Langevin-Based Expectation-Maximization for Noise and Posterior Estimation in Bayesian Inverse Problems

Masterarbeit

Christoph Marcus Jankowsky

Betreuung: Prof. Dr. Claudia Schillings
Zweitgutachten: Jun. Prof. Dr. Ana Djurdjevac

Freie Universität Berlin
FB Mathematik und Informatik

Berlin, 2. November 2024

Eigenständigkeitserklärung

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Masterarbeit selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Arbeit wurde in gleicher oder ähnlicher Form noch bei keiner anderen Universität als Prüfungsleistung eingereicht.

2. November 2024

(Datum) (Unterschrift)

Table of Contents

1. Introduction

The issue of obtaining meaningful data from observations is prevalent in many areas of research and industry and particularly critical in fields where direct measurement is either impossible or impractical, necessitating the use of indirect, often noise-corrupted measurements. Examples such as reconstructing internal structures from MRI Scans or X-Ray images [Marschall et al., 2023, Kolehmainen et al., 2007], inferring geological properties from seismic measurements [Rawlinson et al., 2010] or dating methods used in archaeology [Buck et al., 1996] demonstrate the wide variety of fields where this kind of problem occurs. Measured quantities are generated from hidden parameters through a possibly nonlinear forward model which, in practice, is often described as a partial differential equation that stems from the laws governing the physical system that is being investigated. These inverse problems are generally ill-posed, meaning that solutions might not be unique if they exist at all. In order to address this a regularized problem is considered. Through modifications such as combining several optimization objectives, well-posedness is introduced. A natural and powerful regularization framework for addressing these challenges is found in the setting of so-called *Bayesian inverse problems* (BIPs). Here, parameters of interest and data are regarded as being random variables, where a first a priori guess of the parameter distribution is then updated via Bayes' rule in order to obtain a distribution of the parameters conditioned on the measurement observations. Once the target distribution has been identified, samples are generated which then correspond to solution parameters of the inverse problem. Sampling from these posterior distributions is not a trivial task, especially when the observational noise is also unknown. This thesis will discuss some approaches, namely expectation-maximization algorithms and Langevin samplers, and then propose a combination of the two in order to obtain a robust sampling method for BIPs with the added benefit of an estimation of the measurement noise parameters.

1.1. Related Work

Significant research has been conducted on sampling in the context of BIPs This thesis mainly builds on the following works, each of which will be presented in greater detail in Chapter 3.

[Hagemann et al., 2024] proposes the use of an EM-Algorithm, the details of which will be discussed in Section 3.1, for jointly estimating the posterior distribution and noise parameters of a given BIP. The authors make use of *normalizing flows* (NF), namely neural networks, in order to learn the posterior within the E-step and propose an analytical

M-step for their noise model.

Using Langevin dynamics [Langevin, 1908] as a particle propagator is a common technique for solving BIPs [Gelman et al., 1997, Roberts and Rosenthal, 1998, Bédard, 2007]. Interacting particle Langevin dynamics introduces interactions between particles which can greatly improve convergence of Langevin samplers [Matthews et al., 2016]. A class of affine invariant interacting particle Langevin samplers, called *ALDI*, was introduced in [Garbuno-Inigo et al., 2020]. In addition to the gradient-based approach, a gradient-free alternative is presented that can reduce computational cost. These ensemble samplers have been made more efficient with the introduction of *LIDL* in [Eigel et al., 2022]. Preconditioning and sample enrichment methods are proposed, that can be implemented for a variety of Langevin samplers and improve results especially for complex posterior distributions by accelerating convergence and reducing the number of required forward calls. An overview of Langevin samplers is given in Section 3.2

1.2. Contribution

This thesis demonstrates the advantage of using interacting particle Langevin samplers within an EM-algorithm in order to generate samples from an unknown posterior in a Bayesian inverse problem where noise parameters are not known a priori and are estimated within the algorithm. Some preliminaries on Bayesian inversion will be presented in Chapter 2. Chapter 3 gives a detailed explanation and analysis of the EM-algorithm, Langevin dynamics and samplers such as the *ALDI* - and *LIDL*-methods. These methods are then combined in order to obtain a robust method for the generation of posterior samples with simultaneous estimation of measurement noise covariance for Gaussian noise models. Numerical experiments are presented and evaluated in Chapter 4. Here, the success of the algorithm and the effect of hyperparameter optimization on a variety of problems is studied. In particular, the efficacy of the algorithm on a PDE-constrained forward problem is studied, which serves as a proxy for real-world applications, since these often rely on PDEs. Finally, a summary and notes on possible further research are given in Chapter 5.

2. Preliminaries

2.1. Inverse Problems

Inverse problems arise naturally in many applications as they offer a framework for gaining insight into the nature of latent parameters based on indirect measurements.

Consider the problem of finding $u \in \mathbb{R}^D$ such that

$$y = \mathcal{G}(u)$$

for a given $y \in \mathbb{R}^K$. Here, y can be interpreted as observed data in some experiment and u as the unknown latent variable. Data is generated via a forward map $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^K$. In order to model real-world data, we have to include some observational noise $\eta \in \mathbb{R}^K$, since measurements are never perfectly accurate. This gives rise to the following inverse problem:

$$\text{Find } u \in \mathbb{R}^D \text{ such that } y = \mathcal{G}(u) + \eta. \tag{2.1}$$

An important concept in the definition of such problems is that of *well-posedness* which has been introduced in [Hadamard, 1902].

Definition 2.1 (Well-posed and Ill-posed Problem)**.** *A problem is called well-posed if all of the following conditions are satisfied.*

1. *(Existence): There exists at least one solution.*
2. *(Uniqueness): There is at most one solution.*
3. *(Stability): The solution depends continuously on the data.*

If any of the conditions are not satisfied, the problem is called ill-posed.

Example 2.2 (Linear System)**.** *Consider the inverse problem of finding the latent variable $u \in \mathbb{R}^D$ such that*

$$y = \mathcal{G}(u) = Gu$$

for a given $y \in \mathbb{R}^K$ where $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^K$ is a linear map defined by a matrix $G \in \mathbb{R}^{K \times D}$. For $D < K$ the system is underdetermined and thus the problem is ill-posed, since multiple solutions may exist. △

A more involved example is the following that arises in digital image processing.

Example 2.3 (Image deblurring)**.** *Consider a black and white image represented as a matrix* $U \in [0, 255]^{h \times w} \subset \mathbb{N}^{h \times w}$ *for* $w, h \in \mathbb{N}$*. Each entry represents the brightness of a pixel ranging from* 0 *to* 255*. A blurring kernel* $\mathcal{G}$ *is applied and the result* $\mathcal{G}(U)$ *is the blurred image. Additionally, some noise* $\eta \in [0, 255]^{h \times w}$*, independent of* U*, may be applied. The result*

$$Y = \mathcal{G}(U) + \eta$$

is the blurred and noisy image and the task is to reconstruct U *from* Y*.*

Figure 2.1.: From left to right: Original image U, blurred image $\mathcal{G}(U)$ and blurred and noisy image $Y = \mathcal{G}(U) + \eta$.

△

Note that this problem is *ill-posed*, as small pixel variations in the latent image U can lead to the same resulting image Y after application of the blurring kernel $\mathcal{G}$ and the added noise η. This example also illustrates that the problem definition does not imply a specific algorithm, as many have been proposed for the task of image deblurring [Dong et al., 2011, Chen et al., 2021, Zamir et al., 2021]. Most recent advancements rely on some form of neural network architecture. However, this problem has been studied with many different approaches, many of which predate modern machine learning tools [Fergus et al., 2006, Jia, 2007].

In general, we assume that the forward problem is *well-posed*. However, as has been illustrated in Example 2.2 and Example 2.3, the same cannot be said for the inverse problem. One of the main challenges in solving inverse problems is how to deal with the ill-posedness of the problem. In order to address this, regularization methods are introduced. A deterministic regularization method that is commonly used is called *Thikhonov regularization*. Here, the objective is stated as

$$\min_u \Big(||\mathcal{G}(u) - y||^2 + \lambda ||u||^2 \Big),$$

where the hyperparameter λ has to be chosen in order balance the size of the norm of u with the accuracy of the solution. However, in this thesis we will focus on a stochastic regularization method based on *Bayes' theorem*.

2.2. Bayesian Inverse Problems

As illustrated by example 2.3, the inverse problem is generally *ill-posed*, meaning that solutions u might not be unique if they exists at all. A way to deal with this is via *Bayesian inversion*. In the context of *Bayesian inverse problems* , our primary goal is to approximate the unknown quantity $u \in \mathbb{R}^D$ based on observations $y \in \mathbb{R}^K$, which are both considered to be random variables and which are related through a possibly nonlinear and noisy process. Specifically, we consider the model

$$y = \mathcal{G}(u) + \eta,$$

where $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^K$ is called the forward operator, and η represents noise that is independent of u and follows a possibly unknown distribution.

To estimate u given an observation $y = \tilde{y} \in \mathbb{R}^K$, we utilize Bayes' theorem to derive the posterior distribution of u given $y = \tilde{y}$. The prior distribution of u, denoted as μ_{prior}, is assumed to have a density π_{prior} with respect to the Lebesgue measure. The observational noise η is distributed according to a measure ν_0 with density ρ.

Given a realization $\tilde{u} \in \mathbb{R}^D$ of the random variable u, the corresponding observation y is distributed according to a shifted version of ν_0, denoted as $\nu_{\tilde{u}}$, with the density function

$$\rho_{\tilde{u}}(y) := \rho(y|\tilde{u}) = \rho(y - \mathcal{G}(\tilde{u})).$$

This represents the likelihood of observing $y \in \mathbb{R}^K$ given $u = \tilde{u} \in \mathbb{R}^D$. Consequently, the joint distribution of (u, y) in $\mathbb{R}^D \times \mathbb{R}^K$ has a density given by

$$\phi(u, y) = \rho_u(y)\pi_{\text{prior}}(u) = \rho(y - \mathcal{G}(u))\pi_{\text{prior}}(u).$$

Bayes' theorem allows us to update our prior beliefs about u after observing $y = \tilde{y}$ by computing the posterior distribution. The posterior density $\pi_{\text{post.}}(u) = \pi(u \mid \tilde{y})$ is given by

$$\pi_{\text{post.}}(u) = \frac{1}{Z}\rho(\tilde{y} - \mathcal{G}(u))\pi_{\text{prior}}(u),$$

where the normalization constant Z ensures that $\pi_{\text{post.}}(u)$ integrates to one and is defined as

$$Z = \int_{\mathbb{R}^D} \rho(\tilde{y} - \mathcal{G}(u))\pi_{\text{prior}}(u)\, \mathrm{d}u.$$

To express this in a different manner, let $\mu_{\text{post.}}$ and μ_{prior} denote the posterior and prior measures on $\mathbb{R}^D$, respectively, with densities $\pi_{\text{post.}}$ and π_{prior}. Bayes' theorem can then be formulated as

$$\frac{\mathrm{d}\mu_{\text{post.}}}{\mathrm{d}\mu_{\text{prior}}}(u) = \frac{1}{Z}\rho(\tilde{y} - \mathcal{G}(u)).$$

The posterior measure $\mu_{\text{post.}}$ is absolutely continuous with respect to the prior measure μ_{prior}, with the Radon-Nikodym derivative proportional to the likelihood function.

This formulation demonstrates the Bayesian approach to inverse problems, providing a method for updating our estimates of the unknown parameter u as new data becomes available. The posterior distribution models the trade-off between prior knowledge and the information provided by the observed data, allowing for a probabilistic interpretation of the solution to the inverse problem. Instead of obtaining a single value for the unknown parameter u, in *Bayesian inverse problems* we obtain a distribution $\mu_{\text{post.}}$ of the parameter given the data with the added benefit of encoding uncertainty of the solution.

2.3. The Model Problem

In this thesis we will discuss the methods for solving the *Bayesian inverse problem* as outlined in section 2.2. This section aims to define the model problem that will be discussed in the remainder of the thesis. We consider the relation

$$y = \mathcal{G}(u) + \eta$$

for noise $\eta \in \mathbb{R}^K$ and a forward map $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^K$. Multiple measurements $y_1, \cdots, y_N \in \mathbb{R}^K$, $N \in \mathbb{N}$, can be taken.

The observational noise η is commonly assumed to be distributed according to a centered Gaussian distribution $\eta \sim \mathcal{N}(\mathbf{0}_K, \Sigma)$ with positive-definite covariance matrix $\Sigma \in \mathbb{R}^{K\times K}$ and zero-mean vector $\mathbf{0}_K = (0, \dots, 0) \in \mathbb{R}^K$. In this case, its density ρ can be expressed as

$$\begin{aligned}\rho(y) &= \big((2\pi)^K \det(\Sigma)\big)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} y^\intercal \Sigma^{-1} y \right) \\ &= \big((2\pi)^K \det(\Sigma)\big)^{-\frac{1}{2}} \exp\big(-L(y)\big),\end{aligned}$$

with $L(y) := \frac{1}{2}|y|_\Sigma^2$ where $|\cdot|$ denotes the Euclidean norm and $|\cdot|_\Sigma := |\Sigma^{-\frac{1}{2}} \cdot|$. This allows for a more explicit formulation of the posterior density of the unknown parameter u for data y as

$$\begin{aligned}\pi_{\text{post.}}(u) &= \frac{1}{Z}\rho(y - \mathcal{G}(u))\pi_{\text{prior}}(u) \\ &= \frac{1}{Z}\big((2\pi)^K \det(\Sigma)\big)^{-\frac{1}{2}} \exp\big(-L(y - \mathcal{G}(u))\big)\pi_{\text{prior}}(u) \\ &\propto \exp\big(-L(y - \mathcal{G}(u))\big)\pi_{\text{prior}}(u) \\ &= \exp\big(-\Phi(u)\big), \end{aligned} \tag{2.2}$$

where

$$\Phi(u) = L(y - \mathcal{G}(u)) - \log \pi_{\text{prior}}(u) \tag{2.3}$$

is called the *potential*.

The prior distribution μ_{prior} is generally assumed to be Gaussian with mean $m_{\text{prior}} \in \mathbb{R}^D$ and covariance $\mathcal{C}_{\text{prior}} \in \mathbb{R}^{d\times d}$. Its density π_{prior} can then be expressed explicitly as

$$\pi_{\text{prior}}(u) \propto \exp\left(-\frac{1}{2}(u - m_{\text{prior}})^\intercal {\mathcal{C}_{\text{prior}}}^{-1}(u - m_{\text{prior}})\right). \tag{2.4}$$

This allows for a concrete representation of the potential eq. (2.3) as

$$\begin{aligned}\Phi(u) &= L(y - \mathcal{G}(u)) - \log \pi_{\text{prior}}(u)\\ &= \frac{1}{2}(y - \mathcal{G}(u))^\intercal \Sigma^{-1}(y - \mathcal{G}(u)) + \frac{1}{2}(u - m_{\text{prior}})^\intercal {\mathcal{C}_{\text{prior}}}^{-1}(u - m_{\text{prior}}).\end{aligned}$$

3. Theoretical Aspects

3.1. The Expectation-Maximization Algorithm

Consider the *inverse problem* described in Equation (2.1). Here, data y is generated from hidden parameters u via a forward map $\mathcal{G}$ and additional noise η. The usual convention is to work with centered Gaussian noise $\eta \sim \mathcal{N}(0, \Sigma)$ for some positive definite covariance matrix $\Sigma \in \mathbb{R}^{K \times K}$. Many methods, such as the Langevin samplers that will be described later, rely on the knowledge of the parameters of this noise model, however in practice, this knowledge cannot always be guaranteed due to the nature of the experiments. Thus, methods for estimating these parameters are of interest.

A common approach for this lies in *maximum likelihood estimation* (MLE).

Definition 3.1 (Likelihood function)**.** *Let $\{Y_\theta\}$ be a family of random variables taking values in $\mathbb{R}^K$ with probability density functions p_θ for parameters $\theta \in \Theta$. The likelihood function of a parameter θ for i.i.d. realizations $y_1, \dots, y_N \in \mathbb{R}^K$, $N \in \mathbb{N}$, sampled from Y_{θ^*} for some arbitrary but fixed and (possibly) unknown parameter $\theta^* \in \Theta$ is given by*

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p_\theta(y_i).$$

Alternatively, the log-likelihood

$$\begin{aligned} \log \mathcal{L}(\theta) &= \log \left(\prod_{i=1}^{N} p_\theta(y_i) \right) \\ &= \sum_{i=1}^{N} \log p_\theta(y^{(i)}) \end{aligned}$$

may be considered instead.

In other words the likelihood function $\mathcal{L}(\theta)$ describes how likely the data y_i, $i = 1, \dots, N$ is generated by the random variable Y_θ for any $\theta \in \Theta$. Now the parameter of interest θ^* may be estimated based on the data y_i, $i = 1, \dots, N$ by maximizing the (log-)likelihood function.

Definition 3.2 (Maximum log-likelihood estimate)**.** *Within the setup of Definition 3.1, the maximum log-likelihood estimate of a parameter θ^* given realizations $y_1, \dots, y_N \in \mathbb{R}^K$, $N \in \mathbb{N}$, is given by*

$$\arg\max_\theta \ \log \mathcal{L}(\theta) = \arg\max_\theta \ \sum_{i=1}^{N} \log p_\theta(y_i).$$

The *expectation-maximization* (EM) algorithm was first introduced in [Dempster et al., 1977] as a method for maximum-likelihood (ML) estimation. In the following, we assume $u \in \mathbb{R}^D$. From Bayes' theorem, we have

$$p_\theta(u|y) = \frac{p_\theta(y|u)\ p_\theta(u)}{p_\theta(y)}, \tag{3.1}$$

where $p_\theta(y|u)$ is the *likelihood*, $p_\theta(u)$ is the *prior* on u and $p_\theta(y)$ is the *marginal* density of y. We assume that the prior distribution of u is independent of θ and write $p(u) = p_\theta(u)$. Rearranging for the marginal density of y and introducing an auxiliary approachable density q on u yields

$$\begin{aligned} p_\theta(y) &= \frac{p_\theta(y|u)\ p(u)}{p_\theta(u|y)} \\ &= \frac{p_\theta(y|u)\ p(u)}{q(u)} \frac{q(u)}{p_\theta(u|y)}. \end{aligned}$$

Now, by taking logarithms we obtain

$$\begin{aligned} \log p_\theta(y) &= \log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)} \frac{q(u)}{p_\theta(u|y)}\right) \\ &= \log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)}\right) + \log\left(\frac{q(u)}{p_\theta(u|y)}\right). \end{aligned}$$

Finally, we compute the expectation with respect to q and get

$$\begin{aligned} \mathbb{E}_q[\log p_\theta(y)] &= \mathbb{E}_q\left[\log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)}\right) + \log\left(\frac{q(u)}{p_\theta(u|y)}\right)\right] \\ &= \mathbb{E}_q\left[\log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)}\right)\right] + \mathbb{E}_{q(u)}\left[\log\left(\frac{q(u)}{p_\theta(u|y)}\right)\right] \\ &= \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)}\right) \mathrm{d}u + \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{q(u)}{p_\theta(u|y)}\right) \mathrm{d}u. \end{aligned} \tag{3.2}$$

For the continuation of the construction of the EM-algorithm, we first need an important lemma which is an elementary result and thus will be presented without proof.

Lemma 3.3 (Jensen's inequality)**.** *Let X be a $\mathbb{R}$-valued random variable such that $\mathbb{E}[X]$ exists. Further, let $f : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\mathbb{E}[f(X)]$ also exists. Then the following holds*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

For a concave f, the reversed statement

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$$

holds. Strict convexity/concavity implies strict inequalities.

Now using the concavity of the logarithm and Jensen's inequality we can relate Equation (3.2) to $\log\big(p_\theta(y)\big)$.

$$\begin{aligned}
\log\big(p_\theta(y)\big) &= \log\big(\mathbb{E}_q[p_\theta(y)]\big) \\
&\geq \mathbb{E}_q[\log\big(p_\theta(y)\big)] \\
&= \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)}\right) \mathrm{d}u + \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{q(u)}{p_\theta(u|y)}\right) \mathrm{d}u. \qquad (3.3)
\end{aligned}$$

Definition 3.4 (Kullback–Leibler (KL) divergence)**.** *The Kullback-Leibler divergence of two probability measures P, Q with densities p and q such that $q(x) > 0$ in $\mathbb{R}^D$, respectively, is defined as*

$$KL(P, Q) := \int_{\mathbb{R}^D} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x. \qquad (3.4)$$

Lemma 3.5. *The KL-divergence of two distributions is always non-negative.*

Proof. Let P, Q be probability measures with densities p and q, respectively. Note that $\log x \leq x - 1$ for all $x > 0$. Further note that since P and Q are probability measures it holds $p(x) \geq 0$ for all x, $q(x) \geq 0$ for all x and $\int_{\mathbb{R}^D} p(x)\ \mathrm{d}x = \int_{\mathbb{R}^D} q(x)\ \mathrm{d}x = 1$. Now, we

have

$$\begin{aligned}
-\mathrm{KL}(P,Q) &= -\int_{\mathbb{R}^D} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \\
&= \int_{\mathbb{R}^D} p(x)\left(-\log\left(\frac{p(x)}{q(x)}\right)\right) \mathrm{d}x \\
&= \int_{\mathbb{R}^D} p(x) \log\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x \\
&\leq \int_{\mathbb{R}^D} p(x)\left(\frac{q(x)}{p(x)} - 1\right) \mathrm{d}x \\
&= \int_{\mathbb{R}^D} q(x)\ \mathrm{d}x - \int_{\mathbb{R}^D} p(x)\ \mathrm{d}x \\
&= 0.
\end{aligned}$$

From $-\mathrm{KL}(P,Q) \leq 0$ it follows directly that $\mathrm{KL}(P,Q) \geq 0$. □

Now recall Equation (3.3) from above:

$$\log p_\theta(y) \geq \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)}\right) \mathrm{d}u + \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{q(u)}{p_\theta(u|y)}\right) \mathrm{d}u.$$

Note that the right summand is equal to the *KL-divergence* of $q(u)$ and $p_\theta(u|y)$ and is thus always non-negative. Therefore, it can be omitted for the task of optimization, yielding

$$\begin{aligned}
\log p_\theta(y) &\geq \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)}\right) \mathrm{d}u + \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{q(u)}{p_\theta(u|y)}\right) \mathrm{d}u \\
&= \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)}\right) \mathrm{d}u + \mathrm{KL}(q(u), p_\theta(u|y)) \\
&\geq \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{p_\theta(y|u)\ p(u)}{q(u)}\right) \mathrm{d}u. \qquad (3.5)
\end{aligned}$$

Since the task at hand is to maximize the log-likelihood, one can see that this can be achieved to some extent by maximizing the right hand side of Equation (3.5). The EM-algorithm does this via a two-step process. First, initial estimates of the noise parameter $\theta = \theta^{(0)}$ is established. Using this initial estimate, the so-called *E-step* is performed.

$$\text{E-step:}\quad q^{(r+1)}(u) = \arg\max_q \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{p_{\theta^{(r)}}(y|u)\ p_{\theta^{(r)}}(u)}{q(u)}\right) \mathrm{d}u. \qquad (3.6)$$

Then the *M-step* is performed to update the noise parameter.

$$\text{M-step:}\quad \theta^{(r+1)} = \arg\max_\theta \int_{\mathbb{R}^D} q^{(r+1)}(u)\ \log\left(\frac{p_\theta(y|u)\ p(u)}{q^{(r+1)}(u)}\right) \mathrm{d}u. \qquad (3.7)$$

E-step

The *E-step* of the EM-algorithm can be understood as an estimation of the posterior distribution of the parameter u given data y and the current estimate of the noise parameter $\theta \approx \theta^{(r)}$. This can be seen when rearranging Equation (3.2).

$$\begin{aligned}\log p_{\theta^{(r)}}(y) - \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{p_{\theta^{(r)}}(y|u)\ p_{\theta^{(r)}}(u)}{q(u)}\right) \mathrm{d}u &= \int_{\mathbb{R}^D} q(u)\ \log\left(\frac{q(u)}{p_{\theta^{(r)}}(u|y)}\right) \mathrm{d}u \\ &= \mathrm{KL}(q(u), p_{\theta^{(r)}}(u|y)).\end{aligned}$$

This difference is minimal if and only if $q(u)$ and $p_{\theta^{(r)}}(u|y)$ are equal, which implies that the optimal result of the E-step is the posterior distribution of u given y and the current noise estimate $\theta^{(r)}$.

There are many different approaches to estimating the posterior distribution of a quantity in the setting of Bayesian inverse problems. [Hagemann et al., 2024] propose the use of *conditional normalizing flows* in the form of neural networks, however other approaches may also be used. The use of *interacting particle Langevin samplers* will be discussed in detail in Section 3.3.

M-step

For solving the *M-step* it is important to be aware of the noise model that is decided upon in the problem setup. We will use additive centered Gaussian noise $\eta \sim \mathcal{N}(0, \Sigma)$, such that the noise term is described in totality by the noise parameter $\theta = \Sigma$.

Recalling the definition of the *M-step* from Equation (3.7) we may decompose the right hand side.

$$\begin{aligned}\theta^{(r+1)} &= \arg\max_{\theta} \int_{\mathbb{R}^D} q^{(r+1)}(u)\ \log\left(\frac{p_\theta(y|u)\ p(u)}{q^{(r+1)}(u)}\right) \mathrm{d}u \\ &= \arg\max_{\theta}\left(\int_{\mathbb{R}^D} q^{(r+1)}(u)\ \log\big(p_\theta(y|u)\ p(u)\big)\ \mathrm{d}u - \int_{\mathbb{R}^D} q^{(r+1)}(u)\ \log\big(q^{(r+1)}(u)\big)\ \mathrm{d}u\right) \\ &= \arg\max_{\theta}\Bigg(\mathbb{E}_{q^{(r+1)}}[\log\big(p_\theta(y|u)\ p(u)\big)] - \underbrace{\mathbb{E}_{q^{(r+1)}}[\log\big(q^{(r+1)}(u)\big)]}_{\text{independent of } \theta}\Bigg).\end{aligned}$$

As can be seen immediately, it suffices to maximize the first summand of the right hand side in the *M-step*. Thus our simplified *M-step* is given by

$$\theta^{(r+1)} = \arg\max_{\theta} \mathbb{E}_{q^{(r+1)}}[\log\big(p_\theta(y|u)\ p(u)\big)], \tag{3.8}$$

and for N measurements we have

$$\theta^{(r+1)} = \arg\max_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{q^{(r+1)}}[\log\big(p_\theta(y_i|u)\ p(u)\big)]. \tag{3.9}$$

The discussion from above can be described in a concise manner as an algorithm.

Algorithm 3.1.1 Expectation-Maximization (EM) Algorithm

Require: Initial parameters $\theta^{(0)}$
Ensure: Estimated parameters θ^*
1: Initialize $\theta \leftarrow \theta^{(0)}$
2: **repeat**
3: **E-step:** Compute the posterior distribution $q^{(r+1)}$ given the current noise estimate $\theta^{(r)}$.
4: **M-step:** Update the noise estimate $\theta \leftarrow \theta^{(r+1)}$ using $q^{(r+1)}$.
5: Update $t \leftarrow t + 1$
6: **until** Convergence of θ

Remark. *Note that as a gradient-ascent algorithm, the EM-algorithm does not reduce the log-likelihood* $\log p_\theta(y)$ *at each iteration [Wu, 1983].* △

3.2. Langevin Samplers

3.2.1. Langevin Dynamics

Paul Langevin first introduced Langevin dynamics in [Langevin, 1908] as an alternative analysis of *Brownian motion*, to *Albert Einsteins* previous description in [Einstein, 1905]. The discovery of Brownian motion, the erratic and seemingly random movement of suspended particles, is commonly attributed to *Robert Brown* who published his observations of the movement of pollen suspended in water in [Brown, 1828].

The dynamics of a particle of mass m under Langevin dynamics are described by the Langevin equation

$$\underbrace{m\frac{\mathrm{d}v}{\mathrm{d}t}}_{\text{force acting on the particle}} = \underbrace{-\lambda v}_{\text{deterministic force}} + \underbrace{\eta(t)}_{\text{random force}},$$

where v is the velocity of the particle and λ is a damping coefficient. The force acting on the particle is a combination of a viscous force, which is proportional to the particle's velocity (according to Stokes' law), and a (time-dependent) noise term $\eta(t)$, which represents the effect of collisions with the fluid's molecules. The noise term $\eta(t)$ follows a Gaussian probability distribution.

Observe that

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}^2u}{\mathrm{d}t^2},$$

where u denotes the position of the particle. We may introduce a potential $\Phi(u)$ with

the property

$$\nabla\Phi(u) = m\frac{\mathrm{d}^2u}{\mathrm{d}t^2} = -\lambda v + \eta(t).$$

Rearranging yields a differential equation describing the evolution of the position u of the particle.

$$\lambda v = -\nabla\Phi(u) + \eta(t),$$

which can be expressed as a *stochastic differential equation* (SDE)

$$\mathrm{d}u_t = \underbrace{-\nabla\Phi(u_t)}_{\text{drift term}}\mathrm{d}t + \underbrace{\Gamma}_{\text{diffusion term}}\mathrm{d}W_t, \tag{3.10}$$

where W_t denotes the standard Wiener process and u_t denotes the position of the particle at time $t \geq 0$. There exists a unique invariant measure μ_∞ with density π_∞ for this process, with the property that if u_0 is distributed according to π_∞, then so is u_t for any $t \geq 0$. In order to sample from a target distribution μ, one may set the potential Φ such that $\mu_\infty = \mu$.

3.2.2. Langevin Samplers

Let $\pi_t(u)$ denote the probability density of the random variable u_t in Equation (3.10). The Fokker-Planck equation of the process is given by

$$\frac{\mathrm{d}\pi_t(u)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{\mathrm{d}\Phi(u)}{\mathrm{d}u}\pi_t(u)\right] + \frac{\mathrm{d}^2\pi_t(u)}{\mathrm{d}u^2} \tag{3.11}$$

and describes the time evolution of the density π_t. In order to obtain an invariant density π_∞ under this process, the condition

$$\frac{\mathrm{d}\pi_\infty(u)}{\mathrm{d}t} = 0$$

has to be fulfilled. Together with Equation (3.11), this implies

$$\begin{aligned}
0 &= \frac{\mathrm{d}\pi_\infty(u)}{\mathrm{d}t} \\
&= \frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{\mathrm{d}\Phi(u)}{\mathrm{d}u}\pi_\infty(u)\right] + \frac{\mathrm{d}^2\pi_\infty(u)}{\mathrm{d}u^2} \\
&= \frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{\mathrm{d}\Phi(u)}{\mathrm{d}u}\pi_\infty(u) + \frac{\mathrm{d}\pi_\infty(u)}{\mathrm{d}u}\right] \\
&=: \frac{\mathrm{d}}{\mathrm{d}u}J(u)
\end{aligned} \tag{3.12}$$

and thus $J(u)$ has to be constant. With the boundary condition that $J(u) = 0$ at infinity, we get

$$J(u) = \frac{\mathrm{d}\Phi(u)}{\mathrm{d}u}\pi_\infty(u) + \frac{\mathrm{d}\pi_\infty(u)}{\mathrm{d}u} = 0, \tag{3.13}$$

which has the solution

$$\pi_\infty(u) \propto \exp(-\Phi(u)). \tag{3.14}$$

For an initial $u_0 \in \mathbb{R}^D$ and a step size $\tau > 0$, the Euler-Maruyama discretization of Equation (3.10) is given by

$$u_{t+1} = u_t - \tau \nabla \Phi(u_t) + \sqrt{2\tau}\omega_t, \tag{3.15}$$

with $\omega_t \sim \mathcal{N}(\mathbf{0}_K, I_K) \in \mathbb{R}^K$.

In order to apply this in the context of *Bayesian inverse problems*, one may choose the potential $\Phi(u)$ such that $\pi_\infty = \pi_{\text{post.}}$ in Equation (3.14). Samples can then be generated from an accessible prior distribution μ_{prior} with density π_{prior} and evolved according to the dynamics described above until they are approximately distributed according to $\pi_{\text{post.}}$. The choice of the appropriate potential for the model problem has already been discussed in Section 2.2 with

$$\Phi(u) = L(u) - \log \pi_{\text{prior}}(u). \tag{2.3}$$

Convergence of the Langevin Algorithm for Gaussian Prior and Linear Forward Map

In order to investigate the convergence of Equation (3.15), it is essential to quantify the distance between the distribution of the particles u_t at time $t \geq 0$ and the target distribution, which will be the posterior distribution $\mu_{\text{post.}}$ in the context of *Bayesian inverse problems*. The derivation will follow [Pedregosa, 2023].

Definition 3.6 (p-Wasserstein distance)**.** *Let $p \geq 1$ and denote by $\mathcal{P}_p(\mathbb{R}^D)$ the set of all Borel probability measures μ on $\mathbb{R}^D$ such that $\int_{\mathbb{R}^D} ||x||^p \, \mathrm{d}\mu(x) < \infty$, where $||\cdot||$ is the Euclidean norm on $\mathbb{R}^D$. Given probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^D)$, the p-Wasserstein distance $W_p(\mu, \nu)$ is defined as*

$$W_p(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \left(\mathbb{E}_{(x,y)\sim\gamma} \left[||x - y||^p \right] \right)^p,$$

where $\Pi(\mu, \nu)$ denotes the set of all couplings of μ and ν.

Proposition 3.7 (2-Wasserstein distance for Gaussian distributions)**.** *For Gaussian distributions μ and ν on $\mathbb{R}^D$ with mean m_μ, m_ν and covariance Σ_μ, Σ_ν, respectively, the squared 2-Wasserstein distance $W_2(\mu, \nu)^2$ can be expressed in an explicit formulation as*

$$W_2(\mu, \nu)^2 = \frac{1}{2} \left(||m_\mu - m_\nu||^2 + \mathrm{Tr}(\Sigma_\mu + \Sigma_\nu - 2\sqrt{\Sigma_\mu \Sigma_\nu}) \right).$$

If further $\Sigma_\mu \Sigma_\nu = \Sigma_\nu \Sigma_\mu$ is satisfied, the expression can be simplified as

$$W_2(\mu,\nu)^2 = ||m_\mu - m_\nu||^2 + ||\Sigma_\mu^{\frac{1}{2}} - \Sigma_\nu^{\frac{1}{2}}||_F^2,$$

where $||\cdot||_F$ denotes the Frobenius norm.

Proof. See [Dowson and Landau, 1982]. □

We will assume that the posterior distribution $\mu_{\text{post.}}$ is Gaussian with mean m_{post} and covariance $\mathcal{C}_{\text{post.}}$. This choice is justified in the case of a linear forward operator $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^K$, which is demonstrated in Appendix B. Since the potential Φ is set up such that $\pi_{\text{post.}}(u) \propto \exp(-\Phi(u))$ and $\pi_{\text{post.}} \sim \mathcal{N}(m_{\text{post}}, \mathcal{C}_{\text{post.}})$, Φ can be expressed as

$$\Phi(u) = \frac{1}{2}(u - m_{\text{post}})^\intercal \mathcal{C}_{\text{post.}}^{-1}(u - m_{\text{post}}).$$

The Langevin algorithm relies on the potential gradient, which is simple to compute in this case:

$$\nabla_u \Phi(u) = \mathcal{C}_{\text{post.}}^{-1}(u - m_{\text{post}}).$$

Recalling (3.15), we have

$$\begin{aligned} u_{t+1} &= u_t - \tau \nabla \Phi(u_t) + \sqrt{2\tau}\omega_t \\ &= u_t - \tau \mathcal{C}_{\text{post.}}^{-1}(u_t - m_{\text{post}}) + \sqrt{2\tau}\omega_t \\ &= u_t - m_{\text{post}} + m_{\text{post}} - \tau {\mathcal{C}_{\text{post.}}}^{-1}(u_t - m_{\text{post}}) + \sqrt{2\tau}\omega_t \\ &= m_{\text{post}} + (I_D - \tau \mathcal{C}_{\text{post.}}^{-1})(u_t - m_{\text{post}}) + \sqrt{2\tau}\omega_t \end{aligned}$$

and thus

$$u_{t+1} - m_{\text{post}} = (I_D - \tau \mathcal{C}_{\text{post.}}^{-1})(u_t - m_{\text{post}}) + \sqrt{2\tau}\omega_t,$$

which can be repeated in order to obtain

$$u_{t+1} - m_{\text{post}} = (I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{t+1}(u_0 - m_{\text{post}}) + \sqrt{2\tau}\sum_{i=0}^{t}(I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{t-i}\omega_i. \tag{3.16}$$

Note that $\mathbb{E}[\omega_t] = 0$ for all $t \geq 0$ and assume $u_0 \sim \mathcal{N}(m_0, \sigma^2 I_D)$. Taking expectations of Equation (3.16) yields

$$\begin{aligned} \mathbb{E}[u_{t+1}] - m_{\text{post}} &= m_{t+1} - m_{\text{post}} \\ &= (I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{t+1}(m_0 - m_{\text{post}}), \end{aligned}$$

which implies that $\mathbb{E}[u_t]$ approaches the posterior mean m_{post} exponentially fast as $t \to \infty$, provided that $\rho(I_D - \tau \mathcal{C}_{\text{post.}}^{-1}) < 1$, which can be controlled via the step size τ. Here, $\rho(A)$ denotes the spectral radius of a matrix $A \in \mathbb{R}^{D \times D}$.

The covariance $\mathcal{C}_t$, $t \geq 0$ is given by

$$
\begin{aligned}
\mathcal{C}_{t+1} &= \mathbb{E}[(u_{t+1} - m_{t+1})^2] \\
&= \mathbb{E}\left[\left((I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{t+1}(u_0 - m_{\text{post}}) + \sqrt{2\tau}\sum_{i=0}^{t}(I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{t-i}\omega_i - (m_{t+1} - m_{\text{post}})\right)^2\right] \\
&= \mathbb{E}\left[\left((I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{t+1}(u_0 - m_0) + \sqrt{2\tau}\sum_{i=0}^{t}(I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{t-i}\omega_i\right)^2\right] \\
&= 2\tau \sum_{i=0}^{t}(I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{2i} + \sigma^2 (I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{2(t+1)} \\
&= (\mathcal{C}_{\text{post.}}^{-1} - \frac{\tau}{2}\mathcal{C}_{\text{post.}}^{-2})^{-1} + (I_D - \tau \mathcal{C}_{\text{post.}}^{-1})^{2(t+1)}\big(\sigma^2 I_D - (\mathcal{C}_{\text{post.}}^{-1} - \frac{\tau}{2}\mathcal{C}_{\text{post.}}^{-2})^{-1}\big).
\end{aligned}
$$

Note that the second summand tends to 0 as $t \to \infty$. Thus $\mathcal{P}_t \to (\mathcal{C}_{\text{post.}}^{-1} - \frac{\tau}{2}\mathcal{C}_{\text{post.}}^{-2})^{-1} \neq \mathcal{C}_{\text{post.}}$. However, as $\tau \to 0$, this difference vanishes.

Lemma 3.8. *Let π_t denote the particle distribution of the Langevin algorithm (3.15) with step-size τ and quadratic objective function. Further let $\pi_{prior} \sim \mathcal{N}(m_{prior}, \sigma^2 I_D)$, $\sigma \leq L(1 + \tau 2L)$ with l and L being a lower and upper bound on the eigenvalues of $\mathcal{C}_{post.}^{-1}$. Then, (3.15) converges exponentially fast in the Wasserstein distance towards a Gaussian distribution π_τ, with mean m_{post} and covariance $(\mathcal{C}_{post.}^{-1} - \frac{\tau}{2}\mathcal{C}_{post.}^{-2})^{-1}$. For any $\tau < \frac{2}{L}$ it holds*

$$W_2(\pi_t, \pi_\tau) \leq (\max|1 - \tau L|, |1 - \tau l|)^t W_2(\pi_{prior}, \pi_\tau).$$

Proof. See [Pedregosa, 2023]. □

3.2.3. Interacting Particle Langevin Samplers

The process of *Langevin sampling* may be interpreted as evolving a set of particles in the parameter space moving according to a stochastic differential equation. A *drift term*, which relies on indirect measurements given by the forward map and measurement noise, moves particles to areas of higher probability while a *diffusion term* introduces some randomness in order to prevent the particles from all collapsing towards some mode of the target distribution. In order to further improve the efficiency of this sampling process, interactions between particles can be introduced.

Let u be a random variable taking values in $\mathbb{R}^D$ and having the posterior probability density $\pi_{\text{post.}}(u) = \frac{1}{Z}\exp(-\Phi(u))$ based on a measurement $y \in \mathbb{R}^K$. $\Phi : \mathbb{R}^D \to \mathbb{R}$ is

called *potential* and $Z := \int_{\mathbb{R}^D} \exp(-\Phi(u))du < \infty$ is a normalization constant ensuring that $\pi_{\text{post.}}$ is indeed a probability density.

Sampling is done with M interacting particles moving according to a stochastic process in $\mathbb{R}^D$ such that the marginal distributions approach $\pi_{\text{post.}}$ as $t \to \infty$. For any given time t, the particle positions are collected in a matrix

$$U^{(t)} = (u_1^{(t)}, u_2^{(t)}, \dots, u_M^{(t)}) \in \mathbb{R}^{D \times M},$$

where $u_i^{(t)}$ encodes the position of the i-th particle at time t.

With this notation the evolution equations are given by

$$\mathrm{d}u_i^{(t)} = -\mathcal{A}(U^{(t)})\nabla_{u_i}\mathcal{V}(U^{(t)}) + \Gamma(U^{(t)})\mathrm{d}W_i^{(t)} \tag{3.17}$$

for a positive semi-definite $\mathcal{A}(U) \in \mathbb{R}^{D \times D}$ that may encode particle interactions, a potential $\mathcal{V} : \mathbb{R}^{D \times M} \to \mathbb{R}$, and $\Gamma(U) \in \mathbb{R}^{D \times L}$ with $L \in \mathbb{N}$ typically being equal to D or K, such that the $W_i^{(t)}$ are standard L-dimensional Wiener processes.

The classical example of such a sampler is given by *scaled first-order overdamped Langevin dynamics*

$$\mathrm{d}u_i^{(t)} = -C(U^{(t)})\nabla_{u_i}\Phi(U^{(t)}) + \sqrt{2}C^{\frac{1}{2}}(U^{(t)})\mathrm{d}W_i^{(t)}, \tag{3.18}$$

such that $\mathcal{A}(U) = C(U)$ is the empirical covariance of the particles, $\mathcal{V}(U) = \Phi(U)$ is the potential and $\Gamma(U) = \sqrt{2}C^{\frac{1}{2}}(U)$ with a square root of the empirical covariance.

3.2.4. ALDI

[Garbuno-Inigo et al., 2020] introduces Langevin-based sampler that has the advantage of being affine invariant.

Definition 3.9 (Affine Invariance)**.** *Consider the evolution equations for a collection of parameters* $U^{(t)} = (u_1^{(t)}, u_2^{(t)}, \dots, u_M^{(t)}) \in \mathbb{R}^{D \times M}$ *given in Equation* (3.17)*. That is*

$$\mathrm{d}u_i^{(t)} = -\mathcal{A}(U^{(t)})\nabla_{u_i}\mathcal{V}(U^{(t)}) + \Gamma(U^{(t)})\mathrm{d}W_i^{(t)} \tag{3.17}$$

with $\mathcal{A}(U) \in \mathbb{R}^{D \times D}$ *positive, semi-definite,* $\mathcal{V} : \mathbb{R}^{D \times M} \to \mathbb{R}$ *and* $\Gamma(U) \in \mathbb{R}^{D \times L}$*,* $L \in N$*. Let*

$$u_i = Av_i + b \tag{3.19}$$

be an affine transformation for a non-singular matrix $A \in \mathbb{R}^{D \times D}$ *and a vector* $b \in \mathbb{R}^D$*. Let*

$$V = (v_1, v_2, \cdots, v_M) \in \mathbb{R}^{D \times M}$$

be the vector of $M \in \mathbb{N}$ transformed parameters. The method is called affine invariant if the evolution equations of the transformed parameters are given by

$$\mathrm{d}v_i^{(t)} = -\mathcal{A}(V^{(t)})\nabla_{u_i}\mathcal{V}(AV^{(t)} + b\mathbf{1}_M^\intercal) + \Gamma(V^{(t)}t)\mathrm{d}W_i^{(t)}, \tag{3.20}$$

where $\mathbf{1}_M \in \mathbb{R}^M$ is the vector of $M \in \mathbb{N}$ ones.

In order to give the differential equation that describes the method, a few notational conventions are introduced.

Definition 3.10 (Empirical Covariance and Mean)**.** *For a collection of parameters $U = (u_1, u_2, \ldots, u_M) \in \mathbb{R}^{D\times M}$, their empirical covariance matrix is given by*

$$\mathcal{C}(U) = \frac{1}{M}\sum_{i=1}^{M}(u_i - \bar{U})(u_i - \bar{U})^\intercal, \tag{3.21}$$

where

$$\overline{U} = \frac{1}{M}\sum_{i=1}^{M} u_i \tag{3.22}$$

is their empirical mean.

In order to simplify this notation, particle deviations from the mean are written as

$$\tilde{U} = (u_1 - \overline{U}, u_2 - \overline{U}, \ldots, u_M - \overline{U}) \in \mathbb{R}^{D\times M}, \tag{3.23}$$

such that the empirical covariance may be expressed by

$$\mathcal{C}(U) = \frac{1}{M}\tilde{U}(\tilde{U})^\intercal. \tag{3.24}$$

Utilizing this notation, a square root of the empirical covariance is given by

$$\mathcal{C}^{\frac{1}{2}}(U) = \frac{1}{\sqrt{M}}\tilde{U}, \tag{3.25}$$

such that $\mathcal{C}(U) = \mathcal{C}^{\frac{1}{2}}(U)(\mathcal{C}^{\frac{1}{2}}(U))^\intercal$.

Remark. *Note that in the setting of Definition 3.10, non-singularity of $\mathcal{C}(U)$ is obtained for $M > D$.* △

With these notational conventions, the affine invariant method can be expressed as a set of evolution equations in the vain of Equation (3.17).

Definition 3.11 (ALDI Method)**.** *The particle system for a collection of parameters* $U^{(t)} = (u_1^{(t)}, u_2^{(t)}, \dots, u_M^{(t)}) \in \mathbb{R}^{D\times M}$*,* $t \in \mathbb{N}$*, given by*

$$\mathrm{d}u_i^{(t)} = -\mathcal{C}(U^{(t)})\nabla_{u_i}\Phi(u_i^{(t)})\mathrm{d}t + \frac{D+1}{M}\left(u_i^{(t)} - \overline{U}^{(t)}\right)\mathrm{d}t + \sqrt{2}\mathcal{C}^{\frac{1}{2}}(U^{(t)})\mathrm{d}W_i^{(t)}, \tag{3.26}$$

is called affine invariant Langevin dynamics (ALDI). Here, $\mathcal{C}(U^{(t)})$ *denotes the empirical covariance of* $U^{(t)}$ *with square root* $\mathcal{C}^{\frac{1}{2}}(U^{(t)})$ *and* $\overline{U}^{(t)}$ *denotes the empirical mean of* $U^{(t)}$*.*

Gradient-free ALDI

In addition to the general *ALDI* method described above, [Garbuno-Inigo et al., 2020] also introduces a gradient-free formulation which is computationally less demanding.

Definition 3.12 (Empirical Cross-Correlation)**.** *For a collection of parameters* $U = (u_1, u_2, \dots, u_M) \in \mathbb{R}^{D\times M}$*, the empirical cross-correlation is given by*

$$\mathcal{D}(U) = \frac{1}{M}\sum_{i=1}^{M}\big(u_i - \overline{U}\big)\big(\mathcal{G}(u_i) - \overline{\mathcal{G}(U)}\big)^{\intercal}, \tag{3.27}$$

where

$$\overline{\mathcal{G}(U)} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{G}(u_i) \tag{3.28}$$

is the empirical mean.

The empirical cross-correlation does not rely on differentiation and has relatively low computational cost, since it only consists of summation and basic operations on vectors. Thus it allows for a more efficient direct computation of the drift term for a simple class of forward maps.

Proposition 3.13. *For affine forward maps of the form* $\mathcal{G}(u) = Au + b$ *with* $A \in \mathbb{R}^{D\times D}$ *non-singular and* $b \in \mathbb{R}^D$*, it holds*

$$\mathcal{D}(U) = \mathcal{C}(U)\nabla_u\mathcal{G}(u). \tag{3.29}$$

Proof. Let the assumptions of Definition 3.10 and Definition 3.12 hold and further let $\mathcal{G}(u) = Au + b \in \mathbb{R}^K$ for $u \in \mathbb{R}^D$. Then it holds that $\nabla\mathcal{G}(u) = A$ for all $u \in \mathbb{R}^D$. Note

that for all $u \in \mathbb{R}^D$ further holds

$$\begin{aligned}\mathcal{G}(u) - \overline{\mathcal{G}(U)} &= Au + b - \frac{1}{M}\sum_{i=1}^{M} Au_i + b \\ &= Au - \frac{1}{M}\sum_{i=1}^{M} Au_i \\ &= A\big(u - \overline{U}\big).\end{aligned}$$

Substituting this into the definition of $\mathcal{D}(U)$ yields

$$\begin{aligned}\mathcal{D}(U) &= \frac{1}{M}\sum_{i=1}^{M} \big(u_i - \overline{U}\big)\big(\mathcal{G}(u_i) - \overline{\mathcal{G}(U)}\big)^\intercal \\ &= \frac{1}{M}\sum_{i=1}^{M} \big(u_i - \overline{U}\big)\big(A\big(u_i - \overline{U}\big)\big)^\intercal \\ &= \frac{1}{M}\sum_{i=1}^{M} \big(u_i - \overline{U}\big)\big(u_i - \overline{U}\big)^\intercal A^\intercal \\ &= \mathcal{C}(U)\nabla_u \mathcal{G}(u).\end{aligned}$$

□

This identity is regarded as an approximation for general forward maps, thus allowing for a computationally less intensive method at the cost of accuracy.

Definition 3.14 (Gradient-free ALDI)**.** *Let the assumptions of Definition 3.11 hold. A gradient-free approximation of the particle system (3.26) is given by*

$$\mathrm{d}u_i^{(t)} = -\mathcal{D}(U^{(t)})\Sigma^{-1}\big(\mathcal{G}(u_i^{(t)}) - \tilde{y}\big)\mathrm{d}t + \frac{D+1}{M}\left(u_i^{(t)} - \overline{U}^{(t)}\right)\mathrm{d}t + \sqrt{2}\mathcal{C}^{\frac{1}{2}}(U^{(t)})\mathrm{d}W_i^{(t)}. \tag{3.30}$$

The authors note that affine invariance is maintained for the gradient-free formulation.

3.2.5. LIDL

As additional components to Langevin samplers such as *ALDI*, [Eigel et al., 2022] propose methods to improve convergence to a variety of distributions. A particularly difficult problem for Langevin samplers is generating samples from distributions with multiple modes that are not close to each other. The difficulties are amplified if the prior distribution is close to one of the target distributions.

Example 3.15 (Gaussian Mixture)**.** *Consider an instance of a Bayesian inverse problem where the potential is given by*

$$\Phi(u) : \mathbb{R}^2 \to \mathbb{R}, \quad \Phi(u) = -\log\left(\frac{1}{2\pi 4\sqrt{\det\Sigma}}\sum_{i=1}^{4}\exp\left(-\frac{1}{2}(u-c_i)^\intercal\Sigma^{-1}(u-c_i)\right)\right),$$

where $\Sigma = I_2$ *and* $c_i = \left(\cos i\frac{\pi}{2}, \sin i\frac{\pi}{2}\right), i = 1, \dots, 4$. *The prior density is chosen to be*

$$\pi_{prior} = \frac{1}{2\pi\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}(u-c_3)^\intercal\Sigma^{-1}(u-c_3)\right).$$

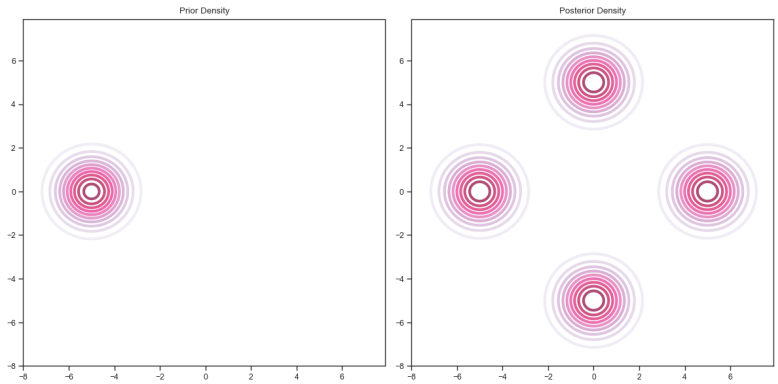

Figure 3.1.: Contour plot of the prior and posterior density of a Gaussian mixture.

Here, the prior density is identical to one of the modes of the posterior density, which results in the potential having a local minimum at this point. Thus, particles are unlikely to escape this minimum in a Langevin sampler. △

The main contributions of [Eigel et al., 2022] consist of two classes of modifications or extensions to Langevin samplers which will be presented briefly.

Preconditioning

In order to mitigate the issue of local modes in the posterior distribution, a preconditioning step is employed. Specifically, an auxiliary distribution μ_{aux} is introduced, which, while not necessarily identical to the prior distribution μ_{prior}, allows for efficient sampling. The corresponding auxiliary potential is denoted by Ψ. Both the prior and auxiliary distributions are utilized in the *LIDL*-method, where the particle propagator is governed by a homotopy between the prior and auxiliary potentials

$$\Upsilon(s) = (1-s)\Psi + s\Phi,$$

for $s \in [0,1]$. This formulation generates intermediate distributions with densities proportional to $\exp(-\Upsilon(s))$. Natural choices for the auxiliary distribution include setting $\mu_{\text{aux}} = \mu_{\text{prior}}$ or employing a Gaussian approximation of the posterior. Another approach would be to use an approximation of the forward map $\mathcal{G}$ in order to define the auxiliary potential Ψ.

Ensemble Enrichment

As evaluations of the forward map can be computationally expensive in many settings, ensemble enrichment aims to reduce such interactions by reducing the number of particles in early stages of the sampling process. These strategies however do not hold special relevance in the context of this thesis and will thus not be presented further.

3.3. Langevin-Based Expectation-Maximization

As seen in Section 3.1, we can employ the *EM-algorithm* in order to estimate both the posterior distribution of u and the measurement noise parameter θ. Considering a simple instance of the model problem from Section 2.3

$$y = \mathcal{G}(u) + \eta,$$

where $\eta \sim \mathcal{N}(0, \Sigma)$, we can set $\theta = \Sigma$. Then the defining steps of the Langevin-based EM-algorithm are as follows:

- **E-step:** Generate samples $u = (u_{i1}, \dots, u_{iM})_{i=1}^N$ from the posterior distribution $\pi_{\text{post.},\Sigma_{(r)}}(u|y_i)$ given the current covariance estimate $\Sigma_{(r)}$ and observations y_i, $i = 1, \dots, N$ using a Langevin-algorithm.
- **M-step:** Update the covariance estimate $\Sigma \leftarrow \Sigma_{(r+1)}$, utilizing the previously attained samples from $\pi_{\text{post.},\Sigma_{(r)}}(u|y_i)$, $i = 1, \dots, N$.

3.3.1. Maximization-Step for Scalar Covariance Matrix

Consider the case where $\mathbb{R}^K \ni \eta \sim \mathcal{N}(\mathbf{0}_K, \Sigma)$ with $\Sigma = \sigma^2 I_K$. We will derive an analytical formulation of the *M-step* of the Langevin-based EM-algorithm. From Equation (3.9) we have

$$\begin{aligned}\sigma^2_{(r+1)} &= \arg\max_{\sigma^2} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{u \sim q^{(r+1)}} \left[\log \left(p_{\sigma^2_{(r)}}(y_i|u)\, p(u) \right) \right] \\ &= \arg\max_{\sigma^2} -\frac{K}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2NM\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{M} ||y_i - \mathcal{G}(u_{ij})||^2 \end{aligned} \tag{3.31}$$

For ease of notation we define

$$Q(\sigma^2, y, u) := -\frac{K}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2NM\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M}||y_i - \mathcal{G}(u_{ij})||^2.$$

Finding the maximizer $\sigma^2_{(r+1)}$ can be achieved by setting the derivative of Q with respect to σ^2 to zero.

$$\begin{aligned} & 0 = \frac{\mathrm{d}}{\mathrm{d}\sigma^2}Q(\sigma^2_{(r+1)}, y, u) \\ \iff \qquad & \sigma^2_{(r+1)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}||y_i - \mathcal{G}(u_{ij})||^2}{NMK}. \end{aligned} \tag{3.32}$$

The derivation of Equation (3.31), Equation (3.32) and the proof that this is indeed a maximizer is discussed in detail in Appendix C.

This leads to an algorithmic description of the process.

Algorithm 3.3.1 Langevin-EM Algorithm for Scalar Covariance Matrix

Require: Initial noise level σ_0^2 of the measurement error, observations $y_1, \cdots, y_N \in \mathbb{R}^K$, prior distribution π_{prior}, forward map $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^K$

Ensure: Estimated parameters σ_*^2, posterior samples $u_{i1}, \cdots, u_{iM} \in \mathbb{R}^D$ for $i = 1, \ldots, N$.

1: Initialize $\theta \leftarrow \sigma_0^2$
2: Initialize prior samples $\{u_{i1}, \cdots, u_{iM}\}_{i=1}^N$.
3: **repeat**
4: **E-step:** Compute posterior samples $u_{i1}, \cdots, u_{iM}$ via an interacting particle Langevin sampler (such as $ALDI$) based on measurements y_i for $i = 1, \ldots, N$.
5: **M-step:** Update the noise estimate $\sigma_t^2 \leftarrow \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}||y_i - \mathcal{G}(u_{ij})||^2}{NMK}$
6: Update $t \leftarrow t + 1$
7: **until** Convergence of σ^2

3.3.2. Maximization-Step for Diagonal Covariance Matrix

For dimensionally dependent noise $\eta \sim \mathcal{N}(\mathbf{0}_K, \Sigma)$, we may propose a covariance matrix that is of diagonal type $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_K^2) \in \mathbb{R}^{K\times K}$. Here, we may examine each component individually by computing each diagonal entry $\sigma_k^2 = \Sigma_{kk}$ for $k = 1, \ldots, K$.

This can be achieved by observing that

$$\begin{aligned}\sigma_k^2 &= \mathbb{E}[(y_k - \mathbb{E}[y_k])^2] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}(y_{ik} - \mathbb{E}[y_{ik}])^2 \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\left(y_{ik} - \frac{1}{M}\sum_{j=1}^{M}\mathcal{G}(u_{ij})_k\right)^2.\end{aligned}$$

Algorithm 3.3.2 Langevin-EM Algorithm for Diagonal Covariance Matrix

Require: Initial noise level σ_0^2 of the measurement error, observations $y_1, \cdots, y_N \in \mathbb{R}^K$, prior distribution π_{prior}, forward map $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^K$

Ensure: Estimated parameters σ_*^2, posterior samples $u_{i1}, \cdots, u_{iM} \in \mathbb{R}^D$ for $i = 1, \ldots, N$.

1: Initialize $\theta \leftarrow \sigma_0^2$
2: Initialize prior samples $\{u_{i1}, \cdots, u_{iM}\}_{i=1}^N$.
3: **repeat**
4: **E-step:** Compute posterior samples $u_{i1}, \cdots, u_{iM}$ via an interacting particle Langevin sampler (such as *ALDI*) based on measurements y_i for $i = 1, \ldots, N$.
5: **M-step:** Update the noise estimate $\Sigma_{kkt} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\left(y_{ik} - \frac{1}{M}\sum_{j=1}^{M}\mathcal{G}(u_{ij})_k\right)^2$
6: Update $t \leftarrow t + 1$
7: **until** Convergence of σ^2

3.3.3. Maximization-Step for Arbitrary Covariance Matrix

For arbitrary covariance matrices $\Sigma \in \mathbb{R}^{K\times K}$ of the observational noise $\eta \sim \mathcal{N}(\mathbf{0}_K, \Sigma)$, the *M-step* can also be expressed explicitly. Since observations y_i, $i = 1, \ldots, N$ are distributed according to a Gaussian distribution centered at the image of the true parameter $\mathcal{G}(u)$ and with covariance Σ, we get for the covariance matrix

$$\begin{aligned}\Sigma &= \mathbb{E}[(y - \mathbb{E}[y])(y - \mathbb{E}[y])^\intercal] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbb{E}[y_i])(y_i - \mathbb{E}[y_i])^\intercal \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \frac{1}{M}\sum_{j=1}^{M}\mathcal{G}(u_{ij})\right)\left(y_i - \frac{1}{M}\sum_{j=1}^{M}\mathcal{G}(u_{ij})\right)^\intercal.\end{aligned}$$

For the implementation of the Langevin-based EM-algorithm, one may employ empirical covariance estimators such as the `cov`-method of Pythons popular *numpy* library.

Algorithm 3.3.3 Langevin-EM Algorithm for arbitrary Covariance Matrix

Require: Initial covariance Σ_0 of the measurement error, observations $y_1, \cdots, y_N \in \mathbb{R}^K$, prior distribution π_{prior}, forward map $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^K$

Ensure: Estimated covariance Σ_*, posterior samples $u_{i1}, \cdots, u_{iM} \in \mathbb{R}^D$ for $i = 1, \ldots, N$.

1: Initialize $\theta \leftarrow \sigma_0^2$
2: Initialize prior samples $\{u_{i1}, \cdots, u_{iM}\}_{i=1}^N$.
3: **repeat**
4: **E-step:** Compute posterior samples $u_{i1}, \cdots, u_{iM}$ via an interacting particle Langevin sampler (such as *ALDI*) based on measurements y_i for $i = 1, \ldots, N$.
5: **M-step:** Update the noise estimate $\Sigma_t \leftarrow \frac{1}{N} \sum_{i=1}^N \left(y_i - \frac{1}{M} \sum_{j=1}^M \mathcal{G}(u_{ij}) \right) \left(y_i - \frac{1}{M} \sum_{j=1}^M \mathcal{G}(u_{ij}) \right)^\intercal$.
6: Update $t \leftarrow t + 1$
7: **until** Convergence of Σ

As samples from previous posterior estimations are used as initial values for the following iteration, this process can be interpreted as a preconditioning step which is performed in the *LIDL*-method as described in Section 3.2.5.

The following chapter will present some numerical results of the application of this algorithm.

4. Numerical Experiments

In this chapter, the Langevin-based EM-algorithm described in Section 3.3 will be tested one two problems. The first one is the relatively simple problem of determining the parameters of a linear function based on noisy measurements of the function value at different points. This will serve as a proof of concept and a way of testing some strategies that may improve convergence or reduce computational time. The problem is set up in a manner as to include a nonlinear forward map. This allows for a comparison between gradient-based and gradient-free sampling methods within the *E-step*. The second problem is a common benchmark for posterior sampling methods [Schillings and Stuart, 2017, Garbuno-Inigo et al., 2019, Garbuno-Inigo et al., 2020, Eigel et al., 2022]. It consists of estimating the permeability field of the one-dimensional Darcy equation based on noise measurements of the solution at discrete points. Such PDE-constrained forward models are very common in the natural sciences [Li et al., 2020, Molesky et al., 2018], which highlights the significance of achieving satisfying results for this class of problems.

4.1. Estimating Parameters of a Linear Map

In this experiment, the inverse problem is designed to approximate the parameters $u = (u_1, u_2) \in \mathbb{R}^2$ of a linear map

$$\begin{aligned} \mathcal{G} : \mathbb{R}^2 &\to \mathbb{R}^K \\ (u_1, u_2) &\mapsto (u_1^2 x_i + u_2)_{i=1}^K, \end{aligned}$$

where the $\{x_i\}_{i=1}^K$ are evenly spaced between -10 and 10 with observations given by

$$y^* = \mathcal{G}(u^*) + \eta^*, \quad \eta \sim \mathcal{N}(\mathbf{0}_K, \sigma^2 I_K)$$

with $I_K \in \mathbb{R}^{K \times K}$ being the identity matrix and $\mathbf{0}_K \in \mathbb{R}^K$ the K-dimensional zero-vector. The resulting measurements for $K = 11$ and the image of the true parameter $u^* = (2, 2)$ under $\mathcal{G}$ are illustrated in Figure 4.1.

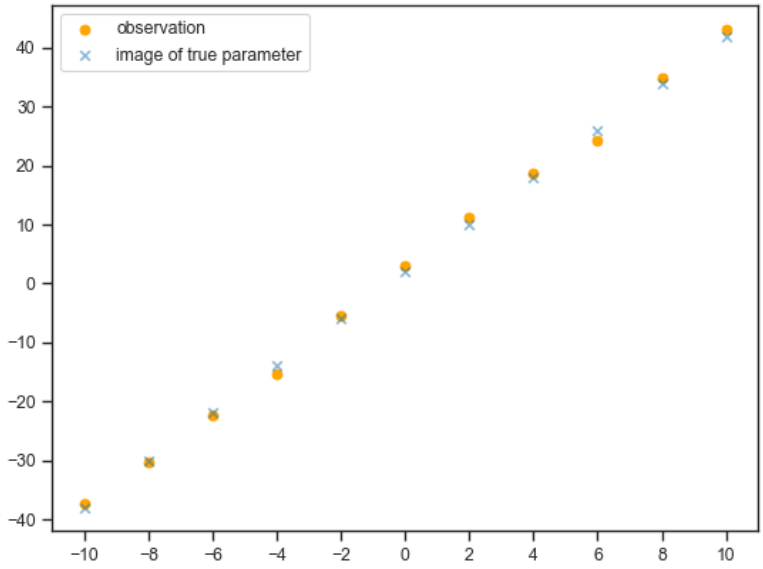

Figure 4.1.: Measurements for $K = 11$ and image of the true parameter u^* under $\mathcal{G}$.

4.1.1. General Results

The Langevin-based EM-algorithm is first run for $L = 30$ iterations with $M = 50$ samples, an initial noise estimate of $\tilde{\sigma}_0^2 = 5$ and measurement dimension $K = 11$ in order to examine the efficacy without any hyperparameter changes. For each experiment, the hyperparameters are documented in a table, with the specific hyperparameters for this experiment listed in Table 4.1.

Table 4.1.: Hyperparameters for Section 4.1.1

Hyperparameter	value
output dimension K	11
# of observations N	1
# of samples per observation M	50
# of EM-iterations L	30
initial noise estimate $\tilde{\sigma}_0^2$	5
prior distribution μ_{prior}	$\mathcal{N}((4,2), 2I_2)$
# sampler steps T	100
sampler step size τ	0.01
# averaging runs R	5

Figure 4.2 shows the resulting samples as approximations of u^* and as points in the sample space $\mathbb{R}^2$ with marginal densities.

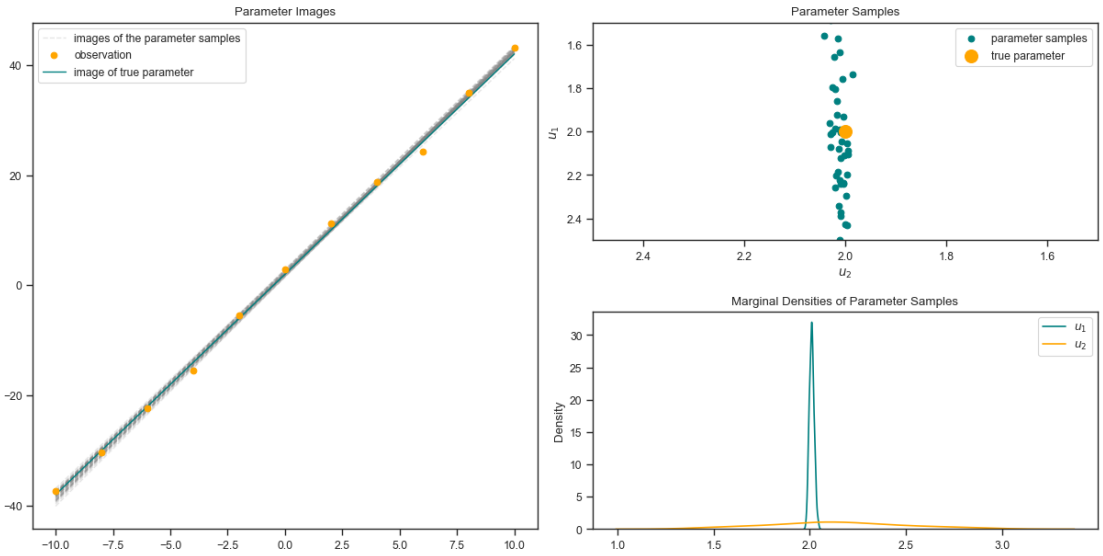

Figure 4.2.: Resulting sample images as approximations of $\mathcal{G}(u^*)$ and as points in the sample space with marginal densities.

Additionally, a plot of the relative distance to the true observational noise level $\frac{|\sigma^2 - \tilde{\sigma}^2|}{\sigma^2}$ is given in Figure 4.3. The initial lack of improvement in the first few iterations is due to the Langevin-EM algorithm's implementation, where a sanity-check is performed at every *M-step*. Here, it is verified that the estimated noise level $\tilde{\sigma}_l^2$ at iteration l is smaller than the initial noise level estimate $\tilde{\sigma}_0^2$. This is valid, since the noise level estimate is initialized as to approximate the true noise level from above. Thus, observation are included within the distributions $\mathcal{N}(\mathcal{G}(u^*), \tilde{\sigma}_l^2)$ for each $l = 0, \ldots, L$. For the first iterations, the estimated noise level does not decrease and thus the particles are propelled using the initial estimate until the update is valid. Although this issue could be addressed by significantly increasing the number of sampler steps during the initial iteration, such an adjustment would yield no substantial difference in results.

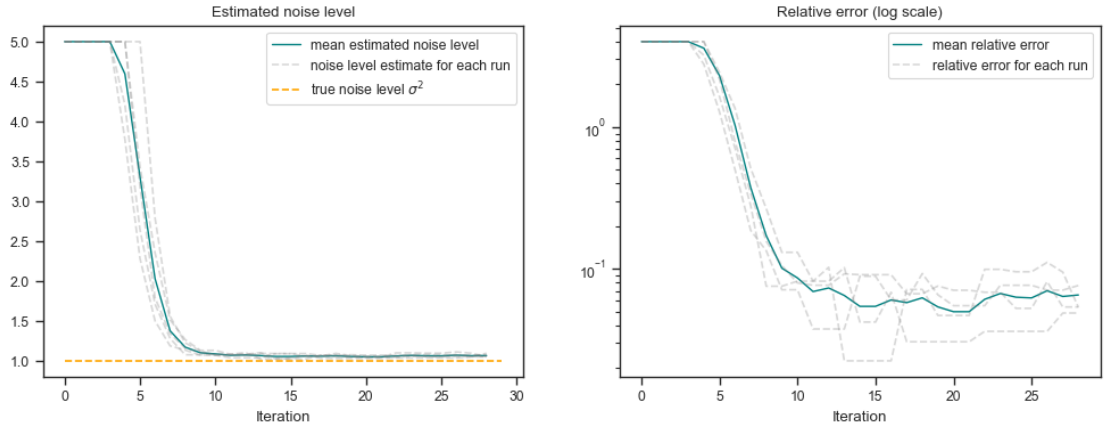

Figure 4.3.: Average estimated noise level $\tilde{\sigma}^2$ (left) and relative distance to the true noise level (right) at each iteration of the Langevin-EM algorithm.

The noise level estimates reach an average relative error of less than 10% within the 10 iterations. Due to the well behaved nature of the *ALDI*-sampler for this problem, this also results in good posterior samples, as demonstrated in Figure 4.2.

Due to the (mild) nonlinearity of the forward map $\mathcal{G}$, we may use this example in order to compare the use of gradient-based and gradient-free *ALDI*-sampling within the *E-step* of the Langevin-based EM-algorithm. All hyperparameters and observations are kept consistent, while only the sampler choice is changed to the gradient-free *ALDI*-method given in Definition 3.14. Comparing Figure 4.3 and Figure 4.4, one can see that the gradient-free sampling method achieves comparable results.

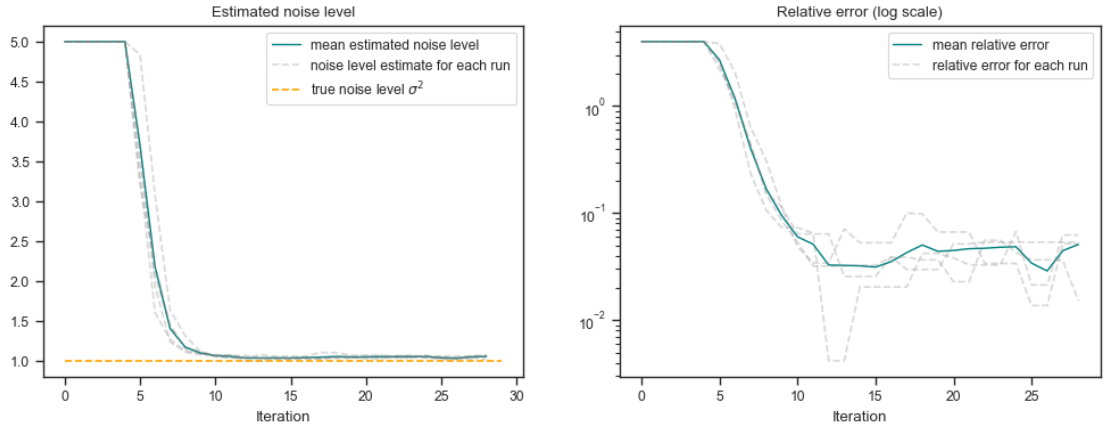

Figure 4.4.: Average estimated noise level $\tilde{\sigma}^2$ (left) and relative distance to the true noise level (right) at each iteration of the Langevin-EM algorithm with gradient-free sampling.

4.1.2. Increase in the Number of Evaluation Points

In this section, we will investigate the effect of an increase in the dimension of observations K, which essentially is an increase in the number of measurement points along the linear graph, on the accuracy of the posterior samples with respect to the true parameter u^*. The dimension of the observations is set to be $K = 2^i$ for $i = 1, \dots, 7$ with the number of samples being kept constant at $M = 50$. The distance to the true parameter is computed individually for both components.

Table 4.2.: Hyperparameters for Section 4.1.2

Hyperparameter	value
K	2^i, $i = 1, \dots, 7$
N	1
M	50
L	20
$\tilde{\sigma}_0^2$	5
μ_{prior}	$\mathcal{N}((4, 2), 2I_2)$
# sampler steps T	300
sampler step size τ	0.01
# averaging runs	5

Figure 4.5 shows steady improvement of the relative approximation error for the noise level estimates.

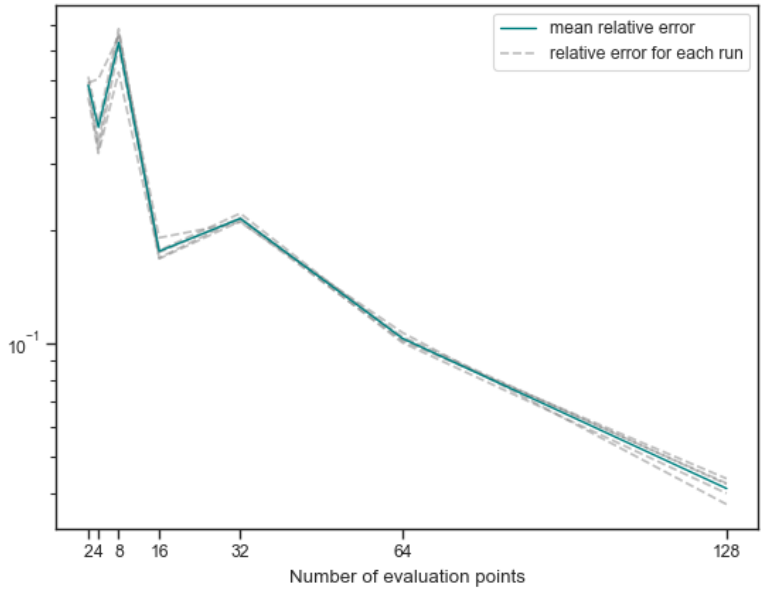

Figure 4.5.: Relative noise level approximation error for an increasing number of evaluation points K.

The average relative distance over all $M = 50$ samples is also recorded for each K and the results are shown in Figure 4.6. Here it is again apparent that even for a drastically higher number of evaluation points K, the second component of the parameter of interest $u \in \mathbb{R}^2$ is approximated less effectively than the first component. It should however be noted that the sample mean does converge towards the true parameter u^*.

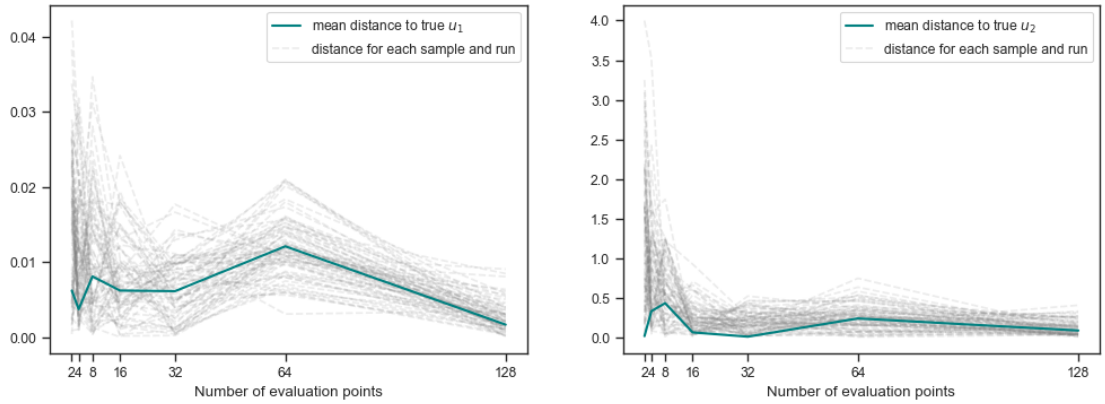

Figure 4.6.: Distance of each component of posterior samples to the true parameter for each K.

4.1.3. Increase in the Number of Observations

In the previous experiment, the number of evaluation points K is increased, while the number of measurements is kept constant at $N = 1$. A different strategy is to reverse this and test the effect of taking multiple noisy measurements, each with the same evaluation points. The number of measurements is increased from $N = 1$ to 7 and the Langevin-based EM-algorithm is run for each of these instances with $K = 11$. All other hyperparameters are also kept constant in order to isolate the effect of an increasing number of measurements, as can be seen in Table 4.3.

Table 4.3.: Hyperparameters for Section 4.1.3

Hyperparameter	value
K	11
N	$1, \dots, 7$
M	50
L	30
$\tilde{\sigma}_0^2$	5
μ_{prior}	$\mathcal{N}((4,2), 2I_2)$
# sampler steps T	300
sampler step size τ	0.01
# averaging runs	5

Figure 4.7 shows the steady decrease of the relative approximation error as the number of measurements increases. Note that for $N = 6$ measurements at $K = 11$ observation points the mean relative error is comparable to the error for $N = 1$ measurement at $K = 64$ observation points from the previous example. One method does not seem to hold a significant advantage over the other.

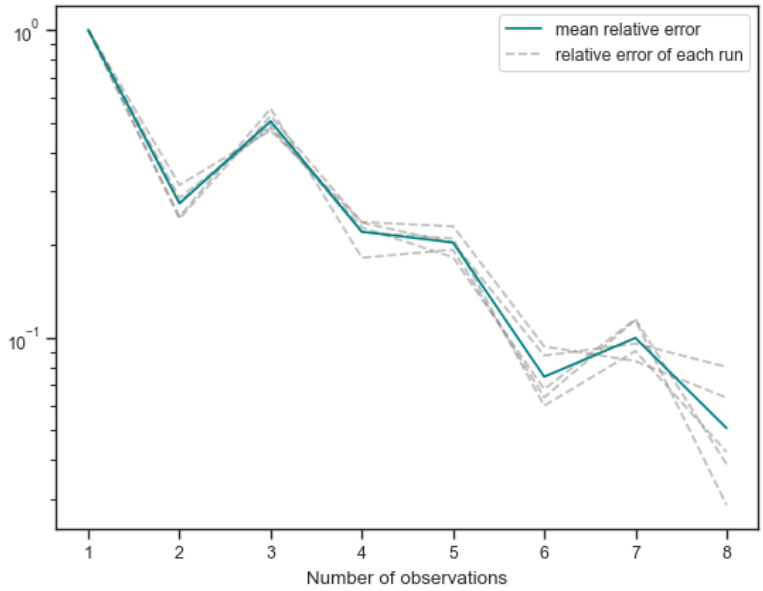

Figure 4.7.: Relative noise level approximation error for an increasing number of noisy measurements.

4.1.4. Increase in the Number of Samples

In order to analyze the effect of the number of samples we run the Langevin-based EM-algorithm with an initial noise estimate of $\tilde{\sigma}^2 = 5$. The number of samples is given by $M = 2^i$ for $i = 1, \ldots, 8$ and the Langevin-based EM-algorithm is run for $L = 10$ iterations given each number of samples. The observation $y \in \mathbb{R}^K$ is the same for each run with $K = 16$. This whole process is done 10 times in order to take a final average, such that the effect of outliers is mitigated.

Table 4.4.: Hyperparameters for Section 4.1.4

Hyperparameter	value
K	11
N	1
M	20
L	30
$\tilde{\sigma}_0^2$	5
μ_{prior}	$\mathcal{N}((4,2), 2I_2)$
# sampler steps T	$500, 1000, 1500, \ldots, 3000$
sampler step size τ	$0.1, 0.015, 0.01, 0.0015, \ldots, 0.0001$
# averaging runs	15

The results are compiled in Figure 4.8. It is evident, that for $M = D = 2$ the approximation of the noise level is significantly hindered. However, significantly increasing the number of samples only marginally reduces the approximation error once an accurate estimate has been obtained.

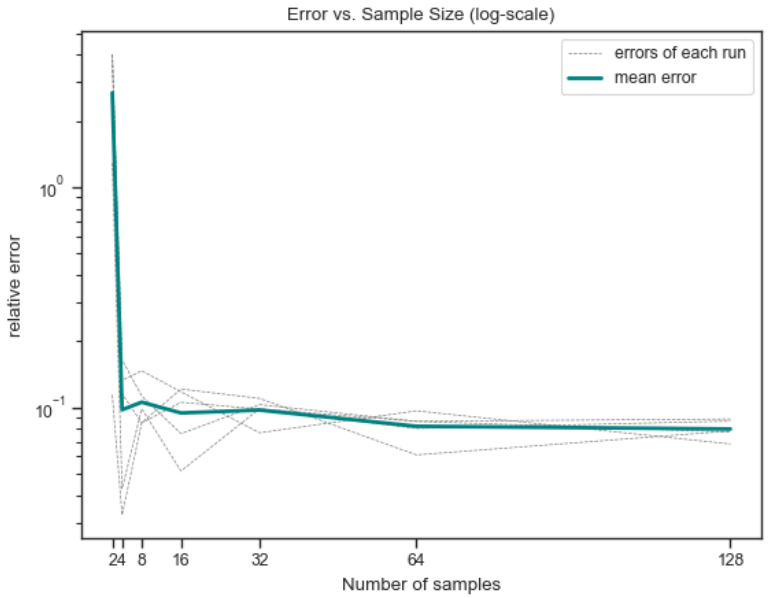

Figure 4.8.: Distance to the true noise level over number of samples.

4.1.5. Adaptive Sampling

As has been demonstrated in Section 4.1.1, the first iterations generally do not improve the noise estimate as the estimated noise level is larger than the initial estimate. This motivates the modification of the sampler step size τ and number of sampling steps T for different stages of the Langevin-based EM-algorithm. In this experiment, the number of sampler steps changes at every fifth iteration of the Langevin-based EM-algorithm. Additionally, at each of these iterations, the sampler step size decreases. This allows the sampler to propagate particles very quickly for the first iterations while reducing the sampling error in later iterations. Together, this results in faster computation due to less interactions with the forward map, which is generally a costly step of the algorithm. As an added benefit, the estimation of the noise level also improves due to the finer resolution of the particle propagator in the final iterations of the EM-loop.

Table 4.5.: Hyperparameters for Section 4.1.5

Hyperparameter	value
K	11
N	1
M	2^i, $i = 1, \dots, 7$
L	30
$\tilde{\sigma}_0^2$	5
μ_{prior}	$\mathcal{N}((4,2), 2I_2)$
# sampler steps T	500, 1000, 1500, 2000, 2500, 3000
sampler step size τ	0.1, 0.015, 0.01, 0.0015, 0.001, 0.00015
# averaging runs	5

Both gradient-based and gradient-free *ALDI* samplers are compared to a gradient-based sampler with constant number of sampler steps $T = 3000$ and constant step size $\tau = 0.0015$. The resulting relative noise level approximation errors and average computations times are compiled in Figure 4.9. Both the gradient-based and gradient-free variable step methods achieve better results than the fixed-step method with significantly less computational effort.

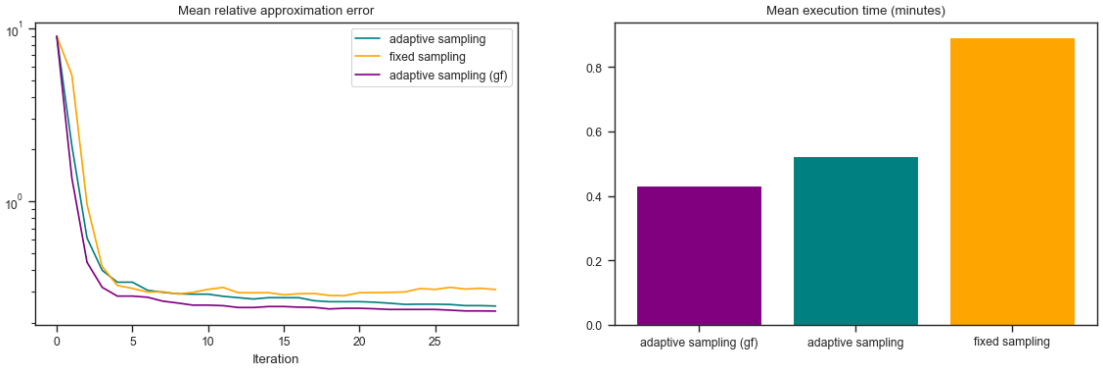

Figure 4.9.: Mean relative error (left) and mean execution time (right) for gradient-based and gradient-free schemes with adaptive steps compared to a gradient-based fixed-step scheme.

4.2. One-Dimensional Darcy Flow

As a surrogate for real-world problems that often include a PDE-constrained forward operator, we consider the one-dimensional elliptic PDE

$$-\nabla \cdot (\exp(u)\nabla p) = f \qquad \text{in } I = (0,1), \tag{4.1}$$

$$p = 0 \qquad \text{on } \partial I, \tag{4.2}$$

where our objective is to infer the log permeability u^* from noisy observations of the form

$$y = \mathcal{O}(p) + \eta = \mathcal{O}(G(u^*)) + \eta = \mathcal{G}(u^*) + \eta,$$

of the solution p. Here, G denotes the solution operator of the elliptic PDE, and $\mathcal{G} := \mathcal{O} \circ G$. $\mathcal{O}$ denotes the observation operator, which returns discrete observations of the PDE solution at $K = 2^{l_{\text{obs}}} - 1$ equidistant points on $I = (0, 1)$. For the experiments, the source term is chosen as $f(x) = 100x$, $x \in I$. The observational noise η is assumed to be normally distributed, i.e., $\eta \sim \mathcal{N}(0, \sigma^2)$ with $0 \neq \sigma \in \mathbb{R}$. The prior distribution of u is taken as a Gaussian prior, $\mu_{\text{prior}} \sim \mathcal{N}(0, \mathcal{C}_{\text{prior}})$, with covariance operator $\mathcal{C}_{\text{prior}} = (-\Delta)^{-1}$, where Δ is the Laplacian on $(0, 1)$ with homogeneous Dirichlet boundary conditions. The forward problem is solved numerically using the Finite Element Method (FEM) with continuous, piecewise linear basis functions on a uniform mesh of width $h = 2^{-l}$, $l \in \mathbb{N}$.

To simulate draws from the prior distribution $\mu_{\text{prior}} \sim \mathcal{N}(0, (-\Delta)^{-1})$, we utilize the Karhunen-Loève expansion of the random field:

$$u(x, \omega) = \sum_{k=1}^{\infty} \frac{\sqrt{2}}{k\pi} \sin(k\pi x)\zeta_k(\omega),$$

where $\zeta_k(\omega)$ are i.i.d. random variables, $\zeta_k \sim \mathcal{N}(0, 1)$. For the numerical simulations, we truncate the KL expansion to:

$$u(x, \omega) = \sum_{k=1}^{n} \frac{\sqrt{2}}{k\pi} \sin(k\pi x)\zeta_k(\omega).$$

Motivated by the results of Section 4.1, we employ a gradient-free sampling approach this section. Previous experiments have shown that the results of using a gradient-free *ALDI*-sampler are comparable to those of the gradient-based version while reducing total computational effort.

4.2.1. General Results

Table 4.6 compiles the hyperparameters used in this experiment.

Table 4.6.: Hyperparameters for Section 4.2.1

Hyperparameter	Value
Discretization level of u: l	4
$D = 2^l - 1$	15
KL-dimension of u^*: n_{true}	5
Discretization level of the observation: l_{obs}	6
$K = 2^{l_{\text{obs}}} - 1$	63
# samples: M	50
KL-dimension of samples: n_{samples}	10
initial noise level estimate: $\tilde{\sigma}_0^2$	0.1
prior distribution: μ_{prior}	$\mathcal{N}(0, -\Delta^{-1})$
# sampler steps: T	100
sampler step size: τ	0.001
# EM-iterations: L	10
# averaging runs	10

Analyzing Figure 4.10, it is evident that the posterior samples approximate the true permeability field u^* reasonably well. The spread of the posterior samples around u^* indicates the uncertainty in the reconstructed permeability field driven by the noisy observations and the prior assumptions.

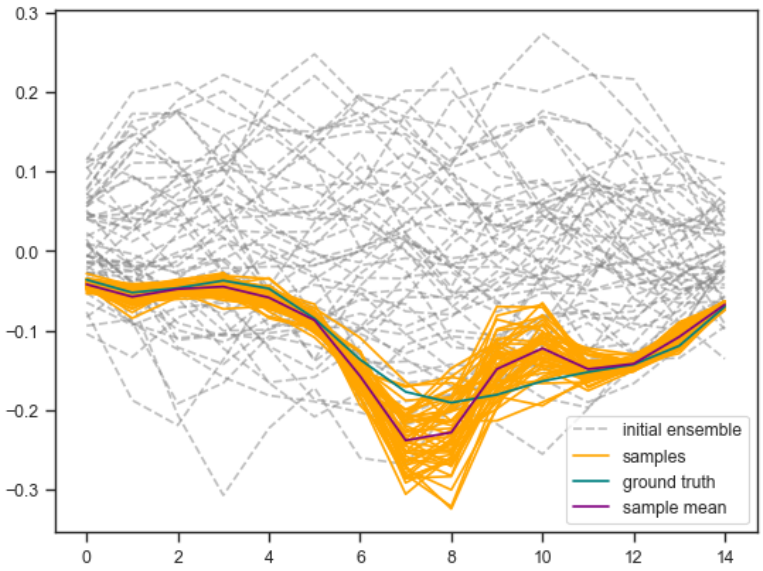

Figure 4.10.: Posterior samples of u compared to true parameter u^* and initial ensemble.

However, it is crucial to compare the forward solutions derived from the posterior samples to the observed data to assess the quality of the solution. Although the sample mean

of the posterior does not perfectly align with the true parameter u^*, we observe that the images under the forward operator $\mathcal{G}$ are nearly identical for both the posterior samples and the true parameter. This observation is evident from Figure 4.11, where the forward solutions from both the sampled and true parameters align closely, suggesting that the posterior samples are effective in reconstructing the underlying physical process represented by the elliptic PDE.

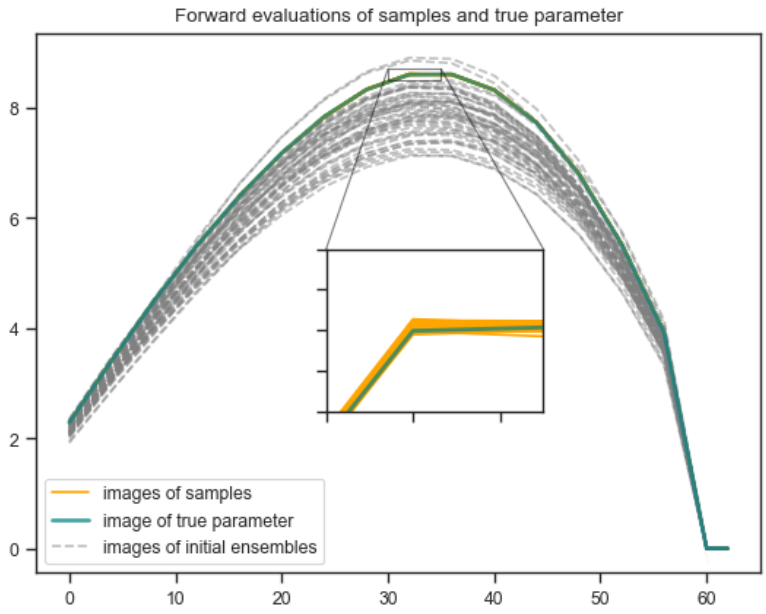

Figure 4.11.: Forward operator images of the initial ensemble, the posterior samples and the true parameter.

Another critical aspect is the estimation of the observational noise level, $\sigma^2 > 0$. The noise level approximation derived through the Langevin-based EM-algorithm exhibits results comparable to those observed in Section 4.1, confirming the robustness of our methodological framework. Figure 4.12 depicts noise level estimates, showing a consistent trend that aligns with the true noise levels used in the simulations. This consistency highlights the efficacy of the Langevin-based EM-algorithm in accurately capturing the uncertainty related to the observational noise.

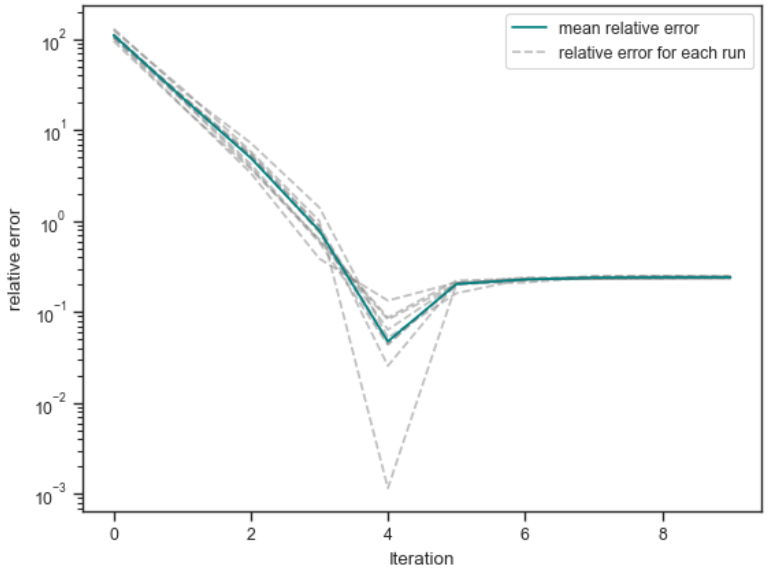

Figure 4.12.: Relative noise level approximation error for each iteration of the main EM-loop.

The pronounced valley observed in Figure 4.12 at iteration 4 is attributed to an underestimation of the noise level, causing it to settle below the true value. Consequently, the relative approximation error momentarily decreases before experiencing a slight increase, ultimately stabilizing around 10^{-1}. While this valley might suggest an optimal iteration point at which to terminate the EM-loop, such information is contingent upon prior knowledge of the true noise level. Thus, in practical applications, this approach is not viable.

4.2.2. Increase in the Number of Evaluation Points

In this experiment, we investigate the effect of increasing the number of evaluation points on the accuracy of the noise level approximation. Specifically, we examine the influence of the parameter l_{obs}, which controls the number of observation points $K = 2^{l_{\text{obs}}} - 1$, on the relative error in the noise level estimation.

Table 4.7.: Hyperparameters for Section 4.2.2

Hyperparameter	Value
l	4
$D = 2^l - 1$	15
n_{true}	5
l_{obs}	$1, \ldots, 7$
$K = 2^{l_{\text{obs}}} - 1$	$1, \ldots, 127$
M	50
n_{samples}	10
$\tilde{\sigma}_0^2$	0.1
μ_{prior}	$\mathcal{N}(0, -\Delta^{-1})$
T	100
τ	0.001
L	10
# averaging runs	10

As illustrated in Figure 4.13, there is a marked trend in the data that reveals a steady decline in the relative noise level approximation error as l_{obs} increases from 1 to 7. This consistent reduction in error highlights the beneficial impact of utilizing a higher number of evaluation points, thereby reducing the discretization error in the FEM solver.

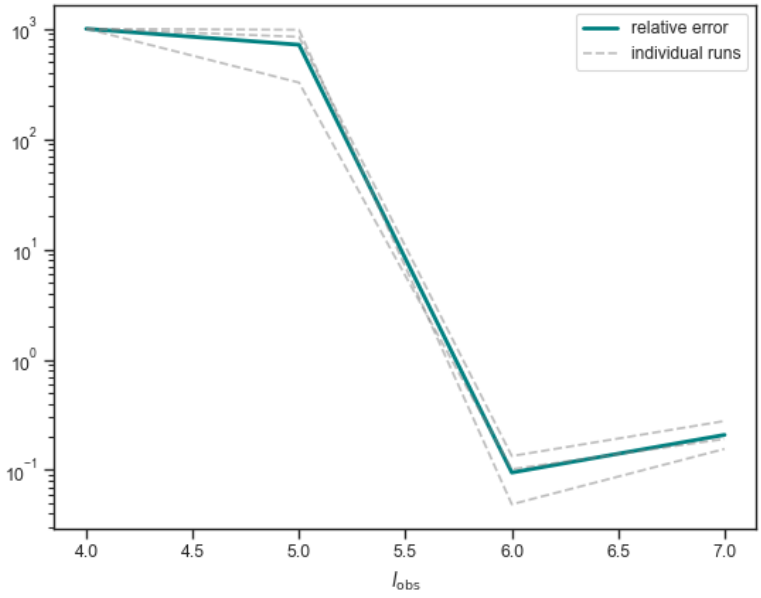

Figure 4.13.: Relative noise level approximation error for an increasing discretization level of the observation operator.

4.2.3. Adaptive Sampling

In Section 4.1, the forward model did not involve great computational effort. Even still, adaptive sampling was shown to have a great effect on computational cost by reducing the number of forward calls drastically. In this more involved PDE-constrained forward model, adaptive sampling promises to lead to even more significant computational benefits while maintaining a noise level approximation comparable to the fixed-sampling scheme.

We implement the adaptive sampling scheme by increasing the number of *ALDI*-steps T at every second iteration of the EM-loop. At these times, the step size τ is decreased for a total of $2 \cdot 6 = 12$ Langevin-based EM-algorithm-iterations. The exact values for T and τ are presented in Table 4.8 The use of an adaptive sampler is compared to a sampler with constant number of sampler steps $T = 300$ and constant step size $\tau = 0.0015$.

Table 4.8.: Hyperparameters for Section 4.2.3

Hyperparameter	Value
l	4
$D = 2^l - 1$	15
n_{true}	5
l_{obs}	6
$K = 2^{l_{\text{obs}}} - 1$	63
M	50
n_{samples}	10
$\tilde{\sigma}_0^2$	0.1
μ_{prior}	$\mathcal{N}(0, -\Delta^{-1})$
T	25, 50, 75, 100, 200, 300
τ	0.1, 0.015, 0.01, 0.0015, 0.001, 0.00015
L	10
# averaging runs	10

From the left panel of Figure 4.14, it is evident that adaptive sampling achieves comparable results to a fixed-sampling method, which is again consistent with the results from Section 4.1.5. The right panel of Figure 4.14 compares the mean execution time required for both approaches. Adaptive sampling demonstrates a marked reduction in computation time, averaging around 0.75 minutes per run, while the fixed sampling method takes approximately 1.75 minutes. This substantial decrease in execution time underscores the efficiency of adaptive sampling, making it a preferable choice for practical applications where computational resources and time are critical constraints.

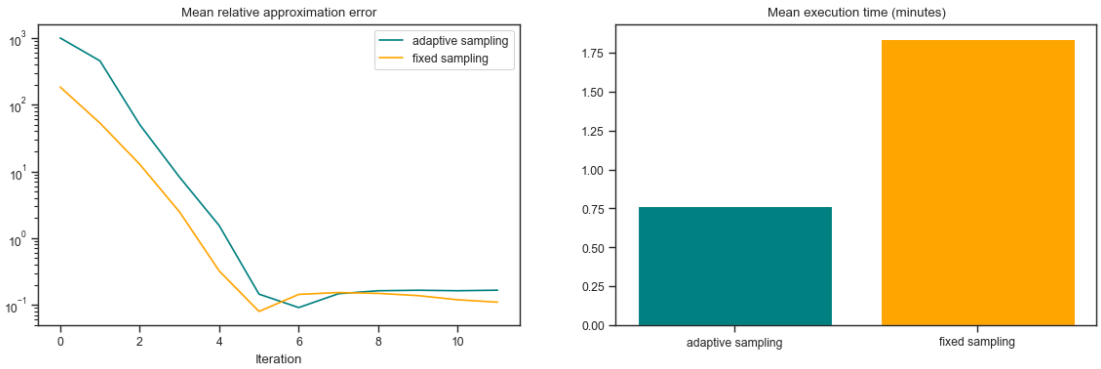

Figure 4.14.: Mean relative approximation error for adaptive and fixed sampling (left) and mean execution time in minutes for adaptive and fixed sampling (right).

5. Conclusion

In this thesis, we propose a method for generating posterior samples and estimating observational noise for the inverse problem

$$y = G(u) + \eta,$$

where $\eta \sim N(0, \Sigma)$ and Σ is an unknown covariance matrix. This consists of an Expectation-Maximization (EM) algorithm that offers an approach to estimate the noise covariance Σ, while simultaneously generating posterior samples during the E-step. A key innovation in the approach is the application of $ALDI$-sampler, which is given in both gradient-based and gradient-free forms. The proposed algorithm was tested on two case studies: a simple illustrative example and a more complex scenario involving a Partial Differential Equation (PDE)-constrained forward map.

Empirical results from these case studies indicate that the Langevin-based EM-algorithm performs robustly, with noise levels estimated to a relative error of approximately 10%. Additionally, the posterior samples generated demonstrated high fidelity to the true posterior distributions. Noteworthy is the performance of the adaptive sampling strategy, where step size and number of sampler-steps are varied across the iterations of the EM-loop, which showed significant promise in reducing computational loads while maintaining or exceeding the accuracy of fixed-step sampling methods.

These promising results open several future research avenues that can further enhance and expand the applicability of the proposed algorithm. Incorporating mixed noise models could make the Langevin-based EM-algorithm more versatile and applicable to a broader range of problems. Another fruitful direction involves extending our algorithm to accommodate normalizing flow noise models in order to extend the observational noise beyond the Gaussian type and allow for arbitrary noise distributions. Incorporating ensemble enrichment techniques, as suggested in [Eigel et al., 2022], could further bolster the algorithm's sampling efficiency and accuracy. While our algorithm empirically demonstrates strong performance, establishing theoretical guarantees of convergence remains an important goal. Formal convergence proofs would not only solidify the algorithm's theoretical foundations but also enhance its credibility and trustworthiness in practical applications.

In conclusion, this thesis provides a robust and efficient algorithm for posterior sampling and noise estimation in Bayesian inverse problems, based on interacting particle Langevin samplers and the EM algorithm.

Bibliography

[Bédard, 2007] Bédard, M. (2007). Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *The Annals of Applied Probability*, 17(4):1222 – 1244.

[Brown, 1828] Brown, R. (1828). A brief account of microscopical observations made in the months of june, july and august, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philosophical Magazine*, 4:161–173.

[Buck et al., 1996] Buck, C. E., Cavanagh, W. G., and Litton, C. D. (1996). *Bayesian approach to interpreting archaeological data*. Wiley.

[Chen et al., 2021] Chen, L., Lu, X., Zhang, J., Chu, X., and Chen, C. (2021). Hinet: Half instance normalization network for image restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 182–192, Los Alamitos, CA, USA. IEEE Computer Society.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

[Dong et al., 2011] Dong, W., Zhang, L., Shi, G., and Wu, X. (2011). Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857.

[Dowson and Landau, 1982] Dowson, D. and Landau, B. (1982). The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455.

[Eigel et al., 2022] Eigel, M., Gruhlke, R., and Sommer, D. (2022). Less interaction with forward models in langevin dynamics. arXiv:2212.11528.

[Einstein, 1905] Einstein, A. (1905). Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der Physik*, 322(8):549–560.

[Fergus et al., 2006] Fergus, R., Singh, B., Hertzmann, A., Roweis, S. T., and Freeman, W. T. (2006). Removing camera shake from a single photograph. *ACM Trans. Graph.*, 25(3):787–794.

[Garbuno-Inigo et al., 2019] Garbuno-Inigo, A., Hoffmann, F., Li, W., and Stuart, A. M. (2019). Interacting langevin diffusions: Gradient structure and ensemble kalman sampler.

[Garbuno-Inigo et al., 2020] Garbuno-Inigo, A., Nüsken, N., and Reich, S. (2020). Affine invariant interacting langevin dynamics for bayesian inference. *SIAM Journal on Applied Dynamical Systems*, 19(3):1633–1658.

[Gelman et al., 1997] Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110 – 120.

[Hadamard, 1902] Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13(4):49–52.

[Hagemann et al., 2024] Hagemann, P., Hertrich, J., Casfor, M., Heidenreich, S., and Steidl, G. (2024). Mixed noise and posterior estimation with conditional deepgem. arXiv:2402.02964.

[Jia, 2007] Jia, J. (2007). Single image motion deblurring using transparency. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

[Kolehmainen et al., 2007] Kolehmainen, V., Vanne, A., Siltanen, S., Järvenpää, S., Kaipio, J., Lassas, M., and Kalke, M. (2007). Bayesian inversion method for 3d dental x-ray imaging. *Elektrotechnik und Informationstechnik*, 124:248–253.

[Langevin, 1908] Langevin, P. (1908). Sur la théorie du mouvement brownien. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 146:530–533.

[Li et al., 2020] Li, D., Xu, K., Harris, J. M., and Darve, E. (2020). Coupled time-lapse full waveform inversion for subsurface flow problems using intrusive automatic differentiation.

[Marschall et al., 2023] Marschall, M., Wübbeler, G., Schmähling, F., and Elster, C. (2023). Machine learning based priors for Bayesian inversion in MR imaging. *Metrologia*, 60(4):044003.

[Matthews et al., 2016] Matthews, C., Weare, J., and Leimkuhler, B. (2016). Ensemble preconditioning for markov chain monte carlo simulation.

[Molesky et al., 2018] Molesky, S., Lin, Z., Piggott, A. Y., Jin, W., Vucković, J., and Rodriguez, A. W. (2018). Inverse design in nanophotonics. *Nature Photonics*, 12(11):659–670.

[Pedregosa, 2023] Pedregosa, F. (2023). On the convergence of the unadjusted langevin algorithm. `https://fa.bianp.net/blog/2023/ulaq`.

[Rawlinson et al., 2010] Rawlinson, N., Pozgay, S., and Fishwick, S. (2010). Seismic tomography: A window into deep earth. *Physics of the Earth and Planetary Interiors*, 178:101–135.

[Roberts and Rosenthal, 1998] Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(1):255–268.

[Rodgers, 2000] Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding: Theory and Practice*. WORLD SCIENTIFIC.

[Schillings and Stuart, 2017] Schillings, C. and Stuart, A. M. (2017). Analysis of the ensemble kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, 55(3):1264–1290.

[Wu, 1983] Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95 – 103.

[Zamir et al., 2021] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. (2021). Multi-stage progressive image restoration. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14816–14826.

A. Measure and Probability Theory

Here we will recall some basic definitions and concepts from measure and probability theory that are needed in order to give a rigorous definition of Bayesian inference.

In probability theory, a probability space is a fundamental construct that is defined as a tuple $(\Omega, \mathcal{F}, \mathbb{P})$. Here, Ω represents the sample space, $\mathcal{F}$ denotes the σ-algebra of events, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is the probability measure. The measure $\mathbb{P}$ is said to be σ-finite if the sample space Ω can be expressed as a countable union of sets within $\mathcal{F}$ each having finite measure. An important example of a σ-finite measure is the Lebesgue measure on $\mathbb{R}^d$.

A σ-algebra is a collection of sets that captures the notion of information, indicating whether an event has occurred or not. A sub-σ-algebra represents partial information about the underlying σ-algebra. Specifically, $\mathcal{B}(X)$, the Borel σ-algebra, is generated by the open sets of a topological space X.

A measurable map $U : \Omega \to X$ is defined such that for every set in $\mathcal{B}(X)$, the preimage under U is in $\mathcal{F}$. When $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, such a map U is often referred to as a random variable. The measure induced by U is given by $\mu(A) = \mathbb{P}(U^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega : U(\omega) \in A\})$ for $A \in \mathcal{B}(X)$. This measure μ is known as the probability distribution of U, denoted as $U \sim \mu$.

If μ and ν are measures on the same space, μ is said to be absolutely continuous with respect to ν (denoted $\mu \ll \nu$) if $\nu(A) = 0$ implies $\mu(A) = 0$. The measures μ and ν are equivalent if both $\mu \ll \nu$ and $\nu \ll \mu$ hold.

Theorem A.1 (Radon-Nikodym)**.** *Let μ and ν be two measures on the same measure space $(\Omega, \mathcal{F})$. If $\mu \ll \nu$ and ν is σ-finite, then there exists a unique function $g \in L^1_\nu$ such that for any measurable set $A \in \mathcal{F}$,*

$$\mu(A) = \int_A g \, d\nu.$$

The function g is called the Radon-Nikodym derivative of μ with respect to ν, denoted by $\frac{d\mu}{d\nu}$.

Example A.2. *Let μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and assume $\mu \ll \nu_L$, where ν_L is the Lebesgue measure on $\mathbb{R}^d$ (which is σ-finite). Then, there exists a unique function $g \in L^1(\mathbb{R}^d)$ such that for all $A \in \mathcal{B}(\mathbb{R}^d)$,*

$$\mu(A) = \int_A g(t) \, dt.$$

The function g is referred to as the probability density of $U \sim \mu$. △

Definition A.3 (Conditional expectation)**.** *Let $\mathcal{S} \subseteq \mathcal{F}$ be a sub-σ-algebra. A function $V : \Omega \to X$ is called a conditional expectation of $U : \Omega \to X$ with respect to $\mathcal{S}$ if V is $\mathcal{S}$-measurable and*

$$\int_S U \, d\mathbb{P} = \int_S V \, d\mathbb{P}$$

for all $S \in \mathcal{S}$. This is denoted as $\mathbb{E}(U \mid \mathcal{S}) = V$.

Definition A.4 (Conditional probability)**.** *Given a sub-σ-algebra $\mathcal{S}$ of $\mathcal{F}$, the conditional probability of $A \in \mathcal{B}(X)$ given $\mathcal{S}$ is defined by*

$$\mathbb{P}(A \mid \mathcal{S}) = \mathbb{E}(\mathbf{1}_A \mid \mathcal{S}),$$

where $\mathbf{1}_A$ is the indicator function of the set A.

Definition A.5 (Regular conditional distribution)**.** *A family of probability distributions $(\mu(\cdot, \omega))_{\omega \in \Omega}$ on $(X, \mathcal{B}(X))$ is termed a regular conditional distribution of U given $\mathcal{S} \subseteq \mathcal{F}$ if*

$$\mu(A, \cdot) = \mathbb{E}(\mathbf{1}_A(U) \mid \mathcal{S}) \text{ a. s.}$$

for every $A \in \mathcal{B}(X)$.

Theorem A.6. *Let $U : \Omega \to X$ be a random variable and $\mathcal{S} \subseteq \mathcal{F}$ a sub-σ-algebra. Then, there exists a regular conditional distribution $(\mu(\cdot, \omega))_{\omega \in \Omega}$ of U given $\mathcal{S}$.*

Consider the σ-algebra $\sigma(V) \subseteq \mathcal{F}$ generated by a random variable V. We can define the regular conditional probability measure as $\pi_{\text{post}}(A, V(\omega)) = \mathbb{E}(\mathbf{1}_A(U) \mid \sigma(V))(\omega)$. This measure serves as the posterior measure and can be identified with $\pi_{\text{post}}(A, v) = \pi_{\text{prior}}(A \mid v)$.

B. Description of the Posterior for Gaussian Prior and Linear Forward Map

A linear map $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^K$ can be represented by a matrix $G \in \mathbb{R}^{D\times K}$ such that $\mathcal{G}(u) = Gu$ for any $u \in R^D$. Assuming a Gaussian prior $\mu_{\text{prior}} \sim \mathcal{N}(m_{\text{prior}}, \mathcal{C}_{\text{prior}})$ and centered Gaussian observational noise $\eta \sim \mathcal{N}(\mathbf{0}_N, \Sigma)$ for observations $y = Gu + \eta$, the likelihood of data y being observed from parameter u is also Gaussian and is distributed according to $\pi(\cdot|u) \sim \mathcal{N}(Gu, \Sigma)$. According to Bayes' rule we have

$$\pi(u|y) = \pi_{\text{post.}}(u) = \frac{1}{Z}\pi(y|u)\pi_{\text{prior}}(u),$$

where $Z = \int_{R^D} \pi(y|u)\pi_{\text{prior}}(u)\,\mathrm{d}u$ is a normalizing constant. Using closed-form representations for the probability density functions of Gaussians we obtain

$$\begin{aligned}\pi_{\text{post.}}(u) &\propto \pi(y|u)\pi_{\text{prior}}(u)\\ &\propto \exp\left(-\frac{1}{2}(y-Gu)^\intercal\Gamma^{-1}(y-Gu)\right)\exp\left(-\frac{1}{2}(u-m_{\text{prior}})^\intercal\Sigma_{\text{prior}}^{-1}(u-m_{\text{prior}})\right)\\ &= \exp\left(-\frac{1}{2}\left((y-Gu)^\intercal\Gamma^{-1}(y-Gu)+(u-m_{\text{prior}})^\intercal\Sigma_{\text{prior}}^{-1}(u-m_{\text{prior}})\right)\right)\end{aligned}$$

Expanding the terms inside the exponent, we get

$$\begin{aligned}\pi_{\text{post}}(u \mid y) \propto \exp\Bigg(-\frac{1}{2}\Big(&u^\intercal G^\intercal\Sigma^{-1}Gu - 2u^\intercal G^\intercal\Sigma^{-1}y + y^\intercal\Sigma^{-1}y + u^\intercal{\mathcal{C}_{\text{prior}}}^{-1}u\\ &- 2u^\intercal{\mathcal{C}_{\text{prior}}}^{-1}m_{\text{prior}} + {m_{\text{prior}}}^\intercal{\mathcal{C}_{\text{prior}}}^{-1}m_{\text{prior}}\Big)\Bigg).\end{aligned}$$

Grouping terms in u, we find that the exponent is quadratic in u and can be rewritten as

$$\pi_{\text{post}}(u \mid y) \propto \exp\left(-\frac{1}{2}\left(u^\intercal(G^\intercal\Sigma^{-1}G + {\mathcal{C}_{\text{prior}}}^{-1})u - 2u^\intercal(G^\intercal\Sigma^{-1}y + {\mathcal{C}_{\text{prior}}}^{-1}m_{\text{prior}})\right)\right).$$

This is the kernel of a Gaussian distribution, and hence $\pi_{\text{post.}}(u|y)$ is Gaussian with covariance

$$\mathcal{C}_{\text{post.}} = (G^\intercal\Sigma^{-1}G + {\mathcal{C}_{\text{prior}}}^{-1})^{-1}$$

and mean

$$m_{\text{post}} = \mathcal{C}_{\text{post.}}(G^{\intercal}\Sigma^{-1}y + {\mathcal{C}_{\text{prior}}}^{-1}m_{\text{prior}}).$$

More details and alternative descriptions of the posterior are given in [Rodgers, 2000].

C. Derivation of the *M-step*

From eq. (3.9) we have

$$
\begin{aligned}
\sigma^2_{(r+1)} &= \arg\max_{\sigma^2} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{u\sim q_i^{(r+1)}}\Big[\log\big(p_{\sigma^2_{(r)}}(y_i|u)\; p(u)\big)\Big] \\
&= \arg\max_{\sigma^2} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{u\sim q_i^{(r+1)}}\Big[\log\big(p_{\sigma^2_{(r)}}(y_i|u)\big) + \log\big(p(u)\big)\Big] \\
&= \arg\max_{\sigma^2} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{u\sim q_i^{(r+1)}}\Big[\log\big(p_{\sigma^2_{(r)}}(y_i|u)\big)\Big] + \underbrace{\mathbb{E}_{u\sim q_i^{(r+1)}}\Big[\log\big(p(u)\big)\Big]}_{\text{independent of } \sigma^2} \\
&= \arg\max_{\sigma^2} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{u\sim q_i^{(r+1)}}\Big[\log\big(p_{\sigma^2_{(r)}}(y_i|u)\big)\Big] \\
&\approx \arg\max_{\sigma^2} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \log\big(p_{\sigma^2_{(r)}}(y_i|u_{ij})\big) \qquad (*) \\
&= \arg\max_{\sigma^2} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \log\Bigg(\frac{1}{\sqrt{(2\pi)^K \det\left(\sigma^2 I\right)}} \\
&\qquad\qquad \exp\Big(-\frac{1}{2}\big(y_i-\mathcal{G}(u_{ij})\big)^\intercal\,(\sigma^2 I)^{-1}\,\big(y_i-\mathcal{G}(u_{ij})\big)\Big)\Bigg) \\
&= \arg\max_{\sigma^2} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \log\Bigg(\frac{1}{\sqrt{(2\pi)^K \sigma^{2K}}} \exp\Big(-\frac{1}{2}\big(y_i-\mathcal{G}(u_{ij})\big)^\intercal\,(\sigma^{-2} I)\,\big(y_i-\mathcal{G}(u_{ij})\big)\Big)\Bigg) \\
&= \arg\max_{\sigma^2} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \log\Bigg((2\pi\sigma^2)^{-\frac{K}{2}} \exp\Big(-\frac{1}{2\sigma^2}\big(y_i-\mathcal{G}(u_{ij})\big)^\intercal\, I\,\big(y_i-\mathcal{G}(u_{ij})\big)\Big)\Bigg) \\
&= \arg\max_{\sigma^2} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \log\Bigg((2\pi\sigma^2)^{-\frac{K}{2}} \exp\Big(-\frac{1}{2\sigma^2}\big(y_i-\mathcal{G}(u_{ij})\big)^\intercal\,\big(y_i-\mathcal{G}(u_{ij})\big)\Big)\Bigg) \\
&= \arg\max_{\sigma^2} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \log\Bigg((2\pi\sigma^2)^{-\frac{K}{2}} \exp\Big(-\frac{1}{2\sigma^2}||y_i-\mathcal{G}(u_{ij})||^2\Big)\Bigg) \\
&= \arg\max_{\sigma^2} \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \log\Big((2\pi\sigma^2)^{-\frac{K}{2}}\Big) + \log\Big(\exp\Big(-\frac{1}{2\sigma^2}||y_i-\mathcal{G}(u_{ij})||^2\Big)\Big) \\
&= \arg\max_{\sigma^2} \frac{NM}{NM}\log\Big((2\pi\sigma^2)^{-\frac{K}{2}}\Big) + \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\Big(-\frac{1}{2\sigma^2}||y_i-\mathcal{G}(u_{ij})||^2\Big) \\
&= \arg\max_{\sigma^2} \log\Big((2\pi\sigma^2)^{-\frac{K}{2}}\Big) - \frac{1}{2NM\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M}||y_i-\mathcal{G}(u_{ij})||^2 \\
&= \arg\max_{\sigma^2} -\frac{K}{2}\log\big(2\pi\sigma^2\big) - \frac{1}{2NM\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M}||y_i-\mathcal{G}(u_{ij})||^2 \\
&= \arg\max_{\sigma^2} Q(\sigma^2, y, u),
\end{aligned}
$$

where in $(*)$ M samples $u_{ij} \in \mathbb{R}^D$ from $q_i^{(r+1)}$ are introduced for each observation y_i, $i = 1, \ldots, N$.

Proposition C.1. *Q has an extremum at*

$$\sigma_*^2 = \frac{\Lambda}{NMK}$$

with $\Lambda = \sum_{i=1}^{N} \sum_{j=1}^{M} ||y_i - \mathcal{G}(u_{ij})||^2$.

Proof.

$$\begin{aligned}
& 0 = \frac{\partial}{\partial \sigma^2} Q(\sigma_*^2, y, u) \\
\iff \quad & 0 = \frac{\partial}{\partial \sigma^2} \left[-\frac{K}{2} \log\left(2\pi\sigma_*^2\right) - \frac{1}{2NM\sigma_*^2} \sum_{i=1}^{N} \sum_{j=1}^{M} ||y_i - \mathcal{G}(u_{ij})||^2 \right] \\
\iff \quad & 0 = -\frac{K}{2\sigma_*^2} + \frac{1}{2NM(\sigma_*^2)^2} \sum_{i=1}^{N} \sum_{j=1}^{M} ||y_i - \mathcal{G}(u_{ij})||^2 \\
\iff \quad & \frac{K}{2\sigma_*^2} = \frac{1}{2NM(\sigma_*^2)^2} \sum_{i=1}^{N} \sum_{j=1}^{M} ||y_i - \mathcal{G}(u_{ij})||^2 \\
\iff \quad & \frac{K\sigma_*^2}{2} = \frac{1}{2NM} \sum_{i=1}^{N} \sum_{j=1}^{M} ||y_i - \mathcal{G}(u_{ij})||^2 \\
\iff \quad & K\sigma_*^2 = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} ||y_i - \mathcal{G}(u_{ij})||^2 \\
\iff \quad & \sigma_*^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} ||y_i - \mathcal{G}(u_{ij})||^2}{NMK} \\
& \qquad = \frac{\Lambda}{NMK}.
\end{aligned}$$

□

Proposition C.2. *$\sigma_*^2 = \frac{\Lambda}{NMK}$ is a maximizer of Q.*

Proof. For this σ_*^2 to be a maximizer of Q, it must satisfy $\frac{\partial^2}{\partial(\sigma^2)^2} Q(\sigma_*^2, y, u) < 0$.

$$\begin{aligned}\frac{\partial^2}{\partial(\sigma^2)^2}Q(\sigma^2, y, u) &= \frac{\partial}{\partial\sigma^2}\Big[-\frac{NK}{2\sigma^2} + \frac{\Lambda}{2(\sigma^2)^2}\Big] \\ &= \frac{NK}{2(\sigma^2)^2} - \frac{\Lambda}{(\sigma^2)^3}.\end{aligned}$$

Evaluating at $\sigma_*^2 = \frac{\Lambda}{NK}$, we get

$$\begin{aligned}\frac{\partial^2}{\partial(\sigma^2)^2}Q(\sigma_*^2, y, u) &= \frac{NK}{2(\sigma_*^2)^2} - \frac{\Lambda}{(\sigma_*^2)^3} \\ &= \frac{NK}{2\big(\frac{\Lambda}{NK}\big)^2} - \frac{\Lambda}{\big(\frac{\Lambda}{NK}\big)^3} \\ &= \frac{(NK)^3}{2\Lambda^2} - \frac{\Lambda(NK)^3}{\Lambda^3} \\ &= \frac{(NK)^3}{2\Lambda^2} - \frac{(NK)^3}{\Lambda^2} \\ &= \frac{(NK)^3}{2\Lambda^2} - \frac{2(NK)^3}{2\Lambda^2} \\ &= -\frac{(NK)^3}{2\Lambda^2} \\ &< 0,\end{aligned}$$

since $N, K, \Lambda > 0$. Thus, σ_*^2 is a maximizer of Q. □