MIRROR DESCENT IN NON-CONVEX STOCHASTIC PROGRAMMING

ZHENGYUAN ZHOU*, PANAYOTIS MERTIKOPOULOS§, NICHOLAS BAMBOS*, STEPHEN BOYD*, AND PETER GLYNN*

Abstract. In this paper, we examine a class of nonconvex stochastic optimization problems which we call variationally coherent, and which properly includes all quasi-convex programs. In view of solving such problems, we focus on the widely used stochastic mirror descent (SMD) family of algorithms, and we establish that the method’s last iterate converges with probability 1. We further introduce a localized version of variational coherence which ensures local convergence of SMD with high probability. These results contribute to the landscape of nonconvex stochastic optimization by showing that quasi-convexity is not essential for convergence: rather, variational coherence, a much weaker requirement, suffices. Finally, building on the above, we reveal an interesting insight regarding the convergence speed of SMD: in variationally coherent problems with sharp minima (e.g. generic linear programs), the last iterate of SMD reaches an exact global optimum in a finite number of steps (a.s.), even in the presence of persistent noise. This result is to be contrasted with existing work on black-box stochastic linear programs which only exhibit asymptotic convergence rates.

1. Introduction

Stochastic mirror descent (SMD) and its variants [2, 8, 10, 24, 26, 28, 40] arguably comprise one of the most widely used families of first-order methods in stochastic optimization – convex and non-convex alike. Roughly speaking, SMD proceeds by aggregating a sequence of i.i.d. gradient samples and then projecting the result back to the problem’s feasible region via a mirror map. In particular, SMD includes as special cases the classical stochastic gradient descent (SGD) algorithm (where Euclidean projections play the role of the mirror map) [30, 32] and various incarnations of the exponential weights (EW) algorithm for online learning (where the mirror map is induced by an entropic regularizer) [19, 22, 35].

Starting with the seminal work of [26], the convergence of SMD has been studied extensively in the context of convex programming (both stochastic and deterministic), saddle-point problems, and monotone variational inequalities. In this context, a recent series of influential papers by Nemirovski et al. [25], Nesterov [28] and Xiao [40] provided tight ergodic convergence bounds for SMD in online/stochastic convex programs, variational inequalities, and saddle-point problems. These results were further boosted by recent work on extra-gradient variants of the algorithm [18, 27], and on an ergodic generalization of the underlying sampling process where the i.i.d.

* Department of Electrical Engineering, Stanford University.
§ Univ. Grenoble Alpes, CNRS, Grenoble INP, Inria, LIG, F-38000, Grenoble, France.

2010 Mathematics Subject Classification. Primary 90C15, 90C26, secondary 90C25, 90C05.
Key words and phrases. Mirror descent; non-convex optimization; stochastic optimization; stochastic approximation; variational coherence.
assumption for the gradient samples is replaced by a uniformly mixing stochastic process [13].

However, all these works place primary emphasis on the convergence of a time-averaged sequence of output points (either unweighted or weighted by the algorithm’s step-size). Despite its successes, this “ergodic convergence” criterion is strictly weaker than the convergence of the algorithm’s last iterate (which is often beneficial in terms of memory and processing power requirements). In addition, most of the analysis focuses on establishing convergence “in expectation” and then employing sophisticated martingale concentration inequalities to derive “large deviations” bounds that hold with high probability. Finally, in all cases, the monotonicity of the gradient operator plays a crucial role: thanks to this monotonicity, it is possible to exploit regret-like bounds and transform them to explicit convergence rates, either for the problem’s convex objective, or the merit function of a variational inequality (which is the standard convergence criterion in the framework of variational inequalities).\footnote{For the role of variational monotonicity in convex programming, see also the recent paper [39].}

By contrast, in this paper, we focus on a class of non-convex programs whose gradient operator is not assumed to satisfy any of the known monotonicity properties: either vanilla monotonicity, quasi-/pseudo-monotonicity, or any of the standard variants encountered in the theory of variational inequalities [15]. Furthermore, we focus squarely on the convergence of the algorithm’s last iterate (which implies ergodic convergence), and we seek to derive global convergence results that hold with probability 1 (as opposed to in the mean). In this setting, the lack of convexity and an inherent averaging mechanism mean that it is no longer possible to employ a regret-based analysis. Instead, to establish convergence, we use the key notion of an asymptotic pseudotrajectory (APT) due to Benaïm and Hirsch [3, 4] to compare the evolution of the iterates of SMD to the flow of a mean, underlying dynamical system.\footnote{For related approaches based on dynamical systems, see also the recent papers [21] and [36]; the latter considers convergence of the last iterate, but only for convex optimization problems.} In so doing, we are able to leverage a series of deterministic, continuous-time convergence results to obtain convergence in a bona fide discrete-time, stochastic framework.

1.1. Our contributions. Our contributions are threefold. First, we introduce a class of non-convex optimization problems, which we call variationally coherent, and which properly includes all quasi-convex programs (and hence all convex and pseudo-convex programs as well). For this class of problems, we show that the last iterate of SMD converges with probability 1 to a global minimum under i.i.d. gradient samples that are only bounded in $L^2$. To the best of our knowledge, this strong convergence guarantee (almost sure convergence of the last iterate of SMD) is not known even for stochastic quasi-convex programs. As such, this result contributes to the landscape of nonconvex stochastic optimization by showing that quasiconvexity is not essential for convergence: rather, variational coherence, a much weaker requirement, suffices.

Our analysis relies on several novel ideas from the theory of stochastic approximation and convex analysis. To begin with, instead of focusing on the discrete-time algorithm directly, we first establish the convergence of an underlying, deterministic dynamical system. We accomplish this by means of a primal-dual analogue
of the Bregman divergence which we call the *Fenchel coupling* and which serves as an energy function for the dynamics of mirror descent. We then connect this continuous-time analysis to SMD via the so-called “ordinary differential equation method” of stochastic approximation, as put forth in [3, 4]. However, even though ODE approximations of discrete-time processes have been studied extensively in control and stochastic optimization, converting the convergence guarantees of an ODE back to the discrete-time process is not automatic and must be done on a case-by-case basis. Here, we achieve this by means of a recurrence result for SMD which we state in Section 3.

Second, we introduce a localized version of variational coherence which includes functions such as the Rosenbrock test function and other instances that fail even local quasi-convexity. Here, in contrast to the variationally coherent case, an “unlucky” sample could throw the SMD process away from the “basin of attraction” of a locally coherent minimizer, possibly never to return; as a result, almost sure convergence seems to be “a bridge too far” in full generality. We obtain the next best thing in this setting: in problems that admit a locally coherent minimizer, the last iterate of SMD converges locally with high probability (for a precise statement, see Section 5).

Finally, to study the algorithm’s convergence speed, we consider a class of optimization problems that admit *sharp* minima, a fundamental notion due to Polyak [29, Chapter 5.2]. In this case, and in stark contrast to existing black-box estimates for the ergodic rate of convergence of SMD (which are asymptotic in nature), we show that the algorithm’s last iterate converges to sharp minima of variationally coherent problems in an almost surely finite number of iterations, provided that the method’s mirror map is surjective. As an important special case of this result, it follows that the last iterate of stochastic gradient descent (or any other incarnation with a surjective mirror map) attains a solution of a stochastic linear program in a finite number of steps (a.s.). This (fairly surprising) result is to be compared to existing work on stochastic linear programming which exhibits asymptotic convergence rates [1, 37]: as opposed to these works, the last iterate of SMD identifies a solution in a finite number of iterations, despite the persistent noise.

2. Setup and preliminaries

Let $\mathcal{X}$ be a convex compact subset of a $d$-dimensional real space $V$ with norm $\|\cdot\|$. Throughout this paper, we focus on the stochastic optimization problem

$$\begin{align*}
\text{minimize} & \quad g(x), \\
\text{subject to} & \quad x \in \mathcal{X},
\end{align*}$$

(Opt)

where the objective function $g: \mathcal{X} \to \mathbb{R}$ is of the form

$$g(x) = E[G(x; \xi)]$$

(2.1)

for some random function $G: \mathcal{X} \times \Xi \to \mathbb{R}$ defined on an underlying (complete) probability space $(\Xi, \mathcal{F}, P)$. We make the following two assumptions regarding (Opt):

**Assumption 1.** $G(x, \xi)$ is continuously differentiable in $x$ for almost all $\xi \in \Xi$. 

Assumption 2. $\nabla G(x; \xi)$ has bounded second moments and is Lipschitz continuous in the mean: $\mathbb{E}[\|\nabla G(x; \xi)\|^2] < \infty$ for all $x \in \mathcal{X}$ and $\mathbb{E}[\nabla G(x; \xi)]$ is Lipschitz on $\mathcal{X}$.\textsuperscript{3}

Assumption 1 is a token regularity assumption which can be relaxed to account for nonsmooth objectives by using subgradient devices (as opposed to gradients). However, this would make the presentation significantly more cumbersome, so we stick with smooth objectives throughout. Assumption 2 is also standard in the stochastic optimization literature: it holds trivially if $\nabla G$ is uniformly Lipschitz (another commonly used condition) and, by the dominated convergence theorem, it further implies that $g$ is smooth and $\nabla g(x) = \nabla \mathbb{E}[G(x; \xi)] = \mathbb{E}[\nabla G(x; \xi)]$ is Lipschitz continuous. As a result, the solution set

$$\mathcal{X}^* = \arg\min g$$

of (Opt) is closed and nonempty (by the compactness of $\mathcal{X}$ and the continuity of $g$).

Remark 2.1. An important special case of (Opt) is when $G(x; \xi) = g(x) + \langle \zeta, x \rangle$ for some $\mathcal{V}^*$-valued random vector $\zeta$ such that $\mathbb{E}[\zeta] = 0$ and $\mathbb{E}[\|\zeta\|^2] < \infty$. This gives $\nabla G(x; \xi) = \nabla g(x) + \zeta$, so (Opt) can also be seen as a model for deterministic optimization problems with noisy gradient observations.

2.1. Variational coherence. With all this at hand, we are now in a position to define the class of \textit{variationally coherent} optimization problems:

**Definition 2.1.** We say that (Opt) is \textit{variationally coherent} if

$$\langle \nabla g(x), x - x^* \rangle \geq 0 \quad \text{for all } x \in \mathcal{X}, \ x^* \in \mathcal{X}^*,$$

with equality if and only if $x \in \mathcal{X}^*$.

Variational coherence will play a central role in our paper so a few remarks and examples are in order:

Remark 2.2. (VC) can be interpreted in two ways. First, as stated, it is a nonrandom condition for $g$, so it applies equally well to \textit{deterministic} optimization problems (with or without noisy gradient observations). Alternatively, by the dominated convergence theorem, (VC) can be written equivalently as

$$\mathbb{E}[\langle \nabla G(x; \xi), x - x^* \rangle] \geq 0.$$  \hspace{1cm} (2.3)

In this form, it can be interpreted as saying that $G$ is variationally coherent “on average”, without any individual realization thereof satisfying (VC). Both interpretations will come in handy later on.

Remark 2.3. Importantly, (VC) does not have to be stated in terms of the solution set of (Opt). Indeed, assume that $\mathcal{C}$ is a nonempty subset of $\mathcal{X}$ such that

$$\langle \nabla g(x), x - p \rangle \geq 0 \quad \text{for all } x \in \mathcal{X}, \ p \in \mathcal{C},$$  \hspace{1cm} (2.4)

with equality if and only if $x \in \mathcal{C}$. Then, as the next lemma indicates, $\mathcal{C} = \arg\min g$:

\textsuperscript{3}In the above, gradients are treated as elements of the dual space $\mathcal{V}^*$ of $\mathcal{V}$ and $\|v\|_* = \sup\{\langle v, x \rangle : \|x\| \leq 1\}$ denotes the dual norm of $v \in \mathcal{V}^*$. We also note that $\nabla G(x; \xi)$ refers to the gradient of $G(x; \xi)$ with respect to $x$; since $\Xi$ need not have a differential structure, there is no danger of confusion.
Lemma 2.2. Suppose that (2.4) holds for some nonempty subset \( C \) of \( X \). Then \( C \) is closed, convex, and it consists precisely of the global minimizers of \( g \).

Corollary 2.3. If \((\text{Opt})\) is variationally coherent, \( \arg\min g \) is convex and compact.

For streamlining purposes, we delegate the proof of Lemma 2.2 to Appendix B and we proceed with some examples of variationally coherent programs:

Example 2.1 (Convex programs). If \( g \) is convex, \( \nabla g \) is a monotone operator \([34]\), i.e.
\[
\langle \nabla g(x) - \nabla g(x'), x - x' \rangle \geq 0 \quad \text{for all } x, x' \in X.
\] (2.5)

By the first-order optimality conditions for \( g \), it follows that
\[
\langle g(x^*), x - x^* \rangle \geq 0 \quad \text{for all } x \in X.
\]

Hence, by monotonicity, we get
\[
\langle \nabla g(x), x - x^* \rangle \geq \langle \nabla g(x^*), x - x^* \rangle \geq 0 \quad \text{for all } x, x^* \in X^*.
\] (2.6)

Finally, by convexity, it follows that
\[
\langle \nabla g(x), x - x^* \rangle < 0 \quad \text{whenever } x^* \in X^* \text{ and } x \in X \setminus X^*,
\]
so equality holds in (2.6) if and only if \( x \in X^* \).

Example 2.2 (Quasi/Pseudo-convex programs). The previous example shows that variational coherence is a weaker and more general notion than convexity and/or operator monotonicity. In fact, as we show below, the class of variationally coherent problems also contains quasi-convex programs:

Proposition 2.4. If \( g \) is quasi-convex, \((\text{Opt})\) is variationally coherent.

Proof. Recall first that \( g \) is quasi-convex if and only if, for all \( x, x' \in X \), \( g(x') \leq g(x) \) implies that \( \langle \nabla g(x), x' - x \rangle \leq 0 \) \([6]\). Hence, if \( x^* \in \arg\min g \), we have \( \langle g(x), x - x^* \rangle \geq 0 \), with the inequality being strict whenever \( x \notin \arg\min g \).

As an additional note, \( g \) is called pseudo-convex if, for all \( x, x' \in X \), we have
\[
\langle \nabla g(x), x - x' \rangle \geq 0 \implies g(x') \geq g(x).
\] (2.7)

Every convex function is pseudo-convex, and every pseudo-convex function is quasi-convex (with both inclusions proper). Consequently, Proposition 2.4 also implies that every pseudo-convex program is variationally coherent.

Example 2.3 (Beyond quasi-convexity). Importantly, quasi-convex programs are properly contained in the class of variationally coherent problems. As a simple example, consider the function
\[
g(x) = 2 \sum_{i=1}^{d} \sqrt{1 + x_i}, \quad x \in [0, 1]^d.
\] (2.8)

When \( d \geq 2 \), it is easy to see \( g \) is not quasi-convex: for instance, taking \( d = 2 \), \( x = (0, 1) \) and \( x' = (1, 0) \) yields \( g(x'/2 + x'/2) = 2\sqrt{6} > 2\sqrt{2} = \max\{g(x), g(x')\} \), so \( g \) is not quasi-convex. On the other hand, to establish (VC), simply note that \( X^* = \{0\} \) and \( \langle \nabla g(x), x - 0 \rangle = \sum_{i=1}^{d} x_i / \sqrt{1 + x_i} > 0 \) for all \( x \in [0, 1]^d \setminus \{0\} \).

For a more elaborate example of a variationally coherent problem that is not quasi-convex, see Fig. 2.
Of course, albeit broad, the class of variationally coherent problems is not exhaustive. A simple example where (VC) fails is the function \( g(x) = \cos(x) \) over \( X = [0, 4\pi] \): since \( \arg \min g = \{\pi, 3\pi\} \) is not convex, Lemma 2.2 shows that \( g \) cannot be variationally coherent. In general, certifying (VC) for a given optimization problem might be difficult, although convenient sufficient conditions can be easily obtained by resorting to ones that ensure convexity or quasi/pseudo-convexity.

2.2. Stochastic mirror descent. To solve (Opt), we will focus on the widely used family of algorithms known as stochastic mirror descent (SMD).

Heuristically, the main idea of the method is as follows: At each iteration, the algorithm takes as input an i.i.d. sample of the gradient of \( G \) at the algorithm’s current state. Subsequently, the method takes a step along this stochastic gradient in the dual space \( \mathcal{Y} \equiv \mathcal{V}^* \) of \( \mathcal{V} \) (where gradients live), the result is “mirrored” back to the problem’s feasible region \( \mathcal{X} \) to obtain a new solution candidate, and the process repeats.

Formally, SMD can be written in pseudocode form as follows (see also Fig. 1):

**Algorithm 1** Stochastic mirror descent (SMD)

Require: Initial score variable \( Y_0 \)

1: \( n \leftarrow 0 \)
2: repeat
3: \( X_n = Q(Y_n) \)
4: \( Y_{n+1} = Y_n - \alpha_{n+1} \nabla G(X_n, \xi_{n+1}) \)
5: \( n \leftarrow n + 1 \)
6: until end
7: return solution candidate \( X_n \)

In the above representation, the key elements of SMD are:

1. The “mirror map” \( Q : \mathcal{Y} \rightarrow \mathcal{X} \) that outputs a solution candidate \( X_n \in \mathcal{X} \) as a function of the auxiliary score variable \( Y_n \in \mathcal{Y} \).
2. The algorithm’s step-size sequence \( \alpha_n > 0 \), chosen to satisfy the “\( \ell^2 - \ell^1 \)” summability condition

\[
\sum_{n=1}^{\infty} \alpha_n^2 < \infty, \quad \sum_{n=1}^{\infty} \alpha_n = \infty. \tag{2.9}
\]

3. A sequence of i.i.d. samples \( \xi_n \in \Xi \) that determine the stochastic objective \( G(x; \xi) \) at each iteration.

In more detail, the algorithm’s mirror map \( Q \) is defined as

\[
Q(y) = \arg \max_{x \in \mathcal{X}} \{ \langle y, x \rangle - h(x) \}, \tag{2.10}
\]

where the “regularization term” \( h(x) \) satisfies the following:

---

4Mirror descent dates back to the original work of Nemirovski and Yudin [26]. More recent treatments include [2, 13, 25, 28, 35] and many others; the specific variant of SMD that we are considering here is most closely related to Nesterov’s “dual averaging” scheme [28].

5The specific indexing convention for \( \xi_n \) has been chosen so that \( Y_n \) and \( X_n \) are both adapted to the natural filtration \( \mathcal{F}_n \) of \( \xi_n \).
Figure 1. Schematic representation of stochastic mirror descent (Algorithm 1).

Definition 2.5. We say that $h : \mathcal{X} \to \mathbb{R}$ is a regularizer (or penalty function) on $\mathcal{X}$ if it is (i) continuous; and (ii) strongly convex, i.e.

$$h(\tau x + (1 - \tau)x') \leq \tau h(x) + (1 - \tau)h(x') - \frac{1}{2}K\tau(1 - \tau)\|x' - x\|^2,$$

for some $K > 0$ and for all $x, x' \in \mathcal{X}$, $\tau \in [0, 1]$. The map $Q : \mathcal{Y} \to \mathcal{X}$ defined in (2.10) is called the mirror map induced by $h$.

We present below two well-known examples of this regularization and their induced mirror maps:

Example 2.4 (Euclidean regularization). Let $h(x) = \frac{1}{2}\|x\|_2^2$. Then, $h$ is 1-strongly convex with respect to $\|\cdot\|_2$ and the induced mirror map is the closest point projection

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

The resulting descent algorithm is known in the literature as stochastic gradient descent (SGD) and we study it in detail in Section 6. For future reference, we also note that $h$ is differentiable throughout $\mathcal{X}$ and $\Pi$ is surjective (i.e. $\text{im } \Pi = \mathcal{X}$).

Example 2.5 (Entropic regularization). Let $\Delta = \{x \in \mathbb{R}^d_+ : \sum_{i=1}^d x_i = 1\}$ denote the unit simplex of $\mathbb{R}^d$, and consider the (negative) Gibbs entropy $h(x) = \sum_{i=1}^d x_i \log x_i$. The function $h(x)$ is 1-strongly convex with respect to the $L^1$-norm on $\mathbb{R}^d$ and a straightforward calculation shows that the induced mirror map is

$$\Lambda(y) = \frac{1}{\sum_{i=1}^d \exp(y_i)}(\exp(y_1), \ldots, \exp(y_d)).$$

This is known as entropic regularization and the resulting algorithm has been studied extensively in the context of linear programming, online learning and game theory [35]. For posterity, we also note that $h$ is differentiable only on the relative interior $\Delta^\circ$ of $\Delta$ and $\text{im } \Lambda = \Delta^\circ$ (i.e. $\Lambda$ is “essentially” surjective).

2.3. Overview of main results. To motivate the analysis to follow, we provide below a brief overview of our main results:

- Global convergence: If $\text{(Opt)}$ is variationally coherent, the last iterate $X_n$ of SMD converges to a global minimizer of $g$ with probability 1.

• **Local convergence:** If \( x^* \) is a locally coherent minimum point of \( g \) (a notion introduced in Section 5), the last iterate \( x_n \) of SMD converges locally to \( x^* \) with high probability.

• **Sharp minima:** If \( Q \) is surjective and \( x^* \) is a sharp minimum of \( g \) (a fundamental notion due to Polyak which we discuss in Section 6), \( x_n \) reaches \( x^* \) in a finite number of iterations (a.s.).

### 3. Recurrence of stochastic mirror descent

As a stepping stone to establish the convergence properties of Algorithm 1, we begin with an interesting recurrence phenomenon that arises in the current context: if \((\text{Opt})\) is variationally coherent, then, with probability 1, \( x_n \) visits any neighborhood of \( x_\infty \) infinitely often. As a result, it follows immediately that at least a subsequence of SMD converges to \( \arg\min g \) (a.s.). Our goal in this section will be to state this result formally and to introduce the analytic machinery used for its proof (and the proofs of our other results).

#### 3.1. The Fenchel coupling

We first define the **Fenchel coupling**, a primal-dual variant of the Bregman divergence \([7]\) that plays an indispensable role as an energy function for SMD:

**Definition 3.1.** Let \( h: \mathcal{X} \to \mathbb{R} \) be a regularizer on \( \mathcal{X} \). The induced Fenchel coupling \( F(p, y) \) between a base-point \( p \in \mathcal{X} \) and a dual vector \( y \in \mathcal{Y} \) is defined as

\[
F(p, y) = h(p) + h^*(y) - \langle y, p \rangle,
\]

where \( h^*(y) = \max_{x \in \mathcal{X}} \{ \langle y, x \rangle - h(x) \} \) denotes the convex conjugate of \( h \).

By Fenchel’s inequality (the namesake of the Fenchel coupling), we have \( h(p) + h^*(y) - \langle y, p \rangle \geq 0 \) with equality if and only if \( p = Q(y) \). As such, \( F(p, y) \) can be seen as a (typically asymmetric) “distance measure” between \( p \in \mathcal{X} \) and \( y \in \mathcal{Y} \). The following lemma provides a stronger version of this positive-definiteness property:

**Lemma 3.2.** 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:

\[
a) \quad F(p, y) \geq \frac{1}{2K} \| Q(y) - p \|^2.
\]

\[
b) \quad F(p, y') \leq F(p, y) + \langle y' - y, Q(y) - p \rangle + \frac{1}{2K} \| y' - y \|^2.
\]

Lemma 3.2 (which we prove in Appendix B) shows that \( Q(y_n) \to p \) whenever \( F(p, y_n) \), so the Fenchel coupling can be used to test the convergence of \( x_n = Q(y_n) \) to a given base point \( p \in \mathcal{X} \). For technical reasons, it will be convenient to also make the converse assumption, namely:

**Assumption 3.** \( F(p, y_n) \to 0 \) whenever \( Q(y_n) \to p \).

Assumption 3 can be seen as a “reciprocity condition”: essentially, it means that the sublevel sets of \( F(p, \cdot) \) are mapped under \( Q \) to neighborhoods of \( p \) in \( \mathcal{X} \) (cf. Appendix B). As such, Assumption 3 is similar to the reciprocity conditions for the Bregman divergence that are widespread in the literature on proximal and forward-backward methods \([9, 20]\). Most common regularizers satisfy this mild technical requirement (including the Euclidean and entropic regularizers of Examples 2.4 and 2.5).
Theorem 3.4. Fix some $B$. Assumptions 1–3 hold, the piece of notation pertaining to measuring distances in Corollary 3.5. There exists a subsequence $\{n_k\}$ of $\{n\}$ such that $X_{n_k} \to X^*$ (a.s.), it must also enter $F$ infinitely often (a.s.), so our claim and Corollary 3.5 follow immediately.

Definition 3.3. Let $S$ be a subset of $X$.

(1) The distance between $S$ and $x \in X$ is defined as $\dist(S, x) = \inf_{x' \in S} \|x - x'\|$, and the corresponding $\varepsilon$-neighborhood of $S$ is

$$\mathbb{B}(S, \varepsilon) = \{x \in X : \dist(S, x) < \varepsilon\}.$$  \hfill (3.3a)

(2) The (setwise) Fenchel coupling between $S$ and $y \in Y$ is defined as $F(S, y) = \inf_{p \in S} F(p, y)$, and the corresponding Fenchel $\delta$-zone of $S$ under $h$ is

$$\mathbb{B}_F(S, \delta) = \{x \in X : x = Q(y) \text{ for some } y \in Y \text{ with } F(S, y) < \delta\}. \hfill (3.3b)$$

We then have the following recurrence result for variationally coherent problems:

Theorem 3.4. Fix some $\varepsilon > 0$ and $\delta > 0$. If $(\text{Opt})$ is variationally coherent and Assumptions 1–3 hold, the (random) iterates $X_n$ of Algorithm 1 enter $\mathbb{B}(X^*, \varepsilon)$ and $\mathbb{B}_F(X^*, \delta)$ infinitely many times (a.s.).

Corollary 3.5. There exists a subsequence $X_{n_k}$ of $X_n$ such that $X_{n_k} \to X^*$ (a.s.).

The proof of Theorem 3.4 consists of three main steps which we outline below:

**Step 1: Martingale properties of $Y_n$.** First, let

$$v(x) = -\mathbb{E}[\nabla G(x; \xi)] = -\nabla g(x)$$ \hfill (3.4)

denote the mean negative gradient of $G$ at $x \in X$. Then, Algorithm 1 may be written in Robbins–Monro form as

$$Y_{n+1} = Y_n + \alpha_{n+1} [v(X_n) + \xi_{n+1}], \hfill (3.5)$$

where

$$\xi_{n+1} = \nabla g(X_n) - \nabla G(X_n; \xi_{n+1}) \hfill (3.6)$$

is the difference between the mean gradient at $X_n$ and the gradient sample that enters Algorithm 1 at stage $n$ (cf. Fig. 1). By construction, $\xi_n$ is a martingale difference sequence adapted to the history (natural filtration) $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$ of $\xi_n$. Furthermore, by Assumption 2, it follows that $\xi_n$ has uniformly bounded second moments; in particular, there exists some finite $V_*$ (cf. (B.14) in Appendix B) such that

$$\mathbb{E}[[\xi_{n+1}]^2 | \mathcal{F}_n] \leq V_*^2 \text{ for all } n, \hfill (3.7)$$

implying in turn that $\xi_n$ is an $L^2$-bounded martingale difference sequence.

**Step 2: Recurrence of $\varepsilon$-neighborhoods.** Invoking the law of large numbers for $L^2$-bounded martingale difference sequences and using the Fenchel coupling as an energy function (cf. Appendix B), we show that if $X_n$ remains outside $\mathbb{B}(X^*, \varepsilon)$ for all sufficiently large $n$, we must also have $F(X^*, Y_n) \to -\infty$ (a.s.). This contradicts the positive-definiteness of $F$, so $X_n$ must enter $\mathbb{B}(X^*, \varepsilon)$ infinitely often (a.s.).

**Step 3: Recurrence of Fenchel zones.** By reciprocity (Assumption 3), $\mathbb{B}_F(X^*, \delta)$ always contains an $\varepsilon$-neighborhood of $X^*$. Since $X_n$ enters $\mathbb{B}(X^*, \varepsilon)$ infinitely often (a.s.), it must also enter $\mathbb{B}_F(X^*, \delta)$ infinitely often (a.s.), so our claim and Corollary 3.5 follow immediately.
Figure 2. Convergence of SMD when $g(r, \theta) = (3 + \sin(5\theta) + \cos(3\theta))r^2(5/3 - r)$ over the unit ball ($0 \leq r \leq 1, 0 \leq \theta \leq 2\pi$). In the left subfigure, we have plotted the graph of $g$; the plot to the right superimposes a typical SMD trajectory over the contours of $g$. Variational coherence follows by noting that the radial part of the gradient of $g$ is positive.

4. GLOBAL CONVERGENCE

The convergence of a subsequence of $X_n$ to the minimum set of (Opt) is one of the crucial steps in establishing our first main result:

**Theorem 4.1** (Almost sure global convergence). Suppose that (Opt) is variationally coherent. Then, under Assumptions 1–3, $\lim_{n \to \infty} \text{dist}(X^*, X_n) = 0$ (a.s.).

**Corollary 4.2.** If $g$ is quasi-convex and Assumptions 1–3 hold, the last iterate of SMD converges to $\text{arg min}_g$ with probability 1.

Theorem 4.1 is our blanket result for variationally coherent problems so, before discussing its proof, some remarks are in order (for a numerical example, see Fig. 2):

**Remark 4.1.** Most of the literature surrounding SMD and its variants (see e.g. [13, 25, 28, 40] and references therein) focuses on the so-called *ergodic average* $\overline{X}_n = \frac{1}{n} \sum_{k=0}^{n} X_k$ of $X_n$. Despite the appealing “self-averaging” properties of $\overline{X}_n$ [25, 28], it is unclear whether the regret-based analysis of SMD used to establish convergence of $\overline{X}_n$ can be extended beyond the standard convex/monotone setting (even to quasi-convex programs). Since convergence of $X_n$ automatically implies that of $\overline{X}_n$, Theorem 4.1 simultaneously establishes the convergence of the last iterate of SMD and extends existing ergodic convergence results to a much wider class of non-convex stochastic programs.

**Remark 4.2.** When there are no random deviations between the samples of $G(x; \xi)$ and their mean (the deterministic case), it can be shown that the last iterate of Algorithm 1 converges to $X^*$ even if $\nabla g$ is not Lipschitz continuous (cf. Assumption 2), and Algorithm 1 is run with a more aggressive class of step-size policies satisfying the milder requirement $\lim_{n \to \infty} \frac{\sum_{k=1}^{n} \alpha_k^2}{\sum_{k=1}^{n} \alpha_k} = 0$. Since these additional assumptions are fairly standard in stochastic optimization, we omit this special case.
due to space constraints; however, we point out that the analysis requires a slightly different proof approach than the one used to prove Theorem 4.1.

In the rest of this section, we discuss the main ideas behind the proof of Theorem 4.1, relegating the details to Appendix B. Our overall approach is to show that SMD can be approximated by a continuous-time dynamical system which converges to the minimum set of (Opt), and then use stochastic approximation tools and techniques to derive the convergence of Algorithm 1. For notational clarity and simplicity, we assume in the discussion to follow that (Opt) admits a unique global minimum, i.e. \( \mathcal{X}^* \) is a singleton; the proof is analogous in the case of multiple minimizers, provided we replace \( x^* \) by \( \mathcal{X}^* \) and use the setwise distance measures of Definition 3.3 instead.

With all this in mind, our proof comprises the following steps:

Step 1: Continuous-time approximation of SMD. Algorithm 1 can be seen as a discretization of the dynamical system

\[
\dot{y} = v(x), \quad x = Q(y),
\]

where, as before, \( v(x) = -\nabla g(x) \). By standard existence and uniqueness results, this system of ODEs is well-posed, i.e. it admits a unique global solution \( y(t) \) for any initial condition. As a result, the family of solutions of (4.1) induces a semiflow on \( \mathcal{Y} \), i.e. a continuous map \( \Phi: \mathbb{R}_+ \times \mathcal{Y} \to \mathcal{Y}, (t,y) \mapsto \Phi_t(y) \), so that \( \Phi_t(y) \) represents the position at time \( t \) of the (unique) solution of (4.1) starting at \( y \).

Note here that \( Q \) is not invertible, so we do not necessarily get an autonomous dynamical system on \( \mathcal{X} \). A similar phenomenon arises in the study of game dynamics based on dual averaging [23].

Step 2: Asymptotic pseudotrajectories. The precise sense in which (4.1) constitutes a continuous-time approximation of Algorithm 1 is that of an asymptotic pseudotrajectory (APT) [3, 4]. To make this precise, connect the iterates \( Y_0, Y_1, \ldots, Y_n, \ldots \) of SMD at times \( 0, \alpha_1, \ldots, \sum_{k=1}^{n-1} \alpha_k, \ldots \) to form a continuous, piecewise affine (random) curve \( Y(t) \). We then show that this linear interpolation of \( Y_n \) is almost surely an APT of (4.1) in the following sense:

**Definition 4.3.** Let \( \Phi \) be a semiflow on \( \mathcal{Y} \). A continuous curve \( Y: \mathbb{R}_+ \to \mathcal{Y} \) is said to be an asymptotic pseudotrajectory (APT) of \( \Phi \) if, for every \( T > 0 \), we have

\[
\lim_{t \to \infty} \sup_{0 \leq h \leq T} \| Y(t + h) - \Phi_h(Y(t)) \|_* = 0.
\]

Step 3: Convergence of the mean dynamics. Having characterized the relation between the iterates of SMD and the flow of the mean dynamics (4.1), we proceed by establishing the convergence of the latter. This is where the energy attributes of the Fenchel coupling manifest themselves: under (4.1), \( F(x^*, y(t)) \) is strictly decreasing unless \( Q(y(t)) = x^* \). As a result, we conclude that the flow of (4.1) converges globally to arg min \( g \), i.e. \( \lim_{t \to \infty} Q(\Phi_t(y)) = x^* \) for every initial condition \( y \in \mathcal{Y} \).

Step 4: Almost uniform convergence. Because \( \mathcal{Y} \) is not compact, it is not reasonable to expect that \( x(t) = Q(y(t)) \) converges uniformly to \( x^* \) because the dynamics'
initial condition could be arbitrarily far from $Q^{-1}(x^*) \subseteq \mathcal{V}$. However, convergence
is “almost” uniform in that, for all $\varepsilon > 0$, there exists some $\tau_\varepsilon > 0$ such that
\[
F(x^*, \Phi_s(y)) \leq \max\{\varepsilon/2, F(x^*, y) - \varepsilon/2\} \quad \text{for all } s \geq \tau_\varepsilon.
\] (4.2)
In other words, after some (uniform) time $\tau_\varepsilon$, all trajectories of (4.1) will lie in an $\varepsilon$-zone around $x^*$, or their distance to $x^*$ (measured in terms of the Fenchel coupling) will have decreased by $\varepsilon$.

**Step 5: Bounding the discretization gap.** We are now in a position to quantify the gap between the interpolated curve $Y(t)$ and the solution orbits of (4.1). To do so, consider a fixed arbitrary horizon $T$. If $F(x^*, Y(t))$ is small, then, by the monotonicity property of the Fenchel coupling (Step 3 above), $F(x^*, \Phi_h(Y(t)))$ will remain small for all $h \in [0, T]$. Since $Y(t)$ is an asymptotic pseudotrajectory of $\Phi$ (by Step 2), $Y(t + h)$ and $\Phi_h(Y(t))$ will be very close for $h \in [0, T]$, at least for $t$
large enough. This means that $F(x^*, Y(t + h))$ should also be small on the entire interval $[0, T]$; more precisely, for all $\varepsilon, T > 0$ there exists some $\tau \equiv \tau(\varepsilon, T)$ such that
\[
F(x^*, Y(t + h)) < F(x^*, \Phi_h(Y(t))) + \varepsilon/2 \quad \text{for all } t \geq \tau, h \in [0, T]
\] (4.3)
with probability 1.

**Step 6: Copy-paste ad infinitum.** Finally, we are ready to put the above pieces together. Step 5 gives us a way to control the amount by which the value of the Fenchel coupling differs between $Y(t)$ and the flow of (4.1) over $[0, T]$. Steps 3 and 4 together allow us to extend such control over successive intervals of the form $[T, 2T], [2T, 3T], \ldots$, thereby establishing that $F(x^*, Y(t + h))$ will remain small for all $h > 0$ if $F(x^*, Y(t))$ is small and $t$ is sufficiently large. In turn, this implies that if $X_n$ visits $\mathbb{B}_F(x^*, \varepsilon)$ for large enough $n$, it will be forever trapped inside a neighborhood of twice that size around $x^*$. Since Theorem 3.4 ensures that $X_n$ visits $\mathbb{B}_F(x^*, \varepsilon)$ infinitively often with probability 1, the above hypothesis is guaranteed to be true. Consequently, this implies that, for all $\varepsilon > 0$, there exists some $\tau_0 \equiv \tau_0(\varepsilon)$ such that
\[
F(x^*, Y(\tau_0 + h)) < \varepsilon \quad \text{for all } h \in [0, \infty).
\] (4.4)

To conclude, Eq. (4.4) implies that $\lim_{n \to \infty} F(x^*, Y_n) = 0$ (a.s.). By Lemma 3.2, this readily implies that $X_n = Q(Y_n) \to x^*$ as $n \to \infty$ (a.s.), thereby completing the proof of Theorem 4.1.

5. Convergence to locally coherent minimizers

We now extend our analysis to account for optimization problems that are only locally variationally coherent. In particular, we only require here that (VC) holds over some neighborhood of a (local) minimizer of $g$:

**Definition 5.1.** Let $\mathcal{C}$ be a set of local minimizers of $g$, viz. $g(x) \geq g(x^*)$ for all $x^* \in \mathcal{C}$ and all $x$ sufficiently close to $\mathcal{C}$. We say that $\mathcal{C}$ is locally coherent if there exists an open neighborhood of $\mathcal{C}$ such that
\[
\langle \nabla g(x), x - x^* \rangle \geq 0 \quad \text{for all } x \in U, x^* \in \mathcal{C},
\] (LVC)
with equality if and only if $x \in \mathcal{C}$. 
In words, (LVC) can be seen as a localized version of (VC). As such, by localizing Lemma 2.2, it readily follows that locally coherent sets are compact and convex sets of local minima of \( g \). We also note that although the minimum set of a globally coherent problem is \textit{a fortiori} locally coherent, the converse need not hold: an example of a function which is not globally coherent but which admits a locally coherent minimum is the standard Rosenbrock benchmark (cf. Fig. 3).\footnote{Local coherence can be proved by a straightforward algebraic calculation (omitted for concision).} Examples such as this show that the profile of \( g \) around a locally coherent set could be highly non-convex and could include a variety of valleys, talwegs and ridges (so even local quasi-convexity may fail to hold; cf. Figs. 2 and 3).

Now, in contrast to globally coherent optimization problems, an “unlucky” gradient sample could throw the SMD algorithm out of the “basin of attraction” of a locally variationally coherent set (the largest neighborhood \( U \) for which (LVC) holds), possibly never to return. As a result, obtaining an almost sure global convergence result seems to be “a bridge too far” for any first-order optimization algorithm. In light of all this, our next result provides a tight relaxation of Theorem 4.1 in the current setting: it shows that SMD converges locally to locally coherent sets with controllably high probability.

**Theorem 5.2** (Local convergence with high probability). Let \( C \) be a locally coherent minimum set of (Opt). Then, under Assumptions 1–3, there exists an open neighborhood \( R \) of \( C \) such that, for every confidence level \( \delta > 0 \), there exists a sufficiently small step-size sequence \( \alpha_n \) such that the last iterate of SMD satisfies

\[
P(X_n \text{ converges to } x^* | X_0 \in R) \geq 1 - \delta. \tag{5.1}
\]

**Remark 5.1.** As a concrete instantiation of Theorem 5.2, fix any \( \beta \in (1/2, 1] \). Then, for every confidence level \( \delta > 0 \), Theorem 5.2 implies that there exists some small enough \( \alpha > 0 \) such that if Algorithm 1 is run with step-size \( \alpha_n = \alpha/n^\beta \), Eq. (5.1) holds. We emphasize the interesting point here: the open neighborhood \( R \) is fixed once for all, and does not depend on the probability threshold \( \delta \). That is, to get convergence with higher probability, one need not start any closer from the optimal solutions: one need only use a smaller step-size sequence satisfying (2.9).

The key idea behind the proof of Theorem 5.2 is as follows: First, without loss of generality, it suffices to consider the case where \( C \) consists of a single minimizer (the argument for the general case carries over identically by using the setwise distance measures of Definition 3.3). Conditioning on the event that \( X_n \) remains sufficiently close to \( x^* \) for all \( n \), convergence can be obtained by invoking Theorem 4.1 and treating (Opt) as a variationally coherent problem over a smaller subset of \( X \) over which (LVC) holds. Therefore, to prove Theorem 5.2, it suffices to show that \( X_n \) remains close to \( x^* \) for all \( n \) with probability no less than \( 1 - \delta \). To achieve this, we rely again on the properties of the Fenchel coupling, and we decompose the stochastic errors affecting each iteration of the algorithm into a first-order \( \mathcal{O}(\alpha_n) \) martingale term and a second-order \( \mathcal{O}(\alpha_n^2) \) submartingale perturbation. Using Doob’s maximal inequality, it can then be shown that the aggregation of both errors remains controllably small with probability at least \( 1 - \delta \), which in turn yields the result.
Figure 3. Convergence of SMD in the case of the Rosenbrock function \( g(x_1, x_2) = (1-x_1)^2 + 100(x_2-x_1^2)^2 \) over the benchmark domain \([-2, 2]^2\).

In the left subfigure, we have plotted the graph of \( g \); the plot to the right superimposes a typical SMD trajectory over the contours of \( g \). Even though the Rosenbrock test function is not variationally coherent, its minimum set is easily checked to be locally coherent, so SMD converges with high probability.

Proof of Theorem 5.2. We break the proof into three steps.

Step 1: Controlling the martingale error. Fix some \( \epsilon > 0 \). As in the proof of Theorem 4.1, let \( \zeta_{n+1} = \nabla g(X_n) - \nabla G(X_n; \xi_{n+1}) \) and set \( \psi_{n+1} = \langle \zeta_{n+1}, X_n - x^* \rangle \).

We show below that there exists a step-size sequence \((\alpha_n)_{n=1}^{\infty}\) such that

\[
P\left( \sup_{n} \sum_{k=1}^{n} \alpha_k \psi_k \leq \epsilon \right) \geq 1 - \frac{\delta}{2}.
\] (5.2)

To show this, we start by noting that, as in the proof of Step 2 of Theorem 3.4, the aggregate process \( S_n = \sum_{k=1}^{n} \alpha_k \psi_k \) is a martingale adapted to the natural filtration \( \mathcal{F}_n \) of \( \xi_n \). Then, letting \( R = \sup_{x \in X} \|x\| \), we can bound the variance of each individual term of \( S_n \) as follows:

\[
E[\psi_k^2] = E[|\langle \zeta_k, X_{k-1} - x^* \rangle|^2 | \mathcal{F}_{k-1}] \leq E[E[|\|\zeta_k\|_2^2 X_{k-1} - x^*\|^2 | \mathcal{F}_{k-1}]]
\]
\[
= E[\|X_{k-1} - x^*\|^2 E[|\zeta_k|^2 | \mathcal{F}_{k-1}]] \leq R^2 V^2_x,
\] (5.3)

where the first inequality follows from the definition of the dual norm and the second one follows from (3.7). Consequently, by Doob’s maximal inequality (Theorem A.4), we have:

\[
P\left( \sup_{0 \leq k \leq n} S_k \geq \epsilon \right) \leq P\left( \sup_{0 \leq k \leq t} |S_k| \geq \epsilon \right) \leq \frac{E[|S_n|^2]}{\epsilon^2} \leq \frac{R^2 V^2_x \sum_{k=1}^{n} \alpha_k^2}{\epsilon^2},
\] (5.4)

where the last inequality follows from expanding \( E[|S_k|^2] \), using Eq. (5.3), and noting that \( E[\zeta_k \zeta_{\ell}] = E[E[\zeta_k \zeta_{\ell}] | \mathcal{F}_{k \lor \ell-1}] = 0 \) whenever \( k \neq \ell \). Therefore, by
picking \( \alpha_n \) so that \( \sum_{k=1}^{\infty} \alpha_k^2 \leq \varepsilon^2 \delta/(2R^2V_n^2) \), Eq. (5.4) gives
\[
\mathbb{P}\left( \sup_{0 \leq k \leq \delta} S_k \geq \varepsilon \right) \leq \frac{R^2V_n^2 \sum_{k=1}^{n} \alpha_k^2}{\varepsilon^2} \leq \frac{R^2V_n^2 \sum_{k=1}^{\infty} \alpha_k^2}{\varepsilon^2} \leq \frac{\delta}{2} \quad \text{for all } n. \tag{5.5}
\]

Since the above holds for all \( n \), our assertion follows.

**Step 2: Controlling the submartingale error.** Again, fix some \( \varepsilon > 0 \) and, with a fair amount of hindsight, let \( R_n = (2K)^{-1} \sum_{k=1}^{n} \alpha_k^2 \| \nabla G(X_{k-1}; \xi_k) \|_a^2 \). By construction, \( R_n \) is a non-negative submartingale adapted to \( \mathcal{F}_n \). We again establish that there exists step-size sequence \( (\alpha_n)_{n=1}^\infty \) satisfying the summability condition (2.9) and such that
\[
\mathbb{P}\left( \sup_{n} R_n \leq \varepsilon \right) \geq 1 - \frac{\delta}{2}. \tag{5.6}
\]

To show this, Doob’s maximal inequality for submartingales (Theorem A.3) yields
\[
\mathbb{P}\left( \sup_{0 \leq k \leq n} R_k \geq \varepsilon \right) \leq \frac{\mathbb{E}[R_n]}{\varepsilon} \leq \frac{B^2 \sum_{k=0}^{n} \alpha_k^2}{2K \varepsilon}, \tag{5.7}
\]
where we used the fact that \( \mathbb{E}[\| \nabla G(X_n; \xi_{n+1}) \|_a^2] \leq B^2 \) for some finite \( B < \infty \) (cf. the proof of the second step of Theorem 3.4 in Appendix B). Consequently, if we choose \( \alpha_n \) so that \( \sum_{k=1}^{\infty} \alpha_k^2 \leq K\delta\varepsilon/B^2 \), Eq. (5.7) readily gives
\[
\mathbb{P}\left( \sup_{0 \leq k \leq n} R_k \geq \varepsilon \right) \leq \frac{B^2 \sum_{k=0}^{\infty} \alpha_k^2}{2K \varepsilon} \leq \frac{\delta}{2} \quad \text{for all } n \tag{5.8}
\]
Since the above is true for all \( n \), Eq. (5.6) follows.

**Step 3: Error aggregation.** We are now ready to put the above two pieces together. To that end, fix some sufficiently small \( \varepsilon > 0 \) so that \( \mathbb{B}_F(x^*, 3\varepsilon) \subset U \), where \( U \) is the open neighborhood given in LVC. Furthermore, let \( \mathcal{R} = \mathbb{B}_F(x^*, \varepsilon) \) and pick a step-size sequence \( \alpha_n \) satisfying (2.9) and such that
\[
\sum_{k=1}^{\infty} \alpha_k^2 \leq \min\left\{ \frac{\delta \varepsilon^2}{2R^2V_n^2}, \frac{K\delta\varepsilon}{B^2} \right\}. \tag{5.9}
\]

Now, if \( X_0 \in \mathcal{R} \), it follows that \( F(x^*, Y_0) < \varepsilon \) by the definition of \( \mathbb{B}_F \) (cf. Definition 3.3). Then, by Eqs. (5.2) and (5.6) above, we have \( \mathbb{P}(\sup_n S_n \geq \varepsilon) \leq \frac{\delta}{2} \) and \( \mathbb{P}(\sup_n R_n \geq \varepsilon) \leq \frac{\delta}{2} \). Consequently, with this choice of \( \alpha_n \), it follows that
\[
\mathbb{P}(\sup_n \max\{S_n, R_n\} \leq \varepsilon) \geq 1 - \delta/2 - \delta/2 = 1 - \delta \tag{5.10}
\]
Then, letting \( F_n = F(x^*, Y_n) \) and using Lemma 3.2 and Eq. (3.5) to expand \( F_{n+1} = F(x^*, Y_{n+1}) \), we get
\[
F_{n+1} = F(x^*, Y_n + \alpha_{n+1}(v(X_n) + \zeta_{n+1}))
\leq F(x^*, Y_n + \alpha_{n+1}(v(X_n) + \zeta_{n+1}, Q(Y_n) - x^*) + \frac{\alpha_{n+1}^2}{2K} \| v(X_n) + \zeta_{n+1} \|_a^2
= F_n + \alpha_{n+1}(v(X_n), X_n - x^*) + \alpha_{n+1} \psi_{n+1} + \frac{\alpha_{n+1}^2}{2K} \| \nabla G(X_n; \xi_{n+1}) \|_a^2 \tag{5.11}
\]
with \( \psi_{n+1} = (\zeta_{n+1}, X_n - x^*) \) defined as in Step 1 of the proof of Theorem 4.1. Telescoping (5.11) then yields

\[
F_n + 1 \leq F_0 + \sum_{k=0}^{n} \alpha_k + 1 \langle v(X_k), X_k - x^* \rangle + S_{n+1} + R_{n+1}
\]

(5.12)

\[
\leq \bar{\varepsilon} + \sum_{k=0}^{n} \alpha_k + 1 \langle v(X_k), X_k - x^* \rangle + \bar{\varepsilon},
\]

(5.13)

with probability at least \( 1 - \delta \). We thus obtain

\[
F(x^*, Y_{n+1}) \leq 3\bar{\varepsilon} + \sum_{k=0}^{n} \alpha_k + 1 \langle v(X_k), X_k - x^* \rangle
\]

(5.14)

for all \( n \), with probability at least \( 1 - \delta \).

Now, if we assume inductively that \( F(x^*, Y_k) \leq 3\bar{\varepsilon} \) (or, equivalently, \( X_k \in \mathbb{B}_F(x^*, 3\bar{\varepsilon}) \)) for all \( k \leq n \) (implying in turn that \( \langle v(X_k), X_k - x^* \rangle \leq 0 \) for all \( k \leq n \)), the above estimate readily yields \( F(x^*, Y_{n+1}) \leq 3\bar{\varepsilon} \) as well. Furthermore, the base case is satisfied automatically since \( X_0 \in \mathcal{R} = \mathbb{B}_F(x^*, \bar{\varepsilon}) \subset \mathbb{B}_F(x^*, 3\bar{\varepsilon}) \). We conclude that \( X_n \) stays in \( \mathbb{B}_F(x^*, 3\bar{\varepsilon}) \subset U \) for all \( n \) with probability at least \( 1 - \delta \); our claim then follows by conditioning on this event and repeating the same steps as in the proof of Theorem 4.1.

6. Sharp minima and applications to stochastic linear programming

Given the randomness involved at each step, obtaining an almost sure (or high probability) bound for the convergence speed of the last iterate of SMD is rather involved. Indeed, in contrast to the ergodic rate analysis of SMD for convex programs, there is no intrinsic averaging in the algorithm’s last iterate, so it does not seem possible to derive a precise black box convergence rate for \( X_n \). Essentially, as in the analysis of Section 5, a single “unlucky” gradient sample could violate any convergence speed estimate that is probabilistically independent of any finite subset of realizations.

Despite this difficulty, if SMD is run with a surjective mirror map, we show that \( X_n \) reaches a minimum point of \((\text{Opt})\) in a finite number of iterations for a large class of optimization problems that includes all generic linear programs.\(^8\) As we noted in Section 2, an important example of a surjective mirror map is the standard Euclidean projection \( \text{proj}_\mathcal{X}(y) = \arg \min_{x \in \mathcal{X}} \|y - x\|_2 \). The resulting descent method is the well-known stochastic gradient descent (SGD) algorithm (cf. Algorithm 2 below), so our results in this section also provide new insights into the behavior of SGD in stochastic linear programs.

6.1. Sharp minima: definition and characterization. The starting point of our analysis is Polyak’s fundamental notion of a sharp minimum [29, Chapter 5.2] which describes functions that grow at least linearly around their minimum points:

**Definition 6.1.** We say that \( x^* \in \mathcal{X} \) is a \( \gamma \)-sharp (local) minimum of \( g \) if

\[
g(x) \geq g(x^*) + \gamma \|x - x^*\| \quad \text{for some } \gamma > 0 \text{ and all } x \text{ sufficiently close to } x^*. 
\]

\(^8\) “Generic linear program” means here that \( \mathcal{X} \) is a polytope, \( g : \mathcal{X} \rightarrow \mathbb{R} \) is affine, and \( g \) is constant only on the zero-dimensional faces of \( \mathcal{X} \).
Definition 6.1 implies that sharp minima are isolated (local) minimizers of \( g \), and they remain invariant under small perturbations of \( g \) (assuming of course that such a minimizer exists in the first place). In what follows, we shall omit the modifier “local” for concision and rely on the context to resolve any ambiguities.

Sharp minima admit a useful geometric interpretation in terms of the polar cone of \( X \). To state it, recall first the following basic facts from convex analysis:

**Definition 6.2.** Let \( X \) be a closed convex subset of \( \mathbb{R}^d \). Then:

1. The tangent cone \( TC(p) \) to \( X \) at \( p \) is defined as the closure of the set of all rays emanating from \( p \) and intersecting \( X \) in at least one other point.
2. The dual cone \( TC^\ast(p) \) to \( X \) at \( p \) is the dual set of \( TC(p) \), viz. \( TC^\ast(p) = \{ y \in \mathbb{R}^d : \langle y, z \rangle \geq 0 \text{ for all } z \in TC(p) \} \).
3. The polar cone \( PC(p) \) to \( X \) at \( p \) is the polar set of \( TC(p) \), viz. \( PC(p) = -TC^\ast(p) = \{ y \in \mathbb{R}^d : \langle y, z \rangle \leq 0 \text{ for all } z \in TC(p) \} \).

We then have the following geometric characterization of sharp minima:

**Lemma 6.3.** If \( x^* \in X \) is a \( \gamma \)-sharp minimum of \( g \), we have
\[
\langle \nabla g(x^*), z \rangle \geq \gamma \| z \| \quad \text{for all } z \in TC(x^*),
\]
so, in particular, \( \nabla g(x^*) \) belongs to the topological interior of \( TC^\ast(x) \). The converse also holds if \( g \) is convex.

The converse part of this lemma immediately implies the following intuitive result:

**Corollary 6.4.** The (necessarily) unique solution of a generic linear program is sharp.

**Proof of Lemma 6.3.** For the direct implication, fix some \( x \in X \) satisfying (6.1), and let \( z = x - x^* \in TC(x^*) \). Then, by the definition of a sharp minimum, we get
\[
g(x^* + \tau z) \geq g(x^*) + \gamma \tau \| z \| \quad \text{for all } \tau \in [0, 1].
\]
In turn, this implies that
\[
g(x^* + tz) - g(x^*) \geq \gamma \| z \| \quad \text{for all sufficiently small } t > 0.
\]
Hence, taking the limit \( \tau \to 0^+ \), we get \( \langle \nabla g(x^*), z \rangle \geq \gamma \| z \| \), and our claim follows from the definition of \( TC(x^*) \) as the closure of the set of all rays emanating from \( x^* \) and intersecting \( X \) in at least one other point. As for the converse implication, simply note that \( g(x) - g(x^*) \geq \langle \nabla g(x^*), x - x^* \rangle \geq \gamma \| x - x^* \| \) if \( g \) is convex.

Sharp minima have several other interesting and useful properties. First, by Lemma 6.3, a sharp minimum is locally variationally coherent. To see this, simply note that for all \( x \in X \) sufficiently close to \( x^* \) (with \( x \neq x^* \)), we have \( z = x - x^* \in TC(x^*) \) and \( \langle \nabla g(x^*), z \rangle \geq \gamma \| z \| \). Consequently, \( \langle \nabla g(x^*), x - x^* \rangle > 0 \), implying by continuity that \( \langle \nabla g(x), x - x^* \rangle > 0 \) for all \( x \) in some open neighborhood of \( x^* \) (excluding \( x^* \)). In addition, if the \( (\text{Opt}) \) is variationally coherent, then a sharp (local) minimum is globally sharp as well.
A second important property is that the dual cone $\text{TC}^*(x^*)$ of a sharp minimum must necessarily have nonempty topological interior – since it contains $\nabla g(x^*)$ by Lemma 6.3. This implies that sharp minima can only occur at corners of $\mathcal{X}$: for instance, if a sharp minimum were an interior point of $\mathcal{X}$, the dual cone to $\mathcal{X}$ at $x^*$ would be a proper linear subspace of the ambient vector space, so it would have no topological content.

6.2. Global convergence in a finite number of iterations. We now turn to showing that, if a variationally coherent program admits a sharp minimum $x^*$, Algorithm 1 reaches $x^*$ in a finite number of iterations (a.s.). The interesting feature here is that convergence is guaranteed to occur in a finite number of iterations: specifically, there exists some (random) $n_0$ such that $X_n = x^*$ for all $n \geq n_0$. In general, this is a fairly surprising property for a first-order descent scheme, even if one considers the ergodic average $n^{-1} \sum_{k=0}^{n} x_k$: a priori, a single “bad” sample could kick $X_n$ away from $x^*$, which is the reason why (ergodic) convergence rates are typically asymptotic.

The key intuition behind our analysis is that sharp minima must occur at corners of $\mathcal{X}$ (as opposed to interior points): while averaging helps accelerate the convergence process when a minimum point is interior, it slows the process down when the minimum occurs at a corner of $\mathcal{X}$. As a further key insight, when the solution of (Opt) occurs at a corner, noisy gradients may still play the role of a random disturbance; however, since they are applied to the dual process $Y_n$, a surjective mirror map would immediately project $Y_n$ back to a corner of $\mathcal{X}$ if $Y_n$ has progressed far enough in the interior of the polar cone to $\mathcal{X}$ at $x$. This ensures that the last iterate $X_n$ of SMD will stay exactly at the optimal point, irrespective of the persistent noise entering Algorithm 1.

Exploiting these insights and the structural properties of sharp minima, we have:

**Theorem 6.5.** Suppose that (Opt) is a variationally coherent problem that admits a sharp minimum $x^*$. If Algorithm 1 is run with a surjective mirror map and Assumptions 1–3 hold, $X_n$ converges to $x^*$ in a finite number of steps (a.s.).

Since quasi-convex problems are variationally coherent, we obtain the following immediate corollary of Theorem 6.5:

**Corollary 6.6.** Suppose that $g$ is quasi-convex and admits a sharp minimum $x^*$. Then, with assumptions as above, $X_n$ converges to $x^*$ in a finite number of steps (a.s.).

The prototypical example of a surjective mirror map is the Euclidean projector $\text{proj}_X(y) = \arg \min_{x \in X} \|y - x\|_2$ induced by the quadratic regularization function $h(x) = \|x\|_2^2/2$ (cf. Example 2.4). The resulting mirror descent scheme is the well-known stochastic gradient descent (SGD) algorithm:

Thus, specializing our results to Euclidean projections and linear programs, we obtain the following novel convergence feature of SGD:

**Corollary 6.7.** In generic linear programs, the last iterate $X_n$ of SGD converges to the problem’s (necessarily unique) solution in a finite number of steps (a.s.).

9Compare for instance the sequence 1, 0, 0, ..., to its ergodic average: the former attains 0 in a single iteration whereas the latter converges to 0 at a rate of $1/n$. 
We used the assumption that... for martingale differences (Theorem A.1 with TC(n) for all n ≥ n0. In turn, this implies that... for all n ≥ n0. Thus, continuing to use the notation v(\(X_n\)) = -\(\nabla g(X_n)\) and \(\zeta_{n+1} = \nabla G(X_n) - \nabla G(X_n; \xi_{n+1})\), we get

\[
\langle Y_{n+1}, z \rangle = \langle Y_{n_0}, z \rangle + \sum_{k=n_0}^{n} \alpha_{k+1} \langle v(X_k), z \rangle + \sum_{k=n_0}^{n} \alpha_{k+1} \langle \zeta_{k+1}, z \rangle 
\]

\[
\leq \|Y_{n_0}\|_\ast - \gamma \frac{n}{2} \sum_{k=n_0}^{n} \alpha_{k+1} + \sum_{k=n_0}^{n} \alpha_{k+1} \langle \zeta_{k+1}, z \rangle,
\]

(6.5)

for all z ∈ TC(\(x^\ast\)) with \(\|z\| \leq 1\).

As we discussed in the proof of Theorem 4.1, \(\alpha_n \zeta_n\) is a martingale difference sequence adapted to the natural filtration of \(\xi_n\). Hence, by the law of large numbers for martingale differences (Theorem A.1 with \(p = 2\) and \(U_n = \sum_{k=0}^{n} \alpha_k\)), we get

\[
\lim_{n \to \infty} \frac{\sum_{k=n_0}^{n} \alpha_{k+1} \zeta_{k+1}}{\sum_{k=n_0}^{n} \alpha_{k+1}} = 0 \quad (\text{a.s.}).
\]

(6.6)

Thus, there exists some \(n^\ast\) such that \(\|\sum_{k=n_0}^{n} \alpha_{k+1} \zeta_{k+1}\|_\ast \leq (\gamma/4) \sum_{k=n_0}^{n} \alpha_{k+1}\) for all \(n \geq n^\ast\) (a.s.). Eq. (6.5) then implies

\[
\langle Y_{n+1}, z \rangle \leq \|Y_{n_0}\|_\ast - \gamma \frac{n}{2} \sum_{k=n_0}^{n} \alpha_{k+1} + \gamma \frac{n}{4} \sum_{k=n_0}^{n} \alpha_{k+1} = \|Y_{n_0}\|_\ast - \gamma \frac{n}{4} \sum_{k=n_0}^{n} \alpha_{k+1},
\]

(6.7)

where we used the assumption that \(\|z\| \leq 1\). Since \(\sum_{k=n_0}^{n} \alpha_{k+1} \to \infty\) as \(n \to \infty\), we get \(\langle Y_n, z \rangle \to -\infty\) with probability 1.

To proceed, we claim that if \(Q(y^\ast) = x^\ast\), then \(y^\ast + \text{PC}(x^\ast) \subseteq Q^{-1}(x^\ast)\), i.e. \(Q^{-1}(x^\ast)\) contains all cones of the form \(y^\ast + \text{PC}(x^\ast)\) for \(y^\ast \in Q^{-1}(x^\ast)\). To see this, note first that \(x^\ast = Q(y^\ast)\) if and only if \(y^\ast \in \partial h(x^\ast)\), where \(\partial h(x^\ast)\) is the set of all subgradients of \(h\) at \(x^\ast\) [33]. Therefore, it suffices to show that \(y^\ast + w \in \partial h(x^\ast)\) whenever \(w \in \text{PC}(x^\ast)\). To show this, note that by the definition of the polar cone, we have

\[
\langle w, x - x^\ast \rangle \leq 0 \quad \text{for all } x \in \mathcal{X}, w \in \text{PC}(x^\ast),
\]

(6.8)
and hence
\[ h(x) \geq h(x^*) + \langle y^*, x - x^* \rangle \geq h(x^*) + \langle y^* + w, x - x^* \rangle. \]  
(6.9)

The above shows that \( y^* + w \in \partial h(x^*) \), as claimed.

With \( Q \) surjective, the set \( Q^{-1}(x^*) \) is nonempty, so it suffices to show that \( Y_n \) lies in the cone \( y^* + PC(x^*) \) for some \( y^* \in Q^{-1}(x^*) \) and all sufficiently large \( n \). To do so, simply note that \( Y_n \in y^* + PC(x^*) \) if and only if \( \langle Y_n - y^*, z \rangle \leq 0 \) for all \( z \in TC(x^*) \) with \( \|z\| = 1 \). Since \( \langle Y_n, z \rangle \) converges to \(-\infty\) (a.s.), our assertion is immediate.

6.3. Local convergence in a finite number of iterations. In extending our convergence analysis to problems with locally coherent minima (cf. Section 5), we showed that SMD converges locally with high probability. Our last result in this section is that even if \((\text{Opt})\) is not variationally coherent, then, with high probability, SMD converges locally to sharp local minima in a finite number of iterations:

**Theorem 6.8.** Suppose that \( g \) admits a sharp (local) minimum \( x^* \). If Algorithm 1 is run with a surjective mirror map and Assumptions 1–3 hold, there exists an open neighborhood \( \mathcal{R} \) of \( x^* \) such that, for every confidence level \( \delta > 0 \), there exists a sufficiently small step-size sequence \( \alpha_n \) such that the last iterate of SMD satisfies
\[ P(X_n \text{ converges to } x^* \text{ in a finite number of steps } | X_0 \in \mathcal{R}) \geq 1 - \delta. \]  
(6.10)

**Proof.** Under the stated assumptions, Theorem 5.2 implies that there exists an open neighborhood \( \mathcal{R} \) of \( x^* \) such that \( P(\lim_{n \to \infty} X_n = x^* | X_0 \in \mathcal{R}) \geq 1 - \delta \). In turn, this means that there exists some (random) \( n_0 \) which is finite with probability at least \( 1 - \delta \) and is such that \( \langle \nabla g(x_n), z \rangle \geq \gamma \|z\|/2 \) for all \( n \geq n_0 \) (by the sharpness assumption). Our assertion then follows by conditioning on this event and proceeding as in the proof of Theorem 6.5.

7. Discussion

To conclude, we first note that our analysis can be extended to the study of stochastic variational inequalities with possibly non-monotone operators. The notion of a variationally coherent problem still make sense for a variational inequality “as is”, and the Fenchel coupling can also be used to establish almost sure convergence to the solution set of a variational inequality. That said, there are several details that need to be adjusted along the way, so we leave this direction to future work.

Second, even though variationally coherent problems contain all quasi-convex optimization problems, they do not include all quasi-concave problems (though, for instance, Example 2.3 is quasi-concave). The recent work [16] presents an interesting algorithm for minimizing a certain class of quasi-concave optimization problems; extending our analysis to this setting is another important direction for future work. Likewise, there are other first-order methods enjoying widespread use and success, particularly in the framework of distributed large-scale optimization problems (ADMM, augmented Lagrangian methods, etc.) [5, 12, 38]. A particularly worthwhile direction would thus be to understand the convergence properties of these methods in the non-convex regime, under the condition of variational coherence.

Finally, we should also mention that another merit of SMD is that, at least for (strongly) convex optimization problems [14, 31], the algorithm is amenable
to asynchronous parallelization. This is an increasingly desirable advantage, especially in the presence of large-scale datasets that are characteristic of “big data” applications requiring the computing power of a massive number of parallel processors. Although we do not take on this question in this paper, the techniques developed here can potentially be leveraged to provide theoretical guarantees for certain non-convex stochastic programs when SMD is run asynchronously and in parallel.

Appendix A. Elements of martingale limit theory

In this appendix, we state for completeness some basic results from martingale limit theory which we use throughout our paper. The statements are adapted from [17] where we refer the reader for detailed proofs.

We begin with a strong law of large numbers for martingale difference sequences:

**Theorem A.1.** Let \( M_n = \sum_{k=0}^{n} d_k \) be a martingale with respect to an underlying stochastic basis \( (\Omega, \mathcal{F}, (\mathcal{F}_n)_{n=0}^{\infty}, \mathbb{P}) \) and let \( (U_n)_{n=0}^{\infty} \) be a nondecreasing sequence of positive numbers with \( \lim_{n \to \infty} U_n = \infty \). If \( \sum_{n=0}^{\infty} U_n^{-p} \mathbb{E}[|d_k|^p | \mathcal{F}_n] < \infty \) for some \( p \in [1, 2] \) (a.s.), we have:

\[
\lim_{n \to \infty} \frac{M_n}{U_n} = 0 \quad (a.s.) \tag{A.1}
\]

The second result we use is Doob’s martingale convergence theorem:

**Theorem A.2.** If \( M_n \) is a submartingale that is bounded in \( L^1 \) (i.e. \( \sup_n \mathbb{E}[|M_n|] < \infty \)), \( M_n \) converges almost surely to a random variable \( M \) with \( \mathbb{E}[|M|] < \infty \).

The next result is also due to Doob, and is known as Doob’s maximal inequality:

**Theorem A.3.** Let \( M_n \) be a non-negative submartingale and fix some \( \varepsilon > 0 \). Then:

\[
\mathbb{P}(\sup_n M_n \geq \varepsilon) \leq \frac{\mathbb{E}[M_n]}{\varepsilon}. \tag{A.2}
\]

Finally, a widely used variant of Doob’s maximal inequality is the following:

**Theorem A.4.** With assumptions and notation as above, we have

\[
\mathbb{P}(\sup_n |M_n| \geq \varepsilon) \leq \frac{\mathbb{E}[M_n^2]}{\varepsilon^2}. \tag{A.3}
\]

Appendix B. Technical proofs

In this appendix, we present the proofs that were omitted from the main text. We begin with the implications of variational coherence for the minimum set of (Opt):

**Proof of Lemma 2.2.** We first show that \( C \) is closed. To do so, fix some \( p \in C \) and let \( (q_j)_{j=1}^{\infty} \) be a sequence in \( C \) converging to some limit point \( q \in \mathcal{X} \) (but not necessarily in \( C \)). Since \( q_j \in C \) for all \( j \in \mathbb{N} \), Eq. (2.4) gives

\[
\langle \nabla g(q_j), q_j - p \rangle = 0. \tag{B.1}
\]

Hence, by taking limits, we get

\[
\langle \nabla g(q), q - p \rangle = \lim_{j \to \infty} \langle \nabla g(q_j), q_j - p \rangle = 0, \tag{B.2}
\]

implying in turn that \( q \in C \). Since the sequence \( q_j \) was chosen arbitrarily, we conclude that \( C \) is closed.
To show that $\mathcal{C}$ is convex, take $p_0, p_1 \in \mathcal{X}$ and let $p_\tau = (1 - \tau)p_0 + \tau p_1$ for $\tau \in [0, 1]$. By substituting in (2.4), we get
\[ 0 \leq \langle \nabla g(p_\tau), p_\tau - p_0 \rangle = \tau \langle \nabla g(p_\tau), p_1 - p_0 \rangle, \tag{B.3a} \]
and, symmetrically:
\[ 0 \leq \langle \nabla g(p_\tau), p_\tau - p_1 \rangle = -(1 - \tau)\langle \nabla g(p_\tau), p_1 - p_0 \rangle. \tag{B.3b} \]
Combining Eqs. (B.3a) and (B.3b) above yields $\langle \nabla g(p_\tau), p_\tau - p_0 \rangle = 0$ and therefore $\langle \nabla g(p_\tau), p_\tau - p_0 \rangle = -\tau \langle \nabla g(p_\tau), p_1 - p_0 \rangle = 0$. By (2.4), this implies that $p_\tau \in \mathcal{C}$, so, with $p_0, p_1$ arbitrary, it follows that $\mathcal{C}$ is convex.

We are left to show that $\mathcal{C} = \arg \min g$. To that end, fix some $p \in \mathcal{C}$, $x \in \mathcal{X}$, and let $z = x - p$. Then, for all $\tau \in [0, 1]$, we have
\[ \frac{d}{d\tau} g(p + \tau z) = \langle \nabla g(p + \tau z), z \rangle = \frac{1}{\tau} \langle \nabla g(p + \tau z), p + \tau z - p \rangle \geq 0, \tag{B.4} \]
where the last inequality follows from (2.4). This shows that $g(p + \tau z)$ is increasing in $\tau$, so $g(x) \geq g(p)$. With $x$ chosen arbitrarily, we conclude that $p \in \arg \min g$, i.e. $\mathcal{C} \subseteq \arg \min g$. For the converse implication, suppose that $x \notin \mathcal{C}$ is a global minimizer of $g$ that does not lie in $\mathcal{C}$. Since $\mathcal{C}$ is closed, we have $(1 - \tau)p + \tau x \notin \mathcal{C}$ for all $\tau$ sufficiently close to 1. Therefore, by (2.4), the inequality (B.4) is strict for $\tau$ close to 1 implying that $g(p + \tau z)$ is strictly increasing for $\tau$ close to 1. This gives $g(x) > g(p)$, contradicting the original assumption that $x \in \arg \min g$. We conclude that $\arg \min g \supseteq \mathcal{C}$ and our proof is complete.

We proceed with the fundamental properties of the Fenchel coupling:

\textbf{Proof of Lemma 3.2.} To prove the first claim, let $x = Q(y) = \arg \max_{x' \in \mathcal{X}} \{ \langle y, x' \rangle - h(x') \}$, so $y \in \partial h(x)$ from standard results in convex analysis [33]. We thus get:
\[ F(p, y) = h(p) + h^*(y) - \langle y, p \rangle = h(p) - h(x) - \langle y, p - x \rangle. \tag{B.5} \]
Since $y \in \partial h(x)$ and $h$ is $K$-strongly convex, we also have
\[ h(x) + \tau \langle y, p - x \rangle \leq h(x + \tau (p - x)) \leq h(x) + \tau h(p) - \frac{1}{2}K\tau (1 - \tau)\|x - p\|^2 \tag{B.6} \]
for all $\tau \in [0, 1]$, thereby leading to the bound
\[ \frac{1}{2}K(1 - \tau)\|x - p\|^2 \leq h(p) - h(x) - \langle y, p - x \rangle = F(p, y). \tag{B.7} \]
Our claim then follows by letting $\tau \to 0^+$ in (B.7).

For our second claim, we start by citing a well-known duality principle for strongly convex functions [34, Theorem 12.60]: If $f: \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ is proper, lower semi-continuous and convex, its convex conjugate $f^*$ is $\sigma$-strongly convex if and only if $f$ is differentiable and satisfies
\[ f(x') \leq f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2\sigma}\|x' - x\|^2 \tag{B.8} \]
for all $x, x' \in \mathbb{R}^d$. In our case, if we extend $h$ to all of $\mathcal{Y}$ by setting $h \equiv +\infty$ outside $\mathcal{X}$, we have that $h$ is $K$-strongly convex, lower semi-continuous and proper, so $(h^*)^* = h$ [34, Theorem 11.1]. Further, it is easy to see that $h^*$ is proper, lower semi-continuous and convex (since it is a point-wise maximum of affine functions by definition), so the $K$-strong convexity of $h = (h^*)^*$ implies that $h^*$ is differentiable and satisfies
\[ h^*(y') \leq h^*(y) + \langle y' - y, \nabla h^*(y) \rangle + \frac{1}{2\kappa} \|y' - y\|^2. \tag{B.9} \]
for all \( y, y' \in \mathcal{Y} \), where the last equality follows from the fact that \( \nabla h^*(y) = Q(y) \). Therefore, substituting the preceding inequality in the definition of the Fenchel coupling, we obtain

\[
F(x, y') = h(x) + h^*(y') - \langle y', x \rangle
\]

\[
\leq h(x) + h^*(y) + \langle y' - y, \nabla h^*(y) \rangle + \frac{1}{2K} \| y' - y \|_*^2 - \langle y', x \rangle
\]

\[
= F(x, y) + \langle y' - y, Q(y) - x \rangle + \frac{1}{2K} \| y' - y \|_*^2,
\]

and our assertion follows. 

Proof of Theorem 3.4. Our proof proceeds step-by-step, as discussed in Section 3:

Step 1: Martingale properties of \( Y_n \). By Assumption 2 and the fact that finiteness of second moments implies finiteness of first moments, we get \( \mathbb{E}[\| g(x; \xi_{n+1}) \|_*] < \infty \). We then claim that \( \zeta_{n+1} = \nabla g(X_n) - \nabla G(X_n; \xi_{n+1}) \) is an \( L^2 \)-bounded martingale difference sequence with respect to the natural filtration of \( \xi_{n+1} \).

(1) Since \( Y_n \) is adapted to \( F_n \) and \( \xi_n \) is i.i.d., we readily get

\[
\mathbb{E}[\zeta_{n+1} | F_n] = \mathbb{E}[\nabla g(X_n) - \nabla G(X_n; \xi_{n+1}) | F_n]
\]

\[
= \mathbb{E}[\nabla g(X_n) - \nabla G(X_n; \xi_{n+1}) | \xi_1, \ldots, \xi_n]
\]

\[
= \nabla g(X_n) - \nabla g(X_n) = 0,
\]

so \( \mathbb{E}[\zeta_{n+1}] = \mathbb{E}[\mathbb{E}[\zeta_{n+1} | F_n]] = 0 \), i.e. \( \zeta_n \) has zero mean.

(2) Furthermore, the \( L^2 \) norm of \( \zeta \) satisfies

\[
\mathbb{E}[\| \zeta_{n+1} \|_*^2 | F_n] = \mathbb{E}[\| \nabla g(X_n) - \nabla G(X_n; \xi_{n+1}) \|_*^2 | F_n]
\]

\[
\leq 2 \mathbb{E}[\| \nabla g(X_n) \|_*^2 | F_n] + 2 \mathbb{E}[\| \nabla G(X_n; \xi_{n+1}) \|_*^2 | F_n]
\]

\[
\leq 2 \| \nabla g(X_n) \|_*^2 + 2 \mathbb{E}[\| \nabla G(X_n; \xi_{n+1}) \|_*^2 | F_n]
\]

\[
= 4 \| \nabla g(X_n) \|_*^2,
\]

where we used the dominated convergence theorem to interchange expectation and differentiation in the second line and Jensen’s inequality in the third. Since \( \mathcal{X} \) is compact, there exists some finite \( B \) such that \( \| \nabla g(x) \|_* \leq B \) for all \( x \in \mathcal{X} \). We thus obtain

\[
\mathbb{E}[\| \zeta_{n+1} \|_*^2] = \mathbb{E}[\mathbb{E}[\| \zeta_{n+1} \|_*^2 | F_n]] \leq V_*^2,
\]

where we set \( V_* = 2B \).

Step 2: Recurrence of \( \varepsilon \)-neighborhoods. We proceed to show that every \( \varepsilon \)-neighborhood \( \mathbb{B}(\mathcal{X}^*, \varepsilon) \) of \( \mathcal{X}^* \) is recurrent under \( X_n \). To do so, fix some \( \varepsilon > 0 \) and assume ad absurdum that \( X_n \) enters \( \mathbb{B}(\mathcal{X}^*, \varepsilon) \) only a finite number of times, so ther exists some finite \( n_0 \) such that \( \text{dist}(\mathcal{X}^*, X_n) \geq \varepsilon \) for all \( n \geq n_0 \). Since \( \mathcal{X} \setminus \mathbb{B}(\mathcal{X}^*, \varepsilon) \) is compact, \( v(x) = -\nabla g(x) \) is continuous in \( x \), and \( \langle v(x), x - x^* \rangle = -\langle \nabla g(x), x - x^* \rangle < 0 \) for all \( x \in \mathcal{X} \setminus \mathbb{B}(\mathcal{X}^*, \varepsilon) \), \( x^* \in \mathcal{X}^* \), there exists some \( c \equiv c(\varepsilon) > 0 \) such that

\[
\langle v(x), x - x^* \rangle \leq -c < 0 \quad \text{for all } x \in \mathcal{X} \setminus \mathbb{B}(\mathcal{X}^*, \varepsilon), \ x^* \in \mathcal{X}^*.
\]
To proceed, let $R = \max_{x \in X} \|x\|$ and set $\beta_n = \alpha_n^2/(2K)$. Then, letting $F_n = F(x^*, Y_n)$ and $\psi_{n+1} = (\zeta_{n+1}, X_n - x^*)$, Lemma 3.2 yields

$$F_{n+1} = F(x^*, Y_{n+1}) = F(x^*, Y_n + \alpha_n+1(v(X_n) + \zeta_{n+1}))$$

$$\leq F(x^*, Y_n) + \alpha_n+1\langle v(X_n) + \zeta_{n+1}, X_n - x^* \rangle + \beta_n+1\|v(X_n) + \zeta_{n+1}\|^2_s$$

$$= F_n + \alpha_n+1\langle v(X_n), X_n - x^* \rangle + \alpha_n+1\psi_n+1 + \beta_n+1\|v(X_n) + \zeta_{n+1}\|^2_s$$

$$\leq F_n - \alpha_n+1c + \alpha_n+1\psi_n+1 + 2\beta_n+1\left[\|v(X_n)\|^2_s + \|\zeta_{n+1}\|^2_s\right]. \tag{B.16}$$

Hence, letting $U_{n+1} = \sum_{k=n_0}^n \alpha_{k+1}$ and telescoping from $n_0$ to $n + 1$, we get

$$F_{n+1} \leq F_n - c \sum_{k=n_0}^n \alpha_{k+1} + \sum_{k=n_0}^n \alpha_{k+1}\psi_{k+1} + 2 \sum_{k=n_0}^n \beta_{k+1}\left[\|v(X_k)\|^2_s + \|\zeta_{k+1}\|^2_s\right]$$

$$= F_n - U_{n+1}\left[c - \sum_{k=n_0}^n \alpha_{k+1}\psi_{k+1}/U_{n+1}\right] + 2 \sum_{k=n_0}^n \left[\|v(X_k)\|^2_s + \|\zeta_{k+1}\|^2_s\right]$$

$$\leq F_n - U_{n+1}\left[c - \sum_{k=n_0}^n \alpha_{k+1}\psi_{k+1}/U_{n+1}\right] + 2B^2 \sum_{k=n_0}^n \beta_{k+1} + 2 \sum_{k=n_0}^n \beta_{k+1}\|\zeta_{k+1}\|^2_s \tag{B.17}$$

We now proceed to bound each term of (B.17). First, by construction, we have

$$E[\psi_{n+1} | F_n] = E[\langle \zeta_{n+1}, X_n - x^* \rangle | F_n] = (E[\zeta_{n+1} | F_n], X_n - x^*) = 0, \tag{B.18}$$

where we used the fact that $X_n$ is $F_n$-measurable. Moreover, by Young’s inequality, we also have

$$|\psi_{n+1}| = |\langle \zeta_{n+1}, X_n - x^* \rangle| \leq \|\zeta_{n+1}\|_s \|X_n - x^*\| \leq 2R\|\zeta_{n+1}\|_s, \tag{B.19}$$

where, as before, $R = \max_{x \in X} \|x\|$. Eq. (B.14) then gives

$$E[\psi_{n+1}^2 | F_n] \leq (2R)^2 \E[\|\zeta_{n+1}\|_s^2 | F_n] \leq 4R^2 V_s^2, \tag{B.20}$$

implying in turn that $\psi_n$ is an $L^2$-bounded martingale difference sequence. It then follows that $\psi_n$ satisfies the summability condition

$$\sum_{n=n_0}^\infty \frac{E[\|\alpha_{n+1}\psi_{n+1}\|^2 | F_n]}{U_{n+1}^2} \leq 4R^2 V_s^2 \sum_{n=n_0}^\infty \frac{\alpha_{n+1}^2}{U_{n+1}^2} < \infty, \tag{B.21}$$

where the last inequality follows from the assumption that $\alpha_n$ is square-summable. Thus, by the law of large numbers for martingale difference sequences (Theorem A.1), it follows that

$$\sum_{k=n_0}^n \frac{\alpha_{k+1}\psi_{k+1}}{U_{n+1}} \to 0 \quad \text{as} \quad n \to \infty \quad \text{(a.s.)}. \tag{B.22}$$

Then, with $\sum_{k=n_0}^\infty \alpha_k = \infty$, we obtain

$$\lim_{n \to \infty} U_{n+1}\left[c - \sum_{k=n_0}^n \frac{\alpha_{k+1}\psi_{k+1}}{U_{n+1}}\right] = \infty \quad \text{(a.s.)}. \tag{B.23}$$
The third term of (B.17) is trivially bounded because $\alpha_n$ is square-summable. Finally, for the last term, let $S_{n+1} = \sum_{k=n_0}^{n} \beta_{k+1} \|\xi_{k+1}\|^2$. Clearly, $S_n$ is $F_n$-measurable and nondecreasing; in addition, we have

$$E[S_n] = \sum_{k=n_0}^{n} \beta_{k+1} E[\|\xi_{k+1}\|^2] \leq V_*^2 \sum_{k=n_0}^{n} \beta_{k+1} < \infty,$$

(B.24)

with the last step following from (B.14). This implies that $S_n$ is an $L^1$-bounded submartingale so, by Doob’s submartingale convergence theorem (Theorem A.2), $S_n$ converges almost surely to some random variable $S_\infty$. This implies that the last term of (B.17) is bounded; hence, combining all of the above, we finally obtain

$$\limsup_{n \to \infty} F_n = -\infty \quad (a.s.),$$

(B.25)

contradicting the positive-definiteness of the Fenchel coupling (cf. Lemma 3.2). We thus conclude that $X_n$ enters $\mathbb{B}(\mathcal{X}^*, \varepsilon)$ infinitely many times (a.s.), as claimed.

**Step 3: Recurrence of Fenchel zones.** Using the reciprocity of the Fenchel coupling (Assumption 3), we show below that every Fenchel zone $\mathbb{B}_F(\mathcal{X}^*, \delta)$ of $\mathcal{X}^*$ contains an $\varepsilon$-neighborhood of $\mathcal{X}^*$. Then, given that $X_n$ enters $\mathbb{B}(\mathcal{X}^*, \varepsilon)$ infinitely often (per the previous step), it will also enter $\mathbb{B}_F(\mathcal{X}^*, \delta)$ infinitely often.

To establish this claim, assume instead that there is no $\varepsilon$-neighborhood $\mathbb{B}(\mathcal{X}^*, \varepsilon)$ contained in some $\delta$-zone $\mathbb{B}_F(\mathcal{X}^*, \delta)$. Then, for all $\ell > 0$ there exists some $y_\ell \in \mathcal{Y}$ such that $\text{dist}(\mathcal{X}^*, Q(y_\ell)) = 1/\ell$ but $F(\mathcal{X}^*, y_\ell) \geq \varepsilon$. This produces a sequence $(y_\ell)_{\ell=1}^\infty$ such that $\text{dist}(\mathcal{X}^*, Q(y_\ell)) \to 0$ but $F(\mathcal{X}^*, y_\ell) \geq \varepsilon$. Since $\mathcal{X}$ is compact and $\mathcal{X}^*$ is closed, we can assume without loss of generality that $Q(y_\ell) \to p$ for some $p \in \mathcal{X}^*$ (at worst, we only need to descend to a subsequence of $y_\ell$). We thus get

$$\varepsilon \leq F(\mathcal{X}^*, y_\ell) \leq F(p, y_\ell).$$

(B.26)

However, since $Q(y_\ell) \to p$, Assumption 3 gives $F(p, y_\ell) \to 0$, a contradiction. We conclude that $\mathbb{B}_F(\mathcal{X}^*, \delta)$ contains an $\varepsilon$-neighborhood of $\mathcal{X}^*$, completing our proof.

We finally turn to the proof of our main global convergence result:

**Proof of Theorem 4.1.** We proceed by proving in turn the 6 steps laid out in Section 4.

**Step 1: Continuous-time approximation of SMD.** Since $h$ is $K$ strongly convex, the induced mirror map $Q$ is $(1/K)$-Lipschitz continuous by standard results in convex analysis [33]. Since $v(x)$ is Lipschitz continuous, $v(Q(\cdot))$ is also Lipschitz continuous, so the mean dynamics (4.1) admit a unique global solution from every initial condition [11]. Letting $\Phi_t(y) \in \mathcal{Y}$ denote the position at time $t$ of the unique solution starting at $y \in \mathcal{Y}$, continuity on initial conditions [11] shows that $\Phi_t$ is jointly continuous in $t$ and $y$, so we get a well-defined continuous semiflow $\Phi : \mathbb{R}_+ \times \mathcal{Y} \to \mathcal{Y}$ mapping $(y, t) \mapsto \Phi_t(y)$.

**Step 2: Asymptotic pseudotrajectories.** As outlined in the main body of the paper, let $Y(t)$ denote the linear interpolation of $Y_n$; specifically, letting $\tau_n = \sum_{k=0}^{n} \alpha_n$,\textsuperscript{10} the process $Y(t)$ is defined as

$$Y(t) = Y_{n-1} + (t - \tau_{n-1}) \frac{Y_n - Y_{n-1}}{\alpha_n} \quad \text{for } t \in [\tau_{n-1}, \tau_n], \ n \in \mathbb{N}.$$  

(B.27)

\textsuperscript{10}For $n = 0$, we have $\tau_n = 0$, in tune with the standard convention for the empty sum.
We now claim that $Y(t)$ is (a.s.) an APT of $\Phi$ in the sense of Definition 4.3. Indeed, as shown in [3, Propositions 4.1 and 4.2], a sufficient condition for the Robbins-Monro algorithm (3.5) to be an APT of (4.1) is:

1. $\zeta_{n+1}$ is a martingale difference sequence with $\sup_n E[\|\zeta_{n+1}\|_q^2] < \infty$ for some $q \geq 2$ such that $\sum_{n=1}^{\infty} \alpha_n^{1+q/2} < \infty$.
2. The vector field of motion $V(y) = v(Q(y))$ of (4.1) is Lipschitz continuous and bounded.

As we showed in the proof of Theorem 3.4, the martingale noise term $\zeta_{n+1}$ of (3.5) satisfies the above requirements for $q = 2$ (recall also that $\alpha_n$ is taken to be square-summable). Since $\mathcal{X}$ is compact, $v = -\nabla g$ is also bounded, so our assertion follows.

**Step 3: Convergence of the mean dynamics.** By standard results in convex analysis [34, Chap. 12], the convex conjugate $h^*$ of $h$ is differentiable and

$$\nabla h^*(y) = Q(y).$$

(B.28)

Thus, fixing a base point $p \in \mathcal{X}$ and a solution trajectory $y(t)$ of (4.1), we obtain

$$\frac{d}{dt} F(p, y(t)) = \frac{d}{dt} [h(p) + h^*(y(t)) - \langle y(t), p \rangle]$$

$$= \langle y(t), \nabla h^*(y(t)) \rangle - \langle y(t), p \rangle$$

$$= \langle v(x(t)), Q(y(t)) \rangle - \langle v(x(t)), p \rangle$$

$$= \langle v(x(t)), x(t) - x^* \rangle,$$

(B.29)

where the second equality follows from (B.28). Applying the above to a minimizer $x^* \in \mathcal{X}^*$ of (2.1) and using (VC), we conclude that

$$\frac{d}{dt} F(x^*, y(t)) = \langle v(x(t)), x(t) - x^* \rangle \leq 0,$$

(B.30)

with equality if and only if $x(t) = x^*$.

Now, let $\hat{x}$ be an $\omega$-limit of $x(t)$ and assume that $\hat{x} \notin \arg \min g$. Since $\mathcal{X}^*$ is closed, there exists a neighborhood $U$ of $\hat{x}$ in $\mathcal{X}$ such that $\langle v(x), x - x^* \rangle \leq -c$ for some $c > 0$ and for all $x^* \in \mathcal{X}^*$. Furthermore, since $\hat{x}$ is an $\omega$-limit of $x(t)$, there exists an increasing sequence of times $t_n \uparrow \infty$ such that $x(t_n) \in U$ for all $n$. Then, for all $\tau > 0$, we have

$$\|x(t_n + \tau) - x(t_n)\| = \|Q(y(t_n + \tau)) - Q(y(t_n))\| \leq \frac{1}{K} \|y(t_n + \tau) - y(t_n)\|_*$$

$$\leq \frac{1}{K} \int_{t_n}^{t_n + \tau} \|v(x(s))\|_* ds \leq \frac{\tau}{K} \max_{x \in \mathcal{X}} \|v(x)\|_*.$$

(B.31)

Since this bound does not depend on $n$, there exists some $\delta > 0$ such that $x(t_n + \tau) \in U$ for all $\tau \in [0, \delta]$ and all $n \in \mathbb{N}$ (so we also have $\langle v(x(t_n + \tau)), x(t_n + \tau) - x^* \rangle \leq -c$).

---

11All the statements made here are true almost surely. For convenience, we will omit repeating this assumption except when confusion may arise. Alternatively, one can treat what follows as a path-by-path argument, with each ensuing statement made for a particular realization of $(\zeta_n)_{n=1}^{\infty}$ drawn from a subset $\Xi_0$ of $\Xi$ with $P(\Xi_0) = 1$.

12By "$\omega$-limit" we mean here that there exists an increasing sequence $t_n \uparrow \infty$ such that $x(t_n) = Q(y(t_n)) \to \hat{x}$. That $x(t)$ admits at least one $\omega$-limit stems from the compactness of $\mathcal{X}$.
Thus, with \( (v(x), x - x^*) \leq 0 \) for all \( x \in X \), \( x^* \in X^* \), Eq. (B.30) yields

\[
F(x^*, y(t_n + \delta)) - F(x^*, y(0)) \leq \sum_{j=1}^{n} \int_{t_j}^{t_{j+1}} \langle v(x(s)), x(s) - x^* \rangle \, ds \leq -nc\delta, \tag{B.32}
\]

showing that \( \liminf_{t \to \infty} F(x^*, y(t)) = -\infty \), a contradiction. Since \( x(t) \) admits at least one \( \omega \)-limit, we conclude that \( x(t) \) converges to \( X^* \).

**Step 4: Almost uniform convergence.** We now proceed to show that \( x(t) \) converges to \( X^* \) in an almost uniform way, as described in the main text; specifically, given \( \varepsilon > 0 \), we will show that there exists some \( T_{\varepsilon} > 0 \) such that \( F(x^*, \Phi_s(y)) \leq \max\{\varepsilon/2, F(x^*, y) - \varepsilon/2\} \) for all \( s \geq T_{\varepsilon} \) (cf. Eq. (4.2) in Section 4).

To do so, pick some \( \hat{\varepsilon} \) such that \( \mathbb{B}(x^*, \hat{\varepsilon}) \subseteq \mathbb{B}_F(x^*, \varepsilon/2) \). Then, invoking (B.30) and reasoning as in the second step of the proof of Theorem 3.4, there exists some positive constant \( c \equiv c(\varepsilon) \) such that

\[
\frac{d}{dt} F(x^*, \Phi_s(y)) = \langle v(x(t)), x(t) - x^* \rangle \leq -c \quad \text{if } x(t) \in X \setminus \mathbb{B}(x^*, \hat{\varepsilon}). \tag{B.33}
\]

With some hindsight, let \( T_{\varepsilon} = \varepsilon/(2c) \). Then, starting at \( y \) and following the flow \( \Phi \) of (4.1) for time \( s \geq T_{\varepsilon} \), there are two possibilities. The first is that \( x(s) \in \mathbb{B}_F(x^*, \varepsilon/2) \), in which case we have

\[
F(x^*, \Phi_s(y)) < \frac{\varepsilon}{2}. \tag{B.34}
\]

The second possibility is that \( x(s) \notin \mathbb{B}_F(x^*, \varepsilon/2) \). This implies that \( x(t) \notin \mathbb{B}(x^*, \hat{\varepsilon}) \) for all \( t \in [0, s] \) because, otherwise, we would have \( x(s_0) \in \mathbb{B}_F(x^*, \varepsilon/2) \) for some \( s_0 \leq s \); by the monotonicity property (B.30) of the Fenchel coupling, this would then imply that \( F(x^*, \Phi_s(y)) \leq F(x^*, \Phi_{s_0}(y)) < \varepsilon/2 \), contradicting the assumption \( x(s) \notin \mathbb{B}_F(x^*, \varepsilon/2) \). Thus, with \( x(t) \notin \mathbb{B}(x^*, \hat{\varepsilon}) \) for all \( t \in [0, s] \), we also have \( x(t) \notin \mathbb{B}(x^*, \hat{\varepsilon}) \) for all \( t \in [0, s] \), and hence, integrating (B.33), we obtain

\[
F(x^*, \Phi_s(y)) \leq F(x^*, y) - cs \leq F(x^*, y) - cT_{\varepsilon} \leq F(x^*, y) - \frac{\varepsilon}{2} \tag{B.35}
\]

Therefore, Eqs. (B.34) and (B.35) together establish that

\[
F(x^*, \Phi_s(y)) \leq \max\{\varepsilon/2, F(x^*, y) - \varepsilon/2\}, \tag{B.36}
\]

as claimed.

**Step 5: Bounding the discretization gap.** As in the proof of Theorem 3.4, let \( R = \sup_{x \in X} \|x\| \). Then, by the definition of the dual norm, we have

\[
\langle Y(t + h) - \Phi_h(Y(t + h)), Q(\Phi_h(Y(t + h))) - x^* \rangle \\
\leq \|Y(t + h) - \Phi_h(Y(t + h))\|_* \cdot \|Q(\Phi_h(Y(t + h))) - x^*\| \\
\leq R\|Y(t + h) - \Phi_h(t + h)\|_* \tag{B.37}
\]

Now, fix some \( T > 0 \), and, after some reverse engineering, take \( \delta = \sqrt{(KR)^2 - \varepsilon K - KR} \). From Step 3 we have

\[
\lim_{t \to \infty} \sup_{0 \leq h \leq T} \|Y(t + h), \Phi_h(Y(t))\|_* = 0, \tag{B.38}
\]

\[13\]That this is possible follows from Assumption 3; see also the proof of the last step of Theorem 3.4.
so there exists some (random) \( \tau \equiv \tau(\delta,T) \) such that \( \|Y(t+h) - \Phi_h(Y(t))\|_* < \delta \)
for all \( t \geq \tau \). Then, using Lemma 3.2 to expand the Fenchel coupling, we obtain
\[
F(x^*, Y(t+h)) = F(x^*, \Phi_h(Y(t))) + Y(t+h) - \Phi_h(Y(t)) \\
\leq F(x^*, \Phi_h(Y(t))) \\
+ \langle Y(t+h) - \Phi_h(Y(t)), Q(\Phi_h(Y(t))) \rangle - x^* \\
+ \frac{1}{2K}\|Y(t+h) - \Phi_h(Y(t))\|^2_2 \\
\leq F(x^*, \Phi_h(Y(t))) + R\delta + \frac{1}{2K}\delta^2 \\
\leq F(x^*, \Phi_h(Y(t))) + RK \left( \sqrt{R^2 + \frac{\varepsilon}{K}} - R \right) + \frac{K}{2} \left[ \sqrt{R^2 + \frac{\varepsilon}{K}} - R \right]^2 \\
= F(x^*, \Phi_h(Y(t))) + \frac{\varepsilon}{2},
\]
where the first inequality follows from Eq. (B.37) and the rest is highschool algebra.

**Step 6: Copy-paste ad infinitum.** We start by fixing an arbitrary \( \varepsilon > 0 \). By Step 5, there exists some \( T_\varepsilon > 0 \) such that Eq. (4.2) holds. Then, by taking \( T = T_\varepsilon \) in Step 6, there exists a \( \tau \) such that (4.3) holds for all \( t \geq \tau \). By Theorem 3.4, \( X_n \) visits \( \mathbb{B}_F(x^*, \delta) \) infinitely often, so there exists some \( \tau_0 \) such that \( Q(Y(\tau_0)) \in \mathbb{B}_F(x^*, \varepsilon/2) \) or, equivalently:
\[
F(x^*, Y(\tau_0)) < \frac{\varepsilon}{2}.
\]

Our goal is to establish that \( F(x^*, Y(\tau_0 + h)) < \varepsilon \) for all \( h \in [0, \infty) \). To that end, partition the \( [0, \infty) \) into disjoint time intervals of the form \(( (n-1)T_\varepsilon, nT_\varepsilon )\) for \( n \in \mathbb{N} \). Per Step 4, the monotonicity property of the Fenchel coupling implies that
\[
F(x^*, \Phi_h(Y(\tau_0))) \leq F(x^*, \Phi_0(Y(\tau_0))) = F(x^*, Y(\tau_0)) < \frac{\varepsilon}{2}
\]
for all \( h \geq 0 \), (B.41)
where the equality in the penultimate step follows from the semiflow properties of \( \Phi \). In this way, Eq. (4.3) yields:
\[
F(x^*, Y(\tau_0 + h)) < F(x^*, \Phi_h(Y(\tau_0))) + \frac{\varepsilon}{2} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,
\]
where the last inequality is a consequence of (B.41).

Now, assume inductively that Eq. (B.42) holds for all \( h \in [(n-1)T_\varepsilon, nT_\varepsilon) \) for some \( n \geq 1 \). Then, for all \( h \in [(n-1)T_\varepsilon, nT_\varepsilon) \), we have
\[
F(x^*, Y(\tau_0 + T_\varepsilon + h)) < F(x^*, \Phi_{T_\varepsilon}(Y(\tau_0 + h))) + \frac{\varepsilon}{2} \\
\leq \max \left\{ \frac{\varepsilon}{2}, F(x^*, Y(\tau_0 + h)) - \frac{\varepsilon}{2} \right\} + \frac{\varepsilon}{2} \\
\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,
\]
where the first inequality follows from (4.3), the second from (4.2), and the third from the induction hypothesis \( F(x^*, Y(\tau_0 + h)) < \varepsilon \). Consequently, Eq. (B.42) holds for all \( h \in [nT_\varepsilon, (n+1)T_\varepsilon) \). This completes the induction and our proof. \( \blacksquare \)

\(^{14}\)Note here that \( T_\varepsilon \) depends only on \( \varepsilon \) because \( T_\varepsilon \) is also defined in terms of \( \varepsilon \).
References


[23] A. Nedic and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted..."