Accelerated, Optimal and Parallel:
Some results on model-based stochastic optimization

Karan Chadha [*1] **Gary Cheng** [*1] **John Duchi** [1 2]

Abstract

The Approximate-Proximal Point (aProx) family of model-based stochastic optimization algorithms improve over standard stochastic gradient methods, as they are robust to step size choices, adaptive to problem difficulty, converge on a broader range of problems than stochastic gradient methods, and converge very fast on interpolation problems, all while retaining nice minibatching properties (Asi & Duchi, 2019b; Asi et al., 2020). In this paper, we propose an acceleration scheme for the aProx family and provide non-asymptotic convergence guarantees, which are order-optimal in all problem-dependent constants and provide even larger minibatching speedups. For interpolation problems where the objective satisfies additional growth conditions, we show that our algorithm achieves linear convergence rates for a wide range of stepsizes. In this setting, we also prove matching lower bounds, identifying new fundamental constants and showing the optimality of the aProx family. We corroborate our theoretical results with empirical testing to demonstrate the gains accurate modeling, acceleration, and minibatching provide.

1. Introduction

We move beyond stochastic and “minibatch”-gradient methods for stochastic optimization problems to extend (Asi et al., 2020)’s work on parallelizable and minibatch aware model-based and (approximate) proximal point methods for the problem

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

Here, $\mathcal{S}$ denotes the sample space, and $S \sim P$ is an $\mathcal{S}$-valued random variable, where for each sample $s \in \mathcal{S}$, $F(\cdot;s) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function, subdifferentiable on the closed convex domain $\mathcal{X}$.

First order stochastic methods are the default choice for solving problem (1). They enjoy numerous convergence guarantees (Zinkevich, 2003; Nemirovski et al., 2009; Bottou & Bousquet, 2007; Shalev-Shwartz et al., 2011), and extensions to parallelism and distributed computing that make them practically attractive (Lan, 2012; Dekel et al., 2012; Duchi et al., 2012). However, they are not robust to noise and hyperparameter tuning (Li et al., 2017; Asi & Duchi, 2019a;b); in fact, they may even diverge with slightly mis-specified stepsizes (Asi & Duchi, 2019b; Nemirovski et al., 2009). Motivated by the limitations of gradient methods, researchers (Bertsekas, 2011; Kulis & Bartlett, 2010; Davis & Drusvyatskiy, 2019; Duchi & Ruan, 2018; Asi & Duchi, 2019b) have developed stochastic (approximate) proximal-point (aProx) and model-based methods as a more robust alternative.

These aProx methods, as we explain in Section 2, construct a model of the function and iterate by minimizing regularized versions of the model. They improve over standard stochastic gradient methods, as they are adaptive to problem difficulty, converge on a broader range of problems than stochastic gradient methods, while retaining nice minibatching properties (Duchi & Ruan, 2018; Asi & Duchi, 2019b; Asi et al., 2020). Argubly most excitingly aProx also converge very fast on *interpolation problems*—that is, problems for which there exists $x^\star \in \mathcal{X}$ minimizing $F(\cdot;s)$ with P-probability 1—for a (very) wide selection of step size choices. Such problems arise in numerous modern machine learning applications (Belkin et al., 2018; 2019)—where one can achieve zero training error—or, for example, in finding a point in the intersection of convex sets $\cap_{i=1}^N C_i$, where one takes $\mathcal{S} = \{1,\dots,N\}$ and $F(x;i) = \text{dist}(x, C_i)$.

In spite of this progress, many questions remain open. aProx does not attain the classical smooth optimization lower bound as the noise tends to 0 (Nesterov, 2004); an improvement here could lead to even larger minibatching speedups. Additionally, the minibatch convergence rates (Asi et al., 2020) shown in the interpolation setting require

[*] Equal contribution; authors listed in alphabetical order. [1] Electrical Engineering Department, Stanford University, Stanford, CA [2] Statistics Department, Stanford University, Stanford, CA. Correspondence to: Karan Chadha <knchadha@stanford.edu>, Gary Cheng <chenggar@stanford.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

a small step size to work, dampening the message of the original APROX paper (Asi & Duchi, 2019b) that APROX is robust to step size choices. Finally, it is not clear whether further improvements to APROX can be made in the interpolation setting, as no optimality results have been shown.

In this paper, we answer these open questions. Like (Asi et al., 2020), we study methods to parallelize the APROX family via minibatched samples $S^{1:m} \in \mathcal{S}^m$, that is, where each iteration of the method receives an independent batch $S^{1:m} \overset{\text{iid}}{\sim} P$, developing several new results for model-based methods the problem (1) more generally along the way. Concretely, we provide the following:

1. *Non-asymptotic rates and accelerated convergence:* We develop an accelerated and minibatched version of the APROX family in Section 3 which improves on previous convergence results and is minimax optimal. This algorithm enjoys linear speedups in minibatch size up to the *cube* of the total number of iterations run.

2. *Optimal convergence and interpolation problems:* In Sections 4 and 5, for interpolation problems, we develop new lower bound results, characterizing (worst-case) problem difficulty based on a particular growth condition (Asi & Duchi, 2019b) introduce, which is (by these results) evidently fundamental; this result also shows that APROX is minimax optimal, with the correct problem dependent constants. We give some sufficient conditions for minibatching to yield improved convergence, and we further improve on the minibatching upper bounds presented in (Asi et al., 2020), by showing that the same (near)-linear convergence rates hold for a (very) wide range of step size schedules.

3. *Experimental evaluation:* We conclude with an experimental evaluation in Section 6, where we study the robustness and acceleration properties of the methods; performance profiles highlight the benefits of using these better models.

1.1. Related work

First order stochastic methods (Robbins & Monro, 1951) are the most popular method for minimizing stochastic objectives; an enormous literature gives numerous convergence results (Polyak, 1987; Polyak & Juditsky, 1992; Zinkevich, 2003; Nemirovski et al., 2009; Zhang, 2004; Kushner & Yin, 2003; Bach & Moulines, 2011). The growth of parallel computing has motivated the development of "minibatch" methods that use multiple samples S in each iteration, where researchers have shown how stochastic gradient-like methods enjoy linear speedups as batch sizes increase (Lan, 2012; Dekel et al., 2012; Duchi et al., 2012; Niu et al., 2011; Chaturapruek et al., 2015). Other work proposes accelerated stochastic optimization methods, showing faster (worst-case optimal) associated convergence rates (Lin et al., 2018; Lan, 2012). In spite of their successes, stochastic gradient methods still suffer a number of drawbacks. For example, misspecified stepsizes may force slow convergence for these methods (Nemirovski et al., 2009); objective functions without appropriate scaling or that grow too quickly may cause divergence (Asi & Duchi, 2019a;b); they can fail to adapt to problem geometry (Duchi et al., 2011; Levy & Duchi, 2019). To circumvent this, (Asi & Duchi, 2019b;a) show how better models in stochastic optimization yield improved stability, robustness, and convergence guarantees over classical stochastic gradient methods. The aim of our development is to extend accelerated convergence rates (as available for gradient-based methods (Lan, 2012; Nesterov, 2004)) to such model-based methods.

2. Preliminaries & Methods

The foundation of our methods is the model-based approximate proximal-point (APROX) framework (Davis & Drusvyatskiy, 2019; Duchi & Ruan, 2018; Asi & Duchi, 2019b), which approximates the functions F via *models* F_x of F localized at x, which satisfy the following conditions:

(C.i) *Convexity:* The function $y \mapsto F_x(y; s)$ is convex and subdifferentiable on $\mathcal{X}$.

(C.ii) *Lower bounds and local accuracy:* For all $y \in \mathcal{X}$,

$$F_x(y; s) \le F(y; s) \text{ and } F_x(x; s) = F(x; s).$$

Note that Condition (C.ii) immediately implies that $\partial F_x(y; s)|_{y=x} \subset \partial F(x; s)$. With such a model, APROX algorithms iteratively sample $S_k \overset{\text{iid}}{\sim} P$ and update

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

Typical choices for the models include the following three:

- *Stochastic gradient methods:* for any $F'(x; s) \in \partial F(x; s)$, use the linear model

$$F_x(y; s) := F(x; s) + \langle F'(x; s), y - x \rangle. \quad (3)$$

- *Stochastic proximal point methods:* use the true function

$$F_x(y; s) := F(y; s). \quad (4)$$

- *Truncated methods:* for any $F'(x; s) \in \partial F(x; s)$, use

$$F_x(y; s) := \max\left\{ F(x; s) + \langle F'(x; s), y - x \rangle, \inf_{z \in \mathcal{X}} F(z; s) \right\}. \quad (5)$$

> The model (5) is often simple to apply: in many applications, the objective is non-negative, so $\inf_{z\in\mathcal{X}} F(z;s) = 0$ and the model is simply the positive part of the linear approximation (3).

Now, we provide extensions to APROX to the minibatch setting, where instead of having one sample function at every iteration, we have m sample functions. (Asi et al., 2020) proposed using the model $\overline{F}(x;S_k^{1:m}) := \frac{1}{m}\sum_{i=1}^m F(x;S_k^i)$ which exhibits improved rates with minibatching.

It minimizes a model of the average at every iteration. With any model $\overline{F}_{x_k}(x;S_k^{1:m})$ of the average satisfying Conditions (C.i) and (C.ii), we can perform the update

$$x_{k+1} := \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ \overline{F}_{x_k}(x;S_k^{1:m}) + \frac{1}{2\alpha_k}\|x-x_k\|_2^2 \right\}. \tag{6}$$

In the case of truncated models (5), this leads to the following two natural algorithms, which can be parallelized efficiently (see (Asi et al., 2020) for details):

Truncated Average (*TruncAv*): For any lower bound $\Lambda(s^{1:m})$ on $\overline{F}(\cdot,s^{1:m})$, choose the model as $\overline{F}_x(y;s^{1:m}) := \max\left\{\overline{F}(x;s^{1:m}) + \langle \overline{F}'(x;s^{1:m}), y-x\rangle, \Lambda(s^{1:m})\right\}$. This results in the update,

$$\begin{aligned} x_{k+1} &= x_k \\ &- \min\left\{\alpha_k, \frac{\overline{F}(x_k;S_k^{1:m}) - \Lambda(S^{1:m})}{\|\overline{F}'(x_k;S_k^{1:m})\|_2^2}\right\} \overline{F}'(x_k;S_k^{1:m}). \end{aligned} \tag{7}$$

Average of Truncated Models (*AvMod*):

$$x_{k+1} := \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ \frac{1}{m}\sum_{i=1}^m F_{x_k}(x;S_k^i) + \frac{1}{2\alpha_k}\|x-x_k\|_2^2 \right\}. \tag{8}$$

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. We follow (Bertsekas, 1973), where we take $F'(x;s)$ to be any measureable selection in $\partial F(x;s)$, that is,

$$F'(x;s) = g(x;s) \in \partial F(x;s)$$

where $s \mapsto g(x,s)$ is P-measurable. We set $f'(x) = \int F'(x;s)dP(s) = \int g(x;s)dP(s)$ accordingly. We let $\mathcal{X}^\star = \operatorname{argmin}_{x\in\mathcal{X}} f(x)$ denote the optimal set of problem (1) and $x^\star \in \mathcal{X}^\star$ denote a single minimizer.

3. Non-Asymptotic Convergence Results

Our first set of theoretical results extends the familiar non-asymptotic rates of convergence for smooth convex stochastic optimization (Lan, 2012) to model-based methods. Here, we show that model-based methods for problem (1) enjoy optimal dependence on the variance of stochastic gradients, and, building off of (Tseng, 2008) and (Lan, 2012), can be accelerated to achieve worst-case optimal complexity. To present our results in the most generality, we allow non-Euclidean geometries to generalize mirror descent (Beck & Teboulle, 2003; Nemirovski et al., 2009).

To that end, recall that a differentiable convex function h is a *distance generating function* for $\mathcal{X}$ if it is strongly convex with respect to a norm $\|\cdot\|$ over $\mathcal{X}$, meaning $h(y) \ge h(x) + \langle \nabla h(x), y-x\rangle + \frac{1}{2}\|x-y\|^2$ for $x,y\in\mathcal{X}$. The associated *Bregman divergence* is then $D_h(x,y) := h(x) - h(y) - \langle \nabla h(y), x-y\rangle$, which evidently satisfies $D_h(x,y) \ge \frac{1}{2}\|x-y\|^2$. Recalling the dual norm $\|z\|_* = \sup_{\|x\|\le 1}\langle z,x\rangle$, throughout this section, we will work with the following standard assumption (Lan, 2012).

Assumption 1. *The function f has L-Lipschitz gradient with respect to the norm $\|\cdot\|$, meaning that*

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x-y\|,$$

and there exists $\sigma_0^2 < \infty$ such that for each $x\in\mathcal{X}$,

$$\mathbb{E}[\|\nabla f(x) - \nabla F(x;S)\|_*^2] \le \sigma_0^2.$$

When $D_h(x,y) \le R^2$ for all $x,y\in\mathcal{X}$ and Assumption 1 holds, mirror descent methods achieve convergence guarantees of the form $\frac{LR^2}{k} + \frac{\sigma_0 R}{\sqrt{k}}$, while accelerated methods (Lan, 2012) can achieve $\frac{LR^2}{k^2} + \frac{\sigma_0 R}{\sqrt{k}}$. The latter is worst-case optimal (Nemirovski & Yudin, 1983). We develop an accelerated analogue of the iteration (2), which gives a leading minimax-optimal $O(1/k^2)$ rate, by building off of the ideas of (Lan, 2012) and (Tseng, 2008). We consider a modified iteration, which augments the model-based update (2) with two auxiliary sequences whose momentum allows accelerated convergence. For full generality and completeness, we consider an augmented version of problem (1), where we wish to minimize

$$f(x) + r(x) = \mathbb{E}_P[F(x;S)] + r(x),$$

where r is a known convex function (typically a regularizer of some type). We require a non-increasing sequence $\theta_k \in [0,1]$ of stepsizes and consider the three term iteration

$$\begin{aligned} y_k &= (1-\theta_k)x_k + \theta_k z_k \\ z_{k+1} &= \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ F_{y_k}(x;S_k) + r(x) + \frac{1}{\alpha_k} D_h(x,z_k) \right\} \\ x_{k+1} &= (1-\theta_k)x_k + \theta_k z_{k+1}. \end{aligned} \tag{9}$$

All our analysis requires is that the additional stepsizes θ_k satisfy $\theta_0 = 1$, $\frac{1-\theta_k}{\theta_k^2} \le \frac{1}{\theta_{k-1}^2}$ for all k, and are non-increasing; for example, our choice $\theta_k = \frac{2}{k+2}$ satisfies these desiderata, as does taking any constant stepsize. We then have the following theorem; proof in Appendix A.1.

Theorem 1. *Let Assumption 1 hold, and assume that $D_h(x^\star, x) \le R^2$ for all $x \in \mathcal{X}$. Let (y_k, z_k, x_k) follow the three term iteration (9) for any model satisfying Conditions (C.i) and (C.ii). Take stepsizes $\theta_k = \frac{2}{k+2}$ and $\alpha_k = \frac{1}{L+\eta_k}$ for $\eta_k = \eta_0\sqrt{k+1}$, where $\eta_0 \ge 0$. Then*

$$\begin{aligned}&\mathbb{E}[f(x_{k+1}) + r(x_{k+1}) - f(x^\star) - r(x^\star)] \\ &\quad \le \frac{4LR^2}{(k+2)^2} + 2\frac{R^2}{\sqrt{k}}\left[\frac{\sigma_0^2}{\eta_0} + \eta_0\right].\end{aligned}$$

Specializing to the "minibatch" setting with $h(x) = \frac{1}{2}\|x\|_2^2$ again yields a minimax optimal algorithm for the class of problems we consider.

Corollary 3.1. *Let the conditions of Theorem 1 hold, except that we use a minibatch $S_k^{1:m} \overset{\text{iid}}{\sim} P$ of size m at each iteration, and $\overline{F}_{y_k}(\cdot; S_k^{1:m})$ is a model of $\frac{1}{m}\sum_{i=1}^m F(\cdot; S_k^i)$ satisfying Conditions (C.i) and (C.ii). Set $\eta_0 = \frac{\sigma_0\sqrt{m}}{R}$. Then*

$$\mathbb{E}[f(x_{k+1})+r(x_{k+1})-f(x^\star)-r(x^\star)] \le \frac{4LR^2}{(k+2)^2}+3\frac{R\sigma_0}{\sqrt{km}}.$$

The error rate $\mathcal{O}(1/k^2+1/\sqrt{km})$ is faster than the $\mathcal{O}(1/k+1/\sqrt{km})$ rate we showed for the basic minibatched APROX algorithm (2), and it is minimax rate optimal.

4. Interpolation Problems

In *interpolation problems*, there exists a consistent solution $x^\star \in \mathcal{X}$ satisfying $F(x^\star; S) = \inf_{z\in\mathcal{X}} F(z; S)$ with probability 1. While this is a strong assumption, it holds in numerous practical scenarios: in machine learning problems, where a perfect predictor (at least on training data) exists (Belkin et al., 2018; 2019; Ma et al., 2018); in problems of finding a point in the intersection $C^\star = \cap_{i=1}^N C_i$ of convex sets C_i, assuming $C^\star \neq \emptyset$, where we may take $F(x; i) = \operatorname{dist}(x, C_i)$ (e.g. (Bauschke & Borwein, 1996)); or in least-squares problems with consistent solutions (Needell et al., 2014; Strohmer & Vershynin, 2009). We show a few results in this section, first that model-based methods (often) enjoy linear convergence on these problems—in analogy to the results available for stochastic gradient methods (Ma et al., 2018)—while also demonstrating improvement via mini-batching and reducing variance. Second, we revisit the convergence guarantees that Asi and Duchi (Asi & Duchi, 2019b) provide, giving a unified treatment and some discussion of the possibilities of parallelism. These conditions appear on their face to be somewhat non-standard, but as we show, they capture the essential difficulty of interpolation problems, and we can provide sharp (matching to within numerical constants) lower bounds for optimization using them.

Definition 4.1. *Let $\mathcal{X}^\star := \operatorname{argmin}_{x\in\mathcal{X}} f(x)$. Then problem (1) is an* interpolation problem *if there exists $x^\star \in \mathcal{X}^\star$ such that for P-almost all $s \in \mathcal{S}$, we have $\inf_{x\in\mathcal{X}} F(x; s) = F(x^\star; s)$.*

We develop two sets of upper bounds for such interpolation problems. The first applies to any model-based method, while the second relies on the models having more fidelity to the functions F.

4.1. Upper bounds under smoothness and quadratic growth

Our first set of upper bounds relies on two assumptions about the growth of the function f at the optimum—which is weaker than typical strong convexity assumptions (Ma et al., 2018) that require quadratic growth everywhere—and the noise in its gradients.

Assumption 2 (Quadratic Population Growth)**.** *There exist $\lambda > 0$ such that for all $x \in \mathcal{X}$,*

$$f(x) - f(x^\star) \ge \lambda \operatorname{dist}(x, \mathcal{X}^\star)^2.$$

Assumption 3. *There exists $\sigma_2^2 < \infty$ such that for every $x \in \mathcal{X}$, we have $\mathbb{E}[\|\nabla f(x) - \nabla F(x; S)\|_2^2] \le \sigma_2^2 \operatorname{dist}(x, \mathcal{X}^\star)^2$.*

It is straightforward to give examples satisfying the assumptions; noiseless linear regression problems provide the simplest such approach.

Example 1: Consider a linear regression problem with data $s = (a, b) \in \mathbb{R}^n \times \mathbb{R}$, where $a^T x^\star = b$ for all (a, b), and set $F(x; (a, b)) = \frac{1}{2}(a^T x - b)^2$. If the data a belong to a subspace $V \subset \mathbb{R}^n$ (which may be $V = \mathbb{R}^n$), then Assumption 2 holds with $\lambda = \inf_{\|v\|_2=1}\{v^T\mathbb{E}[aa^T]v/2 \mid v \in V\}$, and it is immediate that $\operatorname{Var}(F'(x; S)) \le \mathbb{E}[\|a\|_2^2 \langle a, x - x^\star\rangle^2]$, so Assumption 3 holds with $\sigma_2^2 = \lambda_{\max}(\mathbb{E}[\|a\|_2^2 aa^T])$. For example, if a is uniform on the scaled sphere $\sqrt{n}\mathbb{S}^{n-1}$, then $\lambda = 1$ and $\sigma_2^2 = n$. $\diamond$

Alternatively, we may follow (Ma et al., 2018) by considering a problem where the functions have Lipschitz gradients:

Example 2: If $F(\cdot; s)$ has $L(s)$-Lipschitz gradient and problem (1) is an interpolation problem with $x^\star \in \operatorname{int}\mathcal{X}$, then $\nabla F(x^\star; S) = 0$ with probability 1, and so

$$\begin{aligned}&\mathbb{E}[\|\nabla f(x) - \nabla F(x; S)\|_2^2] \\ &= \mathbb{E}[\|\nabla f(x) - \nabla f(x^\star) - (\nabla F(x^\star; S) - \nabla F(x; S))\|_2^2] \\ &\le 2\|\nabla f(x) - \nabla f(x^\star)\|_2^2 \\ &\qquad + 2\mathbb{E}[\|\nabla F(x^\star; S) - \nabla F(x; S)\|_2^2] \\ &\le 4\mathbb{E}[L(S)^2]\|x - x^\star\|_2^2.\end{aligned}$$

We may thus take $\sigma_2^2 \lesssim \mathbb{E}[L(S)^2]$. $\diamond$

We present a linear convergence result (Proposition 1) for model-based methods with constant and decaying stepsize choices. This is a cleaner, generalized, and unified version of Proposition 1 and Theorem 2 from (Asi et al., 2020) that uses a more standard and simpler definition of quadratic growth. We primarily present these results for the completeness of the story and to contrast them with results later in the paper.

Proposition 1. *Assume problem (1) is an interpolation problem (Definition 4.1) and let f have L-Lipschitz gradient and satisfy Assumptions 2 and 3 , where $L \geq \lambda$. Let x_k follow the model-based iteration (6) with any model $\overline{F}_x(y; S^{1:m})$ satisfying conditions (C.i) and (C.ii) with minibatch size m. Then*

(i) Let $\alpha_k = \frac{1}{L+\eta_k}$ for $\eta_k \geq 0$. Then

$$\mathbb{E}[\mathrm{dist}(x_k, \mathcal{X}^\star)^2] \leq \exp\left(-\frac{1}{2}\sum_{i=1}^{k} \lambda\alpha_k + \sum_{i=1}^{k} \frac{\sigma_2^2}{m}\frac{\alpha_i}{\eta_i}\right) \mathrm{dist}(x_0, \mathcal{X}^\star)^2.$$

(ii) With the constant stepsize choice $\alpha_k = (L+\eta)^{-1}$ and $\eta = \max\{L, \frac{8\sigma_2^2}{m\lambda}\}$,

$$\mathbb{E}[\mathrm{dist}(x_k, \mathcal{X}^\star)^2] \leq \exp\left(-k \min\left\{\frac{\lambda}{8L}, \frac{m\lambda^2}{64\sigma_2^2}\right\}\right) \mathbb{E}[\mathrm{dist}(x_0, \mathcal{X}^\star)^2].$$

The results in Proposition 1 imply that when the batch size is large enough that $m \gtrsim \sigma_2^2/(\lambda L)$, we achieve convergence rate $\mathbb{E}[\mathrm{dist}(x_k, \mathcal{X}^\star)] \lesssim (1 - c\frac{\lambda}{L})^k \mathbb{E}[\mathrm{dist}(x_0, \mathcal{X}^\star)]$, where $c > 0$ is a numerical constant, which is the rate of convergence for (deterministic) gradient methods with optimal stepsize choices (Nesterov, 2004). More generally, we see a roughly linear speedup in the batch size m to achieve a given accuracy until $m \geq \frac{\sigma_2^2}{\lambda L}$: to obtain $\mathbb{E}[\mathrm{dist}(x_k, \mathcal{X}^\star)^2] \leq \epsilon$ takes

$$k = O(1) \max\left\{\frac{L}{\lambda}, \frac{\sigma_2^2}{\lambda^2 m}\right\} \log\frac{1}{\epsilon}$$

iterations with appropriately chosen stepsize α. That is, we expect to see a linear improvement in the number of iterations to achieve a given accuracy ϵ until the condition number $\frac{L}{\lambda}$ dominates the variance of the gradient estimates.

4.2. Upper bounds under an expected growth condition

In Proposition 1 above, we restrict the stepsizes to have the form $\alpha_k = \frac{1}{L+\eta_k}$. With more accurate models and an alternative growth assumption on the functions F and f, we can remove this weakness, highlighting the robustness of more accurate models. To that end, we revisit a few results of Asi and Duchi (Asi & Duchi, 2019b), beginning with a slight generalization of their growth assumption (which corresponds to the choices $\gamma \in \{0, 1\}$ below):

Assumption 4 (γ-Growth)**.** *There exist constants $\lambda_0, \lambda_1 > 0$ and $\gamma \in [0, 1]$, such that for all $\alpha \in \mathbb{R}_+, x \in \mathcal{X}, x^\star \in \mathcal{X}^\star$, we have*

$$\mathbb{E}\left[(F(x; S) - F(x^\star; S)) \min\left\{\alpha, \frac{F(x; S) - F(x^\star, S)}{\|F'(x; S)\|_2^2}\right\}\right] \geq \min\{\lambda_0\alpha, \lambda_1 \mathrm{dist}(x, \mathcal{X}^\star)^{1-\gamma}\} \, \mathrm{dist}(x, \mathcal{X}^\star)^{1+\gamma}.$$

As we will show in the coming section, while Assumption 4 looks like a technical assumption, it actually fairly closely governs the complexity of solving interpolation problems, in that the λ_1 parameter describes lower bounds on the convergence of any method. Essentially, the assumption states that the functions F must grow relative to the magnitude of their gradients at a particular rate, so that it provides a type of stochastic growth condition. We shall revisit this in the next section when we prove our lower bounds, for now focusing on algorithms and their convergence under the assumption. First, however, we may again rely on linear regression-type objectives for an example satisfying Assumption 4.

Example 3: Consider a problem with data $s = (a, b) \in \mathbb{R}^n \times \mathbb{R}$, where $b = \langle a, x^\star \rangle$ for all (a, b), and set $F(x; (a, b)) = \frac{1}{1+\gamma} |\langle a, x - x^\star \rangle|^{1+\gamma}$, so $\|F'(x; (a, b))\|_2^2 = \|a\|_2^2 \, |\langle a, x - x^\star \rangle|^{2\gamma}$. If $a \sim \mathsf{N}(0, I_n)$, then $|\langle a, x - x^\star \rangle| \geq \frac{1}{2}\|x - x^\star\|_2$ with probability at least $\frac{3}{5}$, and similarly $\|a\|_2^2 \leq 2n$ with probability at least $\frac{3}{5}$, so that both occur with probability at least $\frac{1}{5}$. We then obtain

$$\mathbb{E}\left[F(x; S) \min\left\{\alpha, \frac{F(x; S)}{\|F'(x; S)\|_2^2}\right\}\right] \geq \frac{1}{5} \frac{\|x - x^\star\|_2^{1+\gamma}}{2^{1+\gamma}(1+\gamma)} \min\left\{\alpha, \frac{\|x - x^\star\|_2^{1-\gamma}}{2^{1-\gamma}(1+\gamma) \cdot 2n}\right\},$$

so that Assumption 4 holds with $\lambda_0 \geq \frac{1}{5(1+\gamma)2^{1+\gamma}}$ and $\lambda_1 \geq \frac{1}{2^{2-\gamma}(1+\gamma)n}$. ◇

To give stronger convergence results under Assumption 4, we require one additional condition on our models, which Asi and Duchi (Asi & Duchi, 2019b) introduce:

(C.iii) For all $s \in \mathcal{S}$, the models $F_x(\cdot; s)$ satisfy

$$F_x(y; s) \geq \inf_{z \in \mathcal{X}} F(z; s).$$

In minibatch settings, where one considers a batch $S^{1:m}$ of samples in each model, the condition (C.iii) can be somewhat challenging to verify, as it requires accuracy for the average $\inf_z \overline{F}(z; s^{1:m})$, though (obviously) proximal methods (4) satisfy this condition, and in typical situations (e.g. linear regression) where the batch size $m \leq n$, the average of truncated models (8) will be similarly accurate.

Theorem 2. *Let Assumption 4 hold, and let x_k be generated by the stochastic iteration (2) for a model satisfying conditions (C.i)–(C.iii). Take stepsizes $\alpha_k = \alpha_0 k^{-\beta}$ for some $\beta \in [0,1]$. Define $K_0 := \lfloor (\lambda_0 \alpha_0 / (\lambda_1 \operatorname{dist}(x_1, \mathcal{X}^\star)^{1-\gamma}))^{1/\beta} \rfloor$. Then*

$$\mathbb{E}[\operatorname{dist}(x_{k+1}, \mathcal{X}^\star)^2] \le \exp\Bigg(-\lambda_1 \min\{k, K_0\} - \frac{\lambda_0}{\operatorname{dist}(x_1, \mathcal{X}^\star)^{1-\gamma}} \sum_{i=K_0+1}^{k} \alpha_i\Bigg) \operatorname{dist}(x_1, \mathcal{X}^\star)^2.$$

In the best case—when the stepsizes $\alpha_k \uparrow \infty$ in Theorem 2—we achieve convergence scaling as $\mathbb{E}[\operatorname{dist}(x_k, \mathcal{X}^\star)^2] \lesssim \exp(-\lambda_1 k) \operatorname{dist}(x_1, \mathcal{X}^\star)^2$, and moreover (as we show in the next section) this dependence on the growth constant λ_1 is unimprovable. With this as motivation, one might hope that increased sampling (minibatching) might increase the growth constant λ_1 in Assumption 4; here we provide a sketch of such a result, which also makes it somewhat easier to check the conditions of Assumption 4, by giving three growth conditions.

(G.i) There exists $\mu > 0$ and a probability $p > 0$ such that for all $x \in \mathcal{X}$, we have

$$\mathbb{P}(F(x;S) - F(x^\star;S) \ge \mu \operatorname{dist}(x, \mathcal{X}^\star)^{1+\gamma}) \ge p.$$

(G.ii) The (sub)gradient f' is (L,γ)-Holder continuous, meaning $\|f'(x) - f'(y)\|_2 \le L \|x-y\|_2^\gamma$, and $0 \in \partial f(x^\star)$.

(G.iii) There exists ρ such that

$$\rho \ge \sup_{g \text{ measurable}} \left\{ \frac{\operatorname{Var}(g(x;S))}{\|f'(x)\|_2^2} \;\middle|\; \begin{array}{l} g(x;s) \in \partial F(x;s), \\ f'(x) = \mathbb{E}[g(x;S)] \end{array} \right\}$$

for all $x \in \mathcal{X}$.

Our typical situation is to think of μ and p numerical constants, where the scaling ρ measures the noise inherent to the problem. In any case, a short calculation shows how Conditions (G.i)–(G.iii) suffice to give Assumption 4.

Lemma 4.1. *Let conditions (G.i)–(G.iii) hold. Then the average $\overline{F}(x; s^{1:m}) = \frac{1}{m}\sum_{i=1}^n F(x; s^i)$ satisfies the γ-growth condition of Assumption 4 with*

$$\lambda_0 = \frac{\lfloor mp \rfloor}{4m}\mu \text{ and } \lambda_1 = \frac{(\lfloor mp \rfloor / m)^2 \mu^2}{16 L^2 (1 + \frac{\rho}{m})}.$$

In brief, we see that mini-batches of size m suggest improved convergence related to the noise-to-signal ratio $\rho := \sup_x \frac{\operatorname{Var}(F'(x;S))}{\|f'(x)\|_2^2}$: once the sample size m is large enough that $\rho/m \lesssim 1$, we expect relatively little improvement, though we *do* see a linear improvement in the growth constant λ_1 as m grows whenever $m \ll \rho$. To see this, let us for simplicity assume that in Conditions (G.i)–(G.iii) we have $p \gtrsim 1$ and $L/\mu \lesssim 1$ (that is, the problem is well-conditioned). Then applying Theorem 2, we see that for large enough stepsizes α,

$$k = O(1)\left(1 + \frac{\rho}{m}\right) \log \frac{1}{\epsilon} \tag{10}$$

iterations of any model-based method (2) with mini-batches of size m—assuming that Conditions (C.i)–(C.iii) hold for the models $\overline{F}_x$—are sufficient to guarantee $\mathbb{E}[\operatorname{dist}(x_k, \mathcal{X}^\star)^2] \le \epsilon$.

5. Optimality in Interpolation Problems

We conclude the theoretical portion of this paper by developing several new optimality results for interpolation problems, that is, those satisfying Definition 4.1. In brief, we shall show that the dependence of Theorem 2 on the growth constant λ_1 is sharp and unimprovable, and that in some cases, the dependence on the signal-to-noise ratio $\rho^{-1} := \inf_x \frac{\|f'(x)\|_2^2}{\operatorname{Var}(F'(x;S))}$ is essentially sharp as well. We do so via information-theoretic lower bounds on estimation of optimal points, the first in a stylized $n = 1$ dimensional problem that gives the correct dependence on the growth constants in Assumption 4, the second in standard regression problems but where we choose the dimension $n \in \mathbb{N}$ more carefully.

We define our *minimax risk* as follows. Let $\mathcal{P}$ be a family of problems, where a problem is a pair (F, P) consisting of a probability distribution P supported on $\mathcal{S}$ and function F as defined in the introduction. We let $\mathcal{X}^\star(F,P) = \operatorname{argmin}_{x \in \mathcal{X}} \mathbb{E}_P[F(x;S)]$ be the collection of minimizers, and define the minimax squared error

$$\mathfrak{M}_k(\mathcal{P}, \mathcal{X}) := \inf_{\widehat{x}^k} \sup_{(F,P) \in \mathcal{P}} \mathbb{E}_{P^k}\left[\operatorname{dist}(\widehat{x}^k, \mathcal{X}^\star(F,P))^2\right], \tag{11}$$

where the infimum is over all measurable $\widehat{x}^k : \mathcal{S}^k \to \mathbb{R}^n$, the supremum is over problems $(F,P) \in \mathcal{P}$, and the inner expectation is over the samples $S_1, \ldots, S_k \overset{\text{iid}}{\sim} P$.

5.1. A lower bound for one-dimensional problems

We first focus on problems for which we can isolate the contributions of the growth constant λ_1 in Assumption 4, letting the dimension $n = 1$ to show that our complexity bounds hold independent of dimension; higher dimensions can only yield increased complexity. We consider a collection of well-conditioned problems, where we analogize the typical condition number of f by defining

$$\lambda_\gamma(f) := \inf_{x \notin \mathcal{X}^\star} \frac{f(x) - f(x^\star)}{\frac{1}{1+\gamma}\operatorname{dist}(x, \mathcal{X}^\star)^{1+\gamma}} \quad \text{and}$$
$$L_\gamma(f) := \sup_{x \neq y} \frac{|f'(x) - f'(y)|}{|x - y|^\gamma},$$

calling $\kappa_\gamma(f) := \frac{L}{\lambda}$ the condition number. We also note in passing that the constant $\lambda_1 \le 1$ in Assumption 4, as by convexity we have

$$\frac{(F(x;s) - F(x^\star;s))^2}{F'(x;s)^2} \le \frac{\langle F'(x;s), x - x^\star\rangle^2}{F'(x;s)^2} \le |x - x^\star|^2,$$

so taking $\alpha \uparrow \infty$ in Assumption 4 guarantees $\lambda_1 \in [0,1]$. Thus, for our first collection of problems, we let $\mathcal{P}_\gamma(\lambda_1)$ be those problems satisfying Assumption 4 with a given $\gamma, \lambda_1 \in [0,1]$, any $\lambda_0 \ge \lambda_1$, our standing assumption of the interpolation condition in Definition 4.1, and condition number $\kappa_\gamma(f) = 1$. The choice of the condition number serves to highlight the difficulties from stochasticity in the problem, eliminating the contributions of hardness from the population (deterministic) objective f; an identical lower bound will of course hold in the coming theorem for more poorly conditioned problems with $\kappa_\gamma(f) \ge 1$, as this is simply a larger collection.

Theorem 3. *Let $\mathcal{P}_\gamma(\lambda_1)$ be the collection $\mathcal{P}_\gamma(\lambda_1)$, assume that $\mathcal{X}$ contains an ℓ_2-ball of radius $R \ge 0$. Then*

$$\mathfrak{M}_k(\mathcal{P}_\gamma(\lambda_1), \mathcal{X}) \ge \frac{R^2}{2}\left[1 - (1+\gamma)^2\lambda_1\right]_+^k.$$

We defer the proof to the appendix and make a few remarks here. First, the convergence guarantees in Section 4.2 show that appropriate model-based methods converge to ϵ accuracy in $O(\frac{1}{\lambda_1}\log\frac{1}{\epsilon})$ iterations, which by the theorem is optimal. Thus, in a strong sense, the *a priori* esoteric-seeming growth condition in Assumption 4 is indeed fundamental.

5.2. A lower bound for well-conditioned regression problems

The proof of Theorem 3 relies on constructing certain power functions and a very careful choice of growth and probability. An alternative approach is to mimic those ideas in proving complexity results for deterministic problems (Nemirovski & Yudin, 1983; Nesterov, 2004; Carmon et al., 2019), where one takes the dimension larger. By allowing high-dimensional problems, we can show that the noise-to-signal ratio $\rho := \sup_x \frac{\operatorname{Var}(F'(x;S))}{\|\nabla f(x)\|^2}$ and growth constant λ_1 from Assumption 4 remain fundamental, even in noiseless linear regression.

To make the proof cleaner we make a slight modification to the class of problems we consider: instead of assuming a bounded domain $\mathcal{X}$, we instead assume $\mathcal{X} = \mathbb{R}^n$, but now we consider a randomized (instead of minimax/worst case) adversary that chooses a problem $(F, P) \in \mathcal{P}$ according to a measure π on the space of problems; in particular, we assume that $\mathbb{E}_\mu[\|x_0 - x^\star\|_2^2] \le R^2$, that is, the expected distance of x_0 to $x^\star$ is at most R. Letting $\mathcal{X}^\star(F,P) = \operatorname{argmin}_x \mathbb{E}_P[F(x;S)]$ be the optimal set for a given problem (F,P), we define the minimum average risk

$$\mathfrak{M}_k(\mathcal{P}, \pi) := \inf_{\widehat{x}^k} \int \mathbb{E}_{P^k}[\operatorname{dist}(\widehat{x}^k, \mathcal{X}^\star(F,P))^2] d\pi(F,P).$$

We note that the minimum average risk defined here naturally lower bounds the minimax risk (11), redefined analogously for our problem.

We specialize this randomized risk for each $n \in \mathbb{N}$, letting $\mathcal{P}_n$ be a collection of noiseless linear regression problems on $\mathbb{R}^n$, where we identify the prior measure π with $x^\star \sim \mathsf{N}(0, \frac{R^2}{n} I_{n\times n})$. Then certainly $\mathbb{E}[\|x^\star\|_2^2] = R^2$. We consider samples s consisting of a pair $A \in \mathbb{R}^{m\times n}$ and $b = Ax^\star$, considering the quadratic loss

$$F(x;s) = F(x;(A,b)) = \frac{1}{2}\|Ax - b\|_2^2, \qquad (12)$$

and we call the resulting objective $f(x) = \mathbb{E}[F(x;S)]$ *perfectly conditioned* if $f(x) = c\|x - x^\star\|_2^2$ for a constant $c \in \mathbb{R}_+$. We have the following theorem.

Theorem 4. *Let $\lambda_1 \in [0, \frac{1}{4}]$ and $\gamma = 1$. Then there exists a collection $\mathcal{P}$ of perfectly conditioned interpolating problems with squared error (12), satisfying Assumption 4 and $\mathbb{E}_\pi[\|x^\star\|_2^2] = R^2$, such that*

$$\mathfrak{M}_k(\mathcal{P}, \pi) \ge R^2(1 - 4\lambda_1)^k.$$

Alternatively, let $\rho \in [1, \infty]$. There exists a collection $\mathcal{P}$ of perfectly conditioned interpolating problems with squared error (12), with noise-to-signal ratio satisfying $\sup_x \frac{\operatorname{Var}(\nabla F(x;S))}{\|\nabla f(x)\|_2^2} \le \rho$, such that

$$\mathfrak{M}_k(\mathcal{P}, \pi) \ge R^2\left(1 - \frac{1}{\rho}\right)^k.$$

Thus, one cannot hope to achieve (much) better convergence even for quadratics than that we have outlined: the dependence on either the growth λ_1 or the signal-to-noise ρ^{-1} is unavoidable, and one must collect at least $k \gtrsim \frac{1}{\lambda_1}\log\frac{1}{\epsilon}$ or $k \gtrsim \rho\log\frac{1}{\epsilon}$ samples S to achieve accuracy ϵ, again highlighting that these quantities—as we (inspired by (Asi & Duchi, 2019b)) identify in Theorem 2 and the iteration bound (10)—are fundamental for interpolation problems.

6. Experiments

We now study and demonstrate the speedup and robustness of aProx methods with minibatches, comparing the relative performance of the proposed methods on several stochastic optimization problems. We consider the following methods in our experiments, where we use both single sample ($m = 1$) and minibatch ($m > 1$) versions: (SGM(3), Proximal(4), IA (averaging individual APROX iterates from (Asi et al., 2020)), TruncAv (7), AvMod (8))

We use stepsizes $\alpha_k = \alpha_0 k^{-1/2}$, varying α_0, and for each algorithm a report the number $T_{\mathsf{a},m}(\alpha_0)$ of total samples used to reach ε accuracy using minibatches of size m; that is, $T_{\mathsf{a},m}(\alpha_0) = km$ where k is the first iteration to satisfy $f(x_k) - f(x^\star) \le \varepsilon$. We also let $T^\star_{\mathsf{a},m} = \min_{\alpha_0} T_{\mathsf{a},m}(\alpha_0)$ denote the fewest iterations to convergence for a method a using batch size m. Each of our experiments involves data $(A, b) \in \mathbb{R}^{N \times n} \times \mathbb{R}^N$, where $f_{A,b}(x) = \frac{1}{N} \sum_{i=1}^N F(x; a_i, b_i)$ for a given loss F, and we vary the condition number of A, taking $N = 10^3$ and $n = 40$. We present three types of results:

1. **Best speedups for minibatching:** For each method a, we plot $\frac{T^\star_{\mathsf{a},1}}{T^\star_{\mathsf{a},m}}$ against the minibatch size m to show the speedup minibatching provides using the best step sizes.
2. **Performance profiles** (Dolan & Moré, 2002)**:** For each method a, we evaluate for each $r \ge 1$ the fraction of the total executed experiments for which the $T_{\mathsf{a},m}(\alpha_0) \le r T_{\mathsf{a}^\star,m}(\alpha_0)$, where $\mathsf{a}^\star$ is the best performing method in each experiment, giving r on the horizontal axis and the proportion on the vertical. Here, to evaluate robustness, we define a single experiment as one execution of each of the 5 methods for a particular step size α_0, minibatch size m, and condition number combination. We discard the experiments where more than 3 of the methods fail to complete before the max number of iterations.
3. **Iterations to convergence w.r.t. stepsize:** For each method a and minibatch size m, we plot $T_{\mathsf{a},m}(\alpha_0)$ against the initial step size α_0.

We use minibatch sizes $m \in \{1, 4, 8, 16, 32, 64\}$ and initial steps $\alpha_0 \in \{10^{i/2}, i \in \{-4, -3, \dots, 5\}\}$. For all experiments we run 30 trials with different seeds and plot the 95% confidence sets. We use and extend the code provided by (Asi et al., 2020). We describe the objective function and noise mechanism for each problem below.

6.1. Linear Regression

We have $f(x) = \frac{1}{2N} \|Ax - b\|_2^2$. We generate rows of A and $x^\star$ i.i.d. $\mathsf{N}(0, I_n)$ and, setting $b = Ax^\star + \sigma v$ with $v \sim \mathsf{N}(0, I_N)$. In the noisy setting for our experiments, we set $\sigma = 0.5$. In Figure 1, we plot the minibatch speedups of the accelerated and non-accelerated methods; acceleration gives a $\sim 16\times$ improvement for AvMod and TruncAv. Figure 2 outlines the performance profiles for the linear regression experiments. The fully proximal, AvMod, and TruncAv methods are noticeably better than IA and SGM.

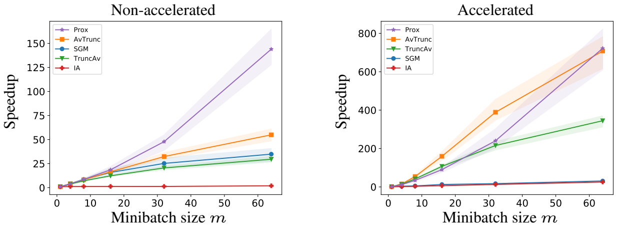

Figure 1: Speed ups vs. batch size for noiseless absolute regression. Note the difference in scales.

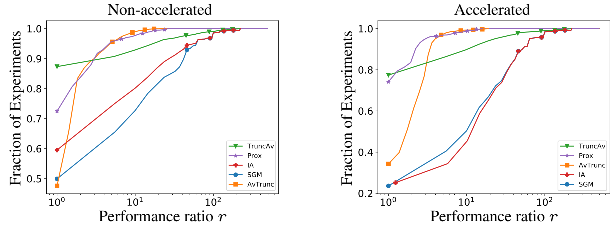

Figure 2: Performance profiles for linear regression.

6.2. Absolute loss regression

We have $f(x) = \frac{1}{2N} \|Ax - b\|_1$. Again we generate rows of A and $x^\star$ i.i.d. $\mathsf{N}(0, I_n)$, setting $b = Ax^\star + \sigma v$ and drawing $v \sim \mathsf{Lap}(1)^N$. In the noisy setting for our experiments, we set $\sigma = 0.5$. In Figure 3, we plot the speedup up of each algorithm (relative to minibatch size $m = 1$) against minibatch size in the noiseless setting. We observe that the speedups in the acclerated setting are more pronounced than in the non-accelerated setting. We provide performance profiles for the non-accelerated and accelerated algorithms in Figure 4. Similar to the linear regression setting, we see that AvMod, TruncAv, and full-prox, outperform IA and SGM.

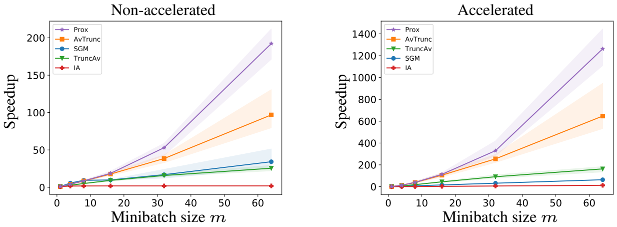

Figure 3: Speed ups vs. batch size for noiseless absolute regression. Note the difference in scales.

6.3. Logistic Regression

We have $f(x) = \frac{1}{2N} \sum_{i=1}^N \log(1 + \exp(-b_i \langle a_i, x \rangle))$. We generate rows of A and $x^\star$ i.i.d. $\mathsf{N}(0, I_n)$, setting $b_i = \text{sign}(\langle a_i, x^\star \rangle)$. To add noise, we flip each label b_i independently with probability $p = .01$. In Figure 5, we plot

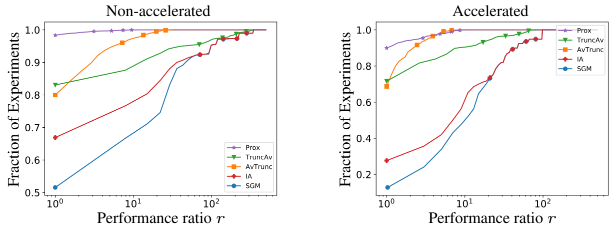

Figure 4: Performance profiles for absolute regression.

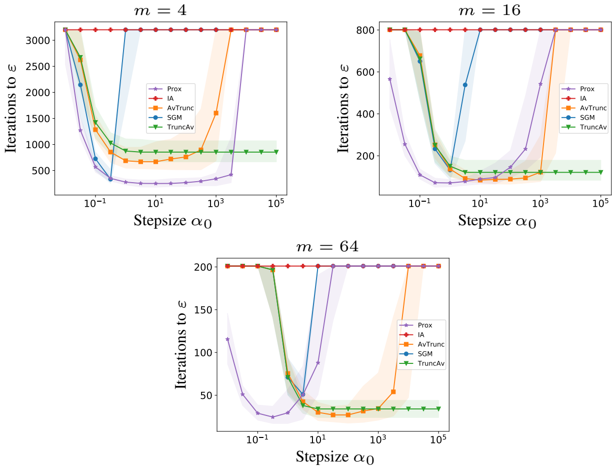

Figure 5: Iterations to convergence of the accelerated methods vs. initial stepsizes for logistic regression

the iterations to convergence against initial step size for accelerated methods in this setting. We observe that even accelerated versions of model-based methods exhibit the attractive property of robustness to problem parameters (initial step size choice) like their non-accelerated counterparts.

References

Asi, H. and Duchi, J. C. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019a. URL https://doi.org/10.1073/pnas.1908018116.

Asi, H. and Duchi, J. C. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019b. URL https://arXiv.org/abs/1810.05633.

Asi, H., Chadha, K., Cheng, G., and Duchi, J. C. Minibatch stochastic approximate proximal point methods. In *Advances in Neural Information Processing Systems 33*, 2020.

Bach, F. and Moulines, E. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, pp. 451–459, 2011.

Bauschke, H. and Borwein, J. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3): 367–426, 1996.

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

Belkin, M., Hsu, D., and Mitra, P. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems 31*, pp. 2300–2311. Curran Associates, Inc., 2018.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1611–1619, 2019.

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

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

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

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points I. *Mathematical Programming, Series A*, 2019. doi: s10107-019-01406-y.

Chaturapruek, S., Duchi, J. C., and Ré, C. Asynchronous stochastic convex optimization: the noise is in the noise and SGD don't care. In *Advances in Neural Information Processing Systems 28*, 2015.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

Dolan, E. D. and Moré, J. J. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.

Duchi, J. C. and Ruan, F. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

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

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

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

Kushner, H. J. and Yin, G. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, second edition, 2003.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming, Series A*, 133 (1–2):365–397, 2012.

Levy, D. and Duchi, J. C. Necessary and sufficient geometries for gradient methods. In *Advances in Neural Information Processing Systems 32*, 2019.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2017.

Lin, H., Mairal, J., and Harchaoui, Z. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18(212), 2018.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Needell, D., Ward, R., and Srebro, N. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Advances in Neural Information Processing Systems 27*, pp. 1017–1025, 2014.

Nemirovski, A. and Yudin, D. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

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

Nesterov, Y. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

Niu, F., Recht, B., Ré, C., and Wright, S. Hogwild: a lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, 2011.

Polyak, B. T. *Introduction to Optimization*. Optimization Software, Inc., 1987.

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

Robbins, H. and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

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

Strohmer, T. and Vershynin, R. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

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

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

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

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

A. Proofs of non-asymptotic upper bounds

We collect our proof of Theorem 1 in this section. It relies on a standard claim on minimizers of sums of convex functions, which we state and prove here for convenience.

Claim A.1. *Let u and ψ be convex, ψ be differentiable on $\mathcal{X}$, and $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$. If x^+ minimizes $u(x) + \psi(x)$ over $x \in \mathcal{X}$, then*

$$u(x^+) + \psi(x^+) \le u(x) + \psi(x) - D_\psi(x, x^+) \text{ for all } x \in \mathcal{X}.$$

Proof By convexity and the optimality of x^+, there exists $u'(x^+) \in \partial u(x^+)$ such that $\langle u'(x^+) + \nabla\psi(x^+), x - x^+\rangle \ge 0$ for all $x \in \mathcal{X}$. Using the standard first-order convexity inequality, we thus obtain

$$\begin{aligned} u(x) &\ge u(x^+) + \langle u'(x^+), x - x^+\rangle \\ &= u(x^+) + \langle u'(x^+) + \nabla\psi(x^+), x - x^+\rangle - \langle \nabla\psi(x^+), x - x^+\rangle \\ &\ge u(x^+) - \langle \nabla\psi(x^+), x - x^+\rangle \\ &= u(x^+) + \psi(x^+) - \psi(x) + D_\psi(x, x^+), \end{aligned}$$

as desired. □

A.1. Proof of Theorem 1

In this proof, we begin with a deterministic one-step progress bound and then iterate the bound. In analogy to Lemma B.1, we rely on the conditionally mean-zero function and gradient errors

$$e_k := F(x^\star; S_k) - f(x^\star) + f(y_k) - F(y_k; S_k) \text{ and } \xi_k := \nabla f(y_k) - \nabla F(y_k; S_k).$$

We have the one-step progress bound

Lemma A.1. *Let $\alpha_k \le \frac{1}{L\theta_k + \eta_k}$ and $\Delta_k = f(x_k) + r(x_k) - f(x^\star) - r(x^\star)$. Then*

$$\begin{aligned} &\Delta_{k+1} \\ &\le (1 - \theta_k)\Delta_k + \theta_k \left[e_k + \langle \xi_k, z_k - y_k\rangle + \frac{\|\xi_k\|_*^2}{2\eta_k} + \frac{1}{\alpha_k}\left(D_h(x^\star, z_k) - D_h(x^\star, z_{k+1})\right)\right]. \end{aligned}$$

Proof We follow the proof of Tseng, Proposition 1 (Tseng, 2008). For shorthand, let

$$\mathsf{lin}_f(x, y) := f(y) + \langle \nabla f(y), x - y\rangle + r(x),$$

which linearly approximates f and does not approximate the additive component r. Then by the L-smoothness of ∇f, we obtain

$$\begin{aligned} f(x_{k+1}) + r(x_{k+1}) &\le \mathsf{lin}_f(x_{k+1}, y_k) + \frac{L}{2}\|x_{k+1} - y_k\|^2 \\ &= \mathsf{lin}_f((1 - \theta_k)x_k + \theta_k z_{k+1}, y_k) + \frac{L\theta_k^2}{2}\|z_{k+1} - z_k\|^2 \\ &\overset{(i)}{\le} (1 - \theta_k)\mathsf{lin}_f(x_k, y_k) + \theta_k \mathsf{lin}_f(z_{k+1}, y_k) + \frac{L\theta_k^2}{2}\|z_{k+1} - z_k\|^2 \\ &\overset{(ii)}{\le} (1 - \theta_k)(f(x_k) + r(x_k)) + \theta_k\left[\mathsf{lin}_f(z_{k+1}, y_k) + \frac{L\theta_k}{2}\|z_{k+1} - z_k\|^2\right], \end{aligned} \tag{13}$$

where the inequality (i) used that r is convex and (ii) that f is convex.

We consider the final two terms in the bound (13), and with function and gradient errors $e_k^{(1)} := f(y_k) - F(y_k; S_k)$ and $\xi_k := \nabla f(y_k) - \nabla F(y_k; S_k)$, we expand the first in terms of the random samples to write

$$\begin{aligned} \mathsf{lin}_f(z_{k+1}, y_k) &= F(y_k; S_k) + \langle \nabla F(y_k; S_k), z_{k+1} - y_k\rangle + r(z_{k+1}) + e_k^{(1)} + \langle \xi_k, z_{k+1} - y_k\rangle \\ &\le F_{y_k}(z_{k+1}; S_k) + r(z_{k+1}) + e_k^{(1)} + \langle \xi_k, z_{k+1} - y_k\rangle, \end{aligned} \tag{14}$$

where the inequality uses that the models F_{y_k} necessarily upper bound the first-order (linear) approximation to F at y_k (recall the discussion following Condition (C.ii)). To control term (14), we apply Claim A.1 with $u(x) = F_{y_k}(x; S_k)$, $\psi(x) = D_h(x, z_k)$, and $x^+ = z_{k+1}$, so that inequality (14) implies

$$\begin{aligned}\mathsf{lin}_f(z_{k+1}, y_k) \le F_{y_k}(x; S_k) + r(x) + \frac{1}{\alpha_k}\left[D_h(x, z_k) - D_h(x, z_{k+1}) - D_h(z_{k+1}, z_k)\right] \\ + e_k^{(1)} + \langle \xi_k, z_{k+1} - y_k\rangle\end{aligned}$$

for any $x \in \mathcal{X}$. Rearranging terms and using the Fenchel-Young inequality to see that

$$\langle \xi_k, z_{k+1} - y_k\rangle = \langle \xi_k, z_k - y_k\rangle + \langle \xi_k, z_{k+1} - z_k\rangle \le \langle \xi_k, z_k - y_k\rangle + \frac{\|\xi_k\|_*^2}{2\eta_k} + \frac{\eta_k}{2}\|z_{k+1} - z_k\|^2$$

and using the strong convexity bound $D_h(z_{k+1}, z_k) \ge \frac{1}{2}\|z_{k+1} - z_k\|^2$ then implies

$$\begin{aligned}\mathsf{lin}_f(z_{k+1}, y_k) \le F_{y_k}(x; S_k) + r(x) + \frac{1}{\alpha_k}\left[D_h(x, z_k) - D_h(x, z_{k+1})\right] + e_k^{(1)} \\ + \langle \xi_k, z_k - y_k\rangle + \frac{\|\xi_k\|_*^2}{2\eta_k} + \frac{\eta_k}{2}\|z_{k+1} - z_k\|^2 - \frac{1}{2\alpha_k}\|z_{k+1} - z_k\|^2.\end{aligned}$$

Our modeling assumptions guarantee that $F_{y_k}(x; S_k) \le F(x; S_k)$, so writing the function error $e_k = F(x; S_k) - f(x) + f(y_k) - F(y_k; S_k)$ and substituting this upper bound on $\mathsf{lin}_f(z_{k+1}, y_k)$ into the bound (13) gives the single-step progress guarantee

$$\begin{aligned}f(x_{k+1}) + r(x_{k+1}) \le (1 - \theta_k)(f(x_k) + r(x_k)) + \theta_k(f(x) + r(x)) \\ + \theta_k\Bigg[e_k + \langle \xi_k, z_k - y_k\rangle + \frac{\|\xi_k\|_*^2}{2\eta_k} + \frac{1}{\alpha_k}\left[D_h(x, z_k) - D_h(x, z_{k+1})\right] \\ + \frac{L\theta_k + \eta_k}{2}\|z_{k+1} - z_k\|^2 - \frac{1}{2\alpha_k}\|z_{k+1} - z_k\|^2\Bigg].\end{aligned}$$

Any stepsize $\alpha_k \le \frac{1}{L\theta_k + \eta_k}$ cancels the the $\|z_{k+1} - z_k\|^2$ terms, and setting $x = x^\star$ gives the lemma. □

Iterating Lemma A.1 with $\Delta_k = f(x_k) + r(x_k) - f(x^\star) - r(x^\star) \ge 0$ yields the following deterministic convergence guarantee.

Lemma A.2. *Let the conditions of Theorem 1 hold. Define the error terms* $\zeta_k \coloneqq e_k + \langle \xi_k, z_k - y_k\rangle + \frac{\|\xi_k\|_*^2 - \sigma_0^2}{2\alpha_k}$. *Then*

$$\frac{1}{\theta_k^2}\left[f(x_{k+1}) + r(x_{k+1}) - f(x^\star) - r(x^\star)\right] \le \sum_{i=0}^{k}\frac{\sigma_0^2}{2\theta_i\eta_i} + \left(L + \frac{\eta_k}{\theta_k}\right)R^2 + \sum_{i=0}^{k}\frac{1}{\theta_i}\zeta_i.$$

Proof Lemma A.1 yields

$$\begin{aligned}\frac{1}{\theta_k^2}\Delta_{k+1} &\le \frac{1 - \theta_k}{\theta_k^2}\Delta_k + \frac{1}{\theta_k\alpha_k}\left[D_h(x^\star, z_k) - D_h(x^\star, z_{k+1})\right] + \frac{\sigma_0^2}{2\eta_k\theta_k} \\ &\quad + \frac{1}{\theta_k}\Bigg[\underbrace{e_k + \langle \xi_k, z_k - y_k\rangle + \frac{\|\xi_k\|_*^2 - \sigma_0^2}{2\eta_k}}_{=:\zeta_k}\Bigg] \\ &\le \frac{1}{\theta_{k-1}^2}\Delta_k + \frac{1}{\theta_k\alpha_k}\left[D_h(x^\star, z_k) - D_h(x^\star, z_{k+1})\right] + \frac{\sigma_0^2}{2\eta_k\theta_k} + \frac{1}{\theta_k}\zeta_k.\end{aligned}$$

where we recalled that $(1 - \theta_k)/\theta_k^2 \le 1/\theta_{k-1}^2$. Iterating the inequality and using $\frac{1-\theta_0}{\theta_0^2} = 0$, we find that

$$\frac{1}{\theta_k^2}\Delta_{k+1} \le \sum_{i=0}^{k}\frac{\sigma_0^2}{2\eta_i\theta_i} + \sum_{i=0}^{k}\frac{1}{\theta_i\alpha_i}\left(D_h(x^\star, z_i) - D_h(x^\star, z_{i+1})\right) + \sum_{i=0}^{k}\frac{1}{\theta_i}\zeta_i. \tag{15}$$

We bound the middle summation above as

$$\begin{aligned}\sum_{i=0}^{k}\frac{1}{\theta_i\alpha_i}\left(D_h(x^\star,z_i)-D_h(x^\star,z_{i+1})\right) &\le \sum_{i=2}^{k}\left(\frac{1}{\theta_i\alpha_i}-\frac{1}{\theta_{i-1}\alpha_{i-1}}\right)D_h(x^\star,x_i)-\frac{1}{\theta_{k+1}\alpha_{k+1}}D_h(x^\star,x_{k+1})\\ &\quad+\frac{1}{\theta_1\alpha_1}D_h(x^\star,x_1)\\ &\le \frac{R^2}{\theta_k\alpha_k}\\ &\le LR^2+\frac{\eta_k}{\theta_k}R^2\end{aligned}$$

where the last step uses $D_h(x^\star,y)\le R^2$ and $\frac{1}{\theta_i\alpha_i}=L+\frac{\eta_i}{\theta_i}$. Substituting this in (15) we get the result. □

Now take expectations in Lemma A.2. We have $\mathbb{E}[\zeta_k]\le 0$, and

$$\begin{aligned}\sum_{i=0}^{k}\frac{i+2}{\sqrt{i+1}} &\le \sum_{i=1}^{k+1}\sqrt{i}+\sum_{i=1}^{k+1}\frac{1}{\sqrt{i}}\\ &\le \int_1^{k+2}\sqrt{t}dt+\int_0^{k+1}\frac{1}{\sqrt{t}}dt=\frac{2}{3}((k+2)^{3/2}-1)+2\sqrt{k+1}\overset{(i)}{\le}(k+2)^{3/2},\end{aligned}$$

where inequality (i) holds for $k>2$. Multiplying by $\theta_k^2=4/(k+2)^2$ and using $\eta_k\theta_k=\eta_0\frac{2\sqrt{k}}{k+2}\le 2\eta_0/\sqrt{k}$ gives the deterministic bound

$$\theta_k^2\sum_{i=0}^{k}\frac{\sigma^2}{2\theta_i\eta_i}+\theta_k^2\left(L+\frac{\eta_k}{\theta_k}\right)R^2\le\frac{4LR^2}{(k+2)^2}+\frac{2R^2\eta_0}{\sqrt{k}}+\frac{2\sigma^2}{\eta_0\sqrt{k+2}},$$

as desired.

B. Proofs from Section 4

B.1. Proof of Proposition 1

We assume without loss of generality that $f(x^\star)=0=F(x^\star;s)$ for notational simplicity. We first begin by proving a single step guarantee in the following lemma (Lemma B.1).

Lemma B.1. *Let the conditions of Theorem 1 hold, and define the function value errors* $e_k=[F(x^\star;S_k)-f(x^\star)]-[F(x_k;S_k)-f(x_k)]$. *Then*

$$\begin{aligned}&f(x_{k+1})-f(x^\star)\\ &\le\frac{1}{\alpha_k}\left[D_h(x^\star,x_k)-D_h(x^\star,x_{k+1})\right]+e_k+\frac{1}{2\eta_k}\left\|\nabla F(x_k;S_k)-\nabla f(x_k)\right\|_*^2.\end{aligned}$$

Proof Setting $u(\cdot)=F_{x_k}(\cdot;S_k)$ and $\psi(x)=\frac{1}{\alpha_k}D_h(x,x_k)$ in Claim A.1, and taking $x^+=x_{k+1}$ and $x=x^\star$, we have the progress bound

$$F_{x_k}(x_{k+1};S_k)+\frac{1}{\alpha_k}D_h(x_{k+1},x_k)\le F_{x_k}(x^\star;S_k)+\frac{1}{\alpha_k}\left[D_h(x^\star,x_k)-D_h(x^\star,x_{k+1})\right]. \tag{16}$$

We turn to bounding the difference $F_{x_k}(x^\star;S_k)-F_{x_k}(x_{k+1};S_k)$. Let $g_k=\nabla F(x_k;S_k)$ and define the gradient error $\xi_k:=g_k-\nabla f(x_k)$. Using the convexity of $F_{x_k}(\cdot;S_k)$ and recalling that $g_k\in\partial F_{x_k}(x_k;S_k)$ as in our discussion following Condition (C.ii), we have $F_{x_k}(x_{k+1};S_k)\ge F_{x_k}(x_k;S_k)+\langle g_k,x_{k+1}-x_k\rangle$. As a consequence, we have

$$\begin{aligned}F_{x_k}(x^\star;S_k)-F_{x_k}(x_{k+1};S_k)&\le F_{x_k}(x^\star;S_k)-F(x_k;S_k)+\langle g_k,x_k-x_{k+1}\rangle\\ &=F_{x_k}(x^\star;S_k)-F(x_k;S_k)+\langle\nabla f(x_k),x_k-x_{k+1}\rangle+\langle\xi_k,x_k-x_{k+1}\rangle\\ &\overset{\text{(C.ii)}}{\le}F(x^\star;S_k)-F(x_k;S_k)+\langle\nabla f(x_k),x_k-x_{k+1}\rangle+\langle\xi_k,x_k-x_{k+1}\rangle\\ &=f(x^\star)-f(x_k)+\langle\nabla f(x_k),x_k-x_{k+1}\rangle+e_k+\langle\xi_k,x_k-x_{k+1}\rangle,\end{aligned}$$

where we used the error $e_k = [F(x^\star; S_k) - f(x^\star)] - [F(x_k; S_k) - f(x_k)]$. Finally, the smoothness of f implies $f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \left\| x_k - x_{k+1} \right\|^2$, so

$$\begin{aligned} &F_{x_k}(x^\star; S_k) - F_{x_k}(x_{k+1}; S_k) \\ &\quad \le f(x^\star) - f(x_{k+1}) + \frac{L}{2} \left\| x_k - x_{k+1} \right\|^2 + e_k + \langle \xi_k, x_k - x_{k+1} \rangle. \end{aligned}$$

Substituting this into inequality (16) and rearranging, we obtain

$$\begin{aligned} f(x_{k+1}) - f(x^\star) \le \frac{1}{\alpha_k} \left[D_h(x^\star, x_k) - D_h(x^\star, x_{k+1}) - D_h(x_k, x_{k+1}) \right] \\ + e_k + \langle \xi_k, x_k - x_{k+1} \rangle + \frac{L}{2} \left\| x_k - x_{k+1} \right\|^2. \end{aligned} \tag{17}$$

We apply the Fenchel-Young inequality to control the error $\langle \xi_k, x_k - x_{k+1} \rangle$: we have $\langle \xi_k, x_k - x_{k+1} \rangle \le \frac{1}{2\eta_k} \left\| \xi_k \right\|_*^2 + \frac{\eta_k}{2} \left\| x_k - x_{k+1} \right\|^2$, so

$$\begin{aligned} f(x_{k+1}) - f(x^\star) \le \frac{1}{\alpha_k} \left[D_h(x^\star, x_k) - D_h(x^\star, x_{k+1}) \right] \\ + e_k + \frac{1}{2\eta_k} \left\| \xi_k \right\|_*^2 + \frac{L + \eta_k}{2} \left\| x_k - x_{k+1} \right\|^2 - \frac{1}{\alpha_k} D_h(x_k, x_{k+1}), \end{aligned}$$

which with $\alpha_k = \frac{1}{L + \eta_k}$ gives the lemma once we apply the strong convexity of h, that is, that $D_h(x_k, x_{k+1}) \ge \frac{1}{2} \left\| x_k - x_{k+1} \right\|^2$. □

To complete the proof of Proposition 1, we proceed as follows. Let $D_k = \mathrm{dist}(x_k, \mathcal{X}^\star)$ for shorthand, and recall our notations $e_k = (F(x^\star; S_k) - f(x^\star)) - (F(x_k; S_k) - f(x_k)) = f(x_k) - F(x_k; S_k)$ (in this case) and $\xi_k = \nabla F(x_k; S_k) - \nabla f(x_k)$. Then Lemma B.1 implies

$$\begin{aligned} \frac{1}{2} D_{k+1}^2 &\le \frac{1}{2} D_k^2 - \alpha_k f(x_{k+1}) + \alpha_k e_k + \frac{\alpha_k}{2\eta_k} \left\| \xi_k \right\|_*^2 \\ &\le \frac{1}{2} D_k^2 - \frac{\alpha_k \lambda}{2} D_{k+1}^2 + \alpha e_k + \frac{\alpha_k}{2\eta_k} \left\| \xi_k \right\|_*^2, \end{aligned}$$

where the second inequality follows from Assumption 2 that $f(x_{k+1}) \ge \frac{\lambda}{2} D_{k+1}^2$. Noting that $\mathbb{E}[\left\| \xi_k \right\|_*^2 \mid x_k] \le \frac{\sigma_2^2}{m} D_k^2$ by Assumption 3, we rearrange and take expectations on both sides to obtain

$$\mathbb{E}[D_{k+1}^2] \le \underbrace{\frac{1}{\alpha_k \lambda + 1}}_{\le \exp(-\alpha_k \lambda / 2)} \underbrace{\left(1 + \frac{\alpha_k \sigma_2^2}{\eta m} \right)}_{\le \exp(\frac{\alpha_k \sigma_2^2}{\eta m})} \mathbb{E}[D_k^2] \le \exp\left(\frac{-\lambda \alpha_k}{2} + \frac{\sigma_2^2 \alpha_k}{\eta_k m} \right) \mathbb{E}[D_k^2].$$

Iterate this inequality to achieve the result (i) in the theorem.

For result (ii), we simply note that if $\alpha_k = \frac{1}{L + \eta}$, then using $2 \max\{L, \eta\} > L + \eta > \eta$, we have

$$\mathbb{E}[D_{k+1}^2] \le \exp\left(\frac{-\lambda}{4 \max\{L, \eta\}} + \frac{\sigma_2^2}{\eta^2 m} \right) \mathbb{E}[D_k^2].$$

Substituting $\eta = \max\{L, \frac{8\sigma_2^2}{m\lambda}\}$ gives the result.

B.2. Proof of Theorem 2

Proof Let $D_k = \mathrm{dist}(x_k, \mathcal{X}^\star)$ and $\mathcal{F}_k = \sigma(S_1, \ldots, S_k)$ be the σ-field generated by the first k samples S_i. Then Lemma 4.1 of the paper (Asi & Duchi, 2019b) immediately yields

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

As $D_1 \geq D_k$ (again, by (Asi & Duchi, 2019b), Lemma 4.1), we in turn obtain

$$\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] \leq \max \left\{ 1 - \lambda_1, 1 - \lambda_0 \alpha_k / D_1^{1-\gamma} \right\} D_k^2.$$

The remainder of the argument is algebraic manipulations, as in the proof of Proposition 2 from (Asi & Duchi, 2019b). □

B.3. Proof of Lemma 4.1

Proof For shorthand, we assume w.l.o.g. that $F(x^\star; S) = 0$ with probability 1. The event that $F(x; S^i) \geq \mu \operatorname{dist}(x, \mathcal{X}^\star)^{1+\gamma}$ has probability at least p, and as the median of a $\mathsf{Binomial}(m, p)$ distribution lies in $\{\lfloor mp \rfloor, \lceil mp \rceil\}$, we have

$$\mathbb{P}\left(\overline{F}(x; S^{1:m}) \geq \frac{\lfloor mp \rfloor}{m} \mu \operatorname{dist}(x, x^\star)^{1+\gamma} \right) \geq \frac{1}{2}. \tag{18}$$

Thus, the event

$$A := \left\{ \left\| \overline{F}'(x; S^{1:m}) \right\|_2^2 \leq 4\mathbb{E}\left[\left\| \overline{F}'(x; S^{1:m}) \right\|_2^2 \right], \ \overline{F}(x; S^{1:m}) \geq \frac{\lfloor mp \rfloor}{m} \mu \operatorname{dist}(x, x^\star)^{1+\gamma} \right\}$$

satisfies

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) \geq 1 - \mathbb{P}\left(\left\| \overline{F}'(x; S^{1:m}) \right\|_2^2 \geq 4\mathbb{E}\left[\left\| \overline{F}'(x; S^{1:m}) \right\|_2^2 \right] \right) - \frac{1}{2} \geq \frac{1}{4},$$

where we use inequality (18). We also have

$$\begin{aligned} \mathbb{E}\left[\left\| \overline{F}'(x; S^{1:m}) \right\|_2^2 \right] = \|f'(x)\|_2^2 \left(1 + \frac{\operatorname{Var}(F'(x; S))}{\|f'(x)\|_2^2} \right) &\leq \left(1 + \frac{\rho}{m} \right) \|f'(x)\|_2^2 \\ &\leq \left(1 + \frac{\rho}{m} \right) L^2 \operatorname{dist}(x, \mathcal{X}^\star)^{2\gamma}, \end{aligned}$$

where we have used Conditions (G.iii) and (G.ii). Applying these observations gives

$$\begin{aligned} &\mathbb{E}\left[\min\left\{ \alpha \overline{F}(x; S^{1:m}), \frac{\overline{F}(x; S^{1:m})^2}{\left\| \overline{F}'(x; S^{1:m}) \right\|_2^2} \right\} \right] \\ &\geq \frac{1}{4} \min\left\{ \alpha \frac{\lfloor mp \rfloor}{m} \mu \operatorname{dist}(x, \mathcal{X}^\star)^{1+\gamma}, \frac{(\lfloor mp \rfloor / m)^2 \mu^2 \operatorname{dist}(x, \mathcal{X}^\star)^{2+2\gamma}}{4\mathbb{E}[\|\overline{F}'(x; S^{1:m})\|_2^2]} \right\} \\ &\geq \frac{1}{4} \min\left\{ \alpha \frac{\lfloor mp \rfloor}{m} \mu \operatorname{dist}(x, \mathcal{X}^\star)^{1+\gamma}, \frac{(\lfloor mp \rfloor / m)^2 \mu^2 \operatorname{dist}(x, \mathcal{X}^\star)^2}{4L^2(1 + \frac{\rho}{m})} \right\}, \end{aligned}$$

as desired. □

C. Proofs from Section 5

C.1. Proof of Theorem 3

Proof Let $\mathcal{P} = \mathcal{P}_\gamma(\lambda_1)$ for short, and assume w.l.o.g. that $\lambda_1 \leq 1/(1+\gamma)^2$, as the result is trivial otherwise. We base our argument on Le Cam's two point method (see, e.g., (Wainwright, 2019), Eq. (15.14)). We consider two probability distributions P_1, P_{-1}, and let $\mathcal{X}_v^\star$ be (for now) arbitrary sets indexed by $v \in \{\pm 1\}$. Then recall the variation distance $\|P - Q\|_{\mathrm{TV}} = \sup_A |P(A) - Q(A)|$ between distributions P and Q, we have Le Cam's two-point method:

Lemma C.1 (Le Cam)**.** *Let $\widehat{x}^k$ be an arbitrary function of $S_1, \ldots, S_k$. Then*

$$\max_{v \in \{-1,1\}} \mathbb{E}_{P_v^k}\left[\operatorname{dist}(\widehat{x}^k, \mathcal{X}_v^\star)^p \right] \geq \frac{1}{8} \operatorname{dist}(\mathcal{X}_{-1}^\star, \mathcal{X}_1^\star)^2 \left(1 - \|P_{-1}^k - P_1^k\|_{\mathrm{TV}} \right).$$

To use Lemma C.1 to lower bound the minimax risk it suffices to choose a pair of problems $(F, P_v) \in \mathcal{P}$ whose optimal sets are well-separated and apply the lemma. To that end, let $\delta \in (0,1)$ to be chosen later, and consider the choices

$$P_{-1} : \begin{cases} S = 0 & \text{w.p. } 1-\delta \\ S = -1 & \text{w.p. } \delta \end{cases} \quad P_1 : \begin{cases} S = 0 & \text{w.p. } 1-\delta \\ S = 1 & \text{w.p. } \delta. \end{cases} \tag{19}$$

Our functions F are trivial to construct: given the radius R, we define

$$F(x;1) = \frac{1}{1+\gamma}|x-R|^{1+\gamma}, \;\; F(x;-1) = \frac{1}{1+\gamma}|x+R|^{1+\gamma}, \;\; F(x;0) = 0. \tag{20}$$

The intuition here is that given a sample $S \in \{-1, 0, 1\}$, we either completely identify the distribution or receive no information.

It remains to show that the pairs $(F, P_v) \in \mathcal{P}$ and to bound the variation distance $\|P_1^k - P_{-1}^k\|_{\mathrm{TV}}$. For the latter, we have

Lemma C.2. *Let P_{-1}, P_1 be as in Eq. (19). Then $\|P_{-1}^k - P_1^k\|_{\mathrm{TV}} = 1 - (1-\delta)^k$.*

Proof For any distributions P, Q, with densities p, q w.r.t. a base measure μ, we have $\|P - Q\|_{\mathrm{TV}} = P(p > q) - Q(p > q)$. For P_{-1}, P_1 as above, we thus have

$$\begin{aligned} \|P_{-1}^k - P_1^k\|_{\mathrm{TV}} &= P_1^k(\text{there exists } i \in [k] \text{ s.t. } S_i = 1) \\ &= 1 - P_1(S_1 = 0, \ldots, S_k = 0) = 1 - (1-\delta)^k. \end{aligned}$$

□

Now, consider the functions

$$f_v(x) := \mathbb{E}_{P_v}[F(x;S)] = \frac{\delta}{1+\gamma}|x - vR|^{1+\gamma}.$$

We have $\kappa_\gamma(f) = 1$, so that the problem is well-conditioned, and the optimal sets $\mathcal{X}_v^\star := \operatorname{argmin}_{x\in\mathcal{X}} f_v(x)$ are the singletons $\mathcal{X}_v^\star = \{x_v^\star = vR\}$. Additionally, we have

$$\begin{aligned} &\mathbb{E}_v\left[(F(x;S) - F(x_v^\star;S)) \min\left\{\alpha, \frac{F(x;S) - F(x_v^\star;S)}{\|F'(x;S)\|_2^2}\right\}\right] \\ &= \frac{\delta}{1+\gamma}|x - vR|^{1+\gamma} \min\left\{\alpha, \frac{|x-vR|^{1+\gamma}}{(1+\gamma)|x - vr|^{2\gamma}}\right\} \\ &= \min\left\{\frac{\delta\alpha}{1+\gamma}, \frac{\delta}{(1+\gamma)^2}\operatorname{dist}(x, \mathcal{X}_v^\star)^{1-\gamma}\right\}\operatorname{dist}(x, \mathcal{X}_v^\star)^{1+\gamma}, \end{aligned}$$

so by choosing $\delta = (1+\gamma)^2\lambda_1 \le 1$, our problems (F, P_v) belong to $\mathcal{P}_\gamma(\lambda_1)$. Le Cam's Lemma C.1 and the variation distance bound in Lemma C.2 imply that

$$\max_{v\in\{\pm 1\}} \mathbb{E}_{P_v^k}\left[|\widehat{x}^k - x_v^\star|^2\right] \ge \frac{1}{8}|x_1^\star - x_{-1}^\star|^2(1-\delta)^k = \frac{R^2}{2}(1-\delta)^k.$$

Substituting $\delta = (1+\gamma)^2\lambda_1$ gives the result. □

C.2. Proof of Theorem 4

Proof Let $U = [u_1 \; \cdots \; u_n] \in \mathbb{R}^{n\times n}$ be an arbitrary orthogonal matrix, so $U^T U = UU^T = I_n$. Let $\mathcal{P}_n$ be the collection of linear regression problems with data matrices $A \in \mathbb{R}^{m\times n}$ chosen by taking $m \le n$ columns $(u_{i(1)}, \ldots, u_{i(m)})$ of U uniformly at random and setting $A = \sqrt{n/m}[u_{i(1)} \; \cdots \; u_{i(m)}]^T$, so that $\mathbb{E}[A^T A] = I_n$ and $(A^T A)^2 = (n/m)A^T A$, and let $b = Ax^\star$, where $x^\star \sim \pi = \mathsf{N}(0, \frac{R^2}{n}I_n)$ follows a Gaussian prior. Each observation S_i corresponds to releasing (perfectly) a random linear projection of $x^\star$, so that given the k observations, if we let $C_k = [A_1 \; \cdots \; A_k] \in \mathbb{R}^{n \times mk}$ denote the concatenated data matrix after k observations, the posterior on $x^\star$ is

$$x^\star \mid (S_1, \ldots, S_k) \sim \mathsf{N}\left(\mathbb{E}[x^\star \mid S_1, \ldots, S_k], \frac{R^2}{n}(I_n - C_k(C_k^T C_k)^{-1} C_k^T)\right),$$

that is, the covariance projects out C_k. By a standard Bayesian argument,

$$\inf_{\widehat{x}^k} \mathbb{E}\left[\left\|\widehat{x}^k - x^\star\right\|_2^2\right] = \mathbb{E}\left[\left\|\mathbb{E}[x^\star \mid S_1^k] - x^\star\right\|_2^2\right] = R^2 \mathbb{E}\left[\frac{n - \operatorname{rank}(C_k)}{n}\right], \tag{21}$$

as $I_n - C_k(C_k^T C_k)^{-1} C_k^T$ is a rank $n - \operatorname{rank}(C_k)$ projection matrix. Let $r_k = \operatorname{rank}(C_k)$ for shorthand. Then we may compute $\mathbb{E}[r_k]$ exactly by noting that

$$\mathbb{E}[r_k \mid r_{k-1}] = r_{k-1} + m\frac{n - r_{k-1}}{n} = \left(1 - \frac{m}{n}\right) r_{k-1} + m,$$

so that with $r_1 = m$ we obtain

$$\mathbb{E}[r_k] = m \sum_{i=1}^k \left(1 - \frac{m}{n}\right)^{k-i} = m \frac{1 - (1 - m/n)^k}{1 - (1 - m/n)} = n - n\left(1 - \frac{m}{n}\right)^k,$$

and substituting this into expression (21) gives

$$\inf_{\widehat{x}^k} \mathbb{E}\left[\left\|\widehat{x}^k - x^\star\right\|_2^2\right] = R^2 \left(1 - \frac{m}{n}\right)^k. \tag{22}$$

We now use expression (22) to prove the two results in the theorem. For the first, we note that for $s = (A, b)$, we have $\nabla F(x; s) = A^T(Ax - b) = A^T A(x - x^\star)$, and as $(A^T A)^2 = \frac{n}{m} A^T A$ by construction and $\mathbb{E}[A^T A] = I_n$,

$$\begin{aligned} &\mathbb{E}\left[(F(x; S) - F(x^\star; S)) \min\left\{\alpha, \frac{F(x; S) - F(x^\star; S)}{\|\nabla F(x; S)\|_2^2}\right\}\right] \\ &= \mathbb{E}\left[\min\left\{\frac{\alpha}{2}\left\|A(x - x^\star)\right\|_2^2, \frac{\|A(x - x^\star)\|_2^4}{4\left\|A^T A(x - x^\star)\right\|_2^2}\right\}\right] = \min\left\{\frac{\alpha}{2}, \frac{m}{4n}\right\}\left\|x - x^\star\right\|_2^2. \end{aligned}$$

In particular, we can choose m, n so that $\frac{m}{4n} \geq \lambda_1$ the problem satisfies Assumption 4 with $\gamma = 1$ and $\lambda_0 = \frac{1}{2}$. This gives the first result by substituting into expression (22) and taking m, n so that $\frac{m}{n}$ is arbitrarily close to $4\lambda_1$.

For the second result, we recognize the noise-to-signal ratio

$$\frac{\operatorname{Var}(\nabla F(x; S))}{\|\nabla f(x)\|_2^2} \leq \frac{\frac{n}{m}\left\|x - x^\star\right\|_2^2}{\left\|x - x^\star\right\|_2^2} = \frac{n}{m}.$$

Making appropriate substitutions by taking $\frac{n}{m} \leq \rho$ gives the second lower bound. □