SIAM J. OPTIM.
Vol. 30, No. 1, pp. 687–716

ON THE CONVERGENCE OF MIRROR DESCENT BEYOND STOCHASTIC CONVEX PROGRAMMING*

ZHENGYUAN ZHOU†, PANAYOTIS MERTIKOPOULOS‡, NICHOLAS BAMBOS§, STEPHEN P. BOYD§, AND PETER W. GLYNN§

Abstract. In this paper, we examine the convergence of mirror descent in a class of stochastic optimization problems that are not necessarily convex (or even quasi-convex) and which we call variationally coherent. Since the standard technique of "ergodic averaging" offers no tangible benefits beyond convex programming, we focus directly on the algorithm's last generated sample (its "last iterate"), and we show that it converges with probabiility 1 if the underlying problem is coherent. We further consider a localized version of variational coherence which ensures local convergence of Stochastic mirror descent (SMD) with high probability. These results contribute to the landscape of nonconvex stochastic optimization by showing that (quasi-)convexity is not essential for convergence to a global minimum: 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 problems with sharp minima (such as generic linear programs or concave minimization problems), SMD reaches a minimum point in a finite number of steps (a.s.), even in the presence of persistent gradient noise. This result is to be contrasted with existing black-box convergence rate estimates that are only asymptotic.

Key words. mirror descent, nonconvex programming, stochastic optimization, stochastic approximation, variational coherence

AMS subject classifications. 90C15, 90C26, 90C25, 90C05

DOI. 10.1137/17M1134925

1. Introduction.

Stochastic mirror descent (SMD) and its variants make up arguably one of the most widely used families of first-order methods in stochastic optimization—convex and nonconvex alike [3, 9, 10, 13, 24, 26, 27, 28, 29, 30, 31, 32, 34, 40, 45]. Heuristically, in the "dual averaging" (or "lazy") incarnation of the method [34, 40, 45, 48], SMD proceeds by aggregating a sequence of independent and identically distributed (i.i.d.) gradient samples and then mapping the result back to the problem's feasible region via a specially constructed "mirror map" (the namesake of the method). In so doing, SMD generalizes and extends the classical stochastic gradient descent (SGD) algorithm (with Euclidean projections playing the role of the mirror map) [33, 35, 36], the exponentiated gradient method of [22], the matrix regularization schemes of [21, 26, 42], and many others.

Starting with the seminal work of Nemirovski and Yudin [32], the convergence of mirror descent has been studied extensively in the context of convex program-

*Received by the editors June 16, 2017; accepted for publication (in revised form) September 27, 2019; published electronically February 27, 2020. Part of this work was presented at the 31st International Conference on Neural Information Processing Systems (NIPS 2017) [45].

https://doi.org/10.1137/17M1134925

Funding: The first author gratefully acknowledges the support of the IBM Goldstine fellowship. The second author was partially supported by French National Research Agency (ANR) grant ORACLESS (ANR-16-CE33-0004-01).

†IBM Research, Yorktown Heights, NY 10598, and Stern School of Business, New York University, New York, NY 10012 (zyzhou@stanford.edu).

‡Université Grenoble Alpes, CNRS, Grenoble INP, Inria, LIG, 38000 Grenoble, France (panayotis.mertikopoulos@imag.fr).

§Department of Electrical Engineering, Department of Management Science & Engineering, Stanford University, Stanford, CA 94305 (bambos@stanford.edu, boyd@stanford.edu, glynn@stanford.edu).

ming (including distributed and stochastic optimization problems) [3, 31, 34, 45], non-cooperative games/saddle-point problems [28, 31, 34, 47], and monotone variational inequality [20, 30, 34]. In this monotone setting, it is customary to consider the so-called ergodic average $\bar{X}_n = \sum_{k=1}^n \gamma_k X_k \big/ \sum_{k=1}^n \gamma_k$ of the algorithm's generated sample points X_n, with γ_n denoting the method's step-size. The reason for this is that, by Jensen's inequality, convexity guarantees that a regret-based analysis can lead to explicit convergence rates for $\bar{X}_n$ [31, 34, 40, 45]. However,(a) this type of averaging provides no tangible benefits in nonconvex programs; and (b) it is antagonistic to sparsity (which plays a major role in applications to signal processing, machine learning, and beyond). In view of this, we focus here directly on the properties of the algorithm's last generated sample—often referred to as its "last iterate."

The long-term behavior of the last iterate of SMD was recently studied by Shamir and Zhang [41] and Nedic and Lee [29] in the context of strongly convex problems. In this case, the algorithm's last iterate achieves the same value convergence rate as its ergodic average, so averaging is not more advantageous. Jiang and Xu [19] also examined the convergence of the last iterate of SGD in a class of (not necessarily monotone) variational inequalities that admit a unique solution, and they showed that it converges to said solution with probability 1. In [11], it was shown that in phase retrieval problems (a special class of nonconvex problems that involve systems of quadratic equations), SGD with random initialization converges to global optimal solutions with probability 1. Finally, in general nonconvex problems, Ghadimi and Lan [15, 16] showed that running SGD with a randomized stopping time guarantees convergence to a critical point in the mean, and they estimated the speed of this convergence. However, beyond these (mostly recent) results, not much is known about the convergence of the individual iterates of mirror descent in nonconvex programs.

Our contributions. In this paper, we examine the asymptotic behavior of mirror descent in a class of stochastic optimization problems that are not necessarily convex (or even quasi-convex). This class of problems, which we call *variationally coherent*, are related to a class of variational inequalities studied by Jiang and Xu [19] and, earlier, by Wang, Xiu, and Wang [44]—though, importantly, we do not assume here the existence of a unique solution. Focusing for concreteness on the *dual averaging* variant of SMD (also known as "lazy" mirror descent) [34, 40, 45], we show that the algorithm's last iterate converges to a global minimum with probability 1 under mild assumptions for the algorithm's gradient oracle (unbiased i.i.d. gradient samples that are bounded in L^2). This result can be seen as the "mirror image" of the analysis of [19] and reaffirms that (quasi-)convexity/monotonicity is not essential for convergence to a global optimum point: the weaker requirement of variational coherence suffices.

To extend the range of our analysis, we also consider a localized version of variational coherence which includes multimodal functions that are not even *locally* (quasi-)convex near their minimum points (so, in particular, an eigenvalue-based analysis cannot be readily applied to such problems). Here, in contrast to the globally coherent case, a single, "unlucky" gradient sample could drive the algorithm away from the "basin of attraction" of a local minimum (even a locally coherent one), possibly never to return. Nevertheless, we show that, with overwhelming probability, the last iterate of SMD converges locally to minimum points that are locally coherent (for a precise statement, see section 5).

Going beyond this "black-box" analysis, we also consider a class of optimization problems that admit *sharp* minima, a fundamental notion due to Polyak [35]. In stark contrast to existing ergodic convergence rates (which are asymptotic in nature), we

show that the last iterate of SMD 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 corollary, it follows that the last iterate of (lazy) SGD attains a solution of a stochastic linear program in a finite number of steps (a.s.). For completeness, we also derive a localized version of this result for problems with sharp local minima that are not globally coherent: in this case, convergence in a finite number of steps is retained but, instead of "almost surely," convergence now occurs with overwhelming probability.

Important classes of problems that admit sharp minima are generic linear programs (for the global case) and concave minimization problems (for the local case). In both instances, the (fairly surprising) fact that SMD attains a minimizer in a finite number of iterations should be contrasted to existing work on stochastic linear programming which exhibits asymptotic convergence rates [2, 43]. We find this result particularly appealing as it highlights an important benefit of working with "lazy" descent schemes: "greedy" methods (such as vanilla gradient descent) always take a gradient step from the last generated sample, so convergence in a finite number of iterations is a priori impossible in the presence of persistent noise. By contrast, the aggregation of gradient steps in "lazy" schemes means that even a "bad" gradient sample might not change the algorithm's sampling point (if the mirror map is surjective), so finite-time convergence *is* possible in this case.

Our analysis hinges on the construction of a primal-dual analogue of the Bregman divergence which we call the *Fenchel coupling* and which tracks the evolution of the algorithm's (dual) gradient aggregation variable relative to a target point in the problem's (primal) feasible region. This energy function allows us to perform a quasi-Fejérian analysis of stochastic mirror descent and, combined with a series of (sub)martingale convergence arguments, ultimately yields the convergence of the algorithm's last iterate—first as a subsequence, then with probability 1.

Notation. Given a finite-dimensional vector space $\mathcal{V}$ with norm $\|\cdot\|$, we 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 $\mathcal{V}^*$. If $\mathcal{C} \subseteq \mathcal{V}$ is convex, we also write $\operatorname{ri}(\mathcal{C})$ for the relative interior of $\mathcal{C}$, $\|\mathcal{C}\| = \sup\{\|x' - x\| : x, x' \in \mathcal{C}\}$ for its diameter, and $\operatorname{dist}(\mathcal{C}, x) = \inf_{x' \in \mathcal{C}} \|x' - x\|$ for the distance between $x \in \mathcal{V}$ and $\mathcal{C}$. For a given $x \in \mathcal{C}$, the *tangent cone* $\operatorname{TC}_{\mathcal{C}}(x)$ is defined as the closure of the set of all rays emanating from x and intersecting $\mathcal{C}$ in at least one other point; dually, the *polar cone* $\operatorname{PC}_{\mathcal{C}}(x)$ to $\mathcal{C}$ at x is defined as $\operatorname{PC}_{\mathcal{C}}(x) = \{y \in \mathcal{V}^* : \langle y, z\rangle \leq 0 \text{ for all } z \in \operatorname{TC}_{\mathcal{C}}(x)\}$. For concision, we will write $\operatorname{TC}(x)$ and $\operatorname{PC}(x)$ instead when $\mathcal{C}$ is clear from the context.

2. Problem setup and basic definitions.

2.1. The main problem.

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

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

where

$$f(x) = \mathbb{E}[F(x;\omega)] \tag{2.1}$$

for some stochastic objective function $F\colon \mathcal{X} \times \Omega \to \mathbb{R}$ defined on an underlying (complete) probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In terms of regularity, our blanket assumptions for (Opt) will be as follows.

Assumption 1. $F(x,\omega)$ is continuously differentiable in x for almost all $\omega \in \Omega$.

Assumption 2. The gradient of F is uniformly bounded in L^2, i.e., $\mathbb{E}[\|\nabla F(x;\omega)\|_*^2] \leq M^2$ for some finite $M \geq 0$ and all $x \in \mathcal{X}$.

Remark 2.1. In the above, gradients are treated as elements of the dual space $\mathcal{Y} \equiv \mathcal{V}^*$ of $\mathcal{V}$. We also note that $\nabla F(x;\omega)$ refers to the gradient of $F(x;\omega)$ with respect to x; since Ω is not assumed to carry a differential structure, there is no danger of confusion.

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 F is uniformly Lipschitz (another commonly used condition) and, by the dominated convergence theorem, it further implies that f is smooth and $\nabla f(x) = \nabla \mathbb{E}[F(x;\omega)] = \mathbb{E}[\nabla F(x;\omega)]$ is bounded. As a result, the solution set

$$\mathcal{X}^* = \arg\min f \tag{2.2}$$

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

We briefly discuss below two important examples of (Opt).

Example 2.1 (distributed optimization). An important special case of (Opt) with high relevance to statistical inference, signal processing, and machine learning is when f is of the special form

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{2.3}$$

for some family of functions (or training samples) $f_i \colon \mathcal{X} \to \mathbb{R}$, $i = 1,\dots,N$. As an example, this setup corresponds to empirical risk minimization with uniform weights, the sample index i being drawn with uniform probability from $\{1,\dots,N\}$.

Example 2.2 (noisy gradient measurements). Another widely studied instance of (Opt) is when

$$F(x;U) = f(x) + \langle U, x\rangle \tag{2.4}$$

for some random vector U such that $\mathbb{E}[U] = 0$ and $\mathbb{E}[\|U\|_*^2] < \infty$. This gives $\nabla F(x;U) = \nabla f(x) + U$, so (Opt) can be seen here as a model for deterministic optimization problems with noisy gradient measurements.

2.2. Variational coherence. We are now in a position to define the class of variationally coherent problems.

Definition 2.1. *We say that* (Opt) *is* variationally coherent *if*

$$\langle \nabla f(x), x - x^*\rangle \geq 0 \quad \textit{for all } x \in \mathcal{X},\ x^* \in \mathcal{X}^*, \tag{VC}$$

and there exists some $x^* \in \mathcal{X}^*$ *such that equality holds in* (VC) *only if* $x \in \mathcal{X}^*$.

In words, (VC) states that solutions of (Opt) can be harvested by solving a (Minty) variational inequality—hence the term "variational coherence." To the best of our knowledge, the closest analogue to this condition first appeared in the classical

paper of Bottou [5] on online learning and stochastic approximation algorithms, but with the added assumptions that (a) the problem (Opt) admits a unique solution x^* and (b) an extra positivity requirement for $\langle \nabla f(x), x - x^* \rangle$ in punctured neighborhoods of x^*. In the context of variational inequalities, a closely related variant of (VC) has been used to establish the convergence of extragradient methods [14, 44] and SGD [19] in (Stampacchia) variational inequalities with a unique solution. By contrast, there is no uniqueness requirement in (VC), an aspect of the definition which we examine in more detail below.

We should also note that, as stated, (VC) is a nonrandom requirement for f so it applies equally well to *deterministic* optimization problems. Alternatively, by the dominated convergence theorem, (VC) can be written equivalently as

$$\mathbb{E}[\langle \nabla F(x;\omega), x - x^* \rangle] \geq 0, \tag{2.5}$$

so it can be interpreted as saying that F is variationally coherent "on average," without any individual realization thereof satisfying (VC). Both interpretations will come in handy later on.

All in all, the notion of variational coherence will play a central role in our paper so a few examples are in order.

Example 2.3 (convex programming). If f is convex, ∇f is *monotone* [38] in the sense that

$$\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \geq 0 \quad \text{for all } x, x' \in \mathcal{X}. \tag{2.6}$$

By the first-order optimality conditions for f, it follows that $\langle f(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$. Hence, by monotonicity, we get

$$\langle \nabla f(x), x - x^* \rangle \geq \langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \text{for all } x \in \mathcal{X},\ x^* \in \mathcal{X}^*. \tag{2.7}$$

By convexity, it further follows that $\langle \nabla f(x), x - x^* \rangle < 0$ whenever $x^* \in \mathcal{X}^*$ and $x \in \mathcal{X} \setminus \mathcal{X}^*$, so equality holds in (2.7) if and only if $x \in \mathcal{X}^*$. This shows that convex programs automatically satisfy (VC).

Example 2.4 (quasi-convex problems). More generally, the above analysis also extends to *quasi-convex* objectives, i.e., when

$$f(x') \leq f(x) \implies \langle \nabla f(x), x' - x \rangle \leq 0 \tag{QC}$$

for all $x, x' \in \mathcal{X}$ [6]. In this case, we have the following.

PROPOSITION 2.2. *Suppose that f is quasi-convex and nondegenerate, i.e.,*

$$\langle \nabla f(x), z \rangle \neq 0 \quad \textit{for all nonzero } z \in \mathrm{TC}(x),\ x \in \mathcal{X} \setminus \mathcal{X}^*. \tag{2.8}$$

Then, f is variationally coherent.

Remark 2.2. The nondegeneracy condition (2.8) is *generic* in that it is satisfied by every quasi-convex function after an arbitrarily small perturbation leaving its minimum set unchanged. In particular, it is automatically satisfied if f is convex or pseudoconvex.

Proof. Take some $x^* \in \mathcal{X}^*$ and $x \in \mathcal{X}$. Then, letting $x' = x^*$ in (QC), we readily obtain $\langle \nabla f(x), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$, $x^* \in \mathcal{X}^*$. Furthermore, if $x \notin \mathcal{X}^*$ but $\langle \nabla f(x), x - x^* \rangle = 0$, the gradient nondegeneracy condition (2.8) would be violated, implying in turn that, for any $x^* \in \mathcal{X}^*$, we have $\langle \nabla f(x), x - x^* \rangle = 0$ only if $x \in \mathcal{X}^*$. This shows that f satisfies (VC). ∎

Example 2.5 (beyond quasi-convexity). A simple example of a function that is variationally coherent without even being quasi-convex is

$$f(x) = 2\sum_{i=1}^{d} \sqrt{1+x_i}, \quad x \in [0,1]^d. \tag{2.9}$$

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

Example 2.6 (a weaker version of coherence). Consider the function

$$f(x) = \frac{1}{2}\prod_{i=1}^{d} x_i^2, \quad x \in [-1,1]^d. \tag{2.10}$$

By inspection, it is easy to see that the minimum set of f is $\mathcal{X}^* = \{x^* \in [-1,1]^d : x_i^* = 0 \text{ for at least one } i = 1,\dots,d\}$.[1] Since $\mathcal{X}^*$ is not convex for $d \geq 2$, f is not quasi-convex. On the other hand, we have $\nabla f(x) = 2f(x) \cdot (1/x_1, \dots, 1/x_d)$, so $\langle \nabla f(x), x - 0\rangle \geq 0$ for all $x \in [-1,1]^d$ with equality only if $x \in \mathcal{X}^*$. Moreover, for any $x^* \in \mathcal{X}^*$ and all $x \in \mathcal{X}$ sufficiently close to x^*, we have

$$\langle \nabla f(x), x - x^*\rangle = 2f(x)\sum_{i=1}^{d}\left[1 - \frac{x_i^*}{x_i}\right] = 2f(x)\left[d - \sum_{i:x_i^*\neq 0}\frac{x_i^*}{x_i}\right] \geq 0. \tag{2.11}$$

We thus conclude that f satisfies the following weaker version of (VC).

DEFINITION 2.3. *We say that $f\colon \mathcal{X} \to \mathbb{R}$ is* weakly coherent *if the following conditions are satisfied:*

(a) *There exists some $p \in \mathcal{X}^*$ such that $\langle \nabla f(x), x - p\rangle \geq 0$ with equality only if $x \in \mathcal{X}^*$.*
(b) *For all $x^* \in \mathcal{X}^*$, $\langle \nabla f(x), x - x^*\rangle \geq 0$ whenever x is close enough to x^*.*

Our analysis also applies to problems satisfying these less stringent requirements, in which case the minimum set $\mathcal{X}^* = \arg\min f$ of f need not even be convex.[2] For simplicity, we will first work with Definition 2.1 and relegate these considerations to section 5.

2.3. Stochastic mirror descent. To solve (Opt), we will focus on the SMD family of algorithms, a class of first-order methods pioneered by Nemirovski and Yudin [32] and studied further by Beck and Teboulle [3], Nesterov [34], Lan, Nemirovski, and Shapiro [24], and many others. Referring to [8, 40] for an overview, the specific variant of SMD that we consider here is usually referred to as *dual averaging* [28, 34, 45] or *lazy* mirror descent [40].

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 F at the algorithm's current state. Subsequently, the method takes a step along this stochastic gradient in the dual space

[1] Linear combinations of functions of this type play an important role in training deep learning models—and, in particular, generative adversarial network [17].

[2] Obviously, Definitions 2.1 and 2.3 coincide if $\mathcal{X}^*$ is a singleton. This highlights the intricacies that arise in problems that do not admit a unique solution.

$\mathcal{Y} \equiv \mathcal{V}^*$ of $\mathcal{V}$ (where gradients live), the result is "mirrored" back to the problem's feasible region $\mathcal{X}$, and the process repeats. Formally, this gives rise to the recursion

$$\text{(SMD)} \qquad \begin{aligned} X_n &= Q(Y_n), \\ Y_{n+1} &= Y_n - \gamma_n \nabla F(X_n; \omega_n), \end{aligned}$$

where

1. $n = 1, 2, \dots$ denotes the algorithm's running counter,
2. $Y_n \in \mathcal{Y}$ is a score variable that aggregates gradient steps up to stage n,
3. $Q\colon \mathcal{Y} \to \mathcal{X}$ is the *mirror map* that outputs a solution candidate $X_n \in \mathcal{X}$ as a function of the score variable $Y_n \in \mathcal{V}^*$,
4. $\omega_n \in \Omega$ is a sequence of i.i.d. samples,[3]
5. $\gamma_n > 0$ is the algorithm's step-size sequence, assumed in what follows to satisfy the Robbins–Monro summability condition

$$\sum_{n=1}^{\infty} \gamma_n^2 < \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma_n = \infty. \tag{2.12}$$

For a schematic illustration and a pseudocode implementation of (SMD), see Figure 1 and Algorithm 1, respectively.

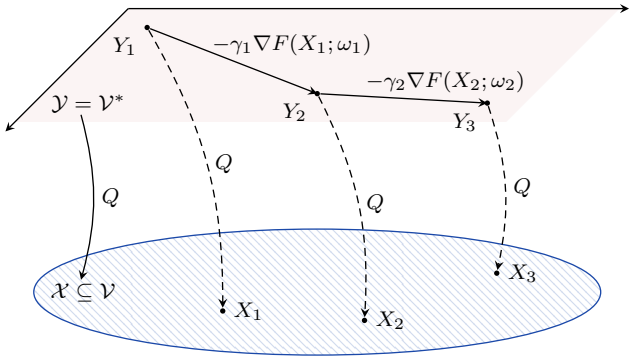


FIG. 1. *Schematic representation of SMD (Algorithm 1).*

Algorithm 1 Stochastic mirror descent.

```
Require: mirror map Q: 𝒴 → 𝒳; step-size sequence γ_n > 0
choose Y ∈ 𝒴 ≡ 𝒱*                                  # initialization
for n = 1, 2, ... do
    set X ← Q(Y)                                    # set state
    draw ω ∈ Ω                                      # gradient sample
    get v̂ = −∇F(X; ω)                               # get oracle feedback
    set Y ← Y + γ_n v̂                               # update score variable
end for
return X                                            # output
```

[3]The indexing convention for ω_n means that Y_n and X_n are *predictable* relative to the natural filtration $\mathcal{F}_n = \sigma(\omega_1, \dots, \omega_n)$ of ω_n, i.e., Y_{n+1} and X_{n+1} are both $\mathcal{F}_n$-measurable. To this history, we also attach the trivial σ-algebra as $\mathcal{F}_0$ for completeness.

In more detail, the algorithm's mirror map $Q\colon \mathcal{Y} \to \mathcal{X}$ is defined as

$$Q(y) = \arg\max_{x\in\mathcal{X}}\{\langle y, x\rangle - h(x)\}, \tag{2.13}$$

where the *regularizer* (or *penalty function*) $h\colon \mathcal{X} \to \mathbb{R}$ is assumed to be continuous and strongly convex on $\mathcal{X}$, i.e., there exists some $K > 0$ such that

$$h(\tau x + (1-\tau)x') \leq \tau h(x) + (1-\tau)h(x') - \tfrac{1}{2}K\tau(1-\tau)\|x' - x\|^2 \tag{2.14}$$

for all $x, x' \in \mathcal{X}$ and all $\tau \in [0,1]$. The mapping $Q\colon \mathcal{V}^* \to \mathcal{X}$ defined by (2.13) is then called the *mirror map induced by* h. For concreteness, we present below some well-known examples of regularizers and mirror maps.

Example 2.7 (Euclidean regularization). Let $h(x) = \frac{1}{2}\|x\|_2^2$. Then, h is 1-strongly convex with respect to the Euclidean norm $\|\cdot\|_2$, and the induced mirror map is the closest point projection

$$\Pi(y) = \arg\max_{x\in\mathcal{X}}\big\{\langle y, x\rangle - \tfrac{1}{2}\|x\|_2^2\big\} = \arg\min_{x\in\mathcal{X}} \|y - x\|_2^2. \tag{2.15}$$

The resulting descent algorithm is known in the literature as (lazy) SGD and we study it in detail in section 6. 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 2.8 (entropic regularization). Let $\Delta = \{x \in \mathbb{R}^d_+ : \sum_{i=1}^d x_i = 1\}$ denote the unit simplex of $\mathbb{R}^d$. A widely used regularizer in this setting is the (negative) Gibbs entropy $h(x) = \sum_{i=1}^d x_i \log x_i$: this regularizer is 1-strongly convex with respect to the L^1-norm and a straightforward calculation shows that the induced mirror map is

$$\Lambda(y) = \frac{1}{\sum_{i=1}^d \exp(y_i)}(\exp(y_1), \dots, \exp(y_d)). \tag{2.16}$$

This example is known as *entropic regularization* and the resulting mirror descent algorithm has been studied extensively in the context of linear programming, online learning, and game theory [1, 40]. For posterity, we also note that h is differentiable *only* on the relative interior Δ° of Δ and $\operatorname{im}\Lambda = \Delta^\circ$ (i.e., Λ is "essentially" surjective).

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

- *Global convergence:* If (Opt) is variationally coherent, the last iterate X_n of (SMD) converges to a global minimizer of f with probability 1.
- *Local convergence:* If x^* is a *locally coherent* minimum point of f (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 f (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. Main tools and first results. As a stepping stone to analyze the long-term behavior of (SMD), we derive in this section a recurrence result which is interesting in its own right. Specifically, we show that if (Opt) is coherent, then, with probability 1, X_n visits any neighborhood of $\mathcal{X}^*$ infinitely often; as a corollary, there exists a (random) subsequence X_{n_k} of X_n that converges to $\arg\min f$ (a.s.). In what follows, our goal will be to state this result formally and to introduce the analytic machinery used for its proof (as well as the analysis of the subsequent sections).

3.1. The Fenchel coupling. The first key ingredient of our analysis will be the *Fenchel coupling*, a primal-dual variant of the Bregman divergence [7] that plays the role of an energy function for (SMD).

DEFINITION 3.1. *Let $h\colon \mathcal{X} \to \mathbb{R}$ 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*

$$F(p,y) = h(p) + h^*(y) - \langle y, p\rangle, \tag{3.1}$$

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 quantifies some basic properties of this coupling.

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*

$$\text{(a)}\quad F(p,y) \geq \frac{K}{2}\|Q(y) - p\|^2, \tag{3.2a}$$

$$\text{(b)}\quad F(p,y') \leq F(p,y) + \langle y' - y, Q(y) - p\rangle + \frac{1}{2K}\|y' - y\|_*^2. \tag{3.2b}$$

Lemma 3.2 (which we prove in Appendix B) shows that $Q(y_n) \to p$ whenever $F(p,y_n) \to 0$, so the Fenchel coupling can be used to test the convergence of the primal sequence $x_n = Q(y_n)$ to a given base point $p \in \mathcal{X}$. For technical reasons, it will be convenient to also make the converse assumption.

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). In this way, Assumption 3 can be seen as a primal-dual analogue of the reciprocity conditions for the Bregman divergence that are widely used in the literature on proximal and forward-backward methods [9, 23]. Most common regularizers satisfy this technical requirement (including the Euclidean and entropic regularizers of Examples 2.7 and 2.8, respectively).

3.2. Main recurrence result. To state our recurrence result, we require one last piece of notation pertaining to measuring distances in $\mathcal{X}$.

DEFINITION 3.3. *Let $\mathcal{C}$ be a subset of $\mathcal{X}$.*

1. *The* distance *between $\mathcal{C}$ and $x \in \mathcal{X}$ is defined as* $\operatorname{dist}(\mathcal{C},x) = \inf_{x'\in\mathcal{C}}\|x' - x\|$, *and the corresponding* ε-neighborhood of $\mathcal{C}$ *is*

$$\mathbb{B}(\mathcal{C},\varepsilon) = \{x \in \mathcal{X} : \operatorname{dist}(\mathcal{C},x) < \varepsilon\}. \tag{3.3a}$$

2. *The (setwise)* Fenchel coupling *between $\mathcal{C}$ and $y \in \mathcal{Y}$ is defined as $F(\mathcal{C},y) = \inf_{x\in\mathcal{C}} F(x,y)$, and the corresponding* Fenchel δ-zone *of $\mathcal{C}$ under h is*

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

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

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

COROLLARY 3.5. *With probability* 1*, there exists a subsequence* X_{n_k} *of* X_n *converging to a* (*random*) *minimum point* x^* *of* (Opt).

The proof of Proposition 3.4 consists of three main steps, which we outline below.

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

$$v(x) = -\mathbb{E}[\nabla F(x;\omega)] = -\nabla f(x) \tag{3.4}$$

denote the negative gradient of f at $x \in \mathcal{X}$, and write

$$\hat{v}_n = -\nabla F(X_n;\omega_n) \tag{3.5}$$

for the corresponding oracle feedback at stage n. Then, Algorithm 1 may be written in Robbins–Monro form as

$$Y_{n+1} = Y_n + \gamma_n \hat{v}_n = Y_n + \gamma_n[v(X_n) + U_n], \tag{3.6}$$

where

$$U_n = \nabla f(X_n) - \nabla F(X_n;\omega_n) \tag{3.7}$$

denotes the difference between the mean gradient of f at X_n and the nth stage gradient sample.[4] By construction, U_n is a martingale difference sequence relative to the history (natural filtration) $\mathcal{F}_n = \sigma(\omega_1, \dots, \omega_n)$ of ω_n, i.e.,

$$\mathbb{E}[U_n \,|\, \mathcal{F}_{n-1}] = 0 \quad \text{for all } n. \tag{3.8a}$$

Furthermore, by Assumption 2, it readily follows that U_n has uniformly bounded second moments, i.e., there exists some finite $\sigma \geq 0$ such that

$$\mathbb{E}[\|U_n\|_*^2 \,|\, \mathcal{F}_{n-1}] \leq \sigma^2 \quad \text{for all } n, \tag{3.8b}$$

implying in turn that U_n is bounded in L^2 (for a more detailed treatment, see Appendix B).

Step 2*: Recurrence of* ε*-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 $\mathbb{B}(\mathcal{X}^*, \varepsilon)$ for sufficiently large n, we must also have $F(\mathcal{X}^*, Y_n) \to -\infty$ (a.s.). This contradicts the nonnegativity of F, so X_n must enter $\mathbb{B}(\mathcal{X}^*, \varepsilon)$ infinitely often (a.s.).

Step 3*: Recurrence of Fenchel zones.* By reciprocity (Assumption 3), $\mathbb{B}_F(\mathcal{X}^*, \delta)$ always contains an ε-neighborhood of $\mathcal{X}^*$. Since X_n enters $\mathbb{B}(\mathcal{X}^*, \varepsilon)$ infinitely many times (a.s.), the same must hold for $\mathbb{B}_F(\mathcal{X}^*, \delta)$. Our claim and Corollary 3.5 then follow immediately.

4. Global convergence under coherence. 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*,* X_n *converges with probability* 1 *to a* (*possibly random*) *minimum point of* (Opt).

[4] Recall here that there is a one-step offset between X_n and ω_{n+1} at the nth iteration of SMD.

COROLLARY 4.2. *If f is a nondegenerate quasi-convex (or pseudoconvex, or convex) function and Assumptions* 1–3 *hold, the last iterate of* (SMD) *converges with probability* 1 *to a (possibly random) minimum point of* (Opt).

Before discussing the proof of Theorem 4.1, it is important to note that most of the literature surrounding (SMD) and its variants (see, e.g., [13, 31, 34, 45] and references therein) focuses on the so-called ergodic average of X_n, i.e.,

$$\bar{X}_n = \frac{\sum_{k=1}^n \gamma_k X_k}{\sum_{k=1}^n \gamma_k}. \tag{4.1}$$

Despite the appealing "self-averaging" properties of $\bar{X}_n$ in convex problems [31, 34], it is not clear how to extend the standard tools used to establish convergence of $\bar{X}_n$ beyond convex/monotone problems (even to pseudoconvex programs). Since convergence of X_n automatically implies that of $\bar{X}_n$, Theorem 4.1 simultaneously establishes the convergence of the last iterate of SMD and extends existing ergodic convergence results to a wider class of nonconvex stochastic programs.

Corollary 4.2 also extends the corresponding results of [29] for the convergence of the last iterate of (SMD) when f is (strongly) convex and h has Lipschitz-continuous gradients (so the induced Bregman divergence can be bounded from above by a quadratic surrogate of the primal norm). Our proof strategy is similar and relies on the following lemma, often attributed to Gladyshev [35, p. 49].[5]

LEMMA 4.3 (Gladyshev). *Let a_n, $n = 1, 2, \dots$, be a sequence of nonnegative random variables such that*

$$\mathbb{E}[a_{n+1} \,|\, a_1, \dots, a_n] \le (1 + \delta_n) a_n + \varepsilon_n, \tag{4.2}$$

where δ_n and ε_n are nonnegative deterministic sequences with

$$\sum_{n=1}^{\infty} \delta_n < \infty \quad \textit{and} \quad \sum_{n=1}^{\infty} \varepsilon_n < \infty. \tag{4.3}$$

Then, a_n converges (a.s.) to some random variable $a_\infty \ge 0$.

As shown below, this "quasi-Fejér" monotonicity property plays a critical part in establishing the convergence of (SMD).

Proof of Theorem 4.1. Let $x^* \in \mathcal{X}^*$ be a minimum point of (Opt). Then, letting $F_n = F(x^*, Y_n)$, Lemma 3.2 gives

$$\begin{aligned} F_{n+1} = F(x^*, Y_{n+1}) &= F(x^*, Y_n + \gamma_n \hat{v}_n) \\ &\le F(x^*, Y_n) + \gamma_n \langle \hat{v}_n, X_n - x^* \rangle + \frac{\gamma_n^2}{2K} \|\hat{v}_n\|_*^2 \\ &= F_n + \gamma_n \langle v(X_n), X_n - x^* \rangle + \gamma_n \xi_n + \frac{\gamma_n^2}{2K} \|\hat{v}_n\|_*^2 \\ &\le F_n + \gamma_n \xi_n + \frac{\gamma_n^2}{2K} \|\hat{v}_n\|_*^2, \end{aligned} \tag{4.4}$$

where we set $\xi_n = \langle U_n, X_n - x^* \rangle$ in the third line and used the fact that f satisfies (VC) in the last one. Since Y_n is predictable relative to $\mathcal{F}_n$ (i.e., Y_n is $\mathcal{F}_{n-1}$-measurable), the process $F_n = F(x^*, Y_n)$ is itself adapted to the shifted filtration

[5] We thank an anonymous reviewer for suggesting this approach. Our original proof strategy relied on the so-called ODE method of stochastic approximation [4] and was considerably more intricate.

$\mathcal{F}'_n = \sigma(\omega_1, Y_2 \dots, \omega_{n-1}, Y_n) = \mathcal{F}_{n-1}$. Thus, taking conditional expectations and invoking Assumption 2, the bound (4.4) becomes

$$\begin{aligned}\mathbb{E}[F_{n+1} \,|\, \mathcal{F}'_n] &\leq F_n + \mathbb{E}[\xi_n \,|\, \mathcal{F}'_n] + \frac{\gamma_n^2}{2K}\,\mathbb{E}[\|\hat{v}_n\|_*^2 \,|\, \mathcal{F}'_n] \\ &= F_n + \mathbb{E}[\xi_n \,|\, \mathcal{F}_{n-1}] + \frac{\gamma_n^2}{2K}\,\mathbb{E}[\|\nabla F(X_n;\omega_n)\|_*^2 \,|\, \mathcal{F}_{n-1}] \\ &\leq F_n + \frac{\gamma_n^2 M^2}{2K}, \qquad (4.5)\end{aligned}$$

where, in the last line, we used Assumption 2 and the fact that U_n is a martingale difference sequence (so $\mathbb{E}[\xi_n \,|\, \mathcal{F}_{n-1}] = 0$; for a detailed derivation, see the proof of Proposition 3.4 in Appendix B). Hence, with $\sum_{n=1}^{\infty} \gamma_n^2 < \infty$, Lemma 4.3 implies that F_n converges (a.s.) to some finite limit F_∞.

Now, by Proposition 3.4, there exists (a.s.) a subsequence Y_{n_k} of Y_n and some (possibly random) $x^* \in \mathcal{X}^*$ such that $\lim_{k\to\infty} F(x^*, Y_{n_k}) = 0$. Since the limit $\lim_{n\to\infty} F(x^*, Y_n)$ exists (a.s.), it follows that $\lim_{n\to\infty} F(x^*, Y_n) = 0$. This shows that, with probability 1, $X_n = Q(Y_n)$ converges to some (random) minimum point x^* of (Opt), as claimed. ∎

In closing this section, we should note that the conclusion of Theorem 4.1 also applies to problems that are "almost" coherent in the sense of Example 2.6, i.e.,

(a) there exists a minimizer $p \in \mathcal{X}^*$ such that $\langle \nabla f(x), x - p\rangle \geq 0$ with equality only if $x \in \mathcal{X}^*$;
(b) for all $x^* \in \mathcal{X}^*$, $\langle \nabla f(x), x - x^*\rangle \geq 0$ whenever x is close enough to x^*.

Proving this more general result requires some of the machinery presented in the following section, so we relegate its discussion until all the requisite tools are in place.

5. Convergence under local/weak coherence. In this section, our goal is to extend the convergence analysis of the previous section to account for optimization problems that are only "locally" coherent. Building on Definition 2.1, these are defined as follows.

DEFINITION 5.1. *Let $\mathcal{C}$ be a closed set of local minimizers of f, viz. $f(x) \geq f(x^*)$ 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 U of $\mathcal{C}$ such that*

$$\langle \nabla f(x), x - x^*\rangle \geq 0 \quad \textit{for all } x \in U,\ x^* \in \mathcal{C}, \qquad \text{(LVC)}$$

and there exists some $x^* \in \mathcal{C}$ such that equality holds in (LVC) *only if $x \in \mathcal{C}$.*

An immediate consequence of Definition 5.1 is that locally coherent sets are isolated components of local minimizers of f. To see this, if $\mathcal{C}$, U, and x^* are as in Definition 5.1 and $x \in U$ is a local minimizer of f, we would have $\langle \nabla f(x), z\rangle \geq 0$ for all tangent $z \in \mathrm{TC}(x)$. Applying this to $z = x^* - x$ gives $\langle \nabla f(x), x - x^*\rangle \geq 0$, so, by the definition of local coherence, we conclude that $x \in \mathcal{C}$.

We also note that although the minimum set of a globally coherent problem is *a fortiori* locally coherent, the converse need not hold. A concrete example of a function which is not globally coherent but which admits a locally coherent minimum is the Rosenbrock test function

$$f(x) = \sum_{i=1}^{d} [100(x_{i+1} - x_i)^2 + (1 - x_i^2)], \quad x \in [-2, 2]^d, \qquad (5.1)$$

which has seen extensive use in the literature as a nonconvex convergence speed benchmark (cf. section 7).[6] From this example, we see that the profile of f around a locally coherent set could be highly nonconvex, possibly including a wide variety of valleys, talwegs, and ridges; in fact, even *quasi*-convexity may fail to hold locally.

Now, in contrast to globally coherent optimization problems, an "unlucky" gradient sample could drive (SMD) out of the "basin of attraction" of a locally coherent set (the largest neighborhood U for which (LVC) holds), possibly never to return. For this reason, instead of focusing on global convergence results with probability one, we will focus on local convergence with high probability. Our main result along these lines is as follows.

THEOREM 5.2 (local convergence with high probability). *Let $\mathcal{C}$ be locally coherent for* (Opt) *and fix some confidence level $\delta > 0$. Then, under Assumptions* 1–3*, there exists an open neighborhood $\mathcal{U}$ of $\mathcal{C}$, independent of δ, such that*

$$\mathbb{P}(X_n \text{ converges to } \mathcal{C} \,|\, X_1 \in \mathcal{U}) \geq 1 - \delta, \tag{5.2}$$

provided that the algorithm's step-size sequence γ_n is small enough.

Remark 5.1. As a concrete application 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 $\gamma > 0$ such that if Algorithm 1 is run with step-size $\gamma_n = \gamma/n^\beta$, (5.2) holds. We emphasize the interesting point here: the open neighborhood $\mathcal{U}$ is fixed once and for all and does not depend on the probability threshold δ. That is, to get convergence with higher probability, it is *not* necessary to assume that X_1 starts closer to $\mathcal{C}$: one need only use a smaller step-size sequence satisfying (2.12).

The key idea behind the proof of Theorem 5.2 is as follows. First, it suffices to consider the case where $\mathcal{C}$ consists of a single local minimizer x^*; the argument for the general case follows the same techniques as in section 4. Then, 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 $\mathcal{X}$ 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 $\mathcal{O}(\gamma_n)$ martingale term and a second-order $\mathcal{O}(\gamma_n^2)$ submartingale perturbation. Using Doob's maximal inequality, we then show that the aggregation of both errors remains controllably small with probability at least $1 - \delta$.

We formalize all this below.

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

Step 1*: Controlling the martingale error.* Fix some $\varepsilon > 0$. As in the proof of Theorem 4.1, let $U_n = \nabla f(X_n) - \nabla F(X_n; \omega_n)$ and set $\xi_n = \langle U_n, X_n - x^* \rangle$, where $x^* \in \mathcal{C}$ is such that (LVC) holds as an equality only if $x \in \mathcal{C}$ (cf. Definition 5.1). We show below that there exists a step-size sequence $(\gamma_n)_{n=1}^\infty$ such that

$$\mathbb{P}\left(\sup_n \sum_{k=1}^n \gamma_k \xi_k \leq \varepsilon \right) \geq 1 - \frac{\delta}{2}. \tag{5.3}$$

To show this, we start by noting that, as in the proof of Proposition 3.4, the aggregate process $S_n = \sum_{k=1}^n \gamma_k \xi_k$ is a martingale relative to the natural filtration $\mathcal{F}_n$ of ω_n.

[6] Local coherence can be proved by a straightforward algebraic calculation (omitted for concision).

Then, letting $R = \sup_{x\in\mathcal{X}}\|x\|$, we can bound the variance of each individual term of S_n as follows:

$$\begin{aligned}\mathbb{E}[\xi_k^2] &= \mathbb{E}[\mathbb{E}[|\langle U_k, X_k - x^*\rangle|^2 \,|\, \mathcal{F}_{k-1}]] \\ &\leq \mathbb{E}[\mathbb{E}[\|U_k\|_*^2 \|X_k - x^*\|^2 \,|\, \mathcal{F}_{k-1}]] \\ &= \mathbb{E}[\|X_k - x^*\|^2 \,\mathbb{E}[\|U_k\|_*^2 \,|\, \mathcal{F}_{k-1}]] \\ &\leq R^2 M^2, \end{aligned} \tag{5.4}$$

where the first inequality follows from the definition of the dual norm and the second one follows from (3.8b). Consequently, by Doob's maximal inequality (Theorem A.4 in Appendix A), we have

$$\mathbb{P}\biggl(\sup_{0\leq k\leq n} S_k \geq \varepsilon\biggr) \leq \mathbb{P}\biggl(\sup_{0\leq k\leq n} |S_k| \geq \varepsilon\biggr) \leq \frac{\mathbb{E}[S_n^2]}{\varepsilon^2} \leq \frac{R^2M^2\sum_{k=1}^n \gamma_k^2}{\varepsilon^2}, \tag{5.5}$$

where the last inequality follows from expanding $\mathbb{E}[|S_n|^2]$, using (5.4), and noting that $\mathbb{E}[\xi_k\xi_\ell] = \mathbb{E}[\mathbb{E}[\xi_k\xi_\ell]\,|\,\mathcal{F}_{k\vee\ell-1}] = 0$ whenever $k \neq \ell$. Therefore, by picking γ_n so that $\sum_{k=1}^\infty \gamma_k^2 \leq \varepsilon^2\delta/(2R^2M^2)$, (5.5) gives

$$\mathbb{P}\biggl(\sup_{0\leq k\leq t} S_k \geq \varepsilon\biggr) \leq \frac{R^2M^2\sum_{k=1}^n \gamma_k^2}{\varepsilon^2} \leq \frac{R^2M^2\sum_{k=1}^\infty \gamma_k^2}{\varepsilon^2} \leq \frac{\delta}{2} \quad \text{for all } n. \tag{5.6}$$

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 foresight, let $R_n = (2K)^{-1}\sum_{k=1}^n \gamma_k^2\|\hat{v}_k\|_*^2$. By construction, R_n is a nonnegative submartingale relative to $\mathcal{F}_n$. We again establish that there exists step-size sequence $(\gamma_n)_{n=1}^\infty$ satisfying the summability condition (2.12) and such that

$$\mathbb{P}\biggl(\sup_n R_n \leq \varepsilon\biggr) \geq 1 - \frac{\delta}{2}. \tag{5.7}$$

To show this, Doob's maximal inequality for submartingales (Theorem A.3) gives

$$\mathbb{P}\biggl(\sup_{0\leq k\leq n} R_k \geq \varepsilon\biggr) \leq \frac{\mathbb{E}[R_n]}{\varepsilon} \leq \frac{M^2\sum_{k=1}^n \gamma_k^2}{2K\varepsilon}, \tag{5.8}$$

where we used the fact that $\mathbb{E}[\|\nabla F(X_n;\omega_n)\|_*^2] \leq M^2$ for some finite $M < \infty$. Consequently, if we choose γ_n so that $\sum_{k=1}^\infty \gamma_k^2 \leq K\delta\varepsilon/M^2$, (5.8) readily gives

$$\mathbb{P}\biggl(\sup_{0\leq k\leq n} R_k \geq \varepsilon\biggr) \leq \frac{M^2\sum_{k=1}^\infty \gamma_k^2}{2K\varepsilon} \leq \frac{\delta}{2} \quad \text{for all } n. \tag{5.9}$$

Since the above is true for all n, (5.7) follows.

Step 3*: Error aggregation.* To combine the above, assume that $\varepsilon > 0$ is sufficiently small so that $\mathbb{B}_F(x^*, 3\varepsilon) \subset U$, where U is the open neighborhood given in (LVC). Furthermore, let $\mathcal{U} = \mathbb{B}_F(x^*, \varepsilon)$ and pick a step-size sequence γ_n satisfying (2.12) and such that

$$\sum_{n=1}^\infty \gamma_n^2 \leq \min\biggl\{\frac{\delta\varepsilon^2}{2R^2M^2}, \frac{K\delta\varepsilon}{M^2}\biggr\}. \tag{5.10}$$

If $X_1 \in \mathcal{U}$, it follows that $F(x^*, Y_1) < \varepsilon$ by the definition of $\mathbb{B}_F$ (cf. Definition 3.3). Then, by (5.3) and (5.7), we get $\mathbb{P}(\sup_n S_n \geq \varepsilon) \leq \delta/2$ and $\mathbb{P}(\sup_n R_n \geq \varepsilon) \leq \delta/2$. Consequently, with this choice of γ_n, it follows that

$$\mathbb{P}(\sup_n \max\{S_n, R_n\} \leq \varepsilon) \geq 1 - \delta/2 - \delta/2 = 1 - \delta. \tag{5.11}$$

Then, letting $F_n = F(x^*, Y_n)$ and arguing as in the proof of Theorem 4.1, we may expand $F_n = F(x^*, Y_n)$ to get

$$\begin{aligned} F_n &= F(x^*, Y_n + \gamma_n \hat{v}_n) \\ &\leq F_n + \gamma_n \langle v(X_n), X_n - x^* \rangle + \gamma_n \xi_n + \frac{\gamma_n^2}{2K} \|\nabla F(X_n; \omega_n)\|_*^2 \end{aligned} \tag{5.12}$$

with $\xi_n = \langle U_n, X_n - x^* \rangle$ defined as above. Telescoping (5.12) then yields

$$\begin{aligned} F_n &\leq F_1 + \sum_{k=1}^n \gamma_k \langle v(X_k), X_k - x^* \rangle + S_n + R_n \\ &\leq \varepsilon + \sum_{k=1}^n \gamma_k \langle v(X_k), X_k - x^* \rangle + \varepsilon + \varepsilon \end{aligned} \tag{5.13}$$

with probability at least $1 - \delta$. Therefore, with probability at least $1 - \delta$, we have

$$F(x^*, Y_n) \leq 3\varepsilon + \sum_{k=1}^n \gamma_k \langle v(X_k), X_k - x^* \rangle. \tag{5.14}$$

Now, assume inductively that, for all $k \leq n$, we have $F(x^*, Y_k) \leq 3\varepsilon$ or, equivalently, $X_k \in \mathbb{B}_F(x^*, 3\varepsilon)$. In turn, this implies that $\langle v(X_k), X_k - x^* \rangle \leq 0$ for all $k \leq n$ and hence, by (5.14), that $F(x^*, Y_n) \leq 3\varepsilon$ as well. Since the base case $X_1 \in \mathcal{U} = \mathbb{B}_F(x^*, \varepsilon) \subset \mathbb{B}_F(x^*, 3\varepsilon)$ is satisfied automatically, we conclude that X_n stays in $\mathbb{B}_F(x^*, 3\varepsilon) \subset U$ 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. ∎

We close this section by revisiting the notion of weak coherence (Definition 2.3). In view of Definition 5.1, we see that weak coherence mixes elements of both global *and* local coherence: on the one hand, it posits the existence of a (global) minimizer $p \in \mathcal{X}^*$ for which (VC) holds globally, thus satisfying the second part of Definition 2.1; on the other hand, minimizers other than p are only required to satisfy (VC) locally (though they need not be locally coherent themselves). From a stability viewpoint, this means that individual elements of a weakly coherent set may be stable but not necessarily attracting (even locally). However, taken as a whole, weakly coherent sets are *globally* attracting.

THEOREM 5.3. *Suppose that* (Opt) *is weakly coherent. Then, under Assumptions* 1–3*,* X_n *converges with probability* 1 *to a* (*possibly random*) *minimum point of* (Opt).

Proof. The proof is essentially a combination of the proofs of Theorems 4.1 and 5.2, so we only provide the main arguments and omit the minor details.

The first observation is that the conclusion of Proposition 3.4 only requires the first part of Definition 2.3 (simply take $x^* = p$ in the proof of Proposition 3.4). From this, we conclude that, with probability 1, the sequence of generated states X_n admits a subsequence that converges to some (possibly random) point in $\mathcal{X}^*$.

To proceed, fix some positive δ and ε as in the proof of Theorem 5.2. Then, analogously to (5.10), there exists some starting index $n_0 \equiv n_0(\delta, \varepsilon)$ such that $\sum_{n=n_0}^\infty \gamma_n^2 \leq$

$M^{-2}\min\{\delta\varepsilon^2/(2R^2), K\delta\varepsilon\}$, implying in turn that $\mathbb{P}(\sup_{n\geq n_0} S_n \geq \varepsilon) \leq \delta/2$ and $\mathbb{P}(\sup_{n\geq n_0} R_n \geq \varepsilon) \leq \delta/2$ (by (5.3) and (5.7), respectively). Then, arguing as in (5.13), we get

$$F_n \leq F_{n_1} + \sum_{k=n_1}^{n} \gamma_k \langle v(X_k), X_k - x^* \rangle + 2\varepsilon \quad \text{for all } n \geq n_1 \geq n_0 \tag{5.15}$$

with probability at least $1-\delta$.

Assume now that $\varepsilon > 0$ is sufficiently small so that $\langle \nabla f(x), x - x^* \rangle \geq 0$ for all $x \in \mathbb{B}_F(x^*, 3\varepsilon)$ and all $x^* \in \mathcal{X}^*$ (that such an ε exists is a consequence of Definition 2.3 and the compactness of $\mathcal{X}^*$). With this ε in hand, if $x^* \in \mathcal{X}^*$ is a limit point of X_n (recall our first observation above), we may instantiate n_1 so that $F_{n_1} = F(x^*, Y_{n_1}) < \varepsilon$. Then, for all $n \geq n_1$, we will have $F_n \leq 3\varepsilon + \sum_{k=n_1}^{n} \gamma_k \langle v(X_k), X_k - x^* \rangle$ with probability at least $1-\delta$. Thus, proceeding inductively as in the proof of Theorem 5.2, we finally get

$$\mathbb{P}(X_n \in \mathbb{B}_F(x^*, 3\varepsilon) \text{ for all } n \geq n_1) \geq 1 - \delta. \tag{5.16}$$

Since ε can be taken arbitrarily small in (5.16), we conclude that X_n converges to a (possibly random) minimizer $x^* \in \mathcal{X}^*$ with probability at least $1-\delta$. Hence, with $\delta > 0$ itself arbitrary, our assertion follows. ∎

6. Sharp minima and applications. 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 fairly 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 below that X_n reaches a minimum point of (Opt) in a *finite* number of iterations for a large class of optimization problems that admit *sharp* minima (see below). As we noted in section 2, an important example of a surjective mirror map is the standard Euclidean projection $\Pi(y) = \arg\min_{x\in\mathcal{X}} \|y - x\|_2$. The resulting descent method is the well-known SGD algorithm (cf. Algorithm 2 below), so our results in this section also provide new insights into the behavior of SGD.

6.1. Definition and characterization. The starting point of our analysis is Polyak's fundamental notion of a *sharp minimum* [35, 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* ρ-sharp *(local) minimum of f if*

$$f(x) \geq f(x^*) + \rho\|x - x^*\| \quad \textit{for some } \rho > 0 \textit{ and all } x \textit{ sufficiently close to } x^*. \tag{6.1}$$

Polyak's original definition concerned global sharp minima of unconstrained (convex) optimization problems; by contrast, the above definition is tailored to local optima of constrained (and possibly nonconvex) programs. In particular, Definition 6.1 implies that sharp minima are isolated (local) minimizers of f, and they remain invariant under small perturbations of f (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.

Algorithm 2 Stochastic gradient descent.

```
Require: step-size sequence γ_n > 0
choose Y ∈ R^d, X = Π(Y)                          # initialization
for n = 1, 2, ... do
    get v̂ = −∇F(X; ω)                             # oracle feedback
    set Y ← Y + γ_n v̂                             # gradient step
    set X ← Π(Y)                                  # set state
end for
return X                                          # output
```

Sharp minima admit a useful geometric interpretation in terms of the polar cone of $\mathcal{X}$. To state it, recall first the following basic definitions (see also the notation in section 1).

DEFINITION 6.2. *Let $\mathcal{X}$ be a closed convex subset of $\mathbb{R}^d$. Then,*

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

The above gives the following geometric characterization of sharp minima.

LEMMA 6.3. *If $x^* \in \mathcal{X}$ is a ρ-sharp minimum of f, we have*

$$\langle \nabla f(x^*), z\rangle \geq \rho\|z\| \quad \text{for all } z \in \mathrm{TC}(x^*). \tag{6.2}$$

In particular, $\nabla f(x^*)$ belongs to the topological interior of $\mathrm{TC}^*(x^*)$. Conversely, if (6.2) holds and f is convex, x^* is sharp.

Proof of Lemma 6.3. For the direct implication, fix some $x \in \mathcal{X}$ satisfying (6.1), and let $z = x - x^* \in \mathrm{TC}(x^*)$. Then, by the definition of a sharp minimum, we get

$$f(x^* + \tau z) \geq f(x^*) + \rho\tau\|z\| \quad \text{for all } \tau \in [0, 1]. \tag{6.3}$$

In turn, this implies that

$$\frac{f(x^* + tz) - f(x^*)}{t} \geq \rho\|z\| \quad \text{for all sufficiently small } t > 0. \tag{6.4}$$

Hence, taking the limit $t \to 0^+$, we get $\langle \nabla f(x^*), z\rangle \geq \rho\|z\|$, and our claim follows from the definition of $\mathrm{TC}(x^*)$ as the closure of the set of all rays emanating from x^* and intersecting $\mathcal{X}$ in at least one other point. Furthermore, if $\nabla f(x^*)$ lies at the boundary of $\mathrm{TC}(x^*)$, there exists some nonzero $z \in \mathrm{TC}(x^*)$ such that $\langle \nabla f(x^*), z\rangle = 0$; this contradicts (6.2), so we conclude that $\nabla f(x^*)$ is interior.

Finally, for the converse implication of the theorem, simply note that $f(x) - f(x^*) \geq \langle \nabla f(x^*), x - x^*\rangle \geq \rho\|x - x^*\|$ if f is convex. ∎

Example 6.1 (linear programs). A first important class of examples of functions that admit sharp minima is that of generic linear programs.[7] Indeed, by definition, a

[7] "Generic" means here that $\mathcal{X}$ is a polytope, $f\colon \mathcal{X} \to \mathbb{R}$ is affine, and f is constant only on the zero-dimensional faces of $\mathcal{X}$. Any linear program can be turned into a generic one after an arbitrarily small perturbation.

linear function grows (exactly) linearly around its minimum points so, by genericity, we have the following.

Example 6.2 (concave minimization). For a nonconvex class of examples, let $f : \mathcal{X} \to \mathbb{R}$ be a strictly *concave* function defined over a convex polytope $\mathcal{X}$ of $\mathbb{R}^d$. Concavity implies that f is superharmonic, i.e.,

$$\Delta f(x) = \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2} \leq 0 \tag{6.5}$$

for all $x \in \mathcal{X}$.[8] By the minimum principle for superharmonic functions, the minimum of f over any connected region $\mathcal{C}$ of $\mathcal{X}$ is attained at the boundary of $\mathcal{C}$. Hence, by strict concavity, we conclude that the local minima of f are attained at zero-dimensional faces of $\mathcal{X}$, and they are de facto sharp (simply note that f is strictly concave along any ray of the form $x^* + tz$, $z \in \mathrm{TC}(x^*)$).

Remark 6.1. Sharp minima have several other interesting and useful properties. First, by Lemma 6.3, sharp minimum points are locally coherent. To see this, simply note that for all $x \in \mathcal{X}$ sufficiently close to x^* (with $x \neq x^*$), we have $z = x - x^* \in \mathrm{TC}(x^*)$ and $\langle \nabla f(x^*), z \rangle \geq \rho \|z\| > 0$. Consequently, $\langle \nabla f(x^*), x - x^* \rangle > 0$, implying by continuity that $\langle \nabla f(x), x - x^* \rangle > 0$ for all x in some open neighborhood of x^* (excluding x^*). In addition, if (Opt) is variationally coherent, then a sharp (local) minimum is globally sharp as well.

A second important property is that the dual cone $\mathrm{TC}^*(x^*)$ of a sharp minimum must necessarily have nonempty topological interior—since it contains $\nabla f(x^*)$ 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 (see also Example 6.2 above).

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 $n^{-1} \sum_{k=1}^{n} X_k$: 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). 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 the following.

[8] We tacitly assume above that f is twice-differentiable but this conclusion still holds even if f is not differentiable.

THEOREM 6.4. *Suppose that* (Opt) *is variationally coherent. If f admits a* (*necessarily unique*) *sharp minimum x^*, and 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.*). *More precisely, we have*

$$\mathbb{P}(X_n = x^* \text{ for all sufficiently large } n) = 1 \tag{6.6}$$

COROLLARY 6.5. *Let f be a nondegenerate quasi-convex* (*or pseudoconvex, or convex*) *function and let x^* be a sharp minimum of f. Then, with assumptions as above, X_n reaches x^* in a finite number of steps* (*a.s.*).

The prototypical example of a surjective mirror map is the Euclidean projector $\Pi(y) = \arg\min_{x\in\mathcal{X}}\|y - x\|_2$ induced by the quadratic regularization function $h(x) = \|x\|_2^2/2$ (cf. Example 2.7). The resulting descent scheme is the well-known SGD algorithm (see Algorithm 2 for a pseudocode implementation), for which we obtain the following novel convergence result.

COROLLARY 6.6. *If* (Opt) *is a generic linear program, the last iterate X_n of SGD reaches its* (*necessarily unique*) *solution in a finite number of steps* (*a.s.*).

Remark 6.2. The sharpness assumption is crucial for obtaining convergence in a finite number of iterations, as is the use of "lazy" versus "greedy" mirror steps. A special case of this result, when the objective is convex, was independently obtained in the context of constraint identification and stochastic optimization in the recent work [12].[9]

With all this said and done, we proceed with the proof of Theorem 6.4.

Proof of Theorem 6.4. Since x^* is a ρ-sharp minimum, there exists a sufficiently small open neighborhood $\mathcal{U}$ of x^* such that $\langle \nabla f(x), z\rangle \geq \rho\|z\|/2$ for all $z \in \mathrm{TC}(x^*)$ and all $x \in \mathcal{U}$ (cf. Remark 6.1). By Theorem 4.1, X_n converges to x^* (a.s.), so there exists some (random) n_0 such that $X_n \in \mathcal{U}$ for all $n \geq n_0$. In turn, this implies that $\langle \nabla f(X_n), z\rangle \geq \rho\|z\|/2$ for all $n \geq n_0$. Thus, continuing to use the notation $v(X_n) = -\nabla f(X_n)$ and $U_n = \nabla f(X_n) - \nabla F(X_n;\omega_n)$, we get for all $z \in \mathrm{TC}(x^*)$ with $\|z\| \leq 1$

$$\begin{aligned}
\langle Y_n, z\rangle &= \left\langle Y_{n_0} + \sum_{k=n_0}^{n} \gamma_k \hat{v}_k, z \right\rangle \\
&= \langle Y_{n_0}, z\rangle + \sum_{k=n_0}^{n} \gamma_k \langle v(X_k), z\rangle + \sum_{k=n_0}^{n} \gamma_k \langle U_k, z\rangle \\
&\leq \|Y_{n_0}\|_* - \frac{\rho}{2}\sum_{k=n_0}^{n} \gamma_k + \sum_{k=n_0}^{n} \gamma_k \langle U_k, z\rangle, \qquad (6.7)
\end{aligned}$$

where, in the last line, we used the fact that $X_k \in \mathcal{U}$ for all $k \geq n_0$.

As we discussed in the proof of Theorem 4.1, $\gamma_n U_n$ is a martingale difference sequence relative to the natural filtration $\mathcal{F}_n$ of ω_n. Hence, by the law of large numbers for martingale differences (cf. Theorem A.1 for $p = 2$ and $\tau_n = \sum_{k=0}^{n}\gamma_k$), we get

[9] We were made aware of the work of [12] during the final stages of the revision of our paper; we are grateful to the authors for bringing it to our attention.

$$\lim_{n\to\infty} \frac{\sum_{k=n_0}^{n} \gamma_k U_k}{\sum_{k=n_0}^{n} \gamma_k} = 0 \quad \text{(a.s.)}. \tag{6.8}$$

Thus, there exists some n^* such that $\|\sum_{k=n_0}^{n} \gamma_k U_k\|_* \leq (\rho/4) \sum_{k=n_0}^{n} \gamma_k$ for all $n \geq n^*$ (a.s.). (6.7) then implies

$$\begin{aligned} \langle Y_n, z\rangle &\leq \|Y_{n_0}\|_* - \frac{\rho}{2} \sum_{k=n_0}^{n} \gamma_k + \sum_{k=n_0}^{n} \gamma_k \langle U_k, z\rangle \\ &\leq \|Y_{n_0}\|_* - \frac{\rho}{2} \sum_{k=n_0}^{n} \gamma_k + \frac{\rho}{4} \sum_{k=n_0}^{n} \gamma_k = \|Y_{n_0}\|_* - \frac{\rho}{4} \sum_{k=n_0}^{n} \gamma_k, \end{aligned} \tag{6.9}$$

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

To proceed, we claim that $y^* + \mathrm{PC}(x^*) \subseteq Q^{-1}(x^*)$ whenever $Q(y^*) = x^*$, i.e., $Q^{-1}(x^*)$ contains all cones of the form $y^* + \mathrm{PC}(x^*)$ for $y^* \in Q^{-1}(x^*)$. Indeed, note first that $x^* = Q(y^*)$ if and only if $y^* \in \partial h(x^*)$, where $\partial h(x^*)$ is the set of all subgradients of h at x^* [37]. Therefore, it suffices to show that $y^* + w \in \partial h(x^*)$ whenever $w \in \mathrm{PC}(x^*)$. To that end, note that the definition of the polar cone gives

$$\langle w, x - x^*\rangle \leq 0 \quad \text{for all } x \in \mathcal{X},\ w \in \mathrm{PC}(x^*), \tag{6.10}$$

and hence

$$h(x) \geq h(x^*) + \langle y^*, x - x^*\rangle \geq h(x^*) + \langle y^* + w, x - x^*\rangle. \tag{6.11}$$

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 $y^* + \mathrm{PC}(x^*)$ for some $y^* \in Q^{-1}(x^*)$ and all sufficiently large n. To do so, simply note that $Y_n \in y^* + \mathrm{PC}(x^*)$ if and only if $\langle Y_n - y^*, z\rangle \leq 0$ for all $z \in \mathrm{TC}(x^*)$ 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. Our convergence analysis for locally coherent sets of minimizers (cf. section 5) showed that SMD converges locally with high probability. Our last result in this section complements this analysis by showing that, with high probability, SMD converges locally to sharp local minima in a *finite* number of iterations.

THEOREM 6.7. *Let x^* be a sharp (local) minimum of f, and fix some confidence level $\delta > 0$. If Algorithm* 1 *is run with a surjective mirror map and Assumptions* 1–3 *hold, there exists an open neighborhood $\mathcal{U}$ of x^*, independent of δ, such that*

$$\mathbb{P}(X_n = x^* \text{ for all sufficiently large } n \,|\, X_1 \in \mathcal{U}) \geq 1 - \delta, \tag{6.12}$$

provided that the algorithm's step-size sequence γ_n is small enough.

Given that local minimizers of concave minimization programs are sharp, an application of Theorem 6.7 gives the following convergence result for SGD.

COROLLARY 6.8. *Suppose that f is strictly concave as in Example* 6.2. *Then, under Assumptions* 1 *and* 2, *the last iterate of SGD converges locally to a local minimum of* (Opt) *with arbitrarily high probability.*

Proof of Theorem 6.7. Under the stated assumptions, Theorem 5.2 implies that there exists an open neighborhood $\mathcal{U}$ of x^* such that $\mathbb{P}(\lim_{n\to\infty} X_n = x^* \,|\, X_0 \in \mathcal{U}) \geq 1-\delta$. In turn, this means that there exists some (random) n_0 which is finite with probability at least $1-\delta$ and is such that $\langle \nabla f(x_n), z\rangle \geq \rho\|z\|/2$ 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 6.4. ∎

We close this section by noting that the convergence of X_n in a finite number of steps is a unique feature of *lazy* descent schemes with a surjective mirror map. For example, if we consider the *greedy* (or *eager*) projected descent scheme

$$X_{n+1} = \Pi(X_n - \gamma_n \nabla F(X_n;\omega_n)), \tag{6.13}$$

it is not possible to obtain a result similar to Theorems 6.4 and 6.7 without further assumptions on the stochasticity affecting (Opt). To see why, assume that x^* is a sharp minimum of (Opt) and $X_n = x^*$ for some n. If the sampled gradient $\nabla F(X_n;\omega_n)$ attains all directions with positive probability (more precisely, if the unit vector $\nabla F(X_n;\omega_n)/\|\nabla F(X_n;\omega_n)\|_*$ is supported on the entire unit sphere $\mathbb{S}^d$ of $\mathbb{R}^d$), there exists some $\delta > 0$ such that

$$\mathbb{P}(\nabla F(X_n;\omega_n) \notin \mathrm{TC}^*(x^*)) \geq \delta \quad \text{for all } n. \tag{6.14}$$

We thus obtain

$$\mathbb{P}(X_{n+1} \neq x^* \,|\, X_n = x^*) \geq \delta \quad \text{for all } n, \tag{6.15}$$

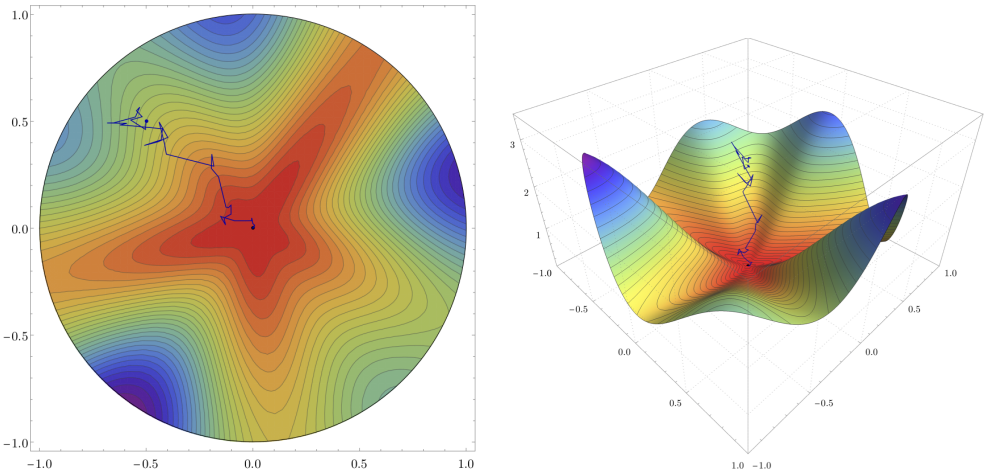
implying in turn that X_n cannot converge to x^* in a finite number of iterations. We find this property of lazy descent schemes particularly appealing as it ensures very fast convergence in the presence of sharp minima.

7. Numerical experiments. In this section, we validate the theoretical analysis of the previous sections via a series of numerical experiments.

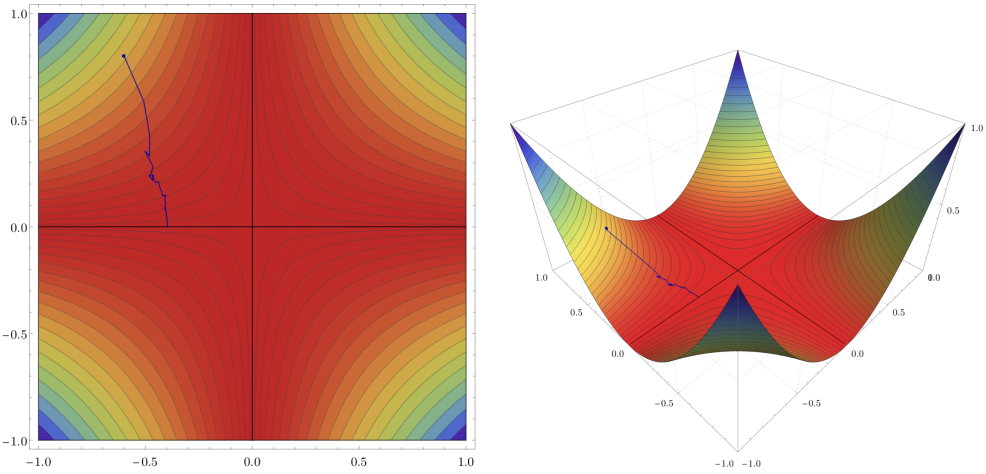
As a first illustration of Theorems 4.1 and 5.3, we begin by plotting the generated trajectories of (SMD) for two nonconvex test functions satisfying the coherence requirements of Definitions 2.1 and 2.3 respectively. Referring to Figure 2 for the detailed expressions, the specific setup is as in Example 2.2 with U following a standard Gaussian distribution; (SMD) was then run with the Euclidean projector of Example 2.7 and a step-size sequence $\gamma_n \propto 1/n$. In both cases, the (randomly) generated trajectories of (SMD) are seen to converge to a minimum point of (Opt), even when the problem's minimum set is nonconvex (as in the second example plotted in Figure 2).

To go beyond globally coherent problems, we also test the convergence of (SMD) against the widely used Rosenbrock benchmark of (5.1). This test function admits a unique global minimum point at $x^* = (1,\dots,1)$; however, this minimum is at the lowest point of a very flat and thin parabolic valley which is notoriously difficult for first-order methods to traverse [39]. Because of the parabolic shape of this valley, the problem is not globally coherent (there are rays emanating from x^* along which f fails to be nondecreasing) but an easy algebraic calculation in the spirit of Example 2.6 shows that x^* is locally coherent.

For illustration purposes, we first focus on a low-dimensional example with $d = 2$ degrees of freedom and algorithmic parameters as in Figure 2. Despite the lack of global coherence, the simulated trajectories of (SMD) quickly reach the Rosenbrock valley and eventually converge to the minimum of f; a typical such trajectory is shown



(a) $f(r,\theta) = (3+\sin(5\theta)+\cos(3\theta))r^2(5/3-r)$ in polar coordinates $(0 \leq r \leq 1,\ 0 \leq \theta \leq 2\pi)$.



(b) $f(x_1,x_2) = x_1^2 x_2^2$, $-1 \leq x_1, x_2 \leq 1$.

FIG. 2. *Convergence of SMD in a coherent problem with a unique minimizer (top) and a weakly coherent problem with a nonconvex minimum set (bottom). In both cases, the minimum set of f is plotted in solid black.*

in Figure 3. Subsequently, in Figure 4, we run a series of tests on the Rosenbrock function for $d = 10^3$ and $d = 10^4$ degrees of freedom. Because the calculation of the gradient becomes increasingly difficult as d grows large, we take the approach of Example 2.1 and, at each iteration $n = 1, 2, \dots$ of the algorithm, we randomly pick an integer between 1 and d and calculate the gradient of $f_i(x) = 100(x_{i+1}-x_i)^2+(1-x_i)^2$.

In so doing, we obtain the plots of Figure 4 where, for statistical significance, we report the findings of $S = 100$ sample realizations. For comparison purposes, we also include in the figure the performance of the ergodic average $\bar{X}_n$ of X_n as defined in (4.1). This sequence is the standard output of mirror descent/dual averaging schemes in convex problems; however, in our nonconvex setting, this averaging offers no tangible benefits. This is seen clearly in Figure 4, where the convergence speed of $\bar{X}_n$ is considerably slower than that of the algorithm's last iterate.

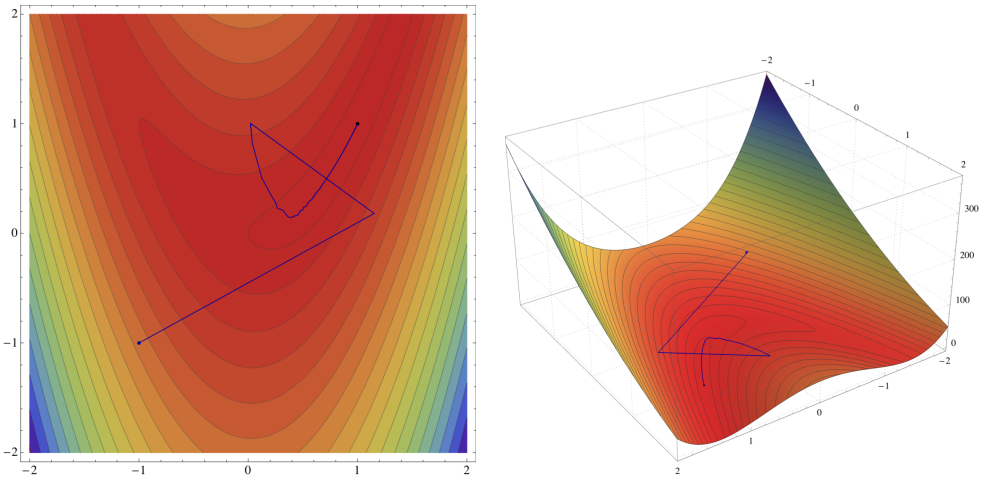


FIG. 3. *Convergence of the SMD algorithm in the Rosenbrock test with $d = 2$ degrees of freedom.*

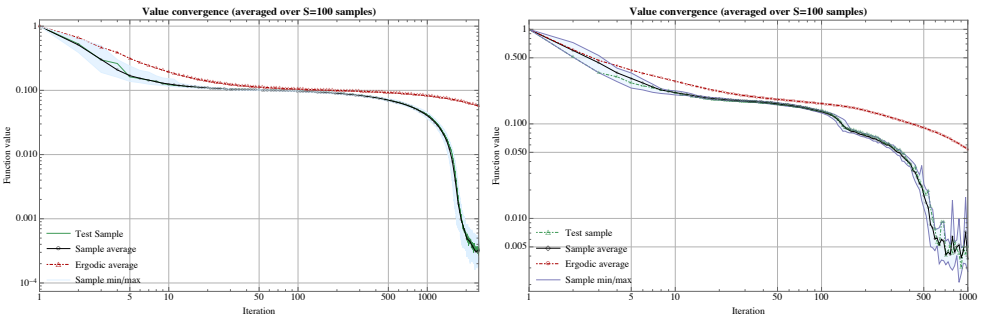


FIG. 4. *Convergence speed of SMD in the Rosenbrock benchmark for $d = 10^3$ and $d = 10^4$ degrees of freedom (left and right, respectively). The lightly shaded envelope indicates the best and worst realizations of the algorithm over $S = 100$ sample runs; the corresponding sample mean is represented by a solid black line. For comparison purposes, we also plot the performance of the ergodic average $\bar{X}_n = \sum_{k=1}^n \gamma_k X_k / \sum_{k=1}^n \gamma_k$ of X_n (dashed red line). Due to lack of convexity, the ergodic average of X_n converges at a significantly slower rate.*

Finally, in Figure 5, we examine the convergence rate of (SMD) for quadratic objective functions of the form

$$f(x) = \frac{1}{2} \sum_{i,j=1}^{d} Q_{ij} x_i x_j + \sum_{i=1}^{d} b_i x_i. \tag{7.1}$$

When x is constrained to lie on the unit simplex of $\mathbb{R}^d$, the minimization of f is related to the maximum weight clique problem [25]: this problem is well known to be NP-hard, so fast convergence to local minima of f is essential in order to get reasonable approximate solutions.

Using again the stochastic setup of Example 2.2, we ran both Algorithm 2 and its greedy variant (6.13) for a general random quadratic objective of the form (7.1) with $d = 100$ and randomly drawn Q and b. As can be seen in Figure 5, the lazy variant of SGD converges within a finite number of iterations, whereas the greedy variant still oscillates within the allotted time window. This behavior is explained by Theorem 6.7

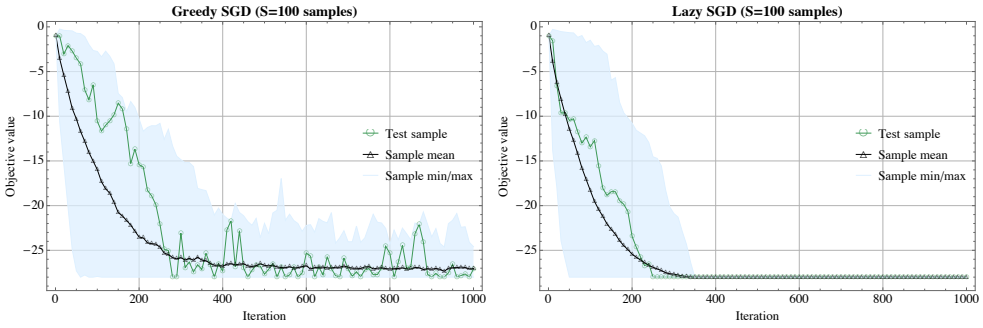


FIG. 5. *Convergence of the lazy and greedy variants of SGD in a quadratic minimization problem of the form* (7.1) *with* $d = 100$. *The lightly shaded area traces the best and worst realizations of the algorithm over* $S = 100$ *sample runs; the corresponding sample mean is drawn as a solid black line. In the dedicated runtime* ($n = 1000$ *iterations), the greedy variant still hasn't converged; by contrast, even the worst realization of lazy SGD has converged within approximately* 300 *iterations.*

and the discussion that follows: because the greedy variant essentially "remembers" only the last state, convergence within a finite number of iterations is not possible; by contrast, the dual averaging that takes place in the lazy variant allows X_n to converge in finite time to a sharp local minimum, despite all the noise.

8. Discussion. To conclude, we first note that our analysis can be extended to the study of stochastic variational inequalities with possibly nonmonotone operators. The notion of a variationally coherent problem still makes 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.

Finally, we should also mention that another merit of SMD is that, at least for (strongly) convex optimization problems, 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 tackle this question in this paper, the techniques developed here can potentially be leveraged to provide theoretical guarantees for certain nonconvex 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 [18], 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* $M_n = \sum_{k=1}^{n} d_k$ *be a martingale with respect to an underlying stochastic basis* $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n=1}^{\infty}, \mathbb{P})$ *and let* $(\tau_n)_{n=1}^{\infty}$ *be a nondecreasing sequence of positive numbers with* $\lim_{n\to\infty} \tau_n = \infty$. *If* $\sum_{n=1}^{\infty} \tau_n^{-p}\, \mathbb{E}[|d_n|^p \,|\, \mathcal{F}_{n-1}] < \infty$ *for some* $p \in [1, 2]$ (*a.s.*), *we have*

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

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 nonnegative submartingale and fix some $\varepsilon > 0$. Then,*

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

Finally, a widely used variant of Doob's maximal inequality is the following.

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

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

Appendix B. Technical proofs. In this appendix, we present the proofs that were omitted from the main text. We begin with the core properties of the Fenchel coupling.

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 [37]. We thus get

$$F(p, y) = h(p) + h^*(y) - \langle y, p \rangle = h(p) - h(x) - \langle y, p - x \rangle. \tag{B.1}$$

Since $y \in \partial h(x)$ and h is K-strongly convex, we also have

$$h(x) + \tau \langle y, p - x \rangle \leq h(x + \tau(p - x)) \leq (1 - \tau)h(x) + \tau h(p) - \tfrac{1}{2} K \tau (1 - \tau) \|x - p\|^2 \tag{B.2}$$

for all $\tau \in [0, 1]$, thereby leading to the bound

$$\tfrac{1}{2} K (1 - \tau) \|x - p\|^2 \leq h(p) - h(x) - \langle y, p - x \rangle = F(p, y). \tag{B.3}$$

Our claim then follows by letting $\tau \to 0^+$ in (B.3).

For our second claim, we start by citing a well-known duality principle for strongly convex functions [38, Theorem 12.60]: If $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ is proper, lower semicontinuous, and convex, its convex conjugate f^* is σ-strongly convex if and only if f is differentiable and satisfies

$$f(x') \leq f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2\sigma} \|x' - x\|^2 \tag{B.4}$$

for all $x, x' \in \mathbb{R}^d$. In our case, if we extend h to all of $\mathcal{V}$ by setting $h \equiv +\infty$ outside $\mathcal{X}$, we have that h is K-strongly convex, lower semicontinuous, and proper, so $(h^*)^* = h$ [38, Theorem 11.1]. It is also easy to see that h^* is proper, lower semicontinuous, and convex (since it is a pointwise maximum of affine functions by definition), so the K-strong convexity of $h = (h^*)^*$ implies that h^* is differentiable and satisfies

$$h^*(y') \leq h^*(y) + \langle y' - y, \nabla h^*(y) \rangle + \frac{1}{2K} \|y' - y\|_*^2 \tag{B.5}$$

$$= h^*(y) + \langle y' - y, Q(y) \rangle + \frac{1}{2K} \|y' - y\|_*^2 \tag{B.6}$$

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

$$\begin{aligned} F(x,y') &= h(x) + h^*(y') - \langle y', x\rangle \\ &\leq h(x) + h^*(y) + \langle y' - y, \nabla h^*(y)\rangle + \frac{1}{2K}\|y'-y\|_*^2 - \langle y', x\rangle \\ &= F(x,y) + \langle y' - y, Q(y) - x\rangle + \frac{1}{2K}\|y'-y\|_*^2, \end{aligned} \tag{B.7}$$

and our assertion follows. ∎

We now turn to the recurrence properties of SMD.

Proof of Proposition 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 $\mathbb{E}[\|F(x;\omega_n)\|_*] < \infty$. We then claim that $U_n = \nabla f(X_n) - \nabla F(X_n;\omega_n)$ is an L^2-bounded martingale difference sequence with respect to the natural filtration of ω_n. Indeed, we have as follows:

1. Since X_n is $\mathcal{F}_{n-1}$-measurable and ω_n is i.i.d., we readily get

$$\begin{aligned} \mathbb{E}[U_n \,|\, \mathcal{F}_{n-1}] &= \mathbb{E}[\nabla f(X_n) - \nabla F(X_n;\omega_n) \,|\, \mathcal{F}_{n-1}] \\ &= \mathbb{E}[\nabla f(X_n) - \nabla F(X_n;\omega_n) \,|\, \omega_1, \dots, \omega_{n-1}] \\ &= \nabla f(X_n) - \nabla f(X_n) \\ &= 0. \end{aligned} \tag{B.8}$$

2. Furthermore, by Assumption 2, the L^2 norm of U satisfies

$$\begin{aligned} \mathbb{E}[\|U_n\|_*^2 \,|\, \mathcal{F}_{n-1}] &= \mathbb{E}[\|\nabla f(X_n) - \nabla F(X_n;\omega_n)\|_*^2 \,|\, \mathcal{F}_{n-1}] \\ &\leq 2\,\mathbb{E}[\|\nabla f(X_n)\|_*^2 \,|\, \mathcal{F}_{n-1}] + 2\,\mathbb{E}[\|\nabla F(X_n;\omega_n)\|_*^2 \,|\, \mathcal{F}_{n-1}] \\ &\leq 2\|\nabla f(X_n)\|_*^2 + 2M^2 \\ &= 2\|\mathbb{E}[\nabla F(X_n;\omega)]\|_*^2 + 2M^2 \\ &\leq 2\,\mathbb{E}[\|\nabla F(X_n;\omega)\|_*^2] + 2M^2 \\ &\leq \sigma^2, \end{aligned} \tag{B.9}$$

where we set $\sigma^2 = 4M^2$, and we used the dominated convergence theorem to interchange expectation and differentiation in the second line, and Jensen's inequality in the penultimate one.

Step 2*: Recurrence of* ε*-neighborhoods.* We proceed to show that every ε-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 there exists some finite n_0 such that $\operatorname{dist}(\mathcal{X}^*, X_n) \geq \varepsilon$ for all $n \geq n_0$. Since $\mathcal{X} \setminus \mathbb{B}(\mathcal{X}^*, \varepsilon)$ is compact, $v(x) = -\nabla f(x)$ is continuous in x; furthermore, letting x^* be such that $\langle \nabla f(x), x - x^*\rangle = 0$ only if $x \in \mathcal{X}^*$ (cf. Definition 2.1), it follows that there exists some $c \equiv c(\varepsilon) > 0$ such that

$$\langle v(x), x - x^*\rangle \leq -c < 0 \quad \text{for all } x \in \mathcal{X} \setminus \mathbb{B}(\mathcal{X}^*, \varepsilon). \tag{B.10}$$

To proceed, let $R = \max_{x\in\mathcal{X}}\|x\|$ and set $\beta_n = \gamma_n^2/(2K)$. Then, letting $F_n = F(x^*, Y_n)$ and $\xi_n = \langle U_n, X_n - x^*\rangle$, Lemma 3.2 yields

$$\begin{aligned} F_{n+1} = F(x^*, Y_{n+1}) &= F(x^*, Y_n + \gamma_n\hat{v}_n) \\ &\leq F(x^*, Y_n) + \gamma_n\langle v(X_n) + U_n, X_n - x^*\rangle + \beta_n\|\hat{v}_n\|_*^2 \\ &= F_n + \gamma_n\langle v(X_n), X_n - x^*\rangle + \gamma_n\xi_n + \beta_n\|\hat{v}_n\|_*^2 \\ &\leq F_n - \gamma_n c + \gamma_n\xi_n + \beta_n\|\hat{v}_n\|_*^2. \end{aligned} \tag{B.11}$$

Hence, letting $\tau_n = \sum_{k=n_0}^n \gamma_k$ and telescoping from n_0 to n, we get

$$\begin{aligned} F_{n+1} &\leq F_{n_0} - c\sum_{k=n_0}^n \gamma_k + \sum_{k=n_0}^n \gamma_k\xi_k + \sum_{k=n_0}^n \beta_k\|\hat{v}_k\|_*^2 \\ &= F_{n_0} - \tau_n\left[c - \frac{\sum_{k=n_0}^n \gamma_k\xi_k}{\tau_n}\right] + \sum_{k=n_0}^n \beta_k\|\hat{v}_k\|_*^2. \end{aligned} \tag{B.12}$$

We now proceed to bound each term of (B.12). First, by construction, we have

$$\mathbb{E}[\xi_n \,|\, \mathcal{F}_{n-1}] = \mathbb{E}[\langle U_n, X_n - x^*\rangle \,|\, \mathcal{F}_{n-1}] = \langle \mathbb{E}[U_n \,|\, \mathcal{F}_{n-1}], X_n - x^*\rangle = 0, \tag{B.13}$$

where we used the fact that X_n is $\mathcal{F}_{n-1}$-measurable. Also, Young's inequality gives

$$|\xi_n| = |\langle U_n, X_n - x^*\rangle| \leq \|U_n\|_*\|X_n - x^*\| \leq 2R\|U_n\|_*, \tag{B.14}$$

where, as before, $R = \max_{x\in\mathcal{X}}\|x\|$. (B.9) then gives

$$\mathbb{E}[\xi_n^2 \,|\, \mathcal{F}_{n-1}] \leq (2R)^2\,\mathbb{E}[\|U_n\|_*^2 \,|\, \mathcal{F}_{n-1}] \leq 4R^2\sigma^2, \tag{B.15}$$

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

$$\sum_{n=n_0}^\infty \frac{\mathbb{E}[|\gamma_n\xi_n|^2 \,|\, \mathcal{F}_{n-1}]}{\tau_n^2} \leq 4R^2\sigma^2 \sum_{n=n_0}^\infty \frac{\gamma_n^2}{\tau_n^2} < \infty, \tag{B.16}$$

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

$$\frac{\sum_{k=n_0}^n \gamma_k\xi_k}{\tau_n} \to 0 \quad \text{as } n \to \infty \text{ (a.s.)}, \tag{B.17}$$

and, with $\sum_{k=n_0}^\infty \gamma_k = \infty$, we finally obtain

$$\lim_{n\to\infty} \tau_n\left[c - \frac{\sum_{k=n_0}^n \gamma_k\xi_k}{\tau_n}\right] = \infty \quad \text{(a.s.)}. \tag{B.18}$$

For the last term of (B.12), let $S_n = \sum_{k=n_0}^n \beta_k\|\hat{v}_k\|_*^2$, so S_n is $\mathcal{F}_n$-measurable and nondecrasing. In addition, we have

$$\mathbb{E}[S_n] = \sum_{k=n_0}^n \beta_k\,\mathbb{E}[\|\hat{v}_k\|_*^2] \leq M^2\sum_{k=n_0}^n \beta_k < \infty, \tag{B.19}$$

with the last step following from (B.9). 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_∞, i.e., the last term of (B.12) is bounded. Hence, combining all of the above, we finally obtain

$$\limsup_{n\to\infty} F_n = -\infty \quad \text{(a.s.)}, \tag{B.20}$$

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 ε-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 ε-ball $\mathbb{B}(\mathcal{X}^*, \varepsilon)$ contained in $\mathbb{B}_F(\mathcal{X}^*, \delta)$. Then, for all $k > 0$ there exists some $y_k \in \mathcal{Y}$ such that $\operatorname{dist}(\mathcal{X}^*, Q(y_k)) = 1/k$ but $F(\mathcal{X}^*, y_\delta) \geq \varepsilon$. This produces a sequence $(y_k)_{k=1}^\infty$ such that $\operatorname{dist}(\mathcal{X}^*, Q(y_k)) \to 0$ but $F(\mathcal{X}^*, y_k) \geq \varepsilon$. Since $\mathcal{X}$ is compact and $\mathcal{X}^*$ is closed, we can assume without loss of generality that $Q(y_k) \to p$ for some $p \in \mathcal{X}^*$ (at worst, we only need to descend to a subsequence of y_k). We thus get $\varepsilon \leq F(\mathcal{X}^*, y_k) \leq F(p, y_k)$. However, since $Q(y_k) \to p$, Assumption 3 gives $F(p, y_k) \to 0$, a contradiction. We conclude that $\mathbb{B}_F(\mathcal{X}^*, \delta)$ contains an ε-neighborhood of $\mathcal{X}^*$, completing our proof. ∎

REFERENCES

[1] S. Arora, E. Hazan, and S. Kale, *The multiplicative weights update method: A meta-algorithm and applications*, Theory Comput., 8 (2012), pp. 121–164.

[2] A. Auslender, R. Cominetti, and M. Haddou, *Asymptotic analysis for penalty and barrier methods in convex and linear programming*, Math. Oper. Res., 22 (1997), pp. 43–62.

[3] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper. Res. Lett., 31 (2003), pp. 167–175.

[4] M. Benaïm, *Dynamics of Stochastic Approximation Algorithms*, in Séminaire de Probabilités XXXIII, J. Azéma, M. Émery, M. Ledoux, and M. Yor, eds., Lecture Notes in Math. 1709, Springer, Berlin, 1999, pp. 1–68.

[5] L. Bottou, *On-line learning and stochastic approximations*, On-line Learning in Neural Networks, 17 (1998), pp. 9–42.

[6] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

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

[8] S. Bubeck, *Convex optimization: Algorithms and complexity*, Found. Trends Machine Learning, 8 (2015), pp. 231–358.

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

[10] X. Chen, Q. Lin, and J. Pena, *Optimal regularized dual averaging methods for stochastic optimization*, in NIPS '12: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012.

[11] Y. Chen, Y. Chi, J. Fan, and C. Ma, *Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval*, in Proceedings of the International Conference on Machine Learning, 2018.

[12] J. Duchi and F. Ruan, *Asymptotic optimality in stochastic optimization*, Ann. Statist., to appear.

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

[14] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer Ser. Oper. Res., Springer, Berlin, 2003.

[15] S. Ghadimi and G. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim., 23 (2013), pp. 2341–2368.
[16] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), pp. 59–99.
[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in NIPS '14: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014.
[18] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Application*, Probab. Math. Statist., Academic Press, New York, 1980.
[19] H. Jiang and H. Xu, *Stochastic approximation approaches to the stochastic variational inequality problem*, IEEE Trans. Automat. Control, 53 (2008), pp. 1462–1475.
[20] A. Juditsky, A. S. Nemirovski, and C. Tauvel, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stoch. Syst., 1 (2011), pp. 17–58.
[21] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, *Regularization techniques for learning with matrices*, J. Mach. Learn. Res., 13 (2012), pp. 1865–1890.
[22] J. Kivinen and M. K. Warmuth, *Exponentiated gradient versus gradient descent for linear predictors*, Inform. Comput., 132 (1997), pp. 1–63.
[23] K. C. Kiwiel, *Free-steering relaxation methods for problems with strictly convex costs and linear constraints*, Math. Oper. Res., 22 (1997), pp. 326–349.
[24] G. Lan, A. S. Nemirovski, and A. Shapiro, *Validation analysis of mirror descent stochastic approximation method*, Math. Program., 134 (2012), pp. 425–458.
[25] A. Massaro, M. Pelillo, and I. M. Bomze, *A complementary pivoting approach to the maximum weight clique problem*, SIAM J. Optim., 12 (2002), pp. 928–948.
[26] 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.
[27] P. Mertikopoulos and M. Staudigl, *On the convergence of gradient-like flows with noisy gradient input*, SIAM J. Optim., 28 (2018), pp. 163–197.
[28] P. Mertikopoulos and Z. Zhou, *Learning in games with continuous action sets and unknown payoff functions*, Math. Program., 173 (2019), pp. 465–507.
[29] A. Nedic and S. Lee, *On stochastic subgradient mirror-descent algorithm with weighted averaging*, SIAM J. Optim., 24 (2014), pp. 84–107.
[30] A. S. Nemirovski, *Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM J. Optim., 15 (2004), pp. 229–251.
[31] A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.
[32] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, 1983.
[33] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer, Dordrecht, 2004.
[34] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., 120 (2009), pp. 221–259.
[35] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.
[36] H. Robbins and S. Monro, *A stochastic approximation method*, Ann Math. Statist., 22 (1951), pp. 400–407.
[37] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[38] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, Berlin, 1998.
[39] H. H. Rosenbrock, *An automatic method for finding the greatest or least value of a function*, Computer J., 3 (1960), pp. 175–184.
[40] S. Shalev-Shwartz, *Online learning and online convex optimization*, Found. Trends Mach. Learn., 4 (2011), pp. 107–194.
[41] O. Shamir and T. Zhang, *Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes*, in ICML '13: Proceedings of the 30th International Conference on Machine Learning, 2013.
[42] K. Tsuda, G. Rätsch, and M. K. Warmuth, *Matrix exponentiated gradient updates for online Bregman projection*, J. Mach. Learn. Res., 6 (2005), pp. 995–1018.
[43] R. J. Vanderbei, *Linear Programming: Foundations and Extensions*, Springer, Berlin, 2007.
[44] Y. Wang, N. Xiu, and C. Wang, *A new version of extragradient method for variational inequality problems*, Comput. Math. Appl., 42 (2001), pp. 969–979.

[45] L. Xiao, *Dual averaging methods for regularized stochastic learning and online optimization*, J. Mach. Learn. Res., 11 (2010), pp. 2543–2596.

[46] Z. Zhou, P. Mertikopoulos, N. Bambos, S. Boyd, and P. W. Glynn, *Stochastic mirror descent for variationally coherent optimization problems*, in NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.

[47] Z. Zhou, P. Mertikopoulos, N. Bambos, P. W. Glynn, and C. Tomlin, *Countering feedback delays in multi-agent learning*, in Proceedings of Advances in Neural Information Processing Systems, 2017, pp. 6171–6181.

[48] Z. Zhou, P. Mertikopoulos, A. L. Moustakas, N. Bambos, and P. Glynn, *Mirror descent learning in continuous games*, in Proceedings of the 2017 IEEE 56th Annual Conference on Decision and Control, IEEE, 2017, pp. 5776–5783.