Modeling simple structures and geometry for better stochastic optimization algorithms

Hilal Asi
Stanford University

John C. Duchi
Stanford University

Abstract

We develop procedures for solving convex stochastic optimization problems that exploit the structure and geometry of the underlying problem. Our procedures build on the model-based aProx framework that we develop in the paper [2], which highlights the importance of more careful structural modeling; as one example of this, if we seek to minimize a non-negative loss, then stochastic optimization methods should use non-negative approximations. We extend this earlier work to improve adaptivity to problem geometry via careful choices of divergence measures, highlighting both the importance of leveraging problem structure and geometry—in the form of the divergence used to define stochastic updates—for strong performance. Our experiments confirm our theoretical results in a range of problems, including deep learning.

1 Introduction

We develop and analyze a family of methods for solving the stochastic convex optimization problem

$$\text{minimize} \;\; F(x) = \mathbb{E}_P[f(x;S)] = \int_{\mathcal{S}} f(x;s)dP(s) \quad \text{subject to} \;\; x \in \mathcal{X}, \tag{1}$$

where the set $\mathcal{S}$ is a sample space, for each $s \in \mathcal{S}$ the function $f(\cdot;s) : \mathbb{R}^n \to \mathbb{R}$ is a closed convex function, and $\mathcal{X} \subset \mathbb{R}^n$ is closed convex. Such problems arise in many scenarios where one can only compute a noisy estimate of the gradient and have applications in numerous fields, including machine learning, statistical estimation, and simulation-based optimization [40, 18, 36]. The standard methodology for these problems is the stochastic (sub)gradient method [42, 40, 27, 9, 35], which begins from an initial point x_1, then iteratively draws $S_k \overset{\text{iid}}{\sim} P$ and updates

$$x_{k+1} := x_k - \alpha_k g_k \;\; \text{for some } g_k \in \partial f(x_k; S_k). \tag{2}$$

In the paper [2], we develop and analyze aProx (Approximate Proximal) framework for stochastic optimization, which exhibits improved robustness over basic stochastic subgradient methods, which can be sensitive to parameter specification [2, 38]. We build on our earlier methodology to extend the aProx family to applications in high-dimensional optimization and estimation [27, 37] and allow adaptivity to underlying problem geometry [12], extending the benefits of better modeling to more general problems.

1.1 Approach and Contribution

Our starting point is to review the stochastic "model-based" minimization approach [15, 10, 2] for stochastic optimization. The key development of our paper [2] is that such methods are robust to stepsize choice, adaptive to problem difficulty, and enjoy optimal convergence over a range of scenarios. The aProx family iteratively minimizes

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}. \tag{3}$$

Here $f_x(\cdot; s)$ is the *model* of $f(\cdot; s)$ at the point x (see examples in Section 2), satisfying the conditions

(C.i) The function $y \mapsto f_x(y; s)$ is convex and subdifferentiable on its domain.

(C.ii) The model f_x satisfies the equality $f_x(x; s) = f(x; s)$ and

$$f_x(y; s) \le f(y; s) \;\; \text{for all } y.$$

(C.iii) At $y = x$, we have the containment

$$\partial_y f_x(y; s)|_{y=x} \subset \partial_x f(x; s).$$

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Many optimization algorithms fall into this framework, including stochastic subgradient methods, with the linear model $f_x(y;s) = f(x;s) + \langle f'(x;s), y - x\rangle$, and stochastic proximal point methods [32, 25, 7, 23, 8, 2], which use the exact model $f_x(y;s) = f(y;s)$. In [2] we argue that using accurate models with certain structural properties in APROX results in procedures with improved robustness to stepsize choice, objective, and problem difficulty. As is well-known in first-order optimization [3, 27], these procedures may perform poorly in "non-Euclidean" problems, for example, high dimensional problems over the ℓ_1-ball.

To that end, we extend the APROX framework to more general divergence measures and non-Euclidean geometries, modifying the iteration (3) to

$$x_{k+1} := \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ f_{x_k}(x; S_k) + \frac{1}{\alpha_k} D_h(x, x_k) \right\}, \quad (4)$$
$$D_h(x,y) := h(x) - h(y) - \langle h'(y), x - y\rangle.$$

Here D_h is the Bregman-divergence generated by $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and we assume that h is strongly convex with respect to a norm $\|\cdot\|$ over $\mathcal{X}$, meaning that $D_h(x,y) \geq \frac{1}{2}\|x - y\|^2$ for $x, y \in \mathcal{X}$.

We investigate the model and divergence-based iteration (4), studying the effects of accurate models and divergences on the behavior of the optimization algorithms. First, we show that the robustness results for accurate models with Euclidean distance [2] hold for general divergences as well, motivating the importance of model choice even with good distance measures. Moreover, we show that even when the approximation f_x is accurate, appropriate choices of divergence D_h can significantly improve the convergence time of our procedures. As we show in the sequel, the best-stepsize performance of simple models with appropriate divergence can outperform those achieved by accurate models with poorly chosen divergence.

We conclude our paper with a substantial experimental investigation of our methods in multiple settings, including deep learning. Our experiments confirm our theoretical results, demonstrating that accurate models can improve robustness for stepsize choice, while a proper divergence choice can enhance the best-stepsize performance.

1.2 Related work

We situate our work in the connection to stochastic model-based optimization methods [15, 2, 10] and mirror descent methods [28]. These two lines of research aim to solve two different questions; the former suggests the model-based framework as a robust extension of standard gradient methods, which are known to be sensitive to stepsize choices as well as the functions themselves being optimized [27], while in the later, researchers aim to develop optimization algorithms that more accurately reflect problem geometry [28, 3, 12].

The first and most well-known model-based minimization approaches are the proximal point methods [32], which minimize regularized version of the true function. In stochastic cases, proximal point methods help to reduce some of the instability inherent to stochastic optimization. We identify a few papers in this line of work leading to ours. Bertsekas [7] analyzes stochastic proximal point methods in an incremental framework, i.e. when $\mathcal{S} = \{1, 2, \ldots, m\}$ is a finite set, and provides convergence results similar to subgradient methods, while Ryu and Boyd [33] investigate the same algorithm and show some cases in which it is more stable than standard stochastic subgradient methods. More recently, Duchi and Ruan [14], followed by Davis and Drusvyatskiy [10] and our work [2], develop a model-based framework allowing for more general models than the linear model, as in basic stochastic subgradient methods, or the exact model, as in stochastic proximal point methods. Duchi and Ruan [14] noted fairly extraordinary performance gains of certain model-based methods over standard subgradient schemes, though they could only provide an asymptotic analysis of convergence (without rates). Davis and Drusvyatskiy [10] developed the first convergence guarantees in non-convex settings, showing how to describe convergence of well-behaved (but potentially non-smooth) non-convex optimization problems. Our earlier paper [2] follows these works and provides evidence—empirical and theoretical—that (we believe) explains the performance benefits of model-based schemes over conventional methods, showing stability, convergence, and adaptivity guarantees for methods based on accurate models.

All of this work focuses on Euclidean (low-dimensional) settings, and in this paper, we develop mirror-descent (non-Euclidean) analogues of our results [2]. This of course builds out of mirror descent convergence guarantees originally developed by Nemirovski and Yudin [28], which others have complemented in stochastic and online optimization [3, 27]. Recently, Davis et al. [11] study model-based optimization methods with Bregman divergences chosen to control one-sided error of the models, providing convergence guarantees similar to standard mirror descent method.

Notation For a convex function f, $\partial f(x)$ denotes its subgradient set at x, and $f'(x) \in \partial f(x)$ denotes an arbitrary element of the subdifferential. Throughout, $x^\star$ denotes a minimizer of problem (1) and

$\mathcal{X}^\star = \operatorname{argmin}_{x\in\mathcal{X}} F(x)$ its optimal set. We let $\mathcal{F}_k := \sigma(S_1, \ldots, S_k)$ be the σ-field generated by the first k random variables S_i, so $x_k \in \mathcal{F}_{k-1}$ for all k under iteration (4). We let $D_h(\cdot,\cdot)$ denote a Bregman divergence generated by a 1-strongly convex function $h(\cdot)$ with respect to $\|\cdot\|_h$. We denote its dual norm by $\|\cdot\|_{h^*}$. For a set $A \subseteq \mathbb{R}^n$, we denote $D_h(A, x) := \inf_{y\in A}\{D_h(y, x)\}$.

2 Models and divergences

We give a brief introduction for the models and divergences we use throughout the paper, adapting the models we introduce [2, Sec. 2] for stochastic optimization with Euclidean distance. We also mention a few divergences we will use later in our examples.

Stochastic subgradient: The model is

$$f_x(y; s) := f(x; s) + \langle f'(x; s), y - x\rangle. \tag{5}$$

Truncated subgradient: The model is

$$f_x(y; s) := \big(f(x; s)+\langle f'(x; s), y-x\rangle\big)\vee \inf_{z\in\mathcal{X}} f(z; s). \tag{6}$$

More generally, we can consider models that satisfy the following condition.

(C.iv) For all $s \in \mathcal{S}$, the models $f_x(\cdot; s)$ satisfy
$$f_x(y; s) \geq \inf_{z\in\mathcal{X}} f(z; s).$$

As a simple example of Condition (C.iv), consider any loss function known to be non-negative, for example, the hinge, logistic, or squred loss. Then the truncated model (6) is simply the positive part $f_x(y; s) = [f(x; s) + \langle f'(x; s), y - x\rangle]_+$.

Proximal point: The model

$$f_x(y; s) := f(y; s). \tag{7}$$

We also consider relatively accurate models, such as the bundle model (see [2]), which satisfy

(C.v) For some $\epsilon > 0$, there exists $C : \mathcal{S} \to \mathbb{R}_+$ with $\mathbb{E}[C(S)] < \infty$ such that for $x_0 \in \mathcal{X}$, the point $x_\alpha = \operatorname{argmin}_{x\in\mathcal{X}}\{f_{x_0}(x; s) + \frac{1}{\alpha}D_h(x, x_0)\}$ satisfies
$$f(x_\alpha; s) \leq f_{x_0}(x_\alpha; s) + \frac{1-\epsilon}{\alpha}D_h(x_\alpha, x_0) + C(s)\alpha.$$

Now, we discuss a few well-known divergences which will be useful for our development.

Euclidean distance: The function $h(x) = \frac{1}{2}\|x\|_2^2$ generates the Euclidean divergence

$$D_h(x, y) = \frac{1}{2}\|x - y\|_2^2. \tag{8}$$

Clearly h is strongly convex with respect to $\|\cdot\|_2$.

Mahalanobis distance: The function $h(x) = \frac{1}{2}x^T Hx$, for a matrix $H \succ 0$, generates the divergence

$$D_h(x, y) = \frac{1}{2}(x - y)^T H(x - y) \tag{9}$$

In this case, h is strongly convex with respect to the Mahalanobis norm $\|x\|_H := \sqrt{x^T Hx}$. The AdaGrad algorithms [12] adaptively choose

$$H_k = \operatorname{diag}\left(\sum_{i=1}^k f'(x_i; S_i)f'(x_i; S_i)^T\right)^{\frac{1}{2}} \tag{10}$$

and use $h_k(x) = \frac{1}{2}x^T H_k x$ in the kth update (4).

The p-norm divergences: When the optimization domain $\mathcal{X}$ is a subset of the ℓ_1-ball or an ℓ_p-ball, $1 < p \leq 2$, is is natural to use the p-norm divergences [16, 34, Sec. 5.1.4 and Ex. 5], where one takes $h(x) = \frac{1}{2}\|x\|_p^2$, which are $(p-1)$-strongly convex with respect to themselves. For later use, we note that

$$h'(x) = \frac{[\operatorname{sign}(x_j)|x_j|^{p-1}]_{j=1}^n}{\|x\|_p^{p-2}} \text{ and } \|h'(x)\|_q = \|x\|_p$$

where $1/p+1/q = 1$. In addition, they have the strong convexity and smoothness(-like) properties

$$\frac{(p-1)}{2}\|x - y\|_p^2 \leq D_h(x, y) \leq \frac{q-1}{2}\cdot\ldots \tag{11}$$
$$\left(\left(\|x\|_p \vee \|y\|_p\right)^{2-p}\|x - y\|_p^{p-1} + (2 - p)\|x - y\|_p\right)^2.$$

(See the supplemental Appendix 6.6 for a proof.)

From inequality (11), we see that $h(x) = \frac{1}{2(p-1)}\|x\|_p^2$ is strongly convex with respect to itself; moreover, choosing $p = 1 + 1/\log(2d)$ yields the lower bound $\|x\|_1 \geq c\|x\|_p$ for a numerical constant c. Alternative choices are possible; if the domain is the simplex $\mathcal{X} = \{x \in \mathbb{R}^n \mid x \geq 0, \mathbf{1}^T x = 1\}$, the negative entropy $h(x) = \sum_{i=1}^n x_i \log x_i$ yields $D_h(x, y) = D_{\text{kl}}(x\|y)$, which is strongly convex with respect to the ℓ_1-norm.

3 Stability and convergence guarantees

We now investigate the effects of models and divergences on the behavior of the method (4). We start with a simple example that illustrates the influence of the model, discussing the importance of the divergence subsequently.

Example 1 (choice of model)**:** Consider the function $f(x) = e^x + e^{-x}$ with $D_h(x, y) = \frac{1}{2}(x - y)^2$. A calculation shows that the iterates of the gradient method, without noise (i.e. the linear model (5)) satisfy $|x_k| > 2^{2^k}|x_1|$ if α_1 is large enough; while Corollary

3.2 of our paper [2] shows that the iterates (4) using the proximal model (7) are bounded by a constant, no matter the value of the initial stepsize α_1. $\diamond$

Example 1 illustrates some of the benefits of more careful and accurate choices for the model; after presenting our main stability and convergence results, we will show how better choices of the divergence can yield similarly important benefits over the Euclidean model-based methods (3).

3.1 Stability

Now we start our theoretical investigation of the effects of divergence and model choice on the stability of the iterates. First, we extend our definitions of stability [2] to any Bregman divergence $D_h(\cdot,\cdot)$. Let $\mathcal{A}$ denote the set of positive stepsize sequences $\{\alpha_k\}$ with $\sum_k \alpha_k^2 < \infty$. The pair $(\mathcal{F},\mathcal{P})$ is a *collection of problems* if $\mathcal{P}$ is a collection of probability measures on a sample space $\mathcal{S}$, and $\mathcal{F}$ is a collection of functions $f : \mathcal{X}\times\mathcal{S}\to\mathbb{R}$, where $f(\cdot;s)$ is convex. We have the following definition of stability.

Definition 3.1. *An algorithm generating iterates x_k according to the model-based update (4) is* stable in L_2 *for the divergence $D_h(\cdot,\cdot)$ for the collection of problems $(\mathcal{F},\mathcal{P})$ if for all $f\in\mathcal{F}$ and $P\in\mathcal{P}$ defining $F(x) = \mathbb{E}_P[f(x;S)]$ and $\mathcal{X}^\star = \operatorname{argmin}_{x\in\mathcal{X}} F(x)$,*

$$\sup_{\alpha\in\mathcal{A}}\sup_{k\in\mathbb{N}} \frac{\mathbb{E}[D_h(\mathcal{X}^\star, x_k,)]}{\sum_k \alpha_k^2} < \infty. \tag{12a}$$

It is stable in probability *for the divergence $D_h(\cdot,\cdot)$ if for all stepsize sequences $\{\alpha_k\}\in\mathcal{A}$,*

$$\sup_k D_h(\mathcal{X}^\star, x_k) < \infty \quad \textit{with probability 1.} \tag{12b}$$

We have previously shown [2] how the iterates of accurate models are stable (in both notions) when one uses Euclidean divergences $D_h(x,y) = \frac{1}{2}\|x-y\|_2^2$. In what follows, we show that this remains true for any Bregman divergence, where the stability bound depends on the choice of divergence. The following theorem provides our basic recursion for proving stability of our methods.

Theorem 1. *Let x_k be generated by the iteration (4) with any model satisfying Conditions (C.i)–(C.iii) and (C.v). Then for all $x^\star\in\mathcal{X}^\star$,*

$$\mathbb{E}\left[D_h(x^\star, x_{k+1}) \mid \mathcal{F}_{k-1}\right] \le D_h(x^\star,x_k) + \alpha_k^2\left(\frac{\mathbb{E}\left[\|f'(x^\star;S_k)\|_{h^*}^2\right]}{\epsilon} + \mathbb{E}[C(S)]\right).$$

By applying the recursion in Theorem 1 iteratively and using standard martingale convergence theorems we have the following stability corollary.

Corollary 3.1. *Let the conditions of Theorem 1 hold. Assume further that $\mathbb{E}[\|f'(x^\star;S)\|_{h^*}^2] \le \sigma_h^2$ for all $x^\star\in\mathcal{X}^\star$. Then, for each $k\in\mathbb{N}$,*

$$\mathbb{E}\left[D_h(\mathcal{X}^\star, x_{k+1})\right] \le \mathbb{E}\left[D_h(\mathcal{X}^\star,x_1)\right] + \left(\frac{\sigma_h^2}{\epsilon} + \mathbb{E}[C(S)]\right)\sum_{i=1}^k \alpha_i^2.$$

If $\sum_k \alpha_k^2 < \infty$, then

$$\sup_{k\in\mathbb{N}} D_h(\mathcal{X}^\star, x_k) < \infty$$

and $D_h(\mathcal{X}^\star,x_k)$ converges to some finite value with probability 1.

Corollary 3.1 shows that the iterates are stable (Def. 3.1) in both senses for the Bregman divergence whenever the model is accurate. In particular, whenever $\mathbb{E}[\|f'(x^\star;S)\|_{h^*}^2] < \infty$, we obtain stability. This dependence on the norm $\|\cdot\|_{h^*}$ is important, as some choice of divergence may not yield finite bounds on the "variance" of $f'(x^\star;S)$. Indeed, consider the following

Example 2 (Mean-like estimation under p-norms)**:** Consider the space $\ell_p(\mathbb{N}) = \{x\in\mathbb{R}^\mathbb{N} \mid \|x\|_p < \infty\}$, and let $\mathcal{X}$ be the 1-ball in $\ell_p(\mathbb{N})$, that is, $\mathcal{X} = \{x\in\ell_p(\mathbb{N}) \mid \|x\|_p \le 1\}$, and consider $f(x;s) = \|x-s\|_p^2$. Then the p-norm divergences (recall inequality (11)), generated by $h(x) = \frac{1}{2(p-1)}\|x\|_p^2$, yield dual norms $\|\cdot\|_{h^*} = \|\cdot\|_q$ for $q = \frac{p}{p-1}$ conjugate to p. For $x\in\ell_p(\mathbb{N})$ we have

$$\|f'(x;s)\|_q = \|x-s\|_p,$$

while $\|f'(x;s)\|_2^2 \ge \|x-s\|_p^{2-p}\sum_i(|x_i|^{p-1} - |s_i|^{p-1})^2$. Assuming $x^\star = 0$, then we have $\|f'(x^\star;s)\|_2^2 = \|s\|_p^{2-p}\sum_i |s_i|^{2(p-1)}$. This may be infinite even when $s\in\ell_p(\mathbb{N})$, so that standard bounds based on Euclidean distances yield no guarantees in this case. $\diamond$

Evidently the choice of divergence can have strong effects on the stability properties of the algorithms.

3.2 Divergence based convergence guarantees

In the previous section, we demonstrated the effects of model and divergence on the stability of the iterates. In this section, we provide convergence guarantees for our models, which show the dependence on the model and divergence. We start with a general convergence result, showing that our family of methods converge under extremely weak assumptions for stable models with general Bregman divergences.

Proposition 1. *Assume there exists an increasing function $\mathsf{G}_{\text{big}} : \mathbb{R}_+ \to [0,\infty)$ such that for all $x\in$*

$\mathcal{X}$, $\mathbb{E}[\|f'(x;S)\|^2] \le \mathsf{G}_{\text{big}}(D_h(\mathcal{X}^\star, x))$. *Let the iterates x_k be generated by any method satisfying Conditions (C.i)–(C.iii), and additionally assume that the iterates are bounded: with probability 1, $\sup_k \|x_k\| < \infty$. Then $\sum_k \alpha_k(F(x_k) - F(x^\star)) < \infty$. If in addition*

$$\Gamma(\epsilon) := \inf_{x \in \mathcal{X}} \{F(x) - F(x^\star) \mid D_h(\mathcal{X}^\star, x) \ge \epsilon\} > 0$$

for all $\epsilon > 0$, then $D_h(\mathcal{X}^\star, x_k) \stackrel{a.s.}{\to} 0$.

Proposition 1 also implies an asymptotic convergence rate on (weighted averages of) the iterates x_k. Indeed, we get the following corollary using Jensen's inequality.

Corollary 3.2. *Let the conditions of Proposition 1 hold. Let $\{\gamma_k\}_{k=1}^\infty \subset \mathbb{R}_+$ be a non-decreasing sequence with $\gamma_k \ge 0$. Then define the weighted averages $\overline{x}_k = \sum_{i=1}^k \gamma_i \alpha_i x_i / (\sum_{i=1}^k \gamma_i \alpha_i)$. Then with probability 1,*

$$\limsup_{k \to \infty} \frac{1}{\gamma_k} \left(\sum_{i=1}^k \gamma_i \alpha_i \right) [F(\overline{x}_k) - F^\star] = 0.$$

In particular, if $\gamma_k = \alpha_k^{-1}$ then $\overline{x}_k = \frac{1}{k} \sum_{i=1}^k x_i$ satisfies

$$k \alpha_k \left(F(\overline{x}_k) - F^\star\right) \stackrel{a.s.}{\to} 0.$$

We proceed now to provide convergence guarantees for any general model that depends on the chosen divergence $D_h(\cdot, \cdot)$. As we demonstrate next, choosing better divergences can result in convergence rates with improved dependence on certain problem parameters such as the dimension. We have the following convergence result for general models and divergences if the functions $f(\cdot; s)$ are Lipschitz.

Proposition 2. *Let the iterates x_k be generated by algorithm (4) by any model satisfing Conditions (C.i)–(C.iii). Define $\overline{x}_k = (\sum_{i=1}^k \alpha_i)^{-1} \sum_{i=1}^k \alpha_i x_i$. Then*

$$\mathbb{E}[F(\overline{x}_k)] - F(x^\star) \le \frac{D_h(x^\star, x_1)}{\sum_{i=1}^k \alpha_i} + \frac{1}{2 \sum_{i=1}^k \alpha_i} \sum_{i=1}^k \alpha_i^2 \mathbb{E}\left[\|f'(x_k; S_k)\|_{h^*}^2 \right].$$

If $\mathcal{X}$ is bounded with $R := \sup_{x \in \mathcal{X}} D_h(x^\star, x)$, and $f(\cdot; S)$ satisfies $\mathbb{E}[\|f'(x;S)\|_{h^*}^2] \le M_h^2$ for all $x \in \mathcal{X}$, then the average $\overline{x}_k := \frac{1}{k} \sum_{i=1}^k x_i$ satisfies

$$\mathbb{E}[F(\overline{x}_k)] - F(x^\star) \le \frac{R}{k \alpha_k} + \frac{M_h^2}{2k} \sum_{i=1}^k \alpha_i.$$

Proposition 2 provides divergence dependent convergence rates for any general model that satisfies Conditions (C.i)–(C.iii). Thus, we have recovered the typical convergence guarantees for stochastic mirror descent methods [27], in complete analogy with the Euclidena case, but applying to this our more general model-based framework.

3.3 Convergence with adaptive divergences

While not without controversy [38], there is theoretical and empirical evidence [12, 26, 13, 29] that adaptive methods—which modify the divergence D_h throughout their iterations—can yield strong performance for stochastic gradient methods. Consequently, in this section, we show how to develop convergence guarantees for such scenarios in the model-based setting (4). In particular, we consider the setting where the function h_k is $\mathcal{F}_k$-measurable, $D_{h_k}(x, y) = \frac{1}{2}(x - y)^T H_k (x - y)$. We show that any model in our framework enjoys the typical convergence guarantees associated with AdaGrad and related algorithms [12, 26]. In particular, we consider updates

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha}(x - x_k)^T H_k (x - x_k) \right\}. \tag{13}$$

Recalling that $\|x\|_H^2 = x^T H x$ for symmetric H, we have have the following convergence guarantee.

Proposition 3. *Let x_k follow algorithm (13) with any model satisfing Conditions (C.i)–(C.iii). Then*

$$\begin{aligned} &\frac{1}{k} \sum_{i=1}^k f(x_i; S_i) - f(x^\star; S_i) \\ &\le \frac{1}{2k\alpha} \|x_1 - x^\star\|_{H_1}^2 + \frac{1}{2k\alpha} \sum_{i=1}^k \|x_i - x^\star\|_{H_i - H_{i-1}}^2 \\ &+ \frac{\alpha}{2k} \sum_{i=1}^k \|f'(x_i; S_i)\|_{H_i^{-1}}^2. \end{aligned}$$

If $H_k = \operatorname{diag}(\sum_{i=1}^k f'(x_i; S_i) f'(x_i; S_i)^T)^{1/2}$ and $\mathcal{X}$ is bounded with $R_\infty := \sup_{x \in \mathcal{X}} \|x - x^\star\|_\infty$, the average $\overline{x}_k := \frac{1}{k} \sum_{i=1}^k x_i$ satisfies

$$\mathbb{E}\left[F(\overline{x}_k)\right] - F(x^\star) \le \left(\frac{R_\infty^2}{2k\alpha} + \frac{\alpha}{k} \right) \mathbb{E}[\operatorname{tr}(H_k)].$$

As a consequence of this proposition, we see, for example, that model-based methods may use the AdaGrad [12] stepping with $H_k = \operatorname{diag}(\sum_{i=1}^k g_i g_i^T)^{1/2}$ for $g_i \in \partial f(x_i; S_i)$, achieving adaptive minimax optimal rates when the domain is an ℓ_∞-box [13].

4 Fast convergence on easier problems

We now study the performance of the mirror extension of the APROX family on problems with additional structure—strong convexity or interpolating solutions—that allow more efficient solution methods. For these, we show that the model based methods exhibit stronger adaptivity properties than standard stochastic gradient methods.

4.1 Problems with consistent solutions

We begin by investigating what we term "easy problems," investigating convergence properties of Bregman-based updates. By an easy problem, we mean a problem for which there exists a shared minimizer for all samples $s \in \mathcal{S}$; given the success of machine learning and engineering applications in which there exist interpolating solutions (for example, in deep learning) [39, 4, 5], it is important to investigate algorithm performance in these regimes. As in our earlier paper [2], we show that models satisfying Conditions (C.i)–(C.iii), and additionally the lower bound Condition (C.iv), enjoy fast (sometimes linear) convergence for these instances. In contrast to our earlier analysis [2], which proves similar results for the Euclidean distance, we allow growth with respect to nontrivial powers of Bregman divergences. These extensions are of interest in situations similar to those for which mirror descent proves effective: when Euclidean distance introduces inappropriate geometry for the problem. We begin with our definition of easy problems [2].

Definition 4.1. *Let $F(x) := \mathbb{E}_P[f(x; S)]$. Then F is* easy *to optimize if for $\mathcal{X}^\star := \operatorname{argmin}_{x\in\mathcal{X}} F(x)$, for each $x^\star \in \mathcal{X}^\star$ and P-almost all $s \in \mathcal{S}$ we have*

$$\inf_{x\in\mathcal{X}} f(x; s) = f(x^\star; s).$$

The following lemma, which generalizes [2, Lemma 4.1], provides a single-step progress guarantee for models satisfying Conditions (C.i)–(C.iv). The lemma shows that our iterates always make progress towards the optimal set for easy problems. Perhaps the most compelling example in this case comes from situations with non-negative losses and a consistent solution—evidently a situation of growing frequency in large-scale machine learning [39, 4, 5]. Here, the truncated model that simply uses the positive part $f_{x_k}(x; S_k) = [f(x_k; S_k) + \langle g_k, x - x_k\rangle]_+$ for $g_k \in \partial f(x_k; S_k)$ is enough to achieve the guarantees.

Lemma 4.1. *Let F be easy to optimize (Definition 4.1). Let x_k be generated by the updates (4) using a model satisfying Conditions (C.i)–(C.iv). Let $\gamma_k = \min\{\alpha_k, \frac{f(x_k;S_k)-f(x^\star;S_k)}{\|f'(x_k;S_k)\|_{h^*}^2}\}$. Then for any $x^\star \in \mathcal{X}^\star$,*

$$D_h(x^\star, x_{k+1}) \le D_h(x^\star, x_k) - \frac{\gamma_k}{2}[f(x_k; S_k) - f(x^\star; S_k)].$$

Given the progress condition in Lemma 4.1, it is natural to make an assumption on the growth of f away from its optimizers.

Assumption A1 (Expected sharp growth)**.** *There exist constants $\lambda_0, \lambda_1 > 0$ and $\rho \in (0, \infty)$ such that for all $\alpha \in \mathbb{R}_+$ and $x \in \mathcal{X}$ and $x^\star \in \mathcal{X}^\star$,*

$$\mathbb{E}\left[\min\left\{\alpha[f(x; S) - f(x^\star; S)], \frac{(f(x; S) - f(x^\star; S))^2}{\|f'(x; S)\|_{h^*}^2}\right\}\right] \ge D_h(\mathcal{X}^\star, x)^{\frac{\rho}{2}} \min\left\{\lambda_0\alpha, \lambda_1 D_h(\mathcal{X}^\star, x)^{\frac{\rho}{2}}\right\}.$$

While Assumption A1 is somewhat complex, it holds when the losses f deviate above f with constant probability. For example, as in our earlier work [2], a simple argument with expectations shows that Assumption A1 holds when there exist $\lambda > 0, p > 0$ such that

$$\mathbb{P}(f(x; S) - f(x^\star; S) \ge \lambda D_h(\mathcal{X}^\star, x)^{\rho/2}) \ge p \qquad (14)$$

and $\mathbb{E}[\|f'(x; S)\|_{h^*}^2] \le M_h^2 < \infty$. We give several examples in our paper [2] in the Euclidean case, as well as examples of interpolating problems; let us here provide a very stylized example to show when Assumption A1 holds for certain Bregman divergences and distance-generating h, but not for Euclidean cases.

Example 3 (Observations of a consistent vector)**:** Let $p \in (1, 2)$ and let $\ell_p(\mathbb{N}) = \{x \in \mathbb{R}^{\mathbb{N}} \mid \sum_{i=1}^\infty |x_i|^p < \infty\}$ as is standard. Fix some $v \in \ell_p(\mathbb{N})$, $\|v\|_p \le 1$, and let $\mathcal{X} = \{x : \|x\|_p \le 2\}$. Consider the case that $S \in \{0, 1\}^{\mathbb{N}}$ has independent coordinates. Then $f(x; s) = \|s \odot (x - v)\|_p$, where $\odot$ denotes Hadamard (elementwise) product, satisfies $\inf_x \mathbb{P}(f(x; S) \ge \lambda \|x - v\|_p) > 0$ (as $x^\star = v$). Now, we use the "smoothness" inequalities (11), which, because $\|x\|_p \le 2$, yields that condition (14) holds in that for some $\lambda = \lambda_p > 0$ we have

$$\inf_x \mathbb{P}(f(x; S) \ge \lambda D_h(x^\star, x)^{\frac{1}{2(p-1)}}) > 0.$$

Evidently Assumption A1 cannot hold with $h(x) = \frac{1}{2}\|x\|_2^2$, as the relevant norms may have infinite expectations. $\diamond$

Finally, if Assumption A1 holds, we can prove linear convergence of our methods, as the next proposition shows.

Proposition 4. *Let F be easy to optimize and Assumption A1 hold. Let x_k be generated by the stochastic iteration (4) using any model satisfying Conditions (C.i)–(C.iv), where the stepsizes $\alpha_k = \alpha_0 k^{-\beta}$, $\beta \in [0, 1)$. For any $\epsilon > 0$, define*

$$k(\epsilon) := \left\lfloor \left(\frac{\lambda_1 D_1 \epsilon^{\rho-1}}{\lambda_0\alpha_0}\right)^{-1/\beta} \right\rfloor$$

for $D_1^2 = D_h(\mathcal{X}^\star, x_1)$. Then

$$\mathbb{E}[D_h(\mathcal{X}^\star, x_{k+1})] \le \max\left\{\epsilon^2, D_1^2 \cdot \ldots \right.$$
$$\left. \exp\left(-\lambda_1\epsilon^{2\rho-2}\min\{k(\epsilon), k\} - \frac{\lambda_0\epsilon^{\rho-1}}{D_1}\sum_{i>k(\epsilon)}^{k}\alpha_i\right)\right\}.$$

If $\beta < \frac{2\rho}{3\rho-1}$ and $\rho > 1$, then with probability 1,

$$\limsup_{k\to\infty} \frac{D_h(\mathcal{X}^\star, x_k)}{k^{\frac{1}{\rho-1}}} < \infty.$$

In the case that the growth condition of Assumption A1 holds with constant $\rho = 1$, we recover Proposition 2 of our earlier result [2], that is, that the procedure converges linearly. As it is, under weaker growth conditions, we see that we have somewhat weaker convergence. Nonetheless, this convergence rate is in fact adaptive and what we might hope for: in a local minimax (instance-specific) sense, stochastic minimization algorithms of the function $f(x) = |x|^\rho$ can have no better convergence than $|\widehat{x}_k - x^\star|^2 \lesssim k^{\frac{1}{\rho-1}}$ (cf. [41]).

Proposition 4 highlights the importance of a good model, i.e. one satisfying Condition (C.iv), and an appropriate divergence, i.e. one satisfying Assumption A1. Under such conditions, the method (4) yields optimization procedures with near optimal convergence guarantees, providing adaptivity to problem difficulty. In contrast, other methods—such as stochastic subgradient—or other divergences (Example 3) do not enjoy similar convergence or adaptivity.

4.2 Strongly convex functions

Strong convexity allows substantially faster convergence for stochastic gradient procedures, including in cases using generalized notions of convexity with respect to divergences [19, 20, 17], though these methods are frequently sensitive to the choice of stepsize (and may exhibit extremely slow convergence [27], even $\Omega(1/\log k)$). To that end, in this section, we revisit the stochastic proximal point models when the functions $f(\cdot; s)$ are strongly convex, showing that the exact model (7) is insensitive to the stepsize choice in this case. We have the following assumption.

Assumption A2 (Strong convexity)**.** *The functions $f(\cdot; s)$ are λ-strongly convex with respect to h, that is,*

$$f(y; s) \geq f(x; s) + \langle f'(x; s), y - x\rangle + \lambda D_h(y, x)$$

for all $f'(x; s) \in \partial f(x; s)$.

If Assumption A2 holds, we have the following convergence guarantees for the exact model.

Proposition 5. *Let Assumptions A2 hold, and let x_k be generated by iteration (4) using the exact model (7) with stepsizes $\alpha_k = \alpha_0 k^{-\beta}$ for some $\beta \in (0, 1)$. Assume further that $\mathbb{E}[\|f'(x^\star; S)\|_{h^*}^2] \leq \sigma_h^2$ for all $x^\star \in \mathcal{X}^\star$, and denote $\lambda_0 = \frac{\lambda}{1+\lambda\alpha_1}$. Then for a numerical constant $C < \infty$,*

$$\mathbb{E}[D_h(x^\star, x_{k+1})] \leq \exp\left(-\lambda_0 \sum_{i=1}^k \alpha_i\right) D_h(x^\star, x_1) + C \cdot \frac{\sigma_h^2}{\lambda_0} \alpha_k \cdot \log k.$$

Proposition 5 demonstrates the two facets of this paper: while accurate models—the exact model in this case—are much more robust to stepsize choice than standard gradient methods for strongly convex functions, the choice of divergence D_h may modify the bounds substantially through the dual (gradient) quantity σ_h^2 and primal distance $D_h(x^\star, x_1)$.

5 Experiments

We conclude our paper with an empirical study. In contrast to [2], which provides experiments for the effects of different models, our experiments test both the choice of model and divergence, emphasizing the importance of both aspects. We consider six procedures:

(i) SGM: uses the linear model (5) with the Euclidean distance $D_h(x, y) = \frac{1}{2}\|x - y\|_2^2$.

(ii) Truncated: uses the lower truncated model (6) with the Euclidean distance.

(iii) SGM-KL: uses the linear model (5) with the KL-divergence $D_h(x, y) = \sum_i x_i \log \frac{x_i}{y_i}$.

(iv) Trunc-KL: uses the lower truncated model (6) with the KL-divergence.

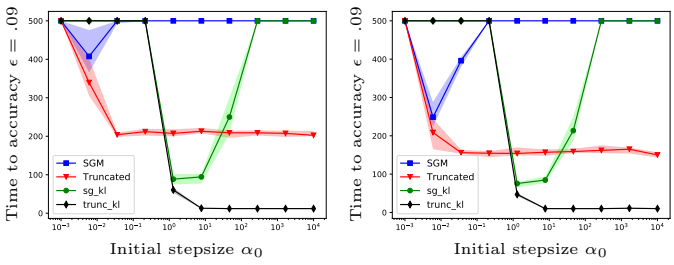
(v) SGM-adagrad: uses the linear model (5) with Mahalanobis divergence (9) using Adagrad matrix (10).

(vi) Trunc-adagrad: uses the lower truncated model (6) with Mahalanobis divergence (9) using Adagrad matrix (10).

In each experiment, we run each procedure for a range of initial stepsizes for a total of K iterations, where we decrease the stepsizes using the rule $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (1/2, 1)$. We calculate the number of iterations to achieve ϵ-accuracy, $F(x_k) - F(x^\star) \leq \epsilon$. We repeat the above process T times, reporting the median number of iterations required to achieve ϵ-accuracy and displaying 90% confidence intervals.

5.1 Robust Linear Regression

In our robust linear regression experiments, we let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $F(x) = \frac{1}{m}\|Ax - b\|_1$, where in each individual experiment we generate $x^\star \in \mathbb{R}^n$ uniformly at random from the simplex with sparsity s, i.e.



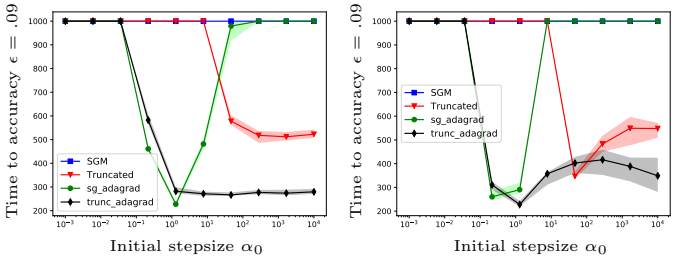
(a) (b)

Figure 1. Robust linear regression with $m = 500$, $n = 3000$, and $s = 20$. (a) The noiseless setting with $\sigma = 0$. (b) Noisy setting with $\sigma = 0.01$.

$\mathcal{X} = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, x_i \geq 0, \|x\|_0 = s\}$, giving rise to the KL-divergence as a natural divergence to consider for this setting. We generate the entires of A independently from $\mathsf{N}(0,1)$ and set $b = Ax^\star + \sigma v$ for $v \sim \mathsf{N}(0, I_m)$. We choose σ differently depending on the experiment, setting $\sigma = 0$ in noiseless experiments and $\sigma = 0.01$ otherwise.

Figure 1 shows our plots for this experiment in the noisy and noiseless settings, illustrating the dependence of APROX on model and divergence choice. The plots show that the robustness of the methods to the stepsize specification is dependent on the accuracy of the model, regardless of the chosen divergence. However, the convergence rate for the best stepsize value significantly depends on the underlying divergence; indeed, when using the KL-divergence, the best stepsize performance of SGM outperforms those achieved by the truncated model with the Euclidean distance.

5.2 Hinge Classification



(a) (b)

Figure 2. Hinge classification with $m = 5000$ and $n = 1000$. (a) The noiseless setting with $\sigma = 0$. (b) Noisy setting with $\sigma = 0.05$.

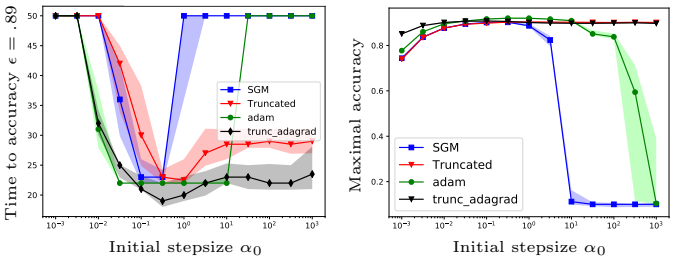
In this section, we test our methods for classification problems, where we have a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $b \in \{-1, +1\}^m$, and our goal is find a classifier $x^\star$ such that $\text{sign}(\langle a_i, x^\star \rangle) = b_i$ for almost every i, where a_i is the i'th row of A. To this end, we minimize

$$F(X) = \frac{1}{m} \sum_{i=1}^{m} [1 - b_i \langle a_i, x \rangle]_+ . \tag{15}$$

We generate the entries A_{ij} as follows: with probability $1 - 1/j$ set $A_{ij} = 0$, and with probability $1/j$ choose A_{ij} uniformly from $\{-1, +1\}$, so that A is sparse (a situation in which we expect AdaGrad to exhibit improved performance [12, 26, 13]). We choose $x^\star$ uniformly at random from $\{-1, +1\}^n$, and set $b_i = \text{sign}(\langle a_i, x^\star \rangle)$ for every $i \in [m]$. In the noisy setting we flip the sign of b_i with probability σ.

We present the results of this experiment in Figure 2. The plots tell a similar story to the previous experiment, that is, model choice can affect robustness to stepsize value, while the divergence can contribute to the performance achieved by the best stepsize choice.

5.3 CIFAR10 classification



(a) (b)

Figure 3. CIFAR10 dataset. (a) The number of iterations to achieve ϵ-accuracy versus initial stepsize α_0. (b) The maximal accuracy achieved after $T = 50$ iterations.

In our last experiment, we test the performance of our models for training neural networks for classification task over the CIFAR10 dataset [24]. We use the Resnet18 architechture [21] (while replacing all Relu activations with Elu), and run each optimization procedure for $T = 50$ iterations. Here, we also compare our optimization methods to Adam, the default optimizer in the TensorFlow package [1].

Figure 3 shows our plots for this experiment, where Figure 3(a) shows the number of iterations required to achieve $\epsilon = 0.89$-classification accuracy, and Figure 3(b) plots the maximal accuracy that each procedure achieve after $T = 50$ iterations. The results confirm our previous insights in this setting as well, showing that using accurate models improves robustness to the stepsize choice. Moreover, using better divergence (i.e. Mahalanobis divergence with Adagrad in this case) can improve the best convergence rates of any given model.

References

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[2] Asi, H. and Duchi, J. C. (2018). Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *arXiv:1810.05633 [math.OC]*.

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

[4] Belkin, M., Hsu, D., and Mitra, P. (2018a). Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. *arXiv:1806.05161 [stat.ML]*.

[5] Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2018b). Does data interpolation contradict statistical optimality? *arXiv:1806.09471 [stat.ML]*.

[6] Bertsekas, D. P. (1973). Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231.

[7] Bertsekas, D. P. (2011). Incremental proximal methods for large scale convex optimization. *Mathematical Programming, Series B*, 129:163–195.

[8] Bianchi, P. (2016). Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260.

[9] Bottou, L. and Bousquet, O. (2007). The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems 20*.

[10] Davis, D. and Drusvyatskiy, D. (2018). Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, To Appear.

[11] Davis, D., Drusvyatskiy, D., and MacPhee, K. (2018). Stochastic model-based minimization under high-order growth. *arXiv:1807.00255 [math.OC]*.

[12] Duchi, J. C., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

[13] Duchi, J. C., Jordan, M. I., and McMahan, H. B. (2013). Estimation, optimization, and parallelism when data is sparse. In *Advances in Neural Information Processing Systems 27*.

[14] Duchi, J. C. and Ruan, F. (2018a). The right complexity measure in locally private estimation: It is not the Fisher information. *arXiv:1806.05756 [stat.TH]*.

[15] Duchi, J. C. and Ruan, F. (2018b). Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, To appear.

[16] Gentile, C. (2003). The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299.

[17] Ghadimi, S. and Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492.

[18] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, second edition.

[19] Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192.

[20] Hazan, E. and Kale, S. (2011). An optimal algorithm for stochastic strongly convex optimization. In *Proceedings of the Twenty Fourth Annual Conference on Computational Learning Theory*.

[21] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[22] Hiriart-Urruty, J. and Lemaréchal, C. (1993). *Convex Analysis and Minimization Algorithms I & II*. Springer, New York.

[23] Karampatziakis, N. and Langford, J. (2011). Online importance weight aware updates. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*.

[24] Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

[25] Kulis, B. and Bartlett, P. (2010). Implicit online learning. In *Proceedings of the 27th International Conference on Machine Learning*.

[26] McMahan, B. and Streeter, M. (2010). Adaptive bound optimization for online convex optimization. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*.

[27] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

[28] Nemirovski, A. and Yudin, D. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley.

[29] Orabona, F., Crammer, K., and Cesa-Bianchi, N. (2015). A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435.

[30] Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

[31] Robbins, H. and Siegmund, D. (1971). A convergence theorem for non-negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York.

[32] Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898.

[33] Ryu, E. and Boyd, S. (2014). Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent.

[34] Shalev-Shwartz, S. (2007). *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem.

[35] Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011). Pegasos: primal estimated subgradient solver for SVM. *Mathematical Programming, Series B*, 127(1):3–30.

[36] Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Programming Society.

[37] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

[38] Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. (2017). The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems 31*.

[39] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *Proceedings of the Fifth International Conference on Learning Representations*.

[40] Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*.

[41] Zhu, Y., Chatterjee, S., Duchi, J., and Lafferty, J. (2016). Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems 30*.

[42] Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*.

6 Proofs of results from Section 3

6.1 Proof of Theorem 1

First, we prove the following lemma.

Lemma 6.1. *Let g be convex and subdifferentiable on a closed convex set $\mathcal{X}$ and let $\beta > 0$. Then for all $x_0, x_1, y \in \mathcal{X}$, and $g'(y) \in \partial g(y)$,*

$$g(y) - g(x_1) \le \langle g'(y), y - x_0 \rangle + \frac{1}{\beta} D_h(x_1, x_0) + \frac{\beta}{2} \left\| g'(y) \right\|_{h^*}^2$$

Proof By the first-order conditions for convexity, we have

$$\begin{aligned} & g(y) - g(x_1) \\ & \le \langle g'(y), y - x_1 \rangle = \langle g'(y), y - x_0 \rangle + \langle g'(y), x_0 - x_1 \rangle \\ & \le \langle g'(y), y - x_0 \rangle + \frac{1}{2\beta} \left\| x_1 - x_0 \right\|_h^2 + \frac{\beta}{2} \left\| g'(y) \right\|_{h^*}^2 \\ & \le \langle g'(y), y - x_0 \rangle + \frac{1}{\beta} D_h(x_1, x_0) + \frac{\beta}{2} \left\| g'(y) \right\|_{h^*}^2, \end{aligned}$$

where the second line uses the Fenchel-Young inequality, and the last one uses the strong convexity of the bregman divergence with respect to $\|\cdot\|_h$. □

We also have the following lemma, which gives a one-step progress guarantee for any algorithm using models satisfying Conditions (C.i)–(C.iii).

Lemma 6.2. *Let Condition (C.i) hold. In each step of the method (4), for any $x \in \mathcal{X}$,*

$$\begin{aligned} D_h(x, x_{k+1}) \le D_h(x, x_k) - \alpha_k \left[f_{x_k}(x_{k+1}; S_k) - f_{x_k}(x; S_k) \right] \\ - D_h(x_{k+1}, x_k) \end{aligned}$$

Proof By the first-order conditions for convex optimization, for some $g_k \in \partial f_{x_k}(x_{k+1}; S_k)$ we have that $\langle \alpha_k g_k + (h'(x_{k+1}) - h'(x_k)), y - x_{k+1} \rangle \ge 0$ for all $y \in \mathcal{X}$. Setting $y = x$, we obtain

$$\begin{aligned} \alpha_k \langle g_k, x_{k+1} - x \rangle & \le \langle h'(x_{k+1}) - h'(x_k), x - x_{k+1} \rangle \\ & = D_h(x, x_k) - D_h(x, x_{k+1}) \\ & \quad - D_h(x_{k+1}, x_k). \end{aligned}$$

As $f_{x_k}(x; S_k) \ge f_{x_k}(x_{k+1}; S_k) + \langle g_k, x - x_{k+1} \rangle$ by Condition (C.i), this gives the result. □

With Lemmas 6.1 and 6.2 in place, we can prove the theorem. Let $x^\star \in \mathcal{X}^\star$ be an otherwise arbitrary optimal point. Applying Lemma 6.2 with $x = x^\star$, we have

$$\begin{aligned} & D_h(x^\star, x_{k+1}) \\ & \le D_h(x^\star, x_k) - \alpha_k \left[f_{x_k}(x_{k+1}; S_k) - f_{x_k}(x^\star; S_k) \right] \\ & \quad - D_h(x_{k+1}, x_k) \\ & \overset{(i)}{\le} D_h(x^\star, x_k) - \alpha_k \left[f(x_{k+1}; S_k) - f_{x_k}(x^\star; S_k) \right] \\ & \quad - \epsilon D_h(x_{k+1}, x_k) + C(S_k) \alpha_k^2 \\ & \overset{(ii)}{\le} D_h(x^\star, x_k) - \alpha_k \left[f(x_{k+1}; S_k) - f(x^\star; S_k) \right] \\ & \quad - \epsilon D_h(x_{k+1}, x_k) + C(S_k) \alpha_k^2, \end{aligned}$$

where inequality (i) is a consequence of the accurate model condition (C.v) and (ii) because $f_x(x^\star; s) \le f(x^\star; s)$ by the lower model condition (C.ii). Now, we apply Lemma 6.1 with $x_1 = x_{k+1}$, $x_0 = x_k$, $y = x^\star$, and $\beta = \frac{\alpha_k}{\epsilon}$ to find

$$\begin{aligned} D_h(x^\star, x_{k+1}) \le D_h(x^\star, x_k) + \alpha_k \langle f'(x^\star; S_k), x^\star - x_k \rangle \\ + \frac{\alpha_k^2}{2\epsilon} \left\| f'(x^\star; S_k) \right\|_{h^*}^2 + C(S_k) \alpha_k^2 \end{aligned}$$

for all $f'(x^\star; S_k) \in \partial f(x^\star; S_k)$.

For some $F'(x^\star) \in \partial F(x^\star)$, we have $\langle F'(x^\star), y - x^\star \rangle \ge 0$ for all $y \in \mathcal{X}$. As our choice of $f'(x^\star; s) \in \partial f(x^\star; s)$ above was arbitrary, we may take $f'(x^\star; S_k)$ so that $\mathbb{E}[f'(x^\star; S_k)] = F'(x^\star)$ for any desired $F'(x^\star) \in \partial F(x^\star)$ (cf. [6]). Thus, taking expectations with respect to $\mathcal{F}_{k-1}$,

$$\begin{aligned} \mathbb{E}[D_h(x^\star, x_{k+1}) \mid \mathcal{F}_{k-1}] \le D_h(x^\star, x_k) + \frac{\alpha_k^2}{2\epsilon} \mathbb{E} \left[\left\| f'(x^\star; S) \right\|_{h^*}^2 \right] \\ + \mathbb{E}[C(S)] \alpha_k^2 + \alpha_k \langle F'(x^\star), x^\star - x_k \rangle. \end{aligned}$$

As $\langle F'(x^\star), x^\star - x_k \rangle \le 0$, we obtain the theorem.

6.2 Proof of Corollary 3.1

The first part of the proof follows by iteratively applying the recursion from Theorem 1.

The second part uses the Robbins-Siegmund almost supermartingale convergence lemma.

Lemma 6.3 ([31])**.** *Let $A_k, B_k, C_k, D_k \ge 0$ be non-negative random variables adapted to the filtration $\mathcal{F}_k$ and satisfying $\mathbb{E}[A_{k+1} \mid \mathcal{F}_k] \le (1 + B_k) A_k + C_k - D_k$. Then on the event $\{\sum_k B_k < \infty, \sum_k C_k < \infty\}$, there is a random $A_\infty < \infty$ such that $A_k \overset{a.s.}{\to} A_\infty$ and $\sum_k D_k < \infty$.*

By applying Theorem 1 with $A_k = D_h(\mathcal{X}^\star, x_{k+1})$, $C_k = \alpha_{k+1}(\sigma^2/\epsilon + \mathbb{E}[C(S)])$, and $B_k = D_k = 0$, we get the corollary.

6.3 Proof of Proposition 1

To prove Proposition 1, we present a lemma giving a one-step progress guarantee for any method satisfying Conditions (C.i)–(C.iii).

Lemma 6.4. *Let Conditions (C.i)–(C.iii) hold and let x_k be generated by the updates (4). Then for any $x \in \mathcal{X}$,*

$$D_h(x, x_{k+1}) \le D_h(x, x_k) - \alpha_k[f(x_k; S_k) - f(x; S_k)] + \frac{\alpha_k^2}{2} \left\|f'(x_k; S_k)\right\|_{h^*}^2 .$$

Proof Using Lemma 6.2, it suffices to show that for any $\alpha > 0$ and $x_0, x_1, x \in \mathcal{X}$

$$-\alpha[f_{x_0}(x_1; s) - f_{x_0}(x; s)] - D_h(x_1, x_0) \le -\alpha[f(x_0; s) - f(x; s)] + \frac{\alpha^2}{2} \left\|f'(x_0; s)\right\|_{h^*}^2 .$$

To see this, note that

$$\begin{aligned} &-f_{x_0}(x_1; s) + f_{x_0}(x; s) \\ &= -[f_{x_0}(x_0; s) - f_{x_0}(x; s)] + f_{x_0}(x_0; s) - f_{x_0}(x_1; s) \\ &\overset{(C.iii)}{\le} -[f_{x_0}(x_0; s) - f_{x_0}(x; s)] + \langle f'(x_0; s), x_0 - x_1\rangle \\ &\overset{(C.ii)}{\le} -[f(x_0; s) - f(x; s)] + \langle f'(x_0; s), x_0 - x_1\rangle. \end{aligned}$$

Then we use the Fenchel-Young inequality and strong convexity of D_h with respect to $\|\cdot\|_h$ to get

$$\begin{aligned} \langle f'(x_0; s), x_0 - x_1\rangle &\le \frac{1}{2\alpha} \left\|x_1 - x_0\right\|_h^2 + \frac{\alpha}{2} \left\|f'(x_0; s)\right\|_{h^*}^2 \\ &\le \frac{1}{\alpha} D_h(x_1, x_0) + \frac{\alpha}{2} \left\|f'(x_0; s)\right\|_{h^*}^2 , \end{aligned}$$

which finishes the proof. □

We are now ready to prove the proposition. We have that $\mathbb{E}[\|f'(x; S)\|_{h^*}^2] \le \mathsf{G}_{\text{big}}(r)$ for all x such that $D_h(x^\star, x) \le r$. Take $x^\star$ as the projection of x_k onto $\mathcal{X}^\star$ with respect to the Bregman divergence $D_h(\cdot, \cdot)$. Then Lemma 6.4 implies that

$$\mathbb{E}[D_h(x^\star, x_{k+1}) \mid \mathcal{F}_{k-1}] \le D_h(\mathcal{X}^\star, x_k)^2 - 2\alpha_k(F(x_k) - F^\star) + \alpha_k^2 \mathsf{G}_{\text{big}}(D_h(x^\star, x_k)).$$

On the event that $\sup_k D_h(\mathcal{X}^\star, x_k) < \infty$, we have $\sum_k \alpha_k^2 \mathsf{G}_{\text{big}}(D_h(\mathcal{X}^\star, x_k)) < \infty$, and so the Robbins-Siegmund Lemma 6.3 implies that $D_h(\mathcal{X}^\star, x_k) \overset{a.s.}{\to} D_\infty$ for some finite random variable D_∞ and $\sum_k \alpha_k(F(x_k) - F(x^\star)) < \infty$. If $\Gamma(\epsilon) > 0$, then a simple argument by contradiction shows that $D_\infty = 0$ with probability 1, as $\sum_k \alpha_k = \infty$.

6.4 Proof of Proposition 2

The proposition follows nearly directly from Lemma 6.4. Indeed, applying that lemma, for any $x^\star \in \mathcal{X}^\star$ we have

$$\mathbb{E}[D_h(x^\star, x_{k+1}) \mid \mathcal{F}_{k-1}] \le D_h(x^\star, x_k) - \alpha_k[F(x_k) - F(x^\star)] + \frac{\alpha_k^2}{2} \left\|f'(x_k; S_k)\right\|_{h^*}^2 ,$$

where we have used that $x_{k+1} \in \mathcal{F}_{k-1}$. This implies that

$$\alpha_k \mathbb{E}[F(x_k) - F(x^\star)] \le \mathbb{E}\left[D_h(x^\star, x_k) - D_h(x^\star, x_{k+1})\right] + \frac{\alpha_k^2}{2} \mathbb{E}\left[\left\|f'(x_k; S_k)\right\|_{h^*}^2\right].$$

Summing and telescoping yields $\sum_{i=1}^k \alpha_i \mathbb{E}[F(x_i) - F(x^\star)] \le D_h(x^\star, x_1) + \frac{1}{2}\sum_{i=1}^k \alpha_i^2 \mathbb{E}\left[\left\|f'(x_k; S_k)\right\|_{h^*}^2\right]$, and dividing by $\sum_{i=1}^k \alpha_i$ and using Jensen's inequality gives the first result of the proposition.

For the second result, we rearrange the first display above to see that

$$\mathbb{E}[F(x_k) - F(x^\star)] \le \frac{1}{\alpha_k} \mathbb{E}[D_h(x^\star, x_k) - D_h(x^\star, x_{k+1})] + \frac{\alpha_k}{2} M_h^2.$$

Summing this quantity, we obtain

$$\sum_{i=1}^k \mathbb{E}[F(x_i) - F(x^\star)] \le \sum_{i=2}^k \left(\frac{1}{\alpha_i} - \frac{1}{\alpha_{i-1}}\right) \mathbb{E}[D_h(x^\star, x_i)] + \frac{1}{\alpha_1} D_h(x^\star, x_1) + \frac{M^2}{2} \sum_{i=1}^k \alpha_i.$$

Noting that $\mathbb{E}[D_h(x^\star, x_i)] \le R$ by assumption, dividing by k and applying Jensen's inequality to $F(\overline{x}_k) \le \frac{1}{k}\sum_{i=1}^k F(x_i)$ gives the result.

6.5 Proof of Proposition 3

For each k, the function $D_k(x, y) = \frac{1}{2}(x-y)^T H_k (x-y)$ is a divergence generated by $h(x) = \frac{1}{2} x^T H_k x$. Thus, we use Lemma 6.4 to get

$$\left\|x_k - x^\star\right\|_{H_k}^2 \le \left\|x_k - x^\star\right\|_{H_k}^2 - 2\alpha[f(x_k; S_k) - f(x^\star; S_k)] + \alpha^2 \left\|f'(x_k; S_k)\right\|_{H_k^{-1}}^2 .$$

This implies that

$$2\alpha \sum_{i=0}^k f(x_i; S_i) - f(x^\star; S_i) \le \left\|x_1 - x^\star\right\|_{H_1}^2 + \sum_{i=1}^k \left\|x_i - x^\star\right\|_{H_i}^2 - \left\|x_i - x^\star\right\|_{H_{i-1}}^2 + \alpha^2 \sum_{i=1}^k \left\|f'(x_i; S_i)\right\|_{H_i^{-1}}^2$$

The first part of the claim now follows from Jensen's inequality. The second part of the proof follows from the first part using standard arguments from stochastic optimization with adaptive step sizes.

6.6 Proof of inequalities (11)

The lower bound that $\frac{p-1}{2}\|x-y\|_p^2 \le D_h(x,y)$ is a standard strong convexity result [34, Example 5, Appendix A]. For the upper bound, we note that (see, e.g., Gentile [16, Lemma 4]) that

$$D_h(x,y) \le \frac{q-1}{2}\|\nabla h(x) - \nabla h(y)\|_q^2 .$$

Now, we compute bounds on the norm error. By the triangle inequality, we have

$$\begin{aligned}
&\|\nabla h(x) - \nabla h(y)\|_q \\
&\le \left\|\nabla h(x)\frac{\|y\|_p^{2-p}}{\|x\|_p^{2-p}} - \nabla h(y)\right\|_q + \|\nabla h(x)\|_q \left|\frac{\|y\|_p^{2-p}}{\|x\|_p^{2-p}} - 1\right| \\
&= \|y\|_p^{2-p}\left(\sum_{j=1}^n |\operatorname{sign}(x_j)|x_j|^{p-1} - \operatorname{sign}(y_j)|y_j|^{p-1}|^q\right)^{1/q} \\
&\quad + \|x\|_p^{p-1}\left|\|y\|_p^{2-p} - \|x\|_p^{2-p}\right| \\
&\le \|y\|_p^{2-p}\left(\sum_{j=1}^n |x_j - y_j|^{q(p-1)}\right)^{1/q} \\
&\quad + \|x\|_p^{p-1}\left|\|y\|_p^{2-p} - \|x\|_p^{2-p}\right| ,
\end{aligned}$$

where we have used that $|s^{p-1} - t^{p-1}| \le |t-s|^{p-1}$. Now, noting that $q(p-1) = p$ and $p/q = 1$ we have

$$\|y\|_p^{2-p}\left(\sum_{j=1}^n |x_j - y_j|^{q(p-1)}\right)^{1/q} = \|y\|_p^{2-p}\|x-y\|^{p-1} .$$

Now, let us assume that $\|y\|_p \ge \|x\|_p$, so that $\|y\|_p = \|x\|_p + \delta$ for some $0 \le \delta \ge \|x-y\|_p$. Letting $u = \|y\|_p$, we obtain

$$\begin{aligned}
&\|x\|_p^{p-1}\left|\|y\|_p^{2-p} - \|x\|_p^{2-p}\right| \\
&= u^{p-1}\left((u+\delta)^{2-p} - u^{2-p}\right) \le (2-p)u^{p-1}\frac{\delta}{u^{p-1}},
\end{aligned}$$

where we have used the concavity of $\delta \mapsto (u+\delta)^{2-p}$, as $p \in [1,2]$. We obtain that if $\|y\|_p \ge \|x\|_p$, then

$$\|\nabla h(x) - \nabla h(y)\|_q \le \|y\|_p^{2-p}\|x-y\|_p^{p-1} + (2-p)\|x-y\|_p .$$

The case that $\|x\|_p \ge \|y\|_p$ is completely similar, and squaring the gives the result.

7 Proofs of fast convergence on easy problems

7.1 Proof of Lemma 4.1

We assume without loss of generality that $f(x^\star; s) = 0$ for all $x^\star \in \mathcal{X}^\star$, as we may replace f with $f - \inf f$. By Lemma 6.2, the update (4) satisfies

$$\begin{aligned}
D_h(x^\star, x_{k+1}) &\le D_h(x^\star, x_k) - D_h(x_{k+1}, x_k) \\
&\quad + \alpha_k\left[f_{x_k}(x^\star; S_k) - f_{x_k}(x_{k+1}; S_k)\right].
\end{aligned}$$

Denote $g_k = f'(x_k; S_k)$, $f_k = f(x_k; S_k)$ and $\widetilde{f}_{x_k}(x) = [f_k + \langle g_k, x_{k+1} - x_k\rangle]_+$. As $f_{x_k}(x^\star; S_k) \le f(x^\star; S_k) = 0$, and by Condition (C.iii) and Condition (C.iv) we have $f_{x_k}(x_{k+1}; S_k) \ge \widetilde{f}_{x_k}(x_{k+1})$, we have

$$\begin{aligned}
&D_h(x^\star, x_{k+1}) \\
&\le D_h(x^\star, x_k) - \alpha_k\widetilde{f}_{x_k}(x_{k+1}) - D_h(x_{k+1}, x_k) \\
&\le D_h(x^\star, x_k) - \alpha_k\widetilde{f}_{x_k}(x_{k+1}) - \frac{1}{2}\|x_{k+1} - x_k\|_h^2 \quad (16) \\
&\le D_h(x^\star, x_k) - \alpha_k \inf_x\left\{\widetilde{f}_{x_k}(x) + \frac{1}{2\alpha_k}\|x - x_k\|_h^2\right\}.
\end{aligned}$$

Let $u(\Delta) = \frac{1}{2}\|\Delta\|_h^2$, so that $u^*(v) = \frac{1}{2}\|v\|_{h^*}^2$ and $\langle v, \nabla u^*(\lambda v)\rangle = \lambda\|v\|_{h^*}^2$ by standard duality calculations [22, Chap. X]. Let $\widetilde{x}_{k+1}$ denote the unconstrained minimizer

$$\widetilde{x}_{k+1} = \operatorname*{argmin}_x\left\{[f_k + \langle g_k, x - x_k\rangle]_+ + \frac{1}{2\alpha_k}\|x - x_k\|_h^2\right\}. \quad (17)$$

Let us now split the proof to two cases, depending on whether $f_k \lesseqgtr \alpha_k\|g_k\|_{h^*}^2$. First, in the case that $f_k \ge \alpha\|g_k\|_{h^*}^2$, the minimizer in Eq. (17) is $\widetilde{x}_{k+1} = \nabla u^*(-\alpha_k g_k)$, because this also attains the infimum in the fully unconstrained minimization

$$\begin{aligned}
&\inf_x\left\{f_k + \langle g_k, x - x_k\rangle + \frac{1}{2\alpha_k}\|x - x_k\|_h^2\right\} \\
&\qquad = f_k - \frac{\alpha_k}{2}\|g_k\|_{h^*}^2 \ge \frac{f_k}{2},
\end{aligned}$$

as

$$f_k + \langle g_k, \nabla u^*(-\alpha_k g_k)\rangle = f_k - \alpha_k\|g_k\|_{h^*}^2 \ge 0$$

and

$$\frac{1}{2\alpha_k}\|\nabla u^*(-\alpha_k g_k)\|_h^2 = \frac{\alpha_k}{2}\|g_k\|_{h^*}^2 .$$

Here evidently the fully unconstrained solution coincides with $\widetilde{x}_{k+1}$ in Eq. (17). In the case that $f_k < \alpha_k\|g_k\|_{h^*}^2$, we must be somewhat more careful. A duality calculation shows that the solution $\Delta = \widetilde{x}_{k+1} - x_k$ satisfies $\Delta = \nabla u^*(-\lambda g_k)$ for the unique λ solving

$$0 = f_k + \langle g_k, \nabla u^*(-\lambda g_k)\rangle = f_k - \lambda\|g_k\|_{h^*}^2 ,$$

or $\lambda = \frac{f_k}{\|g_k\|_{h^*}^2}$ and $\|\Delta\|_h \|g_k\|_{h^*} = \langle g_k, \Delta\rangle$ and $\|\Delta\|_h = \frac{f_k}{\|g_k\|_{h^*}^2}$. For this Δ we evidently have

$$[f_k + \langle g_k, \Delta\rangle]_+ + \frac{1}{2\alpha_k}\|\Delta\|_h^2 = \frac{\lambda^2}{2\alpha_k}\|g_k\|_{h^*}^2 = \frac{f_k^2}{2\alpha_k\|g_k\|_{h^*}^2}.$$

Combining these two cases into inequality (16), we have

$$D_h(x^\star, x_{k+1}) \le D_h(x^\star, x_k) - \frac{1}{2}\min\left\{\alpha_k f_k, \frac{f_k^2}{\|g_k\|_{h^*}^2}\right\},$$

which is the desired result.

7.2 Proof of Proposition 4

Let $D_k^2 = D_h(\mathcal{X}^\star, x_k)$, so $D_k \in \mathcal{F}_{k-1}$. Then Lemma 4.1 implies that under Assumption A1, we have

$$\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] \le D_k^2 - \min\left\{\lambda_0\alpha_k D_k^\rho, \lambda_1 D_k^{2\rho}\right\}.$$

Now, let $\epsilon > 0$. At any iteration, we either have $D_k^2 \le \epsilon^2$, or, because D_k is decreasing, we have

$$\begin{aligned}\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] &\le D_k^2 - \min\{\lambda_0\alpha_k\epsilon^{\rho-1}D_k, \lambda_1\epsilon^{2\rho-2}D_k^2\} \\ &= \max\left\{1 - \lambda_1\epsilon^{2\rho-2}, 1 - \lambda_0\alpha_k\epsilon^{\rho-1}/D_k\right\}D_k^2 \\ &\le \max\left\{1 - \lambda_1\epsilon^{2\rho-2}, 1 - \lambda_0\alpha_k\epsilon^{\rho-1}/D_1\right\}D_k^2\end{aligned}$$

because $D_1 \le D_k$.

Now, we can use an analysis similar to Proposition 2 of our paper [2]. Solving

$$k(\epsilon) = \left(\frac{\lambda_1 D_1\epsilon^{\rho-1}}{\lambda_0\alpha_0}\right)^{-\frac{1}{\beta}}$$

gives $\lambda_0\alpha_k\epsilon^{\rho-1}/D_1 \le \lambda_1\epsilon^{2\rho-2}$ for all $k \ge k(\epsilon)$, and thus using $1 - x \le e^{-x}$ we have that if $D_k^2 \ge \epsilon^2$, then

$$\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] \le \exp\left(-\lambda_1\epsilon^{2\rho-2}\min\{k(\epsilon), k\} - \frac{\lambda_0\epsilon^{\rho-1}}{D_1}\sum_{i>k(\epsilon)}^{k}\alpha_i\right)D_1^2.$$

This gives the first claim of the proposition.

Now we ignore all constants $\lambda_0, \lambda_1, D_1$ to obtain the rate of convergence. Assume that $k > 2k(\epsilon)$. Then for $c > 0$, we have

$$\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] \le \max\left\{\epsilon^2, \exp\left(-c\epsilon^{\rho-1}k^{1-\beta}\right)D_1^2\right\}$$

for large enough k, and choosing $\epsilon = k^{\frac{\beta-1}{\rho-1}}\log^{\frac{2}{\rho-1}}k$ gives

$$\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] \le \max\left\{\frac{\log^{\frac{4}{\rho-1}}k}{k^{\frac{2(1-\beta)}{\rho-1}}}, \exp(-c\log^2 k)D_1^2\right\}.$$

Now, we follow the style of argument as in the proof of Proposition 2 of [2], eliding probabilistic details (see [2]). Eventually $\alpha_k D_k^\rho \le D_k^{2\rho}$ if

$$k^{-\beta}k^{-\frac{2\rho(1-\beta)}{\rho-1}} > k^{-\frac{4\rho(1-\beta)}{\beta-1}} \quad \text{i.e.} \quad \beta < \frac{2\rho}{3\rho-1}.$$

With this, then eventually the recursion is dominated by the $\lambda_1 D_k^{2\rho}$ term, and so we have

$$\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] \le D_k^2 - \lambda_1 D_k^{2\rho}$$

eventually, with probability 1. Then a similar calculation as above yields that eventually, we have

$$\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] \lesssim \max\left\{\epsilon^2, \exp\left(-c\epsilon^{2\rho-2}k\right)\right\},$$

and setting $\epsilon = (\log^2 k/k)^{\frac{1}{\rho-1}}$ gives the result.

7.3 Proof of Proposition 5

First, we prove the following singe-step progress recursion for proximal models.

Lemma 7.1. *Let Assumption A2 hold and the iterates x_k be generated by iteration (4) using the exact model (7). Then*

$$\mathbb{E}\left[D_h(x^\star, x_{k+1}) \mid \mathcal{F}_{k-1}\right] \le \left(1 - \frac{\alpha_k\lambda}{1+\alpha_k\lambda}\right)D_h(x^\star, x_k) + \frac{\alpha_k^2}{2}\mathbb{E}\left[\|f'(x^\star; S)\|_{h^*}^2\right].$$

Proof For all $g_k \in \partial f(x_{k+1}; S_k)$ and $y \in \mathcal{X}$, Assumption A2 implies

$$f(y; S_k) \ge f(x_{k+1}; S_k) + \langle g_k, y - x_{k+1}\rangle + \lambda D_h(y, x_{k+1}).$$

Using this inequality in place of the last step of the proof of Lemma 6.2 yields

$$\begin{aligned}&D_h(x^\star, x_{k+1}) + \alpha_k\lambda D_h(x^\star, x_{k+1}). \qquad (18)\\ &\le D_h(x^\star, x_k) - \alpha_k\left[f(x_{k+1}; S_k) - f(x^\star; S_k)\right] - D_h(x_{k+1}, x_k).\end{aligned}$$

Applying Lemma 6.1 with $y = x^\star$, $x_1 = x_{k+1}$, $x_0 = x_k$, $\beta = \alpha_k$, and $g(\cdot) = f(\cdot; S_k)$ implies

$$D_h(x^\star, x_{k+1})(1 + \lambda\alpha_k) \le D_h(x^\star, x_k) + \alpha_k\langle f'(x^\star; S_k), x^\star - x_k\rangle + \frac{\alpha_k^2}{2}\|f'(x^\star; S_k)\|_{h^*}^2.$$

Taking expectations and using that $\mathbb{E}[\langle f'(x^\star; S), x^\star - x\rangle] \le 0$ for all x (as in the proof of Theorem 1) gives the desired result.

□

Using Lemma 7.1, we are ready to prove the proposition. As the sequence of stepsizes α_k is decreasing, we

get that $\frac{\lambda}{1+\lambda\alpha_k} \geq \frac{\lambda}{1+\lambda\alpha_1}$. Thus, denoting $\lambda_0 = \frac{\lambda}{1+\lambda\alpha_1}$, Lemma 7.1 implies

$$\mathbb{E}\left[D_h(x^\star, x_{k+1}) \mid \mathcal{F}_{k-1}\right] \leq (1-\alpha_k\lambda_0)D_h(x^\star, x_k) + \frac{\alpha_k^2}{2}\sigma_h^2$$

Applying this inequality recursively gives

$$\begin{aligned}\mathbb{E}[D_h(x^\star, x_{k+1})] \leq \prod_{i=1}^{k}(1-\alpha_i\lambda_0)D_h(x^\star, x_1) \\ + \frac{1}{2}\sum_{i=1}^{k}\alpha_i^2\prod_{j=i+1}^{k}(1-\alpha_j\lambda_0)\sigma_h^2,\end{aligned}$$

where we note that $\alpha_j\lambda_0 < 1$ for all j. An inductive argument [30] implies the proposition.