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

PANAYOTIS MERTIKOPOULOS* AND MATHIAS STAUDIGL$^\sharp$

ABSTRACT. In view of solving convex optimization problems with noisy gradient input, we analyze the asymptotic behavior of gradient-like flows under stochastic disturbances. Specifically, we focus on the widely studied class of mirror descent schemes for convex programs with compact feasible regions, and we examine the dynamics' convergence and concentration properties in the presence of noise. In the vanishing noise limit, we show that the dynamics converge to the solution set of the underlying problem (a.s.). Otherwise, when the noise is persistent, we show that the dynamics are concentrated around interior solutions in the long run, and they converge to boundary solutions that are sufficiently "sharp". Finally, we show that a suitably rectified variant of the method converges irrespective of the magnitude of the noise (or the structure of the underlying convex program), and we derive an explicit estimate for its rate of convergence.

1. INTRODUCTION

Consider an unconstrained convex program of the form

$$\text{minimize} \quad f(x), \tag{P_0}$$

where $f: \mathcal{V} \to \mathbb{R}$ is a convex function defined on some finite-dimensional real space $\mathcal{V}$. To solve (P_0), a key role is played by the *gradient flow* of f, i.e. the gradient descent dynamics

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

As is well known, under mild regularity assumptions for f, the solution trajectories of (GD) converge to the solution set of (P_0) – 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.

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

destroyed, even in simple, one-dimensional problems. To see this, take $f(x) = \theta(x-\mu)^2/2$ 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, 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), $X(t)$ *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 $\sigma^2/(2\theta)$ [26, Chap. 5.6]. Thus, in the long run, $X(t)$ will fluctuate around μ with a spread that is roughly proportional to the noise volatility coefficient σ.

More generally, by solving the associated Fokker–Planck equation, it is well known that the perturbed gradient descent system

$$dX = -\nabla f(X)\, dt + \sigma\, dW \tag{1.3}$$

admits a unique invariant measure $e^{-2f(x)/\sigma^2}\, dx$, which gives rise to a (unique) invariant distribution $d\mu_\infty \propto e^{-2f(x)/\sigma^2}\, dx$ (assuming that $\int e^{-2f(x)/\sigma^2}\, dx < \infty$ for normalization purposes). In other words, for large t, $X(t)$ is most likely to be found near $\arg\min f$ and this likelihood is (exponentially) inversely proportional to σ. Moreover by ergodicity, the distribution of the time-averaged process $\bar{X}(t) = t^{-1}\int_0^t X(s)\, ds$ also converges to μ_∞; thus, in general, even the ergodic average of $X(t)$ fails to converge to $\arg\min f$ with positive probability.

Somewhat surprisingly, except for these basic results for unconstrained problems, the long-run behavior of constrained gradient-like flows with noisy input remains largely unexplored. With this in mind, we consider here the widely studied class of *mirror descent* (MD) dynamics that were pioneered by Nemirovski and Yudin [41] for constrained convex programs (and which include gradient descent as a special case), and we examine their convergence properties in the presence of stochastic disturbances.

Concretely, our paper focuses on constrained convex programs 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 compact convex subset of $\mathcal{V}$ and $f \colon \mathcal{X} \to \mathbb{R}$ is a C^1-smooth convex function on $\mathcal{X}$. In continuous time, the dynamics of mirror descent 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 "mirror map" $Q(y) = \arg\max_{x \in \mathcal{X}}\{\langle y | x\rangle - h(x)\}$ is a projection-like mapping defined via a strongly convex "prox-function" $h \colon \mathcal{X} \to \mathbb{R}$. In this way, (MD) is the continuous-time limit of Nesterov's well-known *dual averaging* scheme [43]

$$\begin{aligned} y_{t+1} &= y_t - \gamma_t \nabla f(x_t), \\ x_{t+1} &= Q(\eta y_{t+1}), \end{aligned} \tag{1.4}$$

where $\gamma_t > 0$, $t = 1, 2, \ldots$, is a variable step-size sequence.

The dynamics of mirror descent have recently attracted considerable interest in optimization [3, 4, 14, 32, 54] and machine learning [31, 33], and we summarize some of the convergence results obtained for (MD) in Section 2. As an example, if $h(x) = \frac{1}{2}\|x\|_2^2$, we have $Q(y) = \arg\min_{x \in \mathcal{X}} \|y - x\|$, so (MD) boils down to a (Euclidean) projected gradient descent scheme. Extending this interpretation to general h, the authors of [3, 4, 14] showed that (MD) may be viewed as the gradient flow of f with respect to a certain Riemannian metric on $\mathcal{X}$; thus, in addition to projected (Euclidean) gradient descent, (MD) also covers a very broad class of Riemannian gradient-like flows (cf. Section 5).

Moving beyond this deterministic framework, the study of mirror descent with noisy first-order feedback is a classic topic in optimization (see e.g. [18, 34, 40, 42] and references therein). In view of this, our paper focuses on the *stochastic mirror descent* dynamics

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

where $Z(t)$ is an Itô martingale process (such as Brownian motion) representing the sum of all random disturbances affecting the gradient input to (MD). In this stochastic setting, the simple example (1.1) shows that the deterministic convergence properties of (MD) cannot be carried over to (SMD) in full generality. Accordingly, our paper focuses on the following questions:

(1) If the volatility of $Z(t)$ decays over time, intuition suggests that the good convergence properties of (MD) should also apply to (SMD). In Section 4.1, we make this intuition precise by noting that the solutions of (SMD) correspond to *asymptotic pseudotrajectories* (APTs) [9] of (MD) in the vanishing noise limit.[1] Except for the very recent paper [12], 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, trajectory convergence to interior points is no longer possible. Nonetheless, if f is *strongly* convex and (P) admits an interior solution x^* (a case of particular interest in machine learning and statistics [51]), the long-run behavior of (SMD) can be described by examining the dynamics' invariant distribution. Our analysis in Section 4.2 provides an explicit estimate for this invariant measure and shows that (SMD) spends an arbitrarily large fraction of the time arbitrarily close to x^* if the dynamics' sensitivity parameter η is small enough.

(3) Departing from the interior case, we also consider sharp solutions that arise e.g. in generic linear programs. In this case, if the sensitivity parameter 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 4.3).

(4) Finally, if no assumptions can be made on the structure of (P), we show in Section 4.4 that a suitably rectified variant of (SMD) with a decreasing sensitivity parameter converges with probability 1. Specifically, if $\eta \equiv \eta(t)$

---

[1]For background information on the theory of stochastic approximation and APTs, see [8, 9].

decays as $\Theta(t^{-1/2})$, the ergodic average $\bar{X}(t) = t^{-1}\int_0^t X(s)\,ds$ of $X(t)$ enjoys an almost sure $\mathcal{O}(t^{-1/2}\sqrt{\log\log t})$ value convergence rate.[2]

At a technical level, this paper belongs to the growing literature on dynamical systems that arise in the solution of continuous optimization problems and variational inequalities – see e.g. [1, 13, 20, 24, 32, 41, 52, 52, 54] and references therein. More precisely, the deterministic bedrock of our analysis coincides with the gradient-like dynamics studied in [3, 4, 14]; along with an important dichotomy that arises in the stochastic regime, we make this link precise in Section 5. Otherwise, from a stochastic viewpoint, the work that is closest to our analysis is the recent paper [46] where the authors showed that the ergodic average of an interior-valued subclass of (SMD) converges within $\mathcal{O}(\sigma^2)$ of the solution set of (P) and further provided a variance reduction scheme based on the parallel sampling of multiple trajectories. To the best of our knowledge, this is the only result known for (SMD); our analysis in Section 4.4 shows that this optimality gap can be reduced to 0 if (SMD) is run with a decreasing sensitivity parameter.

There is also a broad and vigorous literature on second-order gradient systems such as Nesterov's accelerated gradient method (cf. [52] and references therein) and Polyak's "heavy ball with friction" dynamics [2, 5, 7, 16, 45]. Up to a dissipative friction term, such systems can be seen as quasi-gradient flows on $\mathcal{X} \times \mathcal{V}$ (the system's phase space) and recent works have considered the limit behavior of Itô perturbations of such flows [19]. Even though they might share some asymptotic properties, these second-order systems are fundamentally different from the first-order systems that we consider here (even in the noiseless, deterministic regime), so there is no overlap of results.

2. Preliminaries

Notation. Given an n-dimensional real space $\mathcal{V}$ with norm $\|\cdot\|$, we will write $\mathcal{V}^*$ for its dual, $\langle y\,|\,x\rangle$ for the pairing between $y \in \mathcal{V}^*$ and $x \in \mathcal{V}$, and $\|y\|_* \equiv \sup\{\langle y\,|\,x\rangle : \|x\| \leq 1\}$ for the dual norm of y in $\mathcal{V}^*$. Also, given an extended-real-valued function $g \colon \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$, its *effective domain* is defined as $\operatorname{dom} g = \{x \in \mathcal{V} : g(x) < \infty\}$ and its *subdifferential* at $x \in \operatorname{dom} g$ is given by $\partial g(x) = \{y \in \mathcal{V}^* : g(x') \geq g(x) + \langle y\,|\,x' - x\rangle$ for all $x' \in \mathcal{V}\}$.

In the rest of the paper, $\mathcal{X}$ will denote a compact convex subset of $\mathcal{V}$ and $f \colon \mathcal{X} \to \mathbb{R}$ will be a C^1-smooth convex function on $\mathcal{X}$; we will also write $\mathcal{X}^\circ \equiv \operatorname{ri}(\mathcal{X})$ for the relative interior of $\mathcal{X}$ and $\|\mathcal{X}\| = \max\{\|x' - x\| : x, x' \in \mathcal{X}\}$ for its diameter. For $x \in \mathcal{X}$, the *tangent cone* $\operatorname{TC}_{\mathcal{X}}(x)$ is the closure of the set of all rays emanating from x and intersecting $\mathcal{X}$ in at least one other point. The *polar cone* $\operatorname{PC}_{\mathcal{X}}(x)$ to $\mathcal{X}$ at x is then defined as $\operatorname{PC}_{\mathcal{X}}(x) = \{y \in \mathcal{V}^* : \langle y\,|\,z\rangle \leq 0$ for all $z \in \operatorname{TC}_{\mathcal{X}}(x)\}$. For concision, when $\mathcal{X}$ is understood from the context, we will drop it altogether and we will write $\operatorname{TC}(x)$ and $\operatorname{PC}(x)$ instead.

Finally, the asymptotic equality notation "$f(t) \sim g(t)$ for large t" means that $\lim_{t\to\infty} f(t)/g(t) = 1$; the symbols "$\lesssim$" and "$\gtrsim$" are defined analogously.

2.1. **Mirror descent.** Dating back to Nemirovski and Yudin [41], the main idea of mirror descent is as follows: Given a smooth convex objective $f \colon \mathcal{X} \to \mathbb{R}$, the optimizer takes an infinitesimal step along the negative gradient of f in the dual space $\mathcal{V}^*$; the output is then "mirrored" back to the problem's feasible region $\mathcal{X} \subseteq \mathcal{V}$

---

[2]That is, $f(\bar{X}(t)) - \min f = \mathcal{O}(t^{-1/2}\sqrt{\log\log t})$ except for a set of measure zero.

and the process continues. More precisely, in continuous time, the dynamics of this process can be represented as

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

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 (see 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).

A key element in the above description of mirror descent is the distinction between primal and dual variables – that is, between candidate solutions $x \in \mathcal{X}$ and score variables $y \in \mathcal{V}^*$. To emphasize this duality, we will write $\mathcal{Y} \equiv \mathcal{V}^*$ for the dual space of $\mathcal{V}$ and, following [43], we will often refer to the dynamics (MD) as *dual averaging*. Also, in terms of regularity, we will assume that

$$v(x) \text{ is Lipschitz continuous on } \mathcal{X}. \tag{$\mathbf{H}_1$}$$

Strictly speaking, Hypothesis ($\mathbf{H}_1$) is not needed for much of the analysis of (MD) and can be replaced e.g. by global integrability of v; however, it simplifies the presentation considerably, so we keep it throughout our paper.

Given that the dual variable y aggregates (negative) gradient steps, a reasonable candidate for the mirror map Q might appear to 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 is set-valued and 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:

Definition 2.1. We say that $h \colon \mathcal{X} \to \mathbb{R}$ is a *regularizer* (or *penalty function*) 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 defined as

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

In view of the above, we have $Q(\eta y) = \arg\max_{x \in \mathcal{X}} \{\langle y | x \rangle - \eta^{-1} h(x)\}$, so η essentially controls the weight of the penalty term $h(x)$ in (2.2). Consequently, 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 $\mathcal{X}$ (implying in turn that the primal variable x becomes less susceptible to changes in y, hence the name "sensitivity").

For concreteness, we discuss below some examples of this construction:

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

$$\Pi(y) = \arg\max_{x \in \mathcal{X}} \{\langle y | x \rangle - \tfrac{1}{2}\|x\|_2^2\} = \arg\min_{x \in \mathcal{X}} \|y - x\|_2^2. \tag{2.3}$$

The dynamics derived from (2.3) may thus be viewed as a continuous-time version of (Euclidean) projected gradient descent [4, 35, 43]. For future reference, we also note that h is differentiable throughout $\mathcal{X}$ and Π is *surjective* (i.e. $\operatorname{im}\Pi = \mathcal{X}$).

Example 2.2 (Entropic regularization). Let $\Delta = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$ denote the unit simplex of $\mathbb{R}^n$ and consider the (negative) Gibbs entropy

$$h(x) = \sum_{i=1}^n x_i \log x_i. \tag{2.4}$$

The function $h(x)$ is 1-strongly convex with respect to the L^1-norm on $\mathbb{R}^n$ and a straightforward calculation shows that the induced mirror map is

$$\Lambda(y) = \frac{1}{\sum_{i=1}^n \exp(y_i)}(\exp(y_1), \ldots, \exp(y_n)). \tag{2.5}$$

This model is known as *logit choice* and the associated dynamics have been studied extensively in linear programming [27], online learning [50] and game theory [22]. In contrast to Example 2.1, h is differentiable *only* on the relative interior Δ° of Δ and $\operatorname{im}\Lambda = \Delta^\circ$ (i.e. Λ is "essentially" surjective).

Example 2.3 (Matrix regularization). Motivated by applications to semidefinite programming, consider the unit spectrahedron $\mathcal{D} = \{\mathbf{X} \in \operatorname{Sym}(\mathbb{R}^n) : \mathbf{X} \succcurlyeq 0, \operatorname{tr}\mathbf{X} \leq 1\}$ of positive-semidefinite matrices with nuclear norm $\|\mathbf{X}\|_1 = \operatorname{tr}\mathbf{X} \leq 1$. A widely used regularizer on $\mathcal{D}$ is provided by the *von Neumann entropy* [53]

$$h(\mathbf{X}) = \operatorname{tr}(\mathbf{X}\log\mathbf{X}) + (1 - \operatorname{tr}\mathbf{X})\log(1 - \operatorname{tr}\mathbf{X}), \tag{2.6}$$

which is $(1/2)$-strongly convex with respect to the nuclear norm [25]. A straightforward calculation [36] then shows that the induced mirror map is given by

$$\mathbf{\Lambda}(\mathbf{Y}) = \frac{\exp(\mathbf{Y})}{1 + \|\exp(\mathbf{Y})\|_1} \quad \text{for all } \mathbf{Y} \in \operatorname{Sym}(\mathbb{R}^n). \tag{2.7}$$

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

The examples above highlight an important relationship between the domain of differentiability of h and the image of the induced mirror map Q. To describe it in detail, extend h to all of $\mathcal{V}$ by setting $h \equiv \infty$ outside $\mathcal{X}$, and let $\operatorname{dom}\partial h \equiv \{x \in \mathcal{X} : \partial h(x) \neq \varnothing\}$ be the domain of subdifferentiability of h. We then have the following characterization of Q:

Proposition 2.2. *Let h be a K-strongly convex regularizer, let $Q\colon \mathcal{Y} \to \mathcal{X}$ be the mirror map induced by h, and let $h^*(y) = \max\{\langle y|x\rangle - h(x) : x \in \mathcal{X}\}$ denote the convex conjugate of h. 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.*

Proof. Standard; see e.g. [48, Theorem 23.5] and [49, Theorem 12.60(b)]. ∎

Since $\mathcal{X}^\circ \subseteq \operatorname{dom}\partial h \subseteq \mathcal{X}$ [48, Chap. 23], Proposition 2.2 shows that Q is "almost" surjective; specifically, the only points of $\mathcal{X}$ that do not belong to $\operatorname{im}Q$ are boundary points of $\mathcal{X}$ where h becomes "infinitely steep". Motivated by this, we say that h is

steep at x if $\partial h(x) = \varnothing$ and *nonsteep* otherwise. As a result, regularizers that are everywhere nonsteep induce mirror maps that are surjective (Example 2.1), while regularizers that are steep throughout $\mathrm{bd}(\mathcal{X})$ give rise to interior-valued mirror maps (Example 2.3).

2.2. Bregman divergences and the Fenchel coupling. Another 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 [29], $D(p, x)$ 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 $h'(x; z) = \lim_{t \to 0^+} t^{-1}[h(x + tz) - h(x)]$ denotes the one-sided derivative of h at x along $z \in \mathrm{TC}(x)$. Given that h is strictly convex, we have $D(p, x) \geq 0$ and $x(t) \to p$ whenever $D(p, x(t)) \to 0$; hence, the convergence of $x(t)$ to p can be checked by means of 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) if $x(t)$ is not interior.[3] To overcome this difficulty, we will instead employ the so-called *Fenchel coupling*

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

so named because it collects all terms of Fenchel's inequality.[4] This "primal-dual" divergence was first introduced in [35, 38] and, as a consequence of Fenchel's inequality, it follows that $F(p, y) \geq 0$ with equality if and only if $p = Q(y)$.

The following proposition (taken from [35]) links the Fenchel coupling with the Bregman divergence and the underlying norm:

Proposition 2.3. *Let h be a K-strongly convex regularizer 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 whenever* $Q(y) \in \mathcal{X}^\circ$. $\tag{2.10a}$

b) $F(p, y) \geq \frac{1}{2} K \|Q(y) - p\|^2$. $\tag{2.10b}$

c) $F(p, y') \leq F(p, y) + \langle y' - y | Q(y) - p \rangle + \frac{1}{2K} \|y' - y\|_*^2$. $\tag{2.10c}$

Proof. See [35, Proposition 4.3]. ∎

An immediate consequence of (2.10b) is that $Q(y_n) \to 0$ for every sequence $(y_n)_{n=0}^\infty$ in $\mathcal{Y}$ such that $F(p, y_n) \to 0$. As a result, the convergence of $x(t) = Q(y(t))$ to $p \in \mathcal{X}$ may be checked by showing that $F(p, y(t)) \to 0$. For technical reasons, it will be convenient to assume that the converse also holds, i.e.

$$F(p, y_n) \to 0 \quad \text{whenever} \quad Q(y_n) \to p. \tag{$\mathbf{H_2}$}$$

---

[3]To understand this, consider the case where $\mathcal{X} = [0, 1]$ and $Q = \Pi$, the Euclidean projector of Example 2.1. If we take the objective $f(x) = x$ and start (MD) at $y_0 = a > 1$, then $x(t)$ would be stuck at 1 for all $t \in [0, a-1]$. The Bregman divergence would not be able to detect the evolution of $y(t)$ in this case (in tune with the fact that (MD) cannot be recast as an autonomous dynamical system in terms of x when h is not steep); for a detailed discussion, see [38].

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

When h is steep, combining Propositions 2.2 and 2.3 gives $F(p, y) = D(p, Q(y))$ for all $y \in \mathcal{Y}$,[5] so ($\mathbf{H}_2$) boils down to the requirement

$$D(p, x_n) \to 0 \quad \text{whenever } x_n \to p. \tag{2.11}$$

This so-called "reciprocity condition" is well known in the theory of Bregman functions [3, 17, 29] and, essentially, it means that the sublevel sets of $D(p, \cdot)$ are neighborhoods of p in $\mathcal{X}$. Hypothesis ($\mathbf{H}_2$) instead posits that the *images* of the sublevel sets of $F(p, \cdot)$ under Q are neighborhoods of p in $\mathcal{X}$, so ($\mathbf{H}_2$) may be seen as a "primal-dual" variant of Bregman reciprocity.

It is easy to verify that Examples 2.1–2.3 all satisfy ($\mathbf{H}_2$). For an in-depth discussion of the geometric implications of Bregman reciprocity, the reader is referred to [29].

2.3. **Deterministic analysis.** Together with Proposition 2.2, the Lipschitz continuity hypothesis ($\mathbf{H}_1$) implies that the driving vector field $v(Q(\eta y))$ 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 $y_0 \in \mathcal{Y}$ [47, Chap. V]. With this in mind, we have:

Theorem 2.4. *Assume* ($\mathbf{H}_1$) *holds and let* $x(t) = Q(\eta y(t))$ *be a solution of* (MD) *initialized at* $y_0 \in \mathcal{Y}$.

(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 *under* (MD), *we have*

$$f_{\min}(t) - \min f \le \bar{f}(t) - \min f = \mathcal{O}(1/t). \tag{2.12}$$

In particular, if (MD) *is initialized at* $y_0 = 0$, *we have*
$$f_{\min}(t) \le \bar{f}(t) \le \min f + \Omega/t, \tag{2.13}$$

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

(2) *If* ($\mathbf{H}_2$) *also holds,* $x(t)$ *converges to some* $x^* \in \arg\min f$ *(possibly depending on* y_0).

Theorem 2.4 is a strong convergence result guaranteeing global trajectory convergence to a solution of (P) and an $\mathcal{O}(1/t)$ value convergence rate for the averaged process $\bar{x}(t) = t^{-1} \int_0^t x(s)\, ds$ (by Jensen's inequality). In Appendix B.1, we provide a Lyapunov-based proof leveraging the fact that the "η-deflated" Fenchel coupling

$$V(t) = \eta^{-1} F(x^*, \eta y(t)), \tag{2.14}$$

is nondecreasing along the solution orbits of (MD) for all $x^* \in \arg\min f$.

The first part of the theorem is well known and essentially dates back to the original work of Nemirovski and Yudin [41]. As for the trajectory convergence properties of (MD), [3, 14] provide a proof for a Hessian Riemannian gradient system which is formally equivalent to (MD) when h is steep (for a detailed discussion, see Section 5); [4] also deals with the singular Riemannian case (corresponding to non-steep h), but requires that $\mathcal{X}$ be polyhedral. Finally, [3, 31] also provide an $\mathcal{O}(1/t)$

---

[5]To be clear, Proposition 2.3 guarantees that $F(p, y) = D(p, Q(y))$ whenever $Q(y)$ is interior. The statement for steep h is sharper because it states that $F(p, y) = D(p, Q(y))$ *for all* y. This is a consequence of the fact that $\operatorname{im} Q = \mathcal{X}^\circ$ for steep h, hence the need to invoke Proposition 2.2.

value convergence rate for $x(t)$; under ($\mathbf{H_2}$), Part (ii) of Theorem 2.4 narrows this convergence down to a *point* $x^* \in \arg\min f$ (instead of the *set* $\arg\min f$).[6]

Building on this basic deterministic result, our aim in the rest of this paper will be to explore how the strong convergence properties of (MD) are affected if the gradient input of (MD) is contaminated by noise.

3. MIRROR DESCENT WITH NOISY GRADIENT INPUT

To account for noise and measurement errors in (MD), our starting point will be the random 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 $t \geq 0$. To write the Langevin equation (3.1) as a formal stochastic differential equation, let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space,[7] and consider the stochastic mirror descent dynamics

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

where $Z(t) = (Z_1(t), \ldots, Z_n(t))$ is a continuous $\mathcal{F}_t$-adapted Itô martingale. More precisely, we assume throughout that $Z(t)$ is of the general form

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

where:

(1) $W = (W_1, \ldots, W_m)$ is an adapted m-dimensional Wiener process.[8]
(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. More formally, we posit that

$$\begin{aligned} \sup_{x,t} |\sigma_{ik}(x,t)| &< \infty, \\ |\sigma_{ik}(x',t) - \sigma_{ik}(x,t)| &\leq \ell \, \|x' - x\|. \end{aligned} \tag{$\mathbf{H_3}$}$$

for some $\ell > 0$ and for all $x, x' \in \mathcal{X}$, $t \geq 0$.

The most straightforward case for the noise is when $m = n$ and $Z(t) = \sigma W(t)$ for constant σ. This case corresponds to i.i.d. increments that are uncorrelated across different components and that do not depend on t or $X(t)$. However, these independence assumptions are not always realistic: in Section 5.1, we discuss an important example with nontrivial correlations that arise in the study of traffic networks, and which necessitate the more general treatment above.

---

[6]If Q is smooth (as opposed to Lipschitz), a simple differentiation shows that $f(x(t))$ is nonincreasing in t. In this case, an $\mathcal{O}(1/t)$ convergence for $f(x(t))$ follows readily from an $\mathcal{O}(1/t)$ upper bound on $\bar{f}(t)$ by noting that $f(x(t)) = t^{-1} \int_0^t f(x(t))\, ds \leq t^{-1} \int_0^t f(x(s))\, ds = \bar{f}(t)$.

[7]We tacitly assume here that $\mathcal{F}_t$ satisfies the usual conditions, i.e. it is complete ($\mathcal{F}_0$ contains all $\mathbb{P}$-null sets) and right-continuous ($\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$).

[8]It is possible to consider even more general continuous semimartingale error terms here, but the presentation would become much more complicated.

More concretely, the correlation structure of the noise process Z can be captured by the *quadratic covariation process* $[Z(t), Z(t)]$,[9] given here by the SDE

$$d[Z_i(t), Z_j(t)] = \sum_{k,\ell=1}^{m} \sigma_{ik}(X(t),t)\sigma_{j\ell}(X(t),t)\, dW_k(t) \cdot dW_\ell(t)$$

$$= \sum_{k=1}^{m} \sigma_{ik}(X(t),t)\sigma_{jk}(X(t),t)\, dt = \Sigma_{ij}(X(t),t)\, dt, \qquad (3.3)$$

where $\Sigma = \sigma\sigma^\top$ is the *infinitesimal covariance matrix* of the process. If Σ is not diagonal, the components of Z exhibit nontrivial correlations quantified by the nonzero off-diagonal elements of Σ. This also highlights the role of the underlying m-dimensional Wiener process $W(t)$ in (SMD): if $m < n$, the induced disturbances are necessarily correlated; if $m = n$ and σ is diagonal, the errors are independent across components; and if $m > n$, the noise in each component may result from the aggregation of several, independent error sources. Obviously, the precise statistics of the noise depend crucially on the application being considered, so, for generality, we maintain an application-agnostic approach and we make no assumptions on the structure of Σ.

For posterity, we also note here that the noise regularity hypothesis ($\mathbf{H}_3$) gives

$$\|\sigma(x,t)\|_F^2 \leq \sigma_*^2 \quad \text{for some } \sigma_* > 0 \text{ and all } x \in \mathcal{X},\, t \geq 0, \qquad (3.4)$$

where

$$\|\sigma\|_F \equiv \sqrt{\operatorname{tr}[\sigma\sigma^\top]} = \sqrt{\operatorname{tr}[\Sigma]} \qquad (3.5)$$

denotes the Frobenius norm of the $n \times m$ matrix σ. In what follows, it will be convenient to measure the magnitude of the noise affecting (SMD) via σ_*;[10] obviously, when $\sigma_* = 0$, we recover the noiseless, deterministic dynamics (MD).

Now, under the Lipschitz continuity hypothesis ($\mathbf{H}_1$) and the noise regularity condition ($\mathbf{H}_3$), standard results from the theory of stochastic differential equations show that (SMD) admits unique strong solutions that exist for all time (see e.g. Theorem 3.21 in [44]).[11] 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 [28, 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. *Assume* ($\mathbf{H}_1$) *and* ($\mathbf{H}_3$) *hold. Then, for all* $y_0 \in \mathcal{Y}$ *and up to a* $\mathbb{P}$*-null set,* (SMD) *admits a unique strong solution* $(Y(t))_{t \geq 0}$ *such that* $Y(0) = y_0$.

In the rest of the paper, when we refer to a solution trajectory of (SMD), we will implicitly invoke the well-posedness result above without making an explicit reference to it.

---

[9]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 $\Pi = \{t_0 = 0 < t_1 < \cdots < t_k = t\}$ of $[0, t]$ with mesh $|\Pi| \equiv \max_j |t_j - t_{j-1}| \to 0$ [26].

[10]Note here that σ_*^2 typically scales with the dimensionality of $\mathcal{X}$ (for instance, if Z is a standard n-dimensional Wiener process).

[11]The Lipschitz continuity of the drift and diffusion terms of (SMD) is key in this regard.

	Hypothesis	Statement
($\mathbf{H}_1$)	Lipschitz gradients	$v(x)$ is Lipschitz continuous
($\mathbf{H}_2$)	Bregman reciprocity	$F(p, y_n) \to 0$ whenever $Q(y_n) \to p$
($\mathbf{H}_3$)	Noise regularity	$\sigma(x, t)$ is bounded and Lipschitz in x

Table 1. Overview of the various hypotheses used in the paper.

4. Convergence results

Despite the strong convergence properties of the deterministic dynamics (MD), the noise-contaminated dynamics (SMD) may fail to converge, even in simple, one-dimensional problems. For an elementary example, take $f(x) = x^2/2$ over $\mathcal{X} = [-1, 1]$, let $Z(t) = W(t)$: since the martingale part of (SMD) does not vanish when $X(t) = 0$, it follows that $X(t)$ cannot converge to $\arg\min f = \{0\}$ with positive probability – and this, independently of the choice of mirror map Q.

In view of this nonconvergent example, our aim in the rest of this section will be to:

(1) Analyze the convergence properties of (SMD) in the "vanishing noise" regime (Section 4.1).
(2) Study the long-run concentration properties of $X(t)$ around interior minimizers (Section 4.2).
(3) Identify classes of convex programs where $X(t)$ *does* converge (Section 4.3).
(4) Examine a convergent variant of (SMD) with a decreasing sensitivity parameter (Section 4.4).

4.1. **The vanishing noise regime.** 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 over time [36]. In this "vanishing noise" limit, intuition suggests that $X(t)$ should asymptotically follow the dynamics (MD), and hence converge (in some sense) to $\arg\min f$.

To make this intuition precise, we first show below that $\arg\min f$ is *recurrent* under $X(t)$, i.e. $X(t)$ visits any neighborhood of $\arg\min f$ infinitely often:

Proposition 4.1. *Assume* ($\mathbf{H}_1$) *and* ($\mathbf{H}_3$) *hold, and let* $X(t) = Q(\eta Y(t))$ *be a solution of* (SMD)*. If* $\lim_{t\to\infty} \sup_{x\in\mathcal{X}} \|\sigma(x, t)\|_F = 0$*, there exists a (random) sequence of times* $t_n \uparrow \infty$ *such that* $X(t_n) \to \arg\min f$ (a.s.)*.*

As in the noiseless case (and much of the analysis to follow), the proof of Proposition 4.1 hinges on the "η-deflated" Fenchel coupling

$$V(t) = \eta^{-1}F(x^*, \eta Y(t)), \tag{4.1}$$

which satisfies the (stochastic) Lyapunov-like property

$$V(t) - V(0) \le \int_0^t \langle v(X(s)) \,|\, X(s) - x^* \rangle \, ds \qquad \text{(drift)}$$

$$+ \frac{\eta}{2K} \int_0^t \text{tr}[\Sigma(X(s), s)] \, ds \qquad \text{(Itô correction)}$$

$$+ \sum_{i=1}^{n} \int_{t_0}^{t} (X_i(s) - x_i^*)\, dZ_i(s) \qquad \text{(martingale noise)} \qquad (4.2)$$

Arguing by contradiction, if $X(t)$ remained a bounded distance away from $\arg\min f$, the drift term in (4.2) would decrease linearly in t for all $x^* \in \arg\min f$ (by convexity). Since the Itô correction and martingale noise terms grow sublinearly in t (by the vanishing noise assumption and the law of large numbers respectively), this would give $V(t) \to -\infty$, contradicting the fact that $V(t) \geq 0$.

Of course, Proposition 4.1 is considerably weaker than its deterministic counterpart (Theorem 2.4), because it does not even imply that $X(t) \to \arg\min f$ with positive probability. Nonetheless, by slightly strengthening the vanishing noise requirement $\sup_x \|\sigma(x,t)\|_F \to 0$, we obtain that $X(t) \to \arg\min f$ *with probability* 1:

Theorem 4.2. *Assume* $(\mathbf{H}_1)$–$(\mathbf{H}_3)$ *hold and let* $X(t) = Q(\eta Y(t))$ *be a solution of* (SMD). *If* $\sup_{x \in \mathcal{X}} \|\sigma(x,t)\|_F = o(1/\sqrt{\log t})$, *we have* $X(t) \to \arg\min f$ *(a.s.).*

The key challenge in obtaining this a.s. convergence result is that, even if we ignored the martingale term in (4.2), it is quite difficult to balance the drift (helpful) and Itô correction (antagonistic) terms. Thus, in lieu of a direct Lyapunov approach, we will show that $X(t)$ "tracks" the deterministic dynamics (MD) in a certain, precise sense (see below), and then leverage the convergence properties of (MD) to deduce that $X(t) \to \arg\min f$.

To quantify what "tracking" means in this context, we use the seminal notion of an *asymptotic pseudotrajectory* (APT) due to Benaïm and Hirsch [8, 9]:

Definition 4.3. Let $(Y(t))_{t \geq 0}$ be a continuous curve in $\mathcal{Y}$ and let $\Phi_t \colon \mathcal{Y} \to \mathcal{Y}$, $t \geq 0$, be the semiflow of (MD) on $\mathcal{Y}$ (i.e. $(\Phi_t(y))_{t \geq 0}$ denotes the solution orbit of (MD) that starts at $y \in \mathcal{Y}$). Then, Y is an *asymptotic pseudotrajectory* (APT) of Φ if

$$\lim_{t \to \infty} \sup_{0 \leq h \leq T} \|Y(t+h) - \Phi_h(Y(t))\|_* = 0 \quad \text{for all } T > 0. \qquad (4.3)$$

Heuristically, an APT of (MD) asymptotically follows the induced semiflow Φ with arbitrary accuracy over windows of arbitrary length. Nonetheless, this "fixed horizon" property does not suffice to establish the convergence of an APT to $\arg\min f$, despite the strong convergence properties of (MD). On that account, the basic steps of our proof are as follows:

i) Using the analysis of [9], we show that the stated decay assumption for $\sigma(x,t)$ implies that solutions of (SMD) are APTs of (MD).

ii) By Proposition 4.1, $\arg\min f$ is recurrent under (SMD), so solutions of (SMD) cannot stray too far from $\arg\min f$ in the long run.

iii) Once a solution of (SMD) gets close enough to $\arg\min f$, the APT property means that it becomes trapped in its vicinity and eventually converges to it.

We make all this precise in Appendix B.2 where we prove Proposition 4.1 and Theorem 4.2.

4.2. **Long-run concentration around solution points.** Beyond the vanishing noise regime, the simple example $f(x) = x^2/2$ with $Z(t) = W(t)$ shows that $X(t)$ may fluctuate around $\arg\min f$ in perpetuity if the noise is persistent. As such, our goal in what follows will be to analyze the long-run concentration properties of (SMD)

and to determine the domain that $X(t)$ occupies with high probability in the long run.

For reasons that will become clear shortly, we focus on strongly convex problems that admit a (necessarily unique) interior solution $x^* \in \mathcal{X}^\circ$. More concretely, this means that there exists some $\alpha > 0$ (related to the convexity of the problem) such that

$$f(x) - f(x^*) \geq \tfrac{1}{2}\alpha\|x - x^*\|^2 \quad \text{for all } x \in \mathcal{X}. \tag{4.4}$$

Our first result in this case is as follows:

Proposition 4.4. *Assume* $(\mathbf{H}_1)$ *and* $(\mathbf{H}_3)$ *hold, and let* f *be an* α-*strongly convex function with an interior minimizer* $x^* \in \mathcal{X}^\circ$. *If* $X(t) = Q(\eta Y(t))$ *is a solution of* (SMD) *initialized at* $y_0 \in \mathcal{Y}$, *we have*

$$\mathbb{E}\left[\frac{1}{t}\int_0^t \|X(s) - x^*\|^2\, ds\right] \leq \frac{2F(x^*, \eta y_0)}{\eta\alpha t} + \frac{\eta\sigma_*^2}{\alpha K}. \tag{4.5}$$

Moreover, if $\tau_\delta = \inf\{t > 0 : \|X(t) - x^*\| \leq \delta\}$ *denotes the first time at which* $X(t)$ *gets within* $\delta > 0$ *of* x^*, *we also have*

$$\mathbb{E}[\tau_\delta] \leq \frac{2KF(x^*, \eta y_0)}{\eta\alpha K\delta^2 - \eta^2\sigma_*^2}, \tag{4.6}$$

provided that $\eta < \alpha K\delta^2/\sigma_*^2$. *In particular, for* $y_0 = 0$, *we have the optimized bound*

$$\mathbb{E}[\tau_\delta] \leq \frac{8\Omega\sigma_*^2}{\alpha^2 K\delta^4}, \tag{4.7}$$

achieved for $\eta = \alpha K\delta^2/(2\sigma_*^2)$.

Remark 4.1. In the above, the constant α has to do with the objective function f and the feasible region $\mathcal{X}$, while K and Ω are linked to the mirror map Q (and, of course, also $\mathcal{X}$). The optimizer has no control over the former, but if its value can be estimated and the geometry of $\mathcal{X}$ is relatively simple, the latter can be finetuned further to sharpen the above bounds.

Remark 4.2. For a value-based analogue of (4.5) when h is steep, see [46, Prop. 4].

Proposition 4.4 (proved in Appendix B.3) provides a basic estimate of the long-run concentration of $X(t)$ around x^*, and also highlights the role of α and σ_*. Specifically, (4.7) shows that $X(t)$ hits a δ-neighborhood of x^* in time which is $\mathcal{O}(1/\delta^4)$ on average; what's more, the multiplicative constant in this bound increases with the noise level in (SMD) and decreases with the sharpness of the minimum point x^* (as quantified by the strong convexity constant α of f).

To obtain finer information regarding the concentration of $X(t)$ around x^*, we need to consider its occupation measure:

Definition 4.5. The *occupation measure* of X at time $t \geq 0$ is given by

$$\mu_t(A) = \frac{1}{t}\int_0^t \mathbb{1}(X(s) \in A)\, ds \quad \text{for every Borel } A \subseteq \mathcal{X}. \tag{4.8}$$

In words, $\mu_t(A)$ is the fraction of time that X spends in A up to time t. As such, the asymptotic concentration of X around x^* can be estimated by the quantity $\mu_t(\mathbb{B}_\delta)$, where

$$\mathbb{B}_\delta \equiv \mathbb{B}_\delta(x^*) = \{x \in \mathcal{X} : \|x - x^*\| \leq \delta\} \tag{4.9}$$

is the intersection of a δ-ball centered at x^* with $\mathcal{X}$. We then have the following concentration result (for a numerical illustration, see Fig. 1):
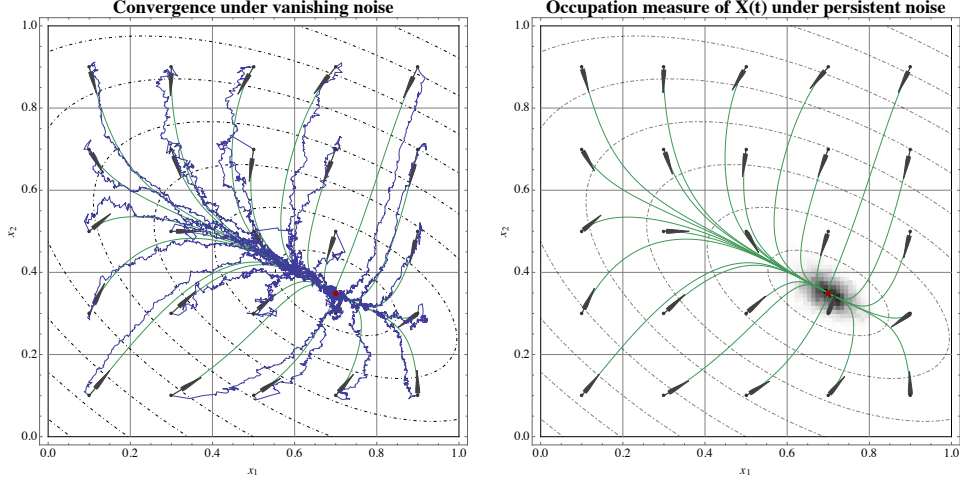
Figure 1. Numerical illustration of (SMD) with $Q(y) = e^y/(1 + e^y)$. The dashed contours represent the level sets of f over $\mathcal{X} = [0, 1]^2$, and the flowlines indicate the flow of (MD). In the first figure, we exhibit the convergence of (SMD) to $\arg\min f$ when the volatility of the noise decays as $\Theta(1/\log t)$. In the second, we estimate the long-run occupation measure of X: darker shades of gray correspond to higher probabilities of observing X in a given region.

Theorem 4.6. *Assume* ($\mathbf{H}_1$) *and* ($\mathbf{H}_3$) *hold, and let f be an α-strongly convex function admitting an interior minimizer $x^* \in \mathcal{X}^\circ$. Moreover, fix some $\delta > 0$ and suppose 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 $\eta < \alpha K \delta^2/\sigma_*^2$, then*

$$\mu_t(\mathbb{B}_\delta) \gtrsim 1 - \frac{\eta \sigma_*^2}{\alpha K \delta^2} \quad \textit{for sufficiently large } t \; (a.s.). \tag{4.10}$$

Corollary 4.7. *Fix some tolerance $\varepsilon > 0$. If (SMD) is run with assumptions as above and $\eta \leq \varepsilon \alpha K \delta^2/\sigma_*^2$, we have $\mu_t(\mathbb{B}_\delta) \geq 1 - \varepsilon$ for all sufficiently large t (a.s.).*

Remark 4.3. 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 $\mathrm{rank}(\sigma) = n$, i.e. the components of Z are not completely correlated. For instance, this condition is trivially satisfied in the baseline case where Z is a Wiener process in $\mathbb{R}^n$.

Remark 4.4. It is also worth noting that the bound (4.10) only depends on the mirror map Q via its inverse Lipschitz constant K (that is, the strong convexity constant of h). Eq. (4.10) suggests that K should be taken as large as possible (to have $\mu_t(\mathbb{B}_\delta) \approx 1$). However, in so doing, the process $X(t)$ will initially spend a much larger amount of time near the prox-center $x_c \equiv \arg\min h$ of $\mathcal{X}$, so there is a trade-off between the sharpness of the asymptotic concentration of $X(t)$ near x^* and the time it takes to attain this asymptotic regime.

In a nutshell, Theorem 4.6 states that the concentration of $X(t)$ around x^* may be arbitrarily sharp if η is taken small enough. Indeed, for $\eta < \alpha K \delta^2/\sigma_*^2$,

Proposition 4.4 shows that $\mathbb{B}_\delta$ is *recurrent*, i.e. $\mathbb{P}(X(t) \in \mathbb{B}_\delta$ for some $t \geq 0) = 1$ for every initial condition $y_0 \in \mathcal{Y}$. Relegating the (fairly intricate) details to Appendix B.3, it can be shown that the stated assumptions guarantee the existence of a unique invariant distribution ν for the dual process $Y(t)$. The pushforward of ν to $\mathcal{X}$ is precisely the limit of the occupation measures μ_t of X as $t \to \infty$, so (4.10) follows by using the mean square bound (4.5) to estimate ν.

We close this section by noting that the assumption that x^* is interior is crucial in the statement of Theorem 4.6. 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 (a.s.) to the smallest face of $\mathcal{X}$ that contains x^* and fluctuates around x^* along the relative interior of said face. However, obtaining a precise result along these lines is fairly cumbersome, so we omit this analysis.

4.3. Sharp solutions and linear programming.
Consider now the elementary linear program

$$\begin{aligned} \text{minimize} \quad & 1 - x, \\ \text{subject to} \quad & 0 \leq x \leq 1. \end{aligned} \tag{4.11}$$

Taking for concreteness $\eta = 1$, $h(x) = x \log x + (1-x) \log(1-x)$ and $Z(t) = \sigma W(t)$ with constant σ, the dynamics (SMD) become

$$\begin{aligned} dY &= dt + \sigma\, dW, \\ X &= e^Y/(1 + e^Y), \end{aligned} \tag{4.12}$$

and, after integrating, we get $Y(t) = t + \sigma W(t)$. By a trivial stochastic estimate, this implies that $Y(t) \geq t/2$ for large t (except possibly on a $\mathbb{P}$-null set), so $\lim_{t \to \infty} X(t) = 1$ (a.s.). In other words, in the simple linear program (4.11), $X(t)$ converges to $\arg\min f$ with probability 1, no matter the level of the noise.

The reason behind this convergence (as opposed to the case of interior minimizers) is that the drift of (4.12) does not vanish when $X(t)$ approaches $\arg\min f$, so it ends up dominating the martingale term $W(t)$. 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. Following Polyak [45, Chapter 5.2], we formalize this idea by focusing on convex programs with *sharp* solutions:

Definition 4.8. We say that $x^* \in \mathcal{X}$ is a *γ-sharp* minimum point of f if

$$f(x) \geq f(x^*) + \gamma \|x - x^*\| \quad \text{for some } \gamma > 0 \text{ and all } x \in \mathcal{X}. \tag{4.13}$$

From Definition 4.8, it is easy to see that a sharp minimum point is the unique minimizer of f and it remains invariant under small perturbations of f (assuming of course that such a minimizer exists in the first place). On top of that, with f assumed smooth,[12] we also have the following geometric characterization:

Lemma 4.9. $x^* \in \mathcal{X}$ *is a γ-sharp solution of* (P) *if and only if*

$$\langle v(x^*) \,|\, z \rangle \leq -\gamma \|z\| \quad \text{for some } \gamma > 0 \text{ and for all } z \in \mathrm{TC}(x^*). \tag{4.14}$$

---

[12]Definition 4.8 is meaningful even if f is not smooth, but we only treat smooth functions here.

Proof. The "if" part follows trivially by convexity. For the "only if" part, let $z \in \mathrm{TC}(x^*)$ and note that (4.13) gives

$$\frac{f(x^* + tz) - f(x^*)}{t} \geq \gamma \|z\| \quad \text{for all sufficiently small } t > 0. \tag{4.15}$$

Hence, taking the limit $t \to 0^+$, we get $\langle \nabla f(x^*) | z \rangle \geq \gamma \|z\|$ and (4.14) follows. ∎

A further consequence of Lemma 4.9 is that $v(x^*) \in \mathrm{int}(\mathrm{PC}(x^*))$, implying that sharp solutions of smooth convex programs can only occur at *corners* of $\mathcal{X}$ (that is, points whose polar cone has nonempty topological interior). In this sense, sharp minimizers constitute the flip side of the interior-point analysis of the previous section, a contrast which is further reflected in the following a.s. convergence result:

Theorem 4.10. *Assume* $(\mathbf{H}_1)$–$(\mathbf{H}_3)$ *hold and suppose that* f *admits a* (*necessarily unique*) *sharp minimum point* x^*. *If* (SMD) *is run with a sufficiently small 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.*).

As an important special case, note that every solution of a (generic) linear program is sharp.[13] Theorem 4.10 then gives:

Corollary 4.11. *If* (P) *is a generic linear program and* (SMD) *is run with Euclidean projections* (*cf. Example 2.1*) *and small enough* η, $X(t)$ *converges to* $\arg\min f$ *in finite time* (*a.s.*).

To gain some insight in the proof of Theorem 4.10, note first that the driving vector field $v(x)$ of (SMD) points towards x^* for all $x \in \mathcal{X}$ (by convexity). Thanks to this basic property, almost every solution of (SMD) visits any neighborhood of x^* infinitely many times (a.s.). However, when $X(t)$ is near x^*, the sharpness of the solution "traps" $X(t)$ near x^* and does not allow any overshoots (as in the interior case) because x^* is a corner of $\mathcal{X}$. By a hitting time argument based on Girsanov's theorem, it is then possible to show that the dual process $Y(t)$ escapes to infinity along a direction contained in the polar cone $\mathrm{PC}(x^*)$ of $\mathcal{X}$ at x^*. Then, the a.s. convergence of $X(t)$ to x^* follows from a straightforward geometric argument.

We make all this precise in Appendix B.4.

4.4. **Rectification.** In this section, we examine a "rectified" variant of (SMD) which is run with a decreasing sensitivity parameter and which takes into account all past information up to time t. Specifically, motivated by Theorem 2.4(i), consider the transformed process

$$\widetilde{X}(t) = \frac{1}{t} \int_0^t X(s) \, ds, \tag{4.16a}$$

or

$$\widetilde{X}(t) = X(s_t) \quad \text{with } s_t \in \arg\min_{0 \leq s \leq t} f(X(s)), \tag{4.16b}$$

corresponding respectively to the long-run average (also known as the "ergodic average" in optimization) and the "best value" of X up to time t.

The results of [46] and the analysis of Section 4.2 indicate that $\widetilde{X}(t)$ is concentrated around interior solutions of $\mathcal{X}$ (in the long run and in probability), provided that (SMD) is run with sufficiently small η. That said, in a black-box setting where

---

[13]"Generic linear program" means here that $\mathcal{X}$ is a polytope, $f \colon \mathcal{X} \to \mathbb{R}$ is affine, and f is constant only along the zero-dimensional faces of $\mathcal{X}$ [45].

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. Thus, a meaningful work-around would be to employ a variable sensitivity parameter $\eta \equiv \eta(t)$ which decreases to 0 as $t \to \infty$.

Since $Y(t) = \mathcal{O}(t)$ by the Lipschitz assumption ($\mathbf{H}_1$), $\eta(t)$ 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 make the following assumption throughout this section:

$$\eta(t) \text{ is Lipschitz continuous, nonincreasing, and } \lim_{t\to\infty} t\eta(t) = \infty. \qquad (\mathbf{H}_4)$$

Under this assumption, we have:

Theorem 4.12. *Assume* ($\mathbf{H}_1$), ($\mathbf{H}_3$) *and* ($\mathbf{H}_4$) *hold. Then, the rectified process* $\widetilde{X}(t)$ *enjoys the performance guarantees*

$$f(\widetilde{X}(t)) \leq \min f + \frac{\Omega}{t\eta(t)} + \frac{\sigma_*^2}{2Kt} \int_0^t \eta(s)\,ds + \mathcal{O}(\sqrt{\log\log t/t}) \quad (a.s.), \quad (4.17)$$

and

$$\mathbb{E}[f(\widetilde{X}(t))] \leq \min f + \frac{\Omega}{t\eta(t)} + \frac{\sigma_*^2}{2Kt} \int_0^t \eta(s)\,ds + \mathcal{O}(1/t), \qquad (4.18)$$

where $\Omega = \max\{h(x') - h(x) : x, x' \in \mathcal{X}\}$. *In particular, if* $\lim_{t\to\infty} \eta(t) = 0$, *we have* $\widetilde{X}(t) \to \arg\min f$ *(a.s.).*

Corollary 4.13. *Suppose that* $\eta(t) \propto t^{-\beta}$ *for some* $\beta \in (0,1)$ *and all* $t \geq 1$. *Then:*

$$f(\widetilde{X}(t)) - \min f = \begin{cases} \mathcal{O}\left(t^{-\beta}\right) & \text{if } 0 < \beta < \frac{1}{2}, \\ \mathcal{O}\left(\sqrt{\log\log t/t}\right) & \text{if } \beta = \frac{1}{2}, \\ \mathcal{O}\left(t^{\beta-1}\right) & \text{if } \frac{1}{2} < \beta < 1. \end{cases} \qquad (4.19)$$

Corollary 4.14. *If* $\eta(t) = \sqrt{\Omega K/\sigma_*^2}\, \min\{1, 1/\sqrt{t}\}$, *we have*

$$\mathbb{E}[f(\widetilde{X}(t))] \leq \min f + 2\sqrt{\Omega\sigma_*^2/(Kt)}. \qquad (4.20)$$

Compared to (2.12), Corollary 4.14 indicates a drop in convergence speed 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 (4.18): balancing this second-order error against the noise-free bound $\Omega/(t\eta(t))$ imposes a $\Theta(1/\sqrt{t})$ schedule for $\eta(t)$ – otherwise, one term would be asymptotically slower than the other. In this regard, (4.20) is reminiscent of the well-known $\mathcal{O}(1/\sqrt{t})$ bounds derived in [40, Section 2.3] and [43, Section 6] for the dual averaging method (1.4) in stochastic environments. As discussed in [33], the drop in performance from $\mathcal{O}(1/t)$ to $\mathcal{O}(1/\sqrt{t})$ in the discrete-time case stems from the gap between continuous and discrete time: specifically, the discretization of the continuous-time dynamics introduces a second-order Taylor term which slows down convergence. In the case of (SMD), the second-order error that appears is not due to discretization, but to the (second-order) Itô correction which has a similar effect.

5. Discussion

In this last section, we discuss some applications and extensions of our analysis so far.

5.1. **The traffic assignment problem: a case study.** We begin with an application of our results to traffic assignment, a key problem in transportation and network science that concerns the optimal selection of paths between origins and destinations in traffic networks. Referring to [6, 10] for a detailed discussion, the core incarnation of the problem is as follows: First, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed multi-graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Assume further that there is an origin-destination (O/D) pair $(o, d) \in \mathcal{V} \times \mathcal{V}$ sending λ units of traffic from o to d via a set of paths $p \in \mathcal{P}$ (that is, a set of simple edge chains joining o to d in $\mathcal{G}$ in the usual way).[14] The set of feasible *routing flows* $x = (x_p)_{p \in \mathcal{P}}$ in the network is then defined as

$$\mathcal{X} = \lambda \, \Delta(\mathcal{P}) = \big\{ (x_p)_{p \in \mathcal{P}} : x_p \geq 0 \text{ and } \textstyle\sum_{p \in \mathcal{P}} x_p = \lambda \big\}. \tag{5.1}$$

Given a routing flow $x \in \mathcal{X}$, the *load* on edge $e \in \mathcal{E}$ is $w_e = \sum_{p \ni e} x_p$ and the *delay* experienced by an infinitesimal traffic element traversing edge e is $c_e(w_e)$, where $c_e \colon \mathbb{R}_+ \to \mathbb{R}_+$ is a nondecreasing convex *cost function* (often a polynomial with positive coefficients). Then, the delay along path $p \in \mathcal{P}$ is given by

$$c_p(x) \equiv \sum_{e \in p} c_e(w_e), \tag{5.2}$$

and the average delay in the network will be

$$C(x) = \sum_{p \in \mathcal{P}} x_p c_p(x) = \sum_{p \in \mathcal{P}} \sum_{e \in \mathcal{E}} x_p c_e(w_e) = \sum_{e \in \mathcal{E}} w_e c_e(w_e). \tag{5.3}$$

In this setting, solving the traffic assignment problem means finding a socially optimum routing flow $x^* \in \arg\min_{x \in \mathcal{X}} C(x)$. Assuming that the controlling O/D pair updates its routing flow at each $t \geq 0$, Theorem 2.4 shows that an optimum flow can be attained in an online manner by following the dynamics (MD). More precisely, if we introduce the *marginal cost*

$$\tilde{c}_e(w_e) = (w_e c_e(w_e))' = c_e(w_e) + w_e c_e'(w_e) \tag{5.4}$$

and its path-based analogue $\tilde{c}_p(x) = \sum_{e \in p} \tilde{c}_e(w_e)$, a simple differentiation yields

$$\frac{\partial C}{\partial x_p} = \sum_{e \in p} \tilde{c}_e(w_e) = \tilde{c}_p(x). \tag{5.5}$$

Thus, the dynamics (MD) take the form

$$\begin{aligned} \dot{y}_p &= -\tilde{c}_p(x), \\ x &= Q(\eta y), \end{aligned} \tag{5.6}$$

and, assuming c and h are sufficiently regular,[15] Theorem 2.4 shows that every solution $x(t)$ of (5.6) converges to an optimum routing flow $x^* \in \arg\min C$.

Now, if the marginal cost of each edge is only observable up to a random error, the scoring step of (5.6) takes the form

$$\begin{aligned} dY_p &= -\sum_{e \in p} [\tilde{c}_e(w_e) \, dt + \sigma_e \, dW_e] = -\tilde{c}_p(X) \, dt + \, dZ_p, \\ X &= Q(\eta Y), \end{aligned} \tag{5.7}$$

---

[14]The extension of the model to networks with multiple O/D pairs requires more elaborate notation, but is otherwise straightforward; for an atomic, nonsplittable variant, see [15].

[15]For instance, this is so if c_e is polynomial and h is the entropic regularizer of Example 2.2.
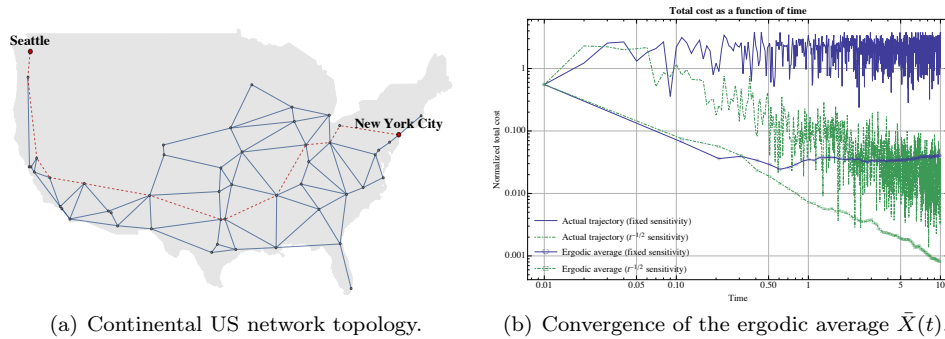
(a) Continental US network topology.

(b) Convergence of the ergodic average $\bar{X}(t)$.

Figure 2. Evolution of the dynamics (SMD) in the traffic assignment problem. Fig. 2(a) shows the underlying fiber network for the 50 largest continental US cities. In Fig. 2(b), we provide a log-log plot of the normalized total cost $C_0(x) = C(x) - \min C$ under (SMD) with logit choice (Example 2.2). When run with a fixed sensitivity, $X(t)$ meanders around without converging (solid blue line) and even the time-averaged process $\bar{X}(t) = t^{-1} \int_0^t X(s)\, ds$ fails to converge (blue line with circle markers). If run with a $t^{-1/2}$ sensitivity schedule, $X(t)$ gets closer to the optimum (dashed green line) and its time-average follows a power law (dashed green line with square markers).

where $dZ_p = -\sum_{e \in p} \sigma_e\, dW_e$ and σ_e is assumed constant (for simplicity). An easy calculation then shows that the infinitesimal covariance matrix Σ of Z is given by

$$\Sigma_{pp'} = \sum_{e,e' \in \mathcal{E}} \sigma_e \sigma_{e'} \delta_{ee'} \mathbb{1}(e \in p)\, \mathbb{1}(e' \in p') = \sum_{e \in p \cap p'} \sigma_e^2 \tag{5.8}$$

i.e. stochastic fluctuations across two different paths $p, p' \in \mathcal{P}$ are correlated over their common edges. This provides an important example where different components of the noise process Z are inherently correlated – here, due to the underlying graph $\mathcal{G}$.

Fig. 2 shows the evolution of the dynamics (5.6) in a data network consisting of the 50 largest continental US cities with noise volatility $\sigma_e = 0.25$ for all $e \in \mathcal{E}$ and affine cost functions of the form $c_e(w_e) = a_e w_e + b_e$ (both a_e and b_e drawn uniformly between 0 and 1). The stochastic system (SMD) was integrated numerically following a standard Euler–Maruyama discretization scheme [30] run for $N = 10^3$ iterations with a step-size of $\delta = 10^{-2}$. Then, in Fig. 2(b), we plotted the normalized total cost $C_0(x) = C(x) - \min C$ in log-log scale: in tune with Theorem 4.12, we see that if (SMD) is run with a decreasing sensitivity parameter, the ergodic average $\bar{X}(t) = t^{-1} \int_0^t X(s)\, ds$ enjoys a power law convergence rate (corresponding to a straight line in log-log scale), even though the unrectified process $X(t)$ fails to converge altogether.

5.2. **Links with Hessian Riemannian gradient flows.** In this last section, we briefly examine some links between (SMD) and the literature on Hessian Riemannian gradient flows [3, 4, 14]. To begin with, when h is steep and $\mathcal{X}$ has nonempty (topological) interior, the differential theory of Legendre transformations [48, 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

$$\dot{x} = \frac{d}{dt}Q(y) = \nabla Q(y) \cdot \dot{y} = \nabla(\nabla h^*(y)) \cdot v(Q(y)) = -\operatorname{Hess}(h^*(y)) \cdot \nabla f(x). \quad (5.9)$$

By Legendre's identity, we also have $\operatorname{Hess}(h^*(y)) = \operatorname{Hess}(h(Q(y)))^{-1}$ for all $y \in \mathcal{Y}$, so (5.9) leads to the *Hessian Riemannian* (HR) dynamics

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

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

As such, a natural question that arises is whether this equivalence between (HD) and (MD) carries over to the stochastic regime analyzed here. To address this issue, assume first that the gradient input to (HD) is perturbed by some random noise function $\epsilon(t)$ as in (3.1), viz.

$$\dot{x} = H(x)^{-1} \cdot (-\nabla f(x) + \epsilon(t)). \qquad\qquad (5.10)$$

Then, writing out (5.10) as a proper (Itô) stochastic differential equation, we get the stochastic Hessian Riemannian dynamics

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

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

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

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 $\frac{1}{2}\Sigma(X) \cdot \operatorname{Hess}(Q(Y)) \, dt$ (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 this, 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

$$dX = -X(1 - X)\left[dt - \sigma \, dW\right], \qquad\qquad (5.11)$$

while, after a routine application of Itô's lemma, (SMD) gives

$$dX = -X(1 - X)\left[dt - \sigma \, dW\right] + \frac{1}{2}X(1 - X)(1 - 2X)\sigma^2 \, dt. \qquad (5.12)$$

We thus see that the primal dynamics (5.11) and (5.12) differ by the Itô correction term $\frac{1}{2}X(1 - X)(1 - 2X) \, dt$. Accordingly, the dynamics' behavior with respect to the minimizer $x^* = 0$ of f is expected to be different as well.

Indeed, the score process $Y(t)$ of (SMD) becomes $Y(t) = Y(0) - t + \sigma W(t) \to -\infty$ (a.s.), implying in turn that $X(t) \to x^*$ under (5.12). On the other hand, under (5.11), it can be shown that $X(t)$ converges to $\arg\max f$ with high probability if σ is large enough. To see this, let $G(x) = \log x - \log(1 - x)$, so $G(X(t)) \to -\infty$ if $X(t) \to 0^+$ and $G(X(t)) \to +\infty$ if $X(t) \to 1^-$. Itô's lemma then yields

$$dG = G'(X) \, dX + \frac{1}{2}(dX)^2 = -dt + \sigma \, dW + (X - 1/2)\sigma^2 \, dt. \qquad (5.13)$$

From (5.13), it is intuitively obvious (and can be shown rigorously) that the drift of (5.13) remains uniformly positive with probability arbitrarily close to 1 if $X(0) > 1/2$ and σ is large.[16] In turn, this implies that $G(X(t)) \to \infty$, i.e. (5.11) converges with high probability to $\arg\max f$ instead of $\arg\min f$!

The above shows that the Hessian Riemannian system (HD) is 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. Given the link between Hessian Riemannian dynamics and the replicator dynamics of evolutionary game theory [3, 22], this is also reminiscent of the different long-run behavior of the replicator dynamics with aggregate shocks [23] and the dynamics of stochastically perturbed exponential learning [37]. We intend to explore these relations at 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. We begin with a structural property 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. ∎

The following technical comparison result is also useful in our analysis:

Lemma A.2. *If $y_2 - y_1 \in \mathrm{PC}(p)$, we have $F(p, y_1) \geq F(p, y_2)$ and*

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

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, (2.10c) readily yields

$$F(p, y_2) - F(p, y_1) \leq \langle y_2 - y_1 \,|\, Q(y) - p \rangle + \frac{1}{2K}\|y_2 - y_1\|_*^2$$

$$\leq \|\mathcal{X}\| \, \|y_2 - y_1\|_* + \frac{1}{2K}\|y_2 - y_1\|_*^2, \tag{A.3}$$

and, after rearranging, we get $\omega^2 + 2K\|\mathcal{X}\|\omega - 2K\delta \geq 0$, where $\omega = \|y_2 - y_1\|_* \geq 0$. 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.3) only holds if $\omega \geq \omega_+$, so (A.2) follows. ∎

Our next result describes the evolution of the η-deflated Fenchel coupling $V(t) = \eta^{-1}F(x^*, \eta y(t))$ under (MD):

---

[16]For a formal argument along these lines, see [39, Theorem 3.3.3].

Lemma A.3. *Fix some $x^* \in \mathcal{X}$. Then, under* (MD), *we have*

$$\dot{V}(t) = \langle v(x(t)) | x(t) - x^* \rangle. \tag{A.4}$$

Consequently, $V(t)$ is nonincreasing for all $x^* \in \arg\min f$.

Proof. By the definition (2.14) of the η-deflated Fenchel coupling and Proposition 2.2, we have

$$\dot{V}(t) = \eta^{-1} [\langle \eta \dot{y} | \nabla h^*(\eta y) \rangle] - \langle \dot{y} | x^* \rangle = \langle v(x) | x - x^* \rangle, \tag{A.5}$$

as claimed. As for our second claim, simply note that $\langle v(x) | x - x^* \rangle \le f(x^*) - f(x) \le 0$ for all $x^* \in \arg\min f$. ∎

We now extend Lemma A.3 to the stochastic dynamics (SMD) with a variable sensitivity parameter $\eta \equiv \eta(t)$:

Lemma A.4. *Fix some $x^* \in \mathcal{X}$. Then, for all $t \ge t_0 \ge 0$, we have*

$$V(t) - V(t_0) \le \int_{t_0}^t \langle v(X(s)) | X(s) - x^* \rangle \, ds \tag{A.6a}$$

$$- \int_{t_0}^t \frac{\dot{\eta}(s)}{\eta(s)^2} [h(x^*) - h(X(s))] \, ds \tag{A.6b}$$

$$+ \frac{1}{2K} \int_{t_0}^t \eta(s) \operatorname{tr}[\Sigma(X(s), s)] \, ds \tag{A.6c}$$

$$+ \sum_{i=1}^n \int_{t_0}^t (X_i(s) - x_i^*) \, dZ_i(s). \tag{A.6d}$$

Proof. By Proposition 2.2, we have $\nabla F(x^*, y) = \nabla h^*(y) - x^* = Q(y) - x^*$ for all $y \in \mathcal{Y}$. Thus, given that $Q = \nabla h^*$ is $(1/K)$-Lipschitz continuous (again by Proposition 2.2), our result follows from Proposition C.2 (see also Remark C.1). ∎

Appendix B. Convergence analysis

In this appendix, we prove the convergence results of Sections 2 and 4.

B.1. **Deterministic analysis.** We begin with the convergence properties of the deterministic dynamics (MD):

Proof of Theorem 2.4. For all $x^* \in \arg\min f$, Lemma A.3 gives

$$V(t) - V(0) = \int_0^t \langle v(x(s)) | x(s) - x^* \rangle \, ds \le t[\min f - \bar{f}(t)]. \tag{B.1}$$

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

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

$$\|x(t_k + \tau) - x(t_k)\| = \|Q(\eta y(t_k + \tau)) - Q(\eta y(t_k))\| \le \frac{\eta}{K} \|y(t_k + \tau) - y(t_k)\|_*$$

$$\leq \frac{\eta}{K} \int_{t_k}^{t_k+\tau} \|v(x(s))\|_* \, ds \leq \frac{\eta\tau}{K} \max_{x\in\mathcal{X}} \|v(x)\|_*. \tag{B.2}$$

Given that the bound (B.2) 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

$$V(t_k + \delta) - V(0) \leq \sum_{j=1}^{k} \int_{t_j}^{t_j+\delta} \langle v(x(s)) | x(s) - x^* \rangle \, ds \leq -ak\delta, \tag{B.3}$$

showing that $\liminf_{t\to\infty} F(x^*, \eta y(t)) = -\infty$, a contradiction. Since $x(t)$ admits at least one ω-limit, we conclude that $x(t)$ converges to $\arg\min f$.

Assuming $x^* \in \arg\min f$ is an ω-limit of $x(t)$, we have $x(t'_k) \to x^*$ for some sequence of times $t'_k \uparrow \infty$. By ($\mathbf{H_2}$), it follows that $V(t'_k) \to 0$ and hence, with $V(t)$ nonincreasing, that $V(t) \to 0$. Since $x(t)$ admits at least one ω-limit (by the compactness of $\mathcal{X}$), we conclude that $\lim_{t\to\infty} x(t) = x^*$, as claimed. $\blacksquare$

B.2. **The vanishing noise limit.** We proceed with the proof of our "vanishing noise" results, namely Proposition 4.1 and Theorem 4.2:

Proof of Proposition 4.1. Arguing by contradiction, assume that $X(t)$ remains a bounded distance away from $\arg\min f$ for large t with positive probability. This implies that there exists some $a > 0$ and a (random) t_0 such that

$$\langle v(X(t)) | X(t) - x^* \rangle \leq -a \quad \text{for all } t \geq t_0, \tag{B.4}$$

again with positive probability. Then, fixing some $x^* \in \arg\min f$ and taking the associated Fenchel coupling $V(t) = \eta^{-1} F(x^*, \eta Y(t))$, Lemma A.4 gives

$$V(t) - V(t_0) \leq \int_{t_0}^{t} \langle v(X(s)) | X(s) - x^* \rangle \, ds$$

$$+ \frac{1}{2K} \int_{t_0}^{t} \operatorname{tr}[\Sigma(X(s), s)] \, ds + \sum_{i=1}^{n} \int_{t_0}^{t} (X_i(s) - x^*) \, dZ_i(s)$$

$$\leq -a(t - t_0) + \frac{1}{2K} \int_{t_0}^{t} \operatorname{tr}[\Sigma(X(s), s)] \, ds + \xi(t), \tag{B.5}$$

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.1 in Appendix C shows that $\xi(t)/t \to 0$ (a.s.). Moreover, we also have $\lim_{t\to\infty} t^{-1} \int_0^t \operatorname{tr}[\Sigma(X(s), s)] \, ds = \lim_{t\to\infty} \|\sigma(X(t), t)\|_F^2 = 0$ (by the vanishing noise assumption and de l'Hôpital's rule), so the last two terms in (B.5) are both sublinear in t. We thus obtain $V(t) \to -\infty$ with positive probability, a contradiction which establishes our claim. $\blacksquare$

We are now in a position to prove Theorem 4.2 under the additional assumption $\sup_{x\in\mathcal{X}} \|\sigma(x, t)\|_F = o(1/\sqrt{\log t})$:

Proof of Theorem 4.2. 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 $U_\varepsilon = \{x = Q(y) : F(x^*, y) < \varepsilon\}$. Our first claim is that there exists a time $T \equiv T(\varepsilon)$ such that $F(x^*, \Phi_T(y)) \leq \max\{\varepsilon, F(x^*, y) - \varepsilon\}$ for all $y \in \mathcal{Y}$. Indeed, by $(\mathbf{H_2})$ and the continuity of $v(x)$, there exists some $a \equiv a(\varepsilon) > 0$ such that

$$\langle v(x) \,|\, x - x^* \rangle \leq -a \quad \text{for all } x \notin U_\varepsilon. \tag{B.6}$$

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 A.3 gives

$$F(x^*, \Phi_t(y)) - F(x^*, y) = \int_0^t \langle v(x(s)) \,|\, x(s) - x^* \rangle \, ds \leq -at \quad \text{for all } t \leq \tau_y. \tag{B.7}$$

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

(1) If $T \leq \tau_y$, (B.7) 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 here 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 4.3. Indeed, by Proposition 4.6 in [8], it suffices to show that $\int_0^\infty e^{-c/\Sigma_{\max}(t)} \, dt < \infty$ where $\Sigma_{\max}(t) = \sup_{x \in \mathcal{X}} \text{tr}[\Sigma(x, t)]$ and $c > 0$ is arbitrary. However, by assumption

$$\Sigma_{\max}(t) = \sup_{x \in \mathcal{X}} \|\sigma(x, t)\|_F^2 = \phi(t)/\log t \tag{B.8}$$

for some $\phi(t)$ with $\lim_{t \to \infty} \phi(t) = 0$. Therefore,

$$e^{-c/\Sigma_{\max}(t)} = \left(e^{\log t}\right)^{-c/\phi(t)} = t^{-c/\phi(t)} = \mathcal{O}(t^{-\beta}) \quad \text{for all } \beta > 1, \tag{B.9}$$

and our assertion follows.

To proceed, fix a solution $Y(t)$ of (SMD) which is an APT of (MD). Moreover, with notation as in Definition 4.3, 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

$$F(x^*, Y(t+h)) \leq F(x^*, \Phi_h(Y(t))) + \langle Y(t+h) - \Phi_h(Y(t)) \,|\, Q(\Phi_h(Y(t))) - x^* \rangle$$

$$+ \frac{1}{2K}\|Y(t+h) - \Phi_h(Y(t))\|_*^2$$

$$\leq F(x^*, \Phi_h(Y(t))) + \delta\|\mathcal{X}\| + \frac{\delta^2}{2K} \leq F(x^*, \Phi_h(Y(t))) + \varepsilon, \tag{B.10}$$

where, in the first line, we used the second-order Taylor estimate for the Fenchel coupling derived in Proposition 2.3 (cf. Appendix A).

By Proposition 4.1, there exists some $T_0 \geq t_0$ such that $F(x^*, Y(T_0)) \leq 2\varepsilon$ (a.s.), implying that $F(x^*, Y(T_0)) \leq 2\varepsilon$ for some $T_0 \geq t_0$. Hence, by (B.10), we get

$$F(x^*, Y(T_0 + h)) \leq F(x^*, \Phi_h(Y(T_0))) + \varepsilon \leq F(x^*, Y(T_0)) + \varepsilon \leq 3\varepsilon \tag{B.11}$$

for all $h \in [0, T]$. However, 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 $X(t) \to x^*$, as claimed. ■

B.3. **Long-run concentration around solution points.** We now turn to the ergodic properties of (SMD) under persistent, nonvanishing noise:

Proof of Proposition 4.4. Let $V(t) \equiv \eta^{-1} F(x^*, \eta Y(t))$ denote the η-deflated Fenchel coupling between x^* and $Y(t)$. Then, by the growth bound (A.6), we get

$$V(t) - V(0) \leq \int_0^t \langle v(X(s)) \mid X(s) - x^* \rangle \, ds + \frac{1}{2K} \int_0^t \eta \, \mathrm{tr}[\Sigma(X(s), s)] \, ds + \xi(t)$$

$$\leq -\frac{\alpha}{2} \int_0^t \|X(s) - x^*\|^2 \, ds + \frac{\eta \sigma_*^2 t}{2K} + \xi(t), \qquad (B.12)$$

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 (4.4) to write $\langle v(x) \mid x - x^* \rangle \leq f(x^*) - f(x) \leq -\frac{1}{2}\alpha\|x - x^*\|^2$ in the second line. Since $V(t) \geq 0$, the bound (4.5) follows by taking expectations, exploiting the fact that $\xi(t)$ has zero mean, and rearranging.

Now, replacing t by $\tau_\delta \wedge t$ in (B.12), we also get

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

$$\leq V(0) + \frac{\eta \sigma_*^2 - \alpha K \delta^2}{2K} \mathbb{E}[\tau_\delta \wedge t], \qquad (B.13)$$

where we used the fact that $\|X(s) - x^*\| \geq \delta$ for all $s \leq \tau_\delta$. Since $V \geq 0$, we conclude that $\mathbb{E}[\tau_\delta \wedge t] \leq 2KV(0)/(\alpha K \delta^2 - \eta \sigma_*^2)$. Our claim then follows by letting $t \to \infty$ (so $\tau_\delta \wedge t \to \tau_\delta$) and invoking the dominated convergence theorem. Finally, the optimized bound (4.7) is obtained by maximizing the denominator of (4.6). ∎

We are now in a position to estimate the occupation measure of $X(t)$:

Proof of Theorem 4.6. We begin by introducing a transformed version of $Y(t)$ which is recurrent under (SMD).[17] 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 $\mathcal{V}_0 = \mathrm{aff}(\mathcal{X} - \mathcal{X}) \subseteq \mathcal{V}$ 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

$$\langle \pi_0(y) \mid z \rangle = \langle y \mid z \rangle \quad \text{for all } z \in \mathcal{V}_0. \qquad (B.14)$$

We then have $\pi_0(y) = 0$ whenever y annihilates $\mathcal{V}_0$ (i.e. $\langle y \mid z \rangle = 0$ for all $z \in \mathcal{V}_0$).[18]

Accordingly, in view of Proposition 4.4, 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 $C_0 \subseteq \mathcal{Y}_0$ 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.

---

[17]Recall here that $Y(t)$ is *recurrent* if there exists a compact set C such that $\mathbb{P}(Y(t) \in C$ for some $t \geq 0) = 1$ for every initial condition y_0 of Y [11, 28]. In our case, the set $Q^{-1}(\mathbb{B}_\delta)$ need not be compact, so the generating process $Y(t)$ need not be recurrent either.

[18]Of course, if $\mathcal{X}$ has nonempty interior as a subset of $\mathcal{V}$, we have $\mathcal{V}_0 = \mathcal{V}$ and π_0 is the identity.

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

$$d\Psi_i = \sum_{k=1}^n \Pi_{ik} \left(v_k(X) \, dt + \, dZ_k \right). \tag{B.15}$$

Moreover, define the "restricted" mirror map $Q_0 \colon \mathcal{Y}_0 \to \mathcal{X}$ as

$$Q_0(w) = \arg\max_{x \in \mathcal{X}} \{ \langle w \,|\, x \rangle - h(x) \}, \tag{B.16}$$

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 (B.15) represents a regular Itô diffusion.

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

$$d[\Psi_i, \Psi_j] = d(\Pi Y)_i \, d(\Pi Y)_j = \sum_{k,\ell=1}^n \Pi_{ik} \Pi_{j\ell} \Sigma_{k\ell} \, dt = \left(\Pi \Sigma \Pi^\top \right)_{ij} dt, \tag{B.17}$$

where we used the definition (3.3) of Σ in the penultimate equality. However, we also have $\Pi \Sigma \Pi^\top \succcurlyeq \lambda \Pi \Pi^\top \succcurlyeq \lambda \pi_{\min}^2 I$, where $\pi_{\min} > 0$ denotes the smallest singular value of $\Pi^\top$ (recall that π_0 has full rank). 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 $\mathbb{B}_\delta \subseteq \mathcal{X}^\circ$ (obviously, this is possibly only if $x^* \in \mathcal{X}^\circ$). Momentarily viewing $\mathcal{X}$ as a convex body of $\mathcal{V}_0$ (and $\mathbb{B}_\delta$ as a ball in $\mathcal{V}_0$), Remark 6.2.3 in [21] implies that the set $C_0 = \eta^{-1} \partial h(\mathbb{B}_\delta)$ is compact.[19] Then, by Proposition 4.4, 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 [28, 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

$$\lim_{t \to \infty} t^{-1} \int_0^t \phi(\Psi(s)) \, ds = \int_{\mathcal{Y}_0} \phi \, d\nu, \tag{B.18}$$

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

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbb{1}(X(s) \in \mathbb{B}_\delta) \, ds = \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbb{1}(\eta \Psi(s) \in Q_0^{-1}(\mathbb{B}_\delta)) \, ds$$

$$= \int_{\mathcal{Y}_0} \mathbb{1}_{\eta^{-1} Q_0^{-1}(\mathbb{B}_\delta)} \, d\nu = \nu(\eta^{-1} Q_0^{-1}(\mathbb{B}_\delta)), \tag{B.19}$$

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 (4.5) also yields

$$1 - \nu(\eta^{-1} Q_0^{-1}(\mathbb{B}_\delta)) = \lim_{t \to \infty} \frac{1}{t} \mathbb{E}\left[\int_0^t \mathbb{1}(X(s) \notin \mathbb{B}_\delta) \, ds \right]$$

---

[19]Strictly speaking, Remark 6.2.3 of [21] 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.

$$\leq \lim_{t\to\infty} \frac{1}{t} \mathbb{E}\left[\int_0^t \frac{\|X(s) - x^*\|^2}{\delta^2}\, ds\right]$$

$$\leq \lim_{t\to\infty} \frac{1}{\delta^2}\left[\frac{2F(x^*, \eta y_0)}{\eta\alpha t} + \frac{\eta\sigma_*^2}{\alpha K}\right] = \frac{\eta\sigma_*^2}{\alpha K \delta^2},$$

as was to be shown. ∎

B.4. **Convergence to sharp solutions.** The proof of our convergence result for sharp solutions is fairly involved, so we encode it in a series of technical lemmas. The first one shows that neighborhoods of sharp solutions are recurrent under (SMD):

Lemma B.1. *Fix $\delta > 0$ and assume that f admits a γ-sharp solution. If (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 $V(t) = \eta^{-1}F(x^*, \eta Y(t))$ for the η-deflated Fenchel coupling between x^* and $Y(t)$, Lemma A.4 yields

$$V(t) \leq V(t_0) + \int_{t_0}^t \langle v(X(s))\,|\,X(s) - x^*\rangle\, ds + \frac{1}{2K}\int_{t_0}^t \eta\,\mathrm{tr}[\Sigma(X(s), s)]\, ds + \xi(t)$$

$$\leq V(t_0) - \left[\gamma\delta - \frac{\eta\sigma_*^2}{2K} - \frac{\xi(t)}{t - t_0}\right](t - t_0), \tag{B.20}$$

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 4.9 in the second. Since $\xi(t)/(t - t_0) \to 0$ by Lemma C.1 in Appendix C, the bound (B.20) yields $\lim_{t\to\infty} V(t) = -\infty$ if $\eta\sigma_*^2 < 2\gamma\delta K$, a contradiction (recall that $V(t) \geq 0$ 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 shows that the dual 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^*.

Lemma B.2. *Suppose that f admits a sharp minimum $x^* \in \mathcal{X}$, and let P be a polyhedral cone such that $v(x^*) \in \mathrm{int}(P)$ and $P \subseteq \mathrm{int}(\mathrm{PC}(x^*)) \cup \{0\}$. 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*

$$\mathbb{P}(Y(t) \in y + P \text{ for all } t \geq 0) \geq 1 - e^{-\kappa\delta/(\eta\sigma_*^2)}, \tag{B.21}$$

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

Proof. Let $P^\perp = \{z \in \mathcal{V} : \langle y\,|\,z\rangle \leq 0 \text{ for all } y \in P\}$ denote the polar cone of P and let $\mathcal{U} = \{u_j\}_{j=1}^d$ be a basis for $P^\perp$ (recall that P is assumed polyhedral). Further, fix a small compact neighborhood L of x^* such that $\langle v(x)\,|\,z\rangle \leq -\gamma_L\|z\|$ for some $\gamma_L > 0$ and all $x \in L$, $z \in P^\perp$;[20] with a fair bit of hindsight, assume also that $\delta < K\|\mathcal{X}\|^2$ is sufficiently small so that $Q(\eta y) \in L$ whenever $F(x^*, \eta y) \leq \varepsilon + \delta$. Finally, invoking Lemma A.2, let $y = y_0 - cv(x^*)$ for some $c > 0$ such that $F(x^*, \eta y) = \varepsilon + \delta$. Then, (A.2) gives

$$\|y_0 - y\|_* = c\|v(x^*)\|_* \geq \frac{K\|\mathcal{X}\|}{\eta}\left[\sqrt{1 + 2\delta/(K\|\mathcal{X}\|^2)} - 1\right] \geq \frac{\delta}{2\eta\|\mathcal{X}\|}, \tag{B.22}$$

---

[20]That such a γ_L exists is a consequence of the continuity of $v(x)$ and Lemma 4.9.
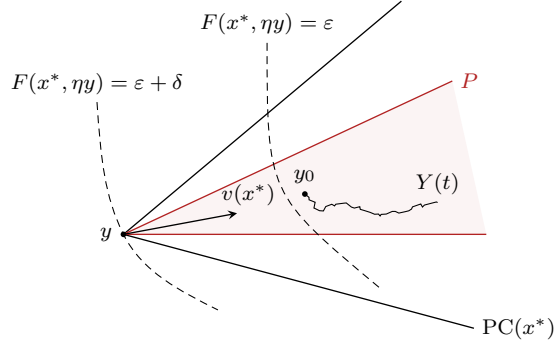
Figure 3. The various sets in the proof of Lemma B.2.

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

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

$$G_u(t) = G_u(0) + \int_0^t \langle v(X(s)) \,|\, u \rangle \, ds + \xi_u(t) \leq -A\|u\| - B\|u\|t + \xi_u(t), \quad \text{(B.23)}$$

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.1, 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 $\rho_u(t) = [\xi_u(t), \xi_u(t)]$ denotes the quadratic variation of ξ_u. By (B.23), this further implies that $G_u(t) \leq 0$ whenever $W_u(\rho_u(t)) \leq A\|u\| + B\|u\|t$; hence, $\tau_P = \infty$ whenever $W_u(\rho_u(t)) \leq A\|u\| + B\|u\|t$.

Moreover, note that

$$d\rho_u = d\xi_u \cdot d\xi_u = \sum_{i,j=1}^n \Sigma_{ij} u_i u_j \, dt, \quad \text{(B.24)}$$

so $\rho_u(t) \leq \sigma_*^2 \|u\|_2^2 t \leq R\sigma_*^2 \|u\|^2 t$ for some constant $R > 0$ that depends only on the choice of primal norm $\|\cdot\|$. Hence, if a trajectory of W_u is such that $W_u(t) \leq A\|u\| + \frac{B}{R\|u\|\sigma_*^2} t$ for all $t \geq 0$, we also get

$$W_u(\rho_u(t)) \leq A\|u\| + \frac{B}{R\|u\|\sigma_*^2}\rho(t) \leq A\|u\| + B\|u\|t \quad \text{for all } t \geq 0. \quad \text{(B.25)}$$

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

To do so, let

$$\tau_P' = \inf\left\{t > 0 : W_u(t) = A\|u\| + \frac{B}{R\|u\|\sigma_*^2} t \text{ for some } u \in \mathcal{U}\right\} \quad \text{(B.26)}$$

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

$$\mathbb{P}(\tau_P' < \infty) = \mathbb{P}\Big(\bigcup_{u \in \mathcal{U}} E_u\Big) \leq \sum_{u \in \mathcal{U}} \mathbb{P}(E_u) = |\mathcal{U}| e^{-2AB/(R\sigma_*^2)}. \quad \text{(B.27)}$$

Now, from the bound (B.22) and the definition of A and B, we have

$$\frac{AB}{R} = \frac{c\gamma_L \min_{u' \in \mathcal{U}} |\langle v(x^*) | u' \rangle|}{R} \geq \frac{\delta}{2\eta \|\mathcal{X}\|} \frac{\gamma_L \min_{u' \in \mathcal{U}} |\langle v(x^*) | u' \rangle|}{R \|v(x^*)\|_*} = \frac{\kappa\delta}{\eta}, \quad \text{(B.28)}$$

where we set $\kappa = \frac{\gamma_L \min_{u' \in \mathcal{U}} |\langle v(x^*) | u' \rangle|}{2R \|v(x^*)\|_* \|\mathcal{X}\|}$. Backtracking then yields $\mathbb{P}(\tau_P = \infty) \geq \mathbb{P}(\tau_P' = \infty) \geq 1 - e^{-\kappa\delta/(\eta\sigma_*^2)}$, provided that $\eta \leq \kappa\delta/(\sigma_*^2 \log|\mathcal{U}|)$. Therefore, with $\mathbb{P}(Y(t) \in y + P \text{ for all } t \geq 0) = \mathbb{P}(\tau_P = \infty)$, 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 B.3. *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 get

$$h(x^*) \geq h(x_n) + \langle y_n | x^* - x_n \rangle \geq h(x_n) - \langle y_n | z_n \rangle \|x_n - x^*\|, \quad \text{(B.29)}$$

where we set $z_n = (x_n - x^*)/\|x_n - x^*\|$. Since z_n lives in the unit sphere of $\|\cdot\|$, compactness (and a descent 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 $\liminf$ on both sides, we obtain $\liminf h(x^*) = \infty$, a contradiction. ∎

We are now in a position to prove our main result for sharp solutions:

Proof of Theorem 4.10. As in the proof of Lemma B.2, 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 $P \subseteq \mathrm{int}(\mathrm{PC}(x^*))$ such that $\langle v(x) | z \rangle \leq -\gamma_L \|z\|$ for all $x \in L$, $z \in P^{\perp}$.

With this in mind, pick $\varepsilon, \delta > 0$ sufficiently small so that the conclusion of Lemma B.2 holds and $Q(\eta y) \in L$ whenever $F(x^*, \eta y) \leq \varepsilon + \delta$. If η is also chosen small enough, combining ($\mathbf{H}_2$) with Lemma B.1 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 B.2 and the strong Markov property of $Y(t)$, there exists some $a > 0$ such that $\mathbb{P}(F(x^*, \eta Y(t_n + t)) \leq \varepsilon + \delta \text{ for all } t \geq 0) \geq 1 - (1-a)^n$ for all n. Thus, with notation as in (B.23), we get

$$G_z(t_n + t) \leq -A\|z\| - B\|z\|t + \xi_z(t) \quad \text{for all } t \geq 0, \quad \text{(B.30)}$$

with probability at least $1 - (1-a)^n$. In turn, Lemma C.1 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 $z \in \mathrm{TC}(x^*)$, so $X(t) \to x^*$ (a.s.) by Lemma B.3.

We are left to show that this convergence occurs in finite time if Q is surjective. 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

$$\langle Y(t) - Y(t_0) | z \rangle = \int_{t_0}^{t} \langle v(X(s)) | z \rangle \, ds + \langle Z(t) | z \rangle \leq -\gamma_L(t - t_0) + \|Z(t)\|_*. \quad \text{(B.31)}$$

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

B.5. **Convergence via rectification.** We now turn to the rectified variants of (SMD) with a decreasing sensitivity parameter:

Proof of Theorem 4.12. For all $x \in \mathcal{X}$ and $x^* \in \mathcal{X}$, we have

$$f(x) - \min f = f(x) - f(x^*) \le \langle \nabla f(x) \,|\, x - x^* \rangle = \langle v(x) \,|\, x^* - x \rangle, \qquad \text{(B.32)}$$

by convexity of f. Hence, by the definition of $\widetilde{X}$ (and Jensen's inequality in the case of (4.16a)), we obtain

$$f(\widetilde{X}(t)) \le \min f + \frac{1}{t} \int_0^t \langle v(X(s)) \,|\, x^* - X(s) \rangle \, ds, \qquad \text{(B.33)}$$

so it suffices to properly majorize the right-hand side of the above equation.

To that end, let $V(t) = \eta(t)^{-1} F(x^*, \eta(t) Y(t))$ denote the η-deflated Fenchel coupling between $Y(t)$ and $x^* \in \arg\min f$. Then, Lemma A.4 yields

$$\int_0^t \langle v(X(s)) \,|\, x^* - X(s) \rangle \, ds \le V(0) - V(t) \qquad \text{(B.34a)}$$

$$- \int_0^t \frac{\dot{\eta}(s)}{\eta^2(s)} \left[h(x^*) - h(X(s)) \right] ds \qquad \text{(B.34b)}$$

$$+ \frac{1}{2K} \int_0^t \eta(s) \, \mathrm{tr}[\Sigma(X(s), s)] \, ds \qquad \text{(B.34c)}$$

$$+ \sum_{i=1}^n \int_0^t (X_i(s) - x_i^*) \, dZ_i(s). \qquad \text{(B.34d)}$$

We now proceed to bound each term of (B.34):

a) Since $V(t) \ge 0$ for all t, the term (B.34a) is bounded from above by $V(0)$, viz.

$$\text{(B.34a)} \le V(0) = \frac{h(x^*) + h^*(\eta(0) Y(0))}{\eta(0)} - \langle Y(0) \,|\, x^* \rangle \qquad \text{(B.35)}$$

b) For (B.34b), we have $h(x^*) - h(X(s)) \le \Omega$ by definition, so, with $\dot{\eta}(t) \le 0$ for almost all t by ($\mathbf{H_4}$), we get

$$\text{(B.34b)} \le -\Omega \int_0^t \frac{\dot{\eta}(s)}{\eta^2(s)} \, ds = \frac{\Omega}{\eta(t)} - \frac{\Omega}{\eta(0)}. \qquad \text{(B.36)}$$

c) For (B.34c), the definition of σ_*^2 gives (B.34c) $\le (2K)^{-1} \sigma_*^2 \int_0^t \eta(s) \, ds$.

d) Finally, for (B.34d), let $\xi(t) = \int_0^t \sum_{i=1}^n (X_i(s) - x_i^*) \, dZ_i(s)$ and write $\rho(t) = [\xi(t), \xi(t)]$ for the quadratic variation of ξ. We then get

$$d[\xi, \xi] = d\xi \cdot d\xi = \sum_{i,j=1}^n \Sigma_{ij} (X_i - x_i^*)(X_j - x_j^*) \, dt \le \sigma_*^2 \|X - x^*\|_2^2 \, dt, \qquad \text{(B.37)}$$

so $\rho(t) \le R \sigma_*^2 \|\mathcal{X}\|^2 t$ for some norm-dependent constant $R > 0$. Arguing as in the proof of Lemma C.1 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, p. 112], we then obtain

$$\limsup_{t\to\infty} \frac{\widetilde{W}(\rho(t))}{\sqrt{2Mt\log\log(Mt)}} \leq \limsup_{t\to\infty} \frac{\widetilde{W}(\rho(t))}{\sqrt{2\rho(t)\log\log\rho(t)}} = 1 \quad \text{(a.s.)}, \qquad \text{(B.38)}$$

where $M = R\sigma_*^2\|\mathcal{X}\|^2$. 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

$$\frac{1}{t}\int_0^t \langle v(X(s))\,|\,x^* - X(s)\rangle ds \leq \frac{\Omega}{t\eta(t)} + \frac{\sigma_*^2}{2Kt}\int_0^t \eta(s)ds + \mathcal{O}(t^{-1/2}\sqrt{\log\log t}), \quad \text{(B.39)}$$

where we have absorbed the $\mathcal{O}(1/t)$ terms from (B.35) and (B.36) in the logarithmic term $\mathcal{O}(\sqrt{t^{-1}\log\log t})$. Our claim then follows from (B.33). Finally, recalling that $\xi(t)$ is a zero-mean local martingale, the mean bound (4.18) follows by taking expectations above. ∎

APPENDIX C. RESULTS FROM STOCHASTIC ANALYSIS

In this last appendix, we collect some results from stochastic analysis that we use throughout the paper. The first such result is a growth estimate for Itô martingales with bounded volatility:

Lemma C.1. *Let $W(t)$ be a Wiener process in $\mathbb{R}^m$ and let $\zeta(t)$ be a bounded, continuous process in $\mathbb{R}^m$. Then, for every function $f\colon [0,\infty) \to (0,\infty)$, we have*

$$f(t) + \int_0^t \zeta(s) \cdot dW(s) \sim f(t) \quad \text{as } t \to \infty \text{ (a.s.)}, \qquad \text{(C.1)}$$

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

Proof of Lemma C.1. Let $\xi(t) = \sum_{i=1}^n \int_0^t \zeta_i(s)\,dW_i(s)$. Letting $\rho(t) = [\xi(t), \xi(t)]$ denote the quadratic variation of $\xi(t)$, we have

$$d\rho = \sum_{i=1}^n \zeta_i\zeta_j\delta_{ij}\,dt \leq M\,dt, \qquad \text{(C.2)}$$

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

$$\tau_\rho(s) = \begin{cases} \inf\{t \geq 0 : \rho(t) > s\} & \text{if } s \leq \rho_\infty, \\ \infty & \text{otherwise.} \end{cases} \qquad \text{(C.3)}$$

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 $t \geq 0$. The rest of the proof then follows by applying the law of the iterated logarithm as in [15, Lemma B.4]. ∎

The second result we report here is a weak version of Itô's formula for differentiable functions with Lipschitz-continuous gradient. For notational convenience, let $\mathbf{C}_L^{1,1}(\mathcal{Y})$ denote the space of functions $\phi\colon \mathcal{Y} \to \mathbb{R}$ such that

$$\|\nabla\phi(y_2) - \nabla\phi(y_1)\| \leq L\|y_2 - y_1\|_* \quad \text{for all } y_1, y_2 \in \mathcal{Y}. \qquad \text{(C.4)}$$

We then have:

Proposition C.2. *Let $Y(t) = (Y_i(t))_{i=1}^n$ be a $\mathcal{Y}$-valued Itô process of the form*

$$Y_i(t) = Y_i(0) + \int_0^t \alpha_i(s)\,ds + \sum_{k=1}^m \int_0^t \beta_{ik}(s)\,dW_k(s), \tag{C.5}$$

where $W(t) = (W_k(t))_{k=1}^m$ is a standard m-dimensional Wiener process. If $\phi \in \mathbf{C}_L^{1,1}(\mathcal{Y})$ is convex, then, for all $t \geq 0$, we have:

$$\phi(Y(t)) \leq \phi(Y_0) + \int_0^t \langle \nabla \phi(Y(s)) \,|\, dY(s) \rangle + \frac{L}{2} \int_0^t \mathrm{tr}[\beta(s)\beta(s)^\top]\,ds. \tag{C.6}$$

The proof of Proposition C.2 is based on the following property of convex functions in $\mathbf{C}_L^{1,1}(\mathbb{R}^n)$:

Lemma C.3. *Let $\phi \in \mathbf{C}_L^{1,1}(\mathcal{Y})$ be convex. Then ϕ is almost everywhere twice differentiable and its Hessian satisfies*

$$0 \preccurlyeq \mathrm{Hess}(\phi(y)) \preccurlyeq LI \quad \text{for (Lebesgue) almost all } y \in \mathcal{Y}. \tag{C.7}$$

Proof. The fact that ϕ is twice differentiable (Lebesgue) a.e. is Alexandrov's theorem. Hence, there exists a Lebesgue-full set $\mathcal{Y}_0 \subseteq \mathcal{Y}$ such that

$$\phi(y+z) = \phi(y) + \langle \nabla\phi(y) \,|\, z \rangle + \frac{1}{2} z^\top \mathrm{Hess}(\phi(y))z + \theta(y,z) \quad \text{for all } y \in \mathcal{Y}_0, \tag{C.8}$$

with $\theta(y,z) = o(\|z^2\|_*)$. Furthermore, by the well-known descent lemma for functions with Lipschitz continuous gradient [42, Theorem 2.1.5], we also have

$$\phi(y+z) \leq \phi(y) + \langle \nabla\phi(y) \,|\, z \rangle + \frac{L}{2}\|z\|_*^2 \quad \text{for all } y, z \in \mathcal{Y}. \tag{C.9}$$

Thus, taking $z = tu$ for some unit vector $u \in \mathcal{Y}$ (i.e. $\|u\|_* = 1$) and combining the above, we readily obtain

$$\frac{t^2}{2} u^\top \mathrm{Hess}(\phi(y))u + \theta(y,tu) \leq \frac{L}{2}t^2 \quad \text{for all } y \in \mathcal{Y}_0, \, t \geq 0. \tag{C.10}$$

Hence, dividing by t and letting $t \to 0^+$ yields

$$u^\top \mathrm{Hess}(\phi(y))u \leq \frac{L}{2} \quad \text{for all } y \in \mathcal{Y}_0, \tag{C.11}$$

implying in turn that $\mathrm{Hess}(\phi(y)) \preccurlyeq LI$ for all $y \in \mathcal{Y}_0$. The bound $\mathrm{Hess}(\phi(y)) \succcurlyeq 0$ is a trivial consequence of convexity, completing our proof. ∎

Proof of Proposition C.2. Our proof relies on smoothing by mollification. To begin, consider the standard unit mollifier

$$\rho(u) = \begin{cases} c\exp\left(-\frac{1}{1-\|u\|_*^2}\right) & \text{if } \|u\|_* < 1, \\ 0 & \text{if } \|u\|_* \geq 1, \end{cases} \tag{C.12}$$

with $c > 0$ chosen so that $\int_{\mathbb{R}^n} \rho(w)\,dw = 1$. Then, for all $\varepsilon > 0$, let

$$\rho_\varepsilon(u) = \varepsilon^{-n}\rho(u/\varepsilon), \tag{C.13a}$$

and

$$\phi_\varepsilon(y) = (\phi * \rho_\varepsilon)(y) = \int_\mathcal{Y} \phi(y-w)\rho_\varepsilon(w)\,dw, \tag{C.13b}$$

with "$*$" above denoting convolution over $\mathbb{R}^n$. We then have $\phi_\varepsilon \in \mathbf{C}^\infty(\mathcal{Y})$, so the standard form of Itô's formula gives us

$$\phi_\varepsilon(Y(t)) = \phi_\varepsilon(Y(s)) + \int_s^t \langle \nabla \phi_\varepsilon(Y(\tau)) \,|\, dY(\tau) \rangle$$

$$+ \frac{1}{2} \int_s^t \mathrm{tr}\big[\mathrm{Hess}(\phi_\varepsilon(Y(\tau)))\beta(\tau)\beta(\tau)^\top\big]\, d\tau$$

$$= \phi_\varepsilon(Y(s)) + \int_s^t \left\langle \int_\mathcal{Y} \nabla\phi(z)\rho_\varepsilon(Y(\tau) - z)\, dz \,\Big|\, dY(\tau) \right\rangle$$

$$+ \frac{1}{2} \int_s^t \int_\mathcal{Y} \mathrm{tr}\big[\mathrm{Hess}(\phi(z))\beta(\tau)\beta(\tau)^\top\big]\rho_\varepsilon(Y(\tau) - z)\, d\tau\, dz, \qquad \text{(C.14)}$$

where the last equality uses the fact that $\mathrm{Hess}(\phi)$ exists for (Lebesgue) almost all y, as established in Lemma C.3. Using Lemma C.3 one more time, we further have $\mathrm{tr}[\mathrm{Hess}(\phi(z))\beta(\tau)\beta(\tau)^\top] \le L\,\mathrm{tr}[\beta(\tau)\beta(\tau)^\top]$, implying in turn that

$$\phi_\varepsilon(Y(t)) - \phi_\varepsilon(Y(s)) \le \int_s^t \left\langle \int_\mathcal{Y} \nabla\phi(z)\rho_\varepsilon(Y(\tau) - z)\, dz \,\Big|\, dY(\tau) \right\rangle$$

$$+ \frac{L}{2} \int_s^t \mathrm{tr}[\beta(\tau)\beta(\tau)^\top]\, d\tau. \qquad \text{(C.15)}$$

Our assertion then follows by letting $\varepsilon \to 0^+$ and invoking the dominated convergence theorem. ∎

Remark C.1. In the main body of the paper, the above result is typically applied to the Fenchel coupling $F(p, y)$ which, as a function of y, is in the class $\mathbf{C}_{1/K}^{1,1}(\mathcal{Y})$ for every $p \in \mathcal{X}$, by Proposition 2.2. Specifically, letting $Y(t)$ denote the unique strong solution to (SMD) and taking $V(t) = F(p, Y(t))$ for some $p \in \mathcal{X}$, Proposition C.2 yields

$$V(t) - V(0) \le \int_0^t \langle \nabla F(p, Y(s)) \,|\, dY(s) \rangle + \frac{1}{2K} \int_0^t \mathrm{tr}[\Sigma(X(s), s)]\, ds, \qquad \text{(C.16)}$$

where we used the definition $\Sigma = \sigma\sigma^\top$ of Σ.

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] F. ALVAREZ, H. ATTOUCH, J. BOLTE, AND P. REDONT, *A second-order gradient-like dissipative dynamical system with Hessian damping. Applications to optimization and mechanics*, Journal des Mathématiques Pures et Appliquées, 81 (2002), pp. 774–779.

[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] H. ATTOUCH, X. GOUDOU, AND P. REDONT, *The heavy ball with friction method, I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system*, Communications in Contemporary Mathematics, 2 (2000), pp. 1–34.

[6] M. BECKMANN, C. B. MCGUIRE, AND C. WINSTEN, *Studies in the Economics of Transportation*, Yale University Press, 1956.

[7] P. Bégout, J. Bolte, and M.-A. Jendoubi, *On damped second-order gradient systems*, Journal of Differential Equations, 259 (2015), pp. 3115–3143.

[8] 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.

[9] 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.

[10] D. P. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, Englewood Cliffs, NJ, 2 ed., 1992.

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

[12] P. Bianchi and W. Hachem, *Dynamical behavior of a stochastic forward–backward algorithm using random monotone operators*, Journal of Optimization Theory and Applications, 171 (2016), pp. 90–120.

[13] 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.

[14] 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.

[15] M. Bravo and P. Mertikopoulos, *On the robustness of learning in games with stochastically perturbed payoff observations*, Games and Economic Behavior, to appear (2017).

[16] A. Cabot, H. Engler, and S. Gadat, *On the long time behavior of second order differential equations with asymptotically small dissipation*, Transactions of the American Mathematical Society, 361 (2009), pp. 5983–6017.

[17] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM Journal on Optimization, 3 (1993), pp. 538–543.

[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] S. Gadat and F. Panloup, *Long time behaviour and stationary regime of memory gradient diffusions*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 50 (2014), pp. 564–601.

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

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

[22] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.

[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] N. Karmarkar, *Riemannian geometry underlying interior point methods for linear programming*, in Mathematical Developments Arising from Linear Programming, no. 114 in Contemporary Mathematics, American Mathematical Society, 1990.

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

[29] K. C. Kiwiel, *Free-steering relaxation methods for problems with strictly convex costs and linear constraints*, Mathematics of Operations Research, 22 (1997), pp. 326–349.

[30] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer Berlin Heidelberg, 1992.

[31] W. Krichene, *Continuous and discrete dynamics for online learning and convex optimization*, PhD thesis, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 2016.

[32] W. Krichene, A. Bayen, and P. Bartlett, *Accelerated mirror descent in continuous and discrete time*, in NIPS '15: Proceedings of the 29th International Conference on Neural Information Processing Systems, 2015.

[33] J. Kwon and P. Mertikopoulos, *A continuous-time approach to online optimization*, Journal of Dynamics and Games, 4 (2017), pp. 125–148.

[34] Q. Li, C. Tai, and W. E, *Dynamics of stochastic gradient algorithms*. https://arxiv.org/abs/1511.06251, 2015.

[35] P. Mertikopoulos, *Learning in games with continuous action sets and unknown payoff functions*. https://arxiv.org/abs/1608.07310, 2016.

[36] P. Mertikopoulos, E. V. Belmega, R. Negrel, and L. Sanguinetti, *Distributed stochastic optimization via matrix exponential learning*, IEEE Trans. Signal Process., 65 (2017), pp. 2277–2290.

[37] 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.

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

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

[40] 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.

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

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

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

[44] E. Pardoux and A. Rascanu, *Stochastic Differential Equations, Backwards SDE, Partial Differential Equations*, Springer - Stochastic Modelling and Applied Probability, 2014.

[45] B. T. Polyak, *Introduction to Optimization*, Optimization Software, 1987.

[46] 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.

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

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

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

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

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

[52] 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.

[53] V. Vedral, *The role of relative entropy in quantum information theory*, Reviews of Modern Physics, 74 (2002), pp. 197–234.

[54] A. WIBISONO, A. C. WILSON, AND M. I. JORDAN, *A variational perspective on accelerated methods in optimization.* https://arxiv.org/abs/1603.04245, 2016.