Multilevel Optimization for Inverse Problems

Simon Weissmann SIMON.WEISSMANN@UNI-HEIDELBERG.DE
Universität Heidelberg, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, D-69120 Heidelberg, Germany

Ashia Wilson ASHIA07@MIT.EDU
Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Sciences, Cambridge, MA, 02139, USA

Jakob Zech JAKOB.ZECH@UNI-HEIDELBERG.DE
Universität Heidelberg, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, D-69120 Heidelberg, Germany

Editors: Po-Ling Loh and Maxim Raginsky

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.

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) Giles (2008); Heinrich (2001) 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 Dodwell et al. (2015) as well as to deterministic quadrature rules such as sparse grid Haji-Ali et al. (2016); Zech et al. (2019) and quasi-Monte Carlo methods Giles and Waterhouse (2009); Dick et al. (2017).

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

$$\min_{x\in\mathcal{X}} \ \Phi(x), \tag{1}$$

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 Φ (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 Φ, 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 Martin and Nobile (2021); Martin et al. (2021) which use multilevel optimization on an optimal control problem, and Alsup et al. (2021), 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 a 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

$$Y = F(x) + \eta, \tag{2}$$

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 Hadamard (1902), 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

$$\Phi(x) := \ell(x, y) + R(x), \qquad \ell(x, y) = \frac{1}{2}\|\Gamma^{-1/2}(F(x) - y)\|^2_{\mathbb{R}^{n_y}}, \tag{3}$$

where ℓ 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 Engl et al. (1989) with $R(x) = \frac{\lambda}{2}\|C_0 x\|^2_{\mathcal{X}}$ and total variation regularization Chambolle et al. (2010); Rudin et al. (1992). 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 Φ in (3). For motivation and further discussion of regularization methods to solve (2), see, e.g., Engl et al. (1996); Benning and Burger (2018) and references therein.

2.2. Bayesian inference

In the Bayesian approach (e.g. Stuart (2010)) 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 (Stuart, 2010, Theorem 6.31), the posterior Q^y_* is then well-defined, absolutely continuous with respect to the prior, and $Q^y_*(\mathrm{d}x) = \frac{1}{Z}\exp(-\ell(x, y))Q_0(\mathrm{d}x)$, where $Z = \int_{\mathcal{X}} \exp(-\ell(x, y))Q_0(\mathrm{d}x) \in \mathbb{R}$ is a normalizing constant and ℓ is given by (3).

Suppose for the moment that $\mathcal{X} = \mathbb{R}^{n_x}$ 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^y_* has Lebesgue density $\rho_*(x) = \frac{1}{Z}\exp(-\ell(x, y) - R(x))$, with $R(x) = \frac{\lambda}{2}\|C_0^{-1/2}x\|^2_{\mathbb{R}^{n_x}}$. In the Bayesian framework, solving the inverse problem amounts to sampling from the posterior. One way to achieve this is by minimizing the objective

$$\Phi(\psi) = \mathrm{KL}(\psi \| \rho_*), \tag{4}$$

for ψ in a given family of (tractable) probability distributions on $\mathcal{X}$. Here KL stands for the Kullback–Leibler divergence Kullback and Leibler (1951). 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, the minimization of either objective requires multiple evaluations of the forward model F, which might be very costly in practice. To explain this further, we now discuss a simple PDE driven inverse problem (i.e. evaluating F requires to solve a PDE) which will serve as our running example throughout. We emphasize, that our analysis has implications far beyond this toy problem, since PDE constrained optimization has a wide range of applications in various fields such as shape optimization, optimal control and—as discussed in the present paper—parameter estimation, see, e.g., Hinze et al. (2008); Belov (2012).

Example 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*

$$\begin{cases} -\Delta u_f(s) + u_f(s) = f(s) & s \in D, \\ u_f(s) = 0 & s \in \partial D, \end{cases} \tag{5}$$

has a unique weak solution $u_f \in H^2(D) \cap H_0^1(D) \subseteq \mathcal{X}$. 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 Φ (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 Φ as in (1), we consider an abstract optimization method described by the fixed point iteration

$$x_{k+1} := \Psi(x_k), \quad x_0 \in \mathcal{X}. \tag{6}$$

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

Remark 1 *In practice Ψ_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 Ψ in the update rule (6) with Ψ_l leads to

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

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 kth iterate x_k by e_k. For example, if x_* is the unique minimizer of Φ, 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 Ψ by Ψ_l. Other forms of convergence, such as polynomial convergence (as occurs, e.g., for non-strongly convex objectives), are work in progress.

Assumption 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* $k \in \mathbb{N}$

$$e_{k+1} \leq c e_k + l_k^{-\alpha}, \tag{8}$$

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

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

We motivate Assumption (2) by verifying conditions (8) and (9) on several examples.

Example 2 *Suppose that the objective* Φ *is* L*-smooth and* μ*-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* $0 \leq \eta < \infty$

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

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

$$x_{k+1} = x_k - \eta_k g_{l_k}(x_k) \tag{11}$$

can be shown to satisfy error decay (8) *with* $e_k = \|x_k - x_*\|_{\mathcal{X}}$*. Interpreting* l_k *as the computational cost of evaluating* g_{l_k}*, the overall cost to compute* x_k *follows our cost model* (9)*. A similar statement holds for stochastic gradient descent after replacing* $g_{l_k}(x)$ *in* (11) *with* $G_{l_k}(x)$ *and setting* $e_k = \mathbb{E}[\|x_{k+1} - x_*\|_{\mathcal{X}}]$*. Moreover, accelerated versions of both algorithms can be shown to satisfy* (8)*.*

Details for Example 2 and further discussion of the implications of our results for gradient descent, accelerated gradient descent and their stochastic versions are given in Appendix B. While we do not provide details for other variants of these basic gradient algorithms (e.g., SVRG Johnson and Zhang (2013), FISTA Beck and Teboulle (2009) and the extragradient method Monteiro and Svaiter (2013)), multilevel formulations of these algorithms are possible under our framework and will be left to future work.

Remark 3 *Assumption 2 states that the relation between computational cost and corresponding error is of the type* "$\text{error} \sim \text{cost}^{-\alpha}$" *for some* $\alpha > 0$*. To clarify, consider the following examples in the context of applying gradient descent to minimize* Φ*:*

- *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 1. Further details are contained in Appendix C.

Example 3 (Continuation of Example 1) *Let $F(f) = \mathcal{O}(u_f)$, where u_f solves (5) and $\mathcal{O} : L^2(D) \to \mathbb{R}^{n_y}$ is bounded linear, i.e. $\mathcal{O}(p) = \int_D \xi p$ for some $\xi \in L^2(D, \mathbb{R}^{n_y})$ and all $p \in L^2(D)$.*

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

$$\nabla\Phi(f) = u_h + \lambda f \in L^2(D), \tag{12}$$

where u_h solves (5) with right-hand side $h(\cdot) = (\mathcal{O}(u_f) - y)^\top \Gamma^{-1}\xi(\cdot) \in L^2(D)$. 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^l satisfying $\|u_f - u_f^l\|_{L^2(D)} \lesssim l^{-1}$, and subsequently with $\tilde{h} = (\mathcal{O}(u_f^l) - y)^\top \Gamma^{-1}\xi$, a FEM approximation $u_{\tilde{h}}^l$ satisfying $\|u_h - u_{\tilde{h}}^l\|_{L^2(D)} \lesssim l^{-1}$. Here l corresponds to the dimension of the FEM space, and can thus be interpreted as the complexity of computing $u_{\tilde{h}}^l$. Note, $g_l(f) := u_{\tilde{h}}^l + \lambda f \in L^2(D)$ yields an approximation to $\nabla\Phi(f)$ s.t. $\|\nabla\Phi(f) - g_l(f)\|_{L^2(D)} \lesssim l^{-1}$. 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

$$e_k \le c(ce_{k-2} + l_{k-1}^{-\alpha}) + l_k^{-\alpha} \le \cdots \le c^k e_0 + \sum_{j=0}^{k-1} c^{k-1-j} l_j^{-\alpha} =: \tilde{e}_k((l_j)_j). \tag{13}$$

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

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

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 $j = 0, \dots, K-1$. By assumption (9), $\mathrm{cost}(x_K) = K\bar{l}_K$. We wish to minimize the cost under the error constraint $\tilde{e}_K \le \varepsilon$. To slightly simplify the problem for the moment, we instead demand both terms in the definition of $\tilde{e}_k$ in (13) to be bounded by $\frac{\varepsilon}{2}$ (so that in particular $\tilde{e}_K \le \varepsilon$). More precisely, $\bar{l}_K > 0$ should be minimal such that

$$c^K e_0 \le \frac{\varepsilon}{2}, \qquad (\bar{l}_K)^{-\alpha}\frac{1 - c^K}{1 - c} \le \frac{\varepsilon}{2}. \tag{14}$$

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

$$\bar{l}_K \ge \left(2\frac{1 - c^K}{(1 - c)\varepsilon}\right)^{1/\alpha}. \tag{15}$$

We choose $\bar{l}_K$ so that (15) holds with equality. Given $K \mapsto \mathrm{cost}(x_K) = Kl_K$ is monotonically increasing, in order to get error ε at possibly small cost the following choices suffice:

$$\bar{l}_K(\varepsilon) := \left(2\frac{1 - \frac{\varepsilon}{2e_0}}{(1 - c)\varepsilon}\right)^{1/\alpha}, \qquad K(\varepsilon) := \left\lceil \frac{\log(\varepsilon/(2e_0))}{\log(c)} \right\rceil. \tag{16}$$

We next introduce a notion of optimality, and then summarize our observations in Theorem 5.

Definition 4 (Quasi-optimal single level choice) *A family of reals* $\bar{l}_K(\varepsilon) > 0$ *and integers* $K(\varepsilon) \in \mathbb{N}$ *satisfying* $\tilde{e}_{K(\varepsilon)}(\bar{l}_K(\varepsilon)) \leq \varepsilon$ *for all* $\varepsilon > 0$*, is a* quasi-optimal single level choice *iff*

$$K(\varepsilon)\bar{l}_K(\varepsilon) = O\big(\inf\{\hat{K}\hat{l} \,:\, \tilde{e}_{\hat{K}}(\hat{l}) \leq \varepsilon\},\ \hat{K} \in \mathbb{N},\ \hat{l} > 0\big) \qquad \textit{as} \quad \varepsilon \to 0.$$

Theorem 5 (Single-level convergence) *Equation* (16) *defines a quasi-optimal single level choice. It holds*

$$\mathrm{cost}_{\mathrm{SL}}(\varepsilon) := K(\varepsilon)\bar{l}_K(\varepsilon) \simeq \log(\varepsilon^{-1})\varepsilon^{-\frac{1}{\alpha}} \qquad \textit{as} \quad \varepsilon \to 0. \tag{17}$$

Proof For the proof see Appendix A.1. ■

Due to the quasi-optimality, the (single-level-) cost behaviour (17) 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 $\mathrm{cost}(x_K) = \sum_{j=0}^{K-1} l_{K,j}$ is minimized under the constraint of both terms in (13) being bounded by $\frac{\varepsilon}{2}$, i.e. such that

$$c^K e_0 \leq \frac{\varepsilon}{2}, \qquad \sum_{j=0}^{K-1} c^{K-1-j} l_{K,j}^{-\alpha} \leq \frac{\varepsilon}{2}. \tag{18}$$

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

Lemma 6 *For every* $K \in \mathbb{N}$, $\varepsilon > 0$

$$l_{K,j}(\varepsilon) = C_{K,\varepsilon} \cdot c^{\frac{K-1-j}{1+\alpha}}, \quad C_{K,\varepsilon} = \left(\frac{\varepsilon}{2}\right)^{-\frac{1}{\alpha}} \left(\frac{1 - c^{\frac{K}{1+\alpha}}}{1 - c^{\frac{1}{1+\alpha}}}\right)^{\frac{1}{\alpha}} \tag{19}$$

minimizes $\sum_{j=0}^{K-1} l_{K,j}$ *under the constraint* $\sum_{j=0}^{K-1} c^{K-1-j} l_{K,j}^{-\alpha} \leq \frac{\varepsilon}{2}$.

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 $\delta = c^{\frac{1}{1+\alpha}} \in (0,1)$,

$$\sum_{j=0}^{K-1} l_{K,j} = C_{K,\varepsilon} \sum_{j=0}^{K-1} c^{\frac{K-1-j}{1+\alpha}} = \left(\frac{\varepsilon}{2}\right)^{-\frac{1}{\alpha}} \left(\frac{1 - c^{\frac{K}{1+\alpha}}}{1 - c^{\frac{1}{1+\alpha}}}\right)^{\frac{1+\alpha}{\alpha}}. \tag{20}$$

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 (18), which leads to

$$K(\varepsilon) = \left\lceil \frac{\log(\varepsilon/(2e_0))}{\log(c)} \right\rceil, \qquad l_{K,j}(\varepsilon) = \left(\frac{\varepsilon}{2}\right)^{-\frac{1}{\alpha}} c^{\frac{K(\varepsilon)}{1+\alpha}} c^{-\frac{1+j}{1+\alpha}} \left(\frac{1 - c^{\frac{1}{1+\alpha}}}{1 - c^{\frac{1}{1+\alpha}}}\right)^{\frac{1+\alpha}{\alpha}} \qquad \forall j < K(\varepsilon). \tag{21}$$

Observing that $c^{\frac{K(\varepsilon)}{1+\alpha}}$ behaves like $\varepsilon^{\frac{1}{1+\alpha}}$, and $1 - c^{\frac{K(\varepsilon)}{1+\alpha}} \to 1$ as $\varepsilon \to 0$, we find $l_{K,j}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha(1+\alpha)}} c^{-\frac{1+j}{1+\alpha}}$, with lower and upper bounds independent of ε, K and j.

Definition 7 (Quasi-optimal multilevel choice) *A family $((l_{K,j}(\varepsilon))_{j<K(\varepsilon)})_{\varepsilon>0}$ of sequences satisfying $\tilde{e}_{K(\varepsilon)}((l_{K,j}(\varepsilon))_{j<K(\varepsilon)}) \le \varepsilon$ for all $\varepsilon > 0$ is a* quasi-optimal multilevel choice*, iff*

$$\sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) = O\Bigg(\inf \Bigg\{ \sum_{j=0}^{\hat{K}-1} \hat{l}_j \ : \ \tilde{e}_{\hat{K}}((\hat{l}_j)_{j<\hat{K}}) \le \varepsilon, \ \hat{K} \in \mathbb{N}, \ \hat{l}_j > 0 \, \forall j < \hat{K} \Bigg\} \Bigg) \qquad \textit{as} \quad \varepsilon \to 0.$$

Theorem 8 (Multilevel convergence) *Equation (21) defines a quasi-optimal multilevel choice for $\varepsilon \in (0, e_0)$. It holds*

$$\text{cost}_{\text{ML}}(\varepsilon) := \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha}} \qquad \textit{as} \quad \varepsilon \to 0. \tag{22}$$

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 $\tilde{e}_K \le \varepsilon$ cannot be improved. Comparing with the single-level method in Theorem 8, 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 8 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 5.

Remark 9 *Suppose that e_k generated with levels l_j satisfies instead of (8) the relaxed condition $e_{k+1} \le ce_k + Cl_k^{-\alpha}$ for some constant $C \ge 1$. Then, $\tilde{e}_k$ generated with the levels $\tilde{l}_k := C^{1/\alpha} l_k$ satisfies $\tilde{e}_{k+1} \le c\tilde{e}_k + C\tilde{l}_k^{-\alpha} = c\tilde{e}_k + l_k^{-\alpha}$. The cost quantity $\sum_{j=0}^{k-1} \tilde{l}_j$ only increases by the constant factor $C^{1/\alpha}$ compared to $\sum_{j=0}^{k-1} l_j$. Hence, the asymptotic cost behaviour stated in Theorem 5 (single-level) and Theorem 8 (multi-level) remains valid also for $C > 1$.*

We next continue our discussion of Example 3, for details see Appendix C.4.

Example 4 (Continuation of Example 3) *It can be shown that the regularized objective Φ in Example 3 is λ-strongly convex and L-smooth with $L = \|\xi\|_{L^2(D)}^2 \|\Gamma^{-1}\| + \lambda$. Consider the multilevel gradient descent method $f_{j+1} = f_j - \eta g_{l_j}(f_j)$, where g_l is the approximation to $\nabla\Phi$ from Example 3. The level choice (21) then yields $\|f_* - f_{K(\varepsilon)}\|_{L^2(D)} \lesssim \varepsilon$, for the unique minimizer f_* of Φ. The cost quantity, which corresponds to the aggregated computational cost of all required FEM approximations to compute $f_{K(\varepsilon)}$, behaves like ε^{-1} as $\varepsilon \to 0$.*

Finally we point out that our analysis and notion of quasi-optimality are based on the constraint $\tilde{e}_k \le \varepsilon$ (rather than $e_k \le \varepsilon$), 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 (22) 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., Schillings and Stuart (2017); Blömker et al. (2019); Kovachki and Stuart (2019). 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 Iglesias et al. (2013), 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

$$\Phi(x) = \frac{1}{2}\|\Sigma^{-1/2}(H(x) - z)\|^2_{\mathbb{R}^{n_z}}. \tag{23}$$

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

$$H = F, \qquad z = y \in \mathbb{R}^{n_y}, \qquad \Sigma = \Gamma, \tag{24}$$

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

$$H = \begin{pmatrix} F \\ \mathrm{Id} \end{pmatrix}, \quad z = \begin{pmatrix} y \\ \mathbf{0}_{\mathbb{R}^{n_x}} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Gamma & 0 \\ 0 & \frac{1}{\lambda}C_0 \end{pmatrix}, \tag{25}$$

yields the regularized objective $\Phi(x)$ in (3) with regularizer $R(x) = \frac{\lambda}{2}\|C_0^{-1/2}x\|^2_{\mathbb{R}^{n_x}}$. We refer to EKI applied to the unregularized and regularized objective as *standard EKI* and *Tikhonov regularized EKI (TEKI)*, respectively. See Chada et al. (2020); Weissmann et al. (2022) for more details on TEKI.

We consider the continuous-time formulation of EKI Blömker et al. (2018); Blömker et al. (2021): For a fixed ensemble size $M \in \mathbb{N}$, let $v_t^{(m)} \in \mathbb{R}^{n_x}$, $m = 1, \dots, M$, satisfy the coupled system of stochastic differential equations (SDEs)

$$\begin{aligned} \mathrm{d}v_t^{(m)} &= C^{v,H}(v_t)\Sigma^{-1}(z - H(v_t^{(m)}))\,\mathrm{d}t + C^{v,H}(v_t)\Sigma^{-1/2}\,\mathrm{d}W_t^{(m)}, \\ v_0^{(m)} &\overset{\text{i.i.d.}}{\sim} Q_0 \qquad m = 1, \dots, M. \end{aligned} \tag{26}$$

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 $C^{v,H} \in \mathbb{R}^{n_x \times n_z}$ denotes a mixed sample covariance, see Appendix E for the precise formula. Under certain assumptions, it can be shown that (26) is well-posed, i.e. existence of unique and strong solutions can be guaranteed, and their average converges to the minimizer of Φ (Blömker et al. (2019)).

To motivate this behaviour, suppress for the moment the diffusion term in (26), and consider a linear forward map $H \in \mathbb{R}^{n_z \times n_x}$. Then

$$\frac{\mathrm{d}v_t^{(m)}}{\mathrm{d}t} = -C(v_t)H^\top\Sigma^{-1}(Hv_t^{(m)} - z) = -C(v_t)\nabla_v\Phi(v_t^{(m)}),$$

for $\Phi(x) = \frac{1}{2}\|\Sigma^{-1/2}(Hx - z)\|^2_{\mathbb{R}^{n_z}}$. Hence, in the linear and deterministic setting the EKI is a preconditioned gradient flow w.r.t. the data misfit ℓ (or w.r.t. the Tikhonov regularized data misfit). We refer to Chada and Tong (2022); Weissmann (2022) 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 (26). Fixing $\tau > 0$, with $t_j := j \cdot \tau$ we denote by $v_{t_{j+1}}^{(m),l_j}$ the solution of the coupled system of SDEs in integral form

$$v_{t_{j+1}}^{(m),l_j} = v_{t_j}^{(m),l_{j-1}} + \int_{t_j}^{t_{j+1}} C^{v,H_{l_j}}(v_t^{l_j})\Sigma^{-1}(z - H_{l_j}(v_t^{(m),l_j}))\,\mathrm{d}t + \int_{t_j}^{t_{j+1}} C^{v,H_{l_j}}(v_t^{l_j})\Sigma^{-1/2}\,\mathrm{d}W_t^{(m)}, \tag{27}$$

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

$$x_j := \frac{1}{M}\sum_{m=1}^{M} v_{t_j}^{(m),l_j}. \tag{28}$$

We discuss a discretization scheme for the SDE (27) 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 (23) 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 $\operatorname{rank}(FF^\top) = n_y$. 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 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 Ern and Guermond (2021); Sauter and Schwab (2011).

Assumption 10 (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 $\|F - F_l\|^2_{\mathbb{R}^{n_y \times n_x}} \le bl^{-\alpha}$.*

In the following we use the notation H_l to denote H as in (24) or (25) 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., Anderson (2007, 2009); Tong et al. (2016). Specifically, for some fixed SPD matrix $B \in \mathbb{R}^{n_x \times n_x}$, (27) is replaced by the stabilized dynamics

$$v_{t_{j+1}}^{(m),l_j} = \int_{t_j}^{t_{j+1}} (C(v_t^{l_j}) + B)H_{l_j}^\top \Sigma^{-1}(z - H_{l_j} v_t^{(m),l_j})\,\mathrm{d}t + \int_{t_j}^{t_{j+1}} C(v_t^{l_j})H_{l_j}^\top \Sigma^{-1/2}\,\mathrm{d}W_t^{(m)}. \tag{29}$$

Assume that an evaluation of F_l has cost $O(l)$. Then, approximating (29) 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 $\sum_{j=0}^{K-1} l_j$ introduced in (9), can be interpreted as the computational cost of computing x_K in (28). We give convergence result for EKI and TEKI in the case of noise-free and noisy data respectively.

Let x_j in (28) be the mean of the particle system driven by (29), with H (and H_l) as in (24).

Proposition 11 (Multilevel EKI) *Let Assumption 10 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 $l_{K,j}(\varepsilon)$, $j = 0, \dots, K(\varepsilon) - 1$, given by (21),*

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

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

Proof For the proof see Appendix G.1. ■

Now let x_j in (28) be the mean of the particle system driven by (29), with H (and H_l) as in (25).

Proposition 12 (Multilevel TEKI) *Let Assumption 10 be satisfied, $x_* \in \mathbb{R}^{n_x}$ be the unique minimizer of Φ in (23) with H given by (25), 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 $l_{K,j}(\varepsilon)$, $j = 0, \dots, K(\varepsilon) - 1$, given by (21),*

$$e_{K(\varepsilon)} := \mathbb{E}\left[\frac{1}{2}\|\Gamma^{-1/2}F(x_{K(\varepsilon)} - x_*)\|^2_{\mathbb{R}^{n_y}} + \frac{\lambda}{2}\|C_0^{-1/2}(x_{K(\varepsilon)} - x_*)\|^2_{\mathbb{R}^{n_x}}\right] \leq \varepsilon,$$

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

Proof For the proof see Appendix G.2. ■

In both cases there holds an analogous statement for the single-level choice in (16) 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. $\mathcal{X} = \mathbb{R}^{n_x}$, 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

$$\rho_*(x) = \frac{1}{Z}\exp(-\ell_R(x)), \qquad \ell_R(x) := \frac{1}{2}\|\Gamma^{-1/2}(F(x) - y)\|^2_{\mathbb{R}^{n_y}} + \frac{\lambda}{2}\|C_0^{-1/2}x\|^2_{\mathbb{R}^{n_x}} \tag{30}$$

with $Z = \int_{\mathbb{R}^{n_x}} \exp(-\ell_R(x))\,\mathrm{d}x$. To explain the idea of approximating ρ_* 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

$$\mathrm{d}v_t = -\nabla_x \ell_R(v_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad v_0 \sim q_0. \tag{31}$$

The evolution of its distribution ρ_t is described by the Fokker-Planck equation, see e.g. Pavliotis (2014),

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla \ell_R) + \Delta \rho_t, \quad \rho_0 = q_0. \tag{32}$$

Under certain assumptions on ℓ_R the Markov process v_t is ergodic and ρ_* is its unique invariant distribution. Furthermore, it is possible to describe the rate of convergence in terms of the gradient flow structure given by (32), see for example (Pavliotis, 2014, Theorem 4.9).

5.1. Interacting Langevin sampler

We consider an interacting Langevin particle system introduced in Garbuno-Inigo et al. (2020). Let $(v_t^{(m)})_{m=1}^M$ solve

$$\mathrm{d}v_t^{(m)} = -C(v_t)\nabla\ell_R(v_t^{(m)})\,\mathrm{d}t + \sqrt{2C(v_t)}\,\mathrm{d}W_t^{(m)}, \tag{33}$$

with $v_0^{(m)} \sim q_0$ i.i.d., where $(W_t^{(m)})_{m=1}^M$ 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 (31)-(32) (e.g. Ding and Li (2021b)):

$$\mathrm{d}v_t = -C(\rho)\nabla\ell_R(v_t)\,\mathrm{d}t + \sqrt{2C(\rho)}\,\mathrm{d}W_t,$$

where

$$m(\rho) = \int_{\mathbb{R}^{n_x}} v\rho(v)\,\mathrm{d}v, \qquad C(\rho) = \int_{\mathbb{R}^{n_x}} \left((v - m(\rho)) \otimes (v - m(\rho))\right)\rho(v)\,\mathrm{d}v \tag{34}$$

and

$$\partial_t\rho_t = \nabla\cdot(\rho_t C(\rho_t)\nabla\ell_R) + \mathrm{Tr}(C(\rho_t)\nabla^2\rho_t), \quad \rho_0 = q_0. \tag{35}$$

5.2. Mean-field and discrete multilevel interacting Langevin sampler

Fix $\tau > 0$ and set $t_j := j\cdot\tau$. In the following let ℓ_R^l be given by (30) with the forward model $F : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ replaced by an approximation F_l. We consider the discrete time process $(\rho_j)_{j=1,\dots,K}$ iteratively defined by $\rho_{j+1} = \rho_{t_{j+1}}^{l_j}$, where $\rho_0^{l_1} = q_0$ and $\rho_t^{l_j}$ solves (35) on a time interval of length τ initialized with the previous distribution, i.e.

$$\partial_t\rho_t^{l_j} = \nabla\cdot(\rho_t^{l_j} C(\rho_t^{l_j})\nabla\ell_R^{l_j}) + \mathrm{Tr}(C(\rho_t^{l_j})\nabla^2\rho_t^{l_j}), \quad \rho_{t_j}^{l_j} = \rho_{t_j}^{l_{j-1}}. \tag{36}$$

While the subsequently discussed analysis will be based on the mean field limit (36), we present a discretized version by introducing a particle based approximation based on (33). To this end let $(\hat\rho_j)_j$ iteratively be defined by

$$\hat\rho_{j+1} = \frac{1}{M}\sum_{m=1}^M \delta_{v_{t_{j+1}}^{(m),l_j}},$$

where δ_v denotes the dirac-measure at v and $v_{t_{j+1}}^{(m),l_j}$ initialized by $v_{t_j}^{(m),l_j} = v_{t_j}^{(m),l_{j-1}}$ solves

$$v_{t_{j+1}}^{(m),l_j} = v_{t_j}^{(m),l_j} - \int_{t_j}^{t_{j+1}} C(v_t^{l_j})\nabla\ell_R(v_t^{(m),l_j})\,\mathrm{d}t + \int_{t_j}^{t_{j+1}} \sqrt{2C(v_t^{l_j})}\,\mathrm{d}W_t^{(m)}, \quad m = 1,\dots,M. \tag{37}$$

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

Under convexity assumptions on ℓ_R^l and assuming that $C(\rho_t^l)$ 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 (Garbuno-Inigo et al., 2020, Proposition 2). These assumptions can for example be verified for linear forward models and Gaussian priors (Garbuno-Inigo et al., 2020, Corollary 5). We make the following additional assumption:

Assumption 13 (Approximation of the forward operator) *There exist $b > 0$ and $\alpha > 0$ such that for each $j \in \mathbb{N}$ there exists $F_{l_j} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ satisfying*

$$\int_{\mathbb{R}^{n_x}} \|F(x) - F_{l_j}(x)\|^2_{\mathbb{R}^{n_y}} \rho_j(x) \, \mathrm{d}x \leq b l_j^{-\alpha}.$$

For linear forward models and Gaussian priors, the pdf ρ_j remains Gaussian and Assumption 13 can be inferred from Assumption 10. Extending the convergence result of Garbuno-Inigo et al. (2020) to error decay (8) allows us to apply our results from Sec. 3. This leads to the following Proposition:

Proposition 14 (Multilevel interacting Langevin sampler) *Suppose $\mathrm{KL}(q_0 \| \rho_*) < \infty$ and let Assumption 13 be satisfied. Suppose there exist σ_1, $\sigma_2 > 0$ such that $\nabla^2 \ell_R^l > \sigma_1 \mathrm{Id}$ and $C(\rho_j) > \sigma_2 \mathrm{Id}$ (cp. (34)) for all $j \geq 0$. For $\tau > 0$ sufficiently small there exists $c \in (0,1)$ depending on σ_1, σ_2, such that $l_{K,j}(\varepsilon)$, $j = 0, \ldots, K(\varepsilon) - 1$ given by (21) yields*

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

for all small enough $\varepsilon > 0$ where ρ_j solves (36). Moreover, $\mathrm{cost}_{\mathrm{ML}}(\varepsilon) \simeq \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha}}$.

Proof For the proof see Appendix I.1. ■

An analogous results holds for the single-level choice in (16) 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 $C_0 = \mathrm{diag}(i^{-2}, i = 1, \ldots, n_x)$.

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 $\mathbb{E}[\frac{1}{2}\|\Gamma^{-1/2} F_{\mathrm{ref}}(x_K(\varepsilon) - x_*)\|^2_{\mathbb{R}^{n_y}} + \frac{\lambda}{2}\|C_0^{-1/2}(x_K(\varepsilon) - x_*)\|^2_{\mathbb{R}^{n_x}}]$ with $F_{\mathrm{ref}} = F_{2^{14}}$. For the interacting Langevin sampler we consider the convergence of the posterior mean, and the error quantity is $\mathbb{E}[\frac{1}{2}\|f(\cdot, x_{K(\varepsilon)}) - f(\cdot, x_*)\|^2_{L^2(D)}]$. 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.

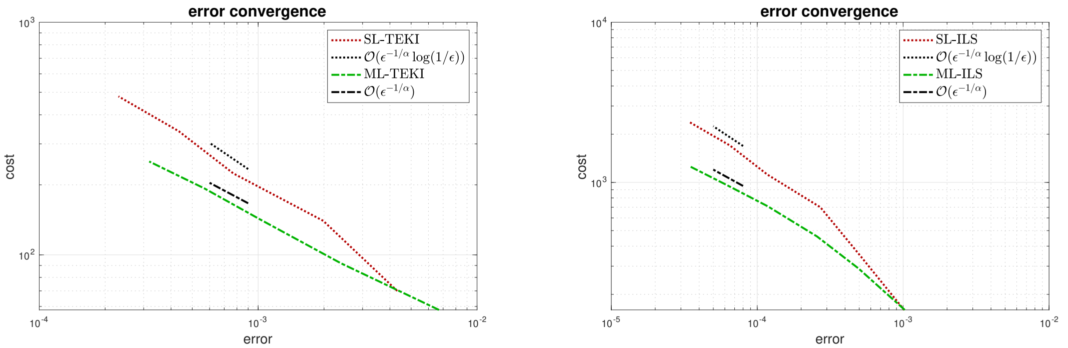

Figure 1: Computational costs vs. estimated error for TEKI (left) and the interacting Langevin sampler (ILS) (right). We ran all Algorithms 100 times to estimate the expected error.

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 14. Moreover, developing a framework for adaptive step sizes and levels would be an interesting extension.

Acknowledgments

This work was finished during the visit of Simon Weissmann and Jakob Zech at MIT. SW and JZ thank Youssef Marzouk and Ashia Wilson for the invitation to visit, for providing a fruitful research environment and for their kind hospitality. Moreover, SW and JZ thank Universität Heidelberg for funding the visit as part of the program “Mobilitätsmassnahmen im Rahmen internationaler Forschungskooperationen”.

References

Terrence Alsup, Luca Venturi, and Benjamin Peherstorfer. Multilevel Stein variational gradient descent with applications to Bayesian inverse problems. Preprint arXiv:2104.01945, 2021.

Jeffrey L. Anderson. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A*, 59(2):210–224, 2007. doi: 10.1111/j.1600-0870.2006.00216.x.

Jeffrey L. Anderson. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A*, 61(1):72–83, 2009. doi: 10.1111/j.1600-0870.2008.00361.x.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Yurii Ya. Belov. *Inverse Problems for Partial Differential Equations*. De Gruyter, 2012. ISBN 9783110944631. doi: doi:10.1515/9783110944631.

Martin Benning and Martin Burger. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111, 2018. doi: 10.1017/S0962492918000016.

Dirk Blömker, Claudia Schillings, Philipp Wacker, and Simon Weissmann. Continuous time limit of the stochastic ensemble Kalman inversion: Strong convergence analysis. Preprint arXiv:2107.14508, 2021.

Dirk Blömker, Claudia Schillings, and Philipp Wacker. A strongly convergent numerical scheme from ensemble Kalman inversion. *SIAM Journal on Numerical Analysis*, 56(4):2537–2562, 2018. doi: 10.1137/17M1132367.

Dirk Blömker, Claudia Schillings, Philipp Wacker, and Simon Weissmann. Well posedness and convergence analysis of the ensemble Kalman inversion. *Inverse Problems*, 35(8):085007, 2019. doi: 10.1088/1361-6420/ab149c.

Neil K. Chada and Xin T. Tong. Convergence acceleration of ensemble Kalman inversion in nonlinear settings. *Math. Comp.*, 91(335):1247–1280, 2022. doi: 10.1090/mcom/3709.

Neil K. Chada, Andrew M. Stuart, and Xin T. Tong. Tikhonov regularization within ensemble Kalman inversion. *SIAM Journal on Numerical Analysis*, 58(2):1263–1294, 2020. doi: 10.1137/19M1242331.

Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. *An Introduction to Total Variation for Image Analysis*. De Gruyter, Berlin, Boston, 2010. ISBN 9783110226157. doi: 10.1515/9783110226157.263.

Josef Dick, Robert N. Gantner, Quoc T. Le Gia, and Christoph Schwab. Multilevel higher-order quasi-Monte Carlo Bayesian estimation. *Math. Models Methods Appl. Sci.*, 27(5):953–995, 2017. ISSN 0218-2025. doi: 10.1142/S021820251750021X.

Zhiyan Ding and Qin Li. Ensemble Kalman inversion: mean-field limit and convergence analysis. *Statistics and Computing*, 31(1):9, 2021a. doi: 10.1007/s11222-020-09976-0.

Zhiyan Ding and Qin Li. Ensemble Kalman sampler: mean-field limit and convergence analysis. *SIAM Journal on Mathematical Analysis*, 53(2):1546–1578, 2021b. doi: 10.1137/20M1339507.

Tim J. Dodwell, Christian Ketelsen, Robert Scheichl, and Aretha L. 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. doi: 10.1137/130915005.

Heinz W. Engl, Karl Kunisch, and Andreas Neubauer. Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. *Inverse Problems*, 5(4):523–540, aug 1989. doi: 10.1088/0266-5611/5/4/007.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer Netherlands, 1996. ISBN 9780792341574. doi: 10.1007/978-94-009-1740-8.

Alexandre Ern and Jean-Luc Guermond. *Finite elements II—Galerkin approximation, elliptic and mixed PDEs*, volume 73 of *Texts in Applied Mathematics*. Springer, Cham, 2021. ISBN 978-3-030-56922-8; 978-3-030-56923-5. doi: 10.1007/978-3-030-56923-5.

Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, and Andrew M. Stuart. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020. doi: 10.1137/19M1251655.

Michael B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008. ISSN 0030-364X. doi: 10.1287/opre.1070.0496.

Michael B. Giles and Benjamin 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. doi: 10.1515/9783110213140.165.

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

Abdul-Lateef Haji-Ali, Fabio Nobile, Lorenzo Tamellini, and Raúl Tempone. Multi-index stochastic collocation for random PDEs. *Comput. Methods Appl. Mech. Engrg.*, 306:95–122, 2016. ISSN 0045-7825. doi: 10.1016/j.cma.2016.03.029.

Stefan Heinrich. *Multilevel Monte Carlo Methods*, pages 58–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-540-45346-8. doi: 10.1007/3-540-45346-6_5.

Michael Hinze, Rene Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE Constraints*. Mathematical Modelling: Theory and Applications. Springer Netherlands, 2008. ISBN 9781402088391. doi: 10.1007/978-1-4020-8839-1.

Marco A. Iglesias, Kody J.H. Law, and Andrew M. Stuart. Ensemble Kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, mar 2013. doi: 10.1088/0266-5611/29/4/045001.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Nikola B. Kovachki and Andrew M. Stuart. Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, 35(9):095005, 2019. doi: 10.1088/1361-6420/ab1c3a.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951. doi: 10.1214/aoms/1177729694.

Matthieu Martin and Fabio Nobile. PDE-constrained optimal control problems with uncertain parameters using SAGA. *SIAM/ASA J. Uncertain. Quantif.*, 9(3):979–1012, 2021. doi: 10.1137/18M1224076.

Matthieu Martin, Sebastian Krumscheid, and Fabio 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. ISSN 0764-583X. doi: 10.1051/m2an/2021025.

Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

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

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Texts in Applied Mathematics. Springer New York, 2014. ISBN 9781493913220. doi: 10.1007/978-1-4939-1323-7.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259 – 268, 1992. ISSN 0167-2789. doi: 10.1016/0167-2789(92)90242-F.

Stefan A. Sauter and Christoph Schwab. *Boundary element methods*, volume 39 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2011. ISBN 978-3-540-68092-5. doi: 10.1007/978-3-540-68093-2. Translated and expanded from the 2004 German original.

Claudia Schillings and Andrew M. Stuart. Analysis of the ensemble Kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, 55(3):1264–1290, 2017. doi: 10.1137/16M105959X.

Andrew M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.

Xin T. Tong, Andrew J. Majda, and David Kelly. Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation. *Communications in Mathematical Sciences*, 14(5):1283–1313, 2016. ISSN 1539-6746. doi: 10.4310/CMS.2016.v14.n5.a5.

Stephen Tu, Shivaram Venkataraman, Ashia C Wilson, Alex Gittens, Michael I Jordan, and Benjamin Recht. Breaking locality accelerates block gauss-seidel. In *International Conference on Machine Learning*, pages 3482–3491. PMLR, 2017.

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

Simon Weissmann, Neil K. Chada, Claudia Schillings, and Xin T. Tong. Adaptive Tikhonov strategies for stochastic ensemble Kalman inversion. *Inverse Problems*, 38(4):045009, 2022. doi: 10.1088/1361-6420/ac5729.

Ashia C. Wilson, Ben Recht, and Michael I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

Jakob Zech, Dinh Dũng, and Christoph Schwab. Multilevel approximation of parametric and stochastic PDEs. *Math. Models Methods Appl. Sci.*, 29(9):1753–1817, 2019. ISSN 0218-2025. doi: 10.1142/S0218202519500349.

Appendix A. Proofs of the main theorems

A.1. Proof of Theorem 5

Proof Let $\varepsilon \in (0, e_0)$. With $\tilde{e}_K$ in (13), denote

$$\tilde{e}_K((l_j)_j) = c^K e_0 + \sum_{j=0}^{K-1} c^{K-1-j} l_j^{-\alpha} =: \tilde{e}_{K,1} + \tilde{e}_{K,2}((l_j)_j). \tag{38}$$

By definition, $K(\varepsilon) \in \mathbb{N}$ in (16) is minimal such that $\tilde{e}_{K(\varepsilon),1} \leq \frac{\varepsilon}{2}$. Moreover, by definition $\bar{l}_K(\varepsilon) > 0$ in (16) is minimal such that $\tilde{e}_{K,2}((\bar{l}_K(\varepsilon))_j) \leq \varepsilon$.

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

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

and by the optimality property of $K(\varepsilon)$ we obtain $\hat{K} \geq K(2\varepsilon)$. Similarly, since

$$\tilde{e}_{\hat{K},2}((\hat{l})_j) \leq \varepsilon, \tag{39}$$

the optimality property of $\bar{l}_K(\varepsilon)$ implies $\hat{l} \geq \bar{l}_K(2\varepsilon)$. Taking the infimum over all such $\hat{l}$ and $\hat{K}$

$$K(2\varepsilon)\bar{l}_K(2\varepsilon) \leq \inf\{\hat{K}\hat{l} \,:\, \tilde{e}_K(\hat{l}) \leq \varepsilon\} \leq K(\varepsilon)\bar{l}_K(\varepsilon), \tag{40}$$

where the second inequality holds due to $\tilde{e}_{K(\varepsilon)}(\bar{l}_K(\varepsilon)) \leq \varepsilon$.

By definition of $K(\varepsilon)$ and $\bar{l}_K(\varepsilon)$ in (16)

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

Since with $\delta := c^{\frac{1}{1+\alpha}} \in (0,1)$ holds $1 \leq \sum_{i=0}^{K-1} c^{\frac{K-1-i}{1+\alpha}} \leq (1-c)^{-1}$ we have $\mathrm{cost}_{\mathrm{SL}}(\varepsilon) \simeq \varepsilon^{-\frac{1}{\alpha}} \log(\varepsilon^{-1})$ as $\varepsilon \to 0$. Together with (40) we find

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

This shows (17) and quasi-optimality in the sense of Def. 4. ■

A.2. Proof of Lemma 6

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 λ leads to the following first order optimality conditions

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

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,

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

This shows (19). ■

Remark 15 *The constant $C_{K,\varepsilon}$ in (19) 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*

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

A.3. Proof of Theorem 8

Proof With $l_{K,j}(\varepsilon)$ and $K(\varepsilon)$ as in (21), the calculation in (20) and Rmk. 15 show

$$\mathrm{cost}_{\mathrm{ML}}(\varepsilon) = \sum_{j=0}^{K(\varepsilon)-1} l_{K,j} = \left(\frac{\varepsilon}{2}\right)^{-\frac{1}{\alpha}} \left(\frac{1 - c^{K(\varepsilon)/(1+\alpha)}}{1 - c^{1/(1+\alpha)}}\right)^{\frac{1+\alpha}{\alpha}} \simeq \varepsilon^{-\frac{1}{\alpha}} \qquad \text{as} \quad \varepsilon \to 0, \tag{41}$$

as claimed.

The proof of quasi-optimality in the sense of Def. 7 follows the same argument as in the proof of Theorem 5: Let $\varepsilon \in (0, e_0)$. Split $\tilde{e}_K((l_j)_j)$ in (13), in the terms $\tilde{e}_{K,1}$ and $\tilde{e}_{K,2}((l_j)_j)$ as in (38).

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

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

which by Lemma 6 minimizes $\sum_{j=0}^{\hat{K}-1} \tilde{l}_{K,j}$ under the constraint $\tilde{e}_{\hat{K},2}((\tilde{l}_{K,j})_j) \leq \varepsilon$. Since also $\tilde{e}_{\hat{K},2}((\hat{l}_j)_j) \leq \tilde{e}_{\hat{K}}((\hat{l}_{K,j})_j) \leq \varepsilon$, this implies

$$\sum_{j=0}^{\hat{K}-1} \hat{l}_j \geq \sum_{j=0}^{\hat{K}-1} \tilde{l}_{K,j} = \varepsilon^{-\frac{1}{\alpha}} \left(\frac{1 - c^{\hat{K}/(1+\alpha)}}{1 - c^{1/(1+\alpha)}}\right)^{\frac{1+\alpha}{\alpha}} \tag{42}$$

where the equality holds by the calculation in (20).

Next observe that by definition, $K(2\varepsilon) \in \mathbb{N}$ in (21) is minimal such that $\tilde{e}_{K(\varepsilon),1} \leq \varepsilon$. Since also $\tilde{e}_{\hat{K},1} \leq \tilde{e}_{\hat{K}}((\hat{l}_j)_j) \leq \varepsilon$, this gives $K(2\varepsilon) \leq \hat{K}$. Therefore

$$\varepsilon^{-\frac{1}{\alpha}} \left(\frac{1 - c^{\hat{K}/(1+\alpha)}}{1 - c^{1/(1+\alpha)}} \right)^{\frac{1+\alpha}{\alpha}} \geq \varepsilon^{-\frac{1}{\alpha}} \left(\frac{1 - c^{K(2\varepsilon)/(1+\alpha)}}{1 - c^{1/(1+\alpha)}} \right)^{\frac{1+\alpha}{\alpha}} = \sum_{j=0}^{K(2\varepsilon)} l_{K,j}(2\varepsilon), \tag{43}$$

where the last inequality holds by the calculation in (20). In all, (42) and (43) yield

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

Since $\hat{K}$ and $\hat{l}_j$ were arbitrary such that $\tilde{e}((\hat{l}_j)_j) \leq \varepsilon$,

$$\sum_{j=0}^{K(2\varepsilon)-1} l_{K,j}(2\varepsilon) \leq \inf \left\{ \sum_{j=0}^{\hat{K}-1} \hat{l}_j \, : \, \tilde{e}_{\hat{K}}((\hat{l}_j)_j) \leq \varepsilon, \; \hat{K} \in \mathbb{N}, \; \hat{l}_j > 0 \, \forall j \right\} \leq \sum_{j=0}^{K(\varepsilon)-1} l_{K,j}(\varepsilon), \tag{44}$$

where the second inequality holds due to $\tilde{e}_{K(\varepsilon)}((l_{K,j}(\varepsilon))) \leq \varepsilon$. By (41) all terms in (44) behave like $\varepsilon^{-1/\alpha}$ as $\varepsilon \to 0$, and this shows quasi-optimality in the sense of Def. 7. ■

Appendix 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, μ-strongly convex, i.e.

$$\Phi(y) \geq \Phi(x) + \langle \nabla\Phi(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_{\mathcal{X}}^2 \qquad \forall x, y \in \mathcal{X}, \tag{45}$$

and satisfies L-smoothness so that

$$\Phi(y) \leq \Phi(x) + \langle \nabla\Phi(x), y - x \rangle + \frac{L}{2} \|y - x\|_{\mathcal{X}}^2 \qquad \forall x, y \in \mathcal{X}. \tag{46}$$

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

For the deterministic algorithms we assume the existence of approximate gradients $g_l : \mathcal{X} \to \mathcal{X}$ such that

$$\|\nabla\Phi(x) - g_l(x)\|_{\mathcal{X}} \leq \frac{l^{-\alpha}}{\eta} \qquad \forall x \in \mathcal{X}, \tag{47}$$

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

$$\mathbb{E}\|\nabla\Phi(x) - G_l(x)\|_{\mathcal{X}} \leq \frac{l^{-\alpha}}{\eta} \qquad \forall x \in \mathcal{X}. \tag{48}$$

B.1. Gradient Descent (GD)

The gradient descent update with step-size η and approximate gradient at level l_k reads

$$x_{k+1} = x_k - \eta g_{l_k}(x_k). \tag{49}$$

Proposition 16 *Let $0 \leq \eta \leq 1/L$ and let g_l satisfy (47). Then x_k generated by (49) satisfies the bound (8) with $e_k = \|x_{k+1} - x_*\|_{\mathcal{X}}$ and $c = \sqrt{1-\eta\mu}$.*

Proof Observe that

$$\begin{aligned}\|x_{k+1} - x_*\|_{\mathcal{X}} &= \|x_k - x_* - \eta\nabla_x\Phi(x_k) + \eta\nabla_x\Phi(x_k) - \eta g_{l_k}(x_k)\|_{\mathcal{X}} \\ &\leq \|x_k - x_* - \eta\nabla_x\Phi(x_k)\|_{\mathcal{X}} + \eta\|\nabla_x\Phi(x_k) - g_{l_k}(x_k)\|_{\mathcal{X}}.\end{aligned}$$

Using μ-strong convexity (45) and L-smoothness (46) we obtain the following upper bound,

$$\begin{aligned}\|x_k - x_* - \eta\nabla_x\Phi(x_k)\|_{\mathcal{X}}^2 &= \|x_k - x_*\|_{\mathcal{X}}^2 - 2\eta\langle x_k - x_*, \nabla_x\Phi(x_k)\rangle + \eta^2\|\nabla_x\Phi(x_k)\|_{\mathcal{X}}^2 \\ &\leq \|x_k - x_*\|_{\mathcal{X}}^2 - \eta\mu\|x_k - x_*\|_{\mathcal{X}}^2 - 2\eta(\Phi(x_k) - \Phi(x_*)) + \eta^2\|\nabla_x\Phi(x_k)\|_{\mathcal{X}}^2 \\ &\leq (1-\eta\mu)\|x_k - x_*\|_{\mathcal{X}}^2 - \eta(1/L - \eta)\|\nabla_x\Phi(x_k)\|_{\mathcal{X}}^2 \\ &\leq (1-\eta\mu)\|x_k - x_*\|_{\mathcal{X}}^2.\end{aligned}$$

Combining these inequalities with assumption (47) leads to desired recursion (8)

$$e_{k+1} = \|x_{k+1} - x_*\|_{\mathcal{X}} \leq \sqrt{1-\eta\mu}\|x_k - x_*\|_{\mathcal{X}} + l_k^{-\alpha} = ce_k + l_k^{-\alpha}.$$

■

It now follows from Thm. 8 that:

Corollary 17 (MLGD) *Consider the setting of Proposition 16. Then with the levels $l_{K,j}(\varepsilon)$ as in (21), x_K generated by (49) satisfies $e_K := \|x_K - x_*\|_{\mathcal{X}} \leq \varepsilon$, and it holds $\sum_{j=0}^{K-1} l_{K,j}(\varepsilon) = O(\varepsilon^{-1/\alpha})$ 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 (48). The stochastic gradient descent update with step-size η and approximate stochastic gradient at level l_k then reads

$$x_{k+1} = x_k - \eta G_{l_k}(x_k). \tag{50}$$

Proposition 18 *Let $0 \leq \eta \leq 1/L$ and let G_l satisfy (48). Then x_k generated by (50) satisfies the bound (8) with $e_k = \mathbb{E}[\|x_{k+1} - x_*\|_{\mathcal{X}}]$ and $c = \sqrt{1-\eta\mu}$.*

Proof The proof proceeds in the exact same manner as for gradient descent (taking expectations and replacing $g_{l_k}(x_k)$ with $G_{l_k}(x_k)$). To be more precise, taking the expectation w.r.t. the filtration up to iteration k, i.e. $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \sigma(G_{l_j}(x_j), j = 1, \dots, k-1)]$, we have

$$\mathbb{E}_k[\|x_{k+1} - x_*\|_{\mathcal{X}}] \leq \sqrt{1-\eta\mu}\|x_k - x_*\|_{\mathcal{X}} + \eta \mathbb{E}_k[\|\nabla_x \Phi(x_k) - G_{l_k}(x_k))\|_{\mathcal{X}}],$$

where we again applied the inequality $\|x_k - x_* - \eta\nabla_x\Phi(x_k)\|_{\mathcal{X}}^2 \leq (1-\eta\mu)\|x_k - x_*\|_{\mathcal{X}}^2$ under μ-strong convexity and L-smoothness. Taking the expectation (this time without conditioning), we obtain

$$e_{k+1} = \mathbb{E}[\|x_{k+1} - x_*\|_{\mathcal{X}}] \leq \sqrt{1-\eta\mu}\mathbb{E}[\|x_k - x_*\|_{\mathcal{X}}] + \eta\mathbb{E}[\|\nabla_x\Phi(x_k) - G_{l_k}(x_k))\|_{\mathcal{X}}] \leq ce_k + l_k^{-\alpha}.$$

■

Corollary 19 (MLSGD) *Consider the setting of Proposition 18. Then with the levels $l_{K,j}(\varepsilon)$ as in (21), x_K generated by (50) satisfies $e_K := \mathbb{E}[\|x_K - x_*\|_{\mathcal{X}}] \leq \varepsilon$, and it holds $\sum_{j=0}^{K-1} l_{K,j}(\varepsilon) = O(\varepsilon^{-1/\alpha})$ as $\varepsilon \to 0$.*

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

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

$$\min_{x\in\mathcal{X}} \Phi(x), \quad \Phi(x) := \mathbb{E}_\xi[\varphi(x,\xi)], \tag{51}$$

where ξ 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 $\{\xi_k^n\}_{n=1}^{l_k}$ of ξ, we can define a stochastic gradient approximation by the Monte Carlo estimator $G_{l_k}(x_k) := \frac{1}{l_k}\sum_{n=1}^{l_k} \nabla_x\varphi(x_k, \xi_k^n)$. The level l_k describes the batch size in iteration k, which correspond to the number of required evaluations of φ. In this sense, l_k describes the computational cost required to evaluate G_{l_k}. Under certain integrability assumptions, G_{l_k} is an unbiased estimator of $\nabla\Phi(x_k)$, i.e. $\nabla\Phi(x) = \mathbb{E}_\xi[\nabla\varphi(x,\xi)]$, and it holds

$$\mathbb{E}[\|\nabla\Phi(x) - G_l(x)\|_{\mathcal{X}}] \leq \sqrt{\mathbb{E}[\|\nabla\Phi(x) - G_l(x)\|_{\mathcal{X}}^2]} \leq \left(\frac{\mathbb{E}_\xi[\|\nabla\varphi(x,\xi) - \mathbb{E}_\xi[\nabla\varphi(x,\xi)]\|_{\mathcal{X}}^2]}{l}\right)^{1/2},$$

i.e. we have the second inequality in (10) (up to a constant, cp. Rmk. 9) with $\alpha = 1/2$. Applying the multilevel strategy (21) then yields a stochastic optimization method with increasing batchsize.

By Corollary 19, achieving error $\mathbb{E}[\|x_K - x_*\|_{\mathcal{X}}] \leq \varepsilon$ 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 $\eta_k \sim \frac{1}{k}$. 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 Nesterov (1983))

$$\begin{aligned} p_{k+1} &= q_k - \alpha\nabla\Phi(q_k) \\ q_{k+1} &= p_{k+1} + \beta(p_{k+1} - p_k), \end{aligned}$$

where $\alpha = 1/L$ and $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. 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. Nesterov (1983) for equivalence):

$$x_k = \frac{\tau}{1+\tau} z_k + \frac{1}{1+\tau} y_k \tag{52a}$$

$$y_{k+1} = x_k - \frac{1}{L} g_{l_k}(x_k) \tag{52b}$$

$$z_{k+1} = z_k + \tau(x_k - z_k) - \frac{\tau}{\mu} g_{l_k}(x_k). \tag{52c}$$

Proposition 20 *Suppose that g_l satisfies (47) with $\eta = 1/\sqrt{L}$ and $\tau = \sqrt{\mu/L}$. Then (y_k, z_k) generated by (52) satisfies (8) with $e_k = \Phi(y_k) - \Phi(x_*) + \frac{\mu}{2}\|z_k - x_*\|_{\mathcal{X}}^2$ exponent 2α, and $c = 1-\tau$, more precisely*

$$\Phi(y_{k+1}) - \Phi(x_*) + \frac{\mu}{2}\|z_{k+1} - x_*\|_{\mathcal{X}}^2 \le (1-\tau)\Big(\Phi(y_k) - \Phi(x_*) + \frac{\mu}{2}\|z_k - x_*\|_{\mathcal{X}}^2\Big) + l_k^{-2\alpha}.$$

Proof We begin by using the L-smoothness of Φ:

$$\begin{aligned} \Phi(y_{k+1}) - \Phi(x_k) &\le \langle\nabla\Phi(x_k), y_{k+1} - x_k\rangle + \frac{L}{2}\|x_k - y_{k+1}\|_{\mathcal{X}}^2 \\ &= -\frac{1}{L}\langle\nabla\Phi(x_k), g_l(x_k)\rangle + \frac{1}{2L}\|g_l(x_k)\|_{\mathcal{X}}^2 \\ &= \frac{1}{2L}\|g_l(x_k) - \nabla\Phi(x_k)\|_{\mathcal{X}}^2 - \frac{1}{2L}\|\nabla\Phi(x_k)\|_{\mathcal{X}}^2 \le \frac{l_k^{-2\alpha}}{2L\eta^2} - \frac{1}{2L}\|\nabla\Phi(x_k)\|_{\mathcal{X}}^2. \end{aligned} \tag{53}$$

Denote $\tilde{z}_{k+1} := z_{k+1} - \frac{\tau}{\mu}(\nabla\Phi(x_k) - \nabla g_{l_k}(x_k))$. The triangle inequality results in the following upper bound

$$\frac{\sqrt{\mu}}{2}\|z_{k+1} - x_*\|_{\mathcal{X}} \le \frac{1}{2\sqrt{L}}\|\nabla\Phi(x_k) - g_l(x_k)\|_{\mathcal{X}} + \frac{\sqrt{\mu}}{2}\|\tilde{z}_{k+1} - x_*\|_{\mathcal{X}} \le \frac{l_k^{-\alpha}}{2\sqrt{L}\eta} + \frac{\sqrt{\mu}}{2}\|\tilde{z}_{k+1} - x_*\|_{\mathcal{X}}.$$

and using Jensen's inequality ($(a/2 + b/2)^2 \le a^2/2 + b^2/2$) we obtain the subsequent identity,

$$\frac{\sqrt{\mu}}{2}\|z_{k+1} - x_*\|_{\mathcal{X}}^2 \le \frac{l_k^{-2\alpha}}{2L\eta^2} + \frac{\mu}{2}\|\tilde{z}_{k+1} - x_*\|_{\mathcal{X}}^2.$$

The rest of the Lyapunov analysis follows as normal (see e.g. Tu et al. (2017) and Wilson et al. (2021)). We use the previous inequality and strong convexity to obtain the bound:

$$\begin{aligned}
&\frac{\mu}{2}\|z_{k+1}-x_*\|^2_{\mathcal{X}} - \frac{\mu}{2}\|z_k - x_*\|^2_{\mathcal{X}} \leq \frac{\mu}{2}\|\tilde{z}_{k+1}-x_*\|^2_{\mathcal{X}} - \frac{\mu}{2}\|z_k - x_*\|^2_{\mathcal{X}} + \frac{l_k^{-2\alpha}}{2L\eta^2}\\
&= \tau\langle\nabla\Phi(x_k), x_* - z_k\rangle - \tau\mu\langle x_k - z_k, x_* - z_k\rangle + \frac{\mu}{2}\|\tilde{z}_{k+1} - z_k\|^2_{\mathcal{X}} + \frac{l_k^{-2\alpha}}{2L\eta^2}\\
&\leq \tau\langle\nabla\Phi(x_k), x_k - z_k\rangle - \tau\mu\langle x_k - z_k, x_* - z_k\rangle + \frac{\mu}{2}\|\tilde{z}_{k+1} - z_k\|^2_{\mathcal{X}}\\
&\quad - \tau(\Phi(x_k) - \Phi(x_*) + \frac{\mu}{2}\|x_k - x_*\|^2_{\mathcal{X}}) + \frac{l_k^{-2\alpha}}{2L\eta^2}\\
&= \langle\nabla\Phi(x_k), x_k - y_k\rangle - \frac{\tau\mu}{2}\|x_k - z_k\|^2_{\mathcal{X}} + \frac{\mu}{2}\|\tilde{z}_{k+1} - z_k\|^2_{\mathcal{X}} + \frac{l_k^{-2\alpha}}{2L\eta^2}\\
&\quad - \tau(\Phi(x_k) - \Phi(x_*) + \frac{\mu}{2}\|z_k - x_*\|^2_{\mathcal{X}}).
\end{aligned}$$

The inequality uses the strong convexity of Φ. Using the strong convexity of Φ again and the descent bound (53) we have:

$$\begin{aligned}
e_{k+1} &\leq (1-\tau)e_k + \Phi(x_k) - \Phi(y_k) + \langle\nabla\Phi(x_k), x_k - y_k\rangle - \frac{1}{2L}\|\nabla\Phi(x_k)\|^2_{\mathcal{X}} + \frac{l_k^{-2\alpha}}{L\eta^2}\\
&\quad - \tau(\Phi(x_k) - \Phi(y_k)) - \frac{\tau\mu}{2}\|x_k - z_k\|^2_{\mathcal{X}} + \frac{\mu}{2}\|\tilde{z}_{k+1} - z_k\|^2_{\mathcal{X}}\\
&\leq ce_k - \frac{\mu}{2}\|x_k - y_k\|^2_{\mathcal{X}} - \frac{1}{2L}\|\nabla\Phi(x_k)\|^2_{\mathcal{X}} - \tau\langle\nabla\Phi(x_k), x_k - y_k\rangle + \frac{\tau L}{2}\|x_k - y_k\|^2_{\mathcal{X}}\\
&\quad - \frac{\tau\mu}{2}\|x_k - z_k\|^2_{\mathcal{X}} + \frac{\mu}{2}\|\tilde{z}_{k+1} - z_k\|^2_{\mathcal{X}} + \frac{l_k^{-2\alpha}}{L\eta^2}\\
&= ce_k + (\frac{\tau^2}{\mu} - \frac{1}{2L})\|\nabla\Phi(x_k)\|^2_{\mathcal{X}} + (\frac{\tau L}{2} - \frac{\mu}{2\tau})\|x_k - y_k\|^2_{\mathcal{X}} + l_k^{-2\alpha}\\
&= ce_k + l_k^{-2\alpha}.
\end{aligned}$$

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

Corollary 21 (MLAGD) *Consider the setting of Proposition 20. Then with the levels $l_{K,j}(\varepsilon)$ as in (21), (y_K, z_K) generated by (52) satisfies $e_K := \Phi(y_K) - \Phi(x_*) + \frac{\mu}{2}\|z_K - x_*\|^2_{\mathcal{X}} \leq \varepsilon$, and it holds $\sum_{j=0}^{K-1} l_{K,j}(\varepsilon) = O(\varepsilon^{-1/(2\alpha)})$ as $\varepsilon \to 0$.*

Note that the cost $\sum_{j=0}^{K-1} l_{K,j}(\varepsilon)$ for AGD in Corollary 21 increases at half the rate compared to GD in Corollary 17. 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 (48). In the following let

$$x_k = \frac{\tau}{1+\tau} z_k + \frac{1}{1+\tau} y_k \tag{54a}$$

$$y_{k+1} = x_k - \frac{1}{L} G_{l_k}(x_k) \tag{54b}$$

$$z_{k+1} = z_k + \tau(x_k - z_k) - \frac{\tau}{\mu} G_{l_k}(x_k). \tag{54c}$$

Proposition 22 *Suppose that G_l satisfies (48) with $\eta = 1/\sqrt{L}$ where $\tau = \sqrt{\mu/L}$. Then (y_k, z_k) generated by (54) satisfies (8) with $e_k = \mathbb{E}[\Phi(y_k)] - \Phi(x_*) + \frac{\mu}{2}\mathbb{E}[\|z_k - x_*\|_{\mathcal{X}}^2]$, exponent 2α and $c = 1-\tau$, more precisely*

$$\mathbb{E}[\Phi(y_{k+1})] - \Phi(x_*) + \frac{\mu}{2}\mathbb{E}[\|z_{k+1} - x_*\|_{\mathcal{X}}^2] \le (1-\tau)\Big(\mathbb{E}[\Phi(y_k)] - \Phi(x_*) + \frac{\mu}{2}\mathbb{E}[\|z_k - x_*\|_{\mathcal{X}}^2]\Big) + l_k^{-2\alpha}.$$

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

$$\begin{aligned}\mathbb{E}_k[\Phi(y_{k+1})] - \Phi(x_k) &\le \frac{1}{2L}\mathbb{E}_k[\|G_l(x_k) - \nabla\Phi(x_k)\|_{\mathcal{X}}^2 - \|\nabla\Phi(x_k)\|_{\mathcal{X}}^2] \\ &\le \frac{l_k^{-2\alpha}}{2L\eta^2} - \frac{1}{2L}\mathbb{E}_k\|\nabla\Phi(x_k)\|_{\mathcal{X}}^2,\end{aligned}$$

and

$$\frac{\sqrt{\mu}}{2}\mathbb{E}_k\|z_{k+1} - x_*\|_{\mathcal{X}} \le \frac{1}{2\sqrt{L}}\mathbb{E}_k\|\nabla\Phi(x_k) - G_l(x_k)\|_{\mathcal{X}} + \frac{\sqrt{\mu}}{2}\mathbb{E}_k\|\tilde{z}_{k+1} - x_*\|_{\mathcal{X}}.$$

Therefore,

$$\frac{\sqrt{\mu}}{2}\mathbb{E}_k\|z_{k+1} - x_*\|_{\mathcal{X}}^2 \le \frac{l_k^{-2\alpha}}{2L\eta^2} + \frac{\mu}{2}\mathbb{E}_k\|\tilde{z}_{k+1} - x_*\|_{\mathcal{X}}^2.$$

Given the remainder of the proof relies on the strong convexity and smoothness of the function and update (54a) we obtain the following recursion following the same line of argumentation:

$$e_{k+1} \le c e_k + l_k^{-2\alpha}.$$

■

Corollary 23 (MLASGD) *Consider the setting of Proposition 22. Then with the levels $l_{K,j}(\varepsilon)$ as in (21), (y_K, z_K) generated by (54) satisfies $e_K := \mathbb{E}[\Phi(y_k)] - \Phi(x_*) + \frac{\mu}{2}\mathbb{E}[\|z_k - x_*\|_{\mathcal{X}}^2] \le \varepsilon$, and it holds $\sum_{j=0}^{K-1} l_{K,j}(\varepsilon) = O(\varepsilon^{-1/(2\alpha)})$ as $\varepsilon \to 0$.*

Remark 24 *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 $\alpha = 1/2$.*

Appendix 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 $u_f \in H_0^1(D) \subseteq L^2(D)$ the unique weak solution to (5) with right-hand side f, i.e.

$$\int_D \nabla u_f^\top(s)\nabla v(x) + u_f(s)v(s)\mathrm{d}s = \int_D f(s)v(s)\mathrm{d}s \qquad \forall v \in H_0^1(D). \tag{55}$$

C.1. Well-definedness and regularity of u_f

Existence and well-definedness of u_f is classical. Moreover, using the test function $v = u_f$ in (55) one has the apriori estimate $\|u_f\|_{H_0^1(D)} \le \|f\|_{L^2(D)}$ and in particular $\|u_f\|_{L^2(D)} \le \|f\|_{L^2(D)}$. We refer for example to (Ern and Guermond, 2021, §25) for more details.

Furthermore convexity of D in fact implies $H^2(D)$ regularity of u_f and the existence of $C > 0$ such that $\|u_f\|_{H^2(D)} \le C\|f\|_{L^2(D)}$, see (Ern and Guermond, 2021, Theorem 31.30) and the references there.

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

Consider the objective

$$\Phi(f) = \frac{1}{2}\|\Gamma^{-1/2}(F(f) - y)\|^2_{\mathbb{R}^{n_y}} + \frac{\lambda}{2}\|f\|^2_{L^2(D)} =: \ell(f,y) + R(f). \tag{56}$$

in Example 3. 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

$$\mathcal{S} := \begin{cases} \mathbb{R}^{n_y} \to \mathbb{R} \\ w \mapsto \frac{1}{2}\|\Gamma^{-1/2}(w-y)\|^2_{\mathbb{R}^{n_y}}, \end{cases} \qquad \mathcal{O} := \begin{cases} L^2(D) \to \mathbb{R}^{n_y} \\ f \mapsto (\int_D \xi_j f)_{j=1}^{n_y}, \end{cases}$$

and

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

Observe that $\ell(f,y) = \mathcal{S}(\mathcal{O}(\mathcal{A}(f)))$. Using that $\mathcal{A} : L^2(D) \to H_0^1(D)$ 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)

$$\nabla_f \ell(f,y) = \nabla(\mathcal{S} \circ \mathcal{O} \circ \mathcal{A})(f) = \mathcal{A}^*(\nabla(\mathcal{S} \circ \mathcal{O})(\mathcal{A}(f))) = \mathcal{A}(\nabla(\mathcal{S} \circ \mathcal{O})(u_f)) \in L^2(D), \tag{57}$$

where $u_f = \mathcal{A}(f) \in H_0^1(D) \subseteq L^2(D)$. We next compute $\nabla(\mathcal{S} \circ \mathcal{O})$. Denoting by $D\mathcal{S}(w)$ the Fréchet derivative, it holds

$$D\mathcal{S}(w)(v) = (w-y)^\top \Gamma^{-1} v \in \mathbb{R} \qquad \forall v \in \mathbb{R}^{n_y}$$

and with $\xi = (\xi_j)_{j=1}^{n_y} \in L^2(D, \mathbb{R}^{n_y})$ since $\mathcal{O}$ is bounded linear,

$$D\mathcal{O}(f)(g) = \mathcal{O}(g) = \int_D g(s)\xi(s)\mathrm{d}s \in \mathbb{R}^{n_y} \qquad \forall g \in L^2(D).$$

Hence for the composition

$$D(\mathcal{S}\circ\mathcal{O})(f)(g) = D\mathcal{S}(\mathcal{O}(f))(D\mathcal{O}(f)(g)) = \int_D (\mathcal{O}(f)-y)^\top \Gamma^{-1}\xi(s)g(s)\mathrm{d}s \qquad \forall g\in L^2(D).$$

This shows

$$\nabla(\mathcal{S}\circ\mathcal{O})(f) = (\mathcal{O}(f)-y)^\top\Gamma^{-1}\xi(\cdot)\in L^2(D)$$

and finally by (57), $\nabla_f \ell(f,y) = \mathcal{A}(h) = u_h$ with $h(\cdot) = (\mathcal{O}(u_f)-y)^\top\Gamma^{-1}\xi(\cdot)$.

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

Next we argue that u_h can be approximated with the rate claimed in Example 3. According to, e.g., (Ern and Guermond, 2021, §26.3.3., §32.3.2), given $f\in L^2(D)$, the FEM approximation u^l_f to $u(f) = \mathcal{A}(f)$ on a uniform simplicial mesh on $D\subseteq\mathbb{R}^2$ with $O(l)$ elements will satisfy $\|u_f - u^l_f\|_{L^2(D)} \lesssim l^{-1}$. Now set $\tilde h(\cdot) := (\mathcal{O}(u^l_f)-y)^\top\Gamma^{-1}\xi(\cdot)\in L^2(D)$. Then, since $\mathcal{O}: L^2(D)\to\mathbb{R}^{n_y}$ is bounded linear and $\xi\in L^2(D,\mathbb{R}^{n_y})$, we have $\|h-\tilde h\|_{L^2(D)}\lesssim l^{-1}$. Using that $\mathcal{A}: L^2(D)\to L^2(D)$ is bounded linear, $\|u_h - u_{\tilde h}\|_{L^2(D)} = \|\mathcal{A}(h-\tilde h)\|_{L^2(D)}\lesssim l^{-1}$. Finally, $\|u_{\tilde h} - u^l_{\tilde h}\|_{L^2(D)}\lesssim l^{-1}$, and thus by the triangle inequality $\|u_h - u^l_{\tilde h}\|\lesssim l^{-1}$ as claimed.

Then

$$g_l(f) := u^l_{\tilde h} + \lambda f \tag{58}$$

is an approximation to $\nabla\Phi(f)$ satisfying $\|g_l(f) - \nabla\Phi(f)\|_{L^2(D)}\lesssim l^{-1}$.

C.4. Multilevel convergence of gradient descent

We first show L-smoothness and μ-strong convexity for the objective in (56).

With the notation from Sec. C.2 the norm of the operator $\mathcal{O}: L^2(D)\to\mathbb{R}^{n_y}$ satisfies $\|\mathcal{O}\| = \|\xi\|_{L^2(D)}$. Hence for all $f,g\in L^2(D)$, using that $\mathcal{A}: L^2(D)\to L^2(D)$ is bounded with norm 1 by Sec. C.1, and using the formula for $\nabla\Phi$ computed in Sec. C.2,

$$\begin{aligned}\|\nabla\Phi(f)-\nabla\Phi(g)\|_{L^2(D)} &= \|\mathcal{A}(\mathcal{O}(u_f-u_g)^\top\Gamma^{-1}\xi(\cdot)) + \lambda(f-g)\|_{L^2(D)}\\ &\le \|\mathcal{O}(u_f-u_g)^\top\Gamma^{-1}\xi(\cdot)\|_{L^2(D)} + \|\lambda(f-g)\|_{L^2(D)}\\ &\le (\|\xi\|^2_{L^2(D)}\|\Gamma^{-1}\|_{\mathbb{R}^{n_y\times n_y}} + \lambda)\|f-g\|_{L^2(D)}.\end{aligned}$$

This shows that Φ is L-smooth with $L = C\|\xi\|^2_{L^2(D)}\|\Gamma^{-1}\|_{\mathbb{R}^{n_y\times n_y}} + \lambda$.

Moreover, since $F: L^2(D)\to\mathbb{R}^{n_y}$ is bounded linear, the term $f\mapsto \frac{1}{2}\|\Gamma^{-1/2}(F(f)-y)\|^2_{\mathbb{R}^{n_y}}$ is convex. Hence, the added regularizer ensures Φ in (56) to be μ-strongly convex with $\mu=\lambda$.

In all this shows that Example 3 is in the setting of Example 2 with $\alpha=1$ (i.e. Φ is L-smooth, μ-strongly convex, and g_l in (58) satisfies the first inequality in (10) with $\alpha = 1$). Thus, for small enough $\eta>0$, the iteration $f_{j+1} = f_j - \eta\nabla g_{l_j(\varepsilon)}(f_k)$ with the approximate gradient from subsection C.3, converges to the unique minimizer $f_*\in L^2(D)$ of Φ as $k\to\infty$. With the multilevel choice $l_{K,j}(\varepsilon)$, $j=1,\dots,K(\varepsilon)$ as in (21), it holds $\|f_{K(\varepsilon)} - f_*\|_{L^2(D)}\lesssim\varepsilon$ 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 8.

Appendix D. Optimality of the multilevel rate for gradient descent

We give a simple example in the setting of Example 2 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 Φ by $x_* := 0$. We consider gradient descent with a fixed step size $\eta \in (0,1)$, which amounts to (cp. (6))

$$\Psi(x) = x - \eta x, \qquad \Psi_l(x) = x - \eta(x - l^{-\alpha}).$$

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

$$x_{k+1} = \Psi_{l_k}(x_k) = (1-\eta)x_k + \eta l_k^{-\alpha} \tag{59}$$

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

Proposition 25 *Let $x_0 \geq 0$, and let x_k be as in (59). 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 $|x_K - x_*| \leq \varepsilon$,*

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

Proof We have

$$x_K = cx_{K-1} + (1-c)l_{K-1}^{-\alpha} = c^K x_0 + \sum_{j=0}^{K-1}(1-c)c^{K-1-j}l_j^{-\alpha} \geq (1-c)\sum_{j=0}^{K-1} c^{K-1-j} l_j^{-\alpha}.$$

The minimizer of $\sum_{j=0}^{K-1} l_j$ under the constraint $(1-c)\sum_{j=0}^{K-1} c^{K-1-j} l_j^{-\alpha} \leq \varepsilon$ satisfies according to Lemma 6 and (20),

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

■

Appendix E. Notation: particle methods

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

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

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

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

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

Appendix F. Algorithm: Multilevel ensemble Kalman inversion

The original EKI method in Iglesias et al. (2013) 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 $\{v_j^{(m)}\}_{m=1}^{M}$ can be written as

$$v_{j+1}^{(m)} = v_j^{(m)} + C^{v,H}(v_j)(C^{H,H}(v_j) + h^{-1}\Sigma)^{-1}(z_{j+1}^{(m)} - H(v_j^{(m)})), \quad j = 1,\dots,J, \tag{60}$$

where $z_{j+1}^{(m)} \sim \mathcal{N}(y, h^{-1}\Sigma)$ are perturbed observations. Viewing $h > 0$ as step size the authors in Schillings and Stuart (2017) motivated to take the limit $h \to 0$ resulting in the system of coupled SDEs (26), which has been analysed rigorously in Blömker et al. (2018); Blömker et al. (2021). 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 (60) as

$$v_{j+1}^{(m),l} = v_j^{(m),l} + \frac{1}{M}\sum_{r=1}^{M} \alpha_j^{(r),(m),l} v_j^{(r)} \tag{61}$$

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

$$\alpha_j^{(r),(m),l} = \langle H_l(v_j^{(r)}) - \bar{H}_l, (C^{H_l,H_l}(v_j^l) + h^{-1}\Sigma)^{-1}(z_{j+1}^{(m)} - H_l(v_t^{(m),l}))\rangle.$$

This observation is useful for an efficient implementation of the multilevel formulation of EKI and TEKI respectively based on the discretization scheme (60) which we summarize as algorithm below. Note that for standard EKI one needs to run the algorithm with the choice $H_l \equiv F_l$, $z = y \in \mathbb{R}^{n_y}$ and $\Sigma = \Gamma$ and for TEKI, with the corresponding choice

$$H_l(\cdot) = \begin{pmatrix} F_l(\cdot) \\ \mathrm{Id} \end{pmatrix}, \quad z = \begin{pmatrix} y \\ \mathbf{0}_{\mathbb{R}^{n_x}} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Gamma & 0 \\ 0 & \frac{1}{\lambda}C_0 \end{pmatrix}.$$

Algorithm 1 Multilevel ensemble Kalman inversion (ML-EKI)

```
Require: initial ensemble $v_0^{(m),0}$, ensemble size $M$, forward model $(H_l)_{l\geq 0}$, 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 $\varepsilon > 0$.
set the number of iteration $K \propto \log(\varepsilon^{-1})$
set $x_0 = \frac{1}{M}\sum\limits_{m=1}^{M} v_0^{(m),0}$
For $j = 0, \dots, K-1$
    compute level $l_j \propto \varepsilon^{-\frac{1}{\alpha}} c^{\frac{K-1-j}{1+\alpha}}$
    If j=0
        $v_0^{(m),l_0} = v_0^{(m),0},\ m = 1, \dots, M$
    Else
        $v_0^{(m),l_j} = v_\tau^{(m),l_{j-1}},\ m = 1, \dots, M$
    EndIf
    For $n = 0, 1, \dots N-1$
        $y_{n+1}^{(m),l_j} \sim \mathcal{N}(y, h^{-1}\Gamma),\ m = 1, \dots, M$
        $\alpha_n^{(r),(m),l_j} = \langle F_{l_j}(v_n^{(r)}) - \bar{F}_{l_j}, (C^{F_{l_j},F_{l_j}}(v_n^{l_j}) + h^{-1}\Gamma)^{-1}(y_{n+1}^{(m),l_j} - F_{l_j}(v_n^{(m),l_j}))\rangle$
        $v_{n+1}^{(m),l_j} = v_n^{(m),l_j} + \frac{1}{M}\sum_{r=1}^{M} \alpha_n^{(r),(m),l_j} v_n^{(r),l_j}$
    EndFor
    set $x_{j+1} = \frac{1}{M}\sum\limits_{m=1}^{M} v_N^{(m),l_j}$
EndFor
```

Appendix G. Proofs of Section 4

G.1. Proof of Proposition 11

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

$$\mathrm{d}v_t^{(m)} = -(C(v_t) + B)F^\top \Gamma^{-1}(Fv_t^{(m)} - y)\,\mathrm{d}t + C(v_t)F^\top \Gamma^{-1/2}\,\mathrm{d}W_t^{(m)}.$$

We define $\bar{\mathfrak{r}}_t := \Gamma^{-1/2}F(\bar{v}_t - v^\dagger) \in \mathbb{R}^{n_y}$ and $\mathfrak{r}_t^{(m)} := \Gamma^{-1/2}F(v_t^{(m)} - v^\dagger) \in \mathbb{R}^{n_y}$ such that $\frac{1}{2}\|\bar{\mathfrak{r}}_t\|^2_{\mathbb{R}^{n_y}} = \ell(\bar{v}_t)$ and

$$\mathrm{d}\bar{\mathfrak{r}}_t = -(C(\mathfrak{r}_t) + \tilde{B})\bar{\mathfrak{r}}_t\,\mathrm{d}t + C(\mathfrak{r}_t)\,\mathrm{d}\bar{W}_t$$

with $\tilde{B} := \Gamma^{-1/2}FBF^\top\Gamma^{-1/2}$. Following Theorem 5.2 in Blömker et al. (2019) we can bound

$$\mathbb{E}[\|\bar{\mathfrak{r}}_{s+\tau}\|^2_{\mathbb{R}^{n_y}}] \le \mathbb{E}[\|\bar{\mathfrak{r}}_s\|^2_{\mathbb{R}^{n_y}}] - \sigma\int_s^{s+\tau}\mathbb{E}[\|\bar{\mathfrak{r}}_u\|^2_{\mathbb{R}^{n_y}}]\,\mathrm{d}u,$$

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

$$\mathbb{E}[\frac{1}{2}\|F_{l_j}\bar{v}_{t_{j+1}}^{l_j} - y\|^2_\Gamma] \le (1-\sigma\cdot\tau)\mathbb{E}[\frac{1}{2}\|F_{l_j}\bar{v}_{t_j}^{l_{j-1}} - y\|^2_\Gamma],$$

where we have defined $\|\cdot\|^2_\Gamma = \|\Gamma^{-1/2}\cdot\|^2_{\mathbb{R}^{n_y}}$. With Assumption 10 and the reverse triangle inequality $|\|Fx\| - \|Fy\|| \le \|Fx - Fy\|$ we have that

$$\begin{aligned}
e_{j+1} = \mathbb{E}[\frac{1}{2}\|F\bar{v}_{t_{j+1}}^{l_j} - y\|^2_\Gamma] &= \mathbb{E}[\frac{1}{2}\|F_{l_j}\bar{v}_{t_{j+1}}^{l_j} - y\|^2_\Gamma] + \mathbb{E}\left[\frac{1}{2}\|F\bar{v}_{t_{j+1}}^{l_j} - y\|^2_\Gamma - \frac{1}{2}\|F_{l_j}\bar{v}_{t_{j+1}}^{l_j} - y\|^2_\Gamma\right]\\
&\le (1-\sigma\cdot\tau)\mathbb{E}[\frac{1}{2}\|F_{l_j}\bar{v}_{t_j}^{l_{j-1}} - y\|^2_\Gamma] + b_1 l_j^{-\alpha}\\
&\le (1-\sigma\cdot\tau)\mathbb{E}[\frac{1}{2}\|F\bar{v}_{t_j}^{l_{j-1}} - y\|^2_\Gamma] + (1-\sigma\cdot\tau)b_1 l_j^{-\alpha} + b_1 l_j^{-\alpha}\\
&\le (1-\sigma\cdot\tau)e_j + b l_j^{-\alpha},
\end{aligned}$$

for some constant $b_1 > 0$ and $b = (2-\sigma\cdot\tau)b_1$. We note that we have used that $\mathbb{E}[\|\bar{v}_{t_{j+1}}^{l_j}\|^2_{\mathbb{R}^{n_x}}]$ remains uniformly bounded, see e.g. Lemma 5 in Ding and Li (2021a). Moreover, since we assumed finite second moments of the initial distribution Q_0, we have that $\mathbb{E}[\|\bar{v}_0\|^2_{\mathbb{R}^{n_x}}] < \infty$ and by local Lipschitz continuity of $x \mapsto \frac{1}{2}\|Fx - y\|^2_\Gamma$ we have that $e_0 = \mathbb{E}[\|F\bar{v}_0 - y\|^2_\Gamma] < \infty$. With the above computations we have verified that the error quantity e satisfies the decay assumption (8) (respectively the generalization in Rmk. 9) and therefore, the assertion follows by application of Theorem 5 and Theorem 8. ■

G.2. Proof of Proposition 12

Proof We again 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 TEKI can be written as

$$\begin{aligned}
\mathrm{d}v_t^{(m)} = &-(C(v_t) + B)(F^\top\Gamma^{-1}(Fv_t^{(m)} - y) + \lambda C_0^{-1}v_t^{(m)})\,\mathrm{d}t\\
&+ C(v_t)F^\top\Gamma^{-1/2}\,\mathrm{d}W_t^{(m)} + \sqrt{\lambda}C(v_t)C_0^{-1/2}\,\mathrm{d}W_t^{(m)}.
\end{aligned}$$

Next, we define $\bar{\mathfrak{r}}_t := \Sigma^{-1/2}H(\bar{v}_t - x_*) \in \mathbb{R}^{n_z}$ and $\mathfrak{r}_t^{(m)} := \Sigma^{-1/2}H(\bar{v}_t - x_*) \in \mathbb{R}^{n_z}$, where $x_* \in \mathcal{X}$ is the unique minimizer of ℓ_R and H, Σ are defined in (25), such that

$$\frac{1}{2}\|\bar{\mathfrak{r}}_t\|^2_{\mathbb{R}^{n_z}} = \frac{1}{2}\|\Gamma^{-1/2}F(\bar{v}_t - x_*)\|^2_{\mathbb{R}^{n_y}} + \frac{\lambda}{2}\|C_0^{-1/2}(\bar{v}_t - x_*)\|^2_{\mathbb{R}^{n_x}}.$$

Since $x_* \in \mathcal{X}$ is the unique minimizer of ℓ_R it holds true that

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

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

$$\mathrm{d}\bar{\mathfrak{r}}_t = -(C(\mathfrak{r}_t) + \tilde{B})\bar{\mathfrak{r}}_t \,\mathrm{d}t + C(\mathfrak{r}_t)\,\mathrm{d}\bar{W}_t.$$

The assertion follows similarly to the proof of Proposition 11. ■

Appendix H. Algorithm: Multilevel interacting Langevin MCMC

The multilevel interacting Langevin sampler is based on its particle approximation (37). 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 (32). We refer to Ding and Li (2021b) 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.

Appendix I. Proofs of Sec. 5

I.1. Proof of Proposition 14

Proof We define $\Phi_l : \mathcal{P} \to \mathbb{R}$ by $\Phi_l(\rho) = \mathrm{KL}(\rho\|\rho_*^l)$ and $\Phi : \mathcal{P} \to \mathbb{R}$ by $\Phi(\rho) = \mathrm{KL}(\rho\|\rho_*)$ and assuming that for $\sigma_1, \sigma_2 > 0$ we have $\nabla^2 \ell_R^l > \sigma_1 \mathrm{Id}$ and $C(\rho_j) > \sigma_2 \mathrm{Id}$. From (Garbuno-Inigo et al., 2020, Proposition 2) it follows that there exists a constant $c \in (0,1)$ such that $\mathrm{KL}(\rho_{j+1}\|\rho_*^{l_j}) \leq \mathrm{KL}(\rho_j\|\rho_*^{l_j})$. Furthermore, under Assumption 13 it holds true that

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

since by definition of the KL divergence we have that

$$\begin{aligned}
|\Phi_l(\rho) - \Phi(\rho)| &= |\mathrm{KL}(\rho\|\rho_*^l) - \mathrm{KL}(\rho\|\rho_*)| \\
&= |\int_{\mathbb{R}^{n_x}} \rho(x)\log(\rho(x))\,\mathrm{d}x - \int_{\mathbb{R}^{n_x}} \rho(x)\log(\rho_*^l(x))\,\mathrm{d}x \\
&\quad - \int_{\mathbb{R}^{n_x}} \rho(x)\log(\rho(x))\,\mathrm{d}x + \int_{\mathbb{R}^{n_x}} \rho(x)\log(\rho_*(x))\,\mathrm{d}x| \\
&\leq \int_{\mathbb{R}^{n_x}} \rho(x)|\ell_R^l(x) - \ell_R(x)|\,\mathrm{d}x.
\end{aligned}$$

Algorithm 2 Multilevel interacting Langevin sampler (ML-ILS)

```
Require: initial distribution $q_0$, ensemble size $M$, gradient of log-likelihood $(\nabla\Phi_R^l)_{l\geq 0}$, 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 $\varepsilon > 0$.
set the number of iteration $K \propto \log(\varepsilon^{-1})$
Initialize particle system as i.i.d. sample $v_0^{(m),0} \sim q_0$
set $\rho_0 = \frac{1}{M}\sum_{m=1}^{M} \delta_{v_0^{(m),0}}$
For $j = 0, \dots, K-1$
    compute level $l_j \propto \varepsilon^{-\frac{1}{\alpha}} c^{\frac{K-1-j}{1+\alpha}}$
    If j=0
        $v_0^{(m),l_0} = v_0^{(m),0},\ m = 1, \dots, M$
    Else
        $v_0^{(m),l_j} = v_\tau^{(m),l_{j-1}},\ m = 1, \dots, M$
    EndIf
    For $n = 0, 1, \dots N-1$
        $\xi_{n+1}^{(m),l_j} \sim \mathcal{N}(0, I)$
        $\Delta W_{n+1}^{(m),l_j} = \sqrt{h}\xi_{k+1}^{(m),l_j}$,
        $g_n^{(m),l_j} = -hC(v_n^{l_j})\nabla\ell_R^{l_j}(v_n^{(m),l_j}) + \sqrt{2C(v_t^{l_j})}\Delta W_{n+1}^{(m),l_j}$
        $v_{n+1}^{(m),l_j} = v_n^{(m),l_j} + g_n^{(m),l_j}$
    EndFor
    set $\rho_{j+1} = \frac{1}{M}\sum_{m=1}^{M} \delta_{v_N^{(m),l_j}}$
EndFor
```

With Assumption 13 it follows that

$$
\begin{aligned}
|\Phi_{l_j}(\rho_j) - \Phi(\rho_j)| &\leq \int_{\mathbb{R}^{n_x}} \rho_j(x)|\ell_R^{l_j}(x) - \ell_R(x)|\,\mathrm{d}x \\
&\leq b\int_{\mathbb{R}^{n_x}} \rho_j(x)\|F(x) - F_l(x)\|^2_{\mathbb{R}^{n_y}}\,\mathrm{d}x \leq bl^{-\alpha}.
\end{aligned}
$$

Finally, we obtain

$$
\begin{aligned}
\mathrm{KL}(\rho_{j+1}\|\rho_*) &= \mathrm{KL}(\rho_{j+1}\|\rho_*^{l_j}) + \Big(\mathrm{KL}(\rho_{j+1}\|\rho_*) - \mathrm{KL}(\rho_{j+1}\|\rho_*^{l_j})\Big) \\
&\leq c\mathrm{KL}(\rho_j\|\rho_*^{l_j}) + l_j^{-\alpha} \\
&\leq c\mathrm{KL}(\rho_j\|\rho_*) + (1+c)l_j^{-\alpha}.
\end{aligned}
$$

With the above computations we have verified that the error quantity e satisfies the decay assumption (8) (respectively the generalization in Rmk. 9) and therefore, the assertion follows by application of Theorem 5 and Theorem 8. ■

Appendix 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 $F_l(f) = \mathcal{O}(u_f^l)$, where u_f^l denotes the finite element solution to (5) using continuous piecewise linear finite elements on a uniform mesh with meshwidth $2^{-\tau(l)}$ where $\tau(l) = \lceil \log(l)/\log(2) \rceil$ such that $l = 2^{\tau(l)}$. Since $\mathcal{O} : H_0^1(D) \to R^{n_y}$ is continuous, it can be shown that $\|F(x) - F_l(x)\|_{\mathbb{R}^{n_y}} \lesssim l^{-1}\|x\|_{\mathbb{R}^{n_x}}$ for all l and all $x \in \mathbb{R}^{n_x}$, i.e. the convergence rate α 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 $C_0 = \mathrm{diag}(i^{-2\beta}, i = 1, \dots, n_x)$ 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 $\beta = 1$ and $n_x = 100$. The ensemble size was size $M = 50$ and the step size of the discretization scheme (60) was chosen a $h = 0.1$. The plotted error quantity is

$$
e_{K(\varepsilon)} = \mathbb{E}\left[\Gamma^{-1/2}\frac{1}{2}\|F_{\mathrm{ref}}(x_K(\varepsilon) - x_*)\|^2_{\mathbb{R}^{n_y}} + \frac{\lambda}{2}\|C_0^{-1/2}(x_K(\varepsilon) - x_*)\|^2_{\mathbb{R}^{n_x}}\right],
$$

with the reference solution x_* computed via

$$
x_* = (F_{\mathrm{ref}}^\top \Gamma^{-1} F_{\mathrm{ref}} + \lambda C_0^{-1})^{-1} {F_{\mathrm{ref}}}^\top \Gamma^{-1} y.
$$

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

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

where

$$x_* = ({F_{\mathrm{ref}}}^\top \Gamma^{-1} F_{\mathrm{ref}} + C_0^{-1})^{-1} {F_{\mathrm{ref}}}^\top \Gamma^{-1} y,$$

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.