Efficient Single-Pass Algorithms for Stochastic Optimization

Chris Junchi Li$^\diamond$

Department of Electrical Engineering and Computer Sciences$^\diamond$
University of California, Berkeley

October 6, 2024

Abstract

In this paper, we introduce an efficient single-pass stochastic optimization algorithm that competes with the empirical risk minimizer (ERM) across a broad range of convex problems. Our algorithm achieves the statistical rate of convergence of the ERM while using linear time and memory, making it scalable for large datasets. Under standard smoothness and strong convexity assumptions, we show that our algorithm's excess risk decreases at a super-polynomial rate with respect to sample size, even matching the ERM rate for linear regression. Additionally, the algorithm is trivially parallelizable, allowing for significant computational speedups. We provide finite-sample analysis for linear models and general M-estimation problems, quantifying the sample size at which our method becomes competitive with the ERM.

Keywords: Stochastic optimization, Empirical risk minimization, Single-pass algorithms, Super-polynomial convergence, Streaming data.

1 Introduction

Empirical Risk Minimization (ERM), which optimizes the empirical loss over the sample, is widely regarded as a statistically optimal estimator in many settings. However, implementing the ERM typically requires multiple passes through the data, which becomes impractical in streaming or memory-constrained environments. This motivates the search for single-pass algorithms that can achieve comparable performance to the ERM while significantly reducing computational and memory overhead.

In this work, we propose a novel single-pass algorithm that achieves the same statistical rate as the ERM across a wide class of convex problems, including linear and logistic regression, generalized linear models, and smoothed Huber losses. Our method leverages a variant of stochastic variance-reduced gradient (SVRG) techniques, combined with carefully tuned gradient estimates, to ensure both rapid convergence and minimal memory usage.

We establish that the algorithm reaches the ERM rate when the sample size exceeds a threshold proportional to the condition number of the optimization problem, quantified by the smoothness and strong convexity parameters. Additionally, we demonstrate that the algorithm is trivially parallelizable, making it particularly well-suited for distributed computing environments.

Our main contributions include:

- The development of a single-pass algorithm that matches the statistical performance of the ERM for a wide variety of convex problems.

- A finite-sample analysis that quantifies the rate at which the algorithm becomes competitive with the ERM.

- The demonstration that the algorithm is highly efficient in terms of both time and space, making it suitable for streaming and distributed environments.

Mathematically, Consider the following optimization problem:

$$\min_{w \in \mathcal{S}} P(w), \quad \text{where} \quad P(w) \stackrel{\text{def}}{=} \mathbb{E}_{\psi \sim \mathcal{D}}[\psi(w)] \tag{1}$$

and $\mathcal{D}$ is a distribution over convex functions from a Euclidean space $\mathcal{S}$ to $\mathbb{R}$ (*e.g.* $\mathcal{S} = \mathbb{R}^d$ in the finite dimensional setting). Let w_* be a minimizer of P and suppose we observe the functions $\psi_1, \psi_2, \ldots, \psi_N$ independently sampled from $\mathcal{D}$. Our objective is to compute an estimator $\widehat{w}_N$ so that the expected *error* (or, equivalently, the *excess risk*):

$$\mathbb{E}[P(\widehat{w}_N) - P(w_*)]$$

is small, where the expectation is over the estimator $\widehat{w}_N$ (which depends on the sampled functions).

Stochastic approximation algorithms, such as stochastic gradient descent (SGD) [Robbins and Monro(1951)], are the most widely used in practice, due to their ease of implementation and their efficiency with regards to runtime and memory. Without consideration for computational constraints, we often wish to compute the *empirical risk minimizer* (ERM; or, equivalently, the *M-estimator*):

$$\widehat{w}_N^{\text{ERM}} \in \arg\min_{w \in \mathcal{S}} \frac{1}{N} \sum_{i=1}^{N} \psi_i(w) \tag{2}$$

In the context of statistical modeling, the ERM is the maximum likelihood estimator (MLE). Under certain regularity conditions, and under correct model specification,[1] the MLE is asymptotically efficient, in that no unbiased estimator can have a lower variance in the limit (see [Lehmann and Casella(1998), van der Vaart(2000)]).[2] Analogous arguments have been made in the stochastic approximation setting, where we do not necessarily have a statistical model of the distribution $\mathcal{D}$ (see [Kushner and Yin(2003)]).

The question we aim to address is as follows. Consider the ratio:

$$\frac{\mathbb{E}[P(\widehat{w}_N^{\text{ERM}}) - P(w_*)]}{\mathbb{E}[P(\widehat{w}_N) - P(w_*)]} \tag{3}$$

We seek an algorithm to compute $\widehat{w}_N$ in which: (1) under sufficient regularity conditions, this ratio approaches 1 on *every* problem $\mathcal{D}$ and (2) it does so quickly, at a rate quantifiable in terms of the number of samples, the dependence on the initial error (and other relevant quantities), and the computational time and space usage.

1.1 This work

Under certain smoothness assumptions on ψ and strong convexity assumptions on P (applicable to linear and logistic regression, generalized linear models, smoothed Huber losses, and various other M-estimation problems), we provide an algorithm where:

1. The algorithm achieves the same statistical rate of convergence as the ERM on every problem, even considering constant factors, and we quantify the sample size at which this occurs.

---

[1] A well specified statistical model is one where the data is generated under some model in the parametric class. See the linear regression Section 3.1.

[2] However, note that biased estimators, such as the James-Stein estimator, can outperform the MLE [Lehmann and Casella(1998)].

Algorithm / analysis	Problem	Step size	Initial error dependence	Parallel-izable	Finite-sample analysis
[Polyak and Juditsky(1992)]	general	decaying: $1/n^c$	?	?	✗
[Polyak and Juditsky(1992)] / [Dieuleveut and Bach(2014)]	linear regression	constant	$\Omega(1/n^2)$	?	✓
This work: Streaming SVRG	general	constant	$1/n^{\omega(1)}$	✓	✓

Table 1. Comparison of known *streaming* algorithms which achieve a *constant* competitive ratio to the ERM. [Polyak and Juditsky(1992)] is an SGD algorithm with iterate averaging. Concurrent to and independently from our work, [Dieuleveut and Bach(2014)] provide a finite-sample analysis for SGD with averaging in the linear regression problem setting (where the learning rate can be taken as constant). In the "problem" column, "general" indicates problems under the regularity assumptions herein. [Polyak and Juditsky(1992)] require the step size to decay with the sample size n, as $1/n^c$ with c strictly in the range $1/2 < c < 1$. The dependence on c in a finite-sample analysis is unclear (and tuning the decay of learning rates is often undesirable in practice). The initial error is $P(w_0) - P(w_*)$, where w_0 is the starting point of the algorithm. We seek algorithms in which the initial error dependence is significantly lower in order, and we write $1/n^{\omega(1)}$ to indicate that it can be driven down to an arbitrarily low-order polynomial. See Remark with regard to parallelization.

2. The algorithm can be implemented in linear time with a single pass of the observed data, using space linear in the size of a single sample.
3. The algorithm decreases the standard notion of initial error at a super-polynomial rate.[3]
4. The algorithm is trivially parallelizable (see Remark).

Table 1 compares previous (and concurrent) algorithms that enjoy the first two guarantees; this work is the first with a finite-sample analysis handling the more general class of problems. Our algorithm is a variant of the stochastic variance reduced gradient procedure of [Johnson and Zhang(2013)].

Importantly, we quantify how fast we obtain a rate comparable to that of the ERM. For the case of linear regression, we have non-trivial guarantees when the sample size N is larger than a constant times what can be interpreted as a condition number, $\kappa = L/\mu$, where μ is a strong convexity parameter of P and where L is a smoothness parameter of each ψ. Critically, after N is larger than κ, the initial error is divided by a factor that can be larger than any polynomial in N/κ.

Finally, in order to address this question on a per-problem basis, we provide both upper and lower bounds for the rate of convergence of the ERM.

1.2 Related work

Stochastic optimization dates back to the work of [Robbins and Monro(1951)] and has seen much subsequent work [Kushner and Clark(1978), Kushner and Yin(2003), Nemirovski and Yudin(1983)]. More recently, questions of how to quantify and compare rates of estimation procedures – with implications to machine learning problems in the streaming and large dataset settings – have been raised and discussed several times (see [Bottou and Bousquet(2008), Agarwal and Bottou(2014)]).

Stochastic approximation. The pioneering work of [Polyak and Juditsky(1992)] and [Ruppert(1988)] provides an asymptotically optimal streaming algorithm, by averaging the iterates of an SGD pro-

_____

[3]A function is super-polynomial if grows faster than any polynomial.

cedure. It is unclear how quickly these algorithms converge to the rate of the ERM in finite sample; the relevant dependencies, such as the dependence on the initial error – that is, $P(w_0) - P(w_*)$ where w_0 is the starting point of the algorithm – are not specified. In particular, they characterize the limiting distribution of $\sqrt{N}(\widehat{w}_N - w_*)$, essentially arguing that the variance of the iterate-averaging procedure matches the asymptotic distribution of the ERM (see [Kushner and Yin(2003)]).

In a series of papers, [Bach and Moulines(2011)], [Bach and Moulines(2013)], [Dieuleveut and Bach(2014)], and [Defossez and Bach(2015)] provide non-asymptotic analysis of the same averaging schemes. Of these, for the specific case of linear least-squares regression, [Dieuleveut and Bach(2014)] and [Defossez and Bach(2015)] provide rates which are competitive with the ERM, concurrently and independent of results presented herein. The work in [Bach and Moulines(2011)] and [Bach and Moulines(2013)] either does not achieve the ERM rate or has a dependence on the initial error which is not lower in order; it is rather in [Dieuleveut and Bach(2014)] and [Defossez and Bach(2015)] that dependence on the initial error decaying as $1/N^2$ is shown.

For the special case of least squares, one could adapt the algorithm and guarantees of [Dieuleveut and Bach(2014), Defossez and Bach(2015)], by replacing global averaging with random restarts, to obtain superpolynomial rates (results comparable to ours when specializing to linear regression). For more general problems, it is unclear how such an adaptation would work – using constant step sizes alone may not suffice. In contrast, as shown in Table 1, our algorithm is identical for a wide variety of cases and does not need decaying rates (whose choices may be difficult in practice).

We should also note that much work has characterized rates of convergence under various assumptions on P and ψ different than our own. Our case of interest is when P is strongly convex. For such P, the rates of convergence of many algorithms are $O(1/N)$, often achieved by averaging the iterates in some way [Nemirovski et al.(2009)Nemirovski, Juditsky, Lan, and Shapiro, Juditsky and Nesterov(2010), Rakhlin et al.(2012)Rakhlin, Shamir, and Sridharan, Hazan and Kale(2014)]. These results do not achieve a constant competitive ratio, for a variety of reasons (they have a leading order dependencies on various quantities, including the initial error along with strong convexity and smoothness parameters). Solely in terms of the dependence on the sample size N, these rates are known to be optimal [Nemirovski and Yudin(1983), Nesterov(2004), Agarwal et al.(2012)Agarwal, Bartlett, Raviku

Empirical risk minimization (M-estimation). In statistics, it is classically argued that the MLE, under certain restrictions, is an asymptotically efficient estimator for well-specified statistical models [Lehmann and Casella(1998), van der Vaart(2000)]. Analogously, in an optimization context, applicable to mis-specified models, similar asymptotic arguments have been made: under certain restrictions, the asymptotically optimal estimator is one which has a limiting variance that is equivalent to that of the ERM [Anbar(1971), Fabian(1973), Kushner and Clark(1978)].

With regards to finite-sample rates, [Agarwal et al.(2012)Agarwal, Bartlett, Ravikumar, and Wainwright] provide information-theoretic lower bounds (for any strategy) for certain stochastic convex optimization problems. This result does not imply our bounds as they do not consider the same smoothness assumptions on ψ. For the special case of linear least-squares regression, there are several upper bounds (for instance, [Caponnetto and De Vito(2007), Hsu et al.(2014)Hsu, Kakade, and Zhang]). Recently, [Shamir(2014)] provides lower bounds specifically for the least-squares estimator, applicable under model mis-specification, and sharp only for specific problems.

Linearly convergent optimization (and approaches based on doubling). There are numerous algorithms for optimizing sums of convex functions that converge linearly, *i.e.* that depend only logarithmically on the target precision. Notably, several recently developed such algorithms

are applicable in the setting where the sample size N becomes large, due to their stochastic nature [Strohmer and Vershynin(2009), Le Roux et al.(2012)Le Roux, Schmidt, and Bach, Shalev-Shwartz and Zhang(20 Johnson and Zhang(2013)]. These procedures minimize a sum of N losses in time (near to) linear in N, provided N is sufficiently large relative to the dimension and the condition number.

Naively, one could attempt to use one of these algorithms to directly compute the ERM. Such an attempt poses two difficulties. First, we would need to prove concentration results for the empirical function $\widehat{P}_N(w) = \frac{1}{N}\sum_{i=1}^{N}\psi_i(w)$; in order to argue that these algorithms perform well in linear time with respect to the objective P, one must relate the condition number of $\widehat{P}_N(w)$ to the condition number of $P(w)$. Second, we would need new generalization analysis in order to relate the in-sample error $\varepsilon_N(\widehat{w}_N)$, where $\varepsilon_N(w) \stackrel{\text{def}}{=} \widehat{P}_N(w) - \min_{w'}\widehat{P}_N(w')$, to the generalization error $\mathbb{E}[P(\widehat{w}_N) - P(w_*)]$. To use existing generalization analyses would demand that $\varepsilon_N(w_N) = \Omega(1/N)$, but the algorithms in question all require at least $\log N$ passes of the data (furthermore scaled by other problem-dependent factors) to achieve such an in-sample error. Hence, this approach would not immediately describe the generalization error obtained in time linear in N. Finally, it requires that entire observed data sample, constituting the sum, be stored in memory.

A second natural question is: can one naively use a doubling trick with an extant algorithm to compete with the ERM? By this we mean to iteratively run such a linearly convergent optimization algorithm, on increasingly larger subsets of the data, with the hope of cutting the error at each iteration by a constant fraction, eventually down to that of the ERM. There are two points to note for this approach. First, the approach is not implementable in a streaming model as one would eventually have to run the algorithm on a constant fraction of the entire dataset size, thus essentially holding the entire dataset in memory. Second, proving such an algorithm succeeds would similarly involve the aforementioned type of generalization argument.

We conjecture that these tight generalization arguments described are attainable, although with a somewhat involved analysis. For linear regression, the bounds in [Hsu et al.(2014)Hsu, Kakade, and Zhang] may suffice. More generally, we believe the detailed ERM analysis provided herein could be used.

In contrast, the statistical convergence analysis of our single-pass algorithm is self-contained and does not go through any generalization arguments about the ERM. In fact, it avoids matrix concentration arguments entirely.

Comparison to related work. To our knowledge, this work provides the first streaming algorithm guaranteed to have a rate that approaches that of the ERM (under certain regularity assumptions on $\mathcal{D}$), where the initial error is decreased at a super-polynomial rate. The previous work, in the general case that we consider, only provides asymptotic convergence guarantees [Polyak and Juditsky(1992)]. For the special case of linear least-squares regression, the concurrent and independent work presented in [Dieuleveut and Bach(2014)] and [Defossez and Bach(2015)] also converges to the rate of the ERM, with a lower-order dependence on the initial error of $\Omega(1/N^2)$. Furthermore, even if we ignored memory constraints and focused solely on computational complexity, our algorithm compares favorably to using state-of-the-art algorithms for minimizing sums of functions (such as the linearly convergent algorithms in [Le Roux et al.(2012)Le Roux, Schmidt, and Bach, Shalev-Shwartz and Zhang(2013), Johnson and Zhang(2013)]); as discussed above, obtaining a convergence rate with these algorithms would entail some further generalization analysis.

It would be interesting if one could quantify an approach of restarting the algorithm of [Polyak and Juditsky(199 to obtain guarantees comparable to our streaming algorithm. Such an analysis could be delicate in settings other than linear regression, as their learning rates do not decay too quickly or too slowly (they must decay strictly faster than $1/\sqrt{N}$, yet more slowly than $1/N$). In contrast, our algorithm

takes a constant learning rate to obtain its constant competitive ratio. Furthermore, our algorithm is easily parallelizable and its analysis, we believe, is relatively transparent.

1.3 Organization

Section 2 summarizes our main results, and Section 3 provides applications to a few standard statistical models. Section 4 provides the main technical claims for our algorithm, Streaming SVRG (Algorithm 1). Section 5 provides finite-sample rates for the ERM, along with proofs for these rates. The Appendix contains various technical lemmas and proofs of our corollaries.

2 Main results

This section summarizes our main results, as corollaries of more general theorems provided later. After providing our assumptions in Section 2.1, Section 2.2 provides the algorithm, along with performance guarantees. Then Section 2.3 provides upper and lower bounds of the statistical rate of the empirical risk minimizer.

First, a few preliminaries and definitions are needed. Denote $\|x\|_M^2 \overset{\text{def}}{=} x^\mathsf{T} M x$ for a vector x and a matrix M of appropriate dimensions. Denote $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ as the maximal and minimal eigenvalues of a matrix M. Let I denote the identity matrix. Also, for positive semidefinite symmetric matrices A and B, $A \preceq B$ if and only if $x^\mathsf{T} A x \leq x^\mathsf{T} B x$ for all x.

Throughout, define σ^2 as:

$$\sigma^2 \overset{\text{def}}{=} \mathbb{E}_{\psi \sim \mathcal{D}} \left[\frac{1}{2} \left\| \nabla \psi(w_*) \right\|_{(\nabla^2 P(w_*))^{-1}}^2 \right] \tag{4}$$

This quantity governs the precise (problem dependent) convergence rate of the ERM. Namely, under certain restrictions on $\mathcal{D}$, we have

$$\lim_{N \to \infty} \frac{\mathbb{E}[P(\widehat{w}_N^{\text{ERM}}) - P(w_*)]}{\sigma^2/N} = 1 \tag{5}$$

This limiting rate is well-established in asymptotic statistics (see, for instance, [van der Vaart(2000)]), whereas Section 2.3 provides upper and lower bounds on this rate for finite sample sizes N. Analogous to the Cramér-Rao lower bound, under certain restrictions, σ^2/N is the asymptotically efficient rate for stochastic approximation problems [Anbar(1971), Fabian(1973), Kushner and Yin(2003)].[4]

The problem dependent rate of σ^2/N sets the benchmark. Statistically, we hope to achieve a leading order dependency of σ^2/N quickly, with rapidly-decaying dependence on the initial error.

2.1 Assumptions

We now provide two assumptions under which we analyze the convergence rate of our streaming algorithm, Algorithm 1. Our first assumption is relatively standard. It provides upper and lower quadratic approximations (the lower approximation is on the full objective P).

Assumption 1. *Suppose that:*

   1. The objective P is twice differentiable.

---

[4]Though, as with Cramér-Rao, this may be improvable with biased estimators.

2. *(Strong convexity) The objective P is μ-strongly convex, i.e. for all $w, w' \in \mathcal{S}$,*

$$P(w) \geq P(w') + \nabla P(w')^{\mathsf{T}}(w - w') + \frac{\mu}{2}\|w - w'\|_2^2 \tag{6}$$

3. *(Smoothness) Each loss ψ is L-smooth (with probability one), i.e. for all $w, w' \in \mathcal{S}$,*

$$\psi(w) \leq \psi(w') + \nabla \psi(w')^{\mathsf{T}}(w - w') + \frac{L}{2}\|w - w'\|_2^2 \tag{7}$$

Our results in fact hold under a slightly weaker version of this assumption – see Remark . Define:

$$\kappa \stackrel{\text{def}}{=} \frac{L}{\mu} \tag{8}$$

The quantity κ can be interpreted as the condition number of the optimization objective (1). The following definition quantifies a global bound on the Hessian.

Definition 1 (α-bounded Hessian). *Let $\alpha \geq 1$ be the smallest value (if it exists) such that for all $w \in \mathcal{S}$, $\nabla^2 P(w_*) \preceq \alpha \nabla^2 P(w)$.*

Under Assumption 1, we have $\alpha \leq \kappa$, because L-smoothness implies $\nabla^2 P(w_*) \preceq LI$ and μ-strong convexity implies $\mu I \preceq \nabla^2 P(w)$. However, α could be much smaller. For instance, $\alpha = 1$ in linear regression, whereas κ is the maximum to minimum eigenvalue ratio of the design matrix.

Our second assumption offers a stronger, local relationship on the objective's Hessian, namely *self-concordance*. A function is self-concordant if its third-order derivative is bounded by a multiple of its second-order derivative. Formally, $f : \mathbb{R} \to \mathbb{R}$ is M self-concordant if and only if f is convex and $|f'''(x)| \leq M f''(x)^{3/2}$. A multivariate function $f : \mathbb{R}^d \to \mathbb{R}$ is M self-concordant if and only if its restriction to any line is M self-concordant.

Assumption 2 (Self-concordance). *Suppose that:*

1. *P is M-self concordant (or that the weaker condition in Equation (30) holds).*

2. *The following kurtosis condition holds:*

$$\frac{\mathbb{E}_{\psi \sim \mathcal{D}}\left[\|\nabla \psi(w_*)\|_2^4\right]}{\left(\mathbb{E}_{\psi \sim \mathcal{D}}\left[\|\nabla \psi(w_*)\|_2^2\right]\right)^2} \leq C$$

Note that these two assumptions are also standard assumptions in the analysis of the two phases of Newton's method (aside from the kurtosis condition): the first phase of Newton's method gets close to the minimizer quickly (based on a global strong convexity assumption) and the second phase obtains quadratic convergence (based on local curvature assumptions on how fast the local Hessian changes, *e.g.* self-concordance). Moreover, our proof of the streaming algorithm follows a similar structure; we use Assumption 1 to analyze the progress of our algorithm when the current point is far away from optimality and Assumption 2 when it is close.

Algorithm 1 Streaming Stochastic Variance Reduced Gradient (Streaming SVRG)

  Input: Initial point $\widetilde{w}_0$, batch sizes $\{k_0, k_1, \ldots\}$, update frequency m, learning rate η, smoothness L
  for each stage $s = 0, 1, 2, \ldots$ **do**
   Sample $\widetilde{\psi}_1, \ldots, \widetilde{\psi}_{k_s}$ from $\mathcal{D}$ and compute the estimate

$$\widehat{\nabla P(\widetilde{w}_s)} = \frac{1}{k_s} \sum_{i \in [k_s]} \nabla \widetilde{\psi}_i(\widetilde{w}_s) \tag{9}$$

   Sample $\widetilde{m}$ uniformly at random from $\{1, 2, \ldots, m\}$.
   $w_0 \leftarrow \widetilde{w}_s$
   for $t = 0, 1, \ldots, \widetilde{m} - 1$ **do**
    Sample ψ_t from $\mathcal{D}$ and set

$$w_{t+1} \leftarrow w_t - \frac{\eta}{L} \left(\nabla \psi_t(w_t) - \nabla \psi_t(\widetilde{w}_s) + \widehat{\nabla P(\widetilde{w}_s)} \right) \tag{10}$$

   end for
   $\widetilde{w}_{s+1} \leftarrow w_{\widetilde{m}}$
  end for

2.2 Algorithm

Here we describe a streaming algorithm and provide its convergence guarantees. Algorithm 1 is inspired by the Stochastic Variance Reduced Gradient (SVRG) algorithm of [Johnson and Zhang(2013)] for minimizing a strongly convex sum of smooth losses. The algorithm follows a simple framework that proceeds in stages. In each stage s we draw k_s samples independently at random from $\mathcal{D}$ and use these samples to obtain an estimate of the gradient of P at the current point, $\widetilde{w}_s$ ((9)). This *stable* gradient, denoted $\widehat{\nabla P(\widetilde{w}_s)}$, is then used to decrease the variance of a gradient descent procedure. For each of $\widetilde{m}$ steps (where $\widetilde{m}$ is chosen uniformly at random from $\{1, 2, \ldots, m\}$), we draw a sample ψ from $\mathcal{D}$ and take a step opposite to its gradient at the current point, plus a zero-bias correction given by $\nabla \psi(\widetilde{w}_s) - \widehat{\nabla P(\widetilde{w}_s)}$ (see (10)).

The remainder of this section shows that, for suitable choices of k_s and m, Algorithm 1 achieves desirable convergence rates under the aforementioned assumptions.

Remark (Generalizing SVRG)**.** Note that Algorithm 1 is a generalization of SVRG. In particular if we chose $k_s = \infty$, *i.e.* if $\widehat{\nabla P(\widetilde{w}_s)} = \nabla P(\widetilde{w}_s)$, then our algorithm coincides with the SVRG algorithm of [Johnson and Zhang(2013)]. Also, note that [Johnson and Zhang(2013)] do not make use of any self-concordance assumptions.

Remark (Non-conformance to stochastic first-order oracle models)**.** Algorithm 1 is not implementable in the standard stochastic first-order oracle model, *e.g.* that which is assumed in order to obtain the lower bounds in [Nemirovski and Yudin(1983)] and [Agarwal et al.(2012)Agarwal, Bartlett, Ravikumar, Streaming SVRG computes the gradient of the randomly drawn ψ at two points, while the oracle model only allows gradient queries at one point.

We have the following algorithmic guarantee under only Assumption 1, which is a corollary of Theorem 1 (also see the Appendix).

Corollary 1 (Convergence under α-bounded Hessians)**.** *Suppose Assumption 1 holds. Fix $\widetilde{w}_0 \in \mathbb{R}^d$.*

For $p \geq 2$ and $b \geq 3$, set $\eta = \frac{1}{20b^{p+1}}$, $m = \frac{20b^{p+1}\kappa}{\eta}$, $k_0 = 20\alpha\kappa b^{p+1}$, and $k_s = bk_{s-1}$. Denote:

$$N_s \overset{\text{def}}{=} \sum_{\tau=0}^{s-1}(k_\tau + m)$$

(N_s is an upper bound on the number of samples drawn up to the end of stage s).

Let $\widehat{w}_{N_s}$ be the parameter returned at iteration s by Algorithm 1. For $N_s \geq b^{p^2+6p}\kappa$ (and so $s > p^2 + 6p$), we have

$$\mathbb{E}[P(\widehat{w}_{N_s}) - P(w_*)] \leq \left(\left(1 + \frac{4}{b}\right) \frac{\sqrt{\alpha}\sigma}{\sqrt{N_s}} + \sqrt{\frac{P(\widetilde{w}_0) - P(w_*)}{\left(\frac{N_s}{\alpha\kappa}\right)^p}} \right)^2$$

When $\alpha = 1$ (such as for least squares regression), the above bound achieves the ERM rate of σ^2/N (up to a constant factor, which can be driven to one, as discussed later). Furthermore, under self-concordance, we can drive the competitive ratio (3) down from α to arbitrarily near to 1. The following is a corollary of Theorem 2 (also see the Appendix):

Corollary 2 (Convergence under self-concordance). *Suppose Assumptions 1 and 2 hold. Consider $\widetilde{w}_0 \in \mathbb{R}^d$. For $p \geq 2$ and $b \geq 3$, set $\eta = \frac{1}{20b^{p+1}}$, $m = \frac{20b^{p+1}\kappa}{\eta}$, $k_0 = \max\{400\kappa^2 b^{2p+3}, 10C\} = \max\{bm\kappa, 10C\}$, and $k_s = bk_{s-1}$. Denote $N_s \overset{\text{def}}{=} \sum_{\tau=0}^{s-1}(k_s + m)$ (an upper bound on the number of samples drawn up to the end of stage s). Let $\widehat{w}_{N_s}$ be the parameter returned at iteration s by Algorithm 1. Then:*

$$\mathbb{E}[P(\widehat{w}_{N_s}) - P(w_*)] \leq \left(\left(1 + \tfrac{5}{b}\right) \tfrac{\sigma}{\sqrt{N_s}} + \left(2 + \tfrac{5}{b}\right) \tfrac{\sqrt{\kappa}\sigma}{\sqrt{N_s}} \min\left\{1, \left(\tfrac{N_s}{2(M\sigma+1)^2 k_0}\right)^{-p/2}\right\} + \sqrt{\tfrac{P(\widetilde{w}_0) - P(w_*)}{\left(\frac{N_s}{2k_0}\right)^{p+1}}} \right)^2$$

Remark (Implementation and parallelization). Note that Algorithm 1 is simple to implement and requires little space. In each iteration, the space usage is linear in the size of a single sample (along with needing to count to k_s and m). Furthermore, the algorithm is easily parallelizable once we have run enough stages. In both Theorem 1 and Theorem 2 as s increases k_s grows geometrically, whereas m remains constant. Hence, the majority of the computation time is spent averaging the gradient, *i.e.* (9), which is easily parallelizable.

Note that the constants in the parameter settings for the Algorithm have not been optimized. Furthermore, we have not attempted to fully optimize the time it takes the algorithm to enter the second phase (in which self-concordance is relevant), and we conjecture that the algorithm in fact enjoys even better dependencies. Our emphasis is on an analysis that is flexible in that it allows for a variety of assumptions in driving the competitive ratio to 1 (as is done in the case of logistic regression in Section 3, where we use a slight variant of self-concordance).

Before providing statistical rates for the ERM, let us remark that the above achieves super-polynomial convergence rates and that the competitive ratio can be driven to 1 (recall that σ^2/N is the rate of the ERM).

Remark (Linear convergence and super-polynomial convergence). Suppose the ratio γ between $P(\widetilde{w}_0) - P(w_*)$ and σ^2 is known approximately (within a multiplicative factor), we can let $k_s = k_0$ for $\log_b \gamma$ number of iterations, then start increasing $k_s = bk_{s-1}$. This way in the first $\log_b \gamma$ iterations $\mathbb{E}[P(\widehat{w}_{N_s}) - P(w_*)]$ is decreasing geometrically. Furthermore, even without knowing the ratio γ, we can can obtain a super-polynomial rate of convergence by setting the parameters as we specify in the next remark. (The dependence on the initial error will then be $2^{-\Omega(\log N/\log\log N)^2}$.)

Remark (Driving the ratio to 1). By choosing b sufficiently large, the competitive ratio (3) can be made close to 1 (on every problem). Furthermore, we can ensure this constant goes to 1 by altering the parameter choices adaptively: let $k_s = 4^s(s!)k_0$, and let $\eta_s = \eta/2^s$, $m_s = m \cdot 4^s$. Intuitively, k grows so fast that $\lim_{s \to \infty} k_s/N_s = 1$; η_s and m_s are also changing fast enough so the initial error vanishes very quickly.

2.3 Competing with the ERM

Now we provide a finite-sample characterization of the rate of convergence of the ERM under regularity conditions. This essentially gives the numerator of (3), allowing us to compare the rate of the ERM against the rate achieved by Streaming SVRG. We provide the more general result in Theorem 3; this section focuses on a corollary.

In the following, we constrain the domain $\mathcal{S}$; so the ERM, as defined in (2), is taken over this restricted set. Further discussion appears in Theorem 3 and the comments thereafter.

Corollary 3 (of Theorem 3). *Suppose $\psi_1, \psi_2, \ldots, \psi_N$ are an independently drawn sample from $\mathcal{D}$. Assume the following regularity conditions hold; see Theorem 3 for weaker conditions.*

1. *$\mathcal{S}$ is compact.*

2. *ψ is convex (with probability one).*

3. *w_* is an interior point of $\mathcal{S}$, and $\nabla^2 P(w_*)$ exists and is positive definite.*

4. *(Smoothness) Assume the first, second, and third derivatives of ψ exist and are uniformly bounded on $\mathcal{S}$.*

Then, for the ERM $\widehat{w}_N^{ERM}$ (as defined in (2)), we have

$$\lim_{N \to \infty} \frac{\mathbb{E}[P(\widehat{w}_N^{ERM}) - P(w_*)]}{\sigma^2/N} = 1$$

In particular, the following lower and upper bounds hold. With problem dependent constants C_0 and C_1 (polynomial in the relevant quantities, as specified in Theorem 3), we have for all $p \geq 2$, if N satisfies $\frac{p \log dN}{N} \leq C_0$, then

$$\left(1 - C_1 \sqrt{\frac{p \log dN}{N}}\right) \frac{\sigma^2}{N} \leq \mathbb{E}[P(\widehat{w}_N^{ERM}) - P(w_*)]$$

$$\leq \left(1 + C_1 \sqrt{\frac{p \log dN}{N}}\right) \frac{\sigma^2}{N} + \frac{\max_{w \in \mathcal{S}}(P(w) - P(w_*))}{N^p}$$

3 Applications: one pass learning and generalization

This section provides applications to a few standard statistical models, in part providing a benchmark for comparison on concrete problems. For the widely studied problem of least-squares regression, we also instantiate upper and lower bounds for the ERM. The applications in this section can be extended to include generalized linear models, some M-estimation problems, and other loss functions (*e.g.* the Huber loss).

3.1 Linear least-squares regression

In linear regression, the goal is to minimize the (possibly ℓ_2-regularized) squared loss $\psi_{X,Y}(w) = (Y - w^\mathsf{T}X)^2 + \lambda\|w\|_2^2$ for a random data point $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$. The objective (1) is

$$P(w) = \mathbb{E}_{X,Y\sim\mathcal{D}}\left[(Y - w^\top X)^2\right] + \lambda\|w\|_2^2 \tag{11}$$

3.1.1 Upper bound for the algorithm

Using that $\alpha = 1$, the following corollary illustrates that Algorithm 1 achieves the rate of the ERM,

Corollary 4 (Least-squares performance of streaming SVRG)**.** *Suppose that $\|X\|^2 \leq L$. Define $\mu = \lambda + \lambda_{\min}(\Sigma)$. Using the parameter settings of Theorem 1 and supposing that $N \geq b^{p^2+6p}\kappa$,*

$$\mathbb{E}[P(\widetilde{w}_N) - P(w_*)] \leq \left(\left(1 + \frac{4}{b}\right)\frac{\sigma}{\sqrt{N}} + \sqrt{\frac{P(\widetilde{w}_0) - P(w_*)}{\left(\frac{N}{\kappa}\right)^p}}\right)^2$$

Remark (When $N \leq \kappa$)**.** If the sample size is less than κ and $\lambda = 0$, there exist distributions on X in which the ERM is not unique (as the sample matrix $\frac{1}{N}\sum X_i X_i^\top$ will not be invertible, with reasonable probability, on these distributions by construction).

Remark (When do the streaming SVRG bounds become meaningful?)**.** Algorithm 1 is competitive with the performance of the ERM when the sample size N is slightly larger than a constant times κ. In particular, as the sample N size grows larger than κ, then the initial error is decreased at an arbitrary polynomial rate in N/κ.

Let us consider a few special cases. First, consider the unregularized setting where $\lambda = 0$. Assume also that the least-squares problem is well-specified. That is, $Y = w_*^\top X + \eta$ where $\mathbb{E}[\eta] = 0$ and $\mathbb{E}[\eta^2] = \sigma_{\text{noise}}^2$. Define $\Sigma = \mathbb{E}[XX^\top]$. Here, we have

$$\sigma^2 = \mathbb{E}\|\eta^2 X\|_{\Sigma^{-1}}^2 = d\sigma_{\text{noise}}^2 \tag{12}$$

In other words, Corollary 4 recovers the classical rate in this case.

In the mis-specified case – where we do not assume the aforementioned model is correct (*i.e.* $\mathbb{E}[Y|X]$ may not equal $w_*^\top X$) – define $Y_*(X) = w_*^\top X$, and we have

$$\sigma^2 = \mathbb{E}\left[(Y - Y_*(X))^2\|X\|_{\Sigma^{-1}}^2\right] \tag{13}$$

$$= \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\|X\|_{\Sigma^{-1}}^2\right] + \mathbb{E}\left[(\mathbb{E}[Y \mid X] - Y_*(X))^2\|X\|_{\Sigma^{-1}}^2\right] \tag{14}$$

$$= \mathbb{E}\left[\text{var}(Y \mid X)\|X\|_{\Sigma^{-1}}^2\right] + \mathbb{E}\left[\text{bias}(X)^2\|X\|_{\Sigma^{-1}}^2\right] \tag{15}$$

where the last equality exposes the effects of the approximation error:

$$\text{var}(Y \mid X) \overset{\text{def}}{=} \mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2 \mid X] \qquad \text{and} \qquad \text{bias}(X) \overset{\text{def}}{=} \mathbb{E}[Y \mid X] - Y_*(X). \tag{16}$$

In the regularized setting (a.k.a. ridge regression) – also not necessarily well-specified – we have

$$\sigma^2 = \mathbb{E}[\|(Y - Y_*(X))X + \lambda w_*\|_{(\Sigma+\lambda\mathrm{I})^{-1}}^2] \tag{17}$$

3.1.2 Statistical upper and lower bounds

For comparison, the following corollary (of Theorem 3) provides lower and upper bounds for the statistical rate of the ERM.

Corollary 5 (Least-squares ERM bounds). *Suppose that $\|X\|^2_{(\Sigma+\lambda I)^{-1}} \leq \widetilde{\kappa}$ and the dimension is d (in the infinite dimensional setting, we may take d to be the intrinsic dimension, as per Remark). Let c be an appropriately chosen universal constant. For all $p > 0$, if $\frac{p \log N}{N} \leq \frac{c}{\widetilde{\kappa}}$, then*

$$\mathbb{E}[P(\widehat{w}_N^{ERM}) - P(w_*)] \geq \left(1 - c\sqrt{\frac{\widetilde{\kappa} p \log dN}{N}}\right)\frac{\sigma^2}{N} - \frac{\sqrt{\mathbb{E}\left[Z^4\right]}}{N^{p/2}}$$

where $Z = \|\nabla \psi(w_*)\|_{(\nabla^2 P(w_*))^{-1}} = \|(Y - w_*^\top X)X + \lambda w_*\|_{(\Sigma+\lambda I)^{-1}}$.
For an upper bound, we have two cases:

- *(Unregularized case) Suppose $\lambda = 0$. Assume that we constrain the ERM to lie in some compact set $\mathcal{S}$ (and supposing $w_* \in \mathcal{S}$). Then for all $p > 0$, if $\frac{p \log N}{N} \leq \frac{c}{\widetilde{\kappa}}$, we have*

$$\mathbb{E}[P(\widehat{w}_N^{ERM}) - P(w_*)] \leq \left(1 + c\sqrt{\frac{\widetilde{\kappa} p \log dN}{N}}\right)\frac{\sigma^2}{N} + \frac{\max_{w \in \mathcal{S}}(P(w) - P(w_*))}{N^p}$$

- *(Regularized case) Suppose $\lambda > 0$. Then for all $p > 0$, if $\frac{p \log N}{N} \leq \frac{c}{\widetilde{\kappa}}$, we have*

$$\mathbb{E}[P(\widehat{w}_N^{ERM}) - P(w_*)] \leq \left(1 + c\sqrt{\frac{\widetilde{\kappa} p \log dN}{N}}\right)\frac{\sigma^2}{N} + \frac{\frac{\lambda_{\max}(\Sigma+\lambda)}{\lambda}\sigma^2}{N^p}$$

(this last equation follows from a modification of the argument in Equation (37)).

Remark (ERM comparisons). Interestingly, for the upper bound (when $\lambda = 0$), we see no way to avoid constraining the ERM to lie in some compact set; this allows us to bound the loss P in the event of some extremely low probability failure (see Theorem 3). The ERM upper bound has a term comparable to the initial error of our algorithm. In contrast, the lower bound is for the usual unconstrained least-squares estimator.

3.2 Logistic regression

In (binary) logistic regression, we have a distribution on $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$. For any w, define

$$\mathbb{P}(Y = y \mid w, X) \overset{\text{def}}{=} \frac{\exp(yX^\top w)}{1 + \exp(X^\top w)} \tag{18}$$

for $X \in \mathbb{R}^d$ and $y \in \{0, 1\}$. We do not assume the best fit model w_* is correct. The loss function is taken to be the regularized log likelihood $\psi_{X,y}(w) = -\log \mathbb{P}(Y \mid w, X) + \lambda \|w\|_2^2$ and the objective (1) instantiates as the negative expected (regularized) log likelihood $P(w) = \mathbb{E}[-\log \mathbb{P}(Y \mid w, X)] + \lambda \|w\|_2^2$. Define $Y_*(X) = \mathbb{P}(Y = 1 \mid w_*, X)$ and $\Sigma_* = \nabla^2 P(w_*) = \mathbb{E}[Y_*(X)(1 - Y_*(X))XX^\top] + \lambda I$. Analogous to the least-squares case, we can interpret $Y_*(X)$ as the conditional expectation of Y under the (possibly mis-specified) best fit model. With this notation, σ^2 is similar to its instantiation under regularized least-squares (Equation (17)):

$$\sigma^2 = \mathbb{E}\left[\frac{1}{2}\|(Y - Y_*(X))X + \lambda w_*\|_{\Sigma_*^{-1}}^2\right] \tag{19}$$

12

Under this definition of σ^2, by Theorem 2 together with the following defined quantities, the single-pass estimator of Algorithm 1 achieves a rate competitive with that of the ERM:

Corollary 6 (Logistic regression performance). *Suppose that $\|X\|^2 \leq L$. Define $\mu = \lambda$ and $M = \alpha \mathbb{E}[\|X\|^3_{(\nabla^2 P(w_*))^{-1}}]$. Under parameters from Theorem 1, we have*

$$\mathbb{E}[P(\widehat{w}_N) - P(w_*)] \leq \left(\left(1 + \tfrac{5}{b}\right) \tfrac{\sigma}{\sqrt{N}} + \left(2 + \tfrac{5}{b}\right) \tfrac{\sqrt{\kappa}\sigma}{\sqrt{N}} \min\left\{1, \left(\tfrac{N}{2(M\sigma+1)^2 k_0}\right)^{-p/2}\right\} + \sqrt{\tfrac{P(\widetilde{w}_0) - P(w_*)}{\left(\tfrac{N}{2k_0}\right)^{p+1}}} \right)^2$$

The corollary uses Lemma 10, a straightforward lemma to handle self-concordance for logistic regression, which is included for completeness. See [Bach(2010)] for techniques for analyzing the self-concordance of logistic regression.

4 Analysis of Streaming SVRG

Here we analyze Algorithm 1. Section 4.1 provides useful common lemmas. Section 4.2 uses these lemmas to characterize the behavior of the Algorithm 1. These are then used to prove convergence in terms of both α-bounded Hessians (Section 4.3) and M-self-concordance (Section 4.4).

4.1 Common lemmas

Our first lemma is a consequence of smoothness. It is the same observation made in [Johnson and Zhang(2013)].

Lemma 1. *If ψ is smooth (with probability one), then*

$$\mathbb{E}_{\psi \sim \mathcal{D}} \left[\|\nabla \psi(w) - \nabla \psi(w_*)\|^2 \right] \leq 2L \left(P(w) - P(w_*) \right) \tag{20}$$

Remark (A weaker smoothness assumption). Instead of the smoothness Assumption 1 in Equation 7, it suffices to directly assume (38) and still have all results hold as presented. In doing so, we incur an additional factor of 2 as in this case we have $\nabla^2 P(w_*) \preceq 2LI$ by Lemma 9. For further explanation see Appendix A.

Proof. For an L-smooth function $f : \mathbb{R}^d \to \mathbb{R}$, we have

$$f(w) - \min_{w'} f(w') \geq \frac{1}{2L} \|\nabla f(w)\|^2 \tag{21}$$

To see this, observe that

$$\min_{w'} f(w') \leq \min_{\eta} f(w - \eta \nabla f(w))$$

$$\leq \min_{\eta} \left(f(w) - \eta \|\nabla f(w)\|^2 + \frac{1}{2} \eta^2 L \|\nabla f(w)\|^2 \right) = f(w) - \frac{1}{2L} \|\nabla f(w)\|^2$$

using the definition of L-smoothness.

Now define:

$$g(w) = \psi(w) - \psi(w_*) - (w - w_*)^\top \nabla \psi(w_*) \tag{22}$$

Since ψ is L-smooth (with probability one) g is L-smooth (with probability one) and it follows that:

$$\|\nabla \psi(w) - \nabla \psi(w_*)\|^2 = \|\nabla g(w)\|^2$$

13

$$\leq 2L(g(w) - \min_{w'} g(w'))$$
$$\leq 2L(g(w) - g(w_*))$$
$$= 2L(\psi(w) - \psi(w_*) - (w - w_*)^\top \nabla \psi(w_*))$$

where the second step follows from smoothness. The proof is completed by taking expectations and noting that $\mathbb{E}[\nabla \psi(w_*)] = \nabla P(w_*) = 0$. $\qquad\square$

Our second lemma bounds the variance of $\psi \sim \mathcal{D}$ in the $(\nabla^2 P(w_*))^{-1}$ norm.

Lemma 2. *Suppose Assumption 1 holds. Let $w \in \mathbb{R}^d$ and let $\psi \sim \mathcal{D}$. Then*

$$\mathbb{E} \left\| \nabla \psi(w) - \nabla P(w) \right\|^2_{(\nabla^2 P(w_*))^{-1}} \leq 2 \left(\sqrt{\kappa \left(P(w) - P(w_*) \right)} + \sigma \right)^2 \tag{23}$$

Proof. For random vectors a and b, we have

$$\mathbb{E}\|a + b\|^2 = \mathbb{E}\|a\|^2 + 2\mathbb{E}a \cdot b + \mathbb{E}\|b\|^2 \leq \mathbb{E}\|a\|^2 + 2\sqrt{\mathbb{E}\|a\|^2 \mathbb{E}\|b\|^2} + \mathbb{E}\|b\|^2 = \left(\sqrt{\mathbb{E}\|a\|^2} + \sqrt{\mathbb{E}\|b\|^2} \right)^2$$

Consequently,

$$\mathbb{E} \left\| \nabla \psi(w) - \nabla P(w) \right\|^2_{(\nabla^2 P(w_*))^{-1}}$$
$$\leq \left(\sqrt{\mathbb{E} \left\| \nabla \psi(w) - \nabla \psi(w_*) - \nabla P(w) \right\|^2_{(\nabla^2 P(w_*))^{-1}}} + \sqrt{\mathbb{E} \left\| \nabla \psi(w_*) \right\|^2_{(\nabla^2 P(w_*))^{-1}}} \right)^2$$
$$\leq \left(\sqrt{\frac{1}{\mu} \mathbb{E} \left\| \nabla \psi(w) - \nabla \psi(w_*) - \nabla P(w) \right\|^2} + \sqrt{2}\sigma \right)^2$$

where the last step uses $\mu \mathrm{I} \preceq \nabla^2 P(w_*)$ and the definition of σ^2.

Observe that
$$\mathbb{E} \left[\nabla \psi(w) - \nabla \psi(w_*) \right] = \nabla P(w) - \nabla P(w_*) = \nabla P(w)$$

Applying Lemma 1 and for random a, that $\mathbb{E}\|a - \mathbb{E}a\|^2 \leq \mathbb{E}\|a\|^2$, we have

$$\mathbb{E} \left\| \nabla \psi(w) - \nabla \psi(w_*) - \nabla P(w) \right\|^2 \leq \mathbb{E} \left\| \nabla \psi(w) - \nabla \psi(w_*) \right\|^2 \leq 2L(P(w) - P(w_*))$$

Combining and using the definition of κ yields the result. $\qquad\square$

4.2 Progress of the algorithm

The following bounds the progress of one step of Algorithm 1.

Lemma 3. *Suppose Assumption 1 holds, $\widetilde{w}_0 \in \mathbb{R}^d$, and $\widetilde{\psi}_1, \ldots \widetilde{\psi}_k$ are functions from $\mathbb{R}^d \to \mathbb{R}$. Suppose $\psi_1, \ldots \psi_m$ are sampled independently from $\mathcal{D}$. Set $w_0 = \widetilde{w}_0$ and for $t \in \{0, 1, \ldots m - 1\}$, set:*

$$w_{t+1} \overset{\text{def}}{=} w_t - \frac{\eta}{L} \left(\nabla \psi_t(w_t) - \nabla \psi_t(\widetilde{w}_0) + \frac{1}{k} \sum_{i \in [k]} \nabla \widetilde{\psi}_i(\widetilde{w}_0) \right)$$

for some $\eta > 0$. Define:

$$\Delta \overset{\text{def}}{=} \frac{1}{k} \sum_{i \in [k]} \nabla \widetilde{\psi}_i(\widetilde{w}_0) - \nabla P(\widetilde{w}_0)$$

14

For all t let α_t be such that

$$P(w_*) \geq P(w_t) + (w_* - w_t)^\top \nabla P(w_t) + \frac{1}{2\alpha_t}\|w_t - w_*\|^2_{\nabla^2 P(w_*)} \tag{24}$$

(note that such an α_t exists by Assumption 1, as $\alpha_t \leq \kappa$).

Then for all t we have

$$\mathbb{E}L\|w_{t+1} - w_*\|^2 \leq \mathbb{E}\left[L\|w_t - w_*\|^2 - 2\eta(1 - 4\eta)\left(P(w_t) - P(w_*)\right) + 8\eta^2\left(P(\widetilde{w}_0) - P(w_*)\right)\right.$$
$$\left. + \left(\alpha_t\eta + 2\eta^2\right)\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}\right] \tag{25}$$

Proof. Letting

$$g_t(w) = \psi_t(w) - w^\top\left(\nabla\psi_t(\widetilde{w}_0) - \frac{1}{k}\sum_{i\in[k]}\nabla\widetilde{\psi}_i(\widetilde{w}_0)\right)$$

and recalling the definition of w_{t+1} and Δ we have

$$
\begin{aligned}
\mathbb{E}_{\psi_t\sim\mathcal{D}}\|w_{t+1} - w_*\|^2 &= \mathbb{E}_{\psi_t\sim\mathcal{D}}\|w_t - w_* - \frac{\eta}{L}\nabla g_t(w_t)\|^2 \\
&= \mathbb{E}_{\psi_t\sim\mathcal{D}}\left[\|w_t - w_*\|^2 - 2\frac{\eta}{L}(w_t - w_*)^\top\nabla g_t(w_t) + \frac{\eta^2}{L^2}\|\nabla g_t(w_t)\|^2\right] \\
&= \|w_t - w_*\|^2 - 2\frac{\eta}{L}(w_t - w_*)^\top(\nabla P(w_t) + \Delta) + \frac{\eta^2}{L^2}\mathbb{E}_{\psi_t\sim\mathcal{D}}\|\nabla g_t(w_t)\|^2 \quad (26)
\end{aligned}
$$

Now by (24) we know that

$$-2(w_t - w_*)^\top\nabla P(w_t) \leq -2(P(w_t) - P(w_*)) - \frac{1}{\alpha_t}\|w_t - w_*\|^2_{\nabla^2 P(w_*)} \tag{27}$$

Using Cauchy-Schwarz and that $2a\cdot b \leq a^2 + b^2$ for scalar a and b, we have

$$-2(w_t - w_*)^\top\Delta \leq \frac{1}{\alpha_t}\|w_t - w_*\|^2_{\nabla^2 P(w_*)} + \alpha_t\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}} \tag{28}$$

Furthermore

$$
\begin{aligned}
\mathbb{E}_{\psi_t\sim\mathcal{D}}\|\nabla g_t(w_t)\|^2 &= \mathbb{E}_{\psi_t\sim\mathcal{D}}\left\|\nabla\psi_t(w_t) - \nabla\psi_t(\widetilde{w}_0) + \frac{1}{k}\sum_{i\in[k]}\nabla\widetilde{\psi}_i(\widetilde{w}_0)\right\|^2 \\
&= \mathbb{E}_{\psi_t\sim\mathcal{D}}\|(\nabla\psi_t(w_t) - \nabla\psi_t(w_*)) - (\nabla\psi_t(\widetilde{w}_0) - \nabla\psi_t(w_*) - \nabla P(\widetilde{w}_0)) + \Delta\|^2 \\
&\leq 2\mathbb{E}_{\psi_t\sim\mathcal{D}}\|(\nabla\psi_t(w_t) - \nabla\psi_t(w_*)) - (\nabla\psi_t(\widetilde{w}_0) - \nabla\psi_t(w_*) - \nabla P(\widetilde{w}_0))\|^2 + 2\|\Delta\|^2 \\
&\leq 4\mathbb{E}_{\psi_t\sim\mathcal{D}}\|\nabla\psi_t(w_t) - \nabla\psi_t(w_*)\|^2 + 4\mathbb{E}_{\psi_t\sim\mathcal{D}}\|\nabla\psi_t(\widetilde{w}_0) - \nabla\psi_t(w_*) - \nabla P(\widetilde{w}_0)\|^2 + 2\|\Delta\|^2 \\
&\leq 4\mathbb{E}_{\psi_t\sim\mathcal{D}}\|\nabla\psi_t(w_t) - \nabla\psi_t(w_*)\|^2 + 4\mathbb{E}_{\psi_t\sim\mathcal{D}}\|\nabla\psi_t(\widetilde{w}_0) - \nabla\psi_t(w_*)\|^2 + 2\|\Delta\|^2
\end{aligned}
$$

where we have used that $\mathbb{E}[\nabla\psi_t(\widetilde{w}_0) - \nabla\psi_t(w_*) - \nabla P(\widetilde{w}_0)] = 0$ and $\mathbb{E}\|a - \mathbb{E}a\|^2 \leq \mathbb{E}\|a\|^2$. Applying Lemma 1 and using $\nabla^2 P(w_*) \preceq LI$ yields

$$\mathbb{E}_{\psi_t\sim\mathcal{D}}\|\nabla g_t(w_t)\|^2 \leq 8L\left(P(w_t) - P(w_*)\right) + 8L\left(P(\widetilde{w}_0) - P(w_*)\right) + 2L\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}} \tag{29}$$

Combining (26), (27), (28), and (29) yields

$$\mathbb{E}_{\psi_t \sim \mathcal{D}} \|w_{t+1} - w_*\|^2 \leq \|w_t - w_*\|^2 - 2\frac{\eta}{L}(1 - 4\eta)(P(w_t) - P(w_*)) + 8\frac{\eta^2}{L}(P(\widetilde{w}_0) - P(w_*))$$
$$+ \left(\alpha_t \frac{\eta}{L} + 2\frac{\eta^2}{L}\right) \|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}$$

and multiplying both sides by L yields the result. □

Finally we bound the progress of one stage of Algorithm 1.

Lemma 4. *Under the same assumptions as Lemma 3, for $\widetilde{m}$ chosen uniformly at random in $\{1, \ldots m\}$ and $\widetilde{w}_1 \stackrel{\text{def}}{=} w_{\widetilde{m}}$, we have*

$$\mathbb{E}[P(\widetilde{w}_1) - P(w_*)] \leq \frac{1}{1 - 4\eta}\left[\left(\frac{\kappa}{m\eta} + 4\eta\right)P(\widetilde{w}_0) - P(w_*) + \mathbb{E}\left[\frac{\alpha_{\widetilde{m}} + 2\eta}{2}\right]\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}\right]$$

where we are conditioning on $\widetilde{w}_0$ and $\widetilde{\psi}_1, \ldots \widetilde{\psi}_k$.

Proof. Taking an unconditional expectation with respect to $\{\psi_t\}$ and summing (25) from Lemma 3 from $t = m - 1$ down to $t = 0$ yields

$$L \cdot \mathbb{E}\|w_m - w_*\|^2 \leq L \cdot \|\widetilde{w}_0 - w_*\|^2 - 2\eta(1 - 4\eta)\sum_{t=0}^{m-1}\mathbb{E}(P(w_t) - P(w_*))$$
$$8m\eta^2(\mathbb{E}P(\widetilde{w}_0) - P(w_*)) + \sum_{t=0}^{m-1}\mathbb{E}\left[(\alpha_t\eta + 2\eta^2)\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}\right]$$

By strong convexity,

$$\|\widetilde{w}_0 - w_*\|^2 \leq \frac{2}{\mu}(P(\widetilde{w}_0) - P(w_*))$$

and a little manipulation yields that:

$$\frac{2\eta(1 - 4\eta)}{m}\sum_{t=0}^{m-1}\mathbb{E}(P(w_t) - P(w_*)) \leq \left(\frac{2\kappa}{m} + 8\eta^2\right)(P(\widetilde{w}_0) - P(w_*))$$
$$+ \sum_{t=0}^{m-1}\mathbb{E}\left[\frac{\alpha_t\eta + 2\eta^2}{m}\mathbb{E}\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}\right]$$

Rearranging terms and applying the definition of $\widetilde{w}_1$ then yields the result. □

4.3   With α-bounded Hessians

Here we prove the progress made by Algorithm 1 in a single stage under only Assumption 1.

Theorem 1 (Stage progress with α-bounded Hessians). *Under Assumption 1, for Algorithm 1, we have for all s:*

$$\mathbb{E}[P(\widetilde{w}_{s+1}) - P(w_*)]$$
$$\leq \frac{1}{1 - 4\eta}\left[\left(\frac{\kappa}{m\eta} + 4\eta\right)\mathbb{E}[P(\widetilde{w}_s) - P(w_*)] + \frac{\alpha + 2\eta}{k}\left(\sqrt{\kappa \cdot \mathbb{E}[P(\widetilde{w}_s) - P(w_*)]} + \sigma\right)^2\right]$$

16

Proof. By definition of α, we have $\alpha_t \leq \alpha$ for all t in Lemma 4 and therefore

$$\mathbb{E}[P(\widetilde{w}_{s+1}) - P(w_*)] \leq \frac{1}{1-4\eta}\left[\left(\frac{\kappa}{m\eta} + 4\eta\right)\mathbb{E}[P(\widetilde{w}_s) - P(w_*)] + \frac{\alpha+2\eta}{2}\mathbb{E}\left[\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}\right]\right]$$

Now using that the $\widetilde{\psi}_i$ are independent and that $\mathbb{E}[\nabla\widetilde{\psi}_i(\widetilde{w}_s)] = \nabla P(\widetilde{w}_s)$ we have

$$\begin{aligned}
\mathbb{E}[\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}] &= \frac{1}{k}\mathbb{E}_{\psi\sim\mathcal{D}}\left[\|\nabla\widetilde{\psi}_1(\widetilde{w}_s) - P(\widetilde{w}_s)\|^2_{(\nabla^2 P(w_*))^{-1}}\right] \\
&\leq \frac{2}{k}\mathbb{E}\left[\kappa(P(\widetilde{w}_s) - P(w_*)) + \sigma\sqrt{\kappa(P(\widetilde{w}_s) - P(w_*))} + \sigma^2\right] \\
&\leq \frac{2}{k}\left[\kappa\mathbb{E}[P(\widetilde{w}_s) - P(w_*)] + \sigma\sqrt{\kappa\mathbb{E}[P(\widetilde{w}_s) - P(w_*)]} + \sigma^2\right] \\
&= \frac{2}{k}\left(\sqrt{\kappa \cdot \mathbb{E}[P(\widetilde{w}_s) - P(w_*)]} + \sigma\right)^2
\end{aligned}$$

where we have also used Lemma 2 and Jensen's inequality. $\qquad\square$

4.4 With M-self-concordance

Our main result in the self-concordant case follows.

Theorem 2 (Convergence under self-concordance). *Suppose Assumption 1 and 2 hold. Under Algorithm 1, for $\eta \leq \frac{1}{8}$, $k \geq 10C$, and all s, we have*

$$\begin{aligned}
\mathbb{E}[P(\widetilde{w}_{s+1}) - P(w_*)] \leq \frac{1}{1-4\eta}&\left[\left(\frac{\kappa}{m\eta} + 4\eta\right)\mathbb{E}[P(\widetilde{w}_s) - P(w_*)]\right.\\
&\left.+ \frac{1}{k}\left((2M\sigma\kappa + 9\kappa)\sqrt{\mathbb{E}[P(\widetilde{w}_s) - P(w_*)]} + \left(1 + 2\sqrt{\eta} + \frac{10M\sigma\kappa}{\sqrt{k}}\right)\sigma\right)^2\right]
\end{aligned}$$

The proof utilizes the following lemmas. First, we show how self concordance implies that there is a better effective strong convexity parameter in $\nabla^2 P(w_*)$ norm when we are close to w_*.

Lemma 5. *If P is M-self-concordant, then*

$$P(w_*) \geq P(w_t) + (w_* - w_t)^\top\nabla P(w_t) + \frac{\|w_t - w_*\|^2_{\nabla^2 P(w_*)}}{2(1 + M\|w_t - w_*\|_{\nabla^2 P(w_*)})^2}. \tag{30}$$

Proof. First we use the property of self-concordant functions: if f is M-self-concordant, then

$$f(t) \geq f(0) + tf'(0) + \frac{4}{M^2}\left(t\frac{M}{2}\sqrt{f''(0)} - \ln\left(1 + t\frac{M}{2}\sqrt{f''(0)}\right)\right)$$

Apply this property to the function P restricted to the line between w_t and w_*, where the 0 point is at w_t and t is $\|w_t - w_*\|_{\nabla^2 P(w_t)}$, then we have

$$P(w_*) \geq P(w_t) + (w_* - w_t)^\top\nabla P(w_t) + \frac{4}{M^2}\left(\frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_t)} - \ln\left(1 + \frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_t)}\right)\right)$$

In order to convert $\nabla^2 P(w_t)$ norm to $\nabla^2 P(w_*)$ norm, we use another property of self-concordant function:

$$f''(t) \geq \frac{f''(0)}{(1 + t\frac{M}{2}\sqrt{f''(0)})^2}$$

17

Again we restrict to the line between w_* and w_t, where 0 point corresponds to w_*, and t is $\|w_t - w_*\|$, and we get

$$\|w_t - w_*\|^2_{\nabla^2 P(w_t)} \geq \frac{\|w_t - w_*\|^2_{\nabla^2 P(w_*)}}{(1 + \frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_*)})^2}$$

Now consider the function let $h(x) = x - \ln(1 + x)$. The function has the following two properties: When $x \geq 0$, $h(x)$ is monotone and $h(x) \geq x^2/2(1 + x)$. This claim can be verified directly by taking derivatives.

Therefore

$$h\left(\frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_t)}\right) \geq h\left(\frac{\frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_*)}}{(1 + \frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_*)})}\right)$$

$$\geq \frac{\frac{M^2}{4}\|w_t - w_*\|^2_{\nabla^2 P(w_*)}}{(1 + \frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_*)})^2} \cdot \frac{1}{2\left(1 + \frac{\frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_*)}}{(1 + \frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_*)})}\right)}$$

$$= \frac{\frac{M^2}{4}\|w_t - w_*\|^2_{\nabla^2 P(w_*)}}{2(1 + \frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_*)})(1 + M\|w_t - w_*\|_{\nabla^2 P(w_*)})}$$

$$\geq \frac{M^2\|w_t - w_*\|^2_{\nabla^2 P(w_*)}}{8(1 + M\|w_t - w_*\|_{\nabla^2 P(w_*)})^2}$$

This concludes the proof. □

Essentially, this means when $\|w_t - w_*\|^2_{\nabla^2 P(w_*)}$ is small the effective strong convexity in $\|\cdot\|_{\nabla^2 P(w_*)}$ is small. In particular,

$$\alpha_t \leq \min\left\{\alpha, \left(1 + \frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_*)}\right)^2\right\} \leq \min\left\{\kappa, \left(1 + \frac{M}{2}\|w_t - w_*\|_{\nabla^2 P(w_*)}\right)^2\right\}$$

Thus we need to bound the residual error $\|w_t - w_*\|^2_{\nabla^2 P(w_*)}$.

Lemma 6 (Crude residual error bound). *Suppose the same assumptions in Lemma 3 hold and that $\eta \leq \frac{1}{8}$. Then for all t, we have*

$$\mathbb{E}\|w_t - w_*\|^2_{\nabla^2 P(w_*)} \leq 3\kappa(P(\widetilde{w}_0) - P(w_*)) + 6\kappa^2 \|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}$$

Proof. Since $\alpha_t \leq \kappa$ and by Lemma 3 we have

$$\mathbb{E}L\|w_{t+1} - w_*\|^2 \leq \mathbb{E}\left[L\|w_t - w_*\|^2 - 2\eta(1 - 4\eta)(P(w_t) - P(w_*)) + 8\eta^2(P(\widetilde{w}_0) - P(w_*))\right.$$
$$\left. + \left(\kappa\eta + 2\eta^2\right)\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}\right]$$

Using that by strong convexity $P(w_t) - P(w_*) \geq \frac{\mu}{2}\|w_t - w_*\|^2$ we have

$$\mathbb{E}L\|w_{t+1} - w_*\|^2 \leq \mathbb{E}\left[\left(1 - \frac{\eta}{2\kappa}\right)L\|w_t - w_*\|^2\right] + \eta(P(\widetilde{w}_0) - P(w_*)) + 2\eta\kappa\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}$$

Solving for the maximum value of $L\|w_t - w_*\|^2_2$ in this recurrence we have, for all t,

$$\mathbb{E}L\|w_t - w_*\|^2 \leq \frac{3\kappa}{\eta}\left(\eta(P(\widetilde{w}_0) - P(w_*)) + 2\eta\kappa\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}\right)$$

Using that $\nabla^2 P(w_*) \preceq LI$ yields the result. □

18

Finally, we end up needing to bound higher moments of the error from Δ. For this we provide two technical lemmas.

Lemma 7. *Suppose Assumption 1 and 2 hold. For $\widetilde{\psi}_i$ sampled independently, we have*

$$\mathbb{E}\left\|\frac{1}{k}\sum_{i\in k}\widetilde{\psi}_i(w_*)\right\|^4_{(\nabla^2 P(w_*))^{-1}} \leq 12\left(1+\frac{C}{k}\right)\left(\frac{\sigma^2}{k}\right)^2$$

Proof. By Assumption 2 we have

$$\mathbb{E}\left\|\frac{1}{k}\sum_{i\in k}\widetilde{\psi}_i(w_*)\right\|^4_{(\nabla^2 P(w_*))^{-1}}$$

$$=\frac{1}{k^4}\left[k\left(\mathbb{E}_{\psi\sim\mathcal{D}}\|\nabla\psi(w_*)\|^4_{(\nabla^2 P(w_*))^{-1}}\right)+3k(k-1)\left(\mathbb{E}_{\psi\sim\mathcal{D}}\|\nabla\psi_i(w_*)\|^2_{(\nabla^2 P(w_*))^{-1}}\right)^2\right]$$

$$\leq\frac{3k(k-1)+Ck}{k^4}\left(\mathbb{E}_{\psi\sim\mathcal{D}}\|\nabla\psi_i(w_*)\|^2_{(\nabla^2 P(w_*))^{-1}}\right)^2$$

Recalling the definition of σ^2 yields the result. $\qquad\square$

Lemma 8. *Suppose a is a random variable such that $\mathbb{E}[a^4]\leq\widetilde{C}\cdot(\mathbb{E}[a^2])^2$, b is a random variable, and c is a constant. We have*

$$\mathbb{E}[a^2\min\{b^2,c\}]\leq 2\mathbb{E}[a^2]\sqrt{\widetilde{C}\cdot c\cdot\mathbb{E}[b^2]} \tag{31}$$

Proof. Let E_1 be the indicator variable for the event $a^2\geq T\mathbb{E}[a^2]$ where T is chosen later. Let $E_2=1-E_1$.

On one hand, we have $\mathbb{E}[a^2 E_1]T\mathbb{E}[a^2]\leq\mathbb{E}[a^4]$, therefore $\mathbb{E}[a^2 E_1]\leq\frac{\widetilde{C}}{T}\mathbb{E}[a^2]$. On the other hand, $\mathbb{E}[\min\{b^2,c\}a^2 E_2]\leq\mathbb{E}[b^2 a^2 E_2]\leq T\mathbb{E}[a^2]\mathbb{E}[b^2]$. Combining these two cases we have:

$$\mathbb{E}[\min\{b^2,c\}a^2]=\mathbb{E}[\min\{b^2,c\}a^2 E_1]+\mathbb{E}[\min\{b^2,c\}a^2 E_2]$$

$$\leq c\mathbb{E}[a^2 E_1]+\mathbb{E}[b^2 a^2 E_2]$$

$$\leq\frac{c\cdot\widetilde{C}}{T}\mathbb{E}[a^2]+T\mathbb{E}[a^2]\mathbb{E}[b^2]$$

$$=2\mathbb{E}[a^2]\sqrt{c\cdot\widetilde{C}\mathbb{E}[b^2]}$$

In the last step we chose $T=\sqrt{\frac{c\cdot\widetilde{C}}{\mathbb{E}[b^2]}}$ to balance the terms. $\qquad\square$

Using these lemmas, we are ready to provide the proof.

Proof. **of Theorem 2.** We analyze stage s of the algorithm. Let us define the variance term (A) as

$$(A)=\mathbb{E}\left[\left(\frac{\alpha_{\widetilde{m}}+2\eta}{2}\right)\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}\right]$$

Our main goal in the proof is to bound (A). First, for all α,x,y and positive semidefinite H we have

$$\mathbb{E}\alpha\|x+y\|^2_H=\mathbb{E}\left[\|H^{-1/2}\sqrt{\alpha}x+H^{-1/2}\sqrt{\alpha}y\|^2_2\right]$$

19

$$\leq \left(\sqrt{\mathbb{E}\|\sqrt{\alpha}H^{-1/2}x\|_2^2} + \sqrt{\mathbb{E}\|\sqrt{\alpha}H^{-1/2}x\|_2^2} \right)^2$$

$$\leq \left(\sqrt{\mathbb{E}\alpha\|x\|_H^2} + \sqrt{\mathbb{E}\alpha\|y\|_H^2} \right)^2 . \tag{32}$$

By the definition of Δ we have

$$(A) \leq \left(\sqrt{(B)} + \sqrt{(C)} \right)^2$$

where (B) and (C) are defined below. Using that $\mathbb{E}\|a - \mathbb{E}[a]\|_H^2 \leq \mathbb{E}\|a\|_H^2$, Lemma 1, and the strong convexity of P we have

$$(B) = \mathbb{E}\left(\frac{\alpha_{\widetilde{m}} + 2\eta}{2} \right) \left\| \frac{1}{k} \sum_{i \in k} \nabla\widetilde{\psi}_i(\widetilde{w}_s) - \nabla\widetilde{\psi}_i(w_*) - \nabla P(\widetilde{w}_s) \right\|_{(\nabla^2 P(w_*))^{-1}}^2$$

$$\leq \left(\frac{\kappa + 2\eta}{2} \right) \cdot \frac{2\kappa}{k} \cdot \mathbb{E}[P(\widetilde{w}_s) - P(w_*)]$$

$$\leq \frac{2\kappa^2}{k} \cdot \mathbb{E}[P(\widetilde{w}_s) - P(w_*)]$$

We use that $\min\{a, b + c\} \leq b + \min\{a, c\}$ (for positive a, b, and c) by Lemma 1, the definition of σ^2, as well as (32)

$$(C) = \mathbb{E}\left(\frac{\alpha_{\widetilde{m}} + 2\eta}{2} \right) \left\| \frac{1}{k} \sum_{i \in k} \nabla\widetilde{\psi}_i(w_*) \right\|_{(\nabla^2 P(w_*))^{-1}}^2$$

$$\leq \frac{2\eta\sigma^2}{k} + \mathbb{E}\left[\frac{\min\{\kappa, (1 + M\|w_t - w_*\|_{\nabla^2 P(w_*)})^2\}}{2} \left\| \frac{1}{k} \sum_{i \in k} \nabla\widetilde{\psi}_i(w_*) \right\|_{(\nabla^2 P(w_*))^{-1}}^2 \right]$$

$$= \frac{2\eta\sigma^2}{k} + \mathbb{E}\left[\left\| \frac{\min\{\sqrt{\kappa}, 1 + M\|w_t - w_*\|_{\nabla^2 P(w_*)}\}}{\sqrt{2}k} \sum_{i \in k} \nabla\widetilde{\psi}_i(w_*) \right\|_{(\nabla^2 P(w_*))^{-1}}^2 \right]$$

$$\leq \frac{2\eta\sigma^2}{k} + \mathbb{E}\left[\left\| \frac{1}{\sqrt{2}k} \sum_{i \in k} \nabla\widetilde{\psi}_i(w_*) + \frac{\min\{\sqrt{\kappa}, M\|w_t - w_*\|_{\nabla^2 P(w_*)}\}}{\sqrt{2}k} \sum_{i \in k} \nabla\widetilde{\psi}_i(w_*) \right\|_{(\nabla^2 P(w_*))^{-1}}^2 \right]$$

$$\leq \frac{2\eta\sigma^2}{k} + \left(\sqrt{\frac{\sigma^2}{k}} + \sqrt{(D)} \right)^2$$

where (D) is defined below. Using Lemma 6 and the independence of the different types of ψ

$$(D) = \mathbb{E}\left[\frac{\min\{\kappa, M^2\|w_t - w_*\|_{\nabla^2 P(w_*)}^2\}}{2} \left\| \frac{1}{k} \sum_{i \in k} \nabla\widetilde{\psi}_i(w_*) \right\|_{(\nabla^2 P(w_*))^{-1}}^2 \right]$$

$$\leq \mathbb{E}\left[\frac{\min\left\{ \kappa, M^2 \left(3\kappa \cdot P(\widetilde{w}_s) - P(w_*) + 6\kappa^2\|\Delta\|_{(\nabla^2 P(w_*))^{-1}}^2 \right) \right\}}{2} \left\| \frac{1}{k} \sum_{i \in k} \nabla\widetilde{\psi}_i(w_*) \right\|_{(\nabla^2 P(w_*))^{-1}}^2 \right]$$

$$\leq \frac{3\kappa^2 M^2\sigma^2}{k} \mathbb{E}[P(\widetilde{w}_s) - P(w_*)] + \frac{\kappa}{2}(E)$$

20

where (E) is defined below. Using kurtosis,

$$\mathbb{E}\left[\left\|\frac{1}{k}\sum_{i\in k}\widetilde{\psi}_i(w_*)\right\|^4_{(\nabla^2 P(w_*))^{-1}}\right] \le 14(\sigma^2/k)^2$$

By Lemma 7 and applying Lemma 8 we have

$$(E) \le \mathbb{E}\left[\min\left\{1, 6\kappa M^2\|\Delta\|^2_{(\nabla^2 P(w_*))^{-1}}\right\}\left\|\frac{1}{k}\sum_{i\in k}\nabla\widetilde{\psi}_i(w_*)\right\|^2_{(\nabla^2 P(w_*))^{-1}}\right]$$

$$\le \mathbb{E}\left[\min\left\{1, 12\kappa M^2\left(\left\|\frac{1}{k}\sum_{i\in k}\nabla\widetilde{\psi}_i(\widetilde{w}_s) - \nabla\widetilde{\psi}_i(w_*) - \nabla P(\widetilde{w}_s)\right\|^2_{(\nabla^2 P(w_*))^{-1}} + \left\|\frac{1}{k}\sum_{i\in k}\nabla\widetilde{\psi}_i(w_*)\right\|^2_{(\nabla^2 P(w_*))^{-1}}\right)\right\}\right.$$

$$\left. \cdot\left\|\frac{1}{k}\sum_{i\in k}\nabla\widetilde{\psi}_i(w_*)\right\|^2_{(\nabla^2 P(w_*))^{-1}}\right]$$

$$\le \mathbb{E}\left[\min\left\{1, 12\kappa M^2\left\|\frac{1}{k}\sum_{i\in k}\nabla\widetilde{\psi}_i(\widetilde{w}_s) - \nabla\widetilde{\psi}_i(w_*) - \nabla P(\widetilde{w}_s)\right\|^2_{(\nabla^2 P(w_*))^{-1}}\right\}\cdot\left\|\frac{1}{k}\sum_{i\in k}\nabla\widetilde{\psi}_i(w_*)\right\|^2_{(\nabla^2 P(w_*))^{-1}}\right]$$

$$+ 170\kappa M^2\left(\frac{\sigma^2}{k}\right)^2$$

$$\le \frac{4\sigma^2}{k}\sqrt{14\cdot 1\cdot 24\kappa^2 M^2\frac{P(\widetilde{w}_s) - P(w_*)}{k}} + 170\kappa M^2\left(\frac{\sigma^2}{k}\right)^2$$

$$\le 2\frac{4\sqrt{\kappa}M\sigma^2}{k}\sqrt{96\kappa\frac{P(\widetilde{w}_s) - P(w_*)}{k}} + 170\kappa M^2\left(\frac{\sigma^2}{k}\right)^2 \qquad \text{by manipulation of constants}$$

$$\le 16\kappa M^2\left(\frac{\sigma^2}{k}\right)^2 + \frac{96\kappa[P(\widetilde{w}_s) - P(w_*)]}{k} + 170\kappa M^2\left(\frac{\sigma^2}{k}\right)^2 \qquad \text{since } 2a\cdot b \le a^2 + b^2$$

$$\le 200 M^2\kappa\left(\frac{\sigma^2}{k}\right)^2 + \frac{96\kappa[P(\widetilde{w}_s) - P(w_*)]}{k}$$

Using that $\sqrt{|x| + |y|} \le \sqrt{|x|} + \sqrt{|y|}$ this implies

$$(A) \le \left(\sqrt{\frac{2\kappa^2}{k}\cdot\mathbb{E}[P(\widetilde{w}_s) - P(w_*)]} + \sqrt{\frac{2\eta\sigma^2}{k} + \left(\sqrt{\frac{\sigma^2}{k}} + \sqrt{(D)}\right)^2}\right)^2$$

$$\le \left(\frac{2\kappa}{\sqrt{k}}\sqrt{\mathbb{E}[P(\widetilde{w}_s) - P(w_*)]} + \frac{2\sigma\sqrt{\eta}}{\sqrt{k}} + \frac{\sigma}{\sqrt{k}} + \sqrt{(D)}\right)^2$$

$$\le \left(\frac{2\kappa}{\sqrt{k}}\sqrt{\mathbb{E}[P(\widetilde{w}_s) - P(w_*)]} + \frac{2\sigma\sqrt{\eta}}{\sqrt{k}} + \frac{\sigma}{\sqrt{k}}\right.$$

$$\left.+ \sqrt{\frac{3\kappa^2 M^2\sigma^2}{k}\mathbb{E}[P(\widetilde{w}_s) - P(w_*)] + \frac{\kappa}{2}\left(200 M^2\kappa\left(\frac{\sigma^2}{k}\right)^2 + \frac{96\kappa[P(\widetilde{w}_s) - P(w_*)]}{k}\right)}\right)^2$$

21

$$\leq \frac{1}{k} \left((2\kappa + 2M\sigma\kappa + 7\kappa) \sqrt{\mathbb{E}[P(\widetilde{w}_s) - P(w_*)]} + \left(1 + 2\sqrt{\eta} + \frac{10M\sigma\kappa}{\sqrt{k}} \right) \sigma \right)^2$$

Using this bound in Lemma 4 then yields the result. $\qquad\qquad\square$

5   Empirical risk minimization (M-estimation) for smooth functions

We now provide finite-sample rates for the ERM. We take the domain $\mathcal{S}$ to be compact in (1) (see Remark). Throughout this section, define:

$$\|A\|_* = \|(\nabla^2 P(w_*))^{-1/2} \cdot A \cdot (\nabla^2 P(w_*))^{-1/2}\|$$

for a matrix A (of appropriate dimensions).

Theorem 3. *Suppose $\psi_1, \psi_2, \ldots$ are an independently drawn sample from $\mathcal{D}$. Assume:*

1. *(Convexity of ψ) Assume that ψ is convex (with probability one).*

2. *(Smoothness of ψ) Assume that ψ is smooth in the following sense: the first, second, and third derivatives exist at all interior points of $\mathcal{S}$ (with probability one).*

3. *(Regularity Conditions) Suppose*

   (a) *$\mathcal{S}$ is compact (so $P(w)$ is bounded on $\mathcal{S}$).*
   (b) *w_* is an interior point of $\mathcal{S}$.*
   (c) *$\nabla^2 P(w_*)$ is positive definite (and, thus, is invertible).*
   (d) *There exists a neighborhood B of w_* and a constant L_3, such that (with probability one) $\nabla^2 \psi(w)$ is L_3-Lipschitz, namely $\|\nabla^2 \psi(w) - \nabla^2 \psi(w')\|_* \leq L_3 \|w - w'\|_{\nabla^2 P(w_*)}$, for w, w' in this neighborhood.*

4. *(Concentration at w_*) Suppose $\|\nabla \psi(w_*)\|_{(\nabla^2 P(w_*))^{-1}} \leq L_1$ and $\|\nabla^2 \psi(w_*)\|_* \leq L_2$ hold with probability one. Suppose the dimension d is finite (or, in the infinite dimensional setting, the intrinsic dimension is bounded, as in Remark).*

   Then:

$$\lim_{N \to \infty} \frac{\mathbb{E}[P(\widehat{w}_N^{ERM}) - P(w_*)]}{\sigma^2/N} = 1$$

In particular, the following lower and upper bounds hold. Define

$$\varepsilon_N := c \left(L_1 L_3 + \sqrt{L_2} \right) \sqrt{\frac{p \log dN}{N}}$$

where c is an appropriately chosen universal constant. Also, let c' be another appropriately chosen universal constant. We have that for all $p \geq 2$, if N is large enough so that $\sqrt{\frac{p \log dN}{N}} \leq c' \min \left\{ \frac{1}{\sqrt{L_2}}, \frac{1}{L_1 L_3}, \frac{1 \cdot diameter(B)}{L_1} \right\}$, then

$$(1 - \varepsilon_N) \frac{\sigma^2}{N} - \frac{\sqrt{\mathbb{E}[Z^4]}}{N^{p/2}} \leq \mathbb{E}[P(\widehat{w}_N^{ERM}) - P(w_*)]$$

$$\leq \quad (1 + \varepsilon_N)\frac{\sigma^2}{N} + \frac{\max_{w \in \mathcal{S}}\left(P(w) - P(w_*)\right)}{N^p}$$

where $Z = \left\|\nabla \widehat{P}_N(w_*)\right\|_{(\nabla^2 P(w_*))^{-1}}$ and so $\sqrt{\mathbb{E}[Z^4]} \leq L_1^2$. The lower bound above holds even if $\mathcal{S}$ is not compact.

Remark (Infinite dimensional setting). Define $M = \nabla^2 \psi(w_*) - \nabla^2 P(w_*)$ and $\widetilde{d} = \frac{\mathrm{Tr}(\mathbb{E}M^2)}{\lambda_{\max}(\mathbb{E}M^2)}$, which we assume to be finite. Here can replace d with $\widetilde{d}$ in the theorems. See Lemma 12.

Remark (Compactness of $\mathcal{S}$). The lower bound holds even if $\mathcal{S}$ is not compact. For the upper bound, the proof technique uses the compactness of $\mathcal{S}$ to bound the contribution to the expected regret due to a (low probability) failure event that the ERM may not lie in the ball B (or even the interior of $\mathcal{S}$). If P is regularized then this last term can be improved, as $\mathcal{S}$ need not be compact.

The basic idea of the proof follows that of [Hsu et al.(2014)Hsu, Kakade, and Zhang], along with various arguments based on Taylor's theorem.

Proof. Throughout the proof use $\widehat{w}_N$ to denote the ERM $\widehat{w}_N^{\mathrm{ERM}}$. Define:

$$\widehat{P}_N(w) = \frac{1}{N}\sum_{i=1}^{N}\psi_i(w)$$

which is convex as it is the average of convex functions.

Throughout the proof we take $t = cp\log(dN)$ in the tail probability bounds in Appendix D (for some universal constant c). This implies a probability of error less than $\frac{1}{N^p}$.

For all $w \in B$, the empirical function $\nabla^2 \widehat{P}_N(w)$ is L_3-Lipschitz. In Lemma 12 in Appendix D, we may take $v \leq 2\sqrt{L_2}$ (as all eigenvalues of of $\nabla^2 P(w_*)$ are one, under the choice of norm). Using Lemma 12 in Appendix D, for $w \in B$, we have:

$$\|\nabla^2 \widehat{P}_N(w) - \nabla^2 P(w_*)\|_* \leq \|\nabla^2 \widehat{P}_N(w) - \nabla^2 \widehat{P}_N(w_*)\|_* + \|\nabla^2 \widehat{P}_N(w_*) - \nabla^2 P(w_*)\|_*$$

$$\leq L_3\|w - w_*\|_{\nabla^2 P(w_*)} + c\sqrt{\frac{L_2 p \log dN}{N}} \tag{33}$$

for some (other) universal constant c. Now we seek to ensure that $\widehat{P}_N(w)$ is a constant spectral approximation to $\nabla^2 P(w_*)$. By choosing a sufficiently smaller ball B_1 (choose B_1 to have radius of $\min\{1/(10L_3), \mathrm{diameter}(B)\}$), the first term can be made small for $w \in B_1$. Also, for sufficiently large N, the second term can be made arbitrarily small (smaller than $1/10$), which occurs if $\sqrt{\frac{p\log dN}{N}} \leq \frac{c'}{\sqrt{L_2}}$. Hence, for such large enough N, we have for $w \in B_1$:

$$\frac{1}{2}\nabla^2 \widehat{P}_N(w) \preceq \nabla^2 P(w_*) \preceq 2\nabla^2 \widehat{P}_N(w) \tag{34}$$

Suppose N is at least this large from now on.

Now let us show that $\widehat{w}_N \in B_1$, with high probability, for N sufficiently large. By Taylor's theorem, for all w in the interior of $\mathcal{S}$, there exists a $\widetilde{w}$, between w_* and w, such that:

$$\widehat{P}_N(w) = \widehat{P}_N(w_*) + \nabla \widehat{P}_N(w_*)^\top(w - w_*) + \frac{1}{2}(w - w_*)^\top \nabla^2 \widehat{P}_N(\widetilde{w})(w - w_*)$$

23

Hence, for all $w \in B_1$ and if Equation 34 holds,

$$
\begin{aligned}
\widehat{P}_N(w) - \widehat{P}_N(w_*) &= \nabla\widehat{P}_N(w_*)^\top(w - w_*) + \frac{1}{2}\|w - w_*\|^2_{\nabla^2 P(\widetilde{w})} \\
&\geq \nabla\widehat{P}_N(w_*)^\top(w - w_*) + \frac{1}{4}\|w - w_*\|^2_{\nabla^2 P(w_*)} \\
&\geq \|w - w_*\|_{\nabla^2 P(w_*)}\left(-\|\nabla\widehat{P}_N(w_*)\|_{(\nabla^2 P(w_*))^{-1}} + \frac{1}{4}\|w - w_*\|_{\nabla^2 P(w_*)}\right)
\end{aligned}
$$

Observe that if the right hand side is positive for some $w \in B_1$, then w is not a local minimum. Also, since $\|\nabla\widehat{P}_N(w_*)\| \to 0$, for a sufficiently small value of $\|\nabla\widehat{P}_N(w_*)\|$, all points on the boundary of B_1 will have values greater than that of w_*. Hence, we must have a local minimum of $\widehat{P}_N(w)$ that is strictly inside B_1 (for N large enough). We can ensure this local minimum condition is achieved by choosing an N large enough so that $\sqrt{\frac{p \log N}{N}} \leq c' \min\left\{\frac{1}{L_1 L_3}, \frac{\mathrm{diameter}(B)}{L_1}\right\}$, using Lemma 11 (and our bound on the diameter of B_1). By convexity, we have that this is the global minimum, $\widehat{w}_N$, and so $\widehat{w}_N \in B_1$ for N large enough. Assume now that N is this large from here on.

For the ERM, $0 = \nabla\widehat{P}_N(\widehat{w}_N)$. Again, by Taylor's theorem if $\widehat{w}_N$ is an interior point, we have:

$$
0 = \nabla\widehat{P}_N(\widehat{w}_N) = \nabla\widehat{P}_N(w_*) + \nabla^2\widehat{P}_N(\widetilde{w}_N)(\widehat{w}_N - w_*)
$$

for some $\widetilde{w}_N$ between w_* and $\widehat{w}_N$. Now observe that $\widetilde{w}_N$ is in B_1 (since, for N large enough, $\widehat{w}_N \in B_1$). Thus,

$$
\widehat{w}_N - w_* = (\nabla^2\widehat{P}_N(\widetilde{w}_N))^{-1}\nabla\widehat{P}_N(w_*) \tag{35}
$$

where the invertibility is guaranteed by Equation 34 and the positive definiteness of $\nabla P(w_*)$. Using Lemma 11 in Appendix D,

$$
\|\widehat{w}_N - w_*\|_{\nabla^2 P(w_*)}
$$
$$
\leq \|(\nabla^2 P(w_*))^{1/2}(\nabla^2\widehat{P}_N(\widetilde{w}_N))^{-1}(\nabla^2 P(w_*))^{1/2}\|\|\nabla\widehat{P}_N(w_*)\|_{(\nabla^2 P(w_*))^{-1}} \leq cL_1\sqrt{\frac{p \log dN}{N}} \tag{36}
$$

for some universal constant c.

Again, by Taylor's theorem, we have that for some $\widetilde{z}_N$:

$$
P(\widehat{w}_N) - P(w_*) = \frac{1}{2}(\widehat{w}_N - w_*)^\top\nabla^2 P(\widetilde{z}_N)(\widehat{w}_N - w_*)
$$

where $\widetilde{z}_N$ is between w_* and $\widehat{w}_N$.

Observe that both $\widetilde{w}_N$ and $\widetilde{z}_N$ are between $\widehat{w}_N$ and w_*, which implies $\widetilde{w}_N \to w_*$ and $\widetilde{z}_N \to w_*$ since $\widehat{w}_N \to w_*$. By Equations 33 and 36 (and the tail inequalities in Appendix D),

$$
\|\nabla^2\widehat{P}_N(\widetilde{w}_N) - \nabla^2 P(w_*)\|_* \leq c\left(L_1 L_3 + \sqrt{L_2}\right)\sqrt{\frac{p \log dN}{N}}
$$

$$
\|\nabla^2 P_N(\widetilde{z}_N) - \nabla^2 P(w_*)\|_* \leq L_3\|\widetilde{z}_N - w_*\|_{\nabla^2 P(w_*)} \leq cL_1 L_3\sqrt{\frac{p \log dN}{N}}
$$

Define:

$$
\varepsilon_N = c\left(L_1 L_3 + \sqrt{L_2}\right)\sqrt{\frac{p \log dN}{N}}
$$

24

Here the universal constant c is chosen so that:

$$(1 - \varepsilon_N) \nabla^2 P(w_*) \preceq \nabla^2 P(\widetilde{z}_N) \preceq (1 + \varepsilon_N) \nabla^2 P(w_*)$$

and

$$(1 - \varepsilon_N) \nabla^2 P(w_*) \preceq \nabla^2 \widehat{P}_N(\widetilde{w}_N) \preceq (1 + \varepsilon_N) \nabla^2 P(w_*)$$

(using standard matrix perturbation results).

Define:

$$
\begin{aligned}
M_{1,N} &= (\nabla^2 P(w_*))^{1/2} (\nabla^2 \widehat{P}_N(\widetilde{w}_N))^{-1} (\nabla^2 P(w_*))^{1/2} \\
M_{2,N} &= (\nabla^2 P(w_*))^{-1/2} \nabla^2 P(\widetilde{z}_N) (\nabla^2 P(w_*))^{-1/2}
\end{aligned}
$$

For a lower bound, observe that:

$$
\begin{aligned}
P(\widehat{w}_N) - P(w_*) &\geq \frac{1}{2} \lambda_{\min}(M_{2,N}) \|\widehat{w}_N - w_*\|^2_{\nabla^2 P(w_*)} \\
&= \frac{1}{2} \lambda_{\min}(M_{2,N}) \left\| \nabla^2 \widehat{P}_N(\widetilde{w}_N)(\widehat{w}_N - w_*) \right\|^2_{(\nabla^2 \widehat{P}_N(\widetilde{w}_N))^{-1} \nabla^2 P(w_*)(\nabla^2 \widehat{P}_N(\widetilde{w}_N))^{-1}} \\
&\geq \frac{1}{2} (\lambda_{\min}(M_{1,N}))^2 \lambda_{\min}(M_{2,N}) \left\| \nabla^2 \widehat{P}_N(\widetilde{w}_N)(\widehat{w}_N - w_*) \right\|^2_{(\nabla^2 P(w_*))^{-1}} \\
&= \frac{1}{2} (\lambda_{\min}(M_{1,N}))^2 \lambda_{\min}(M_{2,N}) \left\| \nabla \widehat{P}_N(w_*) \right\|^2_{(\nabla^2 P(w_*))^{-1}}
\end{aligned}
$$

where we have used the ERM expression in Equation 35.

Let $I(\mathcal{E})$ be the indicator that the desired previous events hold, which we can ensure with probability greater than $1 - \frac{c}{N^p}$. We have:

$$
\begin{aligned}
&\mathbb{E}[P(\widehat{w}_N) - P(w_*)] \\
\geq\ & \mathbb{E}[(P(\widehat{w}_N) - P(w_*))I(\mathcal{E})] \\
\geq\ & \frac{1}{2} \mathbb{E}\left[(\lambda_{\min}(M_{1,N}))^2 \lambda_{\min}(M_{2,N}) \left\| \nabla \widehat{P}_N(w_*) \right\|^2_{(\nabla^2 P(w_*))^{-1}} I(\mathcal{E}) \right] \\
\geq\ & (1 - c'\varepsilon_N) \frac{1}{2} \mathbb{E}\left[\left\| \nabla \widehat{P}_N(w_*) \right\|^2_{(\nabla^2 P(w_*))^{-1}} I(\mathcal{E}) \right] \\
=\ & (1 - c'\varepsilon_N) \frac{1}{2} \mathbb{E}\left[\left\| \nabla \widehat{P}_N(w_*) \right\|^2_{(\nabla^2 P(w_*))^{-1}} (1 - I(\text{not } \mathcal{E})) \right] \\
=\ & (1 - c'\varepsilon_N) \left(\sigma^2 - \frac{1}{2} \mathbb{E}\left[\left\| \nabla \widehat{P}_N(w_*) \right\|^2_{(\nabla^2 P(w_*))^{-1}} I(\text{not } \mathcal{E}) \right] \right) \\
\geq\ & (1 - c'\varepsilon_N)\sigma^2 - \mathbb{E}\left[\left\| \nabla \widehat{P}_N(w_*) \right\|^2_{(\nabla^2 P(w_*))^{-1}} I(\text{not } \mathcal{E}) \right]
\end{aligned}
$$

(for a universal constant c'). Now define the random variable $Z = \left\| \nabla \widehat{P}_N(w_*) \right\|_{(\nabla^2 P(w_*))^{-1}}$. With a failure event probability of less than $\frac{1}{2N^p}$, for any z_0, we have:

$$
\begin{aligned}
\mathbb{E}\left[Z^2 I(\text{not } \mathcal{E}) \right] &= \mathbb{E}\left[Z^2 I(\text{not } \mathcal{E}) I(Z^2 \leq z_0) \right] + \mathbb{E}\left[Z^2 I(\text{not } \mathcal{E}) I(Z^2 \geq z_0) \right] \\
&\leq z_0 \mathbb{E}\left[I(\text{not } \mathcal{E}) \right] + \mathbb{E}\left[Z^2 I(Z^2 \geq z_0) \right]
\end{aligned}
$$

25

$$\leq \frac{z_0}{2N^p} + \mathbb{E}\left[Z^2\frac{Z^2}{z_0}\right]$$

$$\leq \frac{z_0}{2N^p} + \frac{\mathbb{E}[Z^4]}{z_0}$$

$$\leq \frac{\sqrt{\mathbb{E}[Z^4]}}{N^{p/2}}$$

where we have chosen $z_0 = N^{p/2}\sqrt{\mathbb{E}[Z^4]}$.

For an upper bound:

$$
\begin{aligned}
\mathbb{E}[P(\widehat{w}_N) - P(w_*)] &= \mathbb{E}[(P(\widehat{w}_N) - P(w_*))I(\mathcal{E})] + \mathbb{E}[(P(\widehat{w}_N) - P(w_*))I(\text{not } \mathcal{E})] \quad (37) \\
&\leq \mathbb{E}[(P(\widehat{w}_N) - P(w_*))I(\mathcal{E})] + \frac{\max_{w\in\mathcal{S}}(P(w) - P(w_*))}{N^p}
\end{aligned}
$$

since the probability of not $\mathcal{E}$ is less than $\frac{1}{N^p}$.

For an upper bound of the first term, observe that:

$$
\begin{aligned}
&\mathbb{E}[(P(\widehat{w}_N) - P(w_*))I(\mathcal{E})] \\
&\leq \frac{1}{2}\mathbb{E}\left[(\lambda_{\max}(M_{1,N}))^2\lambda_{\max}(M_{2,N})\left\|\nabla\widehat{P}_N(w_*)\right\|^2_{\nabla^2 P(w_*)}I(\mathcal{E})\right] \\
&\leq (1 + c'\varepsilon_N)\frac{1}{2}\mathbb{E}\left[\left\|\nabla\widehat{P}_N(w_*)\right\|^2_{(\nabla^2 P(w_*))^{-1}}I(\mathcal{E})\right] \\
&\leq (1 + c'\varepsilon_N)\frac{1}{2}\mathbb{E}\left[\left\|\nabla\widehat{P}_N(w_*)\right\|^2_{(\nabla^2 P(w_*))^{-1}}\right] \\
&= (1 + c'\varepsilon_N)\frac{\sigma^2}{N}
\end{aligned}
$$

This completes the proof (using a different universal constant c' in ε_N). $\qquad\square$

6 Conclusion

In many estimation problems, such as linear and logistic regression, the objective is to minimize an unknown function given only unbiased samples. Traditionally, the empirical risk minimizer (ERM), or M-estimator, is the preferred strategy due to its strong statistical properties. However, its computational cost—especially in streaming or large-scale settings—makes it less practical when resources such as time and memory are constrained.

In this work, we presented a single-pass streaming algorithm that competes with the ERM on every problem, while requiring significantly fewer computational resources. The algorithm operates in linear time, using memory proportional to the size of a single sample, and matches the ERM's statistical rate of convergence, even in finite-sample scenarios. Additionally, the performance of our method improves super-polynomially with respect to the initial error, and the algorithm is trivially parallelizable, making it highly efficient for large datasets and distributed environments.

We provided a theoretical analysis of the algorithm, quantifying the finite-sample rate at which it becomes competitive with the ERM, and demonstrated its applicability to various convex problems, including linear and logistic regression. This work offers a practical and scalable alternative to traditional empirical risk minimization methods, paving the way for efficient stochastic optimization in streaming and large-scale data settings.

References

[Agarwal and Bottou(2014)] A. Agarwal and L. Bottou. A lower bound for the optimization of finite sums. Technical report, arXiv, 2014. URL http://arxiv.org/abs/1410.0723.

[Agarwal et al.(2012)Agarwal, Bartlett, Ravikumar, and Wainwright] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, May 2012.

[Anbar(1971)] D. Anbar. *On Optimal Estimation Methods Using Stochastic Approximation Procedures.* University of California, 1971. URL http://books.google.com/books?id=MmpHJwAACAAJ.

[Bach(2010)] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

[Bach and Moulines(2011)] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Neural Information Processing Systems (NIPS)*, 2011.

[Bach and Moulines(2013)] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Neural Information Processing Systems (NIPS)*, 2013.

[Bottou and Bousquet(2008)] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Neural Information Processing Systems (NIPS)*, volume 20, pages 161–168, 2008.

[Caponnetto and De Vito(2007)] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[Defossez and Bach(2015)] A. Defossez and F. Bach. Constant step size least-mean-square: Bias-variance trade-offs and optimal sampling distributions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[Dieuleveut and Bach(2014)] A. Dieuleveut and F. Bach. Non-parametric stochastic approximation with large step sizes. Technical report, arXiv, 2014. URL http://arxiv.org/abs/1408.0361.

[Fabian(1973)] V. Fabian. Asymptotically efficient stochastic approximation; the RM case. *Annals of Statistics*, 1(3), 1973.

[Hazan and Kale(2014)] E. Hazan and S. Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15:2489–2512, 2014. URL http://jmlr.org/papers/v15/hazan14a.html.

[Hsu et al.(2011)Hsu, Kakade, and Zhang] D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. Technical report, arXiv, 2011. URL http://arxiv.org/abs/1110.2842.

[Hsu et al.(2012)Hsu, Kakade, and Zhang] D. Hsu, S. M. Kakade, and T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17(14):1–13, 2012.

[Hsu et al.(2014)Hsu, Kakade, and Zhang] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, June 2014.

[Johnson and Zhang(2013)] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Neural Information Processing Systems (NIPS)*, 2013.

[Juditsky and Nesterov(2010)] A. Juditsky and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions, 2010. URL https://hal.archives-ouvertes.fr/hal-00508933.

[Kushner and Clark(1978)] H. J. Kushner and D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer-Verlag, 1978.

[Kushner and Yin(2003)] H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Applications of mathematics. Springer, 2003. ISBN 9780387008943. URL `http://books.google.com/books?id=_ObIieuUJGkC`.

[Le Roux et al.(2012)Le Roux, Schmidt, and Bach] N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Neural Information Processing Systems (NIPS)*, 2012.

[Lehmann and Casella(1998)] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 1998. ISBN 9780387985022.

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

[Nemirovski and Yudin(1983)] A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley (Chichester and New York), 1983.

[Nesterov(2004)] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004. ISBN 9781402075537.

[Polyak and Juditsky(1992)] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, July 1992. ISSN 0363-0129.

[Rakhlin et al.(2012)Rakhlin, Shamir, and Sridharan] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning, (ICML)*, 2012.

[Robbins and Monro(1951)] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, Sept 1951.

[Ruppert(1988)] D. Ruppert. Efficient Estimations from a Slowly Convergent Robbins-Monro Process. Technical report, ORIE, Cornell University, 1988. URL `http://hdl.handle.net/1813/8664`.

[Shalev-Shwartz and Zhang(2013)] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research (JMLR)*, 14(1):567–599, February 2013.

[Shamir(2014)] O. Shamir. The sample complexity of learning linear predictors with the squared loss. Technical report, arXiv, 2014. URL `http://arxiv.org/abs/1406.5143`.

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

[Tarres and Yao(2014)] P. Tarres and Y. Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

[van der Vaart(2000)] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. ISBN 9780521784504.

[Yao(2010)] Y. Yao. On complexity issues of online learning algorithms. *IEEE Transactions on Information Theory*, 56(12), 2010.

[Ying and Pontil(2008)] Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

A A weaker smoothness assumption

Instead, of the smoothness Assumption 1 in Equation 7, we could instead directly assume that:

$$\mathbb{E}_{\psi \sim \mathcal{D}} \left[\|\nabla\psi(w) - \nabla\psi(w_*)\|^2 \right] \leq 2L\left(P(w) - P(w_*)\right) \tag{38}$$

Our proofs only use this condition, as well as an upper bound on the Hessian of P at w_*. However, we can show that this weaker assumption implies such an upper bound as follows.

Lemma 9. *If* (38) *holds then* $\nabla^2 P(w_*) \preceq 2LI$.

Proof. First we note that for all w, by (38), the convexity of P, and Jensen's inequality, we have:

$$\|\nabla P(w)\|^2 = \|\nabla P(w) - \nabla P(w_*)\|^2 \leq \mathbb{E}_{\psi \sim D} \|\nabla\psi(w) - \nabla\psi(w_*)\|^2 \leq 2L(P(w) - P(w_*)) \tag{39}$$

Since P is convex we also know that

$$P(w_*) \geq P(w) + \nabla P(w)^\top (w_* - w) \tag{40}$$

Combining (39) and (40) and using Cauchy-Schwarz yields that for all w,

$$\|\nabla P(w)\|^2 \leq 2L(\nabla P(w)^\top(w_* - w)) \leq 2L\|\nabla P(w)\|\|w_* - w\|$$

Consequently, for all w we have that

$$\|\nabla P(w)\| \leq 2L\|w_* - w\| \tag{41}$$

Now fix any $\Delta \in \mathcal{S}$ and let $g : \mathbb{R} \to \mathbb{R}$ be defined for all $t \in \mathbb{R}$ as

$$g_\Delta(t) \stackrel{\text{def}}{=} f(w_* + t\Delta)$$

By the chain rule we know that

$$g'(t) = \Delta^\top \nabla P(w_* + t\Delta) \quad \text{and} \quad g''(t) = \Delta^\top \nabla^2 P(w_* + t\Delta)\Delta$$

Consequently, by definition and the fact that P is twice differentiable at w_* we have

$$\Delta^\top \nabla^2 P(w_*)\Delta = g''(0) = \lim_{t \to 0} \frac{\Delta^\top \nabla P(w_* + t\Delta) - \Delta^\top \nabla P(w_*)}{t}$$

Applying Cauchy-Schwarz and (41) yields that

$$\Delta^\top \nabla^2 P(w_*)\Delta \leq \lim_{t \to 0} \frac{\|\Delta\| \cdot \|\nabla P(w_* + t\Delta)\|}{|t|} \leq \lim_{t \to 0} \frac{\|\Delta\| \cdot 2L|t|\|\Delta\|}{|t|} \leq 2L\|\Delta\|^2$$

Since Δ was arbitrary we have the desired result. $\qquad\qquad\square$

B Proofs of Corollaries 1 and 2

Throughout, define:

$$E_s = \sqrt{\mathbb{E}[P(\widehat{w}_{N_s}) - P(w_*)]}$$

Proof. **of Corollary 1.** From Theorem 1, we have:

$$\sqrt{\mathbb{E}[P(\widetilde{w}_{s+1}) - P(w_*)]} \leq \frac{1}{\sqrt{1-4\eta}} \left[\left(\sqrt{\frac{\kappa}{m\eta} + 4\eta} + \sqrt{\kappa\frac{\alpha+2\eta}{k}} \right) \sqrt{\mathbb{E}[P(\widetilde{w}_s) - P(w_*)]} + \sqrt{\alpha + 4\eta}\frac{\sigma}{\sqrt{k}} \right]$$

Let us first show that:

$$E_s \leq \frac{E_{s-1}}{(\sqrt{b})^{p+1}} + \left(1 + \frac{1}{(\sqrt{b})^{(p+1)}}\right) \frac{\alpha\sigma}{\sqrt{k_{s-1}}}$$

We do this using Theorem 1 and some explicit calculations as follows. We shall make use of that for $x \leq 1$, $\sqrt{1-x} \geq 1-x$, and for $0 \leq x \leq \frac{1}{2}$, $\frac{1}{1-x} \leq 1 + x + 2x^2 \leq 1 + 2x$. We have:

$$\frac{1}{\sqrt{1-4\eta}} \leq \frac{1}{\sqrt{1-\frac{1}{45}}} \leq 1.03$$

$$\frac{1}{\sqrt{1-4\eta}} \leq 1 - 8\eta \leq 1 + \frac{1}{2b^{p+1}}$$

$$\frac{\sqrt{\alpha+2\eta}}{\sqrt{1-4\eta}} \leq \alpha(1+\sqrt{2\eta})(1+\frac{1}{2(b^{p+1})}) \leq \alpha(1+\frac{1}{3(\sqrt{b})^{p+1}})(1+\frac{1}{2(b^{p+1})}) \leq \alpha\left(1 + \frac{1}{(\sqrt{b})^{p+1}}\right)$$

$$\frac{\sqrt{4\eta + \frac{\kappa}{\eta m}}}{\sqrt{1-4\eta}} \leq 1.03\left(\sqrt{\frac{4}{20(b^{p+1})} + \frac{1}{20(b^{p+1})}}\right) \leq \frac{0.55}{(\sqrt{b})^{p+1}}$$

$$\frac{\sqrt{\kappa\frac{\alpha+2\eta}{k}}}{\sqrt{1-4\eta}} \leq 1.03\sqrt{\kappa\frac{\alpha(1+1/40)}{k}} \leq \frac{0.1}{(\sqrt{b})^{p+1}}$$

$$\frac{\sqrt{\frac{\kappa}{m\eta} + 4\eta} + \sqrt{\kappa\frac{\alpha+2\eta}{k}}}{\sqrt{1-4\eta}} \leq \frac{1}{(\sqrt{b})^{p+1}}$$

This completes the claim, by substitution into Theorem 1.

We now show:

$$E_s \leq \frac{E_0}{(\sqrt{b})^{(p+1)s}} + \left(1 + \frac{2}{(\sqrt{b})^p}\right) \frac{\alpha\sigma}{\sqrt{k_{s-1}}} \tag{42}$$

We do this by induction. The claim is true for $s = 1$. For the inductive argument,

$$E_s \leq \frac{E_{s-1}}{(\sqrt{b})^{p+1}} + \left(1 + \frac{1}{(\sqrt{b})^{(p+1)}}\right) \frac{\alpha\sigma}{\sqrt{k_{s-1}}}$$

$$\leq \frac{E_0}{(\sqrt{b})^{(p+1)s}} + \frac{1}{(\sqrt{b})^{(p+1)}}\left(1 + \frac{2}{(\sqrt{b})^p}\right) \frac{\alpha\sigma}{\sqrt{k_{s-2}}} + \left(1 + \frac{1}{(\sqrt{b})^{p+1}}\right) \frac{\alpha\sigma}{\sqrt{k_{s-1}}}$$

$$= \frac{E_0}{(\sqrt{b})^{(p+1)s}} + \left(1 + \frac{2}{(\sqrt{b})^p}\right) \frac{\alpha\sigma}{(\sqrt{b})^p\sqrt{k_{s-1}}} + \left(1 + \frac{1}{(\sqrt{b})^{p+1}}\right) \frac{\alpha\sigma}{\sqrt{k_{s-1}}}$$

$$\leq \frac{E_0}{(\sqrt{b})^{(p+1)s}} + \left(1 + \frac{2}{(\sqrt{b})^p}\right) \frac{\alpha\sigma}{\sqrt{k_{s-1}}}$$

which completes the inductive argument.

We now relate k_s and $\sqrt{b}^{(p+1)s}$ to the sample size N_s. First, observe that m is bounded as:

$$m = 400b^{(p+1)^2}\kappa \leq 20b^{(p+1)^2+3}\kappa = 20b^{p^2+2p+4}\kappa$$

Using $s > p^2 + 6p$,

$$N_s = \sum_{\tau=1}^{s}(m + k_\tau) \leq s20b^{p^2+2p+4}\kappa + \frac{k_{s-1}}{1 - \frac{1}{b}} = \frac{sk_s b^{p^2+p+3}}{b^s} + \frac{k_s}{1 - \frac{1}{b}} \leq (1 + \frac{1.1}{b})k_s$$

and so:

$$\alpha\left(1 + \frac{2}{(\sqrt{b})^p}\right)\frac{\sigma}{\sqrt{k_{s-1}}} \leq \left(1 + \frac{2}{(\sqrt{b})^p}\right)\sqrt{1 + \frac{1.1}{b}}\frac{\alpha\sigma}{\sqrt{N_s}} \leq \left(1 + \frac{4}{b}\right)\frac{\alpha\sigma}{\sqrt{N_s}} \tag{43}$$

Also,

$$(b^s)^{p+1} = (b^s)^p b^s = \left(k_s\frac{1}{20\alpha b^{p+1}\kappa}\right)^p b^s \geq \left(N_s\frac{1}{20(1 + \frac{1.1}{b})\alpha b^{p+1}\kappa}\right)^p b^s \geq \left(\frac{1}{\alpha\kappa}N_s\right)^p\frac{b^s}{(b^{p+5})^p} \geq \left(\frac{1}{\alpha\kappa}N_s\right)^p$$

where we have used that, for $s > p^2 + 6p$, $\frac{b^s}{(b^{p+5})^p} \geq 1$. Hence,

$$\frac{1}{b^{s(p+1)}} \leq \frac{1}{\left(\frac{N_s}{\alpha\kappa}\right)^p} \tag{44}$$

The proof is completed substituting (43), (44) in (42). $\qquad\square$

Proof. **of Corollary 2.** Under the choice of parameters and using Theorem 1, we have

$$E_s \leq \frac{E_{s-1}}{\sqrt{b}^{p+1}} + \frac{\sigma}{\sqrt{k_{s-1}}}\sqrt{3\kappa + \frac{1}{b}}$$

On the other hand, suppose t_0 is the first time that $k_{t_0} \geq (M\sigma + 1)^2 400\kappa^2 b^{2p+3} = (M\sigma + 1)^2 k_0$. When $s \geq t_0$, we can use Theorem 2, and under the choice of parameters we have:

$$E_s \leq \frac{E_{s-1}}{\sqrt{b}^{p+1}} + \frac{\sigma}{\sqrt{k_{s-1}}}\left(1 + \frac{1}{b}\right)$$

Now we shall prove by induction that

$$E_s \leq \frac{E_0}{\sqrt{b}^{(p+1)s}} + \frac{\sigma}{\sqrt{k_{s-1}}}(1 + 2/b)\sqrt{3\kappa + 1/b} \cdot \min\{1, (\sqrt{b}^{-p(s-t_0)})\} + \frac{\sigma}{\sqrt{k_{s-1}}}(1 + 2/b)$$

Here E_0 is the initial error. When $s = 1$ the statement is true. When $s \leq t_0$ we use the first recursion (from Theorem 1), clearly

$$E_s \leq \frac{E_{s-1}}{\sqrt{b}^{p+1}} + \frac{\sigma}{\sqrt{k_{s-1}}}\sqrt{3\kappa + \frac{1}{b}}$$

$$\leq \left(\frac{E_0}{\sqrt{b}^{(p+1)(s-1)}} + \frac{\sigma}{\sqrt{k_{s-2}}}(1+2/b)\sqrt{3\kappa+1/b} + \frac{\sigma}{\sqrt{k_{s-2}}}(1+2/b) \right) \cdot \frac{1}{\sqrt{b}^{p+1}} + \frac{\sigma}{\sqrt{k_{s-1}}}\sqrt{3\kappa+\frac{1}{b}}$$

$$= \frac{E_0}{\sqrt{b}^{(p+1)s}} + \frac{\sigma}{\sqrt{k_{s-1}}}\sqrt{3\kappa+1/b}\left(\frac{1+2/b}{\sqrt{b}^p} + 1 \right) + \frac{\sigma}{\sqrt{k_{s-1}}}(1+2/b)\cdot\frac{1}{\sqrt{b}^p}$$

$$\leq \frac{E_0}{\sqrt{b}^{(p+1)s}} + \frac{\sigma}{\sqrt{k_{s-1}}}(1+2/b)\sqrt{3\kappa+1/b} + \frac{\sigma}{\sqrt{k_{s-1}}}(1+2/b)$$

Here the second step uses induction hypothesis, and the third step uses the fact that $k_{s-1}/k_{s-2} = b$ and $(1+2/b)/\sqrt{b}^p + 1 \leq 1 + 2/b$.

When $s > t_0$ we can use the second recursion (from Theorem 2), now we have

$$E_s \leq \frac{E_{s-1}}{\sqrt{b}^{p+1}} + \frac{\sigma}{\sqrt{k_{s-1}}}\sqrt{1+\frac{1}{b}}$$

$$\leq \left(\frac{E_0}{\sqrt{b}^{(p+1)(s-1)}} + \frac{\sigma}{\sqrt{k_{s-2}}}(1+2/b)\sqrt{3\kappa+1/b}\cdot(\sqrt{b}^{-p(s-t_0-1)}) + \frac{\sigma}{\sqrt{k_{s-2}}}(1+2/b) \right)\frac{1}{\sqrt{b}^{p+1}}$$

$$+ \frac{\sigma}{\sqrt{k_{s-1}}}\left(1+\frac{1}{b} \right)$$

$$= \frac{E_0}{\sqrt{b}^{(p+1)s}} + \frac{\sigma}{\sqrt{k_{s-1}}}(1+2/b)\sqrt{3\kappa+1/b}\cdot(\sqrt{b}^{-p(s-t_0)})$$

$$+ \frac{\sigma}{\sqrt{k_{s-1}}}\left(\frac{(1+2/b)}{\sqrt{b}^p} + 1 + \frac{1}{b} \right)$$

$$\leq \frac{E_0}{\sqrt{b}^{(p+1)s}} + \frac{\sigma}{\sqrt{k_{s-1}}}(1+2/b)\sqrt{3\kappa+1/b}\cdot(\sqrt{b}^{-p(s-t_0)}) + \frac{\sigma}{\sqrt{k_{s-1}}}(1+2/b)$$

Again, the second step uses induction hypothesis, and final step uses $k_{s-1}/k_{s-2} = b$ and $(1+2/b)/\sqrt{b}^p + 1 + 1/b \leq 1 + 2/b$. This concludes the induction.

Finally, we need to relate the values k_s, $\sqrt{b}^{(p+1)(s+1)}$, $\sqrt{b}^{(p+1)(s-t_0)}$ with N_s.

First, it is clear that $k_s \geq bm$ for all s, therefore

$$N_s \leq (1+1/b)\sum_{t=0}^{s-1}k_s \leq k_{s-1}(1-1/b)^{-1}(1+1/b) \leq k_{s-1}(1+3/b) \leq 2k_{s-1}$$

Therefore we can substitute $1/\sqrt{k_{s-1}}$ with $\sqrt{1+3/b}/\sqrt{N_s}$. Also, we know $N_s/2k_0 \leq b^{s-1}$, therefore $\sqrt{b}^{-(p+1)s} \leq \sqrt{b}^{-(p+1)(s-1)} \leq (N/2k_0)^{-(p+1)/2}$.

Finally, since k increase by a factor of b, we know k_{t_0} is at most $b(M\sigma + 1)^2 k_0$. Therefore $\frac{N_s}{2(M\sigma+1)^2 k_0} \leq bN_s/2k_{t_0} \leq k_s/k_{t_0} = b^{s-t_0}$, which means $\sqrt{b}^{-p(s-t_0)} \leq \left(\frac{N_s}{2(M\sigma+1)^2 k_0} \right)^{p/2}$. $\qquad\square$

C Self-concordance for logistic regression

The following straightforward lemma, to handle self-concordance for logistic regression, is included for completeness (see [Bach(2010)] for a more detailed treatment for analyzing the self-concordance of logistic regression).

Lemma 10. *(Self-Concordance for Logistic Regression) For the logistic regression case (as defined in Section 3), define $M = \alpha \mathbb{E}[\|X\|^3_{(\nabla^2 P(w_*))^{-1}}]$, then*

$$P(w_*) \geq P(w_t) + (w_* - w_t)^\top \nabla P(w_t) + \frac{\|w_t - w_*\|^2_{\nabla^2 P(w_*)}}{2(1 + M\|w_t - w_*\|_{\nabla^2 P(w_*)})^2}$$

Proof. For $\widetilde{M} = \mathbb{E}[\|X\|^3_{(\nabla^2 P(w_*))^{-1}}]$, first let us show that:

$$P(w_*) \geq P(w_t) + (w_* - w_t)^\top \nabla P(w_t) + \frac{1}{2}\|w_t - w_*\|^2_{\nabla^2 P(w_*)} \max\left\{\frac{1}{\alpha}, 1 - \widetilde{M}\|w_t - w_*\|_{\nabla^2 P(w_*)}\right\} \quad (45)$$

By Taylor's theorem,

$$P(w_*) = P(w_t) + (w_* - w_t)^\top \nabla P(w_t) + \frac{1}{2}(w_t - w_*)^\top \nabla^2 P(z_1)(w_t - w_*)$$

where z_1 is between w_* and w_t. Again, by Taylor's theorem,

$$\nabla^2 P(z_1) = \nabla^2 P(w_*) + \nabla^3 P(z_2)(z_1 - w_*)$$

where z_2 is between w_* and z_1.

By taking derivatives, we have that:

$$\begin{aligned}
&(w_t - w_*)^\top \nabla^2 P(z_1)(w_t - w_*)\\
=&(w_t - w_*)^\top \nabla^2 P(w_*)(w_t - w_*)\\
&\quad + \mathbb{E}[\mathbb{P}(Y|z_2, X)(1 - \mathbb{P}(Y|z_2, X))(1 - 2\mathbb{P}(Y|z_2, X))((w_t - w_*)^\top X)^2(z_1 - w_*)^\top X]\\
\geq&\|w_t - w_*\|^2_{\nabla^2 P(w_*)} - \|w_t - w_*\|^2_{\nabla^2 P(w_*)}\|z_1 - w_*\|_{\nabla^2 P(w_*)}\mathbb{E}[\|X\|^3_{(\nabla^2 P(w_*))^{-1}}]\\
\geq&\|w_t - w_*\|^2_{\nabla^2 P(w_*)} - \|w_t - w_*\|^3_{\nabla^2 P(w_*)}\mathbb{E}[\|X\|^3_{(\nabla^2 P(w_*))^{-1}}]\\
=&\|w_t - w_*\|^2_{\nabla^2 P(w_*)}\left(1 - \widetilde{M}\|w_t - w_*\|_{\nabla^2 P(w_*)}\right)
\end{aligned}$$

Using the definition of α, shows that Equation 45 holds.

Now for $Z > 0$, consider the quantity $\max\{\frac{1}{\alpha}, 1 - Z\}$, and observe that the $1 - Z$ term achieves the max when $1 - Z \geq \frac{1}{\alpha}$ or equivalently when $-1 + \alpha - \alpha Z \geq 0$. Hence,

$$\begin{aligned}
\max\left\{\frac{1}{\alpha}, 1 - Z\right\} &= \max\left\{\frac{1}{\alpha}, \frac{(1 - Z)(1 + \alpha Z)}{1 + \alpha Z}\right\}\\
&= \max\left\{\frac{1}{\alpha}, \frac{1 - Z + \alpha Z - \alpha Z^2}{1 + \alpha Z}\right\}\\
&= \max\left\{\frac{1}{\alpha}, \frac{1 + Z(-1 + \alpha - \alpha Z)}{1 + \alpha Z}\right\}\\
&\geq \max\left\{\frac{1}{\alpha}, \frac{1}{1 + \alpha Z}\right\}\\
&\geq \frac{1}{1 + \alpha Z}\\
&\geq \frac{1}{(1 + \alpha Z)^2}
\end{aligned}$$

Using this completes the proof. $\qquad \square$

D   Probability tail inequalities

The following probability tail inequalities are used in our analysis.

The first tail inequality is for sums of bounded random vectors; it is a standard application of Bernstein's inequality.

Lemma 11 (Vector Bernstein bound; *e.g.* see [Hsu et al.(2011)Hsu, Kakade, and Zhang]). *Let* $x_1, x_2, \ldots, x_n$ *be independent random vectors such that*

$$\sum_{i=1}^{n} \mathbb{E}[\|x_i\|^2] \leq v \quad and \quad \|x_i\| \leq r$$

for all $i = 1, 2, \ldots, n$, *almost surely. Let* $s := x_1 + x_2 + \cdots + x_n$. *For all* $t > 0$,

$$\mathbb{P}\left[\|s\| > \sqrt{v}(1 + \sqrt{8t}) + (4/3)rt\right] \leq e^{-t}$$

The next tail inequality concerns the spectral accuracy of an empirical second moment matrix, where we do not assume the dimension is finite.

Lemma 12 (Infinite Dimensional Matrix Bernstein bound; [Hsu et al.(2012)Hsu, Kakade, and Zhang]). *Let* X *be a random matrix, and* $r > 0$, $v > 0$, *and* $\widetilde{d} > 0$ *be such that, almost surely,*

$$\mathbb{E}[X] = 0, \quad \lambda_{\max}[X] \leq r, \quad \lambda_{\max}[\mathbb{E}[X^2]] = v, \quad \mathrm{Tr}(\mathbb{E}[X^2]) = v\widetilde{d}$$

Define $\widetilde{d}$ *as the intrinsic dimension. If* $X_1, X_2, \ldots, X_n$ *are independent copies of* X, *then for any* $t > 0$,

$$\mathbb{P}\left[\lambda_{\max}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] > \sqrt{\frac{2vt}{n}} + \frac{rt}{3n}\right] \leq \widetilde{d}t(e^t - t - 1)^{-1}$$

If $t \geq 2.6$, *then* $t(e^t - t - 1)^{-1} \leq e^{-t/2}$.