# Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression

Francis Bach

# Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression

Francis Bach
INRIA, SIERRA Project-team
Département d'Informatique de l'École Normale Supérieure
Paris, France
francis.bach@ens.fr

March 25, 2013

**Abstract**

In this paper, we consider supervised learning problems such as logistic regression and study the stochastic gradient method with averaging, in the usual stochastic approximation setting where observations are used only once. We show that for self-concordant loss functions, after $n$ iterations, with a constant step-size proportional to $1/R^2\sqrt{n}$ where $n$ is the number of observations and $R$ is the maximum norm of the observations, the convergence rate is always of order $O(1/\sqrt{n})$, and improves to $O(R^2/\mu n)$ where $\mu$ is the lowest eigenvalue of the Hessian at the global optimum (when this eigenvalue is strictly positive). Since $\mu$ does not need to be known in advance, this shows that averaged stochastic gradient is adaptive to *unknown local* strong convexity of the objective function.

## 1 Introduction

The minimization of an objective function which is only available through unbiased estimates of the function values or its gradients is a key methodological problem in many disciplines. Its analysis has been attacked mainly in three communities: stochastic approximation [1, 2, 3, 4, 5, 6], optimization [7, 8], and machine learning [9, 10, 11, 12, 13, 14, 15]. The main algorithms which have emerged are stochastic gradient descent (a.k.a. Robbins-Monro algorithm), as well as a simple modification where iterates are averaged (a.k.a. Polyak-Ruppert averaging).

The convergence rates of these algorithms depends primarily on the potential strong convexity of the objective function [11, 12, 13, 14, 15]. For $\mu$-strongly convex functions, the optimal rate of convergence of function values is $O(1/n)$ while for convex functions the optimal rate is $O(1/\sqrt{n})$ [20, 21]. For smooth functions, averaged stochastic gradient with step sizes proportional to $1/\sqrt{n}$ achieves them up to logarithmic terms [16].

Convex optimization problems coming from supervised machine learning are typically of the form $f(\theta) = \mathbb{E}\big[\ell(y, \langle\theta, x\rangle)\big]$, where $\ell(y, \langle\theta, x\rangle)$ is the loss between the response $y$ and the prediction $\langle\theta, x\rangle$. They may or may not have strongly convex objective functions. This most often depends on (a) the correlations between covariates $x$, and (b) the strong convexity of the loss function $\ell$. The logistic loss $u \mapsto \log(1 + e^{-u})$ is not strongly convex unless restricted to a compact set; moreover, in the sequential observation model, the correlations are not known at training time. Therefore, many theoretical results based on strong convexity to do not apply. The goal of this paper is

to show that with proper assumptions, namely self-concordance, one can readily obtain favorable theoretical guarantees for logistic regression, namely a rate of the form $O(R^2/\mu n)$ where $\mu$ is the lowest eigenvalue of the Hessian at the global optimum, *without any exponentially increasing constant factor*.

Another goal of this paper is to design an algorithm and provide an analysis that benefit from hidden local strong convexity without requiring to know the local strong convexity constant in advance. In smooth situations, the results of [16] implies that the averaged stochastic gradient method with step sizes of the form $O(n^{-1/2})$ is adaptive to the strong convexity of the problem. However the dependence in $\mu$ in the strongly convex case is of the form $O(\mu^{-2}n^{-1})$, which is sub-optimal. Moreover, the final rate is rather complicated, notably because all possible step-sizes are considered. Finally, it does not apply here because even in low-correlation settings, the objective function of logistic regression cannot be globally strongly convex.

In this paper, we provide an analysis for stochastic gradient with averaging for generalized linear models such as logistic regression, with a step size proportional to $(R^2\sqrt{n})^{-1}$ where $R$ is the radius of the data and $n$ the number of observations, showing such adaptivity. In particular, we show that the algorithm can adapt to the *local* strong-convexity constant, i.e., the lowest eigenvalue of the Hessian at the optimum. The analysis is done for a finite horizon $N$ and a constant step size decreasing in $N$ as $1/R^2\sqrt{N}$, since the analysis is then slightly easier, though (a) a decaying stepsize could be considered as well, and (b) it could be classically extended to varying step-sizes by a doubling trick [17].

# 2 Stochastic approximation for generalized linear models

In this section, we present the assumptions our work relies on, as well as related work.

## 2.1 Assumptions

Throughout this paper, we make the following assumptions. We consider a function $f$ defined on a Hilbert space $\mathcal{H}$, and an increasing family of $\sigma$-fields $(\mathcal{F}_n)_{n\geqslant 1}$; we assume that we are given a deterministic $\theta_0 \in \mathcal{H}$, and a sequence of functions $f_n : \mathcal{H} \to \mathbb{R}$, for $n \geqslant 1$. We make the following assumptions:

**(A1) Convexity and differentiability of** $f$: $f$ is convex and three-times differentiable.

**(A2) Generalized self-concordance of** $f$ [18]: for all $\theta_1, \theta_2 \in \mathcal{H}$, the function $\varphi : t \mapsto f\big[\theta_1 + t(\theta_2 - \theta_1)\big]$ satisfies: $\forall t \in \mathbb{R}$, $|\varphi'''(t)| \leqslant R\|\theta_1 - \theta_2\|\varphi''(t)$.

**(A3) Attained global minimum**: $f$ has a global minimum attained at $\theta_* \in \mathcal{H}$.

**(A4) Lipschitz-continuity of** $f_n$ **and** $f$: all gradients of $f$ and $f_n$ are bounded by $R$, that is, for all $\theta \in \mathcal{H}$,

$$\|f'(\theta)\| \leqslant R \text{ and } \forall n \geqslant 1,\ \|f_n'(\theta)\| \leqslant R \text{ almost surely.}$$

**(A5) Adapted measurability**: $\forall n \geqslant 1$, $f_n$ is $\mathcal{F}_n$-measurable.

**(A6) Unbiased gradients**: $\forall n \geqslant 1$, $\mathbb{E}(f_n'(\theta_{n-1})|\mathcal{F}_{n-1}) = f'(\theta_{n-1})$.

**(A7) Stochastic gradient recursion**: $\forall n \geqslant 1$, $\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})$.

Among the seven assumptions above, the non-standard one is **(A2)**: the notion of self-concordance is an important tool for convex optimization and in particular the study of Newton's method. It corresponds to having the third derivative bounded by the $\frac{3}{2}$-th power of the second derivative. For machine learning, [18] has generalized the notion of self-concordance by removing the $\frac{3}{2}$-th power, so that it is applicable to cost functions arising from probabilistic modeling, as shown below.

Our set of assumptions corresponds to the following examples (with i.i.d. data, and $\mathcal{F}_n$ equal to the $\sigma$-field generated by $x_1, y_1, \ldots, x_n, y_n$):

- **Logistic regression**: $f_n(\theta) = \log(1+\exp(-y_n\langle x_n, \theta\rangle))$, with data $x_n$ uniformly almost surely bounded by $R$ and $y_n \in \{-1, 1\}$. Note that this includes other binary classification losses, such as $f_n(\theta) = -y_n\langle x_n, \theta\rangle + \sqrt{1 + \langle x_n, \theta\rangle^2}$.
- **Generalized linear models with uniformly bounded features**: $f_n(\theta) = -\langle \theta, \Phi(x_n, y_n)\rangle + \log \int h(y) \exp\big(\langle \theta, \Phi(x_n, y)\rangle\big) dy$, with $\Phi(x_n, y)$ almost surely bounded in norm by $R$, for all observations $x_n$ and all potential responses $y$. This includes multinomial regression and conditional random fields [19].
- **Robust regression**: we may use $f_n(\theta) = \varphi(y_n - \langle x_n, \theta\rangle)$, with $\varphi(t) = \log\cosh t = \log \frac{e^t + e^{-t}}{2}$, with a similar boundedness assumption on $x_n$.

## 2.2 Related work

**Non-strongly-convex functions.** When only convexity of the objective function is assumed, then several authors [7, 8, 13, 14, 15] have shown that using a step-size proportional to $1/\sqrt{n}$, *together with some form of averaging*, leads to the minimax optimal rate of $O(1/\sqrt{n})$ [20, 21]. Without averaging, the known convergences rates are suboptimal, that is, averaging is key to obtaining the optimal rate [16]. Note that the smoothness of the loss does not change the rate, but may help to obtain better constants, with the potential use of acceleration [22].

The compactness of the domain is often used within the algorithm (by using orthogonal projections) and within the analysis (in particular to optimize the step size and obtain high-probability bounds). In this paper, we do not make such compactness assumptions, since in a machine learning context, the available bound would be loose and hurt practical performance.

Another difference between several analyses is the use of decaying step sizes of the form $\gamma_n \propto 1/\sqrt{n}$ vs. the use of a constant step size of the form $\gamma \propto 1/\sqrt{N}$ for a finite known horizon $N$ of iterations. The use of a "doubling trick" as done by [17] for strongly convex optimization, where a constant step size is used for iterations between $2^p$ and $2^{p+1}$, with a constant that is proportional to $1/\sqrt{2^p}$, would allow to obtain an anytime algorithm from a finite horizon one. In order to simplify our analysis, we only consider a finite horizon $N$ and a constant step-size that will be proportional to $1/\sqrt{N}$.

**Strongly-convex functions.** When the function is $\mu$-strongly convex, i.e., $\theta \mapsto f(\theta) - \frac{\mu}{2}\|\theta\|^2$ is convex, there are essentially two approaches to obtaining the minimax-optimal rate $O(1/\mu n)$ [20, 21]: (a) using a step size proportional to $1/\mu n$ with averaging for non-smooth problems [7, 8, 13, 14, 15, 23] or a step size proportional to $1/(R^2 + n\mu)$, also with averaging, for smooth problems, where $R^2$ is the smoothness constant of the loss of a single observation [24]; (b) for smooth problems, using longer step-sizes proportional to $1/n^\alpha$ for $\alpha \in (1/2, 1)$ *with* averaging [4, 5, 16].

Note that the "historical" step size, i.e., of the form $C/n$ where $C$ is larger than $1/\mu$, leads, without averaging to a convergence rate of $O(1/\mu^2 n)$ [6, 16], hence leads to a worse dependence on $\mu$.

The solution (a) requires to have a good estimate of the strong-convexity constant $\mu$, while the second solution (b) does not require to know such estimate and leads to a convergence rate achieving asymptotically the Cramer-Rao lower bound [4]. Thus, this last solution is adaptive to unknown (but positive) amount of strong convexity. However, unless we take the limiting setting $\alpha = 1/2$, it is not adaptive to lack of strong convexity. While the non-asymptotic analysis of [16] already gives a convergence rate in that situation, the bound is rather complicated and also has a suboptimal dependence on $\mu$. One goal of this paper is to consider a less general result, but more compact (note also that the analysis of [16] only applies for globally strongly convex functions, see below).

Finally, note that unless we restrict the support, the objective function for logistic regression cannot be globally strongly convex (since the Hessian tends to zero when $\|\theta\|$ tends to infinity). Another goal of the paper is to show that stochastic gradient descent with averaging is adaptive to the *local* strong convexity constant, i.e., the lowest eigenvalue of the Hessian of $f$ at the global optimum, without any exponential terms in $RD$ (which would be present if a compact domain of diameter $D$ was imposed and traditional analyses were performed).

**Adaptivity to unknown constants.** The desirable property of adaptivity to the difficulty of an optimization problem has also been studied in several settings. Gradient descent with constant step size is for example naturally adaptive to the strong convexity of the problem (see, e.g., [25]). In the stochastic context, [26] provides another strategy than averaging with longer step sizes, but for uniform convexity constants.

# 3 Non-strongly convex analysis

In this section, we study the averaged stochastic gradient method in the non-strongly convex case, i.e., without any (global or local) strong convexity assumptions. We first recall existing results in Section 3.1, that bound the expectation of the excess risk leading to a bound in $O(1/\sqrt{n})$. We then show using martingale moment inequalities how all higher-order moments may be bounded in Section 3.2, still with a rate of $O(1/\sqrt{n})$. However, in Section 3.3, we consider the convergence of the squared gradient, with now a rate of $O(1/n)$. This last result is key to obtaining the adaptivity to local strong convexity in Section 4.

## 3.1 Existing results

In this section, we review existing results for Lipschitz-continuous non-strongly convex problems [7, 8, 13, 14, 15]. Note that smoothness is not needed here. We consider a constant step size $\gamma_n = \gamma > 0$, for all $n \geqslant 1$. We denote by $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$ the averaged iterate.

We prove the following proposition, which provides a bound on the expectation of $f(\bar{\theta}_n) - f(\theta_*)$ that decays at rate $O(\gamma + 1/\gamma n)$, hence the usual choice $\gamma \propto 1/\sqrt{n}$:

**Proposition 1** *With constant step size equal to $\gamma$, for any $n \geqslant 0$, we have:*

$$\mathbb{E}f\left(\frac{1}{n}\sum_{k=1}^{n}\theta_{k-1}\right) - f(\theta_*) + \frac{1}{2\gamma n}\mathbb{E}\|\theta_n - \theta_*\|^2 \leqslant \frac{1}{2\gamma n}\|\theta_0 - \theta_*\|^2 + \frac{\gamma}{2}R^2. \tag{1}$$

**Proof.** We have the following recursion, obtained from the Lipschitz-continuity of $f_n$:

$$\begin{aligned}\|\theta_n - \theta_*\|^2 &= \|\theta_{n-1} - \theta_*\|^2 - 2\gamma\langle\theta_{n-1} - \theta_*, f_n'(\theta_{n-1})\rangle + \gamma^2\|f_n'(\theta_{n-1})\|^2 \\ &\leqslant \|\theta_{n-1} - \theta_*\|^2 - 2\gamma\langle\theta_{n-1} - \theta_*, f'(\theta_{n-1})\rangle + \gamma^2 R^2 + M_n,\end{aligned}$$

with

$$M_n = -2\gamma\langle \theta_{n-1} - \theta_*, f_n'(\theta_{n-1}) - f'(\theta_{n-1})\rangle.$$

We thus get, using the classical result from convexity $f(\theta_{n-1}) - f(\theta_*) \leqslant \langle \theta_{n-1} - \theta_*, f'(\theta_{n-1})\rangle$:

$$2\gamma\big[f(\theta_{n-1}) - f(\theta_*)\big] \leqslant \|\theta_{n-1} - \theta_*\|^2 - \|\theta_n - \theta_*\|^2 + \gamma^2 R^2 + M_n. \tag{2}$$

Summing over integers less than $n$, this implies:

$$\frac{1}{n}\sum_{k=0}^{n-1} f(\theta_k) - f(\theta_*) + \frac{1}{2\gamma n}\|\theta_n - \theta_*\|^2 \quad \leqslant \quad \frac{1}{2\gamma n}\|\theta_0 - \theta_*\|^2 + \frac{\gamma}{2}R^2 + \frac{1}{2\gamma n}\sum_{k=1}^{n} M_k.$$

We get the desired result by taking expectation in the last inequality, and using the expectation $\mathbb{E}M_k = \mathbb{E}(\mathbb{E}(M_k|\mathcal{F}_{k-1})) = 0$ and $f\big(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\big) \leqslant \frac{1}{n}\sum_{k=0}^{n-1} f(\theta_k)$. ■

The following corollary considers a specific choice of the step size (note that the bound is only true for the last iteration):

**Corollary 1** *With constant step size equal to $\gamma = \frac{1}{2R^2\sqrt{N}}$, we have:*

$$\forall n \in \{1, \dots, N\}, \quad \mathbb{E}\|\theta_n - \theta_*\|^2 \quad \leqslant \quad \|\theta_0 - \theta_*\|^2 + \frac{1}{4R^2}, \tag{3}$$

$$\mathbb{E}f\left(\frac{1}{N}\sum_{k=1}^{N}\theta_{k-1}\right) - f(\theta_*) \quad \leqslant \quad \frac{R^2}{\sqrt{N}}\|\theta_0 - \theta_*\|^2 + \frac{1}{4\sqrt{N}}. \tag{4}$$

Note that if $\|\theta_0 - \theta_*\|^2$ was known, then a better step-size would be $\gamma = \frac{\|\theta_0 - \theta_*\|}{R\sqrt{N}}$, leading to a convergence rate proportional to $\frac{R\|\theta_0 - \theta_*\|}{\sqrt{N}}$. However, this requires an estimate (simply an upper-bound) of $\|\theta_0 - \theta_*\|^2$, which is typically not available.

We are going to improve this result in several ways:

- All moments of $\|\theta_n - \theta_*\|^2$ and $f(\bar{\theta}_n) - f(\theta_*)$ will be bounded, leading to a sub-Gaussian behavior. Note that we do not assume that the iterates are restricted to a predefined bounded set (which is the usual assumption made to derive tail bounds [8, 27]).
- We are going to show that the squared norm of the gradient at $\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^{n}\theta_{k-1}$ converges at rate $O(n^{-1})$, even in the non-strongly convex case. This will allow us to derive finer convergence rates in presence of local strong convexity in Section 4.

## 3.2 Higher-order bound

In this section, we prove higher-order bounds (see the proof in Appendix C, which is based on taking powers of the inequality in Eq. (2) and using martingale moment inequalities), both for any constant step-sizes and then for the specific choice $\gamma = \frac{1}{2R^2\sqrt{N}}$.

**Proposition 2** *With constant step size equal to $\gamma$, for any $n \geqslant 0$ and integer $p \in \{1, \dots, \lfloor n/4 \rfloor\}$, we have:*

$$\mathbb{E}\left(2\gamma n\big[f(\bar{\theta}_n) - f(\theta^*)\big] + \|\theta_n - \theta_*\|^2\right)^p \leqslant \big(3\|\theta_0 - \theta_*\|^2 + 20np\gamma^2R^2\big)^p. \tag{5}$$

**Corollary 2** *With constant step size equal to* $\gamma = \frac{1}{2R^2\sqrt{N}}$*, for any integer* $p \geqslant 1$*, we have:*

$$\mathbb{E}\|\theta_N - \theta_*\|^{2p} \leqslant \big(3\|\theta_0 - \theta_*\|^2 + 5pR^{-2}\big)^p, \tag{6}$$

$$\mathbb{E}\big[f(\bar{\theta}_N) - f(\theta^*)\big]^p \leqslant \big(3\frac{R^2}{\sqrt{N}}\|\theta_0 - \theta_*\|^2 + \frac{5}{\sqrt{N}}\big)^p. \tag{7}$$

Having abound on all moments allows immediately to derive large deviation bounds in the same two cases (by applying Lemma 1 from Appendix A):

**Proposition 3** *With constant step size equal to* $\gamma$*, for any* $n \geqslant 0$ *and* $t \geqslant 0$*, we have:*

$$\mathbb{P}\bigg(f(\bar{\theta}_n) - f(\theta_*) \geqslant 30\gamma R^2 t + \frac{3\|\theta_0 - \theta_*\|^2}{\gamma n}\bigg) \leqslant 2\exp(-t),$$

$$\mathbb{P}\bigg(\|\theta_n - \theta_*\|^2 \geqslant 60n\gamma^2 R^2 t + 6\|\theta_0 - \theta_*\|^2\bigg) \leqslant 2\exp(-t).$$

**Corollary 3** *With constant step size equal to* $\gamma = \frac{1}{2R^2\sqrt{N}}$*, for any* $t \geqslant 0$ *we have:*

$$\mathbb{P}\bigg(f(\bar{\theta}_N) - f(\theta_*) \geqslant \frac{15t}{\sqrt{N}} + \frac{6R^2\|\theta_0 - \theta_*\|^2}{\sqrt{N}}\bigg) \leqslant 2\exp(-t),$$

$$\mathbb{P}\bigg(\|\theta_N - \theta_*\|^2 \geqslant 15R^{-2}t + 6\|\theta_0 - \theta_*\|^2\bigg) \leqslant 2\exp(-t).$$

We can make the following observations:

- The iterates $\theta_n$ and $\bar{\theta}_n$ do not necessarily converge to $\theta_*$ (note that $\theta_*$ may not be unique in general anyway).
- Given that $(\mathbb{E}[f(\bar{\theta}_n) - f(\theta_*)]^p)^{1/p}$ is affine in $p$, we obtain a subexponential behavior, i.e., tail bounds similar to an exponential distribution.
- The proof of Prop. 2 is rather technical and makes heavy use of martingale moment inequalities.
- The constants in the bounds of of Prop. 2 (and thus other results as well) could clearly be improved. In particular, we have, for $p = 1, 2, 3, 4$ (see proof in Appendix E):

$$\mathbb{E}\bigg(2\gamma n\big[f(\bar{\theta}_n) - f(\theta^*)\big] + \|\theta_n - \theta_*\|^2\bigg) \leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2,$$

$$\mathbb{E}\bigg(2\gamma n\big[f(\bar{\theta}_n) - f(\theta^*)\big] + \|\theta_n - \theta_*\|^2\bigg)^2 \leqslant \big(\|\theta_0 - \theta_*\|^2 + 3n\gamma^2 R^2\big)^2,$$

$$\mathbb{E}\bigg(2\gamma n\big[f(\bar{\theta}_n) - f(\theta^*)\big] + \|\theta_n - \theta_*\|^2\bigg)^3 \leqslant \big(\|\theta_0 - \theta_*\|^2 + 6n\gamma^2 R^2\big)^3,$$

$$\mathbb{E}\bigg(2\gamma n\big[f(\bar{\theta}_n) - f(\theta^*)\big] + \|\theta_n - \theta_*\|^2\bigg)^4 \leqslant \big(\|\theta_0 - \theta_*\|^2 + 9n\gamma^2 R^2\big)^4.$$

## 3.3 Convergence of gradients

In this section, we prove higher-order bounds on the convergence of the gradient, with rate $O(n^{-1})$ for $\|f'(\bar{\theta}_n)\|^2$:

**Proposition 4** *With constant step size equal to $\gamma$, for any $n \geqslant 0$ and integer $p \in \{1, \dots, \lfloor n/4 \rfloor\}$, we have:*

$$\left(\mathbb{E}\left\|f'\left(\frac{1}{n}\sum_{k=1}^{n}\theta_{k-1}\right)\right\|^{2p}\right)^{1/2p} \leqslant \frac{R}{\sqrt{n}}\left[10\sqrt{p} + 40R^2\gamma p\sqrt{n} + \frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2 + \frac{2}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|\right]. \quad (8)$$

**Corollary 4** *With constant step size equal to $\gamma = \frac{1}{2R^2\sqrt{N}}$, for any integer $p \in \{1, \dots, \lfloor N/4 \rfloor\}$, we have:*

$$\left(\mathbb{E}\left\|f'\left(\frac{1}{N}\sum_{k=1}^{N}\theta_{k-1}\right)\right\|^{2p}\right)^{1/2p} \quad \leqslant \quad \frac{R}{\sqrt{N}}\left[10\sqrt{p} + 20p + 6R^2\|\theta_0-\theta_*\|^2 + 4R\|\theta_0-\theta_*\|\right]. \quad (9)$$

We can make the following observations:

- The squared norm of the gradient $\|f'(\bar{\theta}_n)\|^2$ converges at rate $O(n^{-1})$.
- Given that $(\mathbb{E}\|f'(\bar{\theta}_n)\|^{2p})^{1/2p}$ is affine in $p$, we obtain a subexponential behavior for $\|f'(\bar{\theta}_n)\|$, i.e., tail bounds similar to an exponential distribution.
- The proof of Prop. 4 makes use of the self-concordance assumption (that allows to upperbound deviations of gradients by deviations of function values) and of the proof technique of [4].
- The various terms may be improved for small $p$. In particular, we have, for $p = 1, 2$:

$$\left(\mathbb{E}\left\|f'\left(\frac{1}{n}\sum_{k=1}^{n}\theta_{k-1}\right)\right\|^{2}\right)^{1/2} \leqslant \frac{R}{\sqrt{n}}\left[3 + 2\gamma\sqrt{n}R^2 + \frac{2}{\gamma\sqrt{n}R^2}R\|\theta_0-\theta_*\| + \frac{1}{\gamma\sqrt{n}R^2}R^2\|\theta_0-\theta_*\|^2\right],$$

$$\left(\mathbb{E}\left\|f'\left(\frac{1}{n}\sum_{k=1}^{n}\theta_{k-1}\right)\right\|^{4}\right)^{1/4} \leqslant \frac{R}{\sqrt{n}}\left[5 + 6\gamma\sqrt{n}R^2 + \frac{2}{\gamma\sqrt{n}R^2}R\|\theta_0-\theta_*\| + \frac{1}{\gamma\sqrt{n}R^2}R^2\|\theta_0-\theta_*\|^2\right].$$

# 4 Self-concordance analysis

In the previous section, we have shown that $\|f'(\bar{\theta}_n)\|^2$ is of order $O(n^{-1})$. If the function $f$ was strongly convex with constant $\mu > 0$, this would immediately lead to the bound $f(\bar{\theta}_n) - f(\theta_*) \leqslant \frac{1}{2\mu}\|f'(\bar{\theta}_n)\|^2$, of order $O(\mu^{-1}n^{-1})$. However, because of the Lipschitz-continuity of $f$ on the full Hilbert space $\mathcal{H}$, it cannot be strongly convex. In this section, we show how the self-concordance assumption may be used to obtain the exact same behavior, but with $\mu$ replaced by the *local* strong convexity constant.

The required property is summarized in the following proposition about (generalized) self-concordant function (see proof in Appendix B.1):

**Proposition 5** *Let $f$ be a convex three-times differentiable function from $\mathcal{H}$ to $\mathbb{R}$, such that for all $\theta_1, \theta_2 \in \mathcal{H}$, the function $\varphi : t \mapsto f\big[\theta_1 + t(\theta_2 - \theta_1)\big]$ satifies: $\forall t \in \mathbb{R}$, $|\varphi'''(t)| \leqslant R\|\theta_1 - \theta_2\|\varphi''(t)$. Let $\theta_*$ be a global minimizer of $f$ and $\mu$ the lowest eigenvalue of $f''(\theta_*)$, which is assumed strictly positive.*

$$\textit{If } \frac{\|f'(\theta)\|R}{\mu} \leqslant \frac{3}{4}\ \textit{, then } \|\theta-\theta_*\|^2 \leqslant 4\frac{\|f'(\theta)\|^2}{\mu^2} \textit{ and } f(\theta) - f(\theta_*) \leqslant 2\frac{\|f'(\theta)\|^2}{\mu}.$$

We may now use this proposition for the averaged stochastic gradient. For simplicity, we only consider the step-size $\gamma = \frac{1}{2R^2\sqrt{N}}$, and the last iterate:

**Proposition 6** *Assume $\gamma = \frac{1}{2R^2\sqrt{N}}$. Let $\mu > 0$ be the lowest eigenvalue of the Hessian of $f$ at the unique global optimum $\theta_*$. Then:*

$$\begin{aligned}
\mathbb{E} f(\bar{\theta}_N) - f(\theta_*) &\leqslant \frac{R^2}{N\mu}\Big(5R\|\theta_0 - \theta_*\| + 15\Big)^4, \\
\mathbb{E}\big\|\bar{\theta}_N - \theta_*\big\|^2 &\leqslant \frac{R^2}{N\mu^2}\Big(5R\|\theta_0 - \theta_*\| + 20\Big)^4.
\end{aligned}$$

We can make the following observations:

- The proof relies on Prop. 5 and requires a control of the probability that $\frac{\|f'(\bar{\theta}_N)\|R}{\mu} \leqslant \frac{3}{4}$, which is obtained from Prop. 4.
- We conjecture a bound of the form $\big(\frac{R^2}{N\mu}(\square R\|\theta_0 - \theta_*\| + \triangle\sqrt{p})^4\big)^p$ for the $p$-th order moment of $f(\bar{\theta}_N) - f(\theta_*)$.
- The key elements in the previous proposition are that (a) the constant $\mu$ is the *local* convexity constant, and (b) the step-size does not depend on that constant $\mu$, hence the claimed adaptivity.
- The bounds are only better than the non-strongly-convex bounds from Prop. 1, when the Hessian lowest eigenvalue is large enough, i.e., $\mu R^2\sqrt{N}$ larger than a fixed constant.

# 5 Conclusion

In this paper, we have provided a novel analysis of averaged stochastic gradient for logistic regression and related problems. The key aspects of our result are (a) the adaptivity to local strong convexity provided by averaging and (b) the use of self-concordance to obtain a simple bound that does not involve a term which is exponential in $R\|\theta_0 - \theta_*\|$, which could be obtained by constraining the domain of the iterates.

Our results could be extended in several ways: (a) with a finite and known horizon $N$, we considered a constant step-size proportional to $1/R^2\sqrt{N}$; it thus seems natural to study the decaying step size $\gamma_n = O(1/R^2\sqrt{n})$, which should, up to logarithmic terms, lead to similar results (and thus likely provide a solution to a a recently posed open problem for online logistic regression [28]); (b) an alternative would be to consider a doubling trick where the step-sizes are piecewise constant; Finally, (c) it may be possible to consider other assumptions, such as exp-concavity [17] or uniform convexity [26], to derive similar or improved results.

# A Probability lemmas

In this appendix, we prove lemmas relating bounds on moments to tail bounds, with the traditional use of Markov's inequality.

**Lemma 1** *Let $X$ be a non-negative random variable such that for some positive constants $A$ and $B$, and all $p \in \{1, \dots, n\}$,*

$$\mathbb{E}X^p \leqslant (A + Bp)^p.$$

*Then, if $t \leqslant \frac{n}{2}$,*

$$\mathbb{P}(X \geqslant 3Bt + 2A) \leqslant 2\exp(-t).$$

**Proof.** We have, by Markov's inequality, for any $p \in \{1, \dots, n\}$:

$$\mathbb{P}(X \geqslant 2Bp + 2A) \leqslant \frac{\mathbb{E}X^p}{(2Bp + 2A)^p} \leqslant \frac{(A + Bp)^p}{(2A + 2Bp)^p} \leqslant \exp(-\log(2)p).$$

For $u \in [1, n]$, we consider $p = \lfloor u \rfloor$, so that

$$\mathbb{P}(X \geqslant 2Bu + 2A) \leqslant \mathbb{P}(X \geqslant 2Bp + 2A) \leqslant \exp(-\log(2)p) \leqslant 2\exp(-\log(2)u).$$

We take $t = \log(2)u$ and use $2/\log 2 \leqslant 3$. This is thus valid if $t \leqslant \frac{n}{2}$. ■

**Lemma 2** *Let $X$ be a non-negative random variable such that for some positive constants $A$, $B$ and $C$, and for all $p \in \{1, \dots, n\}$,*

$$\mathbb{E}X^p \leqslant (A\sqrt{p} + Bp + C)^{2p}.$$

*Then, if $t \leqslant n$,*

$$\mathbb{P}(X \geqslant (2A\sqrt{t} + 2Bt + 2C)^2) \leqslant 4\exp(-t).$$

**Proof.** We have, by Markov's inequality, for any $p \in \{1, \dots, n\}$:

$$\mathbb{P}(X \geqslant (2A\sqrt{p} + 2Bp + 2C)^2) \leqslant \frac{\mathbb{E}X^p}{(2A\sqrt{p} + 2Bp + 2C)^{2p}} \leqslant \frac{(A\sqrt{p} + Bp + C)^{2p}}{(2A\sqrt{p} + 2Bp + 2C)^{2p}} \leqslant \exp(-\log(4)p)$$

For $u \in [1, n]$, we consider $p = \lfloor u \rfloor$, so that

$$\mathbb{P}(X \geqslant (2A\sqrt{u} + 2Bu + 2C)^2) \leqslant \mathbb{P}(X \geqslant (2A\sqrt{u} + 2Bu + 2C)^2) \leqslant \exp(-\log(2)p) \leqslant 4\exp(-\log(4)u)$$

We take $t = \log(4)u$ and use $\log 4 \geqslant 1$. This is thus valid if $t \leqslant n$. ■

# B Self-concordance properties

In this appendix, we show two lemmas regarding our generalized notion of self-concordance, as well as Prop. 5. For more details, see [18] and references therein.

**Lemma 3** *Let $\varphi : [0, 1] \to \mathbb{R}$ a convex function such that for some $S > 0$, $\forall t \in [0, 1]$, $|\varphi'''(t)| \leqslant S\varphi''(t)$. Assume $\varphi'(0) = 0$, $\varphi''(0) > 0$. Then:*

$$\frac{\varphi'(1)}{\varphi''(0)} S \geqslant 1 - e^{-S} \text{ and } \varphi(1) \leqslant \varphi(0) + \frac{\varphi'(1)^2}{\varphi''(0)}(1 + S).$$

*Moreover, if* $\alpha = \frac{\varphi'(1)S}{\varphi''(0)} < 1$*, then* $\varphi(1) \leqslant \varphi(0) + \frac{\varphi'(1)^2}{\varphi''(0)} \frac{1}{\alpha} \log \frac{1}{1-\alpha}$*. If in addition* $\alpha \leqslant \frac{3}{4}$*, then* $\varphi(1) \leqslant \varphi(0) + 2\frac{\varphi'(1)^2}{\varphi''(0)}$ *and* $\varphi''(0) \leqslant 2\varphi'(1)$*.*

**Proof.** By self-concordance, we obtain that the derivative of $u \mapsto \log \varphi''(u)$ is lower-bounded by $-S$. By integrating between 0 and $t \in [0,1]$, we get

$$\log \varphi''(t) - \log \varphi''(0) \geqslant -St \text{ , i.e., } \varphi''(t) \geqslant \varphi''(0) e^{-St},$$

and by integrating between 0 and 1, we obtain (note that $\varphi'(0) = 0$):

$$\varphi'(1) \geqslant \varphi''(0) \frac{1 - e^{-S}}{S}. \tag{10}$$

We then get (with a first inequality from convexity of $\varphi$, and the last inequality from $e^S \geqslant 1 + S$):

$$\varphi(1) - \varphi(0) \leqslant \varphi'(1) \leqslant \varphi'(1) \frac{\varphi'(1)}{\varphi''(0)} \frac{S}{1 - e^{-S}} = \frac{\varphi'(1)^2}{\varphi''(0)} \left( S + \frac{S}{e^S - 1} \right) \leqslant \frac{\varphi'(1)^2}{\varphi''(0)} (1 + S).$$

Eq. (10) implies that $\alpha \geqslant 1 - e^{-S}$, which implies, if $\alpha < 1$, $S \leqslant \log \frac{1}{1-\alpha}$. This implies that

$$\varphi(1) - \varphi(0) \leqslant \varphi'(1) \frac{\varphi'(1)}{\varphi''(0)} \frac{S}{1 - e^{-S}} \leqslant \frac{\varphi'(1)^2}{\varphi''(0)} \frac{1}{\alpha} \log \frac{1}{1-\alpha},$$

using the monotinicity of $S \mapsto \frac{S}{1-e^{-S}}$. Finally the last bounds are a consequence of $\frac{S}{\alpha} \leqslant \frac{1}{\alpha} \log \frac{1}{1-\alpha} \leqslant 2$, which is valid for $\alpha \leqslant \frac{3}{4}$.

■

**Lemma 4** *Let* $f$ *be a convex three-times differentiable function from* $\mathcal{H}$ *to* $\mathbb{R}$*, such that for all* $\theta_1, \theta_2 \in \mathcal{H}$*, the function* $\varphi : t \mapsto f\big[\theta_1 + t(\theta_2 - \theta_1)\big]$ *satifies:* $\forall t \in \mathbb{R}$*,* $|\varphi'''(t)| \leqslant R\|\theta_1 - \theta_2\| \varphi''(t)$*. For any* $\theta_1, \theta_2 \in H$*, we have:*

$$\big\| f'(\theta_1) - f'(\theta_2) - f''(\theta_2)(\theta_2 - \theta_1) \big\| \leqslant R\big[ f(\theta_1) - f(\theta_2) - \langle f'(\theta_2), \theta_2 - \theta_1 \rangle \big].$$

**Proof.** For a given $z \in \mathcal{H}$ of unit norm, let $\varphi(t) = \big\langle z, f'\big(\theta_2 + t(\theta_1 - \theta_2)\big) - f'(\theta_2) - t f''(\theta_2)(\theta_2 - \theta_1) \big\rangle$ and $\psi(t) = R\big[ f(\theta_2 + t(\theta_1 - \theta_2)) - f(\theta_2) - t\langle f'(\theta_2), \theta_2 - \theta_1 \rangle \big]$. We have $\varphi(0) = \psi(0) = 0$ and $\varphi'(0) = \psi'(0) = 0$. Moreover, we have $\varphi''(t) \leqslant \psi''(t)$ (using the same reasoning as in the proofs of [18]). We thus have $\varphi(1) \leqslant \psi(1)$, which leads to the desired result by maximizing with respect to $z$. ■

## B.1 Proof of Prop. 5

Define $\varphi : t \mapsto f\big[\theta_* + t(\theta - \theta_*)\big] - f(\theta_*)$. We have: $\varphi(0) = \varphi'(0) = 0$, $0 \leqslant \varphi'(1) = \langle f'(\theta), \theta - \theta_* \rangle \leqslant \|f'(\theta)\| \|\theta - \theta_*\|$, $\varphi''(0) = \langle \theta - \theta_*, f''(\theta_*)(\theta - \theta_*) \rangle \geqslant \mu \|\theta - \theta_*\|^2$, and $\varphi(t) \geqslant 0$ for all $t \in [0,1]$, and $\varphi'''(t) \leqslant R\|\theta - \theta_*\| \varphi''(t)$ for all $t \in [0,1]$, i.e., $S = R\|\theta - \theta_*\|$. Lemma 3 leads to the desired result, with $\alpha = \frac{\varphi'(1)S}{\varphi''(0)} \leqslant \frac{\|f'(\theta)\| R}{\mu}$. Note that we also have, for all $\theta \in \mathcal{H}$,

$$f(\theta) - f(\theta_*) \leqslant \big(1 + R\|\theta - \theta_*\|\big) \frac{\|f'(\theta)\|^2}{\mu} \text{ and } \|\theta - \theta_*\| \leqslant \big(1 + R\|\theta - \theta_*\|\big) \frac{\|f'(\theta)\|}{\mu}.$$

# C Proof of Prop. 2

We consider a direct proof based on taking powers of the inequality in Eq. (2), and then using the appropriate martingale properties.

## C.1 Derivation of recursion

We have the recursion:

$$2\gamma\big[f(\theta_{n-1}) - f(\theta_*)\big] + \|\theta_n - \theta_*\|^2 \quad \leqslant \quad \|\theta_{n-1} - \theta_*\|^2 + \gamma^2 R^2 + M_n,$$

with

$$M_n = -2\gamma\langle \theta_{n-1} - \theta_*, f_n'(\theta_{n-1}) - f'(\theta_{n-1})\rangle.$$

This leads to

$$2\gamma n f\left(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\right) - 2\gamma n f(\theta^*) + \|\theta_n - \theta_*\|^2 \leqslant A_n,$$

with $A_n = \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + \sum_{k=1}^n M_k$. Note that $\mathbb{E}(M_k|\mathcal{F}_{k-1}) = 0$ and $|M_k| \leqslant 4\gamma R\|\theta_{k-1} - \theta_*\| \leqslant 4\gamma R A_{k-1}^{1/2}$ almost surely. This leads to, by using the binomial expansion formula:

$$\begin{aligned} A_n^p &\leqslant \left(A_{n-1} + \gamma^2 R^2 + M_n\right)^p = \sum_{k=0}^p \binom{p}{k}\big(A_{n-1} + \gamma^2 R^2\big)^{p-k} M_n^k \\ &\leqslant \big(A_{n-1} + \gamma^2 R^2\big)^p + \big(A_{n-1} + \gamma^2 R^2\big)^{p-1} M_n + \sum_{k=2}^p \binom{p}{k}\big(A_{n-1} + \gamma^2 R^2\big)^{p-k}\big(4\gamma R A_{n-1}^{1/2}\big)^k. \end{aligned}$$

This leads to (using $E(M_n|\mathcal{F}_{n-1}) = 0$ and upper bounding $\gamma^2 R^2$ by $4\gamma^2 R^2$):

$$\begin{aligned} \mathbb{E}\big[A_n^p\big|\mathcal{F}_{n-1}\big] &\leqslant \big(A_{n-1} + 4\gamma^2 R^2\big)^p + \sum_{k=2}^p \binom{p}{k}\big(A_{n-1} + 4\gamma^2 R^2\big)^{p-k}\big(4\gamma R A_{n-1}^{1/2}\big)^k \\ &= \big(A_{n-1} + 4\gamma^2 R^2 + 4\gamma R A_{n-1}^{1/2}\big)^p - 4\gamma R p\big(A_{n-1} + 4\gamma^2 R^2\big)^{p-1} A_{n-1}^{1/2} \\ &\leqslant \big(A_{n-1}^{1/2} + 2\gamma R\big)^{2p} - 4\gamma R p\big(A_{n-1} + 4\gamma^2 R^2\big)^{p-1} A_{n-1}^{1/2} \\ &= \sum_{k=0}^{2p} \binom{2p}{k} A_{n-1}^{k/2}(2\gamma R)^{2p-k} - 4\gamma R p A_{n-1}^{1/2} \sum_{k=0}^{p-1} \binom{p-1}{k} A_{n-1}^k (2\gamma R)^{2(p-1-k)} \\ &= \sum_{k=0}^{2p} A_{n-1}^{k/2}(2\gamma R)^{2p-k} C_k, \end{aligned}$$

with

$$\begin{aligned} C_{2q} &= \binom{2p}{2q} \text{ for } q \in \{0, \dots, p\}, \\ C_{2q+1} &= \binom{2p}{2q+1} - 2p\binom{p-1}{q} \text{ for } q \in \{0, \dots, p-1\}. \end{aligned}$$

In particular, $C_0 = 1$, $C_{2p} = 1$, $C_1 = 0$ and $C_{2p-1} = \binom{2p}{2p-1} - 2p\binom{p-1}{p-1} = 0$.

We have, for $q \in \{1, \dots, p-2\}$,

$$\begin{aligned} C_{2q+1}\frac{2q+1}{2p-2q-1} &\leqslant \binom{2p}{2q+1}\frac{2q+1}{2p-2q-1} \\ &= \frac{(2p)!}{(2q+1)!(2p-2q-1)!}\frac{2q+1}{2p-2q-1} \\ &= \frac{(2p)!}{(2q)!(2p-2q)!}\frac{2p-2q}{2p-2q-1} = \binom{2p}{2q}\frac{2p-2q}{2p-2q-1}. \end{aligned}$$

For $q = p-2$, we obtain $C_{2q+1}\frac{2q+1}{2p-2q-1} \leqslant C_{2q}\frac{4}{3}$, while for $q \leqslant p-3$, we obtain $C_{2q+1}\frac{2q+1}{2p-2q-1} \leqslant C_{2q}\frac{6}{5}$. Moreover, for $q \in \{1, \dots, p-2\}$,

$$\begin{aligned} C_{2q+1}\frac{2p-2q-1}{2q+1} &\leqslant \binom{2p}{2q+1}\frac{2p-2q-1}{2q+1} \\ &= \frac{(2p)!}{(2q+1)!(2p-2q-1)!}\frac{2p-2q-1}{2q+1} \\ &= \frac{(2p)!}{(2q+2)!(2p-2q-2)!}\frac{2q+2}{2q+1} = \binom{2p}{2q+2}\frac{2q+2}{2q+1}. \end{aligned}$$

For $q = 1$, we obtain $C_{2q+1}\frac{2p-2q-1}{2q+1} \leqslant C_{2q+2}\frac{4}{3}$, while for $q \geqslant 2$, we obtain $C_{2q+1}\frac{2p-2q-1}{2q+1} \leqslant C_{2q+2}\frac{6}{5}$. We have moreover

$$\begin{aligned} & C_{2q+1}A_{n-1}^{q+1/2}(2\gamma R)^{2p-2q-1} \\ = \; & C_{2q+1}A_{n-1}^{q}(2\gamma R)^{2p-2q-2}A_{n-1}^{1/2}(2\gamma R) \\ \leqslant \; & C_{2q+1}A_{n-1}^{q}(2\gamma R)^{2p-2q-2}\frac{1}{2}\left[\frac{2q+1}{2p-2q-1}(2\gamma R)^2 + \frac{2p-2q-1}{2q+1}A_{n-1}\right]A_{n-1}^{1/2}(2\gamma R) \\ = \; & \frac{1}{2}C_{2q+1}\frac{2p-2q-1}{2q+1}A_{n-1}^{q+1}(2\gamma R)^{2p-2q-2} + \frac{1}{2}C_{2q+1}\frac{2q+1}{2p-2q-1}A_{n-1}^{q}(2\gamma R)^{2p-2q}. \end{aligned}$$

By combining all elements, we get that the terms indexed by $2q+1$ are bounded by the terms indexed by $2q+2$ and $2q$. All terms with $q \in \{2, \dots, p-3\}$ are expanded with constants $\frac{3}{5}$, while for $q = 1$ and $q = p-2$, this is $\frac{2}{3}$. Overall each even term receives a contribution which is less than $\max\{\frac{6}{5}, \frac{3}{5}+\frac{2}{3}, \frac{2}{3}\} = \frac{19}{15}$. This leads to

$$\sum_{q=1}^{p-2} C_{2q+1}A_{n-1}^{q+1/2}(2\gamma R)^{2p-2q-1} \leqslant \frac{19}{15}\sum_{q=0}^{p-1} C_{2q}A_{n-1}^{q}(2\gamma R)^{2p-2q},$$

leading to the recursion that will allow us to derive our result:

$$\mathbb{E}\big[A_n^p \big| \mathcal{F}_{n-1}\big] \leqslant A_{n-1}^p + \frac{34}{15}\sum_{q=0}^{p-1}\binom{2p}{2q}A_{n-1}^{q}(2\gamma R)^{2p-2q}. \tag{11}$$

## C.2 First bound

In this section, we derive an almost sure bound that will be valid for small $n$. Since $\|\theta_n - \theta_*\| \leqslant \|\theta_{n-1} - \theta_*\| + \gamma R$ almost surely, we have $\|\theta_n - \theta_*\| \leqslant \|\theta_0 - \theta_*\| + n\gamma R$ for all $n \geqslant 0$. This in turn

implies that

$$\begin{aligned}
A_n &\leqslant \|\theta_0-\theta_*\|^2 + n\gamma^2R^2 + 4\gamma R\sum_{k=1}^n \|\theta_{k-1}-\theta_*\| \\
A_n &\leqslant \|\theta_0-\theta_*\|^2 + n\gamma^2R^2 + 4\gamma R\sum_{k=1}^n \big[\|\theta_0-\theta_*\| + (k-1)\gamma R\big] \\
&\leqslant \|\theta_0-\theta_*\|^2 + n\gamma^2R^2 + 4\gamma nR\|\theta_0-\theta_*\| + 2\gamma^2R^2n^2 \\
&\leqslant \|\theta_0-\theta_*\|^2 + n\gamma^2R^2 + 2\gamma^2n^2R^2 + 2\|\theta_0-\theta_*\|^2 + 2\gamma^2R^2n^2 \\
&\leqslant 3\|\theta_0-\theta_*\|^2 + 5n\gamma^2R^2 \text{ almost surely.} \qquad (12)
\end{aligned}$$

## C.3 Proof by induction

We now proceed by induction on $p$. If we assume that $\mathbb{E}A_k^q \leqslant \big(3\|\theta_0-\theta_*\|^2 + kq\gamma^2R^2A\big)^q$ for $q<p$ and a certain $A$ (which we will take to be equal to 20). We first note that if $n \leqslant 4p$, then from Eq. (12), we have

$$\begin{aligned}
E_{A_n^p} &\leqslant \big(3\|\theta_0-\theta_*\|^2 + 5n^2\gamma^2R^2\ \big)^p \\
&\leqslant \big(3\|\theta_0-\theta_*\|^2 + 20np\gamma^2R^2\ \big)^p.
\end{aligned}$$

Thus, we only need to consider $n \geqslant 4p$. We then get from Eq. (11):

$$\begin{aligned}
\mathbb{E}\|\theta_n-\theta_*\|^{2p} &\leqslant \|\theta_0-\theta_*\|^{2p} + \frac{34}{15}\sum_{k=0}^{n-1}\sum_{q=0}^{p-1}\binom{2p}{2q}\mathbb{E}A_k^q(2\gamma R)^{2p-2q} \\
&\leqslant \|\theta_0-\theta_*\|^{2p} + \frac{34}{15}\sum_{k=0}^{n-1}\sum_{q=0}^{p-1}\binom{2p}{2q}\big(3\|\theta_0-\theta_*\|^2 + kq\gamma^2R^2A\big)^q(2\gamma R)^{2p-2q} \\
&\quad \text{using the induction hypothesis,} \\
&= \|\theta_0-\theta_*\|^{2p} + \frac{34}{15}\sum_{q=0}^{p-1}\binom{2p}{2q}(2\gamma R)^{2p-2q}\sum_{k=0}^{n-1}\big(3\|\theta_0-\theta_*\|^2 + kq\gamma^2R^2A\big)^q \\
&\leqslant \|\theta_0-\theta_*\|^{2p} + \frac{34}{15}\sum_{q=0}^{p-1}\binom{2p}{2q}(2\gamma R)^{2p-2q}\sum_{j=0}^{q}3^j\|\theta_0-\theta_*\|^{2j}\binom{q}{j}\big(q\gamma^2R^2A\big)^{q-j}\frac{n^{q-j+1}}{q-j+1} \\
&\quad \text{using } \sum_{k=0}^{n-1}k^\alpha \leqslant \frac{n^{\alpha+1}}{\alpha+1} \text{ for any } \alpha>0, \\
&= \|\theta_0-\theta_*\|^{2p} + \frac{34}{15}\sum_{j=0}^{p-1}3^j\|\theta_0-\theta_*\|^{2j}(4\gamma^2R^2n)^{p-j}\sum_{q=j}^{p-1}\binom{2p}{2q}\binom{q}{j}\big(\frac{qA}{4}\big)^{q-j}\frac{n^{q-p+1}}{q-j+1}.
\end{aligned}$$

We want to show that it is less than

$$\big(3\|\theta_0-\theta_*\|^2 + kp\gamma^2R^2A\big)^p = 3^p\|\theta_0-\theta_*\|^{2p} + \sum_{j=0}^{p-1}3^j\|\theta_0-\theta_*\|^{2j}(\gamma^2R^2n)^{p-j}(Ap)^{p-j}\binom{p}{j}.$$

By comparing all terms in $\|\theta_0-\theta_*\|^{2j}$, this is true as soon as for all $j \in \{0,\dots,p-1\}$,

$$\frac{34}{15}\sum_{q=j}^{p-1}\binom{2p}{2q}\binom{q}{j}(qA/4)^{q-j}\frac{1}{q-j+1}\frac{1}{n^{p-q-1}} \leqslant (Ap/4)^{p-j}\binom{p}{j}$$

$$\Leftrightarrow \frac{34}{15}\sum_{k=0}^{p-1-j}\binom{2p}{2k+2}\binom{p-1-k}{j}\big((p-1-k)A/4\big)^{p-1-k-j}\frac{1}{p-k-j}\frac{1}{n^k} \leqslant (Ap/4)^{p-j}\binom{p}{j},$$

This is implied by (if $n \geqslant 4p$):

$$\frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}p^{-k-p+j}\binom{2p}{2k+2}\frac{\binom{p-1-k}{j}}{\binom{p}{j}}\big(p-1-k\big)^{p-1-k-j}\frac{1}{p-k-j} \leqslant 1$$

$$\Leftarrow \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}p^{-k-p+j}\binom{2p}{2k+2}\frac{(p-1-k)\cdots(p-k-j+1)}{p\cdots(p-j+1)}\big(p-1-k\big)^{p-1-k-j} \leqslant 1$$

$$\Leftarrow \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}p^{-k-p+j}\binom{2p}{2k+2}\frac{(p-j)\cdots(p-k-j+1)}{p\cdots(p-k)}\big(p-1-k\big)^{p-1-k-j} \leqslant 1.$$

We have

$$\begin{aligned}
&\frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}p^{-k-p+j}\binom{2p}{2k+2}\frac{(p-j)\cdots(p-k-j+1)}{p\cdots(p-k)}\big(p-1-k\big)^{p-1-k-j}\\
\leqslant\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}p^{-k-p+j}\binom{2p}{2k+2}\frac{p^k}{p\cdots(p-k)}p^{p-1-k-j}\\
=\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}p^{-2k-1}\binom{2p}{2k+2}\frac{1}{p\cdots(p-k)}\\
=\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}\frac{p^{-2k-1}}{(2k+2)!}\frac{2p(2p-1)\cdots(2p-2k-1)}{p\cdots(p-k)}\\
=\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}\frac{p^{-2k-1}2^{2k+2}}{(2k+2)!}\frac{p(p-1/2)\cdots(p-k-1/2)}{p\cdots(p-k)}\\
\leqslant\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}\frac{2^{2k+2}}{(2k+2)!}\\
\leqslant\ & \frac{136}{15}\sum_{k=0}^{+\infty}\frac{(2/\sqrt{A})^{2k+2}}{(2k+2)!} = \frac{136}{15}\big[\cosh(2/\sqrt{A})-1\big] < 1 \text{ if } A \leqslant 20.
\end{aligned}$$

We thus get the desired result

$$\mathbb{E}A_n^p \leqslant \big(3\|\theta_0-\theta_*\|^2 + 20np\gamma^2R^2\big)^p.$$

# D Proof of Prop. 4

The proof is organized in two parts: first show a bound on $\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1})$, then relate it to $f'\left(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\right)$ using self-concordance.

## D.1 Bound on $\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1})$

We have, following [4, 16]:

$$f_n'(\theta_{n-1}) = \frac{1}{\gamma}(\theta_{n-1} - \theta_n),$$

which implies, by summing over all integers between 1 and $n$:

$$\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1}) \quad = \quad \frac{1}{n}\sum_{k=1}^n \big[f'(\theta_{k-1}) - f_k'(\theta_{k-1})\big] + \frac{1}{\gamma n}(\theta_0 - \theta_*) + \frac{1}{\gamma n}(\theta_* - \theta_n).$$

We denote $X_k = \frac{1}{n}\big[f'(\theta_{k-1}) - f_k'(\theta_{k-1})\big]$. We have: $\|X_k\| \leqslant \frac{2R}{n}$ and $\mathbb{E}(X_k|\mathcal{F}_{k-1}) = 0$, with $\big(\sum_{k=1}^n \mathbb{E}(\|X_k\|^2|\mathcal{F}_{k-1})\big)^{1/2} \leqslant \frac{2R}{\sqrt{n}}$. We may thus apply the Burkholder-Rosenthal-Pinelis inequality [29, Theorem 4.1], and get:

$$\mathbb{E}\bigg[\bigg\|\frac{1}{n}\sum_{k=1}^n \big[f'(\theta_{k-1}) - f_k'(\theta_{k-1})\big]\bigg\|^{2p}\bigg]^{1/2p} \quad \leqslant \quad 2p\frac{2R}{n} + \sqrt{2p}\frac{2R}{n^{1/2}}.$$

This leads to, with $p \leqslant \lfloor n/4 \rfloor$:

$$\begin{aligned}
\mathbb{E}\bigg[\bigg\|\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1})\bigg\|^{2p}\bigg]^{1/2p} &\leqslant 2p\frac{2R}{n} + \sqrt{2p}\frac{2R}{n^{1/2}} + \frac{1}{\gamma n}\|\theta_0 - \theta_*\| + \big[\frac{1}{\gamma n}\sqrt{\|\theta_0 - \theta_*\|^2 + 18np\gamma^2R^2}\big] \\
&\leqslant 2p\frac{2R}{n} + \sqrt{2p}\frac{2R}{n^{1/2}} + \frac{1}{\gamma n}\|\theta_0 - \theta_*\| + \big[\frac{1}{\gamma n}\|\theta_0 - \theta_*\| + \frac{1}{\gamma n}\sqrt{18np}\gamma R\big] \\
&\leqslant \sqrt{p}\frac{2R}{\sqrt{n}} + \sqrt{2p}\frac{2R}{n^{1/2}} + \frac{2}{\gamma n}\|\theta_0 - \theta_*\| + \frac{1}{\gamma n}\sqrt{18np}\gamma R \text{ using } \sqrt{p} \leqslant \sqrt{n}/2, \\
&\leqslant \sqrt{p}\frac{R}{\sqrt{n}}\big[2 + 2\sqrt{2} + \sqrt{18}\big] + \frac{2}{\gamma n}\|\theta_0 - \theta_*\| \\
&\leqslant 10\sqrt{p}\frac{R}{\sqrt{n}} + \frac{2}{\gamma n}\|\theta_0 - \theta_*\|. \qquad (13)
\end{aligned}$$

## D.2 Using self-concordance

Using the self-concordance property of Lemma 4, we obtain:

$$\begin{aligned}
&\bigg\|\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1}) - f'\bigg(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\bigg)\bigg\| \\
= &\bigg\|\frac{1}{n}\sum_{k=1}^n \big[f'(\theta_{k-1}) - f'(\theta_*) - f''(\theta_*)(\theta_{k-1} - \theta_*)\big] - f'\bigg(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\bigg) + f'(\theta_*) + f''(\theta_*)\bigg(\frac{1}{n}\sum_{k=1}^n \theta_{k-1} - \theta_*\bigg)\bigg\| \\
\leqslant &\frac{R}{n}\sum_{k=1}^n \big[f(\theta_{k-1}) - f(\theta_*) - \langle f'(\theta_*), \theta_{k-1} - \theta_*\rangle\big] + R\bigg[f\bigg(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\bigg) - f(\theta_*) + \bigg\langle f'(\theta_*), \frac{1}{n}\sum_{k=1}^n \theta_{k-1} - \theta_*\bigg\rangle\bigg] \\
\leqslant &2R\bigg(\frac{1}{n}\sum_{k=1}^n f(\theta_{k-1}) - f(\theta_*)\bigg).
\end{aligned}$$

This leads to, using Prop. 2:

$$\begin{aligned}
&\left(\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^{n} f'(\theta_{k-1}) - f'\left(\frac{1}{n}\sum_{k=1}^{n}\theta_{k-1}\right)\right\|^{2p}\right)^{1/2p} \\
\leqslant\ & 2R\left(\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} f(\theta_{k-1}) - f(\theta_*)\right]^{2p}\right)^{1/2p} \leqslant \frac{R}{\gamma n}\left(3\|\theta_0-\theta_*\|^2 + 40np\gamma^2R^2\right). \qquad (14)
\end{aligned}$$

Summing Eq. (13) and Eq. (14) leads to the desired result.

# E Results for small $p$

In Prop. 2, we may replace the bound $3\|\theta_0-\theta_*\|^2+20np\gamma^2R^2$ with a bound with a smaller constant for $p=1,2,3,4$:

$$\begin{aligned}
\mathbb{E}\left[2\gamma n\big[f(\bar{\theta}_n)-f(\theta^*)\big]+\|\theta_n-\theta_*\|^2\right]^2 &\leqslant \left(\|\theta_0-\theta_*\|^2+3n\gamma^2R^2\right)^2 \\
\mathbb{E}\left[2\gamma n\big[f(\bar{\theta}_n)-f(\theta^*)\big]+\|\theta_n-\theta_*\|^2\right]^3 &\leqslant \left(\|\theta_0-\theta_*\|^2+6n\gamma^2R^2\right)^3 \\
\mathbb{E}\left[2\gamma n\big[f(\bar{\theta}_n)-f(\theta^*)\big]+\|\theta_n-\theta_*\|^2\right]^4 &\leqslant \left(\|\theta_0-\theta_*\|^2+9n\gamma^2R^2\right)^4.
\end{aligned}$$

This is done using the same proof principle but finer derivations, as follows. We denote $\gamma^2 R^2 = a$ and $\|\theta - \theta_*\|^2 = b$. We have

$$
\begin{aligned}
EA_n &\leqslant a + nb, \\
EA_n^2 &\leqslant A_{n-1}^2 + 2A_{n-1}b + b^2 + 4bA_{n-1} \\
&\leqslant a^2 + 6b[na + \frac{n^2}{2}b] + b^2 n \\
&= a^2 + 6bna + b^2(n + 3n^2) \\
&\leqslant (a + 3nb)^2, \\
EA_n^3 &\leqslant (A_{n-1}^3 + 3A_{n-1}^2 b + 3A_{n-1}b^2 + b^3) + 3(A_{n-1} + b)4bA_{n-1} + 8b^{3/2}A_{n-1}^{3/2} \\
&= (A_{n-1}^3 + 3A_{n-1}^2 b + 3A_{n-1}b^2 + b^3) + 3(A_{n-1} + b)4bA_{n-1} + 4bA_{n-1}[2b^{/2}A_{n-1}^{1/2}] \\
&\leqslant (A_{n-1}^3 + 3A_{n-1}^2 b + 3A_{n-1}b^2 + b^3) + 3(A_{n-1} + b)4bA_{n-1} + 4bA_{n-1}[\frac{A_{n-1}}{4} + 4b] \\
&= A_{n-1}^3 + A_{n-1}^2 b[3 + 12 + 1] + A_{n-1}b^2[3 + 12 + 16] + b^3 \\
&= A_{n-1}^3 + 16A_{n-1}^2 b + 31A_{n-1}b^2 + b^3 \\
&\leqslant a^3 + 16b[na^2 + 3bn^2 a + b^2(n^2/2 + n^3)] + 31b^2[na + bn^2/2] + nb^3 \\
&= a^3 + 16nba^2 + b^2 a[48n^2 + 31n] + b^3[8n^2 + 16n^3 + 31/2n^2 + n] \\
&= a^3 + 16nba^2 + b^2 a[48n^2 + 31n] + b^3[47/2n^2 + 16n^3 + n] \\
&\leqslant (a + 6nb)^3, \\
\mathbb{E}A_n^4 &\leqslant A_{n-1}^4 + 4A_{n-1}^3 b + 6A_{n-1}^2 b^2 + 4A_{n-1}b^3 + b^4 \\
&\quad +6[A_{n-1}^2 + 2A_{n-1}b + b^2]4bA_{n-1} + 4[A_{n-1} + b]4bA_{n-1}[2b^{1/2}A_{n-1}^{1/2}] \\
&\leqslant A_{n-1}^4 + 4A_{n-1}^3 b + 6A_{n-1}^2 b^2 + 4A_{n-1}b^3 + b^4 \\
&\quad +6[A_{n-1}^2 + 2A_{n-1}b + b^2]4bA_{n-1} + 4[A_{n-1} + b]4bA_{n-1}[\frac{A_{n-1}}{2} + 2b] + 16b^2 A_{n-1}^2 \\
&= A_{n-1}^4 + A_{n-1}^3 b[4 + 24 + 8] + A_{n-1}^2 b^2[6 + 48 + 16 + 8 + 32] + A_{n-1}b^3[4 + 24 + 32] + b^4 \\
&= A_{n-1}^4 + 36A_{n-1}^3 b + 110A_{n-1}^2 b^2 + 60A_{n-1}b^3 + b^4 \\
&\leqslant a^4 + 36b[na^3 + 8n^2 ba^2 + b^2 a(48n^3/3 + 31n^2/2) + b^3(47/6n^3 + 4n^4 + n^2/2] \\
&\quad +110b^2[na^2 + 3bn^2 a + b^2(n^2/2 + n^3)] \\
&\quad +nb^4 + 60b^3[na + bn^2/2] \\
&\leqslant a^4 + 36bna^3 + b^2 n^2 a^2[36 \times 8 + 110] + b^3 n^3 a[36 \times 48/3 + 36 \times 31/2 + 330 + 60] \\
&\quad +b^4 n^4[6 \times 47 + 36 \times 4 + 18 + 55 + 110 + 1 + 30] \\
&\leqslant (a + 9nb)^4.
\end{aligned}
$$

# F Proof of Prop. 6

The proof follows from Prop. 5 applied to $\bar{\theta}_n$. We thus need to provide a control on the probability that $\|f'(\bar{\theta}_n)\| \geqslant \frac{3\mu}{4R}$.

## F.1 Tail bound for $\|f'(\bar{\theta}_n)\|$

We derive a large deviation bound, as a consequence of Prop. 4 and Lemma 2:

$$\mathbb{P}\left(\|f'(\bar{\theta}_n)\| \geqslant \frac{2R}{\sqrt{n}}\left[10\sqrt{t} + 40R^2\gamma t\sqrt{n} + \frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2 + \frac{2}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|\right]\right) \leqslant 4\exp(-t),$$

which is valid as long as $t \leqslant n$ (condition from Lemma 2). It is valid for all $t$, because for all gradients are bounded by $R$.

## F.2 Bounding the function values

From Prop. 5, if $\|f'(\bar{\theta}_n)\| \geqslant \frac{3\mu}{4R}$, then $f(\bar{\theta}_n) - f(\theta_*) \leqslant 2\frac{\|f'(\bar{\theta}_n)\|^2}{\mu}$. This will allow us to derive a tail bound for $f(\bar{\theta}_n) - f(\theta_*)$, for sufficiently small deviations. For larger deviations, we will use Prop. 2.

We consider the event

$$A_t = \left\{\|f'(\bar{\theta}_n)\| \leqslant \frac{2R}{\sqrt{n}}\left[10\sqrt{t} + 40R^2\gamma t\sqrt{n} + \frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2 + \frac{2}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|\right]\right\}.$$

If we have:

$$\begin{aligned} 10\sqrt{t} + 40R^2\gamma t\sqrt{n} &\leqslant \frac{2}{3}\frac{3\mu}{4R}\frac{\sqrt{n}}{2R} = \frac{\mu\sqrt{n}}{4R^2} \\ \text{and } \frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2 + \frac{2}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\| &\leqslant \frac{1}{3}\frac{3\mu}{4R}\frac{\sqrt{n}}{2R} = \frac{\mu\sqrt{n}}{8R^2}, \end{aligned}$$

then, by Prop. 5, we have:

$$\begin{aligned} A_t &\subset \left\{f(\bar{\theta}_n) - f(\theta_*) \leqslant \frac{8R^2}{\mu n}\left[10\sqrt{t} + 40R^2\gamma t\sqrt{n} + \frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2 + \frac{2}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|\right]^2\right\} \\ &\subset \left\{f(\bar{\theta}_n) - f(\theta_*) \leqslant \frac{8R^2}{\mu n}\left[10\sqrt{t} + 20\square t + \triangle\right]^2\right\}, \end{aligned}$$

with $\square = 2\gamma R^2\sqrt{n}$ and $\triangle = \frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2 + \frac{2}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|$.

This implies that for all $t \geqslant 0$, such that $10\sqrt{t} + 20\square t \leqslant \frac{\mu\sqrt{n}}{4R^2}$,

$$\mathbb{P}\left(f(\bar{\theta}_n) - f(\theta_*) \geqslant \frac{8R^2}{\mu n}\left[10\sqrt{t} + 20\square t + \triangle\right]^2\right) \leqslant 4e^{-t}.$$

Moreover, we have for all $t \geqslant 0$ (from Prop. 2):

$$\mathbb{P}\left(f(\bar{\theta}_n) - f(\theta_*) \geqslant 30\gamma R^2 t + \frac{3\|\theta_0-\theta_*\|^2}{\gamma n}\right) \leqslant 2\exp(-t).$$

We may now use the last two inequalities to bound the expectation $\mathbb{E}[f(\bar{\theta}_n) - f(\theta_*)]$.

We have:

$$
\begin{aligned}
\mathbb{E}\big[f(\bar{\theta}_n) - f(\theta_*)\big] &= \int_0^{+\infty} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du \\
&= \int_0^{\triangle^2 \frac{8R^2}{\mu n}} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du + \int_{\triangle^2 \frac{8R^2}{\mu n}}^{\frac{8R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du \\
&\quad + \int_{\frac{8R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}^{+\infty} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du \\
&\leqslant \int_0^{\triangle^2 \frac{8R^2}{\mu n}} du + \int_0^{\infty} 4e^{-t} d\left(\frac{8R^2}{\mu n}\left[10\sqrt{t} + 20\square t + \triangle\right]^2\right) \\
&\quad +2\int_{\frac{4R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2 - \frac{3}{\gamma n}\|\theta_0-\theta_*\|^2}^{+\infty} \exp\big(-\frac{u}{30\gamma R^2}\big)du \text{ using the two tail bounds,} \\
&= \triangle^2 \frac{8R^2}{n\mu} + \frac{32R^2}{\mu n}\int_0^{\infty} e^{-t}\left(100 + 400\square^2 2t + 400\square\frac{3}{2}t^{1/2} + 20\triangle\frac{1}{2}t^{-1/2} + 40\triangle\square\right)dt \\
&\quad +60\gamma R^2 \exp\left(-\frac{1}{30\gamma R^2}\left[\frac{4R^2}{\mu n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2 - \frac{3}{\gamma n}\|\theta_0-\theta_*\|^2\right]\right) \\
&\leqslant \triangle^2 \frac{8R^2}{n\mu} + \frac{32R^2}{\mu n}\left(100\Gamma(1) + 400\square^2 2\Gamma(2) + 400\square\frac{3}{2}\Gamma(3/2) + 20\triangle\frac{1}{2}\Gamma(1/2) + 40\triangle\square\Gamma(1)\right) \\
&\quad +60\gamma R^2 \exp\left(-\frac{1}{30\gamma R^2}\left[\frac{4R^2}{\mu n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2 - \frac{\mu}{8R^2}\right]\right) \text{ using } \frac{3}{\gamma n}\|\theta_0-\theta_*\|^2 \leqslant \frac{\mu}{8R^2}, \\
&\quad \text{with } \Gamma \text{ denoting the Gamma function,} \\
&\leqslant \triangle^2 \frac{8R^2}{n\mu} + \frac{32R^2}{\mu n}\left(100 + 400\square^2 2 + 400\square\frac{3}{2}\frac{1}{2}\sqrt{\pi} + 20\triangle\frac{1}{2}\sqrt{\pi} + 40\triangle\square\right) \\
&\quad +60\gamma R^2 \exp\left(-\frac{1}{30\gamma R^2}\left[\frac{\mu}{8R^2}\right]\right) \\
&\leqslant \triangle^2 \frac{8R^2}{n\mu} + \frac{32R^2}{\mu n}\left(100 + 400\square^2 2 + 400\square\frac{3}{2}\frac{1}{2}\sqrt{\pi} + 20\triangle\frac{1}{2}\sqrt{\pi} + 40\triangle\square\right) \\
&\quad +60\gamma R^2 \frac{1}{2\mu} 30\times 8\gamma R^4 \text{ using } e^{-\alpha} \leqslant \frac{1}{2\alpha} \text{ for all } \alpha > 0, \\
&\leqslant \frac{32R^2}{n\mu}\left[\frac{1}{4}\triangle^2 + 100 + 800\square^2 + 532\square + 40\triangle\square + 57\square^2\right].
\end{aligned}
$$

For $\gamma = \frac{1}{2R^2\sqrt{N}}$, with $\alpha = R\|\theta_0 - \theta_*\|$, $\square = 1$ and $\triangle = 6\alpha^2 + 4\alpha$, we get

$$
\begin{aligned}
\mathbb{E}\big[f(\bar{\theta}_N) - f(\theta_*)\big] &\leqslant \frac{32R^2}{N\mu}\left[\frac{1}{4}\triangle^2 + 1489 + 40\Delta\right] \\
&\leqslant \frac{32R^2}{N\mu}\left[9\alpha^4 + 12\alpha^3 + 4\alpha^2 + 1489 + 240\alpha^2 + 160\alpha\right] \\
&\leqslant \frac{R^2}{N\mu}\big(5\alpha + 15\big)^4.
\end{aligned}
$$

Note that the previous bound is valid if $\frac{3}{\gamma\sqrt{n}}\|\theta_0 - \theta_*\|^2 + \frac{2}{\gamma R\sqrt{n}}\|\theta_0 - \theta_*\| \leqslant \frac{\mu\sqrt{n}}{8R^2}$, i.e., under the condition $6R^2\|\theta_0 - \theta_*\|^2 + 4R\|\theta_0 - \theta_*\| \leqslant \frac{\mu\sqrt{N}}{8R^2}$. If the condition is not satisfied, then the bound is still valid because of Prop. 1. We thus obtain the desired result.

## F.3 Bound on iterates

Following the same principle as for function values, we have:

$$
\begin{aligned}
\mathbb{E}\|\bar{\theta}_n - \theta_*\|^2 &\leqslant \int_0^{\frac{16R^2}{\mu^2 n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2} \mathbb{P}\big[\|\bar{\theta}_n - \theta_*\|^2 \geqslant u\big]du + \int_{\frac{16R^2}{\mu^2 n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}^{\infty} \mathbb{P}\big[\|\bar{\theta}_n - \theta_*\|^2 \geqslant u\big]du \\
&\leqslant \int_0^{\frac{16R^2}{\mu^2 n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2} \mathbb{P}\big[\|\bar{\theta}_n - \theta_*\|^2 \geqslant u\big]du + \mathbb{E}\left[1_{\|\bar{\theta}_n-\theta_*\|^2 \geqslant \frac{16R^2}{\mu^2 n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}\|\bar{\theta}_n - \theta_*\|^2\right] \\
&\leqslant \int_0^{\frac{16R^2}{\mu^2 n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2} \mathbb{P}\big[\|\bar{\theta}_n - \theta_*\|^2 \geqslant u\big]du \\
&\quad + \mathbb{P}\left[\|\bar{\theta}_n - \theta_*\|^2 \geqslant \frac{16R^2}{\mu^2 n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2\right]^{1/2}\left[\mathbb{E}\big(\|\bar{\theta}_n - \theta_*\|^4\big)\right]^{1/2} \\
&\quad \text{using Cauchy-Schwarz inequality,} \\
&\leqslant \int_0^{\frac{16R^2}{\mu^2 n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2} \mathbb{P}\big[\|\bar{\theta}_n - \theta_*\|^2 \geqslant u\big]du \\
&\quad + \mathbb{P}\left[\|\bar{\theta}_n - \theta_*\|^2 \geqslant \frac{16R^2}{\mu^2 n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2\right]^{1/2}\Big(\|\theta_0 - \theta_*\|^2 + 3\gamma^2 n R^2\Big) \text{ using Prop. 2.}
\end{aligned}
$$

Moreover, if we denote by $t$ the largest solution of $10\sqrt{t} + 20\square t = \frac{\mu\sqrt{n}}{4R^2}$, we have:

$$
\begin{aligned}
\sqrt{t} &= \frac{-10 + \sqrt{100 + 20\square\frac{\mu\sqrt{n}}{R}}}{40\square} = \frac{-10 + 10\sqrt{1 + 20\square\frac{\mu\sqrt{n}}{100R}}}{40\square} \\
&\geqslant \frac{9}{40\square}\sqrt{20\square\frac{\mu\sqrt{n}}{100R}},
\end{aligned}
$$

as soon as $20\square\frac{\mu\sqrt{n}}{100R} \geqslant 100$, since if $q \geqslant 100$, $-1 + \sqrt{1+q} \leqslant \frac{9}{10}\sqrt{t}$.

This leads to:

$$
\begin{aligned}
\mathbb{E}\|\bar{\theta}_n - \theta_*\|^2 &\leqslant \int_0^{\triangle^2 \frac{16R^2}{\mu^2 n}} \mathbb{P}\big[\|\bar{\theta}_n - \theta_*\|^2 \geqslant u\big] du + \int_{\triangle^2 \frac{16R^2}{\mu^2 n}}^{\frac{16R^2}{\mu^2 n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2} \mathbb{P}\big[\|\bar{\theta}_n - \theta_*\|^2 \geqslant u\big] du \\
&\quad + 2\exp\big(-\frac{t}{2}\big)\Big(\|\theta_0 - \theta_*\|^2 + 3\gamma^2 n R^2\Big) \text{ with } t \geqslant \big(\frac{9}{40\square}\big)^2 \frac{20\square}{100}\frac{\mu\sqrt{n}}{R^2}, \text{ using Prop. 5,} \\
&\leqslant \triangle^2 \frac{16R^2}{n\mu^2} + \int_0^\infty 4e^{-t} d\Big(\frac{16R^2}{\mu^2 n}\Big[10\sqrt{t} + 20\square t + \triangle\Big]^2\Big) \\
&\quad + \frac{9}{2t^2}\Big(\|\theta_0 - \theta_*\|^2 + 3\gamma^2 n R^2\Big) \text{ using } \exp(-\alpha) \leqslant \frac{9}{16\alpha^2} \text{ for all } \alpha > 0, \\
&\leqslant \triangle^2 \frac{16R^2}{n\mu^2} + \frac{64R^2}{\mu^2 n}\int_0^\infty e^{-t}\Big(100 + 400\square^2 2t + 400\square\frac{3}{2}t^{1/2} + 20\triangle\frac{1}{2}t^{-1/2} + 40\triangle\square\Big)dt \\
&\quad + \frac{9}{2}\frac{40^4\square^4 100^2 R^4}{9^4 20^2 \square^2 \mu^2 n}\Big[\frac{3}{4}\square^2/R^2 + \frac{1}{2R^2}\triangle\Big] \\
&\leqslant \triangle^2 \frac{16R^2}{n\mu^2} + \frac{64R^2}{\mu^2 n}\Big(100\Gamma(1) + 400\square^2 2\Gamma(2) + 400\square\frac{3}{2}\Gamma(3/2) + 20\triangle\frac{1}{2}\Gamma(1/2) + 40\triangle\square\Gamma(1)\Big)dt \\
&\quad + 686 \times 64\frac{\square^2 R^2}{\mu^2 n}\Big[\frac{3}{4}\square^2 + \frac{1}{2}\triangle\Big] \\
&\leqslant \triangle^2 \frac{16R^2}{n\mu^2} + \frac{64R^2}{\mu^2 n}\Big(100 + 400\square^2 2 + 400\square\frac{3}{2}\frac{1}{2}\sqrt{\pi} + 20\triangle\frac{1}{2}\sqrt{\pi} + 40\triangle\square\Big) \\
&\quad + 686 \times 64\frac{\square^2 R^2}{\mu^2 n}\Big[\frac{3}{4}\square^2 + \frac{1}{2}\triangle\Big] \\
&\leqslant \frac{64R^2}{n\mu^2}\Big[\frac{1}{4}\triangle^2 + 100 + 800\square^2 + 532\square + 32\triangle + 40\triangle\square + 686\frac{3}{4}\square^4 + 686\frac{\triangle\square^2}{2}\Big].
\end{aligned}
$$

For $\gamma = \frac{1}{2R^2\sqrt{N}}$, with $\alpha = R\|\theta_0 - \theta_*\|$, $\square = 1$ and $\triangle = 2\alpha^2 + 4\alpha$, we get

$$
\begin{aligned}
\mathbb{E}\|\theta_N - \theta_*\|^2 &\leqslant \frac{8R^2}{N\mu^2}\Big[2\triangle^2 + 8\triangle(32 + 40 + 343) + 8(100 + 800 + 532 + 515)\Big] \\
&\leqslant \frac{8R^2}{N\mu^2}\Big[2\triangle^2 + 3320\triangle + 15576\Big] \\
&\leqslant \frac{8R^2}{N\mu^2}\Big[8\alpha^4 + 16\alpha^3 + 32\alpha^2 + 3320 \times 2\alpha^2 + 3320 \times 4\alpha + 15576\Big] \\
&\leqslant \frac{R^2}{N\mu^2}\big(5\alpha + 20\big)^4.
\end{aligned}
$$

The previous bound is valid as long as $\frac{\mu\sqrt{N}}{R} \geqslant \frac{10000}{20} = 500$. If it is not satisfied, then Prop. 1 shows that it is still valid.

# References


[1] M. N. Broadie, D. M. Cicek, and A. Zeevi. General bounds and finite-time improvement for stochastic approximation algorithms. Technical report, Columbia University, 2009.

[2] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer-Verlag, second edition, 2003.

[3] O. Yu. Kul′chitskiĭ and A. È. Mozgovoĭ. An estimate for the rate of convergence of recurrent robust identification algorithms. *Kibernet. i Vychisl. Tekhn.*, 89:36–39, 1991.

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

[5] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report 781, Cornell University Operations Research and Industrial Engineering, 1988.

[6] V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

[7] Y. Nesterov and J. P. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

[8] 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.

[9] L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.

[10] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Adv. NIPS*, 2008.

[11] S. Shalev-Shwartz and N. Srebro. SVM optimization: inverse dependence on training set size. In *Proc. ICML*, 2008.

[12] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proc. ICML*, 2007.

[13] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *proc. COLT*, 2009.

[14] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 9:2543–2596, 2010.

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

[16] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Adv. NIPS*, 2011.

[17] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proc. COLT*, 2001.

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

[19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, 2001.

[20] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley & Sons, 1983.

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

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

[23] S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an o (1/t) convergence rate for projected stochastic subgradient descent. Technical Report 1212.2002, ArXiv, 2012.

[24] N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Adv. NIPS*, 2012.

[25] Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

[26] A. Juditsky and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. Technical Report 00508933, HAL, 2010.

[27] S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Adv. NIPS*, 2009.

[28] H. B. McMahan and M. Streeter. Open problem: Better bounds for online logistic regression. In *COLT/ICML Joint Open Problem Session*, 2012.

[29] I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):pp. 1679–1706, 1994.