Multilevel Optimization for Inverse Problems

Simon Weissmann\textsuperscript{1}, Ashia Wilson\textsuperscript{2}, and Jakob Zech\textsuperscript{1}

\textsuperscript{1}Universität Heidelberg, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, D-69120 Heidelberg, Germany, simon.weissmann@uni-heidelberg.de, jakob.zech@uni-heidelberg.de

\textsuperscript{2}Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Sciences, Cambridge, MA, 02139, USA, ashia07@mit.edu

May 2, 2022

Abstract

Inverse problems occur in a variety of parameter identification tasks in engineering. Such problems are challenging in practice, as they require repeated evaluation of computationally expensive forward models. We introduce a unifying framework of multilevel optimization that can be applied to a wide range of optimization-based solvers. Our framework provably reduces the computational cost associated with evaluating the expensive forward maps stemming from various physical models. To demonstrate the versatility of our analysis, we discuss its implications for various methodologies including multilevel (accelerated, stochastic) gradient descent, a multilevel ensemble Kalman inversion and a multilevel Langevin sampler. We also provide numerical experiments to verify our theoretical findings.

\textbf{keywords} multilevel methods, optimization, inverse problems

1 Introduction

Inverse problems are ubiquitous in applied mathematics and modern machine learning. The aim is usually to quantify information about unknown parameters which are indirectly observed through a noisy observation model. Solutions for inverse problems are often found
using optimization and sampling methods and crucially depend on an underlying physical model incorporated through a forward map. The physical models are typically highly complex such that associated numerical approximations come with extensive computational costs. The multilevel Monte Carlo method (MLMC) \cite{19,23} is a well-established variance reduction method, which addresses this issue by shifting a large part of the work to less accurate model evaluations. In the context of Bayesian inference, MLMC methods have been applied to Markov chain Monte Carlo (MCMC) methods \cite{14} as well as to deterministic quadrature rules such as sparse grid \cite{22,40} and quasi-Monte Carlo methods \cite{20,11}.

In this work, we apply similar ideas to the following general optimization problem

$$\min_{x \in \mathcal{X}} \Phi(x),$$

(1)

where $\mathcal{X}$ is a Hilbert space and $\Phi : \mathcal{X} \to \mathbb{R}_+$ an objective. The idea of multilevel optimization is to replace the evaluation of $\Phi$ (or its derivatives) by some approximation that becomes increasingly more accurate as the optimization process converges. Intuitively, when the current state may be far from minimum, it suffices to roughly move in the direction of the minimizer; however, as the state approaches the minimizer of $\Phi$, higher accuracy is required to reduce numerical bias. Multilevel optimization strategies are targeting efficient algorithms with the aim of reducing overall computational costs.

Such ideas have recently been applied in different contexts. The works closest to ours are \cite{28,27} which use multilevel optimization on an optimal control problem, and \cite{1}, where a multilevel version of the Stein variational gradient descent method is introduced. The aim of our manuscript is to formulate a unifying multilevel framework which can be applied to a wide range of optimization and sampling methods with particular focus on inverse problems.

**Contributions** Our principal contributions are three-fold:

- We formulate a multilevel strategy for general iterative optimization methods where each update step depends on an accuracy level. We derive an optimal choice of levels that minimizes computational costs while ensuring to achieve a certain tolerance for the error. Compared to the single-level framework, we prove that the computational cost can be reduced by a log-factor, and we provide an example to show that our results are sharp.

- We use our framework to introduce a multilevel ensemble Kalman inversion method and its extension to Tikhonov regularization. For linear forward models and with the incorporation of variance inflation, we prove convergence rates that reduce the computational costs by the expected log-factor when compared to single-level methods.

- We apply our framework to particle based sampling methods for Bayesian inference. We develop a multilevel formulation of interacting Langevin samplers. Viewing Langevin dynamics as gradient flow in the space of probability measures w.r.t. the Kullback-Leibler divergence, under certain assumptions we show convergence for the mean-field limit and provide a cost analysis that again reduces cost by a log-factor compared to the single-level method.
Outline §2 discusses optimization-based approaches for solving inverse problems while §3 presents our unified multilevel optimization framework. In §4 and §5 we apply our framework to particle based optimization and Bayesian inference respectively, and §6 presents numerical experiments for these examples.

Notation $f \lesssim g$ indicates the existence of $C$ such that $f(x) \leq Cg(x)$, with $C$ independent of $x$ in a certain range that will be clear from context. Moreover $f \simeq g$ iff $f \lesssim g$ and $g \lesssim f$.

2 Inverse Problems

Let $\mathcal{X}$ be a Hilbert space, $n_y \in \mathbb{N}$ and $F : \mathcal{X} \to \mathbb{R}^{n_y}$ the so-called forward model. We consider the task of recovering the unknown quantity $x \in \mathcal{X}$ from a (noisy) observation $y \in \mathbb{R}^{n_y}$ of $F(x)$. Throughout we assume an additive Gaussian noise model, i.e. $y$ is a realization of the random variable

$$Y = F(x) + \eta,$$

with $\eta \sim \mathcal{N}(0, \Gamma)$ Gaussian for a symmetric positive definite (SPD) covariance matrix $\Gamma \in \mathbb{R}^{n_y \times n_y}$.

This problem is typically ill-posed in the sense of [21], for instance because the dimension of the parameter space $\mathcal{X}$ may be much higher than the dimension $n_y$ of the observation space. We now recall two different methodologies to deal with these difficulties, both of which recast the problem into one of optimization.

2.1 Regularized optimization

One classical approach to approximate $x$ is to minimize the objective

$$\Phi(x) := \ell(x, y) + R(x), \quad \ell(x, y) = \frac{1}{2} \| \Gamma^{-1/2}(F(x) - y) \|_{\mathbb{R}^{n_y}}^2,$$

where $\ell$ denotes the least-squares data misfit loss functional and $R : \mathcal{X} \to \mathbb{R}_+$ is a regularizer. Common choices of regularization include Tikhonov regularization [16] with $R(x) = \frac{\lambda}{2} \| C_0x \|^2$ and total variation regularization [10, 31] with $R(x) = \frac{1}{2} \| \nabla x \|^2$. Note that prior information can be incorporated through $C_0 \in \mathcal{L}(\mathcal{X}, \mathcal{X})$. In the following, for fixed $y$, we use the shorthand $\ell(x) := \ell(x, y)$.

We continue this discussion in Sec. 4 where we present a particle based multilevel optimization method to minimize $\Phi$ in (3). For motivation and further discussion of regularization methods to solve (2), see, e.g., [15, 4] and references therein.

2.2 Bayesian inference

In the Bayesian approach (e.g. [34]) the parameter and observation are modeled as a joint random variable $(X, Y)$ on $\mathcal{X} \times \mathbb{R}^{n_y}$. The goal is to determine the posterior, which refers to the conditional distribution of $X$ given the realization $y$ of $Y$ in (2). Assume $X$ and
η to be stochastically independent, and let \( X \sim Q_0 \) for a prior distribution \( Q_0 \). Under certain technical assumptions [34, Theorem 6.31], the posterior \( Q_*^{prior} \) is then well-defined, absolutely continuous with respect to the prior, and \( Q_*^{prior}(dx) = \frac{1}{Z} \exp(-\ell(x,y))Q_0(dx) \), where \( Z = \int_X \exp(-\ell(x,y))Q_0(dx) \in \mathbb{R} \) is a normalizing constant and \( \ell \) is given by (3).

Suppose for the moment that \( X = \mathbb{R}^n \) is finite dimensional and the prior \( Q_0 = \mathcal{N}(0, \frac{1}{\lambda}C_0) \) is Gaussian, \( C_0 \in \mathbb{R}^{n_x \times n_x} \) SPD, \( \lambda > 0 \). Then, the posterior \( Q_*^{post} \) has Lebesgue density \( \rho_*^{post}(x) = \frac{1}{Z} \exp(-\ell(x,y) + R(x)) \), with \( R(x) = \frac{1}{2\lambda} ||C_0^{-1/2}x||^2_{\mathbb{R}^{n_x}} \). In the Bayesian framework, solving the inverse problem amounts to sampling from the posterior. One way to achieve this is by minimizing the objective

\[
\Phi(\psi) = KL(\psi || \rho_*^{post}),
\]

for \( \psi \) in a given family of (tractable) probability distributions on \( \mathcal{X} \). Here KL stands for the Kullback–Leibler divergence [26]. Hence, we end up again with an optimization problem, but this time over a subspace of the probability measures on \( \mathcal{X} \). This discussion will be continued in Sec. 5 where we present a multilevel optimization algorithm to minimize (4).

While the approaches in (3) and (4) are entirely different, we emphasize that the minimization of both objectives requires multiple evaluations of the forward model \( F \), which might be very costly in practice.

**Example 2.1.** Let \( \mathcal{X} = L^2(D) \) for a convex bounded polygonal domain \( D \subseteq \mathbb{R}^2 \). By classical PDE theory, for every \( f \in \mathcal{X} \), the equation

\[
\begin{aligned}
-\Delta u_f(s) + u_f(s) &= f(s) \quad s \in D, \\
u_f(s) &= 0 \quad s \in \partial D,
\end{aligned}
\]

has a unique weak solution \( u_f \in H^2(D) \cap H^1_0(D) \subseteq \mathcal{X} \). Let \( \mathcal{O} : \mathcal{X} \to \mathbb{R}^{n_y} \) be a bounded linear map called the observation operator. The forward model \( F(f) := \mathcal{O}(u_f) \in \mathbb{R}^{n_y} \) then “observes” the solution of (5) through the functional \( \mathcal{O} \).

Given noisy observations \( y = F(f) + \eta \) as in (2), any method minimizing the objectives in (3) or (4) has to access \( \Phi \) (or its derivatives) and thus repeatedly evaluate the forward model \( F \). Each such evaluation requires solving (5). Since (5) has no closed form solution, \( u_f \) can only be approximated using a numerical PDE solver such as the finite element method (FEM).

## 3 A unified multilevel optimization framework

In order to minimize an objective \( \Phi \) as in (1), we consider an abstract optimization method described by the fixed point iteration

\[
x_{k+1} := \Psi(x_k), \quad x_0 \in \mathcal{X}.
\]

For certain applications, an exact evaluation of \( \Psi \) is either not possible, or computationally infeasible. In such cases, typically numerical approximations \( \Psi_t \) to \( \Psi \) are available.
Here the “level” $l$ is a positive real number and can be understood as the accuracy of the approximation—the larger $l$ the closer $\Psi$ and $\Psi_l$ are. The precise meaning of this statement will be quantified in the following. At the same time, higher accuracy comes at higher computational cost, which is accounted for by the assumption that one evaluation of $\Psi_l$ amounts to computational cost $l$.

**Remark 3.1.** In practice $\Psi_l$ might only be available for certain $l \in \mathbb{N}$. For simplicity we allow $l \in \mathbb{R}$, $l > 0$, but mention that our analysis extends to the discrete case by rounding $l$ to the next larger admissible level.

Replacing $\Psi$ in the update rule (6) with $\Psi_l$ leads to

$$x_{k+1} = \Psi_l(x_k), \quad x_0 \in \mathcal{X}. \quad (7)$$

Here, $l_k \in \mathbb{N}$ is the level in iteration $k$ of the optimization process. The goal is to choose levels which minimize the overall computational cost while achieving fast convergence.

We denote the error of the $k$th iterate $x_k$ by $e_k$. For example, if $x_*$ is the unique minimizer of $\Phi$, $e_k$ could stand for $\|x_k - x_*\|_\mathcal{X}$, or for the distance of the objective to the minimum, i.e. $\Phi(x_k) - \Phi(x_*)$. Our analysis is based on the following abstract assumption. It can be understood as a form of linear convergence, up to an additive term stemming from the approximation of $\Psi$ by $\Psi_l$.

**Assumption 3.2.** There exists $c \in (0, 1)$ and $\alpha > 0$ such that for any choice of levels $l_k \geq 1$ and with $x_k$ as in (7),

(i) **error decay:** for all $k \in \mathbb{N}$

$$e_{k+1} \leq ce_k + l_k^{-\alpha}, \quad (8)$$

(ii) **cost model:** for all $k \in \mathbb{N}$, the cost of computing $x_k$ in (7) equals

$$\text{cost}(x_k) = \sum_{j=0}^{k-1} l_j. \quad (9)$$

We motivate Assumption (3.2) by verifying conditions (8) and (9) on several examples.

**Example 3.3.** Suppose that the objective $\Phi$ is $L$-smooth and $\mu$-strongly convex, and that for each $l \in \mathbb{N}$ we have access to functions $g_l : \mathcal{X} \to \mathcal{X}$ or random variables $G_l(x) \in \mathcal{X}$ for all $x \in \mathcal{X}$, such that for some $0 \leq \eta < \infty$

$$\|\nabla \Phi(x) - g_l(x)\| \leq \frac{l^{-\alpha}}{\eta} \quad \text{or} \quad \mathbb{E}\|\nabla \Phi(x) - G_l(x)\| \leq \frac{l^{-\alpha}}{\eta} \quad \forall x \in \mathcal{X}. \quad (10)$$

Then gradient descent using the approximate gradients, i.e. iterates generated by

$$x_{k+1} = x_k - \eta_k g_l(x_k)$$

can be shown to satisfy error decay (8). Interpreting $l_k$ as the computational cost of evaluating $g_l$, the overall cost to compute $x_k$ follows our cost model (9). Similarly, accelerated gradient descent and the stochastic versions of both algorithms with dynamically increasing batch sizes can be shown to satisfy (8).
Details for Example 3.3 and the implications of our results for gradient descent, accelerated gradient descent and their stochastic versions are given in Appendix B.

**Remark 3.4.** Assumption 3.2 states that the relation between computational cost and corresponding error is of the type “error $\sim \text{cost}^{-\alpha}$” for some $\alpha > 0$. To clarify, consider the following examples in the context of applying gradient descent to minimize $\Phi$:

- **Fix $\gamma > 0$.** Suppose we have access to an algorithm, that for $n \in \mathbb{N}$ requires computational cost $f(n) := n^{\gamma}$ to compute $\nabla \Phi$ up to accuracy $n^{-\alpha}$. With $l := n^{\gamma}$, this is equivalent to saying that at level $l$ the error is of order $l^{-\alpha/\gamma}$, which fits our setting. Without loss of generality we can work with $l$ rather than $n$.

- **Suppose we have access to an algorithm that requires time $t > 0$ to approximate $\nabla \Phi$ up to accuracy $t^{-\alpha}$.** Then the level $l_k$ can be understood as the CPU time $t$ invested in the approximate computation of $\nabla \Phi(x_k)$.

Next, we discuss gradient descent for our running Example 2.1. Further details are contained in Appendix C.

**Example 3.5 (Continuation of Example 2.1).** Let $F(f) = \mathcal{O}(u_f)$, where $u_f$ solves (5) and $\mathcal{O} : L^2(D) \to \mathbb{R}^n$ is bounded linear, i.e. $\mathcal{O}(p) = \int_D \xi p$ for some $\xi \in L^2(D, \mathbb{R}^n)$ and all $p \in L^2(D)$.

Let $\Phi(f) = \frac{1}{2} \| \Gamma^{-1/2}(F(f) - y) \|^2_{\mathbb{R}^n} + \frac{1}{2} \| f \|^2_{L^2(D)}$ as in (3). Then

$$\nabla \Phi(f) = u_h + \lambda f \in L^2(D),$$

(11)

where $u_h$ solves (5) with right-hand side $h(\cdot) = (\mathcal{O}(u_f) - y) \Sigma^{-1} \xi(\cdot) \in L^2(D)$. To approximate $\nabla \Phi(f)$, we use linear finite elements on a uniform mesh on $D \subseteq \mathbb{R}^2$ to first obtain an approximation $u'_f$ satisfying $\| u_f - u'_f \|_{L^2(D)} \leq l^{-1}$, and subsequently with $\tilde{h} = (\mathcal{O}(u'_f) - y) \Sigma^{-1} \xi$, a FEM approximation $u'_h$ satisfying $\| u_h - u'_h \|_{L^2(D)} \leq l^{-1}$. Here $l$ corresponds to the dimension of the FEM space, and can thus be interpreted as the complexity of computing $u'_f$. Note, $g_l(f) := u'_h + \lambda f \in L^2(D)$ yields an approximation to $\nabla \Phi(f)$ s.t. $\| \nabla \Phi(f) - g_l(f) \|_{L^2(D)} \leq l^{-1}$. Hence (for fixed $f$ and up to a constant) $g_l(f)$ satisfies the first inequality in (10) with $\alpha = 1$.

Having established the basic setting, we next illustrate how accuracy levels $l_j$ can be chosen optimally to minimize computational costs. Recursively expanding (8), we get the following upper bound on the error

$$e_k \leq c(e_{k-2} + l_{k-1}^{-\alpha}) + l_k^{-\alpha} \leq \ldots \leq c^k e_0 + \sum_{j=0}^{k-1} c^{k-1-j} l_j^{-\alpha} =: \tilde{e}_k((l_j)_j).$$

(12)

In case $l_j = l$ for all $j = 1, \ldots, K - 1$, we will also use the notation $\tilde{e}_k(l)$.

We next determine levels achieving (almost) minimal cost under the constraint $\tilde{e}_k \leq \varepsilon$. 

6
3.1 Single-level

Fix the number of iteration steps $K \in \mathbb{N}$. For the single-level method, the level $l_j$ is fixed at a (single) value $\bar{l}_K > 0$ throughout the whole iteration $j = 0, \ldots, K - 1$. By assumption \eqref{9}, $\text{cost}(x_K) = K \bar{l}_K$. We wish to minimize the cost under the error constraint $\bar{e}_K \leq \varepsilon$. To slightly simplify the problem for the moment, we instead demand both terms in the definition of $\tilde{e}_K$ in \eqref{12} to be bounded by $\varepsilon$ (so that in particular $\bar{e}_K \leq \varepsilon$). More precisely, $\bar{l}_K > 0$ should be minimal such that

$$c^K e_0 \leq \varepsilon, \quad (\bar{l}_K)^{-\alpha} \frac{1 - c^K}{1 - c} \leq \frac{\varepsilon}{2}. \quad (13)$$

The first inequality implies $K \geq \frac{\log(\varepsilon/(2e_0))}{\log(c)}$, and the second inequality implies

$$\bar{l}_K \geq \left(2 \frac{1 - c^K}{(1 - c)\varepsilon} \right)^{1/\alpha}. \quad (14)$$

We choose $\bar{l}_K$ so that \eqref{14} holds with equality. Given $K \mapsto \text{cost}(x_K) = Kl_K$ is monotonically increasing, in order to get error $\varepsilon$ at possibly small cost the following choices suffice:

$$\bar{l}_K(\varepsilon) := \left(2 \frac{1 - \varepsilon}{(1 - c)\varepsilon} \right)^{1/\alpha}, \quad K(\varepsilon) := \left\lceil \frac{\log(\varepsilon/(2e_0))}{\log(c)} \right\rceil. \quad (15)$$

We next introduce a notion of optimality, and then summarize our observations in Theorem 3.7.

**Definition 3.6** (Quasi-optimal single level choice). A family of reals $\bar{l}_K(\varepsilon) > 0$ and integers $K(\varepsilon) \in \mathbb{N}$ satisfying $\bar{e}_K(\varepsilon)(\bar{l}_K(\varepsilon)) \leq \varepsilon$ for all $\varepsilon > 0$, is a quasi-optimal single level choice iff

$$K(\varepsilon)\bar{l}_K(\varepsilon) = O\left(\inf\left\{ \hat{K}\hat{l} : \bar{e}_K(\hat{l}) \leq \varepsilon \right\}, \hat{K} \in \mathbb{N}, \hat{l} > 0 \right) \quad \text{as} \quad \varepsilon \to 0.$$

**Theorem 3.7** (Single-level convergence). Equation \eqref{15} defines a quasi-optimal single level choice. It holds

$$\text{cost}_{\text{SL}}(\varepsilon) := K(\varepsilon)\bar{l}_K(\varepsilon) \simeq \log(\varepsilon^{-\frac{1}{\alpha}}) \varepsilon^{-\frac{1}{\beta}} \quad \text{as} \quad \varepsilon \to 0. \quad (16)$$

**Proof.** For the proof see Appendix A.1.

Due to the quasi-optimality, the (single-level-) cost behaviour \eqref{16} cannot be improved as $\varepsilon \to 0$.

3.2 Multilevel

Fix the number of iterations $K \in \mathbb{N}$. We now allow for varying levels throughout the optimization process. That is, we wish to find $l_{K,j}(\varepsilon) = l_{K,j} > 0$ such that $\text{cost}(x_K) =$
\[ \sum_{j=0}^{K-1} l_{K,j} \] is minimized under the constraint of both terms in (12) being bounded by \( \frac{\varepsilon}{2} \), i.e. such that
\[ c^K e_0 \leq \frac{\varepsilon}{2}, \quad \sum_{j=0}^{K-1} c^{K-1-j} l_{K,j} \leq \frac{\varepsilon}{2} \] (17)

The first condition gives again a lower bound on the number of iterations \( K \) as in Sec. 3.1. Minimizing cost \( (x_k) \) under the second condition gives:

**Lemma 3.8.** For every \( K \in \mathbb{N}, \varepsilon > 0 \)
\[ l_{K,j}(\varepsilon) = C_{K,\varepsilon} \cdot c^{K-1-j} K, j \quad C_{K,\varepsilon} = \left( \frac{\varepsilon}{2} \right) - \frac{1}{\alpha} \left( \frac{1 - c^{\frac{K}{1+\alpha}}}{1 - c^{\frac{1}{1+\alpha}}} \right) \] (18)
minimizes \( \sum_{j=0}^{K-1} l_{K,j} \) under the constraint \( \sum_{j=0}^{K-1} c^{K-1-j} l_{K,j} \leq \frac{\varepsilon}{2} \).

**Proof.** For the proof see Appendix A.2.

Let us compute the cost. Since \( \sum_{j=0}^{K-1} \delta^{K-1-i} = \frac{1-\delta^K}{1-\delta} \) for \( \delta = c^{\frac{1}{1+\alpha}} \in (0, 1) \),
\[ \sum_{j=0}^{K-1} l_{K,j} = C_{K,\varepsilon} \sum_{j=0}^{K-1} c^{K-1-j} K, j \quad C_{K,\varepsilon} = \left( \frac{\varepsilon}{2} \right) - \frac{1}{\alpha} \left( \frac{1 - c^{\frac{K}{1+\alpha}}}{1 - c^{\frac{1}{1+\alpha}}} \right) \] (19)
As in the single-level case, this term increases in \( K \) (although it remains bounded as \( K \to \infty \)). To keep the cost minimal, we choose \( K \) minimal under the first constraint in (17), which leads to
\[ K(\varepsilon) = \left\lceil \frac{\log(\varepsilon/(2e_0))}{\log(c)} \right\rceil, \quad l_{K,j}(\varepsilon) = \left( \frac{\varepsilon}{2} \right) - \frac{1}{\alpha} \left( \frac{1 - c^{\frac{K}{1+\alpha}}}{1 - c^{\frac{1}{1+\alpha}}} \right) \] (20)\[ \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) = O\left( \inf \left\{ \sum_{j=0}^{\hat{K}-1} \hat{l}_j : \tilde{e}_K((\hat{l}_j)_{j<\hat{K}}) \leq \varepsilon, \hat{K} \in \mathbb{N}, \hat{l}_j > 0 \forall j < \hat{K} \right\} \right) \] as \( \varepsilon \to 0 \).

**Definition 3.9 (Quasi-optimal multilevel choice).** A family \( ((l_{K,j}(\varepsilon))_{j<K(\varepsilon)})_{\varepsilon>0} \) of sequences satisfying \( \tilde{e}_K((l_{K,j}(\varepsilon))_{j<\hat{K}}) \leq \varepsilon \) for all \( \varepsilon > 0 \) is a quasi-optimal multilevel choice, iff
\[ \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) = O\left( \inf \left\{ \sum_{j=0}^{\hat{K}-1} \hat{l}_j : \tilde{e}_K((\hat{l}_j)_{j<\hat{K}}) \leq \varepsilon, \hat{K} \in \mathbb{N}, \hat{l}_j > 0 \forall j < \hat{K} \right\} \right) \] as \( \varepsilon \to 0 \).

**Theorem 3.10 (Multilevel convergence).** Equation (20) defines a quasi-optimal multilevel choice for \( \varepsilon \in (0, e_0) \). It holds
\[ \text{cost}_{ML}(\varepsilon) := \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha}} \quad \text{as} \quad \varepsilon \to 0. \] (21)
Proof. For the proof see Appendix A.3.

Due to the quasi-optimality, the asymptotic cost behaviour $O(\varepsilon^{-1/\alpha})$ required to achieve error $\tilde{e}_K \leq \varepsilon$ cannot be improved. Comparing with the single-level method in Theorem 3.10, we observe that the multilevel method decreases the computational cost by a factor $\log(\varepsilon^{-1})$. In practice and for small $\varepsilon > 0$, this can amount to a significant speedup as we will see in our numerical examples.

In Appendix B, we provide further details of the implications of Theorem 3.10 for gradient descent, accelerated gradient descent and the stochastic versions of these algorithms. As an application we discuss a stochastic gradient descent algorithm that uses increasing batch sizes in Example B.5.

Remark 3.11. Suppose that $e_k$ generated with levels $l_j$ satisfies instead of (8) the relaxed condition $e_{k+1} \leq ce_k + Cl^{-\alpha}_k$ for some constant $C \geq 1$. Then, $\tilde{e}_k$ generated with the levels $\tilde{l}_k := C^{1/\alpha}l_k$ satisfies $\tilde{e}_{k+1} \leq c\tilde{e}_k + Cl^{-\alpha}_k = c\tilde{e}_k + l^{-\alpha}_k$. The cost quantity $\sum_{j=0}^{k-1} l_j$ only increases by the constant factor $C^{1/\alpha}$ compared to $\sum_{j=0}^{k-1} \tilde{l}_j$. Hence, the asymptotic cost behaviour stated in Theorem 3.7 (single-level) and Theorem 3.10 (multi-level) remains valid also for $C > 1$.

We next continue our discussion of Example 3.5, for details see Appendix C.4.

Example 3.12 (Continuation of Example 3.5). The regularized objective $\Phi$ in Example 3.5 is $\lambda$-strongly convex and $L$-smooth with $L = \|\xi\|^2_{L^2(D)}\|\Gamma^{-1}\|+\lambda$. Consider the multilevel gradient descent method $f_{j+1} = f_j - \eta g_l(f_j)$, where $g_l$ is the approximation to $\nabla\Phi$ from Example 3.5. The level choice (20) then yields $\|f_\ast - f_{K(\varepsilon)}\|_{L^2(D)} \lesssim \varepsilon$, for the unique minimizer $f_\ast$ of $\Phi$. The cost quantity, which corresponds to the aggregated computational cost of all required FEM approximations to compute $f_{K(\varepsilon)}$, behaves like $\varepsilon^{-1}$ as $\varepsilon \to 0$.

Finally we point out that our analysis and notion of quasi-optimality are based on the constraint $\tilde{e}_k \leq \varepsilon$ (rather than $e_k \leq \varepsilon$), where $\tilde{e}_k$ is an upper bound of the actual error $e_k$. In Appendix D, we give a concrete example of biased gradient descent to show that the cost asymptotics in (21) is in general sharp for the actual error $e_k$ as well.

4 Particle based optimization: A multilevel ensemble Kalman inversion

As our first application, we present a multilevel ensemble Kalman inversion (EKI) to solve the problem presented in Sec. 2.1. EKI is a derivative free particle based optimization method, e.g., [33, 7, 25]. We first recall the method, and subsequently present a multilevel version.

4.1 Ensemble Kalman inversion

EKI refers to a specific dynamical system describing the evolution of an ensemble of particles. By the well-known subspace property [24], these particles remain within the finite
dimensional affine subspace spanned by the ensemble at initialization. Therefore, there is no loss of generality in assuming $\mathcal{X} = \mathbb{R}^{n_x}$ finite dimensional throughout this section.

We formulate the EKI as a method to minimize the objective

$$\Phi(x) = \frac{1}{2} \| \Sigma^{-1/2}(H(x) - z) \|^2. \quad (22)$$

Here $H : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_z}$ is the forward model and $\Sigma \in \mathbb{R}^{n_x \times n_x}$ is SPD. Letting

$$H = F, \quad z = y \in \mathbb{R}^{n_y}, \quad \Sigma = \Gamma, \quad (23)$$

(22) corresponds to the objective in (3) with $R = 0$ (i.e. unregularized). Fixing an SPD matrix $C_0 \in \mathbb{R}^{n_x \times n_x}$,

$$H = \left( \begin{array}{c} F \\ \text{Id} \end{array} \right), \quad z = \left( \begin{array}{c} y \\ 0_{n_z} \end{array} \right), \quad \Sigma = \left( \begin{array}{cc} \Gamma & 0 \\ 0 & \frac{1}{\lambda} C_0 \end{array} \right), \quad (24)$$

yields the regularized objective $\Phi(x)$ in (3) with regularizer $R(x) = \frac{1}{2} \| C_0^{-1/2} x \|^2$. We refer to EKI applied to the unregularized and regularized objective as standard EKI and Tikhonov regularized EKI (TEKI), respectively. See [8, 38] for more details on TEKI.

We consider the continuous-time formulation of EKI [6, 5]: For a fixed ensemble size $M \in \mathbb{N}$, let $v_t^{(m)} \in \mathbb{R}^{n_x}$, $m = 1, \ldots, M$, satisfy the coupled system of stochastic differential equations (SDEs)

$$dv_t^{(m)} = C^{w,H}(v_t) \Sigma^{-1}(z - H(v_t^{(m)})) dt + C^{w,H}(v_t) \Sigma^{-1/2} dW_t^{(m)}, \quad (25)$$

Here $W_t^{(m)}$ are independent $\mathbb{R}^{n_z}$-valued Brownian motions, $Q_0$ is a fixed initial distribution with finite second moment on $\mathbb{R}^{n_x}$, and $C^{w,H} \in \mathbb{R}^{n_x \times n_x}$ denotes a mixed sample covariance, see Appendix E for the precise formula. Under certain assumptions, it can be shown that (25) is well-posed, i.e. existence of unique and strong solutions can be guaranteed, and their average converges to the minimizer of $\Phi$ (7).

To motivate this behaviour, suppress for the moment the diffusion term in (25), and consider a linear forward map $H \in \mathbb{R}^{n_z \times n_x}$. Then

$$\frac{dv_t^{(m)}}{dt} = -C(v_t) H^T \Sigma^{-1}(H v_t^{(m)} - z) = -C(v_t) \nabla_v \Phi(v_t^{(m)}),$$

for $\Phi(x) = \frac{1}{2} \| \Sigma^{-1/2}(H x - z) \|^2$. Hence, in the linear and deterministic setting the EKI is a preconditioned gradient flow w.r.t. the data misfit $\ell$ (or w.r.t. the Tikhonov regularized data misfit). We refer to [9, 37] for more details on the nonlinear setting.
4.2 Multilevel ensemble Kalman inversion

To formulate the multilevel EKI, assume given approximations $H_l$, $l > 0$, to $H$ in (25). Fixing $\tau > 0$, with $t_j := j \cdot \tau$ we denote by $v^{(m),l_j}_{t_j}$ the solution of the coupled system of SDEs in integral form

$$v^{(m),l_j}_{t_{j+1}} = v^{(m),l_j}_{t_j} + \int_{t_j}^{t_{j+1}} C^{v,H_{l_j}}(v^l_t)\Sigma^{-1}(z - H_{l_j}(v^{(m),l_j}_t)) \, dt + \int_{t_j}^{t_{j+1}} C^{v,H_{l_j}}(v^l_t)\Sigma^{-1/2} \, dW^{(m)}_t,$$

initialized via $v^{(m),l_j}_{t_j} = v^{(m),l_j}_{t_{j-1}}$. We then introduce a discrete-time process $(x_j)_j$ as the ensemble mean of the particles at time $t_j$:

$$x_j := \frac{1}{M} \sum_{m=1}^M v^{(m),l_j}_{t_j}.$$

We discuss a discretization scheme for the SDE (26) and summarize the multilevel EKI as an algorithm in Appendix F. The goal in the following is to show that $x_j$ approximately minimizes the objective (22) for large $j$.

4.3 Error analysis for linear forward operators

Assume $F$ to be linear, specifically $F \in \mathbb{R}^{n_y \times n_x}$ with $n_y < n_x$ and $\text{rank}(FF^\top) = n_y$. We make the following assumptions on the forward operator arising through numerical approximation which is typically satisfied for ODE- or PDE-based forward models. We verify this assumption for Example 2.1 in Appendix C, but emphasize that this is typically satisfied and known for numerical approximations (e.g., using finite elements or boundary elements) to forward maps described by PDE solutions, see for example [17, 32].

**Assumption 4.1** (Approximation of the forward operator). There exist $b > 0$ and $\alpha > 0$ such that for each $l > 0$ there exists $F_l \in \mathbb{R}^{n_y \times n_x}$ satisfying $\|F - F_l\|_{\mathbb{R}^{n_y \times n_x}}^2 \leq bl^{-\alpha}$.

In the following we use the notation $H_l$ to denote $H$ as in (23) or (24) but with $F$ replaced by $F_l$, i.e. $H_l \in \mathbb{R}^{n_z \times n_x}$. We consider EKI and TEKI with covariance inflation, which is a standard data assimilation tool to stabilize the scheme, e.g., [2, 3, 35]. Specifically, for some fixed SPD matrix $B \in \mathbb{R}^{n_z \times n_x}$, (26) is replaced by the stabilized dynamics

$$v^{(m),l_j}_{t_{j+1}} = \int_{t_j}^{t_{j+1}} (C(v^l_t) + B)H_{l_j}^\top\Sigma^{-1}(z - H_{l_j}(v^{(m),l_j}_t)) \, dt + \int_{t_j}^{t_{j+1}} C(v^l_t)H_{l_j}^\top\Sigma^{-1/2} \, dW^m_t.$$

Assume that an evaluation of $F_l$ has cost $O(l)$. Then, approximating (28) using for example an Euler-Maruyama scheme with a fixed number of time steps $T$, requires $O(M \cdot T)$ evaluations of $F_l$. With $M$ and $T$ interpreted as constants, this amounts to cost $O(l)$. In this sense the cost quantity $\sum_{i=0}^{K-1} l_j$ introduced in (9), can be interpreted as the computational cost of computing $x_K$ in (27). We give convergence result for EKI and TEKI in the case of noise-free and noisy data respectively.
Let $x_j$ in \((27)\) be the mean of the particle system driven by \((28)\), with $H$ (and $H_l$) as in \((23)\).

**Proposition 4.2 (Multilevel EKI).** Let Assumption 4.1 be satisfied, $y = Fx^\dagger \in \mathbb{R}^{n_y}$ for some truth $x^\dagger \in \mathbb{R}^{n_x}$ and let $Q_0$ have finite second moment on $\mathbb{R}^{n_x}$. For $\tau > 0$ sufficiently small, there exists $c \in (0,1)$ depending on $F$ and $B$ such that for levels $l_{K,j}(\varepsilon)$, $j = 0, \ldots, K(\varepsilon) - 1$, given by \((20)\),

$$e_{K(\varepsilon)} := \mathbb{E}[\Phi(x_{K(\varepsilon)})] - \Phi(x^\dagger) = \mathbb{E}[\|\Gamma^{-1/2}(Fx_{K(\varepsilon)} - y)\|_{R^{n_y}}^2] \leq \varepsilon,$$

for all small enough $\varepsilon > 0$. Furthermore, $\text{cost}_{\text{ML}}(\varepsilon) = \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha}}$.

**Proof.** For the proof see Appendix G.1.

Now let $x_j$ in \((27)\) be the mean of the particle system driven by \((28)\), with $H$ (and $H_l$) as in \((24)\).

**Proposition 4.3 (Multilevel TEKI).** Let Assumption 4.1 be satisfied, $x^* \in \mathbb{R}^{n_x}$ be the unique minimizer of $\Phi$ in \((22)\) with $H$ given by \((24)\), and let $Q_0$ have finite second moment on $\mathbb{R}^{n_x}$. For $\tau > 0$ sufficiently small there exists $c \in (0,1)$ depending on $F$ and $B$ such that for levels $l_{K,j}(\varepsilon)$, $j = 0, \ldots, K(\varepsilon) - 1$, given by \((20)\),

$$e_{K(\varepsilon)} := \mathbb{E}\left[\frac{1}{2}\|\Gamma^{-1/2}F(x_{K(\varepsilon)} - x^*)\|_{R^{n_y}}^2 + \frac{\lambda}{2}\|\Gamma^{-1/2}(x_{K(\varepsilon)} - x^*)\|_{R^{n_x}}^2\right] \leq \varepsilon,$$

for all small enough $\varepsilon > 0$. Furthermore, $\text{cost}_{\text{ML}}(\varepsilon) = \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha}}$.

**Proof.** For the proof see Appendix G.2.

In both cases there holds an analogous statement for the single-level choice in \((15)\) with (the worse) asymptotic cost behaviour $\varepsilon^{-1/\alpha} \log(\varepsilon^{-1})$ as $\varepsilon \to 0$.

### 5 Optimization over probability measures: Multilevel Bayesian inference

As our second application, we consider an interacting particle system, to solve the problem presented in Sec. 2.2. To derive the method, we apply the multilevel optimization of Sec. 3 to a gradient flow in the space of probability measures. Appendix H gives details on the algorithm.

Throughout this section we adopt the assumptions of Sec. 2.2, i.e. $X = \mathbb{R}^{n_x}$, the forward model $F : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ may be nonlinear and the prior $Q_0 \sim \mathcal{N}(0, \frac{1}{\lambda}C_0)$ is Gaussian. We denote it’s density by $q_0$. The posterior density then equals

$$\rho_\varepsilon(x) = \frac{1}{Z} \exp(\ell_R(x)), \quad \ell_R(x) := \frac{1}{2}\|\Gamma^{-1/2}(F(x) - y)\|_{R^{n_y}}^2 + \frac{\lambda}{2}\|\Gamma^{-1/2}x\|_{R^{n_x}}^2$$

\((29)\)
with \( Z = \int_{\mathbb{R}^n_x} \exp(\ell_R(x)) \, dx \). To explain the idea of approximating \( \rho * \) by an ensemble of particles, we first recall the Langevin dynamics. Let \( v_t \in \mathbb{R}^{n_x} \) initialized as \( v_0 \sim q_0 \) solve
\[
dv_t = -\nabla_x \ell_R(v_t) \, dt + \sqrt{2} \, dW_t, \quad v_0 \sim q_0.
\] (30)
The evolution of its distribution \( \rho_t \) is described by the Fokker-Planck equation, see e.g. [30],
\[
\partial_t \rho_t = \nabla \cdot (\rho_t \nabla \ell_R) + \Delta \rho_t, \quad \rho_0 = q_0.
\] (31)
Under certain assumptions on \( \ell_R \) the Markov process \( v_t \) is ergodic and \( \rho * \) is its unique invariant distribution. Furthermore, it is possible to describe the rate of convergence in terms of the gradient flow structure given by [31], see for example [30] Theorem 4.9.

5.1 Interacting Langevin sampler

We consider an interacting Langevin particle system introduced in [18]. Let \( (v^{(m)}_t)_{m=1}^M \) solve
\[
dv^{(m)}_t = -C(v_t) \nabla \ell_R(v^{(m)}_t) \, dt + \sqrt{2C(v_t)} \, dW^{(m)}_t, \]
with \( v^{(m)}_0 \sim q_0 \) i.i.d., where \( (W^{(m)}_t)_{m=1}^M \) denote independent Brownian motions on \( \mathbb{R}^{n_x} \) and with \( C(v) \) the empirical covariance, see Appendix E. Observe that the mean field limit satisfies the following preconditioned version of (30)-(31) (e.g. [13]):
\[
dv_t = -C(\rho) \nabla \ell_R(v_t) \, dt + \sqrt{2C(\rho)} \, dW_t,
\]
where
\[
m(\rho) = \int_{\mathbb{R}^n_x} \nu \rho(v) \, dv, \quad C(\rho) = \int_{\mathbb{R}^n_x} ((v - m(\rho)) \otimes (v - m(\rho))) \rho(v) \, dv
\] (33)
and
\[
\partial_t \rho_t = \nabla \cdot (\rho_t C(\rho_t) \nabla \ell_R) + \text{Tr}(C(\rho_t) D^2 \rho_t), \quad \rho_0 = q_0.
\] (34)

5.2 Mean-field and discrete multilevel interacting Langevin sampler

Fix \( \tau > 0 \) and set \( t_j := j \cdot \tau \). In the following let \( \ell_k^j \) be given by [29] with the forward model \( F : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y} \) replaced by an approximation \( F_k \). We consider the discrete time process \( (\rho_j)_{j=1}^K \) iteratively defined by \( \rho_{j+1} = \rho_{ij_{j+1}} \), where \( \rho_{0j} = \rho_{0} \) and \( \rho_{ij} \) solves (34) on a time interval of length \( \tau \) initialized with the previous distribution, i.e.
\[
\partial_t \rho_{ij} = \nabla \cdot (\rho_{ij} C(\rho_{ij}) \nabla \ell_R) + \text{Tr}(C(\rho_{ij}) D^2 \rho_{ij}), \quad \rho_{ij_{0}} = \rho_{ij_{-1}}.
\] (35)

While the subsequently discussed analysis will be based on the mean field limit (35), we present a discretized version by introducing a particle based approximation based on [32]. To this end let \( \hat{\rho}_{ij} \) iteratively be defined by
\[
\hat{\rho}_{j+1} = \frac{1}{M} \sum_{m=1}^M \delta_{v^{(m)}_{ij_{j+1}}},
\]
where $\delta_v$ denotes the dirac-measure at $v$ and $v^{(m),l_j}_{t_{j+1}}$ initialized by $v^{(m),l_j}_{t_j} = v^{(m),l_j-1}_{t_j}$ solves

$$v^{(m),l_j}_{t_{j+1}} = v^{(m),l_j}_{t_j} - \int_{t_j}^{t_{j+1}} C(v^l_t) \nabla R(v^l_t) \, dt + \int_{t_j}^{t_{j+1}} \sqrt{2C(v^l_t)} \, dW^m_t, \quad m = 1, \ldots, M.$$  \hfill (36)

### 5.3 Mean-field error analysis for the Kullback–Leibler divergence

Under convexity assumptions on $\ell_l^R$ and assuming that $C(\rho^l_j)$ does not degenerate, one can prove exponential convergence to equilibrium for the mean-field limit of the interacting Langevin dynamics for any fixed level $l$ \cite[Proposition 2]{18}. These assumptions can for example be verified for linear forward models and Gaussian priors \cite[Corollary 5]{18}. We make the following additional assumption:

**Assumption 5.1** (Approximation of the forward operator). There exist $b > 0$ and $\alpha > 0$ such that for each $j \in \mathbb{N}$ there exists $F^j : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_y}$ satisfying

$$\int_{\mathbb{R}^{n_x}} \| F(x) - F^j(x) \|^2 \rho^l_j(x) \, dx \leq bl^{-\alpha}_j.$$ 

For linear forward models and Gaussian priors, the pdf $\rho^l_j$ remains Gaussian and Assumption 5.1 can be inferred from Assumption 4.1. Extending the convergence result of \cite{18} to error decay \cite{8} allows us to apply our results from Sec. 3. This leads to the following Proposition:

**Proposition 5.2** (Multilevel interacting Langevin sampler). Suppose $\text{KL}(q_0||\rho_*) < \infty$ and let Assumption 5.1 be satisfied. Suppose there exist $\sigma_1, \sigma_2 > 0$ such that $\nabla^2 \ell^l_R > \sigma_1 \text{Id}$ and $C(\rho^l_j) > \sigma_2 \text{Id}$ (cp. (33)) for all $j \geq 0$. For $\tau > 0$ sufficiently small there exists $c \in (0,1)$ depending on $\sigma_1, \sigma_2$, such that $l_{K,j}(\varepsilon), j = 0, \ldots, K(\varepsilon) - 1$ given by (20) yields

$$\varepsilon_{K(\varepsilon)} := \text{KL}(\rho_{K(\varepsilon)}||\rho_*) \leq \varepsilon,$$

for all small enough $\varepsilon > 0$ where $\rho^l_j$ solves (35). Moreover, $\text{cost}_{ML}(\varepsilon) \simeq \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha}}$.

**Proof.** For the proof see Appendix I.1. \hfill \Box

An analogous results holds for the single-level choice in (15) with cost behavior $\log(\varepsilon^{-1})\varepsilon^{-\frac{1}{\alpha}}$.

### 6 Numerical experiments

We consider an adaptation of our running example in a one-dimensional setting with $D = (0,1)$. Instead of working on $L^2(D)$ directly, we parametrize $L^2(D)$ via

$$f(x) = f(x, \cdot) = \sum_{i \in \mathbb{N}} x_i \frac{\sqrt{2}}{\pi} \sin(i\pi \cdot) \in L^2(D),$$
for $x \in \mathbb{R}^{n_x}$. We used the observation operator $\mathcal{O} : H^2(D) \to \mathbb{R}^{n_y} : u_f \mapsto (u_f(s_i))_{i=1}^{n_y}$ with the equispaced points $s_i = \frac{i}{n_y+1}; i = 1, \ldots, n_y = 15$. The forward model is $F(x) = \mathcal{O}(u_f(x)) \in \mathbb{R}^{n_y}$ for $x \in \mathbb{R}^{n_x}$, where $u_f$ denotes the solution of (5). We used piecewise linear FEM on a uniform mesh to approximate $F$ by $F_l$. This setup corresponds to $\alpha = 1$ in Sec. 3. As a prior on the parameter space $\mathbb{R}^{n_x}$ we chose $Q_0 = \mathcal{N}(0, C_0)$ with $C_0 = \text{diag}(i^{-2}, i = 1, \ldots, n_x)$.

We ran multilevel TEKI (Algorithm 1) and the multilevel interacting Langevin sampler (Algorithm 2) on this problem and plotted the error convergence in Fig. 1. For TEKI, the plotted error quantity is $E\left[ \frac{1}{2} \| F_{\text{ref}}(x_K(\varepsilon) - x_*) \|^2 + \frac{1}{2} \| x_K(\varepsilon) - x_* \|^2_{C_0} \right]$ with $F_{\text{ref}} = F_{24}$. For the interacting Langevin sampler we consider the convergence of the posterior mean, and the error quantity is $E\left[ \frac{1}{2} \| f(\cdot, x_K(\varepsilon)) - f(\cdot, x_*) \|^2_{L^2(D)} \right]$. The observed convergence rates roughly coincide with the ones proven in Sec. 4 and Sec. 5. In particular, the multilevel algorithms are superior to their single-level counterparts. Even though the reduction in cost is only by the factor $\log(\varepsilon^{-1})$, as the plot shows this can amount to significant gains in practice. More details on all chosen parameters and the setup can be found in Appendix J.

![Figure 1: Computational costs vs. estimated error for TEKI (left) and the interacting Langevin sampler (ILS) (right). We ran all Algorithms 100 times to estimate the expected error.](image)

## 7 Conclusion

We have presented an abstract multilevel optimization framework, and provided quasioptimal level choices. We showed improved convergence compared to single-level methods, and demonstrated the wide applicability by introducing a novel multilevel ensemble Kalman inversion, as well as a new multilevel Langevin sampler. While our main focus was on inverse problems, we additionally discussed different versions of multilevel gradient descent, which in principle are applicable to any kind of optimization problem where the evaluation of the (derivatives of the) objective function is expensive or impossible, and demands approximation by numerical methods. Further directions or research could include developing optimization procedures for sampling with respect to other metrics (than KL), such as the Wasserstein distance. The proof could proceed along similar lines as Proposition 5.2.
Moreover, developing a framework for adaptive step sizes and levels would be an interesting extension.

References

[1] T. Alsup, L. Venturi, and B. Peherstorfer. Multilevel Stein variational gradient descent with applications to Bayesian inverse problems. Preprint arXiv:2104.01945, 2021.

[2] J. L. Anderson. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A*, 59(2):210–224, 2007.

[3] J. L. Anderson. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A*, 61(1):72–83, 2009.

[4] M. Benning and M. Burger. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111, 2018.

[5] D. Blömker, C. Schillings, P. Wacker, and S. Weissmann. Continuous time limit of the stochastic ensemble Kalman inversion: Strong convergence analysis. Preprint arXiv:2107.14508, 2021.

[6] D. Blömker, C. Schillings, and P. Wacker. A strongly convergent numerical scheme from ensemble Kalman inversion. *SIAM Journal on Numerical Analysis*, 56(4):2537–2562, 2018.

[7] D. Blömker, C. Schillings, P. Wacker, and S. Weissmann. Well posedness and convergence analysis of the ensemble Kalman inversion. *Inverse Problems*, 35(8):085007, 2019.

[8] N. K. Chada, A. M. Stuart, and X. T. Tong. Tikhonov regularization within ensemble Kalman inversion. *SIAM Journal on Numerical Analysis*, 58(2):1263–1294, 2020.

[9] N. K. Chada and X. T. Tong. Convergence acceleration of ensemble Kalman inversion in nonlinear settings. *Math. Comp.*, 91(335):1247–1280, 2022.

[10] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. *An Introduction to Total Variation for Image Analysis*. De Gruyter, Berlin, Boston, 2010.

[11] J. Dick, R. N. Gantner, Q. T. Le Gia, and C. Schwab. Multilevel higher-order quasi-Monte Carlo Bayesian estimation. *Math. Models Methods Appl. Sci.*, 27(5):953–995, 2017.

[12] Z. Ding and Q. Li. Ensemble Kalman inversion: mean-field limit and convergence analysis. *Statistics and Computing*, 31(1):9, 2021.
[13] Z. Ding and Q. Li. Ensemble Kalman sampler: mean-field limit and convergence analysis. *SIAM Journal on Mathematical Analysis*, 53(2):1546–1578, 2021.

[14] T. Dodwell, C. Ketelsen, R. Scheichl, and A. Teckentrup. A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):1075–1108, 2015.

[15] H. Engl, M. Hanke, and G. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer Netherlands, 1996.

[16] H. W. Engl, K. Kunisch, and A. Neubauer. Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. *Inverse Problems*, 5(4):523–540, aug 1989.

[17] A. Ern and J.-L. Guermond. *Finite elements II—Galerkin approximation, elliptic and mixed PDEs*, volume 73 of *Texts in Applied Mathematics*. Springer, Cham, 2021.

[18] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020.

[19] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[20] M. B. Giles and B. J. Waterhouse. Multilevel quasi-Monte Carlo path simulation. In *Advanced financial modelling*, volume 8 of *Radon Ser. Comput. Appl. Math.*, pages 165–181. Walter de Gruyter, Berlin, 2009.

[21] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.

[22] A.-L. Haji-Ali, F. Nobile, L. Tamellini, and R. Tempone. Multi-index stochastic collocation for random PDEs. *Comput. Methods Appl. Mech. Engrg.*, 306:95–122, 2016.

[23] S. Heinrich. *Multilevel Monte Carlo Methods*, pages 58–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[24] M. A. Iglesias, K. J. H. Law, and A. M. Stuart. Ensemble Kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, mar 2013.

[25] N. B. Kovachki and A. M. Stuart. Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, 35(9):095005, 2019.

[26] R. Kullback, S.; Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[27] M. Martin, S. Krumscheid, and F. Nobile. Complexity analysis of stochastic gradient methods for PDE-constrained optimal control problems with uncertain parameters. *ESAIM Math. Model. Numer. Anal.*, 55(4):1599–1633, 2021.
[28] M. Martin and F. Nobile. PDE-constrained optimal control problems with uncertain parameters using SAGA. *SIAM/ASA J. Uncertain. Quantif.*, 9(3):979–1012, 2021.

[29] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[30] G. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Texts in Applied Mathematics. Springer New York, 2014.

[31] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259 – 268, 1992.

[32] S. A. Sauter and C. Schwab. *Boundary element methods*, volume 39 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2011. Translated and expanded from the 2004 German original.

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

[34] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[35] X. Tong, A. Majda, and D. Kelly. Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation. *Communications in Mathematical Sciences*, 14(5):1283–1313, 2016.

[36] S. Tu, S. Venkataraman, A. C. Wilson, A. Gittens, M. I. Jordan, and B. Recht. Breaking locality accelerates block gauss-seidel. In *International Conference on Machine Learning*, pages 3482–3491. PMLR, 2017.

[37] S. Weissmann. Gradient flow structure and convergence analysis of the ensemble Kalman inversion for nonlinear forward models. Preprint arXiv:2203.17117, 2022.

[38] S. Weissmann, N. K. Chada, C. Schillings, and X. T. Tong. Adaptive Tikhonov strategies for stochastic ensemble Kalman inversion. *Inverse Problems*, 38(4):045009, 2022.

[39] A. C. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

[40] J. Zech, D. Düng, and C. Schwab. Multilevel approximation of parametric and stochastic PDEs. *Math. Models Methods Appl. Sci.*, 29(9):1753–1817, 2019.
A Proofs of the main theorems

A.1 Proof of Theorem 3.7

Proof. Let $\varepsilon \in (0, e_0)$. With $\bar{e}_K$ in (12), denote

$$\bar{e}_K((l_j)_j) = c^K e_0 + \sum_{j=0}^{K-1} c^{K-1-j} l_j^{-\alpha} =: \bar{e}_{K,1} + \bar{e}_{K,2}((l_j)_j).$$  \hspace{1cm} (37)

By definition, $K(\varepsilon) \in \mathbb{N}$ in (15) is minimal such that $\bar{e}_{K(\varepsilon),1} \leq \varepsilon$. Moreover, by definition $\bar{l}_K(\varepsilon) > 0$ in (15) is minimal such that $\bar{e}_{K,2}((\bar{l}_K(\varepsilon))_j) \leq \varepsilon$.

Now let $\hat{K} \in \mathbb{N}$ and $\hat{l} > 0$ be any numbers such that $\bar{e}_{K,2}((\hat{l})_j) \leq \varepsilon$. Then

$$\bar{e}_{K,1} \leq \bar{e}_{K,1} + \bar{e}_{K,2}((\hat{l})_j) = \bar{e}_K((\hat{l})_j) \leq \varepsilon,$$

and by the optimality property of $K(\varepsilon)$ we obtain $\hat{K} \geq K(2\varepsilon)$. Similarly, since $\bar{e}_{K,2}((\hat{l})_j) \leq \varepsilon$, (38)

$$\text{cost}_{\text{SL}}(\varepsilon) = K(\varepsilon) l_K(\varepsilon) = \log(c)^{-1} \log \left( \frac{\varepsilon}{2 e_0} \right) \left( \sum_{i=0}^{K-1} c^{i-K-1} \right)^{\frac{1}{\alpha}} \left( \frac{\varepsilon}{2} \right)^{-\frac{1}{2}}.$$

Since with $\delta := c^{\frac{1}{1+\alpha}} \in (0, 1)$ holds $1 \leq \sum_{i=0}^{K-1} c^{K-1-i} \leq (1 - c)^{-1}$ we have $\text{cost}_{\text{SL}}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha}} \log(\varepsilon^{-1})$ as $\varepsilon \to 0$. Together with (39) we find

$$\text{cost}_{\text{SL}}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha}} \log(\varepsilon^{-1}) \simeq \inf \{ \hat{K} \hat{l} : \bar{e}_{K}(\hat{l}) \leq \varepsilon \} \quad \text{as} \quad \varepsilon \to 0.$$

This shows (16) and quasi-optimality in the sense of Def. 3.6.

A.2 Proof of Lemma 3.8

Proof. Define $a_j := c^{K-1-j}$ for all $j = 0, \ldots, K - 1$. We wish to minimize $\sum_{j=0}^{K-1} l_j$ for $l_j > 0$ and under the constraint $\sum_{j=0}^{K-1} a_j l_j^{-\alpha} - \varepsilon$. To this end we use a Lagrange multiplier and consider

$$\min_{l_1, \ldots, l_K, \lambda} \sum_{j=0}^{K-1} l_j + \lambda \left( \sum_{j=0}^{K-1} a_j l_j^{-\alpha} - \varepsilon \right).$$
Taking the derivatives w.r.t. $l_j$ and $\lambda$ leads to the following first order optimality conditions

$$1 - \lambda \alpha a_j l_j^{-(1+\alpha)} = 0, \quad \sum_{j=0}^{K-1} a_j l_j^{-\alpha} - \varepsilon = 0.$$ 

The first condition gives $l_j = C_{K,\varepsilon} a_j^{\frac{1}{1+\alpha}}$ for some constant $C_{K,\varepsilon}$. Plugging $l_j$ into the second condition we find

$$C_{K,\varepsilon}^{-\alpha} \sum_{j=0}^{K-1} a_j a_j^{-\frac{\alpha}{1+\alpha}} = \varepsilon.$$ 

Hence $C_{K,\varepsilon} = \varepsilon^{-\frac{1}{\alpha}} (\sum_{j=0}^{K-1} a_j^{\frac{1}{1+\alpha}})^{\frac{1}{\alpha}}$. Finally,

$$\sum_{j=0}^{K-1} a_j^{\frac{1}{1+\alpha}} = \sum_{j=0}^{K-1} c_j^{\frac{1}{1+\alpha}} = \frac{1 - c_K^{\frac{1}{1+\alpha}}}{1 - c^{\frac{1}{1+\alpha}}}.$$ 

This shows (18).

**Remark A.1.** The constant $C_{K,\varepsilon}$ in (18) increases for decreasing tolerance $\varepsilon > 0$. Moreover, $C_{K,\varepsilon}$ is bounded from below and above uniformly for all $K \in \mathbb{N}$ as for $c \in (0, 1)$ holds

$$\sum_{j=0}^{K-1} c^{\frac{1}{1+\alpha}} = \frac{1 - c^{K-1}}{1 - c^{1+\alpha}} \in \left(1, \frac{1}{1 - c^{1+\alpha}}\right).$$ 

**A.3 Proof of Theorem 3.10**

**Proof.** With $l_{K,j}(\varepsilon)$ and $K(\varepsilon)$ as in (20), the calculation in (19) and Rmk. A.1 show

$$\text{cost}_{ML}(\varepsilon) = \sum_{j=0}^{K(\varepsilon) - 1} l_{K,j} = \left(\frac{\varepsilon}{2}\right)^{-\frac{1}{\alpha}} \left(1 - c^{K(\varepsilon)/(1+\alpha)}\right)^{\frac{1+\alpha}{\alpha}} \approx \varepsilon^{-\frac{1}{\alpha}} \quad \text{as} \quad \varepsilon \to 0,$$ 

as claimed.

The proof of quasi-optimality in the sense of Def. 3.9 follows the same argument as in the proof of Theorem 3.7. Let $\varepsilon \in (0, e_0)$. Split $\tilde{e}_K((l_j)_j)$ in (41), in the terms $\tilde{e}_{K,1}$ and $\tilde{e}_{K,2}((l_j)_j)$ as in (37).

Now let $K \in \mathbb{N}$ and $(\tilde{l}_j)_{j=0}^{K-1} > 0$ be arbitrary such that $\tilde{e}((\tilde{l}_j)_j) \leq \varepsilon$. Moreover, define

$$\tilde{l}_{K,j} := C_{K,\varepsilon} \cdot c_j^{\frac{K-1-j}{1+\alpha}}, \quad C_{K,\varepsilon} = \varepsilon^{-\frac{1}{\alpha}} \left(\sum_{i=0}^{K-1} C_{K,\varepsilon}^{\frac{K-1-i}{1+\alpha}}\right)^{\frac{1}{\alpha}}.$$ 

20
which by Lemma 3.8 minimizes $\sum_{j=0}^{K-1} \hat{l}_{K,j}$ under the constraint $\hat{e}_{K,2}(\hat{l}_{K,j}) \leq \epsilon$. Since also $\hat{e}_{K,2}(\hat{l}_{j}) \leq \hat{e}_{K}(l_{K,j}) \leq \epsilon$, this implies

$$\sum_{j=0}^{\hat{K}-1} \hat{l}_j \geq \sum_{j=0}^{K-1} \hat{l}_{K,j} = \epsilon^{-\frac{1}{\alpha}} \left( \frac{1 - e^{\hat{K}/(1+\alpha)}}{1 - e^{1/(1+\alpha)}} \right)^{\frac{1+\alpha}{\alpha}} \tag{41}$$

where the equality holds by the calculation in (19).

Next observe that by definition, $\hat{K}(2\epsilon) \in \mathbb{N}$ in (20) is minimal such that $\hat{e}_{K(\epsilon),1} \leq \epsilon$. Since also $\hat{e}_{\hat{K},1} \leq \hat{e}_{\hat{K}}(\hat{l}_{j}) \leq \epsilon$, this implies $K(2\epsilon) \leq \hat{K}$. Therefore

$$\epsilon^{-\frac{1}{\alpha}} \left( \frac{1 - e^{\hat{K}/(1+\alpha)}}{1 - e^{1/(1+\alpha)}} \right)^{\frac{1+\alpha}{\alpha}} \geq \epsilon^{-\frac{1}{\alpha}} \left( \frac{1 - e^{K(2\epsilon)/(1+\alpha)}}{1 - e^{1/(1+\alpha)}} \right)^{\frac{1+\alpha}{\alpha}} = \sum_{j=0}^{K(2\epsilon)-1} l_{K,j}(2\epsilon), \tag{42}$$

where the last inequality holds by the calculation in (19). In all, (41) and (42) yield

$$\sum_{j=0}^{K(2\epsilon)-1} l_{K,j}(2\epsilon) \leq \sum_{j=0}^{\hat{K}} \hat{l}_j. \tag{43}$$

Since $\hat{K}$ and $\hat{l}_j$ were arbitrary such that $\hat{e}(\hat{l}_{j}) \leq \epsilon$,

$$\sum_{j=0}^{K(2\epsilon)-1} l_{K,j}(2\epsilon) \leq \inf \left\{ \sum_{j=0}^{\hat{K}-1} \hat{l}_j : \hat{e}_{K}(\hat{l}_j) \leq \epsilon, \hat{K} \in \mathbb{N}, \hat{l}_j > 0 \forall j \right\} \leq \sum_{j=0}^{K(\epsilon)-1} l_{K,j}(\epsilon), \tag{43}$$

where the second inequality holds due to $\hat{e}_{K(\epsilon)}(l_{K,j}(\epsilon)) \leq \epsilon$. By (40) all terms in (43) behave like $\epsilon^{-1/\alpha}$ as $\epsilon \to 0$, and this shows quasi-optimality in the sense of Def. 3.9.

**B Gradient descent and accelerated gradient descent**

In this appendix we discuss the implications of our results for gradient descent, accelerated gradient descent and the stochastic versions of those algorithms. Throughout this section we assume that $\mathcal{X}$ is a Hilbert space and $\Phi : \mathcal{X} \to \mathbb{R}$ is Fréchet differentiable, $\mu$-strongly convex, i.e.

$$\Phi(y) \geq \Phi(x) + \langle \nabla \Phi(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \quad \forall x, y \in \mathcal{X}, \tag{44}$$

and satisfies $L$-smoothness so that

$$\Phi(y) \leq \Phi(x) + \langle \nabla \Phi(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \quad \forall x, y \in \mathcal{X}. \tag{45}$$

Moreover, we denote by $x_\ast \in \mathcal{X}$ the unique minimizer of $\Phi$. 

21
For the deterministic algorithms we assume the existence of approximate gradients \( g_l : \mathcal{X} \to \mathcal{X} \) such that
\[
\| \nabla \Phi(x) - g_l(x) \| \leq \frac{l^{-\alpha}}{\eta} \quad \forall x \in \mathcal{X},
\] (46)
where \( \eta > 0 \) will denote the step-size in the following. For the stochastic variants we will work under the assumption that for every \( x \in \mathcal{X} \) there exists a random variable \( G_l(x) \in \mathcal{X} \) such that
\[
\mathbb{E}\| \nabla \Phi(x) - G_l(x) \| \leq \frac{l^{-\alpha}}{\eta} \quad \forall x \in \mathcal{X}.
\] (47)

B.1 Gradient Descent (GD)

The gradient descent update with step-size \( \eta \) and approximate gradient at level \( l_k \) reads
\[
x_{k+1} = x_k - \eta g_{l_k}(x_k).
\] (48)

**Proposition B.1.** Let \( 0 \leq \eta \leq 1/L \) and let \( g_l \) satisfy (46). Then \( x_k \) generated by (48) satisfies the bound (8) with \( e_k = \| x_{k+1} - x^* \| \) and \( c = \sqrt{1 - \eta \mu} \).

**Proof.** Observe that
\[
\| x_{k+1} - x^* \| = \| x_k - x^* - \eta \nabla_x \Phi(x_k) + \eta \nabla_x \Phi(x_k) - \eta g_{l_k}(x_k) \|
\leq \| x_k - x^* - \eta \nabla_x \Phi(x_k) \| + \eta \| \nabla_x \Phi(x_k) - g_{l_k}(x_k) \|.
\]
Using \( \mu \)-strong convexity (44) and \( L \)-smoothness (45) we obtain the following upper bound,
\[
\| x_k - x^* - \eta \nabla_x \Phi(x_k) \|^2 = \| x_k - x^* \|^2 - 2\eta \langle x_k - x^*, \nabla_x \Phi(x_k) \rangle + \eta^2 \| \nabla_x \Phi(x_k) \|^2
\leq \| x_k - x^* \|^2 - \eta \mu \| x_k - x^* \|^2 - 2\eta \langle \Phi(x_k) - \Phi(x^*) \rangle + \eta^2 \| \nabla_x \Phi(x_k) \|^2
\leq (1 - \eta \mu) \| x_k - x^* \|^2 - \eta (1/L - \eta) \| \nabla_x \Phi(x_k) \|^2
\leq (1 - \eta \mu) \| x_k - x^* \|^2.
\]
Combining these inequalities with assumption (46) leads to desired recursion (8)
\[
e_{k+1} = \| x_{k+1} - x^* \| \leq \sqrt{1 - \eta \mu} \| x_k - x^* \| + l^{-\alpha}_{-} = ce_k + l^{-\alpha}_{-}.
\]

It now follows from Thm. 3.10 that:

**Corollary B.2 (MLGD).** Consider the setting of Proposition B.1. Then with the levels \( l_{K,j}(\varepsilon) \) as in (20), \( x_K \) generated by (48) satisfies \( e_K := \| x_K - x^* \| \leq \varepsilon \), and it holds \( \sum_{j=0}^{K-1} l_{K,j}(\varepsilon) = O(\varepsilon^{-\alpha}) \) as \( \varepsilon \to 0 \).
B.2 Stochastic Gradient Descent (SGD)

We now consider the stochastic setting, i.e. we assume given random variables \(G_l(x)\) as in \([47]\). The stochastic gradient descent update with step-size \(\eta\) and approximate stochastic gradient at level \(l_k\) then reads

\[
x_{k+1} = x_k - \eta G_{l_k}(x_k).
\]

**Proposition B.3.** Let \(k\) the filtration up to iteration which correspond to the number of required evaluations of \(\varphi\)

\[
\|\| g\|\| \leq \eta \|g\|_\mu + \eta \|\| \|\| l\|\|_\mu
\]

**Corollary B.4**

\[
\sum_{j=0}^{K-1} l_{K,j}(\varepsilon) = O(\varepsilon^{-\alpha}) \text{ as } \varepsilon \to 0.
\]

\[
\text{We next discuss a standard example of stochastic gradient descent, namely with } G_l \text{ being a Monte Carlo estimator. However, we emphasize that other approximation schemes are applicable as well in this setting.}
\]

**Example B.5** (SGD with dynamic sampling). We consider a stochastic optimization problem in the form of \((1)\) by

\[
\min_{x \in \mathcal{X}} \Phi(x), \quad \Phi(x) := \mathbb{E}_\xi[\varphi(x, \xi)],
\]

where \(\xi\) is a random variable on a underlying probability space \((\Omega, \mathcal{F}, \mathbb{P})\) with state space \((E, \mathcal{E})\) and \(\varphi : \mathcal{X} \times E \to \mathbb{R}\) is the stochastic objective function. The expected value \(\mathbb{E}_\xi[\varphi(x, \xi)]\) and its gradient are often not available analytically. Assuming access to \(i.i.d.\) samples \(\{\xi^n_{k,n}\}_{n=1}^{K}\) of \(\xi\), we can define a stochastic gradient approximation by the Monte Carlo estimator \(G_l(x) := \frac{1}{K} \sum_{n=1}^{K} \nabla \varphi(x_k, \xi^n_{k,n})\). The level \(l_k\) describes the batch size in iteration \(k\), which correspond to the number of required evaluations of \(\varphi\). In this sense, \(l_k\) describes the computational cost required to evaluate \(G_l(x)\). Under certain integrability assumptions, \(G_l(x)\) is an unbiased estimator of \(\nabla \Phi(x_k)\), i.e. \(\nabla \Phi(x) = \mathbb{E}_\xi[\nabla \varphi(x, \xi)]\), and it holds

\[
\mathbb{E}[\| \nabla \Phi(x) - G_l(x) \|] \leq \sqrt{\mathbb{E}[\| \nabla \Phi(x) - G_l(x) \|^2]} \leq \left( \mathbb{E}_{\xi}[\| \nabla \varphi(x, \xi) - \mathbb{E}_{\xi}[\nabla \varphi(x, \xi)] \|^2] \right)^{1/2},
\]
i.e. we have the second inequality in (10) (up to a constant, cp. Rmk. [3.11]) with \(\alpha = 1/2\). Applying the multilevel strategy (20) then yields a stochastic optimization method with increasing batchsize.

By Corollary [B.4], achieving error \(\mathbb{E}[\|x_K - x_*\|] \leq \varepsilon\) requires \(O(\varepsilon^{-1/2})\) evaluations of \(\nabla \varphi(x, \xi)\) (since \(\sum_{j=0}^{K-1} l_{K,j}(\varepsilon)\) coincides with the number of evaluations of \(\nabla \varphi(x, \xi)\) in the current setting). This is the same convergence rate that is obtained for batch size 1 and decreasing step size \(\eta_k \sim 1/k\). However, we point out that the present multilevel version, which uses constant step size and increasing batch size, allows for parallelization in the gradient evaluations.

### B.3 Accelerated Gradient Descent (AGD)

We write the accelerated gradient descent algorithm as the following update (see [29])

\[
\begin{align*}
p_{k+1} &= q_k - \alpha \nabla \Phi(q_k) \\
q_{k+1} &= p_{k+1} + \beta (p_{k+1} - p_k),
\end{align*}
\]

where \(\alpha = 1/L\) and \(\beta = \sqrt{L - \mu} / \sqrt{L + \mu}\). Defining \(x_k = [p_k, q_k]\) we can represent AGD as (7) using the operator

\[
\Psi = \begin{bmatrix}
0 & (I - \alpha \nabla \Phi) \\
\beta I & (I + \beta)(I - \alpha \nabla \Phi)
\end{bmatrix}.
\]

AGD using approximated gradients at level \(l_k\) can then be represented as the following three sequence update (see e.g. [29] for equivalence):

\[
\begin{align*}
x_k &= \frac{\tau}{1 + \tau} z_k + \frac{1}{1 + \tau} y_k \quad (51a) \\
y_{k+1} &= x_k - \frac{1}{L} g_l(x_k) \quad (51b) \\
z_{k+1} &= z_k + \tau (x_k - z_k) - \frac{\tau}{\mu} g_l(x_k). \quad (51c)
\end{align*}
\]

**Proposition B.6.** Suppose that \(g_l\) satisfies (46) with \(\eta = 1/\sqrt{L}\) and \(\tau = \sqrt{\mu/L}\). Then \((y_k, z_k)\) generated by (51) satisfies (8) with \(e_k = \Phi(y_k) - \Phi(x_*) + \frac{\mu}{2} \|z_k - x_*\|^2\) exponent 2\(\alpha\), and \(c = 1 - \tau\), more precisely

\[
\Phi(y_{k+1}) - \Phi(x_*) + \frac{\mu}{2} \|z_{k+1} - x_*\|^2 \leq (1 - \tau) \left( \Phi(y_k) - \Phi(x_*) + \frac{\mu}{2} \|z_k - x_*\|^2 \right) + l_k^{-2\alpha}.
\]

**Proof.** We begin by using the \(L\)-smoothness of \(\Phi\):

\[
\Phi(y_{k+1}) - \Phi(x_*) \leq \langle \nabla \Phi(x_k), y_{k+1} - x_k \rangle + \frac{L}{2} \|x_k - y_{k+1}\|^2
\]

\[
= -\frac{1}{L} \langle \nabla \Phi(x_k), g_l(x_k) \rangle + \frac{1}{2L} \|g_l(x_k)\|^2
\]

\[
= \frac{1}{2L} \|g_l(x_k) - \nabla \Phi(x_k)\|^2 - \frac{1}{2L} \|\nabla \Phi(x_k)\|^2 \leq \frac{l_k^{-2\alpha}}{2L\eta^2} - \frac{1}{2L} \|\nabla \Phi(x_k)\|^2.
\]

(52)
Denote $\tilde{z}_{k+1} := z_{k+1} - \frac{1}{\mu} (\nabla\Phi(x_k) - \nabla g_l(x_k))$. The triangle inequality results in the following upper bound
\[ \frac{\sqrt{\mu}}{2} \| \tilde{z}_{k+1} - x_* \| \leq \frac{1}{2\sqrt{L}} \| \nabla\Phi(x_k) - g_l(x_k) \| + \frac{\sqrt{\mu}}{2} \| \tilde{z}_{k+1} - x_* \| \leq \frac{l_k^{-\alpha}}{2\sqrt{L\eta}} + \frac{\sqrt{\mu}}{2} \| \tilde{z}_{k+1} - x_* \|, \]
and using Jensen’s inequality \(((a/2 + b/2)^2) \leq a^2/2 + b^2/2\) we obtain the subsequent identity,
\[ \frac{\sqrt{\mu}}{2} \| z_{k+1} - x_* \|^2 \leq \frac{l_k^{-2\alpha}}{2L\eta^2} + \frac{\mu}{2} \| \tilde{z}_{k+1} - x_* \|^2. \]

The rest of the Lyapunov analysis follows as normal (see e.g. [36] and [39]). We use the previous inequality and strong convexity to obtain the bound:
\[
\begin{align*}
\frac{\mu}{2} \| z_{k+1} - x_* \|^2 - \frac{\mu}{2} \| z_k - x_* \|^2 &\leq \frac{\mu}{2} \| \tilde{z}_{k+1} - x_* \|^2 - \frac{\mu}{2} \| z_k - x_* \|^2 + \frac{l_k^{-2\alpha}}{2L\eta^2} \\
&= \tau \langle \nabla\Phi(x_k), x_* - z_k \rangle - \tau \mu \langle x_k - z_k, x_* - z_k \rangle + \frac{\mu}{2} \| \tilde{z}_{k+1} - z_k \|^2 + \frac{l_k^{-2\alpha}}{2L\eta^2} \\
&\leq \tau \langle \nabla\Phi(x_k), x_k - z_k \rangle - \tau \mu \langle x_k - z_k, x_* - z_k \rangle + \frac{\mu}{2} \| \tilde{z}_{k+1} - z_k \|^2 \\
&\quad - \tau \langle \Phi(x_k) - \Phi(x_*), z_k \rangle + \frac{\mu}{2} \| x_k - x_* \|^2 + \frac{l_k^{-2\alpha}}{2L\eta^2} \\
&= \langle \nabla\Phi(x_k), x_k - y_k \rangle - \tau \mu \| x_k - z_k \|^2 + \frac{\mu}{2} \| z_{k+1} - z_k \|^2 + \frac{l_k^{-2\alpha}}{2L\eta^2} \\
&\quad - \tau \langle \Phi(x_k) - \Phi(x_*), z_k \rangle + \frac{\mu}{2} \| z_{k+1} - z_k \|^2.
\end{align*}
\]

The inequality uses the strong convexity of $\Phi$. Using the strong convexity of $\Phi$ again and the descent bound (52) we have:
\[
\begin{align*}
e_{k+1} &\leq (1 - \tau)e_k + \Phi(x_k) - \Phi(y_k) + \langle \nabla\Phi(x_k), x_k - y_k \rangle - \frac{1}{2L} \| \nabla\Phi(x_k) \|^2 + \frac{l_k^{-2\alpha}}{L\eta^2} \\
&\quad - \tau \langle \Phi(x_k) - \Phi(y_k), z_k \rangle - \frac{\tau \mu}{2} \| x_k - z_k \|^2 + \frac{\mu}{2} \| \tilde{z}_{k+1} - z_k \|^2 \\
&\leq ce_k - \frac{\mu}{2} \| x_k - y_k \|^2 - \frac{1}{2L} \| \nabla\Phi(x_k) \|^2 - \tau \langle \nabla\Phi(x_k), x_k - y_k \rangle + \tau L \| x_k - y_k \|^2 \\
&\quad - \frac{\tau \mu}{2} \| x_k - z_k \|^2 + \frac{\mu}{2} \| \tilde{z}_{k+1} - z_k \|^2 + \frac{l_k^{-2\alpha}}{L\eta^2} \\
&= ce_k + (\frac{\tau^2}{\mu} - \frac{1}{2L}) \| \nabla\Phi(x_k) \|^2 + (\frac{\tau L}{2\tau} - \frac{\mu}{2}) \| x_k - y_k \|^2 + l_k^{-2\alpha} \\
&\quad = ce_k + l_k^{-2\alpha}.
\end{align*}
\]

The second line uses strong convexity and smoothness of $\Phi$. The following line expands the term $\| \tilde{z}_{k+1} - z_k \|^2 = \| y_k - x_k - \frac{\tau}{\mu} \nabla\Phi(x_k) \|^2$ and uses the smoothness of $\Phi$ and identity $\eta = 1/\sqrt{L}$. 

25
Corollary B.7 (MLAGD). Consider the setting of Proposition B.6. Then with the levels $l_{K,j}(\varepsilon)$ as in (20), $(y_K, z_K)$ generated by (51) satisfies $e_K := \Phi(y_K) - \Phi(x^*) + \frac{\mu}{2} \| z_K - x^* \|^2 \leq \varepsilon$, and it holds $\sum_{j=0}^{K-1} l_{K,j}(\varepsilon) = O(\varepsilon^{-2\alpha})$ as $\varepsilon \to 0$.

Note that the cost $\sum_{j=0}^{K-1} l_{K,j}(\varepsilon)$ for AGD in Corollary B.7 increases at twice the rate compared to GD in Corollary B.2. However, the AGD result is formulated for a quadratic cost quantity, so that the resulting convergence rate of the error in terms of the cost is asymptotically the same.

B.4 Accelerated Stochastic Gradient Descent (ASGD)

We now consider again the stochastic setting, i.e. we assume we are given random variables $G_l$ as in (47). In the following let

$$x_k = \frac{\tau}{1 + \tau} z_k + \frac{1}{1 + \tau} y_k$$  \hspace{1cm} (53a)

$$y_{k+1} = x_k - \frac{1}{L} G_{l_k}(x_k)$$  \hspace{1cm} (53b)

$$z_{k+1} = z_k + \tau (x_k - z_k) - \frac{\tau}{\mu} G_{l_k}(x_k).$$  \hspace{1cm} (53c)

Proposition B.8. Suppose that $G_l$ satisfies (47) with $\eta = 1/\sqrt{L}$ where $\tau = \sqrt{\mu/L}$. Then $(y_k, z_k)$ generated by (53) satisfies (8) with $e_k = \mathbb{E}[\Phi(y_k)] - \Phi(x^*) + \frac{\mu}{2} \mathbb{E}[\| z_k - x^* \|^2]$, exponent $2\alpha$ and $c = 1 - \tau$, more precisely

$$\mathbb{E}[\Phi(y_{k+1})] - \Phi(x^*) + \frac{\mu}{2} \mathbb{E}[\| z_{k+1} - x^* \|^2] \leq (1 - \tau) \left( \mathbb{E}[\Phi(y_k)] - \Phi(x^*) + \frac{\mu}{2} \mathbb{E}[\| z_k - x^* \|^2] \right) + l_k^{-2\alpha}.$$

Proof. The proof proceeds in the exact same manner as accelerated gradient descent replacing $g_{l_k}$ with $G_{l_k}$. In particular note that

$$\mathbb{E}_k[\Phi(y_{k+1})] - \Phi(x_k) \leq \frac{1}{2L} \mathbb{E}_k[\| G_l(x_k) - \nabla \Phi(x_k) \|^2 - \| \nabla \Phi(x_k) \|^2] \leq \frac{l_k^{-2\alpha}}{2L\eta^2} - \frac{1}{2L} \mathbb{E}_k[\| \nabla \Phi(x_k) \|^2];$$

and

$$\frac{\sqrt{\mu}}{2} \mathbb{E}_k[\| z_{k+1} - x^* \|] \leq \frac{1}{2\sqrt{L}} \mathbb{E}_k[\| \nabla \Phi(x_k) - G_l(x_k) \| + \frac{\sqrt{\mu}}{2} \mathbb{E}_k[\| \tilde{z}_{k+1} - x^* \|].$$

Therefore,

$$\frac{\sqrt{\mu}}{2} \mathbb{E}_k[\| z_{k+1} - x^* \|^2] \leq \frac{l_k^{-2\alpha}}{2L\eta^2} + \frac{\mu}{2} \mathbb{E}_k[\| \tilde{z}_{k+1} - x^* \|^2].$$

Given the remainder of the proof relies on the strong convexity and smoothness of the function and update (53a) we obtain the following recursion following the same line of argumentation:

$$e_{k+1} \leq c e_k + l_k^{-2\alpha}. \hspace{1cm} \Box$$
Corollary B.9 (MLASGD). Consider the setting of Proposition B.8. Then with the levels \( l_{K,j}(\varepsilon) \) as in (20), \((y_k, z_k)\) generated by (53) satisfies \( e_K := \mathbb{E}[\Phi(y_k)] - \Phi(x_\ast) + \frac{\eta}{2} \mathbb{E}[\|z_k - x_\ast\|^2] \leq \varepsilon \), and it holds \( \sum_{j=0}^{N-1} l_{K,j}(\varepsilon) = O(\varepsilon^{-2\alpha}) \) as \( \varepsilon \to 0 \).

Remark B.10. The Monte Carlo estimator discussed in the context of SGD, for example, can be used for accelerated SGD where we require increasing batchsize to ensure assumption \([10]\) holds with \( \alpha = 1/2 \).

C Details for the running example

For every \( f \in L^2(D) \) on the bounded convex polygonal Lipschitz domain \( D \subseteq \mathbb{R}^2 \), denote by \( u_f \in H^1_0(D) \subseteq L^2(D) \) the unique weak solution to (5) with right-hand side \( f \), i.e.

\[
\int_D \nabla u_f(s) \nabla v(x) + u_f(s)v(s)ds = \int_D f(s)v(s)ds \quad \forall v \in H^1_0(D).
\] (54)

C.1 Well-definedness and regularity of \( u_f \)

Existence and well-definedness of \( u_f \) is classical. Moreover, one has the apriori estimate \( \|u_f\|_{H^1_0(D)} \leq \|f\|_{H^{-1}(D)} \), we refer for example to [17] §26.1.

Furthermore convexity of \( D \) in fact implies \( H^2(D) \) regularity of \( u_f \) and \( \|u_f\|_{H^2(D)} \leq \|f\|_{L^2(D)} \), see [17] Theorem 31.30).

C.2 Formula for \( \nabla \Phi(f) \)

Consider the objective

\[
\Phi(f) = \frac{1}{2} \|\Gamma^{-1/2}(F(f) - y)\|_{\mathbb{R}^n_y}^2 + \frac{\lambda}{2} \|f\|_{L^2(D)}^2 =: \ell(f, y) + R(f).
\] (55)

in Example 3.5. Clearly the gradient of \( R(f) \) equals \( \lambda f \in L^2(D) \). It remains to compute \( \nabla f \ell(f, y) \in L^2(D) \).

Introduce the operators

\[
S := \begin{cases} \mathbb{R}^n_y \to \mathbb{R} \quad \forall w \mapsto \frac{1}{2} \|\Gamma^{-1/2}(w - y)\|_{\mathbb{R}^n_y}^2, \\
\quad \mathcal{O} := \begin{cases} L^2(D) \to \mathbb{R}^n_y \\
\quad f \mapsto (\int_D \xi_jf)_{j=1}^n,
\end{cases}
\end{cases}
\]

and

\[
\mathcal{A} := \begin{cases} L^2(D) \to L^2(D) \\
\quad h \mapsto u_h,
\end{cases}
\]

Observe that \( \ell(f, y) = S(\mathcal{O}(\mathcal{A}(f))) \). Using that \( \mathcal{A} : L^2(D) \to H^1_0(D) \) is bounded linear, it is easily seen that \( \mathcal{A} : L^2(D) \to L^2(D) \) is bounded linear and self-adjoint. Therefore (by the chain rule)

\[
\nabla f \ell(f, y) = \nabla(S \circ \mathcal{O} \circ \mathcal{A})(f) = \mathcal{A}^\ast(\nabla(S \circ \mathcal{O})(\mathcal{A}(f))) = \mathcal{A}(\nabla(S \circ \mathcal{O})(u_f)) \in L^2(D),
\] (56)
where \( u_f = A(f) \in H^1_0(D) \subseteq L^2(D) \). We next compute \( \nabla(S \circ O) \). Denoting by \( DS(w) \) the Fréchet derivative, it holds
\[
DS(w)(v) = (w - y)^\top \Gamma^{-1} v \in \mathbb{R} \quad \forall v \in \mathbb{R}^{n_y}
\]
and with \( \xi = (\xi_j)_{j=1}^{n_y} \in L^2(D, \mathbb{R}^{n_y}) \) since \( O \) is bounded linear,
\[
D\xi(f)(g) = \xi(g) = \int_D g(s) \xi(s) ds \in \mathbb{R}^{n_y} \quad \forall g \in L^2(D).
\]
Hence for the composition
\[
D(S \circ O)(f)(g) = DS(O(f))(D\xi(f)(g)) = \int_D (\xi(O(f) - y) \Gamma^{-1} \xi(s) g(s) ds \quad \forall g \in L^2(D).
\]
This shows
\[
\nabla(S \circ O)(f) = (\xi(O(f) - y) \Gamma^{-1} \xi(\cdot) \in L^2(D)
\]
and finally by \( \text{[56]} \), \( \nabla f \ell(f, y) = A(h) = u_h \) with \( h(\cdot) = (\xi(O(f) - y) \Gamma^{-1} \xi(\cdot) \in L^2(D).
\]

C.3 Finite element approximation of \( \nabla \Phi(f) \)

Next we argue that \( u_h \) can be approximated with the rate claimed in Example \( \text{[3.5]} \). According to, e.g., \( \text{[17]} \) §26.3.3., §32.3.2., given \( f \in L^2(D) \), the FEM approximation \( u_f \) to \( f = A(f) \) on a uniform simplicial mesh on \( D \subseteq \mathbb{R}^2 \) with \( O(l) \) elements will satisfy \( \|u_f - u_f\|_{L^2(D)} \lesssim l^{-1} \).

Now set \( \tilde{h}(\cdot) := (\xi(O(u_f) - y) \Gamma^{-1} \xi(\cdot) \in L^2(D) \). Then, since \( O : L^2(D) \rightarrow \mathbb{R}^{n_y} \) is bounded linear and \( \xi \in L^2(D, \mathbb{R}^{n_y}) \), we have \( \|\tilde{h} - h\|_{L^2(D)} \lesssim l^{-1} \). Using that \( A : L^2(D) \rightarrow L^2(D) \) is bounded linear, \( \|u_h - u_h\|_{L^2(D)} = \|A(h - \tilde{h})\|_{L^2(D)} \lesssim l^{-1} \). Finally, \( \|u_h - u_h\|_{L^2(D)} \lesssim l^{-1} \), and thus by the triangle inequality \( \|u_h - u_f\| \lesssim l^{-1} \) as claimed.

Then
\[
g_f(f) := u_h + \lambda f
\]
is an approximation to \( \nabla \Phi(f) \) satisfying \( \|g_f(f) - \nabla \Phi(f)\|_{L^2(D)} \lesssim l^{-1} \).

C.4 Multilevel convergence of gradient descent

We first show \( L \)-smoothness and \( \mu \)-strong convexity for the objective in \( \text{[55]} \).

With the notation from Sec. \( \text{C.2} \) the norm of the operator \( O : L^2(D) \rightarrow \mathbb{R}^{n_y} \) satisfies \( \|O\| = \|\xi\|_{L^2(D)} \). Hence for all \( f, g \in L^2(D) \)
\[
\|\nabla \Phi(f) - \nabla \Phi(g)\|_{L^2(D)} = \|\xi(u_f - u_g) \Gamma^{-1} \xi(\cdot) + \lambda (f - g)\|
\leq (\|\xi\|_{L^2(D)} \|\Gamma^{-1}\|_{\mathbb{R}^{n_y \times n_y}} \|\xi\|_{\mathbb{R}^{n_y \times n_y}} + \lambda) \|f - g\|_{L^2(D)}.
\]
This shows that \( \Phi \) is \( L \)-smooth with \( L = \|\xi\|_{L^2(D)} \|\Gamma^{-1}\|_{\mathbb{R}^{n_y \times n_y}} \|\xi\|_{\mathbb{R}^{n_y \times n_y}} + \lambda \).

Moreover, since \( F : L^2(D) \rightarrow \mathbb{R}^{n_y} \) is bounded linear, the term \( f \mapsto \frac{1}{2} \|\Gamma^{-1/2}(F(f) - y)\|_{\mathbb{R}^{n_y}}^2 \) is convex. Hence, the added regularizer ensures \( \Phi \) in \( \text{[55]} \) to be \( \mu \)-strongly convex with \( \mu = \lambda \).
In all this shows that Example 3.5 is in the setting of Example 3.3 with \( \alpha = 1 \) (i.e. \( \Phi \) is \( L \)-smooth, \( \mu \)-strongly convex, and \( g \) in (57) satisfies the first inequality in (10) with \( \alpha = 1 \)). Thus, for small enough \( \eta > 0 \), the iteration \( f_{j+1} = f_j - \eta \nabla g_{\varepsilon}(f_k) \) with the approximate gradient from subsection C.3 converges to the unique minimizer \( f_\ast \in L^2(D) \) of \( \Phi \) as \( k \to \infty \). With the multilevel choice \( l_{K,j}(\varepsilon), j = 1, \ldots, K(\varepsilon) \) as in (20), it holds \( \|f_{K(\varepsilon)} - f_\ast\|_{L^2(D)} \lesssim \varepsilon \) and the cost \( \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) \), which (up to a constant) can be interpreted as the computational cost of all required FEM approximations, behaves like \( O(\varepsilon^{-1}) \) as \( \varepsilon \to 0 \) according to Theorem 3.10.

D Optimality of the multilevel rate for gradient descent

We give a simple example in the setting of Example 3.3 to show that our results in Sec. 3 are sharp in general. For some fixed \( \alpha > 0 \), consider the objective function and its approximation

\[
\Phi(x) = \frac{1}{2} x^2, \quad \Phi_l(x) = \frac{1}{2}(x - l^{-\alpha})^2, \quad x \in \mathbb{R}.
\]

Denote the unique minimizer of \( \Phi \) by \( x_\ast := 0 \). We consider gradient descent with a fixed step size \( \eta \in (0, 1) \), which amounts to (cp. (6))

\[
\Psi(x) = x - \eta x, \quad \Psi_l(x) = x - \eta(x - l^{-\alpha}).
\]

Hence for some initial value \( x_0 \in \mathbb{R} \), (7) becomes

\[
x_{k+1} = \Psi_l(x_k) = (1 - \eta) x_k + \eta l^{-\alpha}_k \tag{58}
\]

and thus assumption (8) holds with \( c := 1 - \eta \in (0, 1) \).

Proposition D.1. Let \( x_0 \geq 0 \), and let \( x_k \) be as in (58). Then for every \( \varepsilon > 0 \), for every \( K \in \mathbb{N} \) and for every \( (l_j)_{j=0}^{K-1} \in (0, \infty)^K \) such that \( |x_K - x_\ast| \leq \varepsilon \),

\[
\sum_{j=0}^{K-1} l_j \geq \eta^\alpha \varepsilon^{-\frac{1}{\alpha}}.
\]

Proof. We have

\[
x_K = cx_{K-1} + (1 - c)l^{-\alpha}_{K-1} = c^K x_0 + \sum_{j=0}^{K-1} (1 - c)c^{K-1-j}l^{-\alpha}_j \geq (1 - c) \sum_{j=0}^{K-1} c^{K-1-j}l^{-\alpha}_j.
\]

The minimizer of \( \sum_{j=0}^{K-1} l_j \) under the constraint \( (1 - c) \sum_{j=0}^{K-1} c^{K-1-j}l^{-\alpha}_j \leq \varepsilon \) satisfies according to Lemma 3.8 and (19),

\[
\sum_{j=0}^{K-1} l_j = \left( \frac{\varepsilon}{1 - c} \right)^{-\frac{1}{\alpha}} \left( \frac{1 - c^{K+1}}{1 - c^{1+\alpha}} \right)^{\frac{1+\alpha}{\alpha}} \geq \left( \frac{\varepsilon}{1 - c} \right)^{-\frac{1}{\alpha}} = \frac{1}{\eta^\alpha \varepsilon^{-\frac{1}{\alpha}}}.
\]

29
E  Notation: particle methods

For a particle system \( \{x^{(m)}\}_{m=1}^{M}, x^{(m)} \in \mathbb{R}^{n} \) and forward operator \( H : \mathbb{R}^{n} \to \mathbb{R}^{n} \), we denote the empirical mean and covariance operators by

\[
\bar{x} = \frac{1}{M} \sum_{m=1}^{M} x^{(m)}, \quad \bar{H} = \frac{1}{M} \sum_{m=1}^{M} H(x^{(m)}),
\]

\[
C^{x,H}(x) = \frac{1}{M} \sum_{m=1}^{M} (x^{(m)} - \bar{x}) \otimes (H(x^{(m)}) - \bar{H}),
\]

\[
C^{H,H}(x) = \frac{1}{M} \sum_{m=1}^{M} (H(x^{(m)}) - \bar{H}) \otimes (H(x^{(m)}) - \bar{H}),
\]

\[
C(x) = C^{x,x}(x) = \frac{1}{M} \sum_{m=1}^{M} (x^{(m)} - \bar{x}) \otimes (x^{(m)} - \bar{x}).
\]

F  Algorithm: Multilevel ensemble Kalman inversion

The original EKI method in \([24]\) has been derived through an artificial discrete-time data assimilation problem and was formulated as iterative scheme. For a fixed ensemble size \( M \) the time-dynamical particle system \( \{v^{(m)}_{j}\}_{m=1}^{M} \) can be written as

\[
v^{(m)}_{j+1} = v^{(m)}_{j} + C^{v,H}(v_{j})(C^{H,H}(v_{j}) + h^{-1} \Sigma)^{-1}(z^{(m)}_{j+1} - H(v^{(m)}_{j})), \quad j = 1, \ldots, J,
\]

where \( z^{(m)}_{j+1} \sim \mathcal{N}(y, h^{-1} \Sigma) \) are perturbed observations. Viewing \( h > 0 \) as step size the authors in \([33]\) motivated to take the limit \( h \to 0 \) resulting in the system of coupled SDEs \([25]\), which has been analysed rigorously in \([6, 5]\). To be more precise, under weak assumptions on the general nonlinear forward model \( H \), convergence in probability can be verified, whereas strong convergence can be verified for linear models.

We reformulate the update formula \((59)\) as

\[
v^{(m)}_{j+1} = v^{(m)}_{j} + \frac{1}{M} \sum_{r=1}^{M} \alpha^{(r),(m),l}_{j} v^{(r)}_{j},
\]

verifying the well-known subspace property of the EKI, which will not be violated for our proposed multilevel formulation. This comes from the fact, that the updating force formulated in the coordinate system spanned by the initial ensemble depends on the discretization level only through the scalar valued coordinates

\[
\alpha^{(r),(m),l}_{j} = \langle H_{l}(v^{(r)}_{j}) - \bar{H}_{l}, (C^{H_{l},H_{l}}(v^{(r)}_{j}) + h^{-1} \Sigma)^{-1}(z^{(m)}_{j+1} - H_{l}(v^{(m)}_{j})) \rangle.
\]

This observation is useful for an efficient implementation of the multilevel formulation of EKI and TEKI respectively based on the discretization scheme \((59)\) which we summarize as
We define the evolution equation of the particle system for standard EKI can be written as
\[ H_t(\cdot) = \left( \begin{array}{c} F_t(\cdot) \\ \text{Id} \end{array} \right), \quad z = \left( \begin{array}{c} y \\ 0 \end{array} \right), \quad \Sigma = \left( \begin{array}{cc} \Gamma & 0 \\ 0 & \frac{1}{\chi} C_0 \end{array} \right). \]

**Algorithm 1** Multilevel ensemble Kalman inversion (ML-EKI)

**Require:** initial ensemble \( v_0((m),0) \), ensemble size \( M \), forward model \( (H_k)_{k \geq 0} \), bias parameter \( \alpha > 0 \), rate parameter \( c \in (0,1) \), step size \( h > 0 \), time-interval length \( \tau \) such that \( N = \tau / h \in \mathbb{N} \), tolerance \( \varepsilon > 0 \).

1. set the number of iteration \( K \propto \log(\varepsilon^{-1}) \)
2. set \( x_0 = \frac{1}{M} \sum_{m=1}^{M} v_0((m),0) \)
3. **For** \( j = 0, \ldots, K - 1 \)
4. compute level \( l_j \propto \varepsilon^{-\frac{1}{2}} c^{-\frac{j-1}{2}}) \)
5. **If** \( j=0 \)
6. \( v_0((m),l_0) = v_0((m),0), \quad m = 1, \ldots, M \)
7. **Else**
8. \( v_0((m),l_j) = v_{l_{j-1}}((m),l_{j-1}), \quad m = 1, \ldots, M \)
9. **EndIf**
10. **For** \( n = 0, 1, \ldots, N - 1 \)
11. \( y_{n+1,l_j} \sim \mathcal{N}(y, h^{-1} \Gamma), \quad m = 1, \ldots, M \)
12. \( \alpha_n((r),l_j,m) = \langle F_l(y_{n+1,l_j}, (C^r_l,v_l(m) + h^{-1} \Gamma)^{-1}(y_{n+1,l_j} - F_l(v_n(m),l_j))) \]
13. \( v_{n+1,l_j} = v_n((m),l_j) + \frac{1}{M} \sum_{r=1}^{M} \alpha_n((r),l_j,m) v^{(m),l_j} \)
14. **EndFor**
15. set \( x_{j+1} = \frac{1}{M} \sum_{m=1}^{M} v_{N,l_j} \)
16. **EndFor**

**Appendix G** Proofs of Section 4

**G.1 Proof of Proposition 4.2**

**Proof.** We first fix an iteration \( j \) with level \( l_j \) and suppress the dependence of \( F_{l_j} \) on \( l_j \). The evolution equation of the particle system for standard EKI can be written as
\[
\text{d}v_t^{(m)} = -(C(v_t) + B)F^\top \Gamma^{-1}(Fv_t^{(m)} - y) \, \text{d}t + C(v_t)F^\top \Gamma^{-1/2} \, \text{d}W_t^{(m)}.
\]

We define \( \bar{v}_t := \Gamma^{-1/2} F(\bar{v}_t - v^i) \) and \( v_t^{(m)} := \Gamma^{-1/2} F(v_t^{(m)} - v^i) \) such that \( \frac{1}{2} \| \bar{v}_t \|^2 = \ell(\bar{v}_t) \) and
\[
\text{d}\bar{v}_t = -(C(v_t) + B)\bar{v}_t \, \text{d}t + C(v_t) \, \text{d}W_t
\]
with $\tilde{B} := \Gamma^{-1/2} F B F^\top \Gamma^{-1/2}$. Following Theorem 5.2 in [7] we can bound

$$\mathbb{E}[\|\tilde{t}_{s+\tau}\|^2] \leq \mathbb{E}[\|\tilde{t}_s\|^2] - \sigma \int_s^{s+\tau} \mathbb{E}[\|\tilde{t}_u\|^2] \, du,$$

where $\sigma > 0$ denotes the smallest eigenvalue of $\tilde{B}$. Hence with $\tilde{v}_{ij}^l := \frac{1}{M} \sum_{m=1}^M v_{ij}^{(m)}$ it follows

$$\mathbb{E}[\frac{1}{2} \| F_i \tilde{v}_{ij+1} - y \|^2_2] \leq (1 - \sigma \cdot \tau) \mathbb{E}[\frac{1}{2} \| F_i \tilde{v}_{ij} - y \|^2_2],$$

where we have defined $\| \cdot \|^2_2 = \| \Gamma^{-1/2} \|_2^2$. With Assumption 4.1 and the reverse triangle inequality $\|F x\| - \|F y\| \leq \|F x - F y\|$ we have that

$$e_{j+1} = \mathbb{E}[\frac{1}{2} \| F_i \tilde{v}_{ij+1} - y \|^2_2] = \mathbb{E}[\frac{1}{2} \| F_i \tilde{v}_{ij+1} - y \|^2_2] + \mathbb{E} \left[ \frac{1}{2} \| F v_{ij+1}^l - y \|^2_2 - \frac{1}{2} \| F_i v_{ij+1}^l - y \|^2_2 \right]$$

$$\leq (1 - \sigma \cdot \tau) \mathbb{E}[\frac{1}{2} \| F_i \tilde{v}_{ij}^l - y \|^2_2] + b_1 l_j^{-\alpha}$$

$$\leq (1 - \sigma \cdot \tau) \mathbb{E}[\frac{1}{2} \| F v_{ij+1}^l - y \|^2_2] + (1 - \sigma \cdot \tau) b_1 l_j^{-\alpha} + b_1 l_j^{-\alpha}$$

$$\leq (1 - \sigma \cdot \tau) e_j + b l_j^{-\alpha},$$

for some constant $b_1 > 0$ and $b = (2 - \sigma \cdot \tau) b_1$. We note that we have used that $\mathbb{E}[\|v_{ij+1}^l\|^2]$ remains uniformly bounded, see e.g. Lemma 5 in [12]. Moreover, since we assumed finite second moments of the initial distribution $Q_0$, we have that $\mathbb{E}[\|\tilde{v}_0\|^2] < \infty$ and by local Lipschitz continuity of $x \mapsto \frac{1}{2} \| F x - y \|^2_2$, we have that $e_0 = \mathbb{E}[\| F \tilde{v}_0 - y \|^2_2] < \infty$. With the above computations we have verified that the error quantity $e$ satisfies the decay assumption [8] (respectively the generalization in Rmk. 3.11) and therefore, the assertion follows by application of Theorem 3.7 and Theorem 3.10.

**G.2 Proof of Proposition 4.3**

*Proof.* We again fix an iteration $j$ with level $l_j$ and suppress the dependence of $F_i$ on $l_j$. The evolution equation of the particle system for TEKI can be written as

$$d v_{ij}^{(m)} = -(C(v_i) + B)(F^\top \Gamma^{-1}(F v_{ij}^{(m)} - y) + \lambda C_0^{-1} v_{ij}^{(m)}) \, dt$$

$$+ C(v_i) F^\top \Gamma^{-1/2} dW_t^{(m)} + \sqrt{\lambda} C(v_i) C_0^{-1/2} dW_t^{(m)}.$$

Next, we define $\bar{v}_t := \Sigma^{-1/2} H(\bar{v}_t - x_*)$ and $v_t^{(m)} := \Sigma^{-1/2} H(v_t - x_*)$, where $x_* \in \mathcal{X}$ is the unique minimizer of $\ell_R$ and $H, \Sigma$ are defined in (24), such that $\frac{1}{2} \| v_t \|^2 = \frac{1}{2} \| F(v_t - x_*) \|^2 + \frac{1}{2} \| \bar{v}_t - x_* \|^2_2$. Since $x_* \in \mathcal{X}$ is the unique minimizer of $\ell_R$ it holds true that

$$0 = \nabla_x \ell_R(x_*) = F^\top \Gamma^{-1}(F x_* - y) + \lambda C_0^{-1} x_* = H^\top \Sigma^{-1}(H x_* - z),$$

and with $\tilde{B} := \Sigma^{-1/2} H B H^\top \Sigma^{-1/2}$ we can write

$$d \bar{v}_t = -(C(v_t) + \tilde{B}) \bar{v}_t \, dt + C(v_t) \, d\bar{W}_t.$$

The assertion follows similarly to the proof of Proposition 4.2. \qed

32
The multilevel interacting Langevin sampler is based on its particle approximation \cite{36}. Due to the finite number of ensemble size $M$, the resulting algorithm contains an additional empirical error according to the mean-field limit represented by the Fokker–Planck equation \cite{31}. We refer to \cite{13} for a detailed analysis of large ensemble size limit. We are going to solve these systems of coupled SDEs by a forward Euler-Maruyama method and emphasize that other numerical approximation schemes for SDEs can be applied as well. The resulting multilevel sampling algorithm is summarized in below.

Algorithm 2 Multilevel interacting Langevin sampler (ML-ILS)

**Require:** initial distribution $q_0$, ensemble size $M$, gradient of log-likelihood $(\nabla \Phi^l_R)_{l \geq 0}$, bias parameter $\alpha > 0$, rate parameter $c \in (0, 1)$, step size $h > 0$, time-interval length $\tau$ such that $N = \tau/h \in \mathbb{N}$, tolerance $\varepsilon > 0$.

1: set the number of iteration $K \propto \log(\varepsilon^{-1})$
2: Initialize particle system as i.i.d. sample $v^{(m),0}_0 \sim q_0$
3: set $\rho_0 = \frac{1}{M} \sum_{m=1}^{M} \delta_{v^{(m),0}}$
4: For $j = 0, \ldots, K - 1$
5: compute level $l_j \propto \varepsilon^{-\frac{1}{\alpha} c^{\frac{K-1-j}{1+\alpha}}}$
6: If $j = 0$
7: $v^{(m),l_0}_0 = v^{(m),0}_0$, $m = 1, \ldots, M$
8: Else
9: $v^{(m),l_j}_0 = v_{\tau}^{(m),l_{j-1}}$, $m = 1, \ldots, M$
10: EndIf
11: For $n = 0, 1, \ldots N - 1$
12: $\xi^{(m),l_j}_{n+1} \sim \mathcal{N}(0, I)$
13: $\Delta W^{(m),l_j}_{n+1} = \sqrt{h}\xi^{(m),l_j}_{k+1}$
14: $g^{(m),l_j}_{n} = -hC(v^{(m),l_j}_{n})\nabla \Phi^l_R(v^{(m),l_j}_{n}) + \sqrt{2C(v^{(m),l_j}_{n})}\Delta W^{(m),l_j}_{n+1}$
15: $v^{(m),l_j}_{n+1} = v^{(m),l_j}_{n} + g^{(m),l_j}_{n}$
16: EndFor
17: set $\rho_{j+1} = \frac{1}{M} \sum_{m=1}^{M} \delta_{v^{(m),l_j}_{N}}$
18: EndFor
Proposition 2] it follows that there exists a constant $c \in (0, 1)$ such that $\text{KL}(\rho_{j+1} \| \rho_*^j) \leq \text{KL}(\rho_j \| \rho_*^j)$. Furthermore, under Assumption 5.1 it holds true that

$$|\Phi(\rho_j) - \Phi_l(\rho_j)| \leq bl^{-\alpha}$$

since by definition of the KL divergence we have that

$$|\Phi_l(\rho) - \Phi(\rho)| = |\text{KL}(\rho \| \rho_0) - \text{KL}(\rho \| \rho_*)|$$

$$= |\int_{\mathbb{R}^n_x} \rho(x) \log(\rho(x)) \, dx - \int_{\mathbb{R}^n_x} \rho(x) \log(\rho_0(x)) \, dx|$$

$$- \int_{\mathbb{R}^n_x} \rho(x) \log(\rho(x)) \, dx - \int_{\mathbb{R}^n_x} \rho(x) \log(\rho_0(x)) \, dx|$$

$$\leq \int_{\mathbb{R}^n_x} \rho(x) |\ell_l^j(x) - \ell_R(x)| \, dx.$$}

With Assumption 5.1 it follows that

$$|\Phi_l^j(\rho_j) - \Phi(\rho_j)| \leq \int_{\mathbb{R}^n_x} \rho_j(x) |\ell_l^j(x) - \ell_R(x)| \, dx$$

$$\leq b \int_{\mathbb{R}^n_x} \rho_j(x) \|F(x) - F_l(x)\|^2 \, dx \leq bl^{-\alpha}.$$}

Finally, we obtain

$$\text{KL}(\rho_{j+1} \| \rho_*) = \text{KL}(\rho_{j+1} \| \rho_*^j) + \left(\text{KL}(\rho_{j+1} \| \rho_*) - \text{KL}(\rho_{j+1} \| \rho_*^j)\right)$$

$$\leq c\text{KL}(\rho_j \| \rho_*^j) + l^{-\alpha}_j$$

$$\leq c\text{KL}(\rho_j \| \rho_*) + (1 + c)l^{-\alpha}_j.$$}

With the above computations we have verified that the error quantity $e$ satisfies the decay assumption (8) (respectively the generalization in Rmk. 3.11) and therefore, the assertion follows by application of Theorem 3.7 and Theorem 3.10.

### J Details for Sec. 6

The numerical approximations $F_l$ to $F$ are computed as follows: Given $x \in \mathbb{R}^{n_x}$, we let $F_l : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ be the map defined by $F_l(f) = \mathcal{O}(u^i)$, where $u^i$ denotes the finite element solution to (5) using continuous piecewise linear finite elements on a uniform mesh with meshwidth $2^{-\tau(l)}$ where $\tau(l) = \lfloor \log(l)/\log(2) \rfloor$ such that $l = 2^{x(l)}$. Since $\mathcal{O} : H^1_0(D) \to \mathbb{R}^{n_y}$ is continuous, it can be shown that $\|F(x) - F_l(x)\| \lesssim l^{-1}\|x\|$ for all $l$ and all $x \in \mathbb{R}^{n_x}$, i.e. the convergence rate $\alpha$ in Section 3 equals 1. As a prior on the parameter space $\mathbb{R}^{n_x}$ we chose $Q_0 = \mathcal{N}(0, C_0)$ with $C_0 = \text{diag}(i^{-2\beta}, i = 1, \ldots, n_x)$ for some fixed $\beta > 0$.

The truth $x^\dagger$ was generated as a draw from the prior. Figure 1 shows the error convergence of multilevel and single-level TEKI in Algorithm 1 in Appendix F with the parameters $\beta = 1$.

34
and \( n_x = 100 \). The ensemble size was size \( M = 50 \) and the step size of the discretization scheme (59) was chosen a \( h = 0.1 \). The plotted error quantity is

\[
e_{K(\varepsilon)} = \mathbb{E} \left[ \frac{1}{2} \| F_{\text{ref}}(x_K(\varepsilon) - x_\ast) \|_\Gamma^2 + \frac{\lambda}{2} \| x_K(\varepsilon) - x_\ast \|_{C_0}^2 \right],
\]

with the reference solution \( x_\ast \) computed via

\[
x_\ast = (F_{\text{ref}}^\top \Gamma^{-1} F_{\text{ref}} + \lambda C_0^{-1})^{-1} F_{\text{ref}}^\top \Gamma^{-1} y.
\]

The cost quantity was computed as in (9).

The second plot in Figure 1 shows the convergence of the posterior mean for the multilevel interacting Langevin sampler (ILS) in Algorithm 2 in Appendix H. In this case we chose \( M = 2000 \) particles, \( \tau = 0.1 \) and the step size \( h = 0.001 \) of the Euler-Maruyama scheme. The plotted error quantity shows

\[
\mathbb{E} \left[ \frac{1}{2} \| f(\cdot, x_{K(\varepsilon)}) - f(\cdot, x_\ast) \|_{L^2(D)}^2 \right],
\]

where

\[
x_\ast = (F_{\text{ref}}^\top \Gamma^{-1} F_{\text{ref}} + C_0^{-1})^{-1} F_{\text{ref}}^\top \Gamma^{-1} y,
\]

which is the posterior mean on reference accuracy level \( 2^{14} \), and coincides with the Tikhonov regularized solution. We see a similar complexity gains as for multilevel TEKI.