Stochastic model-based minimization under high-order growth

Damek Davis * Dmitriy Drusvyatskiy † Kellie J. MacPhee ‡

Abstract

Given a nonsmooth, nonconvex minimization problem, we consider algorithms that iteratively sample and minimize stochastic convex models of the objective function. Assuming that the one-sided approximation quality and the variation of the models is controlled by a Bregman divergence, we show that the scheme drives a natural stationarity measure to zero at the rate $O(k^{-1/4})$. Under additional convexity and relative strong convexity assumptions, the function values converge to the minimum at the rate of $O(k^{-1/2})$ and $\widetilde{O}(k^{-1})$, respectively. We discuss consequences for stochastic proximal point, mirror descent, regularized Gauss-Newton, and saddle point algorithms.

1 Introduction

Common stochastic optimization algorithms proceed as follows. Given an iterate x_t, the method samples a model of the objective function formed at x_t and declares the next iterate to be a minimizer of the model regularized by a proximal term. Stochastic proximal point, proximal subgradient, and Gauss-Newton type methods are common examples. Let us formalize this viewpoint, following [15]. Namely, consider the optimization problem

$$\min_{x\in\mathbb{R}^d} \; F(x) := f(x) + r(x). \tag{1.1}$$

where the function $r\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is closed and convex and the only access to $f\colon \mathbb{R}^d \to \mathbb{R}$ is by sampling a *stochastic one-sided model*. That is, for every point x, there exists a family of models $f_x(\cdot,\xi)$ of f, indexed by a random variable $\xi \sim P$. This setup immediately motivates the following algorithm, analyzed in [15]:

$$\left\{ \begin{array}{l} \text{Sample } \xi_t \sim P, \\ \text{Set } x_{t+1} = \operatorname*{argmin}_x \; \left\{ f_{x_t}(x,\xi_t) + r(x) + \dfrac{1}{2\eta_t}\|x - x_t\|_2^2 \right\} \end{array} \right\}, \tag{1.2}$$

*School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850; people.orie.cornell.edu/dsd95/.

†Department of Mathematics, University of Washington, Seattle, WA 98195; sites.math.washington.edu/~ddrusv. Research of Drusvyatskiy was partially supported by the AFOSR YIP award FA9550-15-1-0237 and by the NSF DMS 1651851 and CCF 1740551 awards.

‡Department of Mathematics, University of Washington, Seattle, WA 98195; sites.math.washington.edu/~kmacphee.

where $\eta_t > 0$ is an appropriate control sequence that governs the step-size of the algorithm.

Some thought shows that convergence guarantees of the method (1.2) should rely at least on two factors: (i) control over the approximation quality, $f_x(\cdot, \xi) - f(\cdot)$, and (ii) growth/stability properties of the individual models $f_x(\cdot, \xi)$. With this in mind, the paper [15] isolates the following assumptions:

$$\mathbb{E}_\xi[f_x(x, \xi)] = f(x) \qquad \text{and} \qquad \mathbb{E}_\xi[f_x(y, \xi) - f(y)] \leq \frac{\tau}{2}\|y - x\|_2^2 \qquad \forall x, y, \tag{1.3}$$

and there exists a square integrable function $L(\cdot)$ satisfying

$$f_x(x, \xi) - f_x(y, \xi) \leq L(\xi)\|x - y\|_2 \qquad \forall x, y. \tag{1.4}$$

Condition (1.3) simply says that in expectation, the model $f_x(\cdot, \xi)$ must globally lower bound $f(\cdot)$ up to a quadratic error, while agreeing with f at the base point x; when (1.3) holds, the paper [15] calls the assignment $(x, y, \xi) \mapsto f_x(y, \xi)$ a stochastic one-sided model of f. Property (1.4), in contrast, asserts a Lipschitz type property of the individual models $f_x(\cdot, \xi)$.[1] The main result of [15] shows that under these assumption, the scheme (1.2) drives a natural stationarity measure of the problem to zero at the rate $O(k^{-1/4})$. Indeed, the stationarity measure is simply the gradient of the Moreau envelope

$$F_\lambda(x) := \inf_y \left\{ F(y) + \tfrac{1}{2\lambda}\|y - x\|_2^2 \right\}, \tag{1.5}$$

where $\lambda > 0$ is a smoothing parameter on the order of τ.

The assumptions (1.3) and (1.4) are perfectly aligned with existing literature. Indeed, common first-order algorithms rely on global Lipschitz continuity of the objective function or of its gradient; see for example the monographs [5, 31, 33]. Recent work [2, 8, 26, 29, 30], in contrast, has emphasized that global Lipschitz assumptions can easily fail for well-structured problems. Nonetheless, these papers show that it is indeed possible to develop efficient algorithms even without the global Lipschitz assumption. The key idea, originating in [2, 29, 30], is to model errors in approximation by a Bregman divergence, instead of a norm. The ability to deal with problems that are not globally Lipschitz is especially important in stochastic nonconvex settings, where line-search strategies that exploit local Lipschitz continuity are not well-developed.

Motivated by the recent work on relative continuity/smoothness [2, 29, 30], we extend the results of [15] to non-globally Lipschitzian settings. Formally, we simply replace the squared norm $\frac{1}{2}\|\cdot\|^2$ in the displayed equations (1.2)-(1.5) by a Bregman divergence

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x \rangle,$$

generated by a Legendre function Φ. With this modification and under mild technical conditions, we will show that algorithm (1.2) drives the gradient of the Bregman envelope (1.5) to zero at the rate $O(k^{-1/4})$, where the size of the gradient is measured in the local norm induced by Φ. As a consequence, we obtain new convergence guarantees for stochastic proximal

[1] The stated assumption (A4) in [15] is stronger than (1.4); however, a quick look at the arguments shows that property (1.4) suffices to obtain essentially the same convergence guarantees.

point, mirror descent[2], and regularized Gauss-Newton methods, as well as for an elementary algorithm for stochastic saddle point problems. Perhaps the most important application arena is when the functional components of the problem grow at a polynomial rate. In this setting, we present a simple Legendre function Φ that satisfies the necessary assumptions for the convergence guarantees to take hold. We also note that the stochastic mirror descent algorithm that we present here does not require mini-batching the gradients, in contrast to the previous seminal work [24].

When the stochastic models $f_x(\cdot, \xi)$ are themselves convex and globally under-estimate f in expectation, we prove that the scheme drives the expected functional error to zero at the rate $O(k^{-1/2})$. The rate improves to $\widetilde{O}(k^{-1})$ when the regularizer r is μ-strongly convex relative to Φ in the sense of [30]. In the special case of mirror descent, these guarantees extend the results for convex unconstrained problems in [29] to the proximal setting. Even specializing to the proximal subgradient method, the convergence guarantees appear to be different from those available in the literature. Namely, previous complexity estimates [7, 20] depend on the largest norms of the subgradients of r along the iterate sequence, whereas Theorems 7.2 and 7.4 replace this dependence only by the initial error $r(x_0) - \inf r$.

The outline of the manuscript is as follows. Section 2 reviews the relevant concepts of convex analysis, focusing on Legendre functions and the Bregman divergence. Section 3 introduces the problem class and the algorithmic framework. This section also interprets the assumptions made for the stochastic proximal point, mirror descent, and regularized Gauss-Newton methods, as well as for a stochastic approximation algorithm for saddle point problems. Section 4 discusses the stationarity measure we use to quantify the rate of convergence. Section 5 contains the complete convergence analysis of the stochastic model-based algorithm. Section 6 presents a specialized analysis for the mirror descent algorithm when f is smooth and the stochastic gradient oracle has finite variance. Finally, in Section 7 we prove convergence rates in terms of function values for stochastic model-based algorithms under (relative strong) convexity assumptions.

2 Legendre functions and the Bregman divergence

Throughout, we follow standard notation from convex analysis, as set out for example by Rockafellar [37]. The symbol $\mathbb{R}^d$ will denote an Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$. For any set $Q \subset \mathbb{R}^d$, we let $\operatorname{int} Q$ and $\operatorname{cl} Q$ denote the interior and closure of Q, respectively. Whenever Q is convex, the set $\operatorname{ri} Q$ is the interior of Q relative to its affine hull. The effective domain of any function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, denoted by $\operatorname{dom} f$, consists of all points where f is finite. Abusing notation slightly, we will use the symbol $\operatorname{dom}(\nabla f)$ to denote the set of all points where f is differentiable.

This work analyzes stochastic model-based minimization algorithms, where the "errors" are controlled by a Bregman divergence. For wider uses of the Bregman divergence in first-order methods, we refer the interested reader to the expository articles of Bubeck [10], Juditsky-Nemirovski [27], and Teboulle [40].

[2]This work appears on arXiv a month after a preprint of Zhang and He [42], who provide similar convergence guarantees specifically for the stochastic mirror descent algorithm. The results of the two papers were obtained independently and are complementary to each other.

Henceforth, we fix a *Legendre function* $\Phi\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, meaning:

1. (Convexity) Φ is proper, closed, and strictly convex.
2. (Essential smoothness) The domain of Φ has nonempty interior, Φ is differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$, and for any sequence $\{x_k\} \subset \operatorname{int}(\operatorname{dom}\Phi)$ converging to a boundary point of $\operatorname{dom}\Phi$, it must be the case that $\|\nabla\Phi(x_k)\| \to \infty$.

Typical examples of Legendre functions are the squared Euclidean norm $\Phi(x) = \frac{1}{2}\|x\|_2^2$, the Shannon entropy $\Phi(x) = \sum_{i=1}^d x_i \log(x_i)$ with $\operatorname{dom}\Phi = \mathbb{R}^d_+$, and the Burge function $\Phi(x) = -\sum_{i=1}^d \log(x_i)$ with $\operatorname{dom}\Phi = \mathbb{R}^d_{++}$. For more examples, we refer the reader to the articles [1, 3, 22, 39] and the recent survey [40].

We will often use the observation that the subdifferential of a Legendre function Φ is empty on the boundary of its domain [37, Theorem 26.1]:

$$\partial\Phi(x) = \emptyset \qquad \text{for all } x \notin \operatorname{int}(\operatorname{dom}\Phi).$$

The Legendre function Φ induces the *Bregman divergence*

$$D_\Phi(y, x) := \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle,$$

for all $x \in \operatorname{int}(\operatorname{dom}\Phi),\ y \in \operatorname{dom}\Phi$. Notice that since Φ is strictly convex, equality $D_\Phi(y, x) = 0$ holds for some $x, y \in \operatorname{int}(\operatorname{dom}\Phi)$ if and only if $y = x$. Analysis of algorithms based on the Bregman divergence typically relies on the following three point inequality; see e.g. [41, Property 1].

Lemma 2.1 (Three point inequality)**.** *Consider a closed convex function $g\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ satisfying* $\operatorname{ri}(\operatorname{dom} g) \subset \operatorname{int}(\operatorname{dom}\Phi)$*. Then for any point $z \in \operatorname{int}(\operatorname{dom}\Phi)$, any minimizer z^+ of the problem*

$$\min_x \ g(x) + D_\Phi(x, z),$$

lies in $\operatorname{int}(\operatorname{dom}\Phi)$*, is unique, and satisfies the inequality:*

$$g(x) + D_\Phi(x, z) \geq g(z_+) + D_\Phi(z_+, z) + D_\Phi(x, z_+) \qquad \forall x \in \operatorname{dom}\Phi.$$

Recall that a function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is called ρ-weakly convex if the perturbed function $f + \frac{\rho}{2}\|\cdot\|_2^2$ is convex [34]. By analogy, we will say that f is *ρ-weakly convex relative to Φ* if the perturbed function $f + \rho\Phi$ is convex. This notion is closely related to the relative smoothness condition introduced in [2, 30].

Relative weak convexity, like its classical counterpart, can be caracterized through generalized derivatives. Recall that the *Fréchet subdifferential* of a function f at a point $x \in \operatorname{dom} f$, denoted $\hat{\partial} f(x)$, consists of all vectors $v \in \mathbb{R}^d$ satisfying

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|) \qquad \text{as } y \to x.$$

The *limiting subdifferential* of f at x, denoted $\partial f(x)$, consists of all vectors $v \in \mathbb{R}^d$ such that there exist sequences $x_k \in \mathbb{R}^d$ and $v_k \in \hat{\partial} f(x_k)$ satisfying $(x_k, f(x_k), v_k) \to (x, f(x), v)$.

Lemma 2.2 (Subdifferential characterization)**.**
The following are equivalent for any locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$.

1. *The function f is ρ-weakly convex relative to Φ.*
2. *For any $x \in \operatorname{int}(\operatorname{dom}\Phi), y \in \operatorname{dom}\Phi$ and any $v \in \hat{\partial} f(x)$, the inequality holds:*
$$f(y) \geq f(x) + \langle v, y - x\rangle - \rho D_\Phi(y, x). \tag{2.1}$$
3. *For any $x \in \operatorname{int}(\operatorname{dom}\Phi) \cap \operatorname{dom}(\nabla f)$, and any $y \in \operatorname{dom}\Phi$, the inequality holds:*
$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle - \rho D_\Phi(y, x). \tag{2.2}$$

If f and Φ are C^2-smooth on $\operatorname{int}(\operatorname{dom}\Phi)$, then the three properties above are all equivalent to
$$\nabla^2 f(x) \succeq -\rho\nabla^2\Phi(x) \qquad \forall x \in \operatorname{int}(\operatorname{dom}\Phi). \tag{2.3}$$

Proof. Define the perturbed function $g := f + \rho\Phi$. We prove the implications $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 1$ in order. To this end, suppose 1 holds. Since g is convex, the subgradient inequality holds:
$$g(y) \geq g(x) + \langle w, y - x\rangle \qquad \text{for all } x, y \in \mathbb{R}^d, w \in \partial g(x). \tag{2.4}$$
Taking into account that Φ is differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$, we deduce $\hat{\partial} g(x) = \hat{\partial} f(x) + \rho\nabla\Phi(x)$ for all $x \in \operatorname{int}(\operatorname{dom}\Phi)$; see e.g. [38, Exercise 8.8]. Rewriting (2.4) with this in mind immediately yields 2. The implication $2 \Rightarrow 3$ is immediate since $\hat{\partial} f(x) = \{\nabla f(x)\}$, whenever f is differentiable at x.

Suppose 3 holds. Fix an arbitrary point $x \in \operatorname{int}(\operatorname{dom}\Phi) \cap \operatorname{dom}(\nabla f)$. Algebraic manipulation of inequality (2.2) yields the equivalent description
$$g(y) \geq g(x) + \langle \nabla f(x) + \rho\nabla\Phi(x), y - x\rangle \qquad \text{for all } y \in \operatorname{dom}\Phi. \tag{2.5}$$
It follows that the vector $\nabla f(x) + \rho\nabla\Phi(x)$ lies in the convex subdifferential of g at x. Since f is locally Lipschitz continuous, Rademacher's theorem shows that $\operatorname{dom}(\nabla f)$ has full measure in $\mathbb{R}^d$. In particular, we deduce from (2.5) that the convex subdifferential of g is nonempty on a dense subset of $\operatorname{int}(\operatorname{dom} g)$. Taking limits, it quickly follows that the convex subdifferential of g is nonempty at every point $x \in \operatorname{int}(\operatorname{dom} g)$ Using [9, Exercise 3.1.12(a)], we conclude that g is convex on $\operatorname{int}(\operatorname{dom} g)$. Moreover, appealing to the sum rule [38, Exercise 10.10], we deduce that $\partial g(x) = \emptyset$ for all $x \notin \operatorname{int}(\operatorname{dom}\Phi)$, since $\partial\Phi(x) = \emptyset$ for all $x \notin \operatorname{int}(\operatorname{dom}\Phi)$. Therefore ∂g is a globally monotone map globally. Appealing to [38, Theorem 12.17], we conclude that g is a convex function. Thus item 1 holds. This completes the proof of the equivalences $1 \Leftrightarrow 2 \Leftrightarrow 3$.

Finally suppose that f and Φ are C^2-smooth on $\operatorname{int}(\operatorname{dom}\Phi)$. Clearly, if f is ρ-weakly convex relative to Φ, then second-order characterization of convexity of the function $g = f + \rho\Phi$ directly implies (2.3). Conversely, (2.3) immediately implies that g is convex on the interior of its domain. The same argument using [38, Theorem 12.17], as in the implication $3 \Rightarrow 1$, shows that g is convex on all of $\mathbb{R}^d$. □

Notice that the setup so far has not relied on any predefined norm. Let us for the moment make the common assumption that Φ is 1-strongly convex relative to some norm $\|\cdot\|$ on $\mathbb{R}^d$, which implies

$$D_\Phi(y, x) \geq \tfrac{1}{2}\|y - x\|^2. \tag{2.6}$$

Then using Lemma 2.2, we deduce that to check that f is ρ-weakly convex relative to Φ, it suffices to verify the inequality

$$f(y) \geq f(x) + \langle v, y - x\rangle - \frac{\rho}{2}\|y - x\|^2 \qquad \text{for all } x, y \in \operatorname{dom}\Phi, v \in \partial f(x).$$

Recall that a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called ρ-smooth if it satisfies:

$$\|\nabla f(y) - \nabla f(x)\|_* \leq \rho\|y - x\| \qquad \text{for all } x, y \in \mathbb{R}^d,$$

where $\|\cdot\|_*$ is the dual norm. Thus any ρ-smooth function f is automatically ρ-weakly convex relative to Φ. Our main result will not require Φ to be 1-strongly convex; however, we will impose this assumption in Section 6 where we augment our guarantees for the stochastic mirror descent algorithm under a differentiability assumption.

3 The problem class and the algorithm

We are now ready to introduce the problem class considered in this paper. We will be interested in the optimization problem

$$\min_x \; F(x) := f(x) + r(x) \tag{3.1}$$

where

- $f\colon \mathbb{R}^d \to \mathbb{R}$ is a locally Lipschitz function,
- $r\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a closed function having a convex domain,
- $\Phi\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is some Legendre function satisfying the compatibility conditions:

$$\operatorname{ri}(\operatorname{dom} r) \subseteq \operatorname{int}(\operatorname{dom}\Phi) \qquad \text{and} \qquad \partial(r + \Phi)(x) = \emptyset \text{ for all } x \notin \operatorname{int}(\operatorname{dom}\Phi). \tag{3.2}$$

The first two items are standard and mild. The third stipulates that r must be compatible with Φ. In particular, the inclusion $\operatorname{ri}(\operatorname{dom} r) \subseteq \operatorname{int}(\operatorname{dom}\Phi)$ automatically implies (3.2), whenever r is convex [37, Theorem 23.8], or more generally whenever a standard qualification condition holds.[3] To simplify notation, henceforth set $U := \operatorname{int}(\operatorname{dom}\Phi)$.

[3]Qualification condition: $\partial^\infty r(x) \cap -N_{\operatorname{dom}\Phi}(x) = \{0\}$, for all $x \in \operatorname{dom} r \cap \operatorname{dom}\Phi$; see [38, Proposition 8.12, Corollary 10.9].

3.1 Assumptions and the Algorithm

We now specify the model-based algorithms we will analyze. Fix a probability space $(\Omega, \mathcal{F}, P)$ and equip $\mathbb{R}^d$ with the Borel $\sigma-$algebra. To each point $x \in \operatorname{dom} f$ and each random element $\xi \in \Omega$, we associate a stochastic one-sided model $f_x(\cdot, \xi)$ of the function f. Namely, we assume that there exist $\tau, \rho, \mathsf{L} > 0$ satisfying the following properties.

(A1) **(Sampling)** It is possible to generate i.i.d. realizations $\xi_1, \ldots, \xi_T \sim P$

(A2) **(One-sided accuracy)** There is a measurable function $(x, y, \xi) \mapsto f_x(y, \xi)$ defined on $U \times U \times \Omega$ satisfying both

$$\mathbb{E}_\xi \left[f_x(x, \xi) \right] = f(x), \qquad \forall x \in U \cap \operatorname{dom} r$$

and

$$\mathbb{E}_\xi \left[f_x(y, \xi) - f(y) \right] \leq \tau D_\Phi(y, x), \qquad \forall x, y \in U \cap \operatorname{dom} r. \tag{3.3}$$

(A3) **(Weak convexity of the models)** The functions $f_x(\cdot, \xi) + r(\cdot)$ are ρ-weakly convex relative to Φ for all $x \in U \cap \operatorname{dom} r$, and a.e. $\xi \in \Omega$.

(A4) **(Lipschitzian property)** There exists a square integrable function $L \colon \Omega \to \mathbb{R}_+$ such that for all $x, y \in U \cap \operatorname{dom} r$, the following inequalities hold:

$$\begin{aligned} f_x(x, \xi) - f_x(y, \xi) &\leq L(\xi) \sqrt{D_\Phi(y, x)}, \\ \sqrt{\mathbb{E}_\xi \left[L(\xi)^2 \right]} &\leq \mathsf{L}. \end{aligned} \tag{3.4}$$

Some comments are in order. Assumption (A1) is standard and is necessary for all sampling based algorithms. Assumption (A2) specifies the accuracy of the models. That is, we require the model in expectation to agree with f at the basepoint, and to globally lower-bound f up to an error controlled by the Bregman divergence. Assumption (A3) is very mild, since in most practical circumstances the function $f_x(\cdot, \xi) + r(\cdot)$ is convex, i.e. $\rho = 0$. The final Assumption (A4) controls the order of growth of the individual models $f_x(y, x)$ as the argument y moves away from x.

Notice that the assumptions (A1)-(A4) do not involve any norm on $\mathbb{R}^d$. However, when Φ is 1-strongly convex relative to some norm, the properties (3.3) and (3.4) are implied by standard assumptions. Namely (3.3) holds if the error in the model approximation satisfies

$$\mathbb{E}_\xi \left[f_x(y, \xi) - f(y) \right] \leq \frac{\tau}{2} \|y - x\|^2, \qquad \forall x, y \in U.$$

Similarly (3.4) will hold as long as for every $x \in U \cap \operatorname{dom} r$ and a.e. $\xi \in \Omega$ the models $f_x(\cdot, \xi)$ are $L(\xi)$-Lipschitz continuous on U in the norm $\| \cdot \|$. The use of the Bregman divergence allows for much greater flexibility as it can, for example, model higher order growth of the functions in question. To illustrate, let us look at the following example where the Lipschitz constant $L(\xi)$ of the models $f_x(\cdot, \xi)$ is bounded by a polynomial.

Example 3.1 (Bregman divergence under polynomial growth)**.** Consider a degree n univariate polynomial

$$p(u) = \sum_{i=0}^{n} a_i u^i,$$

with coefficients $a_i \geq 0$. Suppose now that the one-sided Lipschitz constants of the models satisfy the growth property:

$$\frac{f_x(x,\xi) - f_x(y,\xi)}{\|x-y\|_2} \leq L(\xi)\sqrt{\frac{p(\|x\|_2) + p(\|y\|_2)}{2}} \qquad \text{for all distinct } x, y \in \mathbb{R}^d.$$

Motivated by [29, Proposition 5.1], the following proposition constructs a Bregman divergence that is well-adapted to the polynomial $p(\cdot)$. We defer its proof to Appendix A.1. In particular, with the choice of the Legendre function Φ in (3.5), the required estimate (3.4) holds.

Proposition 3.2. *Define the convex function*

$$\Phi(x) = \sum_{i=0}^{n} a_i \left(\frac{3i+7}{i+2}\right) \|x\|_2^{i+2}. \tag{3.5}$$

Then for all $x, y \in \mathbb{R}^d$, we have

$$D_\Phi(y,x) \geq \frac{p(\|x\|_2) + p(\|y\|_2)}{2} \cdot \|x-y\|_2^2,$$

and therefore the estimate (3.4) *holds.*

The final ingredient we need before stating the algorithm is an estimate on the weak convexity constant of F. The following simple lemma shows that Assumptions (A2) and (A3) imply that F itself is $(\tau+\rho)$-weakly convex relative to Φ.

Lemma 3.3. *The function F is $(\tau+\rho)$-weakly convex relative to Φ.*

Proof. We first show that the function $g := F + (\rho+\tau)\Phi$ is convex on $\operatorname{ri}(\operatorname{dom} F)$. To this end, fix arbitrary points $x, y \in \operatorname{ri}(\operatorname{dom} g)$, and note the equality $\operatorname{ri}(\operatorname{dom} g) = U \cap \operatorname{ri}(\operatorname{dom} r)$ [37, Theorem 6.5]. Choose $\lambda \in (0,1)$ and set $\bar{x} = \lambda x + (1-\lambda)y$. Taking into account (A3), we deduce

$$\begin{aligned}
g(\bar{x}) &= f(\bar{x}) + r(\bar{x}) + (\rho+\tau)\Phi(\bar{x}) \\
&= \mathbb{E}_\xi[f_{\bar{x}}(\bar{x},\xi) + r(\bar{x}) + \rho\Phi(\bar{x})] + \tau\Phi(\bar{x}) \\
&\leq \mathbb{E}_\xi[\lambda(f_{\bar{x}}(x,\xi) + r(x) + \rho\Phi(x)) + (1-\lambda)(f_{\bar{x}}(y,\xi) + r(y) + \rho\Phi(y))] + \tau\Phi(\bar{x}) \\
&= \lambda\mathbb{E}_\xi[f_{\bar{x}}(x,\xi) + r(x)] + (1-\lambda)\mathbb{E}_\xi[f_{\bar{x}}(y,\xi) + r(y)] + \tau\Phi(\bar{x}) + \lambda\rho\Phi(x) + (1-\lambda)\rho\Phi(y) \\
&= \lambda\mathbb{E}_\xi[f_{\bar{x}}(x,\xi) + r(x) - \tau D_\Phi(x,\bar{x})] + (1-\lambda)\mathbb{E}_\xi[f_{\bar{x}}(y,\xi) + r(y) - \tau D_\Phi(y,\bar{x})] \\
&\quad + \lambda\tau(\Phi(\bar{x}) + D_\Phi(x,\bar{x})) + (1-\lambda)\tau(\Phi(\bar{x}) + D_\Phi(y,\bar{x})) + \lambda\rho\Phi(x) + (1-\lambda)\rho\Phi(y).
\end{aligned} \tag{3.6}$$

Now observe

$$\Phi(\bar{x}) + D_\Phi(x,\bar{x}) = \Phi(x) - (1-\lambda)\langle \nabla\Phi(\bar{x}), x-y\rangle,$$

and similarly

$$\Phi(\bar{x}) + D_\Phi(y, \bar{x}) = \Phi(y) - \lambda\langle \nabla\Phi(\bar{x}), y - x\rangle.$$

Hence algebraic manipulation of the two equalities above yields the expression

$$\lambda\tau(\Phi(\bar{x}) + D_\Phi(x, \bar{x})) + (1-\lambda)\tau(\Phi(\bar{x}) + D_\Phi(y, \bar{x})) = \lambda\tau\Phi(x) + (1-\lambda)\tau\Phi(y).$$

Continuing with (3.6), we obtain

$$\begin{aligned} g(\bar{x}) &\le \lambda f(x) + r(x) + (1-\lambda)(f(y) + r(y)) \\ &\quad + \lambda\tau\Phi(x) + (1-\lambda)\tau\Phi(y) + \lambda\rho\Phi(x) + (1-\lambda)\rho\Phi(y) \\ &= \lambda[f(x) + r(x) + (\tau+\rho)\Phi(x)] + (1-\lambda)[f(y) + r(y) + (\tau+\rho)\Phi(y)] \\ &\le \lambda g(x) + (1-\lambda) g(y). \end{aligned}$$

We have thus verified that g is convex on $\operatorname{ri}(\operatorname{dom} g)$. Appealing to (3.2) and the sum rule [38, Exercise 10.10], we deduce that the subdifferential $\partial g(x)$ is empty at every point in $x \notin \operatorname{ri}(\operatorname{dom} g)$, and therefore ∂g is a globally monotone map. Using [38, Theorem 12.17], we conclude that g is a convex function, as needed. □

In light of Lemma 3.3, we also make the following additional assumption on the solvability of the Bregman proximal subproblems.

(A5) **(Solvability)** The convex problems

$$\min_y \left\{ F(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\} \qquad \text{and} \qquad \min_y \left\{ f_x(y, \xi) + r(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\},$$

admit a minimizer for any $\lambda < (\tau+\rho)^{-1}$, any $x \in U$, and a.e. $\xi \in \Omega$.[4] The minimizers vary measurably in $(x, \xi) \in U \times \Omega$.

Assumption (A5) is very mild. In particular, it holds automatically if (i) Φ is strongly convex with respect to some norm, or if (ii) the functions $f_x(\cdot, \xi) + r(\cdot) + \rho D_\Phi(\cdot, x)$ and $F + (\tau+\rho)\Phi$ are bounded from below and Φ has bounded sublevel sets [40, Lemma 2.3].

We are now ready to state the stochastic model-based algorithm we analyze—Algorithm 1.

Algorithm 1: Stochastic Model Based Minimization

Data: $x_0 \in U \cap \operatorname{dom} r$, real $\lambda < (\tau+\rho)^{-1}$, a nonincreasing sequence $\{\eta_t\}_{t\ge 0} \subseteq (0, \lambda)$, and iteration count T.

Step $t = 0, \ldots, T$:

$$\left\{ \begin{array}{l} \text{Sample } \xi_t \sim P \\ \text{Set } x_{t+1} = \operatorname*{argmin}_x \ \left\{ f_{x_t}(x, \xi_t) + r(x) + \frac{1}{\eta_t} D_\Phi(x, x_t) \right\} \end{array} \right\},$$

Sample $t^* \in \{0, \ldots, T\}$ according to the discrete probability distribution

$$\mathbb{P}(t^* = t) \propto \frac{\eta_t}{1 - \eta_t \rho}.$$

Return x_{t^*}

[4] Note the minimizers are automatically unique by Lemma 2.1

3.2 Examples

Before delving into the convergence analysis of Algorithm 1, in this section we illustrate the algorithmic framework on four examples. In all cases, assumptions (A1) and (A5) are self-explanatory. Therefore, we only focus on verifying (A2)-(A4). For simplicity, we also assume that $r(\cdot)$ is convex in all examples.

Stochastic Bregman-proximal point. Suppose that the models $(x, y, \xi) \mapsto f_x(y, \xi)$ satisfy

$$\mathbb{E}_\xi[f_x(y, \xi)] = f(y) \qquad \forall x, y \in U \cap \operatorname{dom} r.$$

With this choice of the models, Algorithm 1 becomes the stochastic Bregman-proximal point method. Analysis of the deterministic version of the method for convex problems goes back to [13,14,22]. Observe that Assumption (A2) holds trivially. Assumption (A3) and Assumption (A4) should be verified in particular circumstances, depending on how the models are generated. In particular, one can verify Assumption (A4) under polynomial growth of the Lipschitz constant, by appealing to Example 3.1.

Stochastic mirror descent. Suppose that the models $(x, y, \xi) \mapsto f_x(y, \xi)$ are given by

$$f_x(y, \xi) = f(x) + \langle G(x, \xi), y - x\rangle,$$

for some measurable mapping $G\colon U \times \Omega \to \mathbb{R}^d$ satisfying $\mathbb{E}_\xi[G(x, \xi)] \in \partial f(x)$ for all $x \in U \cap \operatorname{dom} r$. Algorithm 1 then becomes the stochastic mirror descent algorithm, classically studied in [6,31] in the convex setting and more recently analyzed in [2,29,30] under convexity and relative continuity assumptions. Assumption (A2) simply says that f is τ-weakly convex relative to Φ, while Assumption (A3) holds trivially with $\rho = 0$. Assumption (A4) is directly implied by the relative continuity condition of Lu [29]. Namely it suffices to assume that there is a square integrable function $L\colon \Omega \to \mathbb{R}_{++}$ satisfying

$$\|G(x, \xi)\|_* \leq L(\xi)\frac{\sqrt{D(y, x)}}{\|y - x\|} \qquad \forall x, y \in U,$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^d$, and $\|\cdot\|_*$ is the dual norm. We refer to [29] for more details on this condition and examples.

Gauss-Newton method with Bregman regularization. In the next example, suppose that f has the composite form

$$f(x) = \mathbb{E}_\xi[h(c(x, \xi), \xi)],$$

for some measurable function $h(y, \xi)$ that is convex in y for a.e. $\xi \in \Omega$ and a measurable map $c(x, \xi)$ that is C^1-smooth in x for a.e. $\xi \in \Omega$. We may then use the convex models

$$f_x(y, \xi) := h\left(c(x, \xi) + \nabla c(x, \xi)(y - x), \xi\right), \xi\right),$$

which automatically satisfy (A3) with $\rho = 0$. Algorithm 1 then becomes a stochastic Gauss-Newton method with Bregman regularization.

In the Euclidean case $\Phi = \frac{1}{2}\|\cdot\|^2$, the method reduces to the stochastic prox-linear algorithm, introduced in [21] and further analyzed in [15]. The deterministic prox-linear method has classical roots, going back at least to [11,23,36], while a more modern complexity theoretic perspective appears in [12,18,19,28,32]. Even in the deterministic setting, to make progress, one typically assumes that h and ∇c are globally Lipschitz. More generally and in line with our current work, one may introduce a different Legendre function Φ. For example, in the case of polynomial growth, the following propositions construct Legendre functions that are compatible with Assumptions (A2) and (A4). We defer their proofs to Appendix A.3. In the two propositions, we assume that the outer functions $h(\cdot,\xi)$ are globally Lipschitz, while the inner maps $c(\cdot,\xi)$ may have a high order of growth. It is possible to also analyze the setting when $h(\cdot,\xi)$ has polynomial growth, but the resulting statements and assumptions become much more cumbersome; we therefore omit that discussion.

Proposition 3.4 (Satisfying (A2))**.** *Suppose there are square integrable functions $L_1, L_2 \colon \Omega \to \mathbb{R}_+$ and a univariate polynomial $p(u) = \sum_{i=0}^n a_i u^i$ with nonnegative coefficients satisfying*

$$\begin{aligned} \frac{|h(v,\xi) - h(w,\xi)|}{\|v-w\|_2} &\le L_1(\xi) \qquad \forall v \ne w, \\ \frac{\|\nabla c(x,\xi) - \nabla c(y,\xi)\|_{\mathrm{op}}}{\|x-y\|_2} &\le L_2(\xi)(p(\|x\|_2) + p(\|y\|_2)) \qquad \forall x \ne y. \end{aligned}$$

Define the Legendre function $\Phi(x) := \sum_{i=0}^n \frac{a_i(3i+7)}{i+2}\|x\|_2^{i+2}$. Then assumption (A2) holds with $\tau := \frac{4}{3}\mathbb{E}\left[L_1(\xi)L_2(\xi)\right]$.

Proposition 3.5 (Satisfying (A4))**.** *Suppose there are square integrable functions $L_1, L_2 \colon \Omega \to \mathbb{R}_+$ and a univariate polynomial $q(u) = \sum_{i=0}^n b_i u^i$ with nonnegative coefficients satisfying*

$$\begin{aligned} \frac{|h(v,\xi) - h(w,\xi)|}{\|v-w\|_2} &\le L_1(\xi) \qquad \forall v \ne w, \\ \|\nabla c(x,\xi)\|_{\mathrm{op}} &\le L_2(\xi) \cdot \sqrt{q(\|x\|_2)} \qquad \forall x, \xi. \end{aligned}$$

Then with the Legendre function $\Phi(x) = \sum_{i=0}^n \frac{b_i}{i+2}\|x\|_2^{i+2}$, assumption (A4) holds with $L(\xi) = \sqrt{2}L_1(\xi)L_2(\xi)$.

To construct a Bregman function compatible with both (A2) and (A4) simultaneously, one may simply add the two Legendre functions constructed in Propositons 3.4 and 3.5.

Stochastic saddle point problems. As the final example, suppose that f is given in the stochastic conjugate form

$$f(x) = \mathbb{E}\left[\sup_{w \in W} g(x,w,\xi)\right],$$

where W is some auxiliary set and $g \colon \mathbb{R}^d \times W \times \Omega \to \mathbb{R}$ is some function. Thus we are interested in solving the stochastic saddle-point problem

$$\inf_x \ \mathbb{E}\left[\sup_{w \in W} g(x,w,\xi)\right] + r(x). \tag{3.7}$$

Such problems appear often in data science, where the variation of w in the "uncertainty set" W makes the loss function robust. One popular example is adversarial training [25]. In this setting, we have $g(x, w, \xi) = \mathcal{L}(x + w, y, \xi)$, where $\mathcal{L}(\cdot, \cdot)$ is a loss function, y encodes the observed data, and w varies over some uncertainty set W, such as an ℓ_p-ball.

In order to apply our algorithmic framework, we must have access to stochastic one-sided models $f_x(\cdot, \xi)$ of f. It is quite natural to construct such models by using one-sided stochastic models $g_x(\cdot, w, \xi)$ of g. Indeed, it is appealing to simply set

$$f_x(y, \xi) = g_x(y, \widehat{w}(x, \xi), \xi) \qquad \text{for any} \qquad \widehat{w}(x, \xi) \in \operatorname*{argmax}_w \; g_x(x, w, \xi). \tag{3.8}$$

All of the model types in the previous examples could now serve as the models $g_x(\cdot, w, \xi)$, provided they meet the conditions outlined below.

Formally, to ensure that (A1)-(A5) hold for the models $f_x(y, \xi)$, we must make the following assumptions:

1. The mapping $(x, \xi) \to \sup_{w \in W} g(x, w, \xi)$ is measurable and has finite first moment for every fixed $x \in U \cap \operatorname{dom} r$.

2. The function $g_x(\cdot, w, \xi)$ is ρ-weakly convex relative to Φ, for every fixed $x \in U \cap \operatorname{dom} r$, $w \in W$, and a.e. $\xi \in \Omega$.

3. There exists a mapping $\widehat{w} \colon U \times \Omega \to \mathbb{R}^m$ satisfying

   $$\widehat{w}(x, \xi) \in \operatorname*{argmax}_w g_x(x, w, \xi),$$

   for all $x \in U \cap \operatorname{dom} r$ and a.e. $\xi \in \Omega$ with the property that the functions $(x, y, \xi) \mapsto g_x(y, \widehat{w}(x, \xi), \xi)$ and $(x, y, \xi) \mapsto g(y, \widehat{w}(x, \xi), \xi)$ are measurable.

4. For all $x, y \in U \cap \operatorname{dom} r$, we have

   $$\mathbb{E}_\xi \left[g_x(x, \widehat{w}(x, \xi), \xi) \right] = \mathbb{E}_\xi \left[g(x, \widehat{w}(x, \xi), \xi) \right]$$

   and

   $$\mathbb{E} \left[g_x(y, \widehat{w}(x, \xi), \xi) - g(y, \widehat{w}(x, \xi), \xi) \right] \leq \tau D_\Phi(y, x).$$

5. There exists a square integrable function $L \colon \Omega \to \mathbb{R}_+$ such that

   $$g_x(x, \widehat{w}(x, \xi), \xi) - g_x(y, \widehat{w}(x, \xi), \xi) \leq L(\xi) \sqrt{D_\Phi(y, x)}, \qquad \text{for all } x, y \in U \cap \operatorname{dom} r.$$

Given these assumptions, let us define $f_x(y, \xi)$ as in (3.8) We now verify properties (A2)-(A4). Property (A2) follows from Property 4, which implies that $\mathbb{E} \left[f_x(x, \xi) \right] = f(x)$ and

$$\begin{aligned} \mathbb{E}_\xi \left[f_x(y, \xi) - f(y) \right] &= \mathbb{E}_\xi \left[g_x(y, \widehat{w}(x, \xi), \xi) - \sup_{w \in W} g(y, w, \xi) \right] \\ &\leq \mathbb{E}_\xi \left[g_x(y, \widehat{w}(x, \xi), \xi) - g(y, \widehat{w}(x, \xi), \xi) \right] \\ &\leq \tau D_\Phi(y, x). \end{aligned}$$

Property (A3) follows directly from Property 2. Finally, (A4) follows from Property 5.

4 Stationarity measure

In this section, we introduce a natural stationarity measure that we will use to describe the convergence rate of Algorithm 1. The stationarity measure is simply the size of the gradient of an appropriate smooth approximation of the problem (3.1). This idea is completely analogous to the Euclidean setting [15, 16]. Setting the stage, for any $\lambda > 0$, define the Φ-*envelope*

$$F_\lambda^\Phi(x) := \inf_y \left\{ f(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\},$$

and the associated Φ-*proximal map*

$$\operatorname{prox}_{\lambda f}^\Phi(x) := \operatorname*{argmin}_y \left\{ F(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\}.$$

Note that in the Euclidean setting $\Phi = \frac{1}{2}\|\cdot\|^2$, these two constructions reduce to the standard Moreau envelope and the proximity map; see for example the monographs [35, 38] or the note [17] for recent perspectives.

We will measure the convergence guarantees of Algorithm 1 based on the rate at which the quantity

$$\mathbb{E}[D_\Phi\left(\operatorname{prox}_{\lambda F}^\Phi(x_{t^*}), x_{t^*}\right)] \tag{4.1}$$

tends to zero for some fixed $\lambda > 0$. The significance of this quantity becomes apparent after making slightly stronger assumptions on the Legendre function Φ. In this section only, suppose that $\Phi \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is 1-strongly convex with respect to some norm $\|\cdot\|$ and that Φ is twice differentiable at every point in $\operatorname{int}(\operatorname{dom}\Phi)$. With these assumptions, the following result shows that the Φ-envelope is differentiable, with a meaningful gradient. Indeed, this result follows quickly from [4]. For the sake of completeness, we present a self-contained argument in Appendix A.4.

Theorem 4.1 (Smoothness of the Φ-envelope)**.** *For any positive $\lambda < (\tau + \rho)^{-1}$, the envelope F_λ^Φ is differentiable at any point $x \in \operatorname{int}(\operatorname{dom}\Phi)$ with gradient given by*

$$\nabla F_\lambda^\Phi(x) := \frac{1}{\lambda} \nabla^2 \Phi(x) \left(x - \operatorname{prox}_{\lambda F}^\Phi(x)\right).$$

In light of Theorem 4.1, for any point $x \in \operatorname{int}(\operatorname{dom}\Phi)$, we may define the local norm

$$\|y\|_x := \left\|\nabla^2 \Phi(x) y\right\|_*.$$

Then a quick computation shows that the dual norm is given by

$$\|v\|_x^* = \left\|\nabla^2 \Phi(x)^{-1} v\right\|.$$

Therefore appealing to Theorem 4.1, for any positive $\lambda < (\tau + \rho)^{-1}$ and $x \in \operatorname{int}(\operatorname{dom}\Phi)$ we obtain the estimate

$$\sqrt{D_\Phi\left(\operatorname{prox}_{\lambda F}^\Phi(x), x\right)} \geq \frac{\lambda}{\sqrt{2}} \|\nabla F_\lambda^\Phi(x)\|_x^*.$$

Thus the square root of the Bregman divergence, which we will show tends to zero along the iterate sequence at a controlled rate, bounds the local norm of the gradient ∇F_λ^Φ.

5 Convergence analysis

We now present convergence analysis of Algorithm 1 under Assumptions (A1)-(A5). Henceforth, let $\{x_t\}_{t\geq 0}$ be the iterates generated by Algorithm 1 and let $\{\xi_t\}_{t\geq 0}$ be the corresponding samples used. For each index $t \geq 0$, define the Bregman-proximal point

$$\hat{x}_t = \operatorname{prox}_{\lambda F}^{\Phi}(x_t).$$

To simplify notation, we will use the symbol $\mathbb{E}_t[\cdot]$ to denote the expectation conditioned on all the realizations $\xi_0, \xi_1, \ldots, \xi_{t-1}$. The entire argument of Theorem 5.2—our main result—relies on the following lemma.

Lemma 5.1. *For each iteration $t \geq 0$, the iterates of Algorithm 1 satisfy*

$$\mathbb{E}_t\left[D_\Phi(\hat{x}_t, x_{t+1})\right] \leq \tfrac{1+\eta_t\tau-\eta_t/\lambda}{1-\eta_t\rho} D_\Phi(\hat{x}_t, x_t) + \tfrac{(\mathsf{L}\eta_t)^2}{4(1-\eta_t\rho)} + \tfrac{\eta_t}{1-\eta_t\rho}\mathbb{E}_t\left[r(x_t) - r(x_{t+1})\right].$$

Proof. Taking into account assumption (A3), we may apply the three point inequality in Lemma 2.1 with the convex function $g = f_{x_t}(\cdot, \xi_t) + r(\cdot) + \rho D_\Phi(\cdot, x_t)$ and with $(\frac{1}{\eta_t} - \rho) D_\Phi(\cdot, x_t)$ replacing the Bregman divergence. Thus for any point $x \in \operatorname{int}(\operatorname{dom}\Phi)$, we obtain the estimate

$$f_{x_t}(x, \xi_t) + r(x) + \frac{1}{\eta_t} D_\Phi(x, x_t) \geq f_{x_t}(x_{t+1}, \xi_t) + r(x_{t+1}) + \frac{1}{\eta_t} D_\Phi(x_{t+1}, x_t) + \left(\frac{1}{\eta_t} - \rho\right) D_\Phi(x, x_{t+1}). \tag{5.1}$$

Setting $x = \hat{x}_t$, rearranging terms, and taking expectations, we deduce

$$\begin{aligned} &\mathbb{E}_\xi[f_{x_t}(\hat{x}_t, \xi_t) + r(\hat{x}_t) - f_{x_t}(x_{t+1}, \xi_t) - r(x_{t+1})] \\ &\qquad \geq \frac{1}{\eta_t}\mathbb{E}_t\left[(1-\eta_t\rho) D_\Phi(\hat{x}_t, x_{t+1}) - D_\Phi(\hat{x}_t, x_t) + D_\Phi(x_{t+1}, x_t)\right]. \end{aligned} \tag{5.2}$$

We seek to upper bound the left-hand-side of (5.2). Using assumptions (A2) and (A4), we obtain:

$$\begin{aligned} &\mathbb{E}_t\left[f_{x_t}(\hat{x}_t, \xi_t) - f_{x_t}(x_{t+1}, \xi_t)\right] \\ &\quad \leq \mathbb{E}_t\left[f_{x_t}(\hat{x}_t, \xi_t) - f_{x_t}(x_t, \xi_t) + L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)}\right] \\ &\quad = \mathbb{E}_t\left[f_{x_t}(\hat{x}_t, \xi_t) - f(\hat{x}_t)\right] + \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)}\right] - f(x_t) + f(\hat{x}_t) \\ &\quad \leq \tau D_\Phi(\hat{x}_t, x_t) + \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)}\right] - f(x_t) + f(\hat{x}_t). \end{aligned} \tag{5.3}$$

By the definition of $\hat{x}_t$ as the Bregman-proximal point, we have

$$f(\hat{x}_t) + r(\hat{x}_t) + \frac{1}{\lambda} D_\Phi(\hat{x}_t, x_t) \leq f(x_t) + r(x_t). \tag{5.4}$$

The right hand side of (5.2) is thus upper bounded by

$$\begin{aligned} &\tau D_\Phi(\hat{x}_t, x_t) + \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)} - f(x_t) - r(x_{t+1})\right] + f(\hat{x}_t) + r(\hat{x}_t) \\ &\leq \tau D_\Phi(\hat{x}_t, x_t) + \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)} + (r(x_t) - r(x_{t+1}))\right] + f(\hat{x}_t) + r(\hat{x}_t) - f(x_t) - r(x_t) \\ &\leq \left(\tau - \frac{1}{\lambda}\right) D_\Phi(\hat{x}_t, x_t) + \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)} + (r(x_t) - r(x_{t+1}))\right] \end{aligned}$$

where the last inequality follows from (5.4). Combining this estimate with (5.2), we obtain

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1})-D_\Phi(\hat{x}_t,x_t)+D_\Phi(x_{t+1},x_t)\right]\\&\qquad\leq\left(\tau-\frac{1}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}+(r(x_t)-r(x_{t+1}))\right].\end{aligned}$$

Multiplying through by η_t and rearranging yields

$$\begin{aligned}&(1-\eta_t\rho)\mathbb{E}_t\left[D_\Phi(\hat{x}_t,x_{t+1})\right]\\&\qquad\leq\left(1+\eta_t\tau-\frac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\mathbb{E}_t\left[\eta_tL(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}-D_\Phi(x_{t+1},x_t)\right]\\&\qquad\quad+\eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right].\end{aligned}\tag{5.5}$$

Now define $\gamma:=\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}$. Note that Cauchy-Schwarz implies

$$\mathbb{E}_t\left[\eta_tL(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right]\leq\eta_t\mathsf{L}\gamma.$$

Using this estimate in (5.5), we obtain

$$\begin{aligned}(1-\eta_t\rho)\mathbb{E}_t\left[D_\Phi(\hat{x}_t,x_{t+1})\right]\leq&\left(1+\eta_t\tau-\frac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\eta_t\mathsf{L}\gamma-\gamma^2\\&+\eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right].\end{aligned}$$

Maximizing the right hand side in γ (i.e. taking $\gamma=\frac{\mathsf{L}\eta_t}{2}$), yields the guarantee

$$(1-\eta_t\rho)\mathbb{E}_t\left[D_\Phi(\hat{x}_t,x_{t+1})\right]\leq\left(1+\eta_t\tau-\frac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\frac{(\mathsf{L}\eta_t)^2}{4}+\eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right].$$

Dividing through by $1-\eta_t\rho$ completes the proof. □

We can now prove our main theorem.

Theorem 5.2 (Convergence rate)**.** *The point x_{t^*} returned by Algorithm 1 satisfies:*

$$\begin{aligned}&\mathbb{E}\left[D_\Phi\left(\mathrm{prox}^\Phi_{\lambda F}(x_{t^*}),x_{t^*}\right)\right]\\&\qquad\leq\frac{\lambda^2}{1-\lambda(\tau+\rho)}\left(\frac{F^\Phi_\lambda(x_0)-\min F}{\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}}+\frac{\mathsf{L}^2\sum_{t=0}^T\frac{\eta_t^2}{4\lambda(1-\eta_t\rho)}}{\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}}+\frac{\frac{\eta_0}{\lambda(1-\eta_0\rho)}(r(x_0)-\inf r)}{\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}}\right).\end{aligned}$$

Proof. Using the definitions of x_{t+1} and $\hat{x}_t$ along with Lemma 5.1, we obtain

$$\begin{aligned}\mathbb{E}_t\left[F^\Phi_\lambda(x_{t+1})\right]&\leq\mathbb{E}_t\left[F(\hat{x}_t)+\frac{1}{\lambda}D_\Phi(\hat{x}_t,x_{t+1})\right]\\&\leq\mathbb{E}_t\left[F(\hat{x}_t)+\frac{1}{\lambda(1-\eta_t\rho)}\left(\left(1+\eta_t\left(\tau-\frac{1}{\lambda}\right)\right)D_\Phi(\hat{x}_t,x_t)+\frac{(\mathsf{L}\eta_t)^2}{4}\right)\right]\\&\quad+\frac{\eta_t}{\lambda(1-\eta_t\rho)}\mathbb{E}_t\left[(r(x_t)-r(x_{t+1}))\right]\\&=F^\Phi_\lambda(x_t)+\frac{\eta_t}{\lambda}\left(\frac{\tau+\rho-1/\lambda}{1-\eta_t\rho}\right)D_\Phi(\hat{x}_t,x_t)+\frac{(\mathsf{L}\eta_t)^2}{4\lambda(1-\eta_t\rho)}\\&\quad+\frac{\eta_t}{\lambda(1-\eta_t\rho)}\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right].\end{aligned}$$

Recursing and applying the tower rule for expectations, we obtain

$$\begin{aligned}\mathbb{E}\left[F_\lambda^\Phi(x_{T+1})\right] \leq F_\lambda^\Phi(x_0) + \sum_{t=0}^{T}\left(\frac{\eta_t}{\lambda}\left(\frac{\tau+\rho-1/\lambda}{1-\eta_t\rho}\right)\mathbb{E}[D_\Phi(\hat{x}_t,x_t)] + \frac{(\mathsf{L}\eta_t)^2}{4\lambda(1-\eta_t\rho)}\right) \\ + \sum_{t=0}^{T}\frac{\eta_t}{\lambda(1-\eta_t\rho)}\mathbb{E}\left[r(x_t)-r(x_{t+1})\right].\end{aligned} \tag{5.6}$$

Taking into account that η_t is nonincreasing yields the inequality

$$\sum_{t=0}^{T}\tfrac{\eta_t}{\lambda(1-\eta_t\rho)}(r(x_t)-r(x_{t+1})) \leq \tfrac{\eta_0}{\lambda(1-\eta_0\rho)}(r(x_0)-\inf r).$$

See the auxiliary Lemma A.1 for a verification. Combining this bound with (5.6), using the inequality $\mathbb{E}\left[F_\lambda(x_{T+1})\right] \geq \min F$, and rearranging, we conclude

$$\begin{aligned}\frac{1}{\lambda}\left(\frac{1}{\lambda}-\tau-\rho\right)\sum_{t=0}^{T}\frac{\eta_t}{1-\eta_t\rho}\mathbb{E}[D_\Phi(\hat{x}_t,x_t)] \leq F_\lambda^\Phi(x_0) - \min F + \mathsf{L}^2\sum_{t=0}^{T}\frac{\eta_t^2}{4\lambda(1-\eta_t\rho)} \\ + \frac{\eta_0}{\lambda(1-\eta_0\rho)}(r(x_0)-\inf r),\end{aligned}$$

or equivalently

$$\begin{aligned}\sum_{t=0}^{T}\frac{\eta_t}{1-\eta_t\rho}\mathbb{E}\left[D_\Phi(\hat{x}_t,x_t)\right] \leq \frac{\lambda^2(F_\lambda^\Phi(x_0)-\min F)}{1-\lambda(\tau+\rho)} + \frac{\lambda^2\mathsf{L}^2}{1-\lambda(\tau+\rho)}\sum_{t=0}^{T}\frac{\eta_t^2}{4\lambda(1-\eta_t\rho)} \\ + \frac{\lambda^2\eta_0}{\lambda(1-\lambda(\tau+\rho))(1-\eta_0\rho)}(r(x_0)-\inf r).\end{aligned}$$

Dividing through by $\sum_{t=0}^{T}\frac{\eta_t}{1-\eta_t\rho}$ and recognizing the left-hand-side as $\mathbb{E}[D_\Phi(\hat{x}_{t^*},x_{t^*})]$, the result follows. □

As an immediate corollary of Theorem 5.2, we have the following rate of convergence when the stepsize η_t is constant.

Corollary 5.3 (Convergence rate for constant stepsize)**.** *For some $\alpha > 0$, set $\eta_t = \frac{1}{\lambda^{-1}+\alpha^{-1}\sqrt{T+1}}$ for all indices $t = 1,\ldots,T$. Then the point x_{t^*} returned by Algorithm 2 satisfies:*

$$\mathbb{E}\left[D_\Phi\left(\mathrm{prox}_{\lambda F}^\Phi(x_{t^*}),x_{t^*}\right)\right] \leq \frac{\lambda^2(F_\lambda^\Phi(x_0)-\min F)+\frac{\lambda\mathsf{L}^2\alpha^2}{4}+\frac{\lambda((r(x_0)-\inf r))}{\lambda^{-1}-\rho+\alpha^{-1}}}{1-\lambda(\tau+\rho)}\cdot\left(\frac{\lambda^{-1}-\rho}{T+1}+\frac{1}{\alpha\sqrt{T+1}}\right).$$

6 Mirror descent: smoothness and finite variance

Assumptions (A1)-(A5) are reasonable for the examples described in Section 3.2, being in line with standard conditions in the literature. However, in the special case that f is smooth and we apply stochastic mirror descent, Assumption (A4) is nonstandard. Ideally, one would

like to replace this assumption with a bound on the variance of the stochastic estimator of the gradient. In this section, we show that this is indeed possible by slightly modifying the argument in Section 5.

Henceforth, let Φ be a Legendre function and set $U := \operatorname{int}(\operatorname{dom}\Phi)$. In this section, we make the following assumptions:

(B1) **(Sampling)** It is possible to generate i.i.d. realizations $\xi_1, \ldots, \xi_T \sim P$

(B2) **(Stochastic gradient)** There is a measurable mapping $G : U \times \Omega \to \mathbb{R}^d$ satisfying

$$\mathbb{E}_\xi [G(x, \xi)] = \nabla f(x), \qquad \forall x \in U \cap \operatorname{dom} r.$$

(B3) **(Relative Smoothness)** There exist real $\tau, M \geq 0$, such that

$$-\tau D_\Phi(y, x) \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq M D_\Phi(y, x) \qquad \forall x, y \in U \cap \operatorname{dom} r.$$

(B4) **(Relative convexity)** The function r is ρ-weakly convex relative to Φ.

(B5) **(Strong convexity of Φ)** The Legendre function Φ is 1-strongly convex with respect to some norm $\|\cdot\|$.

(B6) **(Finite variance)** The following variance is finite:

$$\mathbb{E}_\xi \left[\|G(x, \xi) - \nabla f(x)\|_*^2 \right] \leq \frac{\sigma^2}{2} < \infty.$$

Henceforth, we denote by $f_x(\cdot, \xi)$ the linear models

$$f_x(y, \xi) := f(x) + \langle G(x, \xi), y - x \rangle,$$

which are built from the stochastic gradient estimator G. With this notation in hand, let us compare Assumptions (B1)-(B5) with Assumptions (A1)-(A4). Evidently, Assumptions (B1) and (A1) are identical. Upon taking expectations, Assumptions (B2) and (B3) imply the stochastic one-sided accuracy property (A2) for the linear models $f_x(\cdot, \xi)$, while (B4) directly implies (A3). Assumptions (B5) and (B6) replace the Lipschitzian property (A4).

Finally, we reiterate that the relative smoothness property in (B3) was recently introduced in [2, 30] for smooth convex minimization, and extended to smooth nonconvex problems in [8] and to nonsmooth stochastic problems in [26, 29]. This property allows for higher order growth than the standard Lipschitz gradient assumptions, commonly analyzed in the literature. We refer the reader to [2,30] for various examples of Bregman functions that arise in applications.

For the sake of clarity, Algorithm 2 instantiates Algorithm 1 in our setting.

Algorithm 2: Mirror descent for smooth minimization

Data: $x_0 \in U \cap \operatorname{dom} r$, positive $\lambda < (\tau+\rho)^{-1}$, a sequence $\{\eta_t\}_{t\geq 0} \subseteq \left(0, \frac{\lambda}{1+\lambda M}\right)$, and iteration count T

Step $t = 0, \ldots, T$:

$$\left\{\begin{array}{l} \text{Sample } \xi_t \sim P \\ \text{Set } x_{t+1} = \operatorname*{argmin}_x \ \left\{\langle G(x_t,\xi_t), x\rangle + r(x) + \frac{1}{\eta_t} D_\Phi(x, x_t)\right\} \end{array}\right\},$$

Sample $t^* \in \{0, \ldots, T\}$ according to the discrete probability distribution

$$\mathbb{P}(t^* = t) \propto \frac{\eta_t}{1-\eta_t\rho}.$$

Return x_{t^*}

As in Section 5, the convergence analysis relies on the following key lemma. We let $\{x_t\}_{t\geq 0}$ be the iterates generated by Algorithm 2 and let $\{\xi_t\}_{t\geq 0}$ be the corresponding samples used. For each index $t \geq 0$, we continue to use the notation $\hat{x}_t = \operatorname{prox}_{\lambda F}^{\Phi}(x)$ and let $\mathbb{E}_t[\cdot]$ to denote the expectation conditioned on all the realizations $\xi_0, \xi_1, \ldots, \xi_{t-1}$.

Lemma 6.1. *For each iteration $t \geq 0$, the iterates of Algorithm 2 satisfy*

$$\mathbb{E}_t\left[D_\Phi(\hat{x}_t, x_{t+1})\right] \leq \frac{1+\eta_t\tau - \eta_t/\lambda}{(1-\eta_t\rho)} \cdot D_\Phi(\hat{x}_t, x_t) + \frac{1}{4} \cdot \frac{(\sigma\eta_t)^2}{(1-\eta_t(M+\frac{1}{\lambda}))(1-\eta_t\rho)}.$$

Proof. Following the initial steps of the proof of Lemma 5.1, we arrive at the estimate (5.2), namely

$$\begin{aligned} &\frac{1}{\eta_t}\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t, x_{t+1}) - D_\Phi(\hat{x}_t, x_t) + D_\Phi(x_{t+1}, x_t)\right] \\ &\qquad \leq \mathbb{E}_t\left[f_{x_t}(\hat{x}_t, \xi_t) + r(\hat{x}_t) - f_{x_t}(x_{t+1}, \xi_t) - r(x_{t+1})\right]. \end{aligned} \tag{6.1}$$

We now seek to bound the right-hand side of (6.1) using (B3)-(B6). To that end, the following bound will be useful:

$$\begin{aligned} f_{x_t}(x_{t+1}, \xi_t) &= f(x_t, \xi_t) + \langle G(x_t, \xi_t), x_{t+1} - x_t\rangle \\ &\geq f(x_t, \xi_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle - \|G(x_t, \xi_t) - \nabla f(x_t)\|_*\|x_{t+1} - x_t\|. \end{aligned}$$

Taking expectations of both sides and applying Cauchy-Schwarz and (B3)-(B6), we obtain

$$\begin{aligned} \mathbb{E}_t\left[f_{x_t}(x_{t+1}, \xi_t)\right] &\geq \mathbb{E}_t\left[f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle\right] - \mathbb{E}_t\left[\|G(x_t,\xi_t) - \nabla f(x_t)\|_*\|x_{t+1} - x_t\|\right] \\ &\geq \mathbb{E}_t\left[f(x_{t+1}) - M D_\Phi(x_{t+1}, x_t)\right] - \sqrt{\mathbb{E}_t\left[\|G(x_t,\xi_t) - \nabla f(x_t)\|_*^2\right]}\sqrt{\mathbb{E}_t\left[\|x_{t+1} - x_t\|^2\right]} \\ &\geq \mathbb{E}_t\left[f(x_{t+1}) - M D_\Phi(x_{t+1}, x_t)\right] - \sigma\sqrt{\mathbb{E}_t\left[\tfrac{1}{2}\|x_{t+1} - x_t\|^2\right]} \\ &\geq \mathbb{E}_t\left[f(x_{t+1}) - M D_\Phi(x_{t+1}, x_t)\right] - \sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1}, x_t)\right]}. \end{aligned} \tag{6.2}$$

Continuing, add $f_{x_t}(\hat{x}_t, \xi_t)$ to both sides of (6.2), rearrange, and apply (B3) to obtain

$$\begin{aligned}
&\mathbb{E}_t\left[f_{x_t}(\hat{x}_t,\xi_t) - f_{x_t}(x_{t+1},\xi_t)\right] \\
&\quad\le \mathbb{E}_t\left[f_{x_t}(\hat{x}_t,\xi_t) - f(x_{t+1}) + MD_\Phi(x_{t+1},x_t)\right] + \sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]} \\
&\quad\le \mathbb{E}_t\left[f(\hat{x}_t) - f(x_{t+1}) + \tau D_\Phi(\hat{x}_t,x_t) + MD_\Phi(x_{t+1},x_t)\right] + \sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}. \qquad (6.3)
\end{aligned}$$

On the other hand, by the definition of $\hat{x}_t$ we have

$$f(\hat{x}_t) + r(\hat{x}_t) + \frac{1}{\lambda}D_\Phi(\hat{x}_t,x_t) \le f(x_{t+1}) + r(x_{t+1}) + \frac{1}{\lambda}D_\Phi(x_{t+1},x_t).$$

Inserting this equation into (6.3), we obtain

$$\begin{aligned}
&\mathbb{E}_t\left[f(\hat{x}_t) + r(\hat{x}_t) - f(x_{t+1}) - r(x_{t+1})\right] \\
&\le \mathbb{E}_t\left[\left(M+\tfrac{1}{\lambda}\right)D_\Phi(x_{t+1},x_t) + \left(\tau - \tfrac{1}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)\right] + \sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}. \qquad (6.4)
\end{aligned}$$

Combining (6.4) with (6.1) gives the estimate

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1}) - D_\Phi(\hat{x}_t,x_t) + D_\Phi(x_{t+1},x_t)\right] \\
&\qquad\le \left(M+\tfrac{1}{\lambda}\right)\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right] + \left(\tau-\tfrac{1}{\lambda}\right)D_\Phi(\hat{x}_t,x_t) + \sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]},
\end{aligned}$$

Multiplying through by η_t and rearranging, we obtain

$$\begin{aligned}
&\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1}) + \left(1-\eta_t\left(M+\tfrac{1}{\lambda}\right)\right)D_\Phi(x_{t+1},x_t)\right] \\
&\qquad\le \left(1+\eta_t(\tau-\tfrac{1}{\lambda})\right)D_\Phi(\hat{x}_t,x_t) + \sigma\eta_t\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}.
\end{aligned}$$

Now define $\gamma := \sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}$, and rewrite the above as

$$\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1})\right] \le \left(1+\eta_t\tau - \tfrac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t) + \sigma\eta_t\gamma - \left(1-\eta_t(M+\tfrac{1}{\lambda})\right)\gamma^2.$$

Maximizing the right hand side in γ, i.e. taking $\gamma = \frac{\sigma\eta_t}{2\left(1-\eta_t\left(M+\frac{1}{\lambda}\right)\right)}$, we conclude

$$\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1})\right] \le \left(1+\eta_t\tau-\tfrac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t) + \frac{1}{4}\cdot\frac{(\sigma\eta_t)^2}{1-\eta_t(M+\frac{1}{\lambda})},$$

as desired. □

With Lemma 6.1 at hand, we can now establish a convergence rate of Algorithm 2.

Theorem 6.2. *The point x_{t^*} returned by Algorithm 2 satisfies:*

$$\mathbb{E}\left[D_\Phi\left(\mathrm{prox}^\Phi_{\lambda F}(x_{t^*}),x_{t^*}\right)\right] \le \frac{\lambda}{(1-(\tau+\rho)\lambda)}\left(\frac{\lambda(F^\Phi_\lambda(x_0)-\min F)}{\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}} + \frac{\sigma^2\sum_{t=0}^T\frac{\eta_t^2}{(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}}{4\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}}\right).$$

Proof. Using Lemma 6.1, we obtain

$$\begin{aligned}\mathbb{E}_t\left[F_\lambda^\Phi(x_{t+1})\right] &\leq \mathbb{E}_t\left[f(\hat{x}_t) + \frac{1}{\lambda}D_\Phi(\hat{x}_t, x_{t+1})\right] \\ &\leq \mathbb{E}_t\left[f(\hat{x}_t) + \frac{1}{\lambda(1-\eta_t\rho)}\left(\left(1+\eta_t\tau-\eta_t/\lambda\right)D_\Phi(\hat{x}_t,x_t) + \frac{1}{4}\cdot\frac{(\sigma\eta_t)^2}{1-\eta_t(M+1/\lambda)}\right)\right] \\ &= F_\lambda^\Phi(x_t) + \frac{\eta_t}{\lambda}\left(\frac{\tau+\rho-1/\lambda}{1-\eta_t\rho}\right)D_\Phi(\hat{x}_t,x_t) + \frac{(\sigma\eta_t)^2}{4\lambda(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}\end{aligned}$$

Recursing and applying the tower rule for expectations, we obtain

$$\mathbb{E}\left[F_\lambda^\Phi(x_{T+1})\right] \leq F_\lambda^\Phi(x_0) + \sum_{t=0}^{T}\left(\frac{\eta_t}{\lambda}\left(\frac{\tau+\rho-1/\lambda}{1-\eta_t\rho}\right)\mathbb{E}\left[D_\Phi(\hat{x}_t,x_t)\right] + \frac{(\sigma\eta_t)^2}{4\lambda(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}\right)$$

Rearranging and using the fact that $\mathbb{E}\left[F_\lambda(x_{T+1})\right] \geq \min F$, we obtain

$$\sum_{t=0}^{T}\frac{\eta_t}{\lambda}\left(\frac{1/\lambda-\tau-\rho}{1-\eta_t\rho}\right)\mathbb{E}\left[D_\Phi(\hat{x}_t,x_t)\right] \leq F_\lambda^\Phi(x_0) - \min F + \frac{\sigma^2}{4\lambda}\sum_{t=0}^{T}\frac{\eta_t^2}{(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}$$

or equivalently

$$\sum_{t=0}^{T}\frac{\eta_t}{1-\eta_t\rho}\mathbb{E}\left[D_\Phi(\hat{x}_t,x_t)\right] \leq \frac{\lambda^2(F_\lambda^\Phi(x_0)-\min F)}{1-(\tau+\rho)\lambda} + \frac{\lambda\sigma^2}{4(1-(\tau+\rho)\lambda)}\sum_{t=0}^{T}\frac{\eta_t^2}{(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}.$$

Dividing through by $\sum_{t=0}^{T}\frac{\eta_t}{1-\eta_t\rho}$ and recognizing the left-hand-side as $\mathbb{E}[D_\Phi(\hat{x}_{t^*},x_{t^*})]$, the result follows. □

As an immediate corollary, we obtain a convergence rate for Algorithm 2 with a constant stepsize.

Corollary 6.3. *For some $\alpha > 0$, set $\eta_t = \frac{1}{M+\lambda^{-1}+\alpha^{-1}\sqrt{T+1}}$ for all indices $t = 1,\ldots,T$. Then the point x_{t^*} returned by Algorithm 2 satisfies:*

$$\mathbb{E}\left[D_\Phi\left(\mathrm{prox}_{\lambda F}^\Phi(x_{t^*}),x_{t^*}\right)\right] \leq \frac{\lambda^2(F_\lambda^\Phi(x_0)-\min F)+\lambda(\frac{\sigma\alpha}{2})^2}{(1-(\tau+\rho)\lambda)}\cdot\left(\frac{M+\lambda^{-1}-\rho}{T+1}+\frac{1}{\alpha\sqrt{T+1}}\right).$$

7 Rates in function value for convex problems

In this final section, we examine convergence rates for stochastic model based minimization under convexity assumptions and prove rates of converge on function values. To this end, we will use the following definition from [30]. A function $g\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is *μ-strongly convex relative to Φ* if the function $g-\mu\Phi$ is convex. Notice that $\mu = 0$ corresponds to plain convexity of g.

In this section, we make the following assumptions:

(C1) **(Sampling)** It is possible to generate i.i.d. realizations $\xi_1,\ldots,\xi_T \sim P$

(C2) **(One-sided accuracy)** There is a measurable function $(x, y, \xi) \mapsto f_x(y, \xi)$ defined on $U \times U \times \Omega$ satisfying both

$$\mathbb{E}_\xi [f_x(x, \xi)] = f(x), \qquad \forall x \in U \cap \operatorname{dom} r$$

and

$$\mathbb{E}_\xi [f_x(y, \xi)] \leq f(y), \qquad \forall x, y \in U \cap \operatorname{dom} r. \tag{7.1}$$

(C3) **(Convexity of the models)** The exists some $\mu \geq 0$ such that the functions $f_x(\cdot, \xi) + r(\cdot)$ are μ-strongly convex relative to Φ for all $x \in U \cap \operatorname{dom} r$ and a.e. $\xi \in \Omega$.

(C4) **(Lipschitz property)** There exists a square integrable function $L \colon \Omega \to \mathbb{R}_+$ such that for all $x, y \in U \cap \operatorname{dom} r$, the following inequalities holds:

$$f_x(x, \xi) - f_x(y, \xi) \leq L(\xi)\sqrt{D_\Phi(y, x)}, \tag{7.2}$$
$$\sqrt{\mathbb{E}_\xi [L(\xi)^2]} \leq \mathsf{L}.$$

(C5) **(Solvability)** The convex problems

$$\min_y \left\{ F(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\} \qquad \text{and} \qquad \min_y \left\{ f_x(y, \xi) + r(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\},$$

admit a minimizer for any $\lambda > 0$, any $x \in U$, and a.e. $\xi \in \Omega$. The minimizers vary measurably in $(x, \xi) \in U \times \Omega$.

Thus the only difference between assumptions (C1)-(C5) and (A1)-(A5) is that in expectation the stochastic models $f(\cdot, \xi)$ are global under-estimators (C2) and the functions $f(\cdot, \xi) + r(\cdot)$ are relatively strongly convex, instead of weakly convex (C3). Note that under assumptions (C1)-(C5), the objective function F is μ-strongly convex relative to Φ; the argument is completely analogous to that of Lemma 3.3.

Henceforth, we let $\{x_t\}_{t \geq 0}$ be the iterates generated by Algorithm 1 (with $\tau = \rho = 0$) and let $\{\xi_t\}_{t \geq 0}$ be the corresponding samples used. For each index $t \geq 0$, we continue to use the notation $\hat{x}_t = \operatorname{prox}^\Phi_{\lambda F}(x)$ and let $\mathbb{E}_t[\cdot]$ to denote the expectation conditioned on all the realizations $\xi_0, \xi_1, \ldots, \xi_{t-1}$. We need the following key lemma, which identifies the Bregman divergence $D_\Phi(x^*, x_t)$, between the iterates and an optimal solution, as a useful potential function. Notice that this is in contrast to the nonconvex setting, where it was the envelope $F^\Phi_\lambda(x_t)$ that served as an appropriate potential function.

Lemma 7.1. *For each iteration $t \geq 0$, the iterates of Algorithm 1 satisfy*

$$\mathbb{E}_t [(1 + \eta_t \mu) D_\Phi(x^*, x_{t+1})] \leq D_\Phi(x^*, x_t) + \frac{(\mathsf{L}\eta_t)^2}{4} + \eta_t \mathbb{E}_t [r(x_t) - r(x_{t+1})] - \eta_t (F(x_t) - F(x^*)),$$

where x^* is any minimizer of F.

Proof. Appealing to the three point inequality in Lemma 2.1 and (C3), we deduce that all points $x \in \operatorname{dom} r$ satisfy

$$f_{x_t}(x,\xi_t)+r(x)+\frac{1}{\eta_t}D_\Phi(x,x_t) \geq f_{x_t}(x_{t+1},\xi_t)+r(x_{t+1})+\frac{1}{\eta_t}D_\Phi(x_{t+1},x_t)+\frac{(1+\eta_t\mu)}{\eta_t}D_\Phi(x,x_{t+1}). \tag{7.3}$$

Setting $x = x^*$, rearranging terms, and taking expectations, we deduce

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}_t\left[(1+\eta_t\mu)D_\Phi(x^*,x_{t+1}) - D_\Phi(x^*,x_t) + D_\Phi(x_{t+1},x_t)\right] \\ &\qquad \leq \mathbb{E}_t\left[f_{x_t}(x^*,\xi_t) + r(x^*) - f_{x_t}(x_{t+1},\xi_t) - r(x_{t+1})\right].\end{aligned} \tag{7.4}$$

We seek to upper bound the right-hand-side of (7.4). Assumptions (C2) and (C4) imply:

$$\begin{aligned}\mathbb{E}_t\left[f_{x_t}(x^*,\xi_t) - f_{x_t}(x_{t+1},\xi_t)\right] &\leq \mathbb{E}_t\left[f_{x_t}(x^*,\xi_t) - f_{x_t}(x_t,\xi_t) + L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right] \\ &= \mathbb{E}_t\left[f_{x_t}(x^*,\xi_t) - f(x^*)\right] + \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right] - f(x_t) + f(x^*) \\ &\leq \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right] - f(x_t) + f(x^*).\end{aligned}$$

The left hand side of (7.4) is therefore upper bounded by

$$\begin{aligned}&\mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)} - f(x_t) - r(x_{t+1})\right] + f(x^*) + r(x^*) \\ &= \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)} + (r(x_t) - r(x_{t+1}))\right] - (F(x_t) - F(x^*)).\end{aligned}$$

Putting everything together, we arrive at

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}_t\left[(1+\eta_t\mu)D_\Phi(x^*,x_{t+1}) - D_\Phi(x^*,x_t) + D_\Phi(x_{t+1},x_t)\right] \\ &\qquad \leq \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)} + (r(x_t) - r(x_{t+1}))\right] - (F(x_t) - F(x^*))\end{aligned}$$

Multiplying through by η_t and rearranging yields

$$\begin{aligned}\mathbb{E}_t\left[(1+\eta_t\mu)D_\Phi(x^*,x_{t+1})\right] \leq D_\Phi(x^*,x_t) &+ \mathbb{E}_t\left[\eta_t L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)} - D_\Phi(x_{t+1},x_t)\right] \\ &+ \eta_t\mathbb{E}_t\left[r(x_t) - r(x_{t+1})\right] - \eta_t(F(x_t) - F(x^*)).\end{aligned}$$

Now define $\gamma := \sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}$. By Cauchy-Schwarz, we have that $\mathbb{E}_t\left[\eta_t L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right] \leq \eta_t \mathsf{L}\gamma$. Thus we obtain

$$\begin{aligned}\mathbb{E}_t\left[(1+\eta_t\mu)D_\Phi(x^*,x_{t+1})\right] &\leq D_\Phi(x^*,x_t) + \eta_t\mathsf{L}\gamma - \gamma^2 + \eta_t\mathbb{E}_t\left[r(x_t) - r(x_{t+1})\right] - \eta_t(F(x_t) - F(x^*)) \\ &\leq D_\Phi(x^*,x_t) + \frac{(\mathsf{L}\eta_t)^2}{4} + \eta_t\mathbb{E}_t\left[r(x_t) - r(x_{t+1})\right] - \eta_t(F(x_t) - F(x^*)),\end{aligned}$$

where the last inequality follows by maximizing the right-hand-side in γ. □

We are now ready to prove convergence guarantees in the case that $\mu = 0$.

Theorem 7.2 (Convergence rate under convexity)**.** *For all $T > 0$, we have*

$$\mathbb{E}\left[F\left(\tfrac{1}{\sum_{t=0}^{T}\eta_t}\sum_{t=0}^{T}\eta_t x_t\right) - F(x^*)\right] \leq \frac{D_\Phi(x^*, x_0) + \sum_{t=0}^{t}\frac{(\eta_t \mathsf{L})^2}{4} + \eta_0(r(x_0) - \inf r)}{\sum_{t=0}^{T}\eta_t},$$

where x^* is any minimizer of F.

Proof. Lower-bounding the left-hand-side of Lemma 7.1 by zero $D_\Phi(x^*, x_{t+1})$, we deduce

$$\eta_t\left[F(x_t) - F(x^*)\right] \leq \frac{(\mathsf{L}\eta_t)^2}{4} + \eta_t\mathbb{E}_t\left[r(x_t) - r(x_{t+1})\right] + \mathbb{E}_t[D_\Phi(x^*, x_t) - D_\Phi(x^*, x_{t+1})]$$

Applying the tower rule for expectations yields

$$\begin{aligned}&\sum_{t=0}^{T}\eta_t\mathbb{E}\left[F(x_t) - F(x^*)\right]\\ &\quad\leq \sum_{t=0}^{T}\frac{(\eta_t\mathsf{L})^2}{4} + \mathbb{E}\left[\sum_{t=0}^{T}\eta_t(r(x_t) - r(x_{t+1}))\right] + \mathbb{E}\left[\sum_{t=0}^{T}(D_\Phi(x^*, x_t) - D_\Phi(x^*, x_{t+1}))\right].\end{aligned}$$

Using Jensen's inequality, telescoping and using the auxiliary Lemma A.1, we conclude

$$\mathbb{E}\left[F\left(\frac{1}{\sum_{t=0}^{T}\eta_t}\sum_{t=0}^{T}\eta_t x_t\right) - F(x^*)\right] \leq \frac{D_\Phi(x^*, x_0) + \sum_{t=0}^{t}\frac{(\eta_t\mathsf{L})^2}{4} + \eta_0(r(x_0) - \inf r)}{\sum_{t=0}^{T}\eta_t},$$

as claimed. □

As an immediate corollary of Theorem 5.2, we have the following rate of convergence when the stepsize η_t is constant.

Corollary 7.3 (Convergence rate under convexity for constant stepsize)**.** *For any $\alpha > 0$ and corresponding constant stepsize $\eta_t = \frac{\alpha}{\sqrt{T+1}}$, we have*

$$\mathbb{E}\left[F\left(\frac{1}{T+1}\sum_{t=0}^{T}x_t\right) - F(x^*)\right] \leq \frac{D_\Phi(x^*, x_0) + \frac{(\alpha\mathsf{L})^2}{4} + \alpha(r(x_0) - \inf r)}{\alpha\sqrt{T+1}},$$

where x^* is any minimizer of F.

The final result of this section proves that Algorithm 1, with an appropriate choice of stepsize, drives the expected error in function values to zero at the rate $\widetilde{O}(\frac{1}{k})$, whenever $\mu > 0$.

Theorem 7.4 (Convergence rate strongly convex case)**.** *Suppose that $\eta_t = \frac{1}{\mu(t+1)}$ for all $t \geq 0$. Then for all $T > 0$, we have*

$$\mathbb{E}\left[F\left(\frac{1}{T+1}\sum_{t=0}^{T}x_t\right) - F(x^*) + \mu D_\Phi(x^*, x_{T+1})\right] \leq \frac{\frac{\mathsf{L}^2(1+\log(T+1))}{4\mu} + r(x_0) - \inf r + \mu D_\Phi(x^*, x_0)}{T+1}$$

where x^* is any minimizer of F.

Proof. Using Lemma 7.1 and the law of total expectation, we have

$$\mathbb{E}\left[F(x_t) - F(x^*)\right] \leq \frac{\eta_t \mathsf{L}^2}{4} + \mathbb{E}\left[(r(x_t) - r(x_{t+1})) + \frac{1}{\eta_t} D_\Phi(x^*, x_t) - \frac{(1+\eta_t\mu)}{\eta_t} D_\Phi(x^*, x_{t+1})\right]$$

Setting $\eta_t = \frac{1}{\mu(t+1)}$, averaging, and applying Jensen's inequality yields

$$\begin{aligned}
\mathbb{E}\left[F\left(\frac{1}{T+1}\sum_{t=0}^{T} x_t\right) - F(x^*)\right] &\leq \frac{1}{T+1}\sum_{t=0}^{T} \frac{\mathsf{L}^2}{4\mu(t+1)} + \frac{\mathbb{E}\left[r(x_0) - r(x_{T+1})\right]}{T+1} \\
&\quad + \frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}\left[\mu(t+1) D_\Phi(x^*, x_t) - \mu(t+2) D_\Phi(x^*, x_{t+1})\right] \\
&\leq \frac{\frac{\mathsf{L}^2(1+\log(T+1))}{4\mu} + r(x_0) - \inf r + \mu D_\Phi(x^*, x_0)}{T+1} \\
&\quad - \mathbb{E}\left[\frac{(T+2)}{(T+1)} \mu D_\Phi(x^*, x_{T+1})\right],
\end{aligned}$$

where the last inequality follows from telescoping the terms in the sum and using the lower bound $r(x_{T+1}) \geq \inf r$. This completes the proof. □

A Proofs of auxilliary results

A.1 Proof of Proposition 3.2

Let us write

$$\Phi = \widehat{\Phi} + \widetilde{\Phi},$$

for the two functions

$$\widehat{\Phi}(x) := \sum_{i=0}^{n} \frac{a_i}{i+2} \|x\|_2^{i+2} \qquad \text{and} \qquad \widetilde{\Phi}(x) := \sum_{i=0}^{n} 3a_i \|x\|_2^{i+2}.$$

The result [29, Equation (25)] yields the estimate

$$D_{\widehat{\Phi}}(y, x) \geq \frac{1}{2}\sum_{i=0}^{n} a_i \|x\|_2^i \cdot \|x - y\|_2^2 \qquad \forall x, y.$$

Thus the proof will be complete once we establish the inequality,

$$D_{\widetilde{\Phi}}(y, x) \geq \frac{1}{2}\sum_{i=0}^{n} a_i \|y\|_2^i \cdot \|x - y\|_2^2 \qquad \forall x, y. \tag{A.1}$$

To this end, fix an index i, and set $\eta := 3(i+2)$ and $\widetilde{\Phi}_i(x) := 3a_i\|x\|_2^{i+2}$. We will show

$$D_{\widetilde{\Phi}_i}(y, x) \geq \frac{a_i}{2}\|y\|_2^i \cdot \|x - y\|_2^2,$$

which together with the identity, $D_{\tilde{\Phi}}(y,x) = \sum_{i=0}^{n} D_{\widetilde{\Phi}_i}(y,x)$, completes the proof of (A.1).

A quick computation shows that

$$D_{\widetilde{\Phi}_i}(y,x) = 3a_i \left(\|y\|_2^{i+2} + (i+1)\|x\|_2^{i+2} - (i+2)\|x\|_2^i \langle x, y\rangle \right).$$

Let us consider two cases. First suppose that $\eta^{1/i}\|x\|_2 \geq \|y\|_2$. In this case, [29, Proposition 5.1] implies

$$D_{\widetilde{\Phi}_i}(y,x) \geq \frac{a_i \eta}{2}\|x\|_2^i \cdot \|x-y\|_2^2 \geq \frac{a_i}{2}\|y\|_2^i \cdot \|x-y\|_2^2,$$

as desired.

Now suppose that $\|y\|_2 \geq \eta^{1/i}\|x\|_2$. We will show that $D_{\widetilde{\Phi}_i}(y,x) \geq \eta^{-1} D_{\widetilde{\Phi}_i}(x,y)$, which will complete the proof since

$$\eta^{-1} D_{\widetilde{\Phi}_i}(x,y) \geq \frac{a_i}{2}\|y\|^i \cdot \|x-y\|_2^2,$$

by [29, Proposition 5.1]. To that end, we compute

$$\begin{aligned}
&D_{\widetilde{\Phi}_i}(y,x) = 3a_i \left(\|y\|_2^{i+2} + (i+1)\|x\|_2^{i+2} - (i+2)\|x\|_2^i \langle x, y\rangle \right) \\
&\geq \eta^{-1} D_{\widetilde{\Phi}_i}(x,y) = \tfrac{a_i}{i+2} \left(\|x\|_2^{i+2} + (i+1)\|y\|_2^{i+2} - (i+2)\|y\|_2^i \langle x, y\rangle \right) \\
\iff &(1-\eta^{-1}(i+1))\|y\|_2^{i+2} + \eta^{-1}(i+2)\|y\|_2^i \langle x, y\rangle \geq (\eta^{-1} - (i+1))\|x\|_2^{i+2} + (i+2)\|x\|_2^i \langle x, y\rangle \\
\impliedby &(1-\eta^{-1}(i+1))\|y\|_2^i \left(\|y\|^2 + \frac{\eta^{-1}(i+2)}{(1-\eta^{-1}(i+1))} \langle x, y\rangle \right) \geq (i+2)\|x\|_2^i \langle x, y\rangle.
\end{aligned}$$

Let us show that the last inequality is true: First, we upper bound the right hand side

$$(i+2)\|x\|_2^i \langle x, y\rangle \leq \frac{(i+2)}{\eta^{(1+i)/i}}\|y\|^{i+2}.$$

Next, we lower bound the left hand side:

$$\begin{aligned}
&(1-\eta^{-1}(i+1))\|y\|_2^i \left(\|y\|^2 + \frac{\eta^{-1}(i+2)}{(1-\eta^{-1}(i+1))} \langle x, y\rangle \right) \\
&\geq (1-\eta^{-1}(i+1)) \left(1 - \frac{\eta^{-1}(i+2)}{\eta^{1/i}(1-\eta^{-1}(i+1))} \right) \|y\|_2^{i+2} \\
&= \left(1 - \eta^{-1}(i+1+\tfrac{(i+2)}{\eta^{1/i}}) \right) \|y\|_2^{i+2}.
\end{aligned}$$

Therefore, we need only verify that η satisfies

$$\begin{aligned}
&\frac{(i+2)}{\eta^{(1+i)/i}} \leq \left(1 - \eta^{-1}(i+1+\frac{(i+2)}{\eta^{1/i}}) \right) \\
\iff &(i+2) \quad \leq \eta^{(1+i)/i} - \eta^{1/i}(i+1) - (i+2) \\
\iff &2(i+2) \leq \eta^{1/i}\left(\eta - (i+1)\right),
\end{aligned}$$

which holds by the definition of η. Thus the result is proved.

A.2 An auxiliary lemma on sequences.

Lemma A.1. *Consider any nonincreasing sequence* $\{a_t\}_{t\geq 0} \subset \mathbb{R}_{++}$ *and any sequence* $\{b_t\}_{t\geq 0} \subset \mathbb{R}$. *Then for any index* $T \in \mathbb{N}$, *we have*

$$\sum_{t=0}^{T} a_t(b_t - b_{t+1}) \leq a_0(b_0 - b^*),$$

where we set $b^* = \inf_{t\geq 0} b_t$.

Proof. We successively deduce

$$\begin{aligned}\sum_{t=0}^{T} a_t(b_t - b_{t+1}) &= \sum_{t=0}^{T} a_t[(b_t - b^*) - (b_{t+1} - b^*)] \\ &= a_0(b_0 - b^*) - a_T(b_{T+1} - b^*) + \sum_{t=0}^{T-1} (a_{t+1} - a_t)\,(b_{t+1} - b^*) \\ &\leq a_0(b_0 - b^*),\end{aligned}$$

as claimed. □

A.3 Proofs of Propositions 3.4 and 3.5

Proof of Proposition 3.4. Using the fundamental theorem of calculus and convexity of the function $x \mapsto p(\|x\|_2)$ we compute

$$\begin{aligned}&\|c(x,\xi) + \nabla c(x,\xi)(y-x) - c(y,\xi)\|_2 \\ &= \left\| \int_0^1 (\nabla c(x + t(y-x),\xi) - \nabla c(x,\xi))\,(y-x)\,dt \right\|_2 \\ &\leq \int_0^1 \|\nabla c(x + t(y-x),\xi) - \nabla c(x,\xi)\|_{\mathrm{op}}\,\|y-x\|_2\,dt \\ &\leq L_2(\xi)\|y-x\|_2^2 \int_0^1 (p(\|x + t(y-x)\|_2) + p(\|x\|_2))\,t\,dt \\ &\leq L_2(\xi)\|y-x\|_2^2 \int_0^1 ((1-t)p(\|x\|_2) + tp(\|y\|_2)) + p(\|x\|_2))\,t\,dt \\ &\leq \frac{2L_2(\xi)}{3}\|y-x\|_2^2 \cdot (p(\|x\|_2) + p(\|y\|_2)).\end{aligned}$$

Hence, we deduce

$$\begin{aligned}h\left(c(x,\xi) + \nabla c(x,\xi)(y-x),\xi\right) - h\left(c(y,\xi),\xi\right) &\leq L_1(\xi) \cdot \|c(x,\xi) + \nabla c(x,\xi)(y-x) - c(y,\xi)\|_2 \\ &\leq \tfrac{2}{3} L_1(\xi) L_2(\xi) \|y-x\|_2^2 \cdot (p(\|x\|_2) + p(\|y\|_2)) \\ &\leq \tfrac{4}{3} L_1(\xi) L_2(\xi) \cdot D_\Phi(y,x),\end{aligned}$$

where the last inequality follows form Proposition 3.2. Taking expectations yields the claimed guarantee. □

Proof of Proposition 3.5. We successively compute

$$\begin{aligned} h\left(c(x,\xi),\xi\right) - h\left(c(x,\xi) + \nabla c(x,\xi)(y-x),\xi\right) &= L_1(\xi)\|\nabla c(x,\xi)(y-x)\|_2 \\ &\leq L_1(\xi)L_3(\xi)\cdot\sqrt{q(\|x\|_2)}\|y-x\|_2 \\ &\leq \sqrt{2}L_1(\xi)L_3(\xi)\cdot\sqrt{D_\Phi(y,x)}, \end{aligned}$$

where the last line follows from [29, Equation (25)]. The result follows. □

A.4 Proof of Theorem 4.1

First we rewrite F_λ^Φ, using the definition of the Bregman divergence, as

$$\begin{aligned} F_\lambda^\Phi(x) &= \inf_y \left\{ F(y) + \frac{1}{\lambda}\Phi(y) - \frac{1}{\lambda}\langle \nabla\Phi(x), y\rangle \right\} - \frac{1}{\lambda}\Phi(x) + \frac{1}{\lambda}\langle\nabla\Phi(x),x\rangle \\ &= -\sup_y \left\{ \langle \frac{1}{\lambda}\nabla\Phi(x), y\rangle - \left(F + \frac{1}{\lambda}\Phi\right)(y) \right\} - \frac{1}{\lambda}\Phi(x) + \frac{1}{\lambda}\langle\nabla\Phi(x),x\rangle \\ &= -\left(F + \frac{1}{\lambda}\Phi\right)^\star\left(\frac{1}{\lambda}\nabla\Phi(x)\right) - \frac{1}{\lambda}\Phi(x) + \frac{1}{\lambda}\langle\nabla\Phi(x),x\rangle. \end{aligned}$$

Note that $F + \frac{1}{\lambda}\Phi$ is closed and $\left(\frac{1}{\lambda} - (\rho+\tau)\right)$-strongly convex. Thus the conjugate $(F+\frac{1}{\lambda}\Phi)^\star$ is differentiable. By the chain and sum rules for differentiation, we have

$$\begin{aligned} \nabla F_\lambda^\Phi(x) &= -\frac{1}{\lambda}\nabla^2\Phi(x)\left[\nabla\left(F+\frac{1}{\lambda}\Phi\right)^\star\right]\left(\frac{1}{\lambda}\nabla\Phi(x)\right) + \frac{1}{\lambda}\nabla^2\Phi(x)x \\ &= \frac{1}{\lambda}\nabla^2\Phi(x)\left(x - \left[\nabla\left(F+\frac{1}{\lambda}\Phi\right)^\star\right]\left(\frac{1}{\lambda}\nabla\Phi(x)\right)\right) \end{aligned}$$

The (sub)gradient of a convex conjugate function is simply the set of maximizers in the supremum defining the conjugate, so that

$$\begin{aligned} \left[\nabla\left(F+\frac{1}{\lambda}\Phi\right)^\star\right]\left(\frac{1}{\lambda}\nabla\Phi(x)\right) &= \operatorname*{argmax}_y \left\{ \langle\frac{1}{\lambda}\nabla\Phi(x),y\rangle - \left(F+\frac{1}{\lambda}\Phi\right)(y)\right\} \\ &= \operatorname*{argmin}_y \left\{ F(y) + \frac{1}{\lambda}D_\Phi(y,x)\right\} \\ &= \operatorname{prox}_{\lambda F}^\Phi(x). \end{aligned}$$

Putting everything together, we obtain, $\nabla F_\lambda^\Phi(x) = \frac{1}{\lambda}\nabla^2\Phi(x)\left(x-\hat{x}\right)$, as desired.

References

[1] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16(3):697–725, 2006.

[2] H.H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

[3] H.H. Bauschke and J.M. Borwein. Legendre functions and the method of random Bregman projections. *J. Convex Anal.*, 4(1):27–67, 1997.

[4] H.H. Bauschke, M.N. Dao, and S.B. Lindstrom. Regularizing with Bregman-Moreau envelopes. *arXiv:1705.06019*, 2017.

[5] A. Beck. *First-order methods in optimization*, volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2017.

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

[7] J.Y. Bello Cruz. On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions. *Set-Valued and Variational Analysis*, 25(2):245–263, Jun 2017.

[8] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *arXiv:1706.06461*, 2017.

[9] J.M. Borwein and A.S. Lewis. *Convex analysis and nonlinear optimization*, volume 3 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer, New York, second edition, 2006. Theory and examples.

[10] S. Bubeck. *Convex Optimization: Algorithms and Complexity*. Foundations and Trends in Machine Learning. Now Publishers, 2015.

[11] J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Math. Programming*, 33(3):260–279, 1985.

[12] C. Cartis, N.I.M. Gould, and P.L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM J. Optim.*, 21(4):1721–1739, 2011.

[13] Y. Censor and S.A. Zenios. Proximal minimization algorithm with D-functions. *J. Optim. Theory Appl.*, 73(3):451–464, 1992.

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

[15] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *arXiv:1803.06523*, 2018.

[16] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv:1802.02988*, 2018.

[17] D. Drusvyatskiy. Proximal algorithms. *SIAG/OPT Views and News*, 26(1):1–8, January 2018.

[18] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *To appear in Math. Oper. Res., arXiv:1602.06661, version 2*, 2016.

[19] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *To appear in Math. Prog., arXiv:1605.00125*, 2018.

[20] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

[21] J.C. Duchi and F. Ruan. Stochastic methods for composite optimization problems. *Preprint arXiv:1703.08570*, 2017.

[22] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Math. Oper. Res.*, 18(1):202–226, 1993.

[23] R. Fletcher. A model algorithm for composite nondifferentiable optimization problems. *Math. Programming Stud.*, (17):67–76, 1982. Nondifferential and variational techniques in optimization (Lexington, Ky., 1980).

[24] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.*, 155(1-2, Ser. A):267–305, 2016.

[25] I.J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

[26] F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. *arXiv:1803.07374*, 2017.

[27] A. Juditsky and A.S. Nemirovski. First order methods for nonsmooth convex large-scale optimization, II: Utilizing problem's structure. In Stephen J. Wright Suvrit Sra, Sebastian Nowozin, editor, *Optimization for Machine Learning*, pages 29–63. MIT Press, August 2010.

[28] A.S. Lewis and S.J. Wright. A proximal method for composite minimization. *Math. Program.*, pages 1–46, 2015.

[29] H. Lu. Relative continuity for non-lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. *arXiv:1710.04718*, 2017.

[30] H. Lu, R. Freund, and Y. Nesterov. Relatively-smooth convex optimization by first-order methods, and applications. *arXiv:1610.05708*, 2016.

[31] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[32] Y. Nesterov. Modified Gauss-Newton scheme with worst case guarantees for global performance. *Optim. Methods Softw.*, 22(3):469–483, 2007.

[33] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[34] E.A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[35] N. Parikh and S. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, January 2014.

[36] M.J.D. Powell. On the global convergence of trust region algorithms for unconstrained minimization. *Math. Programming*, 29(3):297–303, 1984.

[37] R.T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[38] R.T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, third edition, 2009.

[39] M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Math. Oper. Res.*, 17(3):670–690, 1992.

[40] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, May 2018.

[41] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. http://www.mit.edu/ dimitrib/PTseng/papers/apgm.pdf, May 2008.

[42] S. Zhang and N. He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. arXiv:1806.04781, June 2018.