Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

# Computational Uncertainty Quantification in Imaging

## Methods and Applications for Nuclear Decommissioning

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat.
im Fachbereich Mathematik
der Fakultät für Mathematik, Informatik und Naturwissenschaften
der Universität Hamburg

Vorgelegt von:
**Lorenz Kuger**

Hamburg, 2026

| | |
|---|---|
| Gutachter: | Prof. Dr. Martin Burger |
| | Prof. Dr. Luca Calatroni |
| Vorsitzender der Prüfungskommission: | Prof. Dr. Vitalii Konarovskyi |
| Datum der Disputation: | 23. April 2026 |

# Acknowledgement

I am deeply grateful to many people for supporting me during my studies and thus contributing to the creation of this work.

First and foremost, I would like to thank Martin Burger, a fantastic supervisor and group leader. Thank you for moving to Hamburg, and for being so open about it from an early point on. I am very grateful for the many opportunities, ideas, incentives and patience you have given me. Without your supervision, I could not have learnt and seen as much in these last couple years and would not be where I am today.

The same holds for all the amazing people that I got to work with: I thank Matthias Ehrhardt for hosting me in Bath, essentially acting as a co-supervisor over a long period of time and teaching me so much about optimization. Thank you to Carola Schönlieb for hosting me in Cambridge for two months - because of you, I seem to keep returning to live in this small town in the UK over and over. Thank you to Yiqiu Dong and Per Christian Hansen for a great (albeit very rainy) stay at DTU in Lyngby, for your time and interesting discussions. Thank you to Franca Hoffmann and Andrew Stuart for hosting me at Caltech for a (very sunny) month towards the end of my PhD, and to Ricardo Baptista for pointing out a recent paper during a discussion in Pasadena that led to results in this thesis. Thank you to Luca Calatroni for offering to review this thesis.

A big thank you to my research group(s), first in Erlangen, later on in Hamburg: Alex, Ariane, Daniel, Danielle, Jin, Kehan, Lea, Lena, Leon, Lukas, Samira and Tim. Moving across half the country midway through my PhD was not easy, but having you guys around certainly made it a lot less hard. Thank you for the barbecue nights, boardgame sessions, bouldering workouts and everything we got to experience in Erlangen, Hamburg, and while traveling around Germany and the rest of the world. Thank you also to the whole Helmholtz imaging team in Hamburg who made arriving at DESY such a seamless experience. Thank you to Simon, who was among the first to approach 'this new group of mathematicians at DESY': I learned a lot from you, and the two weeks in Leysin were one of the best experiences in the last few years for me—planning this was a big advance of trust and I appreciate you gifting so much of your time to our common work.

To my family, and especially Laura: Thank you for being here and sharing these years with me.

Finally, I thank Fibonacci, Paula, Morse and an unnamed Alster-duck, for involuntarily serving as models whenever I had to test image processing algorithms and ending up published in this thesis and other publications.

# Contents

# List of Figures

# Abstract

This thesis addresses computational methods for Compton camera imaging in nuclear decommissioning applications, combining advances in forward modeling with the development of efficient sampling algorithms for Bayesian uncertainty quantification.

Compton cameras are promising devices for localizing radioactive contaminations without collimation, exploiting the geometry of Compton scattering to provide spatial information about gamma-ray sources. However, translating the well-established physics of gamma-ray interactions into computationally tractable forward models for practical camera geometries presents significant challenges. For cameras with large scintillation detectors, forward models based on first-order scattering processes in the detectors produce unreliable reconstructions. We develop corrected forward models that include second-order scatter corrections, bidirectional detector geometries, and multi-nuclide imaging capabilities. These corrections are shown to be efficiently computable and to improve image reconstruction results in experiments with both simulated and real measurement data from nuclear decommissioning environments.

The image reconstruction task constitutes a linear inverse problem, which we address within the Bayesian framework. While maximum a posteriori estimation can be computed using established convex optimization methods, full uncertainty quantification requires sampling from high-dimensional posterior distributions. For this purpose, we study Langevin Monte Carlo algorithms, which simulate stochastic processes driven by the gradient of the log-posterior and Brownian motion. The non-smooth potentials arising from total variation regularization and domain constraints in imaging applications pose theoretical and algorithmic challenges for these methods.

We provide a unified treatment of convergence theory for several variants of Langevin sampling algorithms. Our contributions include unified convergence results for explicit and semi-implicit schemes, new convergence proofs for proximal Langevin samplers, a rigorous comparison of convergence behavior across different proximal discretizations, and a new optimality result for mirrored Langevin sampling. The latter provides a preconditioning strategy particularly suited for inverse problems with Poisson-distributed data. Finally, we demonstrate the effectiveness of these sampling methods for practical uncertainty quantification in Compton camera imaging.

# Zusammenfassung

Diese Arbeit befasst sich mit Berechnungsmethoden für die Bildgebung mit Compton-Kameras in Anwendungen zum Rückbau kerntechnischer Anlagen und kombiniert dabei Fortschritte in der Vorwärtsmodellierung mit der Entwicklung effizienter Sampling-Algorithmen für die Bayessche Unsicherheitsquantifizierung.

Compton-Kameras sind Geräte zur Lokalisierung radioaktiver Kontaminationen ohne Kollimation. Sie nutzen die geometrischen Eigenschaften der Compton-Streuung, um räumliche Informationen über Gammastrahlenquellen zu liefern. Die Umsetzung der Physik der Gammastrahlenwechselwirkungen in rechnerisch handhabbare Vorwärtsmodelle für praktische Kamerageometrien stellt jedoch eine große Herausforderung dar. Bei Kameras mit großen Szintillationsdetektoren liefern Vorwärtsmodelle, die auf Streuprozessen erster Ordnung in den Detektoren basieren, unzuverlässige Rekonstruktionen. Wir entwickeln korrigierte Vorwärtsmodelle, die Korrekturen für Streuprozesse zweiter Ordnung, bidirektionale Detektorgeometrien und Multinuklid-Bildgebung enthalten. Diese Korrekturen erweisen sich als effizient berechenbar und verbessern die Ergebnisse der Bildrekonstruktion in Experimenten mit simulierten und realen Messdaten aus kerntechnischen Umgebungen.

Die Bildrekonstruktion stellt ein lineares inverses Problem dar, das wir im Rahmen des Bayes'schen Ansatzes behandeln. Während die maximale a-posteriori-Schätzung mit Hilfe etablierter konvexer Optimierungsmethoden berechnet werden kann, erfordert eine vollständige Quantifizierung der Unsicherheit eine Stichprobenentnahme aus hochdimensionalen a-posteriori-Verteilungen. Zu diesem Zweck untersuchen wir Langevin-Monte-Carlo-Algorithmen, die stochastische Prozesse simulieren, die durch einen Potentialterm und die Brownsche Bewegung angetrieben werden. Die nicht glatten Potenziale, die sich aus der Regularisierung mit totaler Variation und den Nebenbedingungen in Bildgebungsanwendungen ergeben, stellen diese Methoden vor theoretische und algorithmische Herausforderungen.

Für mehrere Varianten von Langevin-Sampling-Algorithmen demonstrieren wir eine einheitliche Konvergenz-Theorie. Unsere Beiträge umfassen einheitliche Konvergenzergebnisse für explizite und semi-implizite Algorithmen, neue Konvergenzbeweise für proximale Langevin-Sampler, einen Vergleich des Konvergenzverhaltens verschiedener prox-

imaler Diskretisierungen und ein neues Optimalitätsresultat für gespiegeltes Langevin-Sampling. Letzteres führt zu einer Sampling-Strategie, die sich besonders für inverse Probleme mit Poisson-verteilten Daten eignet. Schließlich demonstrieren wir die Wirksamkeit dieser Sampling-Methoden für die praktische Unsicherheitsquantifizierung in der Compton-Kamera-Bildgebung.

# Declaration of Original Work

Parts of this dissertation are based on the following articles and proceedings.

[42] M. Burger et al. "Multiple scatter correction for single plane Compton camera imaging in nuclear decommissioning." In: *PAMM* 23.3 (2023). DOI: 10.1002/pamm.202300281

[118] L. Kuger and M. Burger. "Bidirectionality and Multiple Scattering Correction in Compton Camera Imaging." In: *Tomographic Inverse Problems: Mathematical Challenges and Novel Applications*. Vol. 20. 2. EMS Press, 2023, pp. 1143–1144. DOI: 10.4171/OWR/2023/21

[77] M. J. Ehrhardt, L. Kuger, and C.-B. Schönlieb. "Proximal Langevin Sampling with Inexact Proximal Mapping." In: *SIAM Journal on Imaging Sciences* 17.3 (July 2024), pp. 1729–1760. DOI: 10.1137/23m1593565

[43] M. Burger et al. *Analysis of Primal-Dual Langevin Algorithms*. 2024. arXiv: 2405.18098 [math.OC] (preprint, under review)

The following papers were co-authored during the period of doctoral studies, but are not discussed in this thesis.

[119] L. Kuger and G. Rigaud. "On multiple scattering in Compton scattering tomography and its impact on fan-beam CT." in: *Inverse Problems and Imaging* 16.5 (2022), p. 1359. DOI: 10.3934/ipi.2022029

[215] S. Welker et al. *Position-Blind Ptychography: Viability of image reconstruction via data-driven variational inference*. 2025. arXiv: 2509.25269 [eess.IV] (preprint, under review)

Contributions in this thesis are due to the author and cooperation partners as follows.

Specialization on the cylindrical detectors in Chapter 3 is due to prototype cameras owned by Hellma Materials GmbH (Hellma). The cameras were used to make measurements in a cooperation between the author, Martin Burger (MB), Hellma, Hochschule Zittau-Görlitz (HSZG) and VKTA - Strahlenschutz, Analytik & Entsorgung Rossendorf

e. V. (VKTA). The FLUKA simulations for Fig. 3.6 were done by HSZG. All modeling work and discussions are the author's (advised by MB) and were reported in [42, 118].

Implementation of all experiments in Chapter 5 is the author's work. The reconstruction algorithms were devised by the author and MB. Simulated data was generated by the author, real data measured by Hellma, HSZG and VKTA. The discussion on learned forward models is due to the master's thesis [38] by Michelle Bruch (MB*), co-supervised by the author, Daniel Tenbrinck (DT) and MB. Polyenergetic models per-emission-line were suggested in the literature, the per-nuclide formulation is the author's work.

In Chapter 7, the algorithms ULA, B*flow*, *flow*B, F*flow*B and MYULA are known in the literature, the clear taxonomy of these schemes and the algorithm FB*flow* are due to the author. The comparisons for Gaussian targets are partly from literature (Lemmas 7.2, 7.4 and 7.5) and were partly generalized by the author. The new results Lemmas 7.3 and 7.6 and Theorem 7.7 are the author's work. The algorithms ULPDA and MLA and the domain constraint strategies are due to earlier literature. The results on inexact proximal steps in Section 7.6 stem from a cooperation between the author, Matthias Ehrhardt (ME) and Carola-Bibane Schönlieb (CBS) and are published in [77].

The general theme of Chapter 8, i.e., using coupling arguments for ergodicity and convergence proofs of Langevin sampling schemes is due to the author, advised by MB. The results for ULA (Theorem 8.2) and F*flow*B (Theorem 8.4) are from the literature. The result Theorem 8.3, the new convergence proof for FB*flow* with Theorem 8.9 and Lemma 8.10 and the discussions of the proofs are the author's work.
All results on ULPDA in Section 8.3 stem from a cooperation between the author, MB, ME and Lukas Weigand (LW) [43]. Showing that the target is not stationary under primal-dual dynamics is due to LW, a correction (not reported here) due to LW and the author. The stability and ergodicity results in continuous time (Theorems 8.11 and 8.12) are due to the author and MB. Proving hypoellipticity (Theorem 8.13) is the author's work. Convergence to the primal limit (Theorem 8.14) was an idea by MB, the detailed proof worked out by the author. The convergence results in discrete time (Theorems 8.15 and 8.16 and Corollary 8.17) are the author's work, partly advised by MB and ME. The numerical experiments are the author's work, advised by ME.
Section 8.4 is due to a cooperation between the author, MB, and Franca Hoffmann (FH). MB and FH advised examining Riemannian gradient flows and Poincaré inequalities, Theorem 8.18 and Lemma 8.19 stem from existing literature. Corollary 8.20 for gamma distributions and all discussions on the connection to inverse problems with Poisson data are the author's work. All implementations and numerical experiments on different mirror functions are the author's work.

In the experiment section Chapter 9 and the additional results section Appendix A, the data is again either simulated by the author or measured by Hellma, HSZG and VKTA. All implementations of forward models and sampling algorithms are the author's work.

The conclusion Chapter 10 is the author's work.

During the writing of this dissertation, AI-assisted chat and software development tools (ChatGPT based on GPT 4.x models, Claude Code based on Sonnet 3.x and Sonnet 4 versions) were used to suggest text formulations, for minor editorial edits and for refactoring of *MATLAB* and *Python* experiment code.

# Chapter 1

# Introduction

## 1.1. Radiation Imaging in Nuclear Applications

**Background.** In March 2011, an undersea megathrust earthquake, the *Tōhoku earthquake*, occurred 72 km off the Japanese Oshika peninsula. The resulting tsunami waves flooded large areas of the Japanese coast and, in doing so, destroyed the electrical grids and many backup energy sources of the *Fukushima Daiichi Nuclear Power Plant*. Despite the immediate shutdown of the power station, the cooling system of the reactor elements failed, causing a core meltdown and the second nuclear energy accident in history rated at the maximum severity by the International Nuclear Event Scale (the other being the Chernobyl accident in 1986). In the aftermath of the disaster, international anti-nuclear power movements gained momentum and several governments re-examined older and high-risk nuclear power plants. The German government even revised its recent decision to extend the operating life of German nuclear power plants, abandoning this energy source entirely in 2023[1]. This leaves Germany with the considerable and unique task of decommissioning approximately thirty shut-down nuclear power plants. While all nuclear power stations must be decommissioned after their finite operational lifetime, the scale of the German project places special demands on efficiency and the technologies used.

In the process of decommissioning a power plant, one of the largest challenges consists in correctly separating harmless, recyclable building materials from radioactively contaminated or activated waste. As most of the power stations have been operated for several decades, some contaminations in the infrastructure, e.g. in pumps, pipework or ventilation ducts, are unavoidable. Additionally, leakages and spillings allow radioactive material to disperse in liquids and through that in porous building material like

[1] Despite ongoing political debate on this matter, several plants are already in advanced stages of decommissioning. For the sake of motivating this thesis, we can safely assume that all nuclear power plants are decommissioned at *some* point in the future.

concrete. In the course of the contaminations' radioactive decay, the emitted ionizing radiation will in turn have activated other lighter-weight materials, leading to secondary contaminations and resulting in potentially complex contamination situations. Finding and quantifying all relevant contaminations is crucial for guaranteeing workers' safety and materials' recyclability and is hence a legal requirement in the release measurement process.

Different measurement techniques are able to assess the various types of penetrating radiation [13, 101, 113]: Alpha and beta contamination monitoring is performed using proportional counters or scintillation detectors, while neutron detection is important for identifying the presence of fissile materials and assessing neutron activation in structural components. For gamma radiation, the most penetrating type in decommissioning environments, several techniques are available: Ionization chambers or Geiger-Müller counters can detect the existence of contaminations and devices are available in small enough size to be carried around easily, but they lack spectrometric accuracy. Precise gamma spectrometry is carried out using scintillation or semiconductor detectors. By measuring counts and gamma energies, these devices can help identify radionuclides and quantify the activity within a predefined field of view. This technique can, however, hardly provide spatial information on the activity source. Placing a collimating material like lead in front of the detector can mitigate this issue, but severely reduces the count rate and in turn increases measurement time.

**Compton Cameras.** For the joint task of spectrometry and spatial activity reconstruction in the gamma-ray range, Compton cameras are promising measurement devices. They consist of multiple, electronically coupled scintillation or semiconductor detectors. By exploiting the inherent geometry of photon interactions, they are able to provide geometrical information without any collimation.

The concept of a Compton camera was proposed in the 1970s, in two different scientific communities at roughly the same time, namely in astronomical imaging [194] and in clinical diagnosis and therapy [206]. As the requirements of these two applications vary slightly in terms of camera size and energy ranges, both fields developed the imaging system further in the following decades, improving on models and algorithms for the image reconstruction. The third prominent application of Compton cameras arises in nuclear research and energy facilities due to their end of lifespan or critical incidents. In such applications, requirements lie between the existing work in astronomy and medicine, so that the technology can successfully borrow from the developments in either fields. The underlying principle remains the same and is successfully employed in all three fields of application: The source of an unknown gamma emitter within a certain energy range has to be localized, be it in astronomical far field and with higher gamma-ray energies, or in closer distance in a power plant or biological sample and more moderate energy ranges.

The cameras inherit their name from the Compton scattering effect of gamma photons. When gamma rays penetrate a scintillation detector, the photons interact with the detector material either through absorption or scattering. For gamma spectroscopy with uncoupled detectors, the absorption is the relevant interaction type in the detector, since the absorbed energy fingerprint provides information about the source nuclide. By matching absorbed energy to known radionuclides, contamination material can be identified, while photon scattering presents mainly a source of noise[2]. A Compton camera, however, explicitly uses Compton scattered photons for the source localization task by combining several, coupled scintillation detectors. When photons are scattered in a first detector, they may enter another detector and interact again. The interactions happen within a short time interval, allowing to relate interactions in two different detectors to another. The detectors still only measure the energy, but one now obtains geometrical information by a relation between energy loss and scattering angle characteristic to Compton scattering. Since any measurement provides two approximate interaction positions and a scattering angle, the source is not determined exactly. The source position can only be backtracked to the surface of a cone, whose parameters are distorted from measurement noise. This makes the source localization task an inverse problem.

## 1.2. Inverse Problems

**Background.** Inverse problems are concerned with reconstructing unknown causes of observed measurement data, using the knowledge about the physical relationship between cause and data. Mathematically, the relationship between unobserved cause $f$ and measurement $g$ in Banach spaces can be formulated by the operator equation $\mathcal{P}(f) = g$. The forward model $\mathcal{P}$ describes our knowledge about the physical measurement system. Within this thesis, we only consider linear inverse problems, i.e., linear operators $\mathcal{P}$ and write $\mathcal{P}f = g$. The terminology "inverse" problem suggests that there exists a corresponding forward, or direct problem, which is typically the reverse direction of obtaining data $g$ from a known cause $f$. This directon consists in carrying out the measurement, or simulating it in a suitable computational setup.

The distinction between direct and inverse problem corresponds to the well-posedness of retrieving either side of the equation: Inverse problems are often ill-posed, as in not well-posed. In the sense of Hadamard, the problem $\mathcal{P}(f) = g$ is called well-posed if it allows for a unique solution $f^*$ for all suitable data instances $g$, and if this solution $f^*$ depends continuously on $g$. Inverse problems are usually ill-posed, while the corresponding direct problem is often well-posed. The notion of existence and uniqueness of solutions is hereby usually less problematic, as we can guarantee both by softening our definition of a solution: The generalized inverse $\mathcal{P}^\dagger g$ does not solve the equation exactly, but satisfies the existence and uniqueness properties. The instability due to the non-existence of a

[2] With the reservation that, for some nuclides, the characteristic Compton edge may also be used for characterization.

continuous $\mathcal{P}^{-1}$, however, is often unavoidable. Since measurement data $g$ is subject to noise in most applications, this instability is the major challenge in the solution of inverse problems. It makes the numerical solution of the inverse problem difficult for any straightforward solver of the operator equation, even given exact data $g$. The solution approach typically consists in regularizing the problem by imposing additional constraints or prior information about $f$.

One possible way to formalize the regularization and the noise in the measurement data is Bayesian inference. If we know the statistical distribution of the noise in the observation $g$, we can formulate a likelihood distribution with density $\rho_{\text{like}}(g \,|\, f)$. Trying to solve for $f$ naively requires maximizing the likelihood to compute an estimate $\hat{f} = \arg\max_f \rho_{\text{like}}(g \,|\, f)$ of the solution. Even in simple cases with, e.g., a log-concave likelihood, the ill-posedness of $\mathcal{P}$ leads to ill-conditioning of this optimization problem. Standard optimization problems may fail, converge slowly, or produce artifacts in estimates of $\hat{f}$. For regularization, additional information on $f$ is incorporated by formulating our statistical knowledge about $f$ (independent of any measurement at hand) in a prior distribution $\rho_{\text{prior}}(f)$. By Bayes' law, assuming all densities exist, the posterior distribution of $f$ after making a measurement $g$ has the density $\rho_{\text{post}}(f \,|\, g) \propto \rho_{\text{like}}(g \,|\, f)\, \rho_{\text{prior}}(f)$. The posterior is the solution of the inference problem, and needs to be inspected for downstream tasks like point estimate and confidence region estimation. This pipeline poses two main methodical questions. We need to formulate the "correct" densities $\rho_{\text{like}}$ and $\rho_{\text{prior}}$, and we need algorithmical tools to inspect the high-dimensional posterior. The former are modeling problems which require study of the physical reconstruction problem at hand—the first, applied part of this thesis covers these for Compton camera imaging. The latter questions will be studied in the second, algorithmical part of this work on posterior sampling algorithms.

**Posterior sampling algorithms.** We study a class of efficient sampling algorithms based on Langevin diffusion that are applicable to general log-concave target distributions. For a target density $\exp(-V(x))/Z$ with a convex potential $V$, these methods take inspiration from statistical physics. They simulate a stochastic process driven by both the force term $\nabla V$ and random fluctuations induced by Brownian motion. These simulations have led to efficient Monte Carlo sampling algorithms, whose convergence properties to smooth targets scale well to high-dimensional applications. Bayesian imaging posteriors are notoriously high-dimensional, but often contain non-smooth potentials and other problem constraints, which raises several theoretical and algorithmical questions. Under non-smoothness of $V$, standard convergence results are not applicable. Due to the unbounded domain of the heat kernel, the Brownian motion makes sampling from bounded domains particularly challenging from a theoretical perspective. In order to still employ Langevin-based samplers, several extended sampling schemes have been proposed in recent literature. Implicit discretizations of the Langevin diffusion process are aimed at guaranteeing convergence properties for non-smooth potentials, and preconditioning and mirroring methods are employed in order to constrain samples to subsets

of the state space. Convergence and optimality results on these algorithms are, however, still limited.

## 1.3. Overview and Contributions

This thesis focuses on advancing the two topics outlined above: Compton camera modeling and posterior sampling algorithms for Bayesian imaging. We now give an overview of its structure and list the main contributions.

**Overview.** Compton camera imaging in nuclear decommisioning poses a typical, linear imaging inverse problem. Concretely, $f$ in the explanations about inverse problems is an activity map of radioactive contaminations in space. The measurement $g$ is acquired by a Compton camera that is placed near the contaminations. Once we know the camera geometry and the imaged region in space, we can formulate the forward model based on our knowledge about the interaction of gamma rays with the camera's detectors. The fundamental physical laws governing interactions between gamma rays and matter will be reviewed in Chapter 2.

In practice, these laws have to be translated into computationally tractable forward models for practical camera geometries. This requires several simplifications since, knowing $f$, computing $\mathcal{P}f$ is not simple, as gamma rays interact with the camera in convoluted ways. Most studies either approximate $\mathcal{P}f$ in simulations or stick to a simple, first-order approximation of $\mathcal{P}$ and subsume all remaining, unexplained measurements as "noise". For the large-detector geometries considered in this work, first-order models produce unreliable reconstructions. Therefore, we dedicate Chapter 3 to several extensions and corrections of the usual first-order model. We develop higher-order scatter corrections that account for multiple scattering events within large detectors—effects that are typically neglected in the literature. Additionally, we extend the modeling framework to bidirectional camera geometries and multi-nuclide imaging scenarios, the latter of which has been previously demonstrated on simulated data.

Once we decide on a model formulation, this defines the likelihood distribution $\rho_{\text{like}}$. In Chapter 4, we explain other typical considerations going into Bayesian inference for image reconstruction problems like prior distributions and other, typical likelihood densities. The two distributions together define the Bayesian posterior, which is used for downstream inference tasks. We introduce two broad classes of algorithms needed for inference on the posterior, specifically, variational optimization methods for maximum a posteriori (MAP) estimation and Monte Carlo methods for posterior sampling.

Chapter 5 is the first of two chapters in this work combining the modeling efforts with algorithmical inference. The different forward models and their viability for image reconstruction in decommissioning are tested on the problem of MAP computation, using established convex optimization algorithms.

The algorithmical part of this thesis covers the design of efficient sampling algorithms for general, log-concave posteriors in Chapter 6 through Chapter 8. The challenges we address—non-smooth potentials from total variation regularization, domain constraints, and convergence guarantees for high-dimensional distributions—are however directly motivated by imaging applications and essential for Compton camera imaging. We study algorithms that are based on Langevin diffusion, requiring us to cover the basics of diffusion processes and their connections to optimization in Wasserstein space in Chapter 6.

The sampling algorithms are time discretizations of stochastic differential equations. Similar to convex optimization based on gradient flows, this leads to a number of possible schemes based on the chosen discretization. In Chapter 7 we give an overview of possible Langevin diffusion discretizations schemes. This covers various combinations of explicit and implicit discretizations and novel schemes based on primal-dual discretization and mirroring techniques. We provide new insights on the comparison between explicit, implicit and smoothing-based discretization for non-smooth target densities. We end the chapter by covering practical aspects on sampling under domain constraints and the treatment of errors in proximal schemes.

In Chapter 8, we revisit the various different schemes once more, with a focus on the ergodicity and convergence guarantees. The chapter makes new contributions to the convergence analysis of Langevin sampling algorithms for non-smooth targets and for mirrored sampling.

In Chapter 9, we circle back to practical imaging problems and apply Langevin sampling algorithms to Compton camera imaging. The experiments confirm our theoretical findings on both the new forward models and the efficient algorithm design for high-dimensional sampling.

We conclude the thesis in Chapter 10, where we review the results and discuss remaining open questions and potential future research directions.

**Contributions.** The primary contributions of this thesis are:

- We develop corrected forward models for Compton cameras with large scintillation detectors. This includes second-order scatter corrections and bidirectional detector geometries in combination with multi-nuclide imaging. We explain the specific challenges of such detector setups due to limited spatial and energy resolution and show our model corrections are efficiently computable and improve image reconstruction results.
- We employ convex optimization methods to reconstruct activity distributions with the developed Compton camera imaging models. Experiments are carried out with simulated and real measurements from typical nuclear decommissioning environments.

- We unify and generalize convergence theory for several variants of Langevin sampling algorithms. As part of this, we
  - unify existing convergence results for the unadjusted Langevin algorithm and a proximal, semi-implicit variant;
  - give a new convergence proof for a second proximal Langevin sampler;
  - compare the convergence behaviour of three different proximal schemes, obtaining a novel, clear ranking of convergence speed between the three in the simplified Gaussian target setting;
  - review a recently proposed primal-dual Langevin sampler, and compare it to the purely primal schemes;
  - provide a new optimality result for the convergence of mirrored Langevin sampling, which provides a preconditioning strategy for inverse problems with Poisson distributed data.

  These contributions significantly extend the current literature on convergence of Langevin sampling algorithms, particularly in the case of non-smooth or domain-constraint targets.
- We demonstrate the effectiveness of the mirrored Langevin schemes for practical uncertainty quantification by applying the sampler to Compton camera imaging.

# Chapter 2

# Physical Principles of Gamma Emission Imaging

The measurements of a Compton camera in nuclear decommissioning are triggered by energy depositions in the camera's scintillation detectors. Each of these energy depositions is caused by an interaction between gamma rays from radioactive sources and the scintillation material that the detector consists of. In this chapter, we review the decay schemes of radioactive nuclides. We describe the kinematics of the predominant interactions between photons and detector in the low gamma energy region—Compton scattering, photoelectric absorption and electron-positron pair production. Afterwards, we review how they are exploited by a suitable detector setup in a Compton camera in order to reconstruct the source distribution. The discussions in this chapter are all based on the textbook [113], decay data are taken from standard tables on radioactive nuclei [21–23].

## 2.1. Gamma Ray Spectra of Radioisotopes

Gamma radiation is produced when excited nuclei transition to lower energy nuclear states. In decommissioning of power stations, the excited states are present as contaminations in the environment, i.e., on or in walls, pipes, ventilation shafts, etc. in the buildings. They are typically daughter nuclides in the decay chain of the original radioactive fuel, or of secondarily activated elements. The decay scheme of the common gamma-ray source Cobalt-60 is depicted in Fig. 2.1. In Compton camera imaging, we aim to measure the 1.17 MeV and 1.33 MeV gamma rays that are emitted in the decay.

Within this work, we concentrate on four radioisotopes that are commonly present in power plants and important to detect in decommissioning work, listed in Table 2.1. In each of the four nuclides, a primary decay (such as $\beta^-$ or electron capture) leads to

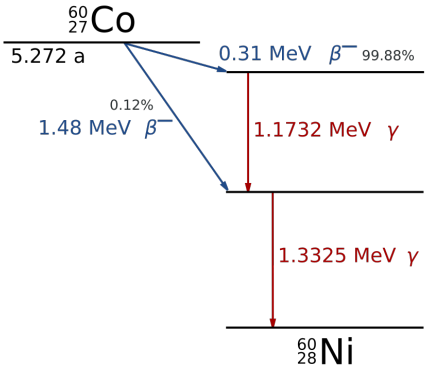



Figure 2.1.: The $\beta^-$ and $\gamma$ decay scheme of Cobalt-60. Source: Wikimedia Commons.

the formation of an excited state in the daughter nucleus. This excited state decays to the ground state, emitting gamma radiation at fixed, characteristic lines. The number and values of characteristic emission energies vary between radioisotopes. In simple cases, like Caesium-137, there is only a single emission energy at 662 keV. Barium-133 and Europium-152 have more complicated decay schemes, with Europium-152 emitting at several dozen characteristic levels, of which approximately seven are realistically detectable. In Table 2.1, we list the four radioisotopes considered in this work, along with their prominent emission lines.

Table 2.1.: Emission lines of standard radioisotopes in nuclear decommissioning.

| Isotope | List of prominent emission lines rounded to keV |
|---|---|
| Caesium-137 | 662 |
| Cobalt-60 | 1173, 1332 |
| Barium-133 | 81, 276, 303, 356, 384 |
| Europium-152 | 40, 122, 244, 344, 778, 1098, 1408 |

The detection of these emission lines with a Compton camera is possible due to several types of interactions that take place between gamma rays and the detector material inside the camera. We next go through the kinematics of these interactions.

## 2.2. Interactions between Gamma Rays and Matter

We start by describing the three types of interactions and their relative predominance depending on the photon energy level and detector material.

**Compton Scattering.** Compton scattering occurs when a gamma photon transfers part of its energy to an electron in an inelastic collision, see Fig. 2.2.

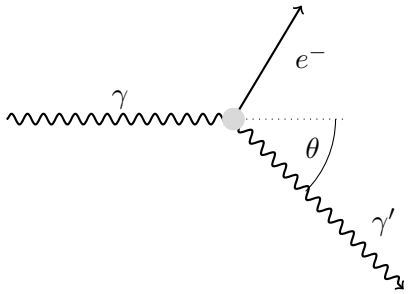



Figure 2.2.: Sketch of a Compton scattering event: Compton scattering is incoherent, meaning the photon $\gamma$ transfers a part of its energy to the electron as kinetic energy. The new photon $\gamma'$ changes direction and has energy $E'$ given by (2.2).

The photon is inelastically scattered in a different direction, and the electron receives kinetic energy. Key to this interaction is the relation between the change in the photon's wavelength and the scattering angle through the Compton equation:

$$\lambda' - \lambda = \frac{h}{m_e c}(1 - \cos\theta). \tag{2.1}$$

Here, $\lambda$ is the wavelength of the incoming photon, $\lambda'$ is the wavelength after scattering. On the right side $h$ is Planck's constant, $m_e$ an electron's mass at rest, $c$ the speed of light, and $\theta$ is the scattering angle of the photon relative to its original direction. By the Compton wavelength relation $\lambda = hc/E$, the energy of the scattered photon is:

$$E' = \frac{E}{1 + \frac{E}{m_e c^2}(1 - \cos\theta)}, \tag{2.2}$$

where $E$ is the energy of the incident gamma photon, and $E'$ is the energy of the scattered photon. The remaining energy is transferred to the electron as kinetic energy. Assuming the electron is at rest, due to energy conservation this is given as

$$E_{\text{electron}} = E - E'. \tag{2.3}$$

Note that this directly implies physical limits on the transferred energy. From $-1 \leq \cos\theta \leq 1$, (2.2) and (2.3) imply the bounds

$$\begin{cases} \frac{E m_e c^2}{m_e c^2 + 2E} \leq E' \leq E, \\ 0 \leq E_{\text{electron}} \leq \frac{2E^2}{m_e c^2 + 2E}. \end{cases} \tag{2.4}$$

In the extreme case of forward scattering $\theta = 0$, close to no energy is transferred to the electron, while the two other bounds correspond to the maximum possible energy loss for backscattering ($\theta = \pi$).

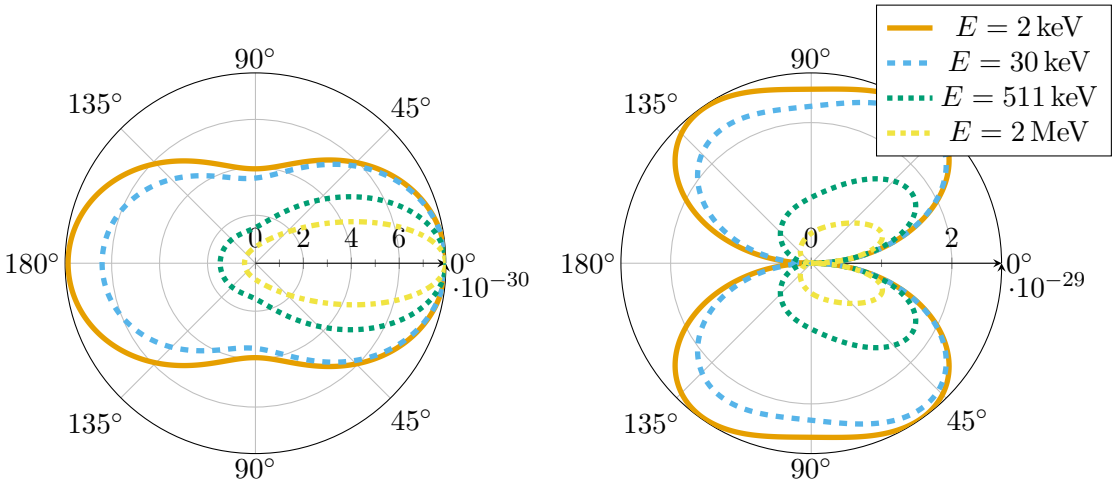



Figure 2.3.: Differential cross sections of Compton scattering. The left plot shows $\frac{\mathrm{d}\sigma^{\mathrm{CS}}}{\mathrm{d}\Omega(\theta)}$ for $\theta \in [0°, 180°]$ and can be interpreted as the scatter probability in a certain direction per unit solid angle. The right plot shows $\frac{\mathrm{d}\sigma^{\mathrm{CS}}}{\mathrm{d}\theta}(\theta)$, i.e., the probability of a given scatter angle $\theta$. Both plots are vertically mirrored to indicate the symmetry in the azimuthal angle $\varphi$. We plot the differential cross sections for different incident energies $E$. Note that while a scatter process within a unit solid angle is most likely around $\theta \approx 0°$, the most likely scatter angles are around $20° - 50°$ for typical gamma ray energies between 50 keV and 2 MeV.

Given a single, unpolarized photon that undergoes Compton scattering, not all scattering angles $\theta$ are equally likely, see Fig. 2.3. The angular distribution is described by the differential cross section of Compton scattering. In simplified terms, the Compton scattering cross section $\sigma^{\mathrm{CS}}$ can be understood as an effective area that quantifies the probability of a photon interacting with a single electron. The differential cross section, $\frac{\mathrm{d}\sigma^{\mathrm{CS}}}{\mathrm{d}\Omega}$, describes how the probability of scattering varies with angle and is expressed per unit solid angle $\mathrm{d}\Omega$. The Compton scattering differential cross section was derived by Klein and Nishina in 1929 [111]. It is given by

$$\frac{\mathrm{d}\sigma^{\mathrm{CS}}}{\mathrm{d}\Omega}(\theta) = \frac{r_e^2}{2}\left(\frac{E'}{E}\right)^2\left(\frac{E}{E'} + \frac{E'}{E} - \sin^2\theta\right), \tag{2.5}$$

where $E'$ depends on $\theta$ and is given by (2.2). Further, $r_e$ denotes the classical electron radius $r_e = e^2/(4\pi\varepsilon_0 m_e c^2)$, where $e$ is the elementary charge and $\varepsilon_0$ the permittivity constant of vacuum. $\mathrm{d}\Omega = \sin\theta\,\mathrm{d}\varphi\,\mathrm{d}\theta$ is a differential solid angle element in the direction of the azimuthal angle $\varphi$ and polar angle $\theta$ relative to the direction of the incident photon.

Note that the right side of (2.5) does not depend on $\varphi$, this implies that the process is symmetric around the incident axis[1]: Just as the energy of the scattered photon (2.2), the Klein–Nishina differential cross section only depends on the scattering angle $\theta$ and not the azimuthal direction $\varphi$ of the new photon after scattering. By integrating (2.5) over all azimuthal angles $\varphi$, we obtain the differential cross section for a given scattering angle $\theta$ as

$$\frac{\mathrm{d}\sigma^{\mathrm{CS}}}{\mathrm{d}\theta}(\theta) = 2\pi \frac{\mathrm{d}\sigma^{\mathrm{CS}}}{\mathrm{d}\Omega}(\theta) \sin\theta. \tag{2.6}$$

Integrating again over all angles $\theta$ between 0 and $\pi$, the total cross section is

$$\sigma^{\mathrm{CS}} = 2\pi r_e^2 \left( \frac{2(1+\alpha)^2}{\alpha^2(1+2\alpha)} + \frac{\log(1+2\alpha)}{\alpha} \left( \frac{1}{2} - \frac{1+\alpha}{\alpha^2} \right) - \frac{1+3\alpha}{(1+2\alpha)^2} \right), \tag{2.7}$$

where $\alpha = E/(m_e c^2)$ is normalized energy. For simulation of Compton scattering effects, the cross section is used to quantify the decrease in intensity. Given it quantifies the probability of interaction with a single electron, the linear attenuation coefficient $\mu^{\mathrm{CS}}$ of Compton scattering is

$$\mu^{\mathrm{CS}} = n_e \sigma^{\mathrm{CS}}, \tag{2.8}$$

where $n_e$ is the electron density in the interacting material.

**Photoelectric Absorption.** In photoelectric absorption events, incident photons lose their complete energy to an electron in the attenuating medium. The photon is absorbed, and the photoelectron is ejected from one of the atom's inner shells, see Fig. 2.4. By energy conservation, the energy of the photoelectron is given by the difference between the energy $E$ of the incident photon and the binding energy $E_{\mathrm{b}}$ of the electron in the atom, i.e.,

$$E_{\mathrm{electron}} = E - E_{\mathrm{b}}. \tag{2.9}$$

In the process, the vacant shell position is rapidly filled by an outer-shell electron, releasing a characteristic X-ray photon of energy $E_{\mathrm{b}}$, which may interact again. This cascade typically continues until the incident photon's full energy is converted to kinetic energy of a large number of freed electrons.

To quantify the expected number of photons absorbed in a medium, we can also give an approximation for the linear attenuation coefficient—or, equivalently, the cross section—of photoelectric absorption. A closed formula does not exist, however, for photons of

[1] This is actually only true for unpolarized incident photons. For polarized photons, the total cross section $\sigma^{\mathrm{CS}}$ does not change, but the differential cross section has to account for the polarization—azimuthal angles $\varphi$ that are approximately perpendicular to the axis of polarization are then more likely. In emission imaging, polarization only has an effect on multiply scattered interactions since the initially emitted photons are unpolarized. The study [53] systematically analyzed the data and reconstruction errors due to polarization and came to the conclusion that they are negligible in Compton camera image reconstruction. We are going to ignore polarization in the modeling part of this thesis.

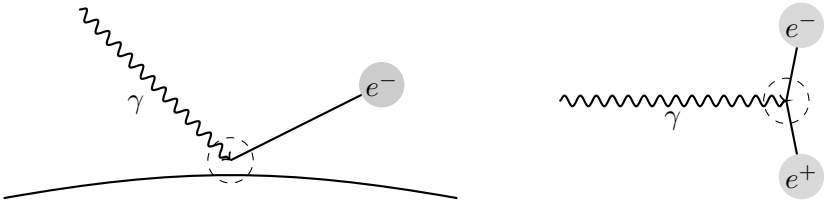



Figure 2.4.: Left: Sketch of a photoelectric absorption event. Right: Sketch of a pair production event.

energy $E$ in interacting material with atomic number $Z$, it is often approximated by

$$\mu^{\mathrm{PE}} \propto Z^5/E^{3.5}, \tag{2.10}$$

or other, more accurate empirical approximations [99]. Note that the cross section exhibits discontinuous jumps (absorption edges) when the photon energy surpasses the binding energies of inner shell electrons, as absorption from these shells becomes energetically possible, see Fig. 2.5. For simulation, it is typically easier to rely on experimentally obtained values, which can be imported from a database such as XCOM [26], allowing to interpolate the cross section at queried energy values.

**Pair Production.** Pair production occurs when a high-energy gamma photon (with energy greater than 1.022 MeV) interacts with the electromagnetic field of a nucleus or an electron, creating an electron-positron pair. The photon's energy is converted into the rest mass of the electron and positron of 511 keV each, any excess energy is shared between the two particles as kinetic energy, see Fig. 2.4.

The electron and positron typically lose their energy rapidly through annihilation of the positron and secondary interaction of the electron. Since pair production requires photons with energies above the 1.022 MeV threshold, the attenuation coefficent $\mu^{\mathrm{PP}}$ for pair production is zero at energies $E$ below the threshold and fairly small for energies slightly above the threshold, so we will also be able to neglect pair production in some of the modeling steps for the Compton camera. The attenuation values are also available in public databases [26]. This interaction type becomes more prevalent above a few MeV and is therefore relevant in high-energy gamma-ray imaging or nuclear medicine (e.g. in positron emission tomography).

**Combining the interaction types.** Neglecting other types of interaction in the gamma energy range [50 keV, 2 MeV], the attenuation coefficient for Compton scattering, photoelectric absorption and pair production combined is simply the sum of the three individual coefficients:

$$\mu^{\mathrm{lin}} = \mu^{\mathrm{CS}} + \mu^{\mathrm{PE}} + \mu^{\mathrm{PP}}. \tag{2.11}$$

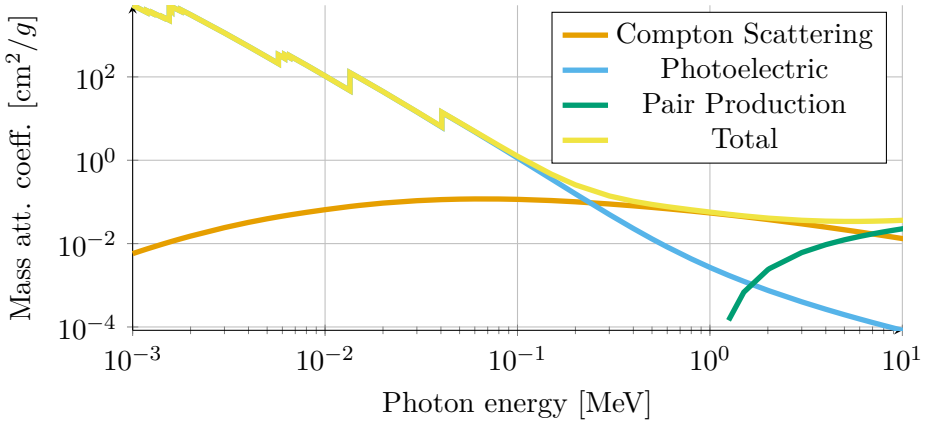



Figure 2.5.: Mass attenuation coefficients $\mu^{\text{mass}} = n\sigma/\rho$ of the scintillation material $CeBr_3$ for Compton scattering, photoelectric absorption and pair production in the range 1 keV to 10 MeV.

Alternatively, one often also uses the mass attenuation coefficient $\mu^{\text{mass}} = \mu^{\text{lin}}/\rho$ in order to compare the attenuation properties of materials, where $\rho$ is the mass density of the interacting material. Figure Fig. 2.5 shows the three attenuation coefficients (normalized by density) for different energy levels in a Cerium Bromide compound, a scintillation detector material.

Knowing the attenuation coefficients of all three types of interactions allows to determine the number of scattered and absorbed photons in the transport of radiation through material. Suppose that a monochromatic beam of photons with initial intensity $I_0$ travels through a material of thickness $z$. Then for a small travelled length $\mathrm{d}z$, the decay $\mathrm{d}I$ in beam intensity is proportional to the probability of an interaction, and hence the linear attenuation coefficient. It therefore holds

$$\frac{\mathrm{d}I}{\mathrm{d}z} = -\mu^{\text{lin}} I. \tag{2.12}$$

If the attenuation coefficient is homogeneous along the whole travelled path of length $z$, integrating (2.12) becomes the Beer-Lambert law for exponential attenuation of beam intensity

$$I = I_0 \exp(-n_e \sigma^{\text{lin}} z). \tag{2.13}$$

For polychromatic beams or inhomogeneous materials, the relation can be generalized by averaging the attenuation coefficient over the present photon energy levels and the local particle densities and cross sections along the travelled path.

## 2.3. Scintillation Coincidence Measurements

Scintillation detectors are among the most widely used instruments for detecting gamma radiation in nuclear imaging and experimental physics. By exploiting the interactions covered in the previous section, they are able to convert incoming high-energy photons into visible light, which can then be transformed into an electrical signal for further processing. We now outline the operating principle of scintillation detectors and show how these detectors can be arranged to detect coincidence events for Compton camera imaging.

**Scintillation and Signal Generation.** When a gamma photon enters a scintillation material, such as $CeBr_3$, NaI, or LYSO, it interacts primarily through photoelectric absorption, Compton scattering, or, at sufficiently high energies, pair production. These interactions deposit energy in the medium, exciting atoms or molecules. As the material de-excites, it emits scintillation light, a cascade of lower-energy (visible or near-visible) photons.

This light is collected by a photodetector, commonly a photomultiplier tube, which converts the incoming photons into an electrical pulse. The process involves photoelectric emission at the photocathode followed by electron amplification, resulting in a current pulse whose total charge $Q$ is, ideally, proportional to the number of scintillation photons and thus to the original gamma photon energy.

**Pulse Charge and Energy Calibration.** The output of a scintillation detector is a voltage or current pulse that rises quickly and then decays over a characteristic timescale. By integrating the light intensity $I^{\text{light}}$ over a whole pulse, one obtains the charge of pulse value $Q$

$$Q \propto \int I^{\text{light}}(t)\,\mathrm{d}t$$

The total light intensity is, in turn, approximately proportional to the energy deposited in the scintillator by the incident photon. Hence, measuring $Q$ provides a means of estimating the photon's energy loss $\Delta E$ that was deposited in the scintillation detector.

Converting $Q$ back to $\Delta E$ requires calibration. To determine the correspondence between measured pulse charge and gamma-ray energy, known radioactive sources with few discrete emission lines (e.g., $^{137}$Cs at 662 keV, $^{22}$Na at 511 keV and 1274 keV) can be used. For each calibration source, the measured pulse height spectrum exhibits peaks around the source's emission energies. By identifying these peaks and assigning them to their known energies, the calibration function $\Delta E = f(Q)$ can be constructed. Unfortunately, proportionality between deposited energy and pulse charge is an idealized relationship. In reality, light output in a scintillation detector can vary quite considerably over the gamma energy range [214]. As a remedy, an extended calibration function can be fitted, e.g., a quadratic $f(Q) = c_2Q^2 + c_1Q + c_0$ or another interpolating model. The complexity

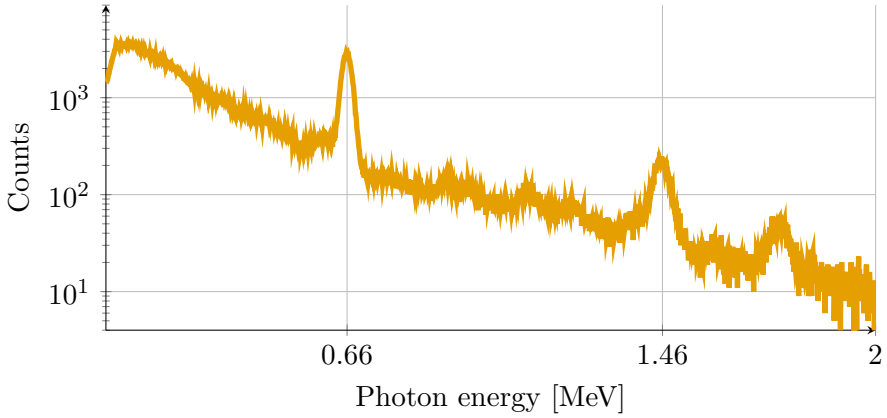



Figure 2.6.: Emission spectrum of a $^{137}$Cs source measured by a $CeBr_3$ detector, which could be used for calibrating the scintillation detector.

of the calibration is, however, severely limited by the amount of calibration peaks that can be identified in practice (a linear model requires two characteristic emission lines, a quadratic three, and so on). For instance, when calibrating with a standard Caesium-137 source, see Fig. 2.6, there is a single characteristic line at 662 keV, and one needs to find at least one other characteristic energy of environmental background sources, like the 1460 keV line of Potassium-40 or the Tellurium-208 line at 2615 keV.

Several other difficulties may arise in this calibration process:

- **Background and Compton continua:** Scattered photons and photons from other environmental sources contribute to a background spectrum, especially below photopeaks, which may obscure weaker peaks or complicate peak fitting. In Fig. 2.6, the background is visible as the flatter plateaus of activity around the strong Caesium-137 and Potassium-40 peaks.
- **Resolution effects:** Due to statistical fluctuations in scintillation light production and photodetection, the energy resolution is finite, typically described by the full width at half maximum (FWHM) of a photopeak. This leads to broadened peaks, visible as well in Fig. 2.6, and complicates precise peak location and identification.
- **Dependence of $f$ on environmental variables:** Even accounting for nonlinearity, the relationship $\Delta E = f(Q)$ is an idealized one, assuming constant detector and environment properties. For Compton cameras in nuclear decommissioning, this means that detectors have to be recalibrated, typically before every measurement. For instance, components of scintillation detectors, like the photomultiplier tube or the electronics, may be temperature-dependent. Since the photomultiplier tube receives thermal energy from the light pulses, it can exhibit calibration instability during the measurement itself. This is particularly true for longer measurements (several minutes or hours) and larger detectors, in which case the calibration has to account for time dependency, fitting instead a model $\Delta E = f(Q, t)$, where $t$ is

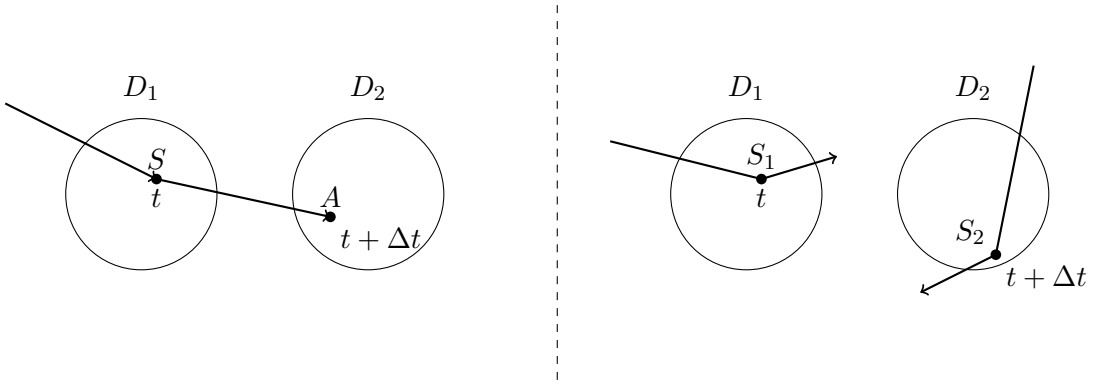



Figure 2.7.: Left: Sketch of a true coincidence event with scattering at $S$ in detector $D_1$ and subsequent absorption at $A$ in another detector $D_2$. Right: Sketch of an accidentally observed coincidence event. If there are two unrelated energy depositions at $S_1$ and $S_2$ in quick succession (i.e. $t + \Delta t < \delta$), this triggers a coincidence measurement. These accidental coincidences present a major source of noise in Compton camera imaging with large scintillation detectors or high total gamma activity.

the time at which a pulse $Q$ is recorded. Without accounting for time dependency, the peaks in a spectrum like Fig. 2.6 will be further broadened and may become even harder to locate and identify.

Since accurate energy measurements are the basis for reliable image reconstruction in Compton cameras, an accurate and robust calibration process is vital in practice.

**Coincidence Measurements.** In Compton camera imaging, it is crucial to identify interactions that originate from the same incident photon—so-called *coincidence events*. These events occur when two or more detectors simultaneously register interactions, such as a gamma photon undergoing Compton scattering in one detector and then being absorbed in another.

Coincidence detection systems rely on precise timing electronics to compare signal arrival times from multiple detectors. If the time difference $\Delta t$ between signals from two detectors lies below a coincidence threshold $\delta$, typically in the nanosecond range, the event is registered as a coincidence.

Accidental coincidences, where unrelated events happen to fall within the coincidence window, represent a significant source of noise, see Fig. 2.7. The rate of such events increases with the activity of the source and the width $\delta$ of the coincidence window. Therefore, optimizing $\delta$ and employing appropriate energy thresholds are critical for ensuring measurement fidelity and reducing model errors in downstream imaging tasks.

# Chapter 3

# Compton Camera Design and Likelihood Modeling

The idea to replace single, collimated detectors in clinical [206] or astronomical [194] applications with a type of "electronic collimation" has inspired numerous works on Compton cameras in the second half of the last century. In medical applications, the developments led to compact cameras with detector sizes of a few centimeters and high spatial resolution [78, 196, 198, 143, 65, 218]. The typical energies lie at the low end of the gamma spectrum (50 keV to 2 MeV). Efforts towards in-vivo imaging of radiotracers have been made in the past years [151, 152, 114, 81, 183, 184]. In the astronomical applications, cameras were developed at much larger size. They typically operate at higher energies (mostly above 1 MeV) and at lower count rates, a prominent example is the Imaging Compton Telescope (COMPTEL) aboard the NASA's Compton Gamma Ray Observatory at the end of the last century [193]. The mathematical principle of both imaging methods is identical. The third prominent application of Compton cameras comes from nuclear research and energy facilities [126, 92, 189, 209, 124], particularly in decommissioning scenarios following their operational lifespan or after incidents. For the exploration of unknown radiation sources in such environments, the same principle can be employed.

The measurement regime in nuclear decommissioning applications sits in the gap between medical and astrophysics applications: The cameras are typically medium sized and image sources in a couple meters distance in the low- to medium gamma-ray range. For decommissioning activities with unclear contamination situations, the cameras are particularly attractive since nuclide identification and localization can be carried out, potentially even in a joint reconstruction without the need to manually identify nuclides in the spectroscopy [222]. In the present chapter, we develop the forward model of the Compton camera image reconstruction problem for monolithic cameras aimed at nuclear decommissioning applications. After reviewing the standard conical Radon transform

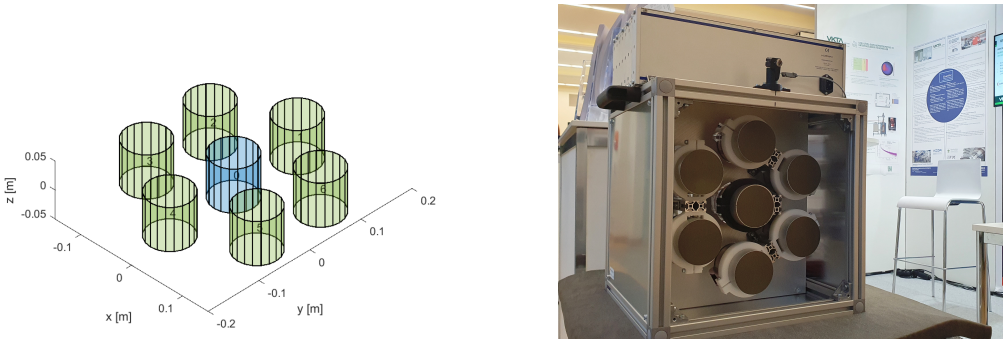



Figure 3.1.: On the left, the relative positions of the seven cylindrical detectors in the camera are shown. The center detector in blue is a $CeBr_3$ crystal detector, the six outer green detectors are PVT plastic scintillators. The right shows the whole camera with encasing and electronics attached at the back.

model, we develop several critical extensions for practical decommissioning scenarios: corrections for second-order scattering in large detectors and a bidirectional model for cameras without predetermined scatterer-absorber assignments. In Chapter 5, we also consider polyenergetic imaging capabilities with these cameras. These extensions address significant sources of model error that are typically neglected in the literature. For extensive overviews of typical models and reconstruction algorithms for Compton camera imaging, we refer to [135, 83].

## 3.1. Camera Design and Experimental Setup

We start by explaining the geometry of the camera that acquired the measurement data used in this thesis.

By a "detector", we refer to a volume of fixed position which registers photon interactions with a certain energy resolution, but *without* spatial resolution. This distinguishes the camera from Compton camera models with pixelated or voxelated detectors, which provide more accurate estimates of photon interaction locations. By analyzing the light distribution, thin plate scintillation detectors can typically achieve spatial resolution up to a few millimeters [92]. In contrast, we use large cylindrical, monolithic detectors of 7.6 cm in both height and diameter. Not requiring spatial resolution makes detector setups cheaper. The large size further maximizes sensitivity and yields photon count several orders of magnitude larger than thin plates, but increases the volume of potential interaction locations by approximately three orders of magnitude.

Most of the measurements we acquire stem from a camera consisting of seven cylindrical detectors. The setup is shown in Fig. 3.1 and is called RSL7 (Radiation Source Locator) in the following. The six outer scintillation detectors are polyvinyltoluene (PVT) plastic scintillation detectors. These are placed on a circle of radius 12.5 cm around the

center $CeBr_3$ detector, an established Compton camera material which has been tested in decommissioning applications before [92].

The six outer PVT detectors have small absorption cross section and lower energy resolution, so coincidences between any two of these are discarded. We only process coincidences between any one PVT detector and the central $CeBr_3$ detector. The detector geometry is similar to an earlier ring shaped design in [143]. Many other Compton cameras have scatterer detectors placed in front of the absorbers, requiring the camera to roughly face contaminations for reliable reconstruction. In contrast, the present setup has scatterers around the absorber, allowing in theory a larger $2\pi$ field of view—half a sphere, since the front and the back of the camera produce indistinguishable data due to the camera's symmetry.

A second, simplistic setup, RSL2, consists of a single pair of $CeBr_3$ detectors placed next to each other. The detectors are cylindrical, but slightly smaller with 5.1 cm diameter and width. This setup has the minimum possible spatial detector resolution of a Compton camera, as the two detectors are spatially unresolved and act as one voxel each. This limited data space prohibits two-dimensional reconstruction. The purpose of this minimal setup is to test the viability of a bidirectional Compton camera, as both detectors serve as scatterer and absorber at the same time, which is typically avioded due to the additional uncertainty in the model, see Section 3.4.

The measurements of both RSL7 and RSL2 are registered by electronics at the back of the photomultiplier tube. The CoP value is converted to energy values in keV by manual calibration using small calibration sources like Caesium-137. In the case of RSL7, the calibration source is first attached to the center $CeBr_3$ detector, which is calibrated using its absorption lines. After that, the source is attached to each of the PVT scintillators, and the coincidence spectrum of PVT and $CeBr_3$ is used for calibration of the plastic detectors.

## 3.2. Modeling the Detector Pair Response

The data acquired by the modeled cameras consists of coincidental energy depositions: If, within short time, it registers two interactions in different detectors, it triggers a write-out of the energy values into a list of events. Each coincidence measurement is therefore a tuple $(i_A, i_B, E_A, E_B)$, where $1 \leq i_A, i_B \leq K$ are indices of detectors in which the respective energy values $0 \leq E_A, E_B \leq E_{\text{max}}$ are deposited. Modeling the forward operator consists of formulating the likelihood of each measurement tuple given a detector setup and a spatial activity distribution.

**Single scattering likelihood for fixed interaction points.** The simplest type of coincidence is triggered by a Compton scattering event in the first detector, followed by an absorption event in the second. For such an event, assume that scattering takes place

in the detector $D_{i_A} \subset \mathbb{R}^3$ at some unknown point $\mathbf{x}$, and absorption in $D_{i_B} \subset \mathbb{R}^3$ at $\mathbf{y}$, also unknown[1]. We further assume for now monochromatic radiation by an isotropic source at point $\mathbf{s} \in \mathbb{R}^3$ with emission energy $E_0$. The source activity at position $\mathbf{s}$ is given by the activity distribution $f : \mathbb{R}^3 \to [0, \infty)$. The task of the inverse problem will be to recover $f$ on a suitable submanifold $\Omega \subset \mathbb{R}^3$. In the more complicated scenario of multi-nuclide imaging, $f$ also depends on the radioisotope $n \in \mathcal{N}$ (detaild in Chapter 5) in a predefined nuclide library $\mathcal{N}$, in which case we recover $|\mathcal{N}|$ different activity maps at once.

In the simplest case of complete absorption, the full remaining photon energy is deposited in $D_{i_B}$, implying $E_B = E_0 - E_A$. For any single photon emitted at $\mathbf{s}$, several physical factors influence the probability of such a coincidence. Neglecting some influences like refraction as well as attenuation and scattering outside of the detector, the most important factors in the likelihood are:

- The probability that the photon is emitted in the direction of a small volume element $\mathrm{d}\mathbf{x}$ around $\mathbf{x}$. Assume that $\mathrm{d}\mathbf{x}$ is a cube of side length $h_{\mathbf{x}}$ around the center point $\mathbf{x}$, with the coordinate axes normal to its faces. Since $h_{\mathbf{x}} \ll \|\mathbf{s} - \mathbf{x}\|$, in a far field approximation the solid angle element subtended by the cube $\mathrm{d}\mathbf{x}$ at $\mathbf{s}$, denoted $\Omega_{\mathbf{s} \to \mathrm{d}\mathbf{x}}$ is $h_{\mathbf{x}}^2 \|\mathbf{s} - \mathbf{x}\|_1 \|\mathbf{s} - \mathbf{x}\|_2^{-3}$. Hence the probability of the photon being emitted towards $\mathrm{d}\mathbf{x}$ is
$$\frac{h_{\mathbf{x}}^2 \|\mathbf{s} - \mathbf{x}\|_1}{4\pi \|\mathbf{s} - \mathbf{x}\|_2^3} \tag{3.1}$$

- The probability that the photon reaches $\mathrm{d}\mathbf{x}$ without interacting with matter on the trajectory between $\mathbf{s}$ and $\mathbf{x}$. Assume for now that there is no other attenuating material between the source and $D_A$. Then we only have to consider the probability of interaction of the photon between the entry point to the detector and $\mathbf{x}$. If the length that the photon travels inside the detector is $l_1$ and the total linear attenuation coefficient of the detector $D_{i_A}$ at photon energy $E_0$ is $\mu_A(E_0)$, then the probability that the photon reaches $\mathrm{d}\mathbf{x}$ is
$$1 - \exp(-l_1 \mu_A(E_0)). \tag{3.2}$$

- The probability that the photon is Compton scattered in $\mathrm{d}\mathbf{x}$. Modeling $\mathrm{d}\mathbf{x}$ as a cube as before, the mean length travelled by the photon inside the cube is in the far field approximation given by $h_{\mathbf{x}} \|\mathbf{s} - \mathbf{x}\|_2 \|\mathbf{s} - \mathbf{x}\|_1^{-1}$. Hence the probability that the photon is Compton scattered in $\mathrm{d}\mathbf{x}$ is
$$\exp\left(-h_{\mathbf{x}} \frac{\|\mathbf{s} - \mathbf{x}\|_2}{\|\mathbf{s} - \mathbf{x}\|_1} \mu_A^{\mathrm{CS}}(E_0)\right). \tag{3.3}$$

[1] Note that the measurements neither tell us whether a measurement came from scattering or absorption, nor the order of measurements. We make an implicit model assumption here that the chronological event sequence was "emission at $\mathbf{s}$" → "interaction in $D_{i_A}$" → "interaction in $D_{i_B}$". One contribution of this work is to consider the order $\mathbf{s} \to D_{i_B} \to D_{i_A}$ as well, see Section 3.4.

- The probability that the photon scatters in the direction of the absorber, i.e., towards the volume element d$\mathbf{y}$ around $\mathbf{y}$, conditioned on a scatter event taking place. For the scattering angle, denote $\pi - \angle(\mathbf{s}, \mathbf{x}, \mathbf{y}) =: \theta$. It is related to the photon's energy loss $E_A$ via (2.2). The probability can be computed as

$$\frac{\Omega_{\mathbf{x}\to \mathrm{d}\mathbf{y}}}{4\pi\sigma_e^{\mathrm{CS}}} \cdot \frac{\mathrm{d}\sigma^{\mathrm{CS}}}{\mathrm{d}\Omega}(\theta). \tag{3.4}$$

  Here $\mathrm{d}\Omega_{\mathbf{x}\to \mathrm{d}\mathbf{y}}$ is the solid angle element subtended by d$\mathbf{y}$ at $\mathbf{x}$ and the differential and total cross sections for Compton scattering are defined in (2.5), (2.7). Assuming again d$\mathbf{y}$ is a small cube, in a far field approximation as before we have

$$\Omega_{\mathbf{x}\to \mathrm{d}\mathbf{y}} = \frac{h_{\mathbf{y}}^2 \|\mathbf{x}-\mathbf{y}\|_1}{\|\mathbf{x}-\mathbf{y}\|_2^3}.$$

- The probability that the photon reaches d$\mathbf{y}$ without interacting between $\mathbf{x}$ and $\mathbf{y}$. The detectors attenuate the photon propagation between $\mathbf{x}$ and $\mathbf{y}$. Assuming the photon's trajectory has length $l_2$ in the scattering detector with total linear attenuation $\mu_A$ and length $l_3$ in the absorption detector with attenuation $\mu_B$ before hitting $\mathbf{y}$, this probability is given by

$$(1 - \exp(-l_2\mu_A(E_0 - E_A)))(1 - \exp(-l_3\mu_B(E_0 - E_A))). \tag{3.5}$$

- The probability that the photon is absorbed in d$\mathbf{y}$, which is given by

$$\exp\left(-h_{\mathbf{y}} \frac{\|\mathbf{x}-\mathbf{y}\|_2}{\|\mathbf{x}-\mathbf{y}\|_1} \mu_B^{\mathrm{PE}}(E_0 - E_A)\right). \tag{3.6}$$

Having formulated the random factors influencing the measurement, we can now give the expected number of counts for a given pair of points $\mathbf{x}, \mathbf{y}$ and energy $E_A$. Note that the measured energy depends on the source position $\mathbf{s}$ via the cosine of the scattering angle $\cos(\theta)$. To obtain the rate of counts with the same energy, we therefore integrate the probabilities above over the locus of all $\mathbf{s}$ giving the same value $\cos(\theta)$. This set is a cone surface

$$\mathcal{C}(\mathbf{x}, \mathbf{y}, \cos(\theta)) = \left\{ \mathbf{s} : \frac{(\mathbf{x}-\mathbf{s})^T(\mathbf{y}-\mathbf{x})}{\|\mathbf{x}-\mathbf{s}\|\|\mathbf{y}-\mathbf{x}\|} = \cos(\theta) \right\}, \tag{3.7}$$

see also Fig. 3.2. This shows that the forward model is the following *conical Radon transform* of $f$ [65]

$$\mathcal{R}_{\mathrm{c}}^{(1)} f(\mathbf{x}, \mathbf{y}, E_A) = C_2(\mathbf{x}, \mathbf{y}, E_A) \int_{\mathbf{s}\in\mathcal{C}(\mathbf{x},\mathbf{y},\cos(\theta))} f(\mathbf{s}) C_1(\mathbf{s}, \mathbf{x}) \, \mathrm{d}S(\mathbf{s}). \tag{3.8}$$

Here $\mathrm{d}S(\mathbf{s})$ is a surface element on the cone $\mathcal{C}(\mathbf{x}, \mathbf{y}, \cos(\theta))$. $C_1(\mathbf{s}, \mathbf{x})$ is the product of the physical factors (3.1) – (3.3) depending on the source position and $C_2(\mathbf{x}, \mathbf{y}, E_A)$ gathers

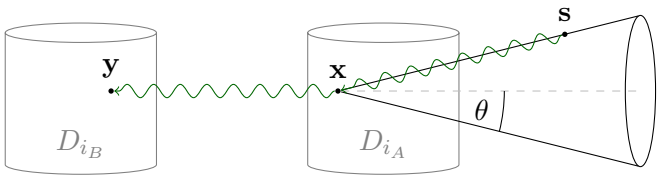



Figure 3.2.: The figure shows a simplified coincidence, with Compton scattering at $\mathbf{x}$ and absorption at $\mathbf{y}$. The locus of possible sources $\mathbf{s}$—and therefore the integration domain of the forward model—is a cone opening with vertex $\mathbf{x}$, central axis $\mathbf{x} - \mathbf{y}$ and opening angle $\theta$ depending on the measurement energy.

the remaining factors defined in (3.4) – (3.6) that depend on the scattering angle (resp. the corresponding energy). The scattering angle $\cos(\theta)$ is related to $E_A$ via the Compton formula (2.2), which can be rearranged to

$$\cos\theta = 1 - m_e c^2 \left( \frac{1}{E_0 - E_A} - \frac{1}{E_0} \right) := \lambda(E_0 - E_A, E_0). \tag{3.9}$$

If $E_0$ is fixed, we will write $\lambda(E_A)$ as well. For a visualization of the integration locus, see the cone in Fig. 3.2.

**The cone transform in the literature.** Since detector geometries and measurement setups vary in applications, the forward model in Compton camera imaging is typically a modification of the cone transform (3.8)—for instance, we will need to integrate (3.8) again in order to account for detectors without spatial resolution. While additional modeling steps complicate analytical inversion, the cone transform itself is well-understood theoretically. One typical approach is to derive filtered backprojection formulas, the first work in this direction being [65]. The first results were impractical, since the operator is inverted from only a subset of the data with cone axes all aligned in parallel, and weight functions $C_1, C_2$ are omitted in several simplifications. The results were improved in [17], albeit only for one-dimensional activity maps $f$. Later on, the works [146, 136, 137, 144, 200, 116, 117, 201] discovered several further relations between (weighted) cone transforms and Fourier and Radon transforms in order to close the gap between theoretical inversion and data from real measurements. Closely related to the backprojection approaches, in another line of works [18, 164, 207, 91] the spectral decomposition of the cone transform was related to spherical harmonics, so that inversion can be reached via truncated expansions. For an overview and further generalizations, we refer to [203, 202].

Analytical inversion formulas are rarely used on real data, due to issues like spatial and energetic resolution, high noise levels and background radiation. In most works mentioned above, simplifying assumptions were made on the detector geometry and availability of data which are not satisfied in practice. Practical reconstructions usually

discretize the forward operator $\mathcal{P}$, leading to a finite dimensional linear system and iterative solution algorithms. This is also the approach taken by the present work, as we found analytical formulas incapable of reconstructing from real data.

**Modeling the unidirectional detector count rate.** In practical measurements, we have to consider both spatial and energetic uncertainty in the detector. In this work, we consider the extreme case of maximal spatial uncertainty, where we do not have the interaction locations $\mathbf{x}, \mathbf{y}$ given. Rather, the camera consists of detector volumes and we only know the interactions took place somewhere in the $D_{i_A}, D_{i_B} \subset \mathbb{R}^3$ involved in a coincidence event.

For the forward model, this implies that the probability of an event anywhere in the detector pair is the expectation of (3.8) over all possible pairs of interaction points in the two detector volumes:

$$\tilde{\mathcal{P}}^{(1)} f(i_A, i_B, E_A) = \int_{D_{i_A}} \int_{D_{i_B}} \mathcal{R}_{\mathrm{c}}^{(1)} f(\mathbf{x}, \mathbf{y}, E_A) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}, \tag{3.10}$$

where we assume a uniform sensitivity of either detector.[2]

Note that the scattering angle varies for different pairs of interaction points in the detector pair, so that $\tilde{\mathcal{P}} f(i_A, i_B, E_A)$ is not an integral over a cone surface anymore. Rather, the integration locus is a union of cones with the same opening angles, but varying vertex points in $D_{i_A}$ and varying central axes.

Due to energetic uncertainty, this model is additionally convolved with a kernel $h_A$ describing the energy resolution of the scintillation material in detector $D_{i_A}$. Let us distinguish between the measurement $E_A$ and the actual energy deposition $\tilde{E}_A$. The expected number of counts for which $D_{i_A}$ *measures* an energy value $E_A$ is

$$\mathcal{P}^{(1)} f(i_A, i_B, E_A) = h_A(\cdot, E_A) * \int_{D_{i_A}} \int_{D_{i_B}} \mathcal{R}_{\mathrm{c}}^{(1)} f(\mathbf{x}, \mathbf{y}, \cdot) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}, \tag{3.11}$$

where $*$ denotes convolution in the energy dimension with a kernel $h_A$. Here $h_A(\tilde{E}_A, \cdot)$ can be interpreted as a probability density function for the measured energy given the true energy $\tilde{E}_A$ that the scattered photon loses in the detector. Equivalently, $h_A(\cdot, E_A)$ is the density function for the true energy loss when the detector measured $E_A$. The probability distribution described by $h_A(\tilde{E}_A, \cdot)$ is often assumed to be Gaussian with standard deviation equal to a constant multiple of $\tilde{E}_A^{1/2}$, see [113, chap. 10. iv], so

$$h_A(\tilde{E}_A, E_A) = \frac{1}{\sqrt{2\pi\eta_A^2 \tilde{E}_A}} \exp\left( -\frac{(E_A - \tilde{E}_A)^2}{2\eta_A^2 \tilde{E}_A} \right), \tag{3.12}$$

[2] In the case of known, non-uniform sensitvities, the integrals would reveive additional weights $\omega_A(\mathbf{x})\omega_B(\mathbf{y})$.

where $\eta_A$ is a constant characteristic to the detector material. Since the convolution with a Gaussian kernel has severe effects on the model's ill-posedness, a good energetic resolution is vital for reliable source reconstruction. Geometrically, the convolution in energy means a convolution over different cone opening angles.

Before moving on, we also remark that we formulated the operators $\mathcal{R}_c^{(1)}$ and $\tilde{P}$ such that they depend on the energy $E_A$. Assuming full absorption, this is of course equivalent to parametrizing it depending on $E_B = E_0 - E_A$. Due to the energy uncertainty (3.12), the forward model $\mathcal{P}$ changes depending on whether we use energy values from $D_{i_A}$ or $D_{i_B}$, since the two detectors might have different energy resolution. When reconstructing from only the fully absorbed events of a measurement, it is therefore important to rely on the energy values from the more accurate detector. In a setup with multiple cheaper PVT detectors and a few better resolved scintillation detectors from a crystal like $CeBr_3$, we have $\eta_{\text{PVT}} > \eta_{\text{CeBr}_3}$ and should therefore opt for the $CeBr_3$ energy measurements.

## 3.3. Poisson Statistics

In photonic imaging modalities like Compton camera imaging, the measurements typically stem from charge-coupled device (CCD) cameras, which read out single measurement events. The detectors of the Compton cameras we consider are instances of such CCD cameras. By the very nature of this quantized measurement of photons, the resulting data $g$ is an element of a countable data space. Apart from background and read-out errors, the quantization may present a main source of noise in many such modalities, e.g., positron emission tomography (PET), single photon emission computed tomography (SPECT), phase retrieval or Compton camera imaging.

For CCD cameras, the noise process can be modelled as follows, see [94]: Given an instance of the unknown $f$, there is a spatial photon density function $\mathcal{A}(f) \in L^1(\mathcal{M})$, where $\mathcal{M} \subset \mathbb{R}^3$ is the measurement manifold, for instance the outer product of detector volumes[3]. In the context of Compton cameras, $\mathcal{A}(f)$ can be approximated by the cone transform model (3.2). During a measurement, denote by $\mathbf{z}_1, \dots, \mathbf{z}_N$ the interaction locations (for instance a tuple $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ in Compton camera model) where photons were measured in $N$ events. Due to the mutual independence of radioactive decay and photon interaction in $\mathcal{M}$, it can be shown that the measure $S = \sum_i \delta_{\mathbf{z}_i}$ is a Poisson point process on $\mathcal{M}$. For any detector volume $\Theta \subset \mathcal{M}$, the random variable $S(\Theta)$ is Poisson distributed with parameter $\lambda = \int_\Theta \mathcal{A}(f)(\mathbf{z})\,\mathrm{d}m$, we write $S(\Theta) \sim \text{Pois}(\lambda)$. It has probability mass function

$$\mathbb{P}[S(\Theta) = k] = \frac{e^{-\lambda}\lambda^k}{k!}.$$

[3] For the simplicity of demonstrating of how Poisson statistics emerge in the measurement, we neglect potential time and energy measurement dimensions. The model can be extended to higher-dimensional data spaces.

The discretization from $\mathcal{A}(f)$ to the forward model $A(f)$ happens since detectors have finite spatial resolution. $\mathcal{M}$ is written as the union $\mathcal{M} = \bigcup_{j=1}^{J} \mathcal{M}_j$, where counts within a single $M_j$ are binned together. $M_j$ are unresolved voxels or volumes within the detector setup, and the expected number of counts in such a detector voxel is given by

$$\mathbb{E}[S(\mathcal{M}_j)] = \int_{\mathcal{M}_j} \mathcal{A}(f)(\mathbf{z}) \, \mathrm{d}\mathbf{z} = A_j(f),$$

where $A_j$ is the $j$-th row of the discretized model matrix $A$. The number of counts in a measurement voxel or bin is precisely the measurement of the inverse problem. Since the counts in the different voxels are independent of each other, this shows that the probability density of observing some data $g$ given $f$ is given by

$$\rho_{\text{like}}(g \mid f) = \prod_{j=1}^{J} \frac{e^{-A_j(f)} A_j(f)^{g_j}}{g_j!}. \tag{3.13}$$

This is the likelihood density of an inverse problem with Poisson data—we will use the likelihood distribution in the Bayesian formulation of image reconstruction problems. From now on, we also denote the independent outer product of Possion distributions in its usual vector notation $g \sim \text{Pois}(A(f))$. The frequently required negative log-likelihood can be computed to be

$$\begin{aligned} -\log \rho_{\text{like}}(g \mid f) &= \sum_{j=1}^{J} A_j(f) - g_j \log A_j(f) + \text{const.} \\ &= \sum_{j=1}^{J} A_j(f) - g_j + g_j \log \left( \frac{g_j}{A_j(f)} \right) + \text{const.}, \end{aligned} \tag{3.14}$$

where constant terms do not depend on $f$. The expression in the second line is frequently referred to as Kullback–Leibler (KL) divergence, since it is a special case of this divergence of two outer products of Poisson distributions with coordinate-wise means $g$ and $A(f)$. We will properly define the KL divergence of probability distributions in a later section.

A special property of the one-dimensional Poisson distribution is that it depends on only the single parameter $\lambda$. It has mean and variance $\lambda$, so the signal-to-noise ratio (SNR) of a Poisson distribution is $\sqrt{\lambda}$. In Fig. 3.3, we see how this affects measurement design: In this scintillation detector measurement, we have $\lambda_j = A_j(f)$ with a linear model $A_j$ in every energy bin $j$. We plot two measurements from two differently scaled versions of $A_j$, created by taking a long and a short measurement of the same source. It is clearly visible that the data from a long measurement time has better SNR than the short measurements, where some weaker peaks in the spectrum become indistinguishable from the inherent statistical variance of the counts.

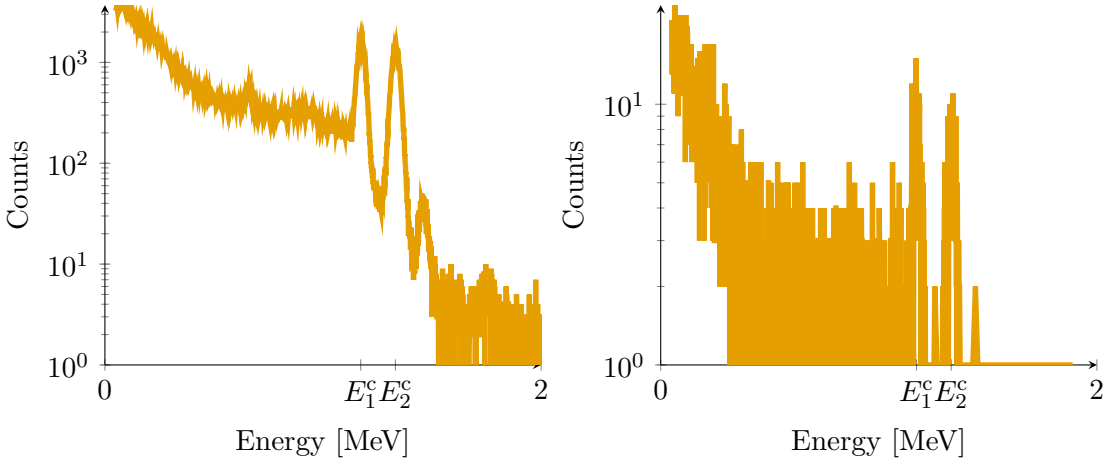



Figure 3.3.: A standard example of discrete-natured data: We show a spectroscopic measurement of a radioactive Cobalt-60 source with characteristic peaks at $E_1^c = 1.17\,\text{MeV}$ and $E_2^c = 1.33\,\text{MeV}$, left is the full measurement and right a random 0.5% of the counts. The pure data consists of single counts, which are then binned into energy intervals on the horizontal axis. By nature of radioactive decay, the count in each bin is Poisson-distributed. Since the signal-to-noise-ratio of a Poisson distribution with mean $\lambda$ grows as $\sqrt{\lambda}$, longer measurements (left) with larger mean counts are less noisy then short measurements (right).

## 3.4. Model Extension to Bidirectional Compton Cameras

So far, we always assumed that we can determine the chronological order of the two registered interactions. In other words, given two interaction points, $\mathbf{x}$ and $\mathbf{y}$, we knew that scattering took place at $\mathbf{x}$ and absorption took place at $\mathbf{y}$. Therefore, we knew the scattering angle and, consequently, the integration cone. For many Compton camera geometries, this is true in practice—at least approximately—as several design factors influence the likelihood of one direction over the other. For example, cameras may be collimated or shielded around one detector. As we have seen in Fig. 2.3, forward scattering with small scattering angles $\theta$ is much more likely than backward scattering with $\theta > 90°$ in certain energy regions. Thus, the orientation of a detector pair towards the source region can strongly influence the count rates. The detector material choice also plays a role. Certain detectors, such as polymerized scintillators like PVT, mostly scatter gamma radiation and rarely absorb it due to their low $Z$-value, see (2.10). We will call such setups, in which the coincidence direction is always certain, *unidirectional* geometries.

**Design and count rate considerations.** Unidirectional designs, while providing directional certainty, inherently limit the total count rate by restricting scatter angles or

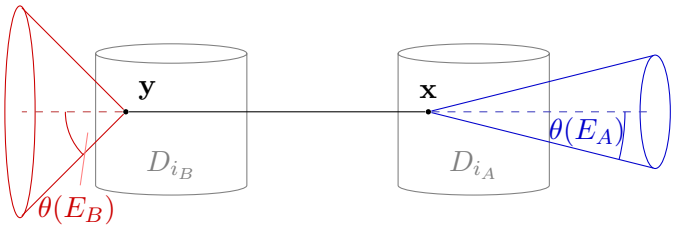



Figure 3.4.: In bidirectional Compton cameras, the chronological order and type of interactions are generally unknown. The forward model therefore generally consists of a weighted sum of two conical Radon transforms, a source on either of these cones could have caused a given measurement. In the sum, the cones are weighted by their respective probabilities $C_1$ and $C_2$, so that the directionality of the coincidence might become more or less ambiguous.

requiring specific material choices. By accepting the directional ambiguity in bidirectional geometries, detectors can be placed relative to each other with more degrees of freedom and collimation is unnecessary. This can increase the field of view and count rates, potentially improving the signal-to-noise ratio given the Poisson statistics of the measurements.

**Modified forward model.** The extension to bidirectional Compton cameras naturally changes the forward model. Consider again a coincidence of events in the Compton camera triggering some measurement output tuple $(i_A, i_B, E_A, E_B)$ . In the bidirectional setup, we can (usually) not distinguish, neither from the data nor from the detector setup, whether a given coincidence was triggered by a scattering event in $D_{i_A}$ and an absorption in $D_{i_B}$ or vice versa. Since both alternatives are possible, the forward model has to account for this additional uncertainty.

If $E_A$ was lost in a scattering event, then the Compton angle is given by $\cos(\theta) = \lambda(E_0 - E_A, E_0)$ as before in (3.9). Note that this is only possible if $E_A$ is a valid energy loss of a Compton scattering event. It needs to satisfy the bounds $E_A \leq 2E_0^2/(m_e c^2 + 2E_0)$, otherwise we would have $\lambda(E_0 - E_A, E_0) \notin [-1, 1]$, see also (2.4). The corresponding integration surface is the cone with vertex $\mathbf{x}$, axis $\mathbf{y} - \mathbf{x}$ and angle $\theta(E_A)$. As before, the expected count rate for given interaction points $\mathbf{x}, \mathbf{y}$ is $\mathcal{R}_c^{(1)} f(\mathbf{x}, \mathbf{y}, E_A)$, defined in (3.8). If the situation was reversed, then the interaction at $\mathbf{y}$ with energy deposition $E_B$ had to be a Compton scattering, if $E_B$ is a valid energy out of a Compton scattering. In that case, the angle was given by $\cos(\theta) = \lambda(E_0 - E_B, E_0)$ and the cone has vertex $\mathbf{y}$ and opens towards the flipped axis $\mathbf{y} - \mathbf{x}$. By the way we formulated the forward model, the corresponding expected counts are given by $\mathcal{R}_c^{(1)} f(\mathbf{y}, \mathbf{x}, E_B)$. Since we are still assuming monochromatic gamma rays and a full absorption event, we can replace $E_B = E_0 - E_A$

and obtain the modified, bidirectional forward model

$$\begin{aligned}\mathcal{R}_{\mathrm{bi}}^{(1)} f(\mathbf{x}, \mathbf{y}, E_A) = \mathbf{1}_{[0,E_{\max}]}(E_A)\mathcal{R}_{\mathrm{c}}^{(1)} f(\mathbf{x}, \mathbf{y}, E_A) \\ + \mathbf{1}_{[0,E_{\max}]}(E_0 - E_A)\mathcal{R}_{\mathrm{c}}^{(1)} f(\mathbf{y}, \mathbf{x}, E_0 - E_A),\end{aligned} \tag{3.15}$$

where $E_{\max} = \frac{2E_0^2}{m_e c^2 + 2E_0}$ is the largest possible energy loss in a Compton scattering event with initial photon energy $E_0$, corresponding to full backscattering with $\theta = 180°$. We will henceforth call $\mathcal{R}_{\mathrm{bi}}^{(1)}$ a biconical Radon transform. Note that the sum of the two unidirectional cone transforms is weighted. Due to the energy-dependent likelihood terms and indicator weights in $\mathcal{R}_{\mathrm{c}}^{(1)}$, some scattering energies are more likely than others (see Fig. 2.3), and some are impossible. Depending on the geometry and energy of a measurement tuple, it can be very ambiguous, or have one unlikely and one likely direction.

The considerations of positional and energy resolution are unchanged by the bidirectional model. The possible pairs of interaction points cover a whole detector pair in our case of no spatial resolution, implying an integral over the outer product of the volumes $D_{i_A}, D_{i_B}$. Regarding energy uncertainty, it is still advisable to use the energy measurement $E$ from the detector with better energy resolution, i.e., the one with smaller constant $\eta$ in the model (3.12). Assuming we pick $E_A$, the expected count rate under this model is given by

$$\mathcal{P}_{\mathrm{bi}}^{(1)} f(i_A, i_B, E_A) = h_A(\cdot, E_A) * \int_{D_{i_A}} \int_{D_{i_B}} \mathcal{R}_{\mathrm{bi}}^{(1)} f(\mathbf{x}, \mathbf{y}, \cdot)\, \mathrm{d}\mathbf{y}\, \mathrm{d}\mathbf{x}. \tag{3.16}$$

The bidirectional model is a natural generalization of the formulations for unidirectional cameras with certain event sequence. To our knowledge, the bidirectional model presented here is novel. Standard approaches attempt "event sequence reconstruction" before source reconstruction, typically choosing the most likely direction or discarding ambiguous events [125]. Our model $\mathcal{P}_{\mathrm{bi}}^{(1)}$ naturally accounts for both directional probabilities in the forward model itself, which from an information-theoretic perspective should reduce model error compared to pre-selection approaches by utilizing all available information.

## 3.5. Higher-Order Scatter Correction

A challenge in Compton camera image reconstruction for nuclear decommissioning applications is the significant level of noise in the measured data. Some of this noise is inherently unavoidable, for instance, coincidences that are caused by background radiation[4]. Even apart from the background, several types of events caused by the source's

[4]Where "background" here not only refers to the natural background, but also to other radioisotopes in the power plant that are not reconstructed.

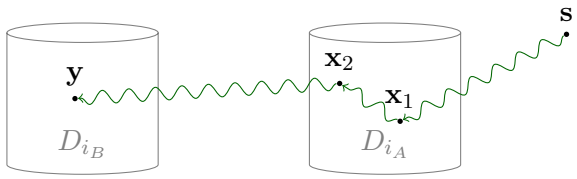



Figure 3.5.: Sketch of a multiply scattered coincidence event.

gamma photons are not covered by the forward model (3.11), but may trigger coincidence measurements.

To understand these events, we go back to the formulation of the forward model, specifically the attenuation factors (3.5) between the scattering and the absorption point in the idealized coincidence model. The first attenuation factor quantifies the number of photons that interacts again in the same detector after the first scattering event. Nothing prevents photons from scattering a second time, and due to the factor $l_2$ this becomes more likely with increasing detector size. The double-scattered photon may still exit the first detector and cause a coincidence by interacting in another detector, see the sketch in Fig. 3.5. The same principle holds for interactions in the second detector. If the first interaction in the second detector is not an absorption, but a second scattering, then the resulting photon may interact again, which is also not reflected by (3.11). Unfortunately for the model, the spatial and time resolution of detectors may not be sufficient to distinguish two scattering events in a single detector voxel. If this is the case, the output $E_A$ in Fig. 3.5 can be expected to be approximately

$$E_A = \Delta E_1 + \Delta E_2, \tag{3.17}$$

where $\Delta E_i$ is the energy deposited in the scattering at $\mathbf{x}_i$. Coincidences of this kind induce model errors, since the energy $E_A$ now does not correspond to a correct cone opening angle to locate $\mathbf{s}$.

The relative frequency of multiply scattered events among all recorded coincidences depends on multiple factors. With increasing size of the detectors, multiple interactions become more likely. Since the attenuation due to Compton scattering is the product of electron density and total cross section (see 2.7), it also depends on the detector material and the photon energy $E_0$. For a fixed camera design, a digital twin of the detector setup and source points can be created in a Monte Carlo simulation toolbox like GEANT4 [1, 4, 5] or FLUKA [79, 31] in order to simulate measurement data. Simulated data allows to track each photon interaction separately, so the multiplicity of interactions within a detector can be examined, see Fig. 3.6. With FLUKA simulations for a camera setup like RSL7, we see that multiple scattering is—in both detectors—more common than single scattering.

The forward model (3.11) ignores modeling the data contribution of multiple scattered photons. We develop a higher-order scatter correction [42] that extends the standard

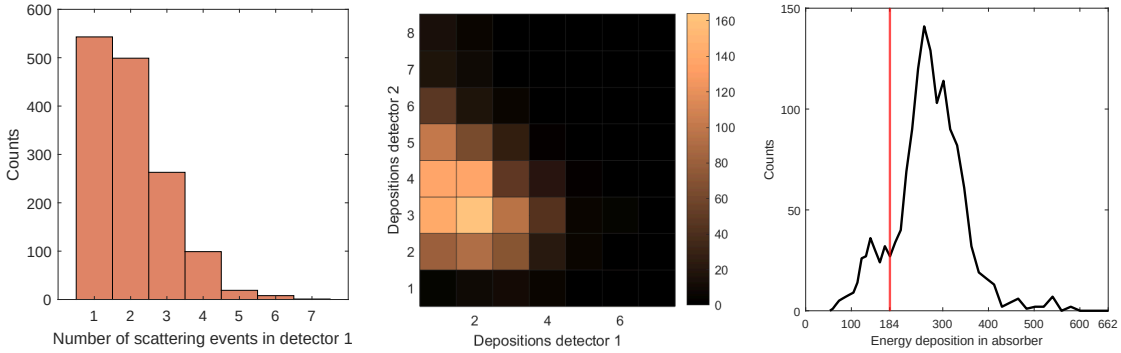



Figure 3.6.: We show coincidence data of a single detector pair in RSL7 from a simulated $^{137}$Cs source with perfect energy resolution and no background, figure from [42]. Left: coincidences grouped by number of scattering events in the first detector. Middle: coincidences grouped by number of interactions in both detectors. Right: Energy deposited in the second detector by fully absorbed photons, showing the model error. For single scattering events, the minimum possible absorbed energy would physically be at least 184 keV corresponding to a scattering angle of 180°, indicated by the red line. The fact that there are many counts around and below the red line indicates a mismatch due to multiple scattering.

single-scatter model to account for double Compton scattering events in the first detector. Twice scattered events allow the same considerations to describe the likelihood of a measurement as the ones in (3.1) – (3.6) for the single scattering model. The necessary changes are additional terms to describe the probability of scattering a second time, and keeping track of the relation between the three energy depositions. Suppose the three interactions take place at $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{y}$ as in Fig. 3.5. The scattering angles are $\theta_1$ and $\theta_2$ and the photon between scattering events has energy $E_1 = E_0 - \Delta E_1$, after the second scattering $E_2 = E_1 - \Delta E_2$. The energy registered in the scatterer $D_{i_A}$ is $E_A = \Delta E_1 + \Delta E_2$. We then assume complete absorption at $\mathbf{y} \in D_{i_B}$, so that $E_B = E_2 = E_0 - E_A$. Since the first measured energy $E_A$ is the sum of both energy depositions during the scattering events, we do not have access to the intermediate energy $E_1$ or the individual angles $\theta_1, \theta_2$. Applying the Compton formula (3.9) twice and summing up gives the relation

$$\cos\theta_1 + \cos\theta_2 = \lambda(E_1, E_0) + \lambda(E_2, E_1) = 2 - m_e c^2 \left( \frac{1}{E_2} - \frac{1}{E_0} \right) \tag{3.18}$$

between the two scattering angles and the measured energies, meaning we have access to the value $\cos\theta_1 + \cos\theta_2$. Geometrically, the energies and interaction positions are now related via two cones. We have $\mathbf{s} \in \mathcal{C}(\mathbf{x}_1, \mathbf{x}_2, \lambda(E_1, E_0))$ with the intermediate energy $E_1$, and $\mathbf{x}_1 \in \mathcal{C}(\mathbf{x}_2, \mathbf{y}, \lambda(E_2, E_1))$. The expected count rate at absorption energy $E_B = E_2$, for which the last two interaction positions were $\mathbf{x}_2, \mathbf{y}$ is therefore a double integral over

both cone surfaces and all possible intermediate energies $E_1 \in (E_2, E_0)$:

$$\mathcal{R}_{\mathrm{c}}^{(2)} f(\mathbf{x}_2, \mathbf{y}, E_A) = C_2(\mathbf{x}_2, \mathbf{y}, E_A) \int_{E_0 - E_A}^{E_0} \int_{\mathcal{C}(\mathbf{x}_2, \mathbf{y}, \lambda(E_0 - E_A, E_1))} \tilde{C}_1(\mathbf{x}_1, \mathbf{x}_2, E_1) \cdot \int_{\mathcal{C}(\mathbf{x}_1, \mathbf{x}_2, \lambda(E_1, E_0))} C_1(\mathbf{s}, \mathbf{x}_1) f(\mathbf{s}) \, \mathrm{d}S(\mathbf{s}) \, \mathrm{d}S(\mathbf{x}_1) \, \mathrm{d}E_1. \tag{3.19}$$

The new factor $\tilde{C}_1$ models the differential cross section of the first scattering event, attenuation in $D_{i_A}$ between $\mathbf{x}_1$ and $\mathbf{x}_2$ and the probability to scatter again at $\mathbf{x}_2$, formulated in a similar way as the factors (3.4), (3.5) and (3.3).

In order to formulate the likelihood of measurement tuples $(i_A, i_B, E_A, E_B)$, the last steps are now the same as in Section 3.2. We integrate the sum of first and second order over the spatial and the energy uncertainty. The spatial uncertainty covers, in our unresolved case, all points $(\mathbf{x}_2, \mathbf{y}) \in D_{i_A} \times D_{i_B}$. This leads to the integral

$$\int_{D_{i_A}} \int_{D_{i_B}} \mathcal{R}_{\mathrm{c}}^{(2)} f(\mathbf{x}, \mathbf{y}, E_A) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}. \tag{3.20}$$

Assuming for simplicity the same energy uncertainty model (3.12) for single and double depositions, the model correction for double scattering with uncertain energy measurements is

$$\mathcal{P}^{(2)} f(i_A, i_B, E_A) = h_A(\cdot, E_A) * \int_{D_{i_A}} \int_{D_{i_B}} \mathcal{R}_{\mathrm{c}}^{(2)} f(\mathbf{x}, \mathbf{y}, \cdot) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}. \tag{3.21}$$

Alternatively, we can again use the measured energy $E_B$ if the detector energy resolution suggests that this is the more accurate value, replacing $h_A$ by $h_B$.

Note that we can generalize the double scattering correction to the bidirectional case as well. To the terms in (3.15), we simply add the double scattering model in both possible event directions. The only detail that changes for the correction term is the maximum energy $E_{\max}$ deposited in the first detector, as this is now the energy loss corresponding to two consecutive 180° backscattering events.

To obtain the corrected model, by linearity of integration and convolution, we can compute $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ separately and obtain the corrected forward model $\mathcal{P}^{(1)} + \mathcal{P}^{(2)}$ for multiply scattered data. The correction significantly reduces the model mismatch between realistic data and the forward model. To show that, we reconsider the example from Fig. 3.6, i.e., a single Caesium-137 point source in a simulated PVT-$CeBr_3$ detector pair with perfect energy resolution. In Fig. 3.7, we show an overlay of the simulated spectrum, and the predicted spectra of the analytical forward models with first order $\mathcal{P}^{(1)}) f(i_A, i_B, \cdot)$, and second order corrected $\mathcal{P}^{(2)} f(i_A, i_B, \cdot)$.

As a final comment on scatter corrected models before moving on, we point out that $\mathcal{P}^{(1)} + \mathcal{P}^{(2)}$ improves over $\mathcal{P}^{(1)}$ only in the number of scattering events in the first detector. As the middle pane of Fig. 3.6 shows, multiple depositions can of course also happen

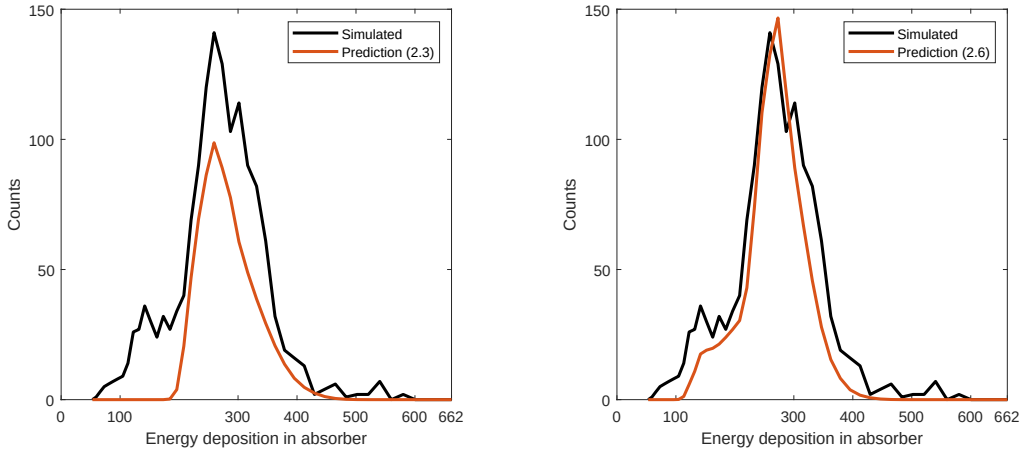



Figure 3.7.: Figure from [42]. For the simulated Caesium-137 point source from Fig. 3.6, we compare the modeled detector response of the single-scattering model $\mathcal{P}^{(1)}f$ (left) with the double-scattering model $(\mathcal{P}^{(1)} + \mathcal{P}^{(2)})f$ (right). Since the largest amount of photons scatters either once or twice in the PVT scatter detector, the double scattering model shows good agreement with the simulated spectrum of energy measurements in the $CeBr_3$ detector.

in the second detector. However, we expect the resulting model error to be smaller overall. Assume a photon scatters in $D_{i_A}$, depositing $E_A$, and then interacts multiple times in $D_{i_B}$, the last event being an absorption. Then the overall deposited energy will be $E_B = E_0 - E_A$, independent of the number of interactions in $D_{i_B}$. The estimated scattering angle is not affected, and the source still lies on a cone surface with opening angle $\theta(E_A)$ and vertex at $\mathbf{x}$. The cone axis $\mathbf{x} - \mathbf{y}$ receives a slight model error, since a corrected model would need to account for the skewed probability distribution in points $\mathbf{y}$ being the first interaction point of multiple interactions in $D_{i_B}$. This error is assumed to be small and seems to have a minor effect in comparisons with practical simulation studies.

## 3.6. Model Errors

Apart from bidirectionality and multiple scattering events, there are multiple persisting error sources that degrade the measurements in Compton camera imaging.

**Coincidences of more than two detectors** The idea that coincidences are triggered whenever two detectors register energy depositions within short time can be generalized to more than two detectors. Depending on the detector geometry and photon energy, these interactions can be fairly common. Factors that promote coincidences with three or more interactions in different detectors are a high photon energy and small detectors

with large Compton scattering attenuation compared to photoelectric absorption. In higher energy regions 2-20 MeV, these interactions have been explicitly considered in the reconstruction in order to increase event rates [170, 154]. A detailed formulation of the modified analytical model for triplet interactions can be found in [181]. A nice property of triplet interactions is the fact that unknown source energies can, in theory, be reconstructed exactly from three interactions, as three interaction points and the first two scatter energy depositions uniquely determine $E_0$ via the Compton formula. In practice, this ability is limited by resolution of positions and energy. The main shortcoming of triplet interactions is the bidirectionality problem. Instead of two possible event sequences, there are six possible paths, complexifying the forward model considerably. In [125], the authors aim to track the sequence of such interactions by extracting the most likely of the six paths from the photon path probabilities in the forward model. In simulation studies for the RSL7 camera, we found that triplet interactions make up a negligible amount below 0.5% of total events, so the electronics of the camera were set up to discard such events.

**Doppler Broadening** Another main source of errors that we have not mentioned in the modeling is Doppler broadening. Strictly speaking, the Compton formula gives the energy of the photon after the interaction. By energy conservation, the electron receives the difference $\Delta E$ as kinetic energy. The model that a camera then measures $\Delta E$, however, is based on the assumption that the electron had no momentum before the interaction, which is not true in practice. Modeling the additional uncertainty generally broadens the peak of observed energies further by convolving with another nonlinear kernel in the energy. The relevance of this additional modeling step has been described in different ways in the literature. While [80] states that "modeling [Doppler broadening] can play a decisive role", [53] concludes that "neither Doppler broadening nor polarization had any significant effect on the sensitivity of the camera". A simple approximate modeling choice can be to increase the energy resolution constant $\eta$ [91].

**Wrongly Triggered Coincidences** In studies with the measurement system described in Section 3.1, a major source of errors are wrongly triggered coincidences, see Fig. 2.7. These measurements are especially prominent in high-activity environments, a common situation in power plants. When there is large background radiation, typically a stream of low-energy (10-80 keV) depositions is recorded in plastic scintillators, orders of magnitude more frequent than the total coincidence rate. For RSL7 measurements, a common type of measurement then consists of a complete or near-complete absorption in a $CeBr_3$ detector and some time-coincidental but unrelated low-energy interaction in a second detector. Due to energy uncertainty, the measurement is indistinguishable from a true coincidence with complete absorption after a small-angle scattering event. These wrongly triggered coincidences significantly reduce the signal-to-noise ratio. Reliable reconstruction was only possible after discarding all events from certain regions of the measurement space. For instance, for Caesium-137, discarding scatter detector

depositions lower than 50 keV reduces the de-facto field of view of the camera to a cone with 69.5° half opening angle around the camera's symmetry axis. This represents a fundamental limitation of the current hardware that future detector designs with better spatial or temporal resolution could mitigate.

## 3.7. Summary of Forward Models

In this chapter, we developed several forward models for the monochromatic Compton camera imaging problem. We saw that photons' interaction paths in the camera can be assumed mostly independent and their measurement probability proportional to the emission activity, leading to linear forward models. For the setups considered in this work, the data models are of the form

$$g(z) = \int_\Omega k(z, \mathbf{s}) f(\mathbf{s}) \, \mathrm{d}\mathbf{s}. \tag{3.22}$$

Here, $g(z)$ is the expected number of photopeak counts in the data bin $z = (i_A, i_B, E_A)$ gathering spatial bins $i_A, i_B$ for both detectors and an energy measurement bin $E_A$. The integration domain $\Omega \subset \mathbb{R}^3$ is the area of emission activity. For the kernel $k$, we proposed three different choices, varying in model complexity and based on requirements of the camera setup.

**Unidirectional single scattering model.** The simplest model is the single scattering model $\mathcal{P}^{(1)}$ given in (3.11). It accounts for photopeak counts of photons that are scattered once in a detector $D_{i_A}$ with energy deposition $E_A$ and then absorbed in another detector $D_{i_B}$. Correcting for the spatial uncertainty in unresolved detectors and an energy noise model (3.12), the integration kernel is given by

$$\begin{aligned} k^{(1)}(i_A, i_B, E_A, \mathbf{s}) &= h_A(\cdot, E_A) * \tilde{k}^{(1)}(i_A, i_B, \cdot, \mathbf{s}) \\ \tilde{k}^{(1)}(i_A, i_B, \tilde{E}_A, \mathbf{s}) &= \int_{D_{i_A}} \int_{D_{i_B}} C(\mathbf{s}, \mathbf{x}, \mathbf{y}, \tilde{E}_A) \delta(\sigma(\mathbf{s}, \mathbf{x}, \mathbf{y}) - \lambda(\tilde{E}_A)) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}, \end{aligned}$$

with constants (3.1) – (3.6) gathered in $C(\mathbf{s}, \mathbf{x}, \mathbf{y}, \tilde{E}_A) = C_2(\mathbf{x}, \mathbf{y}, \tilde{E}_A) C_1(\mathbf{s}, \mathbf{x})$. We further denoted

$$\sigma(\mathbf{s}, \mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} - \mathbf{s})^T (\mathbf{y} - \mathbf{x})}{\|\mathbf{x} - \mathbf{s}\| \|\mathbf{y} - \mathbf{x}\|}$$

for the cosine of the angle spanned by the interaction points and $\lambda(\tilde{E}_A)$ for the scattering angle implied by the deposited energy via Compton's formula, see (3.9).

**Bidirectional model.** The second model we presented addresses camera setups with multiple absorption detectors, in which we cannot tell the nature of interactions from a

measurement tuple $(i_A, i_B, E_A)$. The single scattering, bidirectional kernel is

$$\begin{aligned} k_{\text{bi}}^{(1)}(i_A, i_B, E_A, \mathbf{s}) &= h_A(\cdot, E_A) * \tilde{k}_{\text{bi}}^{(1)}(i_A, i_B, \cdot, \mathbf{s}) \\ \tilde{k}_{\text{bi}}^{(1)}(i_A, i_B, \tilde{E}_A, \mathbf{s}) &= \iint (C_{\text{bi}}(\mathbf{s}, \mathbf{x}, \mathbf{y}, \tilde{E}_A)\delta(\sigma(\mathbf{s}, \mathbf{x}, \mathbf{y}) - \lambda(\tilde{E}_A)) \\ &\qquad + C_{\text{bi}}(\mathbf{s}, \mathbf{y}, \mathbf{x}, E_0 - \tilde{E}_A)\delta(\sigma(\mathbf{s}, \mathbf{y}, \mathbf{x}) - \lambda(E_0 - \tilde{E}_A)))\, \mathrm{d}\mathbf{y}\, \mathrm{d}\mathbf{x}. \end{aligned}$$

We account for the fact that only certain energy ranges are valid Compton scattering angles by setting $C_{\text{bi}}(\mathbf{s}, \mathbf{x}, \mathbf{y}, \tilde{E}_A) = \mathbf{1}_{[0, E_{\max}]}(\tilde{E}_A) C(\mathbf{s}, \mathbf{x}, \mathbf{y}, \tilde{E}_A)$, where $E_{\max} = \frac{2E_0^2}{m_e c^2 + 2E_0}$.

**Double scattering corrected model.** The final proposed model extension accounts for double scattering in the scatter detector, a prominent effect for monolithic detectors and moderate gamma ray energies. The measurement is again a tuple $z = (i_A, i_B, E_A)$, where $E_A$ is the total energy loss of two subsequent scattering events in the same detector $D_{i_A}$. Due to mutual independence of events, the correction is additive, meaning the corrected, unidirectional model is defined by the kernel

$$k^{(1,2)} = k^{(1)} + k^{(2)}$$

where the model correction $k^{(2)}$ is

$$\begin{aligned} k^{(2)}(i_A, i_B, E_A, \mathbf{s}) &= h_A(\cdot, E_A) * \tilde{k}^{(2)}(i_A, i_B, \cdot, \mathbf{s}) \\ \tilde{k}^{(2)}(i_A, i_B, \tilde{E}_A, \mathbf{s}) &= \iiiint_{\mathcal{D}} C^{(2)}(\mathbf{s}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, E_1, \tilde{E}_A)\delta(\sigma(\mathbf{s}, \mathbf{x}_1, \mathbf{x}_2) - \lambda(E_1, E_0)) \\ &\qquad \cdot \delta(\sigma(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) - \lambda(E_0 - \tilde{E}_A, E_0))\, \mathrm{d}(\mathbf{y}, \mathbf{x}_2, \mathbf{x}_1, E_1), \end{aligned}$$

with $\lambda$ as in (3.9) and the weight

$$C^{(2)}(\mathbf{s}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, E_1, \tilde{E}_A) = C_2(\mathbf{x}_2, \mathbf{y}, \tilde{E}_A)\tilde{C}_1(\mathbf{x}_1, \mathbf{x}_2, E_1) C_1(\mathbf{s}, \mathbf{x}_1),$$

see (3.20). The integration domain $\mathcal{D} = [E_0 - \tilde{E}_A, E_0] \times D_{i_A} \times D_{i_A} \times D_{i_B}$ covers all intermediate energies $E_1$ and interaction locations $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{y}$.

In a similar way to the above, it is straightforward to formulate a scatter-corrected kernel $k_{\text{bi}}^{(1)} + k_{\text{bi}}^{(2)}$ for bidirectional cameras, which we omit here. The forward models developed in this chapter—particularly the bidirectional formulation and higher-order scatter corrections—address significant limitations of standard Compton camera models for decommissioning applications. These extensions are validated experimentally in Chapter 5, where we demonstrate improved reconstruction quality compared to first-order unidirectional models.

# Chapter 4

# Bayesian Modeling and Computation

In this chapter, we cover the preliminaries on Bayesian inference and collect necessary tools for its application to imaging inverse problems. In older literature on inverse problems, the reconstruction task was usually approached with a strong focus on the true solution. Within this framework, a successful method would be an efficient and stable algorithm that retrieves this deterministic ground truth or a close approximation to it. The Bayesian viewpoint reframes this concept of a solution by treating only the observed data as a fixed instantiation and all unknown quantities in the problem as random variables. The measurement errors are not subsumed into an unknown noise term, but rather combined with the forward model to form a conditional likelihood distribution with density $\rho_{\text{like}}(y \,|\, x)$ for the measured data $y$. The unknown quantity $x$ that we wish to recover has a marginal distribution with density $\rho_{\text{prior}}(x)$, which is available before the measurement and thus referred to as the prior distribution. Upon making a measurement $y$, the distribution of the unknown $x$ changes by conditioning on $y$. The resulting measure $\rho_{\text{post}}$ is called the posterior distribution. We assume from now on that our problem is finite dimensional, meaning $x \in \mathbb{R}^d$, $y \in \mathbb{R}^m$ for some $d, m \geq 1$ and all involved densities exist and are understood with respect to the Lebesgue measure. Then, by Bayes' law, the posterior density[1] is given by

$$\rho_{\text{post}}(x \,|\, y) = \frac{\rho_{\text{like}}(y \,|\, x)\, \rho_{\text{prior}}(x)}{\rho_{\text{ev}}(y)}, \tag{4.1}$$

where $\rho_{\text{ev}}$ is the marginal density of the data $y$. Upon fixing a measurement $y$, the marginal probability $\rho_{\text{ev}}(y)$ in the denominator is a constant and hence often abbreviated

[1] Note that we will, in common abuse of notation, use the same symbol for a probability measure and its density with respect to the Lebesgue measure, whenever the latter exists. Within the present chapter, we will stick to the letter $\rho$.

as $Z \coloneqq \rho_{\text{ev}}(y)$ in the following. Since $\rho_{\text{post}}$ is a probability density, $Z$ is given by

$$Z = \int \rho_{\text{like}}(y \,|\, \tilde{x})\, \rho_{\text{prior}}(\tilde{x})\, \mathrm{d}\tilde{x}. \tag{4.2}$$

In many applications, particularly in imaging, the integral in (4.2) is high-dimensional and hence intractable. This becomes important when designing inference or sampling algorithms for $\rho_{\text{post}}$: For given $x$, we typically do not have access to density values $\rho_{\text{post}}(x \,|\, y)$ but rather only a constant multiple.

By definition, the posterior $\rho_{\text{post}}$ is the solution of the Bayesian inverse problem and gathers all our information about the unknown $x$. For practical downstream tasks, one can then use the posterior to extract this information. A standard task in inverse problems is for example to compute the instance $x$ that is most likely under $\rho_{\text{post}}$, leading to the maximum a posteriori (MAP) estimate

$$x_{\text{MAP}} \coloneqq \arg\max_x \rho_{\text{post}}(x \,|\, y). \tag{4.3}$$

We will frequently use that MAP computation for posteriors corresponds to variational minimization in the following way. The negative logarithm is monotonic, so

$$x_{\text{MAP}} = \arg\min_x \left\{ -\log \rho_{\text{like}}(y \,|\, x) - \log \rho_{\text{prior}}(x) \right\}. \tag{4.4}$$

This is a standard formulation encountered in the (variational) regularization of standard inverse problems $A(x) = y + \varepsilon$. For instance, if the negative log-likelihood is simply a quadratic weighted by some observation noise variance $\sigma^2$ and we set $V(x) = -\log \rho_{\text{prior}}(x)$, then this is equivalent to a standard regularized least squares problem

$$x_{\text{MAP}} = \arg\min_x \left\{ \frac{1}{2\sigma^2} \|A(x) - y\|_2^2 + V(x) \right\} \tag{4.5}$$

We frequently re-encounter this connection later on in this work, it is commonly viewed as the bridge between the more classical variational approach and the Bayesian perspective on inverse problems.

We now move to the practical instantiation of Bayesian models by reviewing typical choices for prior distributions in Section 4.1, with a focus on imaging applications. We have already defined the likelihood for Poisson data in Chapter 3, and give context and more examples in Section 4.2. We then review typical quantities and computational tasks that arise in Bayesian inference in Section 4.3. We then introduce two classes of typical algorithms that are employed for Bayesian inference with high-dimensional, log-concave posteriors. This includes convex optimization methods for MAP estimation in Section 4.4 and the basics of posterior sampling based on Markov chains in Section 4.5.

## 4.1. Modeling Image Priors

Following the natural order of acquiring information in Bayesian inference, we start by first reviewing typical imaging prior distributions, which encode our knowledge about the unknown before making any measurement. Especially in imaging applications, the understanding of how a likely image may look is often a qualitative one, so that the translation of that information into a quantitative prior density function is a very difficult problem on its own. In some applications, datasets of typical reconstructions from past experiments exist. For instance, the availability of many natural images led researchers to craft more refined distributions that closely match pixel statistics [98]. Another example is magnetic resonance imaging (MRI), where large image datasets significantly accelerated the development of efficient, specialized machine learning algorithms for reconstruction and uncertainty quantification in medical contexts [112].

When no quantitative prior data is available, crafting priors often comes down to formulating energy functionals that penalize unwanted or unlikely features. We will review some of the most common approaches for that in the following.

**Incorporating Hard Constraints.** In many applications, the underlying physical experiments puts strict constraints on the reconstructed unknown. For instance, X-ray attenuation coefficients in computerized tomography (CT) and activity of a radionuclide in emission imaging can only take positive values, so it is important to encode a hard constraint to the positive orthant into the prior $\rho_{\text{prior}}(x)$ in these applications.

If the constraint is encoded by $x \in \mathcal{C}$, then this can be incorporated into the prior knowledge simply by multiplying the remaining prior density $\rho(x)$ with an indicator like

$$\rho_{\text{prior}}(x) \propto \mathbf{1}_{\mathcal{C}}(x)\rho(x), \tag{4.6}$$

where $\mathbf{1}_{\mathcal{C}}(x) = 1$ whenever $x \in \mathcal{C}$ and $\mathbf{1}_{\mathcal{C}}(x) = 0$ otherwise. Note that this requires renormalization of the prior density, if $\int_{\mathcal{C}} \rho(x)\,\mathrm{d}x < 1$.

**Gibbs Priors.** Most of the typical prior distributions used for solving inverse problems fall into the class of Gibbs distributions. These are general measures with a density

$$\rho_{\text{prior}}(x) \propto \exp\left(-\beta V(x)\right), \tag{4.7}$$

where $V(x)$ is a general potential function and $\beta$ a constant. Due to their origin in statistical mechanics, where $V(x)$ is the energy of a system in state $x$ and $\beta = (k_B T)^{-1}$ is the inverse product of Boltzmann constant and thermodynamical temperature, they are sometimes also called Boltzmann distributions. In the context of prior distributions, $T$ controls the distribution's entropy and the potential $V$ is designed such that $V(x)$ is small for likely instances of the unknown $x$, i.e., the mass of the prior concentrates in

regions of small energy. Making a connection to variational regularization, we will see that $\beta V$ can be interpreted as a regularization or penalty functional.

By convention, hard constraints can formally be interpreted as Gibbs priors: By allowing infinite energy $V : \mathbb{R}^d \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$, we can let $V = \tilde{V} + \chi_{\mathcal{C}}$, where $\chi_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$ and $\chi_{\mathcal{C}}(x) = \infty$ otherwise, we arrive[2] at priors of the form (4.6).

One of the simplest Gibbs priors arises from choosing a quadratic $V$. Suppose we have a prior estimate of the mean $m$ and covariance matrix $C$ of $x$, then one could choose

$$V(x) = \frac{1}{2}(x-m)^T C^{-1}(x-m),$$

which corresponds (with $\beta = 1$) to choosing a Gaussian prior $\rho_{\text{prior}} = \mathrm{N}(m, C)$. Assuming Gaussian priors is a common modeling assumption in inverse problems due to their simple structure. In particular, for linear forward models and a Gaussian prior and independent Gaussian noise, we can compute the posterior distribution in closed form, see [106, Thm 3.7].

**Example 4.1: Gaussian Prior for Linear Inverse Problems**

Suppose we want to solve the linear inverse problem $y = Ax + \varepsilon$, with a known matrix $A$ and Gaussian observation noise $\varepsilon \sim \mathrm{N}(m_\varepsilon, C_\varepsilon)$. Suppose that $x$ is mutually independent of $\varepsilon$ and follows a Gaussian prior $x \sim \mathrm{N}(m_x, C_x)$. Then the posterior of $x$ is the following Gaussian:

$$\rho_{\text{post}} = \mathrm{N}(m_{\text{post}}, C_{\text{post}}), \tag{4.8}$$
$$m_{\text{post}} = m_x + C_x A^T (A C_x A^T + C_\varepsilon)^{-1}(y - A m_x - m_e), \tag{4.9}$$
$$C_{\text{post}} = C_x - C_x A^T (A C_x A^T + C_\varepsilon)^{-1} A C_x. \tag{4.10}$$

**Sparsity-Promoting Priors** In signal processing and imaging applications, there is often an emphasis on enforcing sparsity in solutions to inverse problems. The underlying motivation is that ill-conditioned problems can be regularized by assuming that their solution can be approximated as a linear combination of only few elements of a suitable pre-defined basis or dictionary. This idea is intimately connected with the field of compressed sensing [48, 47, 72]. A well-motivated energy in the sense of Gibbs priors to recover sparse solutions would be the $\ell^0$-"norm"

$$\|x\|_0 := \#\{i \in \{1, \dots, d\} \,:\, x_i \neq 0\},$$

which counts the number of non-zero entries in $x$. Sparsity in a basis or frame $D$ could be modeled by energies depending on $\|D^T x\|_0$, where the rows $D_i$ of $D$ are the

[2] Upon adopting the convention $\exp(-\infty) = 0$, which allows to write $\exp(-\chi_{\mathcal{C}}) = \mathbf{1}_{\mathcal{C}}$. We will frequently employ this convention in the context of densities in this work.

basis elements. Note that the $\ell^0$-norm is not a norm on $\mathbb{R}^d$ since it is not absolutely homogeneous: $\|\alpha x\|_0 = \|x\|_0 \neq \alpha\|x\|_0$ for $\alpha \neq 1$ and $0 \neq x \in \mathbb{R}^d$. The $\ell^0$-norm cannot be used directly for prior modeling, due to the violated homogeneity, $\exp(-\beta\|D^T \cdot\|_0) \notin L^1(\mathbb{R}^d)$ is not a valid density. Even though the posterior might still have a well-defined density, computational tasks involving the $\ell^0$-norm are often demanding due to the functional's non-convexity. A common way around this problem is to relax the $\ell^0$-norm by the more convenient $\ell^1$-norm. This is motivated by popular results from optimization and compressed sensing, which state that under certain conditions on the forward model, solutions of $\ell^0$- and $\ell^1$-regularized problems coincide, preserving sparsity [48, 47, 72]. The resulting priors have densities of the form

$$\rho_{\text{prior}}(x) \propto \exp\left(-\beta\|D^T x\|_1\right) = \exp\left(-\beta \sum_{j=1}^{d} |D_j^T x|\right). \tag{4.11}$$

Note that the density above is technically only a density of a probability measure if the columns $D_i$ span the whole $\mathbb{R}^d$. If this is not the case, $\rho_{\text{prior}}$ is called an improper prior and may still be used—since the object of interest is the posterior, we only need that $\rho_{\text{post}} \in L^1(\mathbb{R}^d)$ which can often still be guaranteed even for rank-deficient bases $D$ [106, Thm 3.10].

**Group Sparsity.** If we want to penalize or enforce sparsity of whole groups of parameters at once, it sometimes makes sense to consider group $\ell^1$-norms as potentials. Suppose that $D^T x$ has some physically motivated structure that lets us regroup its coordinates as $D^T x \in \mathbb{R}^{m_1 \times m_2}$. If we want to impose sparsity of whole blocks of size $m_1$, we can consider priors of the form

$$\rho_{\text{prior}}(x) \propto \exp\left(-\beta\|D^T x\|_{p,1}\right) = \exp\left(-\beta \sum_{j=1}^{m_2} \|(D^T x)_j\|_p\right), \tag{4.12}$$

where $\|(D^T x)_j\|_p$ is the $\ell^p$-norm of the $j$-th column $(D^T x)_j \in \mathbb{R}^{m_1}$ of $D^T x$. The exponent $p$ controls the shape of the prior in the within-group subspace, a standard choice is $p = 2$.

**Total Variation Priors for Imaging** A special type of $\ell^1$ or group $\ell^1$ type priors are those based on total variation, a popular regularization energy in imaging applications. Assume that the unknown $x \in \mathbb{R}^d$ is (a vectorized version of) the discretization of a function $u \in \mathcal{U} \coloneqq L^2(\Omega)$, for some compact domain $\Omega \subset \mathbb{R}^2$. The function $u$ represents the unknown as an image at arbitrary spatial resolution. The total variation (TV) of $u$ is defined by

$$\mathrm{TV}(u) \coloneqq \sup_{\substack{\varphi \in C_0^\infty(\Omega;\mathbb{R}^d) \\ \|\varphi\|_\infty \leq 1}} \int_\Omega u \operatorname{div} \varphi \, \mathrm{d}x, \tag{4.13}$$

where $C_0^\infty(\Omega;\mathbb{R}^d)$ is the space of $\mathbb{R}^d$-valued smooth test functions of compact support on $\Omega$. This functional was popularized as a regularization objective for image recovery in the Rudin–Osher–Fatemi (ROF) model named after the authors of [182]. For recovering $u$ from a noisy version $f$ of the image, the ROF model proposed to solve the optimization problem

$$\arg\min_{u\in \mathrm{BV}(\Omega)} \left\{ \frac{1}{2}\|u-f\|^2_{L^2(\Omega)} + \lambda\,\mathrm{TV}(u) \right\}, \tag{4.14}$$

where $\mathrm{BV}(\Omega) = \{u \in L^1(\Omega) : \mathrm{TV}(u) < \infty\}$ is the space of bounded variation and $\lambda > 0$ a regularization parameter. Minimizing the total variation promotes sparse gradients in the image $u$, a prominent sought-for feature in many imaging applications which cannot be realized with smooth regularization. Inspecting (4.14), we can use the connection (4.4) between variational minimization and MAP estimation: It is immediately clear that the ROF model coincides (up to discretization of the domain) with MAP estimation for a Gibbs prior of the form $\exp(-\lambda\,\mathrm{TV}(u))$. In computational practice, the image $u$ is discretized on a grid of pixel values stored in $x$, which requires discretizing the total variation functional as well. For a standard gray-scale image $x$ with pixel values $x_{i,j}$ with rows $1 \le i \le d_1$, columns $1 \le j \le d_2$, there are two typical variants. The first is the isotropic total variation

$$\mathrm{TV_i}(x) = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \sqrt{(x_{i+1,j}-x_{i,j})^2 + (x_{i,j+1}-x_{i,j})^2}, \tag{4.15}$$

and the second the anisotropic total variation

$$\mathrm{TV_a}(x) = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} |x_{i+1,j}-x_{i,j}| + |x_{i,j+1}-x_{i,j}|, \tag{4.16}$$

with Neumann boundary conditions wherever indices of the finite differences are outside the image boundary. The corresponding Gibbs priors are simply $\exp(-\lambda\,\mathrm{TV_i}(x))$ and $\exp(-\lambda\,\mathrm{TV_a}(x))$, respectively, where the ROF regularization parameter $\lambda = \beta$ can as usual be interpreted as inverse temperature. Defining the basis $D$ as a finite difference operator, we also see that $\mathrm{TV_i}$ leads to a special case of a group $\ell^2$-$\ell^1$ sparsity prior (4.12), whereas $\mathrm{TV_a}$ gives an $\ell^1$ prior (4.11). We note here that the two variants differ mostly in the recovery of corners and curved edges. Variational minimization of anisotropic TV leads to sharper corners in images, while the isotropic form tends to smooth out corners, see [41, 63] for more details.

**Total Generalized Variation** The idea of enforcing sparse gradients under total variation priors can be expanded to higher-order sparsity. Like the TV functional, total generalized variation (TGV) allows for discontinuous edges in images, but it generalizes TV in that it also employs higher-order derivatives. Acting as a penalty also on higher order image gradients, TGV-based MAP estimates avoid the typical staircasing effects of

TV in smooth but non-constant regions of images [35, 36]. For the spatially continuous formulation in terms of higher order derivatives, we refer to [35]. In a discretized setting, the second order TGV functional is usually defined as the infimal convolution

$$\begin{aligned} \mathrm{TGV}_\alpha^2(x) &= \inf_{v \in \mathbb{R}^{2d}} J_\alpha^2(x, v), \\ J_\alpha^2(x, v) &= \alpha_1 \|\nabla x - v\|_{2,1} + \alpha_0 \|Ev\|_{2,1}. \end{aligned} \tag{4.17}$$

Here $x \in \mathbb{R}^d$ is an image with $d$ pixels, while $v$ is in the corresponding space $\mathbb{R}^d \times \mathbb{R}^d \cong \mathbb{R}^{2d}$ of (discretized) vertical and horizontal gradients of $x$. $\alpha_0, \alpha_1 > 0$ are regularization parameters. $\nabla = (\nabla_h, \nabla_v) : \mathbb{R}^d \to \mathbb{R}^{2d}$ is a (forward discretized) gradient operator as in the case of TV, and $E : \mathbb{R}^{2d} \to \mathbb{R}^{4d}$ is a symmetrized (backward discretized) gradient operator that acts as second order derivative, see [36] for more details on the discretization and implementation. When constructing a TGV-based prior, it was proposed in experiments in [43] to expand the state variable to $\tilde{x} = (x, v) \in \mathbb{R}^d \times \mathbb{R}^{2d}$ and set

$$\rho_{\text{prior}}(\tilde{x}) \propto \exp(-J_\alpha^2(x, v)).$$

Of the joint posterior in $\rho_{\text{post}}(x, v \,|\, y)$, one is then interested in properties of the marginal $\int_{\mathbb{R}^{2d}} \rho_{\text{post}}(x, v \,|\, y)\, \mathrm{d}v$.

**Hierarchical Priors.** Another popular class of priors that have been proposed to construct refined prior models and enforce sparsity are hierarchical models. The idea consists in adding an additional latent set of parameters $\theta \sim \rho_{\text{hyper}}$ to the model, with a hyperprior $\rho_{\text{hyper}}$ controlling the prior. The likelihood remains unchanged, but the prior may depend on $\theta$ in that $\rho(x, \theta) = \rho_{\text{prior}}(x \,|\, \theta)\, \rho_{\text{hyper}}(\theta)$. One then recovers the joint posterior

$$\rho_{\text{post}}(x, \theta \,|\, y) = \frac{\rho_{\text{like}}(y \,|\, x)\, \rho_{\text{prior}}(x \,|\, \theta)\, \rho_{\text{hyper}}(\theta)}{\rho_{\text{ev}}(y)}.$$

Hierarchical priors have been proposed for sparse reconstruction in [205, 220]. Consider for example a Gaussian prior of the form

$$\rho_{\text{prior}}(x \,|\, \theta) \propto \exp\left(-\frac{1}{2} x^T D B_\theta D^T x\right),$$

where $B_\theta = \mathrm{diag}(\theta_1, \ldots, \theta_m)$. If $D$ is a finite difference operator, this can be interpreted as a kind of TV prior based on $\ell^2$-norms and with spatial variability in the temperature/regularization strength. By imposing suitable hyperpriors on $\theta$, it was shown that images with sparse gradients can be reconstructed using this model [45, 46, 88, 132]. Formally, the TGV approach (4.17) can also be viewed as a hierarchical prior, with a group sparsity hyperprior on the approximate gradients $v$. The hierarchical approach is computationally more demanding than direct priors without hyperparameters, since one has to solve for the unknown $x$ and the hyperparameter $\theta$. Special properties like

conjugacy of the hyperprior can help alleviate some of the additional costs [86, chap. 2].

## 4.2. Likelihood Modeling

When formulating the forward model for Compton camera imaging in the last chapter, we already discussed the Poisson statistics of CCD camera measurements as well as further error or noise sources in the data. The Poisson model defined the likelihood density for the measured data, which can be used for Bayesian inference. In this section, we cover further aspects of likelihood modeling, particularly approximations to the Poisson model and other, typical models in imaging applications. As a simplification, the forward operator $A$ will be assumed deterministic and known, ignoring epistemic uncertainty. Accordingly, we define an "ideal" or "clean" measurement $y^{\text{clean}} = A(x)$ after conditioning on an image $x$. Whenever we know the cause $x$, the formulation of $y^{\text{clean}}$ comes down to describing the physical system which behaves according to known laws like Maxwell's equations. As we saw in the last chapter on Compton cameras, this picture is often an idealized one and we will comment later on how to deal with uncertain parameters in $A$.

**Independent additive noise.** In CCD modeling, we saw that data is typically of discrete nature due to the quantization of events. In practice, it is often sufficient to assume that both $y^{\text{clean}}$ and any noisy instantiation $y$ are elements of an uncountable, continuous space, for instance $y^{\text{clean}}, y \in \mathbb{R}^m$. In the majority of inverse problems literature, the measured data is understood to be an additively corrupted version of the clean data

$$y = y^{\text{clean}} + \varepsilon,$$

where, in the simplest case, the noise $\varepsilon$ is assumed mutually independent of $x$ and hence also of $y^{\text{clean}}$. From an information-theoretic perspective, this is equivalent to saying that the conditional entropy of $y$ given $x$ is completely determined by the entropy in $\varepsilon$. If $\varepsilon$ has a known distribution with density $\rho_{\text{noise}}(\varepsilon)$, then the likelihood can simply be given as

$$\rho_{\text{like}}(y \,|\, x) = \rho_{\text{noise}}\left(y - y^{\text{clean}}\right) = \rho_{\text{noise}}\left(y - A(x)\right).$$

**Example 4.2: Independent Gaussian Noise**

Assume that $\varepsilon \in \mathbb{R}^m$ is additive und mutually independent of the unknown $x$. If $\varepsilon$ follows a normal distribution $\rho_{\text{noise}} = \mathrm{N}(m, C)$, the likelihood obeys

$$\rho_{\text{like}}(y \,|\, x) \propto \exp\left(-\frac{1}{2}(y - A(x) - m)^T C^{-1}(y - A(x) - m)\right).$$

Due to the simple structure of normal distributions, Gaussian noise is a particularly appealing model and common in inverse problems [106]. This is further justified by the central limit theorem, which shows that the sum of multiple independent noise sources may be well approximated by a normal distribution in practice. A Gaussian noise model is fully characterized by its first two moments.

**Independent multiplicative noise.** Similar to additive perturbations, another example that directly yields a likelihood through the noise distribution is multiplicative noise. Suppose $\varepsilon \sim \rho_{\text{noise}}$ and $y^{\text{clean}}$ are real-valued, and $\varepsilon$ is mutually independent of $x$. If the corrupted data is measured as

$$y = (1 + \varepsilon) y^{\text{clean}} \in \mathbb{R},$$

then, due to $\rho(\varepsilon \,|\, x) = \rho_{\text{noise}}(\varepsilon)$, the likelihood is given by

$$\begin{aligned} \rho_{\text{like}}(y \,|\, x) &= \int_{\mathbb{R}} \rho(y \,|\, x, \varepsilon)\, \rho_{\text{noise}}(\varepsilon)\, \mathrm{d}\varepsilon \\ &= \int_{\mathbb{R}} \delta(y - (1+\varepsilon)A(x))\, \rho_{\text{noise}}(\varepsilon)\, \mathrm{d}\varepsilon \\ &= \rho_{\text{noise}}\left(\frac{y}{A(x)} - 1\right). \end{aligned}$$

A typical application of multiplicative noise models is speckle noise in radar or other coherent light imaging [208]. In this case of higher dimensional data like an image with independent, multiplicative noise in every component, the likelihood density just follows by using the density above and taking the outer product over all components (e.g., pixels).

**General additive noise.** In the case where there is no simple dependence of $y$ on $y^{\text{clean}}$ through some noise variable $\varepsilon$ that is mutually independent of $x$, the likelihood may become more involved. Write again $y = y^{\text{clean}} + \varepsilon$ but assume that $\varepsilon \in \mathbb{R}^m$ may depend on $x$, i.e., there is a conditional noise density $\rho_{\text{noise}}(\varepsilon \,|\, x)$. If we have access to this conditional density, then the likelihood is

$$\rho_{\text{like}}(y \,|\, x) = \int_{R^m} \rho(y \,|\, x, e)\, \rho_{\text{noise}}(\varepsilon \,|\, x)\, \mathrm{d}\varepsilon = \rho_{\text{noise}}(y - A(x) \,|\, x).$$

**Poisson likelihood approximations.** While the Poisson likelihood (3.13) is the natural choice for photon counting applications, it can be computationally advantageous to approximate it with simpler models in certain regimes. We briefly discuss two common approximations and their validity conditions.

For large count values, the Poisson distribution can be well approximated by a Gaussian distribution. Specifically, when $\lambda_j = A_j(f) \gg 1$ for all measurement bins $j$, the

approximation $\mathrm{Pois}(\lambda_j) \approx \mathrm{N}(\lambda_j, \lambda_j)$ becomes accurate. Under this approximation, the likelihood takes the form

$$\rho_{\mathrm{like}}(y \,|\, x) \approx \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi A_j(x)}} \exp\left(-\frac{(y_j - A_j(x))^2}{2A_j(x)}\right). \tag{4.18}$$

This approximation is particularly useful when count rates are high, as in long measurement times or strong sources. The resulting optimization and sampling problems often become easier to analyze due to the smoothness of the Gaussian likelihood, though one must account for the heteroscedastic (signal-dependent) noise variance.

A related technique for stabilizing the variance is the Anscombe transform [9], which approximately normalizes Poisson data. The transform $\tilde{y}_j = 2\sqrt{y_j + 3/8}$ converts Poisson-distributed observations to approximately Gaussian with constant variance, allowing the use of standard inverse problem techniques designed for homoscedastic Gaussian noise. After reconstruction, an inverse transform can be applied if needed.

In our Compton camera experiments (Chapter 5), count rates vary significantly across energy bins, with some bins having very low counts while others may have hundreds of counts. Therefore, we use the full Poisson likelihood (3.13) rather than Gaussian approximations, ensuring accurate modeling in all measurement regimes.

**Additional Likelihood Parameters.** For the sake of simplicity, we assume throughout most of this work that the forward model $A$ is deterministic. Note, however, that $A$ may also depend on unknown parameters $\gamma$ in practical applications. The likelihood can be adapted to reflect this model uncertainty, instead of $\rho_{\mathrm{like}}(y \,|\, x)$, one then considers $\rho_{\mathrm{like}}(y \,|\, x, \gamma)$. This introduces an additional prior distribution $\rho(\gamma)$ which depends on the available information about the forward model. The joint estimation of image $x$ and parameter $\gamma$ is a common strategy in Bayesian imaging problems, a standard exemplary task is the estimation of unknown measurement noise levels [168, 88]. Similar ideas have been discussed for various other application-specific unknown parameters, e.g., blur kernels in blind image deconvolution [221], estimating uncertain or unknown scan angles in CT [175, 176], magnetic resonance imaging with unknown fieldmap inhomogeneity [3], error correction in magnetic particle imaging [34], scan position estimation in ptychographic phase retrieval [215], and speed of sound variability in photoacoustic tomography [204]. Although we do not explicitly use such a model, we note here that such an approach could be an option to correct some of the discussed model errors in Compton camera image reconstruction. For example, the measurement data that we consider usually required to recalibrate the detectors before measurement. For practical and efficient measurement routines, this step needs to be automated, requiring a calibration correction before or jointly with the reconstruction step. This could be achieved with a proper Bayesian model for the calibration parameters. We also refer to the discussion of modeling errors in the context of Bayesian inverse problems in [106, chap. 5.8].

## 4.3. Bayesian Estimation and Inference

For a given Bayesian posterior, we typically do not work directly with the recovered distribution or density. In practice, one is instead interested in questions about specific properties of the posterior in order to solve downstream tasks. We already defined the MAP estimate (4.3) as such a quantity: For imaging applications, a MAP estimate allows a quick visual glance at a "typical" (actually, the most typical in the sense of density) reconstruction under this posterior.

**Point estimates.** The MAP estimate is only one example of a point estimate, i.e., a single instance of the unknown $x$ with a predefined property under the posterior. The other prevalent point estimate in Bayesian imaging inverse problems is the conditional mean (CM) estimate

$$x_{\mathrm{CM}} \coloneqq \mathbb{E}[x \mid y] = \int_{\mathbb{R}^d} x \, \rho_{\mathrm{post}}(x \mid y) \, \mathrm{d}x. \tag{4.19}$$

The CM estimate is a special case of an estimator resulting from the Bayes cost method: Suppose that $\tilde{x}$ is some point estimate and $\psi(x, \tilde{x})$ a cost function. Then the general Bayes estimator for the cost $\psi$ is the estimator

$$x_\psi \coloneqq \arg\min_{\tilde{x}} \mathbb{E}_{x \sim \rho_{\mathrm{post}}}[\psi(x, \tilde{x})]. \tag{4.20}$$

By inspecting the optimality condition for $x_\psi$, it is easy to see that the CM results from choosing a squared error cost $\psi(x, \tilde{x}) = \|x - \tilde{x}\|_2^2$. This justifies the term minimum mean squared error (MMSE) which is frequently used equivalently to CM. The MAP is related to, but technically not a real Bayes cost estimator. It arises in the limit of certain estimators: If we set

$$\psi_\delta(x, \tilde{x}) = \begin{cases} 0 & \text{if } \|x - \tilde{x}\|_\infty < \delta \\ 1 & \text{otherwise,} \end{cases}$$

and define $x_\delta = x_{\psi_\delta}$, then $x_{\mathrm{MAP}} = \lim_{\delta \to 0} x_\delta$. Perceptively, the conditional mean is often similar to the MAP estimate for unimodal target distributions. Qualitative differences arise, for instance, in examples with gradient sparsity, where the CM tends to be "smoother" than the MAP, see [139, chap. 6] for detailed discussions about the relation of these estimators. Fig. 4.1 shows an example of a CM estimate in a simulated image deblurring problem.

**Spread estimates.** Apart from point estimates, practicioners may also be interested in higher order moments, for instance to examine variances or standard deviations. The inherent problem of the high dimensionality of imaging applications is the difficult

visualization. If we reconstruct a $d_1$-by-$d_2$ image, then the full posterior covariance matrix

$$C_{\text{post}} \coloneqq \mathbb{E}[(x - x_{\text{CM}})(x - x_{\text{CM}})^T \,|\, y] = \int_{\mathbb{R}^d} (x - x_{\text{CM}})(x - x_{\text{CM}})^T \, \rho_{\text{post}}(x \,|\, y) \, \mathrm{d}x \quad (4.21)$$

has $d_1^2 d_2^2$ entries and is difficult to visualize entirely. Imaging studies therefore often report just the diagonal of $C_{\text{post}}$, i.e., the marginal variance of each pixel under the posterior, see Fig. 4.1. Another example of spread estimate of single pixel values or regions of pixels are credible regions. For a pixel $i$, consider the posterior marginal

$$\rho^i_{\text{post}}(x_i \,|\, y) \coloneqq \int_{\mathbb{R}^{d-1}} \rho_{\text{post}}(x \,|\, y) \, \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1} \, \mathrm{d}x_{i+1} \cdots \mathrm{d}x_d.$$

Then, for a value $0 < \alpha < 1$, a credible set is $C_\alpha \subset \mathbb{R}$ such that $x_i$ has probability $1 - \alpha$ to fall in $C_\alpha$, i.e., $C_\alpha$ has to satisfy

$$\int_{C_\alpha} \rho^i_{\text{post}}(x_i \,|\, y) \, \mathrm{d}x_i = 1 - \alpha. \quad (4.22)$$

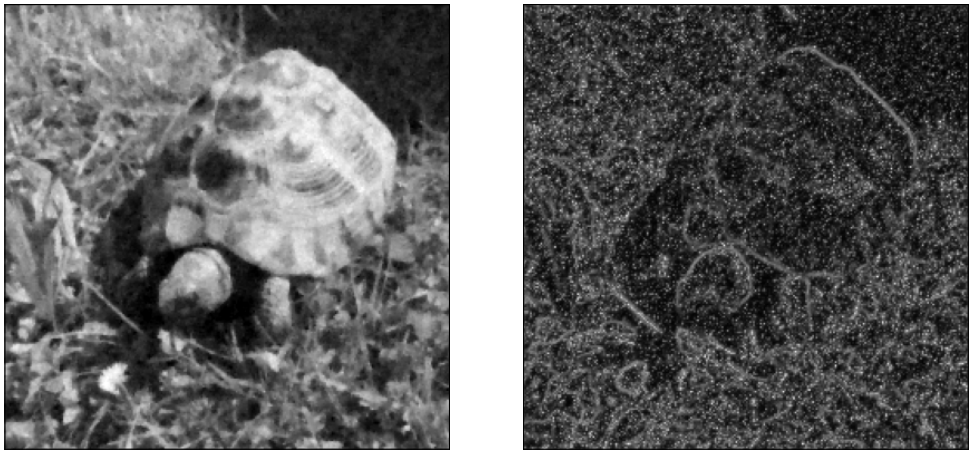
For a single pixel, one is often interested in intervals $C_\alpha = [a(\alpha), b(\alpha)]$, for example centered around the mean of the marginal $\rho^i_{\text{post}}(\cdot \,|\, y)$. For multiple pixels, this can easily be generalized by integrating the corresponding marginal in all selected pixels in (4.22), see [8, 155] for recent applications in image recovery. More generally, we may want to estimate the probability of a certain set $C$ under the posterior, i.e., estimate the value of general integrals of the form

$$\int_C \rho_{\text{post}}(x \,|\, y) \, \mathrm{d}x. \quad (4.23)$$

A standard example for such a task is when we want to compute the probability of some estimator $\psi : \mathbb{R}^d \to \mathbb{R}$ exceeding a threshold $t$—say, a maximum permitted total radioactivity in Compton imaging—under the posterior. This leads to the above formulation (4.23) with $C = \psi^{-1}([t, \infty))$.

**Model Selection** Another task, related to hyperpriors or hyperparameters in likelihoods is Bayesian model selection, see [177, chap. 7]. Frequently, not only the image $x$ is unknown, but we may also be unsure about aspects in the prior or the likelihood. Suppose now, however, that there is not a continuous set of parameters describing this uncertainty, but rather only a few models $\mathcal{M}_i$, from which we want to select one that best fits reality based on the measurements. Making these models explicit in Bayes' law, we see that the posterior changes depending on the modeling assumption

$$\rho_{\text{post}}(x \,|\, y, \mathcal{M}_i) = \frac{\rho_{\text{like}}(y \,|\, x, \mathcal{M}_i) \, \rho_{\text{prior}}(x \,|\, \mathcal{M}_i)}{\rho_{\text{ev}}(y \,|\, \mathcal{M}_i)}.$$



(a) Conditional mean (b) Pixelwise posterior variance

Figure 4.1.: Visualization of first (conditional mean) and second moment (log-variance of the pixelwise marginal distributions) of a posterior distribution in a simulated, TGV-regularized image deblurring problem. Images from [43], computed with the primal-dual Langevin sampling method laid out in Section 7.3.

Without ground truth available, a quantity that allows to compare two models $\mathcal{M}_i, \mathcal{M}_j$ is the Bayes factor

$$B(\mathcal{M}_i, \mathcal{M}_j) = \frac{\rho_{\text{ev}}(y \,|\, \mathcal{M}_i)}{\rho_{\text{ev}}(y \,|\, \mathcal{M}_j)}.$$

The factor $B(\mathcal{M}_i, \mathcal{M}_j)$ can be viewed as a likelihood ratio, telling us that model $\mathcal{M}_i$ should be preferred if $B(\mathcal{M}_i, \mathcal{M}_j) \gg 1$ and $\mathcal{M}_j$ should be chosen if $B(\mathcal{M}_i, \mathcal{M}_j) \ll 1$. Since evaluating the evidence values in the definition of $B(\mathcal{M}_i, \mathcal{M}_j)$ amounts to computing high-dimensional integrals, computing the Bayes factor is computationally challenging in imaging applications. For computational approaches we refer to [61], and [44] for a recent approach using nested sampling in imaging applications.

## 4.4. MAP Computation and Convex Optimization

Due to both the relative simplicity and the ubiquity within classical inverse problems literature, MAP computation (4.3) and, equivalently, variational regularization (4.4) are the de facto standard in the solution of a wide variety of imaging inverse problems [52, 25, 11]. Since many of the theoretical techniques are intimately related to the theory developed in this thesis, we review some relevant computational and algorithmic aspects here. Suppose therefore that we have a standard posterior with negative log-density $V(x) = -\log \rho_{\text{post}}(x \,|\, y) = -\log \rho_{\text{like}}(y \,|\, x) - \log \rho_{\text{prior}}(x)$ and we want to solve the

optimization problem

$$\min_{x \in \mathbb{R}^d} V(x). \tag{4.24}$$

A central assumption in many applications is log-concavity of this posterior, i.e., convexity of the optimization problem (4.24).

The log-concavity assumption is restrictive in the sense that it limits our choice of prior model, but the resulting structure of the posterior allows to use the rich suite of convex optimization algorithms in order to recover the MAP estimate. We will also see in later sections that it is a standard assumption in the convergence analysis of the sampling algorithms that this thesis is concerned with. Broadly speaking, the relevant convex optimization algorithms for imaging can be distinguished into two classes: Gradient methods, which can be seen as discretizations of a gradient flow, and saddle point methods, which arise from the concept of convex duality. We now give a brief overview of some relevant methods from either class and refer to [52] for details in the imaging context.

**Gradient methods.** These algorithms are first-order optimization algorithms that arise from discretizing the gradient flow of $V$, i.e., the ordinary differential equation (ODE)

$$\dot{x}_t = -\nabla V(x_t), \tag{4.25}$$

where $\dot{x}_t$ denotes the time derivative $\mathrm{d}x_t/\mathrm{d}t$ and we assume for simplicity that $V$ is differentiable. By standard stability theory of ODEs, the dynamics are asymptotically stable if $V$ is $\omega$-strongly convex with $\omega > 0$, in which case $x_t$ converges to the unique minimizer $x^*$ of $V$. Further, we can easily obtain stability and convergence rates by assuming $x_t, \tilde{x}_t$ both solve (4.25) with different initializations and computing

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x_t - \tilde{x}_t\|^2 &= \langle \dot{x}_t - \dot{\tilde{x}}_t, x_t - \tilde{x}_t \rangle \\ &= -\langle \nabla V(x_t) - \nabla V(\tilde{x}_t), x_t - \tilde{x}_t \rangle \\ &\leq -\omega \|x_t - \tilde{x}_t\|^2. \end{aligned} \tag{4.26}$$

The quantity in the second line is the symmetrized Bregman distance $\mathrm{D}_V^{\mathrm{sym}}(x_t, \tilde{x}_t) = \mathrm{D}_V(x_t, \tilde{x}_t) + \mathrm{D}_V(\tilde{x}_t, x_t)$, where the Bregman distance $\mathrm{D}_V$ itself is defined by

$$\mathrm{D}_V^{p_0}(x, x_0) = V(x) - V(x_0) - \langle p_0, x - x_0 \rangle \qquad \forall x, x_0 \in \mathbb{R}^d, p_0 \in \partial V(x_0) \tag{4.27}$$

for any general, convex $V$. If the subgradient $p_0$ is unique or irrelevant, we may drop it and denote the Bregman distance simply by $\mathrm{D}_V(x, x_0)$. For the inequality in the third line above, we used the standard quadratic lower bound $\mathrm{D}_V^{\mathrm{sym}}(x, \tilde{x}) \geq \omega \|x - \tilde{x}\|^2$ for $\omega$-strongly convex functionals, see [20, chap. 18]. From the above inequality, an application of Grönwall's inequality then gives the contraction rate

$$\|x_t - \tilde{x}_t\| \leq \|x_0 - \tilde{x}_0\| e^{-\omega t}. \tag{4.28}$$

By inserting the stationary solution $\tilde{x}_0 = \tilde{x}_t = x^*$, we obtain the convergence result. The most straightforward gradient method that can be derived from (4.25) by a forward Euler discretization is gradient descent, characterized by the update

$$x^{(n+1)} = x^{(n)} - \tau \nabla V(x^{(n)}), \tag{4.29}$$

where $\tau > 0$ is a chosen step size. Under the additional assumption of $L$-Lipschitz continuity of $\nabla V$ and a step size restriction like $\tau \leq L^{-1}$, gradient descent converges to $x^*$ and recovers the linear convergence behaviour of the gradient flow, i.e., $\|x^{(n)} - x^*\| = \mathcal{O}(e^{-\kappa n})$ for some constant $\kappa$.

As Section 4.1 already showed, typical sparsity-promoting log-densities $V$ in imaging applications are not differentiable. When $V$ is convex but non-differentiable, the gradient flow (4.25) can be generalized to a subgradient flow $\dot{x}_t \in -\partial V(x_t)$. Indeed, in the stronlgy convex case, the computations (4.26) – (4.28) hold in the same way since the bound on the symmetric Bregman divergence is independent of the subgradient. Similar generalizations hold in the convex case. However, the convergence of discretized gradient—or subgradient—descent with fixed step sizes $\tau$ breaks down for non-differentiable $V$ due to the non-existence and thus violated Lipschitz-continuity of $\nabla V$. This can be mitigated, e.g., by choosing a decreasing sequence of step sizes $(\tau_n)_{n\geq 1}$, see [29]. This is a common technique also known in momentum-based and stochastic variants of gradient descent. As an alternative that is central to this thesis, we may employ a backward discretization of the subgradient flow instead. The implicit update $x^{(n+1)} \in x^{(n)} - \tau \nabla V(x^{(n+1)})$ leads to proximal gradient descent

$$x^{(n+1)} = \operatorname{prox}_V^\tau(x^{(n)}), \tag{4.30}$$

where the proximal mapping or proximal operator is defined by

$$\operatorname{prox}_V^\tau(x) := \operatorname*{arg\,min}_{\tilde{x}\in\mathbb{R}^d} \left\{ V(\tilde{x}) + \frac{1}{2\tau}\|x - \tilde{x}\|^2 \right\}. \tag{4.31}$$

Note that convergence results usually do not depend on the step size $\tau$, but the difficulty of solving (4.31) to evaluate the proximal mapping becomes as hard as the original optimization problem as $\tau$ grows large.

Since $V = -\log \rho_{\text{post}}$ is a sum of likelihood and prior log-densities, we can often employ splitting methods crafted for objectives of the form $V = F + G$. We only mention forward-backward methods here for now and will go into more detail later. If $F$ is differentiable, but $G$ is potentially not, the forward-backward method for solving (4.24) takes the form

$$x^{(n+1)} = \operatorname{prox}_G^\tau(x^{(n)} - \tau \nabla F(x^{(n)})). \tag{4.32}$$

We refer to [20, 161] for theoretical insights on this algorithm and [52] for applications in imaging.

**Acceleration of gradient methods.** In order to speed up the convergence of gradient methods, there are two common techniques to modify the ODE (4.25). The first consists in introducing second-order derivatives, solving instead

$$\ddot{x_t} + \gamma(t)\dot{x_t} = -\nabla V(x_t),$$

with a damping parameter $\gamma(t)$. The nature of the acceleration can be understood by interpreting $x_t$ as the physical location of a particle. Instead of solving the overdamped transport (4.25), the particle is assigned a (time-dependent) mass, which may enable the position to converge faster to a local minimum. The literature often speaks of "momentum"-based algorithms due to this analogy. These momentum-based algorithms were first introduced in [157], the ODE interpretation stems from [199]. Since they only use first-order information just like gradient descent, these methods are particularly useful in the training of neural networks. The other type of acceleration that is particularly relevant to this work are forms of preconditioning. The idea here is to replace (4.25) by

$$\dot{x_t} = -P(x_t)\nabla V(x_t), \tag{4.33}$$

where $P : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a matrix vector field, $P(x)$ positive definite at least in a neighbourhood of $x^*$. The multiplication by $P(x)$ can be understood as a change of geometry of the state space and leads to many relevant optimization procedures, like Newton's method, Quasi-Newton methods or the conjugate gradient method. This includes also methods based on mirror functions [156], which have gained considerable interest in sampling and optimization in machine learning recently. Given some convex mirror function $h : \mathbb{R}^d \to \bar{\mathbb{R}}$, the idea is to map iterates $x^{(n)}$ to a different space via the mirror map $\nabla h$ and run gradient descent in this "mirrored" space. The update is given by

$$x^{(n+1)} = \nabla h^*(\nabla h(x^{(n)}) - \tau\nabla V(x^{(n)})). \tag{4.34}$$

Typically, is it assumed that $h$ is of Legendre type, i.e. strictly convex, differentiable on $\Omega_h := \mathrm{int}(\mathrm{dom}(h))$ and gradient diverging to infinity at the domain's boundary: $\lim_{k\to\infty}\|\nabla h(x_k)\| = \infty$ for any $(x_k)_k \subset \Omega_h$ with $x_k \to \bar{x} \in \partial\Omega_h$, see [180, chap. 26]. A basic choice satisfying these properties is $h(x) = \frac{1}{2}\|x\|_2^2$ with $\Omega_h = \mathbb{R}^d$, which recovers classical gradient descent as a special case of mirror descent. If the assumptions are satisfied, then $\nabla h^* \circ \nabla h = \mathrm{Id}$ and thus also $(\nabla^2 h)^{-1} = \nabla^2 h^* \circ \nabla h$, where $\nabla^2$ denotes the Hessian matrix and $(\nabla^2 h)^{-1}$ the inverse matrix. This shows that, up to a first order Taylor expansion of the mirror map, mirror descent coincides with preconditioned gradient descent

$$\begin{aligned} x^{(n+1)} &= \nabla h^*(\nabla h(x^{(n)}) - \tau\nabla V(x^{(n)})) \\ &\approx \nabla h^*(\nabla h(x^{(n)})) - \tau\nabla^2 h^*(\nabla h(x^{(n)}))\nabla V(x^{(n)}) \\ &= x^{(n)} - \tau\nabla^2 h(x^{(n)})^{-1}\nabla V(x^{(n)}), \end{aligned} \tag{4.35}$$

since the last line coincides with a forward discretization of (4.33). We refer to [213] for further exploration of this connection between mirror descent and preconditioned gradient descent. We will give more geometrical intuition of mirror descent and other mirrored algorithms later. For now, we only mention that the mirror function can be chosen to enforce specific constraints in the optimization problem. If we know, for instance, that $\mathrm{dom}(V) \subset \mathbb{R}^d_+$, i.e. $x^*_j > 0$ for all $j$, there are two typical choices for $h$:

- The first is a logarithmic barrier, also referred to as Burg's entropy, $h(x) = -\sum_j \log(x_j)$, with mirror map $(\nabla h(x))_j = -x_j^{-1}$ and $\Omega_h = \mathbb{R}^d_+$. The mirror map coincides with its inverse $(\nabla h)^{-1} = \nabla h^*$ and the preconditioning matrix in terms of (4.35) is diagonal with $j$-th entry $x_j^{-2}$.

- The second one is the negative entropy[3] $h(x) = \sum_j x_j \log(x_j) - x_j$ with mirror map $(\nabla h(x))_j = \log(x_j)$ and $\Omega_h = \mathbb{R}^d_+$. The inverse mirror map is given by $(\nabla h^*(y))_j = \exp(y_j)$, hence this mirror function typically leads to multiplicative updates that naturally preserve positivity. The Hessian preconditioning is diagonal with $j$-th entry $x_j^{-1}$.

In [19], it was shown that for standard likelihood functions in imaging with positivity constraints, these two mirror functions can lead to stable and convergent optimization algorithms.

**Saddle point methods.** For non-smooth objectives of a specific splitting form, the class of primal-dual hybrid gradient (PDHG) methods has proved efficient to solve compute MAP estimates, or generally solve (4.24). Suppose the posterior has a density of the form

$$\rho_{\mathrm{post}}(x) \propto \exp(-F(Kx) - G(x)), \tag{4.36}$$

where $K : \mathbb{R}^d \to \mathbb{R}^m$ is a linear operator, $F : \mathbb{R}^m \to \bar{\mathbb{R}}$, $G : \mathbb{R}^d \to \bar{\mathbb{R}}$ are both convex and lower semi-continuous but not necessarily differentiable. We also denote the full potential by $F \circ K + G =: V : \mathbb{R}^d \to \bar{\mathbb{R}}$. Then it may help to reformulate the MAP optimization problem to a saddle point problem like

$$\begin{aligned} \min_x H(x) &= \min_x \left\{ G(x) + \max_y \left\{ \langle Kx, y\rangle - F^*(y) \right\} \right\} \\ &= \min_x \max_y S(x,y), \end{aligned} \tag{4.37}$$

with $S(x,y) \coloneqq G(x) + \langle Kx, y\rangle - F^*(y)$. Here $F^*$ denotes the convex conjugate of $F$

$$F^*(y) \coloneqq \sup_x \left\{ \langle x, y\rangle - F(x) \right\}.$$

[3] There are two typical definitions of the negative entropy, the other being $h(x) = \sum_j x_j \log(x_j)$. Since the Bregman divergence $D_h$ and the Hessian $\nabla^2 h$ are invariant to adding affine functions, this does not matter too much.

Under mild assumptions on $G$ and $F^*$, one can prove existence of a saddle point $(x^*, y^*)$ (see section 19 in [20] for details) the primal part $x^*$ of which is a solution to the original problem (4.24). Consequently, instead of minimizing $H$ directly, one can employ primal-dual type algorithms to solve the saddle point problem (4.37). For the context of imaging inverse problems, we refer to [52] for an overview.

The primal-dual optimization method (primal descent and dual ascent with overrelaxation in the primal variable) for (4.37) takes the form

$$\begin{cases} y^{(n+1)} = \operatorname{prox}_{\sigma F^*}\left(y^{(n)} + \sigma K x_\theta^{(n)}\right) \\ x^{(n+1)} = \operatorname{prox}_{\tau G}\left(x^{(n)} - \tau K^T y^{(n+1)}\right) \\ x_\theta^{(n+1)} = x^{(n+1)} + \theta(x^{(n+1)} - x^{(n)}), \end{cases} \tag{4.38}$$

with given step sizes $\tau, \sigma$, overrelaxation parameter $\theta$ and an initial guess $(x^{(0)}, y^{(0)})$. By convention, it is set $x_\theta^{(0)} = x^{(0)}$. For convergence theory of the scheme (4.38) and possible accelerations we refer to [51, 52].

**EM-type methods** Due to its relevance for our work, we recall one other optimization algorithm frequently used to solve (4.24) in photonic imaging inverse problems. As we will see, the maximum likelihood expectation maximization (MLEM) algorithm and variants of it could also be interpreted as preconditioned gradient methods, but are usually statistically motivated due to their origin [71]. Suppose that we measure data $y \in \mathbb{R}^d$ from $y \sim \mathrm{Pois}(Ax)$ for some linear forward model $A$, giving the likelihood $\rho_{\mathrm{like}}$ in (3.14). If we simply want to compute the maximum likelihood estimate (MLE) without imposing any regularizing prior, we solve (4.24) with $V = -\log \rho_{\mathrm{like}}$. The MLEM algorithm can be can be derived as a special case of the expectation maximization (EM) algorithm [71] via introducing the latent variable $z_{ij} \sim \mathrm{Pois}(A_{ij}x_j)$, giving $y_i = \sum_j z_{ij}$. In this special case, the E-step and M-step of the EM iteration can be solved in closed form and combined into the single update

$$x_j^{(n+1)} = \frac{x_j^{(n)}}{a_j} \sum_{i=1}^{m} \frac{A_{ij} y_i}{\sum_{k=1}^{d} A_{ik} x_k^{(n)}}, \tag{4.39}$$

where $a_j = (A^*\mathbf{1})_j = \sum_{i=1}^{m} A_{ij}$. By understanding vector division and multiplication coordinate-wise in this context, this can be written in its common, more compact form

$$x^{(n+1)} = \frac{x^{(n)}}{A^*\mathbf{1}} A^* \left(\frac{y}{Ax^{(n)}}\right). \tag{4.40}$$

For the detailed derivation of the update, we refer to [195], in which EM was first applied to imaging with Poisson distributed data. The more interesting perspective on MLEM for our work is the one as a (preconditioned or mirrored) gradient method, see, e.g. [28]. For a linear inverse problem with Poisson observations, the negative log-likelihood

$V = -\log \rho_{\text{like}}$ was given in (3.14). Its gradient can be written as

$$\nabla V(x) = A^*\mathbf{1} - A^*\left(\frac{y}{Ax}\right). \tag{4.41}$$

Note that by introducing a step size parameter $\tau = 1$, (4.40) can be reformulated to

$$\begin{aligned} x^{(n+1)} &= x^{(n)} - \tau \frac{x^{(n)}}{A^*\mathbf{1}}\left(A^*\mathbf{1} - A^*\left(\frac{y}{Ax^{(n)}}\right)\right), \\ &= x^{(n)} - \tau(\nabla^2 h(x^{(n)}))^{-1}\nabla V(x^{(n)}). \end{aligned} \tag{4.42}$$

This shows that MLEM can also be understood as an instance of preconditioned gradient descent, or, via approximation (4.35), as a first-order approximation of mirror descent with $\tau = 1$. The preconditioning is induced by the weighted negative entropy mirror function

$$h(x) \coloneqq \sum_{j=1}^{d} a_j(x_j \log(x_j) - x_j), \tag{4.43}$$

with the gradient $(\nabla h(x))_j = a_j \log(x_j)$ and Hessian

$$\nabla^2 h(x) = \operatorname{diag}\left(\frac{a_1}{x_1}, \dots, \frac{a_d}{x_d}\right). \tag{4.44}$$

This mirror function is a standard choice to enforce positivity of iterates, and indeed, the preservation of positivity in MLEM is a main feature leading to its popularity in photonic imaging problems, since emission cannot take negative values. Having made this connection to gradient methods, it is easy to add regularizers to MLEM in a similar way as we add a regularization to the likelihood in variational regularization (4.4). The continuous-time limit $\tau \to 0$ of (4.42) is the preconditioned gradient flow

$$\dot{x}_t = -(\nabla^2 h(x_t))^{-1}\nabla V(x), \tag{4.45}$$

Now, instead of setting $V$ as the negative log-likelihood, we simply use the negative log-density of the posterior $V = -\log \rho_{\text{post}}$ in order to derive numerical schemes. Since the inverse Hessian $(\nabla^2 h(x))^{-1}$ is diagonal and simple to evaluate, the preconditioning adds only little computational cost. For instance, if we impose a TV prior $\rho_{\text{prior}} \propto \exp(-\beta\,\mathrm{TV}(\cdot))$ on a photonic imaging problem with linear forward model, set $V = -\log \rho_{\text{post}}$ and discretize (4.45) by a forward-backward method (with the preconditioner evaluated explicitly), this leads to a scheme of the form

$$\begin{cases} x^{(n+\frac{1}{2})} = \frac{x^{(n)}}{A^*\mathbf{1}} A^*\left(\frac{y}{Ax^{(n)}}\right) \\ x^{(n+1)} = x^{(n+\frac{1}{2})} - \frac{\beta x^{(n)}}{A^*\mathbf{1}} p^{(n+1)} \end{cases} \tag{4.46}$$

where $p^{(n+1)} \in \partial\,\mathrm{TV}(x^{(n+1)})$. Here, the forward step is a standard EM-step and the backward step amounts to evaluating a (weighted) proximal mapping for the TV func-

tional. This iteration is known as EM-TV and was proposed and analyzed in [190–192]. In a similar way, we can come up with other regularized modifications of the MLEM algorithm using any of the other Gibbs priors discussed in Section 4.1.

## 4.5. Posterior Sampling and Markov Chains

While MAP computation for log-concave measures can draw from the decades of research in convex optimization, we saw in Section 4.3 that the quantities we need to compute for practical downstream tasks go beyond MAP estimates. A common property of many of the estimators is their representation as an integral with respect to the posterior of the form

$$\int_{\mathbb{R}^d} H(x)\, \rho_{\mathrm{post}}(x \,|\, y)\, \mathrm{d}x, \tag{4.47}$$

where $H$ is a general functional or mapping defined on $\mathrm{supp}(\rho_{\mathrm{post}}(\cdot \,|\, y)) \subset \mathbb{R}^d$. For example, for computing CM, we need to evaluate (4.47) with $H = \mathrm{Id}$. For posterior covariance, we set $H(x) = (x - x_{\mathrm{CM}})(x - x_{\mathrm{CM}})^T$. In the context of extreme value probabilities, we discussed integrals of the form (4.23), which coincide with (4.47) for $H = \mathbf{1}_{\mathcal{C}}$ with a suitable set $\mathcal{C}$. This shows that many tasks in Bayesian inference (apart from MAP computation) come down to finding an efficient quadrature or integration method for (4.47). For the following standard discussions, see, e.g., [106, chap. 3.6].

**Monte Carlo Integration.** Since the integral is high-dimensional and we may not have access to the posterior density $\rho_{\mathrm{post}}(\cdot \,|\, y)$ because of its unknown normalization constant, the standard method is Monte Carlo integration. Assuming that $x_1, \dots, x_N$ are samples drawn from the posterior $\rho_{\mathrm{post}}$, it approximates

$$\int_{\mathbb{R}^d} H(x)\, \rho_{\mathrm{post}}(x \,|\, y)\, \mathrm{d}x \approx \frac{1}{N} \sum_{j=1}^{N} H(x_j). \tag{4.48}$$

If the samples $x_j$ are independently drawn from $\rho_{\mathrm{post}}(\cdot \,|\, y)$, it is a standard result that the approximation (4.48) converges as $N \to \infty$, in the sense that the expected error decreases as $\mathcal{O}(N^{-\frac{1}{2}})$. This is a rather slow convergence, which means that good approximations of estimators require $N$ to be large. Nevertheless, Monte Carlo integration is often used, since it exempts from the need to compute the density's normalization constant and is therefore the single fastest method in inference computations.

**Markov Chains.** The Monte Carlo integration shifts the focus to the sample generation, as the central question is now how to efficiently generate (and potentially store) $x_1, \dots, x_N$ for large $N$. The standard in high-dimensional problems uses Markov chains.

By a Markov chain, we refer to a sequence $(X^{(n)})_n$ of random variables, generated by a transition kernel $P : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ with the property

$$\mathbb{P}[\Gamma \mid X^{(n+1)} \in A] = \mathbb{P}[\Gamma \mid X^{(n+1)} \in A] = P(X^{(n)}, A). \tag{4.49}$$

By definition, $P(\cdot, A)$ has to be measureable for any $A \subset \mathbb{R}^d$, and $P(X, \cdot)$ should be a probability measure over $\mathbb{R}^d$, not necessarily equipped with a Lebesgue density. The first equality above ensures the Markov property, saying that, conditioned on a current state $X^{(n)}$, the next state $X^{(n+1)}$ is mutually independent of any past state. The second equality ensures time homogeneity in the sense that the distribution of any state can be described by the transition kernel. Note that we will use the following, common notation and definitions:

- If $X^{(n)} \sim \mu^{(n)}$ for some probability measure $\mu^{(n)}$, we use the notation $\mu^{(n+1)} = \mu^{(n)} P$ to denote the distribution change under a step along the Markov chain.
- A measure $\mu$ is an invariant or stationary measure of the Markov chain if $\mu = P\mu$.
- We define the $k$-step transition kernel by $P^{(k)}(X, A) = \mathbb{P}[\Gamma \mid X^{(n+k)} \in A]$, which does not depend on $n$ by (4.49).
- The chain is called irreducible with respect to $\mu$ if for all $A \in \mathcal{B}(\mathbb{R}^d)$ with $\mu(A) > 0$ and all $x \in \mathbb{R}^d$, there is some $k$ such that $P^{(k)}(x, A) > 0$.
- An irreducible chain is called periodic if there is some $k \geq 2$ and some sets $\{A_1, \dots, A_k\}, A_j \subset \mathbb{R}^d$ disjoint and nonempty such that $P(x_j, A_{j+1 \,(\mathrm{mod}\, k)}) = 1$ for all $x_j \in A_j$. A chain is called aperiodic if it is not periodic.

The following result justifies the use of Markov chains when generating samples for Monte Carlo integration, see [106, Prop. 3.11]. The resulting framework of employing Markov chain samplers in order to estimate (4.47) is called Markov chain Monte Carlo (MCMC).

**Lemma 4.3: Convergence of MCMC**

Let $(X_j)_j$ be an irreducible, aperiodic Markov chain with transition kernel $P$ and invariant measure $\mu$. Then, for all $x \in \mathbb{R}^d$ and $B \in \mathcal{B}(\mathbb{R}^d)$, it holds

$$\lim_{N \to \infty} P^{(N)}(x, B) = \mu(B),$$

and if $H : \mathbb{R}^d \to \mathbb{R}$ is $\mu$-integrable, it holds

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} H(X^{(n)}) = \int_{\mathbb{R}^d} H(x) \mu(\mathrm{d}x). \tag{4.50}$$

The algorithmically challenging part of MCMC is the design of a suitable Markov chain that guarantees rapid convergence in (4.50). This design question for the special case $\mu = \rho_{\text{post}}(\cdot \,|\, y)$ is the main focus of the remaining chapter.

**Metropolis–Hastings Correction** Our goal is to sample from a posterior $\rho_{\text{post}}(\cdot \,|\, y)$ by simulating a Markov chain, hence one of the basic requirements as per Lemma 4.3 will be that $\mu^* = \rho_{\text{post}}(\cdot \,|\, y)$ is the Markov chain's stationary distribution. As we will see, this is not satisfied by many of the chains that we will consider throughout this thesis. In several of our results, we will show that the stationary distribution is "close" to the posterior in some metric, but for some applications it is vital that the samples follow (at least asymptotically) the correct distribution. A way to guarantee this is by modifying the transition kernel using Metropolis–Hastings correction. Suppose we are given a Markov chain with transition kernel of the form

$$P(x, A) = \int_B k(x, y) \, \mathrm{d}y, \tag{4.51}$$

where we have access to the closed form transition density $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. In order to guarantee $\mu^*$ is stationary, it can be shown that the following detailed balance condition is sufficient

$$\mu^*(x)k(x, y) = \mu^*(y)k(x, y). \tag{4.52}$$

Whenever the so-called proposal kernel $P$ from (4.51) does not satisfy (4.52), we can use the corrected kernel

$$\begin{aligned} \tilde{k}(x, y) &= \alpha(x, y)k(x, y), \\ \alpha(x, y) &= \min\left\{1, \frac{\mu^*(y)k(y, x)}{\mu^*(x)k(x, y)}\right\}. \end{aligned}$$

The corrected kernel $\tilde{P}(x, A) = \int_A \tilde{k}(x, y) \, \mathrm{d}y$ satisfies detailed balance (4.52) and thus has the correct stationary distribution. Algorithmically, it can be implemented by a simple accept-reject step. The Metropolis–Hastings corrected scheme draws samples by alternating the proposal step $P$ and this accept-reject step. Given sample $X^{(n)}$ at the $n$-th step of the algorithm, $X^{(n+1)}$ is generated as follows:

1. Draw a candidate $Y \sim P(X^{(n)}, \cdot)$
2. Compute the acceptance ratio $\alpha(X^{(n)}, Y)$.
3. Draw $r \sim \text{Unif}([0, 1])$ uniformly distributed between 0 and 1.
4. If $r \leq \alpha(X^{(n)}, Y)$, accept the candidate, set $X^{(n+1)} = Y$. Otherwise reject and set $X^{(n+1)} = X^{(n)}$.

Note that a main feature of the correction factor $\alpha$ is its independence of any normalization constants of $\mu^*$, allowing the technique to be applied on high-dimensional Bayesian posteriors. The main disadvantages of Metropolis–Hastings correction are typically its

computational inefficiency and the difficult tuning of the candidate distribution. If the proposal is designed such that acceptance ratio is too low, the algorithm will be stationary for many steps before jumping to the next sample. Conversely, an acceptance ratio close to one often indicates that the proposal distribution itself makes inefficiently small jumps around the state space. The optimal acceptance rate has been computed for very simple proposals like a random walk (optimal acceptance rate $\approx 0.234$, see [85]) or an unadjusted Langevin proposal step (optimal rate $\approx 0.574$, see [178]), this can allow tuning hyperparameters to make the correction as efficient as possible.

# Chapter 5

# Experiments: MAP Estimation in Compton Camera Imaging

In this chapter, we apply the optimization routines outlined in Section 4.4 for MAP computation to the Compton camera image reconstruction problem. This allows us to evaluate the suitability of the developed forward models on data from simulation studies and lab measurements. The resulting imaging problems are ill-posed and underdetermined, requiring us to impose prior distributions on the solution space.

## 5.1. Model Matrix Computation

We formulated the forward model as a linear integral operator (3.22) between Banach spaces. For practical implementation of the model, we describe now how to discretize the image and data spaces and compute the corresponding forward model matrix.

**Discretization of data and reconstruction spaces.** When computing the reconstruction, the source activity is discretized, meaning we approximate $f$ by

$$f(\mathbf{s}) \approx \sum_{j=1}^{N} f_j \psi_j(\mathbf{s}), \tag{5.1}$$

where $\{\psi_j, j = 1, \dots, N\}$ are appropriately normalized basis functions, and $f_j$ the coefficients we aim to reconstruct. The basis functions have compact, disjoint support, and $\Omega = \cup_j \operatorname{supp}(\psi_j)$ is the area in which we reconstruct the activity distribution. In practical decommissioning applications $\Omega$ should cover the region of emission activity of the imaged nuclides in order to avoid background effects. Within this work, we restrict ourselves to simple surfaces for the activity distribution. Assuming two-dimensional,

rectangular areas and choosing $\psi_j$ as pixel basis functions allows to make use of image priors. The corresponding measurement can be thought of as an area on the floor or a wall in the decommissioning building, covering most simple geometries in realisitic experiments.

Inserting (5.1) into the forward model projects the data space to the span of point spread functions

$$\begin{aligned} g(z) &\approx \sum_j f_j (\mathcal{P}\psi_j)(z) \\ &= \sum_j f_j \int_\Omega \psi_j(\mathbf{s}) k(z, \mathbf{s}) \, \mathrm{d}\mathbf{s} \end{aligned}$$

due to linearity of the forward model.

**List-mode versus bin-mode reconstruction.** Due to finite spatial resolution, the space of possible data $z = (i_A, i_B, E_A)$ is already partly discretized, as we integrated the count rates over detector volumes in Eq. (3.10) and only register detector indices $i_A, i_B$ instead of exact positions. With this detector binning, the data $g$ consists of $n_P$ lists of energy measurements, where $n_P$ is the number of detector pairs in the camera.

In this work, we choose a *bin mode* approach in which the energy $E_A$ is also discretized. Since the accuracy of an energy measurement depends on the energy itself, see (3.12), the binning is not chosen uniformly, but rather with increasing bin sizes as $E_A$ grows. According to model (3.12), the energy windows should be chosen of size $\mathcal{O}(\sqrt{E_A})$. The binning $E_{\min} = E_0 < E_1 < \cdots < E_L = E_{\max}$ determines the dimensionality of the discretized inverse problem: After energy binning, the data of a monochromatic source has $n_P L$ elements. The discretized inverse problem reads

$$g = Pf, \tag{5.2}$$

with $g \in \mathbb{R}^{n_P L}$, $f \in \mathbb{R}^N$. The matrix $P$ is computed by estimating the point spread function within one detector pair and energy bin, i.e., by approximating $P_{i,j} = \mathcal{P}\psi_j((i_A)_i, (i_B)_i, (E_A)_i)$. Various computational methods for the forward model are discussed below.

For completeness, let us mention that in detector setups with good spatial and energy resolution, the data dimension $n_P L$ can exceed $10^6$ bins, yielding computation of $P$ prohibitive. In this case, if the measurements are a list of $K$ events and $K \ll n_P L$, it may be benefitial to reconstruct in so-called *list mode*. In list mode, the data space is kept continuous and one instead estimates the values $P_{k,j} = \mathcal{P}\psi_j((i_A)_k, (i_B)_k, (E_A)_k)$ for each event $k = 1, \ldots, K$. Upon proper normalization by the measurement system's total sensitivity, optimization algorithms like MLEM can be modified to list-mode variants, which allow significant speed-up in the low-count case. List-mode MLEM was developed for PET, as it is a similarly underdetermined imaging application with list data [16,

163, 50]. For Compton cameras with good spatial resolution (pairs of interaction points with corresponding energy measurements), list-mode algorithms have successfully been employed in reconstruction [219, 80].

Due to the limited number of pairs in decently sized systems with monolithic detectors, the bin mode is more efficient in our case [103]. With the $CeBr_3$ detectors we employ, the energy resolution allows separation of the spectra into roughly $L \approx 50 - 100$ bins depending on the maximum energy range. In the RSL7 setup, there are $n_P = 6$ scatterer-absorber pairs. This leads to a data space with $300 - 600$ total bins. It is important to note that this makes the imaging problem $g = Pf$ underdetermined, as the sought-for image resolution can be orders of magnitude higher depending on the geometry of the measurement.

**Analytic, Monte Carlo and learned methods.** There are several different ways to compute the entries of the discrete forward model $P_{i,j}$. Inspecting the unidirectional, uncorrected forward model derived in Section 3.2, it appears straightforward to estimate the integral via standard quadrature methods. This is what we refer to as *analytic models*. Building this type of forward model has been explored in [219, 145, 135]. The computationally expensive part of the pipeline is the calculation of the weights $C_1, C_2$, since they involve several terms like attenuation, scatter probabilities and differential cross sections that all depend on the geometry and materials in the system. At moderate resolutions of $10^4 - 10^5$ pixels or point spread functions and quadrature over approximately $10^4$ pairs of interaction points $(\mathbf{x}, \mathbf{y})$ in a detector pair, these are computable in a few seconds on a standard CPU.

The main complications of the analytic approach in our experiments were the resulting large model errors. This was the main motivation of the second order corrected model $\mathcal{R}_c^{(2)}$. Since $\mathcal{R}_c^{(2)}$ involves an additional integration over intermediate energies $E_1$, the resulting quadrature becomes more costly. Considering approximately $10^2$ intermediate energies—equal to the numbers of interaction points in each detector—the evaluation of the forward model becomes $10^2$ times more costly. In experiments, we found that going lower than 50 energy values or interaction points in turn resulted in larger discretization errors in the forward model computation. For monochromatic sources with a single value of $E_0$, the computational cost of the corrected model is still manageable by efficient implementation using parallelization. As we will see in the next section, however, the computational cost also grows quadratically in the number of emission lines, deeming the correction approach impractical for applications with larger nuclide libraries.

An orthogonal approach to analytic matrix computation is the estimation via ray-tracing Monte Carlo. For Compton cameras, this computation has been proposed in [218]. The approach consists in building a digital twin of the camera in a Monte Carlo toolbox like FLUKA or GEANT4, and estimating each point spread function by simulating them one by one. The Monte Carlo computation of the model matrix is much more costly than quadrature-based analytic approaches. However, by its nature, model errors are

only caused by differences between digital twin and real camera, and are typically much smaller than in analytic approaches. If the pixel space remains identical across measurements, Monte Carlo weights can be precomputed once on a larger cluster and therefore be preferred over analytic approaches. Therefore, the approach is widespread for setups with static pixel spaces, e.g., in astronomical Compton telescopes [223] where the pixel space remains static in the far field. In near field applications like our decommissioning approach, the geometry changes between measurements, requiring in theory to recompute the model between measurements. We refer to [145] for a comparison between Monte Carlo simulated and analytic estimation of point spread functions in Compton cameras.

A third, more recent approach is to deploy a hybrid pipeline using Monte Carlo simulations and deep learning methods. Even in a changing measurement geometry in near field, Monte Carlo simulations can be precomputed on a three-dimensional grid and stored in binned format. When the measurement geometry changes, the problem of building the forward model is a question of correctly interpolating between the point spread functions on precomputed grid points, a problem well-suited for deep learning methods [10]. The forward model is continuous in the point position, so operator learning techniques [115] can offer a way to learn parametrized representations of the point spread function. The feasibility of reconstruction from simulated data using learned forward models for the RSL7 camera was explored in [38].

**Polyenergetic imaging.** One further complication in the forward model is the simultaneous existence of multiple source energies $E_0$. In practice, we want to reconstruct the activity of a given nuclide from a predefined library. An example of a monochromatic nuclide in decommissioning is Caesium-137, see Table 2.1. However, most nuclides emit photons at several energy levels $E_0^l, l = 1, \dots, L$. Further, in realistic scenarios, several nuclides are present in the imaged environment at the same time. Robust polyenergetic imaging with Compton cameras is a challenging task which requires accurate forward models and careful reconstruction [148].

The reason for this challenge is mainly that different source emission lines may influence each other and can not be reconstructed separately. For a Cobalt-60 source, measured data has two photopeaks corresponding to the prominent emission lines 1173 keV and 1332 keV. The two energies have approximately equal emission probabilities in a nucleus decay (99.88% and 100%, see the decay scheme Fig. 2.1). The data at the smaller peak experiences *energy leakage* from the higher emission line by incomplete absorption or a third energy loss outside the detectors. This is visible in Fig. 5.1—the displayed spectrum is the sum of scatter energy and absorber energy from all coincidence events. The peak at the lower emission line 1170 keV receives almost twice as many counts as the higher line around 1332 keV. This large difference is not attributed to inconsistencies in the camera's sensitivity, but rather stems from incomplete absorptions of higher emission line photons.

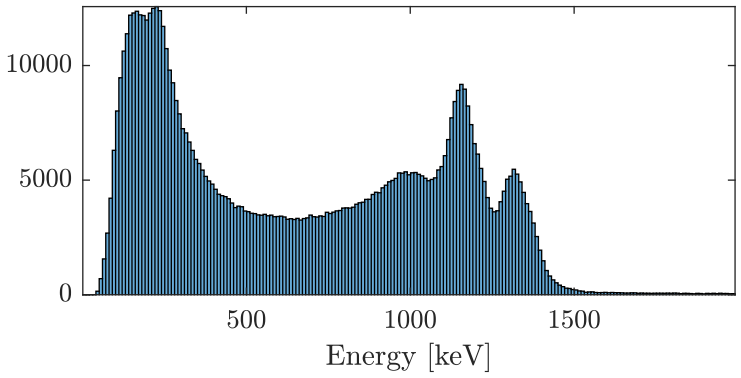



Figure 5.1.: Visualization of the energy leakage phenomenon in measurements of nuclides with multiple emission lines, here Cobalt-60 as an example. The histogram shows the values $\Delta E_1 + \Delta E_2$, i.e., the energy depositions in scatterer and absorber detector added up for each coincidence (the shape resembles a spectroscopic measurement with a single detector, but the underlying distribution of this sum spectrum is different). The data stems from an RSL7 lab measurement of a $^{60}$Co line source, see also the second row of Fig. 5.9.

In order to explain the leakage correction in the forward model, consider for a moment that the counts from different emission lines were independent from each other, i.e., separable in the data space. Assume we are reconstructing the source activity of nuclides $(\mathbf{n}_1, \ldots, \mathbf{n}_M)$, each $\mathbf{n}_m$ with emission lines $E_0^{m,1}, \ldots, E_0^{m,n_m}$. Then we reconstruct $M$ images $f = (f^{(1)}, \ldots, f^{(M)})$, one for each nuclid, from $g = Pf$, where

$$P = \begin{pmatrix} P^{1,1} & & & \\ \vdots & & & \\ P^{1,n_1} & & & \\ & \ddots & & \\ & & P^{M,1} & \\ & & \vdots & \\ & & P^{M,n_M} & \end{pmatrix} \tag{5.3}$$

Here each block $P^{m,l}$ is of size $n_P L \times N$, where $n_P$ is the number of detector pairs, $L$ the number of energy bins, $N$ the number of pixels in a reconstructed image. The number of nonzero blocks is $\sum_{m=1}^{M} n_m$, where $n_m$ is the number of emission peaks for nuclide $\mathbf{n}_m$ and $M$ the total number of nuclides.

The data $g$ is assembled by extracting from all counts the fully absorbed ones at the lines $E_0^{m,l}$, by filtering for the sum of the two deposited energies in each measured event. In this simplified case, the images can be computed independently from each other. The error of this approach can be seen in the assembly of the data: By extracting counts

with $E_A + E_B \approx E_0^{m,l}$ for some emission line $E_0^{m,l}$, we are including partial deposition events from other nuclides with emission lines higher than $E_0^{m,l}$. This leakage acts very similar to "background" noise, but varies strongly across the energy spectrum and is typically not well represented by a single beckground correction parameter [148]. The corrected, polyenergetic forward model would therefore be

$$P + Q = P + \begin{pmatrix} Q_1^{1,1} & \dots & Q_M^{1,1} \\ \vdots & \ddots & \vdots \\ Q_1^{M,n_M} & \dots & Q_M^{M,n_M} \end{pmatrix}. \tag{5.4}$$

Here $Q_k^{m,l}$ is the total leakage of the nuclide $\mathbf{n}_k$ to the emission line $E_0^{m,l}$. This can be computed as $Q_k^{m,l} = Q_{k,1}^{m,l} + \dots + Q_{k,n_k}^{m,l}$, where $Q_{k,i}^{m,l}$ is the leakage of the emission line $E_0^{k,i}$ of $\mathbf{n}_k$ to the emission line $E_0^{m,l}$ of nuclide $\mathbf{n}_m$. The computation of the blocks in $Q$ is very similar to the blocks in $P$ outlined in the modeling section. The only change is that we do not consider probabilities of absorption in the second detector, but rather a second scattering with the energy loss corresponding to the measured energy value. Since leakage always occurs from higher to lower emission lines, approximately half of the terms $Q_{k,i}^{m,l}$ have zero contribution. Still, the number of correction terms $Q$ grows quadratically in the number of total emission lines, significantly increasing the total computational effort. For the total 15 emission lines of the four nuclides in Table 2.1, we need to compute 15 blocks in $P$ for complete absorption at the emission lines, and $15 \cdot 14/2 = 105$ leakage correction blocks in $Q$, so the total computational effort is approximately 120 times the cost of a monochromatic model.

## 5.2. Choice of Optimization Routine

Assume now we placed the Compton camera in a measurement geometry and decided for a discretization of image domain and data space.

**Likelihood formulation.** Having fixed the geometry, we can compute the forward model $P$. Due to Poisson statistics of radioactive decay, the negative log-likelihood is given by

$$-\log \rho_{\text{like}}(g \mid f) = \sum_{j=1}^{J} (Pf)_j - g_j \log((Pf)_j) + \text{const.} \tag{5.5}$$

Ideally, we have access to a background measurement or estimate $b$ taken with the camera in a similar environment. In that case, the background-corrected likelihood is

$$-\log \rho_{\text{like}}(g \mid f) = \sum_{j=1}^{J} (Pf)_j + b_j - g_j \log\left((Pf)_j + b_j\right) + \text{const.} \tag{5.6}$$

As background radiation levels can contribute significantly to the total counts and thereby distort measurements, a good estimate is vital for accurate reconstruction.

**Maximum likelihood estimation.** In a simple, non-Bayesian setting, we can try to reconstruct the radiation map $f$ given $g$ by minimizing (5.5) or (5.6), respectively. For this, we employ the standard bin-mode MLEM iteration

$$f^{(n+1)} = \frac{f^{(n)}}{P^*\mathbf{1}} P^* \left( \frac{g}{Pf^{(n)} + b} \right), \tag{5.7}$$

with an initial guess $f^{(0)}$.

**Choice of prior distribution.** The imaging problem $Pf = g$ with the camera setups in this work is ill-posed and underdetermined. In order to incorporate more information about the solution and regularize the reconstruction, we impose a prior $\rho_{\text{prior}}(f)$ on the image. In realistic decommissioning situations, emission activity is distributed either as larger spots or around cracks or holes in the walls. Contaminations are typically several months or years old and have dispersed, especially so for secondary activations. Due to the broad class of potential emission distributions, we select $\rho_{\text{prior}}$ to be fairly non-restrictive, penalizing only strong variations in the emission distribution by setting $\rho_{\text{prior}}(f) \propto \exp(-\lambda \operatorname{TV}(f))$ with a pre-selected temperature $\lambda > 0$. Given the Poisson distributed nature of measurement data $g$, the Bayesian posterior has negative log-density

$$\begin{aligned} -\log \rho_{\text{post}}(f \,|\, g) &= -\log \rho_{\text{prior}}(f) - \log \rho_{\text{like}}(g \,|\, f) + \text{const.} \\ &= \lambda \operatorname{TV}(f) + \sum_{j=1}^{J} (Pf)_j + b_j - g_j \log\left((Pf)_j + b_j\right) + \text{const.} \end{aligned} \tag{5.8}$$

**Maximum-a-posteriori estimation.** MAP estimation corresponds to minimizing (5.8), a convex variational optimization problem. We employ the following damped variant of the EM-TV algorithm given in (4.46)

$$\begin{cases} f^{(n+\frac{1}{2})} = \frac{f^{(n)}}{A^*\mathbf{1}} A^* \left( \frac{g}{Af^{(n)}} \right) \\ f^{(n+1)} = (1-\omega) f^{(n)} + \omega \left( f^{(n+\frac{1}{2})} - \frac{\lambda f^{(n)}}{A^*\mathbf{1}} p^{(n+1)} \right), \end{cases} \tag{5.9}$$

where $p^{(n+1)} \in \partial \operatorname{TV}(f^{(n+1)})$. The implicit second half step can be realized by solving the weighted total variation denoising problem

$$f^{(n+1)} = \arg\min_f \left\{ \frac{1}{2} \| f - (\omega f^{(n+\frac{1}{2})} + (1-\omega) f^{(n)}) \|^2_{w^{(n)}} + \omega\lambda \operatorname{TV}(f) \right\}, \tag{5.10}$$

with a coordinate-wise weight $w^{(n)} = \frac{K^*\mathbf{1}}{f^{(n)}}$ to the norm. Computationally, we estimate this inner step by running a fixed number of steps of accelerated proximal gradient descent on the dual formulation of (5.10), adapted from [52, Example 4.8]. This variant of the EM-TV algorithm relates the intermediate EM iterate with the previous TV denoised iterate via a convex combination with the damping parameter $\omega$. It was proposed in [192], which proved monotone descent of the objective and superior numerical behaviour of the damped modification.

## 5.3. Simulation and Measurement Setup

Most of our experiments focus on the the RSL7 prototype camera shown in Fig. 3.1, or a digital twin of the camera in simulations. The reconstruction area is a $l \times l$ square on a wall, with $l = 2\,\mathrm{m}$ or $l = 4\,\mathrm{m}$ in most of the two-dimensional results. The camera is placed at a distance of $\Delta z$ in the range of 1–10 m, such that the detector cylinder caps are parallel to the wall and the camera centered relative to the square.

We reconstruct the source distributions at $64 \times 64$ resolution. The low resolution is not due to computational restrictions, but rather practical requirements and the limited angular resolution of the camera: Depending on the energy level and signal-to-noise ratio, prototype experiments showed that a lower limit for the possible angular resolution in simpler, one-dimensional experiments was at least 3°. If $2\Delta z = l$, meaning the image area covers a 90° field of view, a $64 \times 64$ resolution implies that the best resolved pixels at the center of the image cover 1.8°, which is already smaller than the achievable resolution by the camera.

The expected spatial resolution matches with what we can expect in terms of data resolution: The energy space is discretized into $L = 90$ energy bins covering the interval $[E_{\min}, E_{\max}] = [50\,\mathrm{keV}, 2000\,\mathrm{keV}]$. The bins are not of uniform width, but rather controlled by the resolution of the camera—around energy $E$, the width of a bin is $2\eta_A\sqrt{E}$, two standard deviations of the energy measurements under the noise model (3.12). Given a nuclide of interest, we discard all bins above the maximum emission energy, e.g., if we only reconstruct $^{60}$Co, all bins above 1332 keV are discarded as they can only detect background noise, leaving 72 bins. Since we are interested in coincidence events only, the measurements carrying information about the source are usually focused on a narrow band of bins, making the relevant scatter data concentrated on a small subset of the 90 bins. This gives an effective data dimensionality $\ll n_P L = 6 \cdot 90 = 540$ in the monochromatic case. This fact shows the limitation of image resolution without accurate prior knowledge: The image size of $64^2 \approx 4 \cdot 10^3$ pixels makes the problem underdetermined by 1-2 orders of magnitude and is hence already an optimistic choice.

Simulated ray-tracing Monte Carlo data allows us to distinguish counts by the number of depositions in the scatter detector, giving a decomposition of the data of the form

$$g = g^{(1)} + g^{(2)} + \dots,$$

where $g^{(i)}$ is the binned data from all counts that scattered exactly $i$ times in the scatter detector. When referring to "clean data", we mean simulated measurements with exact energy values in each count, no background radiation and no Poisson statistics[1]. Contrary to that, for simulated, "noisy" data, we distort the counts in a number of post-processing steps. First, Poisson count statistics are simulated by drawing the total number of decay events from a Poisson distribution. To the resulting list of measured events, energy noise is added to each count in a second step according to the noise model (3.12). Finally, the counts are randomly mixed with a background spectrum of equal measurement time which was measured by the real, physical RSL7 camera in a lab environment.

## 5.4. Results

In this section, we present results ordered from simple to more complex experiments: We start with a validation experiment for the single scatter model (3.8) with single scattered, simulated data. By "simulated", we mean synthetic measurement data generated by ray-tracing Monte Carlo simulation.

**Single scatter model.** We solve the discretized problem (5.2), where $P = P^{(1)}$ is the quadrature approximation of the single scatter model $\mathcal{P}^{(1)}$. Multiple scattering correction is not considered in the forward model for now, and the problem is simplified by reconstructing from the single scatter data only, meaning we solve

$$g^{(1)} = P^{(1)} f.$$

We reconstruct activity distributions of the approximately monochromatic source nuclide Caesium-137, which means we also do not correct for energy leakage between multiple peaks or nuclides.

Figure 5.2 shows that in the case of clean measurements, maximum likelihood estimation for this model can reconstruct the positions of point sources from their single scatter data. For the experiment displayed in the figure, we chose Caesium-137 sources with activities of 1 MBq with a simulated scan time of 10 minutes at distance $\Delta z = 1\,\mathrm{m}$, leading to $10^4 - 10^5$ count events per source, depending on the source position. In order to compute the reconstructions, we ran $10^4$ iterations of MLEM (5.7).

As Fig. 5.3 shows, however, the MLE reconstructions become distinctively worse if measured energy values are distorted like real scintillation measurements due to varying light output, Poisson statistics and background noise. The presence of multiple point sources at once further makes distinction less reliable. In order to stabilize the reconstruction,

[1] By this, we mean simulating exactly an expected number of decays and taking the resulting events as data. This does not result in the exact data due to Monte Carlo approximation errors, but disregards the additional fluctuation in counts due to Poisson decay statistics.

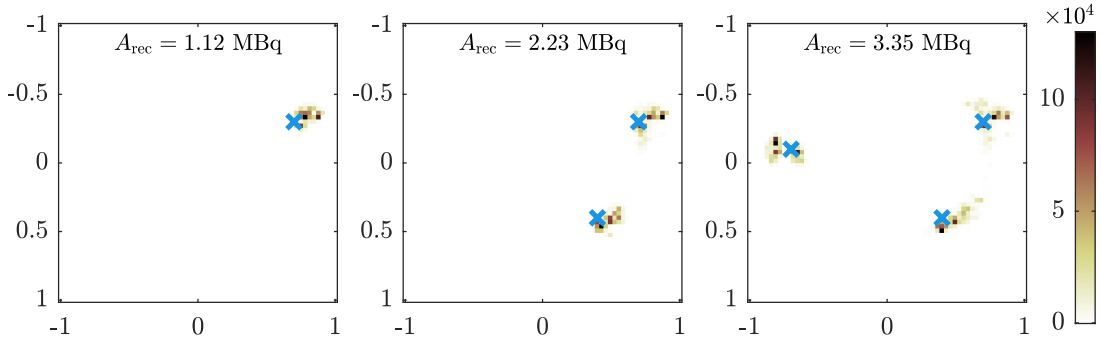



Figure 5.2.: MLE reconstructions of point sources from clean, single scattered data with no Poisson statistics, no background and no energy measurement noise. From left to right, we simulate a single, two and three 1 MBq Caesium-137 point sources (positions indicated in blue) present at the same time.

we correct for background radiation[2] and compute a regularized MAP estimate using $10^4$ iterations of the damped EM-TV algorithm (5.9) from [192]. The second row of Fig. 5.3 shows that this improves source identification and location estimation. With several sources at once and background noise present, however, the separation becomes susceptible to errors due to the low spatial resolution of the camera.

**Multiple scatter correction.** The restriction to single scatter data is unrealistic in practice. In Fig. 5.2 and Fig. 5.3, the snythetic data is filtered to contain only $g^{(1)}$, counts scattered exactly once in the scatter detector. In real applications without spatial resolution in the detectors, we cannot make the distinction between $g^{(1)}$ and $g^{(i)}, i \geq 2$ and have to reconstruct from all registered counts with any kind of deposition in the scatterer. As Fig. 3.6 shows, multiply scattered data can make up a majority of the counted events. The first row of Fig. 5.4 shows reconstructions from solving

$$g^{(1)} + g^{(2)} = P^{(1)} f,$$

for the activity distribution $f$.

These subfigures show how reconstruction quality can suffer even by adding only the second scattering effects when we use the single scattering model $\mathcal{P}^{(1)}$ from (3.11). The model mismatch leads to incorrectly located point sources, already in the simplest case

[2] In the synthetic data experiments, background counts are generated by adding counts of a lab background measurement. The lab measurement covered 30 minutes, so 10 minutes are simulated by randomly sampling a third of the counts and adding them to the simulated data. In order to test realistic background correction, we again sample from the longer measurement so that errors and correction follow the same statistics but are not identical.

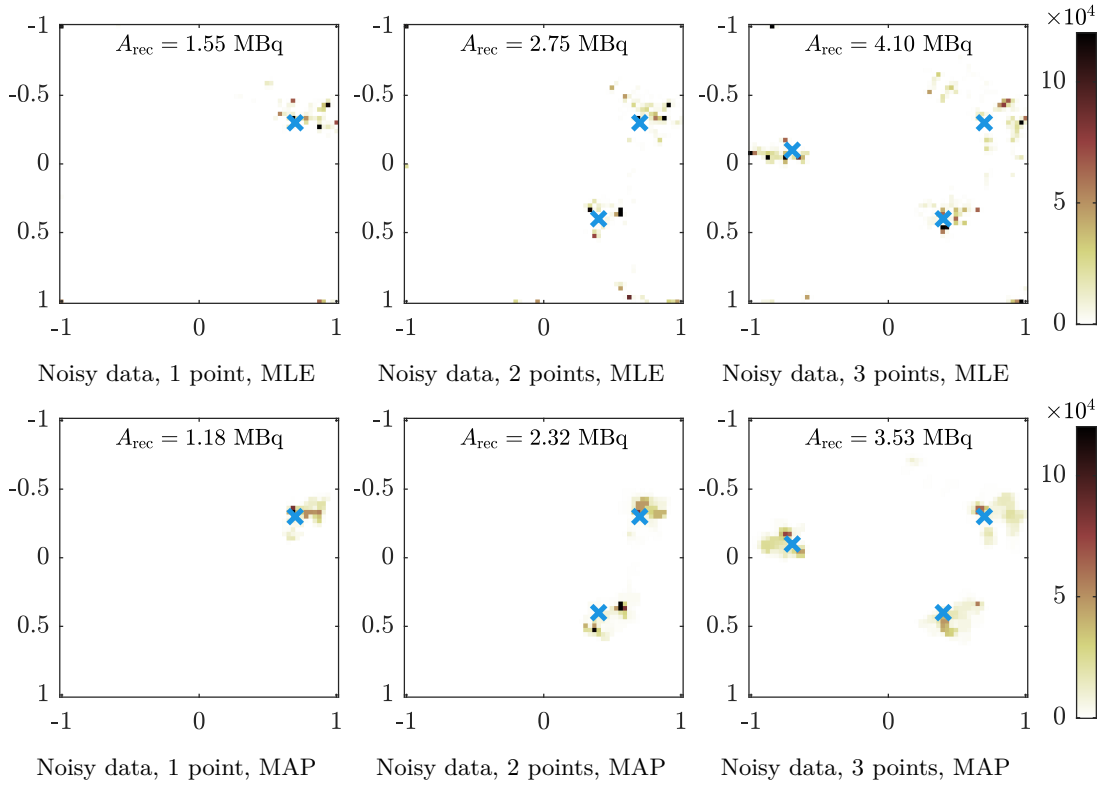



Figure 5.3.: MLEM and EMTV reconstructions of point sources from single scattered data. From left to right, we simulate a single, two and three Caesium-137 point sources (1 MBq each, positions indicated in blue) present at the same time. The data is distorted by adding background radiation, Poisson statistics and energy measurment noise according to (3.12). Top row: MLE reconstructions. Bottom row: MAP estimates with a TV prior. With prior information, the position estimation and total activity gets reconstructed more accurately.

of a single source. Due to the large number of multiply scattered events, they present a major difficulty in the reconstruction.

Due to this model incapacity, we developed the double scattering corrected model $\mathcal{P}^{(1)} + \mathcal{P}^{(2)}$ in in (3.21). The second row of Fig. 5.4 shows reconstructions of

$$g^{(1)} + g^{(2)} = (P^{(1)} + P^{(2)})f,$$

using the corrected model. For one and two point sources, the correction improves the reconstructed source positions substantially. With more than two sources and varying intensities, the model substantially improves over the uncorrected model, but artifacts due to background and Poisson noise are still clearly visible.

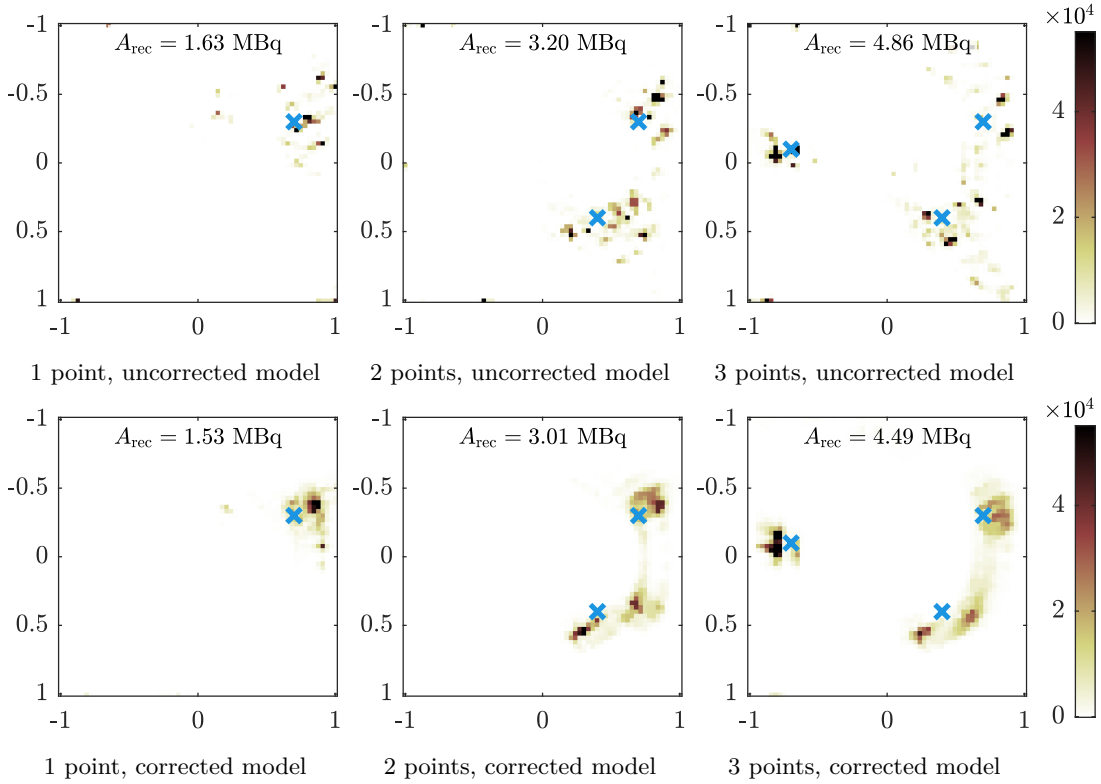



Figure 5.4.: EMTV reconstructions of point sources from noisy data $g^{(1)} + g^{(2)}$. From left to right, we simulate a single, two and three Caesium-137 point sources (1 MBq each, positions in blue) present at the same time. Top row: MAP reconstructions using the model $\mathcal{P}^{(1)}$ from (3.11). Bottom row: MAP estimates with the corrected model $\mathcal{P}^{(1)} + \mathcal{P}^{(2)}$ (3.21). Both rows use the same TV prior.

Empirical evidence from Monte Carlo simulations (see Fig. 3.6) showed that most of the registered counts undergo either single or double scattering. Higher order terms may be rare enough to view them as noise in the reconstruction problem. We test this hypothesis in further reconstructions in Fig. 5.5, where we reconstruct from the full data, solving

$$g^{(1)} + g^{(2)} + \cdots = g = (P^{(1)} + P^{(2)})f.$$

The reconstructions show increased artifacts compared to those of Fig. 5.4, likely due to higher-order scattered data acting as noise. For a single or two point sources, however, the reconstructed locations are still reliable in the present example.

The results show that for correcting second-order scattering in the data, the extended model $\mathcal{P}^{(1)} + \mathcal{P}^{(2)}$ is a promising correction for more accurace reconstruction. This is particularly true for cameras with large detectors, where multiple scattering is the

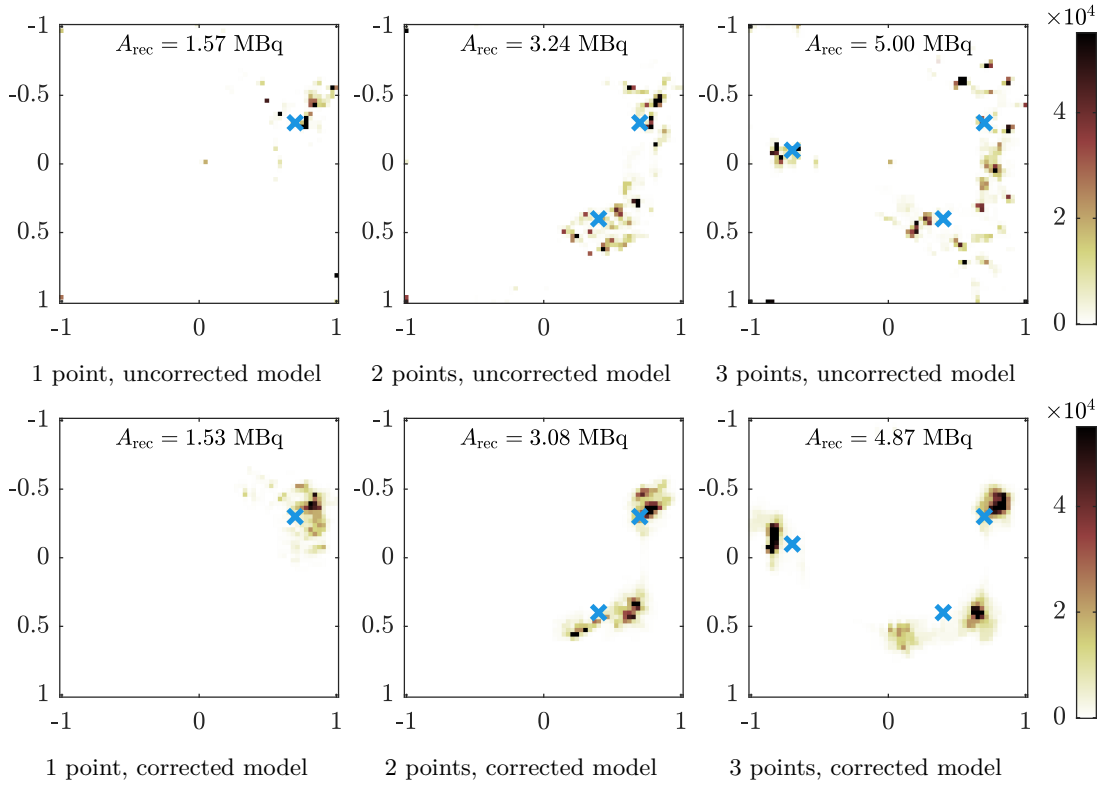



Figure 5.5.: EMTV reconstructions of point sources' full noisy data $g = \sum_i g^{(i)}$. We simulate a single, two and three Caesium-137 point sources (1 MBq each, positions in blue) present at the same time. Top row: MAP reconstructions using the uncorrected model $\mathcal{P}^{(1)}$ from (3.11). Bottom row: MAP estimates from the corrected model $\mathcal{P}^{(1)} + \mathcal{P}^{(2)}$ from (3.21). Both rows use the same TV prior.

rule rather the exception compared to single scattering. The additional integral in the second order model $\mathcal{R}_c^{(2)}$ results in increased computational time, which is also the reason why potential analytical models like $\mathcal{R}_c^{(3)}, \ldots$ for even higher-order scattering become difficult to compute. In simple activation scenarios, however, the higher-order terms can be viewed as an additional noise term, still allowing successful reconstruction.

**Source geometries beyond point sources.** The tests so far only targeted point sources - we now give results on simulated sources with more complex distributions of emission activity. The prevalent distribution of sources in practice spans line sources (e.g., around cracks and seals in building components) and area sources (e.g., contamination on walls, floors, or equipment surfaces). We therefore employ the same unrestrictive, log-concave total variation prior. Figure 5.6 shows MAP reconstructions of simulated Caesium-137 line and area sources at different positions and orientations in the field of

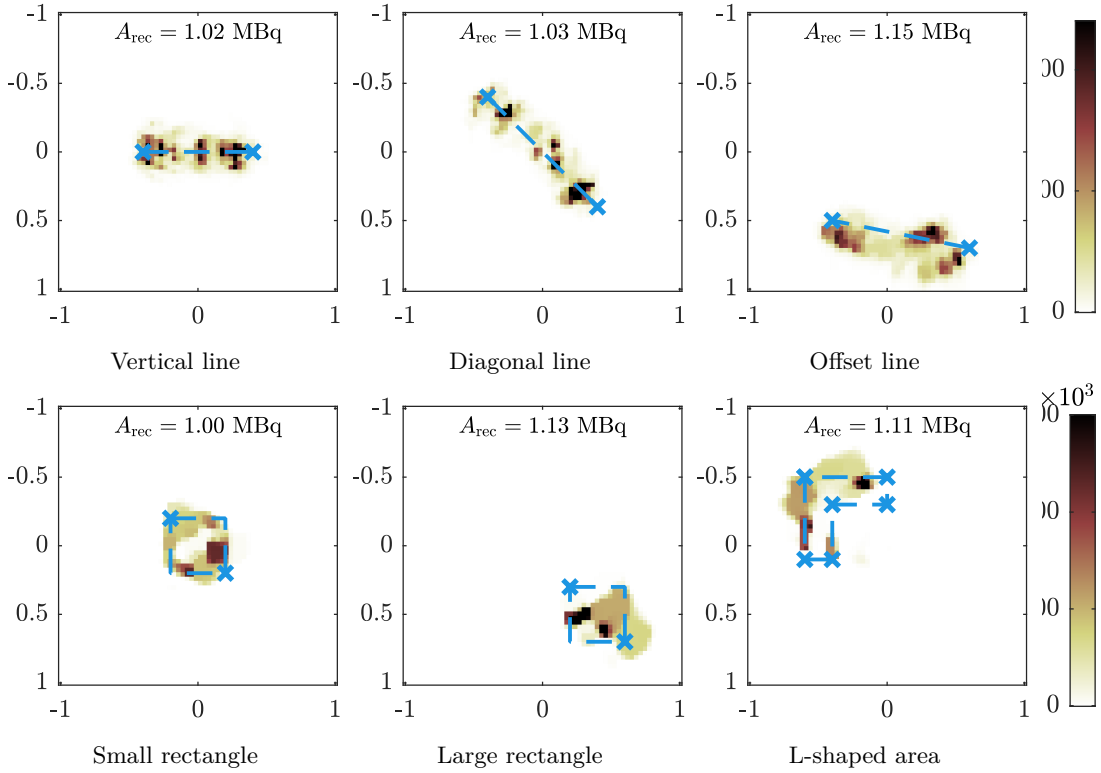



Figure 5.6.: MAP reconstructions of Caesium-137 line and area sources from noisy data without background. Top row: Vertical, diagonal, and offset line sources. Bottom row: Square area sources ($0.4 \times 0.4$ m) in the center of the field of view and offset to the side, and an L-shaped area. Each source has a total activity of 1 MBq uniformly distributed over its extent, with data acquired over 600 seconds. All reconstructions use the scatter corrected model with EMTV and the same TV prior.

view. The top row demonstrates reconstruction of linear source geometries with vertical, diagonal, and offset orientations. The bottom row shows area source reconstructions including two rectangular regions of different sizes and an L-shaped area. Similar to the reconstructions of point sources, the reconstruction truthfully recovers the position of sources that lie central in the field of view. For lines and areas offset to the side, scattering and background effects in the data skew the spectra, resulting in reconstructions that overestimate the angle of sources from the center of the camera. The TV prior effectively promotes the assumed uniform activity distribution.

**Lab measurements.** Moving from simulated measurement data to real measurements with the camera RSL7, we employ the scatter-corrected model to reconstruct measurements of Caesium-137 point sources taken in a lab setting. A 612 kBq Caesium-137

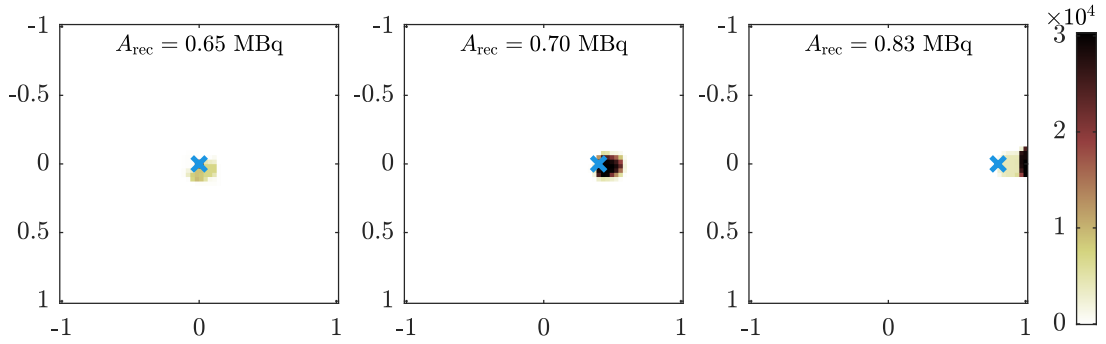



Figure 5.7.: MAP estimates of three Caesium-137 point source (0.61 MBq) lab measurements. The true positions of the sources are marked in blue, they are shifted from the camera's central axis by 0°, 22° and 39°, respectively. Positions in the center of the camera's field of view can be reconstructed reliably. Towards the boundary of the image plane, the signal-to-noise ratio reduces and the spatial errors increase at the same angular resolution.

source was mounted in different $(x, y)$ positions on a wall 1 m from the camera. We show three MAP reconstructions in Fig. 5.7. The measurements show that there are discrepancies between real data and simulated data from the same point source location. Nevertheless, the location and source activity can be reliably reconstructed if the point source lies centrally in the field of view, with artifacts and activity error becoming larger for angles $> 30°$ from the camera's symmetry axis. The remaining model mismatch is likely due to more scattering effects with the camera's housing and stand for sources towards the boundary of the field of view. Additionally, the signal-to-noise ratio is higher in the center due to the smaller distance to the camera and thus higher number of counts. In total, 14 measurements were taken, the full set of data is evaluated and discussed in more detail in Appendix A.

**Leakage correction for a single nuclide.** The next addition to the forward model to test is the energy leakage model (5.4). In the simplest case, we reconstruct the activity of a single nuclide with multiple emission lines, like any of the nuclides apart from Caesium-137 in Table 2.1. For Cobalt-60, we already saw the leakage effect from its higher emission line at 1332 keV to the lower at 1173 keV in Fig. 5.1. The forward models for fully absorbed photon coincidences are $P^{1173}$ and $P^{1332}$ ($P^{1,1}$ and $P^{1,2}$ in the notation of (5.3)). Correcting this with a leakage term $Q_2^{1,1}$ gives the full model $P + Q$ defined in (5.4). Figure 5.8 compares MAP reconstructions of simulated Cobalt-60 point sources with and without leakage correction in the forward model. Cobalt-60 is a rather simple polyenergetic nuclide with only two emission lines, so the effect is rather small here and mostly confined to small artifacts on the image's boundary, which are successfully removed with the corrected model. The total activity also receives a slight correction from including $Q$ into the model.

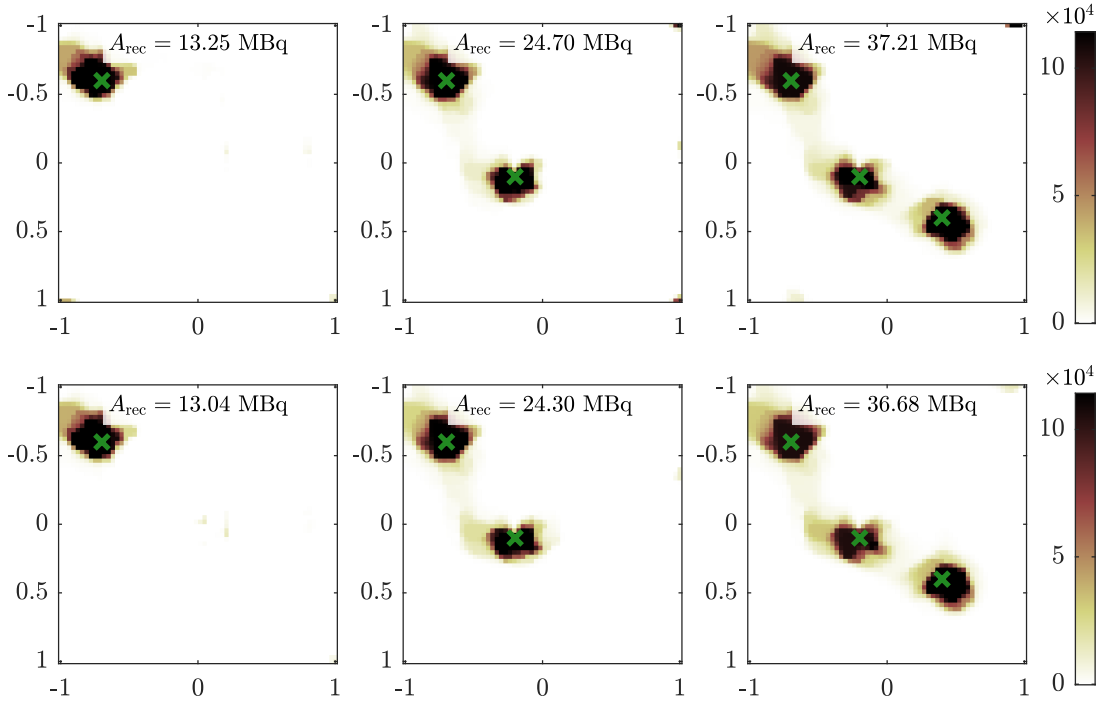



Figure 5.8.: Reconstructions of point sources from simulated Cobalt-60 data. From left to right, we simulated one, two and three point sources, 10 MBq each, positions in green. Top row: reconstructions without energy leakage correction. Bottom row: reconstructions with energy leakage correction.

We use the same model to compute MAP estimates from real measurements of Cobalt-60 sources with the RSL7 camera. In Fig. 5.9, we see reconstructions from a simplified line source at 2 m distance from the camera and small area sources at 4 m distance, all taken in a lab environment. All reconstructions truthfully recover the position of the test source up to small shifts of 10–20 cm.

**Leakage correction for polyenergetic reconstruction.** Beyond single nuclides, realistic decommissioning scenarios typically involve multiple nuclides present simultaneously in the environment. Results on simultaneous reconstruction at various energy levels have been presented in [148]. In contrast, we group emission lines by nuclide, since this reduces the image space dimension, and because the activity by nuclide is ultimately the quantity of interest in decommissioning assuming the nuclide library is exhaustive. We solve for $f = (f_1, \dots, f_M)$, a list of emission images of $M$ different nuclides. The damped EM-TV algorithm has to be modified to correctly regularize the images, the new regularization term is $J(f) = \sum_m \lambda_m \operatorname{TV}(f_m)$. The scalar parameters $\lambda_m$ steer the regularization nuclide-wise based on prior knowledge about individual nuclides. The TV

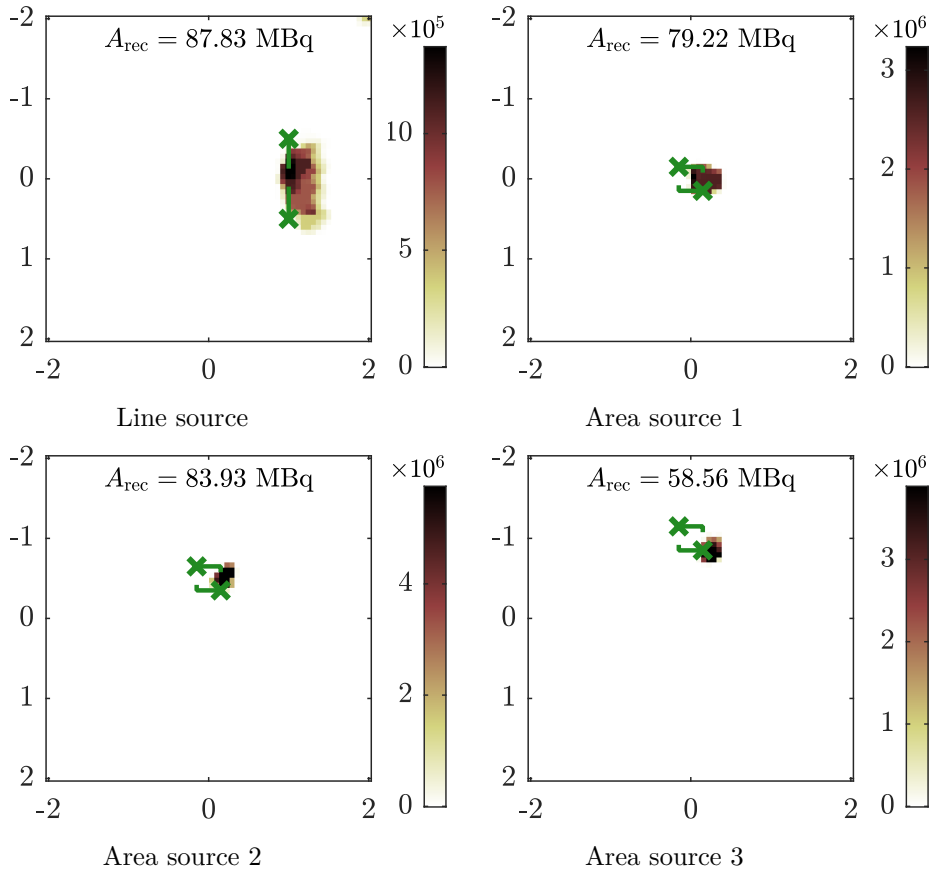



Figure 5.9.: Reconstructions of different types of sources from measured Cobalt-60 data in a lab environment. The top left source is a 1 m line source at distance 2 m from the camera, created by placing 5 point sources (17.6 MBq each, so total 88 MBq) at 25 cm distances in a straight line. The top right and bottom row sources are small areas of contamination, created by forming a $0.3 \times 0.3$ m square of sources, the center shifted by 0, 0.5 and 1 m from the camera's central axis, respectively.

denoising step (5.10) becomes

$$f_m^{(n+1)} = \arg\min_{f_m} \left\{ \frac{1}{2} \| f_m - (\omega f_m^{(n+\frac{1}{2})} + (1-\omega) f_m^{(n)}) \|^2_{w^{(n)}} + \omega\lambda \, \mathrm{TV}(f_m) \right\}, \tag{5.11}$$

where $f_m^{(n+\frac{1}{2})}$ is the $m$-th part of $f^{(n+\frac{1}{2})}$ after the regular first half step.

To test the polyenergetic forward model from (5.4), we reconstruct from combined measurement data by simulating sources and mixing the spectra together. In a first experiment, we consider a Caesium-137 (662 keV) line source and a Cobalt-60 (1173 keV and 1332 keV) point source which overlap spatially. The polyenergetic model reconstructs separate activity images for each nuclide, accounting for energy leakage from higher to lower emission lines across the full spectrum. Figure 5.10 shows that the model successfully separates and localizes both nuclide types from the mixed data, demonstrating the capability to handle realistic multi-nuclide scenarios essential for practical decommissioning applications.

The limits of this type of reconstruction become visible in the second row of Fig. 5.10. When emission spectra of different nuclides influence each other too strongly, the reconstruction becomes challenging. A mixture of Europium-152 and Cobalt-60 serves as an example of this difficulty, since Europium-152 has dominant peaks around 1086 keV, 1112 keV and 1408 keV, which all lie close to the Cobalt-60 peaks at 1173 keV and 1332 keV. These neighboring emission lines interact significantly through incomplete absorption, especially in the case of poor detector calibration where the energy resolution is insufficient to reliably separate the peaks. When multiple nuclides with overlapping spectral features are present, the leakage correction model must account for cross-talk between all emission lines, which can amplify uncertainties and make source separation more difficult.

**Bidirectional setups.** In the course of determining a viable camera setup for nuclear decommissioning, Caesium-137 point source data measurements were conducted using the RSL2 camera, which consists of two 2 inch $CeBr_3$ detectors, see Fig. 5.11.

The idea behind using a bidirectional setup is to allow for more flexibility in the camera design. The PVT scintillation detectors in RSL7 have the advantage of mostly scattering, as their photoelectric absorption cross section is very small due to the low average atomic number $Z$. At the same time, this however means their energy resolution is limited, resulting in a higher constant $\eta_A$ steering the energy uncertainty in (3.12) [113]. By using multiple $CeBr_3$ detectors in a single camera, the energy resolution could be improved, while potentially also increasing the total number of counts. This comes at the cost of additional uncertainty since we cannot track the order of interactions, so we have to use the bidirectional model Section 3.4. As described in (3.15), the forward model accounts for coincidences in both possible interaction sequences, with the probability of each

direction determined by the energy-dependent scattering cross sections and geometric factors.

The bidirectional model was implemented using the same quadrature approximation approach as the other models in this experimental section, with numerical integration over detector volumes and convolution with the energy resolution kernel (3.12). Due to the minimal spatial information provided by only two detectors, the prototype RSL2 camera permits only one-dimensional reconstruction of the source distribution. The reconstruction results are therefore presented as angular profiles from $-90°$ to $90°$, rather than the two-dimensional images shown in previous experiments.

A total of 10 Caesium-137 measurements were acquired on a semicircle around the centrally positioned camera. The first measurement placed the source at the center of the field of view, such that both scatter angles in the bidirectional model are identical. Subsequently, the source was moved in $-10°$ increments down to $-90°$, taking a 20 minute measurement at each position, resulting in 10 measurements spanning the camera's field of view. An additional 20 minute natural background measurement without the Caesium-137 source was taken for background correction in the EM-TV reconstruction algorithm. The resulting sum spectrum (sum of both detector values for each measured coincidence) of the background is shown in Fig. 5.11.

Reconstruction results for $^{137}$Cs point sources at $-10$, $-40$ and $-90°$ are displayed in Fig. 5.12, showing fitted data alongside the corresponding angular reconstructions. The results for all 10 angles can be found in the appendix. The reconstructions can determine the angle of the incoming radiation, even when there is a major contribution of natural background in each spectrum. The reconstructions tend to have high variance due to the limited spatial resolution. For sources at the border of the field of view, reconstruction becomes less accurate. The reason for this is likely shadowing between the two detectors: The analytical forward model so far only takes into account a generic damping factor when one detector blocks the view of a source from another detector. In reality, the scattering and absorption scene will be much more complex due to interactions between gamma rays and the camera housing and electronics. Such a correction could be measured and integrated into the forward model or avoided entirely by using measured or simulated forward models.

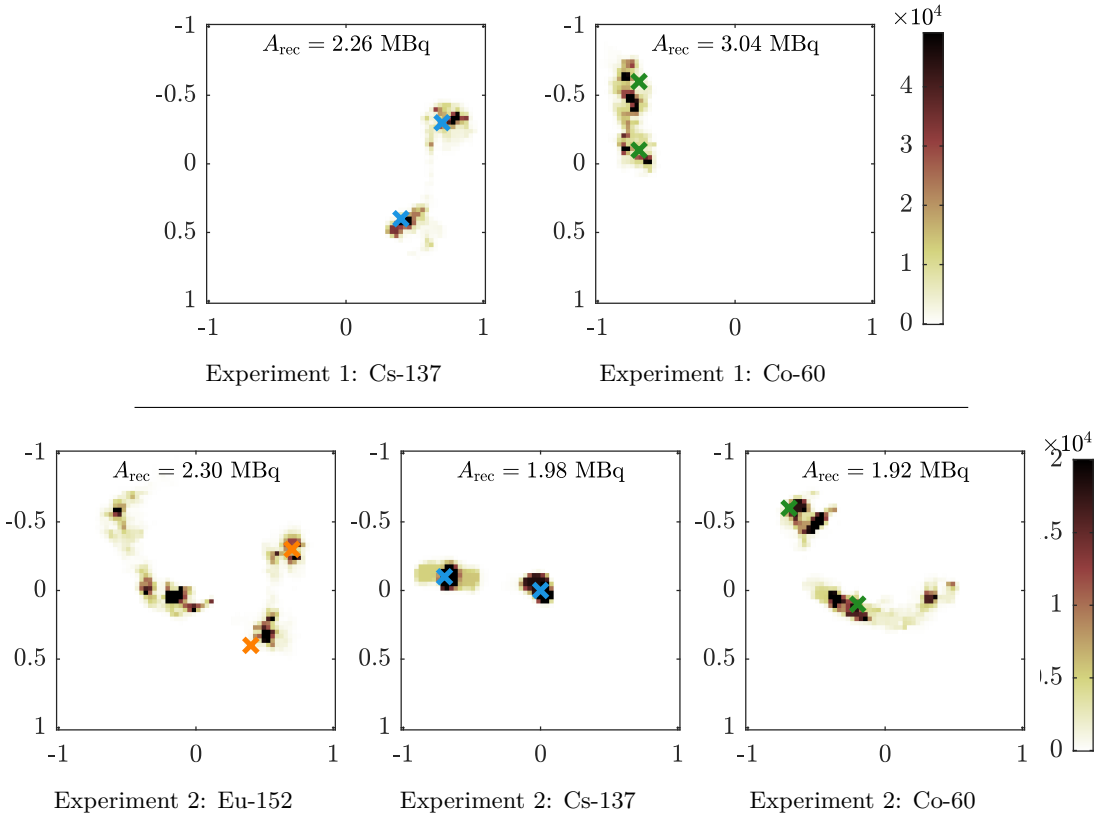
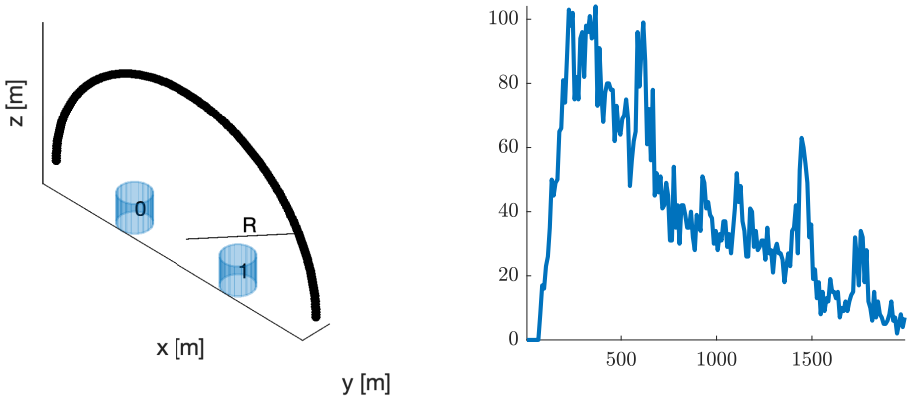



Figure 5.10.: Simultaneous MAP reconstructions of multiple nuclides in the same image plane, simulated by mixing data from different nuclides and background. Top row: A $^{137}$Cs line source (662 keV) together with a $^{60}$Co point source (1173 keV and 1332 keV). The polyenergetic forward model (5.4) successfully separates both nuclides, since the two nuclides are well-separated in the data space. Bottom row: Three-nuclide reconstruction with $^{152}$Eu (8 emission lines from 122–1408 keV), $^{137}$Cs (662 keV), and $^{60}$Co (1173 keV and 1332 keV) sources. The spectral overlap between $^{152}$Eu and $^{60}$Co peaks makes accurate reconstruction very challenging.




(a) One-dimensional RSL2 setup

(b) Background spectrum

Figure 5.11.: A measurement setup with RSL2. Left: The simple setup with only two detectors has a symmetry plane through the middle, so we can only reconstruct a one-dimensional activity distribution, here indicated by the black line. $R$ is chosen as 2 m in our experiment, while the detectors are 2 inches in height and diameter (figure not to scale). Right: A 20 minute natural background measurement (used as correction in the reconstruction) measured with RSL2 in a lab setting. Some natural background peaks are visible, e.g., at 1460 keV from $^{40}$K or at 609 keV, 1120 keV and 1760 keV from $^{214}$Bi.

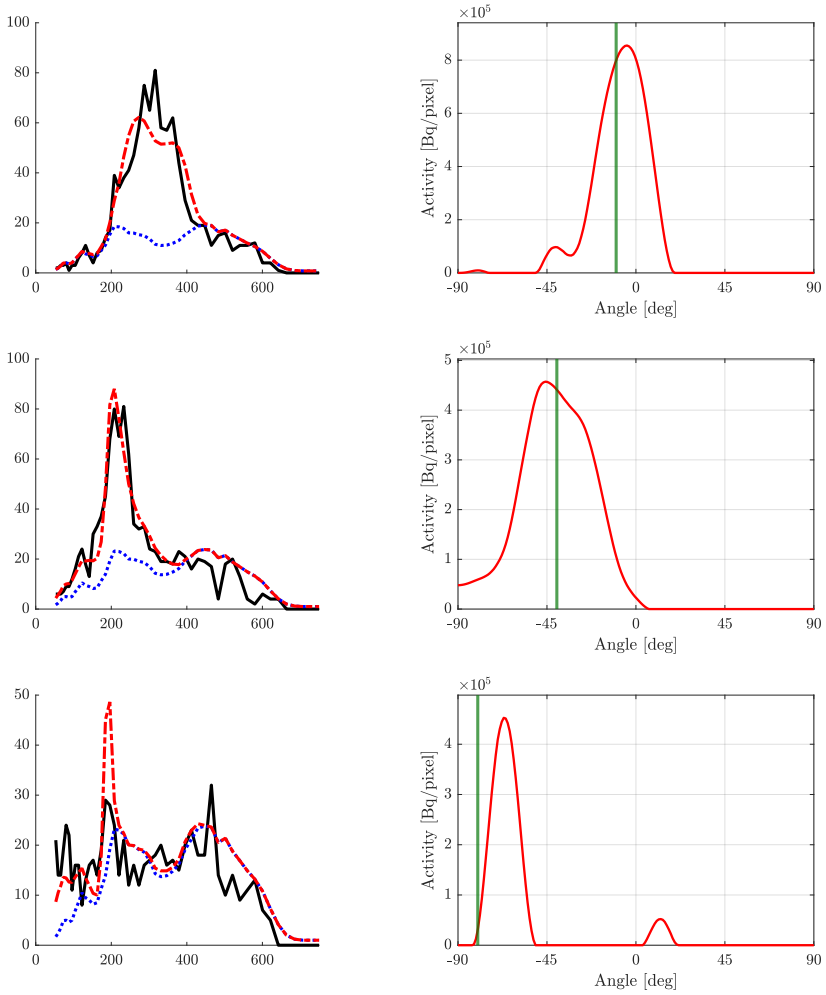



Figure 5.12.: Reconstructions of point sources from lab-measured data acquired with RSL2. From top to bottom, the ground truth positions of the source were located at $-10°$, $-40°$ and $-80°$, taking a 20 minute measurement each. The left shows the data $g$ (black solid, all measured energies in one detector for coincidences around the $^{137}$Cs 662 keV peak), the background spectrum (blue, dotted) and the fitted data $Af^{\text{rec}}$ (red, dashed). The right shows the reconstruction $f^{\text{rec}}$ (red) and the ground truth position of the source (vertical green).

# Chapter 6

# Sampling as an Optimization Problem

We now transition to the algorithmical problem introduced in Section 4.5: Given a posterior or general target distribution $\mu^*$, how do we efficiently draw samples from it in order to solve inference tasks? Upon choosing a suitable state space and loss functional, the problem of sampling from a given distribution can be reformulated as an optimization problem. This approach has inspired important theoretical insights, in particular to sampling algorithms based on Langevin diffusion [216, 73]. We will discuss this connection here and outline how it helps analyzing the algorithms.

Given some target $\mu^*$, the task is to draw i.i.d. samples $X_1, X_2, \dots$ from $\mu^*$. Generally, it is difficult to generate samples from the density function of $\mu^*$ alone—particularly so since the density is high-dimensional and thus only known up to its normalization constant (4.2). Hence, sampling methods are focused on finding a way to sample from a tractable proxy. Suppose an algorithmic approach produces $X_i \sim \hat{\mu}$, where $\hat{\mu}$ allows for easy sampling. Then the challenge lies in guaranteeing that $\hat{\mu}$ is the target, or at least $\hat{\mu} \approx \mu^*$ in a suitable metric. Constructing $\hat{\mu}$ can thus be formulated as the optimization problem

$$\hat{\mu} = \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathcal{D}(\mu \,\|\, \mu^*), \tag{6.1}$$

where $\mathcal{P}(\mathbb{R}^d)$ is the set of probability measures over the state space $\mathbb{R}^d$, and $\mathcal{D}(\cdot \,\|\, \mu^*)$ is an objective functional attaining its global minimum at $\mu = \mu^*$. In practice, instead of solving (6.1) over the entire space $\mathcal{P}(\mathbb{R}^d)$, opting for a specific inference method may restrict the solution to a subclass $\mathcal{M} \subset \mathcal{P}(\mathbb{R}^d)$, e.g. a parametric family of distributions as in variational inference [30]. An important subset of $\mathcal{P}(\mathbb{R}^d)$ will be the set of absolutely continuous measures $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d) \coloneqq \{\mu \in \mathcal{P}(\mathbb{R}^d) \,:\, \mu \ll \mathrm{Leb}\}$, where Leb denotes the Lebesgue measure on $\mathbb{R}^d$. With a slight abuse of notation, we will use $\mu$ for the measure

and the associated density: For Borel sets $\Omega \in \mathcal{B}(\mathbb{R}^d)$, we write $\mu(\Omega)$ and for $x \in \mathbb{R}^d$, we denote $\mu(x)$ for the (Lebesgue) density $\frac{\mathrm{d}\mu}{\mathrm{d}\,\mathrm{Leb}}(x)$ of measures $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$.

The remaining chapter is organized as follows: In order to analyze the optimization problem (6.1), we first suitably constrain the state space and equip it with the Wasserstein metric in Section 6.1. In Section 6.2, we discuss various choices of objective functionals, which lead to different optimization dynamics. The kind of optimization procedures that we are interested in in this work can be interpreted as discretizations of gradient flows in Wasserstein space, so we properly define the latter in Section 6.3. The gradient flows manifest as parabolic partial differential equations (PDEs)—continuity equations which are solved by the density function. For a particular choice of objective, the Kullback–Leibler divergence, we derive the concrete PDE and review typical assumptions that lead to convergence in Section 6.4. Intimately connected to the PDE is the particle implementation via a diffusion stochastic differential equation (SDE), which then leads to implementable sampling algorithms. We therefore cover some basics on diffusion SDEs in Section 6.5. We end the chapter by showing in Section 6.6 that the Kullback–Leibler divergence objective leads to overdamped Langevin diffusion as the corresponding SDE.

## 6.1. Metric Structure of the State Space

In order to provide the optimization problem (6.1) with structure, the state space has to be equipped with a metric. A well-understood and useful choice are Wasserstein spaces, motivated by optimal transport.

**Definition 6.1: Wasserstein Space**

The $p$-Wasserstein distance with $p \in [1, \infty)$ is defined by

$$\mathcal{W}_p(\mu, \tilde{\mu}) \coloneqq \inf_{\pi \in \Gamma(\mu, \tilde{\mu})} \left( \mathbb{E}_{(x,\tilde{x})\sim\pi}[\|x - \tilde{x}\|^p] \right)^{1/p}, \tag{6.2}$$

where $\Gamma(\mu, \tilde{\mu})$ is the set of all couplings of $\mu$ and $\tilde{\mu}$, i.e.,

$$\Gamma(\mu, \tilde{\mu}) = \left\{ \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \;:\; \mu = \int_{\mathbb{R}^d} \pi(\cdot, \tilde{x})\, \mathrm{d}\tilde{x} \text{ and } \tilde{\mu} = \int_{\mathbb{R}^d} \pi(x, \cdot)\, \mathrm{d}x \right\}. \tag{6.3}$$

If we define, for $p \in [1, \infty)$, the set of distributions with finite $p$-th moment by

$$\mathcal{P}_p(\mathbb{R}^d) \coloneqq \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \;:\; \int_{\mathbb{R}^d} \|x\|^p \, \mu(\mathrm{d}x) \right\} \subset \mathcal{P}(\mathbb{R}^d), \tag{6.4}$$

then $(\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p)$ becomes a metric space, called the $p$-Wasserstein space. The subset of measures with a Lebesgue density is $\mathcal{P}_{p,\mathrm{ac}}(\mathbb{R}^d) = \mathcal{P}_p(\mathbb{R}^d) \cap \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$.

We are particularly interested in $p = 2$ due to the rich structure[1] of the space $\mathcal{P}_2$ and the connections to continuum mechanics and the diffusion equations from which our sampling methods later arise. Further, we point out here that Wasserstein distances have their origin in optimal transport, where they correspond to the total cost of transporting mass between two measures. The (Kantorovich) optimal transport problem with the standard $\ell^p$-norm cost is to find the optimal transport plan (instantiated by a coupling $\pi$) that attains the infimum in (6.2). Since $\ell^p$-norms are lower semicontinuous, it is easy to show the existence of an optimal transport plan for (6.2). We will need the following classical result by Brenier [37], which ensures uniqueness and shows that the optimal plan has a particularly simple structure, induced by an optimal transport map.

**Theorem 6.2: Brenier's Theorem**

Let $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, $\tilde{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$. Then the optimal transport plan is unique. It is induced by an optimal transport map, denoted $T_{\mu\to\tilde{\mu}}$, such that the coupling $\pi = (\mathrm{Id}, T_{\mu\to\tilde{\mu}})_{\#}\mu \in \Gamma(\mu, \tilde{\mu})$ attains the infimum in (6.2) as a minimum. The map is further characterized by being the gradient $T_{\mu\to\tilde{\mu}} = \nabla\varphi$ of some convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$, which is called the Kantorovich potential.

For details and applications of optimal transport we refer to the books [211, 187].

## 6.2. Choosing the Sampling Objective

There are various possible choices for $\mathcal{D}$ in (6.1) that lead to practical algorithms for inference. A standard option (and the one we focus on most in this work) is the Kullback–Leibler (KL) divergence (or relative entropy) of $\mu$ with respect to the target $\mu^*$. For any $\mu \in \mathcal{P}(\mathbb{R}^d)$ which is absolutely continuous w.r.t. $\mu^*$, denoted $\mu \ll \mu^*$, it is given by

$$\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) \coloneqq \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\mu^*}(x)\right) \mu(\mathrm{d}x). \tag{6.5}$$

Here $\frac{\mathrm{d}\mu}{\mathrm{d}\mu^*}$ denotes the Radon–Nikodym derivative of $\mu$ w.r.t. $\mu^*$. If $\mu, \mu^* \in \mathcal{P}_{ac}(\mathbb{R}^d)$, we use the equivalent form

$$\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) = \int_{\mathbb{R}^d} \log\left(\frac{\mu(x)}{\mu^*(x)}\right) \mu(x)\,\mathrm{d}x. \tag{6.6}$$

The definition can be extended to all of $\mathcal{P}(\mathbb{R}^d)$ by setting $\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) = \infty$ when $\mu \not\ll \mu^*$. The KL divergence satisfies $\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) \geq 0$ for all $\mu \in \mathcal{P}(\mathbb{R}^d)$ and attains its unique global minimum zero at $\mu = \mu^*$. Since the minimization of KL divergence leads to diffusion equations and thus Langevin sampling algorithms, we will comment on it in more detail below. The KL divergence is sometimes (mis-)interpreted as a distance

[1] This will be justified when we examine the Riemannian structure of $\mathcal{P}_2$ in Section 6.3

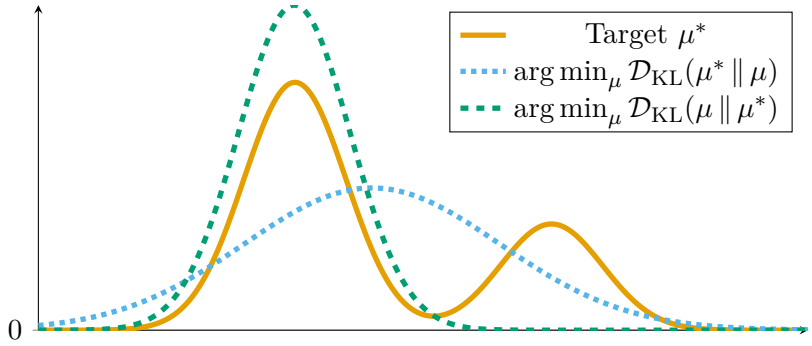



Figure 6.1.: The difference between fitting a Gaussian $\mu = \mathrm{N}(m, \sigma^2)$ with the forward KL divergence $\mathcal{D}_{\mathrm{KL}}(\mu^* \,\|\, \mu)$ and the reverse KL divergence $\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*)$. Fitting a Gaussian with the forward KL reduces to computing maximum likelihood estimates for $(m, \sigma)$ under the target $\mu^*$ and results in a distribution with mode-covering, large variance. The reverse KL divergence tends to concentrate its mass in one of the modes of the multi-modal target $\mu^*$ (here chosen as a Gaussian mixture with two modes). While the difference between the estimates in this case is dramatic, the imaging posteriors modeled in this work are usually log-concave, making the two versions more similar.

between $\mu$ and $\mu^*$, despite not being a metric since it satisfies neither symmetry nor the triangle inequality. Indeed, one can also minimize $\mathcal{D}_{\mathrm{KL}}(\mu^* \,\|\, \mu)$ instead, leading to different algorithms. In variational inference, when $\mu^*$ is potentially complex and is approximated by some simple uni-modal $\mu$ (e.g. a Gaussian), the two variants are referred to as mode-seeking (for $\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*)$, also called reverse KL) and mode-covering (for $\mathcal{D}_{\mathrm{KL}}(\mu^* \,\|\, \mu)$, also forward KL) fitting [30], see Fig. 6.1.

There are further options for choosing $\mathcal{D}$ in (6.1). For instance, it was shown in [59] that Stein variational gradient descent, a sampling method due to [134], can be interpreted as a gradient flow of the chi-squared divergence in 2-Wasserstein space. The chi-squared divergence is defined by

$$\mathcal{D}_{\chi^2}(\mu \,\|\, \mu^*) := \int_{\mathbb{R}^d} \left( \frac{\mathrm{d}\mu}{\mathrm{d}\mu^*}(x) - 1 \right)^2 \mu^*(\mathrm{d}x) = \mathrm{Var}_{\mu^*}\left[ \frac{\mathrm{d}\mu}{\mathrm{d}\mu^*} \right], \tag{6.7}$$

if $\mu \ll \mu^*$ and $\mathcal{D}_{\chi^2}(\mu \,\|\, \mu^*) = \infty$ otherwise. Note that the second equality followed from $\mathbb{E}_{\mu^*}[\frac{\mathrm{d}\mu}{\mathrm{d}\mu^*}] = 1$.

The chi-squared divergence, KL divergence (both forward and reverse) and multiple other relevant functionals of distributions are all instances of $f$-divergences, defined by

$$\mathcal{D}_f(\mu \,\|\, \mu^*) = \int f\left( \frac{\mathrm{d}\mu}{\mathrm{d}\mu^*} \right) \mu^*(\mathrm{d}x). \tag{6.8}$$

If $f$ is convex and satisfies $f(1) = 0$, Jensen's inequality implies that $\mathcal{D}_f(\mu \,\|\, \mu^*) \geq 0$. When it comes to devising a sampling algorithm for high-dimensional Bayesian posteriors, the KL divergence (6.6) has a favourable property among this class, as was shown in [54].

**Theorem 6.3: KL divergence for scaled targets**

Assume $f : (0, \infty) \to \mathbb{R}$ is continuously differentiable with $f(1) = 0$. The KL divergence, corresponding to the choice $f(x) = x \log x$, is the only $f$-divergence for which the quantity $\mathcal{D}_f(\mu \,\|\, \mu^*) - \mathcal{D}_f(\mu \,\|\, c\mu^*)$ is independent of $\mu$ for any $c \in (0, \infty)$ and any $\mu^* \in \mathcal{P}(\mathbb{R}^d)$.

This directly implies that the first variation (in $\mu$) of the KL divergence is independent of constant scalings of $\mu^*$. This property makes the KL divergence a logical choice when constructing first-order algorithms for sampling from $\mu^*$ with unknown normalization constant $Z$ as in (4.2).

## 6.3. Wasserstein Gradient Flows

In order to understand how sampling connects to problems of the form (6.1), we have to refer to several connected works from PDE and optimal transport research.

The main insight is due to Jordan, Kinderlehrer and Otto in the late twentieth century [104]. The authors were interested in analyzing the Fokker–Planck equation, i.e. the mean-field limit of certain diffusion processes. They recognized that the evolution of a density satisfying this Fokker–Planck PDE can be understood as following the gradient flow of the KL divergence in the space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$. In a subsequent work by Otto, this idea was generalized to the porous medium equation and the Riemannian geometry underlying the gradient flow was analyzed further [160]. Since these PDEs are used to model a variety of diffusion processes in physical and biological systems, the works sparked renewed interest in gradient flows and the Riemannian geometry of Wasserstein spaces. A comprehensive treatment of gradient flows in general metric spaces, with a particular focus on Wasserstein spaces, can be found in [7]. We will also employ notation and ideas from [188], but cover only the basic facts necessary to understand the convergence behaviour of the KL divergence and its connections to Langevin Monte Carlo sampling.

We consider the following general problem

$$\operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu), \tag{6.9}$$

with a penalty $\mathcal{F}$, e.g., the KL divergence (6.6). When discussing ways to compute MAP estimates (4.3) by minimizing a convex function $F(x)$ in Euclidean space, we already

encountered gradient flows of the form

$$\dot{x}_t = -\nabla V(x_t).$$

Transferring this to (6.9), one can indeed also simulate a gradient flow, formally

$$\text{“}\dot{\mu}_t = -\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)\text{”}. \tag{6.10}$$

Herein, $\nabla_{\mathcal{W}_2}$ is the gradient in the metric $\mathcal{W}_2$ in $\mathcal{P}_2(\mathbb{R}^d)$. In order to understand these flows, we need to make precise both how curves of measures can be characterized by their time-dependent dynamics, and the Wasserstein gradient of the functional $\mathcal{F}$, as well as transfer the concept of convexity to the new geometry.

The following result [7, chap. 8] provides the concept of continuity equations for curves $\mu_t$ in Wasserstein space.

**Theorem 6.4: Continuity equation for curves in $\mathcal{P}_2$**

Let $(\mu_t)_{t\in(0,T)}$ be an absolutely continuous curve in $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$. Then there exists a time-dependent Borel-measurable velocity field $v_t : \mathbb{R}^d \to \mathbb{R}^d$ with $\|v_t\|_{L^2(\mu_t;\mathbb{R}^d)} \leq |\mu'|(t)$ for almost every $t \in (0,T)$ so that $\mu_t$ solves (in a weak sense) the continuity equation

$$\frac{\partial}{\partial t}\mu_t + \nabla \cdot (v_t\mu_t) = 0 \quad \text{in} \quad \mathbb{R}^d \times (0,T). \tag{6.11}$$

Conversely, if an absolutely continuous $(\mu_t)_{t\in(0,T)}$ solves (6.11) for some velocity field $v_t$ with $\|v_t\|_{L^2(\mu_t;\mathbb{R}^d)} \in L^1((0,T))$, then $|\mu'|(t) \leq \|v_t\|_{L^2(\mu_t;\mathbb{R}^d)}$, where $|\mu'|$ denotes the metric derivative of $\mu$.

This shows that any such curve $(\mu_t)_{t\in(0,T)}$ is entirely characterized by the corresponding velocity fields $(v_t)_{t\in(0,T)}$, with $v_t \in L^2(\mu_t;\mathbb{R}^d)$. For any $t$, the field $v_t$ can be interpreted as a tangent vector to $\mathcal{P}_2(\mathbb{R}^d)$ at $\mu_t$. Note that the same continuity equation also arises in optimal transport, since $v_t$ gives rise to the Benamou–Brenier formula [24] for the Wasserstein distance:

**Theorem 6.5: Wasserstein distance, Benamou–Brenier formula**

Let $\mu, \tilde{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$. Then it holds

$$\mathcal{W}_2^2(\mu,\tilde{\mu}) = \min_{\mu_t, v_t} \left\{ \int_0^1 \|v_t\|^2_{L^2(\mu_t,\mathbb{R}^d)}\,\mathrm{d}t \;:\; \frac{\partial}{\partial t}\mu_t + \nabla\cdot(v_t\mu_t) = 0 \right\}, \tag{6.12}$$

where the minimum is taken over all curves $(\mu_t)_{t\in[0,1]}$ with $\mu_0 = \mu$ and $\mu_1 = \tilde{\mu}$ and corresponding Borel velocity fields. The unique curve $(\mu_t)_{t\in[0,1]}$ attaining the

minimum is the Wasserstein geodesic joining $\mu$ and $\tilde{\mu}$, it can also be given as

$$\mu_t = (t\mathrm{Id} + (1-t)T_{\mu\to\tilde{\mu}})_{\#}\mu, \quad t \in [0,1],$$

where $T_{\mu\to\tilde{\mu}}$ is the optimal transport map from Theorem 6.2.

The transport along Wasserstein geodesics also plays a role when defining Wasserstein gradients of functionals in $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$. Perturbations to $\mu$ are understood as acting along geodesics and hence as elements in the tangent space to $\mathcal{P}_2$ at $\mu$. Due to Theorem 6.4, the tangent space can be identified with the quotient space of $L^2(\mu;\mathbb{R}^d)$ by all divergence-free vector fields, which forms a Hilbert space when equipped with the natural inner product $\langle\cdot,\cdot\rangle_{L^2(\mu;\mathbb{R}^d)}$. This structure allows to define the Wasserstein gradient via a Taylor expansion: $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu)$ is the vector field satisfying

$$\mathcal{F}((\mathrm{Id}+\varepsilon\sigma)_{\#}\mu) = \mathcal{F}(\mu) + \varepsilon\langle\nabla_{\mathcal{W}_2}\mathcal{F}(\mu),\sigma\rangle_{L^2(\mu;\mathbb{R}^d)} + o(\varepsilon) \tag{6.13}$$

for all displacement vector fields $\sigma$ and $\varepsilon \to 0$.

For many relevant functionals, we can characterize this vector field by computing the first variation $\frac{\delta\mathcal{F}}{\delta\mu} : \mathbb{R}^d \to \mathbb{R}$ of $\mathcal{F}$ at $\mu$. The first variation can be understood[2] as a local potential describing the change of $\mathcal{F}$ when perturbing $\mu$ and is defined as the unique (up to additive constants) functional for which

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{F}(\mu+\varepsilon\chi)\big|_{\varepsilon=0} = \int_{\mathbb{R}^d}\chi(x)\frac{\delta\mathcal{F}}{\delta\mu}(x)\,\mathrm{d}x, \tag{6.14}$$

for any perturbation $\chi$ with $\int\chi(x)\,\mathrm{d}x = 0$ such that $\mathcal{F}$ is finite at $\mu+\varepsilon\chi \in \mathcal{P}_2(\mathbb{R}^d)$ at least for $\varepsilon \in [0,\varepsilon_0]$. The Wasserstein gradient is then the unique velocity field $v$ (up to additive, divergence-free fields) satisfying

$$-\int_{\mathbb{R}^d}\frac{\delta\mathcal{F}}{\delta\mu}(x)\nabla\cdot(\mu\sigma)(x)\,\mathrm{d}x = \langle v,\sigma\rangle_{L^2(\mu;\mathbb{R}^d)}, \tag{6.15}$$

which due to integration by parts implies $\nabla_{\mathcal{W}_2}\mathcal{F}(\mu) = \nabla\frac{\delta\mathcal{F}}{\delta\mu}$. Note that the gradient on $\frac{\delta\mathcal{F}}{\delta\mu}$ eliminates additive constants in the first variation.

The above results allow us to rigorously define the Wasserstein gradient flow (WGF) from (6.10). The curve $\mu_t$ is understood to follow the gradient flow if it satisfies the continuity equation

$$\frac{\partial}{\partial t}\mu_t - \nabla\cdot(\mu_t\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)) = 0. \tag{6.16}$$

[2] We are (deliberately) not rigorous here. Formally, the first variation is an element of the cotangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at $\mu$, which acts on the velocity fields in the tangent space. Due to the Riesz representation theorem, the cotangent space can be identified with the tangent space, allowing us to identify the first variation with a functional on $\mathbb{R}^d$. For details, see [7, chap. 8–10].

Using the first variation, we can give an explicit form of the Wasserstein gradient for some classes of important energy functionals. In our analysis of sampling algorithms, we will be particularly interested in potential energies and a a specific internal energy functional. For potential energies, the Wasserstein gradient can be computed as follows.

**Example 6.6: Wasserstein gradient of potential energies**

Let $V : \mathbb{R}^d \to (-\infty, \infty]$ be proper, lower semicontinuous with $V(x) \geq -\alpha - \beta\|x\|^2$ for some $\alpha, \beta \geq 0$. The potential energy of $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with respect to $V$ is

$$\mathcal{E}_V(\mu) \coloneqq \int_{\mathbb{R}^d} V(x)\mu(\mathrm{d}x). \tag{6.17}$$

The functional $\mathcal{E}_V$ is proper and lower semicontinuous. By linearity of the integral, (6.14) directly gives the first variation at any measure $\mu$ as $\frac{\delta \mathcal{E}_V}{\delta \mu} = V$. Hence, the Wasserstein gradient is given by $\nabla_{\mathcal{W}_2}\mathcal{E}_V(\mu) = \nabla V$ for all $\mu$.
Note that we recover standard gradient flows in Euclidean space as a special case: The WGF for $\mathcal{E}_V$ is the continuity equation

$$\frac{\partial}{\partial t}\mu_t(x) = \nabla \cdot (\mu_t(x)\nabla V(x)). \tag{6.18}$$

Interpreted in a weak sense with a concentrated initialization $\mu_0 = \delta_{x_0}$, this simple transport equation has the solution $\mu_t = \delta_{x_t}$, where $x_t$ satisfies the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}x_t = -\nabla V(x_t).$$

Another example of typical objectives in Wasserstein space are internal energies. The entropy of a distribution is a special case of such an internal energy.

**Example 6.7: Wasserstein gradient of the negative entropy**

Define an internal energy functional by

$$\mathcal{H}(\mu) = \begin{cases} \int_{\mathbb{R}^d} \mu(x)\log(\mu(x))\,\mathrm{d}x & \text{if } \mu \ll \text{Leb}, \\ \infty & \text{otherwise.} \end{cases} \tag{6.19}$$

Then $\mathcal{H}(\mu)$ is the negative entropy of $\mu$, and $\mathcal{H}$ is a proper and lower semicontinuous functional. Using that $\frac{\mathrm{d}}{\mathrm{d}x}(x\log x) = 1 + \log x$, we see that the first variation of $\mathcal{H}$ at $\mu$ is $\frac{\delta \mathcal{H}}{\delta \mu} = 1 + \log \mu$. The Wasserstein gradient is $\nabla_{\mathcal{W}_2}\mathcal{H}(\mu) = \nabla \log \mu$. Consequently, the gradient flow with respect to $\mathcal{H}$ is the heat equation $\partial_t \mu_t = \Delta \mu_t$.

In Euclidean gradient flows, an important property to prove convergence to the minimizer is (strong) convexity of the objective functional, see our discussions in Section 4.4. In

the Riemannian structure of $\mathcal{P}_2(\mathbb{R}^d)$, this translates to convexity along geodesics, first introduced in [147].

**Definition 6.8: Geodesic convexity**

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$ be proper and lower semicontinuous. For $\alpha \in \mathbb{R}$, we say that $\mathcal{F}$ is $\omega$-geodesically convex if for all $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, it holds

$$\mathcal{F}(\mu_t) \leq (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1) - \frac{\omega t(1-t)}{2}\mathcal{W}_2^2(\mu_0, \mu_1) \quad \forall t \in [0,1] \tag{6.20}$$

where $(\mu_t)_{t\in(0,1)}$ is the geodesic joining $\mu_0$ and $\mu_1$.

As for gradient flows in Euclidean space, we can exploit (strong) convexity to prove convergence rates of the WGF. A central tool to see this is the evolution variational inequality (EVI). For the WGF, it takes the following form [7, chap. 11]:

**Lemma 6.9: Evolution variational inequality**

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$ be $\omega$-strongly geodesically convex, $\omega \geq 0$. Assume that $(\mu_t)_{t\geq 0}$ solves the WGF (6.16) of $\mathcal{F}$. Then for every $\tilde{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ and every $t > 0$, it holds

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{W}_2^2(\mu_t, \tilde{\mu}) \leq \mathcal{F}(\tilde{\mu}) - \mathcal{F}(\mu_t) - \frac{\omega}{2}\mathcal{W}_2^2(\mu_t, \tilde{\mu}). \tag{6.21}$$

Note that the EVI actually characterizes the WGF entirely, i.e. requiring that $\mu_t$ satisfies EVI for all $t, \tilde{\mu}$ also ensures that it solves (6.16). Since $\nabla_{\mathcal{W}_2}\mathcal{F}$ does not appear in the EVI, this actually provides a different way to characterize (or define) the WGF. From the EVI, we can directly deduce uniqueness and a convergence rate: Since $\mathcal{F}$ is strongly convex, it has a unique minimizer $\mu^*$. Choosing two solutions $\mu_t, \nu_s$, $t, s \geq 0$ of the gradient flow, (6.21) holds for both, with $\tilde{\mu} = \nu_s$ and $\tilde{\mu} = \mu_t$ respectively. Summing up the two equations gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{W}_2^2(\mu_t, \nu_s) + \frac{\mathrm{d}}{\mathrm{d}s}\mathcal{W}_2^2(\mu_t, \nu_s) \leq -2\omega\mathcal{W}_2^2(\mu_t, \nu_s). \tag{6.22}$$

Along $t = s$, it directly follows $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{W}_2^2(\mu_t, \nu_t) \leq -2\omega\mathcal{W}_2^2(\mu_t, \nu_t)$. Grönwall's inequality then implies

$$\mathcal{W}_2(\mu_t, \nu_t) \leq e^{-\omega t}\mathcal{W}_2(\mu_0, \mu_0) \tag{6.23}$$

from which, if $\omega > 0$, we can deduce uniqueness of the gradient flow solution (by setting $\mu_0 = \nu_0$) as well as asymptotic stability and convergence to the stationary minimizer $\mu^*$ (by setting $\nu_0 = \mu^*$).

Since the requirement of strong geodesic convexity is quite restrictive, there is a growing amount of literature that aims at guaranteeing the convergence of the gradient flow under weaker conditions. A standard argument is the following: By the chain rule, (6.16), and integration by parts, it holds

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\mu_t) = \int \frac{\delta \mathcal{F}}{\delta \mu} \nabla \cdot (\mu_t \nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t))\, \mathrm{d}x = -\|\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)\|^2_{L^2(\mu_t;\mathbb{R}^d)} \leq 0 \tag{6.24}$$

This shows that $\mathcal{F}(\mu_t)$ decreases along $\mu_t$ at all times $t$. Let now $\mathcal{F}^* := \min_\mu \mathcal{F}(\mu)$. If $\mathcal{F}$ satisfies a functional inequality of the form

$$\|\nabla_{\mathcal{W}_2}\mathcal{F}(\mu)\|^2_{L^2(\mu;\mathbb{R}^d)} \geq 2\omega(\mathcal{F}(\mu) - \mathcal{F}^*) \quad \forall \mu \in \mathcal{P}_2(\mathbb{R}^d), \tag{6.25}$$

for some $\omega > 0$. This condition (6.25) appears similarly in Euclidean optimization, where it is typically called Polyak–Łojasiewicz inequality [172] and is implied by strong convexity. Together with (6.24) and Grönwall's Lemma, we obtain

$$\mathcal{F}(\mu_t) - \mathcal{F}^* \leq e^{-2\omega t}(\mathcal{F}(\mu_0) - \mathcal{F}^*). \tag{6.26}$$

This implies convergence in objective value, which is generally weaker than the ergodic convergence result (6.23) that followed from EVI.

At the end of this section, we want to point out that the perspective on WGFs presented here is certainly incomplete to a large extent—since our main focus lied on deriving the gradient flow as a continuity equation and on concrete ways to compute the Wasserstein gradient for functionals like internal and potential energies. Indeed, the main novel technique of the seminal work [104] was to derive the WGF as a minimizing movement scheme, i.e., as the continuous-time limit of a sequence of iterated minimization problems. The curve $\mu_t$ in (6.16) is derived as the limit of $(\mu_\tau^{(k)})_{k\geq 0}$ defined by

$$\mu_\tau^{(k+1)} \in \arg\min_\mu \left\{ \mathcal{F}(\mu) + \frac{1}{2\tau}\mathcal{W}_2^2(\mu, \mu_\tau^{(k)}) \right\}, \tag{6.27}$$

where $\tau > 0$ is a time step size. The iteration (6.27) is typically called JKO scheme after the authors of [104] and we refer to [7, chap. 11] and [104, 188] for details. For objective functionals like the reverse KL divergence, employing the iteration (6.27) numerically may appear tempting, but is generally intractable due to the need to compute the $\mathcal{W}_2$-proximal operator of $\mathcal{H}$. However, the scheme is the basis for a line of more numerically focused works that aim to solve (6.16) by approximating the updates of (6.27), e.g. [171, 186, 127, 55].

## 6.4. The Gradient Flow of the Kullback-Leibler Divergence

We now focus on the problem of minimizing (6.6) (and will refer to it simply as KL divergence), setting $\mathcal{F} = \mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \mu^*)$ in the discussions on gradient flows from the pre-

vious section. From now on, we assume that $\mu^* \ll \mathrm{Leb}$ is a fixed target with Lebesgue density $\mu^*(x) = \exp(-V(x))/Z$ for $V : \mathbb{R}^d \to \mathbb{R}$ which is lower semicontinuous and convex (equivalently, we say $\mu^*$ is log-concave). For any $\mu \ll \mu^*$, we can decompose the KL divergence as

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) &= \int_{\mathbb{R}^d} \mu(x) \log\left(\frac{\mu(x)}{\mu^*(x)}\right) \mathrm{d}x && (6.28a)\\ &= \int_{\mathbb{R}^d} \mu(x) \log(\mu(x))\, \mathrm{d}x + \int_{\mathbb{R}^d} V(x)\mu(x)\, \mathrm{d}x && (6.28b)\\ &= \mathcal{H}(\mu) + \mathcal{E}_V(\mu), && (6.28c)\end{aligned}$$

i.e., it is the sum of an internal energy (negative entropy) and a potential energy. The Wasserstein gradient of the KL divergence is therefore

$$\nabla_{\mathcal{W}_2} \mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) = \nabla \log \mu + \nabla V = \nabla \log \frac{\mu}{\mu^*} \tag{6.29}$$

and the WGF of $\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \mu^*)$ is characterized by the continuity equation (6.16)

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla \log \frac{\mu_t}{\mu^*}\right) = -\nabla \cdot (\mu_t \nabla \log \mu^*) + \Delta \mu_t. \tag{6.30}$$

The fact that we are writing $\frac{\mu_t}{\mu^*}$ for the Radon-Nikodym derivative will actually be justified later, where we show that $\mu_t$ has a smooth Lebesgue density for $t > 0$, so that $\mu_t \ll \mu^*$ holds for $t > 0$. Given the WGF for the KL divergence, we are interested in its convergence properties and thus in convexity. It can easily be checked that the convexity of $V$ translates to geodesic convexity of $\mathcal{E}_V$ defined in (6.17): Let $\mu_t$ be a geodesic joining $\mu, \tilde{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$, then

$$\begin{aligned}\mathcal{E}(\mu_t) &= \int_{\mathbb{R}^d} V(x)\, (((1-t)\mathrm{Id} + tT_{\mu\to\tilde{\mu}})_{\#}\mu)\, (\mathrm{d}x)\\ &= \int_{\mathbb{R}^d} V((1-t)x + tT_{\mu\to\tilde{\mu}}(x))\mu(\mathrm{d}x)\\ &\leq (1-t)\mathcal{E}_V(\mu) + t\mathcal{E}_V(\tilde{\mu}).\end{aligned}$$

In the same way, $\omega$-strong convexity of $V$ implies $\omega$-strong geodesic convexity of $\mathcal{E}_V$.

For internal energies of the form $\int f(\mu(x))\, \mathrm{d}x$ with $f$ convex and $f(0) = 0$, the situation is slightly more complicated. Essentially, one needs that the curvature of $f$ around $x = 0$ is sufficiently large, but decays for larger $x$ where $f$ may not be "too convex". a sufficient (but not necessary) condition can be stated as $s \mapsto f(e^{-s})e^s$ being convex and non-increasing for $s \in (0, \infty)$. We omit the proof here, it can be found in [7, Prop. 9.3.9]. Fortunately, the condition is met by the choice $f(x) = x \log x$, guaranteeing that the negative entropy functional $\mathcal{H}$ defined in (6.19) is geodesically convex. As sums of (strongly) geodesically convex functions are again (strongly) geodesically convex, this proves the following auxiliary result [7, sec. 9.4].

**Lemma 6.10: Geodesic convexity of KL divergence**

The KL divergence $\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \mu^*)$ is $\omega$-convex along geodesics in $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ if and only if $\mu^*$ is $\omega$-strongly log-concave.

If the target $\mu^*$ is strongly log-concave, we can employ EVI as in Lemma 6.9 and the arguments following it to deduce convergence.

If $\mu^*$ is not strongly log-concave, the gradient flow may still converge to a global minimizer if a gradient domination condition akin to the Polyak-Łojasiewicz inequality (6.25) is satisfied. Since $\mathcal{D}_{\mathrm{KL}}(\mu^* \,\|\, \mu^*) = 0$ is the unique global minimum of the KL divergence, the condition takes the form

$$\mathrm{FI}(\mu \,\|\, \mu^*) \coloneqq \left\| \nabla \log \left( \frac{\mu}{\mu^*} \right) \right\|^2_{L^2(\mu;\mathbb{R}^d)} \geq 2\omega \, \mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) \quad \forall \mu \in \mathcal{P}_2(\mathbb{R}^d). \tag{6.31}$$

The quantity $\mathrm{FI}(\mu \,\|\, \mu^*)$ is the Fisher information of $\mu$ (with respect to $\mu^*$). Upon slightly rewriting the above condition (6.31), we see that it can be guaranteed for all $\mu$ if we assume that $\mu^*$ satisfies a certain functional inequality called a log-Sobolev inequality [89]. In order to state the latter, we define the following, generalized entropy functional

$$\mathrm{Ent}_\mu(f) \coloneqq \mathbb{E}_\mu[f \log(f)] - \mathbb{E}_\mu[f] \log(\mathbb{E}_\mu[f]). \tag{6.32}$$

We then say $\mu^*$ satisfies the log-Sobolev inequality with constant $C_{\mathrm{LSI}} > 0$ if

$$\mathrm{Ent}_{\mu^*}(f^2) \leq 2C_{\mathrm{LSI}} \, \mathbb{E}_{\mu^*}[\|\nabla f\|^2], \tag{LSI}$$

for all smooth $f : \mathbb{R}^d \to \mathbb{R}$. We gather these derivations in the following result.

**Theorem 6.11: Gradient flow of KL: Ergodicity under (LSI)**

Suppose that $\mu^*$ satisfies (LSI) with $C_{\mathrm{LSI}} > 0$. Then, if $\mu_t$ denotes the solution of the Wasserstein gradient flow (6.30) of KL divergence with respect to $\mu^*$ with initialization $\mu_0$, we have

$$\mathcal{D}_{\mathrm{KL}}(\mu_t \,\|\, \mu^*) \leq \exp\left(-2tC_{\mathrm{LSI}}^{-1}\right) \mathcal{D}_{\mathrm{KL}}(\mu_0 \,\|\, \mu^*).$$

*Proof.* For $f = \sqrt{\mu/\mu^*}$, we have $\mathbb{E}_{\mu^*}[f^2] = 1$ and hence by the definition (6.32)

$$\mathrm{Ent}_{\mu^*}(f^2) = \mathbb{E}_{\mu^*}\left[ \frac{\mu}{\mu^*} \log\left( \frac{\mu}{\mu^*} \right) \right] = \mathbb{E}_\mu\left[ \log\left( \frac{\mu}{\mu^*} \right) \right] = \mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*).$$

Further, we have

$$\mathbb{E}_{\mu^*}[\|\nabla f\|^2] = \frac{1}{4}\,\mathbb{E}_{\mu^*}\left[\frac{\mu^*}{\mu}\left\|\nabla\left(\frac{\mu^*}{\mu}\right)\right\|^2\right] = \frac{1}{4}\,\mathbb{E}_{\mu}\left[\left\|\nabla \log\left(\frac{\mu^*}{\mu}\right)\right\|^2\right] = \frac{1}{4}\,\mathrm{FI}(\mu \,\|\, \mu^*).$$

Thus, if $\mu^*$ satisfies the log-Sobolev inequality, then the gradient domination condition (6.31) is true with $\omega = C_{\mathrm{LSI}}^{-1}$ and, by (6.26), we have exponential decay of the objective along the gradient flow. The statement of the theorem is precisely (6.26) with $\omega = C_{\mathrm{LSI}}^{-1}$ and $\mathcal{F} = \mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \mu^*)$. □

Note that strong convexity implies the log-Sobolev inequality [211, Thm. 21.2], this is also known as the Bakry-Émery criterion due to [14]. The converse is not true: The log-Sobolev inequality is actually weaker than convexity and is, e.g., still satisfied by bounded perturbations of log-concave distributions by the Holley–Stroock perturbation principle [95]. Beyond (LSI), there is a large amount of research trying to identify weak, practically verifyable conditions on $\mu^*$ that guarantee convergence of the gradient flow and its discretizations, see, e.g., [128, 210, 217, 153, 60].

Since we are going to return to this topic later in the context of mirror methods, we mention one other, even weaker functional inequality that implies exponential decay in a divergence measure with respect to $\mu^*$, namely, the Poincaré inequality[3]. The following standard derivation has been used, e.g., in [128, 58, 67]. Recall that the chi-squared divergence is given by $\mathcal{D}_{\chi^2}(\mu \,\|\, \mu^*) = \mathrm{Var}_{\mu^*}(\mathrm{d}\mu/\mathrm{d}\mu^*)$, whenever $\mu \ll \mu^*$. Suppose now that $\mu_t$ solves the WGF of KL divergence with initial distribution $\mu_0$. Similar to (6.24), we can compute the dissipation in chi-squared divergence to be

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\,\mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*) &= \int \frac{\mu_t}{\mu^*}\partial_t \mu_t \,\mathrm{d}x \\
&= \int \frac{\mu_t}{\mu^*}\nabla \cdot \left(\mu_t \nabla \log\left(\frac{\mu_t}{\mu^*}\right)\right) \mathrm{d}x \\
&= -\int \nabla\left(\frac{\mu_t}{\mu^*}\right)\mu_t \nabla \log\left(\frac{\mu_t}{\mu^*}\right) \mathrm{d}x \\
&= -\,\mathbb{E}_{\mu^*}\left[\left\|\nabla\left(\frac{\mu_t}{\mu}\right)\right\|^2\right]
\end{aligned} \tag{6.33}$$

As before, we would like to upper bound the final term by the negative chi-squared divergence for all $\mu_t$, in order to derive convergence using Grönwall's inequality. The necessary inequality is implied if $\mu^*$ satisfies the Poincaré inequality, i.e., if we have

$$\mathrm{Var}_{\mu^*}[f] \le C_{\mathrm{PI}}\,\mathbb{E}_{\mu^*}\Big[\|\nabla f\|^2\Big], \tag{PI}$$

[3] Strictly speaking, the Poincaré inequality is only weaker than (LSI) for log-concave measures, which are, however, the only target measures we care about in this work.

for all smooth $f : \mathbb{R}^d \to \mathbb{R}$ and some constant $C_{\mathrm{PI}} > 0$. We state the implied convergence result in the following theorem.

**Theorem 6.12: Gradient flow of KL: Ergodicity under (PI)**

Suppose that $\mu^*$ satisfies (PI) with $C_{\mathrm{PI}} > 0$. Then, if $\mu_t$ denotes the solution of the Wasserstein gradient flow (6.30) of KL divergence with respect to $\mu^*$ with initialization $\mu_0$, we have

$$\mathcal{D}_{\mathrm{KL}}(\mu_t \,\|\, \mu^*),\ \mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*) \leq \exp\left(-2tC_{\mathrm{PI}}^{-1}\right) \mathcal{D}_{\chi^2}(\mu_0 \,\|\, \mu^*).$$

*Proof.* Using (6.33) and setting $f = \mu_t/\mu^*$ in (PI), we immediately have

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*) \leq -2C_{\mathrm{PI}}^{-1} \mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*),$$

and therefore by Grönwall's inequality

$$\mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*) \leq \exp\left(-2C_{\mathrm{PI}}^{-1}t\right) \mathcal{D}_{\chi^2}(\mu_0 \,\|\, \mu^*). \tag{6.34}$$

This proves ergodicity in chi-squared divergence, which however also implies the same in KL divergence: By Jensen's inequality for the concave function $\log(x)$ and the standard logarithm inequality $\log(1+x) \leq x$, we know that

$$\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) \leq \log\left(\int \frac{\mu^2(x)}{\mu^*(x)}\,\mathrm{d}x\right) = \log\left(1 + \mathcal{D}_{\chi^2}(\mu \,\|\, \mu^*)\right) \leq \mathcal{D}_{\chi^2}(\mu \,\|\, \mu^*).$$

Therefore, if we replace $\mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*)$ by $\mathcal{D}_{\mathrm{KL}}(\mu_t \,\|\, \mu^*)$ on the left side of (6.34), the inequality still holds. □

Before moving on, we mention that for log-concave distributions, the Poincaré inequality is strictly weaker than the log-Sobolev inequality, since (LSI) implies (PI) with the same constant. A proof of this statement, together with a family of functional inequalities that continuously interpolate between the two can be found in [121]. It has recently been shown that the latter, interpolated inequalities can be used to analyze the gradient flows of KL divergence in more general metrics called Rényi divergences [60].

## 6.5. Diffusion Processes

While (6.16) gives a theoretical way to implement a solution algorithm for the optimization problem (6.1), it is not very practical in applications. The Wasserstein gradient flow is a high-dimensional PDE and, as we saw, the velocity fields characterizing Wasserstein gradients may also not be easily computable even if we know the correct density at

some time $t$. Instead, one can often rely on the intimate connection between PDEs at distribution level and SDEs at the particle level. In order to analyze this connection, we have to recall some basics on stochastic diffusion processes. The discussions are based on the books [166, 15].

Diffusion processes are Markov processes with an almost surely continuous sample path. They can be characterized by a drift and a diffusion coefficient. As is common in the field of machine learning applications, we take the perspective of Itō stochastic calculus. This means we call $(X_t)_{t\geq 0}$ a *diffusion process* if it is given by

$$X_t = X_0 + \int_0^t b(X_s, s)\, ds + \int_0^t \sigma(X_s, s)\, dW_s \tag{6.35}$$

with *drift field* $b : [0,\infty)\times\mathbb{R}^d \to \mathbb{R}^d$ and *noise coefficient* $\sigma : [0,\infty)\times\mathbb{R}^d \to \mathbb{R}^{d\times k}$. Here, $W_s$ denotes standard Brownian motion in $\mathbb{R}^k$. We will also call $\Sigma(t,x) = \frac{1}{2}\sigma(t,x)\sigma(t,x)^T$ the *diffusion coefficient*[4] and again refer to the parameter $t$ as *time*. Equivalently, we will say that $X_t$ satisfies the following Itō SDE:

$$\mathrm{d}X_t = b(X_t, t)\,\mathrm{d}t + \sigma(X_t, t)\,\mathrm{d}W_t, \tag{6.36}$$

Since we are mostly interested in time-homogeneous processes, we may drop the coefficients' time dependencies and just use $b(X_t)$, $\sigma(X_t)$, $\Sigma(X_t)$, or $\sigma, \Sigma$ for constant diffusion. When considering solutions $(X_t)_{t\geq 0}$ of such SDEs, we will denote the distribution of $X_t$ at time $t \geq 0$ by $\mu_t := \mathrm{Law}(X_t)$.

The following is a standard existence and uniqueness result on the solutions to the SDE (6.36) [102, chap. 4]:

**Theorem 6.13: SDE - existence and uniqueness of solutions**

Assume that $X_0 \sim \mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $X_0$ is independent of $W_t$. Assume further that the drift and noise coefficient satisfy a Lipschitz condition

$$\|b(x,t)-b(y,t)\| \vee \|\sigma(x,t)-\sigma(y,t)\|_F \leq C\|x-y\| \quad \forall t \in [0,\infty),\ x,y \in \mathbb{R}^d \tag{6.37}$$

and a linear growth condition

$$\|b(x,t)\| \vee \|\sigma(x,t)\|_F \leq C(1+\|x\|) \quad \forall (t,x) \in [0,\infty)\times\mathbb{R}^d \tag{6.38}$$

with some constants $C \in \mathbb{R}$.

The proof is similar to typical Picard iteration arguments known from the solution theory of ordinary differential equations. While these assumptions are sufficient, they are not necessary and can be weakened further. For example, we will deal with SDEs where the

[4] Note that sometimes $\sigma$ is also called diffusion coefficient. In order to avoid confusion, we will call $\sigma$ the noise coefficient instead.

drift $b = -\nabla V$ is the negative gradient of a strongly convex potential $V$. Such a $b$ may violate both conditions (6.37), (6.38) if $V$ grows too rapidly. However, a unique strong solution in $[0, \infty)$ still exists if we replace (6.37) by a local (in $x$) Lipschitz condition and (6.38) by a monotonicity condition in the drift and linear growth in the noise coefficient, i.e.

$$x^T b(x,t) \vee \|\sigma(x,t)\|_F^2 \leq C(1 + \|x\|^2), \tag{6.39}$$

see [142, chap. 3] for details. In this work, we will only consider SDEs that possess a unique strong solution.

Assume from now on that drift and diffusion coefficient are time-homogeneous, meaning $b(X_t), \Sigma(X_t)$ are independent of $t$. The diffusion process $X_t$ given in (6.35) is a Markov process. This means that it can be characterized by a family of Markov semigroup operators $(P_t)_{t\geq 0}$.

**Definition 6.14: Markov semigroup and generator**

The *Markov semigroup operators* $P_t$ are given by

$$(P_t f)(x) \coloneqq \int_{\mathbb{R}^d} f(y) P(t, x, \mathrm{d}y), \tag{6.40}$$

where $P(t, x, \mathrm{d}y) = \mathbb{P}[\Gamma \,|\, X_t \in \mathrm{d}y]$ is the standard Markov transition function. The family $(P_t)_{t\geq 0}$ satisfies the standard semigroup properties $P_0 = I$ and $P_{t+s} = P_t \circ P_s$ for all $t, s \geq 0$. We define the *generator* of $X_t$ as

$$\mathcal{L} f \coloneqq \lim_{t \to 0} \frac{P_t f - f}{t}. \tag{6.41}$$

Considering the semigroup property and (6.41), we can formally write $P_t = e^{t\mathcal{L}}$. A similar relation can be deduced for the (formal) $L^2$-adjoints of $P_t$ and $\mathcal{L}$. We define the adjoint semigroup $P_t^*$ as an operator on distributions over the state space by

$$(P_t^* \mu)(\Omega) \coloneqq \int_{\mathbb{R}^d} P(t, x, \Omega) \mu(\mathrm{d}x), \quad \Omega \in \mathcal{B}(\mathbb{R}^d), \tag{6.42}$$

giving the relation $\langle P_t f, \mu \rangle = \langle f, P_t^* \mu \rangle$. The corresponding $L^2$-adjoint $\mathcal{L}^*$ of the generator with $\langle \mathcal{L} f, \mu \rangle = \langle f, \mathcal{L}^* \mu \rangle$ then satisfies the formal relationship $P_t^* = e^{t\mathcal{L}^*}$. Assume now that $X_0 \sim \mu$ is some initialization of the Markov process and let $\mu_t = P_t^* \mu$. Then we obtain the following time derivative

$$\frac{\partial \mu_t}{\partial t} = \frac{\mathrm{d}}{\mathrm{d}t}(P_t^* \mu) = \frac{\mathrm{d}}{\mathrm{d}t}(e^{t\mathcal{L}^*} \mu) = \mathcal{L}^* \left( e^{t\mathcal{L}^*} \mu \right) = \mathcal{L}^* (P_t \mu) = \mathcal{L}^* \mu_t. \tag{6.43}$$

The partial differential equation $\frac{\partial \mu_t}{\partial t} = \mathcal{L}^* \mu_t$ is also called the *forward Kolmogorov equation*[5], and for diffusion processes typically the *Fokker–Planck equation*. Diffusion processes of the form (6.35) form a special class since their generator $\mathcal{L}$ can be shown to be the following partial differential operator

$$\mathcal{L} = \sum_{j=1}^{d} b_j \frac{\partial}{\partial x_j} + \sum_{j,k=1}^{d} \Sigma_{jk} \frac{\partial^2}{\partial x_j \partial x_k}. \tag{6.44}$$

The formal computations above leading to the Fokker–Planck equation can be made rigorous in the following way, we refer to [166, chap. 2-4] for details.

**Theorem 6.15: Fokker–Planck equation**

Consider a diffusion process $X_t$ given by (6.35), with $b, \sigma$ sufficiently smooth. If $X_0 \sim \mu$, then the density of the distribution $\mu_t = \mathrm{Law}(X_t)$ solves the initial value problem of the Fokker–Planck equation

$$\begin{aligned} \frac{\partial \mu_t}{\partial t} &= -\sum_{j=1}^{d} \frac{\partial}{\partial x_j}(b_j \mu_t) + \sum_{j,k=1}^{d} \frac{\partial^2}{\partial x_j \partial x_k}(\Sigma_{jk} \mu_t)(x), \\ \mu_0 &= \mu. \end{aligned} \tag{6.45}$$

For a constant (in space) diffusion coefficient $\Sigma \in \mathbb{R}^{d \times d}$, the Fokker–Planck equation is frequently written in the simple form

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (-b\mu_t + \Sigma \nabla \mu_t). \tag{6.46}$$

Apart from the Fokker–Planck equation, the other important computational tool for us when dealing with SDEs will be a type of chain rule for stochastic calculus. The following result is called Itō's Lemma or Itō's formula (presented here for vector-valued, time-homogeneous processes).

**Theorem 6.16: Itō's formula**

Suppose $(X_t)_{t \geq 0}$ satisfies the Itō SDE

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t.$$

[5] The "forward" stems from the fact that a similar equation, the *backward Kolmogorov equation*, holds for the generator $\mathcal{L}$. Although the derivation is analogous, it is not particularly relevant for our work, so we omit it here.

Let $f \in C^2(\mathbb{R}^d)$, i.e., $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. Then $f(X_t)$ satisfies the Itō SDE

$$\begin{aligned} \mathrm{d}f(X_t) = &\left( b(X_t)^T \nabla f(X_t, t) + \frac{1}{2} \operatorname{tr}(\sigma(X_t)^T \nabla^2 f(X_t) \sigma(X_t)) \right) \mathrm{d}t \\ &+ \sigma(X_t)^T \nabla f(X_t) \, \mathrm{d}W_t. \end{aligned}$$

For vector-valued maps $f : \mathbb{R}^d \to \mathbb{R}^m$, this result can simply be extended by applying the formula componentwise to $f_j$, $j = 1, \ldots, m$.

## 6.6. Langevin Diffusion

As already mentioned, we are interested in the case when the drift $b$ is the negative gradient of a potential $V$, which we will from now on assume proper and strongly convex, unless stated otherwise. The reason for that now becomes clear in connection with the derivations of Section 6.3. If we set $b = -\nabla V$ and $\Sigma = I$ in (6.46), it coincides with the WGF (6.30) for the KL divergence $\mathcal{D}_{\mathrm{KL}}(\cdot \,\|\, \mu^*)$ with respect to the target $\mu^* \propto e^{-V}$. This connection justifies the derivations on WGFs ex post facto: Given the task to sample from the measure $\mu^*$, the perspective of optimization in Wasserstein space showed that the continuity equation (6.30) converges to $\mu^*$ under certain assumptions on $V$. Due to the gradient flow structure, the dynamics are in a sense optimal in the Wasserstein geometry, but solving the high-dimensional PDE is computationally prohibitive. Theorem 6.15 now justifies that instead of solving the PDE, one can equivalently simulate the diffusion process by solving the SDE (6.36). This SDE is still high-dimensional, but a solver only has to keep track of a single particle, instead of the density on the whole space. If the SDE solver is exact and the resulting Markov chain ergodic and reversible, the particle's distribution will converge to $\mu^*$ as $t \to \infty$ and we can accept the generated particle positions as representative samples from $\mu^*$.

In this special case $b = -\nabla V$, $\sigma = \sqrt{2}I$ (so that $\Sigma = I$), the SDE (6.36) is typically referred to as *overdamped Langevin diffusion*. It is given by

$$\mathrm{d}X_t = -\nabla V(X_t) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}W_t. \tag{6.47}$$

The name *Langevin diffusion* goes back to Langevin's use of Brownian motion to describe the movement of a particle that is subject to deterministic and stochastic forces [120]. The complete Langevin equation with damping parameter $\gamma > 0$ is the second order SDE

$$m \frac{\mathrm{d}^2 X_t}{\mathrm{d}t^2} = -\gamma \frac{\mathrm{d}X_t}{\mathrm{d}t} - \nabla V(X_t) + \sqrt{2\gamma k_B T} \frac{\mathrm{d}W_t}{\mathrm{d}t}, \tag{6.48}$$

where $m$ is the particle's mass and $k_B T = \beta^{-1}$ the thermal energy that we have already seen in Chapter 4 on Bayesian modeling. The "overdamped" setting $\gamma \gg m$ yields (6.47). While we will concentrate on the overdamped case, it is worth pointing out that several works have employed the general case for the construction of sampling algorithms. Recall that in optimization, accelerated gradient descent methods can be understood as discretizations of a second order modification of the gradient flow [199]. In a similar spirit, the (underdamped) Langevin equation can be used to deploy accelerated Langevin sampling algorithms, as a line of recent works has explored [57, 140, 49, 226, 227, 159].

**Example 6.17: Ornstein-Uhlenbeck process**

A simple special case of a Langevin diffusion is the *Ornstein-Uhlenbeck process*. Let $\mu^* = \mathrm{N}(m, C)$ be a normal distribution with mean $m \in \mathbb{R}^d$ and covariance $C \succ 0$. Then the target density obeys $\mu^*(x) \propto \exp(-V(x))$ with $V(x) = \frac{1}{2}(x - m)^T C^{-1}(x - m)$.
The Langevin diffusion process for this target is given by

$$\mathrm{d}X_t = -C^{-1}(X_t - m)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{6.49}$$

and its corresponding Fokker–Planck equation is

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (-\mu_t(x) C^{-1}(x - m)) + \Delta \mu_t. \tag{6.50}$$

Both equations can be solved in closed form. The SDE solution, for instance, can be derived by variation of parameters: Setting $Y_t = X_t e^{tC^{-1}}$ and using the Itō formula shows $\mathrm{d}Y_t = e^{tC^{-1}} C^{-1} \mu\,\mathrm{d}t + \sqrt{2} e^{tC^{-1}}\,\mathrm{d}W_t$, which can simply be solved for $Y_t$ by taking the Itō integral on both sides. It follows that $X_t$ is distributed as

$$X_t \overset{d}{=} m + e^{-tC^{-1}}(X_0 - m) + C^{\frac{1}{2}} \left(I - e^{-2tC^{-1}}\right)^{\frac{1}{2}} Z, \quad Z \sim \mathrm{N}(0, I)$$

meaning that, conditioned on $X_0$, the Gaussian $\mu_t$ converges to $\mathrm{N}(m, C)$ as $t \to \infty$. Likewise, one could solve the Fokker–Planck PDE analytically and arrive at the same result for $\mu_t$, for instance by taking the Fourier transform and using the method of characteristics as was done in a similar linear drift case in [43].

We want to close this section by reiterating on the convergence of overdamped Langevin diffusion processes. By making the connection to Wasserstein gradient flows, we automatically inherited the convergence theory that we discussed in the context of WGFs in Section 6.3. For instance, we there saw that strong convexity of $V$ implies the EVI or weaker functional inequalities, which can be used to derive convergence rates of the measure $\mu_t$ to the invariant $\mu^* \propto e^{-V}$. When approaching convergence analysis from the SDE side and a particle level, we can actually derive equivalent rates in Wasserstein

distance without ever using the Fokker–Planck equation. Since the technique will reappear in our analysis of other types of Langevin dynamics and their discretizations, it is beneficial for the understanding of later sections to demonstrate it in continuous time as well.

**Theorem 6.18: Overdamped Langevin: Convergence by coupling**

Suppose $V$ is $\omega$-strongly convex and $(X_t)_{t\geq 0}$ solves (6.47) with some initial distribution $X_0 \sim \mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then we have

$$\mathcal{W}_2(\mu_t, \mu^*) \leq e^{-\omega t}\mathcal{W}_2(\mu_0, \mu^*).$$

*Proof.* We will employ a coupling argument. Two processes $X_t, \tilde{X}_t$ are called coupled if both solve (6.47) with the same Brownian motion $W_t$, i.e. they solve the joint SDE

$$\begin{pmatrix} \mathrm{d}X_t \\ \mathrm{d}\tilde{X}_t \end{pmatrix} = -\begin{pmatrix} \nabla V(X_t) \\ \nabla V(\tilde{X}_t) \end{pmatrix} + \sqrt{2}\begin{pmatrix} \mathrm{d}W_t \\ \mathrm{d}W_t \end{pmatrix}.$$

Let $X_t, \tilde{X}_t$ be coupled with $X_0 \sim \mu_0$ and $\tilde{X}_0 \sim \tilde{\mu}_0$. Using Theorem 6.15, the joint distribution $\eta_t$ of $(X_t, \tilde{X}_t)$ satisfies the Fokker–Planck equation

$$\frac{\partial \eta_t}{\partial t} = \nabla_x \cdot (\eta_t \nabla V(x)) + \nabla_{\tilde{x}} \cdot (\eta_t \nabla V(\tilde{x})) + (\nabla_x + \nabla_{\tilde{x}}) \cdot (\nabla_x + \nabla_{\tilde{x}})\eta_t. \tag{6.51}$$

Crucially, this holds with a flexible initial value $\eta_0$, since the joint equation holds for any coupling $\eta_0$ with marginals $\mu_0$ and $\tilde{\mu}_0$. We can now rewrite this using $(\nabla_x + \nabla_{\tilde{x}}) \cdot (\nabla_x + \nabla_{\tilde{x}}) = \Delta_{x+\tilde{x}}$ by a coordinate change $(x, \tilde{x}) \mapsto (x + \tilde{x}, x - \tilde{x})$. For any distance function of the form $d(x, \tilde{x}) = \rho(\|x - \tilde{x}\|)$, we directly see $\Delta_{x+\tilde{x}} d(x, \tilde{x}) = 0$. Using this and the joint Fokker–Planck equation, we can show a contraction in Wasserstein distance as follows. We start by estimating

$$\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}_{(x,\tilde{x})\sim\eta_t}[\|x - \tilde{x}\|^2] \\
&= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\int \|x - \tilde{x}\|^2 \eta_t(\mathrm{d}x, \mathrm{d}\tilde{x}) \\
&= \frac{1}{2}\int \|x - \tilde{x}\|^2 \left[\nabla_x \cdot (\eta_t \nabla V(x)) + \nabla_{\tilde{x}} \cdot (\eta_t \nabla V(\tilde{x})) + \Delta_{x+\tilde{x}}\eta_t\right] \mathrm{d}x\,\mathrm{d}\tilde{x} \\
&= -\int (x - \tilde{x}) \cdot (\nabla V(x) - \nabla V(\tilde{x}))\eta_t(\mathrm{d}x, \mathrm{d}\tilde{x}) \\
&\leq -\omega\,\mathbb{E}_{(x,\tilde{x})\sim\eta_t}[\|x - \tilde{x}\|^2],
\end{aligned}$$

where we used the quadratic lower bound of the symmetric Bregman distance

$$(x - \tilde{x}) \cdot (\nabla V(x) - \nabla V(\tilde{x})) \geq \omega\|x - \tilde{x}\|^2.$$

Note that we already used this bound when we sketched the convergence rate proof of Euclidean gradient flow—indeed, the computation here is a straightforward generalization of (4.26). Applying now Grönwall's lemma, we obtain

$$\mathbb{E}_{(x,\tilde{x})\sim\eta_t}[\|x-\tilde{x}\|^2] \leq e^{-2\omega t}\,\mathbb{E}_{(x,\tilde{x})\sim\eta_0}[\|x-\tilde{x}\|^2].$$

The left side is bounded from below by $\mathcal{W}_2^2(\mu_t,\tilde{\mu}_t)$ by definition. By taking the infimum over all couplings $\eta_0$ with marginals $\mu_0$ and $\tilde{\mu}_0$ on the right hand side, we obtain

$$\mathcal{W}_2(\mu_t,\tilde{\mu}_t) \leq e^{-\omega t}\mathcal{W}_2(\mu_0,\tilde{\mu}_0).$$

In particular, if we choose the process $\tilde{X}_t$ to be initialized at the target distribution as $\tilde{\mu}_0 = \mu_*$, then the distribution of $\tilde{X}_t$ is stationary at $\tilde{\mu}_t = \mu^*$. By the above, we obtain the convergence result

$$\mathcal{W}_2(\mu_t,\mu^*) \leq e^{-\omega t}\mathcal{W}_2(\mu_0,\mu^*).$$

□

We will next move to time discretizations and introduce Langevin Monte Carlo sampling algorithms.

# Chapter 7

# Langevin Monte Carlo: Algorithms

The previous two chapters showed that simulating (6.47) provides a fast way to approximate a target distribution $\mu^* \propto e^{-V}$ on particle level. Bringing this to an actually implementable algorithm comes down to discretizing the SDE in time. Some of the possible discretizations are straightforward, the simplest being a standard forward Euler–Maruyama scheme. The latter leads to the most important of Langevin Monte Carlo algorithms, the unadjusted Langevin algorithm (ULA).

Section 7.1 reviews possible direct discretizations of the SDE (6.47), ULA being only one example of such a method. For non-smooth potentials $V$, there are several popular schemes which make use of proximal mappings and all differ slightly, depending on the chosen splitting and order of time-discretized steps. We devote Section 7.2 to the direct comparison of these in a simplified setting of Gaussian targets. In this special case, we exactly know the distribution of the algorithms at each step and at equilibrium, and can prove that certain implicit discretizations consistently achieve smaller non-asymptotic and asymptotic bias than a popular smoothing-based iteration. We then move to more non-standard schemes. Section 7.3 introduces a more recently proposed sampling strategy based on a primal-dual splitting, and Section 7.4 explains how to transfer the strategy of mirroring or preconditioning to gradient-based sampling methods. Sections 7.5 and 7.6 discuss further practical aspects of Langevin sampling needed for experimental results, namely sampling from constraint domains without mirrored schemes and proximal sampling in the case where proximal mappings are available only inexactly.

## 7.1. Overdamped Langevin Schemes and Source of Bias

Similarly as different time discretizations of ODEs may lead to a variety of different optimization algorithms, there is a vast literature on sampling algorithms that arise from time discretizations of (6.47). Many common ideas from convex optimization translate to log-concave sampling tasks: Explicit discretizations require gradient evaluations [179] and introduce growth conditions and step size constraints [68, 74, 216, 69, 75, 210], while implicit schemes lead to proximal methods [167, 217, 73, 185, 186]. More complex discretizations like mid- or multipoint schemes potentially achieve higher order consistency, but come at the cost of increased computational effort [169, 109]. For highly anisotropic potentials, techniques like preconditioning or mirror maps can be employed to accelerate the algorithms [97, 58, 225, 2, 122, 130]. A review of the entire field is beyond the scope of this work, so we will go into detail on some aspects that are relevant for imaging inverse problems. In particular, we are going to put emphasis on algorithms for non-smooth and constrained potentials due to the nature of the considered imaging priors and likelihoods.

**Forward Euler–Maruyama and unadjusted Langevin algorithm.** A standard discretization method for Itō SDEs is the Euler–Maruyama scheme. With a time step size $\tau > 0$, discretizing (6.47) generates the Markov chain $(X^{(n)})_{n\geq 0}$ defined by

$$X^{(n+1)} = X^{(n)} + \tau \nabla \log \mu^*(X^{(n)}) + \sqrt{2\tau}\xi^{(n+1)}, \qquad \xi^{(n+1)} \sim \mathrm{N}(0, I_d) \tag{ULA}$$

with some given initialization $X^{(0)} \sim \mu^{(0)}$. The iteration is typically called unadjusted Langevin algorithm (ULA). In what follows, we will often denote the distribution at the $n$-th step $X^{(n)}$ of the Markov chain by $\mu^{(n)}$.

A major challenge in the analysis of Langevin Monte Carlo algorithms like ULA lies in bounding the discretization error, i.e., guaranteeing that $\mu^{(n)} \approx \mu_{n\tau}$ in a suitable metric, where $\mu_t$ is the distribution of the continuous time Markov process. While the situation is fairly well-understood for ULA (see for example [69], we will go into more detail later), we are also interested in other time discretizations of (6.47). In order to understand the nature of these discretizations, we again take the perspective of Wasserstein gradient flows. ULA can indeed be understood as a splitting scheme of the gradient flow of KL divergence [216] if we write it as two half steps

$$\begin{aligned} X^{(n+\frac{1}{2})} &= X^{(n)} - \tau \nabla V(X^{(n)}) \\ X^{(n+1)} &= X^{(n+\frac{1}{2})} + \sqrt{2\tau}\xi^{(n+1)}, \quad \xi^{(n+1)} \sim \mathrm{N}(0, I). \end{aligned}$$

The objective functional decomposes to $\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) = \mathcal{E}_V(\mu) + \mathcal{H}(\mu)$ as in (6.28). According to Example 6.6, the gradient descent step to $X^{(n+\frac{1}{2})}$ can be understood as an explicit forward step with respect to the energy $\mathcal{E}_V$. Since adding a Gaussian random variable implements the heat kernel, the second half step to $X^{(n+1)}$ actually corresponds

to an exact solution of the heat equation or the gradient flow of $\mathcal{H}$ along a time step of length $\tau$, see Example 6.7. This shows that ULA is a scheme alternating first-order discretized forward steps and exact time steps to solve the WGF of the KL divergence. We now want to formalize this splitting idea with exact and first-order discretized steps and consider both explicit/forward and implicit/backward discretizations. Consider the following nomenclature.

- *Forward* (F) will denote a forward in time discretization of a WGF (6.16). For a Euclidean gradient flow of a functional $f : \mathbb{R}^d \to \mathbb{R}$, a forward step corresponds to approximating $x(t+\tau)-x(t) \approx \tau\dot{x}(t) = -\tau\nabla V(x(t))$. We already saw in (4.29) that this generates updates of the form $x^{(n+1)} = x^{(n)} - \tau\nabla V(x^{(n)}) = (\mathrm{Id} - \tau\nabla V)(x^{(n)})$. By considering the equivalent variational formulation

$$x^{(n+1)} = \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \left\{ V(x^{(n)}) + \langle \nabla V(x^{(n)}), x - x^{(n)}\rangle + \frac{1}{2\tau}\|x - x^{(n)}\|^2 \right\}, \tag{7.2}$$

  we can define the analogue for an objective $\mathcal{F}$ in Wasserstein space by

$$\mu^{(n+1)} = \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{F}(\mu^{(n)}) + \left\langle \frac{\delta\mathcal{F}}{\delta\mu}(\mu^{(n)}), \mu - \mu^{(n)} \right\rangle_{L^2(\mathbb{R}^d)} + \frac{1}{2\tau}\mathcal{W}_2^2(\mu, \mu^{(n)}) \right\}. \tag{7.3}$$

  By our discussions in Section 6.3, the equivalent in Wasserstein space of a direct update on particle level is the transport of a measure under a suitable $L^2$ velocity field. Indeed, rigorously deriving the optimality condition of (7.3) (see [7, chap. 11]) yields $\mu^{(n+1)} = (\mathrm{Id} - \tau\nabla_{\mathcal{W}_2}\mathcal{F}(\mu^{(n)}))_{\#}\mu^{(n)}$. Note the additional dependence of the Wasserstein gradient on the density $\mu^{(n)}$ due to the Riemannian structure of $\mathcal{P}_2(\mathbb{R}^d)$. This dependence vanishes in the special case $\mathcal{F} = \mathcal{E}_V$ where $\nabla_{\mathcal{W}_2}\mathcal{F} = \nabla V$, since the pointwise nature of the potential energy does not encode any particle interaction. In that case, a forward step can be implemented at particle level by $X^{(n+1)} = X^{(n)} - \tau\nabla V(X^{(n)})$, so we recover the standard forward Euler scheme for Euclidean gradient flows. For this special case, we will informally use the notation $\mathrm{F}_V^\tau(x) \coloneqq x - \tau\nabla V(x)$.

- *Backward* (B) denotes a backward in time discretized step of the gradient flow. This corresponds to the implicit approximation $x(t+\tau) - x(t) \approx \tau\dot{x}(t+\tau) = -\tau\nabla V(x(t+\tau))$, and hence the update $x^{(n+1)} = (\mathrm{Id} + \tau\nabla V)^{-1}x^{(n)}$, where the inverse of $\mathrm{Id} + \tau\nabla V$ is well-defined due to the convexity of $V$ [20, chap. 23]. As we already saw in our brief discussion of gradient methods for optimization, this leads to the proximal update

$$x^{(n+1)} = \mathrm{prox}_V^\tau(x^{(n)}) = \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \left\{ V(x) + \frac{1}{2\tau}\|x - x^{(n)}\|^2 \right\}, \tag{7.4}$$

where $\mathrm{prox}_V^\tau$ is the proximal operator. In Wasserstein space, we equivalently define backward steps by

$$\mu^{(n+1)} = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{F}(\mu) + \frac{1}{2\tau} \mathcal{W}_2^2(\mu, \mu^{(n)}) \right\}. \tag{7.5}$$

The resulting transport is implicitly given as $(\mathrm{Id} + \tau \nabla_{\mathcal{W}_2} \mathcal{F}(\mu^{(n+1)}))_\# \mu^{(n+1)} = \mu^{(n)}$. As before, for potential energies the update can be carried out particle-wise, i.e. if $\mathcal{F} = \mathcal{E}_V$ and $X^{(n)} \sim \mu^{(n)}$, then $X^{(n+1)} = \mathrm{prox}_V^\tau(X^{(n)}) = (\mathrm{Id} + \tau \nabla V)^{-1}(X^{(n)})$ has distribution $X^{(n+1)} \sim \mu^{(n+1)}$. To contrast forward and backward steps, we will sometimes also denote $\mathrm{B}_V^\tau = \mathrm{prox}_V^\tau$.

- By *flow*, we denote the exact solution of the Wasserstein gradient flow along one time step of length $\tau$. Since gradient flows are time-homogeneous, this corresponds to assigning $\mu^{(n+1)} = \mu_\tau$, where $(\mu_t)_{t \in [0,\tau]}$ is the solution to the initial value problem

$$\begin{cases} \partial_t \mu_t = \nabla \cdot (\mu_t \nabla_{\mathcal{W}_2} \mathcal{F}(\mu_t)), \\ \mu_0 = \mu^{(n)}. \end{cases} \tag{7.6}$$

In the special case $\mathcal{F} = \mathcal{H}$, the flow step can be carried out exactly on particle-level. In that case, (7.6) is the unconstrained heat equation, so $\mu_0 = \mu^{(n)}$ implies

$$\mu^{(n+1)} = \mu_\tau = \mu_0 * \frac{1}{(4\pi\tau)^{d/2}} \exp(-\|\cdot\|^2/4\tau), \tag{7.7}$$

where $*$ denotes a convolution of the densities. The particle equivalent of this convolution with a heat kernel is adding a zero mean Gaussian random variable $X^{(n+1)} = \mathit{flow}_{\mathcal{H}}^\tau(X^{(n)}) = X^{(n)} + \sqrt{2\tau}\xi$ with $\xi \sim \mathrm{N}(0, I)$.

To summarize, while forward and backward steps are easy when $\mathcal{F} = \mathcal{E}_V$, this does not hold in the case $\mathcal{F} = \mathcal{H}$. There is no closed form representation of discretized forward or backward steps of the heat kernel unless the density has a particularly simple structure, e.g., a Gaussian. Some works implement restricted Gaussian oracles, which are backward steps of $\mathcal{H}$ applied to Dirac point masses, in order to construct consistent sampling schemes, but this is typically computationally costly [186, 127, 55, 131]. Conversely, the WGF of negative entropy $\mathcal{H}$ can easily be solved exactly at particle level, but for potential energies $\mathcal{E}_V$, exact steps are usually intractable.

**Source of bias of Langevin sampling algorithms.** The type of steps applied to a splitting and their order influences the convergence behaviour. We demonstrate this on the simpler case of a forward-backward scheme in Euclidean space.

**Example 7.1: Forward-backward scheme in Euclidean optimization**

Consider the Euclidean gradient flow $\dot{x}_t = -\nabla V(x_t)$ in $\mathbb{R}^d$, where we assume $V = F+G$ with $F$ convex and differentiable on $\mathbb{R}^d$ and $G$ convex, proper and lower semi-continuous. The forward-backward optimization scheme for this objective reads

$$x^{(n+\frac{1}{2})} = x^{(n)} - \tau\nabla f(x^{(n)})$$
$$x^{(n+1)} = \mathrm{prox}_g^\tau(x^{(n+\frac{1}{2})}).$$

A standard result from convex optimization states that alternating forward and backward steps keep stationary points of the objective fixed [161, sec. 4]: Let $x^{(n)} \in \mathbb{R}^d$ such that $0 \in \partial V(x^{(n)}) = \nabla f(x^{(n)}) + \partial g(x^{(n)})$. Writing out the optimality condition of the second half step shows that $x^{(n+1)}$ is the unique point satisfying

$$\frac{x^{(n)} - \tau\nabla f(x^{(n)}) - x^{(n+1)}}{\tau} = \frac{x^{(n+\frac{1}{2})} - x^{(n+1)}}{\tau} \in \partial g(x^{(n+1)}).$$

Since $\nabla f(x^{(n)}) \in -\partial g(x^{(n)})$ due to the stationarity, this is solved by $x^{(n+1)} = x^{(n)}$, so $x^{(n)}$ is a fixed point of the iteration.

The identical argument can be made for discretizations of Wasserstein gradient flows [216, appdx. F]. For instance, a convergent forward-backward discretization of the WGF of KL divergence would be

$$\begin{cases} \mu^{(n+\frac{1}{2})} = (\mathrm{Id} - \tau\nabla V)_\# \mu^{(n)} \\ \mu^{(n+1)} = \arg\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \left\{\mathcal{H}(\mu) + \frac{1}{2\tau}\mathcal{W}_2^2(\mu, \mu^{(n+\frac{1}{2})})\right\}, \end{cases} \tag{7.9}$$

with a forward step in $\mathcal{E}_V$ and a backward step in $\mathcal{H}$.

Unlike the forward-backward scheme, other combinations like a forward-forward or a forward-flow scheme are generally not convergent. As [216] pointed out, this perspective gives an explanation why Langevin algorithms like (ULA) are typically biased. The Markov chain possesses a unique stationary distribution due to its ergodicity (this will be shown rigorously in Chapter 8), but this stationary distribution must necessarily be different from the target $\mu^* \propto e^{-V}$. A convergent discretization would need to balance forward and backward steps. As a potential solution, [216] also proposes such a symmetrized Langevin Monte Carlo scheme as an alternative to (ULA). We will now give an overview of several other discretization schemes that can be applied to the Wasserstein gradient flow for sampling from $\mu^* \propto e^{-V}$.

**Flow-Backward.** For settings in which it is more efficient to evaluate $\mathrm{prox}_V^\tau$ than $\nabla V$, it has been proposed to replace the forward step $\mathrm{F}_V^\tau$ in ULA by a backward step $\mathrm{B}_V^\tau$,

see [167, 27] where this was introduced as proximal unadjusted Langevin algorithm, or P-ULA[1]. This leads to the scheme

$$\begin{aligned} X^{(n+1)} &= (\mathit{flow}_{\mathcal{H}}^{\tau} \circ \mathrm{B}_V^{\tau})(X^{(n)}) \\ &= \mathrm{prox}_V^{\tau}(X^{(n)}) + \sqrt{2\tau}\xi^{(n+1)}, \qquad \xi \sim \mathrm{N}(0, I). \end{aligned} \tag{B\textit{flow}}$$

Similarly, one could extract samples after every backward step. This is the same iteration up to initialization, but since it has been proposed in the literature as well [217, 93], we will briefly compare the two versions later on for practical reasons and state here both;

$$\begin{aligned} X^{(n+1)} &= (\mathrm{B}_V^{\tau} \circ \mathit{flow}_{\mathcal{H}}^{\tau})(X^{(n)}) \\ &= \mathrm{prox}_V^{\tau}(X^{(n)} + \sqrt{2\tau}\xi^{(n+1)}), \qquad \xi \sim \mathrm{N}(0, I). \end{aligned} \tag{\textit{flow}B}$$

Crucially, the (B*flow*) or (*flow*B) schemes can be employed even when $V$ is non-smooth, since $\mathrm{prox}_V^{\tau}$ is well-defined for any convex $V$. While (ULA) can be seen as a sampling equivalent of gradient descent, these methods are closely related to the proximal point method for optimization [161].

**Splitting schemes with two potential terms.** Since we are interested in Bayesian posterior targets whose log-density is the sum of log-likelihood and prior log-density, we will frequently encounter situations where the target is of the form $\mu^* \propto e^{-V}$ with $V = F + G$. Since $\mathcal{E}_V = \mathcal{E}_F + \mathcal{E}_G$, the KL objective decomposes to three terms, and we aim to optimize

$$\arg\min_{\mu} \mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) = \mathcal{E}_F(\mu) + \mathcal{E}_G(\mu) + \mathcal{H}(\mu). \tag{7.10}$$

Often, $F$ and $G$ are functionals with different structural properties, for instance $F$ might be smooth while $G$ is not, as we saw in Sections 4.1 and 4.2 when discussing typical likelihood and prior models in imaging inverse problems. As is common in (Euclidean) optimization [212, 62, 224, 70], it appears natural to consider three-operator splittings for sampling in order to solve (7.10). With the nomenclature from above, we will consider the following two algorithms.

**Forward-Flow-Backward.** Assume that $V = F + G$, where $F$ is convex and differentiable and $G$ is convex. Then the update formula of the forward-flow-backward (F*flow*B) Langevin sampling algorithm is given by

$$\begin{aligned} X^{(n+1)} &= (\mathrm{B}_G^{\tau} \circ \mathit{flow}_{\mathcal{H}}^{\tau} \circ \mathrm{F}_F^{\tau})(X^{(n)}) \\ &= \mathrm{prox}_G^{\tau}(X^{(n)} - \tau\nabla F(X^{(n)}) + \sqrt{2\tau}\xi^{(n+1)}), \qquad \xi \sim \mathrm{N}(0, I). \end{aligned} \tag{F\textit{flow}B}$$

[1] Several of the algorithms that we introduce here have been referred to as a "proximal Langevin algorithm". Due to the ambiguity that this umbrella term has developed, we will be explicit about the type and order of discretization steps and use these to distinguish different algorithms.

This algorithm (and stochastic variants of it) was introduced in [185, 186]. It follows earlier works [217] (which omitted the forward step, essentially doing a *flow*B scheme) [73], where a similar, up to initialization equivalent iteration was considered.

**Forward-Backward-Flow.** The forward-backward-flow (FB*flow*) algorithm is suited for the same composite potential $V = F + G$, with $F$ convex and differentiable and $G$ convex. Compared to F*flow*B, the proximal step and the diffusion step are exchanged, yielding the following iteration update:

$$\begin{aligned} X^{(n+1)} &= (\mathit{flow}_{\mathcal{H}}^{\tau} \circ \mathrm{B}_G^{\tau} \circ \mathrm{F}_F^{\tau})(X^{(n)}) \\ &= \mathrm{prox}_G^{\tau}(X^{(n)} - \tau \nabla F(X^{(n)})) + \sqrt{2\tau}\xi^{(n+1)}, \qquad \xi \sim \mathrm{N}(0, I). \end{aligned} \tag{FB\textit{flow}}$$

**Other forward-backward based splitting schemes.** Naturally, there are four more possible combinations of F, B and *flow* steps for splittings of the form $V = F + G$ which we have not formally introduced now. Indeed, the variant that could be called BF*flow* in our notation was suggested in [73] for non-smooth potentials. However, again each permutation of the three steps results in the iteration being equivalent (up to initialization) to either (F*flow*B) or (FB*flow*) and taking an intermediate iterate. For instance, computing BF*flow* amounts to running (F*flow*B) but extracting the samples after each *flow*-step instead. We are therefore going to present convergence analysis exclusively for (F*flow*B) and (FB*flow*). Practically, choosing between the two algorithms can change samples in a qualitative way. For instance, whenever $G$ is a sparsity-enforcing potential of a prior, the proximal points of $G$ tend to be likely samples under that prior, simply by definition of the proximal mapping. In contrast, after a diffusion step, sample images can look slightly "noisy". These differences naturally become weaker as $\tau$ shrinks to zero and the algorithms approximate the same continuous time dynamics.

**Smoothing-based iterations.** As we saw, the non-smoothness of a potential term can be mitigated by employing an implicit discretization, resulting in updates involving a proximal operator. For potentials of the form $V = F + G$, another method featuring proximal operators that is frequently encountered in the literature is the Moreau–Yosida unadjusted Langevin algorithm (MYULA). The proximal mapping in MYULA, however, does not arise from a backward step of the WGF. Indeed, MYULA is not a discretization of the WGF for the target $\mu^* \propto \exp(-V)$ at all, but rather ULA applied to a modified target. The idea is to separate differentiable terms in $F$ and non-differentiable terms in $G$, and then approximate $G$ by a differentiable regularization $G^\lambda \approx G$. The modified target $\mu^\lambda \propto \exp(-V^\lambda) = \exp(-F - G^\lambda)$ has a smooth density, so ULA can be applied to $\mu^\lambda$. A suitable type of regularization is the Moreau–Yosida envelope defined by

$$G^\lambda(x) = \inf_y \left\{ \frac{1}{2\lambda}\|x - y\|_2^2 + G(y) \right\} \tag{7.11}$$

Since $G$ is convex and the quadratic strongly convex, the infimum is always uniquely attained and the Moreau-Yosida envelope is differentiable everywhere with $\nabla G^\lambda(x) = \frac{1}{\lambda}(x - \mathrm{prox}_G^\lambda(x))$, see [20, chap. 12]. Hence, the MYULA iteration with smoothing parameter $\lambda > 0$ and step size $\tau > 0$ is given by

$$X^{(n+1)} = (1 - \frac{\tau}{\lambda})X^{(n)} - \tau\nabla F(X^{(n)}) + \frac{\tau}{\lambda}\mathrm{prox}_G^\lambda(X^{(n)}) + \sqrt{2\tau}\xi^{(n+1)}, \quad \xi^{(n+1)} \sim \mathrm{N}(0, I_d) \tag{MYULA}$$

Proving convergence of the algorithm comes down to using the triangle inequality and proving two separate bounds. Firstly, one employs a ULA convergence result to show that $\mu^{(n)} \approx \mu^\lambda$ for large enough $n$. Since $\nabla G^\lambda$ is $\lambda^{-1}$-Lipschitz continuous, this typically requires small enough $0 < \tau \leq \lambda$. Secondly, one needs to prove that the approximation error due to smoothing can also be controlled, i.e. $\mu^\lambda \approx \mu^*$ in some metric for small enough $\lambda$. We will omit the details on this second step here and refer to [76] for details. Eventually, the additional approximation error due to smoothing can be mitigated just as any of the standard discretization errors of Langevin samplers, by combining MYULA with a Metropolis–Hastings correction kernel, which was explored in the recent work [66].

## 7.2. Analytical Comparisons for Gaussian Targets

So far, we have presented theoretical arguments on why sampling algorithms based on Langevin diffusion (6.47) may be useful in general, and worked out several different discretization schemes for said SDE. Both the smoothing based (MYULA) and methods based on backward discretizations like (FB*flow*) were introduced in the literature around a similar time and serve a similar purpose, namely, expanding on ULA by using proximal mappings that are able to handle non-smooth log-densities. This leaves us now with multiple different algorithms for the same task and an obvious question:

*If we want to sample from a target with non-smooth density, which of the different proximal Langevin sampling algorithms should we choose?*

Even in the simple case where $V$ is not split up into smooth and non-smooth term, there are three different proximal schemes, (B*flow*), (*flow*B) and, setting $f \equiv 0$ (MYULA). Unfortunately, the stationary distributions of the Markov chains are usually not available in closed form, and the theoretical guarantees for Langevin sampling algorithms not sharp. Thus, comparing the different algorithms can either be done via (i) numerical examples, forming empirical evidence on their comparison, or (ii) computations for simple cases, in which we know the stationary distributions exactly. While (i) has been done to some extent, e.g. for (MYULA) and (F*flow*B) in [123, 84], these works did not provide a clear answer on whether one of the algorithms may be preferable.

In this section, we attempt to form a more complete picture using (ii). The main exception to the inaccessibility of the stationary is made by Gaussian targets $\mu^* =$

$\mathrm{N}(m, C)$. As the target's potential is the quadratic function $V(x) = \frac{1}{2}(x-m)^T C^{-1}(x-m)$, the overdamped Langevin SDE (6.47) reduces to an Ornstein-Uhlenbeck process, see Example 6.17. When discretizing this diffusion process in time, the forward and backward steps of the methods in Section 7.1 are affine-linear, allowing us to derive the densities of the Markov chains' states and equilibria in closed form. This is long known for (ULA) [162, 216] and was used for (*flow*B) in [217], for more complex implicit discretizations of overdamped Langevin diffusion [169, 109] and for simplified score-based methods in a recent paper [100].

For the rest of the section, we set $\mu^* = \mathrm{N}(m, C)$ and consider various discretizations of the corresponding overdamped Langevin SDE

$$\mathrm{d}X_t = -C^{-1}(X_t - m)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t.$$

The states of ULA applied to this special case are characterized as follows [216].

**Lemma 7.2: ULA for Ornstein–Uhlenbeck processes**

For the Gaussian target $\mu^* = \mathrm{N}(m, C)$, the state $X_{\mathrm{ULA}}^{(n)} \sim \mu_{\mathrm{ULA}}^{(n)}$ after $n$ steps of (ULA) with initialization $X_{\mathrm{ULA}}^{(0)} \sim \mu_{\mathrm{ULA}}^{(0)}$ obeys

$$X_{\mathrm{ULA}}^{(n)} - m \overset{d}{=} A_{\mathrm{ULA}}^n(\tau)(X_{\mathrm{ULA}}^{(0)} - m) + \sqrt{2\tau}(I - A_{\mathrm{ULA}}^2(\tau))^{-\frac{1}{2}}(I - A_{\mathrm{ULA}}^{2n}(\tau))^{\frac{1}{2}}\xi, \quad (7.12)$$

where $A_{\mathrm{ULA}}(\tau) = I - \tau C^{-1}$ and $\xi \sim \mathrm{N}(0, I)$ is independent of $X_{\mathrm{ULA}}^{(0)}$. If the step size $\tau$ satisfies $0 < \tau < 2\lambda_{\min}(C)$, then $\lim_{n\to\infty} A_{\mathrm{ULA}}^n(\tau) = 0$ and we obtain convergence to the stationary distribution as $n \to \infty$:

$$X_{\mathrm{ULA}}^{(n)} \xrightarrow{d} m + \sqrt{2\tau}(I - A_{\mathrm{ULA}}^2(\tau))^{-\frac{1}{2}}\xi. \quad (7.13)$$

The stationary distribution of ULA is hence

$$\mu_{\mathrm{ULA}}^{\mathrm{stat}} = \mathrm{N}\left(m, 2\tau(I - A_{\mathrm{ULA}}^2(\tau))^{-1}\right) = \mathrm{N}\left(m, C(I - \frac{\tau}{2}C^{-1})^{-1}\right). \quad (7.14)$$

Since $(I - \frac{\tau}{2}C^{-1})^{-1} \succ I$, the stationary distribution is, compared to the target, overdispersed. The overdispersion error is largest in the direction of $v_{\min}$, the eigenvector of $C$ belonging to $\lambda_{\min}(C)$.

*Proof.* The potential gradient is $\nabla V(x) = C^{-1}(x - m)$. A single update of (ULA), shifted by $m$, gives

$$\begin{aligned} X_{\mathrm{ULA}}^{(n+1)} - m &= X_{\mathrm{ULA}}^{(n)} - \tau\nabla V(X_{\mathrm{ULA}}^{(n)}) - m + \sqrt{2\tau}\xi^{(n+1)} \\ &= (I - \tau C^{-1})(X_{\mathrm{ULA}}^{(n)} - m) + \sqrt{2\tau}\xi^{(n+1)}. \end{aligned}$$

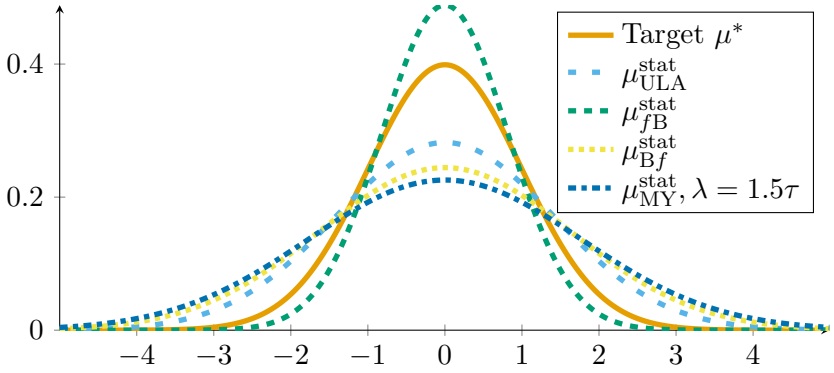



Figure 7.1.: Stationary densities of (ULA), (MYULA), (*flow*B) and (B*flow*) applied to a Gaussian target. For a one-dimensional Gaussian target with $m = 0$ and $C = 1$, we plot the stationary densities of the different algorithms considered in this section. For (MYULA), the smoothing parameter was set to $\lambda = 1.5\tau$ for visualization purpose, note that convergence theory requires $\lambda \geq \tau$ and in the present Gaussian special case, (MYULA) and (B*flow*) coincide for $\lambda = \tau$.

Denote $A_{\text{ULA}}(\tau) \coloneqq (I - \tau C^{-1})$. Unrolling the above update for $n$ steps, one obtains

$$\begin{aligned} X^{(n)}_{\text{ULA}} - m &= A^n_{\text{ULA}}(\tau)(X^{(0)}_{\text{ULA}} - m) + \sqrt{2\tau}\sum_{k=0}^{n-1} A^k_{\text{ULA}}(\tau)\xi^{(n-k)} \\ &\overset{d}{=} A^n_{\text{ULA}}(\tau)(X^{(0)}_{\text{ULA}} - m) + \sqrt{2\tau}\left(\sum_{k=0}^{n-1} A^{2k}_{\text{ULA}}(\tau)\right)^{\frac{1}{2}}\xi \\ &= A^n_{\text{ULA}}(\tau)(X^{(0)}_{\text{ULA}} - m) + \sqrt{2\tau}(I - A^2_{\text{ULA}}(\tau))^{-\frac{1}{2}}(I - A^{2n}_{\text{ULA}}(\tau))^{\frac{1}{2}}\xi, \end{aligned}$$

with $\xi \sim \mathrm{N}(0, I)$, where we have used the summation rule for independent Gaussian distributions in the second line. If $0 < \tau < 2\lambda_{\min}(C)$, we can bound the operator norm $\|A_{\text{ULA}}(\tau)\| < 1$, so that we immediately obtain the convergence result (7.13). The remaining statements follow from computing $2\tau(I - A^2_{\text{ULA}}(\tau))^{-1} = C(I - \frac{\tau}{2}C^{-1})$ and recognizing that the matrix $(I - \frac{\tau}{2}C^{-1})$ is positive definite due to the step size condition $\tau < 2\lambda_{\min}(C)$. □

The comparison between the stationary distribution of (ULA) and a one-dimensional Gaussian target is visualized in Fig. 7.1 and its asymptotic bias in Fig. 7.2.

A similar computation for zero mean Gaussians was carried out for the implicit variant (*flow*B) in [217]. We generalize this computation to arbitrary Gaussians and both (*flow*B) and (B*flow*). The respective stationary densities and biases are also visualized in Figs. 7.1 and 7.2

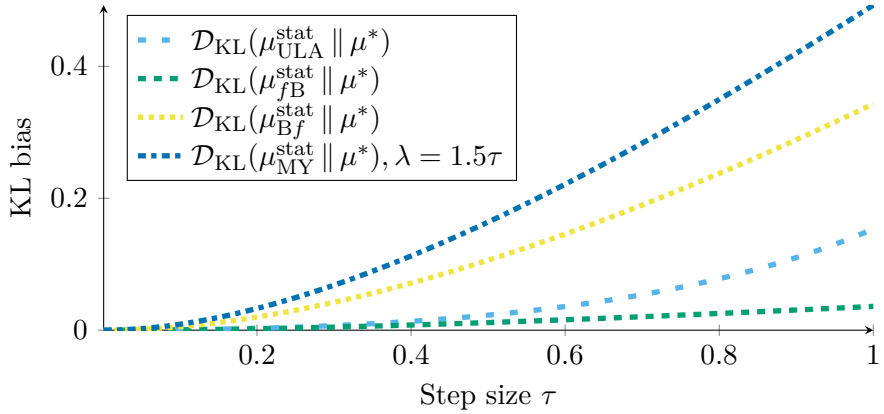



Figure 7.2.: Asymptotic bias in KL divergence of (ULA), (MYULA), (*flow*B) and (B*flow*) applied to a Gaussian target. For a one-dimensional Gaussian target with $m = 0$ and $C = 1$, we plot the KL divergence of the respective stationary distributions from the target $\mu^*$. As in Fig. 7.1, the (MYULA) smoothing parameter was set to $\lambda = 1.5\tau$ for visualization purpose.

**Lemma 7.3: (B*flow*) for Ornstein–Uhlenbeck processes**

For the Gaussian target $\mu^* = \mathrm{N}(m, C)$, the state $X^{(n)}_{\mathrm{B}f} \sim \mu^{(n)}_{\mathrm{B}f}$ after $n$ steps of (B*flow*) with initialization $X^{(0)}_{\mathrm{B}f} \sim \mu^{(0)}_{\mathrm{B}f}$ obeys

$$X^{(n)}_{\mathrm{B}f} - m \overset{d}{=} A^n_{\mathrm{B}f}(\tau)(X^{(0)}_{\mathrm{B}f} - m) + \sqrt{2\tau}(I - A^2_{\mathrm{B}f}(\tau))^{-\frac{1}{2}}(I - A^{2n}_{\mathrm{B}f}(\tau))^{\frac{1}{2}}\xi, \quad (7.15)$$

where $A_{\mathrm{B}f}(\tau) = (I + \tau C^{-1})^{-1}$ and $\xi \sim \mathrm{N}(0, I)$ is independent of $X^{(0)}_{\mathrm{B}f}$. For all $\tau > 0$, it holds $\lim_{n\to\infty} A^n_{\mathrm{B}f}(\tau) = 0$, so, as $n \to \infty$, the state converges to

$$X^{(n)}_{\mathrm{B}f} \overset{d}{\to} m + \sqrt{2\tau}(I - A^2_{\mathrm{B}f}(\tau))^{-\frac{1}{2}}\xi. \quad (7.16)$$

The stationary distribution is thus

$$\mu^{\mathrm{stat}}_{\mathrm{B}f} = \mathrm{N}\left(m, 2\tau(I - A^2_{\mathrm{B}f}(\tau))^{-1}\right) = \mathrm{N}\left(m, C(I + \tau C^{-1})^2(I + \frac{\tau}{2}C^{-1})^{-1}\right). \quad (7.17)$$

It holds $(I+\tau C^{-1})^2(I+\frac{\tau}{2}C^{-1})^{-1} \succ I$, so the stationary distribution is, compared to the target with covariance $C$, overdispersed in every direction in space.

*Proof.* For the potential $\nabla V(x) = \frac{1}{2}(x-m)^T C^{-1}(x-m)$, the proximal mapping is given by $\mathrm{prox}^\tau_V(x) = (I + \tau C^{-1})^{-1}(x + \tau C^{-1}m)$. An update step of (B*flow*) can thus be

rewritten as

$$\begin{aligned} X_{\mathrm{B}f}^{(n+1)} - m &= \mathrm{prox}_V^\tau(X_{\mathrm{B}f}^{(n)}) - m + \sqrt{2\tau}\xi^{(n+1)} \\ &= (I + \tau C^{-1})^{-1}(X_{\mathrm{B}f}^{(n)} - m) + \sqrt{2\tau}\xi^{(n+1)}. \end{aligned}$$

Denote $A_{\mathrm{B}f}(\tau) \coloneqq (I + \tau C^{-1})^{-1}$. By the same steps as in the proof of Lemma 7.2,

$$\begin{aligned} X_{\mathrm{B}f}^{(n)} - m &= A_{\mathrm{B}f}^n(\tau)(X_{\mathrm{B}f}^{(0)} - m) + \sqrt{2\tau}\sum_{k=0}^{n-1} A_{\mathrm{B}f}^k(\tau)\xi^{(n-k)} \\ &\overset{d}{=} A_{\mathrm{B}f}^n(\tau)(X_{\mathrm{B}f}^{(0)} - m) + \sqrt{2\tau}\left(\sum_{k=0}^{n-1} A_{\mathrm{B}f}^{2k}(\tau)\right)^{\frac{1}{2}}\xi \\ &= A_{\mathrm{B}f}^n(\tau)(X_{\mathrm{B}f}^{(0)} - m) + \sqrt{2\tau}(I - A_{\mathrm{B}f}^2(\tau))^{-\frac{1}{2}}(I - A_{\mathrm{B}f}^{2n}(\tau))^{\frac{1}{2}}\xi, \end{aligned}$$

with $\xi \sim \mathrm{N}(0, I)$. Since $C$ is positive definite, we have $\lim_{n\to\infty} A_{\mathrm{B}f}^n = 0$. A simple computation shows that the stationary covariance is

$$2\tau(I - A_{\mathrm{B}f}^2(\tau))^{-1} = C(I + \tau C^{-1})^2(I + \frac{\tau}{2}C^{-1})^{-1}.$$

This makes the stationary distribution overdispersed, since $(I + \tau C^{-1})^2(I + \frac{\tau}{2}C^{-1})^{-1} = I + \tau C^{-1}\frac{3+2\tau C^{-1}}{2+\tau C^{-1}} \succ I$. □

**Lemma 7.4: (*flow*B) for Ornstein–Uhlenbeck processes**

For the Gaussian target $\mu^* = \mathrm{N}(m, C)$, the state $X_{f\mathrm{B}}^{(n)} \sim \mu_{f\mathrm{B}}^{(n)}$ after $n$ steps of (*flow*B) with initialization $X_{f\mathrm{B}}^{(0)} \sim \mu_{f\mathrm{B}}^{(0)}$ obeys

$$X_{f\mathrm{B}}^{(n)} - m \overset{d}{=} A_{f\mathrm{B}}^n(\tau)(X_{f\mathrm{B}}^{(0)} - m) + \sqrt{2\tau}A_{f\mathrm{B}}(\tau)(I - A_{f\mathrm{B}}^2(\tau))^{-\frac{1}{2}}(I - A_{f\mathrm{B}}^{2n}(\tau))^{\frac{1}{2}}\xi, \quad (7.18)$$

where $A_{f\mathrm{B}}(\tau) = A_{\mathrm{B}f}(\tau) = (I + \tau C^{-1})^{-1}$ and $\xi \sim \mathrm{N}(0, I)$ independent of $X_{f\mathrm{B}}^{(0)}$. The state converges in distribution as

$$X_{f\mathrm{B}}^{(n)} \overset{d}{\to} m + \sqrt{2\tau}A_{f\mathrm{B}}(\tau)(I - A_{f\mathrm{B}}^2(\tau))^{-\frac{1}{2}}\xi. \quad (7.19)$$

The stationary distribution is given by

$$\mu_{f\mathrm{B}}^{\mathrm{stat}} = \mathrm{N}\left(m, 2\tau A_{f\mathrm{B}}^2(\tau)(I - A_{f\mathrm{B}}^2(\tau))^{-1}\right) = \mathrm{N}\left(m, C(I + \frac{\tau}{2}C^{-1})^{-1}\right). \quad (7.20)$$

Compared to the target $\mu^*$, the stationary distribution is underdispersed in all directions.

*Proof.* The proof is almost identical to that of Lemma 7.3. The difference between the two algorithms is that the update is

$$X_{fB}^{(n+1)} - m = (I + \tau C^{-1})^{-1}(X_{fB}^{(n)} - m + \sqrt{2\tau}\xi^{(n+1)}),$$

i.e., the diffusion $\xi^{(n+1)}$ is also multiplied by $A_{fB}(\tau)$, giving an additional factor of $A_{fB}^2(\tau)$ in the covariance of the states and the stationary distribution. Crucially, this additional factor makes the stationary distribution underdispersed, while that of (B*flow*) was overdispersed. □

Deriving the closed form of the stationary distributions allows to quantify the bias and compare the different iterations. The following Lemma is (in a slightly modified version) due to [217], who showed that (*flow*B) has a smaller bias than (ULA) in relative entropy.

**Lemma 7.5: (ULA) vs. (flowB) for Gaussian targets**

Denote the eigenvalues of $C$ as $0 < c_1 \leq \cdots \leq c_d$. If $0 < \tau < 2c_1$, then both (ULA) and (*flow*B) converge to the respective stationary distributions $\mu_{\text{ULA}}^{\text{stat}}$ and $\mu_{fB}^{\text{stat}}$ reported in (7.14) and (7.20). It holds

$$\mathcal{D}_{\text{KL}}(\mu_{\text{ULA}}^{\text{stat}} \,\|\, \mu^*) = -\frac{d}{2} + \sum_{j=1}^{d} \left(1 - \frac{\tau}{2c_j}\right)^{-1} + \log\left(1 - \frac{\tau}{2c_j}\right),$$

for (ULA) and

$$\mathcal{D}_{\text{KL}}(\mu_{fB}^{\text{stat}} \,\|\, \mu^*) = -\frac{d}{2} + \sum_{j=1}^{d} \left(1 + \frac{\tau}{2c_j}\right)^{-1} + \log\left(1 + \frac{\tau}{2c_j}\right),$$

for (*flow*B). It holds $\mathcal{D}_{\text{KL}}(\mu_{fB}^{\text{stat}} \,\|\, \mu^*) < \mathcal{D}_{\text{KL}}(\mu_{\text{ULA}}^{\text{stat}} \,\|\, \mu^*)$ for all $\tau > 0$.

Before we prove the result, note that the bias values that we report here differ from the ones in [217], since the author derived the reverse values $\mathcal{D}_{\text{KL}}(\mu^* \,\|\, \mu_{\text{ULA}}^{\text{stat}})$ and $\mathcal{D}_{\text{KL}}(\mu^* \,\|\, \mu_{fB}^{\text{stat}})$ instead. The main conclusion here, namely, that the implicit scheme (*flow*B) has consistently smaller bias than the explicit algorithm ULA, remains unchanged.

*Proof.* We can use the following identity for the KL divergence of two Gaussian distributions: If $\mu_1 = \mathrm{N}(m_1, C_1)$ and $\mu_2 = \mathrm{N}(m_2, C_2)$ in $\mathbb{R}^d$, it holds

$$\mathcal{D}_{\text{KL}}(\mu_1 \,\|\, \mu_2) = \frac{1}{2}\left(\operatorname{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_1 - \mu_2)^T\Sigma_2^{-1}(\mu_1 - \mu_2) - d + \log\frac{\det\Sigma_2}{\det\Sigma_1}\right). \quad (7.21)$$

Since the Ornstein–Uhlenbeck processes in Lemmas 7.2 and 7.4 in the values from (7.14) and (7.20) proves the first two statements. Consider the function $f(x) = \frac{1}{x} + \log x$. By Taylor expanding $\log(1+\varepsilon)$ and $\log(1-\varepsilon)$ around 1, it is easy to see that it obeys the functional inequality $f(1-\varepsilon) > f(1+\varepsilon)$ for all $\varepsilon \in (0,1)$. Applying this inequality to $0 < \varepsilon = \frac{\tau}{2c_j} < 1$ for all $j = 1, \dots, d$ proves the desired result $\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{\mathrm{stat}} \,\|\, \mu^*) < \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{ULA}}^{\mathrm{stat}} \,\|\, \mu^*)$. □

The result Lemma 7.5 compared explicit and implicit schemes and showed that implicit steps may be advantageous—at least when ignoring practical factors like the computational difference between evaluating gradients and proximal mappings. We are now specifically interested in a comparison between the three simplest proximal schemes (B*flow*), (*flow*B) and (MYULA) with $F \equiv 0$. For Gaussian targets, we will show that we have the following two inequalities

$$\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{\mathrm{stat}} \,\|\, \mu^*) < \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{B}f}^{\mathrm{stat}} \,\|\, \mu^*) \leq \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{MY}}^{\mathrm{stat}} \,\|\, \mu^*). \tag{7.22}$$

Here, the second inequality holds if and only if the smoothing parameter $\lambda$ of (MYULA) is chosen as the minimum value $\lambda = \tau$, which results in (MYULA) and (B*flow*) being identical. This partly answers the central question from the beginning of the section: At least for Gaussian targets, (*flow*B) achieves a smaller bias than (B*flow*) or (MYULA). We will further prove that, if all algorithms are initialized with a fixed $X_0$, it holds

$$\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{(n)} \,\|\, \mu_{f\mathrm{B}}^{\mathrm{stat}}) < \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{B}f}^{(n)} \,\|\, \mu_{\mathrm{B}f}^{\mathrm{stat}}) \leq \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{MY}}^{(n)} \,\|\, \mu_{\mathrm{MY}}^{\mathrm{stat}}), \tag{7.23}$$

where equality again holds if and only if $\lambda = \tau$. To summarize, for Gaussian targets, implicit discretizations converge faster to their respective equilibrium states and achieve smaller bias than smoothing-based MYULA. Additionally, extracting the samples after each backward step instead of after the diffusion step further reduces the bias.

In order to prove (7.22) and (7.23), we first derive the stationary distribution of (MYULA). Since it coincides with (ULA) applied to a modified target, we can just use Lemma 7.2. For a visualization of the stationary distribution and the asymptotic bias for variable $\tau$, see Figs. 7.1 and 7.2

**Lemma 7.6: MYULA for Ornstein–Uhlenbeck processes**

For the Gaussian target $\mu^* = \mathrm{N}(m, C)$, the distribution $\mu_{\mathrm{MY}}^{(n)}$ after $n$ steps of (MYULA) with $G = V$ and $F \equiv 0$, conditioned on a fixed $X_{\mathrm{MY}}^0 \sim \mu_{\mathrm{MY}}^0$, is

$$\mu_{\mathrm{MY}}^{(n)} = \mathrm{N}\left(m + A_{\mathrm{MY}}^n(\tau,\lambda)(X_{\mathrm{MY}}^0 - m), 2\tau(I - A_{\mathrm{MY}}^2(\tau,\lambda))^{-1}(I - A_{\mathrm{MY}}^{2n}(\tau,\lambda))\right) \tag{7.24}$$

where $A_{\mathrm{MY}}(\tau,\lambda) = I - \tau C^{-1}(I + \lambda C^{-1})^{-1}$. Under the typical MYULA step size assumption $\tau \leq \lambda$, we have $\lim_{n\to\infty} A_{\mathrm{MY}}^n(\tau,\lambda) = 0$ and the chain converges to its

stationary distribution

$$\begin{aligned}\mu_{\mathrm{MY}}^{\mathrm{stat}} &= \mathrm{N}\left(m, 2\tau(I - A_{\mathrm{MY}}^2(\tau,\lambda))^{-1}\right) \\ &= \mathrm{N}\left(m, C(I+\lambda C^{-1})^2\left(I + (\lambda - \frac{\tau}{2})C^{-1}\right)^{-1}\right). \end{aligned} \tag{7.25}$$

This stationary distribution is, like (B*flow*), overdispersed compared to the target.

*Proof.* Notice that the Moreau–Yosida envelope of $V(x) = \frac{1}{2}(x-m)^T C^{-1}(x-m)$ is given by

$$V^\lambda(x) = \frac{1}{2}(x-m)^T(1+\lambda C^{-1})^{-1}C^{-1}(x-m).$$

From there, one can apply Lemma 7.2, to obtain both statements (7.24) and (7.25). The overdispersion follows from noting

$$\frac{c_j(1+\lambda c_j^{-1})^2}{1+(\lambda-\frac{\tau}{2})c_j^{-1}} \geq \frac{c_j(1+\tau c_j^{-1})^2}{1+\frac{\tau}{2}c_j^{-1}} \geq c_j \tag{7.26}$$

for all $\lambda \geq \tau$, with equality if and only if $\lambda = \tau$. The values in this chain of inequalities are the eigenvalues of the covariance matrices of $\mu_{\mathrm{MY}}^{\mathrm{stat}}$ (left), $\mu_{\mathrm{B}f}^{\mathrm{stat}}$ (middle) and $\mu^*$, respectively. In Lemma 7.3, we already established the second inequality, showing that (B*flow*) exhibits overdispersion compared to the target. The first inequality shows that the effect is even stronger for MYULA. □

We are now able to derive the main result of this section.

**Theorem 7.7: MYULA vs. implicit schemes for Gaussian targets**

Assume that (*flow*B), (B*flow*) and (MYULA) are applied to the target $\mu^* = \mathrm{N}(m, C)$ with the same step size $\tau$ (and $\lambda \geq \tau$ for MYULA). Then

$$\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{\mathrm{stat}} \,\|\, \mu^*) < \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{B}f}^{\mathrm{stat}} \,\|\, \mu^*) \leq \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{MY}}^{\mathrm{stat}} \,\|\, \mu^*),$$

with equality in the second comparison if and only if $\lambda = \tau$.
Assuming further that all three algorithms are initialized with the same $X^{(0)}$, we have

$$\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{(n)} \,\|\, \mu_{f\mathrm{B}}^{\mathrm{stat}}) < \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{B}f}^{(n)} \,\|\, \mu_{\mathrm{B}f}^{\mathrm{stat}}) \leq \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{MY}}^{(n)} \,\|\, \mu_{\mathrm{MY}}^{\mathrm{stat}}).$$

*Proof.* From Lemma 7.5, we know the bias of (*flow*B) is

$$\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{\mathrm{stat}} \,\|\, \mu^*) = -\frac{d}{2} + \sum_{j=1}^{d}\left(1+\frac{\tau}{2}c_j^{-1}\right)^{-1} + \log\left(1+\frac{\tau}{2}c_j^{-1}\right), \tag{7.27}$$

Using formula (7.21) for the KL divergence of Gaussians, we have

$$\mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{B}f}^{\mathrm{stat}} \,\|\, \mu^*) = -\frac{d}{2} + \sum_{j=1}^{d} \frac{(1+\tau c_j^{-1})^2}{1+\frac{\tau}{2}c_j^{-1}} - \log\left(\frac{(1+\tau c_j^{-1})^2}{1+\frac{\tau}{2}c_j^{-1}}\right), \tag{7.28}$$

as well as

$$\mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{MY}}^{\mathrm{stat}} \,\|\, \mu^*) = -\frac{d}{2} + \sum_{j=1}^{d} \frac{(1+\lambda c_j^{-1})^2}{1+(\lambda-\frac{\tau}{2})\,c_j^{-1}} - \log\left(\frac{(1+\lambda c_j^{-1})^2}{1+(\lambda-\frac{\tau}{2})\,c_j^{-1}}\right), \tag{7.29}$$

From inequality (7.26) comparing the eigenvalues of the stationary covariances of (B*flow*) and (MYULA), we obtain

$$\frac{(1+\lambda c_j^{-1})^2}{1+(\lambda-\frac{\tau}{2})c_j^{-1}} \geq \frac{(1+\tau c_j^{-1})^2}{1+\frac{\tau}{2}c_j^{-1}}.$$

Using that $f(x) = x - \log x$ is monotonically increasing for $x \geq 1$, this immediately gives $\mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{B}f}^{\mathrm{stat}} \,\|\, \mu^*) \leq \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{MY}}^{\mathrm{stat}} \,\|\, \mu^*)$ with equality if and only if $\lambda = \tau$. For the first inequality, subtract (7.27) from (7.28) and obtain

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{B}f}^{\mathrm{stat}} \,\|\, \mu^*) - \mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{\mathrm{stat}} \,\|\, \mu^*) &= \sum_{j=1}^{d} \frac{(1+\tau c_j^{-1})^2 - 1}{1+\frac{\tau}{2}c_j^{-1}} - \log\left((1+\tau c_j^{-1})^2\right) \\
&= \sum_{j=1}^{d} 2\tau c_j^{-1} - 2\log\left(1+\tau c_j^{-1}\right) \\
&> 0,
\end{aligned}$$

where we have used that $\log(1+x) < x$ for all $x > 0$ in the last line.

For the second part of the theorem, we again employ formula (7.21) to obtain

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{(n)} \,\|\, \mu_{f\mathrm{B}}^{\mathrm{stat}}) &= -\frac{d}{2} + \mathrm{tr}(I - A_{f\mathrm{B}}^{2n}(\tau)) + s^T A_{f\mathrm{B}}^{2n}(\tau) s - \log\det(I - A_{f\mathrm{B}}^{2n}(\tau)) \\
&= s^T A_{f\mathrm{B}}^{2n}(\tau) s - \sum_{j=1}^{d} \left(a_j^{2n} + \log(1-a_j^{2n})\right),
\end{aligned}$$

where $s \coloneqq C^{-\frac{1}{2}}(I + \frac{\tau}{2}C^{-1})^{\frac{1}{2}}(X^{(0)} - m)$ and $a_j = (1+\tau c_j^{-1})^{-1} < 1$ are the eigenvalues of $A_{f\mathrm{B}}(\tau)$. Noticing that $A_{f\mathrm{B}}(\tau) = A_{\mathrm{B}f}(\tau)$ and employing the same formula, we get

$$\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{(n)} \,\|\, \mu_{f\mathrm{B}}^{\mathrm{stat}}) = s^T(1+\tau C^{-1})^T A_{f\mathrm{B}}^{2n}(\tau)(1+\tau C^{-1})s - \sum_{j=1}^{d} \left(a_j^{2n} + \log(1-a_j^{2n})\right),$$

Since $I + \tau C^{-1}$ and $A_{f\mathrm{B}}$ are symmetric and commute, this implies $\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{(n)} \,\|\, \mu_{f\mathrm{B}}^{\mathrm{stat}}) > \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{B}f}^{(n)} \,\|\, \mu_{\mathrm{B}f}^{\mathrm{stat}})$. For the states of (MYULA), we similarly get

$$\mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{MY}}^{(n)} \,\|\, \mu_{\mathrm{MY}}^{\mathrm{stat}}) = s^T A_{\mathrm{MY}}^{2n}(\tau, \lambda) s - \sum_{j=1}^{d} \left( b_j^{2n} + \log(1 - b_j^{2n}) \right),$$

where $s$ as before and $b_j = \frac{1+(\lambda-\tau)c_j^{-1}}{1+\lambda c_j^{-1}} < 1$ are the eigenvalues of $A_{\mathrm{MY}}(\tau, \lambda)$. It can easily be seen that $a_j \leq b_j$ for all $\lambda \geq \tau$, with equality if and only if $\lambda = \tau$. This implies $s^T A_{f\mathrm{B}}^{2n}(\tau) s \leq s^T A_{\mathrm{MY}}^{2n}(\tau, \lambda) s$. Finally, the function $f(x) = x - \log x$ is monotonically decreasing for $x \in (0, 1)$, so $f(1 - a_j^{2n}) \leq f(1 - b_j^{2n})$, proving the last statement $\mathcal{D}_{\mathrm{KL}}(\mu_{f\mathrm{B}}^{(n)} \,\|\, \mu_{f\mathrm{B}}^{\mathrm{stat}}) \leq \mathcal{D}_{\mathrm{KL}}(\mu_{\mathrm{MY}}^{(n)} \,\|\, \mu_{\mathrm{MY}}^{\mathrm{stat}})$. □

While this shows that implicit discretizations are more effective than applying ULA to a smoothed density in the Gaussian case, we want to mention that simple generalizations of this statement to more general classes of distributions are unlikely and would be difficult to prove due to the intractability of the stationary distribution. Indeed, the paper [179] which originally introduced ULA as an MCMC method, commented on the inconsistent convergence behaviour of ULA for distributions with different tails as being "not very appealing". As the arguments made there extend to the proximal algorithms up to some extent, this could be an explanation for why earlier experimental comparisons of MYULA and implicit algorithms [84, 123] did not conclude that either version is favourable.

## 7.3. Primal-Dual Langevin Schemes

Throughout the last sections, we have repeatedly exploited the intimate connections between convex optimization and log-concave sampling. The parallels between the two fields have inspired relevant advancements in the analysis of some of the sampling algorithms that we already considered [216, 73]. It hence appears logical to try and translate more of the available proof techniques and algorithms from optimization to sampling problems. In the context of MAP computation in Section 4.4, we already discussed the PDHG algorithm (4.38) as a solver whenever the posterior log-density is of the form (4.36)

$$\rho_{\mathrm{post}}(x \,|\, y) \propto \exp(-F(Kx) - G(x)).$$

The optimization routine is particularly relevant for imaging, since this structure is applicable to a broad class of posteriors that arise in imaging inverse problems. For instance, a typical splitting in imaging would be with a negative log-likelihood subsumed in $G$, a finite difference operator in $K$ and an $\ell^1$- or group $\ell^1$-norm for $F$, leading to one of the TV priors we discussed earlier. This could, however, be reversed with $F \circ K$ encoding the log-likelihood and $G$ the prior log-density, this depends entirely on the structure and properties of the chosen model's log-densities. Due to this flexibility in the optimization

approach, a primal-dual type sampler might also be a promising algorithmic candidate for the task of generating samples from these posteriors.

Just as ULA is often interpreted as a gradient descent with additional diffusion term, it appears natural to take a similar perspective on the primal-dual method (4.38) and add a diffusion step at some point in the iteration to construct a sampling algorithm. This could, for instance, lead to an iterative sampler with the following update

$$\begin{cases} \xi^{(n+1)} \sim \mathrm{N}(0, I_d) \\ Y^{(n+1)} = \mathrm{prox}_{\sigma f^*}(Y^{(n)} + \sigma K X_\theta^{(n)}) \\ X^{(n+1)} = \mathrm{prox}_{\tau g}(X^{(n)} - \tau K^T Y^{(n+1)}) + \sqrt{2\tau}\xi^{(n+1)} \\ X_\theta^{(n+1)} = X^{(n+1)} + \theta(X^{(n+1)} - X^{(n)}). \end{cases} \tag{ULPDA}$$

This algorithm was first proposed under the name unadjusted Langevin primal-dual algorithm (ULPDA) in [155] without further convergence analysis. It has since been mentioned in [90, 123], first analytic results were proposed in [43].

The major difference between the previous methods and (ULPDA) is the algorithm's continuous time limit. The convergence theory of ULA depended strongly on the fact that the target $\mu^*$ is the invariant distribution of overdamped Langevin diffusion (6.47), and that overdamped Langevin diffusion can be understood as realizing the Wasserstein gradient flow of the KL divergence, see Chapter 6. For (ULPDA), large parts of this theoretical preparation unfortunately become unusable, because we have neither a gradient flow structure, nor a direct relationship between the algorithm's continuous time limit and the target. To see that, let us derive these limiting dynamics. As there are two step sizes $\tau, \sigma$ in the algorithm that can both be understood as time steps, we assume they are coupled by $\lim_{\tau,\sigma\to 0} \frac{\sigma}{\tau} = \lambda$ where $0 < \lambda < \infty$ is some constant. Taking both step sizes to zero, the formal limiting Markov process $(X_t, Y_t)$ satisfies not the overdamped Langevin diffusion SDE (6.47), but rather the stochastic inclusion

$$\begin{cases} \mathrm{d}X_t \in -(\partial G(X_t) + K^T Y_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t \\ \mathrm{d}Y_t \in -\lambda(\partial F^*(Y_t) - K X_t)\,\mathrm{d}t, \end{cases} \tag{7.30}$$

which we can write in a joint variable $Z_t = (X_t, Y_t)$ as

$$\mathrm{d}Z_t \in -A_{\mathrm{PD}}(Z_t)\,\mathrm{d}t + B_{\mathrm{PD}}\,\mathrm{d}W_t. \tag{7.31}$$

Here we denoted the primal-dual coefficients

$$A_{\mathrm{PD}}(Z_t) \coloneqq \left\{ \begin{pmatrix} p_t + K^T Y_t \\ -\lambda(K X_t + q_t) \end{pmatrix} : p_t \in \partial G(X_t), q_t \in \partial F^*(Y_t) \right\},$$
$$B_{\mathrm{PD}} \coloneqq \begin{pmatrix} \sqrt{2} I_d \\ 0 \end{pmatrix}.$$

There are two main differences between overdamped Langevin diffusion (6.47) and the primal-dual dynamics (7.31) that are relevant in the analysis. Firstly, the diffusion coefficient $\Sigma_{\mathrm{PD}} \coloneqq \frac{1}{2} B_{\mathrm{PD}} B_{\mathrm{PD}}^T$ is only positive semi-definite, in this context sometimes also called degenerate, with an $m$-dimensional null space. Standard existence and uniqueness analysis for kinetic Fokker–Planck equations with positive definite diffusion coefficient thus do not hold anymore. Secondly, the drift coefficient can not be understood as the gradient of any forcing potential, so in particular, the inclusion (7.30) can not be understood as realizing the particle dynamics of a Wasserstein gradient flow. We will reconsider this algorithm's theoretical behaviour in more detail later. We only point out for now that, similar to MYULA, the modified continuous time limit introduces a new stationary distribution $\mu^{(\lambda)}$ which depends on the ratio $\lambda = \frac{\sigma}{\tau}$ of the algorithm's step sizes [43].

## 7.4. Mirror Langevin Sampling

Another class of sampling algorithms based on Langevin diffusion that recently gained popularity in imaging applications are mirrored methods. We already introduced mirror descent (4.34) for optimization, in which the mirror map $\nabla h$, induced by a mirror function or Bregman generator $h$, maps iterates to the mirrored space. In its continuous time limit, we saw in (4.35) that mirrored iterations can be understood as a discretized instance of preconditioned gradient flow

$$\dot{x}_t = -P(x_t)\nabla V(x_t),$$

with a preconditioner $P(x) = (\nabla^2 h(x))^{-1}$ induced by a mirror map $h$. These dynamics are known as a mirror flow or a Hessian Riemannian gradient flow. The "Riemannian" is due to the fact that the inner product $\langle u, v\rangle_{\nabla^2 h(x)} \coloneqq \langle \nabla^2 h(x) u, v\rangle$ induced by the Hessian endows $\Omega_h \coloneqq \mathrm{int}(\mathrm{dom}(h))$ with a Riemannian structure, if $h$ is of Legendre type[2]. Since the gradient $\nabla_{\nabla^2 h(x)} V(x)$ under the Riemannian metric is uniquely defined by $\langle \nabla_{\nabla^2 h(x)} V(x), \sigma\rangle_{\nabla^2 h(x)} = \langle \nabla V(x), \sigma\rangle$ for all $\sigma$, it follows immediately that $\nabla_{\nabla^2 h(x)} V(x) = (\nabla^2 h(x))^{-1} \nabla V(x)$, where $\nabla$ denotes the gradient in the Euclidean geometry. The gradient flow on the Riemannian manifold hence coincides with the preconditioned gradient flow in Euclidean space. In convex optimization, these well-established concepts explain the convergence behaviour of mirrored and preconditioned methods, since these methods often require convexity and smoothness relative to $h$ in order to converge [6, 174, 19, 138, 141].

The Riemannian viewpoint is a central concept in the growing literature aimed at transferring mirror concepts to gradient flows in Wasserstein space and sampling [97, 58, 225,

[2] Note that this and many of the following explanations on the Riemannian structure of mirroring also hold for a general Riemannian metric, i.e., a general $P : \mathbb{R}^d \to \mathcal{S}_+^d$, where $\mathcal{S}_+^d \subset \mathbb{R}^{d\times d}$ is the convex cone of symmetrix positive definite matrices. For practical, algorithmic reasons that become clear later, we will mostly focus on the mirror case $P = (\nabla^2 h)^{-1}$.

2, 122, 130, 173, 32]. Two equivalent approaches easily explain how to derive a diffusion analogue of the mirror optimization dynamics. The first arises by considering the continuity equation. The Liouville equation for the preconditioned gradient flow is

$$\partial_t \mu_t = \nabla \cdot (\mu_t P \nabla V), \tag{7.32}$$

i.e., the velocity field is $v_t(x) = -P(x)\nabla V(x)$ for all $t$. We saw in Section 6.3 that this can be understood as a Wasserstein gradient flow of a potential energy, since $\nabla V = \nabla_{\mathcal{W}_2}\mathcal{E}_V(\mu)$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. The change of geometry through $P = (\nabla^2 h)^{-1}$ can be understood as a locally linear transformation of the velocity field $\nabla_{\mathcal{W}_2}\mathcal{E}_V(\mu)$ for the objective $\mathcal{E}_V$. As in Section 6.3, we can therefore simply consider general continuity equations by replacing $\mathcal{E}_V$ with general objectives $\mathcal{F}$ on the Wasserstein space. For instance, for the KL divergence objective $\mathcal{F}(\mu) = \mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*)$ with $\mu^* \propto \exp(-V)$, this results in the diffusion equation

$$\begin{aligned}\partial_t \mu_t &= \nabla \cdot \left(\mu_t P \nabla \log \frac{\mu_t}{\mu^*}\right) \\ &= \nabla \cdot (\mu_t P(\nabla V + \nabla \mu_t)),\end{aligned} \tag{7.33}$$

with $P = (\nabla^2 h)^{-1}$ as before in the mirror case. The above is hence the Fokker–Planck equation for mirrored Langevin dynamics, from which we can derive a particle implementaion. The other, equivalent derivation of this PDE is via the Riemannian geometry viewpoint from above. Recall that the Hessian $P^{-1}(x) = \nabla^2 h(x)$ induces the inner product $\langle \cdot, \cdot \rangle_{\nabla^2 h(x)}$ and thereby the norm $\|\cdot\|_{\nabla^2 h(x)}$ at every point $x$. This induced Riemannian structure is equipped with the geodesic distance $d_h$ defined by

$$d_h(x, \tilde{x}) = \inf_{\gamma} \left( \int_0^1 \|\dot{\gamma}(t)\|^2_{\nabla^2 h(\gamma(t))} \, \mathrm{d}t \right)^{1/2}, \tag{7.34}$$

where the infimum is taken over all continuously differentiable curves $\gamma : [0, 1] \to \mathbb{R}^d$ with $\gamma(0) = x$, $\gamma(1) = \tilde{x}$. The geodesic distance allows to define generalized Wasserstein distances over the Riemannian manifold $(\Omega_h, d_h)$ via

$$\mathcal{W}_{h,p}(\mu, \tilde{\mu}) \coloneqq \inf_{\pi \in \Gamma(\mu, \tilde{\mu})} \left( \mathbb{E}_{(x,\tilde{x})\sim\pi}[d_h(x, \tilde{x})^p] \right)^{1/p}. \tag{7.35}$$

Note that all these definitions are consistent with the Euclidean case whenever we set $h = \|\cdot\|_2^2$. As before, the Riemannian (Wasserstein) gradients can be expressed as a preconditioned version of the Euclidean (Wasserstein) gradient via the relation $\nabla_{\mathcal{W}_{h,2}}\mathcal{F} = (\nabla^2 h)^{-1}\nabla_{\mathcal{W}_2}\mathcal{F}$. For the KL objective, this results in the same continuity equation (7.33) as before. This shows that the Riemannian Wasserstein $\mathcal{W}_{h,2}$-gradient flow over the manifold coincides with the preconditioned gradient flow in Wasserstein space over $\mathbb{R}^d$—we have derived the same equivalence between preconditioning and change of metric as is common in Euclidean optimization. We also mention that, depending on the definition of $h$, the computation of $\mathcal{W}_{h,2}$ may be cumbersome if there is

no closed form available for $d_h$. It can, however, be shown that $\mathcal{W}_{h,2}^2$ agrees up to second order with the so-called Bregman-Wasserstein distance, a Wasserstein distance with the Bregman divergence cost $\mathrm{D}_h(\cdot,\cdot)$. This has motivated studies of Bregman-Wasserstein spaces and their gradient flows [105].

A lot more can be said about the Riemannian structure, Wasserstein distances and duality concepts in the context of mirror maps. We refer to recent works [32, 173, 105] on this topic. For our purposes, the Fokker–Planck equation is sufficient to derive mirror Langevin sampling: The particle implementation of (7.33) is the SDE

$$\mathrm{d}X_t = ((\nabla \cdot P)(X_t) - P(X_t)\nabla V(X_t))\,\mathrm{d}t + \sqrt{2P(X_t)}\,\mathrm{d}W_t, \tag{7.36}$$

where we again abbreviated $P(x) = (\nabla^2 h(x))^{-1}$. Note that (7.36) is also known as Riemannian Langevin dynamics for a general Riemannian metric $P : \mathbb{R}^d \to \mathcal{S}_+^d$ [87]. Since $P$ is everywhere symmetric and positive definite, we denote by $\sqrt{P(x)}$ the unique symmetric positive definite square root of $P(x)$ for the sake of well-definedness. In distribution $\mu_t$ of $X_t$, it however does not make a difference which square root is chosen, since $\mu_t$ is fully determined by the diffusion coefficient $P$ in the Fokker–Planck equation (7.33). The case in which the Riemannian metric is induced by a Legendre type mirror function $h$ is particularly interesting, since it allows to rewrite the above SDE in a particularly simple way. For $P = (\nabla^2 h)^{-1}$, the divergence term is given by

$$\nabla \cdot ((\nabla^2 h(X_t))^{-1}) = (\nabla^2 h(X_t))^{-1}\,\mathrm{diag}\left(\nabla^3 h(X_t)(\nabla^2 h(X_t))^{-1}\right), \tag{7.37}$$

where the third derivative $\nabla^3 h(X_t) : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ acts as a tensor and we used diag to denote that we extract the diagonal of the matrix argument as a vector. If we now set $X_t = \nabla h^*(Y_t)$, or equivalently $Y_t = \nabla h(X_t)$, by the Itō formula Theorem 6.16, $X_t, Y_t$ satisfy the coupled equations

$$\begin{cases} X_t = \nabla h^*(Y_t) \\ \mathrm{d}Y_t = -\nabla V(X_t) + \sqrt{2\nabla^2 h(X_t)}\,\mathrm{d}W_t. \end{cases} \tag{7.38}$$

These mirror Langevin dynamics have been proposed in [225] and are a special case of the already longer known Riemannian Langevin dynamics [87, 165]. A time discretization of the SDE leads to the iteration

$$X^{(n+1)} = \nabla h^*\left(\nabla h(X^{(n)}) - \tau\nabla V(X^{(n)}) + \sqrt{2\tau\nabla^2 h(X^{(n)})}\xi^{(n+1)}\right), \quad \xi^{(n+1)} \sim \mathrm{N}(0, I), \tag{MLA}$$

which is known as mirror Langevin algorithm (MLA). Importantly for sampling applications, this discretization via the inverse mirror map $\nabla h^*$ is advantageous whenever the mirror function implicitly enforces a constraint on the state variable $X_t$. Say, for instance, we wish to enforce positivity along the sampling iteration, since for instance $\mathrm{dom}(V) = \mathbb{R}_+^d$, and therefore consider the negative entropy mirror function $h(x) = \sum_j x_j \log(x_j) - x_j$ with $\Omega_h = \mathbb{R}_+^d$. Then $(\nabla h^*(x))_j = \exp(x_j)$, so the dis-

cretized scheme (MLA) always produces positive values $X^{(n)}$. This greatly increases interpretability when iterates of the chain are extracted as representative samples of the target distribution. While the continuous time dynamics (7.36) with $P(x) = (\nabla^2 h(x))^{-1}$ automatically confine the state $X_t$ to $\Omega_h$, this would not be true for a forward discretization of (7.36), since the increments of $W_t$ are supported on the whole of $\mathbb{R}^d$.

## 7.5. Sampling under Domain Constraints

A typical, practical difficulty in imaging problems are hard constraints to the image's values. The most common instance is a non-negativity constraint, which typically appears in photonic imaging setups, where the reconstructed object could be an emission intensity, an attenuation coefficient, a density, or some other physical quantity that is by nature required non-negative. Other constraints may come up, e.g., in computational photography, where values are limited by both a minimum and maximum channel exposure.

Sampling methods based on overdamped Langevin diffusion are, by definition of the diffusion process, not well-suited for encoding hard constraints in the state variable. The diffusion SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t,$$

is only well-defined for $V$ differentiable in $\mathbb{R}^d$. In Appendix C, we also discuss the fact that "typically", the density $\mu_t$ of a diffusion process $X_t$ is positive everywhere at all times $t > 0$, since the Gaussian increments of the Brownian motion term are supported everywhere. Standard and overdamped Langevin make no exception to that rule [166, chap. 6]. As [84] recently showed, replacing $\nabla V$ by a singular selection of the subdifferential $\partial V$ is possible. It however still requires the subgradient to be finite everywhere on $\mathbb{R}^d$, preventing us from encoding hard constraints by adding a characteristic function to $V$. Note that this distinguishes diffusion dynamics from typical first-order optimization algorithms based on the Euclidean gradient flow $\dot{x} = -\nabla V(x)$ for an objective $V$. If $V$ is convex with $\mathrm{dom}(V) = \mathcal{C} \subsetneq \mathbb{R}^d$ a convex set, we can simply extend $V(x) = \infty$ for $x \notin \mathcal{C}$, preserving convexity, and consider the subgradient flow $\dot{x} \in -\partial V(x)$, which is still well-defined and convergent.

For sampling dynamics, the way to incorporate state constraints is typically to drastically change the considered diffusion equation. One possible technique that we already covered in Section 7.4 is via a mirror map $h$, resulting in diffusion processes that are constrained to the Riemannian manifold $\mathcal{C} = \mathrm{int}(\mathrm{dom}(h))$.

The second common technique that we now briefly review does not change the geometry of the state space, relying instead on reflected diffusion processes. Suppose we wish to constrain the states to a closed, convex set $\mathcal{C} \subset \mathbb{R}^d$. Instead of the Brownian motion

with drift, one considers a process reflected at the boundary $\partial\mathcal{C}$, solving the SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t - \nu_t L(\mathrm{d}t). \tag{7.39}$$

Here, $\nu_t$ is a unit outer normal vector to the boundary $\partial\mathcal{C}$, and $L$ is a measure supported on $\{t \geq 0 : X_t \in \partial\mathcal{C}\}$. The additional drift $\nu_t L(\mathrm{d}t)$ can be understood to "reflect" the process $X_t$ back to $\partial\mathcal{C}$ whenever the Brownian motion term would drive $X_t$ outside of $\partial\mathcal{C}$. Reflected diffusion processes are closely related to the Skorokhod problem [197], which aims to decompose a given stochastic process into a reflected process and a regulatory process (the term $-\nu_t L(\mathrm{d}t)$ in the example above).

The authors of [39, 40] proposed to use (7.39) for constrained sampling. Suppose $V : \mathcal{C} \to \mathbb{R}$, and the target is given by $\mu^*(x) \propto \exp(-V(x))\chi_{\mathcal{C}}(x)$, where $\chi_{\mathcal{C}}(x) = 1$ whenever $x \in \mathcal{C}$ and $\chi_{\mathcal{C}}(x) = 0$. Then the target's invariance under the reflected SDE is inherited from overdamped Langevin diffusion. Under similar regularity assumptions on $\nabla V$ as in the unconstrained case, the density $\mu_t$ of $X_t$ solves the Fokker-Planck equation, now with homogeneous Neumann boundary conditions

$$\begin{cases} \frac{\partial \mu_t}{\partial t} = \nabla \cdot (\mu_t \nabla V) + \Delta \mu_t & \text{in } \mathcal{C}, \\ \nabla \mu_t \cdot n = 0 & \text{on } \partial\mathcal{C}, \end{cases}$$

where $n$ is a normal vector to the boundary $\mathcal{C}$. The target $\mu^*$ is the invariant solution of the PDE. As proposed in [39, 40], the reflected process can be discretized by projecting samples onto the boundary of $\mathcal{C}$ at the end of every step. This leads to the following iterative scheme

$$X^{(n+1)} = \Pi_{\mathcal{C}}\left(X^{(n)} - \tau\nabla V(X^{(n)}) + \sqrt{2\tau}\xi^{(n+1)}\right), \tag{7.40}$$

with $\xi^{(n+1)} \sim \mathrm{N}(0, I_d)$ independent in every step, where $\Pi_{\mathcal{C}}$ is the orthogonal projection onto $\partial\mathcal{C}$. By carefully estimating the bounds between continuous-time process and the discretization, the authors of [40] are able to derive a polynomial bound on the Markov chain's mixing time in total variation distance.

Note that for any $\tau > 0$, the projection operator can be written as a proximal operator $\Pi_{\mathcal{C}} = \mathrm{prox}^{\tau}_{\chi_{\mathcal{C}}}$ of the characteristic function $\chi_{\mathcal{C}}$. This shows that (7.40) with potential $V$ is a special case of the algorithm (F*flow*B) with potential $V + \chi_{\mathcal{C}}$. In Section 8.2, we will discuss that this exposes a theoretical gap on the convergence of proximal sampling schemes. While (F*flow*B) is only known to converge for finite potentials, it does converge in the special case $G = \chi_{\mathcal{C}}$ due to its relation to (7.40) and reflected diffusion processes. The results on (7.40) act as a motivation to use (F*flow*B) for more general potentials $G$. We stress that this is a promising direction of future research on non-smooth sampling.

We finally mention recent contributions [149, 110], which build on the idea of reflected Langevin dynamics to develop efficient sampling methods tailored for low-photon imaging problems with non-negativity constraints. Similar to the situation in this thesis,

these works assume Bayesian posterior targets of the form

$$\mu^* \propto \exp(-(F+G))\chi_{\mathbb{R}^d_{\geq 0}}.$$

While the basic setup of the reflected SDE follows the same general principle as in (7.39) with $V = F + G$, the work [149] combines the ideas of reflected diffusion with Moreau–Yosida regularization for non-smooth potential terms $G$. This is analogous to the derivation of (MYULA) as an alternative to (ULA) for non-smooth potentials. In the time-discretization, instead of projecting at the end of every step, the authors propose to "reflect" the sample at the boundary. When the constraint is the non-negative orthant $\mathcal{C} = \mathbb{R}^d_{\geq 0}$, a reflection at zero simply amounts to taking the component-wise absolute value[3]. The reflected MYULA iteration then takes the form

$$\begin{aligned} X^{(n+1)} &= \left| X^{(n)} - \tau \nabla V^\lambda(x) + \sqrt{2\tau}\xi^{(n+1)} \right| \\ &= \left| (1 - \frac{\tau}{\lambda}) X^{(n)} - \tau \nabla F(X^{(n)}) + \frac{\tau}{\lambda} \mathrm{prox}_G^\lambda(X^{(n)}) + \sqrt{2\tau}\xi^{(n+1)} \right|, \end{aligned} \tag{7.41}$$

where $V^\lambda = F + G^\lambda$ and $G^\lambda$ is the Moreau–Yosida regularization of $G$ with parameter $\lambda$, cf. Section 7.1. In [110], the authors combine the previous works with ideas to incorporate data-driven priors to Langevin sampling algorithms. As a comparable version of reflected MYULA for data-driven priors, they replace the proximal mapping of $G$ with a trained denoiser, resulting in a plug-and-play approach. From an experimental viewpoint, these works show that the reflected diffusion approach can provide a theoretical framework allowing to adhere to hard non-negativity constraints in imaging inverse problems.

## 7.6. Proximal Operators in Practice

In practical imaging inverse problems, the posterior often involves non-smooth prior terms, such as total variation or sparsity-promoting penalties. The non-smoothness poses challenges for sampling methods based on Langevin dynamics, which traditionally rely on differentiable potentials. Let us assume again that the target is of the form

$$\mu^*(x) \propto \exp\left(-F(x) - G(x)\right),$$

with a smooth potential $F$ and a possibly non-differentiable potential $G$. As we established in Section 7.1, partly implicit schemes like (FB*flow*) or (F*flow*B) can overcome the non-smoothness problem, since they do not require gradients of the potential $G$, but its proximal operator $\mathrm{prox}_G$, which may be available in the non-smooth case.

[3] In the small step size limit $\tau \to 0$, there is no large difference between reflection (like (7.41)) or projection onto the boundary (as in (7.40)). Practically, the projection may lead to a positive mass of the boundary $\partial\mathcal{C}$ while the reflection slightly underestimates the mass in the neighbourhood of $\partial\mathcal{C}$.

In cases where $\mathrm{prox}_G$ is available, this can be very effective. For example, it is simple to derive the proximal mapping of $G = \alpha\|\cdot\|_1$ as

$$(\mathrm{prox}^\tau_{\|\cdot\|_1}(x))_i = \max\{|x_i| - \alpha\tau, 0\}\,\mathrm{sign}(x_i).$$

This allows to replace (ULA) by (F*flow*B) with comparable iteration cost in cases where the non-smooth term is an $\ell^1$-norm or the $\ell^1$-norm of an invertible frame. A practical limitation arises, however, when $\mathrm{prox}^\tau_G$ does not have a closed form and must be computed numerically, for instance via iterative convex solvers. A standard example of such a potential $G$ relevant to imaging is isotropic total variation (4.15). The solution of

$$\mathrm{prox}^\tau_{\mathrm{TV}}(x) = \arg\min_y\{\mathrm{TV}(y) + \frac{1}{2\tau}\|x - y\|^2\}$$

has no accessible, closed form and requires iterative computation on its own [41, 52]. In such cases, each sample update of a proximal Langevin scheme like (F*flow*B) requires an inner iteration, and only has access to an approximation of the proximal point. The resulting scheme is no longer a direct discretization of any SDE and raises the question of how such inexactness impacts convergence to the target distribution $\mu^*$.

By results in [77], the additional bias due to inexactness in proximal points can be controlled, if the inexactness error is of a special type and bounded. Inexact proximal steps can be formalized via the $\varepsilon$-subdifferential of $G$: given $x, y \in \mathbb{R}^d$ and $\varepsilon \geq 0$, we say that $y$ is an $\varepsilon$-approximate proximal point of $x$ if

$$\frac{1}{\tau}(x - y) \in \partial_\varepsilon G(y).$$

Here, $\partial_\varepsilon G(y)$ denotes the $\varepsilon$-subdifferential defined by

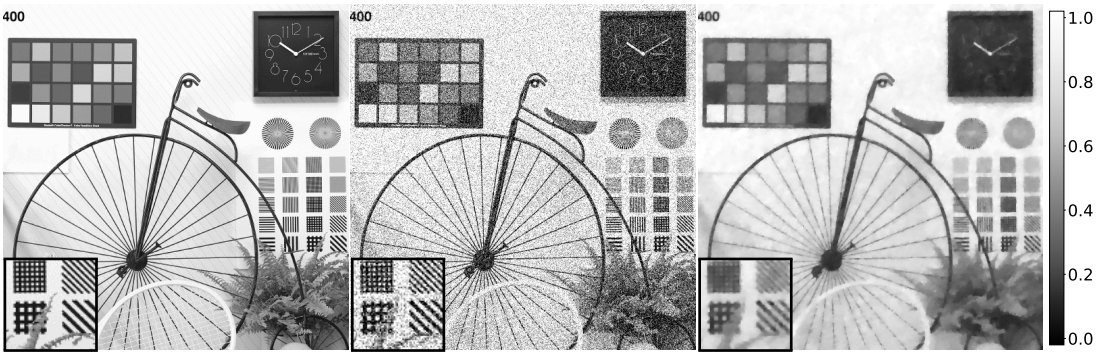
$$\partial_\varepsilon G(x) = \{p \in \mathbb{R}^d \,:\, G(y) \geq G(x) + \langle p, y - x\rangle - \varepsilon\},$$

which generalizes the standard subdifferential as the case $\varepsilon = 0$. This yields an iteration rule of the form

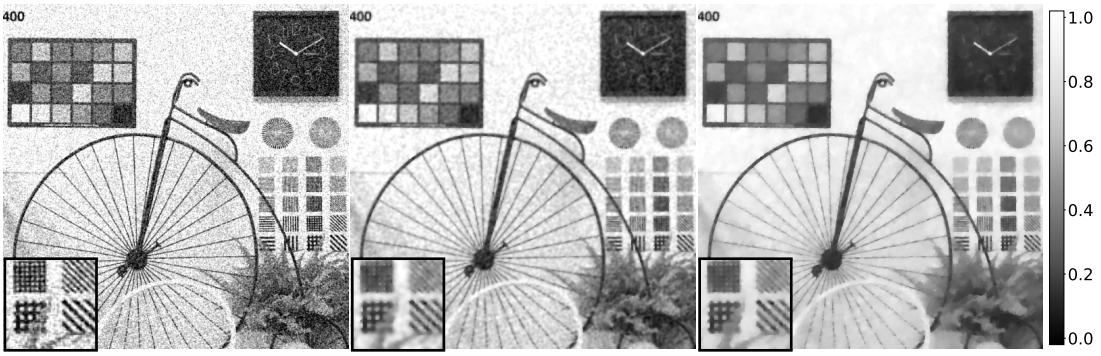
$$X^{(n+1)} \approx_\varepsilon \mathrm{prox}^\tau_G\left(X^{(n)} - \tau\nabla F(X^{(n)}) + \sqrt{2\tau}\xi^{(n+1)}\right). \tag{7.42}$$

By definition of this type of inexact proximal point, the approximation error $\varepsilon$ can be controlled in practice by bounding the duality gap in convex solvers. The error incurs a bias in the stationary distribution of the Markov chain, which [77] quantify. Importantly, under strong convexity of $F$ and bounded errors $\varepsilon$, the additional bias in the sampling distribution remains bounded and can be separated from the bias to the finite step size $\tau$. If errors $\varepsilon_n$ and step sizes $\tau_n$ are allowed to change over the course of the iteration, there are configurations of the algorithm in which $\tau_n \to 0$ and $\varepsilon_n \to 0$ such that the inexact scheme asymptotically recovers the exact target $\mu^*$. Moreover, the convergence analysis of (F*flow*B) with exact proximal points from [186] is recovered in the case $\varepsilon = 0$.

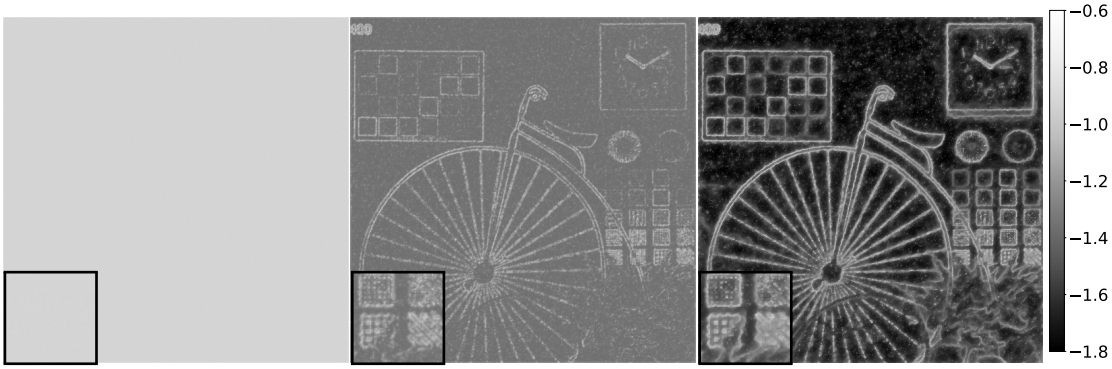
In a broad sense, the results of [77] play an important role in the practical implementation of Langevin schemes for imaging, as they act as the justification to use proximal schemes even in the case where we have to evaluate proximal mappings with an approximate, inner solver. On a more detailed level, the guarantees on inexact (F*flow*B) offer control over the regularizing effect of a non-smooth prior potential. If the proximal point is computed via gradient descent on the dual problem and the diffused image is used as a warm start at each iteration, using a lower accuracy $\varepsilon$ for proximal points can be checked to be equivalent to reducing the proximal step size, and hence to increasing the prior distribution's temperature. This directly results in a trade-off of accuracy and computational cost at each iteration. Figure 7.3 shows this effect on the example of a simple image denoising task with a regularizing TV Gibbs prior. For the lowest accuracy, computing a rough approximation of the TV proximal points is very cheap, but has almost no regularizing effect. Increasing the accuracy requires more iterations, but approximates better the true posterior mean for the chosen prior temperature. In practice, the necessary accuracy for a fixed temperature should therefore be calibrated by comparing with very accurate or Metropolis–Hastings corrected samples for a few iterations, in order not to artificially reduce the prior's strength.




(a) Ground truth (b) Noisy data (c) Posterior mean




(d) Sample mean, $\tilde{\varepsilon} = 10^{0}$ (e) Sample mean, $\tilde{\varepsilon} = 10^{-2}$ (f) Sample mean, $\tilde{\varepsilon} = 10^{-4}$




(g) Logarithm of pixelwise sample variance, $\tilde{\varepsilon} = 10^{0}$ (h) Logarithm of pixelwise sample variance, $\tilde{\varepsilon} = 10^{-2}$ (i) Logarithm of pixelwise sample variance, $\tilde{\varepsilon} = 10^{-4}$

Figure 7.3.: TV-denoising of an image (a) corrupted by pixelwise independent Gaussian noise (b). A close approximation of the MMSE, computed with a long-running, Metropolis–Hastings corrected Markov chain is shown in (c). We show MMSE estimates (d-f) and logarithm of pixelwise standard deviation (g-i) of $10^4$ samples for different proximal accuracy levels $\varepsilon = C_0\tilde{\varepsilon}$.

# Chapter 8

# Langevin Monte Carlo: Ergodicity and Convergence Analysis

In this section, we derive stability and convergence results for different Langevin sampling variants. By "stability", we usually mean ergodicity and specifically geometric ergodicity, for other possible notions of Markov chain stability see [150]. By "convergence", we refer to the distance between the samples' and/or stationary distributions of any algorithm to the target $\mu^*$. As we discussed in Section 7.1, $\mu^*$ and the stationary distribution of Langevin Monte Carlo do not coincide due to the unbalanced discretization.

Some arguments are repetitive for the different algorithms, so we start by giving a the simplest coupling-based convergence proof of ULA in Section 8.1. The section unveils conceptual parallels—firstly, between sampling and optimization from a convex analysis viewpoint, and secondly, between sampling algorithms in discrete time and their continuous time limits. We then revisit these ideas for the more complex algorithms: In Section 8.2, we see that ergodicity statements are easily generalized from ULA to other discretizations of overdamped Langevin diffusion, employing typical convex analysis arguments. For the (F*flow*B) algorithm, we analyze an existing convergence proof from the literature and its assumptions and limitations in detail. We then modify the argument slightly to obtain a new convergence result for (FB*flow*). In Section 8.3, we introduce a recent primal-dual Langevin sampling method and analyze the bias induced by the partial dualization of the potential. We then move to mirror Langevin sampling in Section 8.4, where we again generalize the ergodicity result, since the mirror map changes the underlying geometry of the state space. Here, we use recent theory on the optimality of preconditioners in order to motivate sampling schemes for imaging problems with Poisson data.

## 8.1. Unadjusted Langevin Algorithm

A simple, yet versatile tool to prove convergence results for the Langevin diffusion process and its discretizations is the synchronous coupling method. Using that, ergodicity can be proven by considering two processes or chains with the same stochastic input. It is a standard argument for the analysis of Markov chains and stochastic stability [133, 150] and more recently also diffusion PDE research [82] and has proven efficient in the ergodicity and convergence analysis of various Langevin sampling algorithms. For Langevin diffusion based sampling, the relevant techniques were developed over a series of works that concentrated on ULA with a strongly convex potential [68, 69, 74, 75]. These arguments were subsequently extended to projection-based variants of ULA for distributions over constrained sets [39, 40], mirror-type methods [2] and recently on a primal-dual type method [43] and the F*flow*B algorithm [84]. We already demonstrated its use for overdamped Langevin diffusion in continuous time in Theorem 6.18. It turns out that these arguments can be transferred to the time discretized setting without many changes, leading to "typical" estimates and techniques known from convex analysis.

**Theorem 8.1: ULA: geometric ergodicity by coupling**

Suppose $V$ is $\omega$-strongly convex, differentiable and $\nabla V$ is $L$-Lipschitz continuous. Let $(X^{(n)})_{n\in\mathbb{N}_0}$ be generated by (ULA) with some initialization $X^{(0)} \sim \mu^{(0)}$ and step size $\tau < \frac{2}{L}$. Then ULA has a stationary distribution, denoted as before by $\mu^{\text{stat}}_{\text{ULA}}$, and the distributions $\mu^{(n)}$ of $X^{(n)}$ satisfy

$$\mathcal{W}_2(\mu^{(n)}, \mu^{\text{stat}}_{\text{ULA}}) \le q(\tau)^n \mathcal{W}_2(\mu^{(0)}, \mu^{\text{stat}}_{\text{ULA}}), \tag{8.1}$$

where $q(\tau) = \max\{1 - \omega\tau, L\tau - 1\}$.

*Proof.* Most of the computations below are similar to Lemma 1 in [68], where it was then used to prove convergence of ULA. Let us consider two instances of ULA with the same stochastic input, i.e. $(X^{(n)})_{n\ge0}$ and $(\tilde{X}^{(n)})_{n\ge0}$ defined by initial values $X^{(0)}$ and $\tilde{X}^{(0)}$ and the coupled update

$$\begin{aligned}
\xi^{(n+1)} &\sim \mathrm{N}(0, I),\\
X^{(n+1)} &= X^{(n)} - \tau\nabla V(X^{(n)}) + \sqrt{2\tau}\xi^{(n+1)},\\
\tilde{X}^{(n+1)} &= \tilde{X}^{(n)} - \tau\nabla V(\tilde{X}^{(n)}) + \sqrt{2\tau}\xi^{(n+1)}.
\end{aligned}$$

We now compute

$$\begin{aligned}
&\|X^{(n+1)} - \tilde{X}^{(n+1)}\|^2\\
&= \|X^{(n)} - \tau\nabla V(X^{(n)}) - \tilde{X}^{(n)} + \tau\nabla V(\tilde{X}^{(n)})\|^2\\
&= \|X^{(n)} - \tilde{X}^{n}\|^2 - 2\tau\, \mathrm{D}_V^{\text{sym}}(X^{(n)}, \tilde{X}^{(n)}) + \tau^2\|\nabla V(X^{(n)}) - \nabla V(\tilde{X}^{(n)})\|^2.
\end{aligned}$$

Using (B.13) to lower bound the symmetric Bregman divergence, we obtain

$$\begin{aligned}&\|X^{(n+1)} - \tilde{X}^{(n+1)}\|^2 \\ &\quad\leq \left(1 - \frac{2\tau\omega L}{L+\omega}\right)\|X^{(n)} - \tilde{X}^n\|^2 + \tau\left(\tau - \frac{2}{L+\omega}\right)\|\nabla V(X^{(n)}) - \nabla V(\tilde{X}^{(n)})\|^2.\end{aligned}$$

If $\tau \leq \frac{2}{L+\omega}$, we can use (B.11) and (B.7) in that order to upper bound the second term:

$$\begin{aligned}\|X^{(n+1)} - \tilde{X}^{(n+1)}\|^2 &\leq \left(1 - \frac{2\tau\omega L}{L+\omega} + \omega^2\tau\left(\tau - \frac{2}{L+\omega}\right)\right)\|X^{(n)} - \tilde{X}^n\|^2 \\ &= (1-\tau\omega)^2\|X^{(n)} - \tilde{X}^n\|^2.\end{aligned}$$

Conversely, if $\tau \geq \frac{2}{L+\omega}$, then we use (B.10) and (B.9) in that order and obtain

$$\begin{aligned}\|X^{(n+1)} - \tilde{X}^{(n+1)}\|^2 &\leq \left(1 - \frac{2\tau\omega L}{L+\omega} + L^2\tau\left(\tau - \frac{2}{L+\omega}\right)\right)\|X^{(n)} - \tilde{X}^n\|^2 \\ &= (\tau L - 1)^2\|X^{(n)} - \tilde{X}^n\|^2.\end{aligned}$$

Note that $\tau \leq \frac{2}{L+\omega}$ is equivalent to $1 - \tau\omega \geq \tau L - 1$, so the factor in front of the norm can unifyingly be written as $q(\tau)^2$, where $q(\tau) = \max\{1-\omega\tau, L\tau - 1\}$.

$$\|X^{(n+1)} - \tilde{X}^{(n+1)}\|^2 \leq q(\tau)^2\|X^{(n)} - \tilde{X}^n\|^2.$$

By unrolling that inequality over $n$ iterations, we obtain

$$\|X^{(n)} - \tilde{X}^{(n)}\|^2 \leq q(\tau)^{2n}\|X^{(0)} - \tilde{X}^0\|^2.$$

Since $\omega \leq L$, it is easy to see that $q(\tau) > 0$. The step size condition $\tau \leq \frac{2}{L}$ is equivalent to $q(\tau) < 1$, in which case the Markov chain is ergodic due to the right hand side converging to zero. Taking expectations, the left hand side is lower bounded by $\mathcal{W}_2(\mu^{(n)}, \tilde{\mu}^{(n)})$ by definition of the Wasserstein distance. On the right hand side, we proceed as in Theorem 6.18: Since the inequality holds for any initialization of $X^{(0)}$ and $\tilde{X}^{(0)}$, they hold in particular for the specific coupling for which the joint distribution of $(X^{(0)}, \tilde{X}^{(0)})$ realizes the infimum in the definition of the Wasserstein distance. Taking the square root, we have

$$\mathcal{W}_2(\mu^{(n)}, \tilde{\mu}^{(n)}) \leq q(\tau)^n\mathcal{W}_2(\mu^{(0)}, \tilde{\mu}^{(0)}).$$

Now, as in the continuous time case, we assume that $\tilde{\mu}^{(0)} = \mu^{\text{stat}}_{\text{ULA}}$ is initialized at the stationary distribution of the Markov chain. This implies that it remains stationary throughout the iteration, so $\tilde{\mu}^{(n)} = \mu^{\text{stat}}_{\text{ULA}}$ for all $n \geq 1$, proving the statement. □

Since this ergodicity proof technique for ULA is central to several of the methods we consider, it is helpful to put it into context by comparing it to gradient descent in the optimization case, and to overdamped Langevin diffusion as the continuous time equivalent:

- The proof shows that the coupling technique essentially allows us to transfer typical inequalities from optimization, like a descent lemma or the stronger, lower bound (B.13) on the symmetric Bregman divergence, to the sampling case. This "transfer" is possible whenever we prove stability (or, in Markov chain terms, ergodicity) in a transport distance based on a norm that only depends on the difference of the coupled states, just like the Wasserstein distance. Whenever this is the case, the stochastic terms of the coupled ULA iterations cancel out as in the beginning of the proof above. The differences are then equal to the differences of two parallel instances of gradient descent. Indeed, the computations from there are exactly the same computations that we would have carried out in a stability proof of gradient descent on the functional $V$.

- The other type of conceptual transferability of the coupling method connects the continuous time and discrete time regimes. The central argument to remove the influence of diffusion terms is essentially the same in both: In discrete time, the additive stochastic term got cancelled out by considering the difference $X - \tilde{X}$ of coupled iterations. In continuous time, the diffusion in two coupled processes only acts in the dimension $X + \tilde{X}$, so making the transport distance depend only on the orthogonal difference equally cancelled out the diffusion term in the proof of Theorem 6.18.

While the last proof stressed the similarities in the concepts of stability of ULA and overdamped Langevin diffusion, the discretized time case is to some extent harder: The stability argument only results in ergodicity, and not convergence to the target $\mu^*$. Theorem 8.1 proves a rate that allows to compute the number of iterations needed before the samples' distribution is close to the stationary distribution $\mu_{\text{ULA}}^{\text{stat}}$ of the Markov chain. Since we actually aim to generate samples from $\mu^* \propto e^{-V}$, it remains to bound the distance between $\mu_{\text{ULA}}^{\text{stat}}$ and $\mu^*$ (or between $\mu^{(n)}$ and $\mu^*$ directly).

In order to complete that missing step, we can again draw parallels to the optimization case: Just as we would be able to prove a consistency result that bounds the error of gradient descent compared to the gradient flow $\dot{x}_t = -\nabla V(x_t)$, we can bound the error that arises from discretizing the SDE $\dot{X}_t = -\nabla V(X_t) + \sqrt{2}\dot{W}_t$. While forward Euler for ODEs is a first order method, in the SDE case, the discretization error will be of order $\mathcal{O}(\tau^{\frac{1}{2}})$ at every time step. Due to the unbalanced discretization eluded to in Section 7.1, the errors accumulate and generate a bias between the stationary distributions in continuous and discrete time. As before, the proof can be based on a synchronous coupling argument, this time between continuous and discrete time regimes. We mostly follow the proof of [68], but mention that there are other routes leading to slightly different bounds of the bias in potentially different distance metrics, see, e.g., [74, 73, 56, 210, 60].

**Theorem 8.2: ULA: Wasserstein bias by coupling**

Suppose $V$ is $\omega$-strongly convex, differentiable and $\nabla V$ is $L$-Lipschitz continuous. Let $\tau < \frac{2}{L}$, then by Theorem 8.1, ULA is geometrically ergodic. Then, the distributions $\mu^{(n)}$ generated by ULA satisfy

$$\mathcal{W}_2(\mu^{(n)}, \mu^*) \leq q(\tau)^n \mathcal{W}_2(\mu^{(0)}, \mu^*) + \frac{C\left(\tau^3 L^2 d\right)^{\frac{1}{2}}\left(1 - q(\tau)^n\right)}{1 - q(\tau)}, \tag{8.2}$$

with $q(\tau)$ as in Theorem 8.1 and $C = 1 + \sqrt{2/3}$. In the limit $n \to \infty$, the stationary $\mu^{\text{stat}}_{\text{ULA}}$ satisfies

$$\mathcal{W}_2(\mu^{\text{stat}}_{\text{ULA}}, \mu^*) \leq \begin{cases} C\frac{L}{\omega}(\tau d)^{\frac{1}{2}} & \text{if } 0 < \tau \leq \frac{2}{L+\omega}, \\ C\frac{L\tau}{2-L\tau}(\tau d)^{\frac{1}{2}} & \text{if } \frac{2}{L+\omega} \leq \tau < \frac{2}{L}. \end{cases} \tag{8.3}$$

*Proof.* The standard way to couple the time discretized Markov chain produced by ULA with a diffusion process in continuous time is by understanding the Markov chain as a time continuous object, where we are only interested in the states at times $n\tau$, $n \in \mathbb{N}_0$. Suppose $(X_t)_{t\geq 0}$ solves the overdamped Langevin diffusion (6.47) with $X_0 \sim \mu^*$. By integrating the SDE over $[t, t+\tau]$ for any $t, \tau > 0$, it holds

$$X_{t+\tau} = X_t - \int_t^{t+\tau} \nabla V(X_s)\,\mathrm{d}s + \sqrt{2}(W_{t+\tau} - W_t),$$

and, since the process was initialized at the stationary solution, $X_t \sim \mu^*$ for all times $t \geq 0$. We now set $t_n = n\tau$ and $\xi^{(n+1)} = \tau^{-\frac{1}{2}}(W_{t_n+\tau} - W_{t_n}) \sim \mathrm{N}(0, I)$. Let now $(\tilde{X}^{(n)})_{n\in\mathbb{N}_0}$ be generated by (ULA) with this stochastic input $(\xi^{(n)})_{n\in\mathbb{N}}$. For ease of notation, fix now some $n \in \mathbb{N}_0$ and write $t$ instead of $t_n$. It follows

$$\begin{aligned} &\|X_{t+\tau} - \tilde{X}^{(n+1)}\| \\ &\quad = \left\| X_t - \int_t^{t+\tau} \nabla V(X_s)\,\mathrm{d}s - \tilde{X}^{(n)} + \tau \nabla V(\tilde{X}^{(n)}) \right\| \\ &\quad \leq \left\| X_t - \tilde{X}^{(n)} - \tau(\nabla V(X_t) - \nabla V(\tilde{X}^{(n)})) \right\| + \left\| \int_t^{t+\tau} (\nabla V(X_t) - \nabla V(X_s))\,\mathrm{d}s \right\|. \end{aligned}$$

For the first term, we can use the same type of descent lemma that we derived in the proof of the ergodicity result Theorem 8.1, and obtain

$$\left\| X_t - \tilde{X}^{(n)} - \tau(\nabla V(X_t) - \nabla V(\tilde{X}^{(n)})) \right\| \leq q(\tau) \left\| X_t - \tilde{X}^{(n)} \right\|.$$

Taking a weighted $L^2$-norm on both sides, this implies

$$\begin{aligned}&\left(\mathbb{E}[\|X_{t+\tau} - \tilde{X}^{(n+1)}\|^2]\right)^{\frac{1}{2}} \\ &\quad \leq q(\tau)\left(\mathbb{E}[\|X_t - \tilde{X}^{(n)}\|^2]\right)^{\frac{1}{2}} + \left(\mathbb{E}[\|\int_t^{t+\tau} (\nabla V(X_t) - \nabla V(X_s))\,\mathrm{d}s\|^2]\right)^{\frac{1}{2}}.\end{aligned}$$

For the second term, we use Jensen's inequality and Lipschitz continuity of $\nabla V$ to bound

$$\begin{aligned}\left(\mathbb{E}[\|\int_t^{t+\tau} (\nabla V(X_t) - \nabla V(X_s))\,\mathrm{d}s\|^2]\right)^{\frac{1}{2}} &\leq \left(\tau\,\mathbb{E}[\int_t^{t+\tau} \|\nabla V(X_t) - \nabla V(X_s)\|^2\,\mathrm{d}s]\right)^{\frac{1}{2}} \\ &\leq \left(\tau L^2 \int_t^{t+\tau} \mathbb{E}[\|X_t - X_s\|^2]\,\mathrm{d}s\right)^{\frac{1}{2}}\end{aligned}$$

For any $s \geq t$, we plug in the integrated SDE $X_t - X_s = \int_t^s \nabla V(X_r)\,\mathrm{d}r - \sqrt{2}(W_s - W_t)$. Using the triangle inequality, Jensen's inequality and the fact that $\mathbb{E}[\|W_s - W_t\|^2] = s - t$, we obtain the bound

$$\begin{aligned}&\left(\tau L^2 \int_t^{t+\tau} \mathbb{E}[\|X_t - X_s\|^2]\,\mathrm{d}s\right)^{\frac{1}{2}} \\ &\quad \leq \left(\tau L^2 \int_t^{t+\tau} \mathbb{E}[\|\textstyle\int_t^s \nabla V(X_r)\,\mathrm{d}r\|^2]\,\mathrm{d}s\right)^{\frac{1}{2}} + \left(2\tau L^2 \int_t^{t+\tau} \mathbb{E}[\|W_s - W_t\|^2]\,\mathrm{d}s\right)^{\frac{1}{2}} \\ &\quad \leq \left(\frac{\tau^4 L^2}{3}\,\mathbb{E}[\|\nabla V(X_t)\|^2]\right)^{\frac{1}{2}} + \left(\tau^3 d L^2\right)^{\frac{1}{2}}.\end{aligned}$$

In the last step, we have additionally employed the fact that $X_t$ is stationary in distribution and we therefore know $\mathbb{E}[\|\nabla V(X_r)\|^2] = \mathbb{E}[\|\nabla V(X_t)\|^2]$ for all $r \in [t, t+\tau]$. Since $\nabla V$ is Lipschitz continuous, we can bound this term by $\mathbb{E}[\|\nabla V(X_t)\|^2] \leq dL$. Using the step size condition $\tau \leq 2L^{-1}$, the bound eventually reduces to

$$\left(\tau L^2 \int_t^{t+\tau} \mathbb{E}[\|X_t - X_s\|^2]\,\mathrm{d}s\right)^{\frac{1}{2}} \leq \left(\frac{1}{3}\tau^4 L^3 d\right)^{\frac{1}{2}} + \left(\tau^3 d L^2\right)^{\frac{1}{2}} \leq C\left(\tau^3 L^2 d\right)^{\frac{1}{2}},$$

where $C = 1 + \sqrt{2/3}$. After plugging this bound in, we proceed as in the ergodicity result. The nonasymptotic bound is iterated over $n$ steps, giving

$$\begin{aligned}\left(\mathbb{E}[\|X_t - \tilde{X}^{(n)}\|^2]\right)^{\frac{1}{2}} &\leq q(\tau)^n \left(\mathbb{E}[\|X_0 - \tilde{X}^{(0)}\|^2]\right)^{\frac{1}{2}} + C\left(\tau^3 L^2 d\right)^{\frac{1}{2}} \sum_{k=0}^{n-1} q(\tau)^k \\ &\leq q(\tau)^n \left(\mathbb{E}[\|X_0 - \tilde{X}^{(0)}\|^2]\right)^{\frac{1}{2}} + \frac{C\left(\tau^3 L^2 d\right)^{\frac{1}{2}} (1 - q(\tau)^n)}{1 - q(\tau)}.\end{aligned}$$

By definition of the Wasserstein distance, the left side is lower bounded by $\mathcal{W}_2(\mu^*, \mu_{\mathrm{ULA}}^{\mathrm{stat}})$ for all $n$, since $X_t$ and $\tilde{X}^{(n)}$ are stationary in distribution. The inequality holds for any

joint distribution of $(X_0, X^{(0)})$ with marginals $\mu^*$ and $\mu^{(0)}$, respectively. In particular, it holds for the coupling that realizes the Wasserstein distance, leading to the first statement (8.2). Considering the limit $n \to \infty$ and plugging in the definition of $q(\tau)$ leads to the second inequality (8.3). □

We close this section by mentioning that there are several more relevant results on the convergence behaviour of ULA. For instance, we have seen in Section 6.3 that $\omega$-strong convexity of the potential $V$ is sufficient, but not necessary to guarantee convergence of overdamped Langevin diffusion. This extends to the discretized time setting of ULA, which may also converge under the assumption that $V$ satisfies the weaker log-Sobolev inequality and has a Lipschitz-continuous gradient. Convergence in KL divergence under this weaker set of assumption was shown in [210] and recently extended to general Rényi divergences in [60].

## 8.2. Methods Based on Proximal Steps

Having gone through the central arguments on the example of ULA, it is now easy to show ergodicity for several other discretization schemes. We start this section by proving geometric ergodicity of (F*flow*B) and (FB*flow*) using a coupling argument. Afterwards, for the (biased) convergence of these algorithms to the target $\mu^*$, we however illustrate a different proof strategy due to [73]. The authors there derived a convergence proof for (ULA) slightly differing from the one we presented in the last section, and extended the argument to the (F*flow*B) algorithm. As a main contribution of this section, we will show that some of the arguments can be transferred in order to prove convergence of the (FB*flow*) method. Interestingly, although the two convergence proofs do differ to some extent, the asymptotic bounds on the Wasserstein biases of both methods coincide up to order $\tau$ (the leading order is, as before, $\tau^{\frac{1}{2}}$).

**Theorem 8.3: (F*flow*B) and (FB*flow*): geometric ergodicity**

Suppose $V = F + G : \mathbb{R}^d \to \bar{\mathbb{R}}$, where $F : \mathbb{R}^d \to \mathbb{R}$ is $\omega_F$-strongly convex, has $L_{\nabla F}$-Lipschitz continuous gradient $\nabla F$, and $G : \mathbb{R}^d \to \bar{\mathbb{R}}$ is proper, lower semi-continuous and convex, but potentially not differentiable. Note that we explicitly allow $G$ to take inifinite value. Assume the step size condition $\tau < \frac{2}{L_{\nabla F}}$ holds for both (F*flow*B) and (FB*flow*). Then both algorithms are geometrically ergodic in the sense that

$$\mathcal{W}_2(\mu^{(n)}_{\mathrm{F}f\mathrm{B}}, \mu^{\mathrm{stat}}_{\mathrm{F}f\mathrm{B}}) \leq q_F(\tau)^n \mathcal{W}_2(\mu^{(0)}_{\mathrm{F}f\mathrm{B}}, \mu^{\mathrm{stat}}_{\mathrm{F}f\mathrm{B}})$$
$$\mathcal{W}_2(\mu^{(n)}_{\mathrm{FB}f}, \mu^{\mathrm{stat}}_{\mathrm{FB}f}) \leq q_F(\tau)^n \mathcal{W}_2(\mu^{(0)}_{\mathrm{FB}f}, \mu^{\mathrm{stat}}_{\mathrm{FB}f}),$$

where $\mu_{\mathrm{F}f\mathrm{B}}^{(n)}$ denotes the distribution of (F*flow*B) after $n$ steps, $\mu_{\mathrm{F}f\mathrm{B}}^{\mathrm{stat}}$ denotes the stationary distribution, and analogous notation is used for (FB*flow*) with 'FB$f$'. Similar to Theorem 8.1, we have defined $q_F(\tau) = \max\{1 - \omega_F\tau, L_{\nabla F}\tau - 1\}$.

*Proof.* Consider two synchronously coupled instances of (F*flow*B), i.e., $(X^{(n)})_{n\in\mathbb{N}_0}$ and $(\tilde{X}^{(n)})_{n\in\mathbb{N}_0}$ defined by initial values $X^{(0)}$ and $\tilde{X}^{(0)}$ and the coupled update

$$\begin{aligned}
\xi^{(n+1)} &\sim \mathrm{N}(0, I),\\
X^{(n+1)} &= \mathrm{prox}_G^\tau\left(X^{(n)} - \tau\nabla F(X^{(n)}) + \sqrt{2\tau}\xi^{(n+1)}\right),\\
\tilde{X}^{(n+1)} &= \mathrm{prox}_G^\tau\left(\tilde{X}^{(n)} - \tau\nabla V(\tilde{X}^{(n)}) + \sqrt{2\tau}\xi^{(n+1)}\right).
\end{aligned}$$

Since $G$ is convex, its proximal mapping $\mathrm{prox}_G^\tau$ is firmly non-expansive for all $\tau > 0$, see [20, chap. 23]. In particular, we have

$$\begin{aligned}
\|X^{(n+1)} - \tilde{X}^{(n+1)}\|^2 &= \|\mathrm{prox}_G^\tau\left(X^{(n+\frac{2}{3})}\right) - \mathrm{prox}_G^\tau\left(\tilde{X}^{(n+\frac{2}{3})}\right)\|^2\\
&\leq \|X^{(n+\frac{2}{3})} - \tilde{X}^{(n+\frac{2}{3})}\|^2\\
&= \|X^{(n)} - \tau\nabla F(X^{(n)}) - \tilde{X}^{(n)} + \tau\nabla F(\tilde{X}^{(n)})\|^2, \qquad (8.4)
\end{aligned}$$

where we denoted $X^{(n+\frac{2}{3})} = X^{(n)} - \tau\nabla F(X^{(n)}) + \sqrt{2\tau}\mu^{(n+1)}$. The exact same term was the one that we bounded in the ergodicity proof of ULA in Theorem 8.1, with $F$ taking the place of $V$ under identitical assumptions. We can therefore copy the steps and obtain

$$\|X^{(n+1)} - \tilde{X}^{(n+1)}\|^2 \leq q_F(\tau)^2\|X^{(n)} - \tilde{X}^n\|^2.$$

From there, the steps to geometric ergodicity are equivalent to the proof of Theorem 8.1.

For (FB*flow*), the proof is almost identical. The only difference appears in (8.4): After coupling two iterations, the stochastic terms cancel out right away by considering the difference $X^{(n+1)} - \tilde{X}^{(n+1)}$. We then use firm non-expansivity to upper bound $\|\mathrm{prox}_G^\tau(X^{(n+\frac{1}{3})}) - \mathrm{prox}_G^\tau(\tilde{X}^{(n+\frac{1}{3})})\|^2$ upon setting $X^{(n+\frac{1}{3})} = X^{(n)} - \tau\nabla F(X^{(n)})$ and end up with the same term as in (8.4). From there, we carry on as before. □

As in the case of ULA, the convergence proofs require a little more work, since one needs to bound the SDE discretization error. The proof we present now is not directly based on a synchronous coupling like the ergodicity result, but rather focused on the perspective of viewing sampling as an optimization problem that we presented in Chapter 6. From this point of view, convergence with a bounded bias can be shown by quantifying the decay of the objective functional (the KL divergence with respect to $\mu^*$) along iterations. We loosely sketch the proof idea here for ULA, before going through the details on the more compliated example of (F*flow*B). Recall that we can separate the KL objective

into inner and potential energy as

$$\mathcal{D}_{\mathrm{KL}}(\mu \,\|\, \mu^*) = \mathcal{H}(\mu) + \mathcal{E}_V(\mu).$$

Writing ULA again as two half-steps

$$\begin{aligned} X^{(n+\frac{1}{2})} &= X^{(n)} - \tau \nabla V(X^{(n)}) \\ X^{(n+1)} &= X^{(n+\frac{1}{2})} + \sqrt{2\tau}\xi^{(n+1)}, \quad \xi^{(n+1)} \sim \mathrm{N}(0, I), \end{aligned}$$

now helps to derive bounds. Since the first half step $\mu^{(n)} \to \mu^{(n+\frac{1}{2})}$ is a gradient descent step, it is easy to show that it reduces the potential energy $\mathcal{E}_V$ by a standard descent lemma. Likewise, the diffusion step $\mu^{(n+\frac{1}{2})} \to \mu^{(n+1)}$ increases the entropy. This together gives a certain decay of the objective in every full step. The "error", i.e., the loss in entropy due to the gradient step or vice versa, is shown to remain bounded, resulting in a bounded bias over the course of the iteration.

We now give the complete result for (F*flow*B).

**Theorem 8.4: (F*flow*B): Wasserstein bias**

Suppose $V = F + G : \mathbb{R}^d \to \mathbb{R}$, where $F$ is $\omega_F$-strongly convex, differentiable and $\nabla F$ is $L_{\nabla F}$-Lipschitz continuous. Suppose further $G$ is convex and $L_G$-Lipschitz-continuous. Now assume that (F*flow*B) generates samples with distributions $(\mu^{(n)})_{n\in\mathbb{N}_0}$, where the step size obeys $\tau \leq \frac{1}{L_{\nabla F}}$. Then it holds

$$\mathcal{W}_2^2(\mu^{(n)}, \mu^*) \leq (1 - \omega_F \tau)^n \mathcal{W}_2^2(\mu^{(0)}, \mu^*) + C\tau,$$

with a constant $C = \frac{2}{\omega_F}(L_{\nabla F} d + L_G^2)$. In particular, in the limit $n \to \infty$, we have

$$\mathcal{W}_2(\mu^{\mathrm{stat}}_{\mathrm{F}fB}, \mu^*) \leq (C\tau)^{\frac{1}{2}}. \tag{8.6}$$

In order to state the proof, we collect several auxiliary results. The first is due to [73, Lemma 4] and quantifies the decrease in potential $\mathcal{E}_F$ in a forward step.

**Lemma 8.5: Potential decrease in a gradient step**

Assume $F$ is $\omega_F$-strongly convex, differentiable and $\nabla F$ is $L_{\nabla F}$-Lipschitz continuous. Let $\tau \leq \frac{1}{L_{\nabla F}}$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and denote further $\hat{\mu} = (\mathrm{Id} - \tau\nabla F)_{\#}\mu$. Then it holds

$$\begin{aligned} 2\tau(\mathcal{E}_F(\hat{\mu}) - \mathcal{E}_F(\nu)) \leq & (1 - \omega_F\tau)\mathcal{W}_2^2(\mu, \nu) - \mathcal{W}_2^2(\hat{\mu}, \nu) \\ & - \gamma^2(1 - \gamma L_{\nabla F})\, \mathbb{E}_{X\sim\mu}[\|\nabla U(X)\|^2]. \end{aligned}$$

The next result regards the increase in entropy in every diffusion step [73, Lemma 5]

**Lemma 8.6: Entropy increase in a diffusion step**

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mathcal{H}(\nu) < \infty$. For a fixed $\tau > 0$, denote by $\hat{\mu}$ the solution of the heat flow at time $\tau$ with initial condition $\mu$ at time zero, see (7.7) for the density of $\hat{\mu}$. Then it holds

$$2\tau(\mathcal{H}(\hat{\mu}) - \mathcal{H}(\nu)) \leq \mathcal{W}_2^2(\mu, \nu) - \mathcal{W}_2^2(\hat{\mu}, \nu).$$

For every diffusion step, we can further bound the corresponding change in potential energy [73, Lemma 3].

**Lemma 8.7: Potential increase in a diffusion step**

Suppose $F$ is differentiable and $\nabla F$ is $L_{\nabla F}$-Lipschitz continuous. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\tau > 0$, denote by $\hat{\mu}$ the solution of the heat flow after a time step of length $\tau$ with initial condition $\mu$, exactly as in Lemma 8.6. Then

$$\mathcal{E}_F(\hat{\mu}) - \mathcal{E}_F(\mu) \leq L_{\nabla F} d\tau.$$

Eventually, we control the potential energy $\mathcal{E}_G$ in every backward step [73, Lemma 29].

**Lemma 8.8: Potential decrease in a proximal step**

Suppose $G$ is convex and $L_G$-Lipschitz continuous. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and denote $\hat{\mu} = (\mathrm{prox}_G^\tau)_{\#}\mu$. Then it holds

$$2\tau(\mathcal{E}_G(\mu) - \mathcal{E}_G(\nu)) \leq \mathcal{W}_2^2(\mu, \nu) - \mathcal{W}_2^2(\hat{\mu}, \nu) + 2L_G^2\tau^2.$$

With these auxiliary results, assembling the proof of Theorem 8.4 reduces to applying the lemmata at the correct intermediate steps of the algorithm, see [73].

*Proof of Theorem 8.4.* For ease of notation, write the iteration (F*flow*B) as three substeps, i.e.,

$$\begin{aligned} X^{(n+\frac{1}{3})} &= X^{(n)} - \tau\nabla V(X^{(n)}), \\ X^{(n+\frac{2}{3})} &= X^{(n+\frac{1}{3})} + \sqrt{2\tau}\xi^{(n+1)}, \quad \xi^{(n+1)} \sim \mathrm{N}(0, I) \\ X^{(n+1)} &= \mathrm{prox}_G^\tau(X^{(n+\frac{2}{3})}). \end{aligned}$$

The corresponding distributions at intermediate steps are also denoted as $\mu^{(n)}, \mu^{(n+\frac{1}{3})}$, etc. We first show the following, nonasymptotic result, see [73, Lemma 30]:

$$2\tau\, \mathcal{D}_{\mathrm{KL}}(\mu^{(n+\frac{2}{3})} \,\|\, \mu^*) \leq (1-\omega_F\tau)\mathcal{W}_2^2(\mu^{(n)}, \mu^*) - \mathcal{W}_2^2(\mu^{(n+1)}, \mu^*) + 2\tau^2(L_{\nabla F} d + L_G^2). \quad (8.8)$$

This follows from combining the four auxiliary results. We use Lemma 8.5 with $\mu = \mu^{(n)}$ and $\nu = \mu^*$ to obtain

$$2\tau(\mathcal{E}_F(\mu^{(n+\frac{1}{3})}) - \mathcal{E}_F(\mu^*)) \leq (1-\omega_F\tau)\mathcal{W}_2^2(\mu^{(n)}, \mu^*) - \mathcal{W}_2^2(\mu^{(n+\frac{1}{3})}, \mu^*).$$

We apply Lemma 8.6 with $\mu = \mu^{(n+\frac{1}{3})}$, $\nu = \mu^*$ and get

$$2\tau(\mathcal{H}(\mu^{(n+\frac{2}{3})}) - \mathcal{H}(\mu^*)) \leq \mathcal{W}_2^2(\mu^{(n+\frac{1}{3})}, \mu^*) - \mathcal{W}_2^2(\mu^{(n+\frac{2}{3})}, \mu^*).$$

For Lemma 8.7, setting $\mu = \mu^{(n+\frac{1}{3})}$ as well gives

$$2\tau(\mathcal{E}_F(\mu^{(n+\frac{2}{3})}) - \mathcal{E}_F(\mu^{(n+\frac{1}{3})})) \leq 2L_{\nabla F} d\tau^2.$$

And eventually, we set $\mu = \mu^{(n+\frac{2}{3})}$, $\nu = \mu^*$ in Lemma 8.8 to obtain

$$2\tau(\mathcal{E}_G(\mu^{(n+\frac{2}{3})}) - \mathcal{E}_G(\mu^*)) \leq \mathcal{W}_2^2(\mu^{(n+\frac{2}{3})}, \mu^*) - \mathcal{W}_2^2(\mu^{(n+1)}, \mu^*) + 2L_G^2\tau^2.$$

The central inequality (8.8) follows from adding up the last four inequalities and using $\mathcal{D}_{\mathrm{KL}}(\mu^* \,\|\, \mu^*) = 0$. Denote $\tilde{C} = 2(L_{\nabla F} d + L_G^2)$. Since the KL divergence is nonnegative, we can bound the left side of (8.8) by zero and iterate the inequality over $n$ steps to obtain

$$\begin{aligned}\mathcal{W}_2^2(\mu^{(n)}, \mu^*) &\leq (1-\omega_F\tau)^n \mathcal{W}_2^2(\mu^{(0)}, \mu^*) + \tilde{C}\tau^2 \sum_{k=0}^{n-1} (1-\omega_F\tau)^k \\ &\leq (1-\omega_F\tau)^n \mathcal{W}_2^2(\mu^{(0)}, \mu^*) + C\tau,\end{aligned}$$

where $C = \omega_F^{-1}\tilde{C}$ is the constant in the theorem's statement. □

While we just bounded the KL divergence terms in the proof by zero, note that it would be possible to keep them in the iterated inequality, use convexity of KL in order to bound the ergodic averages and hence obtain convergence results in KL divergence from the target as well. Since these computations are not central to our work, we do not carry out the detailed computations here and refer instead to [73].

We continue by extending the theory to the slightly different (FB*flow*) algorithm. As should be clear from the previous computations, several of the auxiliary results can be recycled when adjusting their order accordingly. The central convergence result is the following.

**Theorem 8.9: (FB*flow*): Wasserstein bias**

Suppose $V = F + G : \mathbb{R}^d \to \mathbb{R}$, where $F$ is $\omega_F$-strongly convex, differentiable and $\nabla F$ is $L_{\nabla F}$-Lipschitz continuous. Suppose further $G$ is convex and $L_G$-Lipschitz-continuous. For a step size $\tau < \frac{1}{L_{\nabla F}}$, let (FB*flow*) generate samples with distributions $(\mu^{(n)})_{n\in\mathbb{N}_0}$. Then it holds

$$\mathcal{W}_2(\mu^{\text{stat}}_{\text{FB}f}, \mu^*) \leq (C\tau)^{\frac{1}{2}} + \mathcal{O}(\tau)$$

where $C = \frac{2}{\omega_F}(L_G^2 + dL_{\nabla F})$ is the same constant as in Theorem 8.4. After $n$ iterations, by applying the triangle inequality and Theorem 8.3, we obtain

$$\mathcal{W}_2(\mu^{(n)}, \mu^*) \leq (1 - \omega_F\tau)^n \mathcal{W}_2(\mu^{(0)}, \mu^{\text{stat}}_{\text{FB}f}) + (C\tau)^{\frac{1}{2}} + \mathcal{O}(\tau).$$

Since the order of the intermediate steps is modified compared to (F*flow*B), we have to prove a simple auxiliary result on the entropy change under a gradient descent step.

**Lemma 8.10: Entropy decrease in a gradient step**

Suppose $F$ is $\omega_F$-strongly convex, differentiable and $\nabla F$ is $L_{\nabla F}$-Lipschitz continuous. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and define $\hat{\mu} = (\mathrm{Id} - \tau\nabla F)_{\#}\mu$, where $\tau < \frac{1}{L_{\nabla F}}$. Then it holds

$$d\log(1 - \tau L_{\nabla F}) \leq \mathcal{H}(\mu) - \mathcal{H}(\hat{\mu}) \leq d\log(1 - \tau\omega_F) < 0.$$

*Proof.* Denote $T = \mathrm{Id} - \tau\nabla F$. The negative differential entropy of $\hat{\mu}$ can be computed by plugging in the pushforward's density

$$\begin{aligned}\mathcal{H}(\hat{\mu}) &= \int \log(\hat{\mu}(x))\hat{\mu}(\mathrm{d}x) \\ &= \int \log\left(\frac{\mu(y)}{\det(DT(y))}\right)\mu(\mathrm{d}y),\end{aligned}$$

where we used that $T$ is invertible since $\tau < L_{\nabla F}^{-1}$. Due to the strong convexity of $F$ and Lipschitz continuity of $\nabla F$, we know that all eigenvalues of $DT(y)$ lie in the interval $[1 - \tau L_{\nabla F}, 1 - \tau\omega_F]$. This implies

$$(1 - \tau L_{\nabla F})^d \leq \det(DT(y)) \leq (1 - \tau\omega_F)^d.$$

For the negative entropy of $\hat{\mu}$, we therefore obtain the bounds

$$\begin{aligned}\mathcal{H}(\hat{\mu}) &= \mathcal{H}(\mu) - \mathbb{E}_{Y\sim\mu}[\log\det(DT(Y))] \\ &\in \left[\mathcal{H}(\mu) - d\log(1 - \tau\omega_F), \mathcal{H}(\mu) - d\log(1 - \tau L_{\nabla F})\right].\end{aligned}$$

This is equivalent to the first two inequalities in the statement. Since $\tau\omega_F \leq \tau L_{\nabla F} < 1$, we obtain $\log(1 - \tau\omega_F) < 0$. □

With this additional result, we are able to prove the convergence result for Theorem 8.9.

*Proof of Theorem 8.9.* We start by deriving a central "descent" inequality that is comparable to the one in the proof of Theorem 8.4. For $n \geq 1$, it takes the form

$$2\tau\,\mathcal{D}_{\mathrm{KL}}(\mu^{(n+\frac{1}{3})} \,\|\, \mu^*) \leq \mathcal{W}_2^2(\mu^{(n-\frac{1}{3})}) - \omega_F\tau\mathcal{W}_2^2(\mu^{(n)}, \mu^*) - \mathcal{W}_2^2(\mu^{(n+\frac{2}{3})}, \mu^*) + 2\tau^2 L_G^2 - 2\tau d\log(1 - \tau L_{\nabla F}), \tag{8.9}$$

where we denoted as before $\mu^{(n+\frac{j}{3})}$, $j \in \{1, 2, 3\}$ for the distribution of intermediate iterates, now in the following order:

$$\begin{aligned} X^{(n+\frac{1}{3})} &= X^{(n)} - \tau\nabla V(X^{(n)}), \\ X^{(n+\frac{2}{3})} &= \mathrm{prox}_G^\tau(X^{(n+\frac{1}{3})}), \\ X^{(n+1)} &= X^{(n+\frac{2}{3})} + \sqrt{2\tau}\xi^{(n+1)}, \quad \xi^{(n+1)} \sim \mathrm{N}(0, I) \end{aligned}$$

As before, the above inequality is the sum of four of the auxiliary results. We start with Lemma 8.5 for $\mu = \mu^{(n)}$ and $\nu = \mu^*$, giving

$$2\tau(\mathcal{E}_F(\mu^{(n+\frac{1}{3})}) - \mathcal{E}_F(\mu^*)) \leq (1 - \omega_F\tau)\mathcal{W}_2^2(\mu^{(n)}, \mu^*) - \mathcal{W}_2^2(\mu^{(n+\frac{1}{3})}, \mu^*).$$

Secondly, we apply Lemma 8.6 with $\mu = \mu^{(n-\frac{1}{3})}$, $\nu = \mu^*$:

$$2\tau(\mathcal{H}(\mu^{(n)}) - \mathcal{H}(\mu^*)) \leq \mathcal{W}_2^2(\mu^{(n-\frac{1}{3})}, \mu^*) - \mathcal{W}_2^2(\mu^{(n)}, \mu^*).$$

Next, we set $\mu = \mu^{(n+\frac{1}{3})}$, $\nu = \mu^*$ in Lemma 8.8 to obtain

$$2\tau(\mathcal{E}_G(\mu^{(n+\frac{1}{3})}) - \mathcal{E}_G(\mu^*)) \leq \mathcal{W}_2^2(\mu^{(n+\frac{1}{3})}, \mu^*) - \mathcal{W}_2^2(\mu^{(n+\frac{2}{3})}, \mu^*) + 2L_G^2\tau^2.$$

Note that we do *not* use Lemma 8.7 this time, but rather have to employ Lemma 8.10 due to the changed order of steps. Setting $\mu = \mu^{(n)}$, its first inequality gives

$$2\tau(\mathcal{H}(\mu^{(n+\frac{1}{3})}) - \mathcal{H}(\mu^{(n)})) \leq -2\tau d\log(1 - \tau L_{\nabla F}).$$

We obtain (8.9) by summing up the last four inequalities. The KL divergence on the left side is bounded below by zero. Iterating the inequality over $n$ steps gives

$$\mathcal{W}_2^2(\mu^{(n+\frac{2}{3})}, \mu^*) \leq \mathcal{W}_2^2(\mu^{(\frac{2}{3})}, \mu^*) - \omega_F\tau\sum_{k=1}^{n}\mathcal{W}_2^2(\mu^{(k)}, \mu^*) + 2n\tau\left(\tau L_G^2 - d\log(1 - \tau L_{\nabla F})\right). \tag{8.11}$$

Since the left side can again be bounded by zero, this inequality would already be sufficient to at least obtain a finite bound on the ergodic average of the squared distances $\frac{1}{n}\sum_{k=1}^{n}\mathcal{W}_2^2(\mu^{(k)},\mu^*)$. In order to obtain a bound on the bias of the stationary distribution, we use the Markov chain's geometric ergodicity: By the reverse triangle inequality and Theorem 8.3, we know that

$$\begin{aligned}\mathcal{W}_2(\mu^{(k)},\mu^*) &\geq \mathcal{W}_2(\mu^{\mathrm{stat}}_{\mathrm{FBf}},\mu^*) - \mathcal{W}_2(\mu^{(k)},\mu^{\mathrm{stat}}_{\mathrm{FBf}}) \\ &\geq \mathcal{W}_2(\mu^{\mathrm{stat}}_{\mathrm{FBf}},\mu^*) - q_F(\tau)^k\mathcal{W}_2(\mu^{(0)},\mu^{\mathrm{stat}}_{\mathrm{FBf}}),\end{aligned}$$

where now $q_F(\tau) = 1-\omega_F\tau$, since we assumed $\tau < \frac{1}{L_{\nabla F}} \leq \frac{2}{\omega_F+L_{\nabla F}}$. After taking the square in the last inequality, we obtain

$$\begin{aligned}\mathcal{W}_2^2(\mu^{(k)},\mu^*) &\geq \mathcal{W}_2^2(\mu^{\mathrm{stat}}_{\mathrm{FBf}},\mu^*) + q_F(\tau)^{2k}\mathcal{W}_2^2(\mu^{(0)},\mu^{\mathrm{stat}}_{\mathrm{FBf}}) \\ &\qquad - 2q_F(\tau)^k\mathcal{W}_2(\mu^{(0)},\mu^{\mathrm{stat}}_{\mathrm{FBf}})\mathcal{W}_2(\mu^{\mathrm{stat}}_{\mathrm{FBf}},\mu^*) \\ &\geq \left(1-\frac{1}{1+q_F(\tau)^{-k}}\right)\mathcal{W}_2^2(\mu^{\mathrm{stat}}_{\mathrm{FBf}},\mu^*) - q_F(\tau)^k\mathcal{W}_2^2(\mu^{(0)},\mu^{\mathrm{stat}}_{\mathrm{FBf}}) \\ &\geq \left(1-q_F(\tau)^k\right)\mathcal{W}_2^2(\mu^{\mathrm{stat}}_{\mathrm{FBf}},\mu^*) - q_F(\tau)^k\mathcal{W}_2^2(\mu^{(0)},\mu^{\mathrm{stat}}_{\mathrm{FBf}}).\end{aligned}$$

Here, in the second step, we used a special case of Young's inequality, $2\alpha\beta \leq c\alpha^2 + c^{-1}\beta^2$ for all $c>0$, and set $c = (1+q_F(\tau)^{-k})$. We can plug this in the iterated inequality (8.11) and obtain

$$\begin{aligned}\omega_F\tau\mathcal{W}_2^2(\mu^{\mathrm{stat}}_{\mathrm{FBf}},\mu^*)\sum_{k=1}^{n}(1-q_F(\tau)^k) \leq \mathcal{W}_2^2(\mu^{(\frac{2}{3})},\mu^*) - \omega_F\tau\mathcal{W}_2^2(\mu^{(0)},\mu^{\mathrm{stat}}_{\mathrm{FBf}})\sum_{k=1}^{n}q_F(\tau)^k \\ + 2n\tau\left(\tau L_G^2 - d\log(1-\tau L_{\nabla F})\right).\end{aligned}$$

We now upper bound the geometrical sum by its series limit on the left, noting again $q_F(\tau) = 1-\omega_F\tau$. For $n \geq 1 + \lceil(\omega_F\tau)^{-1}\rceil$, we divide the inequality by $n\omega_F\tau - 1$ and obtain

$$\begin{aligned}\mathcal{W}_2^2(\mu^{\mathrm{stat}}_{\mathrm{FBf}},\mu^*) \leq \frac{1}{n\omega_F\tau-1}\left(\mathcal{W}_2^2(\mu^{(\frac{2}{3})},\mu^*) - (1-\omega_F\tau-(1-\omega_F\tau)^{n+1})\mathcal{W}_2^2(\mu^{(0)},\mu^{\mathrm{stat}}_{\mathrm{FBf}})\right) \\ + \frac{2n\tau}{n\omega_F\tau-1}\left(\tau L_G^2 - d\log(1-\tau L_{\nabla F})\right).\end{aligned}$$

Note that the left side does not depend on $n$. On the right side, the expression in the first line converges to zero in the limit $n\to\infty$. The second line converges to

$$\frac{2}{\omega_F}\left(\tau L_G^2 - d\log(1-\tau L_{\nabla F})\right) = C\tau + \mathcal{O}(\tau^2),$$

where $C = \frac{2}{\omega_F}(L_G^2 + dL_{\nabla F})$. By taking the limit $n\to\infty$ and a square root of the inequality, we obtain the desired statement. □

Having completed the convergence proofs for (F*flow*B) and (FB*flow*), we want to remark that the assumption that $G$ is Lipschitz continuous[1] in Theorems 8.4 and 8.9 is rather restrictive and not strictly necessary. In at least two works, the statement of Theorem 8.4 due to Durmus et al. has been modified to cover more general potentials:

- In [84], the authors noted that the only time the Lipschitz continuity assumption is used is in the auxiliary result Lemma 8.8. They show the latter essentially still holds if $G$ is convex and grows at most quadratically, i.e., there exist $\alpha, \beta > 0$ such that
$$\|G(x)\| \leq \alpha + \beta\|x\|^2, \qquad \forall x \in \mathbb{R}^d.$$
Under this assumption, the bound in Lemma 8.8 becomes
$$2\tau(\mathcal{E}_G(\mu) - \mathcal{E}_G(\nu)) \leq \mathcal{W}_2^2(\mu, \nu) - \mathcal{W}_2^2(\hat{\mu}, \nu) + \gamma\tau^2,$$
where the constant $\gamma > 0$ depends on the growth of $G$. After replacing $2L_G^2$ by $\gamma$ in the bias bounds in Theorem 8.4 and Theorem 8.9, the statements still hold.

- In [186], the use of Lemma 8.8 was avoided in the first place. Instead of limiting the growth of $G$ in order to bound the quantity $2\tau(\mathcal{E}_G(\mu^{(n+\frac{2}{3})}) - \mathcal{E}_G(\mu^*))$, the authors employ convex duality and bound the quantity
$$2\tau(\mathcal{E}_{G^*}(\nu^{(n+\frac{2}{3})}) - \mathcal{E}_{G^*}(\nu^*)),$$
with pushforward distributions $\nu^{(n+\frac{2}{3})} = \psi^{(n+\frac{2}{3})}{}_{\#}\mu^*$ and $\nu^* = \psi^*{}_{\#}\mu^*$. The mappings $\psi \in L^2(\mu; \mathbb{R}^d)$ can formally be understood as convex dual variables generating pushforward distributions $\nu$ of a synthetic dual iterate. Instead of bounding the KL divergence as in the central inequality (8.8) in the proof of Theorem 8.4, one then obtains decay in a duality gap arising from the generalized Lagrangian
$$\mathcal{L}(\mu, \psi) = \mathcal{H}(\mu) + \mathcal{E}_F(\mu) - \mathcal{E}_{G^*}(\psi_{\#}\mu^*) + \mathbb{E}_{(X,Y)\sim\pi(\mu,\psi)}[\langle X, Y\rangle],$$
where $\pi = (T_{\mu^*\to\mu}, \psi)_{\#}\mu^*$. The main conceptual work of [186] goes into proving a saddle point theorem for $\mathcal{L}$, which is similar in spirit to duality results known from Euclidean optimization. The assumptions on $G$ are relaxed in that an ergodic convergence result can be derived if $V \in S^{1,1}_{\mathrm{loc}}(\mathbb{R}^d)$ and the elements of $\partial G$ have finite variance under $\mu^*$. This precludes $G$ from attaining infinite values, but significantly relaxes the growth assumptions. Improved guarantees hold if $G$ is differentiable and $\nabla G$ Lipschitz continuous.

Up to some extent, these generalizations are still disappointingly restrictive. From the perspective of both Euclidean optimization and Theorem 8.3, one would hope that convexity of $G$, and even potentially a weaker setting where $\mathrm{supp}(G) \subsetneq \mathbb{R}^d$, is sufficient to

[1] Alternatively, in a related paper [185], the authors employed a very similar assumption $\|\partial^0 G(x)\| \leq C$, where $C > 0$ is a constant independent of $x$. Here, $\partial^0 G(x)$ is the subgradient of $G$ at $x$ with minimal norm, i.e., $\partial^0 G(x) \coloneqq \arg\min_{p\in\partial G(x)}\{\|p\|\}$, which is well-defined due to the convexity of $\partial G(x)$.

obtain convergence. Even though the latter case poses new questions on the limiting SDE, the Markov chain is certainly still geometrically ergodic by Theorem 8.3. Similarly, in continuous time, if $F$ is strongly convex or satisfies a log-Sobolev inequality, a convex $G : \mathbb{R}^d \to \mathbb{R}$ is sufficient to guarantee convergence of the continuous time overdamped Langevin diffusion. On the other hand, it is not surprising that the proofs of Theorems 8.4 and 8.9 and those in [73, 186, 84] do not extend the general case of convex $G$. The whole proof technique hinges on the exact order of inequalities for the forward-, diffusion- and backward steps and is therefore laid out to require some form of Lemma 8.8. The latter cannot hold in the general, convex case and requires a bound on the growth of $G$—consider for example the characteristic function $G = \chi_{\mathcal{C}}$ of a compact, convex set $\mathcal{C} \subset \mathbb{R}^d$ to see immediately that the lemma would fail and the KL objective would be infinite at every intermediate step $\mu^{(n+\frac{2}{3})}$. But this indicates exactly the conceptual blind spot of the theory: For $G = \chi_{\mathcal{C}}$, the proximal mapping $\mathrm{prox}_G$ is simply the projection $\Pi_C$. In this special case, (F*flow*B) reduces to projected Langevin Monte Carlo. For this algorithm, it is perfectly possible to derive convergence proofs [39, 40]. Since the theory of sampling from more general posterior distributions and constraint domains is very relevant in applications, this is a promising direction of future research.

## 8.3. Primal-Dual Langevin Sampling

We now assume that the target density is of the form

$$\mu^*(x) \propto \exp(-F(Kx) - G(x)), \tag{8.12}$$

where $F : \mathbb{R}^m \to \bar{\mathbb{R}}$ and $G : \mathbb{R}^d \to \bar{\mathbb{R}}$ are general convex, but potentially non-smooth potentials and $K : \mathbb{R}^d \to \mathbb{R}^m$ a linear operator. In particular, $V = F \circ K + G$ is convex, so $\mu^*$ is assumed to be log-concave. As was already eluded to, this particular structure covers wide ranges of the possible priors and likelihoods that are relevant in imaging inverse problems, see Section 4.4. Just as we did for (ULA) and the forward-backward variants (F*flow*B) and (FB*flow*), we would now like to provide further analysis on the behaviour of the primal-dual sampling method (ULPDA). As before, this entails proving geometric ergodicity and bounding the bias between continuous and discrete time dynamics. However, as we saw, the continuous time dynamics (7.30) do not coincide with (6.47). As we will show next, the stationary state of the new diffusion equation is also not the target distribution. This implies that a quantitative bound on discretization errors does not yet give any bound between the sample distribution and the target.

We start by proving that the continuous time dynamics are contractive, implying that there exists a stationary state. Although we are dealing with a new SDE now, it turns out that the same idea as in the proof of Theorem 6.18 for overdamped Langevin diffusion can be revisited when proving asymptotic stability. A closer inspection of the proof

reveals that the necessary inequality in the calculation of the energy dissipation was

$$\langle x - \tilde{x}, \nabla V(x) - \nabla V(\tilde{x}) \rangle \geq \omega \|x - \tilde{x}\|^2,$$

i.e., strong monotonicity of the drift coefficient $\nabla V$. In the remaining computations, $V$ is never used directly, hence it is easy to generalize the proof to any diffusion SDE with a strongly monotone drift coefficient [43, Sec. 3]. Consider therefore now $(Z_t)_{t\geq 0}$ which satisfies a general stochastic differential inclusion

$$\mathrm{d}Z_t \in -A(Z_t)\,\mathrm{d}t + B(Z_t)\,\mathrm{d}W_t. \tag{8.13}$$

Here $W_t$ is Brownian motion in $\mathbb{R}^k$ for some $k \geq 1$ and $B : \mathbb{R}^d \to \mathbb{R}^{d\times k}$. We assume the drift $A$ to be a set-valued operator $A : \mathbb{R}^d \to 2^{\mathbb{R}^d}$. In order to speak meaningfully about existence and uniqueness in law of solutions to such inclusions, $A$ has to satisfy typical regularity assumptions, i.e., have a closed graph and convex images, as well as satisfy a growth condition similar to the one we gave for SDEs in Theorem 6.13, see [107, 12]. Similarly to Kolmogorov's equations for SDEs, we also obtain solutions by continuous selection to the generalized Fokker–Planck equation with set-valued coefficients, i.e., the density $\pi_t$ of $Z_t$ in (8.13) satisfies

$$\frac{\partial \pi_t}{\partial t} = \nabla_z \cdot (w\pi_t + \Sigma(z)\nabla_z \pi_t), \qquad w \in A(z), \tag{8.14}$$

where $\Sigma(z) = \frac{1}{2}B(z)B(z)^T$. For further details on stochastic inclusions, we also refer to the comprehensive treatment [108]. For us, the relevant property of $A$ to guarantee asymptotic stability is strong monotonicity. We call $A$ monotone if

$$\langle w - \tilde{w}, z - \tilde{z} \rangle \geq 0, \qquad \forall w \in A(z), \tilde{w} \in A(\tilde{z}).$$

Further, $A$ is $\omega$-strongly monotone if the following stronger condition is satisfied:

$$\langle w - \tilde{w}, z - \tilde{z} \rangle \geq \omega \|z - \tilde{z}\|^2, \qquad \forall w \in A(z), \tilde{w} \in A(\tilde{z}).$$

Under strong monotonicity of the drift, we have the following result.

**Theorem 8.11: Stability of diffusion with strongly monotone drift**

Let $A : \mathbb{R}^d \to 2^{\mathbb{R}^d}$ be a set-valued, $\omega$-strongly monotone drift, $B : \mathbb{R}^d \to \mathbb{R}^{d\times k}$ a noise coeffient such that (8.13) admits a solution that is unique in law for any initialization $Z_0 \sim \pi_0$. Then the dynamics (8.13) are contractive in the sense that, if $(Z_t)_{t\geq 0}$ and $(\tilde{Z}_t)_{t\geq 0}$ are two solutions with initializations $\pi_0, \tilde{\pi}_0 \in \mathcal{P}_2(\mathbb{R}^d)$, respectively, then $Z_t \sim \pi_t, \tilde{Z}_t \sim \tilde{\pi}_t$ satisfy

$$\mathcal{W}_2(\pi_t, \tilde{\pi}_t) \leq e^{-\omega t}\mathcal{W}_2(\pi_0, \tilde{\pi}_0).$$

In particular, there is a unique stationary solution $\pi^{\mathrm{stat}}$ and $\lim_{t\to\infty} \pi_t = \pi^{\mathrm{stat}}$ for any $\pi_0 \in \mathcal{P}_2(\mathbb{R}^d)$.

*Proof.* The proof is due to [43, Lemma 3.3], the main idea similar to the proof of Theorem 6.18. We recognize that, if $Z_t$ and $\tilde{Z}_t$ both solve (8.13) with the same Brownian motion, then the coupled process $(Z_t, \tilde{Z}_t) \sim \Pi_t$ solves

$$\begin{pmatrix} \mathrm{d}Z_t \\ \mathrm{d}\tilde{Z}_t \end{pmatrix} \in -\begin{pmatrix} A(Z_t) \\ A(\tilde{Z}_t) \end{pmatrix} \mathrm{d}t + \begin{pmatrix} B(Z_t) \\ B(\tilde{Z}_t) \end{pmatrix} \mathrm{d}W_t, \tag{8.15}$$

initilized at any coupling $\Pi_0 \in \Gamma(\pi_0, \tilde{\pi}_0)$. Similarly, the joint measure $\Pi_t$ solves the generalized Fokker–Planck equation

$$\frac{\partial \Pi_t}{\partial t} = \nabla_z \cdot (w\pi_t) + \nabla_{\tilde{z}} \cdot (\tilde{w}\Pi_t) + (\nabla_z + \nabla_{\tilde{z}}) \cdot (\Sigma(z,\tilde{z})(\nabla_z + \nabla_{\tilde{z}})\Pi_t), \quad w \in A(z), \tilde{w} \in A(\tilde{z}), \tag{8.16}$$

where the diffusion coefficient is given by

$$\Sigma(z,\tilde{z}) = \frac{1}{2}\begin{pmatrix} B(z)B(z)^T & B(z)B(\tilde{z})^T \\ B(\tilde{z})B(z)^T & B(\tilde{z})B(\tilde{z})^T \end{pmatrix}.$$

By a change of variable, we write $\nabla_z + \nabla_{\tilde{z}} = \nabla_{z+\tilde{z}}$. For any $w \in A(z), \tilde{w} \in A(\tilde{z})$, we have

$$\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\, \mathbb{E}_{(z,\tilde{z})\sim\Pi_t}[\|z-\tilde{z}\|^2] \\
&\quad = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} \int \|z-\tilde{z}\|^2 \Pi_t(\mathrm{d}z, \mathrm{d}\tilde{z}) \\
&\quad = \frac{1}{2}\int \|z-\tilde{z}\|^2 \left(\nabla_z \cdot (w\Pi_t) + \nabla_{\tilde{z}} \cdot (\tilde{w}\Pi_t) + (\nabla_{z+\tilde{z}}) \cdot (\Sigma(\nabla_{z+\tilde{z}})\Pi_t)\right) \mathrm{d}z\, \mathrm{d}\tilde{z} \\
&\quad = -\int \left((z-\tilde{z})\cdot w + (\tilde{z}-z)\cdot \tilde{w}\right) \Pi_t(\mathrm{d}z, \mathrm{d}\tilde{z}) \\
&\quad \le -\omega\, \mathbb{E}_{(z,\tilde{z})\sim\Pi_t}[\|z-\tilde{z}\|^2]
\end{aligned}$$

Grönwall's Lemma gives

$$\mathbb{E}_{(z,\tilde{z})\sim\Pi_t}[\|z-\tilde{z}\|^2] \le e^{-2\omega t}\, \mathbb{E}_{(z,\tilde{z})\sim\Pi_0}[\|z-\tilde{z}\|^2]$$

and hence

$$\mathcal{W}_2(\pi_t, \tilde{\pi}_t) \le e^{-\omega t}\, \mathbb{E}_{(z,\tilde{z})\sim\Pi_0}[\|z-\tilde{z}\|^2].$$

Since this inequality holds for all $\Pi_0$, we can take the infimum over all couplings of $\pi_0, \tilde{\pi}_0$ on the right hand side and obtain the statement. □

For the primal-dual dynamics (7.31), the corresponding Fokker–Planck equation is

$$\frac{\partial \pi_t}{\partial t} = \nabla_z \cdot (w\pi_t) + \Delta_x \pi_t, \qquad w \in A_{\mathrm{PD}}(z). \tag{8.17}$$

Proving the existence of a stationary solution and convergence to it now reduces to simply checking $\omega$-strong monotonicity of the drift coefficent $A_{\mathrm{PD}}$. This can be ensured by assuming strong convexity of $G$ and $F^*$ and by suitably changing the metric of the joint primal-dual space. In the context of primal-dual methods, denote from now on by $\langle x, \tilde{x}\rangle_a := a\langle x, \tilde{x}\rangle$ a weighted inner product on $\mathbb{R}^d$ or $\mathbb{R}^m$. On the joint space $\mathbb{R}^d \times \mathbb{R}^m \cong \mathbb{R}^{d+m}$, we want to assign different weights to primal and dual variables, i.e., consider the inner product $\langle z, \tilde{z}\rangle_{a,b} := a\langle x, \tilde{x}\rangle + b\langle y, \tilde{y}\rangle$, for all $z, \tilde{z} \in \mathbb{R}^{d+m}$, where $z = (x, y)$ and $\tilde{z} = (\tilde{x}, \tilde{y})$. By $\|\cdot\|_a$ on $\mathbb{R}^d$ or $\mathbb{R}^m$ and by $\|\cdot\|_{a,b}$ on $\mathbb{R}^{d+m}$, we denote the induced weighted norms. Equivalently, we can define weighted 2-Wasserstein distances. For all $\pi, \tilde{\pi} \in \mathcal{P}_2(\mathbb{R}^{d+m})$, denote

$$\mathcal{T}_{a,b}(\pi, \tilde{\pi}) := \inf_{\Pi \in \Gamma(\pi, \tilde{\pi})} \left(\mathbb{E}_{(z,\tilde{z})\sim\Pi}[\|z - \tilde{z}\|_{a,b}^2]\right)^{1/2}. \tag{8.18}$$

The equivalence of norms on $\mathbb{R}^{d+m}$ implies equivalence of all $\mathcal{T}_{a,b}$ with $a, b > 0$. Suppose now $G$ is $\omega_G$-strongly convex and $F^*$ is $\omega_{F^*}$-strongly convex. Due to the different weights in the off-diagonal terms $K^*y$ and $\lambda K x$ of the primal-dual drift $A_{\mathrm{PD}}$, we move to the weighted inner product $\langle\cdot,\cdot\rangle_{1,\lambda^{-1}}$. Let now $z, \tilde{z} \in \mathcal{Z}$, and $w \in A_{\mathrm{PD}}(z)$, $\tilde{w} \in A_{\mathrm{PD}}(\tilde{z})$. Then there exist $p \in \partial G(x)$ and $q \in \partial F^*(y)$ such that $w = (p + K^T y, \lambda(q - Kx))$, and analogously, $\tilde{w} = (\tilde{p} + K^T\tilde{y}, \lambda(\tilde{q} - K\tilde{x}))$ for some $\tilde{p} \in \partial G(\tilde{x})$ and $\tilde{q} \in \partial F^*(\tilde{y})$. To check strong monotonicity, it is then simple to compute

$$\begin{aligned}
\langle w - \tilde{w}, z - \tilde{z}\rangle_{1,\lambda^{-1}} &= \langle u - \tilde{u}, x - \tilde{x}\rangle + \lambda\langle v - \tilde{v}, y - \tilde{y}\rangle_{\lambda^{-1}} \\
&\quad + \langle K^T(y - \tilde{y}), x - \tilde{x}\rangle - \lambda\langle K(x - \tilde{x}), y - \tilde{y}\rangle_{\lambda^{-1}} \\
&= \langle u - \tilde{u}, x - \tilde{x}\rangle + \langle v - \tilde{v}, y - \tilde{y}\rangle \\
&\geq \omega_G\|x - \tilde{x}\|^2 + \omega_{F^*}\|y - \tilde{y}\|^2 \\
&\geq \min\{\omega_G, \lambda\omega_{F^*}\}\|z - \tilde{z}\|_{1,\lambda^{-1}}^2.
\end{aligned}$$

This shows that $A_{\mathrm{PD}}$ is $\omega$-strongly monotone in the weighed metric if we set $\omega := \min\{\omega_G, \lambda\omega_{F^*}\}$. In view of Theorem 8.11, this computation justifies the following result [43, Thm 3.4].

**Theorem 8.12: Convergence of primal-dual dynamics**

Suppose $G$ is $\omega_G$-strongly convex and $F^*$ is $\omega_{F^*}$-strongly convex. Then (7.31) has a unique stationary measure $\pi_{(\lambda)} \in \mathcal{P}_2(\mathbb{R}^{d+m})$, where we make the dependence on the step size ratio $\lambda$ explicit. Let $(Z_t)_{t\geq 0}$ solve (7.31) with some initial distribution

$Z_0 \sim \pi_0 \in \mathcal{P}_2(\mathbb{R}^{d+m})$. Then

$$\mathcal{T}_{1,\lambda^{-1}}(\pi_t, \pi_{(\lambda)}) \leq e^{-\omega t}\mathcal{T}_{1,\lambda^{-1}}(\pi_0, \pi_{(\lambda)}),$$

with $\omega = \min\{\omega_G, \lambda\omega_{F^*}\}$.

From the sampling perspective, we are mainly interested in the $x$-marginal of $\pi_{(\lambda)}$, i.e. the measure $\mu_{(\lambda)} \in \mathcal{P}_2(\mathbb{R}^d)$ with density $\mu_{(\lambda)}(x) \coloneqq \int_{\mathcal{Y}} \pi_{(\lambda)}(x, y)\,\mathrm{d}y$.

Unfortunately, it generally holds $\mu_{(\lambda)} \neq \mu^*$. Heuristically, this can be seen as a consequence of the dualization with a finite $\lambda$, contrasted with the primal limit $\lambda = \infty$. Due to the presence of the diffusion, the joint process $(X_t, Y_t)$ does not converge to a state which is concentrated at the dual optimality condition $Y_t \in \partial F(KX_t)$. Importantly, this contrasts what one might be used to from primal-dual optimization dynamics, where this joint convergence to the optimal, concentrated state in both primal and dual variable is explicitly designed. Assuming for a moment that $F$ is differentiable, a concentration of the form $Y_t = \nabla F(KX_t)$ would generate a singular joint distribution $\pi_t = (\mathrm{Id}, \nabla F \circ K)_{\#}\mu_t$ for the marginal $X_t \sim \mu_t$. A main contribution of [43] were the following two results. The first shows that the concentration is not possible for general $F, G$ if $\lambda < \infty$.

**Theorem 8.13: Hypoellipticity of primal-dual diffusion**

Assume that $G \in C^\infty(\mathbb{R}^d)$, $F^* \in C^\infty(\mathbb{R}^m)$ and that $m \leq d$ and $K$ has full rank $m$. Then the generator $\mathcal{L}_{\mathrm{PD}} = -A_{\mathrm{PD}}^T(z)\nabla_z + \Delta_x$ satisfies Hörmander's condition, see Lemma C.3. This implies that every solution to (8.17) is $C^\infty((0,\infty)\times\mathbb{R}^{d+m})$-smooth. In particular, the stationary distribution $\pi_\lambda$ also has a $C^\infty(\mathbb{R}^{d+m})$-smooth Lebesgue density.

This result is somewhat negative, as it implies that the ideal joint measure $(\mathrm{Id}, \nabla F \circ K)_{\#}\mu_t$ can generally not occur when simulating the joint dynamics. For the proof, we refer to [43]. In defense of the primal-dual dynamics, the second result shows that the concentrated dynamics are reached in the limit $\lambda \to \infty$, where we formally recover overdamped Langevin diffusion.

**Theorem 8.14: Large $\lambda$ limit of primal-dual diffusion**

Assume that $G$ is $\omega_G$-strongly convex, $F^*$ is $\omega_{F^*}$-strongly convex and $F \in C^3(\mathbb{R}^m)$ satisfies certain, further regularity assumptions[a]. Assume further that $(Z_t)_{t\geq 0}$ solves (7.31), with concentrated initialization $Y_0 = \nabla F(KX_0)$. Then, if $\lambda >$

$\omega_G/\omega_{F^*}$, the primal marginal $\mu_t$ of $\pi_t$ in (8.17) obeys the bound

$$\mathcal{W}_2(\mu_t, \mu^*) \leq (C_1(\lambda)\mathcal{W}_2(\mu_0, \mu^*) + C_2(\lambda))\, e^{-\omega_G t} + C_2(\lambda)$$

with constants $C_1(\lambda) = 1 + \mathcal{O}(\lambda^{-1/2})$ and $C_2(\lambda) = C\lambda^{-1/2} + \mathcal{O}(\lambda^{-3/2})$ with some constant $C > 0$ as $\lambda \to \infty$. In particular, the stationary primal marginal $\mu_{(\lambda)}$ satisfies

$$\mathcal{W}_2(\mu_{(\lambda)}, \mu^*) \leq C_1(\lambda).$$

[a] These additional assumptions guarantee that the concentration manifold $Y = \nabla F(KX)$ in the joint space $\mathbb{R}^{d+m}$ has a certain regularity, such that the pushforward of primal, overdamped Langevin diffusion to this manifold has coefficients with a bounded variance. We refer to [43, Thm 3.11] for details.

For the proof, we again refer to [43]. With the latter result, it is possible to argue that the additional bias from the partial dualization can be controlled by choosing $\lambda$ large enough. This effect is also visualized in Fig. 8.1, where the convergence behaviour of the primal-dual diffusion dynamics is validated in the simple example of a one-dimensional Gaussian target.

From here, we can proceed as in the theoretical considerations of other Langevin sampling algorithms: We need to prove (geometrical) ergodicity in order to guarantee the convergence of the iterates' primal marginal distribution $\mu^{(n)}$ to the stationary limit $\mu_{(\lambda,\tau)}$, which depends on the primal step size $\tau$ and the step size ratio $\lambda$. Afterwards, we need to bound the discretization bias between $\mu_{(\lambda,\tau)}$ and the continuous time stationary $\mu_{(\lambda)}$. We now state the corresponding theorems from [43] and, without going into details, discuss some main proof aspects.

**Theorem 8.15: (ULPDA): geometric ergodicity**

Let $(X^{(n)}, Y^{(n)})_{n\in\mathbb{N}_0}$ and $(\tilde{X}^{(n)}, \tilde{Y}^{(n)})_{n\in\mathbb{N}_0}$ be coupled iterations, generated by (ULPDA) with the same stochastic inputs $(\xi^{(n)})_{n\in\mathbb{N}}$. Denote the joint and primal marginal distributions by $(X^{(n)}, Y^{(n)}) \sim \pi^{(n)}$ and $X^{(n)} \sim \mu^{(n)}$, respectively.
In a first scenario, assume that $\theta = 1$ and $\tau\sigma L^2 \leq 1$, where $L = \|K\|$. Then it holds

$$\mathcal{W}_2(\mu^{(n)}, \tilde{\mu}^{(n)}) \leq \mathcal{T}_{1,\lambda^{-1}}(\pi^{(0)}, \tilde{\pi}^{(0)}).$$

In the second case, we assume $G$ and $F^*$ are $\omega_G$- and $\omega_{F^*}$-strongly convex, respectively. Suppose the extrapolation parameter $\theta$ is chosen such that

$$\max\{(1 + 2\omega_G\tau)^{-1}, (1 + 2\omega_{F^*}\sigma)^{-1}\} \leq \theta < 1,$$

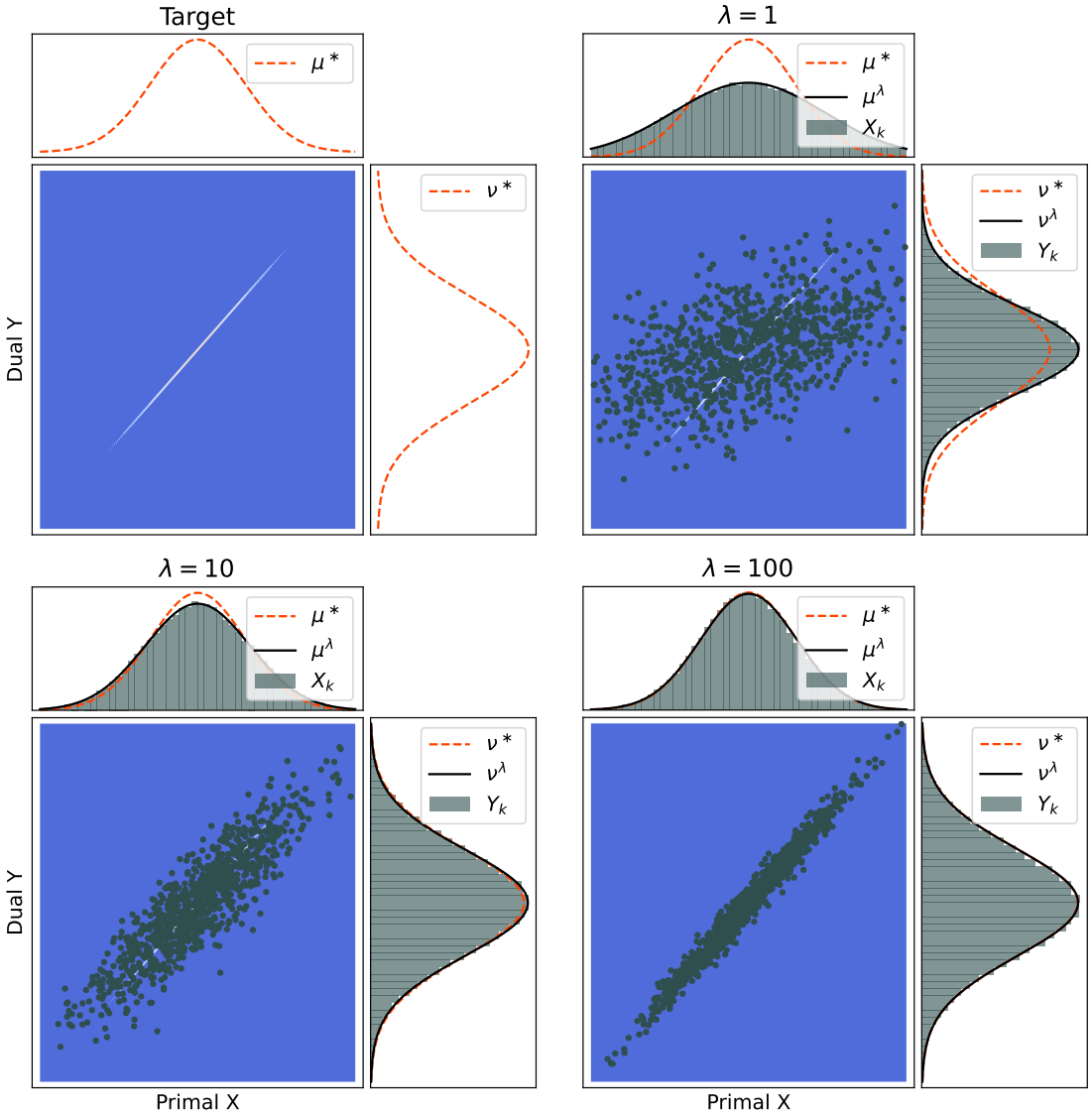



Figure 8.1.: Approximation of $\mu^* = \exp(-V)$ for a quadratic $V$ by primal-dual diffusion (7.30). The top left subfigure shows the primal, dual and joint target. Diffusion results, generated by (ULPDA) for $\lambda = 1$ (top right), $\lambda = 10$ (bottom left) and $\lambda = 100$ (bottom right) are shown. The center scatter plot shows pairs $(X^n, Y^n)$, the background color indicates the mass of the concentrated, joint target $\pi^*$ (zero is blue, the mass is concentrated on the white line). We also show the marginal densities of primal and dual targets $\mu^*$ and $\nu^* = (\partial f \circ K)_{\#}\mu^*$ (dashed red), the stationary marginals $\mu^\lambda, \nu^\lambda$ of primal-dual diffusion (black solid) and drawn samples (grey histograms). With increasing $\lambda$, the joint distribution of samples concentrates more around the target.

and further $\theta\sigma\tau L^2 \leq 1$. Then with the constant $\tilde{\lambda} \coloneqq \frac{1+2\omega_G\tau}{1+2\omega_{F^*}\sigma}\lambda$, it holds

$$\mathcal{W}_2(\mu^{(n)}, \tilde{\mu}^{(n)}) \leq \theta^{N/2}\mathcal{T}_{1,\tilde{\lambda}^{-1}}(\pi^{(0)}, \tilde{\pi}^{(0)}).$$

The ergodicity theorem is again very similar in nature to the ergodicity results Theorem 8.1 for (ULA) and Theorem 8.3 for the forward-backward type samplers. After considering a (pseudo-)metric that only depends on the difference of iterates[2], the stochastic input to the coupled Markov chains cancel out. From there, the argument consists of computations that could equally be used to prove stability of a primal-dual optimization algorithm. Unfortunately, we here only obtain the Wasserstein contraction to a stationary distribution of the Markov chain in the strongly convex case.

With a coupling argument similar to the proof of Theorem 8.2, the discretization errors can also be bounded. This is summarized by the following convergence result.

**Theorem 8.16: (ULPDA): Wasserstein bias to time continuous limit**

Suppose that $F, G$ are $\omega_F$- and $\omega_G$-strongly convex, respectively, differentiable and $\nabla F$, $\nabla G$ are $L_{\nabla F}$- and $L_{\nabla G}$-Lipschitz continuous, respectively. Let the parameters $\tau, \sigma, \theta$ satisfy the previous assumption $\theta\tau\sigma L^2 \leq 1$ and

$$\max\{(1+\omega_G\tau)^{-1}, (1+\omega_{F^*}\sigma)^{-1}\} \leq \theta < 1,$$

and suppose further that the primal-dual drift coefficient has finite variance $\mathbb{E}_{Z\sim\pi_{(\lambda)}}[\|A_{\mathrm{PD}}(Z)\|^2]$. Let now $(X^{(n)}, Y^{(n)})$ be generated by (ULPDA) with joint and primal marginal distributions $\pi^{(n)}$ and $\mu^{(n)}$. Then it holds

$$\mathcal{W}_2(\mu^{(n)}, \mu_{(\lambda)}) \leq \theta^{n/2}\mathcal{T}_{1,\bar{\lambda}^{-1}}(\pi^{(0)}, \pi_{(\lambda)}) + C_3(\tau, \theta),$$

where $\bar{\lambda} = \frac{1+\omega_G\tau}{1+\omega_{F^*}\sigma}\lambda$ and

$$C_3(\tau,\theta) = \frac{1}{(1-\theta)^{1/2}}\left((4dL_{\nabla G}^2\omega_G^{-1} + 6dL^2(1+2\theta^2)L_{\nabla F})^{1/2}\tau + \mathcal{O}(\tau^2)\right).$$

If we assume that $\theta$ is chosen as small as possible, then $1-\theta = \min\{\omega_G, \omega_{F^*}\lambda\}\tau + \mathcal{O}(\tau^2)$, so $C_3(\tau,\theta) = \mathcal{O}(\tau^{1/2})$ as $\tau \to 0$. In particular, for this choice of $\theta$, the primal marginal $\pi_{(\mu,\tau)}$ of the Markov chain's stationary distribution $\pi_{(\lambda,\tau)}$ obeys

$$\mathcal{W}_2(\mu_{(\lambda,\tau)}, \mu_{(\lambda)}) \leq C_3(\tau,\theta) = \mathcal{O}(\tau^{1/2}).$$

[2] In the primal-dual case, we do not choose the norm of the difference, but rather a sum of symmetric Bregman divergences with respect to $G$ and $F^*$.

Combining both Theorems 8.14 and 8.16, we obtain the following result that relates primal samples $X^{(n)}$ and target distribution.

**Corollary 8.17: (ULPDA): Wasserstein bias to target**

Suppose all assumptions from Theorems 8.14 and 8.16 hold, and we choose the smallest admissible extrapolation parameter, which is $\theta = \frac{1}{1+\omega_G \tau}$ for large enough $\lambda$. Then, as $\lambda \to \infty$ and $\tau \to 0$, there is a constant $C > 0$ such that

$$\mathcal{W}_2(\mu^{(n)}, \mu^*) \leq \theta^{n/2} \mathcal{T}_{1,\bar{\lambda}^{-1}}(\pi^{(0)}, \pi_{(\lambda)}) + C(\tau^{1/2} + \lambda^{-1/2}) + \mathcal{O}(\tau^{3/2} + \lambda^{-3/2}).$$

## 8.4. Mirror Langevin Sampling

In Section 7.4, we introduced the Riemannian Langevin dynamics

$$\mathrm{d}X_t = ((\nabla \cdot P)(X_t) - P(X_t)\nabla V(X_t))\,\mathrm{d}t + \sqrt{2P(X_t)}\,\mathrm{d}W_t. \tag{8.19}$$

This SDE was derived as the particle equivalent of its Fokker–Planck equation, the Riemannian Wasserstein gradient flow of the KL divergence with respect to $\mu^*$,

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t P \nabla \log \frac{\mu_t}{\mu^*}\right). \tag{8.20}$$

Recall that in the case $P = (\nabla^2 h)^{-1}$, where the Riemannian metric on $\Omega_h = \mathrm{int}(\mathrm{dom}(h))$ is induced by a mirror function $h : \mathbb{R}^d \to \bar{\mathbb{R}}$ of Legendre type, we have the equivalent and more convenient form

$$\begin{cases} X_t = \nabla h^*(Y_t) \\ \mathrm{d}Y_t = -\nabla V(X_t) + \sqrt{2\nabla^2 h(X_t)}\,\mathrm{d}W_t. \end{cases}$$

with a dual variable $Y_t$. A forward in time discretization gives the mirror Langevin algorithm

$$\begin{cases} \xi^{(n+1)} \sim \mathrm{N}(0, I) \\ X^{(n+1)} = \nabla h^*\left(\nabla h(X^{(n)}) - \tau \nabla V(X^{(n)}) + \sqrt{2\tau \nabla^2 h(X^{(n)})}\,\xi^{(n+1)}\right). \end{cases} \tag{MLA}$$

For the convergence analysis of this method, we again take a step back and analyze the ergodicity of the continuous time dynamics before asking questions about the behaviour of the discretized scheme.

For overdamped Langevin diffusion, we saw in Section 6.4 that a standard gradient domination condition comparable to the ones in convex optimization leads to the log-Sobolev inequality (LSI) for the target distribution. By a very similar computation, we

could however also employ the Poincaré inequality (PI) to derive exponential decay in some divergence measure. Recall that the Poincaré inequality takes the form

$$\mathrm{Var}_{\mu^*}[f] \leq C_{\mathrm{PI}}\, \mathbb{E}_{\mu^*}\Big[\|\nabla f\|^2\Big],$$

and we proved in Theorem 6.12 that it implies decay in chi-squared divergence, and thereby also in the objective KL divergence

$$\mathcal{D}_{\mathrm{KL}}(\mu_t \,\|\, \mu^*),\ \mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*) \leq \exp\Big(-2C_{\mathrm{LSI}}^{-1} t\Big)\, \mathcal{D}_{\chi^2}(\mu_0 \,\|\, \mu^*),$$

if $\mu_t$ is the distribution of overdamped Langevin diffusion. Moving back to the mirror case, it is easy to see that the proof of Theorem 6.12 simply transfers to preconditioned geometries if we slightly modify the Poincaré inequality. Following [67], we say that $\mu^*$ satisfies the Riemannian Poincaré inequality

$$\mathrm{Var}_{\mu^*}[f] \leq C_{\mathrm{RPI}}\, \mathbb{E}_{\mu^*}\Big[\nabla f^T P \nabla f\Big], \tag{RPI}$$

for all smooth $f : \mathbb{R}^d \to \mathbb{R}$, where $P : \mathbb{R}^d \to \mathcal{S}_+^d$ induces the Riemannian metric. The same inequality was introduced in the special case $P = (\nabla^2 h)^{-1}$ in [58] under the name mirror Poincaré inequality. Note that (RPI) is a generalization of (PI), since the standard Poincaré inequality is recovered in the Euclidean setting $P = I$, which corresponds to the mirror map $h = \frac{1}{2}\|\cdot\|^2$. Under (RPI), the ergodicity in chi-squared divergence simply generalizes to the Riemannian case, see [67, Prop. 1].

**Theorem 8.18: Ergodicity of mirror Langevin dynamics**

Let $\mu^*$ satisfy (RPI) with $C_{\mathrm{RPI}} > 0$. Suppose $(\mu_t)_{t\geq 0}$ solves the Fokker–Planck equation of Riemannian Langevin dynamics (7.33). Then it holds

$$\mathcal{D}_{\mathrm{KL}}(\mu_t \,\|\, \mu^*),\ \mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*) \leq \exp\Big(-2C_{\mathrm{RPI}}^{-1} t\Big)\, \mathcal{D}_{\chi^2}(\mu_0 \,\|\, \mu^*).$$

*Proof.* The following proof can be found in [58, 67] and is an immediate generalization of the proof of Theorem 6.12. Similar to (6.33), we compute the dissipation in chi-squared

divergence as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*) &= \int \frac{\mu_t}{\mu^*}\partial_t \mu_t \,\mathrm{d}x \\ &= \int \frac{\mu_t}{\mu^*}\nabla \cdot \left(\mu_t P \nabla \log\left(\frac{\mu_t}{\mu^*}\right)\right) \mathrm{d}x \\ &= -\int \nabla\left(\frac{\mu_t}{\mu^*}\right)\mu_t P \nabla \log\left(\frac{\mu_t}{\mu^*}\right) \mathrm{d}x \\ &= -\,\mathbb{E}_{\mu^*}\left[\nabla\left(\frac{\mu_t}{\mu}\right)^T P \nabla\left(\frac{\mu_t}{\mu}\right)\right] \\ &\geq -C_{\mathrm{RPI}}^{-1}\,\mathcal{D}_{\chi^2}(\mu_t \,\|\, \mu^*),\end{aligned}$$

where we used (RPI) with $f = \mu_t/\mu^*$ in the last step. By Grönwall's inequality, we obtain the statement in chi-squared divergence. The statement in KL divergence follows again from $\mathcal{D}_{\mathrm{KL}} \leq \log(1 + \mathcal{D}_{\chi^2}) \leq \mathcal{D}_{\chi^2}$. □

The obvious question that arises from this ergodicity result is the dependence of the rate constant $C_{\mathrm{RPI}}$ on the choice of $P$, or equivalently in the mirror case, the choice of $h$. In particular, in order to guarantee fast convergence for a given target $\mu^*$, we want to choose $P$ such that $C_{\mathrm{RPI}}$ becomes as small as possible [67, 129]. Making this dependence $C_{\mathrm{RPI}} = C_{\mathrm{RPI}}(\mu^*, P)$ explicit, this leads to the optimization problem

$$\min_{P \in \mathfrak{P}} C_{\mathrm{RPI}}(\mu^*, P), \tag{8.21}$$

where we set $\mathfrak{P} = \{P : \mathbb{R}^d \to \mathcal{S}_+^d \mid \mathbb{E}_{\mu^*}[\mathrm{tr}(P)] = \mathrm{tr}(\mathrm{Cov}_{\mu^*})\}$ as the class of possible metrics as in [67]. The constraint $\mathbb{E}_{\mu^*}[\mathrm{tr}(P)] = \mathrm{tr}(\mathrm{Cov}_{\mu^*})$ acts as a normalization so that a simple time rescaling of the diffusion Poincaré together with $P \mapsto \alpha P$ becomes inadmissible. The constraint implies the lower bound $C_{\mathrm{RPI}}(\mu^*, P) \geq 1$, as can be seen by setting $f(x) = x_i$ in (RPI) and summing the inequality over all $i = 1, \ldots, d$. Note that a different constraint was chosen in [129], leading to potentially different optimal metrics.

As a central contribution of [67], the authors showed that an optimal metric $P^*$ that solves (8.21) with $C(\mu^*, P^*) = 1$ exists under mild regularity assumptions on $\mu^*$. Further, they characterized the solution in that if $C(\mu^*, P^*) = 1$, then $P^*$ is a Stein kernel, meaning it satisfies the relationship

$$\mathbb{E}_{X \sim \mu^*}[f(X)(X - m)] = \mathbb{E}_{X \sim \mu^*}[P^*(X)\nabla f(X)],$$

for all smooth $f : \mathbb{R}^d \to \mathbb{R}$, with $m = \mathbb{E}_{X \sim \mu^*}[X]$ the mean of $\mu^*$. In this case, it directly follows that the Riemannian Langevin dynamics (8.19) take the form

$$\mathrm{d}X_t = -(X_t - m)\,\mathrm{d}t + \sqrt{2P^*(X_t)}\,\mathrm{d}W_t. \tag{8.22}$$

In particular, the preconditioned dynamics with the optimal metric do not directly depend on the potential gradient $\nabla V$ anymore. The linear drift shows that relative to the new geometry induced by the preconditioning, the potential acts like a quadratic with strong convexity constant 1, which matches the convergence rate implied by the the optimal Poincaré constant.

Importantly for our work on imaging problems with Poisson distributed data, the work [67] gave two results that allow a more refined characterization of the optimal metric $P^*$ in special cases. The first case concerns targets that are moment measures. An optimal metric $P^*$ with $C_{\mathrm{RPI}}(\mu^*, P^*) = 1$ exists if the target $\mu^*$ is a moment measure, i.e., $\mu^* = (\nabla\varphi)_{\#}(e^{-\varphi}\,\mathrm{Leb})$, where Leb is the Lebesgue measure. If that is the case, then

$$P^*(x) = \nabla^2\varphi((\nabla\varphi)^{-1}(x)) = (\nabla^2\varphi^*)^{-1}(x),$$

where $\varphi^*$ denotes the convex conjugate of $\varphi$. Note, this shows that for moment measures, the optimal metric is actually also induced by a mirror map. The optimal mirror function $h$ that achieves the fastest possible convergence rate $C(\mu^*, (\nabla^2 h)^{-1}) = 1$ is simply $h = \varphi^*$. The other possible characterization proved in [67, Prop. 2] is in the special case where $\mu^*$ is an outer product of multiple marginal measures:

**Lemma 8.19: Optimal preconditioner for separable targets**

Assume $\mu^* = \mu_1^* \otimes \cdots \otimes \mu_L^* \in \mathcal{P}_2(\mathbb{R}^d)$, where $\mu_l^* \in \mathcal{P}_2(\mathbb{R}^{d_l})$ and $d = \sum_l d_l$. Assume that there are optimal metrics $P_l^* : \mathbb{R}^{d_l} \to \mathcal{S}_+^{d_l}$ which each solve (8.21) for $\mu_l^*$, respectively, with $C(\mu_l^*, P_l^*) = 1$ for all $l$. Then the block-diagonal metric

$$P(x) = \begin{pmatrix} P_1^*(x_{[1]}) & & 0 \\ & \ddots & \\ 0 & & P_L^*(x_{[L]}) \end{pmatrix}$$

solves (8.21) for $\mu^*$ with $C(\mu^*, P^*) = 1$. Here we denoted $x_{[l]} \in \mathbb{R}^{d_l}$ for those coordinates of $x$ that have distribution $\mu_l^*$. Further, for any marginal with $d_l = 1$, mean $m_l^* \in \mathbb{R}$ and connected support on $\mathbb{R}$, one has the closed-form expression

$$P_l^*(x_{[l]}) = \frac{1}{\mu_l^*(x_{[l]})} \int_{x_{[l]}}^{\infty} (t - m_l^*)\mu_l^*(\mathrm{d}t).$$

The authors of [67] give several examples of optimal metrics for different, typical distributions like Gaussians or Poisson distributions. We now state a further example, which allows us to derive optimal convergence rates of posterior sampling in certain Poisson imaging problems, and connect the Riemannian Langevin dynamics back to mirrored methods and the MLEM optimization methods that we introduced in Section 4.4. The proof is possible via both of the above characterizations, i.e., via moment measures, and as an outer product.

**Corollary 8.20: Optimal preconditioning for gamma distributions**

Suppose $\mu^* = \mu_1^* \otimes \cdots \otimes \mu_d^*$ is an outer product of one-dimensional gamma distributions $\mu_j = \text{Gamma}(\alpha_j, \lambda_j)$, with shape and rate parameters $\alpha_j, \lambda_j > 0$. Then, the optimal metric that solves (8.21) is given by

$$P^*(x) = \text{diag}\left(\frac{x_1}{\lambda_1}, \ldots, \frac{x_d}{\lambda_d}\right).$$

This metric is induced by a mirror function, i.e., $P^* = (\nabla^2 h)^{-1}$ with $h$ the weighted negative entropy function

$$h(x) = \sum_{j=1} \lambda_j (x_j \log(x_j) - x_j). \tag{8.23}$$

*Proof.* By Lemma 8.19, the optimal metric $P^*$ is diagonal with $j$-th entry $P_j^*$ given by

$$P_j^*(x_j) = \frac{1}{\mu_j^*(x_j)} \int_{x_j}^{\infty} (t - m_j^*) \mu_j^*(\mathrm{d}t).$$

Since all marginals are analogous, we drop the index $j$ from here in order to keep notation simple. A gamma distribution $\text{Gamma}(\alpha, \lambda)$ with shape $\alpha$ and rate $\lambda$ has density

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \propto x^{\alpha-1} e^{-\lambda x}$$

and mean $\frac{\alpha}{\lambda}$. Since normalizing constants in the density cancel in the above formula, plugging these terms in gives diagonal entries

$$\begin{aligned}
&x^{1-\alpha} e^{\lambda x} \int_x^\infty \left(t - \frac{\alpha}{\lambda}\right) t^{\alpha-1} e^{-\lambda t} \,\mathrm{d}t \\
&\quad = x^{1-\alpha} e^{\lambda x} \left( \int_x^\infty t^\alpha e^{-\lambda t} \,\mathrm{d}t - \frac{\alpha}{\lambda} \int_x^\infty t^{\alpha-1} e^{-\lambda t} \,\mathrm{d}t \right) \\
&\quad = x^{1-\alpha} e^{\lambda x} \left( \frac{1}{\lambda} \int_{\lambda x}^\infty (u/\lambda)^\alpha e^{-u} \,\mathrm{d}u - \frac{\alpha}{\lambda^2} \int_{\lambda x}^\infty (u/\lambda)^{\alpha-1} e^{-u} \,\mathrm{d}t \right) \\
&\quad = \lambda^{-1-\alpha} x^{1-\alpha} e^{\lambda x} \left( \Gamma(\alpha + 1, \lambda x) - \alpha \Gamma(\alpha, \lambda x) \right).
\end{aligned}$$

In the last line, we introduced the incomplete gamma function $\Gamma(s, v) \coloneqq \int_v^\infty u^{s-1} e^{-u} \,\mathrm{d}u$. By partial integration, it is easy to prove the following recurrence relation

$$\Gamma(s + 1, v) = s\Gamma(s, v) + v^s e^{-v}$$

of the incomplete gamma function. Using this with $s = \alpha$ and $v = \lambda x$, we find

$$x^{1-\alpha} e^{\lambda x} \int_x^\infty \left(t - \frac{\alpha}{\lambda}\right) t^{\alpha-1} e^{-\lambda t} \,\mathrm{d}t = \lambda^{-1-\alpha} x^{1-\alpha} e^{\lambda x} (\lambda x)^\alpha e^{\lambda x} = \frac{x}{\lambda}.$$

This proves the first part of the statement. The second part can be seen by simply differentiating $h$. It holds $\partial_{x_j} h(x) = \lambda_j \log(x_j)$, so the Hessian $\nabla^2 h(x_j)$ is a diagonal matrix with $j$-th diagonal entry $\lambda_j x_j^{-1}$. This shows $P^* = (\nabla^2 h)^{-1}$. □

Note that the proof would have led to the same optimal metric via the strategy of moment measures. As is shown in [64], any probability measure with mean zero that is not supported on a lower-dimensional subspace of $\mathbb{R}^d$ is a moment measure of a suitable, strictly convex $\varphi$. By shifting $\mu^*$ by its mean $m = \frac{\alpha}{\lambda}$ to have mean zero, the resulting measure with density $\nu = \mu^*(\cdot + m)$ can be seen to be a moment measure of

$$\varphi(x) = \sum_{j=1}^{d} \lambda_j \exp\left(\frac{x_j}{\lambda_j}\right) - x_j m_j.$$

This shows that the optimal mirror function for sampling $\nu$ is $\varphi^*(x) = \sum_j \lambda_j (x_j + m_j)(\log(x_j + m_j) - 1)$. After translating the mirror function back by the negative mean $-m$, we obtain the same metric as before.

The relevance of this result for imaging problems becomes clear if we recall that gamma distributions are conjugate to Poisson distributions, and thus naturally come up in problems with Poisson observation data. Assume we have measured data $y$ that follows the observation model

$$y \sim \mathrm{Pois}(Ax),$$

with some sought-for signal $x$ and a diagonal matrix $A = \mathrm{diag}(a_{11}, \dots, a_{dd})$. For diagonal $A$, the Poisson likelihood density (3.13) for a given observation $y$ is given by

$$\rho_{\mathrm{like}}(x \,|\, y) = \prod_{j=1}^{d} \frac{e^{-a_{jj}x_j}(a_{jj}x_j)^{y_j}}{y_j!} \propto \prod_{j=1}^{d} x_j^{y_j} e^{-a_{jj}x_j}. \tag{8.24}$$

Interpreting the right side as an unnormalized density function of $x$ shows that all $x_j$ are mutually independent and marginally follow gamma distributions $x_j \sim \mathrm{Gamma}(y_j + 1, a_{jj})$. In view of Corollary 8.20, this means that the optimal preconditioning for sampling this distribution is reached by the weighted negative entropy mirror function

$$h(x) = \sum_{j=1} a_{jj}(x_j \log(x_j) - x_j), \tag{8.25}$$

which is precisely (8.23) with $\lambda_j = a_{jj}$. We can also verify the optimal diffusion equation (8.22): Suppose that $\mu^* = \rho_{\mathrm{like}}(\cdot \,|\, y)$, then $\mu^* \propto \exp(-V)$ with

$$V(x) = \sum_{j} a_{jj}x_j - y_j \log(x_j).$$

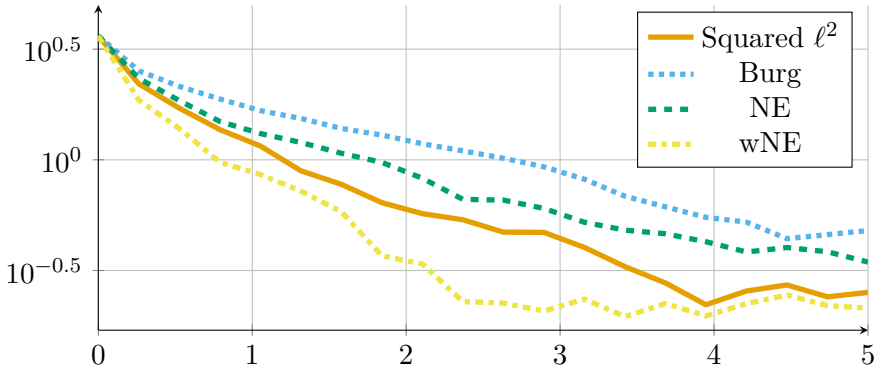



Figure 8.2.: Mirror Langevin dynamics for a gamma distribution and various mirror function choices. The horizontal axis shows SDE time $t$, the vertical axis the Wasserstein distance from the target distribution $\mathcal{W}_2(\mu_t, \mu^*)$, approximated by samples drawn with mirror Langevin algorithm. $\mu^*$ is a highly anisotropic posterior in $d = 2$ dimensions. "Squared $\ell^2$" refers to overdamped Langevin diffusion, implemented via (F*flow*B). "Burg", "NE" and "wNE" correspond to Burg's entropy, negative entropy and weighted negative entropy. The latter is proven to be theoretically optimal, which reflects in the numerical experiment.

The potential's gradient has $j$-th entry $\partial_{x_j} V(x) = a_{jj} - \frac{y_j}{x_j}$. From the mirror function (8.25), we get the inverse Hessian

$$(\nabla^2 h)^{-1}(x) = \operatorname{diag}(a_{11}^{-1} x_1, \dots, a_{dd}^{-1} x_d),$$

the divergence of which has $j$-th entry $(\nabla \cdot (\nabla^2 h)^{-1})_j(x) = a_{jj}^{-1}$. Therefore, the mirrored Langevin dynamics SDE simplifies to

$$\begin{aligned}
\mathrm{d}X_t &= \left( (\nabla \cdot (\nabla^2 h)^{-1})(X_t) - (\nabla^2 h)^{-1}(X_t) \nabla V(X_t) \right) \mathrm{d}t + \sqrt{2 (\nabla^2 h)^{-1}(X_t)}\, \mathrm{d}W_t \\
&= \left( \frac{1}{a} - \frac{X_t}{a} \left( a - \frac{y}{X_t} \right) \right) \mathrm{d}t + \sqrt{2 \frac{X_t}{a}}\, \mathrm{d}W_t \\
&= -\left( X_t - m \right) \mathrm{d}t + \sqrt{2 \frac{X_t}{a}}\, \mathrm{d}W_t,
\end{aligned}$$

where $a = (a_{11}, \dots, a_{dd})^T$, vector division and multiplication is understood componentwise and the mean of the target distribution is $m = \frac{y + \mathbf{1}}{a}$ with $\mathbf{1}$ a vector of constant ones.

We validate the improved convergence behaviour numerically by simulating mirrored Langevin dynamics on a target distribution induced by the likelihood from a Poisson observation model. For that, assume $A$ to be an ill-conditioned forward model, $A = \operatorname{diag}(a_{11}, \dots, a_{dd})$ with $a_{jj} \overset{\text{ind.}}{\sim} \mathrm{Unif}(1, 50)$ randomly drawn and known. We then draw

$y \sim \mathrm{Pois}(Ax^*)$ for some initially chosen, true value $x^* \in \mathbb{R}^d$ and try to solve the Poisson denoising problem by recovering $x$ from the noisy observation $y$. We run this experiment once in a small dimensional example $d = 2$ (where we simply set $A = \mathrm{diag}(1, 50)$), since this allows us to compute estimates of the Wasserstein distances between finite sample distributions and the true distribution of $x$. In Fig. 8.2, we visualize the resulting Wasserstein distances from the true distribution for several instances of preconditioned Langevin diffusion. We compare four different, typical mirror functions:

1. A squared $\ell^2$-norm $h(x) = \frac{\beta}{2}\|x\|^2$. Since the Hessian $\nabla^2 h$ is simply the scaled identity matrix, this leads to overdamped Langevin diffusion. The constant

$$\beta = d\left(\sum_{j=1}^{d} \frac{y_j + 1}{a_{jj}^2}\right)^{-1} \tag{8.26}$$

   effectively leads to a time rescaling and a "fair" comparison with other mirror functions, since the scalar weight ensures that the weighted Euclidean metric satisfies the constraint $\mathbb{E}_{\mu^*}[\mathrm{tr}(P)] = \mathrm{tr}(\mathrm{Cov}_{\mu^*})$ that we optimized the Riemannian Poincaré constant over. We ensure positivity of the samples by including an extra projection step onto $\mathbb{R}^d_{\geq \varepsilon}$ for a small $\varepsilon > 0$, meaning we are effectively running (F*flow*B) with an additional characteristic potential $\chi_{\mathbb{R}^d_{\geq \varepsilon}}$.

2. The weighted negative entropy $h^{\mathrm{wNE}}(x) = \sum_{j=1} a_{jj}(x_j \log(x_j) - x_j)$, which we showed to be theoretically optimal. We also test the more common, unweighted version $h^{\mathrm{NE}}(x) = \beta^{\mathrm{NE}} \sum_{j=1}(x_j \log(x_j) - x_j)$. The weighted version already satisfies the constraint $\mathfrak{P}$, for the unweighted one we can compute that we need to set

$$\beta^{\mathrm{NE}} = \left(\sum_{j=1}^{d} \frac{y_j + 1}{a_{jj}}\right)\left(\sum_{j=1}^{d} \frac{y_j + 1}{a_{jj}^2}\right)^{-1}.$$

3. We also test Burg's entropy $h^{\mathrm{Burg}}(x) = -\beta^{\mathrm{Burg}} \sum_j \log(x_j)$. A simple computation of both side of the constraint $\mathfrak{P}$ shows that we need to normalize with constant

$$\beta^{\mathrm{Burg}} = \left(\sum_{j=1}^{d} \frac{y_j + 1 + (y_j + 1)^2}{a_{jj}^2}\right)\left(\sum_{j=1}^{d} \frac{y_j + 1}{a_{jj}^2}\right)^{-1}.$$

In a second scenario, we set $d = N^2$ and denoise a high-dimensional image. Figure 8.3 shows that for the computation of moment estimates like the sample mean, it also makes a practical difference which preconditioning we choose.

The above denoising scenario is very simplified, since we discarded the prior distribution $\rho_{\mathrm{prior}}$ of $x$ and assumed a perfectly determined denoising setting with diagonal $A$. As we saw in (8.24), this results in an outer product of gamma distributions. In particular, Langevin dynamics are actually not necessary for sampling, as it would be faster

(a) Ground truth $x^*$ (b) Observation $y$

(c) Exact MMSE (d) MMSE, $\ell^2$-squared (e) MMSE, wNE mirror

Figure 8.3.: For a simple denoising setup $y \sim \mathrm{Pois}(Ax)$ with diagonal $A$, we show a ground truth image $x^*$ (a) which is used to generate $y$ (b). The (unregularized) distribution of $x$ implied by the likelihood is an outer product of gamma distributions with known mean (MMSE, c) and variance. We sample from this distribution using mirror Langevin dynamics. At $T = 5$ after $5 \cdot 10^4$ sampling steps, we obtain two different sample mean estimates: (d) with the $\ell^2$-squared mirror (PSNR=11.29 dB to MMSE and (e) with the weighted negative entropy (PSNR=31.08 dB to MMSE).

to directly sample the coordinate-wise gamma distributions marginally. However, the optimality result serves as an algorithmic motivation for more general distributions of $x$: Even if the target does not factorize as a coordinate-wise outer product due to the presence of $\rho_{\text{prior}}$ or a non-diagonal $A$, the weighted negative entropy still ensures positivity and may be an effective preconditioner. Importantly, recall that we showed in Section 4.4 that this same idea leads from the MLEM optimization routine to regularized variants like EM-TV.

For reconstruction problems with Poisson data $y \sim \mathrm{Pois}(Ax)$ and more general $A \in \mathbb{R}^{m\times d}$, the likelihood is

$$\rho_{\mathrm{like}}(x \,|\, y) = \prod_{j=1}^{m} \frac{e^{-(Ax)_j}((Ax)_j)^{y_j}}{y_j!} \propto \prod_{j=1}^{m} ((Ax)_j)^{y_j} e^{-(Ax)_j}.$$

If we now also add prior information again in the form of a Gibbs prior, see (4.7), the resulting target is the posterior

$$\mu^*(x) = \rho_{\mathrm{post}}(x \,|\, y) \propto \rho_{\mathrm{like}}(x \,|\, y)\, \rho_{\mathrm{prior}}(x) \propto \exp(-R(x)) \prod_{j=1}^{m} ((Ax)_j)^{y_j} e^{-(Ax)_j}, \tag{8.27}$$

with a prior potential $R$. Finding the optimal metric for sampling $\mu^*$ is now not as simple as in the diagonal, non-regularized case. We can neither apply Lemma 8.19, nor is there an obvious way to write it as a moment measure of some $\varphi$. We do, however, know a few key facts. If the above distribution of $x$ is non-degenerate with a Lebesgue density (e.g., if the prior is non-degenerate and $A$ has full column rank), it is—after translation by its mean—also a moment measure of some strictly convex function $\varphi$ [64]. This implies that there exists some optimal mirror function $h = \varphi^*$ that achieves $C_{\mathrm{RPI}}(\mu^*, (\nabla^2 h)^{-1}) = 1$. In general, the weighted negative entropy mirror is now sub-optimal in terms of the Riemannian Poincaré constant. It could, however, still lead to fast schemes and is again well-motivated from the optimization viewpoint: As we saw in Section 4.4, algorithms like MLEM or EM-TV can be interpreted as preconditioned gradient methods, where the preconditioner is the inverse Hessian of

$$h(x) = \sum_{j=1}^{d} (A^*\mathbf{1})_j (x_j \log(x_j) - x_j), \tag{8.28}$$

see (4.43). This mirror function is another instance of (8.23) with $\lambda_j = (A^*\mathbf{1})_j$ and is a direct generalization of (8.25) for non-diagonal $A$. In other works focused on inverse problems with Poisson data, it was proposed to replace the weights $(A^*\mathbf{1})_j$ by general weights or potentially even update them dynamically along the iteration [33, 28]. The presence of the prior additionally modifies the geometry of the posterior—whenever it has a strong effect, for example, if the prior is very concentrated compared to the likelihood distribution, this may result in a larger $C_{\mathrm{RPI}}$. On the other hand, the experiments also indicate that (8.28) is a good mirror function choice for the posterior whenever the noise level is small, i.e., whenever the prior density tends to have smaller variation and larger uncertainty than the likelihood. We conjecture that the Riemannian Poincaré inequality is satisfied by (8.28) with a bounded, "small" constant $C_{\mathrm{RPI}}$ for more general, sufficiently regular $A$ and $R$. Equally, under certain regularity assumptions on $A$ like diagonal dominance, it appears intuitive that the Riemannian Poincaré inequality may hold with the mirror function (8.25). Proving these results is beyond the scope of this thesis and will be left for future work.

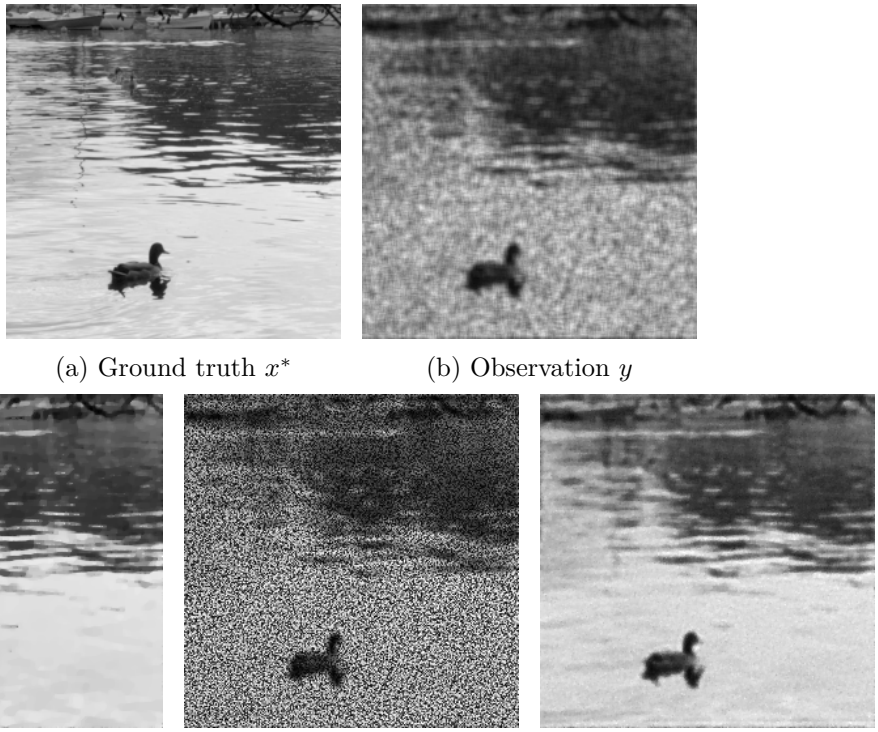
(a) MAP (b) MMSE, $\ell^2$-squared (c) MMSE, wNE mirror

Figure 8.4.: For the same denoising setup $y \sim \mathrm{Pois}(Ax)$ with known diagonal $A$ and a (Huber-smoothed) TV prior, we sample from the posterior of $x$ using mirror Langevin dynamics. The ground truth and noisy data are shown in Fig. 8.3. The MMSE estimate is not available exactly, we show as a rough visual approximation the MAP estimate (a) which was obtained via gradient descent. Mirror Langevin sampling yields the shown sample mean estimates at $T = 5$ after $5 \cdot 10^4$ steps with $\ell^2$-squared mirror (b) and with weighted negative entropy mirror (c).

Before moving on, we also show empirical evidence for the general case from two numerical experiments that suggest that the weighted negative entropy function (8.28) may be a good mirror function choice. Firstly, Fig. 8.4 shows the same denoising experiment as Fig. 8.3, but with an isotropic total variation Gibbs prior applied. Since the mirror Langevin update requires $\log \mu^*$ to be smooth, we smooth the TV regularization term (4.15) in this experiment by replacing the outer $\ell^1$-norm with its Moreau–Yosida envelope, leading to so-called Huberized total variation. The scalar temperature $\beta$ of the potential is pre-selected by hand. As the resulting sample mean estimates after $5 \cdot 10^4$ sampling steps show, the additional TV term does not change the fact the the anisotropy is much better handled by weighted negative entropy preconditioner.

In the second experiment, we also modify the forward model $A$ and replace the denoising setup with a deblurring task. Instead of a diagonal matrix, $A$ is now the product of the former diagonal matrix with entries of varying magnitude, and a uniform blur kernel. This design implies that the likelihood does not factorize as an outer product and is still highly anisotropic. Moving to non-diagonal $A$ requires us to replace (8.25) by its generalization (8.28). As before, the imposed prior is a Huberized isotropic total variation Gibbs prior with a pre-selected temperature. The results still exhibit similar behaviour as in the simpler denoising applications, as reflected by the moment estimates shown in Fig. 8.5.

Before moving on, we briefly outline what is known about the convergence of the time discretized sampling scheme (MLA) that arises from the mirror Langevin dynamics. Expectably, the convergence behaviour now depends on both the potential $V$ in the



(a) Ground truth $x^*$ (b) Observation $y$

(c) MAP (d) MMSE, $\ell^2$-squared (e) MMSE, wNE mirror

Figure 8.5.: We consider a deblurring setup $y \sim \text{Pois}(Ax)$, where $A = UD$, with $U$ a uniform blur matrix and $D$ a diagonal matrix with entries of varying magnitude. Due to the blur kernel, reconstruction from data $y$ (b) is ill-posed. For a (Huber-smoothed) TV prior, we sample from the posterior of $x$ using mirror Langevin dynamics. For reference, the MAP is shown in (c). Mirror Langevin sampling yields the sample mean estimates at $T = 5$ after $5 \cdot 10^4$ steps with $\ell^2$-squared mirror (d) and with weighted negative entropy mirror (e).

target and the chosen mirror function $h$. As for other Langevin sampling methods, typical assumptions are strong convexity and a smooth Lipschitz gradient, however, as is common for mirror descent in optimization, these properties have to be understood relative to the mirror function $h$:

- The potential $V$ is called $\omega$-strongly convex relative to $h$ if

$$\langle \nabla V(x) - \nabla V(\tilde{x}), \nabla h(x) - \nabla h(\tilde{x}) \rangle \geq \omega \|\nabla h(x) - \nabla h(\tilde{x})\|^2, \tag{8.29}$$

  for all $x, \tilde{x} \in \Omega_h$. Via the substitution $z = \nabla h(x), \tilde{z} = \nabla h(\tilde{x})$, this is equivalent to saying that $\nabla V \circ \nabla h^*$ is $\omega$-strongly monotone, which can be understood as $V$ being strongly convex along geodesics on the Riemannian manifold induced by the metric $(\nabla^2 h)^{-1}$.

- The gradient $\nabla V$ is $L$-Lipschitz continuous relative to $h$ if

$$\|\nabla V(x) - \nabla V(\tilde{x})\| \leq L\|\nabla h(x) - \nabla h(\tilde{x})\|^2, \tag{8.30}$$

  for all $x, \tilde{x} \in \Omega_h$. This is equivalent to saying $\nabla V \circ \nabla h^*$ is $L$-Lipschitz continuous.

Unfortunately, these two conditions do not seem sufficient to derive convergence results just yet, since it is in general not enough that the mirror map $h$ is of Legendre type. In [130], the ergodicity proof of (MLA) employs the following, third condition:

- We say that $h$ satisfies a modified self-concordance condition with parameter $\alpha > 0$, if $x \mapsto \sqrt{(\nabla^2 h^*)^{-1}(x)}$ is $\alpha$-Lipschitz continuous on $\mathbb{R}^d$ in the Hilvert-Schmidt norm. This is equivalent to requiring

$$\left\| \sqrt{\nabla^2 h(x)} - \sqrt{\nabla^2 h(\tilde{x})} \right\|_{\mathrm{HS}} \leq \alpha \|\nabla h(x) - \nabla h(\tilde{x})\|, \tag{8.31}$$

  for all $x, \tilde{x} \in \Omega_h$, where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm.

The modified self-concordance condition replaced the similar self-concordance conditions in the earlier works [225, 2]. Under the three assumptions above, we have the following ergodicity result, see [130, Thm. 3.1].

**Theorem 8.21: (MLA): Wasserstein bias**

Assume that $V$ is $\omega$-strongly convex relative to $h$, differentiable with $\nabla V$ $L$-Lipschitz continuous relative to $h$, and $h$ satisfies modified self-concordance with $\alpha < \omega$. Then there exists a maximum step size $\tau_{\max}$ and a constant $C > 0$, such that the distributions $\mu^{(n)}$ generated by (MLA) with $\tau \leq \tau_{\max}$ satisfy the bound

$$\mathcal{W}_{h,2}(\mu^{(n)}, \mu^*) \leq \sqrt{2} e^{-(\omega-\alpha)n\tau} \mathcal{W}_{h,2}(\mu^{(0)}, \mu^*) + C\tau^{\frac{1}{2}}.$$

Here $\mathcal{W}_{h,2}$, defined in (7.35), denotes the Wasserstein distance on the Riemannian manifold induced by the metric $(\nabla^2 h)^{-1}$.

For the full proof, we refer to [130]. It reveals the reason why the third condition on modified self-concordance is assumed. While the weaker Riemannian Poincaré inequality gives ergodicity of the continuous time regime in chi-squared and KL divergence, see Theorem 8.18, the $\omega$-strong convexity and $\alpha$-self-concordance condition together ensure that this also holds in Wasserstein distance, with rate $\omega - \alpha$. The three assumptions together then guarantee that the discrete time algorithm is also ergodic, by a proof that is again based on the now familiar synchronous coupling technique.

# Chapter 9

# Uncertainty Quantification in Compton Camera Imaging

In this chapter, we apply the sampling algorithms studied in the last chapters to the Compton camera imaging problem. In Chapter 5, we have seen that the Compton imaging forward models developed in this work are suitable for MAP estimation in practical nuclear decommissioning scenarios. Given a gamma ray measurement, MAP estimation gives a single estimate of the most likely contamination situation under the posterior.

In practice, however, more complex questions on the structure of the posterior may arise, which cannot be answered by a single point estimate. For example, practitioners may require estimates of the probability that a given nuclide is present in an environment, or the probability that radioactivity exceeds certain thresholds. These questions lead to Bayesian inference tasks like credible set estimation and model selection, introduced in Section 4.3. In order to solve these tasks, algorithms have to go beyond point estimation and draw representative samples from the posterior distribution.

## 9.1. Algorithmic Setup

**Likelihood.** We work with the Compton imaging forward models summarized in Section 3.7. Assume that we have a bin-mode discretized version of the operator $\mathcal{P}$, precomputed as in Section 5.1. Given such a matrix $P$, the Poisson statistics of the data imply that the negative log-likelihood $l$ is given by

$$l(f) \coloneqq -\log \rho_{\text{like}}(g \,|\, f) = \sum_{j=1}^{J} (Pf)_j + b_j - g_j \log\left((Pf)_j + b_j\right) + \text{const.} \tag{9.1}$$

Since the models $P$ are all linear, $l$ is a convex function. The data $g$ is, as before, either simulated by ray-tracing Monte Carlo or acquired with a physical camera in lab settings. Any simulated data is synthetically corrupted by adding lab-measured background radiation. The background correction $b$ is sampled from the same background measurements, ensuring that the corruption and correction data are not identical, but follow the same distribution.

**Prior distribution.** If we impose a log-concave prior distribution on the reconstructed activity image, the Bayesian posteriors we aim to sample from are log-concave. In Chapter 5, we exploited this to apply convex optimization algorithms on MAP computation problems. Due to the broad range of possible sources, we choose a non-restrictive Gibbs distribution with total variation energy and preselected temperature $\lambda$ as image prior. This defines the posterior

$$-\log \rho_{\text{post}}(f \mid g) = \lambda \,\mathrm{TV}(f) + l(f).$$

In Chapter 6 through Chapter 8, we developed extensive theory and practical insights on Langevin algorithms for log-concave sampling. Among these results, our new developments on optimal mirror Langevin sampling for problems with Poisson data are now particularly relevant.

**Choice of mirror maps.** In Section 8.4, we saw that in simplified Poisson data settings, the weighted negative entropy mirror map reaches optimal Poincaré constant. This ensures rapid convergence of mirrored samplers for these target distributions. Experiments confirmed this theoretical finding: with simpler image denoising and deblurring tasks, the weighted negative entropy achieved significant convergence speed-up if the posterior distribution is very anisotropic. Due to the Poisson nature of the data, this makes the mirrored schemes interesting for the Compton imaging problems considered in this work. By construction, the imaging problem in Compton cameras yields highly anisotropic Bayesian posteriors: As the activity map is usually sparse, the posterior variance is very small in regions of low activity. In highly active regions, however, the activity can span several orders of magnitude in the kBq to MBq regime, making the sampling problem highly ill-conditioned.

In order to test mirrored samplers on the Compton imaging problem, we run (MLA), the time-discretization of the mirrored Langevin dynamics (7.38). The sampler takes an initial guess $X^{(0)}$ and is implemented with a dual variable $Y^{(n)}$ giving the update

$$\begin{aligned}
\xi^{(n+1)} &\sim \mathrm{N}(0, I) \\
Y^{(n+1)} &= \nabla h(X^{(n)}) - \tau \nabla \left( l(X^{(n)}) + \lambda \,\mathrm{TV}^{\gamma}(X^{(n)}) \right) + \sqrt{2\tau \nabla^2 h(X^{(n)})}\,\xi^{(n+1)} \\
X^{(n+1)} &= \nabla h^* \left( Y^{(n+1)} \right)
\end{aligned}$$

where $\xi^{(n)}$ are mutually independent. As the total variation functional is non-smooth, we replace it by the regularized Huber-TV functional, which is obtained by replacing the $\ell^1$ norm in the definition $\mathrm{TV}(x) = \|Dx\|_{2,1}$ by its Moreau envelope. This is a very similar strategy to the aforementioned Moreau-Yosida variants of unadjusted Langevin sampling—implicit variants based on backward steps are not directly applicable here due to the mirrored nature of the sampler. The values $X^{(n)}$ are extracted as representative samples and used for downstream computations like posterior mean and variance estimation.

We employ two different mirror maps: we test the squared $\ell^2$-norm as a mirror function

$$h(x) = \frac{1}{2}\|x\|^2,$$

which reduces to the unmirrored special case of unadjusted Langevin diffusion. Positivity of samples is ensured by inserting an additional projection after the second step which maps $Y^{(n+1)}$ to the positive orthant. This ensures that $\nabla l(X^{(n)})$ is well-defined and makes the iteration coincide with (F*flow*B) via adding an indicator on the positive orthant to the potential. The time rescaling coefficient $\beta$ from Eq. (8.26) is absorbed into the step size $\tau$ of the sampler.

The second mirror map that we test is the weighted negative entropy

$$h(x) = \sum_{j=1}^{d} (P^*\mathbf{1})_j (x_j \log(x_j) - x_j),$$

where $P$ is the Compton imaging forward model. This mirror map ensures that samples stay in the positive orthant and greatly improved convergence speed in our earlier experiments on image denoising and deblurring.

**MCMC diagnostics.** To assess the convergence quality of the samplers, we employ standard MCMC diagnostics. For a stationary Markov chain $(X^{(n)})_{n\geq 0}$, the *autocorrelation function* (ACF) at lag $k$ measures the correlation between samples $k$ steps apart:

$$\mathrm{ACF}(k) := \frac{\mathrm{Cov}(X^{(n)}, X^{(n+k)})}{\mathrm{Var}(X^{(n)})}.$$

For independent samples, $\mathrm{ACF}(k) = 0$ for all $k > 0$; correlated MCMC samples exhibit positive autocorrelation that decays with lag. For better mixing, a rapid decay is preferable. The *integrated autocorrelation time* (IACT) $\tau_{\mathrm{int}} = 1 + 2\sum_{k=1}^{\infty} \mathrm{ACF}(k)$ quantifies how many MCMC steps are equivalent to one independent sample. In practical experiments, the infinite sum is truncated. The *effective sample size* $\mathrm{ESS} = N/\tau_{\mathrm{int}}$, where $N$ is the total number of samples, estimates the equivalent number of independent samples obtained from the chain. Higher ESS indicates better mixing; for equal computational budget, a sampler with higher ESS produces more accurate Monte Carlo estimates.

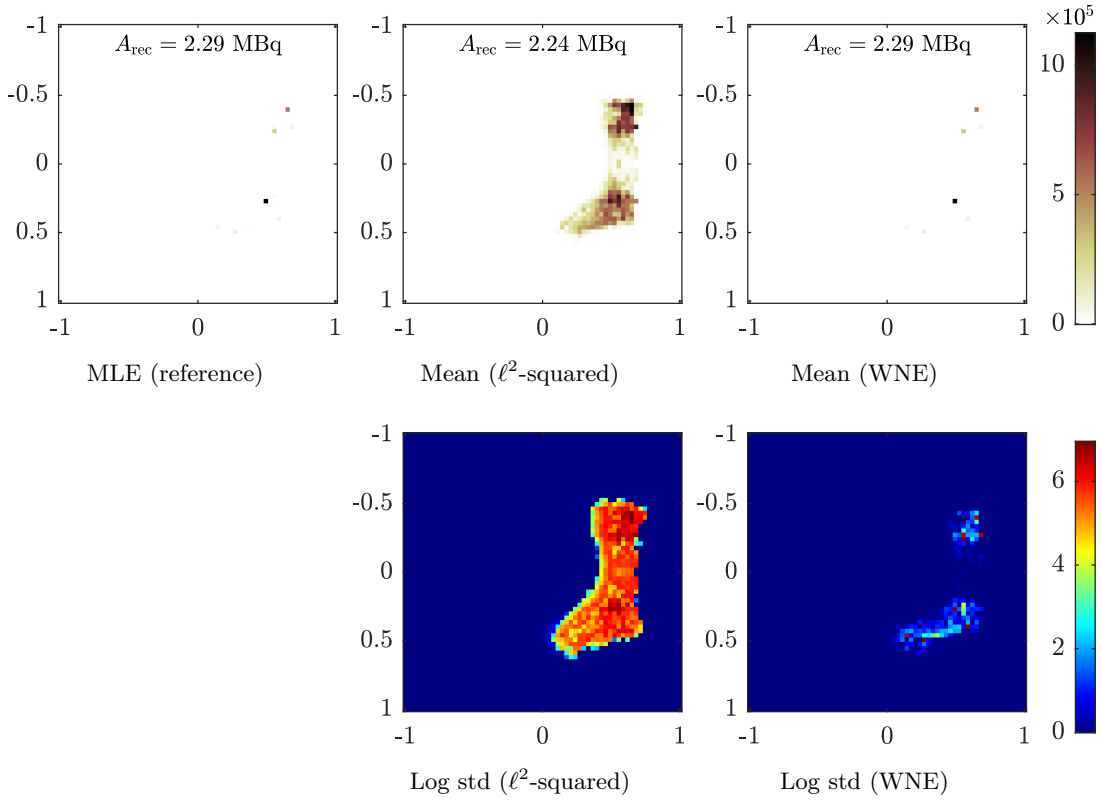
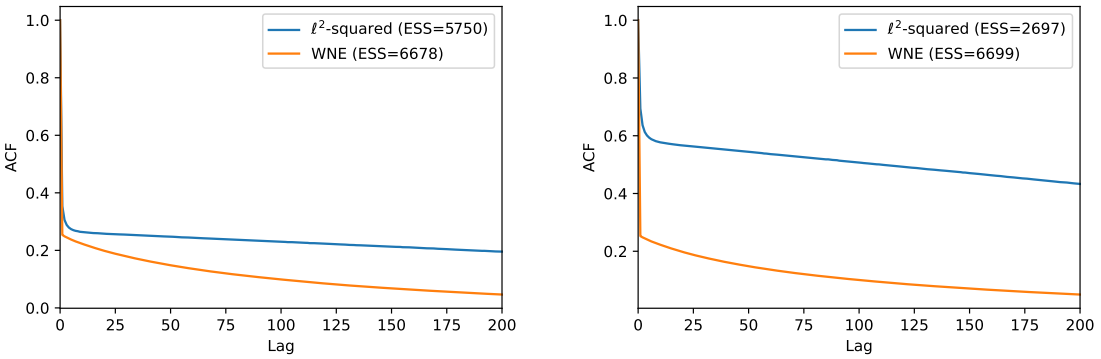



Figure 9.1.: Comparison of mirror functions for likelihood-only sampling from simulated Caesium-137 data. Top row: MLE reconstruction (left) and conditional mean estimates using the $\ell^2$-squared mirror (center) and weighted negative entropy mirror (right). Bottom row: Logarithm of the pixelwise standard deviation. The $\ell^2$-squared sampler shows artifacts from incomplete convergence, while the WNE sampler produces clean estimates matching the MLE.

## 9.2. Results

In this section, we present sampling results for the Compton camera imaging problem. Similar to Chapter 5, we focus on the RSL7 camera setup and reconstruct distributions from simulated and real measurements of the nuclides listed in Table 2.1. The experiments are designed to validate the sampling algorithms and demonstrate their capability to provide uncertainty estimates beyond point reconstructions.

**Mirror function comparison** Before evaluating experiments of varying difficulty, we validate our hypothesis on the mirror function choice, see Figs. 9.1 and 9.3. We reconstruct simple simulated Caesium-137 point sources at 1 m distance from the RSL7 camera, using clean data without energy noise or background radiation and the single




(a) Likelihood-only sampling

(b) TV-regularized posterior

Figure 9.2.: Mean autocorrelation function (ACF) across all pixels for the $\ell^2$-squared mirror and weighted negative entropy (WNE) mirror. Left: Likelihood-only sampling. Right: TV-regularized posterior sampling. The ACF for the WNE mirror decays significantly faster, indicating better mixing. The corresponding estimations of the effective sample sizes (ESS) for $10^4$ samples are reported in the legends.

scattering forward model $P = P^{(1)}$. First, we run the Langevin sampler to draw samples from the likelihood distribution, without imposing a prior. From these samples, we estimate the conditional mean $\mathbb{E}[f \mid g]$ and the pixelwise standard deviation $\sqrt{\mathrm{Var}(f_j \mid g)}$ via Monte Carlo integration.

We compare the weighted negative entropy mirror to the standard $\ell^2$-squared mirror by running both samplers for $10^5$ iterations each. In order to create a fair comparison, the largest possible step sizes guaranteeing ergodicity are used for both samplers. Convergence quality is assessed both visually, through the estimated mean and standard deviation images, and quantitatively through MCMC diagnostics.

For the likelihood-only experiment, the weighted negative entropy mirror achieves substantially faster convergence than the $\ell^2$-squared mirror. This is evident in both the visual quality of the estimates and the diagnostic metrics. When comparing the estimated mean and log-standard deviation images, the $\ell^2$-squared sampler exhibits visible artifacts from incomplete convergence, whereas the mirrored sampler produces clean estimates that closely match the MLE reconstruction (Fig. 9.1). The ACF curves (for the first $10^4$ samples) in Fig. 9.2 quantify this difference: the WNE mirror's ACF decays rapidly to zero, while the $\ell^2$-squared mirror retains significant autocorrelation even at large lags. Adding a TV prior does not fundamentally change this convergence behavior: experiments on the regularized posterior confirm that the weighted negative entropy mirror maintains its advantage over the standard $\ell^2$-squared geometry (Fig. 9.3).

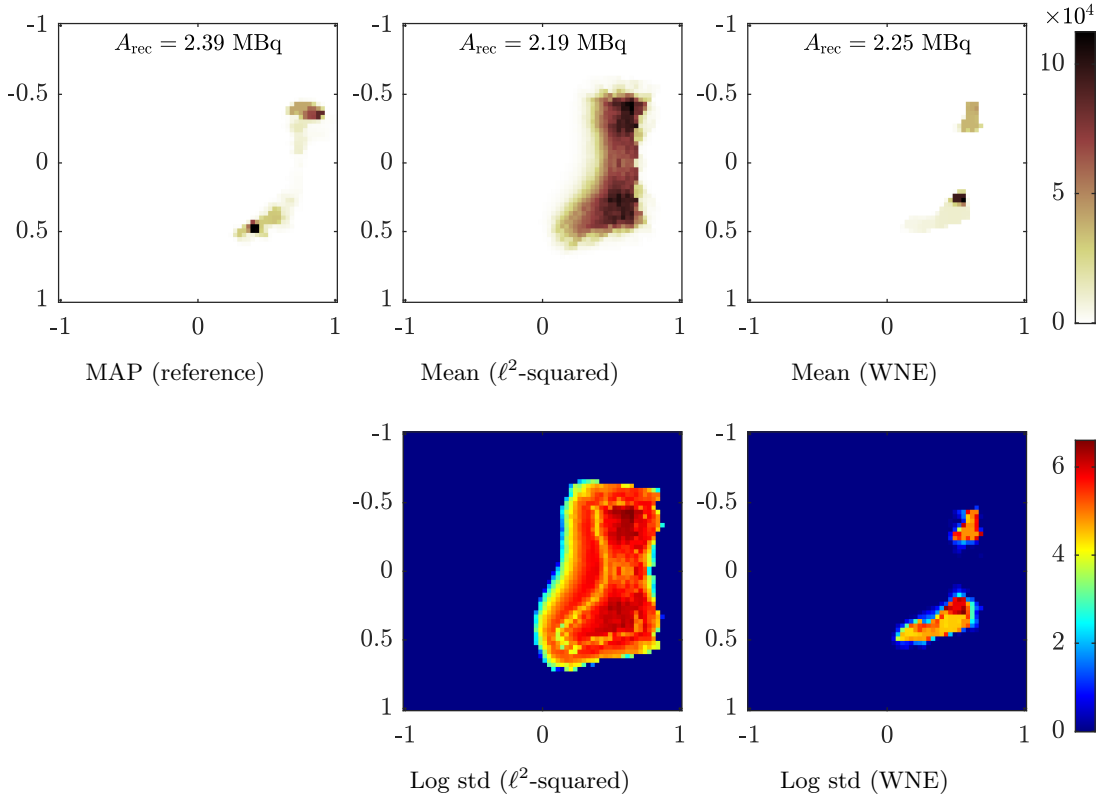



Figure 9.3.: Comparison of mirror functions for TV-regularized posterior sampling. Top row: MAP reconstruction (left) and posterior mean estimates using the $\ell^2$-squared mirror (center) and weighted negative entropy mirror (right). Bottom row: Logarithm of the pixelwise posterior standard deviation. The convergence advantage of the WNE mirror persists when adding the TV prior, confirming that the mirror geometry remains beneficial for the regularized problem.

**Single scattering model** We employ (MLA) with WNE mirror for all further experiments. In the simplest case, we have clean, simulated point source data. Figure 9.4 shows the estimated conditional mean and logarithmic standard deviation in such a setting with Caesium-137 sources. The moments are estimated from $10^5$ samples each. The largest stable step size $\tau$ for this mirror function and forward model is tuned by hand once and kept constant across experiments with the same forward model. The means successfully localize the source positions, resembling the MLE reconstructions from Chapter 5. The logarithmic standard deviation images reveal the spatial structure of uncertainty across several orders of magnitude. For multiple point sources, the standard deviation maps show increased uncertainty in regions between sources, where the overlapping point spread functions make separation more difficult.

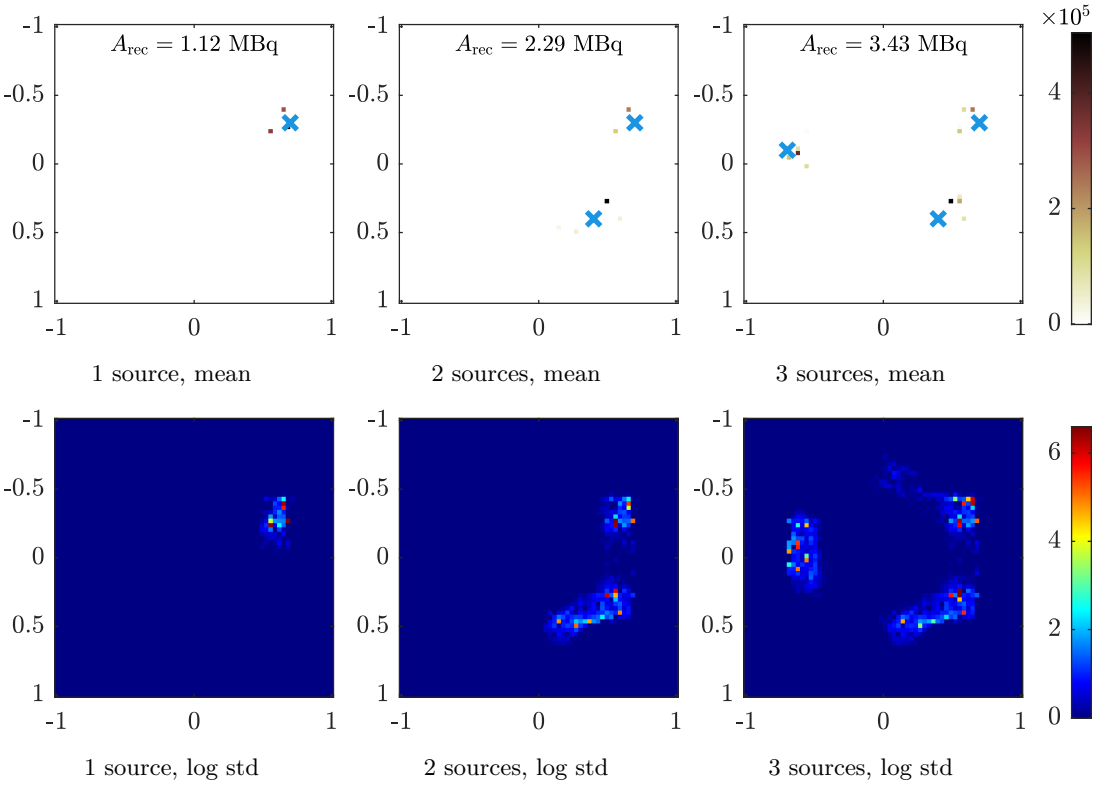



Figure 9.4.: Conditional mean and pixelwise standard deviation for Caesium-137 point source reconstructions from simulated data, using only the likelihood distribution without prior regularization. Top row: Conditional mean estimates for one, two, and three point sources. Bottom row: Logarithm of the pixelwise standard deviation. The mean estimates successfully localize the source positions, while the standard deviation maps reveal the spatial structure of uncertainty.

Next, we consider noisy data subject to Poisson statistics, noisy energy measurements and background radiation. As demonstrated in Chapter 5, reconstructions without regularization are not effective here. We impose a total variation prior on the reconstructed activity, sampling from the posterior distribution $\rho_{\text{post}}(f \mid g) \propto \exp(-\lambda \,\text{TV}(f) - l(f))$ with temperature $\lambda > 0$. This allows us to stabilize the point estimates and quantify how prior assumptions affect the uncertainty estimates.

Figure 9.5 shows the posterior mean and logarithmic standard deviation when incorporating the TV prior. The effect of regularization is clearly visible: The posterior mean and standard deviation estimates are sharply localized in areas around the true source positions, suppressing diffuse activity visible in likelihood-only reconstructions. Just as with MAP estimates, however, the reconstructions do show uncertainty in larger regions around true source positions due to the camera's limited spatial resolution. In the

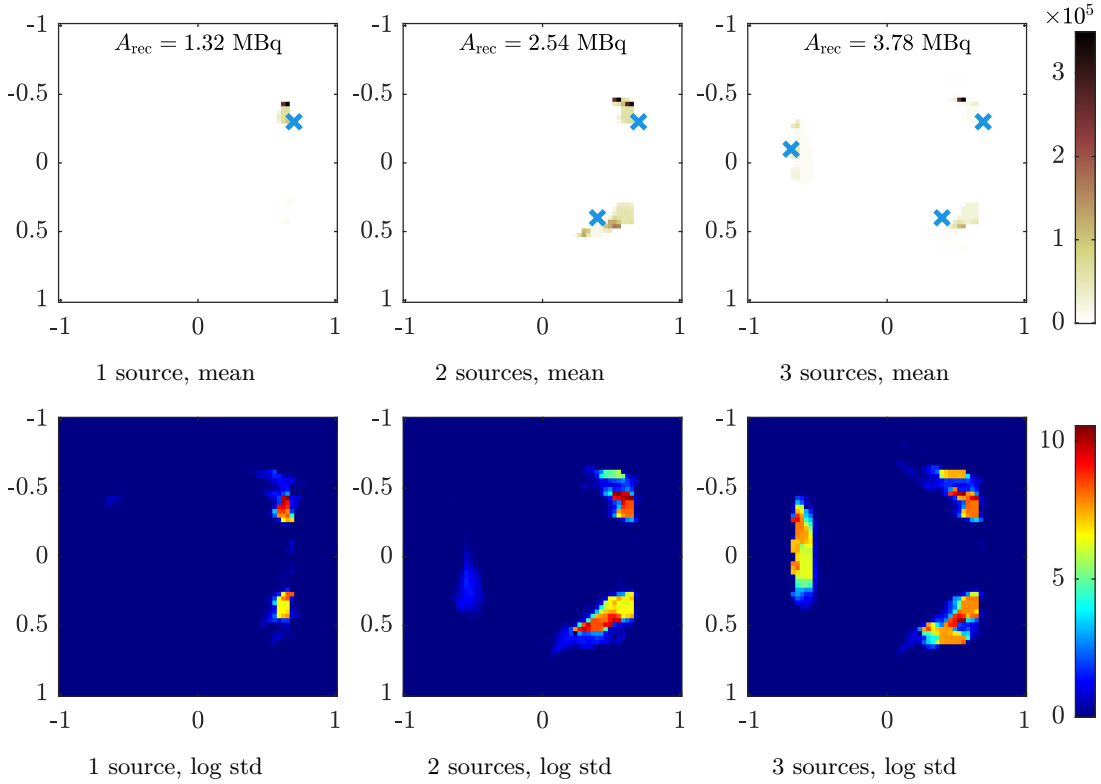



Figure 9.5.: Posterior mean and pixelwise standard deviation for Caesium-137 point source reconstructions with total variation prior. Top row: Posterior mean estimates for one, two, and three point sources. Bottom row: Logarithm of the pixelwise posterior standard deviation. Compared to the likelihood-only reconstruction in Fig. 9.4, the TV prior more clearly demarcates source estimates given the inherent uncertainty of the camera.

single source experiment, the log-variance map exposes an artifact, likely due to noise or background radiation, which could go unnoticed in the MMSE estimate. The results demonstrate that the sampling framework naturally incorporates prior information, providing uncertainty estimates that reflect both data constraints and prior assumptions.

**Double scatter correction.** In Chapter 5, we demonstrated that the scatter-corrected forward model $P^{(1)} + P^{(2)}$ substantially improves MAP reconstruction from data containing multiply scattered photons. We repeat this experiment in the sampling setting. Figure 9.6 shows posterior mean estimates from double scattered data $g^{(1)} + g^{(2)}$ with background radiation. The uncorrected model $P^{(1)}$ either produces mislocalized sources or completely diverges for the same algorithmic configuration. The scatter-corrected model $P^{(1)} + P^{(2)}$, however, accurately recovers the true source positions.

The improvement is consistent with the MAP results from Fig. 5.4, confirming that the scatter correction transfers directly to the framework of uncertainty quantification.

**Full spectrum reconstruction.** Finally, we reconstruct from the full measured spectrum $g = g^{(1)} + g^{(2)} + \ldots$, which includes all scattering orders. As in Chapter 5, we use the scatter-corrected model $P^{(1)} + P^{(2)}$ and treat higher-order scattering as noise. Figure 9.7 shows the posterior mean and log standard deviation. Compared to the double scatter case, artifacts appear due to the unmodeled higher-order terms. The increased uncertainty is attributed to the fact that the extra counts are not explainable by the forward model. Nevertheless, source positions are visibly recovered.

**Caesium-137 lab measurements.** We apply the sampling framework to real Caesium-137 point source measurements from the RSL7 camera, the same data used for MAP estimation in Chapter 5. Figure 9.8 shows posterior mean and log standard deviation for three source positions at varying distances from the camera center. The full set of measurements is evaluated in Appendix A. The generated variance maps reveal more information about the structural uncertainty of reconstructions with this Compton camera: As the images clearly show, measurements in central regions of the FOV reconstruct source positions reliably, but exhibit increased spatial uncertainty for the same prior. This could be attributed to a larger sensitivity to angular changes in this area. On the border of the FOV, the reconstructed variance maps are more localized, but tend to have a structured spatial bias with respect to the true source position, which could be due to modeling errors and secondary scattering effects in real lab environments.

**Cobalt-60 simulations and measurements—leakage corrected model.** For polyenergetic nuclides like Cobalt-60, the energy-leakage-corrected forward model from Section 5.1 is necessary to account for incomplete absorption of higher emission lines. In Chapter 5, we demonstrated that the leakage correction improves MAP reconstructions by removing boundary artifacts caused by unmodeled counts. Figure 9.9 shows posterior mean and log standard deviation estimates from simulated Cobalt-60 point sources using the corrected model. The posterior means successfully localize the source positions, consistent with the MAP results from Fig. 5.8. The log standard deviation maps reveal that uncertainty is concentrated around the true source locations, although artifacts appear for strong contaminations and multiple sources.

We extend the analysis to real Cobalt-60 measurements from the RSL7 camera, using the same lab data as in Fig. 5.9. Figure 9.10 shows posterior mean and log standard deviation for the line source and three area sources. The posterior means recover the source positions consistently with the MAP reconstructions from Chapter 5. The log standard deviation maps reveal the spatial structure of uncertainty in these extended source configurations. For the line source, the uncertainty is distributed along the source

extent, reflecting the elongated geometry. For area sources closer to the camera axis, the variance is more concentrated, while sources at the edge of the field of view exhibit increased spatial uncertainty. These results demonstrate that the sampling framework with energy leakage correction transfers successfully from simulated to real measurement scenarios.

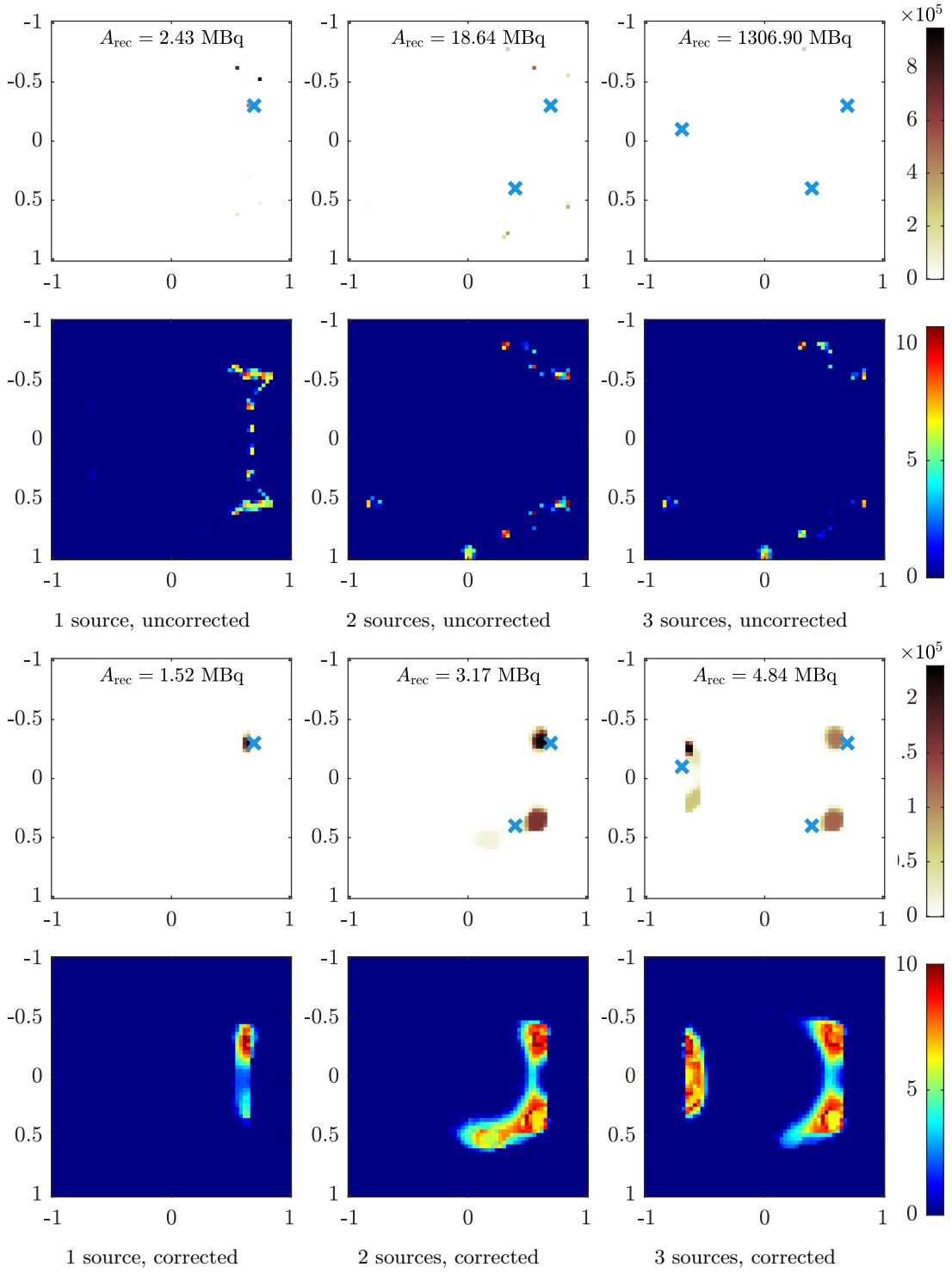



Figure 9.6.: Posterior estimates from double scattered data $g^{(1)} + g^{(2)}$ with background. Top two rows: MMSE reconstruction and log standard deviation using the uncorrected single scatter model $P^{(1)}$ (the large values indicate divergence of the algorithm due to the mismatch between model and data). Bottom two rows: Reconstruction with the scatter-corrected model $P^{(1)} + P^{(2)}$. As with MAP estimation (Fig. 5.4), the corrected model substantially improves source localization.

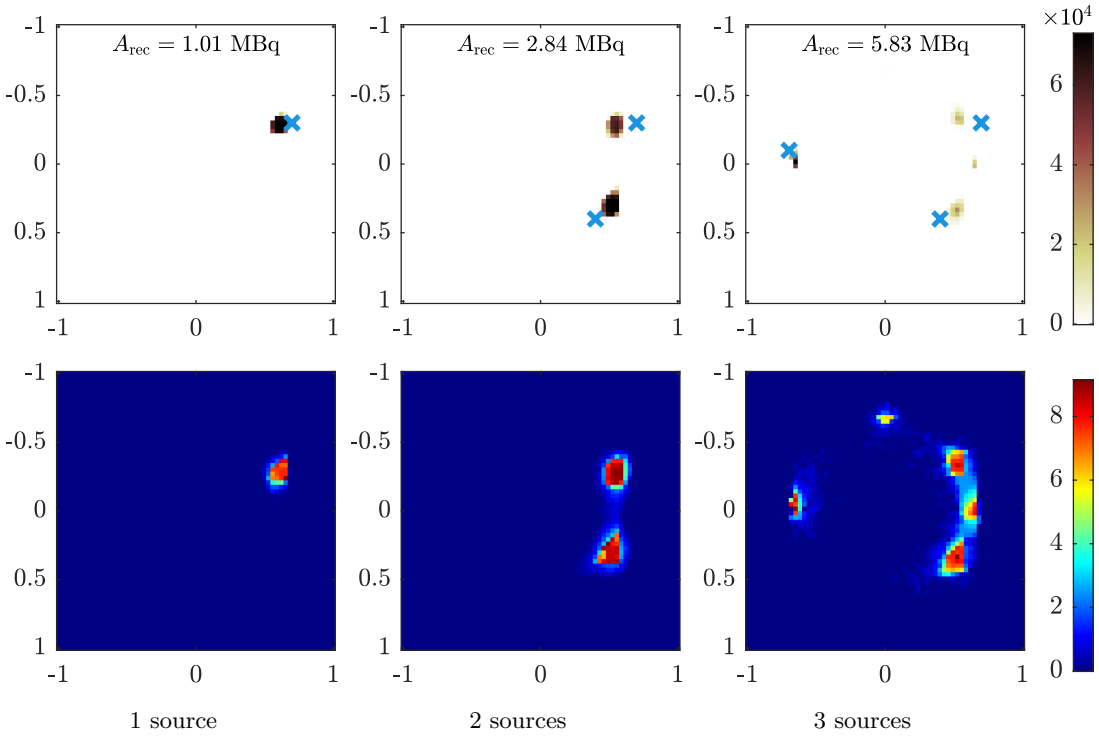



Figure 9.7.: Posterior mean (top) and log standard deviation (bottom) from full spectrum data $g = g^{(1)} + g^{(2)} + \dots$ using the scatter-corrected model $P^{(1)} + P^{(2)}$. Compared to the double scatter case (Fig. 9.6), slight artifacts appear due to higher-order scattering acting as noise, but source localization remains reliable.

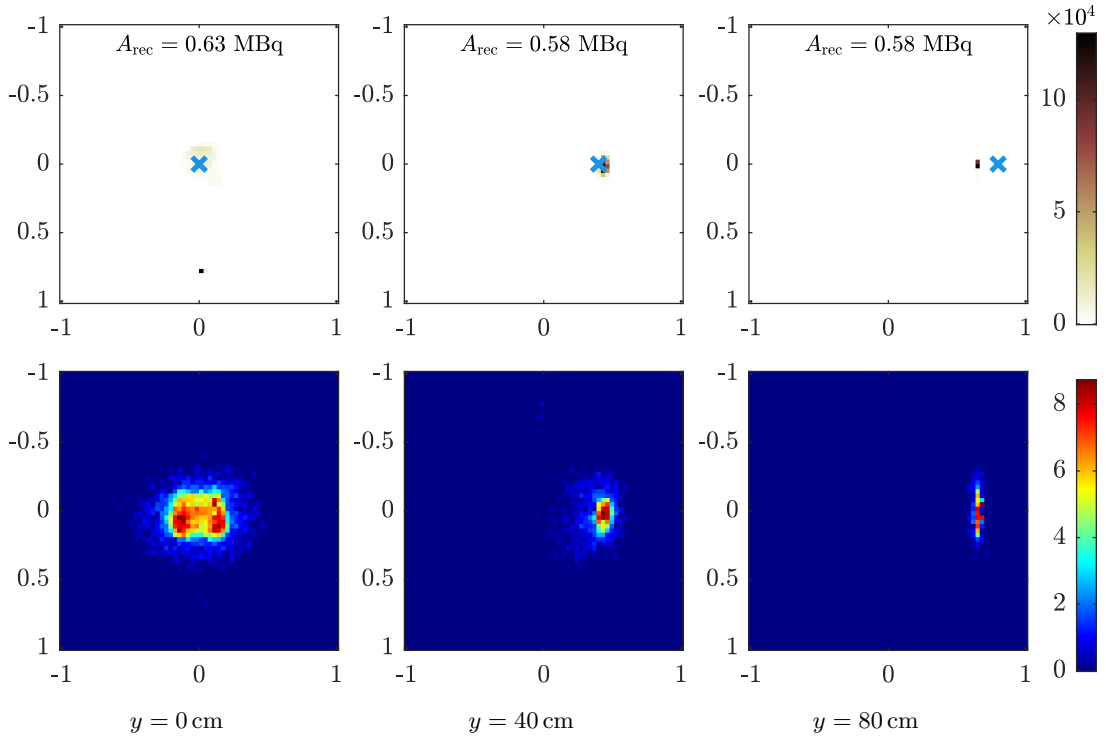



Figure 9.8.: Posterior mean (top) and log standard deviation (bottom) from real Caesium-137 point source (0.61 MBq) measurements at three positions. The true source positions are marked in blue. Consistent with MAP estimates (Fig. 5.7), central positions yield reliable localization while uncertainty increases towards the field of view boundary.

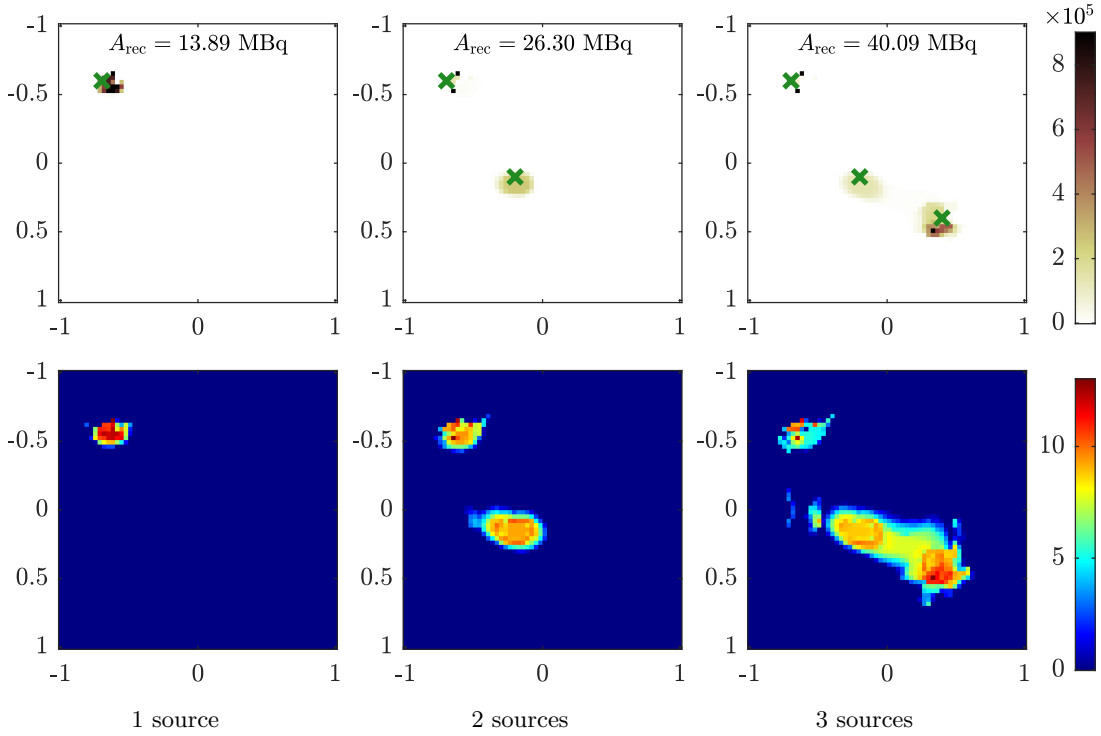



Figure 9.9.: Posterior mean (top) and log standard deviation (bottom) from simulated Cobalt-60 data using the energy leakage corrected forward model. From left to right: one, two and three point sources, 10 MBq each, true positions indicated in green.

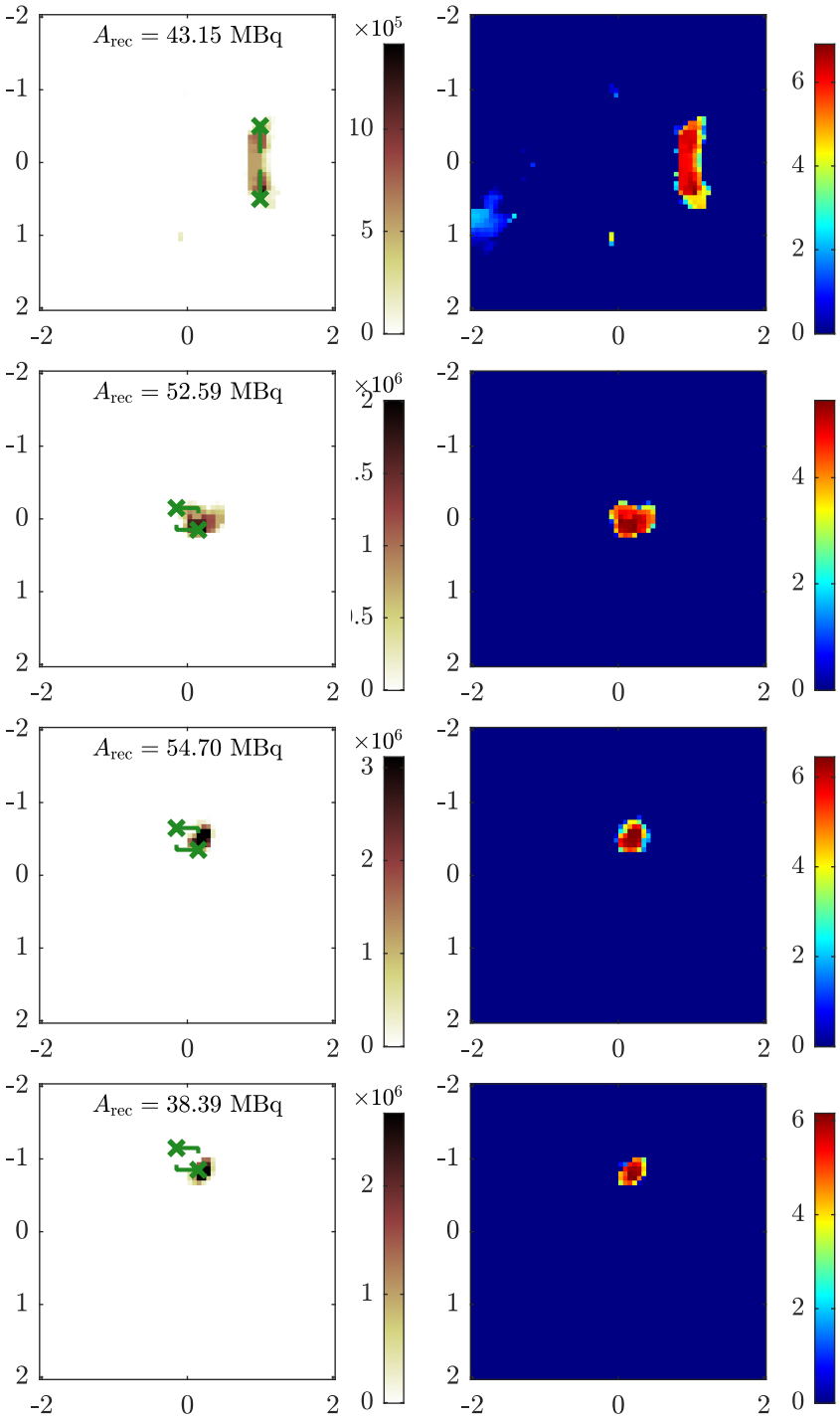



Figure 9.10.: Posterior mean (left) and log standard deviation (right) from measured Cobalt-60 data using the energy leakage corrected forward model. The sources are equal to Fig. 5.9: a 1 m line and three $0.3 \times 0.3$ m area sources.

# Chapter 10

# Conclusion and Open Questions

In this thesis, we have contributed new theory and empirical insights on two broad aspects of image reconstruction: Compton camera imaging for nuclear decommissioning applications, and Langevin sampling algorithms for Bayesian uncertainty quantification in imaging inverse problems.

**Compton Camera modeling and reconstruction.** For the problem of reconstructing contamination activity maps from Compton camera measurements, we formulated the forward model as a linear integral operator based on a conical Radon transform. We showed that the standard single scattering model is computationally convenient and reviewed the literature on its theoretical properties. However, we identified key model errors that prohibit direct application to real-world measurement data from nuclear decommissioning environments.

For cameras with few, large scintillation detectors, Monte Carlo simulations showed that over 50% of measurements can stem from event sequences with multiple Compton scattering interactions in the first detector. We saw that the resulting model mismatch is large enough to prohibit reliable source reconstruction. We developed a model correction that includes second order scattering, covering a large amount of the multiple scattering events.

We gave two further model extensions, enabling bidirectional camera geometries and multi-nuclide imaging capabilities. These model corrections were shown to be efficiently computable and are promising extensions to the forward model for future camera development. For bidirectional camera geometries, our current setup represents a minimalistic prototype. A proper large-scale bidirectional camera system could significantly improve reconstruction quality by exploiting geometric information from multiple viewing angles. The sampling framework developed in this thesis could contribute to this direction through Optimal Experimental Design: with representative samples from the posterior

distribution, one could systematically evaluate different camera configurations and determine optimal detector placements for specific imaging tasks.

Using established convex optimization methods, we demonstrated that MAP estimation with the developed forward models yields reliable activity reconstructions in practical scenarios. Our experiments employed the fully scattered data at realistic activity levels, perturbed by natural background radiation, noisy energy measurements from imperfect detectors, and the standard Poisson statistics underlying radioactive decay. The experiments validated the modeling approach across different source configurations, energy ranges, and measurement conditions, establishing the viability of the corrected models for real-world decommissioning scenarios.

Our experiments focused on a Compton camera with fairly limited spatial resolution within a detector, as the scintillators are assumed to be non-pixelated. This leaves just six pairs of possible interaction "locations", each distributed in a detector volume of roughly $350\,\mathrm{cm}^3$. The resulting data space is small and the image reconstruction problem (depending on the spatial resolution of the detector material) is moderately to severely underdetermined. Considering this, the forward models and algorithms developed in this work are able to obtain good estimates of source activity maps. However, the spatial uncertainty in the reconstruction is necessarily large, as can be seen from the reconstructed variance maps. Pixelated detectors and cameras with smaller and more detectors could lead to improved reconstruction quality and the ability to reliably reconstruct complex contamination scenarios.

The forward models developed in this work follow a physics-based approach, incorporating known scattering physics through analytical or numerical computation. An alternative direction is to explore hybrid or learned forward models. A parallel line of work started investigating approaches where the system matrix could be learned from Monte Carlo simulations and then generated for new camera geometries without rerunning expensive simulations or numerical integration [38]. Such learned models could potentially capture complex detector effects that are difficult to model analytically, while maintaining the necessary computational efficiency.

Another major open direction arises from the fact that the image spaces were chosen to be simple and planar in the present work. In nuclear decommissioning, the contaminated surfaces around the camera take the form of arbitrary combinations of, say, walls, pipework, or barrels. The geometries can quickly get complex enough that proper radiological exploration is only possible from multiple camera viewpoints. In such a scenario, the pipeline should consist of a surface reconstruction process using, e.g., point cloud processing, followed by a joint reconstruction process from multiple camera viewpoints. Such an approach would require recomputation of the forward model per viewpoint and quantification of the remaining uncertainty in a given geometry, making the analytical models and UQ approaches from this work promising.

**Langevin sampling algorithms.** For uncertainty quantification in imaging inverse problems, we provided a unified treatment of Langevin Monte Carlo algorithms. These methods simulate stochastic processes driven by gradient information and Brownian motion, enabling efficient sampling from high-dimensional posterior distributions. The non-smooth potentials arising from total variation regularization and the domain constraints common in imaging applications pose theoretical and algorithmic challenges that we addressed through several contributions.

We unified existing convergence results for the unadjusted Langevin algorithm and its semi-implicit proximal variants, provided new convergence proofs for these proximal schemes, and established a rigorous comparison of convergence behavior across different discretization strategies. For the simplified setting of Gaussian targets, we obtained a clear ranking of convergence speeds among the proximal schemes.

The coupling-based convergence proofs developed in this thesis reveal a deep connection between convex optimization and log-concave sampling. This relationship suggests that other optimization algorithms could be transformed into sampling methods. For example, other splitting methods like ADMM, which have proven highly effective for convex imaging problems, may yield efficient sampling schemes when appropriately stochasticized. The theoretical framework established here provides tools for analyzing such extensions.

We provided a new optimality result for mirrored Langevin sampling with the weighted negative entropy mirror map. We proved that this choice achieves optimal Poincaré constant for posteriors arising from diagonal, linear forward models and Poisson-distributed data. This justifies a preconditioning strategy particularly suited to imaging problems with photon-counting measurements. The practical effectiveness of the mirrored approach was demonstrated in several experiments, where we showed that the mirrored sampler achieved faster convergence than standard schemes.

A promising direction for efficient sampling algorithms is the combination of learned methods with the rigorous foundations of log-concave sampling. The prior distributions imposed on the image space within this work are all model-based Gibbs priors. Plug-and-Play methods, which replace proximal operators with learned denoisers, have achieved remarkable success in optimization-based image reconstruction. Extending these ideas to sampling would require careful theoretical analysis to ensure that the resulting algorithms still target the correct posterior distribution, but could potentially combine the expressiveness of learned priors with the uncertainty quantification capabilities of Bayesian inference.

# Appendix A

# Additional Reconstruction Results

In this appendix chapter, we collect additional reconstruction results from various simualted and real measurements.

**Cs-137 point source measurements with RSL7.** Here we give the full results from a set of RSL7 measurements that was taken with a Caesium-137 source in a lab setting at Hochschule Zittau-Görlitz mid 2023. In the course of the measurements, the camera was mounted on a stand 1 m away (in $z$-direction) from an $x-y-$wall with multiple holes to insert lab sources. A 612 kBq Caesium-137 source was mounted in different $(x, y)$ positions on the wall. For each position, two measurements of 10 minutes measurement time each were taken. The measurement positions were all centered in $x$-direction and shifted relative to the camera on the $y$-axis by $0, 10, 20, 30, 40, 50$ and 80 cm. This resulted in 14 separate measurements to test the spatial resolution of the camera and the ability of different forward models to reconstrct the locations of point sources. Results of this study led to development of the double scatter corrected model (3.19) for this work.

Figure A.1 and Fig. A.2 show MAP estimates, posterior means (MMSE) and log standard deviations for the point sources, all computed with the double scatter corrected model $\mathcal{P}^{(1)} + \mathcal{P}^{(2)}$. It is clearly visible that there is a structural bias between position estimates in the center of the camera's FOV. This effect was visible throughout most experiments with the prototype. Partly, the qualitative differences in the middle can be explained due to different statistics since a central location is closer to the camera. While the $y = 0$ location measurements provided $6.8 \cdot 10^3$ counts around the photopeak on average, the $y = 80$ position only resulted in $5.2 \cdot 10^3$ events. The model does account for this difference, but the SNR is lower at the border of the FOV due to the Poisson statistics.

In the middle, the counts are further more evenly distributed among the detector pairs, so that there are meaningful energy spectra from each spatial bin of the camera—for shifted positions, the scatterer detector further away from it rarely measures valid counts since it is "shadowed" by the detectors partly blocking its field of view.

Apart from different signal-to-noise ratio, a major reason for additional reconstruction errors in real measurements are remaining components in the setup that are not accounted for in the forward model. The aluminium coating of detectors, the housing of the camera and the stand, walls, floor and ceiling in the room all scatter the source, creating potential secondary radiation. Further factors laid out in Section 2.3 can lead to model errors, "fake" events not stemming from real photon coincidences and calibration issues due to inaccurate calibration or temperature shift in the photomultipler during measurements.

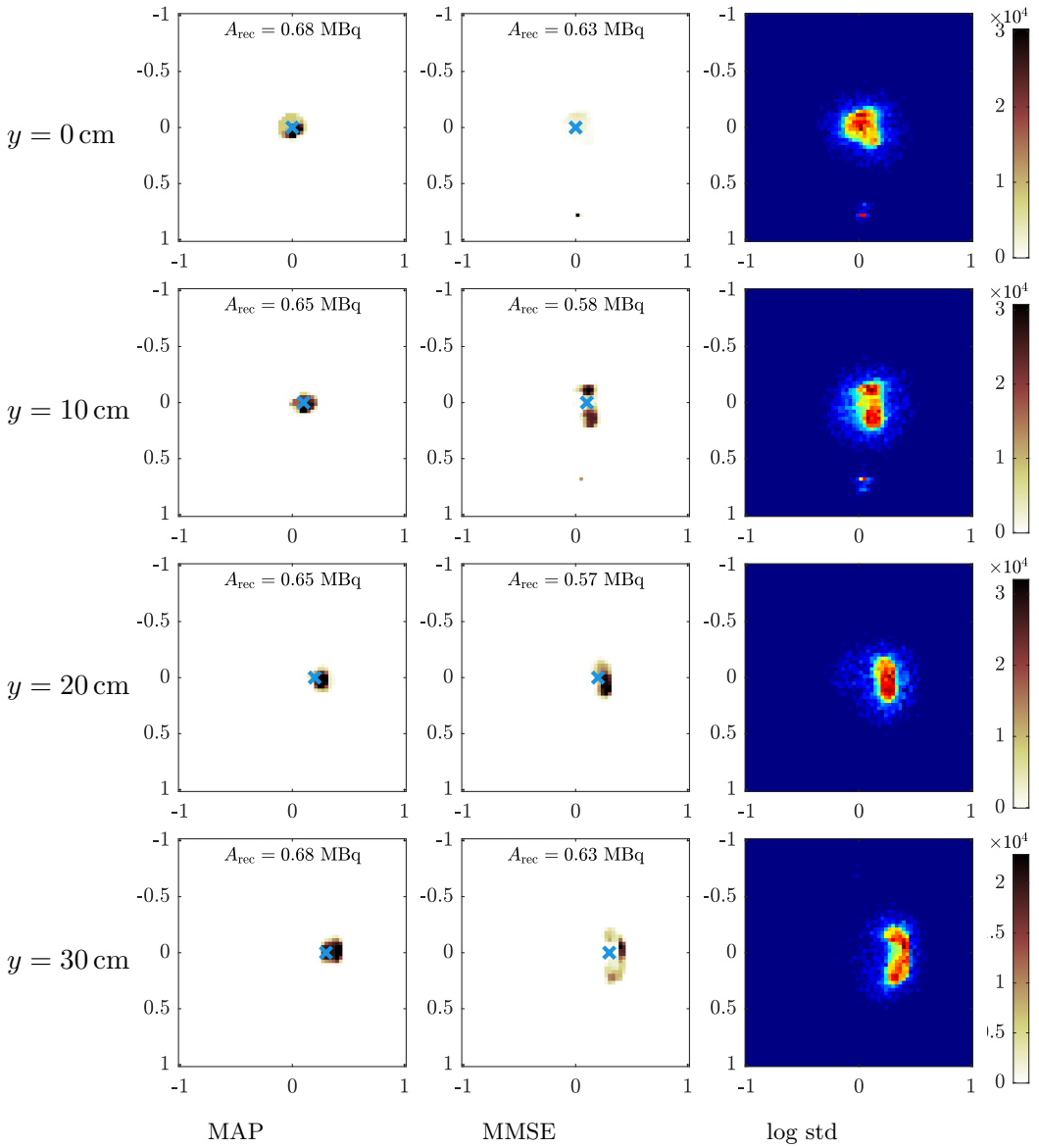



Figure A.1.: MAP, posterior mean (MMSE) and log standard deviation for Caesium-137 point source lab measurements. From top to bottom, the source is positioned 0, 10, 20 and 30 cm shifted from the camera center. True source positions marked in blue.

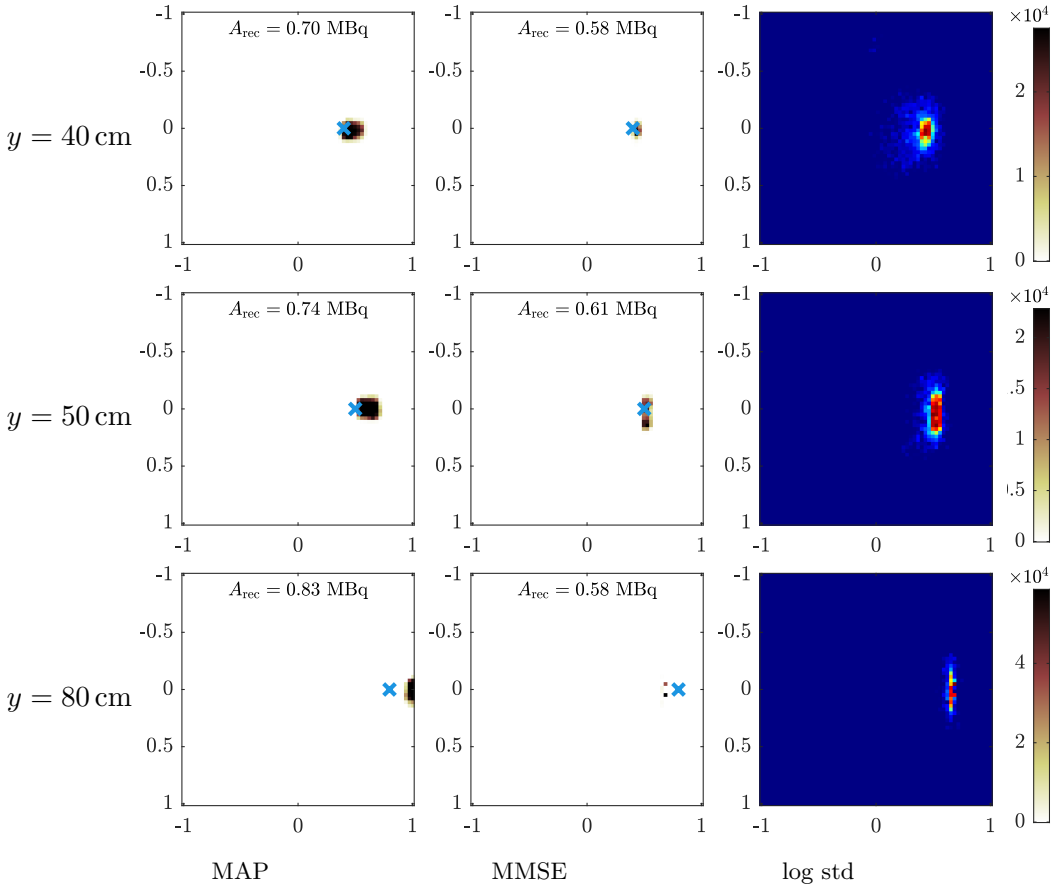



Figure A.2.: MAP, posterior mean (MMSE) and log standard deviation for Caesium-137 point source lab measurements. From top to bottom, the source is positioned 40, 50 and 80 cm shifted from the camera center. True source positions marked in blue.

# Appendix B

# Convex Analysis

In this section, we will recall some basic results from convex analysis that we frequently need throughout this work. The most relevant of these results are functional inequalities of convex functions and their Bregman divergences, those can be found in [20, chap. 18]. In the following, let $\mathcal{X}$ be a general Banach space with dual space $\mathcal{X}^*$. We state the results in their general form, although we will mostly be interested in the case $\mathcal{X} = \mathbb{R}^d$, in which case $\mathcal{X}^*$ is identified with $\mathbb{R}^d$ as well and both spaces are equipped with the Euclidean inner product and norm.

**Convexity.** A functional $V : \mathcal{X} \to \bar{\mathbb{R}}$ is convex if

$$V(tx + (1-t)\tilde{x}) \leq tV(x) + (1-t)V(\tilde{x}), \tag{B.1}$$

for all $t \in [0,1]$ and $x, \tilde{x} \in \mathcal{X}$. Similarly, we call $V$ strictly convex if (B.1) is satisfied as a strict inequality for all $t \in (0,1)$ and $x \neq \tilde{x}$, and $\omega$-strongly convex if, for all $t \in [0,1]$ and $x, \tilde{x} \in \mathcal{C}$, we have the stronger inequality

$$V(tx + (1-t)\tilde{x}) \leq tV(x) + (1-t)V(\tilde{x}) + \frac{\omega t(1-t)}{2}\|x - \tilde{x}\|_{\mathcal{X}}^2, \tag{B.2}$$

with $\omega > 0$. As usual, $\omega$-strong convexity implies strict convexity and strict convexity implies convexity. The functional $V$ is called {–/strictly/$\omega$-strongly} concave if its negative $-V$ is {–/strictly/$\omega$-strongly} convex, respectively. Frequently, we also refer to a probability distribution $\rho$ as log-concave, this means its negative log-density $V = -\log \rho$ is convex.

**Subdifferential.** Let $V : \mathcal{X} \to \bar{\mathbb{R}}$ be convex. We call $V$ subdifferentiable at $x \in \mathcal{X}$ if there exists $p \in \mathcal{X}^*$ with

$$G(\tilde{x}) \geq G(x) + \langle p, \tilde{x} - x \rangle_{\mathcal{X}}, \qquad \forall \tilde{x} \in \mathcal{X}, \tag{B.3}$$

where we use the standard notation $\langle q, u \rangle_{\mathcal{X}} \coloneqq q(u)$ for all $q \in \mathcal{X}^*, u \in \mathcal{X}$. If a $p$ exists that satisfies (B.3), it is called subgradient of $V$ at $x$, denoted $p \in \partial V(x)$, where $\partial V(x)$ is the subdifferential

$$\partial V(x) \coloneqq \{p \in \mathcal{X}^* : G(\tilde{x}) \geq G(x) + \langle p, \tilde{x} - x \rangle_{\mathcal{X}} \, \forall \tilde{x} \in \mathcal{X}\}. \tag{B.4}$$

Note that the subgradient of $V$ at $x$ is unique if $V$ is Fréchet-differentiable at $x$. If that is the case, the subdifferential is $\partial V(x) = \{\nabla V(x)\}$.

**Bregman divergences.** For any $x, \tilde{x} \in \mathcal{X}$ and any $\tilde{p} \in \partial V(\tilde{x})$, we can define the Bregman divergence $\mathrm{D}_V$ (which is also called Bregman distance, although it is not a metric on $\mathcal{X}$) by

$$\mathrm{D}_V^{\tilde{p}}(x, \tilde{x}) = V(x) - V(\tilde{x}) - \langle \tilde{p}, x - \tilde{x} \rangle_{\mathcal{X}}. \tag{B.5}$$

If the subgradient $\tilde{p}$ is unique or irrelevant, we may drop it and denote the Bregman distance simply by $\mathrm{D}_V(x, \tilde{x})$. We also define the symmetrized Bregman divergence

$$\mathrm{D}_V^{p,\tilde{p}}(x, \tilde{x}) = \mathrm{D}_V^{\tilde{p}}(x, \tilde{x}) + \mathrm{D}_V^{p}(\tilde{x}, x) = \langle p - \tilde{p}, x - \tilde{x} \rangle_{\mathcal{X}}, \tag{B.6}$$

where $p \in \partial V(x), \tilde{p} \in \partial V(\tilde{x})$. When $p, \tilde{p}$ are unique or do not matter, we will denote this also by $\mathrm{D}_V^{\mathrm{sym}}(x, \tilde{x})$.

**Bregman divergence bounds.** Throughout this document, we frequently use upper and lower bounds of the Bregman divergences under certain assumptions on $V$. The first of these assumptions is $\omega$-strong convexity (B.2). Under (B.2) with $\omega > 0$, for all $x, \tilde{x} \in \mathcal{X}$ it holds

$$\frac{\omega}{2} \|x - \tilde{x}\|_{\mathcal{X}}^2 \leq \mathrm{D}_V(x, \tilde{x}). \tag{B.7}$$

The second relevant assumption is the existence of a globally $L$-Lipschitz continuous gradient of $V$, i.e. $V$ is Fréchet-differentiable and there is an $L > 0$ such that $\nabla V$ satisfies

$$\|\nabla V(x) - \nabla V(\tilde{x})\|_{\mathcal{X}^*} \leq L \|x - \tilde{x}\|_{\mathcal{X}}. \tag{B.8}$$

By convex duality, $V$ satisfies (B.8) if and only if $V^*$ is $L^{-1}$-strongly convex. The Lipschitz gradient property (B.8) turns out to be equivalent to the following upper bound on the Bregman distance for all $x, \tilde{x} \in \mathcal{X}$

$$\mathrm{D}_V(x, \tilde{x}) \leq \frac{L}{2} \|x - \tilde{x}\|_{\mathcal{X}}^2, \qquad \forall x, \tilde{x} \in \mathcal{X}. \tag{B.9}$$

Two similar bounds can be derived by using the identity $\mathrm{D}_V^{\tilde{p}}(x,\tilde{x}) = \mathrm{D}_{V^*}^{x}(\tilde{p},p)$ for all $p \in \partial V(x)$ and $\tilde{p} \in \partial V(\tilde{x})$. By invoking convex duality, we can apply (B.7) and (B.9) to $V^*$ and obtain, under the Lipschitz gradient assumption

$$\frac{1}{2L}\|\nabla V(x) - \nabla V(\tilde{x})\|_{\mathcal{X}^*}^2 \leq \mathrm{D}_V(x,\tilde{x}), \tag{B.10}$$

and under the $\omega$-strong convexity assumption, for all $p \in \partial V(x)$ and $\tilde{p} \in \partial V(\tilde{x})$

$$\mathrm{D}_V(x,\tilde{x}) \leq \frac{1}{2\omega}\|p - \tilde{p}\|_{\mathcal{X}^*}^2. \tag{B.11}$$

Similar bounds for the symmetric Bregman divergence can be obtained by combining the bounds from above and using symmetry in $x, \tilde{x}$. Assuming $V$ is $\omega$-strongly convex, it directly follows from (B.7) that

$$\omega\|x - \tilde{x}\|_{\mathcal{X}}^2 \leq \mathrm{D}_V^{\mathrm{sym}}(x,\tilde{x}). \tag{B.12}$$

If $V$ is both $\omega$-strongly convex and has an $L$-Lipschitz gradient, we also have the following lower bound due to [158, Thm. 2.1.12]:

$$\frac{\omega L}{\omega + L}\|x - \tilde{x}\|_{\mathcal{X}}^2 + \frac{1}{\omega + L}\|\nabla V(x) - \nabla V(\tilde{x})\|_{\mathcal{X}^*}^2 \leq \mathrm{D}_V^{\mathrm{sym}}(x,\tilde{x}). \tag{B.13}$$

# Appendix C

# Hypoellipticity

An important concept in the study of the properties of Fokker–Planck equations is the smoothness of the solution. Suppose for simplicity that we study (6.46) with no drift $b = 0$ and constant isotropic diffusion $\Sigma = \alpha I$. The resulting PDE is simply a heat equation

$$\partial_t \mu_t = \alpha \Delta \mu_t,$$

which we know has regularizing properties: Even if $\mu_0$ is only measurable, the solution $\mu_t$ is $C^\infty$-smooth for all $t > 0$. This well-known property, which is shared by elliptic PDEs with sufficiently smooth coefficients, is typically called hypoellipticity in a generalizing manner.

**Definition C.1: Hypoellipticity**

Let $\mathcal{L}$ be given by (6.44) with smooth coefficients. If, for every open set $\Omega$ and every distribution $u$ on $\Omega$, $\mathcal{L}u \in C^\infty(\Omega)$ implies $u \in C^\infty(\Omega)$, then $\mathcal{L}$ is called hypoelliptic.

Naturally, every elliptic equation is hypoelliptic, and equally any operator of the form $-\partial_t + \mathcal{L}$, where $\mathcal{L}$ is the generator of a diffusion process with smooth coefficients and uniformly positive definite diffusion coefficient (this is sometimes called parabolic hypoellipticity and includes the heat equation). The latter case includes the heat equation example above.

The purpose of this subsection is to state a weaker result due to Hörmander [96], which shows that the uniform positive definiteness is not necessary. We follow [15, chap. 1.12] and [166, chap. 6]. In order to state the result, we need the definition of the commutator operation and Lie algebras spanned by smooth vector fields.

**Definition C.2: Commutator**

For two vector fields $A, B$ on $\mathbb{R}^d$, the Lie bracket or commutator $[A, B]$ is the vector field on $\mathbb{R}^d$ defined by

$$[A, B](x) = DB(x)A(x) - DA(x)B(x),$$

with the derivative matrices $(DA)_{i,j}(x) = \partial_j A_i(x)$.

**Lemma C.3: Hörmander's hypoellipticity criterion**

Assume that the generator $\mathcal{L}$ of a diffusion process with state space $\mathbb{R}^d$ has $C^\infty$-smooth coefficients and can be written in the form

$$\mathcal{L} = A_0 + \sum_{j=1}^{n} A_j^2$$

with first-order differential operators $A_0, \dots, A_n$. Interpreting the differential operators as vector fields, define a collection of vector fields $\mathcal{A}_k$ by

$$\mathcal{A}_0 \coloneqq \{A_j : j = 1, \dots, n\}, \quad \mathcal{A}_{k+1} = \mathcal{A}_k \cup \{[A, A_j] : A \in \mathcal{A}_k, j = 0, \dots, n\}.$$

If the vector spaces $\mathcal{A}_k(x) \coloneqq \operatorname{span}(\{A(x) : A \in \mathcal{A}_k\})$ satisfy $\bigcup_{k\geq 0} \mathcal{A}_k(x) = \mathbb{R}^d$ for every $x \in \mathbb{R}^d$, then $\mathcal{L}$ is said to satisfy the parabolic Hörmander condition. In that case, $\partial_t - \mathcal{L}$ and $\partial_t - \mathcal{L}^*$ are both hypoelliptic, hence any process generated by $\mathcal{L}$ has $C^\infty$-smooth transition probability. In particular, any solution to the Fokker–Planck equation $\partial_t p = \mathcal{L}^* p$ is $C^\infty((0, \infty) \times \mathbb{R}^d)$-smooth.

Note that this formulation is the parabolic form and the statement can be generalized slightly further when ignoring time dimensions, see the discussions in [15, chap. 1.12]. The stated parabolic form will later allow us to directly apply the result to the diffusion processes considered in this work: Consider the diffusion process generator (6.44) and assume that $b$ is smooth and $\Sigma = \frac{1}{2}\operatorname{diag}(\sigma_1^2, \dots, \sigma_d^2)$ constant. Then $\mathcal{L}$ is in the correct form for Lemma C.3 with

$$\begin{aligned} A_0 &= -b(x)^T \nabla_x, \\ A_j &= \frac{1}{\sqrt{2}} \sigma_j \partial_{x_j}, \qquad j = 1, \dots, d. \end{aligned}$$

Proving the smoothness of the solution will then reduce to showing that $\mathcal{A}_k$ span $\mathbb{R}^d$.

# Bibliography


[1] S. Agostinelli et al. "Geant4—a simulation toolkit." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (July 2003), pp. 250–303. DOI: 10.1016/s0168-9002(03)01368-8.

[2] K. Ahn and S. Chewi. "Efficient constrained sampling via the mirror-Langevin algorithm." In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021.

[3] C. Alkan et al. "Variational Diffusion Models for MRI Blind Inverse Problems." In: *NeurIPS 2023 Workshop on Deep Learning and Inverse Problems*. 2023.

[4] J. Allison et al. "Geant4 developments and applications." In: *IEEE Transactions on Nuclear Science* 53.1 (Feb. 2006), pp. 270–278. DOI: 10.1109/tns.2006.869826.

[5] J. Allison et al. "Recent developments in Geant4." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 835 (Nov. 2016), pp. 186–225. DOI: 10.1016/j.nima.2016.06.125.

[6] F. Alvarez, J. Bolte, and O. Brahic. "Hessian Riemannian Gradient Flows in Convex Programming." In: *SIAM Journal on Control and Optimization* 43.2 (Jan. 2004), pp. 477–501. DOI: 10.1137/s0363012902419977.

[7] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.

[8] A. N. Angelopoulos et al. "Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging." In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 717–730.

[9] F. J. Anscombe. "The Transformation of Poisson, Binomial and Negative-Binomial Data." In: *Biometrika* 35.3/4 (Dec. 1948), p. 246. DOI: 10.2307/2332343.

[10] S. Arridge, A. Hauptmann, and Y. Korolev. *Inverse Problems with Learned Forward Operators*. 2024. arXiv: 2311.12528 [math.NA].

[11] S. Arridge et al. "Solving inverse problems using data-driven models." In: *Acta Numerica* 28 (May 2019), pp. 1–174. DOI: 10.1017/s0962492919000059.

[12] J.-P. Aubin and G. D. Prato. "The viability theorem for stochastic differential inclusions." In: *Stochastic Analysis and Applications* 16.1 (Jan. 1998), pp. 1–15. DOI: 10.1080/07362999808809512.

[13] F. Bach et al. *Handbook on decommissioning of nuclear installations*. European Commission: Directorate-General for Research and Innovation, 1995.

[14] D. Bakry and M. Émery. "Diffusions hypercontractives." In: *Séminaire de Probabilités XIX 1983/84*. Springer Berlin Heidelberg, 1985, pp. 177–206. DOI: 10.1007/bfb0075847.

[15] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer International Publishing, 2014. DOI: 10.1007/978-3-319-00227-9.

[16] H. H. Barrett, T. White, and L. C. Parra. "List-mode likelihood." eng. In: *Journal of the Optical Society of America. A, Optics, Image Science, and Vision* 14.11 (Nov. 1997), pp. 2914–2923. DOI: 10.1364/josaa.14.002914.

[17] R. Basko, G. L. Zeng, and G. T. Gullberg. "Fully three dimensional image reconstruction from "V"-projections acquired by Compton camera with three vertex electronic collimation." In: *1997 IEEE Nuclear Science Symposium Conference Record*. IEEE, 1997. DOI: 10.1109/nssmic.1997.670496.

[18] R. Basko, G. L. Zeng, and G. T. Gullberg. "Application of spherical harmonics to image reconstruction for the Compton camera." en. In: *Physics in Medicine and Biology* 43.4 (Apr. 1998), pp. 887–894. DOI: 10.1088/0031-9155/43/4/016.

[19] H. H. Bauschke, J. Bolte, and M. Teboulle. "A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications." In: *Mathematics of Operations Research* 42.2 (May 2017), pp. 330–348. DOI: 10.1287/moor.2016.0817.

[20] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer New York, 2011. DOI: 10.1007/978-1-4419-9467-7.

[21] M.-M. Bé et al. *Table of Radionuclides*. Vol. 2. Monographie BIPM-5. Pavillon de Breteuil, F-92310 Sèvres, France: Bureau International des Poids et Mesures, 2004.

[22] M.-M. Bé et al. *Table of Radionuclides*. Vol. 3. Monographie BIPM-5. Pavillon de Breteuil, F-92310 Sèvres, France: Bureau International des Poids et Mesures, 2006.

[23] M.-M. Bé et al. *Table of Radionuclides*. Vol. 8. Monographie BIPM-5. Pavillon de Breteuil, F-92310 Sèvres, France: Bureau International des Poids et Mesures, 2016.

[24] J.-D. Benamou and Y. Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem." In: *Numerische Mathematik* 84.3 (Jan. 2000), pp. 375–393. DOI: 10.1007/s002110050002.

[25] M. Benning and M. Burger. "Modern regularization methods for inverse problems." In: *Acta Numerica* 27 (May 2018), pp. 1–111. DOI: 10.1017/s0962492918000016.

[26] M. Berger et al. *XCOM: Photon Cross Section Database*. en. http://physics.nist.gov/xcom. 2010.

[27] E. Bernton. "Langevin Monte Carlo and JKO splitting." In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 1777–1798.

[28] M. Bertero et al. "Image deblurring with Poisson data: from cells to galaxies." In: *Inverse Problems* 25.12 (Nov. 2009), p. 123006. DOI: 10.1088/0266-5611/25/12/123006.

[29] D. P. Bertsekas. *Nonlinear programming*. 2nd ed. Athena scientific optimization and computation series 4. Belmont, Massachusetts: Athena Scientific, 1999. 786 pp.

[30] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians." In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773.

[31] T. Böhlen et al. "The FLUKA Code: Developments and Challenges for High Energy and Medical Applications." In: *Nuclear Data Sheets* 120 (June 2014), pp. 211–214. DOI: 10.1016/j.nds.2014.07.049.

[32] C. Bonet et al. "Mirror and Preconditioned Gradient Descent in Wasserstein Space." In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 25311–25374.

[33] S. Bonettini, R. Zanella, and L. Zanni. "A scaled gradient projection method for constrained image deblurring." In: *Inverse Problems* 25.1 (Nov. 2008), p. 015002. DOI: 10.1088/0266-5611/25/1/015002.

[34] C. Brandt and A. Seppänen. "Recovery from Errors Due to Domain Truncation in Magnetic Particle Imaging: Approximation Error Modeling Approach." In: *Journal of Mathematical Imaging and Vision* 60.8 (Mar. 2018), pp. 1196–1208. DOI: 10.1007/s10851-018-0807-z.

[35] K. Bredies and M. Holler. "A TGV-Based Framework for Variational Image Decompression, Zooming, and Reconstruction. Part I: Analytics." In: *SIAM Journal on Imaging Sciences* 8.4 (Jan. 2015), pp. 2814–2850. DOI: 10.1137/15m1023865.

[36] K. Bredies and M. Holler. "A TGV-Based Framework for Variational Image Decompression, Zooming, and Reconstruction. Part II: Numerics." In: *SIAM Journal on Imaging Sciences* 8.4 (Jan. 2015), pp. 2851–2886. DOI: 10.1137/15m1023877.

[37] Y. Brenier. "Polar factorization and monotone rearrangement of vector-valued functions." In: *Communications on Pure and Applied Mathematics* 44.4 (June 1991), pp. 375–417. DOI: 10.1002/cpa.3160440402.

[38] M. Bruch. "Learned Forward Models for Compton Camera Image Reconstruction." Master's Thesis. Friedrich-Alexander-Universität Erlangen-Nürnberg, 2025.

[39] S. Bubeck, R. Eldan, and J. Lehec. "Finite-Time Analysis of Projected Langevin Monte Carlo." In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015.

[40] S. Bubeck, R. Eldan, and J. Lehec. "Sampling from a Log-Concave Distribution with Projected Langevin Monte Carlo." In: *Discrete & Computational Geometry* 59.4 (Apr. 2018), pp. 757–783. DOI: 10.1007/s00454-018-9992-1.

[41] M. Burger and S. Osher. "A Guide to the TV Zoo." In: *Level Set and PDE Based Reconstruction Methods in Imaging*. Springer International Publishing, 2013, pp. 1–70. DOI: 10.1007/978-3-319-01712-9_1.

[42] M. Burger et al. "Multiple scatter correction for single plane Compton camera imaging in nuclear decommissioning." In: *PAMM* 23.3 (2023). DOI: 10.1002/pamm.202300281.

[43] M. Burger et al. *Analysis of Primal-Dual Langevin Algorithms*. 2024. arXiv: 2405.18098 [math.OC].

[44] X. Cai, J. D. McEwen, and M. Pereyra. "Proximal nested sampling for high-dimensional Bayesian model selection." In: *Statistics and Computing* 32.5 (Oct. 2022). DOI: 10.1007/s11222-022-10152-9.

[45] D. Calvetti and E. Somersalo. "A Gaussian hypermodel to recover blocky objects." In: *Inverse Problems* 23.2 (Mar. 2007), pp. 733–754. DOI: 10.1088/0266-5611/23/2/016.

[46] D. Calvetti and E. Somersalo. "Hypermodels in the Bayesian imaging framework." In: *Inverse Problems* 24.3 (May 2008), p. 034013. DOI: 10.1088/0266-5611/24/3/034013.

[47] E. J. Candes and T. Tao. "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?" In: *IEEE Transactions on Information Theory* 52.12 (Dec. 2006), pp. 5406–5425. DOI: 10.1109/tit.2006.885507.

[48] E. J. Candès, J. K. Romberg, and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements." In: *Communications on Pure and Applied Mathematics* 59.8 (Mar. 2006), pp. 1207–1223. DOI: 10.1002/cpa.20124.

[49] Y. Cao, J. Lu, and L. Wang. "On Explicit $L^2$-Convergence Rate Estimate for Underdamped Langevin Dynamics." In: *Archive for Rational Mechanics and Analysis* 247.5 (Aug. 2023). DOI: 10.1007/s00205-023-01922-4.

[50] L. Caucci et al. “List-mode MLEM Image Reconstruction from 3D ML Position Estimates.” In: *IEEE Nuclear Science Symposium conference record. Nuclear Science Symposium* 2010 (2010), pp. 2643–2647. DOI: 10.1109/NSSMIC.2010.5874269.

[51] A. Chambolle and T. Pock. “A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging.” en. In: *Journal of Mathematical Imaging and Vision* 40.1 (May 2011), pp. 120–145. DOI: 10.1007/s10851-010-0251-1.

[52] A. Chambolle and T. Pock. “An introduction to continuous optimization for imaging.” In: *Acta Numerica* 25 (May 2016), pp. 161–319. DOI: 10.1017/s096249291600009x.

[53] S. Chelikani, J. Gore, and G. Zubal. “Optimizing Compton camera geometries.” In: *Physics in Medicine and Biology* 49.8 (Mar. 2004), pp. 1387–1408. DOI: 10.1088/0031-9155/49/8/002.

[54] Y. Chen et al. *Gradient Flows for Sampling: Mean-Field Models, Gaussian Approximations and Affine Invariance*. 2024. arXiv: 2302.11024 [stat.ML].

[55] Y. Chen et al. “Improved analysis for a proximal algorithm for sampling.” In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 2984–3014.

[56] X. Cheng and P. Bartlett. “Convergence of Langevin MCMC in KL-divergence.” In: *Proceedings of Algorithmic Learning Theory*. Ed. by F. Janoos, M. Mohri, and K. Sridharan. Vol. 83. Proceedings of Machine Learning Research. PMLR, Apr. 2018, pp. 186–211.

[57] X. Cheng et al. “Underdamped Langevin MCMC: A non-asymptotic analysis.” In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 300–323.

[58] S. Chewi et al. “Exponential ergodicity of mirror-Langevin diffusions.” In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 19573–19585.

[59] S. Chewi et al. “SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 2098–2109.

[60] S. Chewi et al. “Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev.” In: *Foundations of Computational Mathematics* (July 2024). DOI: 10.1007/s10208-024-09667-6.

[61] M. A. Clyde et al. “Current Challenges in Bayesian Model Choice.” In: *Statistical Challenges in Modern Astronomy IV*. ASP Conference Series. Vol. 371. 2007.

[62] L. Condat. "A Primal–Dual Splitting Method for Convex Optimization Involving Lipschitzian, Proximable and Linear Composite Terms." In: *Journal of Optimization Theory and Applications* 158.2 (Dec. 2012), pp. 460–479. DOI: 10.1007/s10957-012-0245-9.

[63] L. Condat. "Discrete Total Variation: New Definition and Minimization." In: *SIAM Journal on Imaging Sciences* 10.3 (Jan. 2017), pp. 1258–1290. DOI: 10.1137/16m1075247.

[64] D. Cordero-Erausquin and B. Klartag. "Moment measures." In: *Journal of Functional Analysis* 268.12 (June 2015), pp. 3834–3866. DOI: 10.1016/j.jfa.2015.04.001.

[65] M. J. Cree and P. J. Bones. "Towards direct reconstruction from a gamma camera based on Compton scattering." In: *IEEE Transactions on Medical Imaging* 13.2 (June 1994), pp. 398–407. DOI: 10.1109/42.293932.

[66] F. R. Crucinio et al. *Optimal Scaling Results for Moreau-Yosida Metropolis-adjusted Langevin Algorithms*. 2024. arXiv: 2301.02446 [stat.CO].

[67] T. Cui, X. Tong, and O. Zahm. *Optimal Riemannian metric for Poincaré inequalities and how to ideally precondition Langevin dymanics*. 2025. DOI: 10.48550/ARXIV.2404.02554.

[68] A. Dalalyan. "Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent." In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by S. Kale and O. Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, July 2017, pp. 678–689.

[69] A. S. Dalalyan and A. Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient." en. In: *Stochastic Processes and their Applications* 129.12 (Dec. 2019), pp. 5278–5311. DOI: 10.1016/j.spa.2019.02.016.

[70] D. Davis and W. Yin. "A Three-Operator Splitting Scheme and its Optimization Applications." In: *Set-Valued and Variational Analysis* 25.4 (June 2017), pp. 829–858. DOI: 10.1007/s11228-017-0421-z.

[71] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 39.1 (Sept. 1977), pp. 1–22. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[72] D. L. Donoho. "Compressed sensing." In: *IEEE Transactions on Information Theory* 52.4 (Apr. 2006), pp. 1289–1306. DOI: 10.1109/tit.2006.871582.

[73] A. Durmus, S. Majewski, and B. Miasojedow. "Analysis of Langevin Monte Carlo via Convex Optimization." In: *J. Mach. Learn. Res.* 20.1 (2019), pp. 2666–2711.

[74] A. Durmus and É. Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm." In: *The Annals of Applied Probability* 27.3 (June 2017). DOI: 10.1214/16-aap1238.

[75] A. Durmus and É. Moulines. "High-dimensional Bayesian inference via the unadjusted Langevin algorithm." In: *Bernoulli* 25.4A (Nov. 2019), pp. 2854–2882. DOI: 10.3150/18-BEJ1073.

[76] A. Durmus, É. Moulines, and M. Pereyra. "Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau." In: *SIAM Journal on Imaging Sciences* 11.1 (Jan. 2018), pp. 473–506. DOI: 10.1137/16M1108340.

[77] M. J. Ehrhardt, L. Kuger, and C.-B. Schönlieb. "Proximal Langevin Sampling with Inexact Proximal Mapping." In: *SIAM Journal on Imaging Sciences* 17.3 (July 2024), pp. 1729–1760. DOI: 10.1137/23m1593565.

[78] D. B. Everett et al. "Gamma-radiation imaging system based on the Compton effect." en. In: *Proceedings of the Institution of Electrical Engineers* 124.11 (Nov. 1977), pp. 995–1000. DOI: 10.1049/piee.1977.0203.

[79] A. Fassò et al. *FLUKA: A multi-particle transport code (program version 2005).* en. 2005. DOI: 10.5170/CERN-2005-010.

[80] Y. Feng et al. "Influence of Doppler broadening model accuracy in Compton camera list-mode MLEM reconstruction." In: *Inverse Problems in Science and Engineering* 29.13 (Dec. 2021), pp. 3509–3529. DOI: 10.1080/17415977.2021.2011863.

[81] M. Fontana et al. "Compton camera study for high efficiency SPECT and benchmark with Anger system." en. In: *Physics in Medicine & Biology* 62.23 (Nov. 2017), pp. 8794–8812. DOI: 10.1088/1361-6560/aa926a.

[82] N. Fournier and B. Perthame. "Transport distances for PDEs: the coupling method." In: *EMS Surveys in Mathematical Sciences* 7.1 (Jan. 2021), pp. 1–31. DOI: 10.4171/emss/35.

[83] M. Frandes, B. Timar, and D. Lungeanu. "Image Reconstruction Techniques for Compton Scattering Based Imaging: An Overview [Compton Based Image Reconstruction Approaches]." In: *Current Medical Imaging Reviews* 12.2 (Mar. 2016), pp. 95–105. DOI: 10.2174/1573405612666160128233916.

[84] L. Fruehwirth and A. Habring. *Ergodicity of Langevin Dynamics and its Discretizations for Non-smooth Potentials.* 2024. arXiv: 2411.12051 [math.NA].

[85] A. Gelman, W. R. Gilks, and G. O. Roberts. "Weak convergence and optimal scaling of random walk Metropolis algorithms." In: *The Annals of Applied Probability* 7.1 (Feb. 1997). DOI: 10.1214/aoap/1034625254.

[86] A. Gelman et al. *Bayesian data analysis.* Third edition. A Chapman & Hall book. Boca Raton: CRC Press, Taylor and Francis Group, 2014. 667 pp.

[87] M. Girolami and B. Calderhead. "Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73.2 (Mar. 2011), pp. 123–214. DOI: 10.1111/j.1467-9868.2010.00765.x.

[88] J. Glaubitz, A. Gelb, and G. Song. "Generalized Sparse Bayesian Learning and Application to Image Reconstruction." In: *SIAM/ASA Journal on Uncertainty Quantification* 11.1 (Mar. 2023), pp. 262–284. DOI: 10.1137/22m147236x.

[89] L. Gross. "Logarithmic Sobolev Inequalities." In: *American Journal of Mathematics* 97.4 (1975), p. 1061. DOI: 10.2307/2373688.

[90] A. Habring, M. Holler, and T. Pock. *Subgradient Langevin Methods for Sampling from Non-smooth Potentials*. 2023. DOI: 10.48550/ARXIV.2308.01417.

[91] M. Hirasawa and T. Tomitani. "An analytical image reconstruction algorithm to compensate for scattering angle broadening in Compton cameras." en. In: *Physics in Medicine and Biology* 48.8 (Apr. 2003), pp. 1009–1026. DOI: 10.1088/0031-9155/48/8/304.

[92] M. Z. Hmissi et al. "First images from a CeBr3 /LYSO:Ce Temporal Imaging portable Compton camera at 1.3 MeV." In: *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*. 2018, pp. 1–3. DOI: 10.1109/NSSMIC.2018.8824429.

[93] L. Hodgkinson, R. Salomone, and F. Roosta. "Implicit Langevin Algorithms for Sampling From Log-concave Densities." In: *Journal of Machine Learning Research* 22.136 (2021), pp. 1–30.

[94] T. Hohage and F. Werner. "Inverse problems with Poisson data: statistical regularization theory, applications and algorithms." en. In: *Inverse Problems* 32.9 (July 2016), p. 093001. DOI: 10.1088/0266-5611/32/9/093001.

[95] R. Holley and D. Stroock. "Logarithmic Sobolev inequalities and stochastic Ising models." In: *Journal of Statistical Physics* 46.5–6 (Mar. 1987), pp. 1159–1194. DOI: 10.1007/bf01011161.

[96] L. Hörmander. "Hypoelliptic second order differential equations." In: *Acta Mathematica* 119.0 (1967), pp. 147–171. DOI: 10.1007/bf02392081.

[97] Y.-P. Hsieh et al. "Mirrored Langevin Dynamics." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

[98] J. Huang and D. Mumford. "Statistics of natural images and models." In: *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. CVPR-99. IEEE Comput. Soc, 1999, pp. 541–547. DOI: 10.1109/cvpr.1999.786990.

[99] J. H. Hubbell, H. A. Gimm, and I. O/verbo/. "Pair, Triplet, and Total Atomic Cross Sections (and Mass Attenuation Coefficients) for 1 MeV-100 GeV Photons in Elements $Z = 1$ to 100." In: *Journal of Physical and Chemical Reference Data* 9.4 (Oct. 1980), pp. 1023–1148. DOI: 10.1063/1.555629.

[100] S. Hurault et al. *From Denoising Score Matching to Langevin Sampling: A Fine-Grained Error Analysis in the Gaussian Setting*. 2025. arXiv: 2503.11615 [cs.LG].

[101] IAEA. *Methods for the Minimization of Radioactive Waste from Decontamination and Decommissioning of Nuclear Facilities*. Tech. rep. 401. Vienna, 2001.

[102] N. Ikeda and S. Watanabe. *Stochastic differential Equations and diffusion processes*. 2nd ed. @North-Holland mathematical Library 24. Amsterdam: North-Holland, 2011. 555 pp.

[103] S. Ikeda et al. "Bin mode estimation methods for Compton camera imaging." en. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 760 (Oct. 2014), pp. 46–56. DOI: 10.1016/j.nima.2014.05.081.

[104] R. Jordan, D. Kinderlehrer, and F. Otto. "The Variational Formulation of the Fokker–Planck Equation." In: *SIAM Journal on Mathematical Analysis* 29.1 (Jan. 1998), pp. 1–17. DOI: 10.1137/s0036141096303359.

[105] A. S. Kainth, C. Rankin, and T.-K. L. Wong. *Bregman-Wasserstein divergence: geometry and applications*. 2025. arXiv: 2302.05833 [math.PR].

[106] J. P. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer New York, 2005. DOI: 10.1007/b138659.

[107] M. Kisielewicz. "Strong and weak solutions to stochastic inclusions." eng. In: *Banach Center Publications* 32.1 (1995), pp. 277–286.

[108] M. Kisielewicz. *Stochastic Differential Inclusions and Applications*. Springer New York, 2013. DOI: 10.1007/978-1-4614-6756-4.

[109] T. Klatzer et al. "Accelerated Bayesian Imaging by Relaxed Proximal-Point Langevin Sampling." In: *SIAM Journal on Imaging Sciences* 17.2 (June 2024), pp. 1078–1117. DOI: 10.1137/23m1594832.

[110] T. Klatzer et al. *Efficient Bayesian Computation Using Plug-and-Play Priors for Poisson Inverse Problems*. 2025. arXiv: 2503.16222 [stat.CO].

[111] O. Klein and Y. Nishina. "Über die Streuung von Strahlung durch freie Elektronen nach der neuen relativistischen Quantendynamik von Dirac." In: *Zeitschrift für Physik* 52.11–12 (Nov. 1929), pp. 853–868. DOI: 10.1007/bf01366453.

[112] F. Knoll et al. "fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning." In: *Radiology: Artificial Intelligence* 2.1 (Jan. 2020), e190007. DOI: 10.1148/ryai.2020190007.

[113] G. F. Knoll. *Radiation detection and measurement*. 4th ed. Hoboken, NJ: Wiley, 2010.

[114] T. Kormoll et al. "A Compton imager for in-vivo dosimetry of proton beams—A design study." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 626-627 (Jan. 2011), pp. 114–119. DOI: 10.1016/j.nima.2010.10.031.

[115] N. Kovachki et al. "Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs." In: *Journal of Machine Learning Research* 24.89 (2023), pp. 1–97.

[116] P. Kuchment and F. Terzioglu. "Three-Dimensional Image Reconstruction from Compton Camera Data." In: *SIAM Journal on Imaging Sciences* 9.4 (Jan. 2016), pp. 1708–1725. DOI: 10.1137/16M107476X.

[117] P. Kuchment and F. Terzioglu. "Inversion of weighted divergent beam and cone transforms." en. In: *Inverse Problems & Imaging* 11.6 (2017), p. 1071. DOI: 10.3934/ipi.2017049.

[118] L. Kuger and M. Burger. "Bidirectionality and Multiple Scattering Correction in Compton Camera Imaging." In: *Tomographic Inverse Problems: Mathematical Challenges and Novel Applications*. Vol. 20. 2. EMS Press, 2023, pp. 1143–1144. DOI: 10.4171/OWR/2023/21.

[119] L. Kuger and G. Rigaud. "On multiple scattering in Compton scattering tomography and its impact on fan-beam CT." In: *Inverse Problems and Imaging* 16.5 (2022), p. 1359. DOI: 10.3934/ipi.2022029.

[120] P. Langevin. "Sur la théorie du mouvement brownien." In: *Comptes-rendus de l'Académie des sciences (Paris)* 146 (1908). Ed. by M. Mascart, pp. 530–533.

[121] R. Latała and K. Oleszkiewicz. "Between sobolev and poincaré." In: *Geometric Aspects of Functional Analysis*. Springer Berlin Heidelberg, 2000, pp. 147–168. DOI: 10.1007/bfb0107213.

[122] T. T.-K. Lau and H. Liu. "Bregman Proximal Langevin Monte Carlo via Bregman-Moreau Envelopes." In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by K. Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 12049–12077.

[123] T. T.-K. Lau, H. Liu, and T. Pock. "Non-Log-Concave and Nonsmooth Sampling via Langevin Monte Carlo Algorithms." In: *Advanced Techniques in Optimization for Machine Learning and Imaging*. Springer Nature Singapore, 2024, pp. 83–149. DOI: 10.1007/978-981-97-6769-4_5.

[124] N. Le et al. "An Extended List-Mode MLEM Algorithm for 3D Compton Image Reconstruction from Multi-View Data." In: *EPJ Web of Conferences* 288 (2023). Ed. by A. Lyoussi et al., p. 06003. DOI: 10.1051/epjconf/202328806003.

[125] J. Lee et al. "Evaluation of sequence tracking methods for Compton cameras based on CdZnTe arrays." In: *Nuclear Engineering and Technology* 53.12 (Dec. 2021), pp. 4080–4092. DOI: 10.1016/j.net.2021.06.027.

[126] W. Lee and T. Lee. "A compact Compton camera using scintillators for the investigation of nuclear materials." In: *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*. ISSN: 1082-3654. Oct. 2009, pp. 641–644. DOI: 10.1109/NSSMIC.2009.5402008.

[127] Y. T. Lee, R. Shen, and K. Tian. "Structured Logconcave Sampling with a Restricted Gaussian Oracle." In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 2993–3050.

[128] T. Lelièvre, F. Nier, and G. A. Pavliotis. "Optimal Non-reversible Linear Drift for the Convergence to Equilibrium of a Diffusion." In: *Journal of Statistical Physics* 152.2 (June 2013), pp. 237–274. DOI: 10.1007/s10955-013-0769-x.

[129] T. Lelièvre et al. *Optimizing the diffusion coefficient of overdamped Langevin dynamics*. 2025. arXiv: 2404.12087 [math.NA].

[130] R. Li et al. "The Mirror Langevin Algorithm Converges with Vanishing Bias." In: *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*. Ed. by S. Dasgupta and N. Haghtalab. Vol. 167. Proceedings of Machine Learning Research. PMLR, Apr. 2022, pp. 718–742.

[131] J. Liang and Y. Chen. "A Proximal Algorithm for Sampling from Non-Smooth Potentials." In: *2022 Winter Simulation Conference (WSC)*. IEEE, Dec. 2022, pp. 3229–3240. DOI: 10.1109/wsc57314.2022.10015293.

[132] J. Lindbloom, J. Glaubitz, and A. Gelb. "Efficient sparsity-promoting MAP estimation for Bayesian linear inverse problems." In: *Inverse Problems* 41.2 (Jan. 2025), p. 025001. DOI: 10.1088/1361-6420/ada17f.

[133] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer New York, 2004. DOI: 10.1007/978-0-387-76371-2.

[134] Q. Liu and D. Wang. "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm." In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016.

[135] X. Lojacono. "Image reconstruction for Compton camera with application to hadrontherapy." PhD thesis. INSA de Lyon, 2013.

[136] X. Lojacono et al. "A filtered backprojection reconstruction algorithm for Compton camera." English. In: *11th international meeting on "Fully three-dimensional image reconstruction in radiology and nuclear medicine"*. July 2011.

[137] X. Lojacono et al. "Image reconstruction for Compton camera applied to 3D prompt $\gamma$ imaging during ion beam therapy." In: *2011 IEEE Nuclear Science Symposium Conference Record*. Oct. 2011, pp. 3518–3521. DOI: 10.1109/NSSMIC.2011.6152647.

[138] H. Lu, R. M. Freund, and Y. Nesterov. "Relatively Smooth Convex Optimization by First-Order Methods, and Applications." In: *SIAM Journal on Optimization* 28.1 (Jan. 2018), pp. 333–354. DOI: 10.1137/16m1099546.

[139] F. Lucka. "Bayesian Inversion in Biomedical Imaging." PhD thesis. Westfälische Wilhelms-Universität Münster, 2014.

[140] Y.-A. Ma et al. "Is there an analog of Nesterov acceleration for gradient-based MCMC?" In: *Bernoulli* 27.3 (May 2021). DOI: 10.3150/20-bej1297.

[141] C. J. Maddison et al. "Dual Space Preconditioning for Gradient Descent." In: *SIAM Journal on Optimization* 31.1 (Jan. 2021), pp. 991–1016. DOI: 10.1137/19m130858x.

[142] X. Mao. *Stochastic differential equations and applications*. Second edition, reprinted. Description based upon print version of record. Oxford: Woodhead Publishing, 2011. 1445 pp.

[143] J. B. Martin et al. "A ring Compton scatter camera for imaging medium energy gamma rays." In: *IEEE Transactions on Nuclear Science* 40.4 (Aug. 1993), pp. 972–978. DOI: 10.1109/23.256695.

[144] V. Maxim. "Filtered Backprojection Reconstruction and Redundancy in Compton Camera Imaging." In: *IEEE Transactions on Image Processing* 23.1 (Jan. 2014), pp. 332–341. DOI: 10.1109/TIP.2013.2288143.

[145] V. Maxim et al. "Probabilistic models and numerical calculation of system matrix and sensitivity in list-mode MLEM 3D reconstruction of Compton camera images." en. In: *Physics in Medicine and Biology* 61.1 (2016), pp. 243–264. DOI: 10.1088/0031-9155/61/1/243.

[146] V. Maxim, M. Frandeş, and R. Prost. "Analytical inversion of the Compton transform using the full set of available projections." en. In: *Inverse Problems* 25.9 (Aug. 2009), p. 095001. DOI: 10.1088/0266-5611/25/9/095001.

[147] R. J. McCann. "A Convexity Principle for Interacting Gases." In: *Advances in Mathematics* 128.1 (June 1997), pp. 153–179. DOI: 10.1006/aima.1997.1634.

[148] B. Mehadji et al. "Extension of the List-Mode MLEM Algorithm for Poly-Energetic Imaging with a Compton Camera." In: *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*. ISSN: 2577-0829. Nov. 2018, pp. 1–8. DOI: 10.1109/NSSMIC.2018.8824289.

[149] S. Melidonis et al. "Efficient Bayesian Computation for Low-Photon Imaging Problems." In: *SIAM Journal on Imaging Sciences* 16.3 (July 2023), pp. 1195–1234. DOI: 10.1137/22m1502240.

[150] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. 2nd ed. Springer London, 2009. DOI: 10.1007/978-1-4471-3267-7.

[151] S. Motomura et al. "Gamma-Ray Compton Imaging of Multitracer in Biological Samples Using Strip Germanium Telescope." In: *IEEE Transactions on Nuclear Science* 54.3 (June 2007), pp. 710–717. DOI: 10.1109/TNS.2007.894209.

[152] S. Motomura et al. "Multiple nuclide imaging in live mouse using semiconductor compton camera for multiple molecular imaging." In: *2007 IEEE Nuclear Science Symposium Conference Record*. Vol. 5. ISSN: 1082-3654. Oct. 2007, pp. 3741–3742. DOI: 10.1109/NSSMIC.2007.4436935.

[153] W. Mou et al. "Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity." In: *Bernoulli* 28.3 (Aug. 2022). DOI: 10.3150/21-bej1343.

[154] E. Muñoz et al. “Performance evaluation of MACACO: a multilayer Compton camera.” In: *Physics in Medicine & Biology* 62.18 (Aug. 2017), pp. 7321–7341. DOI: 10.1088/1361-6560/aa8070.

[155] D. Narnhofer et al. “Posterior-Variance–Based Error Quantification for Inverse Problems in Imaging.” In: *SIAM Journal on Imaging Sciences* 17.1 (Feb. 2024), pp. 301–333. DOI: 10.1137/23m1546129.

[156] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Ed. by D. B. Judin and E. R. Dawson. A Wiley-Interscience publication. Includes index. Chichester: Wiley, 1983. 388 pp.

[157] Y. Nesterov. “A method for solving the convex programming problem with convergence rate $O(1/k^2)$.” In: *Proceedings of the USSR Academy of Sciences* 269 (1983), pp. 543–547.

[158] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer US, 2004. DOI: 10.1007/978-1-4419-8853-9.

[159] P. F. V. Oliva and O. D. Akyildiz. *Kinetic Interacting Particle Langevin Monte Carlo*. 2024. arXiv: 2407.05790 [stat.CO].

[160] F. Otto. “The Geometry of Dissipative Evolution Equations: The Porous Medium Equation.” In: *Communications in Partial Differential Equations* 26.1–2 (Jan. 2001), pp. 101–174. DOI: 10.1081/pde-100002243.

[161] N. Parikh and S. Boyd. “Proximal Algorithms.” In: *Foundations and Trends® in Optimization* 1.3 (2014), pp. 127–239. DOI: 10.1561/2400000003.

[162] G. Parisi. “Correlation functions and computer simulations.” In: *Nuclear Physics B* 180.3 (May 1981), pp. 378–384. DOI: 10.1016/0550-3213(81)90056-0.

[163] L. Parra and H. H. Barrett. “List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET.” In: *IEEE Transactions on Medical Imaging* 17.2 (Apr. 1998), pp. 228–235. DOI: 10.1109/42.700734.

[164] L. C. Parra. “Reconstruction of cone-beam projections from Compton scattered data.” In: *IEEE Transactions on Nuclear Science* 47.4 (Aug. 2000), pp. 1543–1550. DOI: 10.1109/23.873014.

[165] S. Patterson and Y. W. Teh. “Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex.” In: *Advances in Neural Information Processing Systems*. Ed. by C. Burges et al. Vol. 26. Curran Associates, Inc., 2013.

[166] G. A. Pavliotis. *Stochastic Processes and Applications*. Springer New York, 2014. DOI: 10.1007/978-1-4939-1323-7.

[167] M. Pereyra. “Proximal Markov chain Monte Carlo algorithms.” en. In: *Statistics and Computing* 26.4 (July 2016), pp. 745–760. DOI: 10.1007/s11222-015-9567-4.

[168] M. Pereyra, J. M. Bioucas-Dias, and M. A. T. Figueiredo. "Maximum-a-posteriori estimation with unknown regularisation parameters." In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, Aug. 2015, pp. 230–234. DOI: 10.1109/eusipco.2015.7362379.

[169] M. Pereyra, L. V. Mieles, and K. C. Zygalakis. "Accelerating Proximal Markov Chain Monte Carlo by Using an Explicit Stabilized Method." In: *SIAM Journal on Imaging Sciences* 13.2 (Jan. 2020), pp. 905–935. DOI: 10.1137/19m1283719.

[170] S. W. Peterson, D. Robertson, and J. Polf. "Optimizing a three-stage Compton camera for measuring prompt gamma rays emitted during proton radiotherapy." In: *Physics in Medicine and Biology* 55.22 (Nov. 2010), pp. 6841–6856. DOI: 10.1088/0031-9155/55/22/015.

[171] G. Peyré. "Entropic Approximation of Wasserstein Gradient Flows." In: *SIAM Journal on Imaging Sciences* 8.4 (Jan. 2015), pp. 2323–2351. DOI: 10.1137/15m1010087.

[172] B. T. Polyak. "Gradient methods for the minimisation of functionals." In: *USSR Computational Mathematics and Mathematical Physics* 3.4 (Jan. 1963), pp. 864–878. DOI: 10.1016/0041-5553(63)90382-3.

[173] C. Rankin and T.-K. L. Wong. *JKO schemes with general transport costs*. 2024. arXiv: 2402.17681 [math.AP].

[174] G. Raskutti and S. Mukherjee. "The Information Geometry of Mirror Descent." In: *IEEE Transactions on Information Theory* 61.3 (Mar. 2015), pp. 1451–1457. DOI: 10.1109/tit.2015.2388583.

[175] N. A. B. Riis, Y. Dong, and P. C. Hansen. "Computed Tomography Reconstruction with Uncertain View Angles by Iteratively Updated Model Discrepancy." In: *Journal of Mathematical Imaging and Vision* 63.2 (July 2020), pp. 133–143. DOI: 10.1007/s10851-020-00972-7.

[176] N. A. B. Riis, Y. Dong, and P. C. Hansen. "Computed tomography with view angle estimation using uncertainty quantification." In: *Inverse Problems* 37.6 (June 2021), p. 065007. DOI: 10.1088/1361-6420/abf5ba.

[177] C. P. Robert. *The Bayesian choice. From decision-theoretic foundations to computational implementation*. 2. ed. Springer texts in statistics. New York: Springer, 2001. 604 pp.

[178] G. O. Roberts and J. S. Rosenthal. "Optimal Scaling of Discrete Approximations to Langevin Diffusions." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 60.1 (Jan. 1998), pp. 255–268. DOI: 10.1111/1467-9868.00123.

[179] G. O. Roberts and R. L. Tweedie. "Exponential Convergence of Langevin Distributions and Their Discrete Approximations." In: *Bernoulli* 2.4 (Dec. 1996), p. 341. DOI: 10.2307/3318418.

[180] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Description based upon print version of record. Princeton: Princeton University Press, 2015. 470 pp.

[181] J. Roser et al. "Image reconstruction for a multi-layer Compton telescope: an analytical model for three interaction events." In: *Physics in Medicine & Biology* 65.14 (July 2020), p. 145005. DOI: 10.1088/1361-6560/ab8cd4.

[182] L. I. Rudin, S. Osher, and E. Fatemi. "Nonlinear total variation based noise removal algorithms." en. In: *Physica D: Nonlinear Phenomena* 60.1 (Nov. 1992), pp. 259–268. DOI: 10.1016/0167-2789(92)90242-F.

[183] M. Sakai et al. "In vivo simultaneous imaging with 99mTc and 18F using a Compton camera." en. In: *Physics in Medicine & Biology* 63.20 (Oct. 2018), p. 205006. DOI: 10.1088/1361-6560/aae1d1.

[184] M. Sakai et al. "Compton imaging with 99m Tc for human imaging." en. In: *Scientific Reports* 9.1 (Sept. 2019), p. 12906. DOI: 10.1038/s41598-019-49130-z.

[185] A. Salim, D. Kovalev, and P. Richtarik. "Stochastic Proximal Langevin Algorithm: Potential Splitting and Nonasymptotic Rates." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[186] A. Salim and P. Richtarik. "Primal Dual Interpretation of the Proximal Stochastic Gradient Langevin Algorithm." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 3786–3796.

[187] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer International Publishing, 2015. DOI: 10.1007/978-3-319-20828-2.

[188] F. Santambrogio. "Euclidean, metric, and Wasserstein gradient flows: an overview." In: *Bulletin of Mathematical Sciences* 7.1 (Mar. 2017), pp. 87–154. DOI: 10.1007/s13373-017-0101-1.

[189] Y. Sato et al. "Radiation imaging using a compact Compton camera inside the Fukushima Daiichi Nuclear Power Station building." In: *Journal of Nuclear Science and Technology* 55.9 (Sept. 2018), pp. 965–970. DOI: 10.1080/00223131.2018.1473171.

[190] A. Sawatzky et al. "Accurate EM-TV algorithm in PET with low SNR." In: *2008 IEEE Nuclear Science Symposium Conference Record*. ISSN: 1082-3654. Oct. 2008, pp. 5133–5137. DOI: 10.1109/NSSMIC.2008.4774392.

[191] A. Sawatzky et al. "Total Variation Processing of Images with Poisson Statistics." en. In: *Computer Analysis of Images and Patterns*. Ed. by X. Jiang and N. Petkov. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 533–540. DOI: 10.1007/978-3-642-03767-2_65.

[192] A. Sawatzky et al. "EM-TV Methods for Inverse Problems with Poisson Noise." en. In: *Level Set and PDE Based Reconstruction Methods in Imaging*. Ed. by M. Burger et al. Lecture Notes in Mathematics. Cham: Springer International Publishing, 2013, pp. 71–142. DOI: 10.1007/978-3-319-01712-9_2.

[193] V. Schoenfelder et al. "COMPTEL overview: Achievements and expectations." In: *Astronomy and Astrophysics Supplement* 120 (Nov. 1996), pp. 13–21.

[194] V. Schönfelder, A. Hirner, and K. Schneider. "A telescope for soft gamma ray astronomy." en. In: *Nuclear Instruments and Methods* 107.2 (Mar. 1973), pp. 385–394. DOI: 10.1016/0029-554X(73)90257-7.

[195] L. A. Shepp and Y. Vardi. "Maximum Likelihood Reconstruction for Emission Tomography." In: *IEEE Transactions on Medical Imaging* 1.2 (Oct. 1982), pp. 113–122. DOI: 10.1109/TMI.1982.4307558.

[196] M. Singh and D. Doria. "An electronically collimated gamma camera for single photon emission computed tomography. Part II: Image reconstruction and preliminary experimental measurements." eng. In: *Medical Physics* 10.4 (Aug. 1983), pp. 428–435. DOI: 10.1118/1.595314.

[197] A. V. Skorokhod. "Stochastic Equations for Diffusion Processes in a Bounded Region." In: *Theory of Probability & Its Applications* 6.3 (Jan. 1961), pp. 264–274. DOI: 10.1137/1106035.

[198] C. J. Solomon and R. J. Ott. "Gamma ray imaging with silicon detectors — A Compton camera for radionuclide imaging in medicine." en. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 273.2 (Dec. 1988), pp. 787–792. DOI: 10.1016/0168-9002(88)90097-6.

[199] W. Su, S. Boyd, and E. J. Candès. "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights." In: *Journal of Machine Learning Research* 17.153 (2016), pp. 1–43.

[200] F. Terzioglu. "Some inversion formulas for the cone transform." en. In: *Inverse Problems* 31.11 (Oct. 2015), p. 115010. DOI: 10.1088/0266-5611/31/11/115010.

[201] F. Terzioglu. "Exact inversion of an integral transform arising in Compton camera imaging." eng. In: *Journal of Medical Imaging (Bellingham, Wash.)* 7.3 (May 2020), p. 032504. DOI: 10.1117/1.JMI.7.3.032504.

[202] F. Terzioglu. "Recovering a function from its integrals over conical surfaces through relations with the Radon transform." In: *Inverse Problems* 39.2 (Jan. 2023), p. 024005. DOI: 10.1088/1361-6420/acad24.

[203] F. Terzioglu, P. Kuchment, and L. Kunyansky. "Compton camera imaging and the cone transform: a brief overview." en. In: *Inverse Problems* 34.5 (Apr. 2018), p. 054002. DOI: 10.1088/1361-6420/aab0ab.

[204] J. Tick, A. Pulkkinen, and T. Tarvainen. "Modelling of errors due to speed of sound variations in photoacoustic tomography using a Bayesian framework." In: *Biomedical Physics & Engineering Express* 6.1 (Nov. 2019), p. 015003. DOI: 10.1088/2057-1976/ab57d1.

[205] M. E. Tipping. "Sparse bayesian learning and the relevance vector machine." In: *J. Mach. Learn. Res.* 1 (Sept. 2001), pp. 211–244.

[206] R. W. Todd, J. M. Nightingale, and D. B. Everett. "A proposed $\gamma$ camera." en. In: *Nature* 251.5471 (Sept. 1974), pp. 132–134. DOI: 10.1038/251132a0.

[207] T. Tomitani and M. Hirasawa. "Image reconstruction from limited angle Compton camera data." In: *Physics in Medicine and Biology* 47.12 (June 2002), pp. 2129–2145. DOI: 10.1088/0031-9155/47/12/309.

[208] M. Tur, K. C. Chin, and J. W. Goodman. "When is speckle noise multiplicative?" In: *Applied Optics* 21.7 (Apr. 1982), p. 1157. DOI: 10.1364/ao.21.001157.

[209] M. Uenomachi et al. "Double photon emission coincidence imaging with GAGG-SiPM Compton camera." en. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. Symposium on Radiation Measurements and Applications XVII 954 (Feb. 2020), p. 161682. DOI: 10.1016/j.nima.2018.11.141.

[210] S. Vempala and A. Wibisono. "Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[211] C. Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009. DOI: 10.1007/978-3-540-71050-9.

[212] B. C. Vũ. "A splitting algorithm for dual monotone inclusions involving cocoercive operators." In: *Advances in Computational Mathematics* 38.3 (Nov. 2011), pp. 667–681. DOI: 10.1007/s10444-011-9254-8.

[213] L. Wang and M. Yan. "Hessian Informed Mirror Descent." In: *Journal of Scientific Computing* 92.3 (July 2022). DOI: 10.1007/s10915-022-01933-5.

[214] L. R. Wayne et al. "Response of NaI(Tl) to X-rays and electrons." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 411.2–3 (July 1998), pp. 351–364. DOI: 10.1016/s0168-9002(98)00193-4.

[215] S. Welker et al. *Position-Blind Ptychography: Viability of image reconstruction via data-driven variational inference*. 2025. arXiv: 2509.25269 [eess.IV].

[216] A. Wibisono. "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem." en. In: *Proceedings of the 31st Conference On Learning Theory*. PMLR, July 2018, pp. 2093–3027.

[217] A. Wibisono. *Proximal Langevin Algorithm: Rapid Convergence Under Isoperimetry*. 2019. arXiv: 1911.01469 [stat.ML].

[218] S. J. Wilderman et al. "Monte Carlo calculation of point spread functions of Compton scatter cameras." In: *1995 IEEE Nuclear Science Symposium and Medical Imaging Conference Record*. Vol. 3. Oct. 1995, 1538–1542 vol.3. DOI: 10.1109/NSSMIC.1995.500319.

[219] S. J. Wilderman et al. "List-mode maximum likelihood reconstruction of Compton scatter camera images in nuclear medicine." In: *1998 IEEE Nuclear Science Symposium Conference Record. 1998 IEEE Nuclear Science Symposium and Medical Imaging Conference (Cat. No.98CH36255)*. Vol. 3. ISSN: 1082-3654. Nov. 1998, 1716–1720 vol.3. DOI: 10.1109/NSSMIC.1998.773871.

[220] D. Wipf and B. Rao. "Sparse Bayesian Learning for Basis Selection." In: *IEEE Transactions on Signal Processing* 52.8 (Aug. 2004), pp. 2153–2164. DOI: 10.1109/tsp.2004.831016.

[221] D. Wipf and H. Zhang. "Revisiting Bayesian Blind Deconvolution." In: *Journal of Machine Learning Research* 15.111 (2014), pp. 3775–3814.

[222] D. Xu and Z. He. "Gamma-ray energy-imaging integrated spectral deconvolution." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 574.1 (Apr. 2007), pp. 98–109. DOI: 10.1016/j.nima.2007.01.171.

[223] H. Yoneda et al. "Enhancing Compton telescope imaging with maximum a posteriori estimation: A modified Richardson–Lucy algorithm for the Compton Spectrometer and Imager." In: *Astronomy & Astrophysics* 697 (May 2025), A117. DOI: 10.1051/0004-6361/202453528.

[224] A. Yurtsever, B. C. Vu, and V. Cevher. "Stochastic Three-Composite Convex Minimization." In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016.

[225] K. S. Zhang et al. "Wasserstein Control of Mirror Langevin Monte Carlo." en. In: *Proceedings of the 33rd Conference on Learning Theory*. PMLR, July 2020, pp. 3814–3841.

[226] S. Zhang et al. "Improved Discretization Analysis for Underdamped Langevin Monte Carlo." In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Ed. by G. Neu and L. Rosasco. Vol. 195. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 36–71.

[227] X. Zuo, S. Osher, and W. Li. *Gradient-adjusted underdamped Langevin dynamics for sampling*. 2024. arXiv: 2410.08987 [math.OC].

## Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben.
Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

## Affidavit

I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated.
If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

Hamburg, 28. Januar 2026 .....

Lorenz Kuger