ESAIM: COCV 31 (2025) 28
https://doi.org/10.1051/cocv/2025018

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

MINI-BATCH DESCENT IN SEMIFLOWS

Alberto Domínguez Corella and Martín Hernández*

Abstract. This paper investigates the application of mini-batch gradient descent to semiflows (gradient flows). Given a loss function (potential), we introduce a continuous version of mini-batch gradient descent by randomly selecting sub-loss functions over time, defining a piecewise flow. We prove that, under suitable assumptions on the potential generating the semiflow, the *mini-batch descent flow* trajectory closely approximates the original semiflow trajectory on average. In addition, we study a randomized minimizing movement scheme that also approximates the semiflow of the full loss function. We illustrate the versatility of this approach across various problems, including constrained optimization, sparse inversion, and domain decomposition. Finally, we validate our results with several numerical examples.

Mathematics Subject Classification. 34G25, 49J52, 37C10, 35K55, 37L05.

Received July 11, 2024. Accepted February 6, 2025.

1. Introduction

Probability theory has significantly impacted various areas of mathematics, particularly algorithms and combinatorics. The use of randomness in algorithms dates back to Monte Carlo methods [1, 2]. In subsequent years, random algorithms gained strength in optimization, notably with the emergence of techniques such as simulated annealing [3] and genetic algorithms [4], which were used in complex optimization problems. Recently, stochastic gradient descent algorithms have attracted attention in artificial intelligence; their development has been crucial in training large-scale models, especially machine learning algorithms [5, 6].

A notable variant of stochastic descent algorithms is the so-called mini-batch gradient descent. Unlike vanilla stochastic gradient descent, mini-batch implementations typically aggregate gradients over small batches, reducing gradient variance [7]. This variant balances robustness and efficiency and has become one of the most widely used implementations of gradient descent in deep learning. The continuous equivalent of gradient descent is the so-called gradient flow. With small learning rates, mini-batch gradient descent mimics the gradient flow trajectory of the full batch loss function. This paper is concerned with mini-batch descent applied to *semiflows* (gradient flows of convex non-necessarily differentiable functions). Given a loss function expressed as the average of *sub-loss* functions, we compare the gradient flow of the full loss function with a flow that does not strictly follow the steepest descent of the full loss function but instead follows portions of it, determined by randomly chosen batches of sub-loss functions changing over time. We refer to this process as *mini-batch descent flow*. A full description of how these trajectories are generated is given in the next section.

Keywords and phrases: Gradient flow, mini-batch, stochastic gradient descent, domain decomposition.

Chair for Dynamics, Control, Machine Learning, and Numerics, Alexander von Humboldt-Professorship, Department of Mathematics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany.

* Corresponding author: martin.hernandez@fau.de

Our main interest lies in comparing trajectories rather than finding minima of the full loss function through the mini-batch descent-generated trajectory. Provided that the replacement of batches is done sufficiently often, we prove that the trajectory generated by the mini-batch descent flow is close (in expectation) to the original trajectory generated by the gradient flow of the full loss function. By examining the paths taken by the trajectories of the mini-batch descent flow, it is possible to gain insights into their convergence properties and robustness to noise.

For simplicity, we focus on proper lower semicontinuous convex functionals defined on Hilbert spaces. With some standard modifications, the results also hold for semi-convex functionals by replacing the convex subdifferential with the limiting one. An advantage of considering gradient flows in Hilbert spaces is that this framework allows several solutions of evolution equations to be interpreted as gradient flows of suitable functionals. To address non-single-valued subdifferentials, we introduce an appropriate concept of a non-biased estimator, ensuring that the expected value of the randomly chosen subgradient aligns with the minimal norm subdifferential of the loss function at each step. We quantify the discrepancy of the mini-batch descent using a measure function analogous to the one used in stochastic gradient-based methods to quantify algorithm variance.

Additionally, we introduce a suitable randomized version of the *minimizing movement scheme* and study its relationship with the gradient flow of the full loss function. Through rigorous analysis, we provide convergence estimates and conditions under which this randomized minimizing movement scheme can approximate the gradient flow. This has significant implications for the long-term behavior of the trajectories generated by this mini-batch descent, demonstrating that they can reach arbitrarily good approximate minimizers of the full loss function.

A full account of the results is given in the next section. Let us now provide an overview of the related literature and highlight our contributions.

The idea of comparing the dynamics generated by an equation with a dynamic changing over time (as in stochastic algorithms) was first employed in [8] for solving particle interacting dynamics under the name of random batch methods. These methods have been further extended to address other problems [9–12]. The mini-batch descent scheme we use is based on [8], Algorithm 2b. This algorithm was further used in [13] to generate approximate optimal control trajectories. We extend the convergence results for state trajectories previously obtained in [13] from positive definite matrices to general convex functions and from Euclidean to Hilbert spaces. Infinite-dimensional spaces offer several advantages for problems where the *optimize then discretize* paradigm is more appropriate. Results for optimal control problems with gradient flow dynamics can potentially be derived by adjusting the proofs in [13] with the convergence results presented here. Additionally, by allowing set-valued operators from a subdifferential, further results for control-constrained problems can be considered. We give an example of how trajectories look when constraints are induced by means of indicator functions in Section 4.1.

In [14], a stochastic process was introduced as a continuous-time representation of the stochastic gradient descent algorithm. This process is characterized as a dynamical system coupled with a continuous-time index process living on a finite state space, wherein the dynamical system, representing the gradient descent part, is coupled with the process on the finite state space, representing the random sub-sampling. It is proved there that the process converges weakly to the gradient flow with respect to the full target function as the step size approaches zero. In contrast, the process studied in this paper does not randomly select when to change from following the steepest descent of a sub-loss function to follow another, *i.e.*, the waiting time is fixed and deterministic. However, we obtain much more robust estimates for the trajectories. We also mention [15], where sparse inversion and classification problems were studied from this same continuous-time perspective. Inspired by this, we also consider a sparse inversion problem in Section 4.2 for illustrative purposes of how our abstract results can be applied.

Let us now mention [16], where an abstract framework for solving evolutionary equations is considered. This framework is also inspired by stochastic algorithms used in the machine learning community. The main idea there is to consider a splitting of an operator and then construct a sequence that solves an implicit Euler scheme that at each step randomly selects operators from the splitting. This is in the same spirit as in [8, 12, 13]. Our results for the minimizing movements are inspired by [16] and are formulated in a somewhat different functional setting. Though the framework in [16] is very general, it is also very abstract and requires a lot of verification even

for simple problems. In comparison, our framework does not require much verification of spaces and operators; one just needs to provide the potential/loss function. Also, gradient flows are very flexible, as, *e.g.*, the heat flow can be seen as an L^2 flow or as an H^{-1} flow. Moreover, while we do require operators to be monotone, we allow them to be non-single-valued. Thus, different problems can be considered, such as the porous media equation and the one-phase Stefan problem. One of the seminal contributions of [16] is that they identify that mini-batch descent trajectories can be used for random domain decomposition algorithms to solve evolutionary equations; they present as an example a parabolic equation with the p-Laplacian. We present an example of domain decomposition for a parabolic obstacle problem, which comes from an operator that is non-single-valued.

The paper is organized as follows. Section 2 contains an overview of the problem and the main results. The proofs are carried out in Section 3. In Section 4, we show how the proposed randomized schemes can be applied in different instances: constrained optimization, sparse inversion, and domain decomposition. Finally, in the Appendix, we provide further details on the decomposition of the minimal norm subgradient and the proposed variance measure based on it.

2. Formulation of the problem and main results

Let $\mathcal{H}$ be a Hilbert space and $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ a convex, lower semicontinuous, and proper function. For an initial datum $u_0 \in \operatorname{dom} \partial\Phi$, we consider the gradient flow equation

$$\begin{cases} \dot{u}(t) \in -\partial\Phi(u(t)) & t \in (0,\infty), \\ u(0) = u_0, \end{cases} \tag{2.1}$$

where u_0 is the initial data. A locally absolutely continuous function $u : [0,+\infty) \to \mathcal{H}$ is said to be a strong solution of (2.1) if $u(0) = u_0$ and $0 \in \dot{u}(t) + \partial\Phi(u(t))$ for almost every $t \in [0,\infty)$. When $u_0 \in \operatorname{dom} \partial\Phi$, equation (2.1) has a unique strong solution [17], Theorem 17.2.2.

Let $p_1, \dots, p_n$ be positive numbers such that $\sum_{i=1}^n p_i = 1$. We assume that Φ can be represented as the average

$$\Phi = \sum_{i=1}^n p_i \Phi_i, \tag{2.2}$$

where $\Phi_1, \dots, \Phi_n : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are convex, lower semicontinuous, and proper functions satisfying

$$\inf_{u\in\mathcal{H}} \Phi_i(u) > -\infty \qquad \text{and} \qquad \operatorname{dom}\Phi \subset \operatorname{dom}\Phi_i \subset \overline{\operatorname{dom}\Phi}, \quad \forall i \in \{1,\dots,n\}. \tag{2.3}$$

The idea behind representation (2.2) is that the potential Φ can be obtained as the expected value of a random variable that takes the value Φ_i with probability p_i for each $i \in \{1,\dots,n\}$.

Let $\mathcal{B}_1, \dots, \mathcal{B}_m$ be nonempty sets such that $\bigcup_{j=1}^m \mathcal{B}_j = \{1,\dots,n\}$; these are called batches. To each batch $\mathcal{B}_j$, there is an associated positive probability π_j in such a way that

$$p_i = \sum_{j : i \in \mathcal{B}_j} \frac{\pi_j}{|\mathcal{B}_j|} \quad \text{for each} \quad i \in \{1,\dots,n\}.$$

The idea behind mini-batch descent is that the flow does not follow strictly the steepest descent of the full potential, but a portion of it determined by a randomly chosen batch.

For a set $\mathcal{B} \subset \{1,\dots,n\}$, let $\Phi_{\mathcal{B}} : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be given by

$$\Phi_{\mathcal{B}}(u) := \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \Phi_i(u).$$

In order to describe the process formally, consider a positive parameter $\varepsilon > 0$ and a sequence $\{j_k\}_{k\in\mathbb{N}}$ of independent random variables. The positive parameter represents the waiting time at which the flow changes from following one subgradient batch to another at each step. On the other hand, each j_k represents the choice of batch $\mathcal{B}_{j_k}$ at the k^{th} step. The sequence $\{j_k\}_{k\in\mathbb{N}}$ is assumed to satisfy $\mathbb{P}(j_k = j) = \pi_j$ for each $j \in \{1,\dots,m\}$ and $k \in \mathbb{N}$. Set $t_0 := 0$ and define recursively $t_k := t_{k-1} + \varepsilon$ for $k \in \mathbb{N}$.

We will say that a continuous function $v_\varepsilon : [0,+\infty) \to \mathcal{H}$ is a solution of the *mini-batch descent flow* scheme if $v_\varepsilon(0) = u_0$, v_ε is locally absolutely continuous in (t_{k-1}, t_k), and

$$0 \in \dot{v}_\varepsilon + \partial\Phi_{\mathcal{B}_{j_k}}(v_\varepsilon) \quad \text{for almost every} \quad t \in [t_{k-1}, t_k) \quad \text{and for all} \quad k \in \mathbb{N}. \tag{2.4}$$

Before proceeding further, let us first address the well-posedness of this scheme, *i.e.*, ensuring existence and uniqueness for the piecewise flow (2.4).

Theorem 2.1. *There exists a unique solution $v_\varepsilon : [0,+\infty) \to \mathcal{H}$ of the mini-batch descent flow scheme. Moreover,*

(i) If $\operatorname{dom}\partial\Phi_{\mathcal{B}_j} \subset \operatorname{dom}\Phi$ for all $j \in \{1,\dots,m\}$, then $v_\varepsilon : [0,+\infty) \to \mathcal{H}$ is locally absolutely continuous;
(ii) If $\operatorname{dom}\partial\Phi_{\mathcal{B}_j} \subset \operatorname{dom}\partial\Phi$ for all $j \in \{1,\dots,m\}$, then $v_\varepsilon : [0,+\infty) \to \mathcal{H}$ is locally Lipschitz continuous.

Proof. This follows from Theorem 3.2 and Proposition 3.3. □

To illustrate our mini-batch gradient descent method, Algorithm 1 presents a step-by-step description of the dynamics construction. For the sake of clarity, we present the algorithm in a bounded time interval $(0,T)$. However, it can be simply modified to the interval $(0,\infty)$ by choosing $K = \infty$.

Algorithm 1 mini-batch descent flow (finite time)

Require: Initial data u_0, step size ε, final time T, batches $\{\mathcal{B}_1,\dots,\mathcal{B}_m\}$, probabilities $\{\pi_1,\dots,\pi_m\}$
1: Initialize $v_\varepsilon(0) \leftarrow u_0$
2: Define $K \leftarrow \lfloor T/\varepsilon \rfloor$
3: **for** $k \leftarrow 1$ to K **do**
4: Sample a batch index j_k according to probabilities $\{\pi_1,\dots,\pi_m\}$
5: Let $\mathcal{B} \leftarrow \mathcal{B}_{j_k}$
6: Compute the subgradient $\partial\Phi_{\mathcal{B}}(u) \leftarrow \partial\left(\frac{1}{|B|}\sum_{i\in B}\Phi_i(u)\right)$
7: Solve $\dot{v}_\varepsilon \in -\partial\Phi_{\mathcal{B}}(v_\varepsilon)$ in $[t_{k-1}, t_k)$
8: **end for**
return v_ε

We mention that a uniform grid was chosen for notational convenience; the arguments of this article should still work if an unstructured time mesh we to be used.

2.1. Convergence

For each $t \in [0,+\infty)$, define $k_t := \min\{k \in \mathbb{N} : \ t < t_k\}$. We can rewrite the mini-batch descent flow as

$$\begin{cases} \dot{v}_\varepsilon(t) \in -\partial\Phi_{\mathcal{B}_{j_{k_t}}}(v_\varepsilon(t)) & t \in (0,\infty), \\ v_\varepsilon(0) = u_0. \end{cases} \tag{2.5}$$

In order for the mini-batch descent to be a non-biased estimator, in the sense that the expected value of the randomly chosen subgradient coincides with the subdifferential of Φ at each step, the sum rule for subdifferentials

is assumed to hold, *i.e.*,

$$\partial\Phi(u) = \sum_{j=1}^{m} \pi_j \, \partial\Phi_{\mathcal{B}_j}(u) \quad \text{for all } u \in \operatorname{dom}\partial\Phi. \tag{2.6}$$

This ensures that for each $j \in \{1, \dots, m\}$ there exists $\xi_j : \operatorname{dom}\partial\Phi \to \operatorname{dom}\partial\Phi_{\mathcal{B}_j}$ such that

$$\partial\Phi(u)^\circ = \sum_{j=1}^{m} \pi_j \xi_j(u) \quad \forall u \in \operatorname{dom}\partial\Phi, \tag{2.7}$$

where $\partial\Phi(u)^\circ = \operatorname{argmin}\{\|\xi\|_{\mathcal{H}} : \xi \in \partial\Phi(u)\}$ for all $u \in \operatorname{dom}\partial\Phi$. We see then that for each $k \in \mathbb{N}$,

$$\mathbb{E}\xi_{j_k}(u) = \partial\Phi(u)^\circ \quad \forall u \in \operatorname{dom}\partial\Phi.$$

In order to provide a variance measure for mini-batch descent, consider the function $\Lambda : \mathcal{H} \to \mathbb{R}$ given by

$$\Lambda(u) := \sum_{j=1}^{m} \pi_j \|\xi_j(u) - \partial\Phi(u)^\circ\|_{\mathcal{H}}^2. \tag{2.8}$$

This function is the usual quantifier of variance used to provide bounds and estimates in stochastic gradient descent algorithms. We observe that $\operatorname{Var}[\xi_{j_k}] = \Lambda$ for all $k \in \mathbb{N}$. Moreover, this function can be used to bound the gap between the gradient and mini-batch descent flows.

Proposition 2.2. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ the solution of the mini-batch descent equation (2.5). Then,*

$$\|v_\varepsilon(t) - u(t)\|_{\mathcal{H}} \leq \max_{j \in \{1,\dots,m\}} \pi_j^{-\frac{1}{2}} t^{\frac{1}{2}} \Big(\int_0^t \Lambda\big(u(s)\big) \, \mathrm{d}s \Big)^{\frac{1}{2}} \quad \forall t \in [0, +\infty).$$

In particular, if $\Lambda : [0, +\infty) \to \mathbb{R}$ is locally integrable, then the v_ε is locally bounded in $L^2\big([0, +\infty); \mathcal{H}\big)$ for every $\varepsilon > 0$.

Proof. It follows from Proposition 3.4 and Corollary 3.5. □

A natural question that arises is whether the function $\Lambda \circ u : [0, +\infty) \to \mathbb{R}$ is locally bounded or at least locally integrable. We discuss these issues in the Appendix A as well as the dependence of Λ on the chosen decomposition of the minimal norm subgradient.

Under additional assumptions on the mini-batch potentials $\Phi_{\mathcal{B}_1}, \dots, \Phi_{\mathcal{B}_m}$, we can provide a better estimate in expected value. The following two theorems guarantee, with different assumptions, the convergence in expectation of the mini-batch descent flow to the gradient flow as $\varepsilon \to 0^+$.

Theorem 2.3. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ the solution of the mini-batch descent flow equation (2.5). Suppose that $\Lambda \circ u : [0, +\infty) \to \mathcal{H}$ is locally integrable, and that for each $j \in \{1, \dots, m\}$, $\Phi_{\mathcal{B}_j}$ is locally bounded in its effective domain. Then, for each any $t \in [0, +\infty)$, $\{\varepsilon^{\frac{1}{2}} \dot{v}_\varepsilon\}_{\varepsilon > 0}$ is bounded in $L^2\big([0, t); \mathcal{H}\big)$ and*

$$\mathbb{E}\|v_\varepsilon(t) - u(t)\|_{\mathcal{H}}^2 \leq 2\varepsilon^{\frac{1}{2}} \|\varepsilon^{\frac{1}{2}} \dot{v}_\varepsilon\|_{L^2([0,t);\mathcal{H})} \Big(\int_0^t \Lambda\big(u(s)\big) \, \mathrm{d}s \Big)^{\frac{1}{2}} \quad \forall \varepsilon > 0.$$

In particular, for any $T > 0$ there exists $c_T > 0$ such that

$$\sup_{t\in[0,T]} \mathbb{E}\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \leq c_T \varepsilon^{\frac{1}{2}} \quad \forall \varepsilon > 0.$$

Proof. It follows from Lemma 3.6 and Corollary 3.8. □

It is possible to improve the rate of convergence by replacing the boundedness assumption on the mini-batch potentials with local Lipschitz continuity. We mention that this is not a big assumption in many cases, as for example, convex functions defined on a finite dimensional space are automatically locally Lipschitz continuous; however this is not the case for infinite dimensional spaces (due to the celebrated Hahn-Banach Theorem there exist unbounded linear functionals). Also note that there are convex functions that are locally bounded in its effective domain, but not locally Lipschitz, even if the underlying Hilbert space is finite dimensional[1].

Theorem 2.4. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ the solution of the mini-batch descent flow equation (2.5). Suppose that $\Lambda \circ u : [0, +\infty) \to \mathcal{H}$ is locally integrable, and that for each $j \in \{1, \dots, m\}$, $\Phi_{\mathcal{B}_j}$ is locally Lipschitz in its effective domain. Then, for any $t \in [0, +\infty)$, $\{\dot{v}_\varepsilon\}_{\varepsilon>0}$ is bounded in $L^2\big([0,t);\mathcal{H}\big)$ and*

$$\mathbb{E}\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \leq 2\varepsilon \big\|\dot{v}_\varepsilon\big\|_{L^2([0,t);\mathcal{H})} \Big(\int_0^t \Lambda\big(u(s)\big)\, \mathrm{d}s \Big)^{\frac{1}{2}} \quad \forall \varepsilon > 0.$$

In particular, for any $T > 0$ there exists $c_T > 0$ such that

$$\sup_{t\in[0,T]} \mathbb{E}\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \leq c_T \varepsilon \quad \forall \varepsilon > 0.$$

Proof. It follows from Lemma 3.6 and Theorem 3.7. □

Previous results will be employed in more particular settings, especially projected dynamics resulting from constrained optimization (Sect. 4.1), and sparse inversion problems arising in data science (Sect. 4.2).

2.2. Mini-batch minimizing movement

For some applications, such as domain decomposition [16], it is preferable to remove assumptions on the potential, such as local boundedness. The possible lack of absolute continuity of the solution of the mini-batch descent flow generates several difficulties in the analysis; therefore, it is more convenient to analyze it in a discrete setting. We begin by introducing a randomized variant of the so-called minimizing movement scheme; this is essentially a stochastic proximal point algorithm.

It can be proved that there exists a unique $w_1 \in \operatorname{argmin}_{w\in\mathcal{H}}\{\Phi_{\mathcal{B}_{j_1}}(w) + \frac{1}{2\varepsilon}\big\|w - u_0\big\|_{\mathcal{H}}^2\}$; see, *e.g.*, [17], Proposition 17.2.1. This procedure can continue in an iterative way, yielding a sequence $\{w_k\}_{k\in\mathbb{N}}$ satisfying

$$w_{k+1} \in \operatorname*{argmin}_{w\in\mathcal{H}} \left\{ \Phi_{\mathcal{B}_{j_{k+1}}}(w) + \frac{1}{2\varepsilon}\big\|w - w_k\big\|_{\mathcal{H}}^2 \right\} \quad \forall k \in \mathbb{N}. \tag{2.9}$$

Consider the function $w_\varepsilon : [0, +\infty) \to \mathcal{H}$ given by $w_\varepsilon(t) := w_{k_t}$. We see that for each $k \in \mathbb{N}$,

$$w_\varepsilon(t) = w_k \quad \forall t \in [t_{k-1}, t_k). \tag{2.10}$$

[1] The function $\psi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, given by $\psi(u) := -2\sqrt{u}$ if $u \geq 0$ and $+\infty$ otherwise, furnishes an example.

We say that $w : [0, +\infty) \to \mathcal{H}$ is a mini-batch minimizing movement associated with the mini-batch potentials $\Phi_{\mathcal{B}_1}, \dots, \Phi_{\mathcal{B}_m}$ if for all $t \in [0, +\infty)$,

$$\mathbb{E}\big\|w_\varepsilon(t) - w(t)\big\|_{\mathcal{H}}^2 \longrightarrow 0 \quad \text{as} \quad \varepsilon \to 0^+.$$

Below, Algorithm 2 illustrate the construction of the sequence $\{w_k\}_{k=1}^K$, for some fixed $K > 0$. Here, for the sake of clarity, we illustrate the construction for a fixed K. However, the same iteration can be done for $K = \infty$.

Algorithm 2 Mini-Batch Minimizing Movement (finite sequence)

Require: Initial data u_0, fixed ε, iterations K, batches $\{\mathcal{B}_1, \dots, \mathcal{B}_m\}$, probabilities $\{\pi_1, \dots, \pi_m\}$

Initialize $w_0 \leftarrow u_0$
for $k \leftarrow 1$ to K **do**
  Sample a batch index j_k according to probabilities $\{\pi_1, \dots, \pi_m\}$
  Let $\mathcal{B} \leftarrow \mathcal{B}_{j_k}$
  Let $\Phi_{\mathcal{B}}(u) \leftarrow \frac{1}{|B|} \sum_{i \in B} \Phi_i(u)$
  Solve $w_k \in \operatorname{argmin}_{w \in \mathcal{H}} \left\{ \Phi_{\mathcal{B}}(w) + \frac{1}{2\varepsilon} \big\|w - w_{k-1}\big\|_{\mathcal{H}}^2 \right\}$
  Let $w_\varepsilon(t) \leftarrow w_k$ in $[t_{k-1}, t_k)$
end for
return w_ε

The following result gives sufficient conditions for a gradient flow solution to be a mini-batch minimizing movement and, moreover, provides an error estimate.

Theorem 2.5. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $w_\varepsilon : [0, +\infty) \to \mathcal{H}$ the function generated by the proximal sequence (2.9). Assume that $\Lambda \circ u : [0, +\infty) \to \mathcal{H}$ is locally bounded. If u belongs to $C^1\big([0, +\infty); \mathcal{H}\big)$, then u is a mini-batch minimizing movement. Moreover, if there exists $\alpha \in (0, 1]$ such that $\dot{u}$ is locally α-Hölder continuous, then for any $T > 0$ there exists $c_T > 0$ such that*

$$\sup_{t \in [0,T]} \mathbb{E}\big\|w_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \le c_T\big(\varepsilon^{2\alpha} + \varepsilon^2 + \varepsilon\big) \quad \forall \varepsilon > 0.$$

Proof. It follows from Corollary 3.14. □

We highlight that to ensure the convergence of the random minimizing movement to the gradient flow, the above result requires only regularity assumptions on the gradient flow and not on the random scheme.

2.3. Asymptotic behavior

We have seen that under some hypotheses, the trajectory of mini-batch descent closely approximates the path of the gradient flow. It is natural to expect that one can say something about the minimization of the full potential. The following result analyzes the asymptotic behavior of the mini-batch descent flow.

Theorem 2.6. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ the solution of the mini-batch descent flow equation (2.5). Suppose that $\Lambda \circ u : [0, +\infty) \to \mathcal{H}$ is locally integrable and that there exists $\alpha \in (0, 1]$ such that, for each $j \in \{1, \dots, m\}$, $\Phi_{\mathcal{B}_j}$ is locally α-Hölder continuous in its effective domain. Then, the following statements hold.*

(i) If $\operatorname{dom} \partial\Phi_{\mathcal{B}_j} \subset \operatorname{dom} \Phi$ for all $j \in \{1, \dots, m\}$, then for every $\eta > 0$ there exists $T > 0$ such that

$$\mathbb{E}\Phi(v_\varepsilon(T)) \le \inf_{v \in \mathcal{H}} \Phi(v) + \eta \quad \text{for all } \varepsilon > 0 \text{ small enough.}$$

(ii) If $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is inf-compact, then there exists $u^* \in \operatorname{argmin}_{v\in\mathcal{H}} \Phi(v)$ satisfying that for every $\eta > 0$ there exists $T > 0$ such that

$$\mathbb{E}\left\|v_\varepsilon(T) - u^*\right\|_{\mathcal{H}}^2 \leq \eta \quad \text{for all } \varepsilon > 0 \text{ small enough.}$$

Proof. This is a particular case of Theorem 3.9. □

For the minimizing movement scheme, it is also possible to provide an analogous result on the asymptotic behavior of trajectories.

Theorem 2.7. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) $w_\varepsilon : [0, +\infty) \to \mathcal{H}$ the function generated by the proximal sequence (2.9). Suppose that $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is inf-compact, $\Lambda \circ u : [0, +\infty) \to \mathcal{H}$ is locally bounded and that $\dot{u}$ is continuous. Then, there exists $u^* \in \operatorname{argmin}_{v\in\mathcal{H}} \Phi(v)$ satisfying that for every $\eta > 0$ there exists $T > 0$ such that*

$$\mathbb{E}\left\|w_\varepsilon(T) - u^*\right\|_{\mathcal{H}}^2 \leq \eta \quad \text{for all } \varepsilon > 0 \text{ small enough.}$$

Proof. The proof is analogous to that of Theorem 3.9. □

3. Well-posedness and convergence of mini-batch descent

In the remainder of this paper, we assume that $u_0 \in \operatorname{dom}(\partial\Phi)$, condition (2.3) holds, and the sum rule (2.6) is satisfied.

3.1. Existence and uniqueness

The existence of the mini-batch descent flow (2.5) is established through an induction argument that ensures existence within each interval. The following argument encapsulates this process.

Lemma 3.1. *Let $k \in \mathbb{N}$. There exists a unique continuous function $v_k : [0, t_k] \to \mathcal{H}$ such that*

(i) $v_k(0) = u_0$;
(ii) v_k is locally absolutely continuous in (t_{l-1}, t_l) for all $l \in \{1, \dots, k\}$;
(iii) $0 \in \dot{v}_k + \partial\Phi_{\mathcal{B}_{j_l}}(v_k)$ a.e. in $[t_{l-1}, t_l]$ for all $l \in \{1, \dots, k\}$.

Proof. Let S denote the set of all natural numbers $k \in \mathbb{N}$ such that there exists a continuous function $v_k : [0, t_k] \to \mathcal{H}$ satisfying items (i)-(iii). We argue by mathematical induction. Since $u_0 \in \operatorname{dom}\Phi \subset \operatorname{dom}\Phi_{\mathcal{B}_{j_1}}$, by [17], Theorem 17.2.3 there exists a unique continuous function $w_1 : [0, t_1] \to \mathcal{H}$ such that $w_1(0) = u_0$, w_1 is locally absolutely continuous in $(0, t_1)$, and $0 \in w_1(t) + \partial\Phi_{\mathcal{B}_{j_1}}(w_1(t))$ for a.e. $t \in [0, t_1]$; thus $1 \in S$.

We now proceed with the inductive step. Let $k \in S$, then there exists a continuous function $v_k : [0, t_k] \to \mathcal{H}$ satisfying items (i)-(iii). Since $v_k(t_k) \in \operatorname{dom}\Phi_{\mathcal{B}_{j_k}} \subset \overline{\operatorname{dom}\Phi} \subset \overline{\operatorname{dom}\Phi_{\mathcal{B}_{j_{k+1}}}}$, by [17], Theorem 17.2.3, there exists a unique continuous function $w_{k+1} : [t_k, t_{k+1}] \to \mathcal{H}$ such that $w_{k+1}(t_k) = v_k(t_k)$, w_{k+1} is locally absolutely continuous in (t_k, t_{k+1}) and $0 \in \dot{w}_{k+1}(t) + \partial\Phi_{\mathcal{B}_{j_{k+1}}}(w_{k+1}(t))$ for a.e. $t \in [t_k, t_{k+1}]$. Define $v_{k+1} : [0, t_{k+1}] \to \mathcal{H}$ by

$$v_{k+1}(t) := \begin{cases} v_k(t) & \text{if} \quad t \in [0, t_k], \\ w_{k+1}(t) & \text{if} \quad t \in (t_k, t_{k+1}]. \end{cases}$$

By construction, v_{k+1} is a continuous function satisfying items (i)-(iii); thus $k + 1 \in S$. This completes the induction, and hence $S = \mathbb{N}$. □

Theorem 3.2. *There exists a unique continuous function $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ such that*

(i) $v_\varepsilon(0) = u_0$;

(ii) v_ε *is locally absolutely continuous in* (t_{k-1}, t_k) *for all* $k \in \mathbb{N}$*;*
(iii) $0 \in \dot{v}_\varepsilon + \partial\Phi_{\mathcal{B}_{j_k}}(v_\varepsilon)$ *a.e. in* $[t_{k-1}, t_k]$ *for all* $k \in \mathbb{N}$.

Proof. By Lemma 3.1, for each $k \in \mathbb{N}$ there exists a unique continuous function $v_k : [0, t_k] \to \mathcal{H}$ such that $v_k(0) = u_0$, v_k is locally absolutely continuous in (t_{k-1}, t_k) and $0 \in \dot{v}_k(t) + \partial\Phi_{j_k}(v_k(t))$ for a.e. $t \in [t_{k-1}, t_k]$. Define $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ by $v_\varepsilon(t) := v_{k_t}(t)$; clearly v_ε is continuous, and satisfies items (i)-(iii). Its uniqueness follows from the uniqueness of the functions $\{v_k\}_{k\in\mathbb{N}}$. □

Proposition 3.3. *Let* $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ *be the solution of the mini-batch descent flow equation (2.5). Then,*

(i) if $\operatorname{dom}\partial\Phi_{\mathcal{B}_j} \subset \operatorname{dom}\Phi$ *for all* $j \in \{1, \dots, m\}$*, then* $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ *is locally absolutely continuous;*
(ii) if $\operatorname{dom}\partial\Phi_{\mathcal{B}_j} \subset \operatorname{dom}\partial\Phi$ *for all* $j \in \{1, \dots, m\}$*, then* $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ *is locally Lipschitz continuous.*

Proof. Given $k \in \mathbb{N}$, by [18], Theorem 4.11, if $v_\varepsilon(t_{k-1}) \in \operatorname{dom}\Phi \subset \operatorname{dom}\Phi_{\mathcal{B}_{j_k}}$, then the restriction of v_ε to the subinterval $[t_{k-1}, t_k]$ is absolutely continuous; we conclude that v_ε is absolutely continuous in any compact subset of $[0, +\infty)$. This proves item (i); the proof of item (ii) follows the same argument replacing [18], Theorem 4.11 by [17], Theorem 17.2.2. □

3.2. Convergence

Proposition 3.4. *Let* $u : [0, +\infty) \to \mathcal{H}$ *be the solution of the gradient flow equation (2.1) and* $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ *the solution of the mini-batch descent flow equation (2.5). Then,*

$$\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}} \le \max_{j\in\{1,\dots,m\}} \pi_j^{-\frac{1}{2}} t^{\frac{1}{2}} \Big(\int_0^t \Lambda\big(u(s)\big)\,\mathrm{d}s \Big)^{\frac{1}{2}} \quad \forall t \in [0, +\infty).$$

Proof. Let $k \in \mathbb{N}$. By monotonicity of the subdifferential,

$$\langle -\dot{v}_\varepsilon(\tau) - \xi_{j_k}(u(\tau)),\, v_\varepsilon(\tau) - u(\tau)\rangle \ge 0 \quad \text{for a.e. } \tau \in (t_{k-1}, t_k),$$

where the functions ξ_{j_k} are given as in (2.7). Using that $\dot{u} = -\partial\Phi(u)^\circ$ a.e. in $[0, +\infty)$, this can be rewritten as

$$0 \le \langle -\dot{v}_\varepsilon(\tau) + \dot{u}(\tau),\, v_\varepsilon(\tau) - u(\tau)\rangle + \langle \partial\Phi(u(\tau))^\circ - \xi_{j_k}(u(\tau)),\, v_\varepsilon(\tau) - u(\tau)\rangle\,,$$

for a.e. $\tau \in (t_{k-1}, t_k)$. This implies

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\tau}\big\|v_\varepsilon(\tau) - u(\tau)\big\|_{\mathcal{H}}^2 &\le \big\|\partial\Phi(u(\tau))^\circ - \xi_{j_k}(u(\tau))\big\|_{\mathcal{H}}\big\|v_\varepsilon(\tau) - u(\tau)\big\|_{\mathcal{H}} \\ &\le \max_{j\in\{1,\dots,m\}} \pi_j^{-\frac{1}{2}} \sqrt{\Lambda\big(u(\tau)\big)}\big\|v_\varepsilon(\tau) - u(\tau)\big\|_{\mathcal{H}},\end{aligned}$$

for a.e. $\tau \in (t_{k-1}, t_k)$. Due to Theorem 3.2, the function $t \longmapsto \big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}$ is locally absolutely continuous in (t_{k-1}, t_k). Then, we can employ Grönwall's inequality [19], Theorem 5.2.2 to conclude that

$$\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}} \le \big\|v_\varepsilon(s) - u(s)\big\|_{\mathcal{H}} + \max_{j\in\{1,\dots,m\}} \pi_j^{-\frac{1}{2}} \int_s^t \sqrt{\Lambda\big(u(\tau)\big)}\,\mathrm{d}\tau \quad \forall s, t \in (t_{k-1}, t_k).$$

Set $\gamma := \max_{j\in\{1,\dots,m\}} \pi_j^{-\frac{1}{2}}$. By continuity of v_ε and u, we conclude that

$$\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}} \le \big\|v_\varepsilon(t_{k-1}) - u(t_{k-1})\big\|_{\mathcal{H}} + \gamma \int_{t_{k-1}}^t \sqrt{\Lambda\big(u(\tau)\big)}\,\mathrm{d}\tau \quad \forall t \in [t_{k-1}, t_k]. \tag{3.1}$$

Since (3.1) holds for arbitrary $k \in \mathbb{N}$, this implies

$$\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}} \leq \gamma\Big[\int_{t_{k_t-1}}^{t} \sqrt{\Lambda\big(u(\tau)\big)}\,\mathrm{d}\tau + \sum_{l=1}^{k_t-1}\int_{t_{l-1}}^{t_l} \sqrt{\Lambda\big(u(\tau)\big)}\,\mathrm{d}s\Big] = \gamma\int_0^t \sqrt{\Lambda\big(u(\tau)\big)}\,\mathrm{d}\tau,$$

for all $t \in [0,+\infty)$. The result follows then from Hölder's inequality. □

Corollary 3.5. *Let $u : [0,+\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $v_\varepsilon : [0,+\infty) \to \mathcal{H}$ the solution of the mini-batch descent flow equation (2.5). Then,*

$$\big\|v_\varepsilon(t)\big\|_{\mathcal{H}} \leq \max_{j\in\{1,\dots,m\}} \pi_j^{-\frac{1}{2}} t^{\frac{1}{2}}\Big(\int_0^t \Lambda\big(u(s)\big)\,\mathrm{d}s\Big)^{\frac{1}{2}} + t\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}} + \big\|u_0\big\|_{\mathcal{H}} \quad \forall t \in [0,+\infty).$$

Proof. Since u is Lipschitz continuous with constant $\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}}$ in $[0,+\infty)$, we have that

$$\big\|v_\varepsilon(t)\big\|_{\mathcal{H}} \leq \big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}} + \big\|u(t) - u_0\big\|_{\mathcal{H}} + \big\|u_0\big\|_{\mathcal{H}} \leq \big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}} + t\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}} + \big\|u_0\big\|_{\mathcal{H}},$$

for all $t \in [0,+\infty)$. The result follows then from Proposition 3.4. □

Lemma 3.6. *Let $u : [0,+\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $v_\varepsilon : [0,+\infty) \to \mathcal{H}$ the solution of the mini-batch descent flow equation (2.5). Suppose that $\Lambda \circ u : [0,+\infty) \to \mathcal{H}$ is locally integrable. Then, the following statements hold.*

(i) If, for each $j \in \{1,\dots,m\}$, $\Phi_{\mathcal{B}_j}$ is locally bounded in its effective domain, then for any $T > 0$, $\{\varepsilon^{\frac{1}{2}}\dot{v}_\varepsilon\}_{\varepsilon>0}$ is bounded in $L^2\big([0,T];\mathcal{H}\big)$.

(ii) If there exists $\alpha \in (0,1]$ such that, for each $j \in \{1,\dots,m\}$, $\Phi_{\mathcal{B}_j}$ is locally α-Hölder continuous in its effective domain, then for any $T > 0$, $\{\varepsilon^{\frac{1-\alpha}{2-\alpha}}\dot{v}_\varepsilon\}_{\varepsilon>0}$ is bounded in $L^2\big([0,T];\mathcal{H}\big)$.

Proof. Let $T > 0$ be given. Both items (i) and (ii) can be treated at the same time, allowing α to be zero; in this way a function is locally bounded if and only if it is locally 0-Hölder continuous. Define

$$r_T := \max_{j\in\{1,\dots,m\}} \pi_j^{-\frac{1}{2}} T^{\frac{1}{2}}\Big(\int_0^T \Lambda\big(u(s)\big)\,\mathrm{d}s\Big)^{\frac{1}{2}} + T\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}} + \big\|u_0\big\|_{\mathcal{H}}.$$

Since $\Lambda \circ u \in L^1\big([0,T];\mathbb{R}\big)$, $r_T < +\infty$. By assumption, there exists $L_T > 0$ such that, for any $j \in \{1,\dots,m\}$,

$$\frac{|\Phi_{\mathcal{B}_j}(v) - \Phi_{\mathcal{B}_j}(w)|}{\big\|v - w\big\|_{\mathcal{H}}^{\alpha}} \leq L_T \quad \text{for all } v, w \in \mathbb{B}_{\mathcal{H}}\big(0, r_T\big) \text{ satisfying } v \neq w.$$

Let $k \in \mathbb{N}$. By [17], Theorem 17.2.3, $\Phi_{\mathcal{B}_{j_k}} \circ v_\varepsilon$ is locally absolutely continuous in (t_{k-1}, t_k), and

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\Phi_{\mathcal{B}_{j_k}}\big(v_\varepsilon(\tau)\big) = -\big\|\dot{v}_\varepsilon(\tau)\big\|_{\mathcal{H}}^2 \quad \text{for a.e. } \tau \in (t_{k-1}, t_k). \tag{3.2}$$

From this and Corollary 3.5, for all $s, t \in (t_{k-1}, t_k)$,

$$\begin{aligned}\int_s^t \big\|\dot{v}_\varepsilon(\tau)\big\|_{\mathcal{H}}^2\,\mathrm{d}\tau &= \Phi_{\mathcal{B}_{j_k}}(v_\varepsilon(s)) - \Phi_{\mathcal{B}_{j_k}}(v_\varepsilon(t)) \leq L_T\big\|v_\varepsilon(s) - v_\varepsilon(t)\big\|_{\mathcal{H}}^{\alpha} \\ &\leq L_T\Big(\int_s^t \big\|\dot{v}_\varepsilon(\tau)\big\|_{\mathcal{H}}\,\mathrm{d}\tau\Big)^{\alpha} \leq L_T(t-s)^{\frac{\alpha}{2}}\Big(\int_s^t \big\|\dot{v}_\varepsilon(\tau)\big\|_{\mathcal{H}}^2\,\mathrm{d}\tau\Big)^{\frac{\alpha}{2}}.\end{aligned}$$

From where we conclude that, for all $s,t \in (t_{k-1},t_k)$,

$$\int_s^t \big\|\dot v_\varepsilon(\tau)\big\|_{\mathcal H}^2 \,\mathrm{d}\tau \le L_T^{\frac{2}{2-\alpha}}(t-s)^{\frac{\alpha}{2-\alpha}} = L_T^{\frac{2}{2-\alpha}}(t-s)^{\frac{2(\alpha-1)}{2-\alpha}}(t-s). \tag{3.3}$$

Since $k \in \mathbb{N}$ was arbitrary, we conclude from (3.3) that

$$\int_0^T \big\|\dot v_\varepsilon(\tau)\big\|_{\mathcal H}^2 \,\mathrm{d}\tau \le \int_{t_{k_T-1}}^T \big\|\dot v_\varepsilon(\tau)\big\|_{\mathcal H}^2 \,\mathrm{d}\tau + \sum_{l=1}^{k_T-1}\int_{t_{l-1}}^{t_l} \big\|\dot v_\varepsilon(\tau)\big\|_{\mathcal H}^2 \,\mathrm{d}\tau \le L_T^{\frac{2}{2-\alpha}}\varepsilon^{\frac{2(\alpha-1)}{2-\alpha}}\Big[T - t_{k_T-1} + \sum_{l=1}^{k_T-1}(t_l - t_{l-1})\Big].$$

We can then conclude that $\displaystyle\int_0^T \big\|\varepsilon^{\frac{1-\alpha}{2-\alpha}}\dot v_\varepsilon(\tau)\big\|_{\mathcal H}^2 \,d\tau = \varepsilon^{\frac{2(1-\alpha)}{2-\alpha}}\int_0^T \big\|\dot v_\varepsilon(\tau)\big\|_{\mathcal H}^2 \,d\tau \le L_T^{\frac{2}{2-\alpha}}T.$ □

Theorem 3.7. *Let $u:[0,+\infty)\to\mathcal H$ be the solution of the gradient flow equation (2.1) and $v_\varepsilon:[0,+\infty)\to\mathcal H$ the solution of the mini-batch descent flow equation (2.5). Then,*

$$\mathbb{E}\big\|v_\varepsilon(t)-u(t)\big\|_{\mathcal H}^2 \le 2\varepsilon\Big(\int_0^t \mathbb{E}\big\|\dot v_\varepsilon(\tau)\big\|_{\mathcal H}^2\,d\tau\Big)^{\frac12}\Big(\int_0^t \Lambda\big(u(s)\big)\,\mathrm{d}s\Big)^{\frac12} \quad \forall t\in[0,+\infty).$$

Proof. Let $k\in\mathbb{N}$. By monotonicity of the subdifferential,

$$\langle -\dot v_\varepsilon(\tau) - \xi_{j_k}(u(\tau)),\, v_\varepsilon(\tau)-u(\tau)\rangle \ge 0 \quad \text{for a.e. } \tau\in(t_{k-1},t_k).$$

Using that $\dot u = -\partial\Phi(u)^\circ$ a.e. in $[0,+\infty)$, this can be rewritten as

$$0 \le \langle -\dot v_\varepsilon(\tau)+\dot u(\tau),\, v_\varepsilon(\tau)-u(\tau)\rangle + \langle \partial\Phi(u(\tau))^\circ - \xi_{j_k}(u(\tau)),\, v_\varepsilon(\tau)-u(\tau)\rangle\,,$$

for a.e. $\tau\in(t_{k-1},t_k)$. This implies that for a.e. $\tau\in(t_{k-1},t_k)$,

$$\frac12\frac{\mathrm{d}}{\mathrm{d}\tau}\big\|v_\varepsilon(\tau)-u(\tau)\big\|_{\mathcal H}^2 \le \langle \partial\Phi(u(\tau))^\circ - \xi_{j_k}(u(\tau)),\, v_\varepsilon(\tau)-u(\tau)\rangle\,. \tag{3.4}$$

Let $\chi:[0,+\infty)\to\mathcal H$ be given by $\chi(t)=\partial\Phi(u(t))^\circ - \xi_{j_k}(u(t))$. Since, $\mathbb{E}\chi(\tau)=0$ for all $\tau\in(t_{k-1},t_k)$,

$$\mathbb{E}\left[\langle\chi(\tau),\, v_\varepsilon(t_{k-1})-u(\tau)\rangle\right] = \langle \mathbb{E}\chi(\tau), \mathbb{E}v_\varepsilon(t_{k-1})-u(\tau)\rangle = 0 \quad \forall\tau\in(t_{k-1},t_k).$$

Combining this with (3.4), we get, for a.e. $\tau\in(t_{k-1},t_k)$,

$$\frac12\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}\big\|v_\varepsilon(\tau)-u(\tau)\big\|_{\mathcal H}^2 \le \mathbb{E}\left[\langle\chi(\tau),\, v_\varepsilon(\tau)-v_\varepsilon(t_{k-1})\rangle\right] \le \sqrt{\mathbb{E}\big\|\chi(\tau)\big\|_{\mathcal H}^2}\sqrt{\mathbb{E}\big\|v_\varepsilon(\tau)-v_\varepsilon(t_{k-1})\big\|_{\mathcal H}^2}. \tag{3.5}$$

Observe that $\mathbb{E}\big\|\chi(t)\big\|_{\mathcal H}^2 = \Lambda\big(u(t)\big)$ for all $t\in[0,+\infty)$. Now, since v_ε is locally absolutely continuous in (t_{k-1},t_k), so is $t\longmapsto \big\|v_\varepsilon(t)-u(t)\big\|_{\mathcal H}$; thus, integrating yields

$$\mathbb{E}\big\|v_\varepsilon(t)-u(t)\big\|_{\mathcal H}^2 \le \mathbb{E}\big\|v_\varepsilon(s)-u(s)\big\|_{\mathcal H}^2 + 2\int_s^t \sqrt{\Lambda\big(u(\tau)\big)}\sqrt{\mathbb{E}\big\|v_\varepsilon(\tau)-v_\varepsilon(t_{k-1})\big\|_{\mathcal H}^2}\,\mathrm{d}\tau \quad \forall s,t\in(t_{k-1},t_k).$$

Hence, by Hölder's inequality, for all $t \in [t_{k-1}, t_k]$,

$$\mathbb{E}\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \leq \mathbb{E}\big\|v_\varepsilon(t_{k-1}) - u(t_{k-1})\big\|_{\mathcal{H}}^2 + 2\Big(\int_{t_k}^t \Lambda\big(u(\tau)\big)\,\mathrm{d}\tau\Big)^{\frac{1}{2}}\Big(\int_{t_k}^t \mathbb{E}\big\|v_\varepsilon(\tau) - v_\varepsilon(t_{k-1})\big\|_{\mathcal{H}}^2\,\mathrm{d}\tau\Big)^{\frac{1}{2}}. \tag{3.6}$$

For each $l \in \mathbb{N}$, let $h_l : [t_{l-1}, t_l] \to \mathbb{R}$ be given by $h_l(t) := \mathbb{E}\big\|v_\varepsilon(t) - v_\varepsilon(t_{l-1})\big\|_{\mathcal{H}}^2$. Since $k \in \mathbb{N}$ was arbitrary, by (3.6) and Hölder's inequality,

$$\begin{aligned}
\mathbb{E}\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 &\leq 2\Big(\int_{t_{k_t-1}}^t \Lambda\circ u\Big)^{\frac{1}{2}}\Big(\int_{t_{k_t-1}}^t h_{k_t}\Big)^{\frac{1}{2}} + 2\sum_{l=1}^{k_t-1}\Big(\int_{t_{l-1}}^{t_l}\Lambda\circ u\Big)^{\frac{1}{2}}\Big(\int_{t_{l-1}}^{t_l} h_l\Big)^{\frac{1}{2}} \\
&\leq 2\Big(\int_{t_{k_t-1}}^t \Lambda\circ u + \sum_{l=1}^{k_t-1}\int_{t_{l-1}}^{t_l}\Lambda\circ u\Big)^{\frac{1}{2}}\Big(\int_{t_{k_t-1}}^t h_{k_t} + \sum_{l=1}^{k_t-1}\int_{t_{l-1}}^{t_l} h_l\Big)^{\frac{1}{2}} \\
&= 2\Big(\int_0^t \Lambda\circ u\Big)^{\frac{1}{2}}\Big(\int_{t_{k_t-1}}^t h_{k_t} + \sum_{l=1}^{k_t-1}\int_{t_{l-1}}^{t_l} h_l\Big)^{\frac{1}{2}},
\end{aligned} \tag{3.7}$$

for every $t \in [0, +\infty)$. Now observe that, for each $l \in \mathbb{N}$

$$h_l(t) \leq \mathbb{E}\Big(\int_{t_{l-1}}^t \big\|\dot v_\varepsilon\big\|_{\mathcal{H}}\Big)^2 \leq (t - t_l)\,\mathbb{E}\int_{t_{l-1}}^t \big\|\dot v_\varepsilon\big\|_{\mathcal{H}}^2 \leq \varepsilon\int_{l-1}^t \mathbb{E}\big\|\dot v_\varepsilon\big\|_{\mathcal{H}}^2\,\mathrm{d}\tau \quad \forall t \in [t_{l-1}, t_l).$$

From this, we get that for every $t \in [0, +\infty)$,

$$\begin{aligned}
\int_{t_{k_t-1}}^t h_{k_t} + \sum_{l=1}^{k_t-1}\int_{t_{l-1}}^{t_l} h_l &\leq \varepsilon\Big[\int_{t_{k_t-1}}^t\Big(\int_{t_{k_t-1}}^s \mathbb{E}\big\|\dot v_\varepsilon\big\|_{\mathcal{H}}^2\Big)\,\mathrm{d}s + \sum_{l=1}^{k_t-1}\int_{t_{l-1}}^{t_l}\Big(\int_{t_{l-1}}^s \mathbb{E}\big\|\dot v_\varepsilon\big\|_{\mathcal{H}}^2\Big)\,\mathrm{d}s\Big] \\
&\leq \varepsilon\Big[\int_{t_{k_t-1}}^t\Big(\int_{t_{k_t-1}}^t \mathbb{E}\big\|\dot v_\varepsilon\big\|_{\mathcal{H}}^2\Big)\,\mathrm{d}s + \sum_{l=1}^{k_t-1}\int_{t_{l-1}}^{t_l}\Big(\int_{t_{l-1}}^{t_l} \mathbb{E}\big\|\dot v_\varepsilon\big\|_{\mathcal{H}}^2\Big)\,\mathrm{d}s\Big] \\
&\leq \varepsilon\Big[(t - t_{k_t-1})\int_{t_{k_t-1}}^t \mathbb{E}\big\|\dot v_\varepsilon\big\|_{\mathcal{H}}^2 + \varepsilon\sum_{l=1}^{k_t}\int_{t_{l-1}}^{t_l}\mathbb{E}\big\|\dot v_\varepsilon\big\|_{\mathcal{H}}^2\Big] \\
&\leq \varepsilon^2\int_0^t \mathbb{E}\big\|\dot v_\varepsilon\big\|_{\mathcal{H}}^2.
\end{aligned}$$

Combining this with (3.7) yields the result. □

Corollary 3.8. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ the solution of the mini-batch descent flow equation (2.5). Then, for any $\alpha \in [0, 1]$,*

$$\mathbb{E}\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \leq 2\varepsilon^{\frac{1}{2-\alpha}}\Big(\int_0^t \mathbb{E}\big\|\varepsilon^{\frac{1-\alpha}{2-\alpha}}\dot v_\varepsilon(\tau)\big\|_{\mathcal{H}}^2\Big)^{\frac{1}{2}}\Big(\int_0^t \Lambda\big(u(s)\big)\,\mathrm{d}s\Big)^{\frac{1}{2}} \quad \forall t \in [0, +\infty).$$

From previous results, it is possible to say something about the expected asymptotic behavior of the mini-batch descent solution.

Theorem 3.9. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ the solution of the mini-batch descent flow equation (2.5). Suppose that $\Lambda \circ u : [0, +\infty) \to \mathcal{H}$ is locally integrable*

and that there exists $\alpha \in (0,1]$ such that, for each $j \in \{1,\dots,m\}$, $\Phi_{\mathcal{B}_j}$ is locally α-Hölder continuous in its effective domain. Then the following statements hold.

(i) Suppose $\operatorname{dom}\partial\Phi_{\mathcal{B}_j} \subset \operatorname{dom}\Phi$ for all $j \in \{1,\dots,m\}$. Then, for every $\eta > 0$ there exists $T > 0$ such that

$$\mathbb{E}\Phi(v_\varepsilon(T)) \leq \inf_{v\in\mathcal{H}} \Phi(v) + \eta \quad \text{for all } \varepsilon > 0 \text{ small enough.}$$

(ii) Suppose $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is inf-compact. Then there exists $u^* \in \operatorname{argmin}_{v\in\mathcal{H}} \Phi(v)$ such that for every $\eta > 0$ there exists $T > 0$ such that

$$\mathbb{E}\big\|v_\varepsilon(T) - u^*\big\|_{\mathcal{H}}^2 \leq \eta \quad \text{for all } \varepsilon > 0 \text{ small enough.}$$

Proof. Let $\eta > 0$ be given.

(i) By [17], Proposition 17.2.7, there exists $T > 0$ such that $\Phi(u(T)) \leq \inf_{v\in\mathcal{H}} \Phi(v) + \eta/2$. Since by Corollary 3.5, $\{v_\varepsilon(T)\}_{\varepsilon>0}$ is bounded, there exists $L_T > 0$ such that

$$|\Phi(v_\varepsilon(T)) - \Phi(u(T))| \leq L_T \big\|v_\varepsilon(T) - u(T)\big\|_{\mathcal{H}}^{\alpha} \quad \forall \varepsilon > 0. \tag{3.8}$$

By Lemma 3.6 and Corollary 3.8, there exists $c_T > 0$ such that $\mathbb{E}\big\|v_\varepsilon(T) - u(T)\big\|_{\mathcal{H}}^2 \leq c_T \varepsilon^{\frac{1}{2-\alpha}}$ for all $\varepsilon > 0$. By Hölder's inequality, for all $\varepsilon > 0$,

$$\mathbb{E}\big\|v_\varepsilon(T) - u(T)\big\|_{\mathcal{H}}^{\alpha} \leq \Big(\mathbb{E}\big\|v_\varepsilon(T) - u(T)\big\|_{\mathcal{H}}^2\Big)^{\frac{\alpha}{2}} \leq c_T^{\frac{\alpha}{2}} \varepsilon^{\frac{\alpha}{2(2-\alpha)}}. \tag{3.9}$$

Finally, from (3.8) and (3.9),

$$\mathbb{E}\Phi(v_\varepsilon(T)) \leq \Phi(u(T)) + \mathbb{E}|\Phi(v_\varepsilon(T)) - \Phi(u(T))| \leq \inf_{v\in\mathcal{H}} \Phi(v) + \frac{\eta}{2} + L_T c_T^{\frac{\alpha}{2}} \varepsilon^{\frac{\alpha}{2(2-\alpha)}},$$

for all $\varepsilon > 0$. It is then enough to take ε such that $L_T c_T^{\frac{\alpha}{2}} \varepsilon^{\frac{\alpha}{2(2-\alpha)}} < \eta/2$.

(ii) As Φ is inf-compact and lower semicontinuous, $\operatorname{argmin}_{v\in\mathcal{H}} \Phi \neq \emptyset$. By [17], Corollary 17.2.1, there exists $u^* \in \operatorname{argmin}_{v\in\mathcal{H}} \Phi$ such that $u(t) \rightharpoonup u^*$ weakly in $\mathcal{H}$ as $t \to +\infty$. Since on the bounded subsets of the lower level sets of Φ, weak and strong convergence coincide, $u(t) \to u^*$ strongly in $\mathcal{H}$ as $t \to +\infty$. Therefore, there exists $T > 0$ such that $\big\|u(T) - u^*\big\|_{\mathcal{H}} \leq \sqrt{\eta}/2$. By Lemma 3.6 and Corollary 3.8, there exists $\varepsilon_0 > 0$ such that $\mathbb{E}\big\|v_\varepsilon(T) - u(T)\big\|_{\mathcal{H}}^2 \leq \eta/4$ for all $\varepsilon \in (0, \varepsilon_0)$. Therefore,

$$\mathbb{E}\big\|v_\varepsilon(T) - u^*\big\|_{\mathcal{H}}^2 \leq 2\mathbb{E}\big\|v_\varepsilon(T) - u(T)\big\|_{\mathcal{H}}^2 + 2\mathbb{E}\big\|u(T) - u^*\big\|_{\mathcal{H}}^2 \leq \eta \quad \forall \varepsilon \in (0, \varepsilon_0).$$

□

3.3. Random minimizing movement

It follows from [17], Theorem 17.2.2 that the right derivative $\mathrm{d}^+u/\mathrm{d}t$ of the solution of gradient flow equation (2.1) exists everywhere. For each $k \in \mathbb{N}$, set

$$\omega_k^+ := \Big\| \frac{u(t_k) - u(t_{k-1})}{\varepsilon} - \frac{\mathrm{d}^+u}{\mathrm{d}t}(t_k) \Big\|_{\mathcal{H}}^2.$$

Observe that the sequence $\{\omega_k^+\}_{k\in\mathbb{N}}$ remains bounded by $4\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}}^2$.

Lemma 3.10. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1). Then,*

$$\mathbb{E}\Big\| \frac{u(t_k) - u(t_{k-1})}{\varepsilon} + \xi_{j_k}(u(t_k))\Big\|_{\mathcal{H}}^2 \leq 2\omega_k^+ + 2\Lambda\big(u(t_k)\big) \quad \forall k \in \mathbb{N}.$$

Proof. Since, by [17], Theorem 17.2.2, $\mathrm{d}^+u/\mathrm{d}t = \partial\Phi(u)^\circ$ everywhere in $[0, +\infty)$; for all $k \in \mathbb{N}$,

$$\Big\| \frac{u(t_k) - u(t_{k-1})}{\varepsilon} + \xi_{j_k}(u(t_k))\Big\|_{\mathcal{H}}^2 \leq 2\Big\| \frac{u(t_k) - u(t_{k-1})}{\varepsilon} - \frac{\mathrm{d}^+u}{\mathrm{d}t}(t_k)\Big\|_{\mathcal{H}}^2 + 2\Big\| - \partial\Phi(u(t_k))^\circ + \xi_{j_k}(u(t_k))\Big\|_{\mathcal{H}}^2.$$

The result follows taking expectation on both sides. □

Lemma 3.11. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $\{w_k\}_{k\in\mathbb{N}}$ the proximal sequence (2.9). Then, for all $k \in \mathbb{N}$,*

$$\mathbb{E}\big\|w_k - u(t_k)\big\|_{\mathcal{H}}^2 \leq 2\varepsilon\Big(\sum_{l=1}^{k} \varepsilon\omega_k^+ + \sum_{l=1}^{k} \varepsilon\Lambda\big(u(t_l)\big) + \sum_{l=1}^{k} \sqrt{\omega_l^+}\, \sqrt{\mathbb{E}\big\|w_{l-1} - u(t_{k-1})\big\|_{\mathcal{H}}^2}\Big) \quad \forall k \in \mathbb{N}.$$

Proof. It is known that proximal sequence (2.9) satisfies the differential inclusions

$$-\frac{w_l - w_{l-1}}{\varepsilon} \in \partial\Phi_{\mathcal{B}_{j_l}}(w_l) \quad \forall l \in \mathbb{N},$$

where $w_0 := u_0$. For each $l \in \mathbb{N}$, define $f_l := -\Big[\varepsilon^{-1}\big(u(t_l) - u(t_{l-1})\big) + \xi_{j_l}(u(t_l))\Big]$. Observe that

$$-f_l - \frac{u(t_l) - u(t_{l-1})}{\varepsilon} \in \partial\Phi_{\mathcal{B}_{j_l}}(u(t_l)) \quad \forall l \in \mathbb{N}.$$

By monotonicity of the subdifferential,

$$\Big\langle -\frac{w_l - w_{l-1}}{\varepsilon} + f_l + \frac{u(t_l) - u(t_{l-1})}{\varepsilon}, w_l - u(t_l)\Big\rangle \geq 0 \quad \forall l \in \mathbb{N}. \tag{3.10}$$

Set $\mathcal{E}_0 := 0$, and for each $l \in \mathbb{N}$, denote $\mathcal{E}_l := w_l - u(t_l)$. Then (3.10) simplifies to $\langle \mathcal{E}_l - \mathcal{E}_{l-1}, \mathcal{E}_l\rangle \leq \varepsilon\langle f_l, \mathcal{E}_l\rangle$ for all $l \in \mathbb{N}$. Using the identity

$$\|a\|_{\mathcal{H}}^2 - \|b\|_{\mathcal{H}}^2 + \|a - b\|_{\mathcal{H}}^2 = 2\langle a - b, a\rangle \quad \forall a, b \in \mathcal{H},$$

we get $\|\mathcal{E}_l\|_{\mathcal{H}}^2 - \|\mathcal{E}_{l-1}\|_{\mathcal{H}}^2 + \|\mathcal{E}_l - \mathcal{E}_{l-1}\|_{\mathcal{H}}^2 \leq 2\varepsilon\langle \mathcal{E}_l, f_l\rangle$ for all $l \in \mathbb{N}$, and hence

$$\|\mathcal{E}_l\|_{\mathcal{H}}^2 - \|\mathcal{E}_{l-1}\|_{\mathcal{H}}^2 + \|\mathcal{E}_l - \mathcal{E}_{l-1}\|_{\mathcal{H}}^2 \leq 2\varepsilon\langle \mathcal{E}_l, f_l\rangle \leq 2\varepsilon\langle \mathcal{E}_l - \mathcal{E}_{l-1}, f_l\rangle + 2\varepsilon\langle \mathcal{E}_{l-1}, f_l\rangle \quad \forall l \in \mathbb{N}.$$

Employing Fenchel-Young inequality yields

$$\|\mathcal{E}_l\|_{\mathcal{H}}^2 - \|\mathcal{E}_{l-1}\|_{\mathcal{H}}^2 + \|\mathcal{E}_l - \mathcal{E}_{l-1}\|_{\mathcal{H}}^2 \leq 2\Big(\frac{\|\mathcal{E}_l - \mathcal{E}_{l-1}\|_{\mathcal{H}}^2}{2} + \frac{\varepsilon^2\|f_l\|_{\mathcal{H}}^2}{2}\Big) + 2\varepsilon\langle \mathcal{E}_{l-1}, f_l\rangle,$$

for all $l \in \mathbb{N}$. Hence,

$$\|\mathcal{E}_l\|_{\mathcal{H}}^2 - \|\mathcal{E}_{l-1}\|_{\mathcal{H}}^2 + \leq \varepsilon^2\|f_l\|_{\mathcal{H}}^2 + 2\varepsilon\langle \mathcal{E}_{l-1}, f_l\rangle \quad \forall l \in \mathbb{N}.$$

Summing up and taking expectation,

$$\mathbb{E}\big\|\mathcal{E}_k\big\|_{\mathcal{H}}^2 \leq \varepsilon^2 \sum_{l=1}^{k} \mathbb{E}\big\|f_k\big\|_{\mathcal{H}}^2 + 2\varepsilon \sum_{l=1}^{k} \mathbb{E}\langle \mathcal{E}_{k-1}, f_k\rangle \quad \forall k \in \mathbb{N}.$$

Now, observe that $\mathbb{E}\langle \mathcal{E}_{l-1}, f_l\rangle = \langle \mathbb{E}[\mathcal{E}_{l-1}], \mathbb{E}f_l\rangle = \langle \mathbb{E}[\mathcal{E}_{l-1}], \mathrm{d}^+u/\mathrm{d}t(t_l) - \varepsilon^{-1}\big(u(t_l) - u(t_{l-1})\big)\rangle$ for all $l \in \mathbb{N}$, and that, by Lemma 3.10, $\mathbb{E}\big\|f_l\big\|_{\mathcal{H}}^2 \leq 2\omega_l^+ + 2\Lambda\big(u(t_l)\big)$ for all $l \in \mathbb{N}$. Thus,

$$\mathbb{E}\big\|\mathcal{E}_k\big\|_{\mathcal{H}}^2 \leq 2\varepsilon^2\Big(\sum_{l=1}^{k} \omega_l^+ + \sum_{l=1}^{k} \Lambda\big(u(t_l)\big)\Big) + 2\varepsilon \sum_{l=1}^{k} \sqrt{\omega_l^+}\, \mathbb{E}\big\|\mathcal{E}_{l-1}\big\|_{\mathcal{H}} \quad \forall k \in \mathbb{N}.$$

Applying Cauchy-Schwartz inequality and rearranging yields the desired estimate. □

Lemma 3.12. *Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of the gradient flow equation (2.1) and $\{w_k\}_{k\in\mathbb{N}}$ the proximal sequence (2.9). Then,*

$$\mathbb{E}\big\|w_k - u(t_k)\big\|_{\mathcal{H}}^2 \leq 4\varepsilon\Big(2t_k \sum_{l=1}^{k} \omega_l^+ + \sum_{l=1}^{k} \varepsilon\Lambda\big(u(t_l)\big)\Big) \quad \forall k \in \mathbb{N}.$$

Proof. Let $k \in \mathbb{N}$ be given. For each $l \in \mathbb{N}$, denote $\mathcal{E}_l := w_l - u(t_l)$. By Lemma 3.11, for all $l \in \{1, \dots, k\}$,

$$\begin{aligned}
\mathbb{E}\big\|\mathcal{E}_l\big\|^2 &\leq 2\varepsilon\Big(\sum_{i=1}^{l} \varepsilon\omega_i^+ + \sum_{i=1}^{l} \varepsilon\Lambda\big(u(t_i)\big) + \sum_{i=1}^{l} \sqrt{\omega_i^+}\, \sqrt{\mathbb{E}\big\|\mathcal{E}_{i-1}\big\|_{\mathcal{H}}^2}\Big) \\
&\leq 2\varepsilon\Big(\sum_{i=1}^{k} \varepsilon\omega_i^+ + \sum_{i=1}^{k} \varepsilon\Lambda\big(u(t_i)\big) + \sum_{i=2}^{k} \sqrt{\omega_i^+}\, \sqrt{\mathbb{E}\big\|\mathcal{E}_{i-1}\big\|_{\mathcal{H}}^2}\Big) \\
&\leq 2\varepsilon\Big(\sum_{i=1}^{k} \varepsilon\omega_i^+ + \sum_{i=1}^{k} \varepsilon\Lambda\big(u(t_i)\big) + \Big[\max_{\{1,\dots,k\}} \mathbb{E}\big\|\mathcal{E}_i\big\|_{\mathcal{H}}^2\Big]^{\frac{1}{2}} \sum_{i=2}^{k} \sqrt{\omega_i^+}\Big).
\end{aligned}$$

Let $x_k := \Big[\max_{\{1,\dots,k\}} \mathbb{E}\big\|\mathcal{E}_i\big\|_{\mathcal{H}}^2\Big]^{\frac{1}{2}}$, $a_k := \varepsilon \sum_{i=2}^{k} \sqrt{\omega_i^+}$, and

$$b_k := \sqrt{2\varepsilon}\Big(\sum_{i=1}^{k} \varepsilon\omega_i^+ + 4\varepsilon \sum_{i=1}^{k} \varepsilon\Lambda\big(u(t_i)\big)\Big)^{\frac{1}{2}}.$$

Then, $x_k^2 \leq 2a_k x_k + b_k^2$; this implies $x_k \leq 2a_k + b_k$, that is,

$$\begin{aligned}
\max_{\{1,\dots,k\}} \mathbb{E}\big\|\mathcal{E}_i\big\|_{\mathcal{H}}^2 &\leq \Big(2\varepsilon \sum_{i=2}^{k} \sqrt{\omega_i^+} + \sqrt{2\varepsilon}\Big(\sum_{i=1}^{k} \varepsilon\omega_i^+ + 4\varepsilon \sum_{i=1}^{k} \varepsilon\Lambda\big(u(t_i)\big)\Big)^{\frac{1}{2}}\Big)^2 \\
&\leq 8\varepsilon^2\Big(\sum_{i=2}^{k} \sqrt{\omega_i^+}\Big)^2 + 4\varepsilon\Big(\sum_{i=1}^{k} \varepsilon\omega_i^+ + 4\varepsilon \sum_{i=1}^{k} \varepsilon\Lambda\big(u(t_i)\big)\Big).
\end{aligned}$$

Applying Cauchy-Schwartz inequality, and rearranging

$$\max_{\{1,\dots,k\}} \mathbb{E}\big\|\mathcal{E}_i\big\|_{\mathcal{H}}^2 \leq 4\varepsilon^2\Big(2(k-1)\sum_{i=2}^{k}\omega_i^+ + \sum_{i=1}^{k}\omega_i^+ + \sum_{i=1}^{k}\Lambda\big(u(t_i)\big)\Big) \leq 4\varepsilon^2\Big(2k\sum_{i=1}^{k}\omega_i^+ + \sum_{i=1}^{k}\Lambda\big(u(t_i)\big)\Big).$$

□

Theorem 3.13. *Let $u:[0,+\infty)\to\mathcal{H}$ be the solution of the gradient flow equation (2.1) and $w_\varepsilon:[0,+\infty)\to\mathcal{H}$ the function generated by the proximal sequence (2.9). Let $T>0$, and set $N:=\lfloor T/\varepsilon\rfloor$. Then,*

$$\mathbb{E}\big\|w_\varepsilon(t)-u(t)\big\|_{\mathcal{H}}^2 \leq 4\varepsilon\Big(2T\sum_{i=1}^{N}\omega_i^+ + \sum_{i=1}^{N}\varepsilon\Lambda\big(u(t_i)\big)\Big) + 2\varepsilon^2\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}}^2. \quad \forall t\in[0,T].$$

Proof. Let $k\in\mathbb{N}$. For $t\in[t_{k-1},t_k)$, $w_\varepsilon(t)=w_k$, and hence

$$\big\|w_\varepsilon(t)-u(t)\big\|_{\mathcal{H}}^2 \leq 2\big\|w_k-u(t_k)\big\|_{\mathcal{H}}^2 + 2\big\|u(t_k)-u(t)\big\|_{\mathcal{H}}^2 \leq 2\big\|w_k-u(t_k)\big\|_{\mathcal{H}}^2 + 2\varepsilon^2\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}}^2.$$

The result follows then from Lemma 3.12. □

Corollary 3.14. *Let $u:[0,+\infty)\to\mathcal{H}$ be the solution of the gradient flow equation (2.1) and $w_\varepsilon:[0,+\infty)\to\mathcal{H}$ the function generated by the proximal sequence (2.9). Let $T>0$. Assume that $\Lambda\circ u$ is bounded in $[0,T]$. If u belongs to $C^1\big([0,T];\mathcal{H}\big)$, then*

$$\sup_{t\in[0,T]}\mathbb{E}\big\|w_\varepsilon(t)-u(t)\big\|_{\mathcal{H}}^2 \longrightarrow 0 \quad as \quad \varepsilon\longrightarrow 0^+.$$

Moreover, if there exists $\alpha\in(0,1]$ such that $\dot u$ is α-Hölder continuous in $[0,T]$, then there exists $c_T>0$ such that

$$\sup_{t\in[0,T]}\mathbb{E}\big\|w_\varepsilon(t)-u(t)\big\|_{\mathcal{H}}^2 \leq c_T\big(\varepsilon^{2\alpha}+\varepsilon^2+\varepsilon\big) \quad \forall\varepsilon>0.$$

Proof. Since $\dot u$ is continuous in the compact set $[0,T]$, it is uniformly continuous over it; hence there exists $\eta:[0,T]\to[0,+\infty)$ such that

$$\big\|\dot u(t)-\dot u(s)\big\|_{\mathcal{H}} \leq \eta\big(|t-s|\big) \quad \forall s,t\in[0,T].$$

Set $N:=\lfloor T/\varepsilon\rfloor$, for all $k\in\{1,\dots,N\}$,

$$\sqrt{\omega_k^+} = \Big\|\frac{u(t_k)-u(t_{k-1})}{\varepsilon} - \frac{\mathrm{d}^+u}{\mathrm{d}t}(t_k)\Big\|_{\mathcal{H}} = \Big\|\frac{1}{\varepsilon}\int_{t_{k-1}}^{t_k}\big(\dot u(s)-\dot u(t_k)\big)\,\mathrm{d}s\Big\|_{\mathcal{H}} \leq \frac{1}{\varepsilon}\int_{t_{k-1}}^{t_k}\big\|\dot u(s)-\dot u(t_k)\big\|_{\mathcal{H}}\,\mathrm{d}s.$$

This implies $\omega_k^+\leq\eta(\varepsilon)^2$ for all $k\in\{1,\dots,N\}$. Now, from Theorem 3.13,

$$\mathbb{E}\big\|w_\varepsilon(t)-u(t)\big\|_{\mathcal{H}}^2 \leq 8T^2\eta(\varepsilon)^2 + \varepsilon\sum_{i=1}^{N}\varepsilon\Lambda\big(u(t_i)\big) + 2\varepsilon^2\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}}^2 \quad \forall t\in[0,T].$$

Let $M := \sup_{[0,T]} |\Lambda(u(t))|$. Then,

$$\sup_{t\in[0,T]} \mathbb{E}\big\|w_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \leq 8T^2\eta(\varepsilon)^2 + M\varepsilon + 2\varepsilon^2\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}}^2. \tag{3.11}$$

Letting $\varepsilon \to 0^+$, yields the first part the result. Now, if $\dot{u}$ is α-Hölder continuous, the modulus of continuity of $\dot{u}$ over $[0,T]$ can be taken as $\eta(t) = L_T t^\alpha$. Then, (3.11) becomes

$$\sup_{t\in[0,T]} \mathbb{E}\big\|w_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \leq 8T^2\varepsilon^{2\alpha} + M\varepsilon + 2\varepsilon^2\big\|\partial\Phi(u_0)^\circ\big\|_{\mathcal{H}}^2.$$

□

4. Applications to optimization and partial differential equations

In this section, we will show how our main results can be applied to problems related to optimization and partial differential equations. We will also provide some numerical examples to illustrate our results.

4.1. Constrained optimization

Let $\mathcal{H}$ be a Hilbert space and C a closed convex bounded subset of $\mathcal{H}$. Let $\Psi_1, \dots, \Psi_n : \mathcal{H} \to \mathbb{R}$ be convex continuously differentiable functions. Let $\pi_1, \dots, \pi_m$ be positive numbers such that $\pi_1 + \cdots + \pi_m = 1$. Consider the average potential $\Psi : \mathcal{H} \to \mathbb{R}$ given by

$$\Psi(u) := \sum_{j=1}^m \pi_j \Psi_j(u).$$

In this subsection, we are interested in the following constrained optimization problem:

$$\min_{u\in C} \Psi(u). \tag{4.1}$$

It is well know that if $u^* \in C$ is a minimizer of problem (4.1), then $\langle \nabla\Psi(u^*), v\rangle \geq 0$ for all $v \in T_C(u^*)$, where $T_C : \mathcal{H} \twoheadrightarrow \mathcal{H}$ is the tangent cone mapping of C. From [17], Proposition 17.2.12, given an initial datum $u_0 \in C$, the gradient flow associated with the problem (4.1) is given by

$$\begin{cases} \dot{u}(t) = -\operatorname{proj}_{T_C(u(t))}\big(\nabla\Psi(u(t))\big), \\ u(0) = u_0. \end{cases} \tag{4.2}$$

Following the procedure described in Section 2, it is possible to construct a continuous function, depending on a parameter $\varepsilon > 0$, $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ such that $v_\varepsilon(0) = u_0$, and for each $k \in \mathbb{N}$,

$$0 = \dot{v}_\varepsilon(t) + \operatorname*{proj}_{T_C(u(t))}\big(\nabla\Psi_{j_k}(v_\varepsilon(t))\big) \quad \text{for a.e. } t \in [(k-1)\varepsilon, k\varepsilon). \tag{4.3}$$

Here, $\{j_l\}_{k\in\mathbb{N}}$ is a sequence of random variables taking value $j \in \{1, \dots, m\}$ with probability π_j.

4.1.1. Convergence and asymptotic behavior

In order to quantify the variance induced by the replacement of gradients over time, consider the function $\Gamma : \mathcal{H} \to \mathbb{R}$ given by

$$\Gamma(u) := \sum_{j=1}^{m} \pi_j \left\| \nabla\Psi_j(u) - \nabla\Psi(u) \right\|_{\mathcal{H}}^2.$$

Observe that $\mathbb{E}\left[\nabla\Psi_{j_k}\right] = \nabla\Psi$ and $\operatorname{Var}\left[\nabla\Psi_{j_k}\right] = \Gamma$ for all $k \in \mathbb{N}$. Also, since Γ is continuous, $\Gamma \circ u : [0, +\infty) \to \mathbb{R}$ is locally bounded.

Theorem 4.1. *There exists a unique locally absolutely continuous function $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ satisfying $v_\varepsilon(0) = u_0$ and (4.3). Moreover, v_ε is locally Lipschitz, and the following statements hold.*

(i) For every $T > 0$ there exists $c_T > 0$ such that

$$\sup_{t\in[0,T]} \mathbb{E}\left\| v_\varepsilon(t) - u(t) \right\|_{\mathcal{H}}^2 \leq c_T \varepsilon \int_0^T \Gamma\big(u(s)\big)\, \mathrm{d}s \quad \forall \varepsilon > 0.$$

(ii) For every $\eta > 0$ there exists $T > 0$ such that

$$\mathbb{E}\Phi(v_\varepsilon(T)) \leq \inf_{v\in\mathcal{H}} \Phi(v) + \eta \quad \textit{for all } \varepsilon > 0 \textit{ small enough.}$$

(iii) There exists $u^* \in \operatorname{argmin}_{v\in C} \Psi(v)$ such that for every $\eta > 0$ there exists $T > 0$ satisfying

$$\mathbb{E}\left\| v_\varepsilon(T) - u^* \right\|_{\mathcal{H}}^2 \leq \eta \quad \textit{for all } \varepsilon > 0 \textit{ small enough.}$$

Proof. For each $j \in \{1, \dots, m\}$, define $\Phi_j := \Psi_j + \delta_C$, where $\delta_C : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is the indicator function of C. By Moreau–Rockafellar subdifferential additivity rule, $\partial\Phi_j = \nabla\Psi_j + N_C(u)$ for each $j \in \{1, \dots, m\}$. Let $\Phi := \Psi + \delta_C$; we see that $\partial\Phi = \sum_{j=1}^m \pi_j \partial\Phi_j$. For each $u \in C$, denote by $\eta^*(u)$ the unique element in $N_C(u)$ such that $\partial\Phi(u)^\circ = \nabla\Psi(u) + \eta^*(u)$. For each $j \in \{1, \dots, m\}$, define $\xi_j : C \to \mathcal{H}$ by $\xi_j(u) := \nabla\Psi_j(u) + \eta^*(u)$. We see that $\sum_{j=1}^m \pi_j \xi_j(u) = \nabla\Psi(u) + \eta^*(u) = \partial\Phi(u)^\circ$ for all $u \in C$. Consider now, the function $\Lambda : C \to \mathbb{R}$ in (2.8) based on the previous decomposition of the minimal norm subdifferential; we see then that

$$\Lambda(u) = \sum_{j=1}^{m} \pi_j \left\| \xi_j(u) - \partial\Phi(u)^\circ \right\|_{\mathcal{H}}^2 = \sum_{j=1}^{m} \pi_j \left\| \nabla\Psi_j(u) - \nabla\Psi(u) \right\|_{\mathcal{H}}^2 = \Gamma(u).$$

Now, observe that since, for each $j \in \{1, \dots, m\}$, Ψ_j is locally Lipschitz, so is Φ_j over C. We can then employ Theorems 2.1 and 2.4 to conclude the result. □

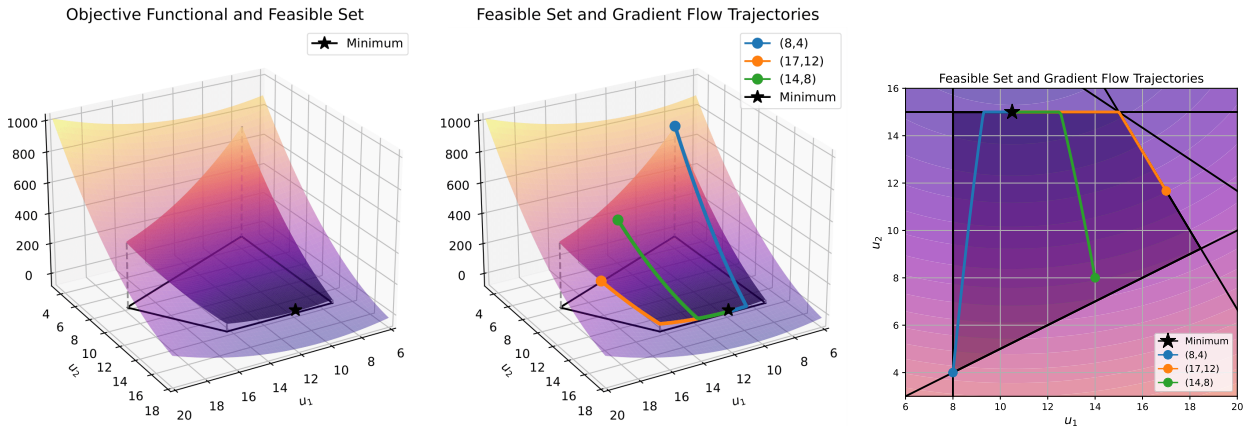
4.1.2. Illustrative numerical example

To illustrate Theorem 4.1, let $u = (u^1, u^2) \in \mathbb{R}^2$ and $u_d = (u_d^1, u_d^2)$, we consider the quadratic programming problem given by

$$\min_{u\in\mathcal{C}} \left\{ \Psi(u) = \left(2|u^1 - u_d^1| + 3|u^2 - y_d|^2 - 2u^1 - 3u^2 \right) \right\}, \tag{4.4}$$

where the feasible set $\mathcal{C}$ is a convex set, defined by the $u = (u^1, u^2)$ such that

$$\mathcal{C} = \{ u = (u^1, u^2) \in \mathbb{R}^2 \,:\, 5u^1 + 3u^2 \leq 120,\ 4u^1 + 6u^2 \leq 150,\ u^1 - 2u^2 \leq 0,\ u^1 \geq 7,\ u^2 \leq 15 \}.$$



(A) (B) (C)

FIGURE 1. (A) Illustration of the cost function for $(u_1, u_2) \in [6, 10] \times [4, 18]$. We can also observe the feasible set (and its projection onto $\mathbb{R}^2$), illustrated with a darker color. The star denotes the minimum of the functional in the feasible set. (B) Trajectories of the gradient flow for different initial points. (C) Projection of the feasible set onto the plane, along with the trajectories of the gradient flow. The contour lines of the functional are also illustrated.

Now, we consider the gradient flow associated with (4.4), which corresponds to the system (4.2). To implement it, we consider the dynamic in a fixed time interval $[0, T]$ with $T = 10$ and $h = 0.01$ as time steps. We fix $u_d = (10, 20)$.We look at three initial points: $(8, 4)$, $(13, 8)$, and $(20, 14)$, which are at a corner, inside, and on the boundary of the feasible set, respectively. We use an explicit Euler method to compute the gradient flow. The projection onto the feasible set $\mathcal{C}$ is performed through an iterative process in which, after each gradient descent step, any violated constraints are sequentially enforced by adjusting the solution in the direction opposite to the normal vectors of the feasible set boundary. The solution of this system is illustrated in Figure 1.

In Figure 1, we observe that starting the dynamics from point $(8, 4)$, placed at a corner, the dynamics evolve inside the feasible set until it reaches the boundary and then move to the minima of the functional Φ. A similar behavior is seen in the dynamics with initial point $(13, 8)$. For the point $(20, 14)$, which is placed on the boundary of the feasible set, the trajectory stays on the boundary until it reaches the minimum.

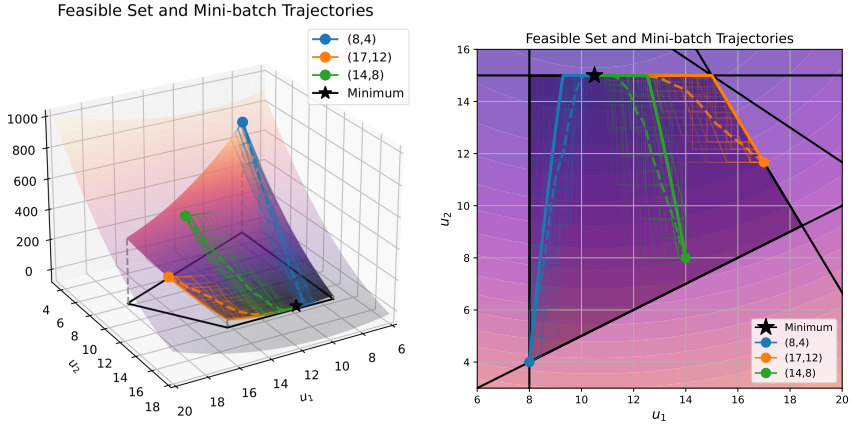
For each $j \in \{1, 2, 3\}$ let us consider the functionals Ψ_j given by

$$\Psi_1(u) = \pi_1 \Psi, \quad \Psi_2(u) = \pi_2(4|u^1 - u_d^1| - 4u^1), \quad \Psi_3(u) = \pi_3(6|u^2 - y_d|^2 - 6u^2),$$

where $\pi_1 = 1/2$ and $\pi_2 = \pi_3 = 1/4$. Observe that $\Phi(u) = \sum_{j=1}^{3} \pi_j \Phi_j(u)$. Then, choosing $\varepsilon = 0.04$, we can introduce the mini-batch gradient descend dynamics defined by the system (4.3) that evolves in the time interval $[0, T]$. To solve numerically this problem, we consider the same setting as the gradient flow.

In Figure 2, several realizations are considered to estimate the average mini-batch descent flow. We observe that the estimated average is close to the gradient flow trajectories. Additionally, due to the construction of the sub-functionals, we observe that Ψ_2 only depends on u^1 and Ψ_3 only depends on u^2. Therefore, when the second or third sub-functional is randomly chosen, the mini-batch trajectories move only in the u^1 or u^2 directions, respectively.

Finally, to corroborate Theorem 4.1, we denote by K the number of batch switching in the interval $[0, T]$, of the mini-batch flow, that is, $K = 1/\varepsilon$. Then, taking $K \to \infty$ (equivalent $\varepsilon \to 0$) we can observe the linear convergence (guaranteed by Thm. 4.1) in Figure 3.



(A) (B)

FIGURE 2. (A) Thin lines show different realizations of the mini-batch descent flow. The dashed line illustrates the average of the different realizations to approximate the average mini-batch descent flow. The solid line shows the previously calculated gradient flow. (B) We illustrate the projection onto the $\mathbb{R}^2$ plane of the trajectories mentioned in (A).

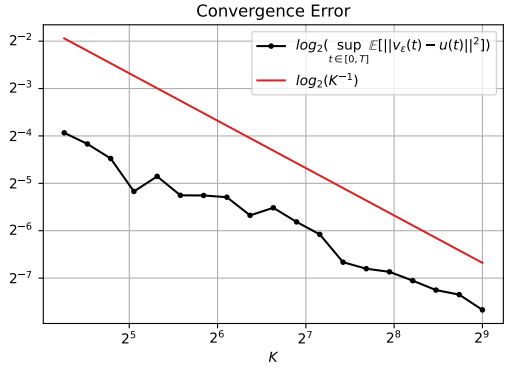


FIGURE 3. We illustrate the convergence as $K \to \infty$ (equivalent to $\varepsilon \to 0$) of the mini-batch descent flow.

4.2. Sparse inversion

Let r, d positive integers and $\lambda \geq 0$. Consider the matrices $A \in \mathbb{R}^{r\times d}$ and $b \in \mathbb{R}^r$. We are concerned with the following *sparse inversion problem*.

$$\min_{u\in\mathbb{R}^d} \left\{ \frac{1}{2}\|Au - b\|_2^2 + \lambda\|u\|_1 \right\}. \tag{4.5}$$

The idea behind problem (4.2) is to find a "sparse" $u \in \mathbb{R}^d$ such that $Au \approx b$.

Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be the objective functional associated with problem (4.2), that is,

$$\Phi(u) := \frac{1}{2}\|Au - b\|_2^2 + \lambda\|u\|_1.$$

We can readily see that $\partial\Phi(u) = A^\top\big(Au - b\big) + \Theta(u)$, where $\Theta : \mathbb{R}^d \twoheadrightarrow \mathbb{R}^d$ is the set-valued mapping given by

$$\Theta(u) := \left\{ \eta \in \mathbb{R}^d : \eta_i(x) \in \begin{cases} \{1\} & \text{if } u_i(x) > 0 \\ [-1,1] & \text{if } u_i(x) = 0 \\ \{-1\} & \text{if } u_i(x) < 0 \end{cases} \text{for all } i \in \{1,\dots,d\} \right\}.$$

Given an initial datum $u_0 \in \mathbb{R}^d$, the gradient flow associated with problem (4.2), from now on called *sparse inversion flow*, is given by

$$\begin{cases} -\dot{u}(t) \in A^\top\big(Au(t) - b\big) + \lambda\Theta(u(t)), \\ u(0) = u_0. \end{cases} \tag{4.6}$$

Following the procedure described in Section 2, it is possible to construct a continuous function, depending on a parameter $\varepsilon > 0$, $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ such that $v_\varepsilon(0) = u_0$, and for each $k \in \mathbb{N}$,

$$-\dot{v}_\varepsilon(t) = \frac{1}{\pi_1} A^\top\big(Av_\varepsilon(t) - b\big) \quad \text{if } j_k = 1, \tag{4.7}$$

$$-\dot{v}_\varepsilon(t) \in \frac{\lambda}{\pi_2}\Theta\big(v_\varepsilon(t)\big) \quad \text{if } j_k = 2. \tag{4.8}$$

for a.e. $t \in [(k-1)\varepsilon, k\varepsilon)$. Here, $\{j_l\}_{k\in\mathbb{N}}$ is a sequence of random variables taking values $j_k = 1$ with probability π_1, and value $j_k = 2$ with probability π_2. We see then that for $k \in \mathbb{N}$,

$$v_\varepsilon(t) = \begin{cases} \mathrm{e}^{-\pi_1^{-1}A^\top At} v_\varepsilon(t_{k-1}) + \pi_1^{-1} \displaystyle\int_{t_{k-1}}^{t} \mathrm{e}^{-\pi_1^{-1}A^\top A(t-s)} A^\top b \,\mathrm{d}s & \text{if } j_k = 1 \\ \big(\operatorname{sgn} v_\varepsilon^i(t_{k-1}) \max\{|v_\varepsilon^i(t_{k-1})| - \pi_2^{-1}\lambda t, 0\}\big)_{i=1}^d & \text{if } j_k = 2 \end{cases} \quad \forall t \in [t_{k-1}, t_k]. \tag{4.9}$$

At step $k \in \mathbb{N}$, if $j_k = 1$, there is a closed form solution of (4.7); on the other hand, if $j_k = 2$, the solution of (4.8) possesses a reasonable expression and can be computed by linearly reducing the vector components of the preceding step to zero.

4.2.1. Convergence and asymptotic behavior

In order to quantify the variance induced by replacing the gradients over a time interval, consider the function $\Gamma : \mathcal{H} \to \mathbb{R}$ given by

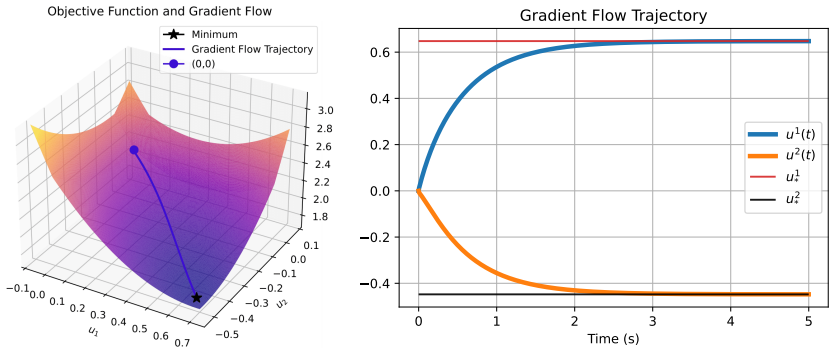
$$\Gamma(u) := \frac{\pi_2^2}{\pi_1}\big\|A^\top\big(Au - b\big)\big\|_2^2 + \frac{\pi_1^2}{\pi_2}(\lambda d)^2.$$

It is not hard to see that $\Gamma \circ u : [0, +\infty) \to \mathbb{R}$ is bounded.

Theorem 4.2. *There exists a unique locally absolutely continuous function $v_\varepsilon : [0, +\infty) \to \mathcal{H}$ satisfying $v_\varepsilon(0) = u_0$ and (4.7)-(4.8). Moreover, v_ε is locally Lipschitz, and the following statements hold.*

(i) For every $T > 0$ there exists $c_T > 0$ such that

$$\sup_{t\in[0,T]} \mathbb{E}\big\|v_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \le c_T\varepsilon \int_0^T \Gamma\big(u(s)\big)\,\mathrm{d}s \quad \forall \varepsilon > 0.$$



(A) (B)

FIGURE 4. (A) Illustration of the gradient descent trajectories defined by system (4.6). The black star denotes the minimum of Φ. (B) Each curve corresponds to a different coordinate of the trajectory. Horizontal lines mark the optimal of Φ.).

(ii) For every $\eta > 0$ there exists $T > 0$ such that

$$\mathbb{E}\Phi(v_\varepsilon(T)) \leq \inf_{v\in\mathcal{H}} \Phi(v) + \eta \quad \textit{for all } \varepsilon > 0 \textit{ small enough.}$$

(iii) There exists $u^* \in \operatorname{argmin}_{v\in\mathbb{R}^d} \Phi(v)$ such that for every $\eta > 0$ there exists $T > 0$ such that

$$\mathbb{E}\left\|v_\varepsilon(T) - u^*\right\|_{\mathcal{H}}^2 \leq \eta \quad \textit{for all } \varepsilon > 0 \textit{ small enough.}$$

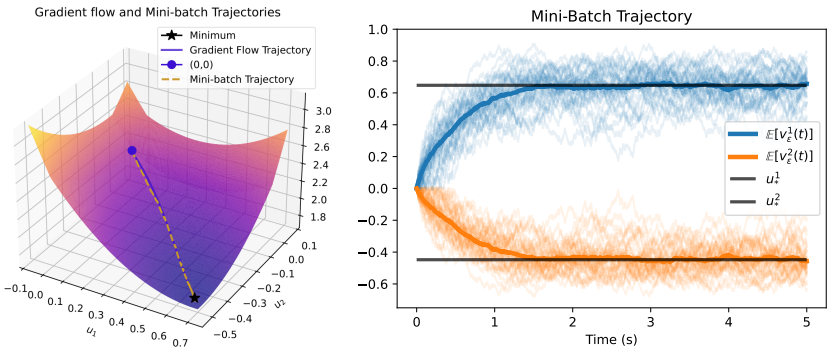
Proof. Define $\Phi_1 : \mathbb{R}^d \to \mathbb{R}$ by $\Phi_1(u) := (2\pi_1)^{-1}\left\|Au-b\right\|_2^2$ and $\Phi_2 : \mathbb{R}^d \to \mathbb{R}$ by $\Phi_2(u) := \pi_2^{-1}\lambda\left\|u\right\|_1$. By Moreau–Rockafellar subdifferential additivity rule, $\partial\Phi = \pi_1\partial\Phi_1 + \pi_2\partial\Phi_2$. For each $u \in \mathbb{R}^d$, denote by $\eta^*(u)$ the unique element in $\Theta(u)$ such that $\partial\Phi(u)^\circ = A^\top(Au - b) + \eta^*(u)$. Let $\xi_1 : \mathbb{R}^d \to \mathbb{R}$ be given by $\xi_1(u) = A^\top(Au - b)$, and $\xi_2 : \mathbb{R}^d \to \mathbb{R}$ by $\xi_2(u) = \eta^*(u)$. We see that $\pi_1\xi_j(u) + \pi_2\xi_2(u) = \partial\Phi(u)^\circ$ for all $u \in \mathbb{R}^d$. Consider now, the function $\Lambda : \mathbb{R}^d \to \mathbb{R}$ in (2.8) based on the previous decomposition of the minimal norm subdifferential. It is not hard to see that $\Lambda(u) \leq \Gamma(u)$ for all $u \in \mathbb{R}^d$. We can then employ Theorems 2.1 and 2.4 to conclude the result. □

4.2.2. Illustrative numerical example

To illustrate Theorem 4.2, consider the following numerical example. Let $A \in \mathbb{R}^{2\times 2}$ and $b \in \mathbb{R}^2$ $(d = r = 2)$ given by

$$A = \begin{pmatrix} 1.76 & 0.4 \\ 0.98 & 2.24 \end{pmatrix}, \quad b = \begin{pmatrix} 1.87 \\ -0.98 \end{pmatrix},$$

and we take $\lambda = 1$. Let us denote by $u_* = (u_*^1, u_*^2)$ the optimal of the sparse inversion problem (4.5). In this particular case $u_* = (0.65, -0.45)$. Now, let us consider the gradient flow (4.6), associated with the sparse inversion problem in the fixed time interval $[0, T]$. We fix the time horizon $T = 5$ and the initial condition $u_0 = (0, 0)$. For the implementation, we consider a time step $h = 0.01$, and using an explicit Euler discretization, we numerically solve (4.6). The trajectory computed is shown in Figure 4. We can observe in Figure 4 how the gradient flow converges to the minimum of the objective functional. The horizontal lines in Figure 4B denote the optimal value $u_* = (u_*^1, u_*^2)$.



(A) (B)

FIGURE 5. (A) Illustration of the expected value of the mini-batch descent flow trajectory. This trajectory star is from the same value of the gradient flow. (B) Illustration of the mini-batch descent flow. Thin curves represent different realizations for each coordinate, while thick lines indicate the average outcomes of the mini-batch descent flow.

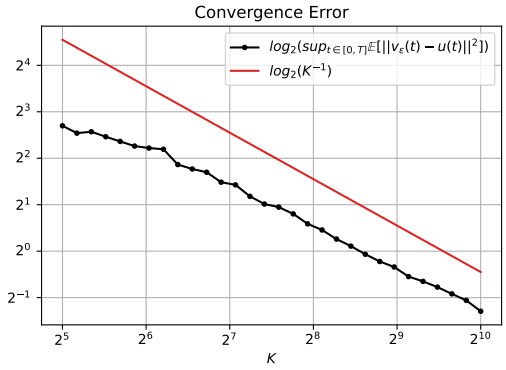


FIGURE 6. Illustration of the convergence as $K \to \infty$ (equivalent to $\varepsilon \to 0$.

On the other hand, for the mini-batch descent flow, we consider $\varepsilon = 0.04$. Then, using the same setting of the gradient flow, we compute the solution of (4.7)–(4.8). To approximate the expected trajectory of the mini-batch descent flow, we compute different realizations of the trajectory and then average them. In Figure 5, we observe different realizations of the mini-batch descent flow and their expected value.

To illustrate the rate of convergence guaranteed by Theorem 4.2, we take $\varepsilon = 1/K$ with K being the number of batch switching in $[0, T]$. Then, Figure 6 shows the convergence of the mini-batch descent flow to the gradient flow as $K \to \infty$ (equivalent $\varepsilon \to 0$).

In the previous example, we considered $d = r = 2$. However, it is possible to consider a larger matrix size and compare the computation time taken to solve the gradient flow versus the mini-batch descent flow. This is illustrated in Table 1, where the speedup corresponds to the ratio between the execution time of the gradient flow and the mini-batch descent flow.

Table 1 shows that as the matrix size increases, the mini-batch descent flow maintains a consistent speedup, demonstrating improved computational efficiency compared to the gradient flow.

TABLE 1. Computational Time Between Gradient and Mini-batch Algorithms. Here, we have considered $r = \lceil d/2 \rceil$. The values of the matrices are chosen randomly from a uniform distribution $U(0,1)$.

Matrix Size A ($d \times d$)	Gradient flow (s)	Mini-batch descent flow (s)	Speedup
5×5	0.005	0.004	1.289
50×50	0.006	0.005	1.248
100×100	0.008	0.006	1.316
200×200	0.017	0.014	1.236
400×400	0.040	0.033	1.206

4.3. Domain decomposition for the parabolic obstacle problem

In this section, we describe a random domain decomposition algorithm for parabolic-type equations that can be rewritten as gradient flows.

In order to make the presentation simpler, we focus on the particular case of the obstacle problems.

4.3.1. Problem formulation

Let $\Omega \subset \mathbb{R}^d$ be an open bounded set. Let $\psi \in H^2(\Omega)$ be a given function satisfying $\psi \leq 0$ a.e. on $\partial\Omega$, from now on called *the obstacle*, and let $K(\psi)$ be given by

$$K(\psi) = \left\{ u \in L^2(\Omega) \mid u(x) \geq \psi(x) \quad \text{for a.e. } x \in \Omega \right\}.$$

Note that $K \neq \emptyset$ because $\psi^+ = \max(\psi, 0) \in K$. Moreover, observe that K is a convex subset of $L^2(\Omega)$. Let $T > 0$ and consider the problem of find $u(t) \in K(\psi) \cap H_0^2(\Omega)$ for a.e. $t \in (0,T)$ such that for all $\phi \in K(\psi)$ the variational inequality

$$\begin{cases} \int_\Omega \dot{u}(\phi - u)\mathrm{d}x - \int_\Omega \Delta u(\phi - u) - \int_\Omega f(\phi - u)\mathrm{d}x \geq 0 & \text{a.e. } t \in (0,T), \\ u(x,0) = u_0(x) & \text{in } \Omega, \end{cases} \tag{4.10}$$

holds. This problem is known as the parabolic obstacle problem; in this case, we are considering a stationary obstacle. Concerning the well-posedness, we have the following result.

Theorem 4.3. *Let $u_0 \in K(\psi)$ and $f \in L^2(0,T;L^2(\Omega))$. Then, there exists a unique absolute continuous solution $u \in C([0,T];L^2(\Omega))$ of (4.10). Moreover, $u(t) \in \{\phi \in K(\psi) \cap H_0^1(\Omega) : \Delta\phi \in L^2(\Omega)\}$ for a.e. $t \in (0,T)$.*

Proof. Let us consider $\Phi : L^2(\Omega) \to \mathbb{R}$ defined as $\Phi = \Psi + \delta_{K(\psi)}$, where $\delta_{K(\psi)}$ is the indicator function of the closed convex $K(\psi)$, and

$$\Psi(u) := \begin{cases} \dfrac{1}{2}\displaystyle\int_\Omega |\nabla u(x)|^2 \,\mathrm{d}x - \int_\Omega f(x)u(x)\,\mathrm{d}x & \text{if } u \in H_0^1(\Omega), \\ +\infty & \text{if } u \in L^2(\Omega),\ u \notin H_0^1(\Omega). \end{cases}$$

Since Ψ is a convex, lower semicontinuous, and proper function, it defines a maximal monotone operator $\partial\Psi$. On the other hand, we have that $\partial\Psi + N_{K(\psi)} \subset \partial(\Psi + \delta_{K(\psi)}) = \partial\Phi$, where $N_{K(\psi)}$ denote the normal cone of the set $K(\psi)$. Moreover, since Minty's Theorem [5], Theorem 17.2.1 $\partial\Psi + N_{K(\psi)}$ is a maximal monotone operator, if only if, $R(I + \partial\Psi + \partial\delta_{K(\psi)}) = L^2(\Omega)$ which is equivalent to the existence of solution of the equation $u - \Delta u + N_{K(\psi)}(u) \ni f$, with $u = 0$ on $\partial\Omega$. The well-posedness is ensured by [18], Proposition 2.11. Consequently, by maximality, $\partial\Psi + N_{K(\psi)} = \partial\Phi$. Observe that dom $\partial\Phi = \{u \in K(\phi) \cap H_0^1(\Omega) : \Delta u \in L^2(\Omega)\}$.

Now let us observe that the gradient flow associated with Φ is given by

$$\begin{cases} 0 \in \dot{u} - \Delta u - f + N_{K(\psi)} & \text{in } \Omega \times (0,T), \\ u(x,t) = 0 & \text{on } \partial\Omega \times (0,T), \\ u(x,0) = u_0(x) & \text{in } \Omega. \end{cases} \tag{4.11}$$

Using the definition of the normal cone mapping, observe that the variational inequality (4.10) is equivalent to (4.11). Therefore, due [17], Theorem 17.2.5 there exists a unique strong solution of (4.10) in $C([0,T];L^2(\Omega))$ such that $u(t) \in \operatorname{dom} \partial\Phi$ a.e. $t \in (0,T)$. □

For simplicity and to avoid formulating additional assumptions, we will assume that the solution of problem (4.10) belongs to $C^1([0,+\infty);L^2(\Omega))$. It is known from classic results that $\max\{\dot{u},0\}$ belongs to $C^{1,1}([0,T]\times\bar{\Omega})$, see [20], Section 4 for further regularity results. If $\dot{u}$ is nonnegative, its continuity follows; see, *e.g.*, [21]. The nonnegativity of the time derivative can be established in some special cases (special initial conditions, boundary conditions, and time-independent coefficients); see, for example, [21, 22].

4.3.2. Random domain decomposition and convergence

In order to decompose the domain, let $n \in \mathbb{N}$ and consider a non-overlapping partition $\{\Omega_i\}_{i=1}^n$ of Ω. Let us introduce $\{\chi_{\Omega_i}\}_{i=1}^n \subset W^{1,\infty}(\Omega)$ a partition of unity subordinate to $\{\Omega_i\}_{i=1}^n$, that is,

$$\sum_{i=1}^{n} \chi_{\Omega_i} = 1 \quad \text{in } \Omega, \quad \text{and} \quad \operatorname{supp}(\chi_{\Omega_i}) \subset \Omega_i, \quad \text{for every } i \in \{1,\dots,n\}. \tag{4.12}$$

In a completely analogous way to Section 2 we introduce the batches $\{B_j\}_{j=1}^m$ such that $\cup_{j=1}^m B_j = \{1,\dots,n\}$, the probabilities $\{p_i\}_{i=1}^n$ and $\{\pi_i\}_{i=1}^n$, and the sequence of independent random variables $\{j_k\}_{k\in\mathbb{N}}$. Therefore, let us consider the function $w_\varepsilon : [0,+\infty) \to L^2(\Omega)$ solution of

$$\begin{cases} \int_\Omega \left(\frac{w_k - w_{k-1}}{\varepsilon}\right)(\phi - w_k)\mathrm{d}x + A(w_k,\phi,j_k) - F(\phi,j_k) \geq 0 & \text{a.e. } k \in \mathbb{N}, \\ w_0(x) = u_0(x) & \text{in } \Omega. \end{cases} \tag{4.13}$$

where $A(w_k,\phi,j_k)$ and $F(\phi,j_k)$ are given by

$$A(w_k,\phi,j_k) = \frac{1}{|B_{j_{k_t}}|} \sum_{i \in B_{j_{k_t}}} \int_\Omega \operatorname{div}(\chi_{\Omega_i}\nabla w_k)(\phi - w_k), \tag{4.14}$$

and

$$F(\phi,j_k) = \frac{1}{|B_{j_{k_t}}|} \sum_{i \in B_{j_{k_t}}} \int_\Omega \chi_{\Omega_i} f(\phi - w_k)\mathrm{d}x. \tag{4.15}$$

To embed sequence $\{w_k\}$ into $L^2_{loc}([0,+\infty);L^2(\Omega))$, we consider the function $w : [0,+\infty) \to L^2(\Omega)$ given by $w(t) := w_k$ if $t \in [t_{k-1},t_k)$. The previous system can be understood as a randomization of both the principal operator and the source. It is important to note that for a $k \in \mathbb{N}$ in which the subdomain Ω_* has not been selected, the solution w_k of the (4.13) remains constant for the next step; that is, $w_{k+1} = w_k$ on Ω_*. We have the following theorem. In the following, we assume that $\dot{u}$ is continuous then the following result holds.

Theorem 4.4. *Let $u : [0,+\infty) \to L^2(\Omega)$ be the solution of the obstacle problem (4.10) and $w_\varepsilon : [0,+\infty) \to L^2(\Omega)$ the solution of (4.13). Then, for each $t \in [0,+\infty)$,*

$$\mathbb{E}\big\|w_\varepsilon(t) - u(t)\big\|_{\mathcal{H}}^2 \longrightarrow 0 \quad as \quad \varepsilon \longrightarrow 0^+. \tag{4.16}$$

Moreover, let $u^* \in H_0^1(\Omega)$ be the solution of the (stationary) obstacle problem

$$\int_\Omega \nabla u^* \nabla(\phi - u^*) - \int_\Omega f(\phi - u^*)\mathrm{d}x \geq 0.$$

Then, for every $\eta > 0$ there exists $T > 0$ such that

$$\mathbb{E}\big\|w_\varepsilon(T) - u^*\big\|_{\mathcal{H}}^2 \leq \eta \quad for\ all\ \varepsilon > 0\ small\ enough. \tag{4.17}$$

Proof. Let us consider Φ the functional defined in the proof of Theorem 4.3. Then, the gradient flow associated with Φ is given by (4.11). Using the definition of the normal cone mapping, observe that the variational inequality (4.10) is equivalent to (4.11). On the other hand, for each $i \in \{1,\dots,n\}$, we introduce

$$\Psi_i(u) := \begin{cases} \dfrac{1}{2}\displaystyle\int_\Omega \chi_{\Omega_i}(x)|\nabla u(x)|^2\,\mathrm{d}x - \int_\Omega \chi_{\Omega_i} f(x)u(x)\,\mathrm{d}x & if\ u \in H_0^1(\Omega), \\ +\infty & if\ u \in L^2(\Omega),\ u \notin H_0^1(\Omega). \end{cases}$$

Let us define the functional $\Phi_i : L^2(\Omega) \to \mathbb{R}$ given by $\Phi_i = \Psi_i + \delta_{K(\psi)}$. Analogous to Theorem 4.3, we can deduce that $\partial\Phi = \partial\Psi + N_{K(\psi)}$. Observe that $\operatorname{dom}\partial\Phi_i = \{u \in K(\psi) \cap H_0^1(\Omega) : \operatorname{div}\left(\chi_{\Omega_i}\nabla u\right) \in L^2(\Omega)\}$ and $\partial\Phi_i(u) = -\operatorname{div}\left(\chi_{\Omega_i}\nabla u\right) - \chi_{\Omega_i} f + N_K(u)$. Therefore, in this case, the random minimizing movement introduced in Section 2.2 is given by the solution of

$$-\frac{w_k - w_{k-1}}{\varepsilon} + A_{j_k}(w_k) \in N_{K(\psi)}(w_k), \quad \forall k \in \mathbb{N}. \tag{4.18}$$

where $A_{j_{k_t}} : H_0^1(\Omega) \subset L^2(\Omega) \to L^2(\Omega)$ is defined by

$$A_{j_{k_t}}(v_\varepsilon) = \frac{1}{|B_{j_{k_t}}|}\sum_{i \in \mathcal{B}_{j_{k_t}}} \big(\operatorname{div}\left(\chi_{\Omega_i}\nabla v_\varepsilon\right) - \chi_{\Omega_i} f\big), \quad \text{for every } k \in \mathbb{N}.$$

Observe that $\Lambda \circ u : [0,+\infty) \to L^2(\Omega)$ corresponde to the norm $H_0^2(\Omega)$ of u and is locally bounded due Theorem 4.3. Therefore, since Theorem 2.5 we deduce (4.16). Moreover, as a consequence of Theorem 2.7, we deduce that for every $\eta > 0$, there exists $T > 0$ such that (4.17) holds. □

4.3.3. Illustrative numerical example

In this section, we will illustrate Theorem 4.4 with a numerical example. For this purpose, let us consider $\Omega = [-1,1] \times [-1,1] \subset \mathbb{R}^2$ and the obstacle

$$\psi(x) = \begin{cases} -4(x+0.5)^2 - 4y^2 & \text{if } 4(x+0.5)^2 + 4y^2 < 1, \\ -4(x-0.5)^2 - 4y^2 & \text{if } 4(x-0.5)^2 + 4y^2 < 1, \\ 0 & \text{otherwise.} \end{cases}$$

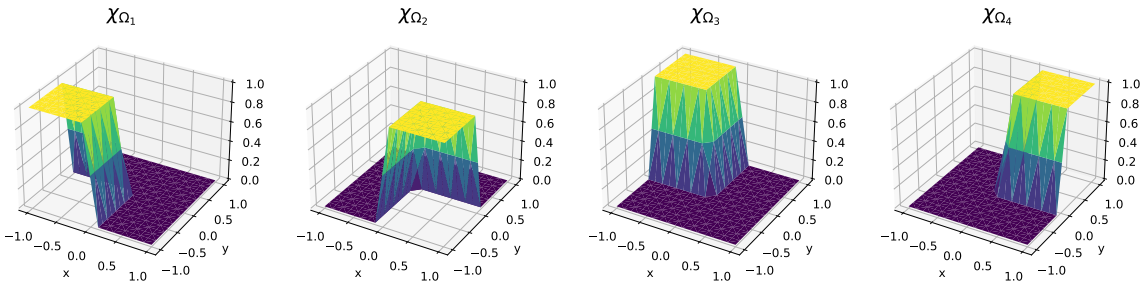


FIGURE 7. Illustration of funtions χ_{Ω_i}. Observe that $\{\chi_i\}_{i=1}^4$ is a partition of unity suborditnate to $\{\Omega_i\}_{i=1}^4$.

We take $u_0(x) = 0$ as the initial condition and $f = -1$ as the source term. This source term acts as a gravity force, pushing u to be close to the obstacle. For the numerical implementation of (4.10), for every $\delta > 0$ we consider the penalized problem

$$\begin{cases} \int_\Omega \dot{u}^\delta \phi \,\mathrm{d}x + \int_\Omega \Delta u^\delta \phi \,\mathrm{d}x - \frac{1}{\delta}\int_\Omega \max\{-u^\delta + \psi, 0\}\phi \,\mathrm{d}x - \int_\Omega f\phi \,\mathrm{d}x = 0 & \text{a.e. } t \in (0,T), \\ u^\delta(x,t) = 0 & \text{on } \partial\Omega \times (0,T), \\ u^\delta(x,0) = u_0(x) & \text{in } \Omega, \end{cases} \tag{4.19}$$

As is shown in [23], Proposition 2.2 $u^\delta \to u$ in $L^2(0,T;L^2(\Omega))$ as $\delta \to 0$. On the other hand, we consider the following subdomains

$$\begin{gathered} \Omega_1 = [-1, 0.1] \times [-1, 0.1], \quad \Omega_2 = [-0.1, 1] \times [-1, 0.1], \\ \Omega_3 = [-0.1, 1] \times [-1, 0.1], \quad \text{and} \quad \Omega_4 = [-0.1, 1] \times [-0.1, 1], \end{gathered}$$

and the function

$$h(x) = \begin{cases} 0 & \text{if } x \geq -0.1 \\ \frac{r+0.1}{0.2} & \text{if } x \in [-0.1, 0.1], \\ 1 & \text{otherwise.} \end{cases}$$

The partition is then given by the functions

$$\begin{gathered} \chi_{\Omega_1}(x,y) = (1-h(x))(1-h(y)), \quad \chi_{\Omega_2}(x,y) = h(x)(1-h(y)), \\ \chi_{\Omega_3}(x,y) = (1-h(x))h(y), \quad \text{and} \quad \chi_{\Omega_4}(x,y) = h(x)h(y). \end{gathered} \tag{4.20}$$

These functions are illustrated Figure 7. We can then condider batches $B_i = i$ and its respective probability for each $i \in \{1, \ldots, 4\}$. Then, we introduce a random minimizing movement scheme of (4.19) as

$$\begin{cases} \int_\Omega \left(\frac{w_k^\delta - w_{k-1}^\delta}{\varepsilon}\right)(\phi - w_k^\delta)\mathrm{d}x + A(w_k^\delta, \phi, j_k) - \frac{1}{\delta}\int_\Omega \max\{-w^\delta + \psi, 0\}\phi \,\mathrm{d}x - F(\phi, j_k) = 0 & \text{a.e. } k \in \mathbb{N}, \\ w_0^\delta(x) = u_0(x) & \text{in } \Omega, \end{cases} \tag{4.21}$$

where A and F are defined as in (4.14) and (4.15) respectively.

We take $T = 0.5$ as the time horizon. For the implementation of (4.19), we consider an implicit Euler discretization of the time interval, using 60 discretization points. For space discretization, we use finite elements,

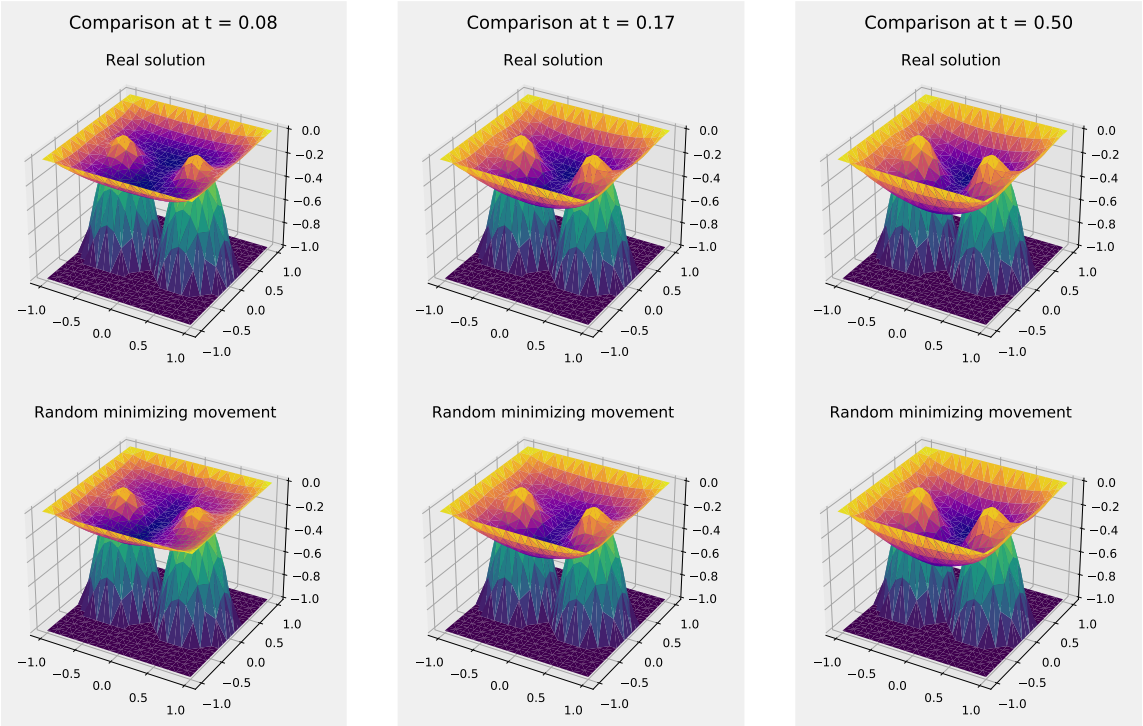


FIGURE 8. Comparison between the obstacle and random minimizing movement solutions (*i.e.*, systems (4.19) and (4.21)) for different times.

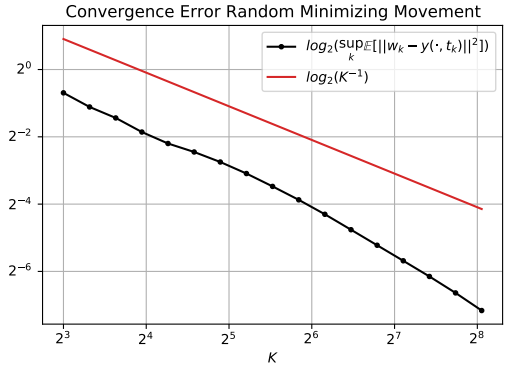


FIGURE 9. We illustrate the convergence as $K \to \infty$ (equivalent to $\varepsilon \to 0$) of the random minimizing movement.

with basis functions given by the standard continuous Galerkin functions (piecewise polynomial functions of order one). Also, the penalization parameter $\delta = 10^{-8}$ and the max function have been regularized using a smooth max function (see [24] for a rigorous analysis). We have used 20 elements. The implementation was carried out using FEniCS, an open-source computing platform finite elements.

For the implementation of (4.21), we used FEniCS to compute the solution. To approximate the average of the solution, we used 8 realizations and then averaged them. The result can be seen in Figure 8.

To illustrate the convergence of the *Random domain decomposition*, we use 8 realizations to approximate the average, and 25 elements fixed for space discretization. Denoting by K the number of time steps that we are considering ($\varepsilon = 1/K$ in (4.21)), Figure 9 illustrates the convergence of this scheme. Observe that we can guarantee convergence, but there has been no convergence rate since our hypothesis. However, the numerical evidence suggests that the convergence order of *Random domain decomposition* is $O(\varepsilon)$.

Acknowledgments

The authors wish to express their gratitude to Enrique Zuazua for insightful discussions, and for his constant and generous support.

Funding

A. Domínguez Corella is supported by the Alexander von Humboldt Foundation with an Alexander von Humboldt research fellowship and by the Emerging Talents Initiative (FAUeti) funding. M. Hernández has been funded by the Transregio 154 Project, Mathematical Modelling, Simulation, and Optimization Using the Example of Gas Networks of the DFG, project C07, the fellowship "ANID-DAAD bilateral agreement", and the DFG and NRF. Südkorea-NRF-DFG-2023 programme number 530756074. Both authors have been partially supported by the DAAD/CAPES Programs for Project-Related Personal, grant 57703041 'Control and numerical analysis of complex system'.

Data availability statement

The research data associated with this article are included in the article.

References

[1] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** (1953) 1087–1092.
[2] N. Metropolis and S. Ulam, The Monte Carlo method. *J. Amer. Statist. Assoc.* **44** (1949) 335–341.
[3] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, Optimization by simulated annealing. *Science* **220** (1983) 671–680.
[4] J.H. Holland, Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI (1975).
[5] L. Bottou, Large-scale machine learning with stochastic gradient descent, in *Proceedings of COMPSTAT'2010*. Physica-Verlag/Springer, Heidelberg (2010) 177–186.
[6] L. Bottou, F.E. Curtis and J. Nocedal, Optimization methods for large-scale machine learning. *SIAM Rev.* **60** (2018) 223–311.
[7] X. Qian and D. Klabjan, The impact of the mini-batch size on the variance of gradients in stochastic gradient descent. arXiv preprint arXiv:2004.13146 (2020).
[8] S. Jin, L. Li and J.-G. Liu, Random batch methods (RBM) for interacting particle systems. *J. Comput. Phys.* **400** (2020) 108877, 30.
[9] S. Jin, L. Li and J.-G. Liu, Convergence of the random batch method for interacting particles with disparate species and weights. *SIAM J. Numer. Anal.* **59** (2021) 746–768.
[10] D. Ko, S.-Y. Ha, S. Jin and D. Kim, Uniform error estimates for the random batch method to the first-order consensus models with antisymmetric interaction kernels. *Stud. Appl. Math.* **146** (2021) 983–1022.
[11] D. Ko and E. Zuazua, Model predictive control with random batch methods for a guiding problem. *Math. Models Methods Appl. Sci.* **31** (2021) 1569–1592.
[12] D.W.M. Veldman, A. Borkowski and E. Zuazua, Stability and convergence of a randomized model predictive control strategy. *IEEE Trans. Automatic Control* **69** (2024) 6253–6260.
[13] D.W.M. Veldman and E. Zuazua, A framework for randomized time-splitting in linear-quadratic optimal control. *Numer. Math.* **151** (2022) 495–549.
[14] J. Latz, Analysis of stochastic gradient descent in continuous time. *Stat. Comput.* **31** (2021) Paper No. 39, 25.
[15] J. Latz, Gradient flows and randomised thresholding: sparse inversion and classification. *Inverse Probl.* **38** (2022) Paper No. 124006, 31.
[16] M. Eisenmann and T. Stillfjord, A randomized operator splitting scheme inspired by stochastic optimization methods. *Numer. Math.* **156** (2024) 435–461.

[17] H. Attouch, G. Buttazzo and G. Michaille, Variational analysis in Sobolev and BV spaces. Vol. 17 of *MOS-SIAM Series on Optimization*, 2nd edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA (2014).
[18] V. Barbu, Nonlinear Differential Equations of Monotone Types in Banach Spaces. Springer Monographs in Mathematics. Springer, New York (2010).
[19] A. Chirilă, M. Marin and A. Öchsner, Distribution Theory Applied to Differential Equations. Springer, Cham (2021).
[20] A. Petrosyan and H. Shahgholian, Parabolic obstacle problems applied to finance, in Recent developments in nonlinear partial differential equations. Vol. 439 of *Contemporary Mathematics*. American Mathematics Society, Providence, RI (2007) 117–133.
[21] A. Blanchet, J. Dolbeault and R. Monneau, On the continuity of the time derivative of the solution to the parabolic obstacle problem with variable coefficients. *J. Math. Pures Appl.* **85** (2006) 371–414.
[22] A. Friedman, Parabolic variational inequalities in one space dimension and smoothness of the free boundary. *J. Funct. Anal.* **18** (1975) 151–176.
[23] D. R. Adams and S. Lenhart, Optimal control of the obstacle for a parabolic variational inequality. *J. Math. Anal. Appl.* **268** (2002) 602–614.
[24] M. Hintermüller and I. Kopacka, A smooth penalty approach and a nonlinear multigrid algorithm for elliptic MPECs. *Comput. Optim. Appl.* **50** (2011) 111–145.
[25] H.H. Bauschke and P.L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert spaces. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 2nd edn. Springer, Cham (2017).
[26] R.S. Burachik and V. Jeyakumar, A dual condition for the convex subdifferential sum formula with applications. *J. Convex Anal.* **12** (2005) 279–290.
[27] S.L. Smith, B. Dherin, D. Barrett and S. De, On the origin of implicit regularization in stochastic gradient Descent, in *International Conference on Learning Representations* (2020).

Appendix A. Appendix: Variance of mini-batch descent

In the following, we present some commentaries concerning the variance of mini-batch descent. Let $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be the convex proper lower semicontinuous functional in Section 2. Due to the definition of probabilities $\pi_1, \dots, \pi_m$ it is possible to see that

$$\Phi = \sum_{j=1}^{m} \pi_j \Phi_{\mathcal{B}_j}. \tag{A.1}$$

This, in turn, says that Φ can be seen as the expected value of a random variable taking value $\Phi_{\mathcal{B}_j}$ with probability π_j. Consider the sequence of independent random variables $\{j_k\}_{k\in\mathbb{N}}$ in Section 2. We see from (A.1) that $\mathbb{E}\Phi_{\mathcal{B}_{j_k}} = \Phi$ for all $k \in \mathbb{N}$.

In order for the mini-batch descent to be a non-biased estimator, in the sense that the expected value of the randomly chosen subgradient coincides with the subdifferential of Φ at each step, the sum rule for subdifferentials was assumed to

hold in Section 2, *i.e.*,

$$\partial\Phi(u) = \sum_{j=1}^{m} \pi_j \, \partial\Phi_{\mathcal{B}_j}(u) \quad \text{for all } u \in \operatorname{dom} \partial\Phi. \tag{A.2}$$

This rule can be thought of as a structural assumption allowing the set-valued variable taking value $\partial\Phi_{\mathcal{B}_j}$ with probability π_j can have expected value $\partial\Phi(u)$. For differentiable potentials, this holds automatically since the gradient operator is linear. For the validity of (A.1), we refer to the general result [25], Corollary 16.50; see also [26], Theorem 3.1 for a sufficient dual condition.

1. Splitting of the minimal norm subdifferential: Additivity assumption (A.2) ensures that for each $u \in \operatorname{dom}\Phi$ there exist $\xi_1(u), \dots, \xi_m(u) \in \mathcal{H}$ such that $\xi(u) \in \partial\Phi_{\mathcal{B}_j}(u)$ for $j \in \{1, \dots, m\}$, and $\partial\Phi(u)^\circ = \sum_{j=1}^m \pi_j \xi_j(u)$. In sparse inversion problem we presented the splitting of the minimal norm subdifferential was unique. However, in the general case, without further assumption on mini-batch potentials, there is no reason for this decomposition to be unique; thus, it should be chosen according to the problem at hand. From now on, we assume that functions $\xi_1, \dots, \xi_m : \operatorname{dom}\Phi :\to \mathcal{H}$ are fixed.

It can be readily seen that for each $k \in \mathbb{N}$, $\mathbb{E}\xi_{j_k}(u) = \partial\Phi(u)^\circ$ for all $u \in \operatorname{dom}\Phi$.

In Section 2, we introduced function $\Lambda : \mathcal{H} \to \mathbb{R}$ given by $\Lambda(u) := \sum_{j=1}^m \pi_j |\xi_j(u) - \partial\Phi(u)^\circ|^2_{\mathcal{H}}$ to provide a quantitative measure of variance for mini-batch descent. When reduced to the case where mini-batch potentials are differentiable, Λ is the usual quantifier of variance used to provide bounds and estimates in stochastic gradient algorithms. The key feature of Λ is that $\operatorname{Var}[\xi_{j_k}] = \Lambda$ for all $k \in \mathbb{N}$.

2. Local boundedness of Λ: Let $u : [0, +\infty) \to \mathcal{H}$ be the solution of gradient flow equation (2.1). All hypotheses made in Section 2 require at least local integrability of $\Lambda \circ u : [0, +\infty) \to \mathbb{R}$. In all the examples we provided (sparse inversion, constrained optimization, and domain decomposition), this function was locally bounded.

In general, there is a criterion to assess the local boundedness of the subdifferential; see [25]. This can be employed to give a sufficient condition under which Λ is locally bounded (this is stronger than $\Lambda \circ u$ being locally bounded). For any $v \in \mathcal{H}$,

$$\text{if } v \in \operatorname{int}\operatorname{dom}\Phi_{\mathcal{B}_j} \ \forall j \in \{1, \dots, m\}, \text{ then } \exists \mathcal{V} \in \mathcal{N}(v) \text{ such that } \Lambda(\mathcal{V}) \text{ is bounded.}$$

Here, $\mathcal{N}(v)$ denotes the neighborhood filter of $v \in \mathcal{H}$. Therefore, the openness of the effective domains of mini-batch potentials is, in general, enough to grant the local boundedness of $\Lambda \circ u$. However, in general, and due to the properties of u, it is much easier to bound $\Lambda \circ u$, as shown in the example in Section 4.2, where $\Lambda \circ u$ is bounded all over $[0, +\infty)$.

3. Another interpretation of function Λ: For simplicity, suppose that m divides n, and that batches contain exactly $\frac{n}{m}$ elements. Moreover, assume that they are selected with equal probability, *i.e.*, $\pi_j = m^{-1}$ for all $j \in \{1, \dots, m\}$.

In this case, the potential can be represented as

$$\Phi(u) = \frac{1}{n} \sum_{j=1}^{m} \sum_{i \in \mathcal{B}_j} \Phi_i(u) \quad \forall u \in \operatorname{dom}\Phi.$$

Assuming that $\partial\Phi_{\mathcal{B}_j} = m^{-1} \sum_{i \in \mathcal{B}_j} \partial\Phi_i$, it is possible to find functions $\xi_{i,j} : \operatorname{dom}\Phi \to \operatorname{dom}\Phi_i$ such that $\partial\Phi(u)^\circ = n^{-1} \sum_{j=1}^m \sum_{i \in \mathcal{B}_j} \xi_{i,j}(u)$ for all $u \in \operatorname{dom}\Phi$. Consider the function $\Gamma : \mathcal{H} \to \mathbb{R}$ given by

$$\Gamma(u) := \frac{1}{n} \sum_{j=1}^{m} \sum_{i \in \mathcal{B}_j} \left\| \xi_{i,j}(u) - \partial\Phi(u)^\circ \right\|^2_{\mathcal{H}}$$

This function can be readily identified as the trace of the empirical covariance matrix of per-example subgradients.

Following [27], Appendix A, it is possible to find that

$$\Lambda(u) = \frac{n - m}{m(n-1)} \Gamma(u) \quad \forall u \in \operatorname{dom}\Phi.$$