(f) = \frac{1}{\left(\frac{1}{f} - 1\right) \exp\left(\frac{v}{\varepsilon_0}\left(\hat{t}_0 - t_0\right)\right) + 1}$$

is a valid parameterization of S with $\|s'(f)\|_2 = \sqrt{1 + \tilde{c}'(f)^2}$. Here, $f_{\min}$ and $f_{\max}$ are considered to be the interval limits introduced in Proposition 5.101 again.

Proof. The probability measure $P_{\mathcal{F}\times\mathcal{C}}$ is concentrated on the curve S in $\mathcal{F} \times \mathcal{C}$ that is defined by (f, c) combinations for which the contour lines match in the image domain on Ω. If the contour lines of a certain $(f, c) \in \mathcal{F} \times \mathcal{C}$ match, it follows directly that the corresponding radii depending on the feature

value f and classification value c are identical. Based on the radius functions $\boldsymbol{r}_{\mathcal{F}}$ and $\boldsymbol{r}_{\mathcal{C}}$ introduced in Proposition 5.101, we derive the following:

$$\begin{aligned}
& \boldsymbol{r}_{\mathcal{F}}(f) = \boldsymbol{r}_{\mathcal{C}}(c) \\
\Leftrightarrow \quad & \varepsilon_0 \log(1/c - 1) + v(t - t_0) = \varepsilon_0 \log(1/c - 1) + v\left(t - \hat{t}_0\right) \\
\Leftrightarrow \quad & \log\left(\frac{\frac{1}{f} - 1}{\frac{1}{c} - 1}\right) = \frac{v}{\varepsilon_0}\left(t_0 - \hat{t}_0\right) \\
\Leftrightarrow \quad & \left(\frac{\frac{1}{f} - 1}{\frac{1}{c} - 1}\right) = \exp\left(\frac{v}{\varepsilon_0}\left(t_0 - \hat{t}_0\right)\right) \\
\Leftrightarrow \quad & \frac{1}{c} = \left(\frac{1}{f} - 1\right)\exp\left(\frac{v}{\varepsilon_0}\left(\hat{t}_0 - t_0\right)\right) + 1 \\
\Leftrightarrow \quad & c = \left[\left(\frac{1}{f} - 1\right)\exp\left(\frac{v}{\varepsilon_0}\left(\hat{t}_0 - t_0\right)\right) + 1\right]^{-1} =: \tilde{\boldsymbol{c}}(f).
\end{aligned}$$

This defines the classification function $\tilde{\boldsymbol{c}} : \boldsymbol{\mathcal{F}} \to \boldsymbol{\mathcal{C}}$. Furthermore, we derive

$$s'(f) = \left(1, \tilde{\boldsymbol{c}}'(f)\right)$$

which naturally leads to $\|s'(f)\|_2 = \sqrt{1 + \tilde{\boldsymbol{c}}'(f)^2}$ with

$$\begin{aligned}
\tilde{\boldsymbol{c}}'(f) &= (-1)\left[\left(\frac{1}{f} - 1\right)\exp\left(\frac{v}{\varepsilon_0}\left(\hat{t}_0 - t_0\right)\right) + 1\right]^{-2}(-1)\frac{1}{f^2}\exp\left(\frac{v}{\varepsilon_0}\left(\hat{t}_0 - t_0\right)\right) \\
&= \frac{\exp\left(\frac{v}{\varepsilon_0}\left(\hat{t}_0 - t_0\right)\right)}{f^2\left[\left(\frac{1}{f} - 1\right)\exp\left(\frac{v}{\varepsilon_0}\left(\hat{t}_0 - t_0\right)\right) + 1\right]^2}
\end{aligned}$$

This completes the proof. □

Remark 5.106. For the sake of completeness, we state that it is also possible to equivalently derive a parameterization of the curve S based on classification values by matching the corresponding contour lines for unique (f, c) combinations again. More precisely, with the help of $t : \boldsymbol{\mathcal{C}} \to \boldsymbol{\mathcal{F}} \times \boldsymbol{\mathcal{C}}$ defined as $t(c) = \left(\tilde{\boldsymbol{f}}(c), c\right)$ and with

$$\tilde{\boldsymbol{f}} : [c_{\min}, c_{\max}] \to \boldsymbol{\mathcal{F}}, \qquad \tilde{\boldsymbol{f}}(c) = \frac{1}{\left(\frac{1}{c} - 1\right)\exp\left(\frac{v}{\varepsilon_0}\left(t_0 - \hat{t}_0\right)\right) + 1},$$

we can parameterize the support of the probability measure by setting

$$S = \left\{t(c) \mid c \in [c_{\min}, c_{\max}]\right\}.$$

Here, we implement the classification limits $c_{\min}$ and $c_{\max}$ introduced in Proposition 5.101. Lastly, we state the corresponding derivative term for the used parameterization function t by $t'(c) = \left(1, \tilde{\boldsymbol{f}}'(c)\right)$ and with

$$\tilde{\boldsymbol{f}}'(c) := \frac{\exp\left(\frac{v}{\varepsilon_0}\left(t_0 - \hat{t}_0\right)\right)}{c^2\left[\left(\frac{1}{c} - 1\right)\exp\left(\frac{v}{\varepsilon_0}\left(t_0 - \hat{t}_0\right)\right) + 1\right]^2}.$$

Such parameterizations for the support of the probability measure $P_{\mathcal{F}\times\mathcal{C}}$ can be used to derive an expression for a probability density function $p_{\mathcal{F}\times\mathcal{C}}$ when considering the integration along the curve with respect to the $\mathcal{H}^1{}_{\llcorner S}$ measure.

Theorem 5.107 (Joint probability density function $p_{\mathcal{F}\times\mathcal{C}}$)
A joint probability density function $p_{\mathcal{F}\times\mathcal{C}}$ living on the one dimensional curve $S := \operatorname{supp}(P_{\mathcal{F}\times\mathcal{C}})$ can be derived such that

$$\int_{S_1} p_{\mathcal{F}\times\mathcal{C}}(f,c)\ \mathrm{d}\mathcal{H}^1{}_{\llcorner S}(f,c) = P_{\mathcal{F}\times\mathcal{C}}(S_1)$$

holds for any $S_1 \subset S$.
If the curve S is parameterized by a function $t : \mathcal{C} \to \mathcal{F}\times\mathcal{C}$, i.e., $S = \{t(c) \mid c \in [c_{\min}, c_{\max}]\}$ (cf. Remark 5.106) the probability density function is given as

$$p_{\mathcal{F}\times\mathcal{C}}(t(c)) = p_{\mathcal{C}}(c)\,\frac{1}{\|t'(c)\|_2} \tag{5.61}$$

whereas a parameterization by a function $s : \mathcal{F} \to \mathcal{F}\times\mathcal{C}$, i.e., $S = \{s(f) \mid, f \in [f_{\min}, f_{\max}]\}$ (cf. Proposition 5.105) leads to

$$p_{\mathcal{F}\times\mathcal{C}}(s(f)) = p_{\mathcal{F}}(f)\,\frac{1}{\|s'(f)\|_2}. \tag{5.62}$$

Proof. We carry out the proof exemplarily for S being parametrized by a function depending on feature values as introduced in Proposition 5.105.
Let S_1 be an arbitrary subset of $S = \{s(f) \mid, f \in [f_{\min}, f_{\max}]\}$. Then exist $f_1, f_2 \in \mathcal{F}$ with $f_{\min} \le f_1 < f_2 \le f_{\max}$ such that $S_1 = \{s(f) \mid, f \in [f_1, f_2]\}$ holds. We stress here the strict inequality of f_1 and f_2 to infer that S_1 is not a $\mathcal{H}^1$ null set.
We assume that $p_{\mathcal{F}\times\mathcal{C}} : S \to \mathbb{R}$ exists such that

$$\int_S p_{\mathcal{F}\times\mathcal{C}}(f,c)\ \mathrm{d}\mathcal{H}^1{}_{\llcorner S}(f,c) = P_{\mathcal{F}\times\mathcal{C}}(S) = 1,$$

i.e. $p_{\mathcal{F}\times\mathcal{C}}$ is the probability density function when considering the integration with respect to $\mathcal{H}^1{}_{\llcorner S}$. We apply the line integration formula stated in Definition 5.104 and integrate only over the subset S_1 now. Then, we receive

$$\begin{aligned}\int_{S_1} p_{\mathcal{F}\times\mathcal{C}}(f,c)\ \mathrm{d}\mathcal{H}^1{}_{\llcorner S}(f,c) &= \int_{f_1}^{f_2} p_{\mathcal{F}\times\mathcal{C}}(f,\tilde{c}(f))\,\|s'(f)\|_2\,\mathrm{d}f \\ &\overset{\text{projection on } \mathcal{F}}{=} \int_{f_1}^{f_2} p_{\mathcal{F}}(f)\ \mathrm{d}f.\end{aligned}$$

Since $S_1 = \{s(f) \mid, f \in [f_1, f_2]\}$ is an arbitrary subset of S, it follows that the previous equation holds *for all* subsets of S. This implies that the integrands need to be identical to evaluate to the same values

when integrating over any subset $S_1 = \{s(f) \mid, f \in [f_1, f_2]\}$ of S with arbitrary f_1 and f_2. Consequently, it holds for all $(f, \tilde{\boldsymbol{c}}(f)) \in S$ that

$$p_{\mathcal{F}\times\mathcal{C}}(f, \tilde{\boldsymbol{c}}(f)) = p_{\mathcal{F}}(f) \frac{1}{\|s'(f)\|_2}$$
$$\overset{\text{Propositions 5.103 and 5.105}}{=} -8\boldsymbol{r}_{\mathcal{F}}(f)\, \boldsymbol{r}'_{\mathcal{F}}(f) \frac{1}{\sqrt{1+\tilde{\boldsymbol{c}}'(f)^2}}.$$

□

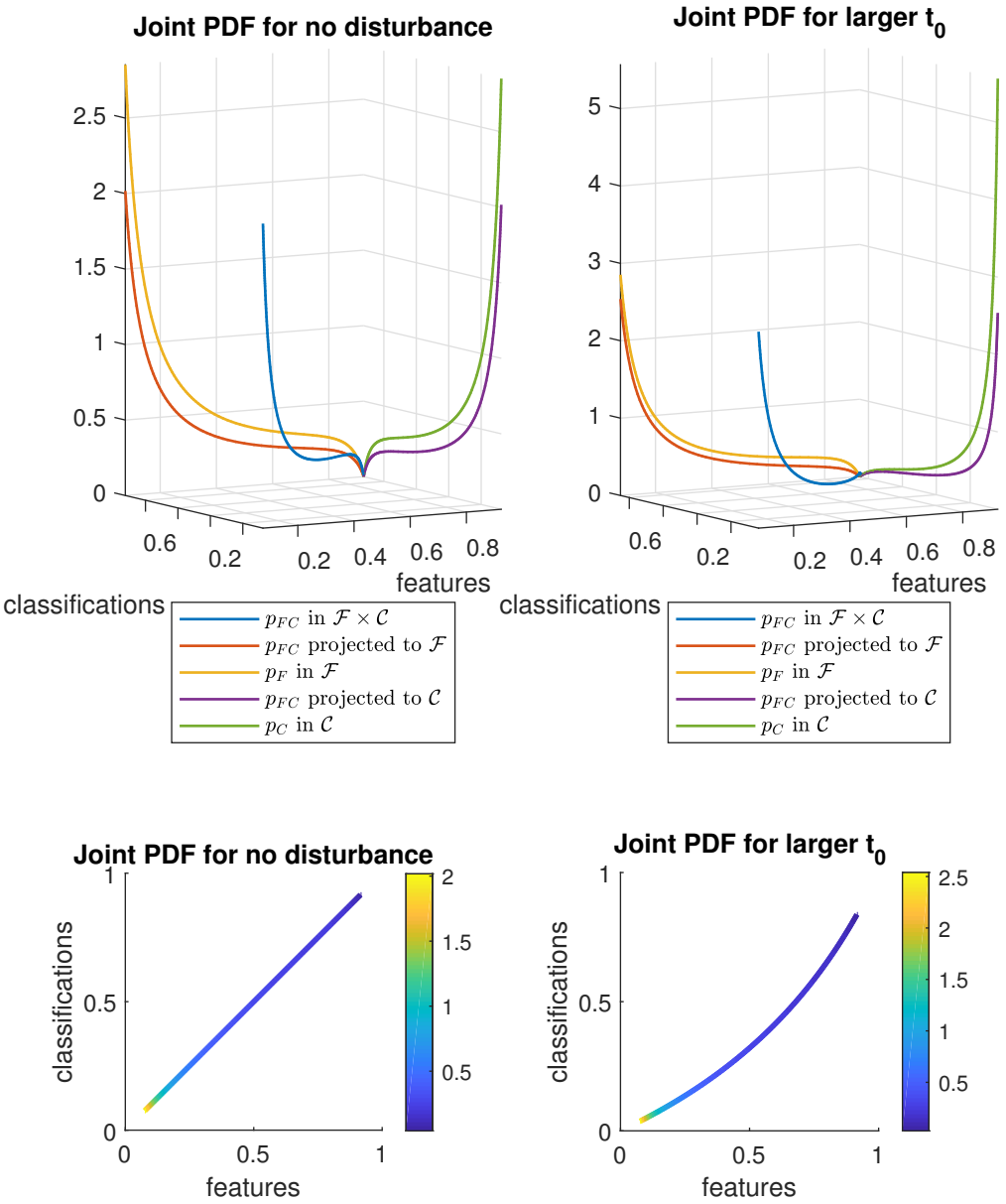


Figure 5.13: Continuous probability density function in the joint space $\mathcal{F} \times \mathcal{C}$.

We illustrate the joint probability density function in Figure 5.13. We plot the individual probability density functions $p_{\mathcal{F}}$ and $p_{\mathcal{C}}$ compared to $p_{\mathcal{F}\times\mathcal{C}}$ and its projections onto $\mathcal{F}$ and $\mathcal{C}$ in the first row of subplots. Here, we consider again parameter settings for the classification space for "no disturbance"

and for a "later t_0". The shown functions are evaluated at the time point $t = t_1 = 0.5$, but we considered a normalization by the domain related to two time frames. This is to keep consistency when comparing the probability density function $p_{\mathcal{F}}$ from this figure with the one on the left hand side of the second row in Figure 5.11 for the first time stamp $t = 0.5$. We point out that $p_{\mathcal{F}}$ and the projection of $p_{\mathcal{F}\times\mathcal{C}}$ to $\mathcal{F}$ are not identical and, similarly, $p_{\mathcal{C}}$ does not coincide with the projection of $p_{\mathcal{F}\times\mathcal{C}}$ to $\mathcal{C}$ either. We refer to Equations (5.61) and (5.62) to stress the differences lying in the scaling effects by $\frac{1}{\|t'(c)\|_2}$ or $\frac{1}{\|s'(f)\|_2}$, respectively.
To stress the fact of $p_{\mathcal{F}\times\mathcal{C}}$ lying on a curve in the joint space, we plotted in the second row the corresponding probability density function in a top view perspective for both chosen parameter settings. The probability density measure has indeed a *curve* as a support which is here presented in an enlarged widths for illustration aspects. Furthermore, we can again see in the color scaling the transition from higher accumulations near small feature and classification values to less combinations towards the upper right corner at $(1, 1)$ which relate to smaller circles in the spatial domain.

After highlighting that the derived probability density function $p_{\mathcal{F}\times\mathcal{C}}$ is leaving on the $\mathcal{H}^1$ curve S and already stating earlier that the joint probability measure $P_{\mathcal{F}\times\mathcal{C}}$ is not absolutely continuous with respect to the Lebesgue measure on the joint space and even not to the product measure $P_{\mathcal{F}} \otimes P_{\mathcal{C}}$, we highlight that we cannot calculate the MI based on the given definitions. For once, we want to infer that the mutual information equals infinity when the joint probability measure is concentrated on a lower dimensional curve. This can be derived by approximation arguments and by using a support of the joint probability measure converging to such a lower dimensional curve. In formulas this means that the joint measure is living in a n dimensional space, but is concentrated on a $\mathcal{H}^{n-1}$ curve which has the measure zero when considering the Lebesgue measure of the related higher dimensional space.
We do not delve into this issue further or perform exact proofs for the stated arguments. When coming back to our numerical toy problem, we can assume the joint probability measure not to be concentrated on a lower dimensional curve. Since we are facing discretization issues when performing calculations on a computer, we are anyway dealing with a joint probability measure with a support that is not concentrated on a thin curve. Depending on the binning width to calculate the histograms, of course, we can approximate such a lower dimensional curve, but will never reach it up to a certain error.
In the next section, we focus on such example discretizations and perform the optimization for our toy problem considering parameter initializations for the classification image that match the parameters considered for the example case of "a later t_0".

5.5.1.3 Numerical convergence tests

In the final section on our toy problem described in Definition 5.99, we focus on the numerical solution of the related optimization problem. Therefore, we first introduce different discretization stages that our numerical test is based on and present some implementation details. Secondly, we showcase the numerical results for our toy example to complete the proof of concept section.

According to the convergence results in the previous Sections 5.3 and 5.4, especially in Theorem 5.98, we design our discretization stages for our toy problem. For the spatial domain of the unit square, we apply a certain pixel width h to simulate pixel width occurring in real data. On the coarsest scale, we consider a pixel width of $h = 0.1$ resulting in 100 pixels per time stamp. For the temporal discretization, we use again the time points $t_1 = 0.5$ and $t_2 = 1$ as two discrete time stamps.
Furthermore, we consider a binning width for the histograms based on the feature and classification images. As we model both images in the same manner and it holds that $\mathcal{C} = \mathcal{F} = \mathbb{R}$, we apply identical binning widths in the feature space $\mathcal{F}$ and in the classification space $\mathcal{C}$. More precisely, we set $\Delta b := \Delta f = \Delta c$. On the coarsest scale, we use $\Delta b = 0.2$. Since $\mathcal{C} = \mathcal{F} \subset [0, 2]$ holds, we only consider this interval for the binning of the feature and classification spaces which results in five bins on both axis. In the joint space, we get 25 bins for this binning widths and this is in line with the requirement to have much less bins compared to the pixels which are 200 in total on the coarsest scale.
For the mollification effect in the classification domain, we apply convolution with a cubic B-spline. The B-spline is normalized to fit the mollification property that its integral over the whole space $\mathbb{R}$ equals to 1. As the scaling factor, we consider here the support of the B-spline and denote it with ε_1 again. We already indicated in the previous sections, that it is important to avoid the mollification to converge faster than the binning width (cf. Remark 5.84). Otherwise, the mollification would get lost in the histogram discretization, if the support of the mollification function got smaller then the discretization width Δb in the classification space. We set the initial kernel's support widths to four times the initial binning width to ensure that there are enough discrete evaluation points available when considering a discrete convolution.
To calculate the discretization parameters for several stages and to comply with the stated convergence requirements for the binning sizes in Theorem 5.98, we consider ε to scale the different parameters according to

$$h = \varepsilon h_0, \quad \Delta b = \varepsilon^\beta \Delta b_0, \quad \varepsilon_1 = \varepsilon^\alpha \varepsilon_{10} \tag{5.63}$$

We consider ε to converge to zero, more precisely we use $\varepsilon \in \{1, \frac{1}{16}, \frac{1}{81}, \frac{1}{256}\}$ to scale the discretization widths. In this sense, the parameters with zero in the subscript are equal to the corresponding coarsest sizes. Furthermore, we use $\beta = \frac{1}{2}$ and $\alpha = \frac{1}{4}$ to ensure the wanted convergence order for the discretization parameters. Consequently, it holds that

$$\frac{h}{\Delta b} = \frac{\varepsilon h_0}{\varepsilon^\beta \Delta b_0} = \frac{h_0}{\Delta b_0} \varepsilon^{\frac{1}{2}},$$
$$\frac{\Delta c}{\varepsilon_1} = \frac{\varepsilon^\beta \Delta b_0}{\varepsilon^\alpha \varepsilon_{10}} = \frac{\Delta b_0}{\varepsilon_{10}} \varepsilon^{\frac{1}{2}}$$

which converge to 0 for $\varepsilon \to 0$. This is in line with Equation (5.59) in Theorem 5.98. In Table 5.2, we state the used discretization parameter in our numerical test. We calculate for four stages the pixel width, binning width and the size of the mollification kernel's support based on the approach given in Equation (5.63). Before we continue, we remark that we used the discretizations on the finest scale for the visualizations in the previous sections.
For each discretization stage, we calculate the feature and classification images on a discrete pixel grid, perform a histogram binning and convolve the joint histogram along the classification axis.

ε	ε^{β}	ε^{α}	pixel width h $h_0 = 0.1$	binning width Δb $\Delta b_0 = 0.2$	width of mollification kernel's support ε_1 $\varepsilon_{10} = 0.8\,(= 4 \cdot \Delta b_0)$
1	1	1	$\frac{1}{10} = 0.1$	$\frac{1}{5} = 0.2$	$\frac{4}{5} = 0.8$
$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{160} \approx 0.0063$	$\frac{1}{20} = 0.05$	$\frac{2}{5} = 0.4$
$\frac{1}{81}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{810} \approx 0.0012$	$\frac{1}{45} \approx 0.0222$	$\frac{4}{15} \approx 0.2667$
$\frac{1}{256}$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{2560} \approx 0.0004$	$\frac{1}{80} = 0.0125$	$\frac{1}{5} = 0.2$

Table 5.2: Discretization stages for the pixel widths, binning widths and widths of the mollification kernel's support in relation to the convergence parameter ε and with $\alpha = \frac{1}{4}$, $\beta = \frac{1}{2}$.

In Figures 5.14 and 5.15, we show the joint histograms for the four discretization stages row-wise. We compare the joint histograms before and after mollification in the classification domain (left vs. right column). In Figure 5.14, we use the joint histograms based on optimal classification parameters whereas in Figure 5.15, we present the results based on the histograms related to the disturbed classification parameters with later t_0.
We want to point out three important observations. To begin with, we note that the *bending effect* due to the disturbed parameters is present in the second figure (Figure 5.15) for all discretization stages. In comparison to this, we get a joint histogram concentrated only on the (discretized) main diagonal for all discretizations in the left column of the first figure (Figure 5.14) presenting the optimal parameter case.
Comparing the left and right column, the *vertical smoothing effect* due to the mollification in the classification space is significantly perceivable on the right hand side. We stress that this is observable in both figures or, respectively, for both parameter combination. Of course, this smoothing effect will still be present if the classification image is generated with another parameter combination.
Finally, we highlight the *converging support* of the histograms. In a top-down perspective, it is obvious that the support of the joint histograms is converging towards a thin curve, especially in the first column without any smoothing effects. Also in the right columns of both figures, the support of the mollified joint histogram is strikingly shrinking. It is also worth mentioning that with the above chosen discretization stages in Table 5.2, we indeed ensure that the mollification kernel's support is always larger than the binning width in the classification space and, consequently, the mollification will always have an effect and cannot be hidden in the classification discretizations.

After having spent some words on the different discretization stages and also presented the mollification effect visually, we delve next into some more implementation details. Based on the smoothed joint histogram, we compute the negative mutual information as our target optimization functional. Furthermore, we provide an approximation of the gradient of mutual information when considering the different spreading parameters influencing the classification image. Both terms are required for our gradient-based optimization approach. We apply the MATLAB optimization function *fmincon*. This solver is used to "find [the] minimum of [a] constrained nonlinear multivariable function" [52]. The constraints we are using for our optimization problem are lower and upper bounds for our pa-

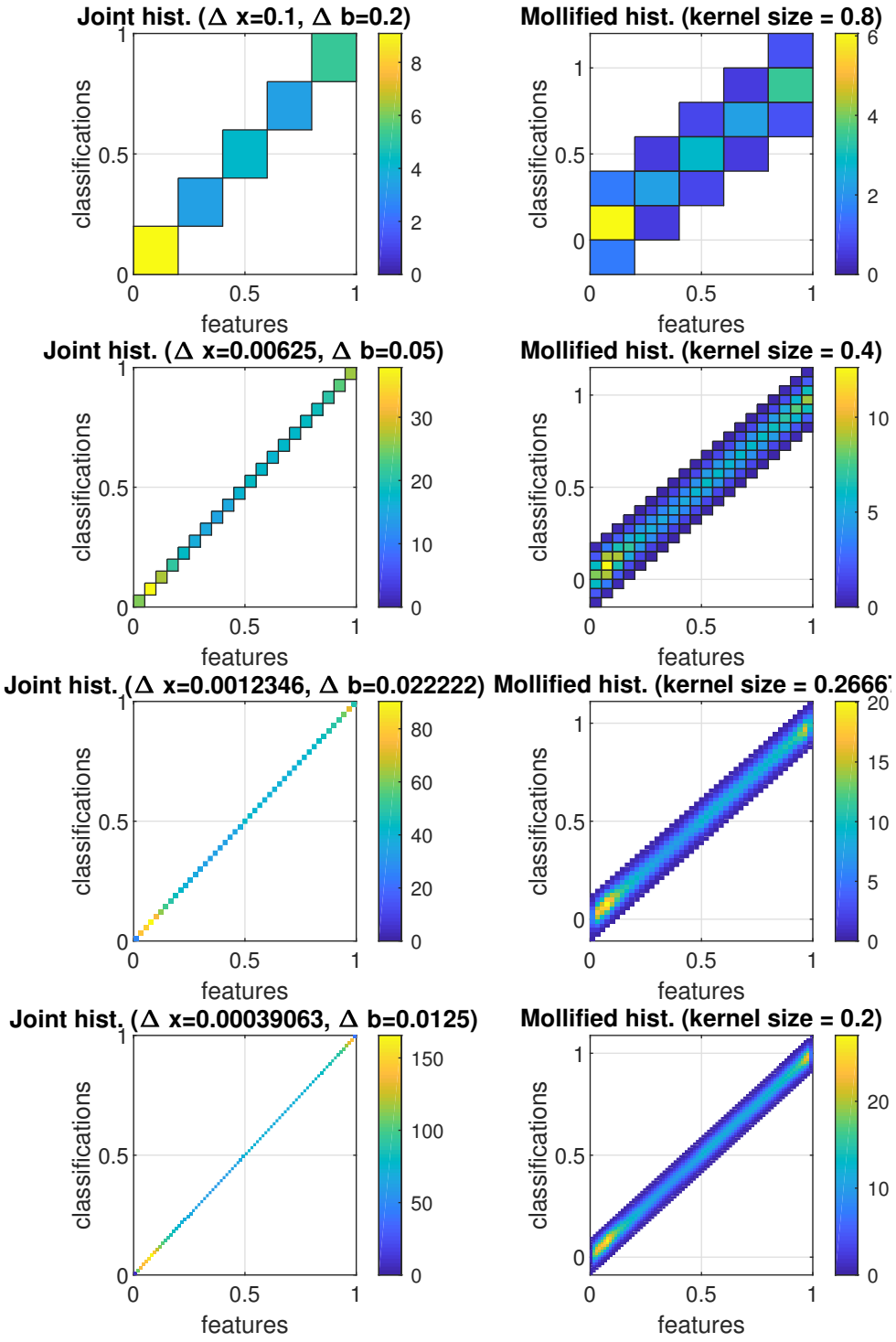


Figure 5.14: Smoothing effect in the classification domain on the joint probability density function $p_{\mathcal{F}\times\mathcal{C}}$ considering optimal parameter settings.

rameter settings to ensure that the origin $\hat{x}_0$ lies within our domain, $\hat{t}_0$ is in the interval $[0, 1]$, the spreading speed is positive and such that for the smallest time step (*here: 0.5*) the moving front does not move further than the length of our square domain. We use the parameters related to "a later t_0" (cf. Table 5.1) for the initialization of our spreading properties for the classification image. For the optimization algorithm we use the default method "interior-point". The performance on our optimiza-

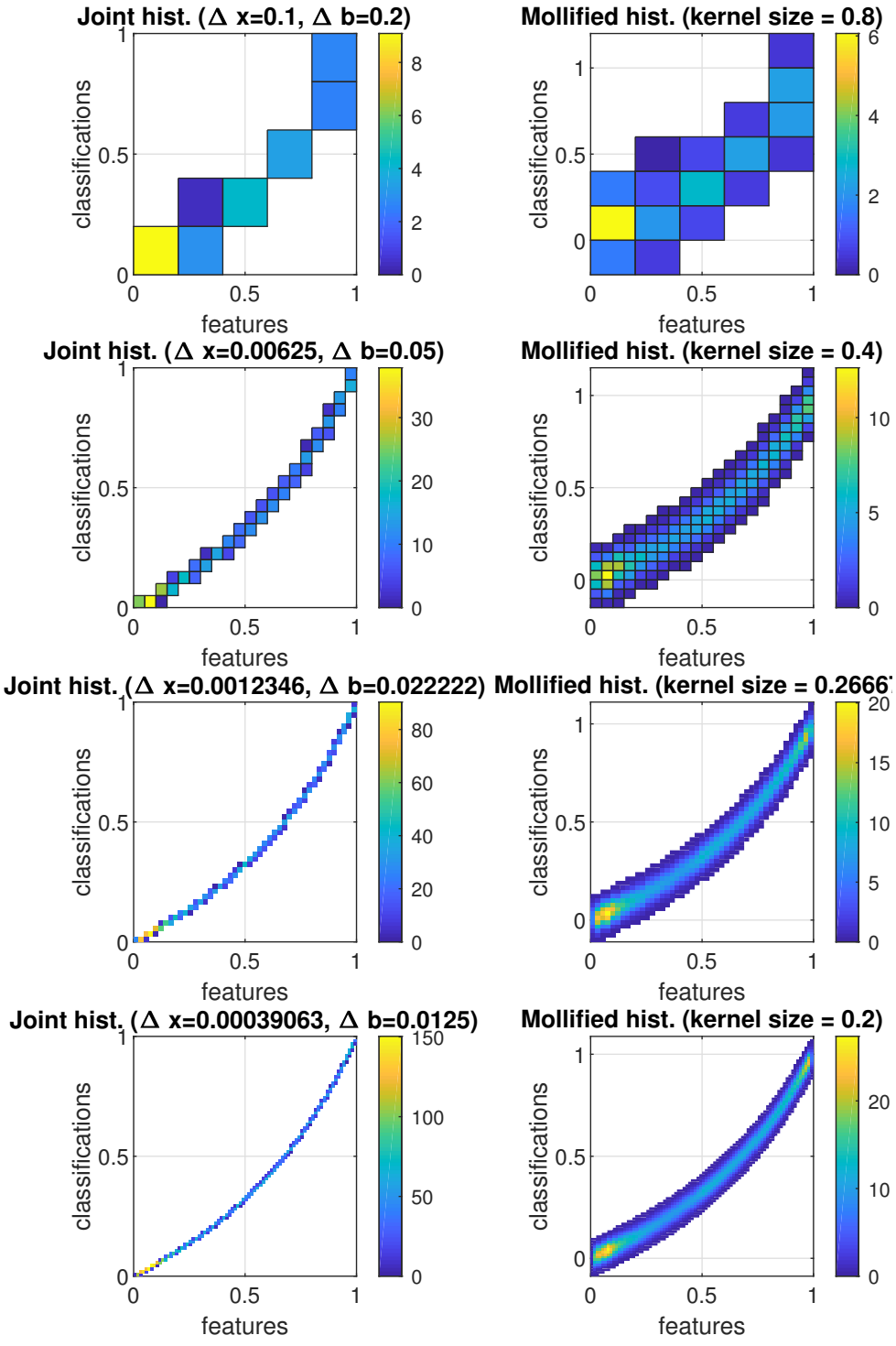


Figure 5.15: Smoothing effect in the classification domain on the joint probability density function $p_{\mathcal{F}\times\mathcal{C}}$ considering disturbed parameter settings.

tion problem was appropriate as it solved our optimization problem well in a reasonable computation time. Of course, for the finer discretization scales, the optimization process takes longer. Still, all four discretization stages were processed in less than half an hour on a local machine without a GPU or parallel computing. For every discretization level the optimization approach stopped because a limit for the step tolerance was reached while a tolerance on the constraints was satisfied. We apply the

default tolerance values for the step size of 10−10 and for the constraint tolerance of 10^{-6}.

$\hat{\boldsymbol{p}}$	initial parameters	optimized parameters				optimal parameters
		$\varepsilon = 1$	$\varepsilon = \frac{1}{16}$	$\varepsilon = \frac{1}{81}$	$\varepsilon = \frac{1}{256}$	
$\hat{\boldsymbol{x}}_{0,1}$	0.5	0.5	0.50009	0.5	0.49992	0.5
$\hat{\boldsymbol{x}}_{0,2}$	0.5	0.5	0.50009	0.5	0.50009	0.5
$\hat{t}_0$	0.15	0.20247	0.012612	0.02103	0.022289	0
$\hat{v}$	0.5	0.64686	0.50579	0.50264	0.50146	0.5

Table 5.3: Spreading parameter $\hat{\boldsymbol{p}} = \left(\hat{\boldsymbol{x}}_0, \hat{t}_0, \hat{v}\right)$.

MI for	initial parameters	optimized parameters				optimal parameters
		$\varepsilon = 1$	$\varepsilon = \frac{1}{16}$	$\varepsilon = \frac{1}{81}$	$\varepsilon = \frac{1}{256}$	
$\varepsilon = 1$	−0.7559	−0.8103				−0.9421
$\varepsilon = \frac{1}{16}$	−1.4267		−1.5262			−1.5465
$\varepsilon = \frac{1}{81}$	−1.8013			−1.8683		−1.8968
$\varepsilon = \frac{1}{256}$	−2.0609				−2.1356	−2.1539

Table 5.4: Resulting negative MI for the different parameter settings on the four discretization stages.

In Table 5.3, we collect the optimization results stating the optimized parameter settings for the different discretization stages. For the sake of completeness and to facilitate comparisons, we include the initial parameter settings (left column) and the optimal parameter settings (right column), too. We stress that we the *optimal parameters* correspond to the ground truth parameters used to generate the feature images while the *optimized parameters* are the parameters calculated by the numerical optimization solver.

In Table 5.4, we complete our optimization results by stating the related negative value of the mutual information for the different discretization stages and for the initial, optimized and optimal parameter settings. Here, we observe that for each discretization stage, the value related to the optimal parameter setting is the smallest and the one for the parameter initialization is the largest. The value for the negative MI approaches the corresponding value for the optimal settings on the same stage and also the optimized parameter setting converges to the optimal parameters for the decreasing discretization quantities (rightest column for optimized parameters in Table 5.3).

In Figure 5.16, we illustrate this convergence statements. We plot the values of the negative MI for the different parameter settings when considering ε to scale the discretization stages as described above, cf. Equation (5.63). The trend for the function value for decreasing discretization scales is obvious: For discretization scales converging to zero the negative MI value drops faster which is in line with the expectation of the MI being infinite when concentrated on a thin line. Moreover, the negative MI is on every discretization stage larger for the initial parameter settings (blue curve) compared to the other two parameter settings. In addition, the curve for the optimized settings (orange curve) converges towards the curve related to the optimal parameter choice (red curve).

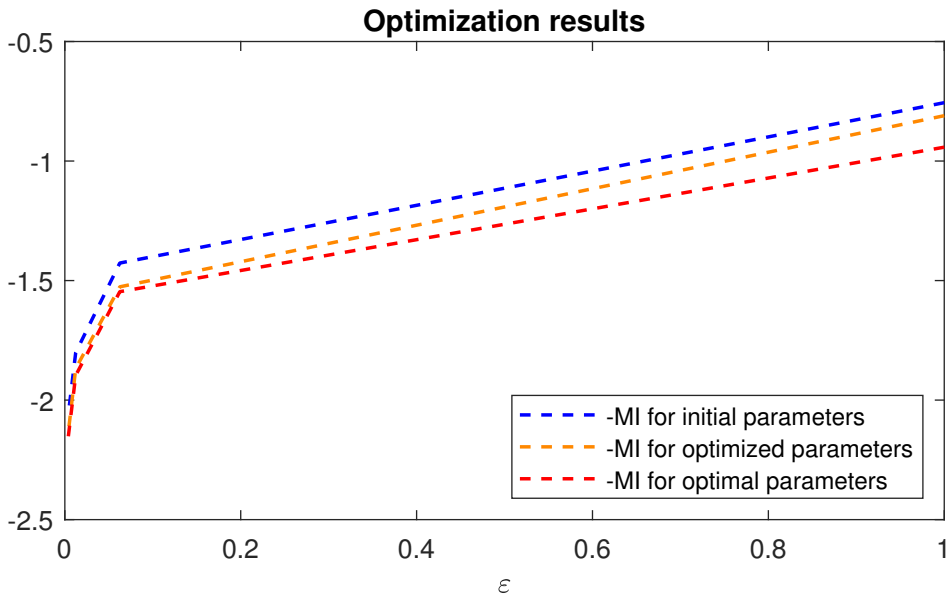


Figure 5.16: Converging values for the negative MI calculated with the optimization approach and compared to the negative MI calculated with the initial and optimal parameter settings.

In Figure 5.17, we plot the joint histogram for the different parameter settings in our optimization approach. Column-wise we present the histograms for the initial parameters, the optimized parameters and the optimal parameters. From top to bottum, the discretization scales decreases row-wise. We present again the four different discretization scales.
In the first column, we observe the largest distortion of the support of the histogram whereas in the last column for the optimal parameters the histogram entries are accumulated close to the main diagonal. Again, for smaller discretization scales the support approaches a lower dimensional curve located near the main diagonal and which is mollified in the direction of the classification axis. We observe row-wise that for the optimized parameter settings the histogram appearance converges towards the appearance for the optimal parameters and gets more and more condensed near the main diagonal. The curvature prominent in the first column which is due to the parameter disturbance of t_0, is in the middle column reduced.
To conclude, the optimization applied to the toy example serves as a proof of concept. It shows that based on feature images combined with classification images, we are able to extract spreading properties, i.e., the parameter setting to model an optimized classification image, by applying a gradient-based optimizer on the mutual information term depending on the related histograms. The resulting spreading parameters approximate spreading properties which relate to a spreading phenomenon present in the feature images.

In the next section, we focus on a real data set again. As a second proof of concept, we present the optimization approach applied to feature images extracted from the microscopy data provided by AstraZeneca and aim for spreading properties with which we can map the occurring spreading phenomena present in the microscopy images to classification images. Indeed, this spreading information is of main interest here to facilitate a high-throughput analysis of microscopy images capturing cell colony development.

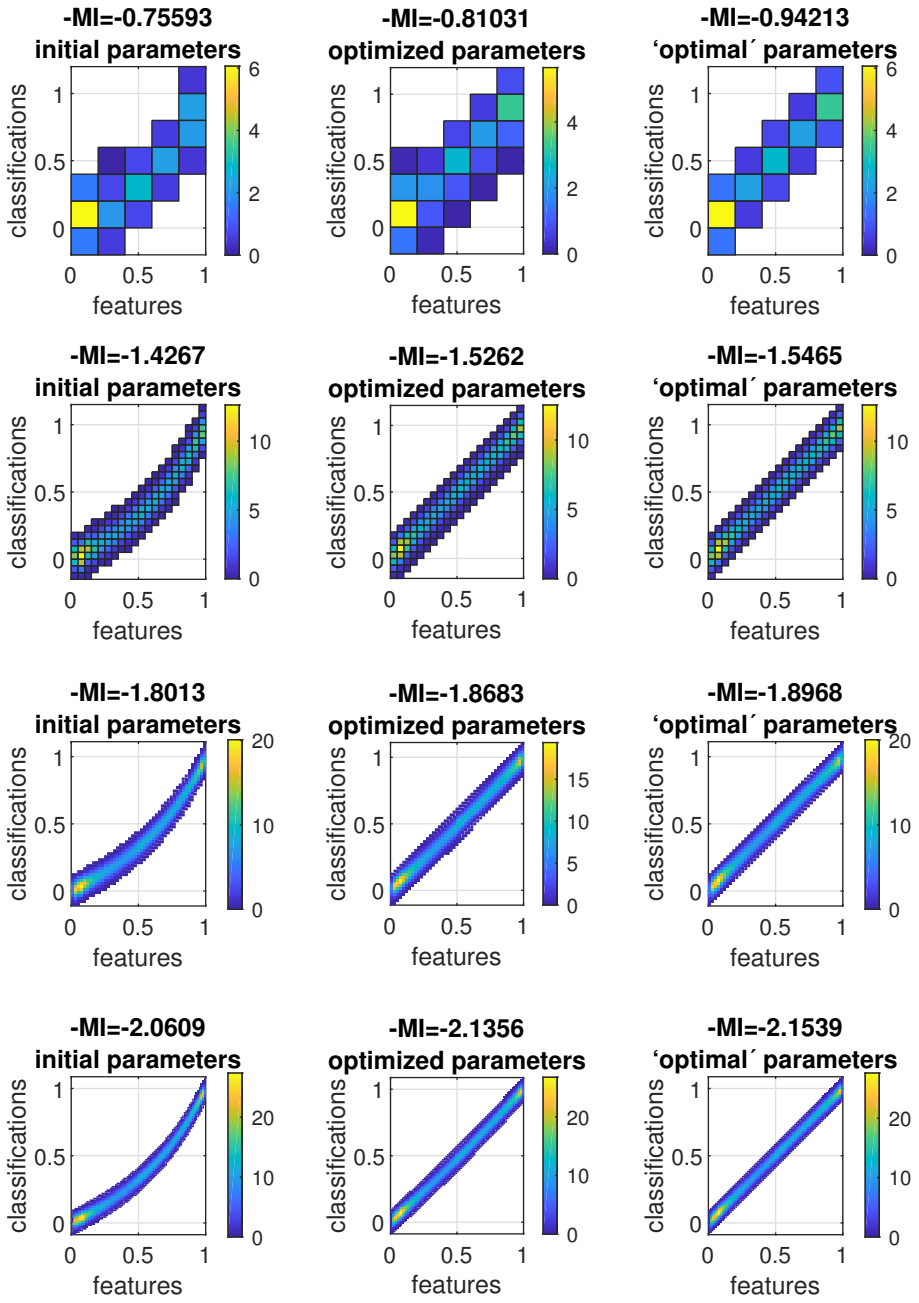


Figure 5.17: Joint histograms and MI for different discretizations based on the initial parameter setting, the optimized parameter setting and the expected, optimal parameter setting.

5.5.2 Numerical optimization for AstraZeneca's data

In this section, we apply the model fitting via the mutual information based optimization approach to microscopy data of AstraZeneca. We choose two example wells with developing cell colonies and extract spreading properties from the related texture data with the implemented software tool. We refer to Figure 3.1 to recapitulate the spreading phenomenon in the first example well B4 of the first plate. In Figures 5.18 and 5.19, we present two example time points showing the colony growth for the second example well I11 from the first plate. We apply a limited color range to these microscopy images for contrast enhancements within the well highlighting the spreading cell colonies.

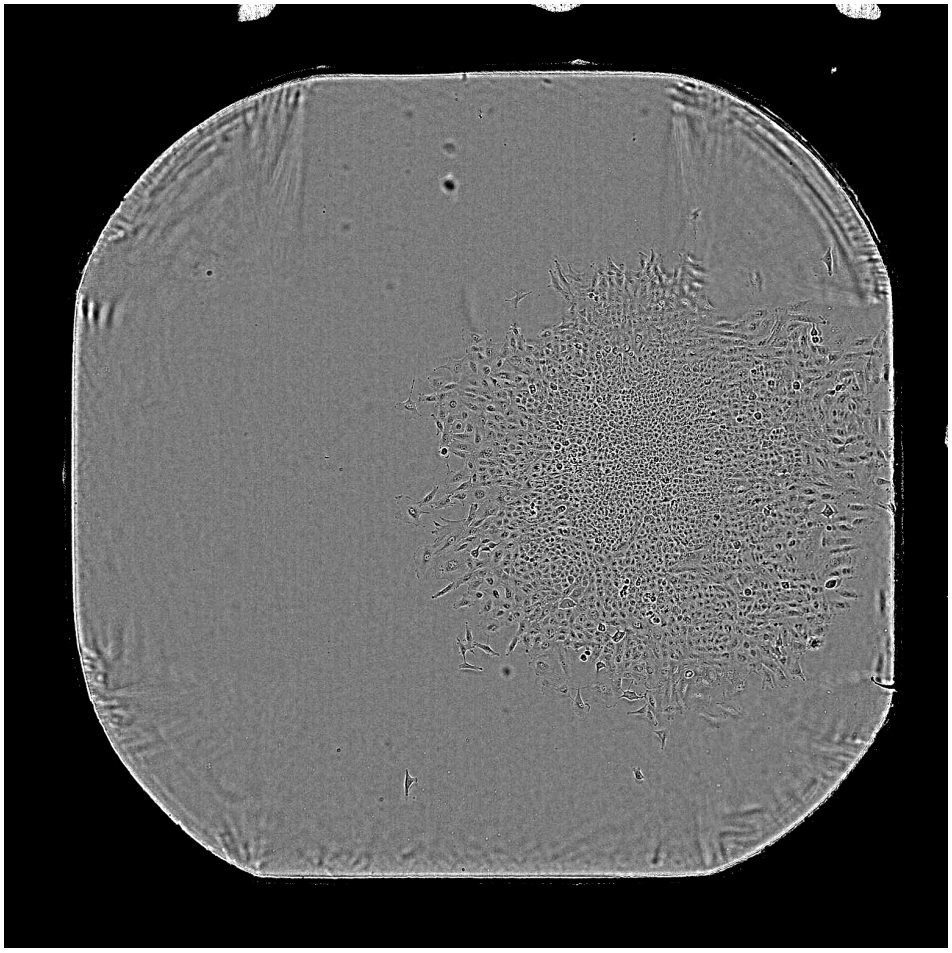


Figure 5.18: Colony development state five after approx. seven days for well I11 of plate 1 from the AstraZeneca data set (limited color range for contrast enhancement).

For extracting spreading properties for the growing cell colonies in well B4 and I11, we first process the time series of both wells *individually*. The analysis of the separate runs for extracting independent spreading information is the subject of Section 5.5.2.2. In this context, we present a preview of the classification images for each time point next to the related feature images based on basic texture descriptors.
In Section 5.5.2.3, we discuss the results of processing both wells jointly. The idea of the mutual information based model fitting is to apply an approach which allows to match similar texture regions

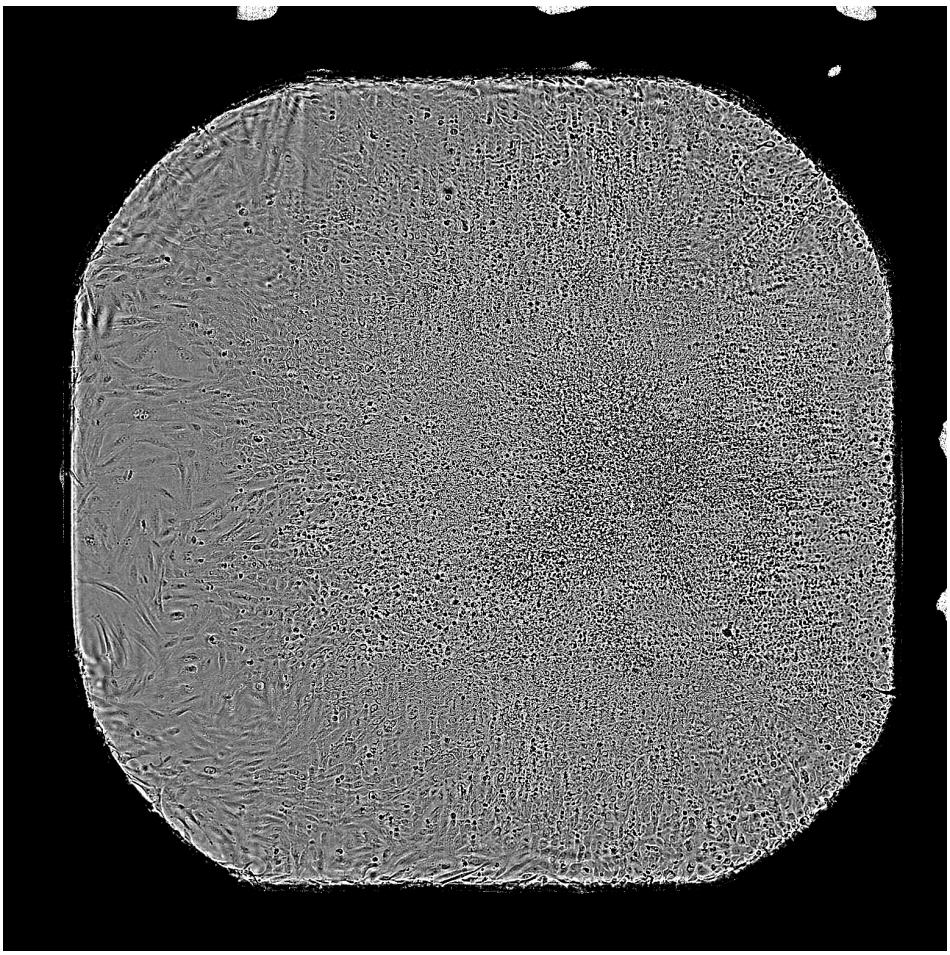


Figure 5.19: Colony development state seven after approx. fourteen days for well I11 of plate 1 from the AstraZeneca data set (limited color range for contrast enhancement).

of different wells' time series into the same classes. Consequently, we intuitively expect that the *joint* processing of the wells is crucial to allow comparisons between cell colonies of different wells. Hence, we use the multi-dimensional feature point clouds of matched texture regions in the two example wells to compare the approaches of the *separate* and the *joint* optimization. However, it turns out that both approaches do not differ significantly in the end. We achieve similar spreading results and comparable MI values for the joint and the separate optimizations. Moreover, the identification of two different subpopulations within a growing cell colony proves to be challenging. With the implemented optimization considering feature images based on simple texture properties, we rather get a differentiation between the inner cell colony and the moving front, i.e., the transition between the cell colony and background regions rather than *normal* versus *abnormal* cellular appearances. We summarize our findings and discuss the results on the optimization for the AstraZeneca data in Section 5.5.2.4.

Before we delve into the numerical results, we start in Section 5.5.2.1 by giving some introductory information on the settings used in the optimization. For consistency reasons, these settings are used in both approaches, i.e., in the separate and joint runs.

5.5.2.1 Settings used in numerical experiments

In this introductory section, we present the basic information on the settings we use for our numerical experiments on two example wells of the AstraZeneca microscopy data set. We choose the time series of well B4 and I11 on the first plate for which we observe developing cell colonies in the phase contrast images. Even by manual investigation the colony growth is observable (cf. Figures 3.1, 5.18 and 5.19) and the time series are serving for our proof of concept test for real data. Instead of using the raw microscopy images, we include the extracted feature images in the optimization approach. We recapitulate that we focus on three basic texture properties, namely the local minima, local maxima and the interquartile ranges of the gray values in small neighborhoods. We refer the reader to Section 3.3.2 in which we introduced the used texture descriptors in more detail.
We use down-sampled versions of the feature images for computational aspects. Rather than using the full image size of 1548×1548 pixels, we reduce the size to 387×387 pixels by applying bilinear down-sampling on the extracted feature images. Instead of aiming for more accuracy, we accept the lower resolution images to speed up the numerical optimization. For the numerical solution, we concentrate on the pixels which lie within a down-sampled cropping frame. As a cropping frame we use again the segmentation mask of a reference well (cf. Section 3.1).
For the further discretizations related to the histogram generation, we apply binning widths of $\Delta c = 0.25$ for a total classification range of $[0, 2]$ and $\Delta f = 0.05$ for each feature dimension with maximal range of $[0, 1]$ each (cf. Equation (3.8)). The smoothing step of the joint histograms in the classification direction is based on convolution with a discrete B-spline kernel stretching over seven bins in the discretized classification domain, i.e., the kernel's width is $7 \cdot \Delta c = 1.75$.
The classification image itself is generated based on the circular spreading model combined with the approximation of the Heaviside step function (cf. Section 4.2). We use $\varepsilon_0 = 10$ as the inherent model parameter to achieve smoothed approximations between the main classifications of 0 related to *background regions*, 1 to *areas of normal cells* and 2 related to *abnormal cell regions*.
Finally, we comment on constraints and solver settings applied for the numerical solution. We use again the MATLAB solver *fmincon* [52] with the algorithm *interior-point*. The *interior-point* method is the recommended algorithm [51] and resulted in adequate numerical results with respect to computation time and convergence. For the stopping criteria, we reduced the default step size tolerance and apply 10^{-14} as the minimal relative step size tolerance which is close to machine accuracy. In Section 5.5.2.3, we comment on the stopping criteria in more detail.
For the minimization problem, we include lower and upper bounds for our spreading properties. For the colony's origin $\boldsymbol{x}_0$, we enforce that it lies *within* our spatial domain Ω. In this context, when calculating with pixels, we make sure that the origin's coordinates take values between 0 and 387. For the starting time points $t_{0,n}$ and $t_{0,a}$ when the "normal" front and the "abnormal" front start to emerge, we apply bounds as well. We claim that the time point for the normal front $t_{0,n}$ takes values between 0 and the final time point. For the second temporal property $t_{0,a}$, we set the upper bound to infinity. With this, we make sure that a normal front is always emerging and an abnormal one might be emerging within the applied time window. Furthermore, we use a linear constraint to incorporate that the time point for the normal front must be smaller than the one for the abnormal front. With this, we ensure that an abnormal front can only emerge *after* a normal one and, consequently, can

only take nonnegative values as well. We stress that we consider the time points to be given in *days*. Last but not least, we add a constraint for the spreading velocity. For physical consistency, we require the velocity to be nonnegative. A maximal velocity is derived from the domain width divided by the maximal time step. The interpretation is that we expect the colony to move slower than this because we suppose that the colony cannot spread from one border of the domain to the opposite one between two consecutive time frames. As the imaging does not take place at equidistantly distributed time points, we use the largest time step to enforce this constraint to hold for all discrete consecutively recorded time frames. For the sake of completeness, we recapitulate in Table 5.5 the approximate time stamps which are introduced in Section 2.1, more precisely in Table 2.1. In the last column of the table we add an approximation of the time step widths between consecutive frames. The table serves as a reference for mentioned time points and time step widths in the following Sections 5.5.2.2 and 5.5.2.3.

time stamp number	approx. time after initial time point	approx. time step
1	0	-
2	2 hours	2 hours
3	17 hours	15 hours
4	5 days and 15 hours	4 days and 22 hours
5	7 days and 17 hours	2 days and 2 hours
6	11 days and 15 hours	3 days and 22 hours
7	14 days and 15 hours	3 days
8	18 days and 15 hours	4 days

Table 5.5: Approximate time stamps relative to initial time point and the related approximate time steps between consecutive time frames.

After having introduced the numerical setting for solving the MI-based optimization problem, we have now all prerequisites at hand. In the following sections, we deal with the results when processing the two example wells B4 and I11. We start in the next Section 5.5.2.2 by applying the optimization approach to the wells separately and focus on the joint processing in Section 5.5.2.3 afterwards.

5.5.2.2 Extraction of spreading properties for individual wells

In this section, we solve the optimization problem for the two example wells individually. By maximizing the mutual information between the extracted texture features, i.e., the feature images, and the classfication images modeling the circular spreading phenomenon, we aim for new insights on the present growth process. To be more precise, with the individual optimization we derive spreading properties with which we can simulate a circular spreading for each well independently.
In Table 5.6, we state the parameter initializations we use for the optimization. The initializations are based on manual estimates for the spreading parameters derived from visual inspection. For well B4, we use a colony origin $\boldsymbol{x}_0$ horizontally centered in the upper first third of the domain while we guess that the spreading in I11 starts off in an origin vertically centered but closer to the right boundary

well	property	unit	parameter initializations	optimized parameters
B4	$x_{0,1}$	pixels	129	68.939
	$x_{0,2}$	pixels	193.5	194.89
	$t_{0,n}$	days	7.7083	5.5563
	$t_{0,a}$	days	14.625	8.5106
	v	$\frac{\text{pixels}}{\text{days}}$	18.651	17.973
	$-$ MI		-0.2974	-0.4245
I11	$x_{0,1}$	pixels	193.5	195.56
	$x_{0,2}$	pixels	290.25	263.05
	$t_{0,n}$	days	0.70833	2.749
	$t_{0,a}$	days	7.7083	5.8975
	v	$\frac{\text{pixels}}{\text{days}}$	18.429	22.455
	$-$ MI		-0.5889	-0.6571

Table 5.6: Comparison of parameter initializations, optimized parameters and resulting negative MI values.

of the domain. Based on the fractions $\frac{1}{2}, \frac{1}{3}$ and $\frac{3}{4}$ of the total number of pixels 387 per dimension, we estimate the related initial center coordinates. For the spreading time points, we apply estimates based on the time frames when we manually observe a spreading colony for the first front and also the change of texture in the colony for the second front. For well B4, the first front's time point is approximated with the fifth time frame whereas we apply the third one for well I11. We estimate the second front to be emerging two frames later than the initial front for both wells. Moreover, we stress that the values for the spreading time points $t_{0,n}$ and $t_{0,a}$ are given in *days*. The rational numbers are due to approximating the time span between the initial recording and the given time frame in days. The initial spreading velocity is based on the ratio between the approximated traveled distance of the first front ($\approx \frac{1}{3}$ of domain width in pixels) between the two given time points $t_{0,n}$ and $t_{0,a}$.
We apply these initializations to get a faster convergence of our numerical solver for the optimization problem. We include here a priori information on the spreading properties which we derived manually by visual inspection. When thinking about processing a larger data set, we suggest to use the same initializations for all wells, e.g. one could use initial parameter settings located in the center of our parameter space. Alternatively, one could implement a more sophisticated approach based on image segmentation for example to derive estimates of the different parameters for each well individually. As this increases the implementation and computation efforts for image pre-processing, we could also imagine to run the optimization for each well repeatedly by considering several different pre-selected initializations. The optimum is then based on the numerical solution corresponding to the smallest negative MI value. Since we are interested in this section in a proof of concept test of our optimization problem used to extract spreading properties for real data, we use the aforementioned hand-crafted initializations in the further course and do not include a deeper qualitative comparison of different initialization strategies.

Coming back to Table 5.6, we stress that we directly include the optimized parameters returned from the numerical experiment next to the parameter initializations for better comparability. We point out that we can observe the function value of the negative MI to drop significantly when comparing it for the initial parameter settings and the one for the optimized parameters for both wells. We emphasize that some of the initial guesses for the spreading properties are already matching quite well. For example, for both wells one of the spatial coordinates is already a good estimate. While the time points significantly improved during the optimization process, the spreading velocity seems to be already well initialized. Focusing on the improved time points for the emerging colonies, we observe that they are smaller than the initial guesses. Plus, the difference between both time points is smaller than in the initialization which corresponds to the second front moving closer behind the first one and, consequently, results in a thinner ring shaped area related to normal cell regions.
We point out that both optimization procedures stopped because the relative change in each spreading property is less than the step size tolerance of 10^{-14}. Our MI function and also the gradient term are only approximating the true, continuous functions. For this reason, we consider the optimized parameters to be also *approximating* the local optima of our continuous problem. Due to the discretization effects we may encounter small oscillations on a very fine scale of the function values and cannot assume a smooth function. In line with those discretizations, we state that the gradient term is only approximated with the given *discrete* convolution. Without assessing the small imperfections of our optimization function more closely, we state that the small oscillations only correspond to inter-pixel changes for the spatial parameters and also to differences of very small scale for the time dependent spreading parameters.

In Figures 5.20 and 5.21, we present the feature images compared to the resulting classification images based on the optimized spreading parameters for each well. In the first three columns, the different texture features are presented, i.e., one minus the local minima, local maxima and local interquartile ranges (cf. Section 3.3.2). In the fourth column, the classification images are calculated for the different time points based on the optimized parameters (cf. Table 5.6) and the circular spreading model introduced in Section 4.2. We stress here that we only include pixels *within* the cropping frame of a reference well and pixels outside this frame are assigned to 0 in the classification images matching areas without any present cell colony. The temporal development is observable comparing the images vertically as we present the images row-wise for the eight discrete time points (cf. Table 5.5).
For well B4 shown in Figure 5.20, the model captures a growing cell colony emerging close to the upper boundary of the well and spreading circularly through the domain. Similarly, the circular spreading is prominent for the cell population in well I11 depicted in Figure 5.21 where the colony emerges closer to the center position of the well. For the first well B4 the colony's border is captured well. For the second well I11, it seems like the colony area is estimated too large and as though the algorithm could not detect the border of the colony accurately enough. Here, the whole domain is covered with classification values near 2, i.e., is classified as abnormal cell regions in the final frame and even already in the previous frame almost the total area is covered with cell regions classified as abnormal. By manual investigation, we would expect the *abnormal* colony to not touch the left boundary of the well I11 for the last frames (cf. Figure 5.19). However, we can actually observe cells moving towards the left border when it comes to *normal* cell appearances already in the last two

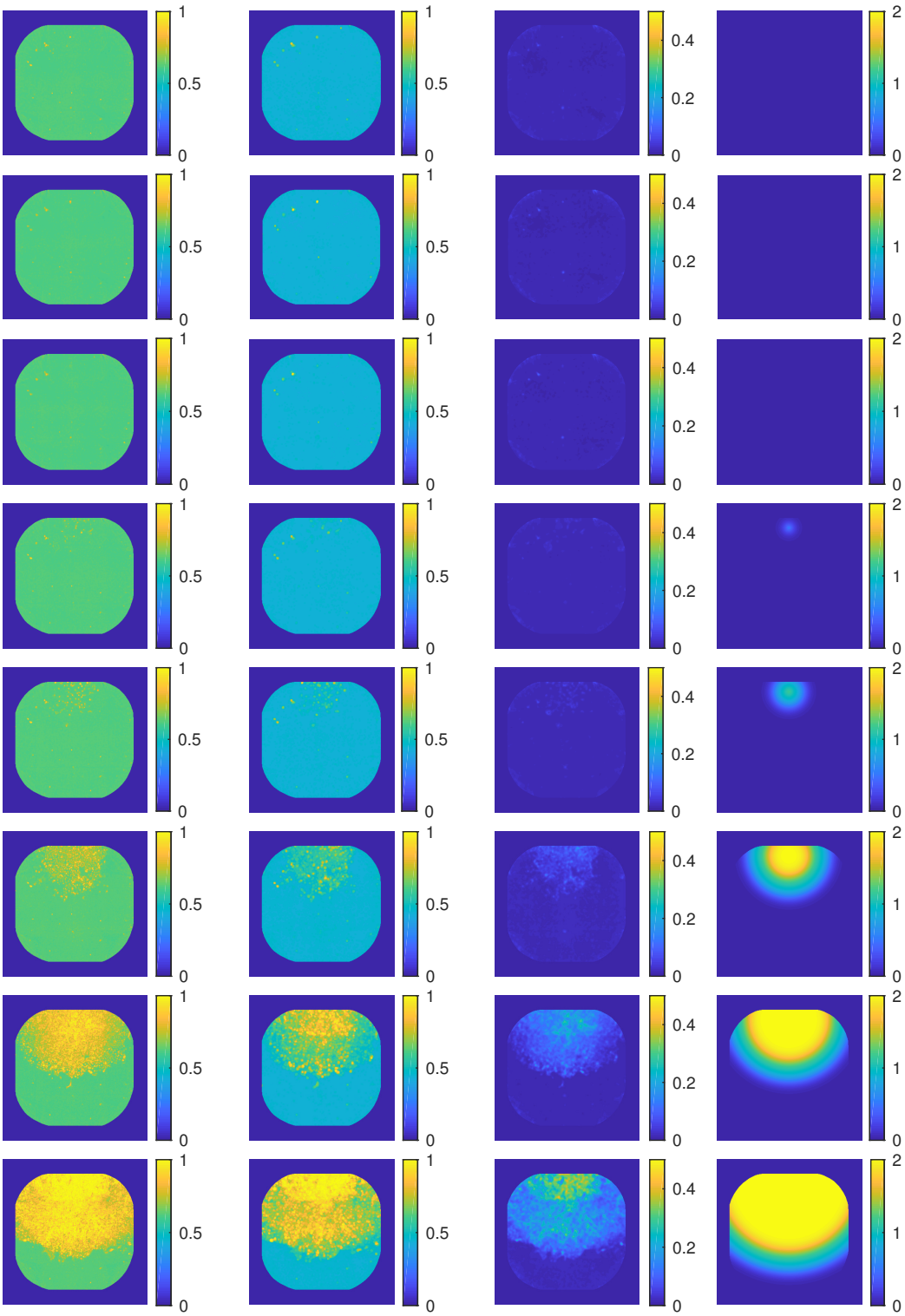


Figure 5.20: Features highlighting spreading in experimental data with estimated circular spreading for well B4. Each row corresponds to one time point. In the first three columns the features based on local texture information, i.e., (1 - local maxima), local minima and local interquartile ranges of occurring grayscale values in small neighborhoods, are presented while the optimized classification images are presented in the last column.

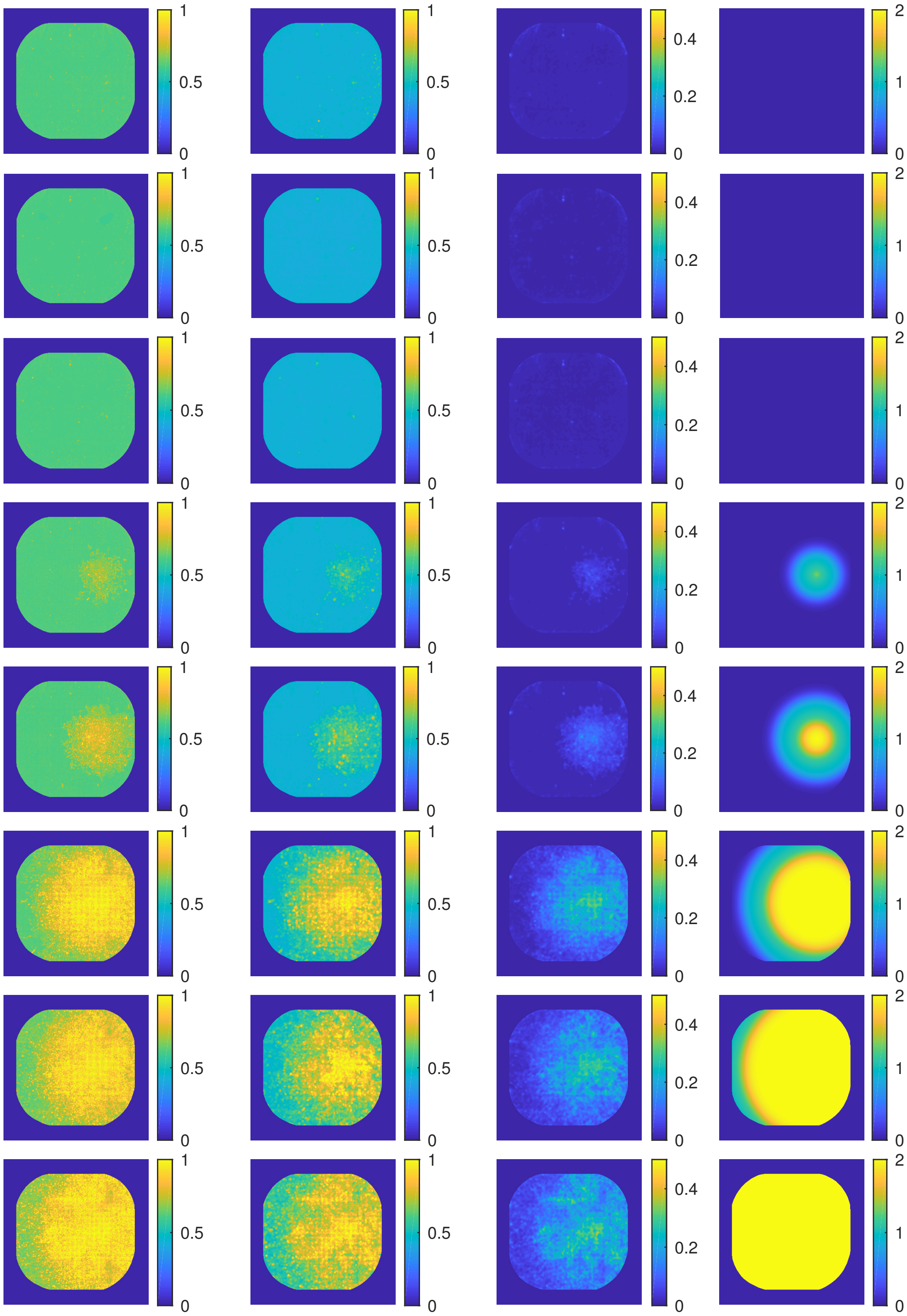


Figure 5.21: Features highlighting spreading in experimental data with estimated circular spreading for well I11. Each row corresponds to one time point. In the first three columns the features based on local texture information, i.e., (1 - local maxima), local minima and local interquartile ranges of occurring grayscale values in small neighborhoods, are presented while the optimized classification images are presented in the last column.

frames. We point out that, we observe changes in the texture features in the first three columns of the last two rows when comparing them to features related to background regions as present in the first three rows. Additionally, we refer to Figure 5.19 where we observe cells close to the left boundary in the original microscopy image for the seventh time frame. We stress that for both wells B4 and I11, it looks like the colony area for the second subcolony classified close to 2 is estimated too large. Actually, it seems as though the optimization approach fails to detect the inner bulk of the cell colony accurately which we consider to be related to the remarkable texture change. It seems like the algorithm approximates the inner *closed* part of the colony with classifications near 2 and the *transition regions* where only few cells rather than a dense colony are present is assigned with classifications near the subclass close to 1.
However, we do not focus on this shortcoming here further, since we are primarily interested in achieving a comparability between different wells rather than capturing the spreading process of each well individually. More precisely, we are interested in an approach for model fitting applied to the whole data set of time series for different wells and thus want to exemplarily apply the optimization approach to both wells simultaneously. Accordingly, we deal with the joint optimization of both example wells in the next section and introduce measures to compare the different classification results with respect to matching texture features.

5.5.2.3 Joint extraction of spreading properties for two example wells

In this section, we focus on the MI optimization when processing the two example wells B4 and I11 jointly. This is to highlight the approach of model fitting via MI-based optimization to gain a certain comparability between spreading colonies captured in time-lapse imaging of different wells. Considering the mutual information between feature data and classification images, we aim for a method to match texture characteristics of the different wells' time series.

We use again the intuitive parameter initializations based on manual assessments for the spreading properties related to well B4 and well I11 (cf. Table 5.6) for a first run. When applying the optimization to both wells simultaneously, we solve the problem directly for two sets of spreading properties. In a more general case when we are processing n wells, for example, we would solve the minimization problem for the five spreading characteristics $x_{0,1}$, $x_{0,2}$, $t_{0,n}$, $t_{0,a}$ and v as vectors of length n. This means that we are also including directly the feature data of both wells and the classification images corresponding to the current parameter settings of each well in the optimization. Again, we solve the corresponding optimization problem with the MATLAB solver *fmincon* [52] and apply the upper and lower bounds as well as the linear constraint introduced in Section 5.5.2.1.
In Table 5.7, we present the initial and the optimized parameter settings for the two example wells B4 and I11. We present in the upper part the results for the first test run when considering the naive, hand-crafted initializations. For comparison effects, we include a second run which is based on some prior knowledge. We recycle here the optimized parameter settings from the individual optimizations in the previous Section 5.5.2.2 (cf. Table 5.6). Concentrating on the lower part of the table for the second test run, we observe that the optimized parameters from the previous section are already good

test run	property	parameter initializations		optimized parameters	
1		**well B4**	**well I11**	**well B4**	**well I11**
	$x_{0,1}$	129	193.5	68.998	195.09
	$x_{0,2}$	193.5	290.25	192.94	263.28
	$t_{0,n}$	7.7083	0.70833	5.6658	2.5709
	$t_{0,a}$	14.625	7.7083	8.4333	5.7129
	v	18.651	18.429	17.825	22.258
	− MI	−0.4653		**−0.5687**	
2		**well B4**	**well I11**	**well B4**	**well I11**
	$x_{0,1}$	68.939	195.56	68.933	195.56
	$x_{0,2}$	194.89	263.05	194.89	263.05
	$t_{0,n}$	5.5563	2.749	5.6687	2.6364
	$t_{0,a}$	8.5106	5.8975	8.5656	5.8366
	v	17.973	22.455	17.898	22.499
	− MI	**−0.5681**		−0.5686	

Table 5.7: Comparison of parameter initializations, optimized parameters and resulting negative MI values. The first test run is related to hand-crafted parameter initializations whereas the second run considers the optimized parameter settings returned from the individual optimizations in Section 5.5.2.2.

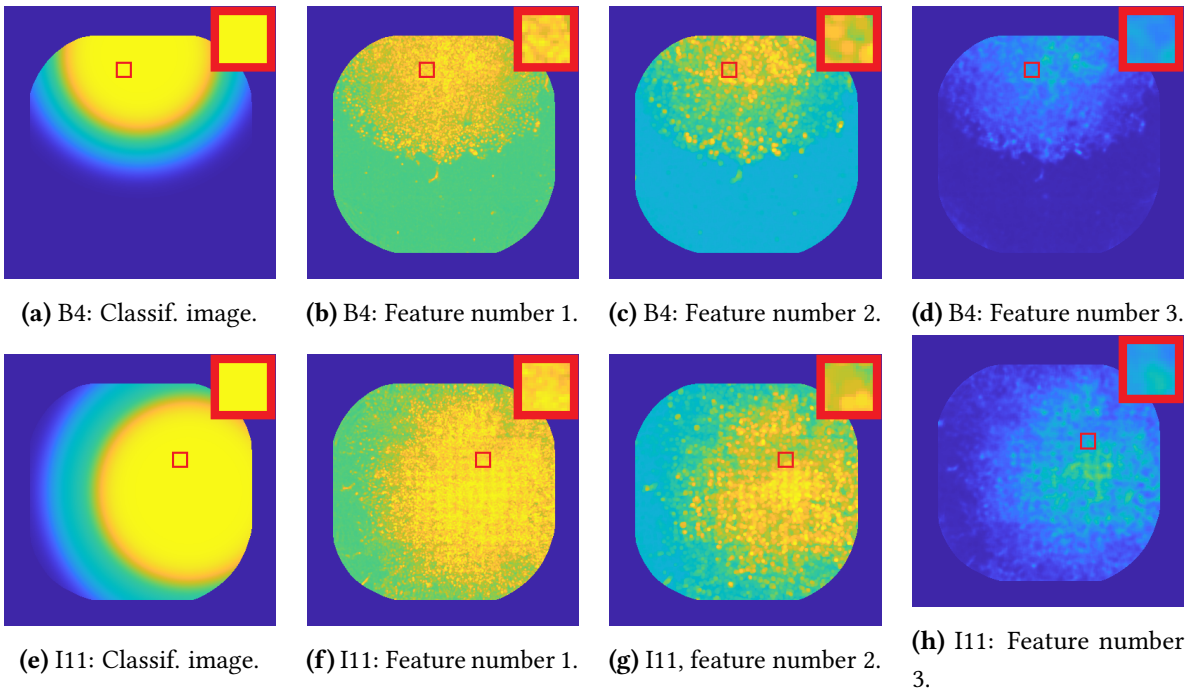
guesses for the optimized parameters when considering the time series of well B4 and I11 jointly. We only observed very few iterations for the solver to terminate and the different spreading properties vary only slightly when we compare the optimized settings for the two runs. As the first run with the intuitive parameter initializations results in a better negative MI value, i.e., its function value is smaller than the one of the second run, it is an open question why the second run does not converge further. However, the optimized parameter settings differ only very slightly when comparing the results for the two test runs in the right part of Table 5.7. We consider that this is due to many local extrema due to oscillation effects on a very fine scale.

In this context, we want to point out that we checked the course of the values of the minimization functional by zooming in and out near to the solution of the first run when moving in the positive and negative direction of the corresponding gradient. We observed oscillations on the smaller scales. To be more precise, we saw kinks or discontinuous jumps in the course of the function value of magnitudes between 10^{-5} and 10^{-6} when zooming in so that we could observe only one bin entry changing in the related joint histograms of the corresponding parameter settings. For the sake of completeness, we state that we scaled the gradient for this by 0.1 and the components of the gradient are by themselves already quite small for this possible local minimum with absolute values smaller than 10^{-4}. We do not present the oscillations of the functions visually or illustrate the changing bin entries for the corresponding joint histograms in more detail as this is not the main focus of our study.

In [62], the authors also present oscillation effects in the course of the optimization functional when

optimizing mutual information. While in their context the oscillations correspond to grid alignment structures when registering two medical images, we consider the oscillations in the course of our function to relate to approximations based on certain discretization aspects. As an example we state that this jumping effect of the MI function value occurs when we observe only *one* changing histogram entry. Since we use a discrete mollification kernel to smooth the joint histogram in the classification direction, we cannot expect a continuous approximation effect here. We are rather applying a discrete approximation which is naturally bound to certain approximation errors.
Without diving further into this analysis, we want to concentrate on the main theme of this section again, namely the comparison between the classified features for the texture point clouds of well B4 and well I11 when considering the separate optimization approach and the joint processing. To keep this analysis simple, we include from the joint processing only the results of the first test run based on the intuitive manual estimated parameter initialization as this run yielded the smaller negative MI value. In line with this, we start with the functional value as a first measure to compare the joint and the separate processing.
We consider the value of our optimization functional when processing both wells jointly. We take the optimized parameters from the *separate* processing in the previous Section 5.5.2.2. We then calculate the negative MI based on the features of both wells together with their classification images related to these spreading parameters. Actually, this value can be found in Table 5.7 as it is the function value corresponding to the initialization of the second run. We compare it with the negative MI value for the optimized spreading properties of the first run for the joint processing approach. Both values are highlighted in Table 5.7 with bold font. We observe that the function value based on the parameters corresponding to the separate run is slightly larger with -0.5681 than the function value corresponding to the optimized settings of the first run with -0.5687. This is already a small hint that the joint processing is the better approach to get matching classifications for the texture features in both wells as this is also reflected in the MI value based on the joint histograms of feature and classification data.

Comparing solely the function values in this context is not enough to get a profound idea if the texture features corresponding to the main classification values 0 (background regions), 1 (normal cell regions) and 2 (abnormal cell regions) are indeed matching better when applying the joint optimization. To motivate the comparison of the different texture regions when concentrating on a certain classification value in both wells, we present example images of well B4 and I11 at two discrete time points and show-case a small patch which is classified to be in an abnormal region, i.e., with classification value near 2. We use the classification images based on the optimized parameters of the first test run when processing both wells jointly as these parameters correspond to the smallest negative MI value.
In Figure 5.22, we present in the first row data related to well B4 at the seventh time point. In the second row, similar data is depicted for well I11 at the sixth time point. We choose these time points because here evolving colonies can be observed in the feature and classification images easily. Column-wise, we start with the classification images at the selected time points for both wells. The three feature images at the given time frames are presented in the next three columns. The images are only used for motivational aspects to present the idea of matching (texture) features across different wells and time points for regions of the same classification values. Later on, we include of course *all*



(a) B4: Classif. image. **(b)** B4: Feature number 1. **(c)** B4: Feature number 2. **(d)** B4: Feature number 3.

(e) I11: Classif. image. **(f)** I11: Feature number 1. **(g)** I11, feature number 2. **(h)** I11: Feature number 3.

Figure 5.22: Close-up of an example frame corresponding to classification values near 2 for well B4 at time point 7 and for well I11 at time point 6. The classification values are shown in scaled colors for the interval $[0, 2]$, the first two features in scaled colors for the interval $[0, 1]$ and the third feature in scaled colors for the interval $[0, 0.5]$.

time frames in the comparison analysis. We point out that the images for well B4 are already shown in Figure 5.20 in row 7 and the ones for well I11 in Figure 5.21 in row 6 with a slightly different order with the classification images in the last column. We refer to the former figures for the related color scaling and color bars.

We highlight in Figure 5.22 a small subpatch within the colony where the classification image reveals an abnormal region. This small window is arbitrarily chosen near the center part of the colony and serves as a motivation for the upcoming comparison analysis of similarly classified texture regions. The patch is marked with a red-frame and we present an enlarged view in the top right corner of each image. The point of our texture analysis is to ensure that similar texture features are classified similarly for both wells. For this example, this means that the point clouds corresponding to the selected features in the red frames of both wells need to be "close" to each other in the three dimensional feature space. Based on such point clouds in the feature space, we introduce in the further course distance measures to get a precise notion of this "closeness". To illustrate this, we start with the corresponding point clouds for the small red frames in Figure 5.23. In the first subplot, we plot exactly those features in the red-frames which are classified as abnormal colony area. For well B4, we draw the cloud with orange points whereas the point cloud for well I11 is shown in green. In the second subplot, we show the point clouds for *all* features classified close to 1 for the two example time frames for well B4 and I11.

In Figure 5.24, we visualize the feature images of well B4 (first row) and well I11 (second row) for the selected time points when focusing on a small classification range near 1. We observe that the

Figure 5.23: Comparison of point clouds corresponding to selected features for the example time points of well B4 and I11. On the left, features related to the red-framed windows in Figure 5.22 are presented and on the right, much denser point clouds of all features classified near 1 are shown (cf. Figure 5.24).

related point clouds in the right subplot of Figure 5.23 naturally contain many more sample points than the clouds of the first subplot which impedes visual inspection and manual comparisons. The related point clouds to features classified close to 2 in the selected frames are even more dense when we do not apply the restriction to the red frame.

In Figure 5.25, we present the feature images of well B4 and I11 for the selected time points. Here, we concentrate only on the features classified close to 2. The subpatches framed in red and introduced in Figure 5.22 are *within* the presented circular areas related to abnormal cell regions. As the pixel area of the selected subdomains for well B4 and I11 are greater than the subregions to normal cell classification close to 1 presented in Figure 5.24, it is a valid conclusion to expect many more sample points in the related point clouds.

Comparing Figure 5.24 and Figure 5.25, we observe that also for the joint processing, it seems like our approach does not accurately enough detect two different subpopulations. It rather distinguishes an inner colony part and a transition region where cells are located more separated from each other. This effect is very prominently observable in Figure 5.24, where subdomains related to features classified near 1 are shown. We recall that the *width* of this detected transition area — or area classified near 1 — is expected to be constant due to the constant spreading velocities we consider in our model assumptions.

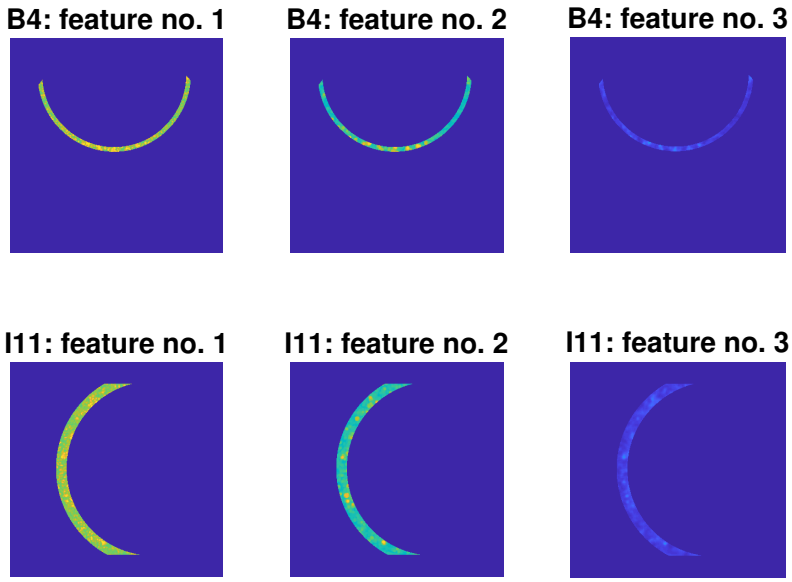


Figure 5.24: Features of well B4 and I11 which are classified close to 1, i.e., which are less than $\frac{\Delta c}{4}$ apart from 1, for two selected example time points as in Figure 5.22. The first two features are shown in scaled colors for the interval $[0, 1]$ and the third feature in scaled colors for the interval $[0, 0.5]$.

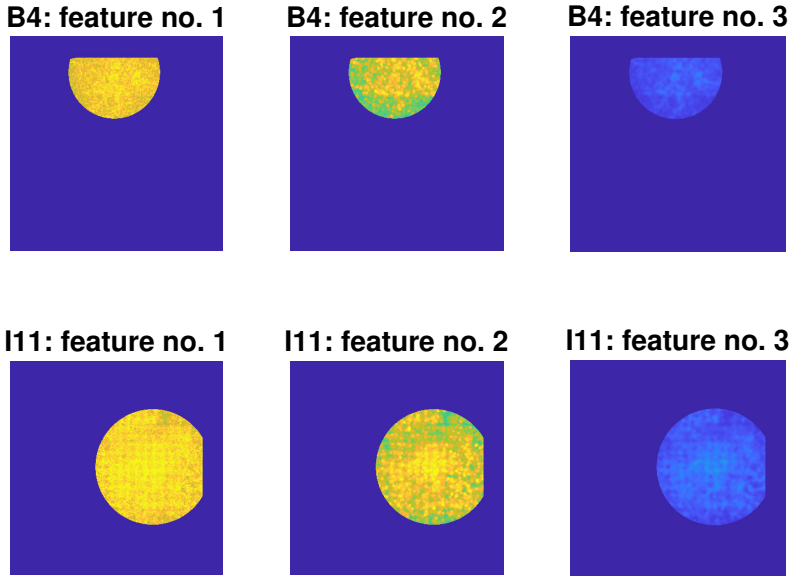


Figure 5.25: Features of well B4 and I11 which are classified close to 2, i.e., which are less than $\frac{\Delta c}{4}$ apart from 2, for two selected example time points as in Figure 5.22. The first two features are shown in scaled colors for the interval $[0, 1]$ and the third feature in scaled colors for the interval $[0, 0.5]$.

Coming back to the feature point clouds, we point out that it is hard to measure the density of such clouds or even compare the clouds for the different wells by visual inspections. When we include all time frames, we naturally expect even more sample points in the clouds and also a changing density which further impedes manual assessments. This emphasizes the need for special metrics to compare the point clouds. We aim for distance metrics measuring the difference between the point clouds of well B4 and I11 when applying on the one hand the separately optimized parameter settings and on the other hand the jointly optimized spreading properties. For this purpose, we start with the Hausdorff measure to compare the feature point clouds that correspond to classifications near the main classes. Before we introduce this distance measure, we first define our notion of features classified "close" to a certain value of our main classes.

Definition 5.108 (Feature point clouds classified near main classes)
We define feature point clouds based on the related classification values as follows: We introduce the main classes in the classification space $\boldsymbol{\mathcal{C}}$ as

$$C_{\text{main}} := \{0, 1, 2\}$$

which correspond originally to the interpretations of background areas, normal cell regions and abnormal cell regions. Based on theses, we define classification ranges on three scales and close to the main classes by setting

$$C_{c_m,s} := \left[c_m - s \cdot \Delta c, c_m + s \cdot \Delta c\right] \qquad \text{centered at } c_m \in C_{\text{main}}$$

with scale values $s = 1, \frac{1}{2}, \frac{1}{4}$ and Δc being the classification bin width again. For each main class $c_m \in C_{\text{main}}$, this results in three intervals centered at the main class and of width $2s\Delta c$ with $s = 1, \frac{1}{2}, \frac{1}{4}$.
Moreover, we introduce point clouds in the feature space by

$$P_{(c_m,s),\,\text{well}} := \{I_1(\boldsymbol{x}, t) \in \boldsymbol{\mathcal{F}} \,|\, I_2(\boldsymbol{p}, \boldsymbol{x}, t) \in C_{c_m,s}\}$$

with $(\boldsymbol{x}, t)$ in the spatio-temporal domain Ω_T, I_1 being the feature image in the feature space $\boldsymbol{\mathcal{F}}$, I_2 the classification image living in the classification space $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{p}$ denoting the spreading properties for the current well.

For a brief recapitulation of the feature and classification images, we refer the reader to Definitions 3.2 and 4.4. With the definition for a point cloud at hand, we get nine feature point clouds in total for well B4 and well I11 each. More precisely, we get for each main class of 0, 1 and 2 three clouds corresponding to the different scale ranges. We are mostly interested in features that are classified identically to the main classes. However, as we are using the approximation of the classification function based on a smoothed Heaviside function, we include small ranges around these main classes. To account for effects based on the choice of the interval widths, we include three different scale ranges: a *larger* widths of twice the classification tolerance size Δc for the interval, a *medium* width of exactly the bin width Δc and a *small* width of only one half the classification tolerance Δc. Depending on this width, we neglect certain features in the point cloud comparisons which are located in transition regions between the main classes and the selected classification ranges.

In preparation of the point cloud comparisons, we introduce the Mahalanobis distance. Our definition is inspired by equation (10) in [16].

Definition 5.109 (Mahalanobis distance)
For a feature $f \in \mathcal{F}$ and a point cloud P in the feature space, we define the Mahalanobis distance between the feature and the point cloud by

$$d_M(f, P) := \sqrt{(f - \bar{g})\, S^{-1}\, (f - \bar{g})^T}$$

where S denotes the covariance matrix and $\bar{g} = \frac{1}{|P|} \sum_{f \in P} f$ is the mean value of the point cloud P with cardinality $|P|$.

In our context, we measure the distance between a feature $f \in \mathcal{F}$ and a point cloud defined as in Definition 5.108. To compare clouds near the main classes for the separate and joint processing approach, we start with the Hausdorff distance as a first distance criterion. We introduce this measure oriented on Definition 2 in [36] and use the Mahalanobis distance as the inner distance measure.

Definition 5.110 (Hausdorff distance for point clouds)
We introduce the Hausdorff distance for two point clouds P_1, P_2 in the feature space $\mathcal{F}$ as

$$D_H(P_1, P_2) := \max\left\{\max\left\{d_M(f, P_2) \mid f \in P_1\right\}, \max\left\{d_M(g, P_1) \mid g \in P_2\right\}\right\}$$

with d_M denoting the Mahalanobis distance of the feature $f \in P_1$ to the point cloud P_2 and between $g \in P_2$ and the point cloud P_1, respectively.

classification range	width of range	optimization for wells	$D_H\left(P_{(c_m,s),\,\mathrm{B4}}, P_{(c_m,s),\,\mathrm{I11}}\right)$		
			$c_m = 0$	$c_m = 1$	$c_m = 2$
$C_{c_m,s=1}$	$\Delta c = \frac{1}{4}$	separate	50.2107	11.4709	6.4483
	(large)	joint	51.3464	11.8405	6.4702
$C_{c_m,s=\frac{1}{2}}$	$\frac{1}{2}\Delta c = \frac{1}{8}$	separate	52.8624	10.7763	6.3944
	(medium)	joint	53.2339	10.6472	6.4138
$C_{c_m,s=\frac{1}{4}}$	$\frac{1}{4}\Delta c = \frac{1}{16}$	separate	53.4131	8.4489	6.3499
	(small)	joint	53.8514	8.5840	6.3657

Table 5.8: Hausdorff distances comparing joint feature point clouds of well B4 and well I11 of the separate wells' optimization and the joint optimization run. Point clouds are generated for three different classification ranges centered at the main classification values 0, 1 and 2.

We show the Hausdorff distances between the feature clouds of well B4 and well I11 for the separate and the joint processing approach in Table 5.8. We concentrate on features corresponding to the previously introduced classification ranges near the main classes (cf. Definition 5.108). We observe that the distances for the classification ranges related to the background area (main class 0) are significantly higher than for the other two classes with values larger than 50 while for the normal cell

regions the distances are approximately 10 and for the abnormal cells are about 6. This is due to more outlier features for the background regions. This measure supports the fact that the inner bulk of a cell colony is detected with highest accuracy because the measure is here smallest. We stress that in this interpretation, we consider classification 2 to relate to the inner bulk and classification 1 to mark transition regions where only few cells are present but are more detached from each other rather than identifying normal and abnormal cell regions with the classification values 1 and 2. Recalling Figures 5.20 and 5.21, we observe that this interpretation might be more appropriate as interpreting the second class as abnormal cells. We point out that if we had performed the classification manually, we would have drawn a circle for texture features corresponding to abnormal regions much smaller than the one present in the previews based on the optimization results. Instead of focusing on this reflection on accurate classifications and model validation here, we return to the comparison of the distances between the point clouds when considering the two processing approaches.

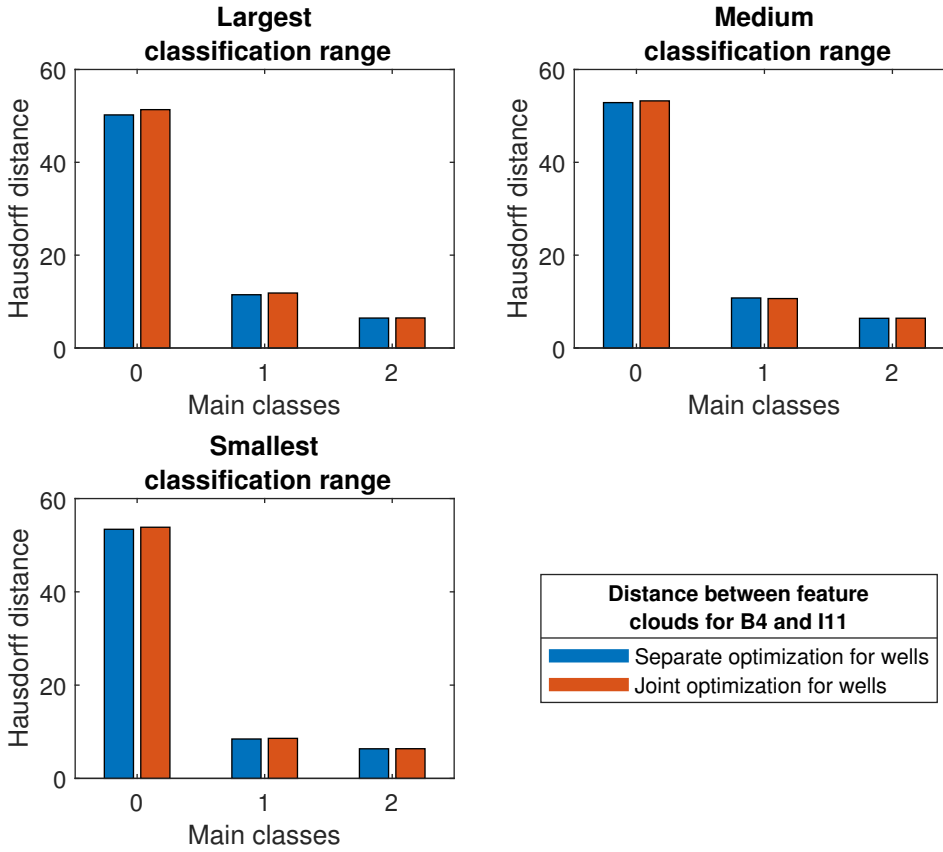


Figure 5.26: Comparison of feature point clouds corresponding to specific classification ranges using the Hausdorff distance. Bar plots for the different classification ranges highlight the small differences comparing the separate and joint processing approach.

For a better visual inspection of the distance measures comparing the two processing approaches, we present in Figure 5.26 bar plots reflecting the stated Hausdorff measures from Table 5.8. The three subplots correspond to the three different classification ranges. In each subplot, we compare the bars corresponding to the Hausdorff distance for the separate optimization approach (blue bars) with those for the joint processing (red bars). On the horizontal axis we show three subgroups to compare the distances for the three main classifications. The plots and also the corresponding values in Table 5.8

show that the differences for the main class 2 are marginal for all three classification ranges. Thus, we decide to neglect them in the evaluation of the different approaches. For the classification value 1, the differences are still small. However, what is more striking is that the bars for the largest and the smallest classification range for the joint processing (red) are exceeding the bars for the separate optimization (red). This trend is more prominent for the bars corresponding to the background class 0. Here, the red bar is for all three classification ranges exceeding the blue one. This is contrary to our expectation that the distance between the point clouds of well B4 and well I11 should be smaller for the joint processing compared to the separate processing. However, the differences turn out to be marginal in general in this comparison.

The Hausdorff measure was an initial distance measure to compare the feature point clouds. When dealing with real data, we can expect in each classification region outlier features as the present colony is not spreading accurately in a circular way. Since the concentric spreading model is merely an approximation of the growth process observable in the microscopy images, it is a natural conclusion that in this sense extracted texture features might get "misclassified" as the transition areas between different subclasses are not always particularly well pronounced. If we then consider the maximal values of the Mahalanobis distances as it is the case in the definition of the Hausdorff measure (cf. Definition 5.110), it follows directly that this measure preferably captures the distances of outlier features of one cloud compared to the other cloud.

As a remedy of this effect, we consider to use averaged Mahalanobis distances instead of the maximal Mahalanobis distances as implemented in the Hausdorff distance. Thus, we now introduce another distance measure based on the averaged Mahalanobis distance of one feature to the other point cloud to compare the point clouds again. The Chamfer distance includes exactly this effect when considering again the Mahalanobis distance as the inner distance measure. Our definition is inspired by the distance measure defined in equation (1) in [26]. In the referenced paper, the measure is used to compare features of one image with features of a template image. When adapting the measure to our context of feature point clouds, it reads

$$D_C\left(P_1, P_2\right) = \frac{1}{|P_2|} \sum_{f \in P_1} d_M\left(f, P_2\right)$$

with $|P_1|$ and $|P_2|$ denoting the cardinality of the point clouds P_1 and P_2. Compared to this original Chamfer distance, we introduce two modifications. Firstly, we use the median Mahalanobis distance instead of the averaged one to account even more for outlier features described above. Secondly, we introduce a certain symmetry in the distance measure by calculating this median Mahalanobis distance for features of point clouds for well B4 compared to the clouds corresponding to well I11 and vice versa. We apply this as a balancing effect for point clouds which may vary in their cardinality. This results finally in the following modified Chamfer measure:

Definition 5.111 (Modified Chamfer measure for point clouds)
We introduce a symmetric version of the Chamfer distance for two point clouds P_1, P_2 in the feature space $\mathcal{F}$ by using the median Mahalanobis distances as follows

$$D_C\left(P_1, P_2\right) := \frac{1}{2}\text{median}\left\{d_M\left(f, P_2\right) \mid f \in P_1\right\} + \frac{1}{2}\text{median}\left\{d_M\left(g, P_1\right) \mid g \in P_2\right\}$$

considering all features $f \in P_1$ in relation to point cloud P_2 and vice versa all features $g \in P_2$ in relation to point cloud P_1.

With this definition at hand, we once more calculate distance measures between the feature point clouds of well B4 and well I11 corresponding to specific classification ranges centered at the main classes. In Table 5.9, the distance measures are given for the different processing approaches (separate vs. joint processing). At first glance, we see that the distances are now of magnitude around 1 and are thus much smaller than the Hausdorff distances before. In particular, we observe that the value for the background region (main class $c_m = 0$) is now for each row the smallest. This is inline with our assumption that the background region corresponds to smooth texture regions without significant changes in the local maxima, minima or interquartile ranges. This highlights that this measure is more appropriate to compare our feature point clouds.

classification range	width of range	optimization for wells	$D_C\left(P_{(c_m,s),\,\mathrm{B4}}, P_{(c_m,s),\,\mathrm{I11}}\right)$		
			$c_m = 0$	$c_m = 1$	$c_m = 2$
$C_{c_m,1}$	$\Delta c = \frac{1}{4}$	separate	0.9182	1.3713	1.5596
	(large)	joint	0.9254	1.2949	1.5604
$C_{c_m,\frac{1}{2}}$	$\frac{1}{2}\Delta c = \frac{1}{8}$	separate	0.9329	1.4371	1.5764
	(medium)	joint	0.9389	1.3212	1.5764
$C_{c_m,\frac{1}{4}}$	$\frac{1}{4}\Delta c = \frac{1}{16}$	separate	0.9425	1.4665	1.5965
	(small)	joint	0.9501	1.3656	1.5953

Table 5.9: Modified Chamfer distances comparing joint feature point clouds of well B4 and well I11 of the separate wells' optimization and the joint optimization run. Point clouds are generated for three different classification ranges centered at the main classification values 0, 1 and 2.

Based on the listed distance values in Table 5.9 and the plots in Figure 5.27, we observe that for all three classification ranges the distance measures do not vary much for the background class 0 and for the abnormal classifications near 2. Only for the classifications close to 1, we observe bigger differences in the bars when comparing the separate approach (blue bars) with the joint processing (red bars). Even more, we see that indeed the distances between the point clouds of well B4 and well I11 differ *less* for the joint processing approach. This is in line with our expectation that it is better to process the wells jointly when aiming for similar texture regions of both wells to be classified alike.

Finally, we point out that even with this modified Chamfer distance the difference between the two processing approaches when considering both wells individually or jointly is not very large. To support this, we draw the reader's attention again to the optimized parameter settings. We recapitulate the optimized spreading properties for the separate and joint approach used in the above comparisons in Table 5.10. For each well's spreading properties, we observe only small differences when comparing the separate and the joint processing approach. The largest difference is observable for the second spatial coordinate of well B4 for which we observe a shifting effect of about two pixels. All other spatial coordinates are differing less than half a pixel. Also the starting time points $t_{0,n}$ and

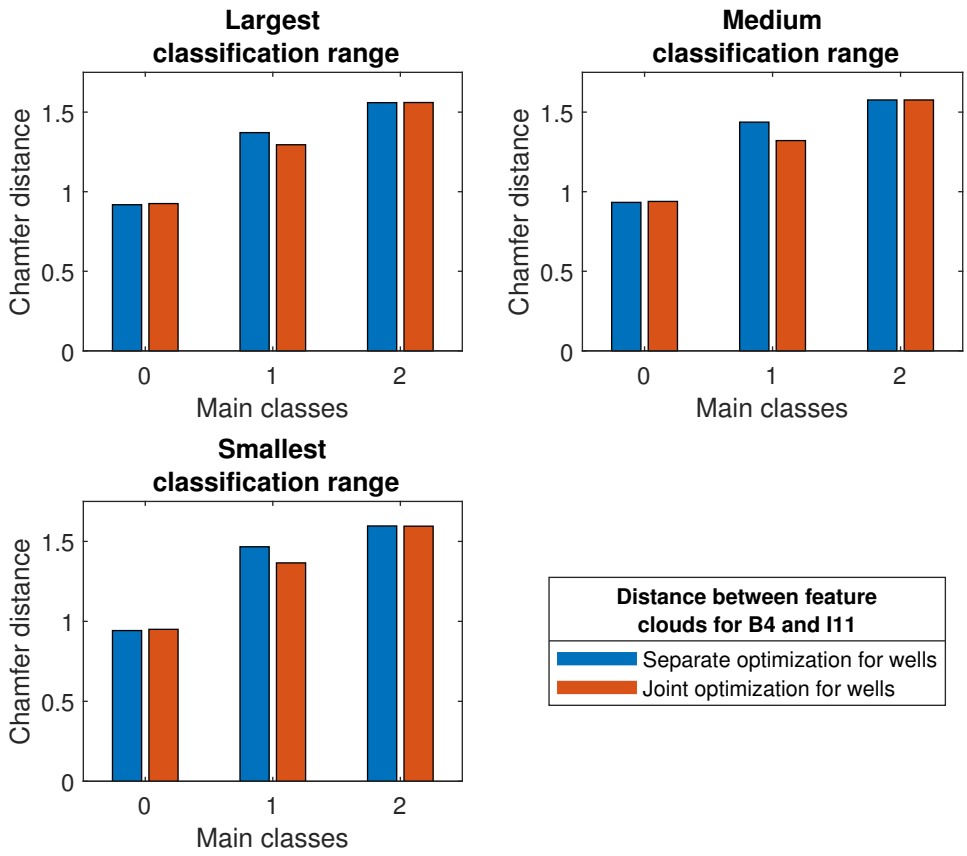


Figure 5.27: Comparison of feature point clouds corresponding to specific classification ranges using the Chamfer distance. Bar plots for the different classification ranges highlight that the differences between point clouds are most prominent for the classification ranges close to 1 when comparing the separate and the joint processing approach.

$t_{0,a}$ are differing less than 0.2 days which is indeed small as well when considering the spreading velocities of about 18 pixels per day for well B4 and approximately 22 pixels per day for well I11. The spreading velocities themselves differ less than 0.2 pixels per day when comparing the different optimization approaches. All in all, this shows that the extraction of spreading properties is working already quite robustly when considering only one well's time series. We justify this based on the fact

property	separate processing		joint processing	
	well B4	**well I11**	**well B4**	**well I11**
$x_{0,1}$	68.939	195.56	68.998	195.09
$x_{0,2}$	194.89	263.05	192.94	263.28
$t_{0,n}$	5.5563	2.749	5.6658	2.5709
$t_{0,a}$	8.5106	5.8975	8.4333	5.7129
v	17.973	22.455	17.825	22.258

Table 5.10: Optimized parameter settings when considering the separate processing and the joint processing of the two example wells B4 and I11.

that we have already for each individual well a huge amount of feature vectors available to estimate the spreading properties. In fact, we have for each time frame a pixel grid of 387×387 pixels where we extract $88,076$ pixels which are located within the cropping frame of a reference well domain (cf. Figure 5.28). Considering then eight time frames for each well, we have $704,608$ feature vectors at hand for the MI-based optimization per well's time series. Because this results already in a huge data set for optimizing the spreading parameters for each well, we guess that the difference between matching feature point clouds would be more pronounced if we had far less features at hand, e.g., only two time frames per well, for the optimization process.

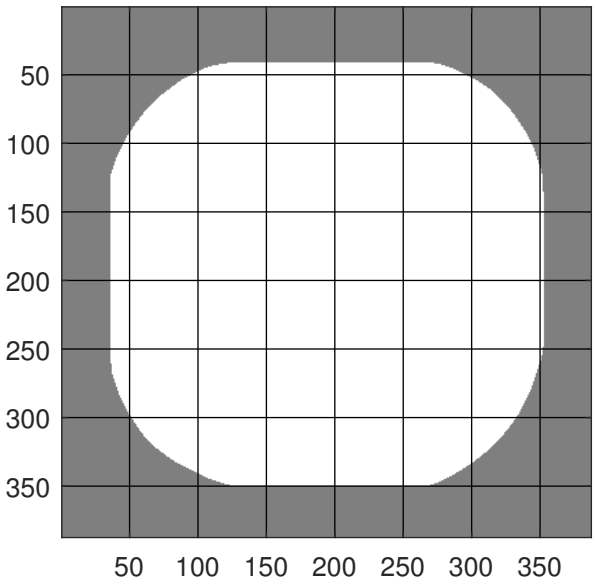


Figure 5.28: Visualization of downsampled cropping frame with the white region corresponding to the area *within* the well's frame and the *background* area shown in gray. Instead of highlighting each pixel individually, only intermediate grid lines are presented.

In the end, we comment on the fact that the differences of the measured Chamfer distances for the point cloud comparison are most prominently observable for the classification ranges close to 1 (cf. Figure 5.27). In Figures 5.24 and 5.25, we observe that the moving colony front of classifications close to 1 is rather thin compared to the inner colony bulk classified close to 2. In these figures, we of course only show one example time point for each well. However, as we use a constant spreading velocity in our model, we infer that in total we expect less features to be classified closed to 1 than to the other two main classes 0 and 2 for the example wells B4 and I11. Exemplarily, we state the number of features for the point clouds of the different classification ranges for the spreading based on the optimal parameter settings for the joint processing of the wells in Table 5.11. For each scale of the classification ranges, the number of feature vectors considered to be classified close to 1 is significantly smaller than for the other two classes. When now considering only small changes of the spreading parameters as for example by applying the spreading based on the optimized properties from the individual runs, this results in changes of the point clouds as well. These small changes in the point clouds result in more significant changes of the distance measure based on the median Mahalanobis distances for the class 1 related to the significantly smaller feature clouds. This is because the greater point clouds for the other two classes are considered to be more robust for small shifting

effects in the underlying spreading parameters and possible outlier features. Eventually, when only few pixels are changing the classification ranges, the variance of smaller point clouds would be more affected than the one for larger point clouds.

classification range	width of range	well	main classes		
			0	1	2
$C_{c_m,1}$	$\Delta c = \frac{1}{4}$	B4	535, 540	18, 696	66, 902
	(large)	I11	370011	39, 885	197, 718
$C_{c_m,\frac{1}{2}}$	$\frac{1}{2}\Delta c = \frac{1}{8}$	B4	52, 4139	9, 740	60, 028
	(medium)	I11	359, 541	23, 679	189, 657
$C_{c_m,\frac{1}{4}}$	$\frac{1}{4}\Delta c = \frac{1}{16}$	B4	513, 118	4, 918	53, 703
	(small)	I11	350, 739	12, 761	182, 260

Table 5.11: Number of feature vectors in point clouds for wells B4 and I11 when considering the different classification ranges.

To summarize our findings from the comparison of the joint and separate processing approaches, we point out that in the end the results do not differ a lot. It turns out that the extracted spreading properties for both example wells are quite robust when comparing both approaches with respect to the extracted spreading parameters, their MI values and when considering the point cloud analysis. We traced this back to having several time frames available which leads to a high number of classification and feature value combinations to generate the joint histograms from. This results in our recommendation to prefer the separate processing approach over the joint one.
Our analysis for the real AstraZeneca data revealed a shortcoming of our applied optimization problem. The maximization of the mutual information between our classification images based on a concentric spreading model and the generated feature images based on simplified texture characteristics failed to detect reliably the second subcolony considered to represent significant cellular texture changes within the colony. It is an open research question to improve for example the feature images to facilitate the differentiation between different subcolonies. Still, we emphasize that we can already derive spreading properties capturing the *total* cell colony's growth process per well with respect to estimates for its spreading velocity and its spatial and temporal origin when assuming a concentric spreading phenomenon. This highlights the major benefit of our approach for biologists in the pharmaceutical field.
For a final reflection on the given model in combination with the AstraZeneca data, we conclude the main insights of this Section 5.5.2 in the next subsection.

5.5.2.4 Conclusion of model fitting for AstraZeneca data

For a conclusion on the whole Section 5.5.2 on MI-based optimization for AstraZeneca data to determine spreading properties for a concentric growth model based on texture information, we highlight

the most important aspects again.
In the previous experiments on extracting spreading information for two example wells, the optimization for each well individually already resulted in robust properties differing only slightly from spreading information when processing both wells jointly. This is why we suggest to handle each well individually which allows to include parallel well processing for improving computation times. In addition, we applied the joint approach only on *two* example wells. To process the whole data set, considerably more memory capacity would be required and the resulting optimization problem would grow significantly. For clarification, we recall that on each plate we could monitor 24×16, i.e., in total 384, wells to capture spreading colonies. With the concentric spreading model, we are optimizing for five spreading parameters per well. Consequently, if each well is occupied by a cell colony, the joint optimization approach for a whole well plate results in an optimization for $5 \times 384 = 1920$ parameters. Considering then again almost $90,000$ pixels for each time point within the down-sampled wells domain, the data to be processed indeed needs a lot of memory space and we suggest to test an optimization on this scale on a high performance cluster. In the end, we recommend for future studies to apply the individual approach as long as we have several time frames in which the growing colony is visible.
Moreover, we underline that we originally introduced the simplified concentric spreading model for two different subcolonies aiming for detecting an inner bulk of cells with significantly differing texture properties. As it turns out, our optimization does not capture accurately enough this region. Rather than determining a normal cell population and another one with differing texture features, it looks like our model succeeds to capture a circular colony region which is classified with 2 and the other cell classification close to 1 corresponds to a transition region between the cell colony and the background area. Of course, one could consider now a model extension or think about strategies to include prior knowledge to ensure that the optimization indeed detects the prominent texture regions occurring near the heart of some colonies and classifies them close to 2 and also preserves another subcolony classified with approximately 1 which captures cells with "normal" appearance. However, this lies beyond the scope of this thesis and might rather serve as a starting point for future work.
Furthermore, the concentric spreading model might not well enough capture the original spreading of the investigated cancer cell populations. We stress that we applied a simplified spreading model to test the MI-based model fitting approach in our studies. In our simple model, we enforce that both subcolonies spread with the same and constant spreading velocity and also have the same origin. Of course, it is possible that the origins differ. In an improved concentric model, one could ask for the origin of the second colony to be within the first subcolony and also time-dependent spreading velocities might be possible or at least the consideration of two differing constant velocities might be beneficial. When considering two fixed but not necessarily identical spreading velocities of the two different colony fronts, one could ask for the second colony to be spreading slower than the first one to prevent the second front from overtaking the first one.
Of course, many different adjustments are possible to extend the simple concentric spreading model. Instead of arguing for more improvements, we remind the reader on the originally introduced PDE model capturing two different subcolonies by including diffusion and reaction terms (cf. Section 4.1). So finally, we recommend to test model fitting based on the given texture data by considering a more advanced model like this PDE model. Similar to the classification images based on adding two

concentric spreading areas via an approximated Heaviside function, we suggest to generate similar classification images by combining the concentrations of the different subcolonies. The MI-based optimization should then lead to more sophisticated spreading information on diffusivity of “normal” cells as well as reproduction and mortality rates to capture the transition between “background“ vs. “normal cells” and “normal cells” vs. “abnormal cells”. Without diving deeper into this topic, we state that for a numerical solution of the system of PDEs one would need to discretize the occurring temporal and spatial derivatives of the different cell concentrations first, before aiming for a solution of the more advanced spreading properties like the diffusivity as well as the mortality and reproduction rates.

With this we end the chapter on MI-based model fitting. In the upcoming final chapter of this thesis, we summarize the findings of this work. Additionally, we briefly comment on remaining challenges and give an outlook on possible future studies.

6

Conclusion & Outlook

In the final chapter of this dissertation, we reflect on the applied optimization approach for extracting spreading properties of developing cell colonies from microscopy data. Before we discuss some difficulties and challenges related to our MI-based model fitting, we highlight particular benefits and the potential we see in the developed concept.

We introduced a novel approach for model fitting by using texture information extracted from phase contrast images together with classification images based on an underlying model definition. By maximizing the mutual information between the feature images and classification images, we were able to extract the related spreading parameters numerically. In Section 5.5, we showcased the applicability of our approach for a toy example and also for real world data provided by AstraZeneca. For the second case, we also observed that our approach proved to be robust enough to match similar texture regions when comparing different wells without requiring joint processing and optimization. We link this effect to the fact that we already have eight discrete time points for each well which results in a great deal of feature vectors and related classification values when considering almost $90,000$ pixels in the down-sampled well domains per time stamp (cf. Section 5.5.2.4). As long as we can assume that the colony is present in several time points, we consider the MI-based approach to already work well for an individual well without considering joint processing of several wells' time series. We stress that this comparability was given at least for the two selected example wells B4 and I11. As tumor cells by themselves are already quite different to each other, we do not claim that this comparability is true for all wells in the data set. If the spreading phenomenon differs a lot from the example wells, or more precisely, if the extracted texture features are quite different, this inter-well comparability is not necessarily given. However, in these cases the joint model fitting could be more beneficial to ensure that similar texture regions are classified alike. Again, we point out that we chose two arbitrary wells for which colony growth was manually observed and which also reveal a pronounced change in texture towards the later time points and in the center part of the colony. As a first proof of concept for the introduced model fitting approach for real data, we consider this as a valid baseline.

As already pointed out in the conclusion of the previous chapter, our approach does not reliably identify the expected second subcolony with the striking texture changes. In this context, we want to elaborate on possible improvements.
Firstly, we suggest to incorporate the model parameter ε_0 determining the steepness of the approximated Heaviside function (cf. Section 4.2) into the optimization. We think that it could be beneficial

to optimize this model parameter simultaneously with the other spreading parameters because with changing steepness we get wider transition areas between the main classification values which in turn facilitate the identification of transition areas in changing texture regions. Of course, in this setting it might also be important to allow two differently "steep" transition areas: One between background regions and the first cell colony's region ("normal" cells) versus another one between the normal cell colony and the second subcolony ("abnormal" cells). In this sense, we suggest for more flexibility to also allow varying spreading velocities for the two estimated colony fronts, plus even different colony origins. We already elaborated on this in the conclusion of the previous chapter. We recapitulate that it is questionable that both colony fronts indeed have the same spreading velocity and also the same origin. Consequently, we suggested to allow different spreading velocities and origins while still ensuring that the second front starts off *within* the first colony's area and also that its front travels *behind* the first one. In the same context, we mention the possibility of another model adaption by allowing more than two subcolonies — or even only one colony at all. The latter one could accelerate the numerical solution as long as our approach fails to detect reliably the inner bulk of cells with the remarkable texture properties. Even more, we stress that with a flexible steepness parameter ε_0, it would still be possible to track the transition areas as it is currently the case with the colony areas classified close to 1.
Instead of diving deeper into possible model extensions for the simplified model, we remind the reader that we initially introduced a system of partial differential equations to model the colony spreading based on two different cell concentrations. For the two cell groups related to the two concentrations, we considered specific spreading properties, e.g., that only the "normal" cells migrate through the spatial domain and could perform mitosis, i.e., divide into daughter cells. In contrast to this cell group, the other one for the "abnormal" cells was considered to be immobile and could not perform mitosis. The only reaction term considered for this subgroup was based on the idea that they only develop due to a kind of mortality term for the first subcolony. So this reaction term modeled the transition between the two subgroups. From a biological perspective, it would be even better to solve an optimization problem revealing the related diffusion coefficient, mortality and proliferation rates in comparison to our simplified, naive concentric spreading model. This is why we suggest to concentrate on model fitting for such a PDE model in a further step to allow the extraction of more sophisticated spreading properties.

Before we elaborate on future studies following a different path, we emphasize that we performed a thorough and profound analysis of the considered optimization problem. We proved various convergence statements when considering different discretization stages we were facing when aiming for a numerical solution. Moreover, we proved statements dealing with the existence and the convergence of minimizers (cf. Theorems 5.97 and 5.98). Without these statements, we would not attach such importance to the numerical results and their significance would be questionable. Anyway, we want to point out the relevance of a thorough mathematical analysis. We highlight that a profound analysis would be required again if the optimization problem for extracting spreading information was changed fundamentally, e.g., by applying a different approach than the MI-based one.

In Section 3.3, we introduced the feature images used in our optimization approach and which are

based on local texture information. Here, we used quite basic texture features with the local minimal and maximal gray values as well as local interquartile ranges. In future studies, one could incorporate more advanced texture descriptors like for example local entropy values or even test features extracted by a convolutional neural network. Anyway, we suggest to include more biological input to validate interpretations related to different texture features. In the scope of this thesis, we only used a basic differentiation into "normal" and "abnormal" colony areas. However, we already indicated in Section 3.1 that additional imaging could facilitate a more profound analysis of the different colony areas. For this reason, we suggest to incorporate staining of cell nuclei to test cell counting based on the prominent nuclei in the new color channel. By this, we could validate the interpretation of significant changes in cell densities when we observe the changing texture in the microscopy images. If a second color channel was accessible, we could imagine a staining which is only expressed in living cells. With this further color channel, we could even differentiate between living cells and apoptotic cells and test the hypothesis of a bulk of cell debris in the center part of some colonies in which we encounter the distinct texture changes.

If we were to analyze the spreading phenomena of the cancer cells further and had another color channel available, we would suggest an enhanced two-step approach. First, we propose to generate a different version of a feature image by including the new insights based on the color channel. For example, we can imagine to generate a feature image based on nuclei counting and thresholding in the color channel images to incorporate information on *cell densities* in our optimization approach. In a second step, we would apply the MI-based model fitting again to optimize spreading characteristics related to a classification image again which is based on a mathematical model.

Since this approach, or more precisely, the first step in this approach would fundamentally change the experiment's design and lead to extensively more imaging as well as more pre-processing to segment and count cell nuclei in the new color channel, we recommend another improvement which only requires a new color channel to allow nuclei counting for *some* example colonies which contain the different texture regions. For these color channel images, we would perform again nuclei counting to estimate cell densities first. Then, we could use this information to train a neural network on estimating cell densities based on local texture regions. By using small image patches with local texture information in combination with the information on cell densities in the related patch from the color channel, we could provide our own training data and thereby overcome the current lack of appropriate, labeled training data. As a conclusion, we consider that having time-lapse color channel images for only a few wells at hand could already be enough to train the network for good cell density estimates as we are breaking down the images into very small patches anyway and, consequently, increasing the training data by this significantly. In the end, we could generate with the neural network new feature images which estimate local cell densities. We suggest to derive spreading properties with the help our our implemented MI-based model fitting approach in a next step by including these feature images.

Nowadays, machine learning and especially neural networks are getting more popular in many different research fields connected to imaging data. Testing a neural network-based approach for the given cell investigation problem is clearly out of the scope of this thesis. Still, we recommend to consider a joint approach of including a neural network *and* the MI-based model fitting as described in the last paragraph, eventually. With this we could incorporate the power of neural networks to

translate texture information to cell density estimates, plus include the mathematically profoundly analyzed MI-based optimization problem for the extraction of spreading characteristics of growing cell populations. As this seems as a quite promising approach, we suggest to definitely consider this in upcoming future studies.

All results of this thesis considered, we state that the MI-based model fitting proves to be a powerful approach to derive spreading information directly from imaging data. While we focused in our work on basic texture descriptors and a simplified spreading model to evaluate the novel approach, we see for both concepts potential future enhancements. We emphasize that we are already able to reveal new insights on the cell colony spreading with our framework as shown in the numerical test scenarios.

List of Symbols

Model and data parameters

$\mathcal{C} = \mathbb{R}$	classification space
$\mathcal{F} = \mathbb{R}^n$	feature space which is in our setting $\mathcal{F} = \mathbb{R}^3$
$\mathcal{F} \times \mathcal{C}$	joint feature-classification space
$\mathcal{F}'$	reduced feature space neglecting features related to very low probabilities when considering the feature image mapping influenced by noise I_1^{d}
$\mathcal{C}'$	reduced classification space which is a subset of $[0, 2]$ and which refers to $\mathrm{supp}\,(p_{\mathcal{C}})$ or, equivalently, to the image of the classification mapping I_2
Ω	spatial domain
L	length of spatial domain
W	width of spatial domain
$[0, T]$	temporal domain
$\{t_1, \ldots, t_{n_T}\}$	discrete time points
Ω_T	spatio-temporal domain
Δt	temporal step width
h	pixel width and height
Ω^{h}	discretized spatial domain
$\tilde{p}$, $\Omega_{\tilde{p}}$	pixel in discretized spatial domain
Ω_1	unit domain $(0, 1)$
Ω_{h}	scaled domain, $\mathrm{h} \cdot (0, 1)$
n_T	number of discrete time points in time interval
$n_{\tilde{p}}$	number of pixels in discretized spatial domain
I_1	feature image
I_1^{h}	discretized feature image living on pixel grid
$\tilde{I}_1$	feature image constantly extended for probability space
I_1^{d}	disturbed feature image
$\left(I_1^{\mathrm{d}}\right)^{-1}$	inverted disturbed feature image
$\hat{I}_1^{\mathrm{d,h}}$	discretized disturbed feature image (condensed in an array)
I_2	classification image
I_2^{h}	discretized classification image living on pixel grid
$\hat{I}_2^{\mathrm{h}}$	discretized classification image (condensed in an array)
$\boldsymbol{I}$	image mapping based on feature and classification images
$\boldsymbol{I}^{\mathrm{h}}$	discretized image mapping with feature and classification images based on a pixel grid of width h
v	speed of traveling wave front, spreading velocity

$t_{0,n}$	starting time point of spreading normal cells
$t_{0,a}$	starting time point of spreading abnormal cells
$\boldsymbol{x}_0$	colony center, starting location of spreading cells
$\boldsymbol{p}$	condensed model parameters $(\boldsymbol{x}_0, t_{0,n}, t_{0,a}, v)$
$\boldsymbol{p}_\varepsilon$	approximation of model parameters, e.g., used to denote a sequence converging to a parameter set $\boldsymbol{p}$ with $\boldsymbol{p}_\varepsilon = (\boldsymbol{x}_{0,\varepsilon}, t_{0,n,\varepsilon}, t_{0,a,\varepsilon}, v_\varepsilon)$
$\boldsymbol{P}$	parameter space
p_N	probability density function for Gaussian noise
P_N	probability measure for Gaussian normal distribution
P_{Ω_T}	uniform probability distribution related to spatio-temporal domain Ω_T
$p_{\mathcal{F}}$	probability density function for features
$P_{\mathcal{F}}$	probability measure in feature space $\mathcal{F}$
$P_{\mathcal{F}}^{\mathrm{d}}$	probability measure in feature space $\mathcal{F}$ considering corrupted feature images due to additive Gaussian noise
$p_{\mathcal{F}}^{\mathrm{d}}$	probability density function for occurring features corrupted by additive Gaussian noise
$p_{\mathcal{F}}^{\mathrm{d},\varepsilon}$	probability density function for occurring features corrupted by additive Gaussian noise when considering discretized setting
$p_{\mathcal{F}}^{\mathrm{d},\tilde{\varepsilon}}$	pointwise converging probability density function for occurring features corrupted by additive Gaussian noise when considering discretized setting
$p_{\mathcal{C}}$	probability density function for occurring classifications
$p_{\mathcal{C}}^{\varepsilon}$	probability density function for occurring classifications when considering discretized setting
$p_{\mathcal{C}}^{\tilde{\varepsilon}}$	pointwise converging probability density function for occurring classifications when considering discretized setting
$P_{\mathcal{C}}$	probability measure in classification space $\mathcal{C}$
$p_{\mathcal{F}\times\mathcal{C}}$	probability density function for occurring feature-classification combinations
$P_{\mathcal{F}\times\mathcal{C}}$	probability measure in joint space $\mathcal{F}\times\mathcal{C}$
$P_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}}$	probability measure in joint space $\mathcal{F}\times\mathcal{C}$ considering corrupted feature images due to additive Gaussian noise
$p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d}}$	probability density function for occurring feature-classification combinations considering corrupted feature images due to additive Gaussian noise
$p_{\mathcal{F}\times\mathcal{C}}^{\mathrm{d},\varepsilon}$	probability density function for occurring feature-classification combinations considering corrupted feature images due to additive Gaussian noise when considering discretized setting

$p^{\mathrm{d},\tilde{\varepsilon}}_{\mathcal{F}\times\mathcal{C}}$	pointwise converging probability density function for occurring feature-classification combinations considering corrupted feature images due to additive Gaussian noise when considering discretized setting
$H_{\mathcal{F}}$	histogram in the feature space $\mathcal{F}$
$H_{\mathcal{C}}$	histogram in the classification space $\mathcal{C}$
$H_{\mathcal{F}\times\mathcal{C}}$	histogram in the joint space $\mathcal{F}\times\mathcal{C}$
$H^{\mathrm{h}}_{\mathcal{F}}$	histogram in the feature space $\mathcal{F}$ for discretized feature images $I_1^{\mathrm{d,h}}$
$H^{\mathrm{h}}_{\mathcal{C}}$	histogram in the classification space $\mathcal{C}$ for discretized classification images I_2^{h}
$H^{\mathrm{h}}_{\mathcal{F}\times\mathcal{C}}$	histogram in the joint space $\mathcal{F}\times\mathcal{C}$ for discretized feature and classification images $I_1^{\mathrm{d,h}}$ and I_2^{h}
$H^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}$	smoothed histogram in the joint space $\mathcal{F}\times\mathcal{C}$
$\overline{H}_{\mathcal{F}}$	discrete histogram for features $\mathcal{F}$ as an array when considering binned space
$\overline{H}_{\mathcal{C}}$	discrete histogram for classifications $\mathcal{C}$ as an array when considering binned space
$\overline{H}_{\mathcal{F}\times\mathcal{C}}$	discrete histogram for feature-classification combinations $\mathcal{F}\times\mathcal{C}$ as an array when considering binned space
$\hat{H}_{\mathcal{F}}$	discrete histogram measure in the feature space $\mathcal{F}$ when considering binned space
$\hat{H}_{\mathcal{C}}$	discrete histogram measure in the classification space $\mathcal{C}$ when considering binned space
$\hat{H}_{\mathcal{F}\times\mathcal{C}}$	discrete histogram measure in the joint space $\mathcal{F}\times\mathcal{C}$ when considering binned space
$\hat{H}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}$	discrete histogram measure in the joint space $\mathcal{F}\times\mathcal{C}$ when considering binned space and with smoothing mollification along $\mathcal{C}$
$\hat{H}^{\mathrm{h}}_{\mathcal{F}\times\mathcal{C}}$	discrete histogram measure in the joint space $\mathcal{F}\times\mathcal{C}$ for discretized feature and classification images $I_1^{\mathrm{d,h}}$ and I_2^{h}
$\hat{H}^{\mathrm{h}}_{\mathcal{C}}$	discrete histogram measure in the classification space $\mathcal{C}$ for discretized classification images I_2^{h}
$\hat{H}^{\mathrm{h}}_{\mathcal{F}}$	discrete histogram measure in the feature space $\mathcal{F}$ for discretized feature images $I_1^{\mathrm{d,h}}$
$\hat{H}^{\mathrm{h},\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}$	smoothed discrete histogram measure in the joint space $\mathcal{F}\times\mathcal{C}$ for discretized feature and classification images $I_1^{\mathrm{d,h}}$ and I_2^{h}
$\hat{H}^{\mathrm{h},\varepsilon_1}_{\mathcal{C}}$	smoothed discrete histogram measure in the classification space $\mathcal{C}$ for discretized classification images I_2^{h}
$h_{\mathcal{F}}$	histogram density function in the feature space $\mathcal{F}$
$h_{\mathcal{C}}$	histogram density function in the classification space $\mathcal{C}$
$h_{\mathcal{F}\times\mathcal{C}}$	histogram density function in the joint space $\mathcal{F}\times\mathcal{C}$

$\hat{h}_{\mathcal{F}}$	piecewise constant discrete histogram density function in the feature space $\mathcal{F}$
$\hat{p}_{\mathcal{F}}$	piecewise constant probability density function in the feature space $\mathcal{F}$
$\hat{h}_{\mathcal{C}}$	piecewise constant discrete histogram density function in the classification space $\mathcal{C}$
$\hat{p}_{\mathcal{C}}$	piecewise constant probability histogram density function in the classification space $\mathcal{C}$
$\hat{h}_{\mathcal{F}\times\mathcal{C}}$	piecewise constant discrete histogram density function in the joint space $\mathcal{F}\times\mathcal{C}$
$\hat{p}_{\mathcal{F}\times\mathcal{C}}$	piecewise constant probability density function in the joint space $\mathcal{F}\times\mathcal{C}$
$\hat{h}^{\mathrm{h}}_{\mathcal{F}}$	piecewise constant discrete histogram density function in the feature space $\mathcal{F}$ for discretized feature images $I_1^{\mathrm{d,h}}$
$\hat{h}^{\mathrm{h}}_{\mathcal{C}}$	piecewise constant discrete histogram density function in the classification space $\mathcal{C}$ for discretized classification images I_2^{h}
$\hat{h}^{\mathrm{h}}_{\mathcal{F}\times\mathcal{C}}$	piecewise constant discrete histogram density function in the joint space $\mathcal{F}\times\mathcal{C}$ for discretized feature and classification images $I_1^{\mathrm{d,h}}$ and I_2^{h}
$h^{\mathrm{h}}_{\mathcal{F}}$	histogram density function in the feature space $\mathcal{F}$ for discretized feature images $I_1^{\mathrm{d,h}}$
$h^{\mathrm{h}}_{\mathcal{C}}$	histogram density function in the classification space $\mathcal{C}$ for discretized classification images I_2^{h}
$h^{\mathrm{h}}_{\mathcal{F}\times\mathcal{C}}$	histogram density function in the joint space $\mathcal{F}\times\mathcal{C}$ for discretized feature and classification images $I_1^{\mathrm{d,h}}$ and I_2^{h}
$\hat{h}^{\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}$	smoothed piecewise constant discrete histogram density function in the joint space $\mathcal{F}\times\mathcal{C}$
$\hat{h}^{\mathrm{h},\varepsilon_1}_{\mathcal{F}\times\mathcal{C}}$	smoothed piecewise constant discrete histogram density function in the joint space $\mathcal{F}\times\mathcal{C}$ for discretized feature and classification images $I_1^{\mathrm{d,h}}$ and I_2^{h}
$\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}}$	discrete histogram for features $\mathcal{F}$ as an array based on pixel counting
$\overset{\circ}{\boldsymbol{H}}_{\mathcal{C}}$	discrete histogram for features $\mathcal{C}$ as an array based on pixel counting
$\overset{\circ}{\boldsymbol{H}}_{\mathcal{F}\times\mathcal{C}}$	discrete histogram for features $\mathcal{F}\times\mathcal{C}$ as an array based on pixel counting
Δc	binning width for the classification space $\mathcal{C}$
Δf	binning width for the feature space $\mathcal{F}$
$\boldsymbol{r}_{\mathcal{C}}$	radius function living in the classification space $\mathcal{C}$
$\boldsymbol{r}_{\mathcal{F}}$	radius function living in the feature space $\mathcal{F}$
$\boldsymbol{c}$	function determining the classification value in $\mathcal{C}$ based on a radius

f	function determining the feature value in $\mathcal{F}$ based on a radius
$\tilde{c}$	function determining the classification value in $\mathcal{C}$ based on a feature value
$\tilde{f}$	function determining the feature value in $\mathcal{F}$ based on a classification value

Other symbols

$\cdot_\varepsilon$	subscript ε indicating a sequence or a disturbance of original parameter
$.^{\mathrm{h}}$	superscript h indicating a discretization based on a discrete step width h
$\mathbf{1}_{n\times n}$	identity matrix of dimenions $n \times n$
$\mathbf{1}_{\mathrm{set}}$	indicator function of a set which is denoted in the subscript
$\mathbb{R}_{>0}$	positive reel numbers
$\mathbb{R}_{+}$	non-negative reel numbers

Abbreviations

MI	mutual information
PDF(s)	probability density function(s)
NoMADS	Program on “Nonlocal Methods for Arbitrary Data Sources” funded by the European Union
A1,A2,...,P15	Alpha-numeric column- and row-wise labelling for different wells on a well plate
DNN	deep neural network
CNN	convolutional neural network
ODE(s)	ordinary differential equation(s)
PDE(s)	partial differential equation(s)
PDG	Partielle Differentialgleichung(en)
SIR-model	an epidemic model to describe subpopulations consisting of susceptible (S), infective (I) and recovered (R) people

List of Figures

List of Tables

Bibliography

[1] C. Aggarwal. *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing, 2018 (cited on page 42).

[2] H. W. Alt. *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung*. Springer-Verlag Berlin Heidelberg, 2006 (cited on pages 118, 138, 140).

[3] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Science Publications. Clarendon Press, 2000 (cited on page 100).

[4] L. Ambrosio, N. Gigli, and G. Savare. "Gradient Flows in Metric Spaces and in the Space of Probability Measures". In: (Jan. 2005) (cited on pages 30, 77).

[5] AstraZeneca UK Ltd. *AstraZeneca in the UK (Webpage)*. Accessed: 2020-06-03. URL: https://www.astrazeneca.co.uk/about-us.html (cited on page 7).

[6] K. B. Athreya and S. N. Lahiri. *Measure Theory and Probability Theory*. Springer-Verlag New York, 2006 (cited on page 78).

[7] M. Bergounioux. "On Poincare-Wirtinger inequalities in spaces of functions of bounded variation". In: *Control and Cybernetics* 40.4 (Jan. 2011), pages 921–929 (cited on page 101).

[8] L. von Bertalanffy. "Problems of Organic Growth". In: *Nature* (1949) (cited on page 47).

[9] M. H. Bharati, J. J. Liu, and J. F. MacGregor. "Image texture analysis: methods and comparisons". In: *Chemometrics and Intelligent Laboratory Systems* 72.1 (2004), pages 57–71 (cited on page 42).

[10] A. Braides. *Gamma-Convergence for Beginners*. Oxford University Press, July 2002 (cited on pages 139, 140, 143, 144).

[11] K. Bredies and D. Lorenz. *Mathematical Image Processing*. Birkhäuser Basel, 2018 (cited on pages 20, 25, 41, 61, 63).

[12] E.-M. Brinkmann. "Novel Aspects of Total Variation-Type Regularization in Imaging". PhD thesis. Aug. 2019 (cited on pages 25, 65, 100).

[13] J. Bucevičius, G. Lukinavičius, and R. Gerasimaitė. "The Use of Hoechst Dyes for DNA Staining and Beyond". In: *Chemosensors* 6.2 (2018) (cited on page 12).

[14] T. F. Chan and L. A. Vese. "Active contours without edges". In: *IEEE Transactions on Image Processing* 10.2 (2001), pages 266–277 (cited on page 15).

[15] B.-M. Chang, H.-H. Tsai, and C.-Y. Yen. "SVM-PSO Based Rotation-Invariant Image Texture Classification in SVD and DWT Domains". In: 52.C (June 2016), pages 96–107 (cited on page 43).

[16] R. De Maesschalck, D. Jouan-Rimbaud, and D. Massart. "The Mahalanobis distance". In: *Chemometrics and Intelligent Laboratory Systems* 50.1 (2000), pages 1–18 (cited on page 192).

[17] E. Dougherty. *An Introduction to Morphological Image Processing*. Books in the Spie Tutorial Texts Series. SPIE Optical Engineering Press, 1992 (cited on page 21).

[18] S. Du, Y. Yan, and Y. Ma. “Local spiking pattern and its application to rotation- and illumination-invariant texture classification”. In: *Optik* 127.16 (2016), pages 6583–6589 (cited on page 43).

[19] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. 2nd edition. New York: Wiley, 2001 (cited on page 20).

[20] T. E. Duncan. “On the Calculation of Mutual Information”. In: *SIAM Journal on Applied Mathematics* 19.1 (1970), pages 215–220 (cited on pages v, vii, 3, 68, 69).

[21] J. Elstrodt. *Maß- und Integrationstheorie*. 6th edition. Springer-Lehrbuch. Springer Berlin Heidelberg, 2009 (cited on page 75).

[22] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 1991 (cited on page 80).

[23] O. Forster. “Implizite Funktionen”. In: *Analysis 2: Differentialrechnung im IRn, gewöhnliche Differentialgleichungen*. Wiesbaden: Vieweg+Teubner, 2010, pages 90–103 (cited on pages 127, 130).

[24] O. Forster. *Analysis 3: Maß- und Integrationstheorie, Integralsätze im IRn und Anwendungen*. Springer Spektrum, 2017 (cited on page 164).

[25] F. Gaede. “Efficient variational graph methods in imaging and 3D data”. PhD thesis. Westfälische Wilhelms-Universität Münster, 2020 (cited on pages 57, 58).

[26] D. Gavrila. “Traffic Sign Recognition Revisited”. In: *Proc. of the 21st DAGM Symposium für Mustererkennung* (1999), pages 86–93 (cited on page 194).

[27] L. Gay, A.-M. Baker, and T. A. Graham. “Tumour Cell Heterogeneity”. In: *F1000Research* 5 (2016) (cited on page 9).

[28] J. Geiping et al. *Fast Convex Relaxations using Graph Discretizations*. 2020 (cited on page 58).

[29] A. Gerisch and M. A. J. Chaplain. “Mathematical modelling of cancer cell invasion of tissue: Local and non-local models and the effect of adhesion”. In: *Journal of Theoretical Biology* 250.4 (2008), pages 684–704 (cited on page 46).

[30] I. C. Ghiran. “Introduction to Fluorescence Microscopy”. In: *Light Microscopy*. Volume 689. Methods in Molecular Biology. Humana Press, Totowa, NJ, 2011, pages 93–136 (cited on page 12).

[31] R. C. Gonzalez, E. W. Woods, and S. L. Eddins. “Digital Image processing using MATLAB”. In: Prentice Hall, 2003. Chapter 11 (cited on page 41).

[32] J. S. Grah et al. “Mathematical imaging methods for mitosis analysis in live-cell phase contrast microscopy”. In: *Methods* 115 (Feb. 2017), pages 91–99 (cited on page 17).

[33] C. Guiot et al. “Does tumor growth follow a ’universal law’?” In: *Journal of theoretical biology* 225 (Nov. 2003), pages 147–51 (cited on page 47).

[34] P. Haridas et al. “Quantifying rates of cell migration and cell proliferation in co-culture barrier assays reveals how skin and melanoma cells interact during melanoma spreading and invasion”. In: *Journal of Theoretical Biology* 423 (2017), pages 13–25 (cited on pages 3, 46, 47, 56).

[35] N. Hartung et al. “Mathematical Modeling of Tumor Growth and Metastatic Spreading: Validation in Tumor-Bearing Mice”. In: *Cancer Research* 74.22 (2014), pages 6397–6407 (cited on page 47).

[36] J. Henrikson. “Completeness and total boundedness of the Hausdorff metric”. In: *MIT Undergraduate Journal of Mathematics* (1999) (cited on page 192).

[37] M. S. Hirsch and J. Watkins. “A Comprehensive Review of Biomarker Use in the Gynecologic Tract Including Differential Diagnoses and Diagnostic Pitfalls”. In: *Advances In Anatomic Pathology* 27 (May 2020) (cited on page 7).

[38] X. Jiang et al. “Distance Measures for Image Segmentation Evaluation”. In: *EURASIP Journal on Advances in Signal Processing* 2006 (Dec. 2006) (cited on page 57).

[39] J. Kapuscinski. “DAPI: a DNA-Specific Fluorescent Probe”. In: *Biotechnic & Histochemistry* 70.5 (1995), pages 220–233 (cited on page 12).

[40] A. Klenke. *Wahrscheinlichkeitstheorie*. Springer Spektrum, 2013 (cited on pages 27–29, 33, 65, 74, 75).

[41] Z. Kratka. “Mathematical-Statistical Models in Insurance”. In: *European Scientific Journal, ESJ* 11.7 (Mar. 2015) (cited on page 45).

[42] D. Lamberton and B. Lapeyre. *Introduction to Stochastic Calculus Applied to Finance, Second Edition*. Chapman & Hall/CRC Financial Mathematics Series. Taylor & Francis, 1996 (cited on page 45).

[43] L. Landrieu and G. Obozinski. “Cut Pursuit: Fast Algorithms to Learn Piecewise Constant Functions on General Weighted Graphs”. In: *SIAM Journal on Imaging Sciences* 10.4 (2017), pages 1724–1766 (cited on pages 57, 58).

[44] L. Landrieu and M. Simonovsky. “Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs”. In: *CoRR* abs/1711.09869 (2017) (cited on page 58).

[45] Z. Long et al. “PDE-Net: Learning PDEs from Data”. In: *Proceedings of the 35th International Conference on Machine Learning*. Edited by J. Dy and A. Krause. Volume 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, Oct. 2018, pages 3208–3216 (cited on page 56).

[46] S. Ltd. *Cell Metric*. Accessed: 2021-03-21. Mar. 2021. URL: https://www.solentim.com/products/cell-metric/ (cited on page 8).

[47] S. Ltd. *Cell Metric Datasheet*. Accessed: 2021-03-21. URL: https://www.trio-biotech.com/wp-content/uploads/2016/09/Cell_Metric_Datasheet_11_2013.pdf (cited on page 8).

[48] J. Luo. “Predictive Monitoring of COVID-19”. 2020 (cited on page 46).

[49] F. Maes, D. Vandermeulen, and P. Suetens. “Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information.” In: *Medical Image Analysis* 3.4 (1999), pages 373–386 (cited on pages 93–95).

[50] D. C. Markham et al. “Comparing methods for modelling spreading cell fronts”. In: *Journal of Theoretical Biology* 353 (2014), pages 95–103 (cited on page 49).

[51] MathWorks. *Algorithm for fmincon in MATLAB R2018a*. Accessed: 2021-02-28. Feb. 2021. URL: https://uk.mathworks.com/help/optim/ug/choosing-the-algorithm.html (cited on page 179).

[52] MathWorks. *fmincon in MATLAB R2018a*. Accessed: 2021-01-26. Jan. 2021. URL: https://uk.mathworks.com/help/releases/R2018a/optim/ug/fmincon.html?s_tid=doc_ta (cited on pages 171, 179, 185).

[53] MathWorks. *imregtform in MATLAB R2018a*. Accessed: 2021-02-23. Feb. 2021. URL: https://uk.mathworks.com/help/releases/R2018a/images/ref/imregtform.html?s_tid=doc_ta (cited on page 15).

[54] A. Munk. *Maß- und Integrationstheorie*. Göttingen: Universitätsverlag Göttingen, 2010 (cited on pages 64–66, 74, 75, 140, 145).

[55] J. D. Murray. *Mathematical Biology I. An Introduction*. Springer-Verlag New York, 2002 (cited on pages 3, 45, 46, 49).

[56] J. D. Murray. *Mathematical Biology II: Spatial Models and Biomedical Applications*. Springer-Verlag New York, 2003 (cited on pages 3, 45, 46, 51).

[57] NobelPrize.org. *The Nobel Prize in Chemistry 2008*. Accessed: 2021-03-16. URL: https://www.nobelprize.org/prizes/chemistry/2008/summary/ (cited on page 12).

[58] *NoMADS - Nonlocal Methods for Arbitrary Data Sources*. Accessed: 2020-06-05. URL: https://www.uni-muenster.de/NoMADS/ (cited on page 1).

[59] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. 6th. Springer, Jan. 2014 (cited on page 45).

[60] G. Peyré and M. Cuturi. *Computational Optimal Transport*. 2020 (cited on page 68).

[61] J. Pluim, J. Maintz, and M. Viergever. "Mutual-Information-Based Registration of Medical Images: A Survey". In: *Medical Imaging, IEEE Transactions on* 22 (Sept. 2003), pages 986–1004 (cited on pages 4, 41, 58, 62, 63).

[62] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. "Interpolation Artefacts in Mutual Information-Based Image Registration". In: *Computer Vision and Image Understanding* 77.2 (2000), pages 211–232 (cited on page 186).

[63] A. Ramola, A. K. Shakya, and D. Van Pham. "Study of statistical methods for texture analysis and their modern evolutions". In: *Engineering Reports* 2.4 (2020), e12149 (cited on page 43).

[64] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (May 2015) (cited on page 42).

[65] D. W. Scott. *Statistics: A Concise Mathematical Introduction for Students, Scientists, and Engineers*. John Wiley & Sons Ltd, 2020 (cited on pages 2, 37, 56).

[66] C. E. Shannon. "A mathematical theory of communication." In: *Bell Syst. Tech. J.* 27.3 (1948), pages 379–423 (cited on page 62).

[67] M. Sharma and S. Singh. "Evaluation of texture methods for image analysis". In: *The Seventh Australian and New Zealand Intelligent Information Systems Conference, 2001.* 2001, pages 117–121 (cited on page 42).

[68] P. Sharma and A. Singh. "Era of deep neural networks: A review". In: *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (2017), pages 1–5 (cited on page 41).

[69] R. L. Siegel, K. D. Miller, and A. Jemal. "Cancer statistics, 2020". In: *CA: A Cancer Journal for Clinicians* 70.1 (2020), pages 7–30 (cited on page 10).

[70] S. Skansi. *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence.* Undergraduate Topics in Computer Science. Springer International Publishing, 2018 (cited on page 42).

[71] P. Stapor et al. "PESTO: Parameter EStimation TOolbox". In: *Bioinformatics* 34.4 (Oct. 2017), pages 705–707 (cited on page 56).

[72] A. Strehl, J. Ghosh, and R. Mooney. "Impact of Similarity Measures on Web-page Clustering". In: *Proceedings of the AAAI Workshop on AI for Web Search (AAAI 2000).* Austin, TX, USA, 2000, pages 58–64 (cited on page 57).

[73] A. Strickaert et al. "Cancer heterogeneity is not compatible with one unique cancer cell metabolic map". In: *Oncogene* 36 (2017), pages 2637–2642 (cited on page 9).

[74] H. Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* n/a.n/a () (cited on page 1).

[75] E. Syrimi et al. "Analysis of Global Pediatric Cancer Research and Publications". In: *JCO Global Oncology* 6 (2020). PMID: 32031437, pages 9–18 (cited on page 1).

[76] R. Temam and A. Miranville. *Mathematical Modeling in Continuum Mechanics.* 2nd edition. Cambridge University Press, 2005 (cited on page 45).

[77] D. Tenbrinck, F. Gaede, and M. Burger. *Variational Graph Methods for Efficient Point Cloud Sparsification.* 2019 (cited on pages 57, 58).

[78] ThermoFisher. *DMEM - Dulbecco's Modified Eagle Medium.* Accessed: 2021-03-21. URL: https://www.thermofisher.com/de/de/home/life-science/cell-culture/mammalian-cell-culture/classical-media/dmem.html (cited on page 8).

[79] K. K. Treloar et al. "Are in vitro estimates of cell diffusivity and cell proliferation rate sensitive to assay geometry?" In: *Journal of Theoretical Biology* 356 (2014), pages 71–84 (cited on page 47).

[80] R. Y. Tsien. "THE GREEN FLUORESCENT PROTEIN". In: *Annual Review of Biochemistry* 67.1 (1998). PMID: 9759496, pages 509–544 (cited on page 12).

[81] Y. Wang. *Imaging Processing and Challenges at AstraZeneca.* Presentation at the NoMADS Kickoff Meeting. Mar. 2018 (cited on page 9).

[82] G. B. West, J. H. Brown, and B. J. Enquist. "A General Model for the Origin of Allometric Scaling Laws in Biology". In: *Science* 276 (1997), pages 122–126 (cited on page 47).

[83] T. E. Wheldon. *Mathematical models in cancer research*. Medical science series. Bristol: Adam Hilger, 1988 (cited on page 47).

[84] X. Xun et al. "Parameter Estimation of Partial Differential Equation Models". In: *Journal of the American Statistical Association* 108.503 (2013), pages 1009–1020 (cited on page 56).

Curriculum Vitae