

Energy-based models for inverse imaging problems

Andreas Habring Martin Holler Thomas Pock Martin Zach

September 17, 2025

Abstract

In this chapter we provide a thorough overview of the use of energy-based models (EBMs) in the context of inverse imaging problems. EBMs are probability distributions modeled via Gibbs densities $p(x) \propto \exp(-E(x))$ with an appropriate energy functional E . Within this chapter we present a rigorous theoretical introduction to Bayesian inverse problems that includes results on well-posedness and stability in the finite-dimensional and infinite-dimensional setting. Afterwards we discuss the use of EBMs for Bayesian inverse problems and explain the most relevant techniques for learning EBMs from data. As a crucial part of Bayesian inverse problems, we cover several popular algorithms for sampling from EBMs, namely the Metropolis-Hastings algorithm, Gibbs sampling, Langevin Monte Carlo, and Hamiltonian Monte Carlo. Moreover, we present numerical results for the resolution of several inverse imaging problems obtained by leveraging an EBM that allows for the explicit verification of those properties that are needed for valid energy-based modeling.

1 Introduction

In this article we consider the resolution of inverse problems of the form

$$\text{given } y \in \mathcal{Y}, \text{ find } x \in \mathcal{X} \text{ such that: } y = \mathcal{P}(\mathcal{F}(x)) \quad (1)$$

with appropriate spaces \mathcal{X} and \mathcal{Y} , a *forward* or *measurement* operator $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ and a noise corruption \mathcal{P} , such as, *e.g.*, additive Gaussian or Poisson distributed noise [17]. Relevant examples of inverse imaging problems include computed tomography (CT), magnetic resonance imaging (MRI), image deblurring or denoising, and many more [6, 69, 99, 98].

In the variational framework [15, 16, 34], inverse problems are usually tackled by considering a minimization problem of the form

$$\min_{x \in \mathcal{X}} \{E_y(x) := \mathcal{D}_y(\mathcal{F}(x)) + \lambda \mathcal{R}(x)\} \quad (2)$$

where \mathcal{D}_y measures the data fit, *i.e.*, the discrepancy between a potential solution x and the measurement y and \mathcal{R} is a regularizer that ensures well-posedness and stability of (2). In contrast, in the Bayesian framework, x and y in (1) are modeled as random variables X and Y with some joint distribution $\mathbb{P}_{X,Y}$. In this case the solution of the inverse problem is simply the so-called *posterior distribution* $\mathbb{P}_{X|Y}$, that is, the distribution of the variable of interest X after observing the measurement Y . Via Bayes' theorem, under mild conditions (see section 2 below) it holds true that

$$\mathbb{P}_{X|Y} = \frac{\mathbb{P}_{Y|X} \mathbb{P}_X}{\mathbb{P}_Y}. \quad (3)$$

Identifying $y \mapsto \mathcal{D}_y(\mathcal{F}(x))$ with a density of $-\log \mathbb{P}_{Y|X}$ and $x \mapsto \lambda \mathcal{R}(x)$ with a density of $-\log \mathbb{P}_X$ relates the variational and Bayesian approach by equating the variational solution with

the maximum a-posteriori (MAP) estimate of the Bayesian solution. The Bayesian perspective for inverse problems provides some interesting benefits compared to the variational approach. Most notably, the probabilistic setting yields a natural framework for modeling uncertainty of a reconstruction [69, 70] as well as for the use of data driven priors. The latter is due to the fact that the use of training data—which constitutes a random sample of some population—almost by default induces a probabilistic treatment.

We now briefly analyze the individual terms of (3). The *likelihood* $\mathbb{P}_{Y|X}$ is usually known and relates to the forward operator and noise distribution. The *evidence* \mathbb{P}_Y typically does not concern us since, after observing Y , it constitutes a constant multiplicative factor with respect to the posterior. Such constant factors need not be known, neither for determining the MAP nor—as we will see later—for sampling. As a consequence, substantial research in the context of Bayesian inverse problems has been dedicated to the remaining task of modeling the *prior* distribution \mathbb{P}_X [39, 47, 82, 98, 99]. EBMs provide a particularly useful approach to do so which complements many of the issues encountered in the context of Bayesian inverse problems. In short, the term EBM (see definition 2.5) refers to modeling a probability distribution (usually the prior \mathbb{P}_X) as a Gibbs distribution via its density

$$p_X(x) = \frac{\exp(-E(x))}{\int \exp(-E(z)) dz} \quad (4)$$

where E is an appropriate *energy* functional or *potential* that may be hand-crafted (*e.g.*, the total variation [40, 70]) or learned from data [70, 98, 99]. Energy based modeling is a natural framework due to two main properties:

1. By definition, $\exp(-E(x))$ is positive for any $x \in \mathcal{X}$, so that we only have to ensure integrability in order to obtain a valid probability density in (4).
2. As mentioned above, many tasks that frequently arise in Bayesian inverse problems (most prominently, sampling from various distributions) often do not require knowledge of the normalization constant $Z = \int \exp(-E(y)) dy$, but rather require the knowledge of the density up to a multiplicative factor, or even only of the score $\nabla \log p_X = \nabla E$. This is particularly relevant, as the computation of Z requires the estimation of a high-dimensional integral, which is numerically infeasible for the problem sizes that are encountered in imaging.

The combination of a well suited theoretical framework and practical flexibility renders EBMs a powerful tool for the application to Bayesian imaging. Within this chapter we provide a concise overview of the most relevant aspects of EBMs for Bayesian inverse imaging problems.

1.1 Overview of the remaining article

In the subsequent sections we cover the following content: In section 2 we introduce the most relevant concepts in Bayesian inverse problems and provide several central theoretical results. In particular, we present results for well-posedness and stability when the space \mathcal{X} is finite-dimensional and infinite-dimensional. In section 3 we present the most relevant training strategies, including different types of divergence minimization as well as bilevel learning, followed by a discussion about popular architectures for data-driven EBMs. In section 4 we analyze some of the most popular algorithms for sampling from EBMs, that cover prior as well as posterior sampling. Lastly, we showcase some numerical results of the application of EBMs for solving inverse imaging problems in section 5.

1.2 Notation

Spaces are denoted with calligraphic letters such as $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. σ -algebras are denoted as $\Sigma_{\mathcal{X}}$ with the corresponding space in the subscript or simply as Σ if there is no risk of ambiguities. The Borel σ -algebra on a space \mathcal{X} is denoted as $\mathcal{B}(\mathcal{X})$.¹ We use capital letters for random variables and greek letters—mostly μ, ν, π —for (probability) measures. The distribution of a random variable $X \in \mathcal{X}$ is written as \mathbb{P}_X . If the distribution \mathbb{P}_X admits a density (with respect to some measure on \mathcal{X}) we denote the density as p_X . For measures μ and ν we write $\mu \ll \nu$ if μ is absolutely continuous with respect to ν and denote the Radon-Nikodým derivative of μ with respect to ν as $\frac{d\mu}{d\nu}$, or, if ν is the Lebesgue measure, as $\frac{d\mu}{dx}$. We denote the set of all probability measures over a space \mathcal{X} as $\mathcal{P}(\mathcal{X})$ without explicitly adding the underlying σ -algebra since we do not equip spaces with more than one σ -algebra. For $m \geq 1$ we define the set of all probability measures with finite m -th moment as $\mathcal{P}_m(\mathcal{X})$. An integral without a domain is interpreted as integration over the entire space, *e.g.*, for μ a measure on \mathcal{X} and f a μ -measurable function,

$$\int f(x)d\mu(x) := \int_{\mathcal{X}} f(x)d\mu(x).$$

Moreover, we may sometimes write $\mu(dx)$ instead of $d\mu(x)$ to denote a measure μ , respectively integration with respect to this measure.

2 Bayesian Inverse Problems

Recall that we are interested in solving general inverse imaging problems of the form

$$\text{given } y, \text{ find } x \text{ such that: } y = \mathcal{P}(\mathcal{F}(x)) \tag{5}$$

where $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ denotes a forward operator and \mathcal{P} a pollution operator that, for any given $x \in \mathcal{X}$, corrupts $\mathcal{F}(x)$ with random noise following a certain distribution that is usually known. We always assume that $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are *separable and complete metric spaces* and that \mathcal{Y} is a subset of a finite-dimensional space. This setup covers almost all practically relevant cases. In particular, the assumption that \mathcal{Y} is a subset of a finite-dimensional space is not restrictive since any measurement device typically gives a finite number of measurements. For the space \mathcal{X} , this section explicitly also covers the infinite-dimensional case. To talk about probability we always equip \mathcal{X} and \mathcal{Y} with the generic Borel σ -algebras $\Sigma_{\mathcal{X}} = \mathcal{B}(\mathcal{X})$ and $\Sigma_{\mathcal{Y}} = \mathcal{B}(\mathcal{Y})$ such that $(\mathcal{X}, \Sigma_{\mathcal{X}})$, $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ are measurable spaces. The starting point for our probabilistic framework of inverse problems is a probability measure $\mathbb{P}_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$ (equipped with the generic product σ -algebra $\Sigma_{\mathcal{X} \times \mathcal{Y}} = \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y})$), which describes the distribution of the data. Based on this, in a Bayesian framework, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ are modeled as random variables $X \sim \mathbb{P}_X$, $Y \sim \mathbb{P}_Y$, formally given as $X : \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{X}$, $(x, y) \mapsto x$ and likewise for Y , and with their distributions \mathbb{P}_X and \mathbb{P}_Y being induced by the joint distribution $\mathbb{P}_{X,Y}$ as pushforward measures. Solving the inverse problem then amounts to determining the conditional distribution of $X|Y$ (see definition 2.2 below), referred to as the *posterior distribution*.

For instructive purposes let us briefly assume $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^d$ for $n, d \in \mathbb{N}$ and that both X and Y admit densities p_X and p_Y with respect to the Lebesgue measure. In this case, using Bayes theorem, the posterior is typically expressed as

$$p_X(x|Y = y) = \frac{p_Y(y|X = x)p_X(x)}{p_Y(y)}.$$

¹assuming the used topology is obvious

This representation is beneficial as it leads to interpretable terms which can be modeled: The *likelihood* $p_Y(\cdot | X = x)$ corresponds to the density of the distribution of the measurements given some fixed x and is modeled based on the forward operator \mathcal{F} and the pollution \mathcal{P} . The *prior* p_X represents the density of the distribution of the variable of interest, X , and is modeled based on prior knowledge or beliefs about X or desirable properties of the solution. The prior can be handcrafted or learned if we have access to samples of X (see section 3). Knowledge of the (*model*) *evidence* p_Y is typically not necessary as most relevant inference techniques (see section 4 below) only rely on the *score* of the *posterior* $p_X(\cdot | Y = y)$ for some fixed y , that is, the gradient of the *log-posterior* $\nabla \log p_X(\cdot | Y = y)$, which is independent of $p_Y(y)$. In conclusion: Bayesian inverse problems mainly revolve around proper modeling of the likelihood and the prior. While this elaboration was restricted to the finite-dimensional case, we will in the following provide a rigorous treatment for the general setting introduced above. This part of our work is strongly inspired by [58].

We start by defining the posterior, which necessitates the notion of a Markov kernel, which we now define.

Definition 2.1. *Let $(\mathcal{Z}_1, \Sigma_{\mathcal{Z}_1})$ and $(\mathcal{Z}_2, \Sigma_{\mathcal{Z}_2})$ be measurable spaces. A Markov kernel is a function $M : \mathcal{Z}_1 \times \Sigma_{\mathcal{Z}_2} \rightarrow [0, 1]$ such that*

- $M(\cdot, A) : \mathcal{Z}_1 \rightarrow [0, 1]$ is measurable for any $A \in \Sigma_{\mathcal{Z}_2}$, and
- $M(z, \cdot) : \Sigma_{\mathcal{Z}_2} \rightarrow [0, 1]$ is a probability measure for any $z \in \mathcal{Z}_1$.

In order to formalize the notion of the posterior distribution as the *distribution of X given Y* we further make the following definition.

Definition 2.2. *If there exists a Markov kernel $M : \mathcal{Y} \times \Sigma_{\mathcal{X}} \rightarrow [0, 1]$ such that for all $A \in \Sigma_{\mathcal{X}}$ and all $B \in \Sigma_{\mathcal{Y}}$*

$$\mathbb{P}(X \in A, Y \in B) = \int_B M(y, A) d\mathbb{P}_Y(y),$$

we call M the conditional distribution of $X|Y$ and denote $M(y, A) = \mathbb{P}_X(A|Y = y)$. Similarly, the conditional distribution of $Y|X$ is defined as above with the roles of X and Y being exchanged.

The following result, which is a standard result from probability theory, shows that in our setting a conditional distribution always exists.

Lemma 2.1. (*Existence of solutions*) *The conditional distribution of $X|Y$, denoted by $\mathbb{P}_X(\cdot | Y = \cdot)$ exists and is \mathbb{P}_Y -a.s. unique. Furthermore, for \mathbb{P}_Y -a.e. $y \in Y$, the measure $\mathbb{P}_X(\cdot | Y = y)$ is concentrated on $X(\{Y = y\})$, i.e., $\mathbb{P}_X(X(\{Y = y\})^c | Y = y) = 0$, for \mathbb{P}_Y -a.e. $y \in Y$.*

Proof. Since $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are complete and separable metric spaces, both $X \times Y$ and Y are Souslin spaces according to [11, Definition 6.6.1]. Thus, [11, Example 10.4.11] implies the existence of a Markov Kernel $M : \mathcal{Y} \times \Sigma_{\mathcal{X} \times \mathcal{Y}} \rightarrow [0, 1]$ such that

$$\mathbb{P}(C, Y \in B) = \int_B M(y, C) d\mathbb{P}_Y(y)$$

for all $C \in \Sigma_{\mathcal{X} \times \mathcal{Y}}$ and $B \in \Sigma_{\mathcal{Y}}$, and such that

$$M(y, \mathcal{X} \times (\mathcal{Y} \setminus \{y\})) = 0$$

for \mathbb{P}_Y -a.e. $y \in Y$. It is then easy to see that

$$\mathbb{P}_X(A | Y = y) := M(y, X^{-1}(A))$$

the conditional distribution as claimed. Regarding the concentration on $X(\{Y = y\})$, we note that $(x, y) \in X^{-1}((X\{Y = y\})^c)$ implies that $(x, y) \in \mathcal{X} \times (\mathcal{Y} \setminus \{y\})$ and, consequently, that

$$\mathbb{P}_X(X(\{Y = y\})^c | Y = y) \leq M(y, \mathcal{X} \times (\mathcal{Y} \setminus \{y\})) = 0.$$

Uniqueness finally follows as in [11, Lemma 10.4.3], using that $\Sigma_{\mathcal{X}}$ is countably generated. \square

Remark 2.1. *Note that, in addition to the classical notion of conditional distribution (see, e.g., [9, Theorem 33.3]), we also obtain that $\mathbb{P}_X(A|Y = y)$ is concentrated on $X(\{Y = y\})$, a property that one would naturally expect from a conditional distribution. This concentration property is a consequence of deriving the conditional distribution via the disintegration of measures rather than the classical Kolmogorov approach as Radon derivatives, see [20] for a discussion. While the former has slightly more restrictive assumptions, those are fulfilled in our setting, which is why we derive this additional property of conditional distribution in this work. Regarding a concentration of $\mathbb{P}_X(A|Y = y)$ for all $y \in \mathcal{Y}$ (not just \mathbb{P}_Y -almost all $y \in \mathcal{Y}$), we refer to [11, Proposition 10.4.12].*

Remark 2.2. *It seems that existence and uniqueness of the conditional distribution, the main quantity of our interest, is, thus, guaranteed without any requiring any assumptions in addition to those made on the underlying spaces \mathcal{X} and \mathcal{Y} . We will see later, however, that the main assumption that the prior distribution is actually a probability distribution, in the sense that it integrates to one, is already a strong assumption that is related to coercivity of the energy functional in case of energy based models.*

With the same argument as in the previous result, we also obtain existence of the conditional distribution of $Y|X$.

Lemma 2.2. *(Existence of conditional data distribution) The conditional distribution of $Y|X$, denoted by $\mathbb{P}_Y(\cdot | X = \cdot)$ exists and is \mathbb{P}_X -a.s. unique.*

The conditional distribution $\mathbb{P}_Y(\cdot | X = \cdot)$ allows us to introduce a model for the measurement noise. Two frequently used and practically relevant examples are given as follows.

Example 2.1 (Gaussian noise). *The assumption of additive, isotropic Gaussian noise on the measurements corresponds to the situation that $\mathcal{Y} = \mathbb{R}^d$ and $\mathbb{P}_Y(\cdot | X = x)$ admits a density $(x, y) \mapsto L(y|X = x)$ with respect to the Lebesgue measure of the form*

$$L(y|X = x) = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|y - \mathcal{F}(x)\|^2}{2\sigma^2}\right),$$

where $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ is the forward model.

Example 2.2 (Poisson noise). *Poisson noise constitutes a relevant example of non-additive noise corruption. In this case, $\mathcal{Y} = \{1, 2, 3, \dots\}^d$, the Borel σ -algebra $\Sigma_{\mathcal{Y}}$ equals the power set, and the conditional distribution $\mathbb{P}(\cdot | X = x)$ is assumed to admit a density $L(y|X = x)$ w.r.t. the counting measure given as*

$$L(y|X = x) = \prod_{i=1}^d \frac{\mathcal{F}(x)_i^{y_i} \exp(-\mathcal{F}(x)_i)}{y_i!}$$

where $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ is again the forward model.

Since we assume that \mathcal{Y} is a subset of a finite-dimensional space, it is not restrictive to assume that $\mathbb{P}_Y(\cdot | X = x)$ admits a density $L(\cdot | X = x)$ as above (generically, with respect to the Lebesgue or the counting measure). Interestingly, under this non-restrictive assumption, we already get a first version of Bayes theorem.

Theorem 2.1 (Bayes theorem, general version). *Assume there exists a probability measure $\mu_{\mathcal{Y}} : \Sigma_{\mathcal{Y}} \rightarrow [0, 1]$ and a function $(x, y) \mapsto L(y|X = x) \in [0, \infty]$ that is jointly measurable in x and y such that*

$$\mathbb{P}_Y(B|X = x) = \int_B L(y|X = x) d\mu_{\mathcal{Y}}(y),$$

i.e., $\mathbb{P}_Y(\cdot|X = x)$ admits the density $L(\cdot|X = x)$ with respect to $\mu_{\mathcal{Y}}$.

Then, also \mathbb{P}_Y has a density p_Y w.r.t. $\mu_{\mathcal{Y}}$ and it holds for almost all $y \in \mathcal{Y}$ with $p_Y(y) \neq 0$ and for all $A \in \Sigma_{\mathcal{X}}$ that

$$\mathbb{P}_X(A|Y = y) = \frac{\int_A L(y|X = x) d\mathbb{P}_X(x)}{p_Y(y)}.$$

Further, $\mathbb{P}_X(\cdot|Y = y)$ is absolutely continuous w.r.t. \mathbb{P}_X with density

$$\frac{d\mathbb{P}_X(\cdot|Y = y)}{d\mathbb{P}_X} = \frac{L(y|X = x)}{p_Y(y)}.$$

Proof. By the properties of $\mathbb{P}_Y(\cdot|X = \cdot)$ we have for every $B \in \Sigma_{\mathcal{Y}}$ that

$$\begin{aligned} \mathbb{P}_Y(B) &= \mathbb{P}_{X \times Y}(\mathcal{X} \times B) = \int_{\mathcal{X}} \mathbb{P}_Y(B|X = x) d\mathbb{P}_X(x) \\ &= \int_{\mathcal{X}} \int_B L(y|X = x) d\mu_{\mathcal{Y}}(y) d\mathbb{P}_X(x). \end{aligned}$$

Using Fubini's theorem (see, e.g., [25, Theorem 6.14]) we obtain that

$$\mathbb{P}_Y(B) = \int_B \int_{\mathcal{X}} L(y|X = x) d\mathbb{P}_X(x) d\mu_{\mathcal{Y}}(y),$$

and, consequently, that \mathbb{P}_Y is absolutely continuous w.r.t. $\mu_{\mathcal{Y}}$ with density $y \mapsto p_Y(y) := \int_{\mathcal{X}} L(y|X = x) d\mathbb{P}_X(x)$.

Again by the properties of the condition distributions we obtain for all $A \in \Sigma_{\mathcal{X}}$ and $B \in \Sigma_{\mathcal{Y}}$ that

$$\int_B \mathbb{P}_X(A|Y = y) d\mathbb{P}_Y(y) = \mathbb{P}(X \in A, Y \in B) = \int_A \mathbb{P}_Y(B|X = x) d\mathbb{P}_X(x).$$

Plugging in the above densities, we obtain that

$$\begin{aligned} \int_B \mathbb{P}_X(A|Y = y) p_Y(y) d\mu_{\mathcal{Y}}(y) &= \int_A \int_B L(y|X = x) d\mu_{\mathcal{Y}}(y) d\mathbb{P}_X(x) \\ &= \int_B \int_A L(y|X = x) d\mathbb{P}_X(x) d\mu_{\mathcal{Y}}(y) \end{aligned}$$

where the last equality is again due to Fubini's theorem. This yields the pointwise almost-everywhere equality

$$\mathbb{P}_X(A|Y = y) p_Y(y) = \int_A L(y|X = x) d\mathbb{P}_X(x).$$

from which the result follows via division by $p_Y(y)$. □

Remark 2.3. *We note the following:*

- Instead of explicitly requiring the existence of the measurable density $(x, y) \mapsto L(y|X = x)$, one could also require that there exists a probability measure μ_Y (or, more generally, a Markov Kernel) such that $\mathbb{P}_Y(B|X = x)$ is absolutely continuous w.r.t. μ_Y for every x . Existence of a measurable density $(x, y) \mapsto L(y|X = x)$ would then follow from Doob's theorem for families of measures, see [25, Theorem 4.44]. In practice, however, existence of the dominating measure μ_Y is usually shown directly by providing the likelihood $(x, y) \mapsto L(y|X = x)$ (as modeling choice), hence we believe it is more useful to use this setting also in the theorem.
- The above result assumes a dominating measure for $\mathbb{P}_Y(\cdot|X = x)$, and concludes that \mathbb{P}_Y is also dominated by that measure. The other direction does not hold true, but in any case it is more practical to impose this assumption on $\mathbb{P}_Y(\cdot|X = x)$ (which again is usually available as modeling choice) rather than \mathbb{P}_Y , which is usually not available.

Finally, if also \mathbb{P}_X admits a density, we get the following result as direct consequence.

Theorem 2.2 (Bayes Theorem, density version). *In the setting of theorem 2.1, assume that \mathbb{P}_X admits a density w.r.t. some measure μ_X that we denote by p_X . Then, for almost all $y \in \mathcal{Y}$ with $p_Y(y) \neq 0$, $\mathbb{P}_X(\cdot|Y = y)$ admits a density w.r.t μ_X that is given as*

$$p_X(x|Y = y) = \frac{L(y|X = x)p_X(x)}{p_Y(y)}.$$

The last theorem is particularly relevant in the finite-dimensional case, where the Lebesgue measure can be chosen as reference measure μ_X on \mathcal{X} .

Having established existence of the posterior distributions, we now move towards a continuous dependency of the posterior $\mathbb{P}_X(\cdot|Y = y)$ on $y \in \mathcal{Y}$. Given the form of the posterior as in theorem 2.1, it is clear that its continuous dependency on y requires a suitable continuity of both $y \mapsto L(y|X = x)$ and $y \mapsto p_Y(y)$. The following lemma shows that the latter is a direct consequence of the former. This is again preferred from the modeling perspective since the likelihood $L(\cdot|X = \cdot)$ is usually available explicitly as modeling choice, while p_Y is generally not explicit.

Lemma 2.3. *In the setting of theorem 2.1, assume that $y \mapsto L(y|X = x)$ is continuous for \mathbb{P}_X -a.e. $x \in \mathcal{X}$ and that there exists $g \in L^1(\mathcal{X}, \mathbb{P}_X)$ such that $L(y|X = x) \leq g(x)$ for all $y \in \mathcal{Y}$ and \mathbb{P}_X -a.e. $x \in \mathcal{X}$. Then, $y \mapsto p_Y(y)$ is also continuous.*

Proof. This follows from pointwise convergence and the dominated convergence theorem: Let $(y_n)_n$ converge to y . Then

$$\begin{aligned} |p_Y(y_n) - p_Y(y)| &= \left| \int_{\mathcal{X}} L(y_n|X = x) d\mathbb{P}_X(x) - \int_{\mathcal{X}} L(y|X = x) d\mathbb{P}_X(x) \right| \\ &\leq \int_{\mathcal{X}} |L(y_n|X = x) - L(y|X = x)| d\mathbb{P}_X(x) \end{aligned}$$

with the right-hand side converging to zero as y_n converges to y due to the assumptions on L . \square

The above assumptions on the likelihood are not very restrictive. First we note that they trivially hold in the case of Gaussian measurement noise.

Example 2.3 (Gaussian noise). *In case $\mathcal{Y} = \mathbb{R}^d$ and $\mathbb{P}_Y(\cdot|X = \cdot)$ admits a density $(x, y) \mapsto L(y|X = x)$ with respect to the Lebesgue measure of the form*

$$L(y|X = x) = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|y - \mathcal{F}(x)\|^2}{2\sigma^2}\right),$$

where $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ is a measurable forward model, $(x, y) \mapsto L(y|X = x)$ fulfills the assumptions of lemma 2.3 since $y \mapsto L(y|X = x)$ is obviously continuous for every x and $L(y|X = x) \leq (2\pi\sigma^2)^{-d/2}$.

Regarding Poisson noise, we can observe that whenever \mathcal{Y} is countable and $\mathbb{P}_Y(\cdot|X = \cdot)$ admits a density $L(\cdot|X = \cdot)$ with respect to the counting measure, we have for every x, y that

$$L(y|X = x) \leq \sum_{\tilde{y} \in \mathcal{Y}} L(\tilde{y}|X = x) \leq 1,$$

such that there always exists some $g \in L^1(\mathcal{X}, \mathbb{P}_X)$ with $L(y|X = x) \leq g(x)$. The continuity of $y \mapsto L(y|X = x)$ again depends on the application, but trivially holds in the case of Poisson noise:

Example 2.4 (Poisson noise). *In case that $\mathcal{Y} = \{1, 2, 3, \dots\}^d$, the Borel σ -algebra $\Sigma_{\mathcal{Y}}$ equals the power set, and the conditional distribution $\mathbb{P}(\cdot|X = x)$ admits a density $L(y|X = x)$ w.r.t. the counting measure given as*

$$L(y|X = x) = \prod_{i=1}^d \frac{\mathcal{F}(x)_i^{y_i} \exp(-\mathcal{F}(x)_i)}{y_i!},$$

the likelihood $L(y|X = x)$ fulfills the assumptions of lemma 2.3 whenever the forward model $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ is measurable.

We now move towards providing a stability result which will be with respect to the Hellinger distance, that we define as follows.

Definition 2.3 (Hellinger distance). *Let (Ω, Σ, μ) be a probability space and let μ_1 and μ_2 be two probability measures on (Ω, Σ) such that $\mu_1, \mu_2 \ll \mu$. We define the Hellinger distance between μ_1 and μ_2 as*

$$d_{\text{Hel}}(\mu_1, \mu_2) = \left(\frac{1}{2} \int_{\Omega} \left(\sqrt{\frac{d\mu_1}{d\mu}} - \sqrt{\frac{d\mu_2}{d\mu}} \right)^2 d\mu \right)^{1/2}$$

Under the same (non-restrictive) assumptions as in lemma 2.3, continuity w.r.t. the Hellinger distance follows.

Theorem 2.3 (Distributional stability [58]). *In the setting of theorem 2.1, assume that*

1. *there exists $g \in L^1(\mathcal{X}, \mathbb{P}_X)$ such that $L(y'|X = x) \leq g(x)$ for all $y' \in \mathcal{Y}$ and \mathbb{P}_X -a.e. $x \in \mathcal{X}$,*
2. *and that $y \mapsto L(y|X = x)$ is continuous for \mathbb{P}_X -a.e. $x \in \mathcal{X}$.*

Then, $y \mapsto \mathbb{P}_X(\cdot|Y = y)$ is continuous with respect to the Hellinger distance at every point $\hat{y} \in \mathcal{Y}$ with $p_Y(\hat{y}) \neq 0$.

Proof. First note that, by the dominated convergence theorem, for any sequence $(y_n)_n$ in \mathcal{Y} that converges to some $y \in \mathcal{Y}$ we obtain from our assumptions that

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X}} |L(y_n|X = x) - L(y|X = x)| d\mathbb{P}_X(x) = 0.$$

²*i.e., μ_1, μ_2 are absolutely continuous with respect to μ*

Now take $\hat{y}, y \in \mathcal{Y}$ with $p_Y(\hat{y}), p_Y(y) \neq 0$. Using the distributional Bayes theorem (theorem 2.1) we can compute that

$$\begin{aligned} & d_{\text{Hel}}(\mathbb{P}_X(\cdot|Y = \hat{y}), \mathbb{P}_X(\cdot|Y = y)) \\ &= \frac{1}{\sqrt{2}} \left\| \frac{\sqrt{L(\hat{y}|X = \cdot)}}{\sqrt{p_Y(\hat{y})}} - \frac{\sqrt{L(y|X = \cdot)}}{\sqrt{p_Y(y)}} \right\|_{L^2(X, \mathbb{P}_X)} \\ &\leq \frac{1}{\sqrt{2p_Y(\hat{y})}} \left\| \sqrt{L(\hat{y}|X = \cdot)} - \sqrt{L(y|X = \cdot)} \right\|_{L^2(X, \mathbb{P}_X)} \\ &\quad + \sqrt{\frac{1}{2}} \left| \frac{1}{\sqrt{p_Y(\hat{y})}} - \frac{1}{\sqrt{p_Y(y)}} \right| \left\| \sqrt{L(y|X = \cdot)} \right\|_{L^2(X, \mathbb{P}_X)}. \end{aligned}$$

The second term on the right hand side converges zero as $y \rightarrow \hat{y}$ due to continuity of p_Y . For the first expression, we note that

$$\begin{aligned} & \left\| \sqrt{L(\hat{y}|X = \cdot)} - \sqrt{L(y|X = \cdot)} \right\|_{L^2(X, \mathbb{P}_X)}^2 \\ &= \int_{\mathcal{X}} \left(\sqrt{L(\hat{y}|X = x)} - \sqrt{L(y|X = x)} \right)^2 d\mathbb{P}_X(x) \\ &\leq \int_{\mathcal{X}} \left| \sqrt{L(\hat{y}|X = x)} - \sqrt{L(y|X = x)} \right| \cdot \left| \sqrt{L(\hat{y}|X = x)} + \sqrt{L(y|X = x)} \right| d\mathbb{P}_X(x) \\ &= \int_{\mathcal{X}} |L(\hat{y}|X = x) - L(y|X = x)| d\mathbb{P}_X(x), \end{aligned}$$

which converges to zero as $y \rightarrow \hat{y}$ as argued at the beginning of this proof. \square

Another commonly used distance function for measures is the total variation distance. For probability measures, it is usually defined as follows.

Definition 2.4 (Total variation distance). *Let (Ω, Σ) be a measure space and let μ_1 and μ_2 be two probability measures on (Ω, Σ) . We define the total variation distance between μ_1 and μ_2 as*

$$d_{\text{TV}}(\mu_1, \mu_2) = \sup_{B \in \Sigma} |\mu_1(B) - \mu_2(B)|.$$

Moreover, it is well known that, if $\mu_1, \mu_2 \ll \mu$, the total variation distance admits the representation

$$d_{\text{TV}}(\mu_1, \mu_2) = \frac{1}{2} \int \left| \frac{d\mu_1}{d\mu}(x) - \frac{d\mu_2}{d\mu}(x) \right| d\mu(x).$$

The following lemma shows that continuity w.r.t. the Hellinger distance is stronger than continuity w.r.t. to the total variation distance.

Lemma 2.4. *Let (Ω, Σ, μ) be a measure space and μ_1, μ_2 be two measures on (Ω, Σ) that are absolutely continuous w.r.t μ . Then*

$$d_{\text{TV}}(\mu_1, \mu_2) \leq \sqrt{2} d_{\text{Hel}}(\mu_1, \mu_2).$$

Proof. Define

$$\delta(x) = \sqrt{\frac{d\mu_1}{d\mu}(x)} - \sqrt{\frac{d\mu_2}{d\mu}(x)}$$

Then

$$d_{\text{Hel}}(\mu_1, \mu_2)^2 = \frac{1}{2} \int \delta(x)^2 d\mu(x).$$

Also, since

$$\left| \frac{d\mu_1}{d\mu}(x) - \frac{d\mu_2}{d\mu}(x) \right| = |\delta(x)| \left(\sqrt{\frac{d\mu_1}{d\mu}(x)} + \sqrt{\frac{d\mu_2}{d\mu}(x)} \right),$$

we obtain that

$$d_{\text{TV}}(\mu_1, \mu_2) = \frac{1}{2} \int |\delta(x)| \left(\sqrt{\frac{d\mu_1}{d\mu}(x)} + \sqrt{\frac{d\mu_2}{d\mu}(x)} \right) d\mu(x).$$

Now, by the Cauchy-Schwarz inequality

$$d_{\text{TV}}(\mu_1, \mu_2) \leq \frac{1}{2} \left(\int \delta(x)^2 d\mu(x) \right)^{1/2} \left(\int \left(\sqrt{\frac{d\mu_1}{d\mu}(x)} + \sqrt{\frac{d\mu_2}{d\mu}(x)} \right)^2 d\mu(x) \right)^{1/2}.$$

Since, again using Cauchy-Schwarz,

$$\begin{aligned} & \int \left(\sqrt{\frac{d\mu_1}{d\mu}(x)} + \sqrt{\frac{d\mu_2}{d\mu}(x)} \right)^2 d\mu(x) \\ &= \int \left(\frac{d\mu_1}{d\mu}(x) + \frac{d\mu_2}{d\mu}(x) + 2\sqrt{\frac{d\mu_1}{d\mu}(x)\frac{d\mu_2}{d\mu}(x)} \right) d\mu(x) \\ &= \int \frac{d\mu_1}{d\mu}(x) d\mu(x) + \int \frac{d\mu_2}{d\mu}(x) d\mu(x) + 2 \int \sqrt{\frac{d\mu_1}{d\mu}(x)\frac{d\mu_2}{d\mu}(x)} d\mu(x) \\ &= 1 + 1 + 2 \cdot \int \sqrt{\frac{d\mu_1}{d\mu}(x)\frac{d\mu_2}{d\mu}(x)} d\mu(x) \leq 4 \end{aligned}$$

we finally obtain that

$$d_{\text{TV}}(\mu_1, \mu_2) \leq \frac{1}{2} \cdot \sqrt{2d_{\text{Hel}}(\mu_1, \mu_2)^2} \cdot \sqrt{4} = \sqrt{2} \cdot d_{\text{Hel}}(\mu_1, \mu_2). \quad \square$$

Corollary 2.1 (Distributional stability in total variation). *In the setting of theorem 2.3, $y \mapsto \mathbb{P}_X(\cdot | Y = y)$ is continuous with respect to the total variation distance at every $y \in \mathcal{Y}$ with $p_Y(\hat{y}) \neq 0$.*

The last result is interesting in particular in view of sampling algorithms, which often show convergence of the distribution of samples to the original distribution in terms of the total variation distance. In this case, by the triangle inequality, the last result implies also stability of sampling from the posterior.

A second metric that is frequently used to analyze sampling algorithms is the Wasserstein metric. Under stronger assumptions, stability of the posterior distribution with respect to the Wasserstein metric can also be obtained [58, Section 3.5]. While the solution to the Bayesian inverse problem is formally defined as the posterior distribution, in practical applications one typically is interested in various quantities derived from the posterior distribution, such as its expectation, its modes, or the probability or quantiles of certain quantities of interest. In the

following we consider when such derived quantities also depend continuously on the data. The following lemma is a central result, as it provides stability of every quantity derived from the posterior via a function that is bounded on the support of \mathbb{P}_X .

Lemma 2.5. *In the setting of theorem 2.3, let $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$ a normed space equipped with the Borel σ -algebra and let $f : \mathcal{X} \rightarrow \mathcal{Z}$ be measurable and bounded on the support of \mathbb{P}_X (i.e., there exist $A \in \Sigma_{\mathcal{X}}$ with $\mathbb{P}_X(A) = 0$ and $C > 0$ such $\sup_{x \in \mathcal{X} \setminus A} \|f(x)\|_{\mathcal{Z}} < C$). Then*

$$y \mapsto \mathbb{E}_{\mathbb{P}_X(\cdot|Y=y)}[f]$$

is continuous at every $y \in \mathcal{Y}$ with $p_Y(y) > 0$. In particular, at every such y , the mapping

$$y \mapsto \mathbb{P}_X(A|Y=y)$$

is continuous for every $A \in \Sigma_{\mathcal{X}}$.

Proof. The proof is given by the following inequality chain for $y, \hat{y} \in \mathcal{Y}$ with $p_Y(y), p_Y(\hat{y}) \neq 0$:

$$\begin{aligned} & \left\| \int_{\mathcal{X}} f(x) d\mathbb{P}_X(\cdot|Y=y) - \int_{\mathcal{X}} f(x) d\mathbb{P}_X(\cdot|Y=\hat{y}) \right\|_{\mathcal{Z}} \\ & \leq \int_{\mathcal{X}} \left\| f(x) \frac{L(y|X=x)}{p_Y(y)} - f(x) \frac{L(\hat{y}|X=x)}{p_Y(\hat{y})} \right\|_{\mathcal{Z}} d\mathbb{P}_X(x) \\ & \leq C \int_{\mathcal{X}} \left| \frac{L(y|X=x)}{p_Y(y)} - \frac{L(\hat{y}|X=x)}{p_Y(\hat{y})} \right| d\mathbb{P}_X(x) \\ & = Cd_{\text{TV}}(\mathbb{P}_X(\cdot|Y=y), \mathbb{P}_X(\cdot|Y=\hat{y})). \end{aligned} \tag{6}$$

The second assertion follows by choosing $f = \chi_A$. \square

This result already implies that the posterior expectation depends continuously on the data if there exists some $C > 0$ such that $\mathbb{P}_X(\{x \in \mathcal{X} \mid \|x\|_{\mathcal{X}} > C\}) = 0$ (since in this case $f(x) = x$ fulfills the above assumptions), which is a reasonable assumption for instance in the case that X models image data where the pixel values are usually within a predefined maximal range. In this case, the above result with $f(x) = x^m$ also implies that all moments $m \in \mathbb{N}$, and in particular also the variance, of $\mathbb{P}_X(\cdot|Y=y)$ are finite.

Alternatively, we obtain continuity of posterior expectation under the minimal assumption that X is integrable and by assuming that the likelihood is bounded.

Lemma 2.6. *In the setting of theorem 2.3, assume in addition that there exists $C > 0$ such that $|L(y|X=x)| \leq C$ for all $y \in \mathcal{Y}$ and \mathbb{P}_X -a.e. $x \in \mathcal{X}$, and that*

$$\int_{\mathcal{X}} \|x\|_{\mathcal{X}} d\mathbb{P}_X(x) < \infty.$$

Then, the posterior expectation is continuous in all points $y \in \mathcal{Y}$ with $p_Y(y) \neq 0$, that is,

$$\mathbb{E}_{\mathbb{P}_X(\cdot|Y=y)}[x] \rightarrow \mathbb{E}_{\mathbb{P}_X(\cdot|Y=\hat{y})}[x]$$

as $y \rightarrow \hat{y}$ with $p_Y(\hat{y}) \neq 0$.

Proof. Let $y \rightarrow \hat{y}$. Then also $p_Y(y) \rightarrow p_Y(\hat{y}) > 0$ by lemma 2.3, and in particular it is bounded from below by $\delta > 0$. Then, since

$$\left\| x \frac{L(y|X=x)}{p_Y(y)} \right\|_{\mathcal{X}} \leq \frac{C}{\delta} \|x\|_{\mathcal{X}}$$

and the latter is \mathbb{P}_X -integrable, the result follows from continuity of the dominated convergence theorem. \square

Note that, as discussed in the examples above, the likelihood $L(\cdot|X = \cdot)$ is always bounded in case of Gaussian- or Poisson measurement noise, such that in these cases, the expectation depends continuously on the data as claimed.

In contrast to the expectation of the posterior, which corresponds to the minimum mean-squared-error (MMSE) estimator, it is not trivial to define and analyze a MAP estimator when \mathcal{X} is infinite-dimensional since in that case the Lebesgue measure is not available as a space-invariant underlying measure w.r.t. which the densities of the prior and the posterior can be defined. We refer to [27] for the analysis of the MAP estimator in infinite-dimensional inverse problems with Gaussian priors, to [45] for an extension to non-Gaussian priors and to [59] for an analysis of the connection of different notions of MAP estimator in infinite-dimensional inverse problems.

In the finite-dimensional setting, the MAP estimator can be defined as the maximizer of posterior density w.r.t. the Lebesgue measure. But even in this case, the MAP estimator cannot exist without further assumptions: For example, $f: \mathbb{R} \rightarrow \mathbb{R}$ with

$$f(x) = \begin{cases} \frac{n}{Z}, & \text{if } x \in [n, n + \frac{1}{n^3}[, \\ 0, & \text{otherwise,} \end{cases} \quad \text{with } Z = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

defines a probability density whose maximum does not exist. Furthermore, since the MAP estimator is defined as pointwise maximum of the posterior density, it depends on the choice of representative for this density, which is only defined Lebesgue-almost-everywhere. To avoid this, one usually requires that the density has a continuous representative, which is then unique.

Using these assumptions, a stability result for the MAP estimator in case of finite-dimensional $\mathcal{X} = \mathbb{R}^d$ is the following.

Lemma 2.7. *In the setting of theorem 2.3, assume that \mathcal{X} is a subset of a finite-dimensional space, that \mathbb{P}_X has a density w.r.t. the Lebesgue measure that we denote by p_X , that p_X is continuous, that $p_X(x_n) \rightarrow 0$ for $\|x_n\| \rightarrow \infty$ and that $(x, y) \mapsto L(y|X = x)$ is (jointly) continuous and bounded. Then, the MAP estimator defined as*

$$\hat{x}_{\text{MAP}}(y) := \arg \max_{x \in \mathcal{X}} L(y|X = x)p_X(x)$$

exists and depends continuously on y in the sense that

$$\hat{x}_{\text{MAP}}(y_n) \rightarrow \hat{x}_{\text{MAP}}(\hat{y})$$

for any sequence $(y_n)_n$ that converges to \hat{y} such that none of the functions $x \mapsto L(y_n|X = x)p_X(x)$ and $x \mapsto L(\hat{y}|X = x)p_X(x)$ is identically zero.

Proof. In case

$$x \mapsto L(\hat{y}|X = x)p_X(x)$$

is identically zero, any $x \in \mathcal{X}$ is a maximizer. In the other case, existence of the MAP estimator follows by the direct method: Take $(x_n)_n$ to be a maximizing sequence. By the assumptions on p_X and boundedness of $L(y|X = x)$ it is clear that $(x_n)_n$ must be bounded, such that there exists a subsequence that converges to some \hat{x} . The continuity of $L(\cdot|X = \cdot)$ and of p_X then imply that \hat{x} is a maximizer.

Stability follows in a very similar way: Given $(y_n)_n$ converging to \hat{y} as in the statement of this lemma, take $(x_n)_n$ to be the corresponding MAP estimators. Take \hat{x} to be a maximizer of $x \mapsto L(\hat{y}|X = x)p_X(x)$, such that in particular $L(\hat{y}|X = \hat{x})p_X(\hat{x}) > 0$. Then

$$L(y_n|X = x_n)p_X(x_n) \geq L(y_n|X = \hat{x})p_X(\hat{x}) \rightarrow L(\hat{y}|X = \hat{x})p_X(\hat{x}) > 0. \quad (7)$$

Since $0 < L(y_n|X = x_n) \rightarrow L(\hat{y}|X = \hat{x}) > 0$ there must further be a constant $C > 0$ such that $L(y_n|X = x_n) > C$ for all n , which, by dividing the above by $L(y_n|X = x_n)$, implies that $p_X(x_n)$ is bounded uniformly away from zero (for sufficiently large n). Hence $(x_n)_n$ must be bounded and, consequently, admits a convergent subsequence. The continuity of all involved quantities finally implies that $\hat{x} = \hat{x}_{\text{MAP}}(\hat{y})$. \square

Remark 2.4. *A crucial difference between the assumptions necessary for the stability of the posterior distribution and those necessary for the stability of the MAP is that the latter requires joint continuity of $(x, y) \mapsto L(y|X = x)$. While this assumption might be replaced by an appropriate semi-continuity assumption, it is still noteworthy that stability of the posterior distribution does not require any such assumption regarding the dependence of $L(y|X = x)$ on x and, consequently, does not require any specific assumption on the forward model \mathcal{F} other than measurability.*

2.1 EBMs for Bayesian Inverse Problems

Using machine learning for Bayesian inverse problems requires a practically realizable way to parametrize the prior distribution \mathbb{P}_X . Usually, this is done in the finite-dimensional setting and under the assumption that \mathbb{P}_X admits a density p_X with respect to the Lebesgue measure, which is then parametrized explicitly or implicitly. One type of parametrization of \mathbb{P}_X is via energy based models:

Definition 2.5 (Energy based models). *Given \mathbb{P}_X to be the distribution of the unknowns $x \in \mathcal{X}$ of interest, and μ_X to be a σ -finite Borel measure μ_X on \mathcal{X} such that \mathbb{P}_X is absolutely continuous w.r.t. μ_X with density p_X , an EBM is a representation of p_X via an energy functional $E : \mathcal{X} \rightarrow \mathbb{R}$ that is given as*

$$p_X(x) := \frac{\exp(-E(x))}{\int \exp(-E(x')) d\mu_X(x')}, \quad (8)$$

where $x \mapsto \exp(-E(x))$ is required to be measurable with $\int_{\mathcal{X}} \exp(-E(x)) d\mu_X(x) < \infty$. We refer to the functional E as the energy, or potential.

Remark 2.5 (The finite-dimensional case). *When $\mathcal{X} = \mathbb{R}^d$, the most relevant setting is the one where μ_X is the Lebesgue measure on \mathbb{R}^d . In this case, we write*

$$p_X(x) := \frac{\exp(-E(x))}{\int \exp(-E(y)) dy}. \quad (9)$$

Remark 2.6 (The infinite-dimensional case). *In case \mathcal{X} is infinite-dimensional, the reference measure μ_X is often the Gaussian measure, see [91] for a classical work in this context and [91, 90, 57, 50] for some works on different models.*

In most cases, especially in the context of learning based methods, the potential E will depend on a set of parameters $\theta \in \Theta$ that lie in some suitable space Θ . In such cases denote the energy as E_θ , the density as p_θ , and the corresponding distribution as \mathbb{P}_θ .

In view of the general theory for Bayesian inverse problems as outlined in section 2, it is interesting to note that, beyond non-restrictive assumptions on the measurement noise, those results do not pose any additional assumption on the (energy-based) model of the prior at all. Nevertheless, it is important to note that the most restrictive assumption on the energy is already hidden in the definition of energy-based models themselves: It is integrability of $x \mapsto \exp(-E(x))$ which already requires some coercivity of $x \mapsto E(x)$ in order to limit the tails of the distribution.

We conclude this section by providing some basic, finite-dimensional examples of energy based models; for an overview of different architectures that are used to parametrize EBMs in machine learning we refer to section 3.3 below.

Example 2.5 (Gaussian Mixture). *With $\mathcal{X} = \mathbb{R}^n$, a classical example is a Gaussian mixture model of the form*

$$E(x) = -\log \left(\sum_{i=1}^W w_i \exp \left(-\frac{(x - \mu_i) C_i^{-1} (x - \mu_i)}{2} \right) \right),$$

where the covariance matrices $C_i \in \mathbb{R}^{n \times n}$ are positive definite and $w_i \geq 0$ for all i with at least one $w_{i_0} > 0$.

Example 2.6 (Coercive energy functionals). *Again with $\mathcal{X} = \mathbb{R}^n$, any functional $E : \mathbb{R}^n \rightarrow [0, \infty)$ such that $E(x) \geq C\|x\|^q$ for all x with $C > 0$, $\|\cdot\|$ a norm on \mathbb{R}^d and $q > 0$ defines a valid energy-based model, since $x \mapsto \exp(-E(x))$ is integrable. A popular example in this context is the sparsity prior*

$$E(x) = \|Wx\|_1,$$

where $\|\cdot\|_1$ denotes the ℓ^1 norm and $W : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $m \geq n$ is an injective linear mapping such as a basis transform (e.g., Wavelets) or a frame.

Example 2.7 (Non-coercive energy functionals). *In case $\mathcal{X} = \mathbb{R}^n$ and*

$$E(x) = R(Wx)$$

with $W : \mathbb{R}^n \rightarrow \mathbb{R}^m$ linear and $R : \mathbb{R}^m \rightarrow [0, \infty)$ such that $R(x) \geq C\|x\|^q$ with $C, q > 0$, E constitutes a valid energy-based model (i.e., is integrable) if and only if W has a trivial nullspace.

However, many frequently used energies, such as the total variation [83], higher-order extensions [15] or convolutional-filter-based energy-based models [82] have a non-trivial nullspace, in particular vanish on constants or affine functions. From the perspective that $p(x) \simeq \exp(-E(x))$ is a general prior for images, this seems also reasonable since the probability of $x \in \mathbb{R}^n$ being perceived as generic, valid image should not depend on constant shifts (possibly even affine transformation) of x .

In order to still ensure validity of the resulting energy models for a concrete application at hand, one option is to infer information on the projection of the unknown to the nullspace of W (e.g., its mean) from the from modeling or data, and to penalize deviates from it as part of the prior. This would correspond to augmenting the energy-based models, e.g., via

$$E(x) = R(Wx) + \|P_{\ker(W)}x - p_0\|_2^2$$

where p_0 is inferred from modeling or the measurement data and $P_{\ker(W)}$ is the orthogonal projection to $\ker(W)$. In this way, E constitutes a well-defined energy-based model that is fine-tuned to the specific image characteristics of the problem.

An alternative that is valid in case that the forward model also vanishes on $\ker(W)$ is to realize the above-discussed Bayesian framework not on \mathbb{R}^n but on the orthogonal complement of $\ker(W)$ in \mathbb{R}^n .

3 Learning paradigms and architectures

In this section we survey prominent strategies for learning the parameters θ of a suitably parametrized energy E_θ from a finite sample of X . Specifically, we assume access to a finite

sample $x_1, x_2, \dots, x_N \in \mathcal{X}$ that is drawn i.i.d. from an unknown distribution \mathbb{P}_X and that forms the empirical distribution

$$\hat{\mathbb{P}}_X = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}. \quad (10)$$

The objective of learning is to choose θ so that the model density induced by E_θ approximates \mathbb{P}_X beyond the observed examples and, at the same time, supports downstream tasks such as the synthesis of unseen photorealistic images or the resolution of inverse problems. Importantly, the empirical distribution $\hat{\mathbb{P}}_X$ is entirely unhelpful for the synthesis of unseen images or the resolution of inverse problems since it assigns probability mass only to the training samples. The hope is that well-designed energy architectures—some of which we review in section 3.3—“generalize” meaningfully outside the training set. In practice, this generalization is typically judged by performance on downstream tasks, for such as recovering unseen samples from \mathbb{P}_X from incomplete data through the resolution of an inverse problem.

Throughout this section we restrict attention to the practical setting of finite-dimensions and always have the Lebesgue measure as the reference measure.

3.1 Divergence minimization

In such a setup, a natural way to approach parameter identification is by minimizing some divergence measure between densities. The most commonly employed divergence is the Kullback-Leibler divergence, primarily due to its connections with maximum-likelihood and maximum-entropy estimation. This approach was notably utilized in seminal works [47, 82, 104] and the Kullback-Leibler divergence has been termed the “standard” divergence by Teh, Welling, Osindero, and Hinton [94]. For two distributions \mathbb{P}_X and \mathbb{P}_Z that admit the densities p_X and p_Z , respectively, the Kullback-Leibler divergence is formally defined as

$$d_{\text{KL}}(\mathbb{P}_X, \mathbb{P}_Z) = \int_{\mathcal{X}} p_X(x) \log \frac{p_X(x)}{p_Z(x)} dx, \quad (11)$$

with the convention $0 \log \frac{0}{0} = 0$.

The Kullback-Leibler divergence places relatively strong assumptions on the involved probability distributions, notably absolute continuity. This presents a practical difficulty as the empirical distribution $\hat{\mathbb{P}}_X = (\frac{1}{N} \sum_{i=1}^N \delta_{x_i})$ is supported only on a Lebesgue-null set. This issue is typically addressed by modifying the target density through a convolution with a Gaussian kernel

$$g_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) \quad (12)$$

with variance σ^2 . The resulting smoothed empirical distribution $\hat{\mathbb{P}}_X^\sigma = g_\sigma * \hat{\mathbb{P}}_X$ then admits the density

$$p_X^\sigma = g_\sigma * \left(\frac{1}{N} \sum_{n=1}^N \delta_{x_n} \right) = \frac{1}{N} \sum_{n=1}^N g_\sigma(\cdot - x_n) \quad (13)$$

with respect to the Lebesgue measure, which is supported on the whole space and infinitely often differentiable. Since this will be the target density in learning, selecting an appropriate variance σ^2 is crucial: On the one hand, it should be sufficiently large to ensure stable training. On the other hand, it should be chosen as small as possible to prevent the loss of important structural features in the empirical distributions due to the excessive smoothing. In practice, finding an appropriate variance usually necessitates hand-tuning.

The minimization of the Kullback-Leibler divergence is directly related to maximum-likelihood estimation. Simple algebraic manipulations show that the learning objective associated with the Kullback-Leibler minimization is equivalent to the classical maximum-likelihood objective

$$\arg \min_{\theta} d_{\text{KL}}(\hat{\mathbb{P}}_X^{\sigma}, \mathbb{P}_{\theta}) = \arg \min_{\theta} \mathbb{E}_{X \sim \hat{\mathbb{P}}_X^{\sigma}} [-\log p_{\theta}(X)]. \quad (14)$$

Furthermore, substituting the Gibbs density from eq. (9) into this expression reveals the classical difference-of-expectation objective,

$$\arg \min_{\theta} \{ \mathbb{E}_{X \sim \hat{\mathbb{P}}_X^{\sigma}} [E_{\theta}(X)] - \mathbb{E}_{X \sim \mathbb{P}_{\theta}} [E_{\theta}(X)] \}. \quad (15)$$

This derivation has been used in the context of parameter identification of densities at least since the eighties [1] and was termed the *wake-sleep* algorithm [46, 36] in the context of Boltzmann machines.

The computation of the objective function in (15) necessitates the evaluation of two expectations, neither of which is typically tractable in closed form for practically relevant cases such as, for instance, when the energy is a shallow or deep neural network. In such cases, both expectations are approximated with Monte Carlo integration, and the optimization problem is resolved by employing stochastic optimizers. However, the computational cost of obtaining some fixed number of samples from each distribution differs significantly. On the one hand, sampling from the smoothed empirical distribution $\hat{\mathbb{P}}_X^{\sigma}$ is relatively inexpensive, as it can be done by ancestral sampling: drawing $\tilde{X} \sim \hat{\mathbb{P}}_X$ (*i.e.*, sampling from the dataset uniformly) and then setting $X = \tilde{X} + \sigma N$ where $N \sim \mathcal{N}(0, \mathbf{I})$. On the other hand, sampling from the model distribution is computationally intensive and generally requires Markov chain Monte Carlo (MCMC) methods. In the high-dimensional setting that is encountered in imaging problems, gradient-based methods such as Langevin or Hamiltonian dynamics are popular for this purpose. Prominent examples include Roth and Black’s fields-of-experts (FoE) model [82] which utilized Hamiltonian Monte Carlo, and more recent works by Du and Mordatch [28] and Nijkamp et al. [73] which popularized the unadjusted Langevin algorithm (ULA) in generative modeling.³ The practical utility of generative models as priors in inverse problems in imaging was further demonstrated recently in [99, 98]. These and other sampling methods that are popular in this context are discussed in more detail later in section 4.

While the sampling of the model density via such standard techniques is theoretically sound, it is often considered computationally prohibitive in practice. This has motivated research into energy functions with special structures that admit efficient sampling methods. In the context of Boltzmann machines, *restricted* Boltzmann machines [84] address this by restricting the architecture of the learned functions. In [86], Schmidt, Gao, and Roth propose a classical FoE-type energy function that utilizes Gaussian scale mixtures and enables the use of efficient auxiliary variable Gibbs methods for the sampling. For a similar energy, Weiss and Freeman [35] imposed constraints such that the normalization constant of the EBM becomes independent of the parameters. This restricts the learnable parameters to a rotation of some predefined filters with and the weights in the Gaussian scale mixture which they learn with an efficient expectation-maximization algorithm.

Despite these advances, computational challenges persist for general, unstructured energies. Hinton’s *contrastive divergence* [47] partly addresses this by abandoning exact sampling of the model distribution and instead relying on short-chain MCMC initialized from empirical samples.

³In principle, all models discussed in this chapter are “generative models” in the sense that they model a distribution and one can sample from this distribution. However, models can be better suited for the purpose of generating photo-realistic images depending on the architecture. The generative models in these papers are capable of synthesizing photorealistic images with sizes of up to 512×512 pixels.

Another strategy that is tightly linked to maximum-likelihood learning is *adversarial regularization* [62, 67, 87, 101]; We debated whether this strategy is best placed here or in the next section but decided to put it here due to the similarity of the objective function that eventually arises. However, this approach is fundamentally different from the other approaches discussed in this section in that it does not admit a straightforward probabilistic interpretation. The adversarial regularization approach can be motivated by the following interpretation of the maximum-likelihood objective expressed in (15): The first term in (15) encourages parameter configurations θ such that the energy E_θ assigns low values to samples drawn from p_X or, equivalently, high likelihood under the model. However, this term alone is not sufficient for the training of an EBM since it has many trivial solutions such as setting $E_\theta \equiv -\infty$. Consequently, the second term in (15) is crucial. This term, which evaluates the energy on samples drawn from the model distribution p_θ , ensures that trivial solutions are avoided by encouraging high energies in regions not supported by the data. Adversarial regularization builds upon this insight but generalizes it by substituting the model distribution in the second term with a suitable *adversarial* surrogate distribution \mathbb{P}_A . The resulting optimization problem becomes

$$\arg \min_{\theta} \left\{ \mathbb{E}_{X \sim \hat{\mathbb{P}}_X^\sigma} [E_\theta(X)] - \mathbb{E}_{X \sim \mathbb{P}_A} [E_\theta(X)] \right\}. \quad (16)$$

Typically, the adversarial distribution \mathbb{P}_A is chosen specifically to represent undesirable artifacts that appear in downstream inference tasks. For example, consider an inverse problem where measurements y are related to the unknown image via $y = Fx + n$ where F is some linear forward operator. A suitable adversarial distribution \mathbb{P}_A in this context might be the distribution of artifact-ridden reconstructions given by $(F^\dagger)_{\#} \mathbb{P}_Y$, where F^\dagger is some (possibly regularized) pseudo-inverse of F and \mathbb{P}_Y is the distribution of the measurements that is obtained through the measurement model. By training the energy to assign low values to clean data samples and high values to artifact-ridden samples, subsequent inference based on variational formulations that leverage E_θ tends to yield artifact-free reconstructions. Though the objective is similar, adversarial regularization does not generally admit a straightforward probabilistic interpretation. Moreover, as the training explicitly incorporates the forward operator, the learned model becomes task-specific which breaks the clear Bayesian separation between likelihood and prior.

Another training method that has gained significant popularity in recent years, partly due to its explicit alignment with Bayesian principle, is based on minimizing the Fisher divergence. For two distributions \mathbb{P}_X and \mathbb{P}_Z that admit the continuously differentiable densities p_X and p_Z with respect to the Lebesgue measure, respectively, the Fisher divergence is formally defined as

$$d_F(\mathbb{P}_X, \mathbb{P}_Z) = \mathbb{E}_{X \sim \mathbb{P}_X} [\|\nabla \log p_X(X) - \nabla \log p_Z(X)\|^2]. \quad (17)$$

Since the formal definition requires differentiable densities, like with the Kullback-Leibler divergence, the empirical distribution $\hat{\mathbb{P}}_X$ is unsuited and we define the target distribution as $\hat{\mathbb{P}}_X^\sigma$ that has the density p_X^σ , given in (13), with respect to the Lebesgue measure. We will see later that this choice of smoothing is actually critical for the derivation of a practical and efficient learning objective. Learning a parametrized energy model then involves finding parameters θ that minimize the Fisher divergence between the target distribution and the model distribution. Explicitly, the training objective becomes

$$\arg \min_{\theta \in \Theta} d_F(\hat{\mathbb{P}}_X^\sigma, \mathbb{P}_\theta) = \arg \min_{\theta \in \Theta} \mathbb{E}_{X \sim \hat{\mathbb{P}}_X^\sigma} [\|\nabla \log p_X^\sigma(X) + \nabla E_\theta(X)\|^2]. \quad (18)$$

Thus, the objective aims to match the gradient of the log-density—also called the (Stein) score—of the model to that of the smoothed data density.

A direct evaluation of the objective in (18) is computationally expensive since the evaluation of $\nabla \log p_X^\sigma = \nabla(x \mapsto \log((g_\sigma * (\frac{1}{N} \sum_{n=1}^N \delta_{x_n}))(x)))$ at any point involves computations on the whole dataset. A critical observation due to Vincent [96] is that (18) can be reformulated equivalently in a substantially simpler manner. Specifically, (18) is equivalent to solving

$$\arg \min_{\theta \in \Theta} \mathbb{E}_{X \sim \hat{P}_X, N \sim \mathcal{N}(0,1)} [\|\sigma \nabla E_\theta(X + \sigma N) - N\|^2]. \quad (19)$$

This reformulation reveals an intuitive interpretation: the energy functional is trained so that its scaled gradient acts as an mean-squared error (MSE)-optimal one-step denoiser, and the objective is typically referred to as denoising score-matching.

To better understand this, denote $Z = X + \sigma N$. The expression within the norm in (19) can be rewritten as

$$\sigma \nabla E_\theta(Z) - N = \frac{1}{\sigma}(X - Z + \sigma^2 \nabla E_\theta(Z)). \quad (20)$$

From this viewpoint, the learning objective seeks parameters such that, for every data point X sampled from the empirical distribution and corresponding noisy observation $Z = X + \sigma N$, the estimator

$$Z - \sigma^2 \nabla E_\theta(Z) \quad (21)$$

approximates the MMSE estimator of the original data point X . Conversely, the classical result known as Tweedie’s formula shows that given the density of Y , one can directly construct the MMSE estimator of X via the relationship

$$Z \mapsto Z + \sigma^2 \nabla \log p_Z(Z). \quad (22)$$

This deep connection between denoising and density estimation underpins some of today’s most effective approaches to image reconstruction, prominently plug-and-play methods, and image generation algorithms such as diffusion models [26, 51].

3.2 Alternative learning methods

In contrast to the previous section, where learning was formulated explicitly as a divergence minimization problem that aligns closely with the Bayesian paradigm that we consider in this work, the approaches described in this section diverge from strict Bayesian interpretations. Instead, the purpose of this section is purely *operational*: we present several training procedures that are designed to give a useful E_θ that can later be exploited as a prior in the resolution of inverse problems. Importantly, the methods discussed in this section should *not* be viewed as attempts to faithfully learn a true underlying density; rather, they constitute pragmatic procedures for obtaining practically useful energies.

Bilevel approaches, pioneered by Samuel [85] and Tappen [93], tackle the problem of identifying optimal parameters of an energy by formulating a nested optimization problem. The lower-level problem is the standard variational formulation that aims to recover some clean image from incomplete measurements, where the energy that we aim to learn serves as a regularizer. The resolution of this lower-level problem yields a parameter-dependent estimate. The upper-level problem is to minimize some loss function (*e.g.*, the norm of the difference between the estimate and the clean image) that involves this parameter-dependent estimate with respect to the parameters of the energy. More rigorously, such methods are formulated as the optimization problem of finding

$$\arg \min_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^N L(x_n^*(\theta), x_n) \quad (23)$$

where $x_n^*(\theta) \in \arg \min_{x \in \mathcal{X}} \{J(x, \theta) = \mathcal{D}_{y_n}(\mathcal{F}(x)) + \lambda E_\theta(x)\}$ for $n = 1, 2, \dots, N$,

where y_1, y_2, \dots, y_N are the data that are typically constructed by utilizing the forward model \mathcal{F} and some pollution operator, and \mathcal{D} is an appropriate discrepancy measure.

An interesting connection between bilevel approaches (or discriminative approaches in general) and Bayesian approaches is that bilevel approaches explicitly seek energies whose MAP estimators approximate the Bayes estimate with respect to the loss L , *i.e.*, after training for any data y we have

$$\arg \min_{x \in \mathcal{X}} \{\mathcal{D}_y(\mathcal{F}(x)) + \lambda E_\theta(x)\} \approx \arg \min_{x \in \mathcal{X}} \mathbb{E}_{X \sim \mathbb{P}_X(\cdot | Y=y)} [L(x, X)]$$

where $\mathbb{P}_X(\cdot | Y=y)$ is the posterior distribution of the inverse problem of interest after observing the data y . As a practically relevant example, when $L(x, y) = \|y - x\|^2/2$, the solution of the lower-level problem should approximate the MSE-optimal Bayes estimator. While the resulting energy functional is thus optimal in a MAP sense relative to a chosen discrepancy measure, it is important to recognize that this optimality is *task-specific* and does not necessarily reflect a correct probabilistic interpretation of the learned model.

Bilevel learning necessitates the resolution of the lower-level problem—typically to very high precision (for a visualization of the resulting test PSNR depending on the precision of the resolution of the lower-level problem, see [23])—as well as the computation of the gradient of the upper-level loss with respect to the parameters. When the lower-level problem is sufficiently smooth, researchers typically use accelerated first-order methods for its resolution, such as Nesterov’s accelerated gradient algorithm, which is also utilized in the numerical sections of this work and given in algorithm 8. The computation of the gradient of the upper-level loss function with respect to the parameters can be tackled by various approaches that differ in their precision and memory footprint. In the following, we only provide a narrow discussion of one of the existing methods which, in particular, covers the most frequent applications which include the ones in this paper. A broader perspective that also includes discussions on potentially nonsmooth lower-level problems is provided in [52, 105, 13]. The following derivation is provided for $N = 1$ and, therefore, we omit the subscript that specifies the data index. The result can be readily generalized to $N > 1$ through appropriate summation of the quantities.

A popular method for the resolution of bilevel problems arises through the differentiation of the optimality condition

$$\nabla_x J(x^*(\theta), \theta) = 0 \tag{24}$$

of the lower-level problem in eq. (23). When J is twice continuously differentiable and its Hessian H is invertible at $(x^*(\theta), \theta)$, then the implicit function theorem implies that $x^*(\theta)$ is locally unique. Differentiating (24) with respect to θ yields

$$0 = H(\theta)(x^*(\theta))' + \nabla_\theta \nabla_x J(x^*(\theta), \theta) \tag{25}$$

which enables the computation of $(x^*(\theta))'$ as

$$(x^*(\theta))' = (H(\theta))^{-1} \nabla_\theta \nabla_x J(x^*(\theta), \theta). \tag{26}$$

Plugging this into the gradient of the upper-level problem, that can be computed with the chain rule as $\nabla L(\theta) = ((x^*(\theta))')^\top \nabla_x L(x^*(\theta))$ yields

$$\nabla L(\theta) = (\nabla_\theta \nabla_x J(x^*(\theta), \theta))^\top (H(\theta))^{-1} \nabla_x L(x^*(\theta)). \tag{27}$$

This strategy based on the implicit function theorem has the benefit that—in contrast to, *e.g.*, unrolling approaches—the complexity of the gradient computation is independent of the algorithm (and, in particular, the number of iterations of that algorithm) that was used to resolve the lower-level problem. A drawback of this strategy is that the Hessian-vector products can be

costly, in particular when the energy is complex and the computation of the involved quantities by hand is infeasible and one has to resort to automatic differentiation for that computation.

Many other operational strategies have been explored in the literature. Many of these also deviate from the pure Bayesian setting by directly optimizing a performance-oriented criterion. Notably, algorithm unrolling approaches such as [42, 64, 75, 66] explicitly differentiate through iterative solvers of variational problems, thereby directly aligning the learning objective with practical reconstruction quality. Related approaches [55, 32] reframe image reconstruction as an optimal control problem and learn an optimal stopping time in the continuous gradient flow associated with the variational problem. These operational strategies underscore the practical value of performance-driven learning, even as they deliberately forego a strict Bayesian interpretation.

3.3 Architectures of EBMs

The training methods discussed in the previous section can, in principle, be applied to any function E_θ that maps from $\mathcal{X} = \mathbb{R}^n$ to the real line. In recent years, several specific architectures with varying characteristics have emerged.

A common starting point for many energy-based architectures is the anisotropic total variation defined by

$$x \mapsto \sum_{i=1}^n \sum_{j=1}^2 |(D_j x)_i| \quad (28)$$

where $D_1, D_2 : \mathbb{R}^n \rightarrow \mathbb{R}^n$ are finite-difference operators in the first and second spatial dimensions. Intuitively, this energy measures the sum of the absolute values of image gradients across all pixels in the image. An extremely common generalization of the anisotropic total variation energy is the FoE energy

$$x \mapsto \sum_{i=1}^n \sum_{j=1}^k \phi_j((K_j x)_i) \quad (29)$$

proposed in [82], where $K_1, K_2, \dots, K_k \in \mathbb{R}^{n \times n}$ are convolution matrices and $\phi_1, \phi_2, \dots, \phi_k : \mathbb{R} \rightarrow \mathbb{R}$ are the associated nonlinear *potentials*.

A large body of work deals with this model and works vary mostly in the training routine and the parametrization of the potentials. Early works, including the original original FoE publication [82], drew inspiration from classical regularization theory [102] and employed rigid parametric potentials such as those derived from negative-log Gaussian, generalized Laplacian, or Student-t densities. These potentials typically feature a global minimum at zero and increase monotonically away from it.

However, from the Bayesian perspective adopted in this work—where the energy should model the negative log-prior—such parametric forms are too restrictive. This limitation was highlighted by Zhu and Mumford [102], who proposed piecewise-constant potentials with arbitrary shapes to capture richer statistics of natural images. They recover potentials that sometimes feature a *maximum* at zero and decrease monotonically away from zero [102, fig. 9]. However, piecewise constant potentials hinder the application of first-order optimization-based image reconstruction approaches and sampling methods that rely on gradient information.

In [86] Schmidt, Gao, and Roth revisited this model and identified that it could not reproduce the marginal statistics of the responses of gradient filters, despite the maximum-likelihood training. They attribute this largely to the inefficient MCMC sampler that was used in the training of the original model. To remedy this, they propose to use potentials derived from Gaussian scale mixtures, that enable efficient sampling via an auxiliary variable Gibbs sampler. Although this improved results, they still fell short of reproducing marginal statistics of random filters, likely

due to the restrictive choice of Gaussian scale mixtures. While more general than those derived from Student-t distributions (indeed, Student-t distributions can be represented by Gaussian scale mixtures with infinitely many components), the potentials derived from Gaussian scale mixtures are still monotonically increasing away from zero. In unrelated work, Heess, Williams, and Hinton [44] also identified the limitation of the choice of the potentials in the original FoE model. The authors propose to replace the unimodal potentials derived from the Student-t distribution with slightly more general bimodal potentials, and showed improved performance on texture synthesis tasks.

Similar observations were made in the context of data-driven discriminative approaches: For example, Chen and Pock’s trainable nonlinear reaction diffusion model [22], which falls under the class of learned optimization schemes, employs a general parametrization of the potentials using Gaussian basis functions. They recover more complex potentials, such as negative Mexican-hat-type or double-well-type potentials [22, fig. 5] with multiple local minima that do not contain zero.

In the context of learning the parameters of an energy through optimal control, the authors of [32] parameterize the potentials of an FoE with B-splines and recover complex potentials with multiple local minima and often times maxima at zero. In the context of diffusion models, the authors of [100] leverage potentials parametrized by negative-log Gaussian mixture models which enables them to implement the Gaussian smoothing of the prior by adapting the variances of the one-dimensional Gaussian mixture models. The obtained potentials also differ significantly to the standard ones.

Beyond improvements in potential parametrizations, recent architectures have generalized the basic structure of the FoE model by integrating deeper, learned feature representations. Kobler et al. [55] introduced an energy termed the *total deep variation*, defined as

$$x \mapsto \sum_{i=1}^n \phi((\mathcal{N}(x))_i) \quad (30)$$

where $\mathcal{N} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a neural network. This model preserved the pixel-wise summation structure but replaces linear convolutional filters with *deep neural features*. Coupled with a learning strategy based on optimal control, the model achieved state-of-the-art results on various imaging tasks.

More recently, researchers have abandoned pixel-wise energies summations entirely. Nijkamp et al. [73, 74] and Du and Mordatch [28] advocated architectures based on fully convolutional neural networks defined as

$$E_\theta = L_l \circ L_{l-1} \circ \dots \circ L_2 \circ L_1, \quad (31)$$

where each layer L_1, L_2, \dots, L_l has the form

$$L_i : x \mapsto \tilde{\Phi}_i(\tilde{W}_i \Phi_i(W_i x)). \quad (32)$$

Here, W_1, W_2, \dots, W_l and $\tilde{W}_1, \tilde{W}_2, \dots, \tilde{W}_l$ are multi-channel convolution operators and $\tilde{W}_1, \tilde{W}_2, \dots, \tilde{W}_l$ operate with a stride that is greater than one and thereby reduce the size of the feature maps in deeper layers, eventually mapping the image to the scalar energy value through a 1×1 convolution with one output channel. The functions $\tilde{\Phi}_i$ and Φ_i apply some nonlinearity point-wise, typically standard neural-network-activation function such as the rectified linear unit, its leaky variant, the softplus, or others. In contrast to the FoE model, there typically are no learnable parameters associated with these nonlinearities.

Such architectures are also loosely linked to the FoE architecture: Instead of replacing the linear features that are fed into the potential with deep neural features—like in the total deep variation—such architectures can be thought of as replacing the linear pixel-wise sum with a *deep*

neural sum \mathcal{S} , since the first layer L_1 has the familiar structure of applying a nonlinearity to linear features:

$$x \mapsto \mathcal{S}(\Phi_1(W_1x)), \quad (33)$$

where $\mathcal{S} = L_l \circ \dots \circ L_1 \circ \tilde{\Phi}_1 \circ \tilde{W}_1$. These architectures also have a strong relationship to U-Net type architectures, which are an extremely popular choice for the direct modeling of the gradient of an energy. Indeed, the gradient of such architectures resembles a U-Net, see [97, fig. 5.6] for an illustration. A similar architecture was used in the context of inverse problems by the authors of [62] in adversarial learning and by the authors of [98, 99] for the purpose of learning an energy that is suitable for medical image reconstruction.

This trend toward increasingly flexible architectures underscores a broader principle: *any* function from the image domain to real numbers can serve as an energy. Indeed, modern architectures encompass various deep neural networks including pixel-wise output sums, U-Net structures, and transformer-based models.

4 Sampling from EBMs

The task of sampling from distributions that are represented as EBMs arises at many occasions in the context of Bayesian inverse problems. As an example, the estimation of various quantities derived from the posterior, such as its expectation or its marginal variance, via Monte-Carlo integration (see section 2, or [98, 99, 70]) necessitates sampling. When the energy is learned from data, sampling is often already an integral part of model training (see section 3.1). Consequently, efficient and flexible sampling algorithms are of the utmost practical relevance.

For the remainder of this section we again restrict ourselves to the finite-dimensional setting. We assume that we are given a distribution π on $\mathcal{B}(\mathbb{R}^d)$ that admits a density p with respect to the Lebesgue measure that is modeled through the energy E , *i.e.*,

$$\frac{d\pi}{dx}(x) = p(x) = \frac{\exp(-E(x))}{\int \exp(-E(\tilde{x})) d\tilde{x}}. \quad (34)$$

Further, we assume that the energy $E : \mathbb{R}^d \rightarrow \mathbb{R}$ is continuously differentiable. This setting covers sampling from priors as well as posteriors simply by setting $\pi = \mathbb{P}_X$ in the former and $\pi = \mathbb{P}_X(\cdot | Y = y)$ in the latter case. We will also sometimes add subscripts to the probability distributions, *e.g.*, π_X , π_V , π_Z , if necessary to distinguish distributions of different random variables. The corresponding Lebesgue densities will be denoted as p_X , p_V , and p_Z , respectively.

Sampling algorithms may be separated into two distinct categories: (i) *direct samplers* which simply yield a random variable $X \sim \pi$ of the target distribution and (ii) *Markov chain* based samplers, where a Markov chain (MC) $(X_k)_k$ is constructed whose stationary measure is the target π or an approximation thereof so that an approximate sample of π may be obtained by simulating $(X_k)_k$ for sufficiently many iterations. For the remainder of this section we will first briefly discuss several direct sampling techniques. Due to the complexity of the energy $E(x)$ in imaging applications, however, direct sampling is often not possible and, therefore, MC based methods are significantly more popular. The main part of the section will consequently be devoted to the most important MC samplers in the context of EBMs and Bayesian imaging. Section 4.1 may be skipped in case the reader is interested in MC based methods.

4.1 Direct sampling techniques

4.1.1 Accept-reject sampling

Accept-reject sampling [80, Section 2.3] can be understood as a specific implementation of the fundamental theorem of simulation, which states that simulating a random variable $X \sim \pi_X$ where π_X admits a density p_X is equivalent to simulating a pair (X, U) which is distributed uniformly on the set $\{(x, u) \mid 0 \leq u \leq p_X(x)\}$.

Theorem 4.1 (Fundamental theorem of simulation). *Let $\pi_X(dx) = p_X(x)dx$ be a distribution on \mathbb{R}^d and (X, U) be uniformly distributed on $\{(x, u) \mid 0 \leq u \leq p_X(x)\}$, then $X \sim \pi_X$.*

Proof. First note that

$$\int_{\mathbb{R}^d} \int_0^{p(x)} du dx = 1$$

so that the density of (X, U) with respect to the Lebesgue measure is simply the indicator function on $\{(x, u) \mid 0 \leq u \leq p(x)(x)\}$. Thus, we can conclude

$$\mathbb{P}[X \in A] = \int_{\mathbb{R}^d} \int_0^{p(x)} \mathbb{1}_A(x) du dx = \int_{\mathbb{R}^d} \mathbb{1}_A(x) p(x) dx. \quad (35)$$

The difficulty in utilizing the fundamental theorem of simulation for practical purposes of course lies in the task of generating a uniform sample on $\{(x, u) \mid 0 \leq u \leq p_X(x)\}$. A possible approach would be to sample $X \sim \pi_X$ and then $U|X \sim \mathcal{U}([0, p_X(X)])$ ⁴. However, this would, of course, be pointless as our goal is precisely to sample $X \sim \pi_X$ which would then be achieved already. In other words, the fundamental theorem of simulation only constitutes an advantage if we can find a way of sampling from $\mathcal{U}(\{(x, u) \mid 0 \leq u \leq p_X(x)\})$ which is easier than the original problem of sampling from π_X .

Accept-reject sampling may achieve this task in some cases. For accept-reject sampling it is sufficient to have access to an unnormalized version of the density p_X , *i.e.*, to a function \tilde{p}_X such that

$$p_X(x) = \frac{1}{Z_X} \tilde{p}_X(x)$$

with $Z_X = \int \tilde{p}_X(z) dz$. Let us assume we have access to a second probability distribution $\pi_Z(dz) = p_Z(z) dz$ which (a) is easier to sample from than π_X and (b) satisfies that

$$\sup_x \frac{\tilde{p}_X(x)}{p_Z(x)} \leq M < \infty. \quad (36)$$

In this case we may simulate a sample from $\mathcal{U}(\{(x, u) \mid 0 \leq u \leq p_X(x)\})$ only through sampling from π_Z which in turn yields a sample from π_X . The procedure is summarized in algorithm 1 and we provide a formal proof of its consistency in theorem 4.2. Within rejection sampling, the challenge is therefore shifted to finding a easy-to-sample distribution π_Z which satisfies (36) together with the respective bound M .

Theorem 4.2. *Assume that p_X and p_Z are both strictly positive and let X be generated via the accept-reject sampler (1). Then $X \sim \pi_X$.*

⁴ $\mathcal{U}(A)$ denotes the uniform distribution on the set A

Algorithm 1 Accept-reject sampling.

Require: Distribution π_Z and $M > 0$ satisfying (36).

- 1: Draw $Z \sim \pi_Z$
 - 2: Draw $U \sim \mathcal{U}([0, 1])$
 - 3: **if** $U \leq \frac{\tilde{p}_X(Z)}{Mp_Z(Z)}$ **then**
 - 4: $X = Z$
 - 5: **else**
 - 6: Go to line 1.
 - 7: **end if**
 - 8: **return** X
-

Proof. We can equivalently phrase the algorithm as sampling first $Z \sim \pi_Z$, then $U|Z \sim \mathcal{U}([0, Mp_Z(Z)])$, and accepting $X = Z$ if and only if $U \leq \tilde{p}_X(Z)$. If we simply show now that $(X, U) \sim \mathcal{U}(\{(x, u) \mid 0 \leq u \leq p_X(x)\})$ we can conclude using theorem 4.1. First note that, since $\tilde{p}_X \leq Mp_Z$

$$\begin{aligned} \mathbb{P}[U \leq \tilde{p}_X(Z)] &= \int_{\mathbb{R}^d} p_Z(z) \frac{1}{Mp_Z(z)} \int_0^{Mp_Z(z)} \mathbb{1}_{\tilde{p}_X(z)}(u) du dz \\ &= \int_{\mathbb{R}^d} \frac{1}{M} \int_0^{\tilde{p}_X(z)} du dz = \frac{Z_X}{M}. \end{aligned} \quad (37)$$

We can conclude for any Borel measurable $A \subset \mathbb{R}^d$ and $B \subset \mathbb{R}$

$$\begin{aligned} \mathbb{P}[(X, U) \in A \times B] &= \mathbb{P}[(Z, U) \in A \times B \mid U \leq \tilde{p}_X(Z)] \\ &= \frac{\mathbb{P}[(Z, U) \in A \times B, U \leq \tilde{p}_X(Z)]}{\mathbb{P}[U \leq \tilde{p}_X(Z)]} \\ &= \frac{M}{Z_X} \int_{\mathbb{R}^d} \mathbb{1}_A(z) p_Z(z) \frac{1}{Mp_Z(z)} \int_0^{Mp_Z(z)} \mathbb{1}_B(u) \mathbb{1}_{[0, \tilde{p}_X(z)]}(u) du dz \\ &= \frac{1}{Z_X} \int_{\mathbb{R}^d} \mathbb{1}_A(z) \int_0^{\tilde{p}_X(z)} \mathbb{1}_B(u) du dz \\ &= \int_{\mathbb{R}^d} \mathbb{1}_A(z) \int_0^{p_X(z)} \mathbb{1}_B(\tilde{u}) d\tilde{u} dz \end{aligned} \quad (38)$$

where the last equality follows from the transformation $\tilde{u} = \frac{u}{Z_X}$. \square

4.1.2 Importance sampling

Contrary to the name, importance sampling (IS) is mostly not used for *sampling* itself but instead as a method for estimating statistics of one distribution using a sample of a *different* distribution. Similar to the accept-reject sampler, let us assume we have access to a second distribution π_Z which is easier to sample from than π_X and such that $\pi_X \ll \pi_Z$ ⁵. IS relies on the following elementary observation that for any function f such that the corresponding integral exists, we have

$$\begin{aligned} \mathbb{E}_{X \sim \pi_X} [f(X)] &= \int f(x) d\pi_X(x) = \int f(x) \frac{d\pi_X}{d\pi_Z}(x) d\pi_Z(x) \\ &= \mathbb{E}_{X \sim \pi_Z} \left[f(X) \frac{d\pi_X}{d\pi_Z}(X) \right]. \end{aligned} \quad (39)$$

⁵recall, that \ll denotes absolute continuity

That is the expectation with respect to π_X can be transformed to an expectation with respect to π_Z by re-weighting the integrand f with the corresponding Radon-Nikodým derivative. In the following we assume for simplicity that both π_X and π_Z admit densities, so that the Radon-Nikodým derivative reduces to the ratio $\frac{p_X}{p_Z}$. For a specific Monte Carlo estimate, (39) reads as

$$\mathbb{E}_{X \sim \pi_X} [f(X)] \approx \frac{1}{N} \sum_{i=1}^N f(Z_i) \frac{p_X(Z_i)}{p_Z(Z_i)} \quad (40)$$

where $(Z_i)_i$ is an i.i.d sample with $Z_i \sim \pi_Z$. The approximation (40) is known as *unnormalized* IS as we do not normalize the weights $W_i = \frac{p_X(Z_i)}{p_Z(Z_i)}$. One may easily check that (40) is, in fact, an unbiased estimator for $\mathbb{E}_{X \sim \pi_X} [f(X)]$. Assuming unnormalized versions of the densities, *i.e.*, \tilde{p}_X , and \tilde{p}_Z such that

$$p_X(x) = \frac{1}{Z_X} \tilde{p}_X(x)$$

and analogously for p_Z with partition function Z_Z instead of Z_X we can rewrite (40) as

$$\mathbb{E}_{X \sim \pi_X} [f(X)] \approx \frac{1}{N} \sum_{i=1}^N f(Z_i) \frac{Z_Z \tilde{p}_X(Z_i)}{Z_X \tilde{p}_Z(Z_i)} \quad (41)$$

where we denote the weights $\tilde{W}_i = \frac{\tilde{p}_X(Z_i)}{\tilde{p}_Z(Z_i)}$. In order to estimate the ratio of the partition functions $\frac{Z_Z}{Z_X}$ we compute

$$\begin{aligned} \frac{Z_X}{Z_Z} &= \frac{1}{Z_Z} \int \tilde{p}_X(x) dx = \frac{1}{Z_Z} \int \frac{\tilde{p}_X(x)}{\tilde{p}_Z(x)} \tilde{p}_Z(x) dx = \int \frac{\tilde{p}_X(x)}{\tilde{p}_Z(x)} p_Z(x) dx \\ &\approx \frac{1}{N} \sum_{i=1}^N \frac{\tilde{p}_X(Z_i)}{\tilde{p}_Z(Z_i)} = \frac{1}{N} \sum_{i=1}^N \tilde{W}_i. \end{aligned} \quad (42)$$

Collecting these results we arrive at the *self-normalized* IS,

$$\mathbb{E}_{X \sim \pi_X} [f(X)] \approx \frac{\sum_{i=1}^N f(Z_i) \tilde{W}_i}{\sum_{i=1}^N \tilde{W}_i} \quad (43)$$

where we recall, $(Z_i)_i$ is i.i.d, $Z_i \sim \pi_Z$ and $\tilde{W}_i = \frac{\tilde{p}_X(Z_i)}{\tilde{p}_Z(Z_i)}$. While, as mentioned above, IS is primarily used for the approximation of a specific expectation, it is also possible to obtain samples from π_X via a strategy coined sampling importance resampling (SIR) [18]. In SIR an approximate sample from π_X is obtained by first sampling from π_Z and afterwards drawing from this sample using the self-normalized importance weights. The procedure is summarized in algorithm 2 and we refer to

4.2 Some preliminaries for Markov chains

Before discussing specific MC based samplers we have to introduce some basic terminology. Let $R : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1]$ be a Markov kernel. We define the action of R on a probability measure μ as

$$\mu R(A) := \int R(x, A) \mu(dx), \quad A \in \mathcal{B}(\mathbb{R}^d). \quad (44)$$

Algorithm 2 Sampling importance resampling.

Require: Distribution π_Z , $N, M > 0$.

- 1: Draw $Z_1, \dots, Z_M \sim \pi_Z$ i.i.d
- 2: Compute self-normalized importance weights

$$W_i = \frac{\tilde{p}_X(Z_i)/\tilde{p}_Z(Z_i)}{\sum_{j=1}^M \tilde{p}_X(Z_j)/\tilde{p}_Z(Z_j)}$$

- 3: Draw X_1, \dots, X_N i.i.d from the set $\{Z_1, \dots, Z_M\}$ with distribution

$$\mathbb{P}[X_1 = Z_i] = W_i, \quad \text{for } i = 1, \dots, M.$$

- 4: **return** X_1, \dots, X_N
-

In the following we restrict our discussions to time-homogeneous MCs. Therefore, we can identify a Markov chain with its Markov or transition kernel R defined via

$$\mathbb{P}(X_{k+1} \in A \mid X_k = x) = R(x, A), \quad x \in \mathbb{R}^d, A \in \mathcal{B}(\mathbb{R}^d),$$

which holds for all k due to the assumed time-homogeneity. If $(X_k)_k$ is a Markov chain with transition kernel R and initial distribution $X_0 \sim \mu$, it, thus, follows $X_k \sim \mu R^k$. The convergence results of the various sampling algorithms will moreover rely on the notion of a stationary distribution.

Definition 4.1 (Stationary distribution). *Let $R : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1]$ be a Markov kernel. We call a probability distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$ stationary or invariant for R if for any $A \in \mathcal{B}(\mathbb{R}^d)$, $\mu R(A) = \mu(A)$, that is,*

$$\mu(A) = \int R(x, A) \mu(dx).$$

Via the equivalence between (time-homogeneous) MCs and Markov kernels, the notion of a stationary distribution applies to MCs as well. If a MC converges to its unique stationary distribution (in some metric) we will refer to the chain as *ergodic*.

4.3 The Metropolis-Hastings algorithm

We begin this exposition with the famous Metropolis-Hastings (MH) algorithm which constitutes a highly flexible and simple method for sampling from a broad range of distributions. The MH algorithm endows any Markov transition kernel with a subsequent accept or reject step that ensures that the resulting Markov chain admits the target as its stationary distribution. Formally, let $Q : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1]$ be a *proposal* transition kernel with density q , i.e., $Q(x, A) = \int_A q(x, v) dv$ for all $x \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$. Given the current iterate X_k , the MH algorithm proceeds by sampling $V_{k+1} \sim Q(X_k, \cdot)$ from the proposal and subsequently setting $X_{k+1} = V_{k+1}$ with probability $\rho(X_k, V_{k+1})$ where

$$\rho(x, v) = \begin{cases} \min \left\{ \frac{p(v)q(v, x)}{p(x)q(x, v)}, 1 \right\}, & \text{if } p(x)q(x, v) > 0 \\ 1, & \text{else.} \end{cases}$$

and otherwise setting $X_{k+1} = X_k$. The algorithm is summarized in algorithm 3.

Algorithm 3 The Metropolis-Hastings algorithm.

Require: Initial value X_0 , proposal density q

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: $V_{k+1} \sim Q(X_k, \cdot)$
 - 3: $X_{k+1} = \begin{cases} V_{k+1} & \text{with probability } \rho(X_k, V_{k+1}) \\ X_k & \text{else.} \end{cases}$
 - 4: **end for**
-

Remark 4.1. For an EBM, $\frac{p(v)}{p(x)} = \exp(E(x) - E(v))$, which enables an efficient computation of the acceptance probability without knowledge of the normalization constant $\int \exp(-E(x)) dx$.

Let us denote the transition kernel of the MH chain as R . The following theorem shows that π is a stationary distribution of the MH chain without making any specific additional assumptions.

Theorem 4.3. The transition kernel $R : \mathcal{X} \times \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1]$ of a MH chain satisfies the so-called detailed balance condition with π . That is, for any measurable and bounded function $f : \mathbb{R}^{2d} \rightarrow \mathbb{R}$

$$\iint f(x, v) R(x, dv) d\pi(x) = \iint f(x, v) R(v, dx) d\pi(v). \quad (45)$$

As a consequence, the MH chain admits π as a stationary distribution.

Proof. First of all, by distinguishing the cases $x = v$ and $x \neq v$ it is easy to check that the transition kernel of the chain reads as

$$R(x, dv) = q(x, v)\rho(x, v)dv + \delta_x(dv) \underbrace{\left(1 - \int q(x, z)\rho(x, z)dz\right)}_{=\mathbb{P}[\text{make any proposal and reject it} \mid x]}. \quad (46)$$

Now let $f : \mathbb{R}^{2d} \rightarrow \mathbb{R}$ be any bounded and measurable function. Since $p(x)q(x, v)\rho(x, v) = p(v)q(v, x)\rho(v, x)$ and recalling that p is the density of π with respect to the Lebesgue measure, we find that

$$\begin{aligned} & \iint f(x, v)p(x)R(x, dv)dx \\ &= \iint f(x, v)p(x)q(x, v)\rho(x, v)dvdx \\ & \quad + \int f(x, x)p(x) \left(1 - \int q(x, z)\rho(x, z)dz\right) dx \\ &= \iint f(x, v)p(v)q(v, x)\rho(v, x)dvdx \\ & \quad + \int f(v, v)p(v) \left(1 - \int q(v, z)\rho(v, z)dz\right) dv \\ &= \iint f(x, v)p(v)R(v, dx)dv. \end{aligned} \quad (47)$$

Detailed balance implies stationarity since for $A \in \mathcal{B}(\mathbb{R}^d)$ by setting $f(x, v) = \mathbb{1}_A(v)$ it follows using Fubini's theorem that

$$\begin{aligned} \pi R(A) &= \iint \mathbb{1}_A(v)p(x)R(x, dv)dx = \iint \mathbb{1}_A(v)p(v)R(v, dx)dv \\ &= \int_A p(v)dv = \pi(A). \end{aligned} \quad (48)$$

□

Note that π being a stationary measure does not automatically imply ergodicity, that is, convergence of MC generated by the MH algorithm to π . However, ergodicity is also obtained under rather mild conditions. Since the underlying theory on ergodicity of Markov chains is, however, quite involved we only give the result without proof here and refer to [65, 80] for more details.

Theorem 4.4. [80, Corollary 7.5] *Assume that the proposal density q satisfies*

$$\mathbb{P}(p(X_k)q(X_k, V_{k+1}) \leq p(V_{k+1})q(V_{k+1}, X_k)) < 1 \quad (49)$$

and that $q(x, v) > 0$ for any x, v . Then, the MH chain is ergodic, that is, for any initial distribution μ ,

$$\lim_{n \rightarrow \infty} d_{\text{TV}}(\mu R^n, \pi) = 0$$

where μR^n denotes the application of n MH steps to the initial distribution.

The condition (49) implies that the proposal density need not satisfy detailed balance itself. Indeed, if that were the case, the inclusion of the MH correction step would be unnecessary and the chain generated by q can be analyzed directly. Altogether, the conditions on the proposal density q are not too restrictive and several MH algorithms utilizing different proposal densities have been proposed in the literature. Many of the proposal densities are based on the discretization of continuous-time diffusion stochastic differential equations (SDEs). For instance, correcting the unadjusted Langevin algorithm (ULA) that is discussed later via a Metropolis-Hastings step leads to the popular Metropolis-adjusted Langevin algorithm (MALA) [81] algorithm. Another example is Prox-MALA, a proximal variant thereof [78].

Remark 4.2 (Acceptance rate). *In order to obtain a method with reasonable convergence behavior in practice, the acceptance rate should be within a certain range. Acceptance rates close to zero indicate that the chain is barely moving which will yield slow convergence. On the other hand, acceptance rates close to one are often a consequence of the proposed step being close to zero, again, yielding a barely moving chain. For high dimensional problems, acceptance rates of approximately 1/4 are advised in the literature [80, Section 7.8.4]. Achieving such a rate in general requires to perform multiple runs of the algorithm.*

4.4 Gibbs sampling

Assume we aim to sample from a distribution $\pi_{X,V}$ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ with density $p_{X,V}$.⁶ The two-block Gibbs sampler (algorithm 4) is as simple as alternately sampling from the conditionals $\pi_X(\cdot | V = \cdot)$ and $\pi_V(\cdot | X = \cdot)$. The scheme can easily be extended to a multi-block samplers (see, e.g., [80, Chapter 10]), but we restrict our discussion to the two-block case for simplicity. As Gibbs sampling requires us to sample from the conditionals $\pi_X(\cdot | V = \cdot)$ and $\pi_V(\cdot | X = \cdot)$ the method may be considered as a *meta* algorithm shifting the issue from directly sampling from $\pi_{X,V}$ to sampling from the conditionals. Significant improvements in computational complexity are primarily achieved in cases where sampling from the conditionals is easy—or even possible directly—whereas sampling from the joint distribution is hard and/or only feasible iteratively.

Let us denote the transition kernel of the joint chain $(X_k, V_k)_k$ as R and the corresponding kernels of first and second variables as R_X and R_V , respectively.

⁶That is, Gibbs sampling requires a joint distribution of two random variables. This can be achieved by splitting the variable of interest into two blocks or via latent variable models, see Paragraph 4.4.

Algorithm 4 The Gibbs sampling algorithm for two variable blocks.

Require: Initial values (X_0, V_0) .

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: $V^{k+1} \sim \pi_V(\cdot | X = X^k)$
 - 3: $X^{k+1} \sim \pi_X(\cdot | V = V^{k+1})$
 - 4: **end for**
-

Lemma 4.1. *The distribution $\pi_{X,V}$ with Lebesgue-density $p_{X,V}$ is invariant for the kernel R and the marginal distributions π_X and π_V with Lebesgue-densities p_X and p_V , respectively, are invariant for R_X and R_V , respectively.*

Proof. Let $A \subset \mathbb{R}^{d_1}$ and $B \subset \mathbb{R}^{d_2}$ be arbitrary measurable sets. A repeated application of Bayes' theorem shows that

$$\begin{aligned}
& \pi R(A \times B) \\
&= \iiint \mathbb{1}_A(\tilde{x}) \mathbb{1}_B(\tilde{v}) p_V(\tilde{v} | X = x) p_X(\tilde{x} | V = \tilde{v}) p_{X,V}(x, v) d\tilde{x} d\tilde{v} dx dv \\
&= \iiint \mathbb{1}_A(\tilde{x}) \mathbb{1}_B(\tilde{v}) p_X(\tilde{x} | V = \tilde{v}) p_V(\tilde{v} | X = x) p_X(x) d\tilde{x} d\tilde{v} dx \\
&= \iint \mathbb{1}_A(\tilde{x}) \mathbb{1}_B(\tilde{v}) p_X(\tilde{x} | V = \tilde{v}) p_V(\tilde{v}) d\tilde{x} d\tilde{v} \\
&= \iint \mathbb{1}_A(\tilde{x}) \mathbb{1}_B(\tilde{v}) p_{X,V}(\tilde{x}, \tilde{v}) d\tilde{x} d\tilde{v} \\
&= \pi(A \times B).
\end{aligned} \tag{50}$$

The proofs for the conditionals are analogous. □

Ergodicity of the Gibbs sampler can be ensured, *e.g.*, using the following positivity condition.

Definition 4.2 (Positivity condition). *The density $p_{X,V}$ satisfies the positivity condition if for all x, v it holds that*

$$[p_X(x) > 0 \text{ and } p_V(v) > 0] \implies p_{X,V}(x, v) > 0. \tag{51}$$

That is, the support of the joint density is the Cartesian product of the supports of its marginals.

Theorem 4.5. *Assume that the transition kernel R is absolutely continuous with respect to the Lebesgue measure and that $p_{X,V}$ satisfies the positivity condition (51). Then the Gibbs sampler is ergodic, *i.e.*, for every initial distribution μ ,*

$$\lim_{n \rightarrow \infty} d_{\text{TV}}(\mu R^n, \pi_{X,V}) = 0. \tag{52}$$

Proof. A proof can be found in [80, Theorems 9.6, 10.10]. □

It turns out that the Gibbs sampler has a strong connection to the MH algorithm.

Theorem 4.6. *The subchains $(X_k)_k$ and $(V_k)_k$ of the Gibbs sampler constitute MH algorithms with acceptance probability equal to one.*

Proof. Due to symmetry it is sufficient to prove the result for the X_k chain. The corresponding proposal density of the chain reads as

$$q(x, \tilde{x}) = \int p_V(\tilde{v} | X = x) p_X(\tilde{x} | V = \tilde{v}) d\tilde{v}.$$

Using the definition of the conditional distribution, elementary computations yield

$$\begin{aligned}
p_X(x)q(x, \tilde{x}) &= \int p_X(x)p_V(\tilde{v}|X=x)p_X(\tilde{x}|V=\tilde{v})d\tilde{v} \\
&= \int p_{X,V}(x, \tilde{v})p_X(\tilde{x}|V=\tilde{v})d\tilde{v} \\
&= \int p_X(x|V=\tilde{v})p_V(\tilde{v})p_X(\tilde{x}|V=\tilde{v})d\tilde{v} \\
&= \int p_X(x|V=\tilde{v})p_X(\tilde{x})p_V(\tilde{v}|X=\tilde{x})d\tilde{v} \\
&= p_X(\tilde{x})q(\tilde{x}, x),
\end{aligned} \tag{53}$$

which shows that the corresponding MH acceptance probability

$$\min \left\{ \frac{p_X(\tilde{x})q(\tilde{x}, x)}{p_X(x)q(x, \tilde{x})}, 1 \right\}$$

is always equal to one. □

Application to latent variable models Gibbs sampling as introduced above builds on a joint distribution $\pi_{X,V}$ of two random variables X and V . Using latent variable models, however, it is easy to see that the approach may also be valuable when working with a distribution of only *one* random variable π_X [56].

Definition 4.3 (Latent variable model). *Let $(\mathcal{X}, \Sigma_{\mathcal{X}})$ and $(\mathcal{V}, \Sigma_{\mathcal{V}})$ be measurable spaces and let π_X be a probability distribution on \mathcal{X} . We call the distribution $\pi_{X,V}$ on $\mathcal{X} \times \mathcal{V}$ a latent variable model for π_X if $\pi_{X,V}$ admits π_X as its marginal, i.e., for any measurable $A \in \Sigma_{\mathcal{X}}$ it holds true that*

$$\pi_X(A) = \pi_{X,V}(A \times \mathcal{V}).$$

In the case $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{V} = \mathbb{R}^\ell$ and all involved distributions admitting densities with respect to the Lebesgue measure, this means that these densities satisfy

$$p_X(x) = \int p_{X,V}(x, v)dv.$$

Given such a latent variable model, sampling from π_X can be simply realized by sampling from $\pi_{X,V}$ and subsequently dropping the latent variable. In [56, Table 1] the authors provide a list of densities that admit specifically favorable latent variable models, where the conditional distribution $\pi_{X|V}$ is Gaussian with mean $\mu(v)$ and covariance $C(v)$ being functions of v so that

$$\begin{aligned}
p_X(x) &= \int p_X(x|V=v)p_V(v)dv \\
&= \int \frac{1}{\sqrt{(2\pi)^d \det(C(v))}} \exp\left(-\|x - \mu(v)\|_{C(v)^{-1}}^2\right) p_Z(v)dv
\end{aligned} \tag{54}$$

is, in fact, a Gaussian mixture. In this case, for some given v , sampling from $\pi_X(\cdot | V=v)$ reduces the sampling from a multivariate Gaussian which is possible via standard techniques such as the Cholesky decomposition of the corresponding covariance matrix, or alternative methods that might be more efficient in high dimensions, such as Perturb-and-MAP (see [56, Section 3.3.1] for details). Also sampling from $p_{V|X}$ is computationally cheap under certain assumptions on

the structure of π_X [56, Section 3.3.2]. As shown in [56], Gibbs sampling has the advantage of providing chains with rapidly decaying autocorrelation. This comes at the cost of computationally more demanding iterations. The results in [56] still show an extremely strong preference for the Gibbs sampler.

4.5 Langevin sampling

For the high-dimensional priors and posteriors encountered in imaging problems, the most widely used class of algorithms for sampling from EBMs are methods based on variants of the Langevin diffusion process. The reason for the popularity of Langevin based sampling is the simple implementation as well as high flexibility of the corresponding algorithms. These methods constitute discretizations of continuous time stochastic processes $(Y_t)_{t \geq 0}$ for which it is known that $\text{Law}(Y_t) \rightarrow \pi$ as $t \rightarrow \infty$. Ensuring that the discretization error grows sufficiently slowly allows to balance the convergence of the continuous time process to the target density with the error of the discretization in order to obtain approximate samples of the target. In the following we will present the *overdamped* and the *underdamped* Langevin algorithms as two specific instances.

While several results can be proven in significantly more general settings, for this section we assume always the following in order to provide rather self contained results.

Assumption 4.1. *We make the following assumptions on the energy E :*

1. E is continuously differentiable and ∇E is L -Lipschitz continuous.
2. E is m -strongly convex, i.e., for any $x, y \in \mathbb{R}^d$ and $\lambda \in (0, 1)$ we have that $E(\lambda x + (1-\lambda)\tilde{x}) \leq \lambda E(x) + (1-\lambda)E(\tilde{x}) - \frac{m\lambda(1-\lambda)}{2} \|x - \tilde{x}\|^2$.

Moreover, we define the condition number as $\kappa = \frac{L}{m}$. Convergence results of various discretizations of the Langevin diffusion in the non-convex case are provided in, e.g., [30, 79, 41]. Non-differentiable energies E and the extensions of the convergence results to such cases are treated in, e.g., [29, 78, 37, 40, 41, 33, 79, 41].

The handling of continuous-time stochastic processes, moreover requires some additional terminology. In duality to (44), we can define the action of a Markov kernel $R : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1]$ on a bounded and measurable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ as

$$Rf(x) = \int f(\tilde{x})R(x, d\tilde{x}), \quad x \in \mathbb{R}^d. \quad (55)$$

We call a family of Markov kernels $(P_t)_{t \geq 0}$, $P_t : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1]$ a *Markov semi-group* if—as an operator according to (55)— P_0 is the identity and $(P_t)_t$ is a semi-group, i.e., $P_{t+s} = P_t \circ P_s$ for any $s, t \geq 0$.⁷ The notion of a stationary measure allows for a straightforward adaptation to the continuous-time setting.

Definition 4.4 (Stationary distribution, continuous time). *Let $(P_t)_t$ be a Markov semi-group on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We call $\mu \in \mathcal{P}(\mathbb{R}^d)$ stationary for $(P_t)_t$ if for any $t > 0$ and $A \in \mathcal{B}(\mathbb{R}^d)$, $\mu P_t(A) = \mu(A)$.*

⁷The term *Markov semi-group* typically entails additional properties which are, however, not relevant for our purposes, see [5] for more details.

Algorithm 5 The unadjusted Langevin algorithm.

Require: Initial value X_0 , step size $\tau > 0$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: $Z_k \sim \mathcal{N}(0, \mathbf{I})$
 - 3: $X_{k+1}^\tau = X_k^\tau - \tau \nabla E(X_k^\tau) + \sqrt{2\tau} Z_k$
 - 4: **end for**
-

4.5.1 Overdamped Langevin sampling

The first approach to model $(X_k)_k$ is as a discretization of the so-called overdamped Langevin diffusion process which is defined via the SDE

$$dX_t = -\nabla E(X_t)dt + \sqrt{2}dW_t \quad (56)$$

where $(W_t)_{t \geq 0}$ denotes Brownian motion. The simplest way to obtain a Markov chain that approximates the continuous-time diffusion process is by employing a first-order Euler-Maruyama (EM) discretization which results in the update rule

$$X_{k+1}^\tau = X_k^\tau - \tau \nabla E(X_k^\tau) + \sqrt{2\tau} Z_k \quad (57)$$

for $k = 0, 1, 2, \dots$ and some suitable X_0 , where $(Z_k)_k$ are i.i.d. standard normal distributed random vectors and $\tau > 0$ is the step size of the discretization. The step size in the superscript emphasizes that the distribution of the chain as well as its stationary distribution depend on the step size. The resulting algorithm which is summarized in algorithm 5 is commonly referred to as the unadjusted Langevin algorithm (ULA).⁸

We will denote the semi-group that generates (56) as $(P_t)_{t \geq 0}$ and the Markov kernel that represents one step in algorithm 5 as R_τ . That is, for any $x \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$,

$$R_\tau(x, A) = \frac{1}{(4\pi\tau)^{d/2}} \int_A \exp\left(-\frac{1}{2} \frac{\|z - (x - \tau \nabla E(x))\|^2}{2\tau}\right) dz.$$

Therefore, if we denote the distribution of X_k^τ for some $k = 0, 1, \dots$ in algorithm 5 as μ_k , then $\mu_{k+1} = \mu_k R_\tau$ and $\mu_k = \mu_0 R_\tau^k$ with R_τ^k the k -fold composition of R_τ . We begin this section with a standard result about the existence of solutions of the Langevin diffusion SDE for all time.

Theorem 4.7. *Let $Z \sim \mu \in \mathcal{P}_2(\mathbb{R}^d)$ be independent of $(W_t)_{t \geq 0}$. Then the SDE (56) with initial condition $X_0 = Z$ admits a unique strong solution $(X_t)_{t \geq 0}$. Moreover, the solution satisfies $\int_0^t \mathbb{E}[\|X_s\|^2] ds < \infty$ for any $t \geq 0$.*

Proof. This is a standard result under the assumptions in this sections. A proof that utilizes a fixed point argument is provided in [76, Theorem 5.2.1]. \square

In the following proposition, we use Ito's formula to derive a weak formulation that describes the distribution of $(X_t)_t$.

Proposition 4.1. *Let $t, h > 0$ and $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be twice continuously differentiable and such that $\int_t^{t+h} \mathbb{E}[\|\nabla \phi(X_s)\|^2 | \mathcal{F}_t] ds < \infty$, where $(\mathcal{F}_t)_{t \geq 0}$ denotes the filtration to which $(W_t)_t$ is adapted. Then the solution $(X_t)_t$ of (56) satisfies*

$$\mathbb{E}[\phi(X_{t+h})] - \mathbb{E}[\phi(X_t)] = \int_t^{t+h} \mathbb{E}[-\langle \nabla E(X_s), \nabla \phi(X_s) \rangle + \Delta \phi(X_s)] ds. \quad (58)$$

⁸The word *unadjusted* emphasizes that, due to discretization errors, ULA only provides biased samples which is not *adjusted* within the algorithm.

That is, if $X_0 \sim \mu$, then

$$\begin{aligned} & \int_{\mathbb{R}^d} \phi(x) (\mu P_{t+h} - \mu P_t) (dx) \\ &= \int_t^{t+h} \int_{\mathbb{R}^d} -\langle \nabla E(x), \nabla \phi(x) \rangle + \Delta \phi(x) \mu P_s(dx) ds. \end{aligned} \quad (59)$$

Proof. The proof is a simple consequence of Ito's lemma [76, Theorem 4.2.1]. Let $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be twice continuously differentiable. Then, by Ito's lemma the process $\phi(X_t)$ satisfies

$$\begin{aligned} & \phi(X_{t+h}) - \phi(X_t) \\ &= \int_t^{t+h} -\langle \nabla \phi(X_s), \nabla E(X_s) \rangle + \Delta \phi(X_s) ds + \sqrt{2} \int_t^{t+h} \nabla \phi^T(X_s) dW_s. \end{aligned} \quad (60)$$

The expected value of the last integral with respect to Brownian motion in (60) is zero (cf. [53, Section 3.3], [38, Chapter 8, Section 2]) and, consequently, the desired result follows. \square

Remark 4.3. If X_t admits a smooth density with respect to the Lebesgue measure—denoted as $p(x, t)$ —the above implies that this density satisfies the Fokker-Planck equation

$$\partial_t p(x, t) = \operatorname{div}(\nabla E(x)p(x, t)) + \Delta p(x, t). \quad (61)$$

Using the above result we can derive a distributional partial differential equation characterizing the stationary measure of the Langevin diffusion.

Corollary 4.1. Assume that the SDE in (56) admits a stationary measure μ . Then, this measure satisfies that

$$\int_{\mathbb{R}^d} -\langle \nabla E(x), \nabla \phi(x) \rangle + \Delta \phi(x) d\mu(x) = 0 \quad (62)$$

for any $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ as in proposition 4.1 and such that $\nabla \phi \in L^2(\mathbb{R}^d, \mu)$.

We can now prove the main result about the ergodicity of the continuous time process to the target distribution π with respect to the Wasserstein-2 distance which we first formally define.

Definition 4.5 (Wasserstein-2 distance). Let μ, ν be two probability measures on \mathbb{R}^d . A coupling γ is a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ such that for any $A \in \mathcal{B}(\mathbb{R}^d)$, $\gamma(\mathbb{R}^d \times A) = \nu(A)$ and $\gamma(A \times \mathbb{R}^d) = \mu(A)$. We denote the set of all couplings of μ and ν as $\Pi(\mu, \nu)$. For two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ we define the Wasserstein-2 distance between μ and ν as

$$d_W(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu, \nu)} \int \|x - \tilde{x}\|^2 d\gamma(x, \tilde{x}) \right)^{\frac{1}{2}}. \quad (63)$$

Note that for any two $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ there exists a coupling $\hat{\gamma} \in \Pi(\mu, \nu)$ realizing the infimum [95] and we refer to it as an optimal coupling. In a slight abuse of terminology we will also refer to two random variables $X \sim \mu$ and $\tilde{X} \sim \nu$ such that $(X, \tilde{X}) \sim \hat{\gamma}$ as an optimal coupling. In this case $d_W(\mu, \nu)^2 = \mathbb{E}[\|X - \tilde{X}\|^2]$.

Theorem 4.8. The measure π is the unique invariant probability measure for P_t . That is, $\pi P_t = \pi$ for all $t \geq 0$. Moreover, for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ it holds that

$$d_W(\mu P_t, \pi) \leq \exp(-mt) d_W(\mu, \pi).$$

Proof. The proof consists of two steps: First we show that the Langevin diffusion induces a contraction with respect to the Wasserstein-2 distance, *i.e.*, that $d_W(\mu P_t, \nu P_t)^2 \leq d_W(\mu, \nu)^2 \exp(-2mt)$. This implies the existence of and convergence to a unique fixed point. Afterwards we identify this fixed point as the target density π .

Let X_t and \tilde{X}_t be two coupled processes, that is, both are solutions of the SDE (56) with the same Brownian motion $(W_t)_t$. Let $X_0 \sim \mu$ and $\tilde{X}_0 \sim \nu$. Then, the process $\mathbf{X}_t = (X_t, \tilde{X}_t)_t$ satisfies the SDE

$$d\mathbf{X}_t = \begin{bmatrix} -\nabla E(X_t) \\ -\nabla E(\tilde{X}_t) \end{bmatrix} dt + \sqrt{2} \begin{bmatrix} I \\ I \end{bmatrix} dW_t. \quad (64)$$

Let us define $\phi(\mathbf{x}) = \phi(x, \tilde{x}) = \frac{1}{2}\|x - \tilde{x}\|^2$. Then using Ito's lemma [53, Theorem 3.3], we find that ϕ satisfies the ordinary differential equation (ODE)⁹

$$d\phi(\mathbf{X}_t) = -\langle X_t - \tilde{X}_t, \nabla E(X_t) - \nabla E(\tilde{X}_t) \rangle dt.$$

Integrating over $(t, t+h)$ for some $h > 0$ and taking the expectation leads to

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{2} \|X_{t+h} - \tilde{X}_{t+h}\|^2 \right] - \mathbb{E} \left[\frac{1}{2} \|X_t - \tilde{X}_t\|^2 \right] \\ &= \mathbb{E} \left[- \int_t^{t+h} \langle X_s - \tilde{X}_s, \nabla E(X_s) - \nabla E(\tilde{X}_s) \rangle ds \right] \\ &\leq -m \int_t^{t+h} \mathbb{E} \left[\|X_s - \tilde{X}_s\|^2 \right] ds. \end{aligned} \quad (65)$$

Defining $\psi(t) = \mathbb{E} \left[\|X_t - \tilde{X}_t\|^2 \right]$ we can rewrite the above equation more compactly as $\psi(t+h) - \psi(t) \leq -2m \int_t^{t+h} \psi(s) ds$. Since

$$t \mapsto \mathbb{E} \left[\langle X_t - \tilde{X}_t, \nabla E(X_t) - \nabla E(\tilde{X}_t) \rangle \right]$$

is continuous, Lebesgue's dominated convergence theorem implies the differentiability of ψ due to the first equality in (65). By dividing by h and letting $h \rightarrow 0$ from above in (65) it follows that

$$\psi'(t) \leq -2m\psi(t).$$

Grönwall's inequality then implies that

$$d_W(\mu P_t, \nu P_t)^2 \leq \mathbb{E} \left[\|X_t - \tilde{X}_t\|^2 \right] \leq \mathbb{E} \left[\|X_0 - \tilde{X}_0\|^2 \right] \exp(-2mt).$$

Plugging in an optimal coupling (X_0, \tilde{X}_0) for μ, ν on the right-hand side shows that the Langevin diffusion provides a contraction on $\mathcal{P}(\mathbb{R}^d)$ equipped with the Wasserstein-2 distance. A variation of Banach's fixed point theorem [37, Theorem 2] then readily shows that (56) admits a unique stationary measure π_∞ and any solution converges to this measure at the rate

$$d_W(\mu P_t, \pi_\infty)^2 \leq d_W(\mu, \pi_\infty)^2 \exp(-2mt).$$

Lastly, we have to identify the stationary distribution as the target measure π . To this end first note that the density of π_∞ has to be a solution of (62) for any ϕ . However, also the target π satisfies (62) for any ϕ . As shown in [12, Example 5.1], [37, Theorem 4.10] (62) admits at most one solution in the space of Borel probability measures. Thus, $\pi_\infty = \pi$. \square

⁹The stochastic integral vanishes since the Brownian motion cancels.

Similarly, we can show that the MC defined by ULA is ergodic.

Theorem 4.9. *For each $\tau < \frac{2m}{L^2}$ the ULA chain is geometrically ergodic, that is, there exists a unique stationary distribution $\pi^\tau \in \mathcal{P}_2(\mathbb{R}^d)$ such that for any initial distribution $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ it holds that*

$$d_W(\mu R_\tau^k, \pi^\tau)^2 \leq (1 - 2m\tau + \tau^2 L^2)^k d_W(\mu, \pi^\tau)^2.$$

Proof. Let $X \sim \mu$ and $\tilde{X} \sim \nu$ and define $X^+ = X - \tau \nabla E(X) + \sqrt{2\tau}Z$ and $\tilde{X}^+ = \tilde{X} - \tau \nabla E(\tilde{X}) + \sqrt{2\tau}Z$ where $Z \sim \mathcal{N}(0, I)$. We can compute that

$$\begin{aligned} \|X^+ - \tilde{X}^+\|^2 &= \|X - \tau \nabla E(X) - \tilde{X} + \tau \nabla E(\tilde{X})\|^2 \\ &= \|X - \tilde{X}\|^2 - 2\tau \langle X - \tilde{X}, \nabla E(X) - \nabla E(\tilde{X}) \rangle \\ &\quad + \tau^2 \|\nabla E(X) - \nabla E(\tilde{X})\|^2 \\ &\leq \|X - \tilde{X}\|^2 (1 - 2m\tau + \tau^2 L^2). \end{aligned} \tag{66}$$

For $\tau < \frac{2m}{L^2}$, $1 - 2m\tau + \tau^2 L^2 < 1$ and, thus, taking the expectation above and minimizing over all couplings (X, \tilde{X}) shows that R_τ is a contraction on $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the Wasserstein-2 distance. By Banach's fixed point theorem, there exists a unique fixed point π^τ which is the unique stationary distribution of the chain and it follows for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ that

$$d_W(\mu R_\tau^k, \pi^\tau)^2 \leq (1 - 2m\tau + \tau^2 L^2)^k d_W(\mu, \pi^\tau)^2. \quad \square$$

We now show that the distribution π^τ that is stationary with respect to the ULA kernel indeed approximates the target π . To do so, we estimate the discretization error introduced by the Euler-Maruyama discretization. This error bound will rely on a uniform bound of the second moments of the iterates of ULA and we first prove the following auxiliary result about these second moments.

Lemma 4.2. *The second moments of the chains $(X_k^\tau)_k$ are bounded uniformly in $\tau \in [0, \bar{\tau}]$ for $\bar{\tau} < \frac{2m}{L^2}$ as long as the initial distribution of the chain has bounded second moment. More precisely, if x^* is the minimizer of E it holds that*

$$\sup_{\tau \in [0, \bar{\tau}]} \sup_{k \in \mathbb{N}} \mathbb{E} [\|X_k^\tau - x^*\|^2] \leq \mathbb{E} [\|X_0^\tau - x^*\|^2] + \frac{1}{2m - L\bar{\tau}} < \infty. \tag{67}$$

Proof. The boundedness of the second moments for some fixed τ trivially follows from the convergence of the chain with respect to the Wasserstein-2 distance. The challenge is to derive a bound uniformly in τ . A proof of this fact can be found in [37, Lemma 5.6], but we provide it here for completeness. Let $x^* \in \mathbb{R}^d$ be the unique minimizer of E , which exists since E is strongly convex. Since $\nabla E(x^*) = 0$ the strong convexity of E and the Lipschitz continuity of ∇E imply that for any $\tau < \bar{\tau}$ it holds that

$$\begin{aligned} \|X_k^\tau - \tau \nabla E(X_k^\tau) - x^*\|^2 &= \|X_k^\tau - x^*\|^2 - 2\tau \langle X_k^\tau - x^*, \nabla E(X_k^\tau) \rangle + \tau^2 \|\nabla E(X_k^\tau)\|^2 \\ &= \|X_k^\tau - x^*\|^2 - 2\tau \langle X_k^\tau - x^*, \nabla E(X_k^\tau) - \nabla E(x^*) \rangle \\ &\quad + \tau^2 \|\nabla E(X_k^\tau) - \nabla E(x^*)\|^2 \\ &\leq \|X_k^\tau - x^*\|^2 (1 - 2m\tau + L\tau^2) \end{aligned} \tag{68}$$

where $(1 - 2m\tau + L\tau^2) := \rho < 1$ by definition of $\bar{\tau}$. It follows that

$$\begin{aligned} \|X_{k+1}^\tau - x^*\|^2 &= \|X_k^\tau - \tau \nabla E(X_k^\tau) - x^*\|^2 \\ &\quad + 2\langle X_k^\tau - \tau \nabla E(X_k^\tau) - x^*, \sqrt{2\tau}Z_k \rangle + 2\tau \|Z_k\|^2 \\ &\leq \|X_k^\tau - x^*\|^2 \rho + 2\langle X_k^\tau - \tau \nabla E(X_k^\tau) - x^*, \sqrt{2\tau}Z_k \rangle + 2\tau \|Z_k\|^2. \end{aligned} \tag{69}$$

Taking the expectation on both sides and noting that Z_k and X_k^τ are independent it follows that

$$\mathbb{E} [\|X_{k+1}^\tau - x^*\|^2] \leq \mathbb{E} [\|X_k^\tau - x^*\|^2] \rho + 2\tau.$$

Solving this recursion leads to

$$\begin{aligned} \mathbb{E} [\|X_k^\tau - x^*\|^2] &\leq \rho^k \mathbb{E} [\|X_0^\tau - x^*\|^2] + 2\tau \sum_{i=0}^{k-1} \rho^i \\ &\leq \mathbb{E} [\|X_0^\tau - x^*\|^2] + \frac{2\tau}{1-\rho} \end{aligned} \tag{70}$$

where $0 \leq \frac{\tau}{1-\rho} = \frac{1}{2m-L\tau} \leq \frac{1}{2m-L\bar{\tau}}$ which is bounded by the constraint on $\bar{\tau}$. \square

We can now present the error bound between the stationary distribution of the continuous time Langevin diffusion (56) and the stationary distribution of its discretization ULA.

Theorem 4.10. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\bar{\tau} < \frac{2m}{L^2}$. Then there exists $c > 0$ such that the Wasserstein-2 error between the distribution of the continuous time diffusion (56) and that of the ULA with $\tau \leq \bar{\tau}$ satisfies for any $n \in \mathbb{N}$ that*

$$d_W(\mu P_{n\tau}, \mu R_\tau^n)^2 \leq cn\tau^2. \tag{71}$$

Proof. Note that one iteration of ULA can be interpolated (in distribution) as $X_k^\tau = \bar{X}_{k\tau}^\tau$ where the process $(\bar{X}_t^\tau)_t$ is defined via

$$d\bar{X}_t^\tau = -\nabla E(\bar{X}_{k\tau}^\tau) dt + \sqrt{2} dW_t, \quad t \in (k\tau, (k+1)\tau)$$

with a constant drift term. As before, we consider the coupled SDE

$$d\mathbf{X}_t = \begin{bmatrix} dX_t \\ d\bar{X}_t^\tau \end{bmatrix} = \begin{bmatrix} -\nabla E(X_t) \\ -\nabla E(\bar{X}_{k\tau}^\tau) \end{bmatrix} dt + \sqrt{2} \begin{bmatrix} I \\ I \end{bmatrix} dW_t, \quad t \in [k\tau, (k+1)\tau),$$

and, again via Ito's lemma, we find that it satisfies the ODE $d\phi(\mathbf{X}_t) = -\langle \nabla E(X_t) - \nabla E(\bar{X}_{k\tau}^\tau), X_t - \bar{X}_t^\tau \rangle dt$ for $\phi(Z_t) = \frac{1}{2} \|X_t - \bar{X}_t^\tau\|^2$. Therefore,

$$\begin{aligned} &\frac{1}{2} \|X_{(k+1)\tau} - X_{k+1}^\tau\|^2 - \frac{1}{2} \|X_{k\tau} - X_k^\tau\|^2 \\ &= - \int_{k\tau}^{(k+1)\tau} \langle \nabla E(X_t) - \nabla E(\bar{X}_{k\tau}^\tau), X_t - \bar{X}_t^\tau \rangle dt \\ &= - \int_{k\tau}^{(k+1)\tau} \langle \nabla E(X_t) - \nabla E(\bar{X}_t^\tau), X_t - \bar{X}_t^\tau \rangle dt \\ &\quad - \int_{k\tau}^{(k+1)\tau} \langle \nabla E(\bar{X}_t^\tau) - \nabla E(\bar{X}_{k\tau}^\tau), X_t - \bar{X}_t^\tau \rangle dt \\ &\stackrel{*}{\leq} -m \int_{k\tau}^{(k+1)\tau} \|X_t - \bar{X}_t^\tau\|^2 dt \\ &\quad + \int_{k\tau}^{(k+1)\tau} L \left(\frac{\alpha}{2} \|\bar{X}_t^\tau - \bar{X}_{k\tau}^\tau\|^2 + \frac{1}{2\alpha} \|X_t - \bar{X}_t^\tau\|^2 \right) dt \\ &\stackrel{**}{\leq} \int_{k\tau}^{(k+1)\tau} \frac{L^2}{4m} \|\bar{X}_t^\tau - \bar{X}_{k\tau}^\tau\|^2 dt \\ &\leq \int_{k\tau}^{(k+1)\tau} \frac{L^2}{4m} \|-(t-k\tau)\nabla E(\bar{X}_{k\tau}^\tau) + \sqrt{2}(W_t - W_{k\tau})\|^2 dt \end{aligned} \tag{72}$$

where we used the elementary inequality $ab \leq \frac{1}{2\alpha}a^2 + \frac{\alpha}{2}b^2$ for any $\alpha > 0$, $a, b \in \mathbb{R}$ in the inequality marked with $*$ and $\alpha = \frac{L}{2m}$ in the inequality marked with $**$. Taking the expected value on both sides and noting that $\bar{X}_{k\tau}^\tau$ and $(W_t - W_{k\tau})$ are independent we obtain that

$$\begin{aligned} & \frac{1}{2}\mathbb{E}[\|X_{(k+1)\tau} - X_{k+1}^\tau\|^2] - \frac{1}{2}\mathbb{E}[\|X_{k\tau} - X_k^\tau\|^2] \\ & \leq \frac{L^2}{4m} \left(\frac{\tau^3}{3} \mathbb{E}[\|\nabla E(\bar{X}_{k\tau}^\tau)\|^2] + \tau^2 \right) \\ & \leq \frac{L^2}{4m} \left(\frac{\tau^3}{3} L^2 (\mathbb{E}[\|X_0^\tau - x^*\|^2] + \eta) + \tau^2 \right) \end{aligned} \quad (73)$$

where we applied lemma 4.2 after setting $\eta = \frac{1}{2m-L\bar{\tau}}$. Since τ is bounded from above, the cubic term τ^3 can be bounded by the quadratic and it follows that there exists $c > 0$ such that

$$\mathbb{E}[\|X_{(k+1)\tau} - X_{k+1}^\tau\|^2] - \mathbb{E}[\|X_{k\tau} - X_k^\tau\|^2] \leq c\tau^2. \quad (74)$$

Taking the expectation and optimizing over all couplings of $\mu P_{k\tau}$ and μR_τ^k leads to $d_W(\mu P_{(k+1)\tau}, \mu R_\tau^{k+1})^2 - d_W(\mu P_{k\tau}, \mu R_\tau^k)^2 \leq c\tau^2$. Finally, the desired result follows from summing up the inequality over k . \square

The combination of the exponential ergodicity of the continuous time process with the bound on the discretization error implies the following result on the approximation of the target π by ULA.

Theorem 4.11. *There exists a $\rho > 0$ such that the ULA with $\tau \leq \bar{\tau}$ satisfies for any $k, n \in \mathbb{N}$, $k \geq n$*

$$d_W(\delta_x R_\tau^k, \pi) \leq \sqrt{nc\tau^2} + \rho \exp(-mn\tau). \quad (75)$$

In particular, let $\epsilon > 0$ be arbitrary. Then it follows for $T = \frac{-\log(\frac{\epsilon}{2\rho})}{m} + 1$, $\tau < \min(1, \frac{\epsilon^2}{4cT})$, and $k \geq \lfloor \frac{T}{\tau} \rfloor$ that $d_W(\delta_x R_\tau^k, \pi) \leq \epsilon$ and, as a consequence, also $d_W(\pi^\tau, \pi) \leq \epsilon$.

Proof. By the triangle inequality and theorem 4.10 we find for any $p, k \in \mathbb{N}$, $p \leq k$ that

$$\begin{aligned} d_W(\delta_x R_\tau^k, \pi) & \leq d_W(\delta_x R_\tau^p R_\tau^{k-p}, \delta_x R_\tau^p P_{(k-p)\tau}) + d_W(\delta_x R_\tau^p P_{(k-p)\tau}, \pi) \\ & \leq \sqrt{(k-p)c\tau^2} + d_W(\delta_x R_\tau^p, \pi) \exp(-m(k-p)\tau). \end{aligned} \quad (76)$$

The boundedness of the second moments ensured by lemma 4.2 implies the existence of a constant $\rho > 0$ such that $d_W(\delta_x R_\tau^p, \pi) \leq \rho$ uniformly in τ and p . This concludes the proof of the first assertion. Now let $\epsilon > 0$ be arbitrary. Denote $n := k - p$ and fix $T = \frac{-\log(\frac{\epsilon}{2\rho})}{m} + 1$ so that $\rho \exp(-m(T-1)) = \frac{\epsilon}{2}$. Choose a step size $\tau < \min(1, \frac{\epsilon^2}{4cT})$. Then choose $n = \lfloor \frac{T}{\tau} \rfloor$ so that $n\tau \geq \tau(\frac{T}{\tau} - 1) = T - \tau \geq T - 1$. It follows for any $k \geq n$

$$\begin{aligned} d_W(\delta_x R_\tau^k, \pi) & \leq \sqrt{nc\tau^2} + \rho \exp(-mn\tau) \\ & \leq \sqrt{cT\tau} + \rho \exp(-m(T-1)) \\ & \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned} \quad (77)$$

\square

Remark 4.4. *In particular, we obtain the following convergence rates: In order to obtain accuracy ϵ in Wasserstein-2 distance, we need to run the chain for $\mathcal{O}(\log(\epsilon)^2/\epsilon^2)$ iterations.*

Remark 4.5. Note that ULA differs conceptually from MH and Gibbs sampling in so far, as the generated Markov chain does not admit the target π as its invariant measure, but only an approximation thereof, that is, we obtain biased samples. The same is true for the underdamped Langevin algorithm presented in the next section. The bias may be mitigated, e.g., by using a vanishing step-size which, however, leads to a time-inhomogeneous MC [29, 40, 33], for which concepts like stationary distributions are not directly applicable. Another approach to obtain unbiased samples is to correct ULA to target π directly by adding a MH accept/reject step which leads to MALA [81].

4.5.2 Underdamped Langevin sampling

An alternative to the overdamped Langevin diffusion that can lead to improved convergence rates is given by the underdamped Langevin diffusion which is defined by the SDE

$$\begin{cases} dX_t = V_t dt, \\ dV_t = (-\alpha V_t - \beta \nabla E(X_t)) dt + \sqrt{2\alpha\beta} dW_t, \end{cases} \quad (78)$$

which, has a physical interpretation of modeling friction in addition to the potential force given by ∇E . The influence of this friction is tuned by the parameter $\alpha > 0$. In addition to this physical interpretation, the system also has tight links to optimization since it corresponds to the stochastic version of the second order ODE for Nesterov's accelerated gradient descent algorithm, see [92].

The process induced by (78) admits the stationary distribution $\pi_{X,V}$ with Lebesgue density $p_{X,V}(x, v) \propto \exp\left(-E(x) - \frac{\|v\|^2}{2\beta}\right)$ [24]. The density $p_{X,V}$ indeed solves the corresponding Fokker-Planck equation

$$0 = -\nabla_x p_{X,V} \cdot v + \alpha \nabla_v p_{X,V} \cdot v + \beta \nabla_v p_{X,V} \cdot \nabla E + \alpha dp + \beta \alpha \Delta_v p_{X,V} \quad (79)$$

which can be derived using Ito's lemma in a very similar fashion to the overdamped setting.

The structure of the Langevin diffusion given in (78) lends itself to a partial discretization that yields improved convergence rates compared to ULA [24]. The scheme is obtained via

$$\begin{cases} d\bar{X}_t = \bar{V}_t dt \\ d\bar{V}_t = (-\alpha \bar{V}_t - \beta \nabla E(\bar{X}_{k\tau}^t)) dt + \sqrt{2\alpha\beta} dW_t \end{cases} \quad (80)$$

for $t \in (k\tau, (k+1)\tau]$, where $\tau > 0$ denotes the discretization step size [24]. That is, we only fix the argument of ∇E , but otherwise solve the SDE exactly. Fortunately, the solution of (80) is known in distribution. In particular, the process V_t is an Ornstein-Uhlenbeck process with exact solution

$$\begin{aligned} \bar{V}_{k\tau+h} &= \bar{V}_{k\tau} \exp(-\alpha h) - \frac{\beta}{\alpha} \nabla E(\bar{X}_{k\tau}^\tau) (1 - \exp(-\alpha h)) \\ &\quad + \sqrt{2\alpha\beta} \int_{k\tau}^{k\tau+h} \exp(-\alpha(k\tau+h-s)) dW_s. \end{aligned} \quad (81)$$

Consequently, the distribution of the random vector $(\bar{X}_{k\tau+h}, \bar{V}_{k\tau+h})$ given some $(\bar{X}_{k\tau}, \bar{V}_{k\tau})$ is a Gaussian with mean $\mu_h(\bar{X}_{k\tau+h}, \bar{V}_{k\tau+h})$ where

$$\mu_h(x, v) = \begin{bmatrix} x + v \frac{1 - \exp(-\alpha h)}{\alpha} - \frac{\beta}{\alpha} \nabla E(x) \left(h - \frac{1 - \exp(-\alpha h)}{\alpha} \right) \\ v \exp(-\alpha h) - \frac{\beta}{\alpha} \nabla E(x) (1 - \exp(-\alpha h)) \end{bmatrix} \quad (82)$$

Algorithm 6 The underdamped Langevin algorithm.

Require: Initial values $X_0, V_0, \alpha, \beta > 0$, step size $\tau > 0$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: Compute $\mu_\tau(X_k^\tau, V_k^\tau)$ and C_τ according to (82), (83)
 - 3: $(X_{k+1}^\tau, V_{k+1}^\tau) \sim \mathcal{N}(\mu_\tau(X_k^\tau, V_k^\tau), C_\tau)$
 - 4: **end for**
-

and covariance

$$C_h = \begin{bmatrix} \frac{2\beta}{\alpha} \left(h - \frac{2(1-\exp(-\alpha h))}{\alpha} + \frac{1-\exp(-2\alpha h)}{2\alpha} \right) \mathbf{I} & \frac{\beta}{\alpha} (1 - 2\exp(-\alpha h) + \exp(-2\alpha h)) \mathbf{I} \\ \frac{\beta}{\alpha} (1 - 2\exp(-\alpha h) + \exp(-2\alpha h)) \mathbf{I} & \beta(1 - \exp(-2\alpha h)) \mathbf{I} \end{bmatrix}. \quad (83)$$

The resulting algorithm is summarized in algorithm 6. The exponential ergodicity of the continuous-time SDE to the target distribution $\pi_{X,V}$ as well as the approximation of $\pi_{X,V}$ by the discretization with sufficiently small step size can be obtained by similar techniques as those used in the overdamped case.

Theorem 4.12. [24, Theorem 1 and 5, Lemma 8] *Assume E is twice continuously differentiable, strongly convex with parameter m , and admits an L -Lipschitz continuous gradient and denote $\kappa = \frac{L}{m}$. Choose $\alpha = 2$ and $\beta = \frac{1}{L}$. Then the continuous-time underdamped Langevin dynamics satisfy that*

$$d_W(\mu P_t, \pi_{X,V}) \leq 4 \exp\left(-\frac{t}{2\kappa}\right) d_W(\mu, \pi_{X,V}). \quad (84)$$

Moreover, for any initial distribution $\mu \in \mathcal{P}_2(\mathbb{R}^{2d})$ there exists $C > 0$ such that

$$d_W(\mu R_\tau^k, \pi_{X,V}) \leq 4 \exp\left(-\frac{k\tau}{2\kappa}\right) d_W(\mu, \pi_{X,V}) + \frac{\tau^2}{1 - \exp\left(-\frac{\tau}{2\kappa}\right)} C. \quad (85)$$

In particular, for any $\epsilon > 0$ we can choose $\tau = \mathcal{O}(\epsilon)$ and $K = \mathcal{O}(\tau^{-1} \log(\epsilon^{-1}))$ such that for $k \geq K$, $d_W(\mu R_\tau^k, \pi_{X,V}) \leq \epsilon$.

Remark 4.6. *The underdamped Langevin algorithm improves the complexity from $\mathcal{O}(\epsilon^{-2} \log(\epsilon^{-1})^2)$ to $\mathcal{O}(\epsilon^{-1} \log(\epsilon^{-1}))$ iterations to reach accuracy ϵ in Wasserstein-2 in comparison the overdamped case.*

Remark 4.7 (Step size choice). *While the convergence theory for Langevin based sampling typically provides step size constraints depending on the Lipschitz constant of ∇E , this constant might sometimes be hard to compute in practice. Moreover, while these constraints ensure ergodicity, the remaining bias of the invariant measure of the chain might still be too large, thus, necessitating smaller step sizes. Empirically, step sizes in the range of $1e-5$ up to $1e-2$ provide reasonable results. However, we advise performing multiple runs with different step sizes in order to balance convergence speed and remaining bias. In particular, for smaller step sizes, the convergence speed becomes prohibitively slow, thus, alleviating any potential reduction in bias.*

4.6 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) sampling [72, 7] is a specific type of MH algorithm which aims at providing a Markov chain with faster mixing, *i.e.*, faster convergence to the target. Similarly to the underdamped Langevin sampling we begin by introducing an auxiliary random variable V

with values in \mathbb{R}^d . We interpret the random variable of interest X as the *position* and V as a *momentum* or *velocity*. Introducing also a functional $K : \mathbb{R}^d \rightarrow \mathbb{R}$ representing kinetic energy, we define the joint distribution of $(X; V)$, $\pi_{X,V}$ via its density with respect to the Lebesgue measure

$$\frac{d\pi_{X,V}}{d(x,v)}(x,v) = p_{X,V}(x,v) \propto \exp(-E(x) - K(v)).$$

Since $p_{X,V}(x,v)$ factorizes, the marginal distribution of X remains unchanged and the kinetic energy K can therefore be chosen freely as long as $v \mapsto \exp(-K(v))$ is integrable. Throughout this section we further assume the following:

Assumption 4.2. 1. Both $E, K : \mathbb{R}^d \rightarrow \mathbb{R}$ are twice continuously differentiable with Lipschitz continuous gradient.

2. K is symmetric, i.e., $K(v) = K(-v)$ for any $v \in \mathbb{R}^d$ and such that $\pi_V \in \mathcal{P}_2(\mathbb{R}^d)$ where

$$\pi_V(dv) := \frac{\exp(-K(v))}{\int \exp(K(w)) dw} dv.$$

A popular choice is $K(v) = v^T M^{-1} v / 2$ with a symmetric and positive definite matrix $M \in \mathbb{R}^{d \times d}$, which leads to a Gaussian marginal distribution for V . In HMC we make use of Hamiltonian dynamics to sample from $\pi_{X,V}$. Dropping the velocity variable leads to a sample $X \sim \pi_X(x) \propto \exp(-E(x))$. As a prerequisite, let us therefore discuss Hamiltonian dynamics.

4.6.1 Hamilton dynamics

Hamiltonian dynamics refer to a system of differential equations modeling phenomena in classical mechanics. Specifically, given the functional $H(x,v) = E(x) + K(v)$ (referred to as the *Hamiltonian*), Hamiltonian dynamics read as

$$\begin{cases} \frac{dx_i}{dt} = \frac{\partial H}{\partial v_i} \\ \frac{dv_i}{dt} = -\frac{\partial H}{\partial x_i} \end{cases} \quad (86)$$

We introduce the following notation.

Definition 4.6. For any $t > 0$ we define the solution operator of Hamiltonian dynamics as

$$\begin{aligned} \phi^t : \mathbb{R}^{2d} &\rightarrow \mathbb{R}^{2d} \\ (x,v) &\mapsto (\phi_x^t(x,v), \phi_v^t(x,v)) = (x(t), v(t)) \end{aligned} \quad (87)$$

where $(x(s), v(s))$ solves (86) with initial condition $(x(0), v(0)) = (x, v)$.

Using standard results on ODEs we can show that ϕ^t is, in fact, well defined as well as continuously differentiable.

Theorem 4.13. Let assumption 4.2 hold. Then ϕ^t is well defined for any $t \geq 0$ and continuously differentiable in (x, v, t) .

Proof. Under assumption 4.2 the right-hand side of (86) is Lipschitz continuous so that for any initial condition existence and uniqueness of a solution for Hamiltonian dynamics is guaranteed for all time by the Picard-Lindelöf theorem [43, 2]. Moreover $(x, v, t) \mapsto \phi^t(x, v)$ is continuous: One can easily check that by Lipschitz-continuity for some $K \geq 0$

$$\|\phi^t(x, v) - \phi^t(\tilde{x}, \tilde{v})\| \leq \|(x, v) - (\tilde{x}, \tilde{v})\| + \int_0^t K \|\phi^s(x, v) - \phi^s(\tilde{x}, \tilde{v})\| ds \quad (88)$$

so that Grönwall's inequality yields continuity of $(x, v) \mapsto \phi^t(x, v)$ for some fixed t . The continuity with respect to t is obvious. Regarding differentiability of $\phi^t(x, v)$, for the time derivative we immediately see that

$$\frac{\partial}{\partial t} \phi^t(x, v) = \begin{bmatrix} \frac{\partial H}{\partial v_i}(\phi^t(x, v)) \\ -\frac{\partial H}{\partial x_i}(\phi^t(x, v)) \end{bmatrix} \quad (89)$$

which is continuous in (x, v) and t . Let us denote

$$F(x, v) = \begin{bmatrix} \frac{\partial H}{\partial v_i}(x, v) \\ -\frac{\partial H}{\partial x_i}(x, v) \end{bmatrix}.$$

Since $\phi^t(x, v) = (x, v) + \int_0^t F(\phi^s(x, v)) ds$ one would expect the derivative (if it existed) to satisfy

$$\frac{\partial}{\partial(x, v)} \phi^t(x, v) = \mathbf{I} + \int_0^t DF(\phi^s(x, v)) \frac{\partial}{\partial(x, v)} \phi^s(x, v) ds.$$

Thus, denote moreover, $A(t) = DF(\phi^t(x, v))$ with D the Jacobian and consider the ODE

$$\frac{d}{dt} u(t) = A(t)u(t). \quad (90)$$

Since the right-hand side $(u, t) \mapsto A(t)u$ is linear in u and continuous in t the ODE admits a unique solution for all time again [48, Chapter 17]. One can check that $u(t)$ with initial condition $u(0) = \mathbf{I}$ is precisely the derivative of $\phi^t(x, v)$ with respect to (x, v) [48, Chapter 17.6]. \square

Hamiltonian dynamics exhibit three crucial properties which HMC relies on:

1. energy conservation: the Hamiltonian $H(x, v)$ is invariant under ϕ^t
2. volume preservice: ϕ^t is symplectic in (x, v) space; it does not change the volume of a set, and
3. reversibility: flipping the momentum variable reverses the trajectory.

We will provide theoretical justifications for all three of these properties as they build the foundation of HMC.

Energy conservation can easily be seen by computing the total derivative with respect to the time along a curve governed by (86):

$$\frac{dH(x, v)}{dt} = \sum_i \frac{\partial H(x, v)}{\partial x_i} \frac{dx_i}{dt} + \frac{\partial H(x, v)}{\partial v_i} \frac{dv_i}{dt} = 0 \quad (91)$$

Next we consider reversibility of ϕ which, in particular, establishes that ϕ^t is a diffeomorphism for all $t > 0$.

Theorem 4.14 (Reversibility of Hamiltonian dynamics). *Let $t > 0$. Hamiltonian dynamics satisfy that*

$$\phi^t(\phi_x^t(x, v), -\phi_v^t(x, v)) = (x, -v). \quad (92)$$

That is, the mapping $(x, v) \mapsto (\phi_x^t(x, v), -\phi_v^t(x, v))$ is an involution.

Proof. Denote $(\tilde{x}(s), \tilde{v}(s)) = (x(t-s), -v(t-s))$. Since H is symmetric with respect to the momentum, *i.e.*, $H(x, v) = H(x, -v)$ for any $x, v \in \mathbb{R}^d$ it holds

$$\frac{\partial H}{\partial x}(x, v) = \frac{\partial H}{\partial x}(x, -v), \quad \text{and} \quad \frac{\partial H}{\partial v}(x, v) = -\frac{\partial H}{\partial v}(x, -v). \quad (93)$$

As a result, $(\tilde{x}(s), \tilde{v}(s))$ satisfies Hamilton's equations as well. By uniqueness of solutions it follows $(\tilde{x}(s), \tilde{v}(s)) = \phi^s(x(t), -v(t))$. As a result,

$$\begin{aligned} \phi^t(\phi_x^t(x, v), -\phi_v^t(x, v)) &= \phi^t(x(t), -v(t)) \\ &= (\tilde{x}(t), \tilde{v}(t)) = (x(0), -v(0)) = (x, -v). \end{aligned} \quad (94)$$

□

Finally, the property of volume preservation is known as *Liouville's theorem* [4, Section 16] and is proven in the following.

Theorem 4.15 (Liouville's theorem). *Hamiltonian dynamics preserve volume. More precisely, for any $t > 0$ and any measurable set $A \in \mathcal{B}(\mathbb{R}^{2d})$ it holds that*

$$\iint \mathbb{1}_A(x, v) dx dv = \iint \mathbb{1}_{\phi^t(A)}(x, v) dx dv. \quad (95)$$

Proof. We provide the proof of a slightly more general result: Let ψ^t be the solution operator for the general ordinary differential equation $\frac{d}{dt}y = F(y)$, *i.e.*, $\psi^t(y) = y(t)$ where $y(0) = y$ and $\frac{d}{dt}y(t) = F(y(t))$ for $t > 0$. Assume that F is such that ψ^t is a diffeomorphism for any $t > 0$. Then, if $\text{div}(F) = 0$, ψ^t is volume preserving. The desired result then immediately follows by considering

$$F(x, v) = \begin{bmatrix} \frac{\partial H}{\partial v_i}(x, v) \\ -\frac{\partial H}{\partial x_i}(x, v) \end{bmatrix}.$$

Fix a measurable set A and let us denote the volume of $\psi^t(A)$ as $v(t)$, *i.e.*, $v(t) = \int \mathbb{1}_{\psi^t(A)}(y) dy$. We will show that the time derivative of v equals 0. By the transformation theorem for integrals we have that

$$\begin{aligned} \frac{d}{dt}v(t) &= \frac{d}{dt} \int \mathbb{1}_{\psi^t(A)}(y) dy \\ &= \frac{d}{dt} \int \mathbb{1}_{\psi^t(A)}(\psi^t(z)) |\det(D\psi^t(z))| dz \\ &= \frac{d}{dt} \int \mathbb{1}_A(z) |\det(D\psi^t(z))| dz \\ &= \int \mathbb{1}_A(z) \frac{d}{dt} |\det(D\psi^t(z))| dz. \end{aligned} \quad (96)$$

By continuity if we assume that t is sufficiently small, we can omit the absolute value as $\det(D\psi^t(z))|_{t=0} = 1$. Moreover, from Jacobi's formula it follows that

$$\frac{d}{dt} \det(D\psi^t) = \text{trace}\left((D\psi^t)^T \frac{d}{dt} D\psi^t\right). \quad (97)$$

In particular, using symmetry of the second derivatives it follows for $t = 0$,

$$\frac{d}{dt} \det(D\psi^t(y))|_{t=0} = \text{trace}\left(\frac{d}{dt} D\psi^t(y)\right)|_{t=0} = \text{trace}(DF(y)) = \text{div}(F)(y) = 0$$

Algorithm 7 The Hamiltonian Monte Carlo algorithm.

Require: Initial values (X_0, V_0) , parameters $T > 0$

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: Sample $\tilde{V}_k \sim \exp(-K(v)) \, dv$
- 3: Simulate Hamiltonian dynamics and set

$$\begin{cases} \bar{X}_{k+1} = \phi_x^T(X_k, \tilde{V}_k) \\ \bar{V}_{k+1} = -\phi_v^T(X_k, \tilde{V}_k) \end{cases} \quad (99)$$

- 4: Set the new iterate according to

$$(X_{k+1}, V_{k+1}) = \begin{cases} (\bar{X}_{k+1}, \bar{V}_{k+1}) & \text{with probability } \rho\left((X_k, \tilde{V}_k), (\bar{X}_{k+1}, \bar{V}_{k+1})\right) \\ (X^k, \tilde{V}_k) & \text{else.} \end{cases}$$

- 5: **end for**
-

and, thus, $\frac{d}{dt}v(0) = 0$. For $t > 0$, since for any $s, t > 0$, $\psi^{s+t}(y) = (\psi^s \circ \psi^t)(y)$ we can deduce that

$$\begin{aligned} v(t+s) &= \int_{\mathbb{R}^d} \mathbb{1}_{\psi^{t+s}(A)}(y) \, dy \\ &= \int_{\mathbb{R}^d} \mathbb{1}_{\psi^{t+s}(A)}(\psi^s(z)) |\det(D\psi^s(z))| \, dz \\ &= \int_{\mathbb{R}^d} \mathbb{1}_{\psi^t(A)}(z) |\det(D\psi^s(z))| \, dz. \end{aligned} \quad (98)$$

Using the same arguments as above we can deduce that $\frac{d}{dt}v(t) = \frac{d}{ds}v(s+t)|_{s=0} = 0$, which implies that v is constant. \square

4.6.2 The HMC algorithm

The HMC algorithm consists of multiple steps as depicted in algorithm 7. Given the previous iterate (X_k, V_k) , first we sample $\tilde{V}_k \sim \pi_V$ where π_V admits a density that is proportional to $\exp(-K(v)) \, dv$, which is possible directly if π_V is chosen, *e.g.*, as a Gaussian. Secondly, we simulate Hamiltonian dynamics¹⁰ for a time $T > 0$ and afterwards flip the sign of the velocity component resulting in $(\bar{X}_{k+1}, \bar{V}_{k+1})$. Flipping the sign has no impact on the value of the Hamiltonian and leaves $\pi_{X,V}$ unchanged, but it renders the dynamics reversible (*cf.*, theorem 4.14). Lastly, $(\bar{X}_{k+1}, \bar{V}_{k+1})$ is accepted as a new sample with the probability $\rho\left((X_k, \tilde{V}_k), (\bar{X}_{k+1}, \bar{V}_{k+1})\right)$ where

$$\rho((x, v), (\tilde{x}, \tilde{v})) = \min\{\exp(H(x, v) - H(\tilde{x}, \tilde{v})), 1\}$$

Otherwise we set $(X_{k+1}, V_{k+1}) = (X_k, \tilde{V}_k)$.

Remark 4.8. *Energy conservation would render the accept/reject step in algorithm 7 obsolete if we could simulate Hamiltonian dynamics exactly as $\exp\left(H(X_k, \tilde{V}_k) - H(\bar{X}_{k+1}, \bar{V}_{k+1})\right) = 1$, that is, the acceptance rate always equals one. In practice, however, the acceptance criterion compensates for errors induced by the used discretization scheme.*

¹⁰In practice, the discretization of the Hamiltonian dynamics has to maintain volume preservice and reversibility, while conservation of the Hamiltonian H is accounted for via a MH step and we refer the reader to remark 4.10 for a feasible method.

Remark 4.9. *Randomized as well as deterministic choices have been proposed for the simulation time T [31, 72]. In practice, it is often necessary to tune T by trial and error to obtain good performance [72, Section 4.2]. If a discrete scheme for simulating ϕ^t is performed for one step, the scheme reduces to MALA [31, Section 1].*

The following result shows that, indeed, $\pi_{X,V}(dx, dv) \propto \exp(-E(x) - K(v)) dx dv$ is invariant for HMC.

Theorem 4.16. *HMC admits $\pi_{X,V}(dx, dv) \propto \exp(-E(x) - K(v)) dx dv$ as a stationary measure if the Hamiltonian dynamics are simulated either exactly or using a scheme which is volume preserving and reversible in the sense of theorem 4.14.*

Proof. The HMC algorithm can be separated in two distinct steps: Sampling the proposal momentum $\tilde{V}_k \sim \exp(-K(v))$ and, afterwards, simulating the Hamiltonian dynamics and performing an accept/reject step. We can show that $\pi_{X,V}$ is invariant with respect to the entire HMC update by showing that it is invariant with respect to both of these steps individually. For the first step this is trivially satisfied: If $(X, V) \sim \pi_{X,V}$ then X and V are independent and for $\tilde{V} \sim \pi_V$ which is again independent of X it holds that $(X, V) \sim (X, \tilde{V}) \sim \pi_{X,V}$. For the second step, we prove invariance by showing that the target distribution satisfies the detailed balance conditions. Let $f : \mathbb{R}^{2d} \rightarrow \mathbb{R}$ be an arbitrary bounded and measurable function and let Q be the transition kernel induced by the Hamiltonian dynamics, *i.e.*,

$$Q((x, v), \cdot) = \delta_{(\phi_x^T(x, v), -\phi_v^T(x, v))} \cdot$$

We assume for now that the Hamiltonian dynamics are simulated exactly so that the acceptance probability is always equal to one. Then, for any bounded and measurable f we obtain that

$$\begin{aligned} & \iint f((\tilde{x}, \tilde{v}), (x, v)) p_{X,V}(x, v) Q((x, v), d(\tilde{x}, \tilde{v})) d(x, v) \\ & \propto \int f((\phi_x^T(x, v), -\phi_v^T(x, v)), (x, v)) \exp(-E(x) - K(v)) d(x, v) \\ & \stackrel{*}{=} \int f((\phi_x^T(x, v), -\phi_v^T(x, v)), (x, v)) \exp(-E(\phi_x^T(x, v)) - K(-\phi_v^T(x, v))) d(x, v) \quad (100) \\ & \stackrel{**}{=} \int f((x, v), (\phi_x^T(x, v), -\phi_v^T(x, v))) \exp(-E(x) - K(-v)) d(x, v) \\ & = \int f((x, v), (\tilde{x}, \tilde{v})) p_{X,V}(x, v) Q((x, v), d(\tilde{x}, \tilde{v})) d(x, v) \end{aligned}$$

where the equality marked with $*$ is due to the Hamiltonian H being invariant under Hamiltonian dynamics as well as under flipping the sign of the momentum component. The equality marked with $**$ follows after replacing (x, v) with $(\phi_x^T(x, v), -\phi_v^T(x, v))$ using the transformation theorem for integrals and employing reversibility of Hamiltonian dynamics, which was proved in theorem 4.14. Since Hamiltonian dynamics as well as flipping the sign are volume preserving, the determinant of the Jacobian of the transformation is equal to one. As a result we obtain detailed balance and, thus, invariance of $\pi_{X,V}$.

If ϕ^T only constitutes an approximate scheme for simulating the Hamiltonian dynamics which is, however, volume preserving and reversible (*e.g.*, the leapfrog scheme) similar arguments apply: In this case, the transition kernel of the Hamiltonian dynamics including the Metropolis acceptance step reads as

$$\begin{aligned} R((x, v), d(\tilde{x}, \tilde{v})) & = \delta_{(\phi_x^T(x, v), -\phi_v^T(x, v))}(\tilde{x}, \tilde{v}) \rho((x, v), (\tilde{x}, \tilde{v})) d(\tilde{x}, \tilde{v}) \\ & \quad + \delta_{(x, v)}(\tilde{x}, \tilde{v}) (1 - \rho((x, v), (\phi_x^T(x, v), -\phi_v^T(x, v)))) d(\tilde{x}, \tilde{v}) \end{aligned} \quad (101)$$

Using the fact that for any $x, v, \tilde{x}, \tilde{v}$ it holds that

$$p_{X,V}(x, v)\rho((x, v), (\tilde{x}, \tilde{v})) = p_{X,V}(\tilde{x}, \tilde{v})\rho((\tilde{x}, \tilde{v})(x, v))$$

and again volume preserverence and reversibility of the scheme ϕ^T we find for any bounded and measurable f that

$$\begin{aligned} & \iint f((x, v), (\tilde{x}, \tilde{v}))p_{X,V}(x, v)R((x, v), d(\tilde{x}, \tilde{v}))d(x, v) \\ &= \int f((x, v), (\phi_x^T(x, v), -\phi_v^T(x, v)))p_{X,V}(x, v)\rho((x, v), (\phi_x^T(x, v), -\phi_v^T(x, v)))d(x, v) \\ & \quad + \int f((x, v), (x, v))p_{X,V}(x, v) (1 - \rho((x, v), (\phi_x^T(x, v), -\phi_v^T(x, v)))) d(x, v) \\ &= \int f((\phi_x^T(x, v), -\phi_v^T(x, v)), (x, v))p_{X,V}(x, v)\rho((x, v), (\phi_x^T(x, v), -\phi_v^T(x, v)))d(x, v) \\ & \quad + \int f((x, v), (x, v))p_{X,V}(x, v) (1 - \rho((x, v), (\phi_x^T(x, v), -\phi_v^T(x, v)))) d(x, v) \\ &= \iint f((\tilde{x}, \tilde{v}), (x, v))p_{X,V}(x, v)R((x, v), d(\tilde{x}, \tilde{v}))d(x, v) \end{aligned} \tag{102}$$

concluding the proof. \square

Remark 4.10. *Theorem 4.16 requires us to choose a discretization of Hamiltonian dynamics which maintains reversibility and volume preserverence whereas conservation of the total energy is accounted for by the MH acceptance step within the algorithm. In practice the most popular choice for such a discretization is the Störmer-Verlet—or leapfrog—method [72, 31]. Its update rule with step size $h > 0$ reads as*

$$\begin{cases} v_{n+\frac{1}{2}} = v_n - \frac{h}{2}\nabla_x E(x_n) \\ x_{n+1} = x_n + hM^{-1}v_{n+\frac{1}{2}} \\ v_{n+1} = v_{n+\frac{1}{2}} - \frac{h}{2}\nabla_x E(x_{n+1}). \end{cases} \tag{103}$$

The choice of the discretization step size $h > 0$ as well as the number of steps performed in each iteration then become crucial parameters of the method, cf., remark 4.11.

Ergodicity of the HMC algorithm has been established in various ways [61, 31, 14, 63]. We present a result here and refer to [31, Theorem 2] for a proof.

Theorem 4.17. *Let $K(v) = \frac{\|v\|^2}{2}$ and let E be continuously differentiable with Lipschitz continuous gradient. Let ϕ^T be the leapfrog scheme for the simulation of Hamiltonian dynamics with step size $h > 0$ and number of steps T . Assume in addition that either*

- *there exist $c > 0$ and $\beta \in [0, 1)$ such that for all x , $\|\nabla E(x)\| \leq c(1 + \|x\|^\beta)$, or*
- *there exists $c > 0$ such that $\|\nabla E(x)\| \leq c(1 + \|x\|)$ and $(1 + hL^{\frac{1}{2}} + \nu(hL^{\frac{1}{2}}))^T - 1 < 1$ where ν is defined as $\nu(s) = 1 + \frac{s}{2} + \frac{s^2}{4}$.*

Then for π -a.e. $x \in \mathbb{R}^d$, $d_{TV}(\delta_x R_X^n, \pi) \rightarrow 0$ as $n \rightarrow \infty$, where R_X denotes the transition kernel on the position variable.

A proof of geometric ergodicity of HMC is provided under more technical conditions in [31].

Remark 4.11 (Parameter choices). *Restricting to the practically most relevant case of $K(v) = \frac{1}{2}\|v\|^2$, the remaining parameters to choose within HMC are the step size $h > 0$ as well as the number of steps $L \in \mathbb{N}$ of the discretization of the Hamiltonian dynamics (cf., remark 4.10). Regarding the step size h , too large values might lead to low acceptance rates. On the other hand, small h will lead to high computational efforts (large L) or a slowly moving chain (small L). Regarding the number of steps L on the other hand, low values might lead to less exploring and, thus, slower mixing of the chain, larger values to high computational cost. As elaborated in [72], it is advised to perform several tuning runs of HMC, first setting h and afterwards L . As explained in [72, Section 4.2] the region of stability for h is governed roughly by the square root of the smallest eigenvalue of the covariance matrix of π_X and $L = 100$ is a reasonable starting point for the number of steps for complex problems. On the other hand, the no-U-turn sampler (NUTS) [49] offers a possible alternative. In NUTS the need for choosing a number of steps L is alleviated. Moreover, in [49] approaches for an automatic choice of the step size h are provided.*

4.7 Further reading

4.7.1 Time-inhomogeneous chains

Many approaches for sampling combine existing sampling techniques with some type of *annealing* or *tempering*. That is, instead of directly targeting the distribution π with a MC, one considers a family of distributions $(\pi_n)_{n=1}^N$ such that $\pi_0 = \pi$ and π_N is a simple reference distribution. Then sampling is performed by sampling successively from π_n for $n = N, \dots, 1$. Examples include geometric tempering [71, 21], annealed Langevin sampling [88], or diffusion at absolute zero [41]. Subsequently, annealed Langevin sampling led to diffusion models [89].

In adaptive MCMC [3] it is assumed that we have access to a family of Markov transition kernels R_θ which are parametrized by θ and such that for any θ , R_θ is ergodic with stationary distribution θ . During the simulation, the parameter θ is chosen adaptively and afterwards one step is performed using the kernel R_θ . In order to ensure that the resulting chain is still ergodic with stationary measure π the adaptation of θ diminishes over time.

4.7.2 Deterministic approximation of π

Instead of directly trying to sample from our target distribution π , the idea of deterministic approximation [10, Chapter 10] is to approximate π by a tractable distribution π_θ which is easy to sample from. A popular approach is to choose a family of distributions $(\pi_\theta)_\theta$ parametrized by θ and then find the value of θ which yields the best approximation of π . For instance in *variational inference* [10, Section 10.1] the parameter θ is determined by minimizing the KL divergence

$$\min_{\theta} d_{\text{KL}}(\pi_\theta, \pi). \quad (104)$$

Assuming that the occurring distributions admit strictly positive densities with respect to the Lebesgue measure as $\frac{d\pi}{dx}(x) = p(x)$ and $\frac{d\pi_\theta}{dx}(x) = p_\theta(x)$, the gradient of this objective may be computed as

$$\begin{aligned} \nabla_{\theta} d_{\text{KL}}(\pi_\theta, \pi) &= \nabla_{\theta} \left[- \int p_\theta(x) \log \left(\frac{p(x)}{p_\theta(x)} \right) dx \right] \\ &= \int -\nabla_{\theta} p_\theta(x) \log(p(x)) + \nabla_{\theta} p_\theta(x) \log(p_\theta(x)) + \nabla_{\theta} p_\theta(x) dx \\ &= \mathbb{E}_{X \sim \pi_\theta} [\nabla_{\theta} \log(p_\theta(X)) \{ \log(p_\theta(X)) - \log(p(X)) \}]. \end{aligned} \quad (105)$$

In the last equality we used the elementary equality $\nabla_{\theta} p_{\theta} = p_{\theta} \nabla \log(p_{\theta})$ and the fact that by integration by parts

$$\int \nabla_{\theta} p_{\theta}(x) dx = 0.$$

Note that the expectation in (105) is with respect to the tractable distribution p_{θ} and, thus, the gradient can be approximated effectively.

In expectation propagation, on the other hand, the objective from variational inference is simply changed by flipping the arguments of the KL divergence leading to

$$\min_{\theta} d_{\text{KL}}(\pi, \pi_{\theta}). \quad (106)$$

For more information we refer the interested reader to [10, Section 10.7].

A particularly interesting approach connected to variational inference is posed by *Stein variational gradient descent* [60]. There, it is proposed to approximate the target distribution π_X as the push-forward measure $T_{\#}\pi_Z$ with some fixed simple reference distribution π_Z and a transformation T to be determined. The transformation T is ideally chosen to minimize $T \mapsto d_{\text{KL}}(T_{\#}\pi_Z, \pi)$. In [60], for the specific case that the transformation T is a perturbation of the identity $T(x) = x + f(x)$ for some function $f \in \mathcal{H}^d$ where \mathcal{H}^d is a reproducing kernel Hilbert space with reproducing kernel $k(\cdot, \cdot)$ the authors derive a formula for the gradient of $d_{\text{KL}}(T_{\#}\pi_Z, \pi)$ with respect to the perturbation f which reads as

$$\begin{aligned} \nabla_f d_{\text{KL}}(T_{\#}\pi_Z, \pi)|_{f=0} &= -\mathbb{E}_{Z \sim \pi_Z} [k(Z, \cdot) \nabla_x \log(p_X(Z)) + \nabla_x k(Z, \cdot)] \\ &=: -\phi_{\pi_Z, \pi}. \end{aligned}$$

Therefore, in a gradient descent fashion, we can reduce the KL divergence using the iteration $\pi^0 = \pi_Z$ and for $k = 1, 2, \dots$

$$\pi^{k+1} = T_{\#}^k \pi^k, \quad \text{where } T^k(x) = x + \epsilon_k \phi_{\pi^k, \pi}(x) \quad (107)$$

where $(\epsilon_k)_k$ is a sequence of step sizes. Interestingly, this iteration is implemented not on the space of probability distributions but in sample space, by initializing randomly $X_0 \sim \pi_Z$ and updating the sample according to

$$X_{k+1} = X_k + \epsilon_k \phi_{\pi^k, \pi}(X_k).$$

Therefore, while initially motivated using variational inference, the method, in fact, yields a sampling algorithm. For details we refer to [60].

5 Numerical experiments

In this section, we turn to practical experiments with models that explicitly allow us to verify key theoretical properties that are required for valid energy-based modeling.

5.1 Energy model

Although we discussed numerous potential architectures for constructing suitable energy functionals in section 3.3, verifying essential properties such as integrability and appropriate growth

conditions of the associated Gibbs distributions remains challenging for many architectures—especially those based on deep neural networks. Consequently, we restrict our focus here to the classical FoE model defined as

$$E_{\theta}(x) = \sum_{i=1}^n \sum_{j=1}^o \phi_j((K_j x)_i) \quad (108)$$

which corresponds to a Gibbs distribution with density

$$p_{\theta}(x) \propto \prod_{i=1}^n \prod_{j=1}^o \exp(-\phi_j((K_j x)_i)). \quad (109)$$

In this formulation, the parameters are the weights in the linear operators K_1, K_2, \dots, K_o and any potential parameters of the potentials $\phi_1, \phi_2, \dots, \phi_o$. The linear operators are typically chosen to encode the common assumption that natural images are stationary, *i.e.*, that the likelihood of any feature in the image is independent of its spatial location. Although we do not give a rigorous proof here, the model is stationary if the linear operators encode convolutions with circular boundary conditions, which is our choice in this work.

Regarding the potentials, Roth and Black [82] originally utilized the potentials that correspond to the leptokurtic Student-t distribution; a choice motivated by the empirical observation the responses of natural images to filters follow a leptokurtic distribution. The belief that the factors coincide with the corresponding filter marginals is a misconception that is sometimes found in publications to this day, even though the seminal works of Zhu, Wu, and Mumford [103, 102, 104] clarified that potentials serve instead as dual variables in a maximum entropy problem that ensure that the model marginals match target statistics rather than directly mirroring empirical marginals.

Indeed, from a principled standpoint, potentials should have finite support since natural images and their filter responses lie within bounded intervals. Thus, for a truly representative FoE model, potentials should ideally reflect finite support dependent on the chosen filters. To approximate potentials with finite support while simultaneously satisfying the smoothness requirements of the various optimization and sampling algorithms, we utilize negative-log Gaussian mixture model (GMM) as potentials. Explicitly, each potential is modeled as

$$\phi_j(x) = -\log\left(\sum_{i=1}^W w_{j,i} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma^2}\right)\right), \quad (110)$$

where the means $\mu_1, \mu_2, \dots, \mu_W$ are positioned on an equidistant grid within an interval $[-\nu, \nu]$. Here, the parameter $\nu \in \mathbb{R}$ must be set large enough to account for strong filter responses but should be small enough such that the potential can exhibit small-scale features where needed, without the number of components becoming excessively large. The variance σ^2 is chosen a-priori and fixed.

In fig. 1, we demonstrate how negative-log GMMs can effectively approximate common potentials such as those derived from the Laplace distribution, the Student-t distribution, or the Mexican hat on the chosen interval. Outside of this interval, the potentials grow quadratically towards infinity. This behaviour is intentional and simulates the theoretically desirable finite-support property while remaining sufficiently smooth to support inference via first-order optimization and sampling methods that are essential for during learning and during the resolution of the inverse problems.

The structure of the model and the choice of the negative-log Gaussian mixture potentials facilitate the explicit verification of the properties that are essential for valid energy-based

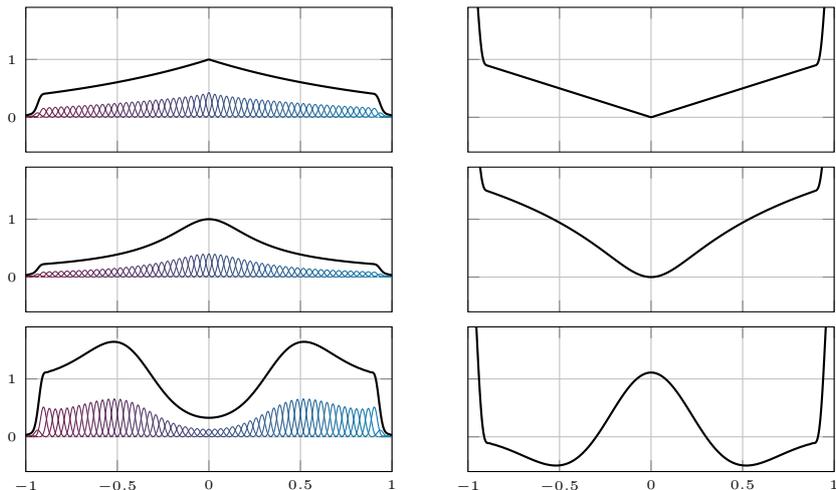


Figure 1: Approximation of various densities via GMMs (left) and the corresponding potentials (right). From top to bottom: Laplace; Student-t; Mexican-hat. To avoid clutter we only plot every second component.

modeling. Specifically, the density defined by (109) can be shown to be a Gaussian mixture model whose precision matrix is given by

$$\frac{1}{\sigma^2} \sum_{i=1}^o (K_i)^\top K_i. \quad (111)$$

The rank of this precision matrix critically depends on the properties of the convolution matrices K_1, K_2, \dots, K_o . A full-rank precision matrix implies a valid Gaussian mixture density with respect to the Lebesgue measure, whereas a rank-deficient matrix results in a distribution supported on a lower-dimensional subspace without a valid density with respect to the Lebesgue measure. To ensure a full-rank precision matrix, we adopt the approach from [23] that constructs the filters k_1, k_2, \dots, k_o that correspond to the convolution matrices K_1, K_2, \dots, K_o from a linear combination of the basis filters b_1, b_2, \dots, b_b by

$$k_i = \sum_{j=1}^b \beta_{i,j} b_j \quad (112)$$

where the coefficients $(\beta_{i,j})_{i=1,j=1}^{b,o}$ are learnable. In particular, the basis filters b_1, b_2, \dots, b_o are given by the discrete cosine transform of size 5×5 , which results in $o = 25$ filters and ensures that the precision matrix has full rank by a straightforward application of the convolution theorem so long as all coefficients are different from zero. In practice, however, the constant basis vector of the DCT is typically excluded during training to enforce equivariance concerning radiometric shifts. To reconcile theoretical requirements with this practical consideration, we implicitly incorporate smooth constraints on the excluded constant component to prevent radiometric biases, though these constraints are rarely active during training.

5.2 Parameter estimation

We consider two different methods for estimating the model parameters: score-matching and bilevel optimization. In detail, we define the target density as $g_\sigma * (\frac{1}{N} \sum_{i=1}^N \delta_{x_i})$ with $\sigma = 2 \times 10^{-2}$

and where N is the number of overlapping patches of size 96×96 in the BSDS500 training dataset and $x_1, \dots, x_N \in \mathbb{R}^{96 \times 96}$ are those patches.¹¹ The parameters of the bilevel model and the score-matching model were both found by optimizing the respective objectives with the Adam optimizer [54]. The parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ of the Adam optimizer were set to standard choices, but we found it necessary to tune the learning rates of the coefficients of the filters and the weights of the potentials separately. For both learning methods, we used a learning rate of 1×10^{-5} for the weights of the potentials, and used 2×10^{-4} and 5×10^{-4} for the coefficients of the filters for the score-matching training and the bilevel training, respectively. The parameter ν was set to 0.8, which is informed by the largest magnitude of any filter response prior to training that was 0.8185.¹² The negative-log GMM potentials were given $W = 123$ components and the variance was chose as $\sigma_{\text{sm}}^2 = \frac{2\nu}{(W-1)}$ for the score-matching training and as $1.5\sigma_{\text{sm}}^2$ for the bilevel training. The weights of all $o = 24$ potential functions were initialized with the vector

$$\text{proj}_{\Delta^W}(w) \tag{113}$$

where $w \in \mathbb{R}^W$ with entries $w_i = -\log(|\mu_i| + 0.001)/10000$, and proj_{Δ^W} is the projection onto the W -dimensional simplex. Since we do not require a normalized model and in order to give the model more freedom, we do not project the weights onto the simplex during training. Each of the coefficients $(\beta_{i,j})_{i=1,j=1}^{b,o}$ of the filters is initialized with γ/oz where z is a standard normal random variable and $\gamma = 2.5$ for the score-matching training and $\gamma = 1.5$ for the bilevel training. In addition, for the bilevel training we introduced a learnable scalar λ that acts as the standard tradeoff parameter in the variational problem. It was crucial to tune this parameter such that initial reconstructions were reasonable; it was initialized with $1/25$ and learned with learning rate 1×10^{-4} .

The objective and the gradient with respect to the parameters for the minimization of the Fisher divergence can be readily computed by plugging in the energy (108) into the denoising score-matching loss (19) and the utilization of automatic differentiation frameworks such as PyTorch [77]. In contrast, the bilevel learning approach necessitates the choice of the lower-level problem and upper-level loss function, the resolution of the parameter-dependent lower-level problem, and the subsequent computation of the gradient of the upper-level loss function with respect to those parameters. For the upper-level loss function we stick to the standard choice $L(x, y) = \|x - y\|^2/2$ due to its smoothness and its relationship to MMSE estimation. Motivated by the significance of denoising algorithms for generative modeling as well as the resolution of inverse problems in the form of regularization by denoising, plug-and-play methods, and diffusion models, we choose denoising as the lower-level problem with variance 0.1 and resolve the lower-level problem with the accelerated gradient descent with Lipschitz backtracking given in algorithm 8. We compute the gradient of the upper-level loss function by utilizing the implicit function theorem approach that we outlined in section 3.2. The Hessian-vector product $(H(\theta))^{-1}\nabla_x L(x^*(\theta))$ is computed via 200 iterations of the conjugate gradient algorithm and the Jacobian-vector product is computed via automatic differentiation.

5.3 Results

In contrast to energies parametrized by deep neural networks, the FoE model given in (108) has a structure that lends itself towards interpretation through the plotting of the various learned components, namely the learned convolution kernels and the learned potentials. We show the models obtained by denoising score-matching and bilevel learning in fig. 2 and fig. 3, respectively.

¹¹There was no noise added to the reference images in bilvel learning.

¹²As is well known, the filter responses are highly leptokurtic: The first and 99th percentile of the filter responses prior to training were -0.1131 and 0.1161 , respectively.

Algorithm 8 Accelerated proximal gradient descent algorithm with Lipschitz backtracking.

Require: Number of iterations K , initial solution x^0 , initial L_0 , number of backtracking iterations

$J, \beta \in (0, 1), \gamma > 1$, relative tolerance r

- 1: $x^{-1} = x^0$
- 2: **for** $k = 0, 1, \dots, K - 1$ **do**
- 3: $\bar{x} = x^k + (x^k - x^{k-1})/\sqrt{2}$
- 4: **for** $j = 0, 1, \dots, J - 1$ **do** ▷ Lipschitz backtracking procedure [68]
- 5: $x^{k+1} = \text{prox}_{\frac{1}{L_k}g}(x - \nabla f(\bar{x}/L_k))$
- 6: **if** $f(x^{k+1}) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), x^{k+1} - \bar{x} \rangle + \frac{L_k}{2} \|\bar{x} - x^{k+1}\|^2$ **then**
- 7: $L_k = \beta L_k$
- 8: **break**
- 9: **end if**
- 10: $L_k = \gamma L_k$
- 11: **end for**
- 12: **end for**
- 13: **return** x^k

It is evident that both methods of obtaining the parameters suffer from spurious low-energy regions that are a consequence of the fact that filter responses rarely or never land in these regions during training.¹³ This could be remedied by careful hand-tuning of the parameter ν for each filter at the beginning of the training, which is very laborious. An alternative would be to simply discard those components of the GMM the resulting potentials that are responsible for the spurious low-energy regions. This could be done by, *e.g.*, recording the extreme positions of the filter responses on the training set and setting the weights of components centered around more extreme positions to zero.

After the models were learned, they can be used as priors in the resolution of inverse problems. The inverse problems we consider are denoising, reconstruction from Fourier samples and reconstruction from Radon samples. In detail, for all three tasks we construct the eight data $y_1, y_2, \dots, y_8 \in \mathbb{K}^d$ where \mathbb{K} is either \mathbb{R} or \mathbb{C} as

$$y_i = Ax_i + \gamma \tag{114}$$

for $i = 1, 2, \dots, 8$ where $A \in \mathbb{K}^{n \times d}$ is the matrix-representation of various linear forward operators \mathcal{F} that are described later and $x_1, x_2, \dots, x_8 \in \mathbb{R}^n$ are the first eight images (lexicographical ordering of the filenames) in the BSDS500 validation data set. We restrict our evaluation to those eight images due to computational reasons. For denoising, the forward operator A is the identity, which results in $d = n$. For Fourier sampling, $A = MF : \mathbb{R}^n \rightarrow \mathbb{C}^d$ where $F : \mathbb{R}^n \rightarrow \mathbb{C}^{\lfloor n/2 \rfloor + 1}$ describes the Fourier transform that accounts for the conjugate symmetry of the spectrum of a real signal and $M : \mathbb{C}^{\lfloor n/2 \rfloor + 1} \rightarrow \mathbb{C}^d$ samples d entries of a vector with $\lfloor n/2 \rfloor + 1$ entries. More specifically, it retains 10% of the low-frequency components and randomly discards 75% of the remaining components. For Radon sampling, A is the parallel beam projector provided by the TIGRE library [8] with 800 detectors that are 0.8 pixels wide and acquires 150 projections whose rotation angles are equispaced in the interval $[0, \pi]$. In all cases, γ is a vector of dimension d whose entries are i.i.d. (possibly complex) Gaussian noise with fixed variance. In particular, the variance

¹³This “locality” of the denoising score-matching loss is a common criticism. Indeed, it sparked the invention of diffusion models, which an ensemble of models trained with denoising score-matching with varying noise variance. The locality becomes less and less of an issues as the variance increases.

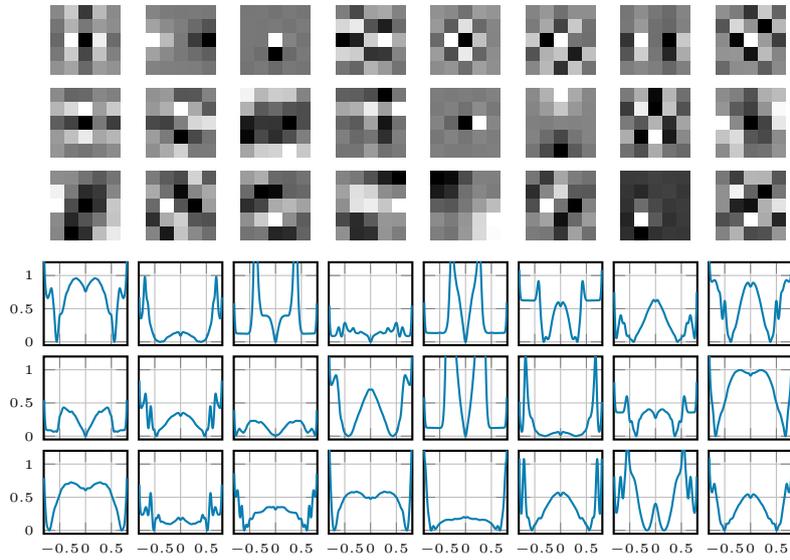


Figure 2: Filters (top) and corresponding potentials (bottom) learned via score-matching.

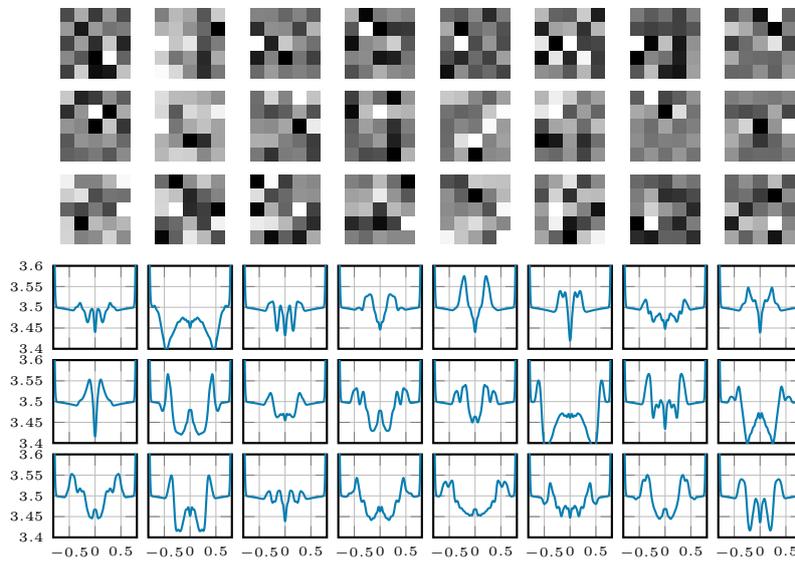


Figure 3: Filters (top) and corresponding potentials (bottom) learned via bilevel optimization.

was chosen as 0.1 for denoising, 2×10^{-3} for Fourier sampling, and 15 for Radon sampling.¹⁴

For the resolution of the inverse problem, we consider the posterior distribution

$$p_X(x | Y = y) \propto p_Y(y | X = x)p_X(x). \quad (115)$$

The likelihood $p_Y(y | X = x)$ is derived from the measurement model and the noise distribution, given explicitly by

$$p_Y(y|X = x) = (2\pi\sigma^2)^{(-d/2)} \exp\left(-\frac{\|y - \mathcal{F}(x)\|^2}{2\sigma^2}\right) \quad (116)$$

based on the assumption of the additive white Gaussian noise in (114).

In the variational treatment of inverse problems, it is common practice to consider a modified posterior

$$p_x^\lambda(x | Y = y) \propto p_Y(y | X = x)(p_X(x))^\lambda, \quad (117)$$

where $\lambda > 0$ is as a tunable parameter. This parameter is an additional degree of freedom to compensate for modeling mismatches between the learned prior and the underlying distribution, to compensate for approximate inference schemes, or to fine-tune the performance of the model with respect to some quality metric. In addition, it is common to consider a rescaled posterior of the form

$$p_x^\lambda(x | y) \propto (p_Y(y | x)(p_X(x))^\lambda)^{T^{-1}} \quad (118)$$

where $T > 0$ acts as a rescaling parameter that is analogous to a physical temperature. While temperature rescaling can theoretically facilitate convergence by controlling the variance of the distribution—such that as $T \rightarrow 0$, the posterior increasingly concentrates around its highest-density regions—it is introduced here explicitly to address practical numerical issues. Specifically, in later experiments we find that using the ULA without temperature scaling leads to exploration of spurious high-likelihood regions that arise due to artifacts from the training process. Introducing the temperature parameter ensures the ULA sampler reliably explores regions near genuine modes, effectively stabilizing the sampling and improving the quality of the posterior inference.

Although both MMSE and MAP inference strategies are feasible for models trained with either bilevel learning or score-matching, a mismatch between the training objectives and the inference methods typically leads to suboptimal results.¹⁵ Consequently, we pursue a MAP inference strategy specifically for the model trained with bilevel learning and an MMSE inference strategy for the model trained with score-matching. For the MAP inference, we use the accelerated gradient descent with Lipschitz backtracking, as detailed in algorithm 8 and as used during training. For the MMSE inference, we use the ULA detailed in algorithm 6 with the temperature set to $T = 1 \times 10^{-1}$ for denoising tasks and $T = 5 \times 10^{-2}$ for reconstructions from Fourier and Radon samples. We use the ULA since it is easy to implement and we are not necessarily interested in obtaining unbiased samples due to the temperature rescaling, the introduction of the weighting parameter λ , and the need to avoid spurious high-density regions. In summary, we obtain a practical sampling scheme that is loosely tuned with respect to the visual quality of the reconstruction rather than a sampling scheme that provably gives unbiased samples from the posterior. As a baseline method for comparison, we use the anisotropic total variation regularizer and solve the variational problem using the primal dual hybrid gradient algorithm [19]. Various

¹⁴These variances are chosen such that $\frac{1}{8} \sum_{j=1}^8 \frac{\sum_{i=1}^d |(Ax_j)_i|^2/d}{10^{\text{SNR}/10}}$ is approximately equal to a prescribed signal-to-noise ratio SNR, namely 10 for denoising and 30 for Fourier and Radon sampling. For Fourier and Radon sampling, the large difference magnitude of the variances is due to different normalizations of the forward operators.

¹⁵We observed these suboptimal results in our experiments but do not provide any results here for the sake of conciseness.



Figure 4: The `watercastle` image.

	Backprojection	TV	Bilevel	Score-matching
Denoising	20.01± 0.02	27.05± 1.60	28.03± 1.86	27.97± 2.05
Fourier	24.46± 1.97	26.30± 3.02	26.72± 3.16	26.88± 3.41
Radon	-87.68± 2.95	24.79± 2.16	25.40± 2.28	25.32± 2.60

Table 1: PSNR in dB (mean ± standard deviation) of the reconstructions obtained by the various methods, rounded to two decimals.

choices of parameters that are not discussed in detail in this manuscript can be found in the online repository <https://github.com/zacmar/ebm-inverse>.

Qualitative results for denoising, reconstruction from Fourier samples, and reconstruction from Radon samples of the popular `watercastle` image (shown in fig. 4) are presented in figs. 5 to 7, respectively. Since the marginal standard deviation is effectively determined by the choice of the temperature, we omit a colorbar. Quantitative results in terms of peak signal-to-noise ratio over the test set are provided in table 1.

The reconstructions obtained by the learned models are consistently of better quality than those obtained by total variation regularization. The addition of the pixel-wise marginal standard deviation serves as an illustration of one of the great benefits of the Bayesian approach, which is that many quantities of interest, such as point estimators or indicators of uncertainty, can be derived from the formal solution of the inverse problem, which is the posterior distribution. Nevertheless, these superior results required extensive parameter tuning. Specifically, introducing the temperature parameter in the ULA algorithm was essential to prevent exploration of spurious low-energy regions arising from the locality inherent in score-matching training, effectively restricting sampling to modes of the posterior density. We anticipate that results can be considerably improved by employing alternative, more robust training methodologies, such as combining Kullback-Leibler divergence minimization with highly efficient Gibbs samplers, as demonstrated successfully in [56]. Overall, these findings underscore the persistent challenges in conducting rigorous Bayesian inference for imaging tasks, even with rapidly expanding computational resources.

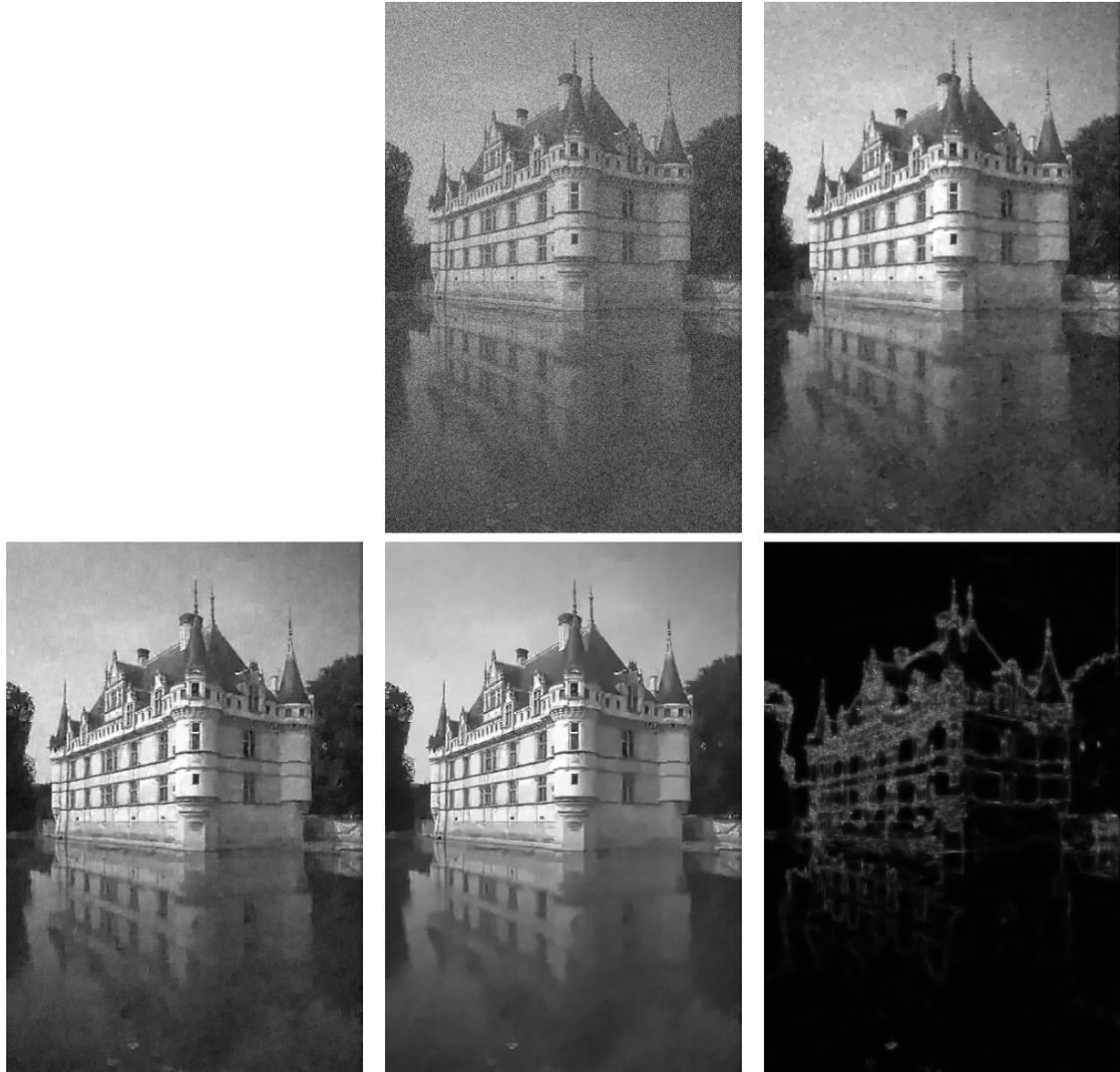


Figure 5: Qualitative denoising results: The top row shows the data, which coincides with the naive reconstruction, and the reconstruction obtained through total variation regularization. The bottom row shows the reconstruction obtained through regularization with the bilevel model, the MMSE estimate obtained through the sampling of the posterior of the score-matching prior, and the corresponding pixel-wise marginal standard deviation.

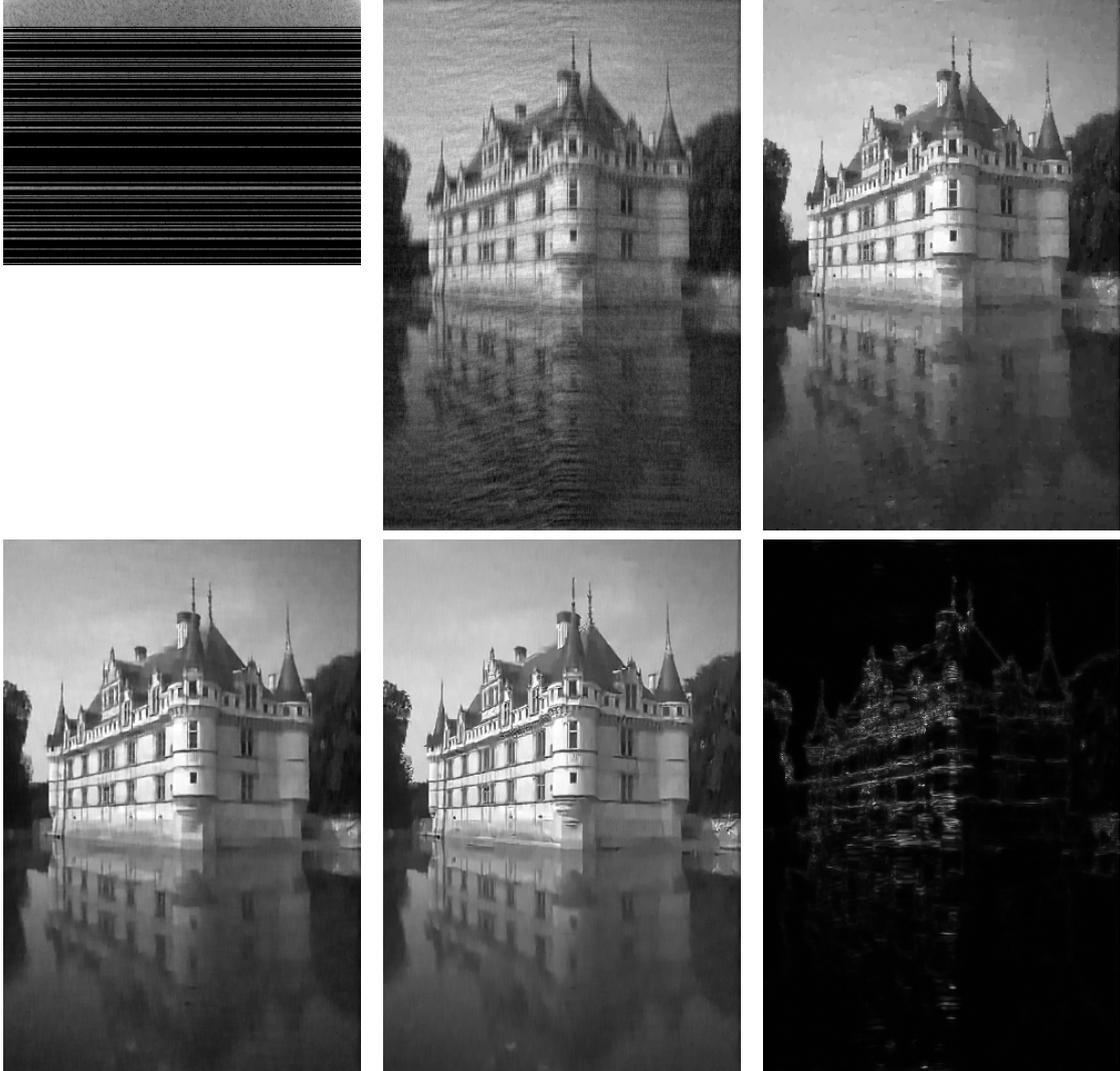


Figure 6: Qualitative results of reconstruction from Fourier samples: The top row shows a visualization of the data, the naive reconstruction obtained by backprojection, and the reconstruction obtained through total variation regularization. The bottom row shows the reconstruction obtained through regularization with the bilevel model, the MMSE estimate obtained through the sampling of the posterior of the score-matching prior, and the corresponding pixel-wise marginal standard deviation.

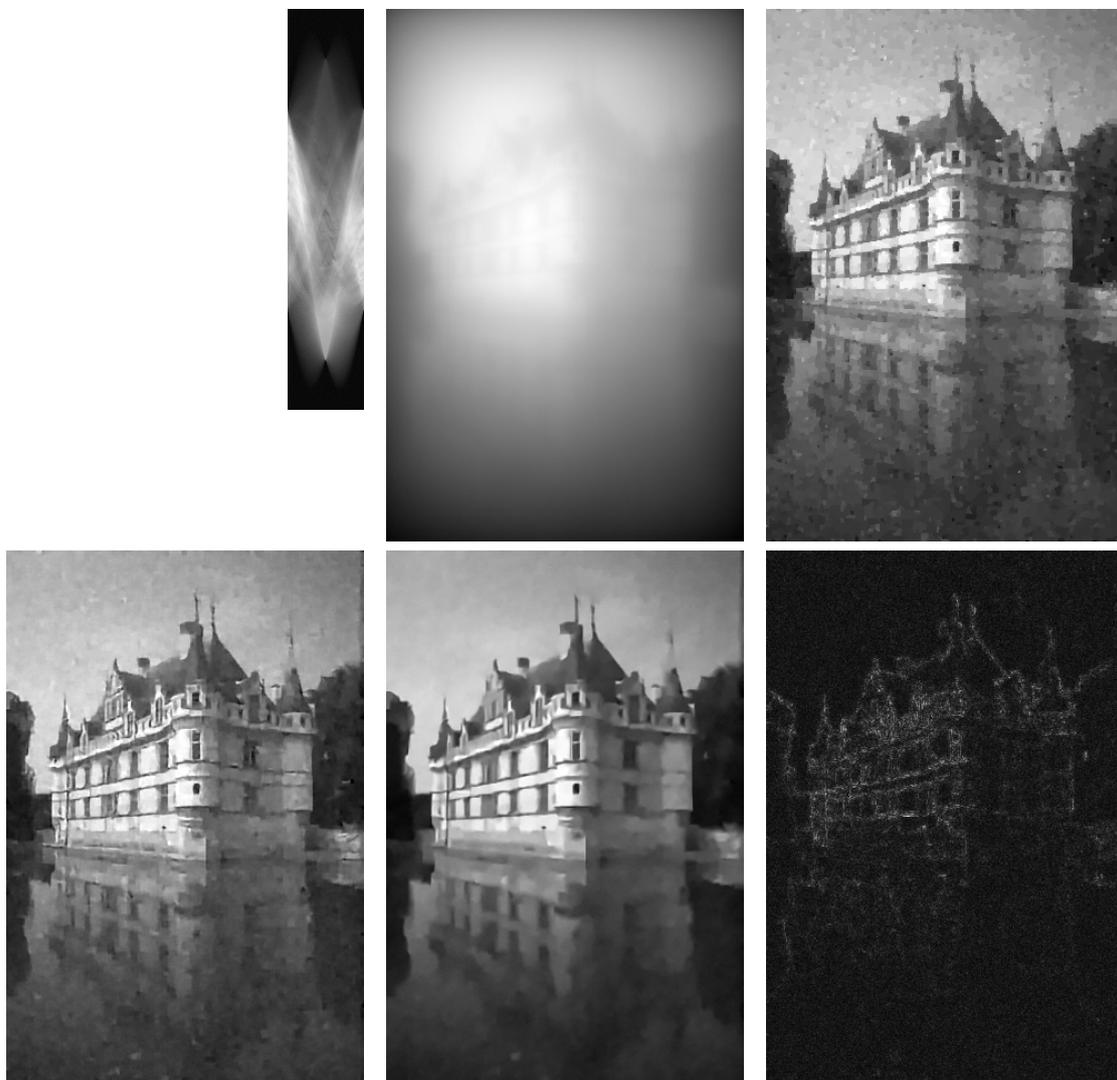


Figure 7: Qualitative results of reconstruction from Radon samples: Top row shows a visualization of the data, the naive reconstruction obtained by backprojection, and the reconstruction obtained through total variation regularization. The bottom row shows the reconstruction obtained through regularization with the bilevel model, the MMSE estimate obtained through the sampling of the posterior of the score-matching prior, and the corresponding pixel-wise marginal standard deviation.

Therefore, continued research into efficient sampling methods for high-dimensional probability densities remains critically important.

6 Acknowledgements

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/F100800.

References

- [1] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. “A Learning Algorithm for Boltzmann Machines”. In: *Cognitive Science* 9.1 (1985), pp. 147–169. ISSN: 0364-0213. DOI: [https://doi.org/10.1016/S0364-0213\(85\)80012-4](https://doi.org/10.1016/S0364-0213(85)80012-4).
- [2] Herbert Amann. *Ordinary differential equations: an introduction to nonlinear analysis*. Vol. 13. Walter de Gruyter, 2011.
- [3] Christophe Andrieu and Johannes Thoms. “A tutorial on adaptive MCMC”. In: *Statistics and computing* 18.4 (2008), pp. 343–373.
- [4] Vladimir I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer New York, 1989. ISBN: 9781475720631. DOI: 10.1007/978-1-4757-2063-1.
- [5] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer International Publishing, 2014. ISBN: 9783319002279. DOI: 10.1007/978-3-319-00227-9.
- [6] Mario Bertero and Patrizia Boccacci. *Introduction to Inverse Problems in Imaging*. CRC Press, Aug. 2020. ISBN: 9780367806941. DOI: 10.1201/9780367806941.
- [7] Michael Betancourt. “A Conceptual Introduction to Hamiltonian Monte Carlo”. In: *arXiv preprint arXiv:1701.02434* (2017).
- [8] Ander Biguri et al. “TIGRE: A MATLAB-GPU Toolbox for CBCT Image Reconstruction”. In: *Biomedical Physics & Engineering Express* 2.5 (Sept. 2016), p. 055010. ISSN: 2057-1976. DOI: 10.1088/2057-1976/2/5/055010.
- [9] Patrick Billingsley. *Probability and Measure*. en. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Blackwell, Feb. 2012.
- [10] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.
- [11] Vladimir I. Bogachev. *Measure Theory*. Vol. II. Springer, 2007.
- [12] Vladimir I. Bogachev, Michael Röckner, and Wilhelm Stannat. “Uniqueness of Solutions of Elliptic Equations and Uniqueness of Invariant Measures of Diffusions”. In: *Sbornik: Mathematics* 193.7 (Aug. 2002), p. 945. DOI: 10.1070/SM2002v193n07ABEH000665.
- [13] Jérôme Bolte, Edouard Pauwels, and Antonio Silveti-Falls. “Differentiating Nonsmooth Solutions to Parametric Monotone Inclusion Problems”. In: *SIAM Journal on Optimization* 34.1 (2024), pp. 71–97. DOI: 10.1137/22M1541630.
- [14] Nawaf Bou-Rabee and Jesús María Sanz-Serna. “Randomized Hamiltonian Monte Carlo”. In: *The Annals of Applied Probability* 27.4 (Aug. 2017). ISSN: 1050-5164. DOI: 10.1214/16-aap1255.

- [15] Kristian Bredies and Martin Holler. “Higher-Order Total Variation Approaches and Generalisations”. In: *Inverse Problems* 36.12 (2020), p. 123001. DOI: 10.1088/1361-6420/ab8f80.
- [16] Martin Burger and Stanley Osher. “Convergence Rates of Convex Variational Regularization”. In: *Inverse Problems* 20.5 (July 2004), pp. 1411–1421. ISSN: 1361-6420. DOI: 10.1088/0266-5611/20/5/005.
- [17] Martin Burger et al. “EM-TV Methods for Inverse Problems with Poisson Noise”. In: *Level Set and PDE Based Reconstruction Methods in Imaging: Cetraro, Italy 2008*, Editors: Martin Burger, Stanley Osher (2013), pp. 71–142.
- [18] Olivier Cappe, Simon J. Godsill, and Eric Moulines. “An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo”. In: *Proceedings of the IEEE* 95.5 (2007), pp. 899–924. DOI: 10.1109/JPROC.2007.893250.
- [19] Antonin Chambolle and Thomas Pock. “A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging”. In: *Journal of Mathematical Imaging and Vision* 40.1 (Dec. 2010), pp. 120–145. ISSN: 1573-7683. DOI: 10.1007/s10851-010-0251-1.
- [20] Joseph T. Chang and David Pollard. “Conditioning as Disintegration”. In: *Statistica Neerlandica* 51.3 (Nov. 1997), pp. 287–317. ISSN: 1467-9574. DOI: 10.1111/1467-9574.00056.
- [21] Omar Chehab et al. “Provable convergence and limitations of geometric tempering for Langevin dynamics”. In: *arXiv preprint arXiv:2410.09697* (2024).
- [22] Yunjin Chen and Thomas Pock. “Trainable Nonlinear Reaction Diffusion: A Flexible Framework for Fast and Effective Image Restoration”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1256–1272. DOI: 10.1109/TPAMI.2016.2596743.
- [23] Yunjin Chen et al. “Revisiting Loss-Specific Training of Filter-Based MRFs for Image Restoration”. In: *Pattern Recognition*. Ed. by Joachim Weickert, Matthias Hein, and Bernt Schiele. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 271–281. ISBN: 978-3-642-40602-7.
- [24] Xiang Cheng et al. “Underdamped Langevin MCMC: A Non-Asymptotic Analysis”. In: *Proc. of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 300–323.
- [25] Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.
- [26] Regev Cohen et al. “It Has Potential: Gradient-Driven Denoisers for Convergent Solutions to Inverse Problems”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 18152–18164.
- [27] Masoumeh Dashti et al. “MAP Estimators and Their Consistency in Bayesian Nonparametric Inverse Problems”. In: *Inverse Problems* 29.9 (2013), p. 095017. DOI: 10.1088/0266-5611/29/9/095017.
- [28] Yilun Du and Igor Mordatch. “Implicit Generation and Modeling with Energy Based Models”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Red Hook, NY, USA: Curran Associates, Inc., Nov. 6, 2019.
- [29] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. “Analysis of Langevin Monte Carlo via Convex Optimization”. In: *Journal of Machine Learning Research* 20.73 (2019), pp. 1–46.

- [30] Alain Durmus and Éric Moulines. “Nonasymptotic Convergence Analysis for the Unadjusted Langevin Algorithm”. In: *The Annals of Applied Probability* 27.3 (2017), pp. 1551–1587. DOI: 10.1214/16-AAP1238.
- [31] Alain Durmus, Eric Moulines, and Eero Saksman. “Irreducibility and Geometric Ergodicity of Hamiltonian Monte Carlo”. In: *The Annals of Statistics* 48.6 (2020), pp. 3545–3564. ISSN: 00905364, 21688966.
- [32] Alexander Effland et al. “Variational Networks: An Optimal Control Approach to Early Stopping Variational Methods for Image Restoration”. In: *Journal of Mathematical Imaging and Vision* 62.3 (Mar. 2020), pp. 396–416. ISSN: 1573-7683. DOI: 10.1007/s10851-019-00926-8.
- [33] Matthias J. Ehrhardt, Lorenz Kuger, and Carola-Bibiane Schönlieb. “Proximal Langevin Sampling with Inexact Proximal Mapping”. In: *SIAM Journal on Imaging Sciences* 17.3 (2024), pp. 1729–1760. DOI: 10.1137/23M1593565.
- [34] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Vol. 375. Springer Science & Business Media, 1996.
- [35] W. T. Freeman and Y. Weiss. “What Makes a Good Model of Natural Images?” In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos, CA, USA: IEEE Computer Society, June 2007, pp. 1–8.
- [36] Yoav Freund and David Haussler. “Unsupervised Learning of Distributions on Binary Vectors Using Two Layer Networks”. In: *Advances in Neural Information Processing Systems* 4 (1991).
- [37] Lorenz Fruehwirth and Andreas Habring. “Ergodicity of Langevin Dynamics and its Discretizations for Non-smooth Potentials”. In: *arXiv preprint arXiv:2411.12051* (2024).
- [38] Iosif Ilyich Gikhman and Skorokhod Anatolij Volodymyrovyc. *Introduction to the Theory of Random Processes*. W.B.Saunders, 1965.
- [39] Andreas Habring and Martin Holler. “A Generative Variational Model for Inverse Problems in Imaging”. In: *SIAM Journal on Mathematics of Data Science* 4.1 (2022), pp. 306–335. DOI: 10.1137/21M1414978.
- [40] Andreas Habring, Martin Holler, and Thomas Pock. “Subgradient Langevin Methods for Sampling from Nonsmooth Potentials”. In: *SIAM Journal on Mathematics of Data Science* 6.4 (2024), pp. 897–925. DOI: 10.1137/23M1591451.
- [41] Andreas Habring et al. “Diffusion at Absolute Zero: Langevin Sampling Using Successive Moreau Envelopes”. In: *arXiv preprint arXiv:2503.22258* (2025).
- [42] Kerstin Hammernik et al. “Learning a Variational Network for Reconstruction of Accelerated MRI Data”. In: *Magnetic Resonance in Medicine* 79.6 (Nov. 2017), pp. 3055–3071. ISSN: 1522-2594. DOI: 10.1002/mrm.26977.
- [43] Philip Hartman. *Ordinary differential equations*. SIAM, 2002.
- [44] Nicolas Heess, Christopher Williams, and Geoffrey Hinton. “Learning Generative Texture Models with Extended Fields-of-Experts.” In: Jan. 2009. DOI: 10.5244/C.23.115.
- [45] Tapio Helin and Martin Burger. “Maximum A-Posteriori Probability Estimates in Infinite-Dimensional Bayesian Inverse Problems”. In: *Inverse Problems* 31.8 (July 2015), p. 085009. ISSN: 1361-6420. DOI: 10.1088/0266-5611/31/8/085009.
- [46] Geoffrey E Hinton et al. “The ‘Wake-Sleep’ Algorithm for Unsupervised Neural Networks”. In: *Science* 268.5214 (1995), pp. 1158–1161. DOI: 10.1126/science.7761831.

- [47] Geoffrey E. Hinton. “Training Products of Experts by Minimizing Contrastive Divergence”. In: *Neural Computation* 14.8 (2002), pp. 1771–1800. DOI: 10.1162/089976602760128018.
- [48] Morris W. Hirsch, Stephen Smale, and Robert L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic press, 2013.
- [49] Matthew D Hoffman, Andrew Gelman, et al. “The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.” In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.
- [50] Bamdad Hosseini. “Well-posed Bayesian Inverse Problems with Infinitely Divisible and Heavy-Tailed Prior Measures”. In: *SIAM/ASA Journal on Uncertainty Quantification* 5.1 (2017), pp. 1024–1060. DOI: 10.1137/16M1096372.
- [51] Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. “Gradient Step Denoiser for convergent Plug-and-Play”. In: *International Conference on Learning Representations*. 2022.
- [52] Kaiyi Ji, Junjie Yang, and Yingbin Liang. “Bilevel Optimization: Convergence Analysis and Enhanced Design”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4882–4892.
- [53] Rafail Khasminskii. *Stochastic Stability of Differential Equations*. Springer, 2012. DOI: 10.1007/978-3-642-23280-0.
- [54] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015.
- [55] Erich Kobler et al. “Total Deep Variation: A Stable Regularization Method for Inverse Problems”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (2022), pp. 9163–9180. DOI: 10.1109/TPAMI.2021.3124086.
- [56] Muhamed Kuric et al. “The Gaussian Latent Machine: Efficient Prior and Posterior Sampling for Inverse Problems”. In: *arXiv preprint arXiv:2505.12836* (2025).
- [57] Matti Lassas, Eero Saksman, and Samuli Siltanen. “Discretization-Invariant Bayesian Inversion and Besov Space Priors”. In: *Inverse Problems and Imaging* 3.1 (2009), pp. 87–122. ISSN: 1930-8337. DOI: 10.3934/ipi.2009.3.87.
- [58] Jonas Latz. “Bayesian Inverse Problems are Usually Well-Posed”. In: *SIAM Review* 65.3 (2023), pp. 831–865. DOI: 10.1137/23M1556435.
- [59] Han Cheng Lie and TJ Sullivan. “Equivalence of Weak and Strong Modes of Measures on Topological Vector Spaces”. In: *Inverse Problems* 34.11 (2018), p. 115013. DOI: 10.1088/1361-6420/aadef2.
- [60] Qiang Liu and Dilin 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.
- [61] Samuel Livingstone et al. “On the Geometric Ergodicity of Hamiltonian Monte Carlo”. In: *Bernoulli* 25.4A (2019), pp. 3109–3138. DOI: 10.3150/18-BEJ1083.
- [62] Sebastian Lunz, Ozan Öktem, and Carola-Bibiane Schönlieb. “Adversarial Regularizers in Inverse Problems”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

- [63] Oren Mangoubi and Nisheeth K. Vishnoi. “Dimensionally Tight Bounds for Second-Order Hamiltonian Monte Carlo”. In: *Advances in Neural Information Processing Systems*. NIPS’18. Montréal, Canada: Curran Associates Inc., 2018, pp. 6030–6040.
- [64] Sheheryar Mehmood and Peter Ochs. “Automatic Differentiation of some First-Order Methods in Parametric Optimization”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1584–1594.
- [65] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.
- [66] Vishal Monga, Yuelong Li, and Yonina C. Eldar. “Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing”. In: *IEEE Signal Processing Magazine* 38.2 (2021), pp. 18–44. DOI: 10.1109/MSP.2020.3016905.
- [67] Subhadip Mukherjee et al. “Learned Convex Regularizers for Inverse Problems”. In: *arXiv:2008.02839* (2021).
- [68] Mahesh Chandra Makkamala et al. “Convex-Concave Backtracking for Inertial Bregman Proximal Gradient Algorithms in Nonconvex Optimization”. In: *SIAM Journal on Mathematics of Data Science* 2.3 (2020), pp. 658–682. DOI: 10.1137/19M1298007.
- [69] Dominik Narnhofer et al. “Bayesian Uncertainty Estimation of Learned Variational MRI Reconstruction”. In: *IEEE Transactions on Medical Imaging* 41.2 (2022), pp. 279–291. DOI: 10.1109/TMI.2021.3112040.
- [70] Dominik Narnhofer et al. “Posterior-Variance-Based Error Quantification for Inverse Problems in Imaging”. In: *SIAM Journal on Imaging Sciences* 17.1 (2024), pp. 301–333. DOI: 10.1137/23M1546129.
- [71] Radford M Neal. “Annealed importance sampling”. In: *Statistics and computing* 11.2 (2001), pp. 125–139.
- [72] Radford M. Neal. “MCMC Using Hamiltonian Dynamics”. In: *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011. DOI: 10.1201/b10905.
- [73] Erik Nijkamp et al. “Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.
- [74] Erik Nijkamp et al. “On the Anatomy of MCMC-Based Maximum Likelihood Learning of Energy-Based Models”. In: *Proc. of the AAAI Conference on Artificial Intelligence* 34.04 (Apr. 2020), pp. 5272–5280. DOI: 10.1609/aaai.v34i04.5973.
- [75] Peter Ochs et al. “Techniques for Gradient-Based Bilevel Optimization with Non-Smooth Lower Level Problems”. In: *Journal of Mathematical Imaging and Vision* 56 (2016), pp. 175–194. DOI: 10.1007/s10851-016-0663-7.
- [76] Bernt Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media, 2013.
- [77] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035.
- [78] Marcelo Pereyra. “Proximal Markov Chain Monte Carlo Algorithms”. In: *Statistics and Computing* 26 (2016), pp. 745–760. DOI: 10.1007/s11222-015-9567-4.
- [79] Marien Renaud et al. “From Stability of Langevin Diffusion to Convergence of Proximal MCMC for Non-log-Concave Sampling”. In: *arXiv preprint arXiv:2505.14177* (2025).

- [80] Christian P Robert, George Casella, and George Casella. *Monte Carlo Statistical Methods*. Vol. 2. Springer, 1999.
- [81] Gareth O. Roberts and Richard L. Tweedie. “Exponential Convergence of Langevin Distributions and Their Discrete Approximations”. In: *Bernoulli* 2.4 (1996), pp. 341–363.
- [82] Stefan Roth and Michael J. Black. “Fields of Experts”. In: *International Journal of Computer Vision* 82.2 (Jan. 2009), pp. 205–229. ISSN: 1573-1405. DOI: 10.1007/s11263-008-0197-6.
- [83] Leonid I Rudin, Stanley Osher, and Emad Fatemi. “Nonlinear Total Variation Based Noise Removal Algorithms”. In: *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268. DOI: 10.1016/0167-2789(92)90242-F.
- [84] David E. Rumelhart and James L. McClelland. “Parallel Distributed Processing”. In: Cambridge, MA: MIT Press, 1986. Chap. Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281.
- [85] Kegan G. G. Samuel and Marshall F. Tappen. “Learning optimized MAP estimates in continuously-valued MRF models”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 477–484. DOI: 10.1109/CVPR.2009.5206774.
- [86] Uwe Schmidt, Qi Gao, and Stefan Roth. “A Generative Perspective on MRFs in Low-Level Vision”. In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2010, pp. 1751–1758.
- [87] Zakhar Shumaylov et al. “Weakly Convex Regularisers for Inverse Problems: Convergence of Critical Points and Primal-Dual Optimisation”. In: *Forty-first International Conference on Machine Learning*. 2024.
- [88] Yang Song and Stefano Ermon. “Generative modeling by estimating gradients of the data distribution”. In: *Advances in neural information processing systems* 32 (2019).
- [89] Yang Song et al. “Score-based generative modeling through stochastic differential equations”. In: *arXiv preprint arXiv:2011.13456* (2020).
- [90] AM Stuart, Stephen Harris, and Masoumeh Dashti. “Besov Priors for Bayesian Inverse Problems”. In: *Inverse Problems and Imaging* 6.2 (2012), pp. 183–200. DOI: 10.3934/ipi.2012.6.183.
- [91] Andrew M Stuart. “Inverse Problems: A Bayesian Perspective”. In: *Acta numerica* 19 (2010), pp. 451–559. DOI: 10.1017/S0962492910000061.
- [92] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. “A Differential Equation for Modeling Nesterov’s Accelerated Gradient Method: Theory and Insights”. In: *Journal of Machine Learning Research* 17.1 (Jan. 2016), pp. 5312–5354. ISSN: 1532-4435.
- [93] Marshall F. Tappen. “Utilizing Variational Optimization to Learn Markov Random Fields”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383037.
- [94] Yee Whye Teh et al. “Energy-Based Models for Sparse Overcomplete Representations”. In: *J. Mach. Learn. Res.* 4.null (Dec. 2003), pp. 1235–1260. ISSN: 1532-4435.
- [95] Cédric Villani. *Topics in Optimal Transportation*. Vol. 58. American Mathematical Soc., 2021.
- [96] Pascal Vincent. “A Connection between Score Matching and Denoising Autoencoders”. In: *Neural Computation* 23.7 (2011), pp. 1661–1674. DOI: 10.1162/NECO_a_00142.

- [97] Martin Zach. *Generative Models as Regularizers for Inverse Problems in Imaging*. 2024. DOI: 10.3217/6N86E-QQK03.
- [98] Martin Zach, Florian Knoll, and Thomas Pock. “Stable Deep MRI Reconstruction Using Generative Priors”. In: *IEEE Transactions on Medical Imaging* 42.12 (2023), pp. 3817–3832. DOI: 10.1109/TMI.2023.3311345.
- [99] Martin Zach, Erich Kobler, and Thomas Pock. “Computed Tomography Reconstruction Using Generative Energy-Based Priors”. English. In: *Proc. of the OAGM Workshop 2021*. Ed. by Markus Seidl, Matthias Zeppelzauer, and Peter M. Roth. Verlag der Technischen Universität Graz, Dec. 2021, pp. 52–58. DOI: 10.3217/978-3-85125-869-1-09.
- [100] Martin Zach et al. “Product of Gaussian Mixture Diffusion Models”. In: *Journal of Mathematical Imaging and Vision* 66.4 (Mar. 2024), pp. 504–528. ISSN: 1573-7683. DOI: 10.1007/s10851-024-01180-3.
- [101] Yasi Zhang and Oscar Leong. “Learning Difference-of-Convex Regularizers for Inverse Problems: A Flexible Framework with Theoretical Guarantees”. In: *arXiv preprint arXiv:2502.00240* (2025).
- [102] Song Chun Zhu and David Mumford. “Prior Learning and Gibbs Reaction-Diffusion”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.11 (1997), pp. 1236–1250. DOI: 10.1109/34.632983.
- [103] Song Chun Zhu, Ying Nian Wu, and David Mumford. “Minimax Entropy Principle and Its Application to Texture Modeling”. In: *Neural Computation* 9.8 (1997), pp. 1627–1660. DOI: 10.1162/neco.1997.9.8.1627.
- [104] Song Chun Zhu, Yingnian Wu, and David Mumford. “Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling”. In: *International Journal of Computer Vision* 27.2 (1998), pp. 107–126. DOI: 10.1023/A:1007925832420.
- [105] Nicolas Zucchet and João Sacramento. “Beyond Backpropagation: Bilevel Optimization Through Implicit Differentiation and Equilibrium Propagation”. In: *Neural Computation* 34.12 (2022), pp. 2309–2346. DOI: 10.1162/neco_a_01547.