HAL
open science

On the convergence of gradient-like flows with noisy gradient input

Panayotis Mertikopoulos, Mathias Staudigl

To cite this version:

Panayotis Mertikopoulos, Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. SIAM Journal on Optimization, 2018, 28 (1), pp.163-197. 10.1137/16m1105682 . hal-01404586

HAL Id: hal-01404586

https://inria.hal.science/hal-01404586v1

Submitted on 28 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ON THE CONVERGENCE OF GRADIENT-LIKE FLOWS WITH NOISY GRADIENT INPUT

PANAYOTIS MERTIKOPOULOS* AND MATHIAS STAUDIGL♯

Abstract. In view of solving convex optimization problems with noisy gradient input, we analyze the asymptotic behavior of gradient-like flows that are subject to random disturbances. Specifically, we focus on the widely studied class of mirror descent methods for constrained convex programming and we examine the dynamics' convergence and concentration properties in the presence of noise. In the small noise limit, we show that the dynamics converge to the solution set of the underlying problem (a.s.). Otherwise, if the noise is persistent, we estimate the measure of the dynamics' long-run concentration around interior solutions and their convergence to boundary solutions that are sufficiently "robust". Finally, we show that a rectified variant of the method with a decreasing sensitivity parameter converges irrespective of the magnitude of the noise and/or the structure of the underlying convex program.

1. Introduction

Consider a general convex program of the form

$$\begin{aligned} \text{minimize} \quad & f(x), \\ \text{subject to} \quad & x \in \mathcal{X}, \end{aligned} \tag{P}$$

where $\mathcal{X}$ is a closed convex subset of a finite-dimensional vector space $\mathcal{V}$ and the objective function $f\colon \mathcal{X} \to \mathbb{R}$ is itself convex. In the base case where $\mathcal{X} = \mathcal{V}$ and f is smooth, the *gradient flow* of f is given by the gradient descent dynamics

$$\dot{x} = -\nabla f(x), \tag{GD}$$

where $\nabla f(x)$ denotes the (Euclidean) gradient of f at x. As is well known, under mild regularity assumptions for f, the orbits of (GD) converge to $\arg\min f$ (provided of course that said set is nonempty). Thus, building on this "quick-and-easy" convergence result, (GD) and its variants have become the starting point for a vast corpus of literature in convex optimization and control.

On the other hand, if the gradient input to (GD) is contaminated by noise (e.g. due to faulty measurements or other exogenous factors), this convergence is

* French National Center for Scientific Research (CNRS) and Laboratoire d'Informatique de Grenoble (LIG), F-38000 Grenoble, France

♯ Maastricht University, Department of Quantitative Economics, P.O. Box 616, NL–6200 MD Maastricht, The Netherlands.

E-mail addresses: panayotis.mertikopoulos@imag.fr, m.staudigl@maastrichtuniversity.nl.

2010 *Mathematics Subject Classification.* Primary 90C25, 60H10; secondary 90C15.

Key words and phrases. Convex programming; dynamical systems; mirror descent; noisy feedback; stochastic differential equations.

This research was supported by the CNRS exploratory project REAL.NET (contract no. PEPS–JCJC–INS2I–07RECH1124–S), and the French National Research Agency (ANR) grant ORACLESS (contract no. ANR–16–CE33–0004–01).

destroyed even in simple, one-dimensional problems. To see this, take the quadratic objective $f(x) = \theta(x-\mu)^2/2$ with parameters $\mu \in \mathbb{R}$ and $\theta > 0$, and consider the perturbed dynamics

$$dX = -\theta(X-\mu)\,dt + \sigma\,dW, \tag{1.1}$$

where $W(t)$ is a one-dimensional Wiener process (Brownian motion) with volatility $\sigma > 0$. This system describes an Ornstein–Uhlenbeck (OU) process with mean μ and reversion rate θ [29], leading to the explicit solution formula

$$X(t) = X(0)e^{-\theta t} + \mu(1-e^{-\theta t}) + \sigma \int_0^t e^{-\theta(t-s)}\,dW(s). \tag{1.2}$$

Thanks to this expression, several conclusions can be drawn regarding (1.1). First, even though the drift of the dynamics (1.1) vanishes at μ (and only at μ), the process $X(t)$ *does not* converge to μ with positive probability; instead, $X(t)$ converges in distribution to a Gaussian random variable X_∞ with mean μ and variance $\sigma^2/(2\theta)$. Thus, in the long run, $X(t)$ will fluctuate around μ with a spread that is roughly proportional to the noise volatility coefficient σ. In addition, by ergodicity, the process $\bar{X}(t) = t^{-1}\int_0^t X(s)\,ds$ also converges to X_∞ (a.s.), so even the long-run average of $X(t)$ fails to converge to μ with positive probability.

More generally, by solving the associated Fokker–Planck equation, it is easy to see that, under suitable growth assumptions for f, the stochastic gradient descent system

$$dX = -\nabla f(X)dt + \sigma\,dW \tag{SGD}$$

admits a unique invariant measure μ_∞ with density $\rho_\infty(x) \propto e^{-2f(x)/\sigma^2}$ [14]. In other words, for large t, $X(t)$ is most likely to be found near $\arg\min f$, and this likelihood is inversely proportional to σ. However, apart from this basic concentration result for unconstrained problems, the long-run behavior of gradient-like flows in the presence of stochastic disturbances remains largely unexplored – especially in the context of constrained optimization.

To address this issue in a concrete manner, we focus on the class of *mirror descent* (MD) dynamics that were pioneered by Nemirovski and Yudin [36] and which have since given rise to an extensive literature in mathematical optimization – see e.g. [6, 7, 18, 35, 37, 38, 43] and references therein. Specifically, in the absence of noise, these dynamics take the form

$$\begin{aligned} \dot{y} &= -\nabla f(x), \\ x &= Q(\eta y), \end{aligned} \tag{MD}$$

where, referring to Section 2 for the details, $\eta > 0$ is a sensitivity parameter while the so-called "mirror map" Q is a projection-like mapping induced by a strongly-convex "prox-function" $h\colon \mathcal{X} \to \mathbb{R}$; more precisely, $Q(y) = \arg\max_{x\in\mathcal{X}}\{\langle y\,|\,x\rangle - h(x)\}$. In this way, (MD) represents the continuous-time limit of the well known dual averaging scheme [38]

$$\begin{aligned} y_{n+1} &= y_n + \gamma_n v(x_n), \\ x_{n+1} &= Q(\eta y_{n+1}), \end{aligned} \tag{1.3}$$

where $\gamma_n > 0$ denotes the method's variable step-size sequence.

As an example, if $\mathcal{X} = \mathcal{V}$ and $h(x) = \frac{1}{2}\|x\|_2^2$, we have $Q(y) = y$, so (MD) boils down to the gradient descent dynamics (GD). More generally, as was shown in [3, 4, 13], (MD) can also be viewed as the gradient flow of f with respect to a certain Riemannian metric on $\mathcal{X}$. Thus, in addition to (GD) and its projected

variants, (MD) also covers a very broad class of gradient-like flows. We analyze these deterministic dynamics in Section 2, where we establish the convergence of trajectories to the solution set of (P) with an $\mathcal{O}(1/t)$ value convergence rate for the averaged process $\bar{x}(t) = t^{-1} \int_0^t x(s)\, ds$.

Moving beyond this deterministic setting, the study of mirror descent with stochastic first-order feedback is a classic topic in optimization (see e.g. [8, 18, 35, 37] and references therein), and our paper addresses precisely this question. In particular, to analyze the impact of noise, we focus on the dynamics (MD) perturbed by a general martingale process such as the Brownian motion term in (SGD) above. This leads us to consider the *stochastic mirror descent* dynamics

$$\begin{aligned} dY &= -\nabla f(X)\, dt + dZ, \\ X &= Q(\eta Y), \end{aligned} \tag{SMD}$$

where $Z(t)$ is a continuous Itô martingale representing the noise in the gradient input process.

Of course, in this stochastic setting, the simple example (1.1) shows that the deterministic convergence properties of the dynamics (MD) cannot be carried over to (SMD) in full generality. In view of this, we focus on the following questions:

(1) If the infinitesimal variance of $Z(t)$ decays over time, the solutions of (SMD) can be viewed as *asymptotic pseudotrajectories* (APTs) [10] of the gradient-like flow (MD).[1] Heuristically, this suggests that the good trajectory convergence properties of (MD) can be transferred to (SMD) in the small noise limit, an intuition which we make precise in Section 4. We are not aware of a similar APT-based analysis in optimization, and this interesting link between deterministic and stochastic mirror descent only becomes transparent in continuous time.

(2) If the noise is persistent, the APT property is lost, so trajectory convergence to interior solutions is no longer possible. Nonetheless, if f is strongly convex and (P) admits a (necessarily) unique interior solution x^* (a case of particular interest in machine learning and statistics [44]), the long-run behavior of (SMD) can be described by examining the dynamics' invariant distribution. Our analysis in Section 5 provides an explicit concentration estimate for this distribution and shows that, if the sensitivity parameter η of (SMD) is taken small enough, $X(t)$ lies arbitrarily close to x^* for an arbitrarily large fraction of the time.

(3) Departing from the interior case, we also consider "robust" boundary solutions that satisfy a certain gradient normality condition which is characteristic of generic linear programs (and which always holds for such problems). In this case, if the sensitivity η of (SMD) is taken sufficiently small, $X(t)$ converges (a.s.), and this convergence occurs in finite time if the mirror map Q is surjective (cf. Section 6).

(4) Finally, if no assumptions can be made on the structure of (P) or its solution set, we show in Section 7 that a rectified variant of (SMD) with a decreasing sensitivity parameter $\eta \equiv \eta(t)$ converges (a.s.). Specifically, if $\eta(t)$ decays to zero as $\Theta(t^{-1/2})$, we obtain an almost sure $\mathcal{O}(t^{-1/2}\sqrt{\log\log t})$ value convergence bound – which drops down to $\mathcal{O}(t^{-1/2})$ in expectation.

[1] For background information on the theory of stochastic approximation and APTs, see [9, 10].

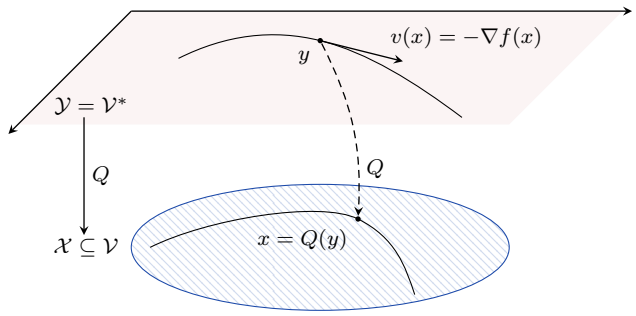

Figure 1. Schematic representation of (MD).

At a technical level, this paper belongs to the extensive literature on dynamical systems that arise in the solution of continuous optimization problems and variational inequalities – see e.g. [1, 5, 12, 21, 24, 36, 45] and references therein. The deterministic bedrock of our analysis consists of the gradient-like dynamics studied in [3, 4, 13, 36]; along with an important dichotomy that arises between these dynamics in the presence of noise, we examine this link at depth in Section 8. The work that is closest to our stochastic analysis is the recent paper [39], where the authors showed that the time-average of an interior-valued subclass of (SMD) converges within $\mathcal{O}(\sigma_*^2)$ of the solution set of (P) when the volatility of the noise is bounded from above by σ_*^2. To the best of our knowledge, this is the only result known for (SMD), and our analysis in Section 7 shows that the optimality gap predicted by [39] can be reduced to 0 if (SMD) is employed with a decreasing sensitivity parameter.

2. Preliminaries

2.1. **Mirror descent.** Following the original approach of Nemirovski and Yudin [36], the main idea of mirror descent is as follows: at each $t \geq 0$, the optimizer takes an infinitesimal step along the negative gradient of f in the dual space $\mathcal{V}^*$ (where gradients live)[2] and the output is "mirrored" back to the problem's feasible region $\mathcal{X} \subseteq \mathcal{V}$. More precisely, this process can be represented in continuous time as

$$\begin{aligned} \dot{y} &= v(x), \\ x &= Q(\eta y), \end{aligned} \tag{MD}$$

where:

1. $v(x) = -\nabla f(x)$ denotes the negative gradient of f at x.
2. $y \in \mathcal{V}^*$ is an auxiliary "score" variable that aggregates gradient steps.
3. $\eta > 0$ is a sensitivity parameter (discussed in detail below).
4. $Q\colon \mathcal{V}^* \to \mathcal{X}$ is the *mirror* (or *choice*) map that outputs a solution candidate $x \in \mathcal{X}$ as a function of the score variable $y \in \mathcal{V}^*$ (also discussed below).

For a schematic illustration of these dynamics, see Fig. 1.

[2]Recall here that $\nabla f(x)$ acts naturally on vectors $z \in \mathcal{V}$ via the directional derivative mapping $z \mapsto \langle \nabla f(x) \,|\, z \rangle \equiv f'(x; z)$. For convenience, we also assume that f is tacitly defined on an open neighborhood of $\mathcal{X}$; doing so allows us to use the ordinary definition of $\nabla f(x)$ but none of our results depend on this device.

A key element in the above description of mirror descent is the distinction between primal and dual variables – that is, between the candidate solution x and the auxiliary score vector y. To emphasize this duality, we will write $\mathcal{Y} \equiv \mathcal{V}^*$ for the dual space of $\mathcal{V}$ and, following [38], we will often refer to (MD) as a *dual averaging* method. Also, in terms of regularity, we will assume throughout that

$$v(x) \text{ is Lipschitz continuous on } \mathcal{X}. \tag{H1}$$

Given that the dual variable y aggregates (negative) gradient steps, a natural choice for Q would be the arg max correspondence $y \mapsto \arg\max_{x\in\mathcal{X}}\langle y\,|\,x\rangle$ whose output is most closely aligned with y. However, this assignment generically selects only extreme points of $\mathcal{X}$, so it is ill-suited for general, nonlinear convex programs. On that account, (MD) is typically run with "regularized" mirror maps of the form $y \mapsto \arg\max_{x\in\mathcal{X}}\{\langle y\,|\,x\rangle - h(x)\}$ where the "penalty term" $h(x)$ satisfies the following requirements:

Definition 2.1. We say that $h\colon \mathcal{X} \to \mathbb{R}$ is a *penalty function* (or *regularizer*) on $\mathcal{X}$ if it is continuous and *strongly convex*, i.e. there exists some $K > 0$ such that

$$h(\lambda x + (1-\lambda)x') \leq \lambda h(x) + (1-\lambda)h(x') - \tfrac{1}{2}K\lambda(1-\lambda)\|x'-x\|^2 \tag{2.1}$$

for all $x, x' \in \mathcal{X}$ and all $\lambda \in [0,1]$. The *mirror map* induced by h is then given by

$$Q(y) = \arg\max_{x\in\mathcal{X}}\{\langle y\,|\,x\rangle - h(x)\}. \tag{2.2}$$

In view of the above, the sensitivity parameter η of (MD) essentially controls the weight of the penalty term $h(x)$ in (2.2). Specifically, we have

$$Q(\eta y) = \arg\max_{x\in\mathcal{X}}\{\eta\langle y\,|\,x\rangle - h(x)\} = \arg\max_{x\in\mathcal{X}}\{\langle y\,|\,x\rangle - \eta^{-1}h(x)\}, \tag{2.3}$$

i.e. running (MD) with $Q(\eta y)$ instead of $Q(y)$ amounts to rescaling h by $1/\eta$. In particular, as $\eta \to 0$, the "η-deflated" mirror map $Q(\eta y)$ tends to select points that are closer to the "prox-center" $x_c \equiv \arg\min h$ of $\mathcal{X}$, so the primal variable $x \in \mathcal{X}$ becomes less susceptible to changes in $y \in \mathcal{Y}$.

For concreteness, we discuss below two examples of this regularization process:

Example 2.1 (Euclidean projections). Let $h(x) = \frac{1}{2}\|x\|_2^2$. Then, h is 1-strongly convex with respect to $\|\cdot\|_2$ and the corresponding mirror map is the closest point projection

$$\Pi_{\mathcal{X}}(y) = \arg\max_{x\in\mathcal{X}}\big\{\langle y\,|\,x\rangle - \tfrac{1}{2}\|x\|_2^2\big\} = \arg\min_{x\in\mathcal{X}}\|y-x\|_2^2. \tag{2.4}$$

Accordingly, the dynamics derived from (2.4) may be viewed as a projected gradient descent scheme; for instance, if $\mathcal{X} = \mathcal{V}$, we obtain precisely (GD). For future reference, we also note that h is differentiable throughout $\mathcal{X}$ and $\Pi_{\mathcal{X}}$ is *surjective* (i.e. $\operatorname{im}\Pi_{\mathcal{X}} = \mathcal{X}$).

Example 2.2 (Matrix regularization). Motivated by applications to semidefinite programming, consider the unit spectrahedron $\mathcal{D} = \{X \in \operatorname{Sym}(\mathbb{R}^n) : X \succcurlyeq 0,\ \|X\|_1 \leq 1\}$ of (symmetric) positive-semidefinite matrices $X \in \operatorname{Sym}_+(\mathbb{R}^n)$ such that $\|X\|_1 \equiv \sum_{i=1}^n |\operatorname{eig}_i(X)| \leq 1$. Then, a commonly employed penalty function on $\mathcal{D}$ is provided by the *von Neumann* (or *quantum*) *entropy*

$$h(X) = \operatorname{tr}[X \log X]. \tag{2.5}$$

This penalty function is (1/2)-strongly convex with respect to the nuclear norm [25], and a straightforward calculation [31, Prop. A.1] shows that the associated mirror map is given by

$$\Lambda(Y) = \frac{\exp(Y)}{1 + \|\exp(Y)\|_1} \quad \text{for all } Y \in \operatorname{Sym}(\mathbb{R}^n). \tag{2.6}$$

In contrast to Example 2.1, h is differentiable only on the relative interior $\mathcal{D}^\circ$ of $\mathcal{D}$; furthermore, since $\exp(Y) \succ 0$ for all $Y \in \operatorname{Sym}(\mathbb{R}^n)$, we have $\operatorname{im}\Lambda = \mathcal{D}^\circ$ (i.e. Λ is "essentially" surjective).

The two examples above highlight an important link between the domain of differentiability of h and the image of the induced mirror map Q. To describe this relationship in detail, extend h to all of $\mathcal{V}$ by setting $h \equiv \infty$ outside $\mathcal{X}$, and let $\partial h(x)$ denote the subdifferential of h at $x \in \mathcal{X}$. We then have $\mathcal{X}^\circ \subseteq \operatorname{dom}\partial h \equiv \{x \in \mathcal{X} : \partial h(x) \neq \varnothing\} \subseteq \mathcal{X}$, meaning that h may fail to be subdifferentiable only on the boundary of $\mathcal{X}$ [41, Chap. 26]. Intuitively, this can happen only if h becomes "infinitely steep" near $x \in \operatorname{bd}(\mathcal{X})$, so we say that h is *steep* at x if $\partial h(x) = \varnothing$ and *nonsteep* otherwise.

The following proposition shows that penalty functions that are everywhere nonsteep induce mirror maps that are surjective; on the flip side, penalty functions that are steep throughout $\operatorname{bd}(\mathcal{X})$ give rise to interior-valued mirror maps (compare with Examples 2.1 and 2.2 respectively):

Proposition 2.2. *Let h be a K-strongly convex penalty function, let $Q\colon \mathcal{Y} \to \mathcal{X}$ be the induced mirror map, and let $h^*\colon \mathcal{Y} \to \mathbb{R}$ be the convex conjugate of h, i.e.*

$$h^*(y) = \max\{\langle y|x\rangle - h(x) : x \in \mathcal{X}\} \quad \textit{for all } y \in \mathcal{Y}. \tag{2.7}$$

Then:

1) *$x = Q(y)$ if and only if $y \in \partial h(x)$; in particular, $\operatorname{im} Q = \operatorname{dom}\partial h$.*
2) *h^* is differentiable on $\mathcal{Y}$ and $\nabla h^*(y) = Q(y)$ for all $y \in \mathcal{Y}$.*
3) *Q is $(1/K)$-Lipschitz continuous.*

In what follows, we will use the above statements freely; for a proof, see [41, Theorem 23.5] and [42, Theorem 12.60(b)].

2.2. **Bregman divergences and the Fenchel coupling.** In addition to Proposition 2.2, a key tool in the convergence analysis of mirror descent (at least when h is steep) is the *Bregman divergence* $D(p,x)$ between $x \in \mathcal{X}$ and a target point $p \in \mathcal{X}$. Following [16], $D(p,x)$ is defined as the difference between $h(p)$ and the best linear approximation of $h(p)$ starting from x, viz.

$$D(p,x) = h(p) - h(x) - h'(x; p - x), \tag{2.8}$$

where $h'(x;z) = \lim_{t\to 0^+} t^{-1}[h(x+tz) - h(x)]$ is the one-sided derivative of h at x along $z \in \operatorname{TC}(x)$. Given that h is (strongly) convex, we have $D(p,x) \geq 0$ and $x(t) \to p$ whenever $D(p,x(t)) \to 0$; consequently, the convergence of $x(t)$ to p can be checked via the associated divergence $D(p,x(t))$.

Notwithstanding, if h is not steep, it is often impossible to obtain information about $D(p,x(t))$ from (MD) unless $x(t)$ is interior. To overcome this difficulty, we will instead employ the so-called *Fenchel coupling*

$$F(p,y) = h(p) + h^*(y) - \langle y|p\rangle \quad \text{for all } p \in \mathcal{X},\ y \in \mathcal{Y}. \tag{2.9}$$

This "primal-dual" divergence was first introduced in [30, 33] and its name alludes to the fact that (2.9) collects all terms of Fenchel's inequality.[3] As a result, $F(p, y)$ is nonnegative and strictly convex in both arguments. Moreover, as we show in Proposition A.2, we have $Q(y_n) \to p$ for every sequence $(y_n)_{n=0}^{\infty}$ in $\mathcal{Y}$ such that $F(p, y_n) \to 0$. For technical reasons, it will often be convenient to assume that the converse also holds, namely

$$F(p, y_n) \to 0 \quad \text{whenever} \quad Q(y_n) \to p. \tag{H2}$$

Clearly, when this is the case, we have $Q(y_n) \to p$ if and *only if* $F(p, y_n) \to 0$.

2.3. **Deterministic convergence analysis.** Together with Proposition 2.2, the Lipschitz continuity hypothesis (H1) implies that the vector field $\mathcal{Y} \ni y \mapsto v(Q(\eta y))$ of (MD) is itself Lipschitz continuous in y. Hence, by standard results in the theory of differential equations, (MD) is *well-posed*, i.e. it admits a unique global solution for every initial condition $y_0 \in \mathcal{Y}$ [40, Chap. V].

This existence and uniqueness result allows us to study the long-term behavior of the solutions of (MD) with respect to the underlying convex program (P). Focusing for simplicity on the case where $\mathcal{X}$ is compact (so $\arg\min f \neq \varnothing$), we have:

Theorem 2.3. *Assume that $\mathcal{X}$ is compact and (H1) holds.*

(1) *If $f_{\min}(t) = \min_{0 \le s \le t} f(x(s))$ and $\bar{f}(t) = t^{-1} \int_0^t f(x(s))$ respectively denote the minimum and mean value of f achieved by (MD), we have*

$$f_{\min}(t) - \min f \le \bar{f}(t) - \min f = \mathcal{O}(1/t). \tag{2.10}$$

In particular, if (MD) is initialized at $y_0 = 0$, we obtain

$$f_{\min}(t) \le \bar{f}(t) \le \min f + \Omega/t, \tag{2.11}$$

where $\Omega = \max\{h(x') - h(x) : x, x' \in \mathcal{X}\}$.

(2) *If (H2) also holds, $x(t)$ converges to some $x^* \in \arg\min f$.*

Theorem 2.3 is a strong convergence result for (MD) as it guarantees global trajectory convergence to a solution of (P) and an $\mathcal{O}(1/t)$ value convergence rate. Up to minor technical differences, the first part of the theorem essentially dates back to Nemirovski and Yudin [36] who introduced the method in the first place. As for the trajectory convergence properties of (MD), [3, 13] provide a proof for a Hessian Riemannian gradient system which is equivalent to (MD) when h is steep (for a detailed discussion, see Section 8). Finally, [4] deals with the singular Riemannian case (corresponding to nonsteep h), but requires that $\mathcal{X}$ be polyhedral.

In Appendix B, we provide a proof of Theorem 2.3 hinging on the fact that the "η-deflated" coupling

$$F_\eta(x^*, y) = \eta^{-1} F(x^*, \eta y) \tag{2.12}$$

is a strict Lyapunov function for (MD) whenever $x^* \in \arg\min f$. Building on this, our aim in the rest of this paper will be to explore how the convergence guarantees of Theorem 2.3 are affected if the gradient input of (MD) is contaminated by noise and measurement errors.

[3] For a related, trajectory-based variant of F, see also [4, p. 444].

3. Noisy mirror descent

To account for the effects of noise in (MD), our point of departure is the stochastic disturbance model

$$\dot{y}(t) = v(x(t)) + \epsilon(t), \tag{3.1}$$

where $\epsilon(t)$ is a random function of time representing the noise in the gradient input $v(x(t))$ at each instance $t \geq 0$. Accordingly, writing out the Langevin equation (3.1) as a formal stochastic differential equation, we will focus on the stochastic mirror descent dynamics

$$\begin{aligned} dY &= v(X)\,dt + dZ, \\ X &= Q(\eta Y), \end{aligned} \tag{SMD}$$

where $Z(t) = (Z_1(t), \dots, Z_n(t))$ is a continuous Itô martingale representing the noise in the gradient input process. To be more precise, we assume throughout that $Z(t)$ is of the general form

$$dZ_i(t) = \sum_{k=1}^{m} \sigma_{ik}(X(t), t)\,dW_k(t), \quad i = 1, \dots, n, \tag{3.2}$$

where:

(1) $W = (W_1, \dots, W_m)$ is a standard m-dimensional Wiener process with respect to some underlying stochastic basis $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$.[4]

(2) The $n\times m$ *volatility matrix* $\sigma_{ik} \colon \mathcal{X}\times\mathbb{R}_+ \to \mathbb{R}$ of $Z(t)$ is assumed measurable, bounded, and Lipschitz continuous in the first argument. Formally, we posit that, for all $x, x' \in \mathcal{X}$ and all $t \geq 0$, we have:

$$\begin{aligned} \sup\nolimits_{x,t} \|\sigma(x,t)\| &< \infty, \\ \|\sigma(x',t) - \sigma(x,t)\| &= \mathcal{O}(\|x' - x\|). \end{aligned} \tag{H3}$$

In the general setting of (SMD), the noise may depend on t and $X(t)$ in a fairly arbitrary manner: for instance, the increments of $Z(t)$ need not be i.i.d. and the noise in different components need not be independent either. This is reflected in the quadratic covariation process of $Z(t)$,[5] given here by

$$d[Z_i, Z_j] = \sum_{k,\ell=1}^{m} \sigma_{ik}\sigma_{j\ell}\,dW_k \cdot dW_\ell = \sum_{k=1}^{m} \sigma_{ik}\sigma_{jk}\,dt = \Sigma_{ij}\,dt, \tag{3.3}$$

where

$$\Sigma = \sigma\sigma^\top. \tag{3.4}$$

Clearly, if the infinitesimal covariance matrix Σ is not diagonal, the components of Z will exhibit nontrivial correlations along the nonzero off-diagonal elements of Σ. The above highlights the role of the underlying m-dimensional Wiener process $W(t)$ in (SMD): if $m < n$, the induced disturbances are necessarily correlated across different components of Z; if $m = n$ and σ is diagonal, the errors affecting v are independent across components; and if $m > n$, the noise in each component of v may result from the aggregation of several independent errors. Obviously, the precise statistics of the noise depend on the application being considered, so, for generality,

[4] In particular, we do not assume here that $m = n$; we discuss this point in detail below.

[5] Recall here that the covariation of two processes X and Y is defined as $[X(t), Y(t)] = \lim_{|\Pi|\to 0} \sum_{1\leq j\leq k} (X(t_j) - X(t_{j-1}))(Y(t_j) - Y(t_{j-1}))$, where the limit is taken over all partitions $\Pi = \{t_0 = 0 < t_1 < \dots < t_k = t\}$ of $[0, t]$ with mesh $|\Pi| \equiv \max_j |t_j - t_{j-1}| \to 0$ [29].

we maintain an application-agnostic approach and we make no assumptions on the structure of Σ.[6]

Remark 3.1. For posterity, note that the noise regularity hypothesis (H3) gives

$$\|\Sigma(x,t)\|_* \leq \sigma_*^2 \quad \text{for some finite } \sigma_* > 0 \text{ and for all } x \in \mathcal{X},\, t \geq 0, \tag{3.5}$$

where $\|\Sigma\|_* \equiv \sup\{\|\Sigma y\| : y \in \mathcal{Y},\, \|y\|_* \leq 1\}$ denotes the induced matrix norm on $\mathcal{V} \otimes \mathcal{V} \cong \mathbb{R}^{n\times n}$. In what follows, it will be convenient to measure the magnitude of the noise affecting (SMD) via σ_*; obviously, when $\sigma_* = 0$, we recover the noiseless, deterministic dynamics (MD).

Now, under the Lipschitz continuity hypothesis (H1) and the noise regularity condition (H3), standard results from the theory of stochastic differential equations show that (SMD) admits unique strong solutions that exist for all time. Specifically, for every (random) $\mathcal{F}_0$-measurable initial condition Y_0 with $\mathbb{E}[\|Y_0\|_*^2] < \infty$, there exists an almost surely continuous stochastic process $Y(t)$ satisfying (SMD) for all $t \geq 0$ and such that $Y(0) = Y_0$. Furthermore, up to redefinition on a $\mathbb{P}$-null set, $Y(t)$ is the unique $\mathcal{F}_t$-adapted process with these properties [27, Theorem 3.4].

For concreteness, we will focus only on non-random initial conditions of the form $Y(0) = y_0$ for a fixed $y_0 \in \mathcal{Y}$. In this case, the second moment condition $\mathbb{E}[\|Y(0)\|_*^2] < \infty$ is satisfied automatically, so we have:

Proposition 3.1. *For all $y_0 \in \mathcal{Y}$, and up to a $\mathbb{P}$-null set,* (SMD) *admits a unique continuous solution $Y(t)$, $t \geq 0$, such that $Y(0) = y_0$.*

With this well-posedness result at hand, we show below that, if $X(t)$ converges, it does so to a minimum point of f:

Proposition 3.2. *If $\lim_{t\to\infty} X(t) = x^*$, then $x^* \in \arg\min f$ (a.s.).*

Proof. Let $v^* = v(x^*) = -\nabla f(x^*)$, and assume to the contrary that $x^* \notin \arg\min f$ so there exists some $q \in \mathcal{X}$ such that $\langle v^* | q - x^*\rangle > 0$. Since $X(t) = Q(\eta Y(t))$, we also have $\eta Y(t) \in \partial h(X(t))$ by Proposition 2.2; hence, for large t, we get

$$h(q) - h(X(t)) \geq \langle \eta Y(t) | q - X(t)\rangle \sim \eta\langle Y(t) | q - x^*\rangle. \tag{3.6}$$

Now, by assumption, we also have $v(X(t)) \to v^*$, implying in turn that $\bar{v}(t) = t^{-1}\int_0^t v(X(s))\,ds \to v^*$. Since $Y(t) = Y(0) + t\bar{v}(t) + Z(t)$ by (SMD), the asymptotic growth estimate (C.4) in Appendix C readily gives

$$h(q) - h(X(t)) \gtrsim \eta\langle Y(0) + t\bar{v}(t) + Z(t) | q - x^*\rangle \sim t\eta\langle v^* | q - x^*\rangle. \tag{3.7}$$

Therefore, with $\langle v^* | q - x^*\rangle > 0$, we conclude that $h(q) - h(X(t)) \to \infty$ (a.s.). Since $\sup_{t\geq 0}\{h(q) - h(X(t))\} < \infty$, this contradicts our initial claim that $x^* \notin \arg\min f$, and our proof is complete. ∎

Proposition 3.2 seems fairly encouraging because it shows that (SMD) can terminate only at solutions of (P). Nevertheless, as we already saw in Section 1, $X(t)$ may fail to converge even in simple, one-dimensional problems. Even worse, this nonconvergent behavior may arise independently of the choice of mirror map Q: for instance, the objective function $f(x) = x^2/2$ over $\mathcal{X} = [-1, 1]$ yields the dynamics

$$dY = Q(\eta Y)\,dt + \sigma\,dW, \tag{3.8}$$

[6] For a concrete example of a nontrivial correlation structure arising in the study of traffic routing and congestion games, see [15].

	Hypothesis	Precise statement
(H1)	Lipschitz gradients	$v(x)$ is Lipschitz continuous
(H2)	Mirror map regularity	$F(p, y_n) \to 0$ whenever $Q(y_n) \to p$
(H3)	Noise regularity	$\sigma(x,t)$ is bounded and Lipschitz in x

Table 1. Overview of the regularity assumptions used in the paper.

where $\sigma(x,t)$ has been assumed constant for simplicity. In this example, even though the drift of (3.8) vanishes when $Q(\eta y) = 0$ (and only then), the martingale component does not, so $\arg\min f = \{0\}$ is not even invariant under (SMD). As a result, with probability 1, $X(t)$ fails to converge to a solution of (P).

In the rest of this paper, we will seek to bypass these basic impossibility results by examining the following questions:

(1) Analyze the "small noise" regime of (SMD), corresponding to cases where the input to (SMD) becomes more accurate over time.
(2) Study the ergodic properties of $X(t)$ and determine where the process spends most time with high probability.
(3) Establish the convergence of $X(t)$ in relevant subclasses of convex optimization problems (such as linear programs).
(4) Focus on a suitably rectified variant of (SMD) which anneals the adverse effects of $Z(t)$ by means of a decreasing sensitivity parameter.

These questions are treated in detail in Sections 4–7 below. For concision (and unless stated otherwise), we will be assuming throughout that (H1)–(H3) hold and that the feasible region $\mathcal{X}$ of (P) is compact. The noncompact case is discussed in Section 8; Hypotheses (H1)–(H3) can also be relaxed at certain points in our analysis, but we keep them "as is" for simplicity.

4. The small noise limit

We begin with the case where the gradient input to (SMD) becomes more accurate as measurements accrue over time – for instance, as in applications to wireless communications where the accumulation of pilot signals allows users to better sense their channel [31]. Our main result in this context is as follows:

Theorem 4.1. *If $\sigma(x,t) = o(1/\sqrt{\log t})$, then $X(t) \to \arg\min f$ (a.s.).*

The main idea of our proof is that, under the stated assumption for the volatility matrix $\sigma(x,t)$ of $Z(t)$, the process $Y(t)$ is an *asymptotic pseudotrajectory* (APT) of the deterministic dynamics (MD) [9, 10]. Heuristically, this means that $Y(t)$ shadows the flow of (MD) with arbitrary accuracy over any fixed horizon in the long run; more formally, we have:

Definition 4.2. Let $\Phi_t \colon \mathcal{Y} \to \mathcal{Y}$, $t \geq 0$, denote the semiflow induced by (MD) on $\mathcal{Y}$ (i.e. $(\Phi_t(y))_{t\geq 0}$ is the solution orbit of (MD) that starts at $y \in \mathcal{Y}$). Then, we say that $(Y(t))_{t\geq 0}$ is an *asymptotic pseudotrajectory* (APT) of Φ_t if

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \|Y(t+h) - \Phi_h(Y(t))\|_* = 0 \quad \text{for all } T > 0. \tag{4.1}$$

Of course, even though the solution orbits $x(t) = Q(y(t))$ of (MD) converge to a solution of (P) for any initial condition $y(0) \in \mathcal{Y}$, the same cannot be said

for any APT of (MD) – for instance, an APT of (MD) might escape to infinity. To overcome this obstacle, we show below that (MD) converges to $\arg\min f$ in a certain "uniform" sense, and we then use an induction argument to show that the same holds for every APT of (MD) that is induced by (SMD).

Proof of Theorem 4.1. Without loss of generality, assume that $\eta = 1$; otherwise, simply replace h by $\eta^{-1}h$ in the definition of (SMD). Also, for simplicity, we only prove the case where f admits a unique minimizer $x^* \in \mathcal{X}$; the general argument is similar (but more cumbersome to write down), so we omit it.

To begin, fix some $\varepsilon > 0$ and let $U_\varepsilon = \{x = Q(y) : F(x^*, y) < \varepsilon\}$. We first claim that there exists some finite $T \equiv T(\varepsilon)$ such that $F(x^*, \Phi_T(y)) \leq \max\{\varepsilon, F(x^*, y) - \varepsilon\}$ for all $y \in \mathcal{Y}$. Indeed, since x^* is the (unique) minimizer of f, there exists some $a \equiv a(\varepsilon) > 0$ such that

$$\langle v(x) \,|\, x - x^* \rangle \leq -a \quad \text{for all } x^* \in \mathcal{X}^*,\ x \notin U_\varepsilon. \tag{4.2}$$

Consequently, if $\tau_y = \inf\{t > 0 : Q(\Phi_t(y)) \in U_\varepsilon\}$ is the first time at which an orbit of (MD) hits U_ε, Lemma B.1 in Appendix B gives

$$F(x^*, \Phi_t(y)) - F(x^*, y) = \int_0^t \langle v(x(s)) \,|\, x(s) - x^* \rangle \, ds \leq -at \quad \text{for all } t \leq \tau_y. \tag{4.3}$$

In view of this, set $T = \varepsilon/a$ and consider the following cases:

(1) If $T \leq \tau_y$, (4.3) gives $F(x^*, \Phi_T(y)) \leq F(x^*, y) - \varepsilon$.
(2) If $T > \tau_y$, we have $F(x^*, \Phi_T(y)) \leq F(x^*, \Phi_{\tau_y}(y)) = \varepsilon$ (recall that $F(x^*, \Phi_t(y))$ is weakly decreasing in t).

In both cases we have $F(x^*, \Phi_T(y)) \leq \max\{\varepsilon, F(x^*, y) - \varepsilon\}$, as claimed.

Now, let $(Y(t))_{t\geq 0}$ be a solution of (SMD); we then claim that $Y(t)$ is (a.s.) an asymptotic pseudotrajectory of (MD) in the sense of Definition 4.2. Indeed, by Proposition 4.6 in [9], it suffices to show that $\int_0^\infty e^{-c/\Sigma_{\max}(t)}\, dt < \infty$ for all $c > 0$, where $\Sigma_{\max}(t) = \max_{x\in\mathcal{X}} \|\Sigma(x,t)\|_*$. However, by assumption, $\Sigma_{\max}(t) = \|\sigma(x,t)\sigma(x,t)^\top\|_* = \phi(t)/\log t$ for some $\phi(t)$ with $\lim_{t\to\infty} \phi(t) = 0$. Therefore,

$$e^{-c/\Sigma_{\max}(t)} = \left(e^{\log t}\right)^{-c/\phi(t)} = t^{-c/\phi(t)} = \mathcal{O}(t^{-\beta}) \quad \text{for all } \beta > 1, \tag{4.4}$$

and our assertion follows.

To proceed, fix a solution orbit $Y(t)$ of (SMD) which is an APT of (MD).[7] Moreover, with notation as in Definition 4.2 (and a bit of hindsight), let $\delta \equiv \delta(\varepsilon)$ be such that $\delta\|\mathcal{X}\| + \delta^2/(2K) \leq \varepsilon$ and choose some (random) $t_0 \equiv t_0(\varepsilon)$ such that $\sup_{0\leq h\leq T} \|Y(t+h) - \Phi_h(Y(t))\|_* \leq \delta$ for all $t \geq t_0$. Then, for all $t \geq t_0$, we get

$$\begin{aligned}
F(x^*, Y(t+h)) &\leq F(x^*, \Phi_h(Y(t))) \\
&\quad + \langle Y(t+h) - \Phi_h(Y(t)) \,|\, Q(\Phi_h(Y(t))) - x^* \rangle \\
&\quad + \frac{1}{2K} \|Y(t+h) - \Phi_h(Y(t))\|_*^2 \\
&\leq F(x^*, \Phi_h(Y(t))) + \delta\|\mathcal{X}\| + \frac{\delta^2}{2K} \\
&\leq F(x^*, \Phi_h(Y(t))) + \varepsilon,
\end{aligned} \tag{4.5}$$

[7]Formally, let $\Omega_0 = \{\omega \in \Omega : (Y(t))_{t\geq 0} \text{ is an APT of (MD)}\}$. By "fixing an orbit", we mean here "fix some $\omega \in \Omega_0$"; since $\mathbb{P}(\Omega_0) = 1$, what follows applies to $\mathbb{P}$-a.a. solution orbits of (SMD).

where, in the first line, we used the second-order Taylor estimate for the Fenchel coupling derived in Proposition A.2 (cf. Appendix A).

We now claim that there exists some (random) $T_0 \geq t_0$ such that $F(x^*, Y(T_0)) \leq 2\varepsilon$ (a.s.). Indeed, if this is not the case, we would have $\langle v(X(t)) \,|\, X(t) - x^* \rangle \leq -a$ for all $t \geq t_0$ with positive probability. Then, by Lemma B.1 in Appendix B, we would also have

$$\begin{aligned} F(x^*, Y(t)) &\leq F(x^*, Y(t_0)) \\ &\quad + \int_{t_0}^{t} \langle v(X(s)) \,|\, X(s) - x^* \rangle \, ds \\ &\quad + \frac{1}{2K} \int_{t_0}^{t} \|\Sigma(X(s), s)\|_* \, ds \\ &\quad + \sum_{i=1}^{n} \int_{t_0}^{t} (X_i(s) - x^*) \, dZ_j(s) \\ &\leq F(x^*, Y(t_0)) - a(t - t_0) + \frac{1}{2K} \int_{t_0}^{t} \|\Sigma(X(s), s)\|_* \, ds + \xi(t), \end{aligned} \tag{4.6}$$

where $\xi(t)$ denotes the martingale term $\sum_{i=1}^{n} \int_{t_0}^{t} (X_i(s) - x_i^*) \, dZ_i(s)$. Since $\|X(s) - x^*\| \leq \|\mathcal{X}\| < \infty$, Lemma C.2 in Appendix C shows that $\xi(t)/t \to 0$ (a.s.). Moreover, we also have $\lim_{t\to\infty} t^{-1} \int_0^t \|\Sigma(X(s), s)\|_* \, ds = \lim_{t\to\infty} \|\Sigma(X(t), t)\|_* = 0$ (by the small noise assumption and de l'Hôpital's rule), so the last two terms in Eq. (4.6) are both sublinear in t. We thus obtain $F(x^*, Y(t)) \to -\infty$ with positive probability, a contradiction.

On account of the above, we conclude that $F(x^*, Y(T_0)) \leq 2\varepsilon$ for some $T_0 \geq t_0$; hence, by (4.5), we get

$$F(x^*, Y(T_0 + h)) \leq F(x^*, \Phi_h(Y(T_0))) + \varepsilon \leq F(x^*, Y(T_0)) + \varepsilon \leq 3\varepsilon \tag{4.7}$$

for all $h \in [0, T]$. However, by the definition of T, we also have $F(x^*, \Phi_T(Y(T_0))) \leq \max\{\varepsilon, F(x^*, Y(T_0)) - \varepsilon\} \leq \varepsilon$, so $F(x^*, Y(T_0 + T)) \leq F(x^*, \Phi_T(Y(T_0))) + \varepsilon \leq 2\varepsilon$. Therefore, repeating the above argument at $T_0 + T$ (instead of T_0) and proceeding inductively, we get $F(x^*, Y(T_0 + h)) \leq 3\varepsilon$ for all $h \in [kT, (k+1)T]$, $k \in \mathbb{N}$. With ε arbitrary, we conclude that $F(x^*, Y(t)) \to 0$, so $X(t) \to x^*$ and our proof is complete. ■

For a numerical illustration of Theorem 4.1, see Fig. 2.

5. Long-run concentration properties

In contrast to the analysis of the previous section, if the noise in (SMD) is persistent, the simple examples (1.1) and (3.8) show that there is no hope of obtaining individual trajectory convergence for (SMD) with positive probability. Nevertheless, as we discussed in Section 1, $X(t)$ may converge to a random variable X_∞ whose distribution is concentrated around $\arg\min f$ – meaning in turn that $X(t)$ will be spending most of its time in a neighborhood of $\arg\min f$. With this in mind, our goal in this section will be to analyze the ergodicity properties of (SMD) and to determine the "concentration basin" of $X(t)$.

For reasons that will become clear in Section 6, we focus here on strongly convex problems that admit a (necessarily unique) interior solution $x^* \in \mathcal{X}^\circ$. More

concretely, this means that there exists some $L > 0$ such that

$$f(x) - f(x^*) \geq \tfrac{1}{2}L\|x - x^*\|^2 \quad \text{for all } x \in \mathcal{X}. \tag{5.1}$$

When this is the case, we have the following preliminary result:

Proposition 5.1. *For every initial condition $y_0 \in \mathcal{Y}$, we have*

$$\mathbb{E}\left[\frac{1}{t}\int_0^t \|X(s) - x^*\|^2\, ds\right] \leq \frac{2F(x^*, \eta y_0)}{\eta L t} + \frac{\eta\sigma_*^2}{KL}, \tag{5.2}$$

i.e. the long-run average of $X(t)$ lies within $\eta\sigma_*^2/(KL)$ of x^*. Moreover, let

$$\tau_\delta = \inf\{t > 0 : \|X(t) - x^*\| \leq \delta\} \tag{5.3}$$

be the first time at which $X(t)$ gets within $\delta > 0$ of x^*. If $\eta < KL\delta^2/\sigma_*^2$, we have

$$\mathbb{E}[\tau_\delta] \leq \frac{2KF(x^*, \eta y_0)}{\eta KL\delta^2 - \eta^2\sigma_*^2}. \tag{5.4}$$

In particular, if $y_0 = 0$ and $\eta = KL\delta^2/(2\sigma_*^2)$, we have

$$\mathbb{E}[\tau_\delta] \leq \frac{8\sigma_*^2\Omega}{KL^2\delta^4}, \tag{5.5}$$

where $\Omega = \max\{h(x') - h(x) : x, x' \in \mathcal{X}\}$.

Remark 5.1. For a value-based analogue of (5.2) in the steep case, see [39, Prop. 4].

Proof. Let $F_\eta(t) \equiv \eta^{-1}F(x^*, \eta Y(t))$ denote the η-deflated Fenchel coupling between x^* and $Y(t)$. Then, by the growth bound (C.2) in Appendix C, we get

$$\begin{aligned} F_\eta(t) - F_\eta(0) &\leq \int_0^t \langle v(X(s)) \,|\, X(s) - x^*\rangle\, ds + \frac{1}{2K}\int_0^t \eta\|\Sigma(X(s), s)\|_*\, ds + \xi(t) \\ &\leq -\frac{L}{2}\int_0^t \|X(s) - x^*\|^2\, ds + \frac{\eta\sigma_*^2 t}{2K} + \xi(t), \end{aligned} \tag{5.6}$$

where $\xi(t) = \sum_{i=1}^n \int_0^t (X_i(s) - x_i^*)\, dZ_i(s)$ and we used the strong convexity bound (5.1) to write $\langle v(x) \,|\, x - x^*\rangle \leq f(x^*) - f(x) \leq -\frac{1}{2}L\|x - x^*\|^2$ in the second line. Since $F_\eta(t) \geq 0$, the bound (5.2) follows by taking expectations and rearranging.

Now, replacing t by $\tau_\delta \wedge t$ in (5.6), we also get

$$\mathbb{E}[F_\eta(\tau_\delta \wedge t)] \leq F_\eta(0) - \frac{L}{2}\,\mathbb{E}\left[\int_0^{\tau_\delta \wedge t} \|X(s) - x^*\|^2\, ds\right] + \frac{\eta\sigma_*^2}{2K}\,\mathbb{E}[\tau_\delta \wedge t] \tag{5.7}$$

$$\leq F_\eta(0) + \frac{\eta\sigma_*^2 - KL\delta^2}{2K}\,\mathbb{E}[\tau_\delta \wedge t], \tag{5.8}$$

where we used the fact that $\|X(s) - x^*\| \geq \delta$ for all $s \leq \tau_\delta$. Since $F_\eta \geq 0$, we conclude that

$$\frac{KL\delta^2 - \eta\sigma_*^2}{2K}\,\mathbb{E}[\tau_\delta \wedge t] \leq F_\eta(0), \tag{5.9}$$

and our claim follows by taking $t \to \infty$ (so $\tau_\delta \wedge t \to \tau_\delta$) and using the dominated convergence theorem to interchange limits and expectations. ∎

By Jensen's inequality, the mean squared error estimate (5.2) implies that the time-averaged process $\bar{X}(t) = t^{-1}\int_0^t X(s)\, ds$ of $X(t)$ enjoys the bound

$$\mathbb{E}[\|\bar{X}(t) - x^*\|] \leq \sqrt{\mathbb{E}\left[\frac{1}{t}\int_0^t \|X(s) - x^*\|^2\, ds\right]} \lesssim \sigma_*\sqrt{\frac{\eta}{KL}}. \tag{5.10}$$

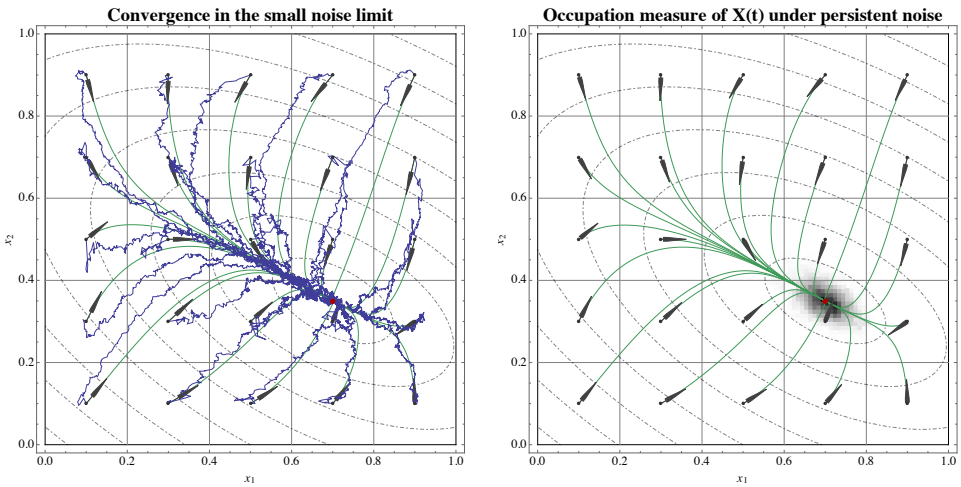

Figure 2. Numerical illustration of the behavior of (SMD) with mirror map $Q(y) = e^y/(1+e^y)$. In both figures, the dashed contours represent the level sets of f over $\mathcal{X} = [0,1]^2$, and the stream lines indicate the flow of (MD). In the first figure, we exhibit the convergence of (SMD) to $\arg\min f$ when the volatility of the noise decays as $\Theta(1/t)$. In the second, we estimate the long-run occupation measure $\mu_\infty = \lim_{t\to\infty} t^{-1}\int_0^t \mathbb{1}_{X(s)}\,ds$ of X: darker shades of gray correspond to higher probabilities of observing X in a given region.

Setting $\delta_\eta = \sigma_*\sqrt{\eta/(KL)}$ allows us to conclude that $\bar{X}(t)$ converges in L^1 to a δ_η–neighborhood of x^*, but we cannot deduce much more than that from (5.2). Instead, to obtain finer information regarding the concentration of the actual process $X(t)$ around x^*, we consider below its *occupation measure*:

Definition 5.2. The *occupation measure* of X up to time $t \geq 0$ is defined as

$$\mu_t(A) = \frac{1}{t}\int_0^t \mathbb{1}(X(s)\in A)\,ds \quad \text{for every Borel } A \subseteq \mathcal{X}. \tag{5.11}$$

In words, $\mu_t(A)$ is simply the fraction of time that X spends in the set A. As such, the asymptotic concentration of X around x^* can be estimated by the quantities $\mu_t(\mathbb{B}_\delta)$ where $\mathbb{B}_\delta = \{x \in \mathcal{X} : \|x - x^*\| \leq \delta\}$ is a δ-ball centered at x^* in $\mathcal{X}$. In this direction, our main result is as follows:

Theorem 5.3. *Fix some $\delta > 0$ and assume that the infinitesimal covariance matrix Σ of (SMD) is time-homogeneous and uniformly positive-definite (i.e. $\Sigma(x,t) \equiv \Sigma(x) \succcurlyeq \lambda I$ for some $\lambda > 0$). If (SMD) is run with sensitivity $\eta < KL\delta^2/\sigma_*^2$, then*

$$\mu_t(\mathbb{B}_\delta) \gtrsim 1 - \frac{\eta\sigma_*^2}{KL\delta^2} \quad \textit{for sufficiently large } t \ (\textit{a.s.}). \tag{5.12}$$

Corollary 5.4. *Fix some tolerance level $\varepsilon > 0$. If (SMD) is run with sensitivity $\eta \leq \varepsilon KL\delta^2/\sigma_*^2$, we have $\mu_t(\mathbb{B}_\delta) \geq 1 - \varepsilon$ for all sufficiently large t (a.s.).*

Remark 5.2. Since $\Sigma = \sigma\sigma^\top$, it follows that Σ is nonnegative-definite by default. The stronger assumption $\Sigma \succcurlyeq \lambda I$ essentially posits that the volatility matrix σ of Z has $\operatorname{rank}(\sigma) = n$, i.e. that the components of Z are not totally correlated.

Heuristically, the proof of Theorem 5.3 hinges on the following reasoning: For $\eta < KL\delta^2/\sigma_*^2$, Proposition 5.1 shows that the ball $\mathbb{B}_\delta$ is *recurrent* in the sense that $\mathbb{P}(X(t) \in \mathbb{B}_\delta \text{ for some } t \geq 0) = 1$ for every initial condition $y_0 \in \mathcal{Y}$. However, since the set $Q^{-1}(\mathbb{B}_\delta)$ is not compact in general,[8] the generating process $Y(t)$ need not be itself recurrent.[9] Nonetheless, under a suitable transformation which "mods out" the directions in $\mathcal{Y}$ that annihilate the linear hull of $\mathcal{X}$, the process $Y(t)$ *becomes* recurrent and admits a unique invariant distribution ν.[10] The pushforward of ν to $\mathcal{X}$ is precisely the limit of the occupation measures μ_t as $t \to \infty$, so (5.12) follows by using the mean square bound (5.2) to estimate ν.

Proof of Theorem 5.3. We begin by introducing a transformed version of $Y(t)$ which is recurrent under (SMD). To that end, note first that $Q^{-1}(x)$ always contains a translate of the polar cone $\mathrm{PC}(x)$ of $\mathcal{X}$ at x (cf. Lemma A.1); in particular, if $\mathcal{X}$ is not full-dimensional, $Q^{-1}(\mathbb{B}_\delta)$ contains a nonzero affine subspace of $\mathcal{Y}$. To mod out this subspace, let $\mathcal{V}_0 = \operatorname{aff}(\mathcal{X} - \mathcal{X}) \subseteq \mathcal{V}$ denote the smallest subspace of $\mathcal{V}$ that contains $\mathcal{X}$ when translated to the origin (so $\mathcal{X}$ may be considered as a convex body of $\mathcal{V}_0$). Then, writing $\mathcal{Y}_0 \equiv \mathcal{V}_0^*$ for the dual space of $\mathcal{V}_0$, define the restriction map $\pi_0 \colon \mathcal{Y} \to \mathcal{Y}_0$ as

$$\langle \pi_0(y) \,|\, z \rangle = \langle y \,|\, z \rangle \quad \text{for all } z \in \mathcal{V}_0. \tag{5.13}$$

We then have $\pi_0(y) = 0$ whenever y annihilates $\mathcal{V}_0$ (i.e. $\langle y \,|\, z \rangle = 0$ for all $z \in \mathcal{V}_0$), so π_0 essentially "forgets" the annihilator of $\mathcal{V}_0$ in $\mathcal{Y}$.[11]

Accordingly, in view of Proposition 5.1, it stands to reason that the transformed process $\Psi(t) = \pi_0(Y(t))$ is recurrent. Indeed, from [11, Proposition 3.1], it suffices to show that *a*) $\Psi(t)$ is an Itô diffusion whose infinitesimal generator is uniformly elliptic; and *b*) there exists some compact set $C_0 \in \mathcal{Y}_0$ such that $\mathbb{P}(\Psi(t) \in C_0 \text{ for some } t \geq 0) = 1$ for every initial condition $\psi_0 \in \mathcal{Y}_0$. The rest of our proof is devoted to establishing these two requirements.

For the first, write $\pi_0(y)$ in coordinates as $(\pi_0(y))_i = \sum_{k=1}^n \Pi_{ik} y_k$. Then, with $\Psi = \Pi \cdot Y$, we get

$$d\Psi_i = \sum_{k=1}^n \Pi_{ik} \left(v_k(X)\, dt + \, dZ_k \right), \tag{5.14}$$

where $X(t) = Q(\eta Y(t))$. Moreover, define the "restricted" mirror map $Q_0 \colon \mathcal{Y}_0 \to \mathcal{X}$ as

$$Q_0(w) = \arg\max\{\langle w \,|\, x \rangle - h(x) : x \in \mathcal{X}\} \quad \text{for all } w \in \mathcal{Y}_0, \tag{5.15}$$

where, in a slight abuse of notation, $\mathcal{X}$ is treated as a subset of $\mathcal{V}_0$. By definition, we have $\langle y \,|\, x \rangle = \langle \pi_0(y) \,|\, x \rangle$ for all $x \in \mathcal{X}$, so $\arg\max\{\langle y \,|\, x \rangle - h(x)\} = \arg\max\{\langle \pi_0(y) \,|\, x \rangle - h(x)\}$ for all $y \in \mathcal{Y}$. This shows that $X(t)$ can be expressed as $X(t) = Q_0(\eta \pi_0(Y(t))) = Q_0(\eta \Psi(t))$, so (5.14) represents a regular Itô diffusion.

---

[8] For instance, take $\mathcal{X} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1, x_2 \geq 0,\ x_1 + x_2 = 1\}$ and consider the mirror map $Q(y_1, y_2) = (e^{y_1}, e^{y_2})/(e^{y_1} + e^{y_2})$ generated by the regularizer $h(x_1, x_2) = x_1 \log x_1 + x_2 \log x_2$. Then, the inverse image $Q^{-1}(1/2, 1/2) = \{(y_1, y_2) \in \mathbb{R}^2 : y_1 = y_2\}$ of $(1/2, 1/2)$ is unbounded.

[9] Recall here that $Y(t)$ is called *recurrent* if there exists a compact set C such that $\mathbb{P}(Y(t) \in C \text{ for some } t \geq 0) = 1$ for every initial condition y_0 of Y [11, 27].

[10] The reason that this procedure cannot be applied directly to the primal process $X(t)$ is that $X(t)$ may fail to be an Itô diffusion if h is nonsteep.

[11] Of course, if $\mathcal{X}$ has nonempty interior as a subset of $\mathcal{V}$ (so $\mathcal{V}_0 = \mathcal{V}$), there is nothing to forget and π_0 is simply the identity function on $\mathcal{Y} = \mathcal{Y}_0$.

We now claim that the infinitesimal generator $\mathcal{L}_\Psi$ of Ψ is uniformly elliptic. Indeed, the quadratic covariation of Ψ is given by

$$d[\Psi_i, \Psi_j] = d(\Pi Y)_i \, d(\Pi Y)_j = \sum_{k,\ell=1}^{n} \Pi_{ik}\Pi_{j\ell}\Sigma_{k\ell} \, dt = \left(\Pi\Sigma\Pi^\top\right)_{ij} dt, \tag{5.16}$$

where we used the definition (3.3) of Σ in the penultimate equality. However, we also have

$$\Pi\Sigma\Pi^\top \succcurlyeq \lambda\Pi\Pi^\top \succcurlyeq \lambda\pi_{\min}^2 I, \tag{5.17}$$

where $\pi_{\min} > 0$ denotes the smallest singular value of $\Pi^\top$ (recall that π_0 has full rank by construction). This shows that the principal symbol $\Pi\Sigma\Pi^\top$ of $\mathcal{L}_\Psi$ is uniformly positive-definite, so $\mathcal{L}_\Psi$ is uniformly elliptic.

For the second component of our proof, assume without loss of generality that δ is sufficiently small so that $\mathbb{B}_\delta \subseteq \mathcal{X}^\circ$. Since $\mathcal{X}$ may be viewed as a convex body of $\mathcal{V}_0$, Remark 6.2.3 in [22] implies that the set $C_0 = \eta^{-1}\partial h(\mathbb{B}_\delta)$ is compact.[12] Then, by Proposition 5.1, it follows that $\Psi(t)$ hits C_0 in finite time (a.s.) for every initial condition $\psi_0 = \pi_0(y_0) \in \mathcal{Y}_0$.

Since the generator $\mathcal{L}_\Psi$ of Ψ is uniformly elliptic and C_0 is compact, Proposition 3.1 in [11] shows that $\Psi(t)$ is recurrent. Hence, from standard results in the theory of Itô diffusions [27, Theorem 4.4.1, Theorem 4.4.2 and Corollary 4.4.4], we conclude that $\Psi(t)$ admits a unique invariant distribution ν which satisfies the law of large numbers

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \phi(\Psi(s)) \, ds = \int_{\mathcal{Y}_0} \phi \, d\nu \quad \text{(a.s.)}, \tag{5.18}$$

for every ν-integrable test function ϕ on $\mathcal{Y}_0$. We thus obtain

$$\begin{aligned}\lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{1}(X(s) \in \mathbb{B}_\delta) \, ds &= \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{1}(\eta\Psi(s) \in Q_0^{-1}(\mathbb{B}_\delta)) \, ds \\ &= \int_{\mathcal{Y}_0} \mathbb{1}_{\eta^{-1}Q_0^{-1}(\mathbb{B}_\delta)} \, d\nu \\ &= \nu(\eta^{-1}Q_0^{-1}(\mathbb{B}_\delta)),\end{aligned} \tag{5.19}$$

i.e. $\mu_t(\mathbb{B}_\delta) \to \nu(\eta^{-1}Q_0^{-1}(\mathbb{B}_\delta))$ as $t \to \infty$ (a.s.). Similarly, given that the limit of μ_t is deterministic and finite, the mean square bound (5.2) also yields

$$\begin{aligned}1 - \nu(\eta^{-1}Q_0^{-1}(\mathbb{B}_\delta)) &= \lim_{t\to\infty} \frac{1}{t}\,\mathbb{E}\left[\int_0^t \mathbb{1}(X(s) \notin \mathbb{B}_\delta) \, ds\right] \\ &\le \lim_{t\to\infty} \frac{1}{t}\,\mathbb{E}\left[\int_0^t \frac{\|X(s) - x^*\|^2}{\delta^2} \, ds\right] \le \frac{\eta\sigma_*^2}{KL\delta^2},\end{aligned} \tag{5.20}$$

and our assertion follows. ■

We close this section with a few remarks regarding Theorem 5.3 (for a numerical illustration, see Fig. 2). First, the parameters of the problem are reflected in a particularly suggestive way in the sensitivity threshold $\eta_* = KL\delta^2/\sigma_*^2$ (which in turn determines how aggressively we can choose η while maintaining an acceptable concentration level around x^*). Indeed, if the noise variance σ_*^2 is small and the

[12] Strictly speaking, Remark 6.2.3 of [22] applies to convex functions that are defined on all of $\mathcal{V}_0$, but since this is a local property, it is trivial to extend it to our case.

strong convexity constant L is large, it should be easier to approximate a solution of (P). Thus, keeping in mind that η essentially acts as an accelerant for $X(t)$, it stands to reason that η_* is decreasing in σ_*^2 and increasing in L and δ.

Finally, the assumption that x^* is interior is crucial in the statement of Theorem 5.3. As we shall see in the next section, if x^* is a corner of $\mathcal{X}$ (i.e. $\mathrm{PC}(x^*)$ has nonempty interior), $Y(t)$ is *transient* (not recurrent) and $X(t)$ *converges* to x^* (instead of fluctuating in a small neighborhood thereof). Otherwise, when x^* belongs to a nontrivial face of $\mathcal{X}$, the dynamics (SMD) exhibit a hybrid behavior: $X(t)$ converges to the smallest face of $\mathcal{X}$ that contains x^* and fluctuates around x^* along the relative interior of said face. That said, obtaining a precise result along these lines is fairly cumbersome, so we only discuss the flip side of Theorem 5.3 in Section 6.

6. Linear programming and robust solutions

In contrast to the non-convergent gradient descent example (1.1), the linear objective $f(x) = 1 - x$ over $\mathcal{X} = [0, 1]$ yields

$$dY = dt + \sigma\, dW, \tag{6.1}$$

or, after integrating, $Y(t) = t + W(t)$. A trivial estimate then yields $Y(t) \geq t/2$ for all sufficiently large t (except possibly on a $\mathbb{P}$-null set), so $\lim_{t\to\infty} X(t) = 1$ (a.s.), independently of the choice of mirror map Q (cf. Lemma 6.7 below). In other words, in this simple linear program, $X(t)$ converges to $\arg\min f$ with probability 1, no matter the magnitude of the noise.

The main reason for this disparity between (1.1) and (6.1) is that the drift of the latter does not vanish when $X(t)$ approaches $\arg\min f$, so it ends up dominating the dynamics' long-term behavior. A nonvanishing gradient is typical of (generic) linear programs, so one would optimistically expect comparable results to hold whenever (P) can be locally approximated by a linear program. With this in mind, we focus here on solutions of (P) that are "robust" in the following sense:

Definition 6.1. We say that $x^* \in \mathcal{X}$ is a *robust solution* of (P) if $\langle v(x^*) \,|\, z\rangle < 0$ for every nonzero tangent vector $z \in \mathrm{TC}(x^*)$.

In the above, the term "robustness" alludes to the fact that if the objective f of (P) is perturbed by a function g with sufficiently small gradient, x^* remains a minimizer of $f + g$. We then have:

Theorem 6.2. *If x^* is a robust solution of* (P) *and* (SMD) *is run with a sufficiently low sensitivity parameter η, $X(t)$ converges to x^* (a.s.). In addition, if the mirror map Q is surjective, this convergence occurs in finite time (a.s.).*

Corollary 6.3. *If* (P) *is a generic linear program, $X(t)$ converges to* $\arg\min f$ *with probability* 1 *(and in finite time if Q is surjective).*

The proof of Theorem 6.2 is somewhat involved so we provide a rough sketch below. First, if x^* is a robust solution of (P), note that the negative gradient field $v(x)$ is "strictly outward pointing" near x^* in the sense that $v(x) \in \mathrm{int}(\mathrm{PC}(x^*))$ for all x close to x^*. Using this property, we show that almost every solution trajectory of (SMD) has an ω-limit arbitrarily close to x^* if η is chosen small enough. In terms of the generating dual process $Y(t)$, this means that affine translates of the polar cone $\mathrm{PC}(x^*)$ in the dual space $\mathcal{Y}$ are recurrent under (SMD). Combining this

recurrence property with a hitting time argument based on Girsanov's theorem, it then follows that $Y(t)$ escapes to infinity along a proper subcone of $\mathrm{PC}(x^*)$ with probability 1. Having established all this, a straightforward geometric argument can then be used to show that $X(t)$ converges to x^* (a.s.).

In what follows, we encode the above in a series of technical lemmas. We begin with the following immediate observation:

Lemma 6.4. *$x^* \in \mathcal{X}$ is a robust solution of (P) if and only if*

$$\langle v(x^*) \,|\, z \rangle \leq -\gamma \|z\| \quad \textit{for some } \gamma > 0 \textit{ and for all } z \in \mathrm{TC}(x^*). \tag{6.2}$$

Proof. The "if" part is trivial. For the "only if" part, let $\gamma = \min\{|\langle v(x^*) \,|\, z\rangle| : z \in \mathrm{TC}(x^*),\, \|z\| = 1\}$ so $\langle v(x^*) \,|\, z\rangle \leq -\gamma\|z\|$ for all $z \in \mathrm{TC}(x^*)$. Since x^* is a robust solution, it follows that $\gamma > 0$, as was to be shown. ∎

Remark 6.1. In view of Lemma 6.4, a point that satisfies (6.2) is called *γ-robust*.

We now show that neighborhoods of robust solutions are recurrent under (SMD):

Lemma 6.5. *Fix some precision tolerance $\delta > 0$. If x^* is a γ-robust solution of (P) and (SMD) is run with sensitivity parameter $\eta < 2\gamma\delta K/\sigma_*^2$, there exists a (random) sequence of times $t_n \uparrow \infty$ such that $\|X(t_n) - x^*\| < \delta$ for all n (a.s.).*

Proof. Suppose there exists some (random) t_0 such that $\|X(t) - x^*\| \geq \delta$ for all $t \geq t_0$. Then, writing $F_\eta(t) = \eta^{-1} F(x^*, \eta Y(t))$ for the η-deflated Fenchel coupling between x^* and $Y(t)$, Lemma C.1 yields

$$\begin{aligned} F_\eta(t) &\leq F_\eta(t_0) + \int_{t_0}^t \langle v(X(s)) \,|\, X(s) - x^* \rangle \, ds + \frac{1}{2K} \int_{t_0}^t \eta \|\Sigma(X(s), s)\|_* \, ds + \xi(t) \\ &\leq F_\eta(t_0) - \gamma\delta\,(t - t_0) + \frac{\eta\sigma_*^2}{2K}(t - t_0) + \xi(t) \\ &\leq F_\eta(t_0) - \left[\gamma\delta - \frac{\eta\sigma_*^2}{2K} - \frac{\xi(t)}{t - t_0}\right](t - t_0), \end{aligned} \tag{6.3}$$

where we set $\xi(t) = \sum_{i=1}^n \int_{t_0}^t (X_i(s) - x_i^*) \, dZ_i(s)$ in the first line and we used Lemma 6.4 in the second. Since $\xi(t)/(t - t_0) \to 0$ by the asymptotic estimate (C.4) in Appendix C, the bound (6.3) yields $\lim_{t\to\infty} F_\eta(t) = -\infty$ if $\eta\sigma_*^2 < 2\gamma\delta K$, a contradiction (recall that $F_\eta(t) \geq 0$ for all $t \geq 0$). This shows that $t_0 = \infty$ (a.s.), so there exists a sequence $t_n \uparrow \infty$ such that $\|X(t_n) - x^*\| < \delta$ for all n. ∎

Our next result is a technical lemma which shows that the generating process $Y(t)$ keeps moving roughly along the direction of $v(x^*)$ with probability arbitrarily close to 1 if η is chosen small enough and $X(0)$ starts sufficiently close to x^* (cf. Fig. 3):

Lemma 6.6. *Let x^* be a robust solution of (P) and let C be a polyhedral cone such that $v(x^*) \in \mathrm{int}(C)$ and $C \subseteq \mathrm{int}(\mathrm{PC}(x^*)) \cup \{0\}$. Then, for small enough $\eta, \varepsilon, \delta > 0$, and for every initial condition $y_0 \in \mathcal{Y}$ with $F(x^*, \eta y_0) < \varepsilon$, there exists some $y \in \mathcal{Y}$ such that $F(x^*, \eta y) = \varepsilon + \delta$ and*

$$\mathbb{P}(Y(t) \in y + C \textit{ for all } t \geq 0) \geq 1 - e^{-\kappa\delta/(\eta\sigma_*^2)}, \tag{6.4}$$

where $\kappa > 0$ is a constant that depends only on C and (P).

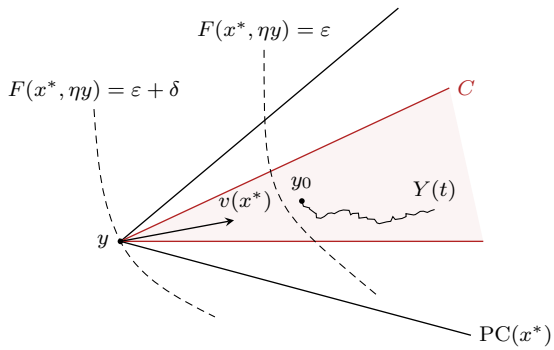

Figure 3. The various sets in the proof of Lemma 6.6.

Proof. Let $C^\circ = \{z \in \mathcal{V} : \langle y \,|\, z\rangle \leq 0 \text{ for all } y \in C\}$ denote the polar cone of C and let $\mathcal{U} = \{u_j\}_{j=1}^d$ be a basis for C° (recall that C is assumed polyhedral). Further, fix a small compact neighborhood L of x^* such that $v(L) \subseteq \operatorname{int}(C)$, pick $\gamma_L > 0$ so that $\langle v(x) \,|\, z\rangle \leq -\gamma_L \|z\|$ for all $x \in L$, $z \in C^\circ$,[13] and with a fair bit of hindsight, take $\delta < K\|\mathcal{X}\|^2$ sufficiently small so that $Q(\eta y) \in L$ whenever $F(x^*, \eta y) \leq \varepsilon + \delta$. Finally, invoking Lemma A.3 in Appendix A, let $y = y_0 - cv(x^*)$ for some $c > 0$ such that $F(x^*, \eta y) = \varepsilon + \delta$; then, by (A.3), we will also have

$$\|y_0 - y\|_* = c\|v(x^*)\|_* \geq \frac{K\|\mathcal{X}\|}{\eta}\left[\sqrt{1 + 2\delta/(K\|\mathcal{X}\|^2)} - 1\right] \geq \frac{\delta}{2\eta\|\mathcal{X}\|}, \tag{6.5}$$

where we used the fact that $\delta < K\|\mathcal{X}\|^2$ in the last inequality.

To proceed, set $\tau_C = \inf\{t \geq 0 : Y(t) \notin y + C\}$ and let $G_u(t) = \langle Y(t) - y \,|\, u\rangle$, so $\tau_C = \inf\{t \geq 0 : G_u(t) > 0 \text{ for some } u \in \mathcal{U}\}$. Then, for all $t \leq \tau_C$, we have

$$G_u(t) = G_u(0) + \int_0^t \langle v(X(s)) \,|\, u\rangle \, ds + \xi_u(t) \leq -A\|u\| - B\|u\|t + \xi_u(t), \tag{6.6}$$

where we have set $A = c \min_{u' \in \mathcal{U}} |\langle v(x^*) \,|\, u'\rangle|$, $B = \gamma_L$, and $\xi_u(t) = \langle Z(t) \,|\, u\rangle$. Arguing as in the proof of Lemma C.2, the Dambis–Dubins–Schwarz time-change theorem for martingales [26, Theorem 3.4.6] implies that there exists a standard Wiener process $W_u(t)$ such that $\xi_u(t) = W_u(\rho_u(t))$, where $\rho_u = [\xi_u, \xi_u]$ denotes the quadratic variation of ξ_u. By (6.6), this further implies that $G_u(t) \leq 0$ whenever $W_u(\rho_u(t)) \leq A\|u\| + B\|u\|t$; hence, $\tau_C = \infty$ whenever $W_u(\rho_u(t)) \leq A\|u\| + B\|u\|t$. Moreover, note that

$$d\rho_u = d\xi_u \cdot d\xi_u = \sum_{i,j=1}^n \Sigma_{ij} u_i u_j \, dt, \tag{6.7}$$

so $\rho_u(t) \leq \sigma_*^2 \|u\|^2 t$. Hence, if a trajectory of W_u is such that $W_u(t) \leq A\|u\| + \frac{B}{\|u\|\sigma_*^2} t$ for all $t \geq 0$, we will also have

$$W_u(\rho_u(t)) \leq A\|u\| + \frac{B}{\|u\|\sigma_*^2}\rho(t) \leq A\|u\| + B\|u\|t \quad \text{for all } t \geq 0. \tag{6.8}$$

[13] That such a γ_L exists is a consequence of the continuity of $v(x)$ and Lemma 6.4.

Therefore, to prove the lemma, it suffices to establish a suitable lower bound for the probability $\mathbb{P}(W_u(t) \leq A\|u\| + Bt/(\|u\|\sigma_*^2)$ for all $t \geq 0)$. To do so, let

$$\tau'_C = \inf\left\{t > 0 : W_u(t) = A\|u\| + \frac{B}{\|u\|\sigma_*^2}t \text{ for some } u \in \mathcal{U}\right\} \tag{6.9}$$

and write E_u for the event "$W_u(t) \geq A\|u\| + Bt/(\|u\|\sigma_*^2)$ for some finite $t \geq 0$". By a standard application of Girsanov's theorem for Wiener processes with drift [26, p. 197], we get $\mathbb{P}(E_u) = e^{-2AB/\sigma_*^2}$ and hence

$$\mathbb{P}(\tau'_C < \infty) = \mathbb{P}\Big(\bigcup\nolimits_{u\in\mathcal{U}} E_u\Big) \leq \sum\nolimits_{u\in\mathcal{U}} \mathbb{P}(E_u) = |\mathcal{U}|e^{-2AB/\sigma_*^2}. \tag{6.10}$$

Now, from the bound (6.5) and the definition of A and B, we have

$$AB = c\gamma_L \min_{u'\in\mathcal{U}} |\langle v(x^*)\,|\,u'\rangle| \geq \frac{\delta}{2\eta\|\mathcal{X}\|}\frac{\gamma_L \min_{u'\in\mathcal{U}}|\langle v(x^*)\,|\,u'\rangle|}{\|v(x^*)\|_*} = \frac{\kappa\delta}{\eta}, \tag{6.11}$$

where we set $\kappa = \frac{\gamma_L \min_{u'\in\mathcal{U}}|\langle v(x^*)\,|\,u'\rangle|}{2\|v(x^*)\|_*\|\mathcal{X}\|}$. By backtracking, we then get

$$\mathbb{P}(\tau_C = \infty) \geq \mathbb{P}(\tau'_C = \infty) \geq 1 - e^{-\kappa\delta/(\eta\sigma_*^2)} \tag{6.12}$$

provided that $\eta \leq \kappa\delta/(\sigma_*^2 \log|\mathcal{U}|)$. Therefore, with $\mathbb{P}(Y(t) \in y + C$ for all $t \geq 0) = \mathbb{P}(\tau_C = \infty)$, our proof is complete. ■

The final ingredient of our proof is that if $Y(t)$ moves deep within $\mathrm{PC}(x^*)$, the induced trajectory $X(t) = Q(\eta Y(t))$ converges to x^*:

Lemma 6.7. *Let $(y_n)_{n=1}^\infty$ be a sequence in $\mathcal{Y}$ such that $\langle y_n\,|\,z\rangle \to -\infty$ for all $z \in \mathrm{TC}(x^*)$. Then, $\lim Q(y_n) = x^*$.*

Proof. By compactness of $\mathcal{X}$ (and passing to a subsequence if necessary), we may assume that $x_n \equiv Q(y_n)$ converges in $\mathcal{X}$. Assume therefore that $x_n \to x' \neq x^*$, so $\liminf\|x_n - x^*\| > 0$. Then, with $y_n \in \partial h(x_n)$ by Proposition 2.2, we obtain

$$h(x^*) \geq h(x_n) + \langle y_n\,|\,x^* - x_n\rangle \geq h(x_n) - \langle y_n\,|\,z_n\rangle\|x_n - x^*\|, \tag{6.13}$$

where we have set $z_n = (x_n - x^*)/\|x_n - x^*\|$. Since z_n lives in the unit sphere of $\|\cdot\|$, compactness (and passing to a further subsequence if necessary) guarantees the existence of some $z \in \mathrm{TC}(x^*)$ with $\|z\| = 1$ and such that $\langle y_n\,|\,z_n\rangle \leq \langle y_n\,|\,z\rangle$ for all n (recall that $\mathrm{TC}(x^*)$ is closed). We thus get $h(x^*) \geq h(x_n) - \langle y_n\,|\,z\rangle\|x_n - x^*\|$ and, taking lim inf on both sides, we obtain $\liminf h(x^*) = \infty$, a contradiction. ■

We are now in a position (finally!) to prove the main result of this section:

Proof of Theorem 6.2. As in the proof of Lemma 6.6, let L be a sufficiently small compact neighborhood of x^* such that $v(L) \subseteq \mathrm{int}(\mathrm{PC}(x^*))$, i.e. $\langle v(x)\,|\,z\rangle \leq -\gamma_L\|z\|$ for some $\gamma_L > 0$ and for all $x \in L$, $z \in \mathrm{TC}(x^*)$. Then, by compactness, there exists a convex cone $C \subseteq \mathrm{int}(\mathrm{PC}(x^*))$ such that $\langle v(x)\,|\,z\rangle \leq -\gamma_L\|z\|$ for all $x \in L$ and for all $z \in C^\circ$.

With this in mind, pick $\varepsilon, \delta > 0$ sufficiently small so that the conclusion of Lemma 6.6 holds and $Q(\eta y) \in L$ whenever $F(x^*, \eta y) \leq \varepsilon + \delta$. If η is also chosen small enough, combining (H2) with Lemma 6.5 shows that there exists a (random) sequence of times $t_n \uparrow \infty$ such that $F(x^*, \eta Y(t_n)) \leq \varepsilon$ for all n (a.s.). Hence, by Lemma 6.6 and the strong Markov property of $Y(t)$, there exists some $a > 0$ such

that $\mathbb{P}(F(x^*, \eta Y(t_n + t)) \leq \varepsilon + \delta$ for all $t \geq 0) \geq 1-(1-a)^n$ for all n.[14] Thus, with notation as in (6.6), we get

$$G_z(t_n + t) \leq -A\|z\| - B\|z\|t + \xi_z(t) \quad \text{for all } t \geq 0, \tag{6.14}$$

with probability at least $1-(1-a)^n$. In turn, Lemma C.2 yields $\xi_z(t)/t \to 0$ (a.s.), showing that $\lim_{t\to\infty} G_z(t_n+t) = -\infty$. Since the above holds for all n, we conclude that $\langle Y(t) \,|\, z\rangle \to -\infty$ for all $z \in \mathrm{TC}(x^*)$ with probability 1, so $X(t) \to x^*$ (a.s.) by Lemma 6.7.

We are left to show that this convergence occurs in finite time if Q is surjective.[15] To that end, note first that if $x^* = Q(\eta y^*)$, we also have $Q(\eta(y^* + v)) = x^*$ for all $v \in \mathrm{PC}(x^*)$ by Lemma A.1. Therefore, it suffices to show that, for some y^* such that $Q(\eta y^*) = x^*$, we have $Y(t) \in y^* + \mathrm{PC}(x^*)$ for all sufficiently large t (a.s.). However, since $X(t) \to x^*$, there exists some t_0 such that $X(t) \in L$ for all $t \geq t_0$. Thus, for all $z \in \mathrm{TC}(x^*)$ with $\|z\| = 1$, we get

$$\begin{aligned}\langle Y(t) \,|\, z\rangle &= \langle Y(t_0) \,|\, z\rangle + \int_{t_0}^{t} \langle v(X(s)) \,|\, z\rangle \, ds + \langle Z(t) \,|\, z\rangle \\ &\leq \|Y(t_0)\|_* - \gamma_L(t - t_0) + \|Z(t)\|_*. \end{aligned} \tag{6.15}$$

Since $Z(t)/t \to 0$ (a.s.) by Lemma C.2, we conclude that $\langle Y(t) \,|\, z\rangle \to -\infty$ uniformly in z (a.s.). Consequently, there exists some t_0' such that $\langle Y(t) - y^* \,|\, z\rangle \leq 0$ for all $t \geq t_0'$ and all $z \in \mathrm{TC}(x^*)$ with $\|z\| = 1$. In turn, this implies that $Y(t) \in y^* + \mathrm{PC}(x^*)$ for all $t \geq t_0'$ and our proof is complete. ■

7. Convergence via rectification

The analysis of the previous sections suggests that the main limiting factor in attaining a solution of (P) via (SMD) is the lack of inherent averaging in $X(t)$. When (P) admits a robust solution, this problem is overcome by the nonvanishing drift of (SMD) which pushes $X(t)$ towards a corner of $\mathcal{X}$ (so random wiggles are automatically contained by the geometry of $\mathcal{X}$). However, as we saw in the previous sections, $X(t)$ cannot converge to a solution of a general (nonlinear) convex program except in a certain, "weakly averaged" sense.

In this section, we counter this by examining a "rectified" variant of (SMD) that is run with a decreasing sensitivity parameter and which takes into account all past realizations of X up to time t. Specifically, motivated by Theorem 2.3(i) in the noise-free regime, consider the following transformation of $X(t)$:

$$\widetilde{X}(t) = \frac{1}{t} \int_0^t X(s) \, ds, \tag{7.1a}$$

and/or

$$\widetilde{X}(t) = X(s_t) \quad \text{where } s_t \in \arg\min_{0 \leq s \leq t} f(X(s)), \tag{7.1b}$$

corresponding respectively to the empirical average and the "best instance" of X up to time t.

Already, the results of [39] and the analysis of Section 5 indicate that the variant (7.1a) is concentrated around interior solutions of $\mathcal{X}$ (in the long run and in probability) if (SMD) is run with sufficiently low η. However, in a black-box setting

[14] Note here that if y has $F(x^*, \eta y) \leq \varepsilon + \delta$ and $C \subseteq \mathrm{PC}(x^*)$, we also have $F(x^*, \eta y') \leq \varepsilon + \delta$ for all $y' \in y + C$ by Lemma A.3.

[15] In fact, it suffices to have $x^* \in \operatorname{im} Q$ – or, equivalently, that h be subdifferentiable at x^*.

where knowledge about (P) and the noise process $Z(t)$ is not readily available, the choice of η would essentially become a matter of trial and error. In turn, this suggests using a variable sensitivity parameter $\eta \equiv \eta(t)$ which decreases to 0 as $t \to \infty$. However, since $Y(t) = \mathcal{O}(t)$, $\eta(t)$ should not decrease to zero faster than $1/t$; otherwise, $X(t) = Q(\eta(t)Y(t))$ would converge to the prox-center $x_c \equiv \arg\min_{x\in\mathcal{X}} h(x)$ of $\mathcal{X}$ with probability 1.

With this in mind, we will assume throughout this section that

$$\eta(t) \text{ is Lipschitz continuous, nonincreasing, and } \lim_{t\to\infty} t\eta(t) = \infty. \tag{7.2}$$

From an algorithmic point of view, (7.2) should be contrasted with the "annealing schedule" of simulated annealing methods, with η interpreted as the method's "inverse temperature" [17, 28].[16] In such schemes, the system typically starts at a high initial temperature ($\eta \approx 0$) and subsequently freezes to low temperatures ($\eta \to \infty$) in order to approach the system's "ground state". In our case however, $\eta(t)$ is nonincreasing, indicating that (SMD) is *heated* over time, not cooled. The reason for this disparity is that the system's "energy levels" $Y(t)$ evolve themselves over time and do not remain fixed (as in simulated annealing), so increasing η might quench prematurely a realized trajectory of the system to a suboptimal state.

If we take either definition of $\widetilde{X}(t)$ as a candidate solution for (P), we obtain:

Theorem 7.1. *The process $\widetilde{X}(t)$ enjoys the performance guarantee*

$$f(\widetilde{X}(t)) \leq \min f + \frac{\Omega}{t\eta(t)} + \frac{\sigma_*^2}{2Kt}\int_0^t \eta(s)\,ds + \mathcal{O}(\sqrt{\log\log t/t}) \quad (a.s.), \tag{7.3}$$

where $\Omega = \max\{h(x') - h(x) : x, x' \in \mathcal{X}\}$.

Corollary 7.2. *If $\lim_{t\to\infty}\eta(t) = 0$, then $\widetilde{X}(t) \to \arg\min f$ (a.s.). In particular, if $\eta(t) \propto \min\{1, t^{-\beta}\}$ for some $\beta \in (0,1)$, we have*

$$f(\widetilde{X}(t)) - \min f = \begin{cases} \mathcal{O}\left(t^{-\beta}\right) & \text{if } 0 < \beta < \frac{1}{2}, \\ \mathcal{O}\left(\sqrt{\log\log t/t}\right) & \text{if } \beta = \frac{1}{2}, \\ \mathcal{O}\left(t^{\beta-1}\right) & \text{if } \frac{1}{2} < \beta < 1. \end{cases} \tag{7.4}$$

Proof of Theorem 7.1. Let $F_\eta(t) = \eta(t)^{-1}F(x^*, \eta(t)Y(t))$ denote the η-deflated Fenchel coupling between $Y(t)$ and $x^* \in \arg\min f$. Then, after a small rearrangement, Lemma C.1 in Appendix C yields

$$\int_0^t \langle v(X(s)) \,|\, x^* - X(s)\rangle\,ds \leq F_\eta(0) - F_\eta(t) \tag{7.5a}$$

$$- \int_0^t \frac{\dot\eta(s)}{\eta^2(s)}\left[h(x^*) - h(X(s))\right]ds \tag{7.5b}$$

$$+ \frac{1}{2K}\int_0^t \eta(s)\|\Sigma(X(s), s)\|_*\,ds \tag{7.5c}$$

$$+ \sum_{i=1}^n \int_0^t (X_i(s) - x_i^*)\,dZ_i(s) \tag{7.5d}$$

We now proceed to bound each term of (7.5):

[16] This comparison becomes manifest when $h(x) = \sum_{j=1}^n x_j \log x_j$ is the entropic regularizer over the unit simplex $\mathcal{X} = \Delta_n$. In this case, we have $Q(\eta y) = (e^{\eta y_1}, \dots, e^{\eta y_n})/\sum_{j=1}^n e^{\eta y_j}$, which is precisely the Gibbs probability distribution employed in simulated annealing methods.

a) Since $F_\eta(t) \geq 0$ for all t, the term (7.5a) is bounded from above by $F_\eta(0)$, viz.

$$(7.5a) \leq F_\eta(0) = \frac{h(x^*) + h^*(\eta(0)Y(0))}{\eta(0)} - \langle Y(0) \,|\, x^* \rangle \tag{7.6}$$

b) For (7.5b), we have $h(x^*) - h(X(s)) \leq \Omega$ by definition, so, with $\dot{\eta}(t) \leq 0$ for almost all t by (7.2), we get

$$(7.5b) \leq -\Omega \int_0^t \frac{\dot{\eta}(s)}{\eta^2(s)}\, ds = \frac{\Omega}{\eta(t)} - \frac{\Omega}{\eta(0)}. \tag{7.7}$$

c) For (7.5c), the definition of σ_*^2 readily gives

$$(7.5c) \leq \frac{\sigma_*^2}{2K} \int_0^t \eta(s)\, ds. \tag{7.8}$$

d) Finally, for (7.5d), let $\xi(t) = \int_0^t \sum_{i=1}^n (X_i(s) - x_i^*)\, dZ_i(s)$ and write $\rho = [\xi, \xi]$ for the quadratic variation of ξ. We then get

$$\begin{aligned} d[\xi,\xi] = d\xi \cdot d\xi &= \sum_{i,j=1}^n (X_i - x_i^*)(X_j - x_j^*)\, dZ_i \cdot dZ_j \\ &= \sum_{i,j=1}^n \Sigma_{ij}(X_i - x_i^*)(X_j - x_j^*)\, dt \leq \sigma_*^2 \|X(s) - x^*\|^2\, dt, \end{aligned} \tag{7.9}$$

implying that $\rho(t) \leq \sigma_*^2 \|\mathcal{X}\|^2 t$. Now, arguing as in the proof of Lemma C.2 in Appendix C, the Dambis–Dubins–Schwarz time-change theorem for martingales [26, Theorem 3.4.6 and Problem 3.4.7] shows that there exists a one-dimensional Wiener process $\widetilde{W}(t)$ with induced filtration $\widetilde{\mathcal{F}}_s = \mathcal{F}_{\tau_\rho(s)}$ and such that $\widetilde{W}(\rho(t)) = \xi(t)$ for all $t \geq 0$. By the law of the iterated logarithm [26], we then obtain

$$\limsup_{t\to\infty} \frac{\widetilde{W}(\rho(t))}{\sqrt{2Mt \log\log(Mt)}} \leq \limsup_{t\to\infty} \frac{\widetilde{W}(\rho(t))}{\sqrt{2\rho(t) \log\log \rho(t)}} = 1 \quad \text{(a.s.)}, \tag{7.10}$$

where $M = \sigma_*^2 \|\mathcal{X}\|^2$. Thus, with probability 1, we have $\xi(t) = \mathcal{O}(\sqrt{t \log\log t})$.

Putting together all of the above and dividing by t, we get

$$\frac{1}{t}\int_0^t \langle v(X(s)) \,|\, x^* - X(s) \rangle ds \leq \frac{\Omega}{t\eta(t)} + \frac{\sigma_*^2}{2Kt}\int_0^t \eta(s) ds + \mathcal{O}(t^{-1/2}\sqrt{\log\log t}), \tag{7.11}$$

where we have absorbed the $\mathcal{O}(1/t)$ terms from (7.6) and (7.7) in the logarithmic term $\mathcal{O}(\sqrt{t^{-1}\log\log t})$. However, since $x^* \in \arg\min f$, we also have $\langle v(x) \,|\, x^* - x \rangle = \langle \nabla f(x) \,|\, x - x^* \rangle \geq f(x) - f(x^*) = f(x) - \min f$ for all $x \in \mathcal{X}$. Hence, by the definition of $\widetilde{X}$ (and Jensen's inequality in the case of (7.1a)), we finally obtain

$$f(\widetilde{X}(t)) \leq \min f + \frac{1}{t}\int_0^t \langle v(X(s)) \,|\, x^* - X(s) \rangle\, ds, \tag{7.12}$$

and our assertion follows. ■

In addition to the almost sure bound (7.3), we also have the following mean value guarantee:

Proposition 7.3. *If $Y(0) = 0$, $\widetilde{X}(t)$ enjoys the mean performance guarantee*

$$\mathbb{E}[f(\widetilde{X}(t))] \leq \min f + \frac{\Omega}{t\eta(t)} + \frac{\sigma_*^2}{2Kt} \int_0^t \eta(s)\, ds. \tag{7.13}$$

In particular, if $\eta(t) = \sqrt{\Omega K/\sigma_*^2} \min\{1, 1/\sqrt{t}\}$, we have

$$\mathbb{E}[f(\widetilde{X}(t))] \leq \min f + 2\sqrt{\Omega\sigma_*^2/(Kt)}. \tag{7.14}$$

Proof. If $Y(0) = 0$, (7.6) gives (7.5a) $\leq \eta(0)^{-1}\,[h(x^*) - h(Q(0))] \leq \Omega/\eta(0)$. With this in mind, (7.11) becomes

$$\frac{1}{t}\int_0^t \langle v(X(s)) \,|\, x^* - X(s)\rangle\, ds \leq \frac{\Omega}{t\eta(t)} + \frac{\sigma_*^2}{2Kt}\int_0^t \eta(s)\, ds + \xi(t), \tag{7.15}$$

where $\xi(t)$ denotes the martingale term (7.5d), and we used the fact that the constant terms of (7.6) and (7.7) cancel out. The bound (7.13) then follows by taking expectations in (7.15), dividing by t, and arguing as in the proof of Theorem 7.1. Finally, (7.14) is obtained by substituting in (7.13) and integrating. ∎

Compared to the value convergence rate (2.10) of the noiseless regime, Proposition 7.3 indicates a drop from $\mathcal{O}(1/t)$ to $\mathcal{O}(1/\sqrt{t})$. This is due to the Itô correction term $\sigma_*^2/(2Kt)\int_0^t \eta(s)\, ds$ in (7.13): balancing this second-order error against the noise-free bound $\Omega/(t\eta(t))$ is what imposes a $\Theta(t^{-1/2})$ sensitivity schedule – otherwise, one term would be slower than the other for large t. In this regard, (7.14) is reminiscent of the $\mathcal{O}(n^{-1/2})$ bounds derived in [35, Section 2.3] and [38, Section 6] for the dual averaging method (1.3) in stochastic environments. There, the drop in performance is due to the gap between continuous and discrete time, whereas here it is due to the gap between ordinary and (Itô) stochastic calculus.

8. Discussion

In this last section, we discuss briefly the case where the feasible region $\mathcal{X}$ of (P) is unbounded and we examine some further links with the literature on Hessian Riemannian gradient flows [3, 4, 13].

8.1. Noncompact feasible regions.

If the feasible region $\mathcal{X}$ of (P) is unbounded, the first difficulty that arises is that the well-posedness of (SMD) becomes harder to verify. To be sure, under the Lipschitz continuity condition (H1), Dynkin's existence and uniqueness theorem [27, Theorem 3.4] shows that (SMD) admits unique global solutions provided that the following growth condition also holds:

$$\|v(x)\|_* + \|\sigma(x,t)\|_* \leq C(1 + \|x\|) \quad \text{for some } C > 0 \text{ and all } x \in \mathcal{X},\, t \geq 0. \tag{H1$'$}$$

However, in the context of convex programming, these growth requirements can be fairly restrictive: for instance, $f(x) = x^2(1 + x^2)$ is analytic and strongly convex over $\mathcal{X} = \mathbb{R}$, but both (H1) and (H1′) fail.

On the other hand, given that f is convex, it is intuitively clear that the drift $v(x) = -\nabla f(x)$ of (SMD) directs $X(t)$ *away* from unbounded regions where (H1) and (H1′) fail, so one would expect (SMD) to remain well-posed. Using more refined regularity notions from stochastic analysis [27, Section 3.4], it is possible to show that this is indeed the case; however, the details of this analysis are fairly cumbersome, so we will be assuming in what follows that (SMD) is well-posed.

A further difficulty is that $\inf f$ need not be finite if $\mathcal{X}$ is not compact – and even if $\inf f > -\infty$, $\arg\min f$ may still be empty. To avoid the complications that

this imposes on a trajectory-based analysis,[17] we only extend below the value-based results of Section 7 to the unbounded case:

Proposition 8.1. *Suppose that* (SMD) *is initialized at* $Y(0) = 0$. *Then, for all* $p \in \mathcal{X}$, *the process* $\widetilde{X}(t)$ *enjoys the mean performance guarantee*

$$\mathbb{E}[f(\widetilde{X}(t))] \leq f(p) + \frac{h(p) - h_{\min}}{t\eta(t)} + \frac{\sigma_*^2}{2Kt} \int_0^t \eta(s)\, ds. \tag{8.1}$$

Consequently, if $\eta(t) \to 0$, *we have* $\mathbb{E}[f(\widetilde{X}(t))] \to \inf f$; *in particular, if* $\inf f > -\infty$ *and* $\eta(t) \propto \min\{1, 1/\sqrt{t}\}$, *then*

$$\mathbb{E}[f(\widetilde{X}(t))] - \inf f = \mathcal{O}(1/\sqrt{t}). \tag{8.2}$$

Remark 8.1. Since h is strongly convex and $\mathcal{X}$ is closed, $h_{\min} > -\infty$ is attained at the prox-center $x_c = Q(0)$ of $\mathcal{X}$.

Proof. Fix some $p \in \mathcal{X}$. Then, arguing as in the proof of Theorem 7.1, we obtain

$$\begin{aligned}\int_0^t \langle v(X(s)) \,|\, p - X(s)\rangle\, ds \leq F_\eta(0) - F_\eta(t) \\ &- (h(p) - h_{\min}) \int_0^t \frac{\dot\eta(s)}{\eta^2(s)}\, ds \\ &+ \frac{\sigma_*^2}{2K} \int_0^t \eta(s)\, ds \\ &+ \sum_{i=1}^n \int_0^t (X_i(s) - p_i)\, dZ_i(s),\end{aligned} \tag{8.3}$$

where we used the fact that $h(X(s)) \geq h_{\min}$ for all $s \geq 0$ in the second line (note also that Lemma C.1 does not require $\mathcal{X}$ to be compact). Since $F(p, 0) = h(p) + h^*(0)$ by definition, the first two terms of (8.3) become

$$F_\eta(0) - F_\eta(t) - (h(p) - h_{\min}) \int_0^t \frac{\dot\eta(s)}{\eta^2(s)}\, ds \leq \frac{h(p) - h_{\min}}{\eta(t)}. \tag{8.4}$$

Our claim then follows as in the proof of Proposition 7.3. ■

As in the compact case, we can get a finer description of the behavior of $X(t)$ under additional assumptions for f. For instance, if f is strongly convex, it can be shown that $\sup_{t\geq 0}\|X(t)\| < \infty$ (a.s.), so the analysis of Section 5 can be carried over to the noncompact regime. Likewise, if (P) admits a robust solution or f is linear, it can be shown that $f(X(t)) \to \inf f$ with probability 1 if (SMD) is run with sufficiently low sensitivity parameter η. However, this analysis would take us too far afield so we postpone it to future work.

8.2. **Perturbed Hessian Riemannian flows.** When h is steep and $\mathcal{X}$ has non-empty (topological) interior, the differential theory of Legendre transformations [41, Chapter 26] shows that the mirror map $Q = \nabla h^*$ is a homeomorphism between $\mathcal{Y} = \mathcal{V}^*$ and $\mathcal{X}^\circ = \operatorname{dom} \partial h$. In this case, the system (MD) induces a semiflow on $\mathcal{X}^\circ$ via the dynamics

$$\dot x = \frac{d}{dt} Q(y) = \nabla Q(y) \cdot \dot y = \nabla(\nabla h^*(y)) \cdot v(Q(y)) = -\operatorname{Hess}(h^*(y)) \cdot \nabla f(x). \tag{8.5}$$

[17] For instance, the notion of a robust solution becomes void if $\mathcal{X}$ is an affine subspace of $\mathcal{V}$.

By Lagrange's identity, we also have $\operatorname{Hess}(h^*(y)) = \operatorname{Hess}(h(Q(y)))^{-1}$ for all $y \in \mathcal{Y}$, so (8.5) leads to the *Hessian Riemannian* (HR) dynamics

$$\dot{x} = -H(x)^{-1} \cdot \nabla f(x), \tag{HD}$$

where $H(x) \equiv \operatorname{Hess}(h(x))$ denotes the Hessian of h evaluated at $x = Q(y)$.

The adjective "Riemannian" above reflects the fact that $H(x)^{-1}\nabla f(x)$ can be viewed as the gradient of f with respect to the Riemannian metric $g = \operatorname{Hess}(h)$ on $\mathcal{X}^\circ$; in words, (HD) simply represents the gradient flow of f with respect to a modified geometry on $\mathcal{X}$. In the context of convex programming, this system has been studied extensively by [3, 13] who examined its well-posedness and convergence properties in a wide array of convex programming applications (see also [4] for the singular Riemannian case corresponding to nonsteep h). As such, a natural question that arises is whether this equivalence between (HD) and (MD) carries over to the stochastic regime analyzed in this paper.

To address this question, assume that the gradient input $\nabla f(x)$ to (HD) is perturbed by some random noise function $\epsilon(t)$ as in (3.1), viz.

$$\dot{x} = H(x)^{-1} \cdot (-\nabla f(x) + \epsilon(t)). \tag{8.6}$$

Then, writing out (8.6) as a stochastic differential equation, we get the stochastic Hessian Riemannian dynamics

$$dX = -H(X)^{-1} \cdot \nabla(f(X))\, dt + H(X)^{-1} \cdot dZ, \tag{SHD}$$

with $Z(t)$ defined as in (3.2). On the other hand, if h is sufficiently smooth, Itô's formula shows that the primal dynamics generated by (SMD) on $\mathcal{X}$ are given by

$$dX = \nabla(Q(Y)) \cdot v(X)\, dt + \nabla(Q(Y)) \cdot dZ + \frac{1}{2}\Sigma(X) \cdot \nabla^2(Q(Y))\, dt, \tag{SMD-P}$$

with the last term corresponding to the second-order Itô correction induced by the nonlinearity of Q (we have also taken $\eta = 1$ for simplicity).

Comparing these two systems, we see that the first two terms of (SMD-P) correspond precisely to the drift and diffusion coefficients of (SHD). However, the Itô correction term $\frac{1}{2}\Sigma(X) \cdot \nabla^2(Q(Y))\, dt$ (which involves the third derivatives of h^*) has no equivalent in (SHD), meaning that (SHD) and (SMD-P) *do not coincide in general* – that is, unless the mirror map $Q\colon \mathcal{Y} \to \mathcal{X}$ happens to be linear.

To illustrate the above, take the linear objective $f(x) = x$ over $\mathcal{X} = [0,1]$ and consider the dynamics generated by the entropic penalty function $h(x) = x\log x + (1-x)\log(1-x)$ with induced mirror map $Q(y) = e^y/(1+e^y)$. Then, (SHD) becomes

$$dX = -X(1-X)\left[dt - \sigma\, dW\right]. \tag{8.7}$$

while (SMD) gives

$$\begin{aligned} dY &= -dt + \sigma\, dW, \\ X &= e^Y/(1+e^Y). \end{aligned} \tag{8.8}$$

Hence, after a short calculation, we obtain:

$$dX = -X(1-X)\left[dt - \sigma\, dW\right] + \frac{1}{2}X(1-X)(1-2X)\sigma^2\, dt. \tag{8.9}$$

We thus see that the primal dynamics (8.7) and (8.9) differ by the term $\frac{1}{2}X(1-X)(1-2X)\, dt$ which is precisely the Itô correction induced on X via Q. Because of this structural difference, the dynamics (8.7) and (8.9) also behave very differently with respect to the minimizer $x^* = 0$ of f. Specifically, (8.8) gives $Y(t) = Y(0) -$

$t+\sigma W(t) \to -\infty$ (a.s.), implying in turn that $X(t) \to x^*$ (a.s.). On the other hand, under (8.7), $X(t)$ may converge to $\arg\max f$ with arbitrarily high probability if σ is large enough. To see this, let $V(x) = \log x - \log(1-x)$, so $V(X(t)) \to -\infty$ if $X(t) \to 0^+$ and $V(X(t)) \to +\infty$ if $X(t) \to 1^-$. Itô's lemma then yields

$$dV = V'(X)\,dX + \frac{1}{2}(dX)^2 = -\,dt + \sigma\,dW + (X-1/2)\sigma^2\,dt. \tag{8.10}$$

From (8.10), it is intuitively obvious (and can be shown rigorously) that the drift of (8.10) remains uniformly positive with probability arbitrarily close to 1 if $X(0) > 1/2$ and σ is large.[18] In turn, this implies that $V(X(t)) \to \infty$, i.e. (8.7) converges with high probability to $\arg\max f$ instead of $\arg\min f$!

The above shows that the Hessian Riemannian system (HD) is far more vulnerable to noise compared to (MD). Intuitively, this failure is due to the fact that (HD) lacks an inherent "averaging" mechanism capable of dissipating the noise in the long run – in (MD), this role is played by the direct aggregation of gradient steps up to time t. At a more formal level, this primal/dual disparity is due to the second-order Itô correction term which appears in (SMD-P) and which acts as a "silver bullet" for the noise. Given the link between Hessian Riemannian dynamics and the replicator dynamics of evolutionary game theory [2, 3], this is also reminiscent of the different long-run behavior of the replicator dynamics with aggregate shocks [20, 23] and the dynamics of stochastically perturbed exponential learning [32]. We intend to explore these relations in depth in a future paper.

Appendix A. Mirror maps and the Fenchel coupling

In this appendix, we collect some basic properties of mirror maps and the Fenchel coupling. Our first result describes the structure of the inverse images of Q:

Lemma A.1. *If $Q(y) = x$, then $Q(y+v) = x$ for all $v \in \mathrm{PC}(x)$.*

Proof. By Proposition 2.2, it suffices to show that $y + v \in \partial h(x)$ for all $v \in \mathrm{PC}(x)$. However, since $v \in \mathrm{PC}(x)$, we also have $\langle v \,|\, x' - x\rangle \leq 0$ for all $x' \in \mathcal{X}$, and hence

$$h(x') \geq h(x) + \eta\langle y \,|\, x' - x\rangle \geq h(x) + \eta\langle y + v \,|\, x' - x\rangle, \tag{A.1}$$

where the first inequality follows from the fact that $y \in \partial h(x)$. The above shows that $y + v \in \partial h(x)$, so $Q(y+v) = x$, as claimed. ∎

Our next result provides a comparison between the Fenchel coupling with the Bregman divergence and the underlying norm on $\mathcal{V}$:

Proposition A.2. *Let h be a K-strongly convex penalty function on $\mathcal{X}$. Then, for all $p \in \mathcal{X}$ and all $y, y' \in \mathcal{Y}$, we have:*

a) $F(p,y) \geq D(p, Q(y))$ *with equality if* $Q(y) \in \mathcal{X}^\circ$. (A.2a)

b) $F(p,y) \geq \frac{1}{2}K\,\|Q(y) - p\|^2$. (A.2b)

c) $F(p,y') \leq F(p,y) + \langle y' - y \,|\, Q(y) - p\rangle + \frac{1}{2K}\|y' - y\|_*^2$. (A.2c)

Proof. See [30, Proposition 4.3]. ∎

The following corollary of Proposition A.2 is also useful in our analysis:

[18] For a formal argument along these lines, see [34, Theorem 3.3.3].

Lemma A.3. *If $y_2 - y_1 \in \mathrm{PC}(p)$, we have $F(p, y_1) \geq F(p, y_2)$ and*

$$\|y_2 - y_1\|_* \geq K\|\mathcal{X}\| \left[\sqrt{1 + 2\delta/(K\|\mathcal{X}\|^2)} - 1\right], \tag{A.3}$$

where $\delta = F(p, y_1) - F(p, y_2)$.

Proof. Let $v = y_2 - y_1$ and set $g(t) = F(p, y_1 + tv)$, $t \in [0, 1]$. Differentiating yields $g'(t) = \langle v \,|\, Q(y_1+tv)-p\rangle \leq 0$ for all t because $v \in \mathrm{PC}(p)$ and $Q(y_1+tv)-p \in \mathrm{TC}(p)$. We thus get $F(p, y_2) = F(p, y_1 + v) \leq F(p, y_1)$, as claimed.

For our second assertion, (A.2c) readily yields

$$\begin{aligned} F(p, y_2) - F(p, y_1) &\leq \langle y_2 - y_1 \,|\, Q(y) - p\rangle + \frac{1}{2K}\|y_2 - y_1\|_*^2 \\ &\leq \|\mathcal{X}\| \, \|y_2 - y_1\|_* + \frac{1}{2K}\|y_2 - y_1\|_*^2, \end{aligned} \tag{A.4}$$

and, after rearranging:

$$\omega^2 + 2K\|\mathcal{X}\|\omega - 2K\delta \geq 0, \tag{A.5}$$

where we set $\omega = \|y' - y\|_* \geq 0$. The roots of this inequality are $\omega_\pm = -K\|\mathcal{X}\| \pm \sqrt{K^2\|\mathcal{X}\|^2 + 2K\delta}$, so $\omega_- < 0 \leq \omega_+$. This implies that (A.5) can only hold if $\omega \geq \omega_+$ and (A.3) follows. ∎

Appendix B. Deterministic results

As we discussed in Section 2, Proposition A.2 allows us to use the Fenchel coupling as a convergence test (in the sense that $x(t) \to x^*$ whenever $F(x^*, y(t)) \to 0$). The suitability of the Fenchel coupling in this regard is owed to its Lyapunov-like behavior under (MD):

Lemma B.1. *Fix some base point $p \in \mathcal{X}$. Then, under (MD), we have*

$$\frac{d}{dt}F_\eta(p, y(t)) = \langle v(x(t)) \,|\, x(t) - p\rangle. \tag{B.1}$$

In particular, $F_\eta(x^*, y(t))$ is nonincreasing for all $x^* \in \arg\min f$.

Proof. By the definition (2.12) of the η-deflated Fenchel coupling, we have

$$\frac{dF_\eta}{dt} = \eta^{-1}\left[\langle \eta\dot{y} \,|\, \nabla h^*(\eta y)\rangle\right] - \langle \dot{y} \,|\, p\rangle = \langle v(x) \,|\, x - p\rangle, \tag{B.2}$$

where the first equality is a consequence of Proposition 2.2. Our second claim then follows by noting that $\langle v(x) \,|\, x - x^*\rangle \leq f(x^*) - f(x) \leq 0$ for all $x^* \in \arg\min f$. ∎

Proof of Theorem 2.3. For all $x^* \in \arg\min f$, Lemma B.1 gives

$$\begin{aligned} F_\eta(x^*, y(t)) - F_\eta(x^*, y_0) &= \int_0^t \langle v(x(s)) \,|\, x(s) - x^*\rangle \, ds \\ &\leq \int_0^t [f(x^*) - f(x(s))] \, ds = t[\min f - \bar{f}(t)]. \end{aligned} \tag{B.3}$$

A simple rearrangement then yields $\bar{f}(t) - \min f \leq F_\eta(x^*, y_0)/t$, so the bound for $f_{\min}(t)$ follows trivially. As for the specific rate Ω/t, it suffices to note that $F(x^*, 0) = h(x^*) + h^*(0) = h(x^*) - h(Q(0)) \leq \max\{h(x') - h(x) : x, x' \in \mathcal{X}\}$.

For our second assertion, let x^ω be an ω-limit of $x(t)$ and assume that $x^\omega \notin \arg\min f$. Since $\arg\min f$ is closed, there exists a neighborhood U of x^ω in $\mathcal{X}$ such that $\langle v(x) \,|\, x - x^*\rangle \leq -a$ for some $a > 0$ and for all $x^* \in \arg\min f$. Furthermore,

since x^ω is an ω-limit of $x(t)$, there exists an increasing sequence of times $t_k \uparrow \infty$ such that $x(t_k) \in U$ for all k. Then, for all $\tau > 0$, Proposition 2.2 gives

$$\begin{aligned}\|x(t_k+\tau) - x(t_k)\| &= \|Q(\eta y(t_k+\tau)) - Q(\eta y(t_k))\| \leq \frac{\eta}{K}\|y(t_k+\tau) - y(t_k)\|_* \\ &\leq \frac{\eta}{K}\int_{t_k}^{t_k+\tau} \|v(x(s))\|_*\,ds \leq \frac{\eta\tau}{K}\max_{x\in\mathcal{X}}\|v(x)\|_*. \end{aligned} \tag{B.4}$$

Now, given that the bound (B.4) does not depend on k, there exists some sufficiently small $\delta > 0$ such that $x(t_k+\tau) \in U$ for all $\tau \in [0,\delta]$, $k \in \mathbb{N}$ (so we also have $\langle v(x(t_k+\tau)) \,|\, x(t_k+\tau) - x^*\rangle \leq -a$). Therefore, given that $\langle v(x) \,|\, x - x^*\rangle \leq 0$ for all $x \in \mathcal{X}$, $x^* \in \arg\min f$, we get

$$\begin{aligned}F_\eta(x^*, y(t_k+\delta)) - F_\eta(x^*, y_0) &= \int_0^{t_k+\delta} \langle v(x(s)) \,|\, x(s) - x^*\rangle\,ds \\ &\leq \sum_{j=1}^k \int_{t_j}^{t_j+\delta} \langle v(x(s)) \,|\, x(s) - x^*\rangle\,ds \leq -ak\delta, \end{aligned} \tag{B.5}$$

showing that $\liminf_{t\to\infty} F(x^*, \eta y(t)) = -\infty$, a contradiction (recall that $F(x^*, \eta y) \geq 0$ for all $y \in \mathcal{Y}$).

Since $x(t)$ admits at least one ω-limit (by compactness of $\mathcal{X}$), we conclude that $x(t)$ converges to $\arg\min f$. Thus, if $x^* \in \arg\min f$ is an ω-limit of $x(t)$, Lemma B.1 shows that $F_\eta(x^*, y(t))$ is nonincreasing. Finally, given that $x(t'_k) \to x^*$ for some increasing sequence of times $t'_k \uparrow \infty$, (H2) further shows that $F_\eta(x^*, y(t'_k)) \to 0$; in turn, this implies that $F(x^*, \eta y(t)) \to 0$, i.e. $\lim_{t\to\infty} x(t) = x^*$. ■

Appendix C. Stochastic calculations

Our main goal in this appendix is to show that the η-deflated coupling $F_\eta(t) = \eta(t)^{-1}F(p, \eta(t)Y(t))$ between p and $Y(t)$ satisfies a noisy version of Lemma B.1:

Lemma C.1. *Let $p \in \mathcal{X}$. Then, for all $t \geq t_0 \geq 0$, we have*

$$F_\eta(t) - F_\eta(t_0) \leq \int_{t_0}^t \langle v(X(s)) \,|\, X(s) - p\rangle\,ds \tag{C.1a}$$

$$- \int_{t_0}^t \frac{\dot\eta(s)}{\eta(s)^2}[h(p) - h(X(s))]\,ds \tag{C.1b}$$

$$+ \frac{1}{2K}\int_{t_0}^t \eta(s)\|\Sigma(X(s), s)\|_*\,ds \tag{C.1c}$$

$$+ \sum_{i=1}^n \int_{t_0}^t (X_i(s) - p_i)\,dZ_i(s). \tag{C.1d}$$

In particular, if the sensitivity parameter η of (SMD) is constant, we have

$$F_\eta(t) - F_\eta(t_0) \leq \int_{t_0}^t \langle v(X(s)) \,|\, X(s) - p\rangle\,ds + \frac{\eta}{2K}\int_{t_0}^t \|\Sigma(X(s), s)\|_*\,ds + \xi(t), \tag{C.2}$$

where $\xi(t) = \sum_{i=1}^n \int_{t_0}^t (X_i(s) - p_i)\,dZ_i(s)$ denotes the martingale term (C.1d).

Proof. By Proposition 2.2, we have $\nabla_y F(p,y) = \nabla h^*(y) - p = Q(y) - p$ for all $y \in \mathcal{Y}$. Thus, given that $Q = \nabla h^*$ is Lipschitz continuous (again by Proposition 2.2), Itô's formula for functions with Lipschitz continuous first derivatives [19] yields

$$\begin{aligned} dF_\eta &= -\frac{\dot\eta}{\eta} F_\eta \, dt + \frac{1}{\eta} \sum_{i=1}^n \frac{\partial F}{\partial y_i}\bigg|_{\eta Y} d(\eta Y_i) + \frac{1}{2\eta} \sum_{i,j=1}^n \frac{\partial^2 F}{\partial y_i \partial y_j}\bigg|_{\eta Y} d(\eta Y_i) \cdot d(\eta Y_j) \\ &= \langle v(X) \,|\, X - p \rangle \, dt + \langle dZ \,|\, X - p \rangle + \frac{\eta}{2} \sum_{i,j=1}^n \frac{\partial^2 h^*}{\partial y_i \partial y_j}\bigg|_{\eta Y} dZ_i \cdot dZ_j, \qquad \text{(C.3a)} \\ &\quad - \frac{\dot\eta}{\eta^2} [h(p) + h^*(\eta Y) - \langle \eta Y \,|\, p \rangle] \, dt + \frac{\dot\eta}{\eta} \langle Y \,|\, X - p \rangle \, dt \qquad \text{(C.3b)} \end{aligned}$$

where we have used the definition of (SMD) to write $d(\eta Y) = \dot\eta Y \, dt + \eta v(X) dt + \eta dZ$ in the second and third lines.

The above expression immediately yields the terms (C.1a) and (C.1d) of (C.1). The Itô correction term (C.1c) follows from the definition $dZ_i \cdot dZ_j = \Sigma_{ij} \, dt$ of the infinitesimal covariance matrix Σ and the fact that $\operatorname{tr}[\operatorname{Hess}(h^*(Y(s)))\Sigma(X(s),s)] \leq K^{-1} \|\Sigma(X(s),s)\|_*$ by the strong convexity of h. Finally, for (C.1b), recall that $h^*(\eta Y) = \langle \eta Y \,|\, X \rangle - h(X)$ by the definition of h^*; we then get $\langle Y \,|\, X \rangle = \eta^{-1}[h^*(\eta Y) + h(X)]$ and (C.1) follows by substituting in (C.3b) and rearranging. ∎

Our final result is a growth estimate for Itô martingales with bounded volatility:

Lemma C.2. *Let $W(t) = (W_1(t), \dots, W_m(t))$, $t \geq 0$, be a Wiener process in $\mathbb{R}^m$ and let $\zeta(t)$ be a bounded, continuous process in $\mathbb{R}^m$. Then, for every continuous function $f \colon [0,\infty) \to (0,\infty)$, we have*

$$f(t) + \int_0^t \zeta(s) \cdot dW(s) \sim f(t) \quad \text{as } t \to \infty \text{ (a.s.)}, \tag{C.4}$$

whenever $\lim_{t\to\infty} (t \log\log t)^{-1/2} f(t) = +\infty$.

Proof of Lemma C.2. Let $\xi(t) = \int_0^t \zeta(s) \cdot dW(s) = \sum_{i=1}^n \int_0^t \zeta_i(s) \, dW_i(s)$. Then, the quadratic variation $\rho = [\xi, \xi]$ of ξ satisfies:

$$d[\xi,\xi] = d\xi \cdot d\xi = \sum\nolimits_{i=1}^n \zeta_i \zeta_j \delta_{ij} \, dt \leq M \, dt, \tag{C.5}$$

where $M = \sup_{t\geq 0} \|\zeta(t)\|^2 < \infty$ (recall that $\zeta(t)$ is bounded by assumption).

Now, let $\rho_\infty = \lim_{t\to\infty} \rho(t) \in [0,\infty]$ and set

$$\tau_\rho(s) = \begin{cases} \inf\{t \geq 0 : \rho(t) > s\} & \text{if } s \leq \rho_\infty, \\ \infty & \text{otherwise}, \end{cases} \tag{C.6}$$

The process $\tau_\rho(s)$ is finite, non-negative, non-decreasing, and right-continuous on $[0,\rho_\infty)$; moreover, it is easy to check that $\rho(\tau_\rho(s)) = s \wedge \rho_\infty$ and $\tau_\rho(\rho(t)) = t$ [26, Problem 3.4.5]. Therefore, by the Dambis–Dubins–Schwarz time-change theorem for martingales [26, Thm. 3.4.6 and Pb. 3.4.7], there exists a standard, one-dimensional Wiener process $\widetilde{W}(t)$ with induced filtration $\widetilde{\mathcal{F}}_s = \mathcal{F}_{\tau_\rho(s)}$ and such that $\widetilde{W}(\rho(t)) = \xi(t)$ for all $t \geq 0$. We thus get:

$$\frac{f(t) + \xi(t)}{f(t)} = 1 + \frac{\widetilde{W}(\rho(t))}{f(t)}. \tag{C.7}$$

Obviously, if $\lim_{t\to\infty}\rho(t) < \infty$, we have $\limsup_{t\to\infty}|\widetilde{W}(\rho(t))| < \infty$ (a.s.) so there is nothing to show. Otherwise, if $\lim_{t\to\infty}\rho(t) = \infty$, the quadratic variation bound (C.5) and the law of the iterated logarithm yield:

$$\frac{\left|\widetilde{W}(\rho(t))\right|}{f(t)} \leq \frac{\left|\widetilde{W}(\rho(t))\right|}{\sqrt{2\rho(t)\log\log\rho(t)}} \times \frac{\sqrt{2Mt\log\log Mt}}{f(t)} \to 0 \quad \text{as } t\to\infty, \tag{C.8}$$

and our claim follows. ■

References

[1] B. Abbas and H. Attouch, *Dynamical systems and forward-backward algorithms associated with the sum of a convex subdifferential and a monotone cocoercive operator*, Optimization, 64 (2015), pp. 2223–2252.

[2] E. Akin, *The geometry of population genetics*, no. 31 in Lecture Notes in Biomathematics, Springer-Verlag, 1979.

[3] F. Alvarez, J. Bolte, and O. Brahic, *Hessian Riemannian gradient flows in convex programming*, SIAM Journal on Control and Optimization, 43 (2004), pp. 477–501.

[4] H. Attouch, J. Bolte, P. Redont, and M. Teboulle, *Singular Riemannian barrier methods and gradient-projection dynamical systems for constrained optimization*, Optimization, 53 (2004), pp. 435–454.

[5] D. A. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming I. Affine and projective scaling trajectories*, Transactions of the American Mathematical Society, 314 (1989), pp. 499–526.

[6] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters, 31 (2003), pp. 167–175.

[7] A. Ben-Tal, T. Margalit, and A. S. Nemirovski, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM Journal on Optimization, 12 (2001), pp. 79–108.

[8] A. Ben-Tal and A. S. Nemirovski, *Lectures on Modern Convex Optimization*, MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[9] M. Benaïm, *Dynamics of stochastic approximation algorithms*, in Séminaire de Probabilités XXXIII, J. Azéma, M. Émery, M. Ledoux, and M. Yor, eds., vol. 1709 of Lecture Notes in Mathematics, Springer Berlin Heidelberg, 1999, pp. 1–68.

[10] M. Benaïm and M. W. Hirsch, *Asymptotic pseudotrajectories and chain recurrent flows, with applications*, Journal of Dynamics and Differential Equations, 8 (1996), pp. 141–176.

[11] R. N. Bhattacharya, *Criteria for recurrence and existence of invariant measures for multidimensional diffusions*, Annals of Probability, (1978), pp. 541–553.

[12] R. I. Boţ and E. R. Csetnek, *Approaching the solving of constrained variational inequalities via penalty term-based dynamical systems*, Journal of Mathematical Analysis and Applications, 435 (2016), pp. 1688–1700.

[13] J. Bolte and M. Teboulle, *Barrier operators and associated gradient-like dynamical systems for constrained minimization problems*, SIAM Journal on Control and Optimization, 42 (2003), pp. 1266–1292.

[14] A. Bovier and F. den Hollander, *Metastability: A Potential-Theoretic Approach*, no. 351 in Grundlehren der mathematischen Wissenschaften, Springer, 2015.

[15] M. Bravo and P. Mertikopoulos, *On the robustness of learning in games with stochastically perturbed payoff observations*, Games and Economic Behavior, to appear (2016).

[16] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Computational Mathematics and Mathematical Physics, 7 (1967), pp. 200–217.

[17] V. Černý, *Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm*, Journal of Optimization Theory and Applications, 45 (1985), pp. 41–51.

[18] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan, *Ergodic mirror descent*, SIAM Journal on Optimization, 22 (2012), pp. 1549–1578.

[19] M. Errami, F. Russo, and P. Vallois, *Itô's formula for $C^{1,\lambda}$-functions of a càdlàg process and related calculus*, Probability Theory and Related Fields, 122 (2002), pp. 191–221.

[20] D. Fudenberg and C. Harris, *Evolutionary dynamics with aggregate shocks*, Journal of Economic Theory, 57 (1992), pp. 420–441.

[21] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, 1996.

[22] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer, Berlin, 2001.

[23] L. A. Imhof, *The long-run behavior of the stochastic replicator dynamics*, The Annals of Applied Probability, 15 (2005), pp. 1019–1045.

[24] A. N. Iusem, B. F. Svaiter, and J. X. da Cruz Neto, *Central paths, generalized proximal point methods, and Cauchy trajectories in Riemannian manifolds*, SIAM Journal on Control and Optimization, 37 (1999), pp. 566–588.

[25] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, *Regularization techniques for learning with matrices*, The Journal of Machine Learning Research, 13 (2012), pp. 1865–1890.

[26] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, Berlin, 1998.

[27] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, no. 66 in Stochastic Modelling and Applied Probability, Springer-Verlag, Berlin, 2 ed., 2012.

[28] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.

[29] H.-H. Kuo, *Introduction to Stochastic Integration*, Springer, Berlin, 2006.

[30] P. Mertikopoulos, *Learning in concave games with imperfect information*. https://arxiv.org/abs/1608.07310, 2016.

[31] P. Mertikopoulos, E. V. Belmega, R. Negrel, and L. Sanguinetti, *Distributed stochastic optimization via matrix exponential learning*. http://arxiv.org/abs/1606.01190, 2016.

[32] P. Mertikopoulos and A. L. Moustakas, *The emergence of rational behavior in the presence of stochastic perturbations*, The Annals of Applied Probability, 20 (2010), pp. 1359–1388.

[33] P. Mertikopoulos and W. H. Sandholm, *Learning in games via reinforcement and regularization*, Mathematics of Operations Research, 41 (2016), pp. 1297–1324.

[34] P. Mertikopoulos and Y. Viossat, *Imitation dynamics with payoff shocks*, International Journal of Game Theory, 45 (2016), pp. 291–320.

[35] A. S. Nemirovski, A. Juditsky, G. G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization, 19 (2009), pp. 1574–1609.

[36] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, NY, 1983.

[37] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, no. 87 in Applied Optimization, Kluwer Academic Publishers, 2004.

[38] ———, *Primal-dual subgradient methods for convex problems*, Mathematical Programming, 120 (2009), pp. 221–259.

[39] M. Raginsky and J. Bouvrie, *Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence*, in CDC '13: Proceedings of the 51st IEEE Annual Conference on Decision and Control, 2013.

[40] C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, CRC Press, Boca Raton, FL, 1995.

[41] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[42] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, vol. 317 of A Series of Comprehensive Studies in Mathematics, Springer-Verlag, Berlin, 1998.

[43] S. Shalev-Shwartz, *Online learning and online convex optimization*, Foundations and Trends in Machine Learning, 4 (2011), pp. 107–194.

[44] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for Machine Learning*, MIT Press, Cambridge, MA, USA, 2012.

[45] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, in NIPS '14: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, pp. 2510–2518.