Course notes for Asymptotic and Non-Asymptotic Analyses of Stochastic Approximation Algorithms (with applications for machine learning)

Théophile Cantelobre

Sorbonne Université
Mines ParisTech

This document contains a presentation of convergence analyses of Stochastic Gradient Descent and its averaged counterpart, with a focus on non-asymptotic bounds. It is based on Bach and Moulines (2013) but makes use of other references.

A complete proof of a non-asymptotic bound for SGD under strong convexity is presented. The results in Bach and Moulines (2013) are discussed to examine the effect of strong-convexity and averaging on non-asymptotic convergence rates. They are compared and contrasted to aymptotic analyses (based on course notes by Antoine Godichon-Baggioni Godichon-Baggioni (2021)).

Finally, numerical experiments are presented and the corresponding source code is available at https://github.com/theophilec/stochastic-optimisation.

These notes contain exercises in the form of verification of hypotheses for cocnrete models and proofs of technical lemmas. Solutions are given at the end of the text.

This presentation is self-contained, but some pre-requisites are familiarity with basic convex analysis and probability theory.

Contents

1 Introduction

In these notes, we consider the following problem

$$\min_{\theta \in \mathbb{R}^d} f(\theta)$$

where f may not be in closed-form (for example, a complex expectation) and we only have access to stochastic estimates of its gradients (for example, corrupted by noise).

Before making the setting precise in Section 2, consider the following motivating examples:

Stochastic approximation Historically Stochastic approximation was the driving force in the development of iterative methods for stochastic optimization. In this setting at each iteration we observe $\nabla f_n(\theta) = \nabla f(\theta) + \varepsilon_n$ where ε_n is conditionally centered and has a second moment.

Online learning In the online learning setting, the goal is to minimize the risk of a predictor θ over the data-generating distribution. If ℓ is the corresponding error metric or *loss* then this problem can be formalized as finding an estimator θ_n of θ^*, a solution of the following optimisation problem

$$\min_{\theta \in \mathcal{H}} f(\theta) := \mathbb{E}\left[\ell(\theta, X)\right] \tag{1}$$

from a stream $x_1, \ldots, x_n$ of i.i.d. observation of the random variable X.

1.1 Iterative methods

Minimizing a sufficiently regular function is a common task in statistics and machine learning, for example to estimate a parameter, the median of a set of points, ... However, an iterative method is necessary when full information is not available (for example f has no closed form or only noisy gradients are available). Such a method is also interesting when a closed form solution is not available, in a streaming context (when predictions are needed as data arrives in real-time) or even if the problem does not fit in memory.

In these notes, we consider two algorithms: Stochastic Gradient Descent (the Robbins-Monro algorithm Robbins and Monro (1951)) and its averaged counterpart (so-called Polyak averaging Ruppert (1988); Polyak (1990)). We introduce both of these recursive algorithms precisely in Section 2 but introduce an intuition behind SGD in the online learning setting.

Given a stream of iid samples $x_1, ..., x_n$, SGD is defined by the initial iterate θ_0 (which can also be random) and the recursion (where γ_n is a sequence on decreasing non-negative step sizes):

$$\forall n \geq k \geq 1,\ \theta_k = \theta_{k-1} - \gamma_k \nabla_\theta \ell(x_k, \theta_{k-1})$$

where γ_k is of the form $Ck^{-\alpha}$, $C > 0$.

In other words, at each step, SGD takes a gradient step in the "right direction" for the sample x_k.

Stochastic Gradient Descent knows many variations such as its Polyak-averaged counterpart, in which we consider the estimator $\bar{\theta}$ the average of the iterates. Indeed, if (θ_n) is generated from the above recursion (under certain hypotheses made precise below), the Polyak-averaged estimator is defined for any $n \geq 1$ by

$$\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \tag{2}$$

.

Polyak averaging is by no means the only alternative to SGD. Others include distributed methods (see e.g., Recht et al. (2011)) or variance-reduced methods (see Gower et al. (2020) for a survey), which we do not cover in these notes. Finally, it is interesting to note that SGD and variants have met great practical success, including with non-convex objectives common in machine learning.

Broadly speaking, there are several ways of studying the convergence of SGD for convex stochastic optimisation. In these notes, we focus on two approaches: the non-asymptotic approach and the asymptotic approach. In the following sections we present these two methods (and related concepts such as the oracle model) at a high level. Of course, a major difference between both approaches is self-explanatory: asymptotic results are valid in the limit of $n \to \infty$ while non-asymptotic results hold for all n. Thus, we'll see that SGD and variants demonstrate transient behavior while n is small, which is captured by non-asymptotic results.

1.2 Oracle model, minimax rates & the non-asymptotic approach

In order to tie together the problem formulation (which is agnostic to the method used to solve it) and the choice to use an iterative method (motivated by the stochastic approximation and online learning formulations), we introduce the first-order oracle model for convex optimization, as in Agarwal et al. (2009). Without introducing too much notation, for a query point θ and a convex function f, the first-order stochastic oracle for convex optimisation yields noisy estimates of $f(\theta)$ and $\nabla f(\theta)$. The chosen optimisation method (for example, SGD) then chooses the next query point (i.e. the next estimate of an optimal point). Now, $\epsilon(\mathcal{M}, f, S) = f(\theta_n) - f(\theta^*) \geq 0$ is the optimization error after n queries with method $\mathcal{M}$ over convex set S.

In these notes we focus on upper-bounds on ϵ where $\mathcal{M}$ is the SGD method given properties on f. Agarwal et al. (2009) shows that if S is the ℓ_∞-unit ball in $\mathbb{R}^d$ and $\mathcal{F}^C$ is the set of convex 1-Lipschitz functions and $\mathcal{F}^S$ the set of strongly-convex and bounded functions, then

$$\inf_{\mathcal{M} \in \mathbb{M}} \sup_{f \in \mathcal{F}^C} \mathbb{E}\left[\epsilon(\mathcal{M}, f, S)\right] \geq c\sqrt{\frac{d}{n}} \tag{3}$$

$$\inf_{\mathcal{M} \in \mathbb{M}} \sup_{f \in \mathcal{F}^S} \mathbb{E}\left[\epsilon(\mathcal{M}, f, S)\right] \geq c\frac{d}{n} \tag{4}$$

where $\mathbb{M}$ is the set of optimization methods. In other words, even if we could adapt our method to the function f we are trying to optimize we could not do any better than $O\left(n^{-1/2}\right)$ or $O\left(n^{-1}\right)$ depending on the regularity class we consider.

A non-asymptotic analysis provides bounds on $\mathbb{E}\left\|x_n - x^*\right\|^2$ and $\mathbb{E}f(x_n) - f(x^*)$. Using the non-asymptotic approach, we'll show that the bound in Eqs. (3) and (4) is achieved (up to constants) for SGD and Polyak-averaged SGD under certain hypotheses. This shows that these methods are minimax optimal for convex optimisation with n queries to a stochastic first-order oracle, and that the non-asymptotic results we present are tight.

1.3 Asymptotic approach & asymptotic efficiency

Although non-asymptotic results show that stochastic gradient estimators θ_n converge in quadratic mean to θ^* and at a certain speed, they do not *a priori* prove from them that θ_n converges almost surely to θ^* (and at which speed), which can be studied through an asymptotic analysis.

Furthermore, we can consider an asymptotic efficiency result related to the first-order oracle. If we have n iid observations, then we can consider the so-called M-estimator defined by:

$$\theta_n = \arg\min_{\theta \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} f_i(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, x_i) \tag{5}$$

Note that in the maching learning community, this choice of estimator is known as Empirical Risk Minimization.

Considering the M-estimator is interesting for two reasons.

1. First, if we were to accumulate data in the online setting, we would be able to do at least as well as the M-estimator. Thus, we would want SGD and variants to achieve the same asymptotic efficiency.

2. Second, because under certain regularity hypotheses, the M-estimator is asymptotically efficient. We state Theorem 1.1 adapted from Godichon-Baggioni (2021) without proof (the interested reader should consult (Van Der Vaart, 1998) for more details). Thus, regarding the previous point, if SGD achieves asymptotic normality with the same covariance matrix, then it is an asymptotically efficient method.

Theorem 1.1. *Let θ_n be the M-estimator for ℓ with random variable X.*

Assume :

- *θ_n converges in probability to θ*
- *For almost all $x \in \mathcal{X}$, $\ell(\cdot, \theta)$ is twice continuously differentiable,*
- *For almost all $x \in \mathcal{X}$, its Hessian $\nabla_\theta^2 \ell$ is $L(x)$-Lipschitz (in operator norm),*
- *$L(X)$ and $\nabla_\theta^2 \ell(X, \theta^*)$ have second moments,*
- *$H = \nabla^2 f(\theta^*)$ is inversible.*

Then,

$$\sqrt{n}(\theta_n - \theta^*) \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{N}(0, H^{-1}\Sigma H) \tag{6}$$

where

$$\Sigma = \mathbb{E}\left[\nabla \ell(X, \theta^*) \nabla \ell(X, \theta^*)^T\right]. \tag{7}$$

1.4 Organization

We will progress in the following order. First, we introduce a formalisation of the stochastic optimisation problem and introduce central hypotheses in notations in Section 2. In Section 3 we present and discuss a non-asymptotic bound for SGD with strongly-convex functions. We prove the result in Section 4. We discuss the effects of strong convexity and averaging on rates and robustness to learning rate choice in Section 5. Finally, in Section 6, we illustrate the results presented in these notes in numerical experiments.

While this document is self-contained, the proofs of some claims are left as an exercise to the reader. They are marked with the ❓ symbol. All solutions are deffered to Section 7, but we encourage the reader to try and prove them before looking at the solution.

2 Problem statement & Notation

In this section, we give a formal problem statement and make all our notations precise. In particular, we introduce some assumptions and discuss their relevance in the machine learning context.

2.1 Stochastic optimization with first order oracle

We consider Eq. (8), which contains both the We consider the following problem:

$$\min_{\theta \in \mathcal{H}} f(\theta) \tag{8}$$

where $\mathcal{H} = \mathbb{R}^d$. Note that the results presented in these notes hold for $\mathcal{H}$ an arbitrary Hilbert space. This generalization is of interest in the context of kernel methods in machine learning, in which cas $\mathcal{H}$ is the Reproducing Kernel Hilbert Space (RKHS) of a positive definite kernel. The interested reader can consult Shawe-Taylor and Cristianini (2004) or Mairal and Vert (2021) for an introduction to kernel methods in the context of machine learning.

We saw that two ways of seeing Eq. (8) were an intractable function f or access to noisy gradients of f (or both). Our goal is to study the convergence of the Stochastic Gradient Descent (SGD) estimator θ_n (and its averaged version $\bar{\theta}_n$ counterpart) to θ^*, a minimizer of f.

First, let us precisely introduce the stochastic setting we use, in a generic framework (Bach and Moulines, 2013). Then, we explain how the concrete situations we presented in the introduction fit into it.

Assumption 2.1. *Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration. Let θ_0 be $\mathcal{F}_0$-measurable. Then, assume that $\forall n \geq 1$, f_n is convex and differentiable, ∇f_n is $\mathcal{F}_n$-measurable, square-integrable, and*

$$\forall \theta \in \mathcal{H},\ \mathbb{E}\left[f_n(\theta) | \mathcal{F}_{n-1}\right] = f(\theta)\ a.s. \qquad \forall \theta \in \mathcal{H},\ \mathbb{E}\left[\nabla f_n(\theta) | \mathcal{F}_{n-1}\right] = \nabla f(\theta)\ a.s. \tag{9}$$

Given Assumption 2.1[1], we can introduce the sequence of estimators $(\theta_n)_n$ defined by following iteration:

$$\theta_n = \theta_{n-1} - \gamma_n \nabla f_n(\theta_{n-1}), \tag{10}$$

and where θ_0 is $\mathcal{F}_0$-measurable.

Remark 2.2. *Let us note several things about Assumption 2.1 and the iteration defined in Eq. (10).*

1. *If $\mathcal{F}_0$ is larger than the trivial σ-algebra, then θ_0 can be random. We'll return to the (non)-importance of the choice of θ_0 below.*

2. *Eq. (10) is known both as Stochastic Gradient Descent (in the optimization and machine learning communities) and the Robbins-Monro algorithm (in the stochastic approximation community). Note the resemblance with gradient descent for convex functions.*

3. *We previously remarked that an intuition behind SGD is to consider that at each iteration a step is taken in the "right direction" for f_n (e.g. for a data point x_n). In fact, the step is in the "right direction" on average (conditionally on $\mathcal{F}_n$), since θ_n is $\mathcal{F}_n$ measurable:*

$$\mathbb{E}\left[\theta_{n+1} | \mathcal{F}_n\right] = \theta_n - \gamma_{n+1} \nabla f(\theta_n)$$

Example 2.3 (Online learning (cont.))**.** *If $\ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}^+$ is a loss function then $f_n(\theta) = \ell(\theta, X_n)$ verifies Assumption 2.1 with $\mathcal{F}_n = \sigma(X_1, ..., X_n)$ by independence and assuming we can interchange derivation and expectation.*

Example 2.4 (Stochastic approximation (cont.))**.** *Let $f_n(\theta) = f(\theta) + \langle \varepsilon_n, \theta \rangle$ where $\mathbb{E}\left[\varepsilon_n | \mathcal{F}_{n-1}\right] = 0$. Then we observe noisy gradients, i.e. $\nabla f_n(\theta) = \nabla f(\theta) + \varepsilon_n$.*

2.2 Smoothness assumptions

Recall that a differentiable function g is L-smooth ($L > 0$) if its gradients are L-Lipschitz, i.e. if

$$\forall x, y, \ \left\| \nabla g(x) - \nabla g(y) \right\| \leq L \left\| x - y \right\|. \tag{11}$$

We relax this assumption in two alternative ways in the stochastic setting: smoothness in quadratic mean (which implies that f is L-smooth, see remark) and almost sure L-smoothness for f_n.

Assumption 2.5.

$$\forall \theta_1, \theta_2 \in \mathcal{H}, \ \mathbb{E}\left[\left\| \nabla f_n(\theta_1) - \nabla f_n(\theta_2) \right\|^2 | \mathcal{F}_{n-1} \right] \leq L^2 \left\| \theta_1 - \theta_2 \right\|^2 \ a.s.$$

Remark 2.6. *By convexity of $x \mapsto \|x\|^2$ and Jensen's inequality, we can see that Assumption 2.5 implies that f is L-smooth:*

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)|\|^2 = \|\mathbb{E}\left[\nabla f_n(\theta_1) - \nabla f_n(\theta_2) | \mathcal{F}_{n-1}\right]\|^2 \leq \mathbb{E}\left[\left\| \nabla f_n(\theta_1) - \nabla f_n(\theta_2) \right\|^2 | \mathcal{F}_{n-1} \right] \ a.s. \tag{12}$$

However, note that this hypothesis is much more relevant because it can be verified when, for example, f_n has a known expression (as is the case for least-squares and logistic regression, see below). Because f is L-smooth, we could apply gradient descent to solve Eq. (8) if we were able to compute its gradient... we'll examine what results we obtain with gradient estimates ∇f_n.

Assumption 2.7.

$$\forall \theta_1, \theta_2 \in \mathcal{H}, \ \left\| \nabla f_n(\theta_1) - \nabla f_n(\theta_2) \right\|^2 \leq L^2 \left\| \theta_1 - \theta_2 \right\|^2 \ a.s.$$

Remark 2.8 (Translation when f (resp. f_n) is twice-differentiable)**.** *If f (resp. f_n) is twice-differentiable, then Assumption 2.5 (resp. Assumption 2.7) is equivalent to $H \preceq LId$ (resp. $H_n \preceq LId$) where H (resp. H_n) is the Hessian of f (resp. f_n).*

[1] Note that we added the conditional unbiasedness of f_n as it is implicit in Bach and Moulines (2013) and is useful in Section 7.1. However, it changes nothing in terms of the algorithmic nature of f_n, since only the gradients are used.

❓ When are Assumption 2.5 and Assumption 2.7 verified?

- Least-squares regression: $f_n(\theta) = 0.5(\langle x_n, \theta\rangle - y_n)^2$, i.e. $\nabla f_n(\theta) = x_n(\langle x_n, \theta\rangle - y_n)$. $\|\nabla f_n(\theta_1) - \nabla f_n(\theta_2)\|^2 = \|x_n\|^2 \langle x_n, \theta_1 - \theta_2\rangle^2 \leq \|x_n\|^4 \|\theta_1 - \theta_2\|^2$. So with $L = \left(\mathbb{E}\|x_n\|^4\right)^{1/2}$, least-squares regression verifies Assumption 2.5. Assumption 2.5 is verified if L is such that almost surely $\|x_n\|^4 \leq L^2$.

- Logistic regression: $f_n(\theta) = \log(1 + \exp(-y_n\langle x_n, \theta\rangle))$, i.e. $\nabla f_n(\theta) = \frac{-y_n x_n \exp(-y_n\langle x_n,\theta\rangle)}{1+\exp(-y_n\langle x_n,\theta\rangle)}$ and $\|\nabla f_n(\theta_1) - \nabla f_n(\theta_2)\|^2 = \left|\frac{\exp(-y_n\langle x_n,\theta_1\rangle)}{1+\exp(-y_n\langle x_n,\theta_1\rangle)} - \frac{\exp(-y_n\langle x_n,\theta_2\rangle)}{1+\exp(-y_n\langle x_n,\theta_2\rangle)}\right|^2 \|x_n\|^2 \leq \frac{1}{4}|\langle x_n, \theta_1 - \theta_2\rangle|^2 \|x_n\|^2 \leq \frac{\|x_n\|^2}{4}\|\theta_1 - \theta_2\|^2$ because $\pi(x)$ is $\frac{1}{2}$-Lipschitz.

  Thus, Assumption 2.5 is verified for $L = \frac{\sqrt{\mathbb{E}\|x_n\|^4}}{4}$. Assumption 2.7 is verified if $\frac{\sqrt{\|x_n\|^4}}{4} \leq L$ almost surely.

Remark 2.9 (Data-generating distribution and L)**.** *Note that Assumption 2.7 is strong: in particular, the data must have bounded norm (and a bound known). For example, this is not the case for data with Gaussian noise! However, Assumption 2.5 comes down to assuming x_n has a fourth moment.*

3 Non-Asymptotic convergence & rates analysis for SGD

In this section, we present a non-asymptotic bound for SGD with strongly-convex functions. We discuss the result and contrast it to similar asymptotic results.

3.1 Assumptions

We introduce the following assumption:

Assumption 3.1. *f is μ-strongly convex, i.e.*

$$\forall \theta_1, \theta_2 \in \mathcal{H},\ \ f(\theta_1) \geq f(\theta_2) + \langle \nabla f(\theta_2), \theta_1 - \theta_2\rangle + \frac{\mu}{2}\|\theta_1 - \theta_2\|^2 \tag{13}$$

Notice that Assumption 3.1 implies that f achieves a unique minimum at θ^*, and that $\nabla f(\theta^*) = 0$. Also, we do not assume that f_n is strongly convex (which does imply Assumption 3.1 though). A common practice in machine learning is to add regularization (which also has learning-theoretic virtues), often of the form $\frac{\mu}{2}\|\theta\|^2$. One can also consider the restriction of a loss function to a bounded domain where it is strongly convex (see Bach and Moulines (2013) for more heuristics).

Finally, we assume that the noise has finite power. More generically:

Assumption 3.2. *There exists $\sigma^2 \geq 0$ such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|\nabla f_n(\theta^*)\|^2\right] \leq \sigma^2 \tag{14}$$

In the stochastic approximation setting where $f_n(\theta) = f(\theta) + \langle \varepsilon_n, \theta\rangle$, since $\nabla f(\theta^*) = 0$, Assumption 3.2 is equivalent to $\mathbb{E}\left[\|\varepsilon_n\|^2\right] \leq \sigma^2$.

3.2 Non-asymptotic bound

Theorem 3.3. *Under Assumptions 2.1, 2.5, 3.1 and 3.2 and with $\gamma_n = Cn^{-\alpha}$,*

$$\delta_n := \mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] \leq \begin{cases} 2\exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right)\exp\left(-\frac{\mu C}{4}n^{1-\alpha}\right)\left(\delta_0 + \frac{\sigma^2}{L^2}\right) + \frac{4C\sigma^2}{\mu n^\alpha}, & 0 \leq \alpha < 1 \\ \frac{\exp(2L^2C^2)}{n^{\mu C}}\left(\delta_0 + \frac{\sigma^2}{L^2}\right) + 2\sigma^2C^2\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}}, & \alpha = 1, \end{cases}$$

where $\varphi_\beta(n)$ is defined as:

$$\varphi_\beta(n) = \begin{cases} \frac{n^\beta - 1}{\beta}, & \beta \neq 0 \\ \log(n), & \beta = 0. \end{cases} \tag{15}$$

We can immediatly deduce that θ_n converges in quadratic mean to θ^* for any $\alpha \in (0, 1]$, while in general almost sure convergence is possible only is $\alpha > 1/2$.

Furthermore, Theorem 3.3 gives that for $\alpha < 1$:

$$\delta_n = O\left(n^{-\alpha}\right). \tag{16}$$

In fact, by the smoothness hypothesis, we can relate δ_n and $\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right]$ (❓, see Section 7.1 for a proof).

Let us also highlight that for $\alpha = 1$, the rate depends on the product μC, which we seek to illustrate in Section 6.2. If $C = \frac{2}{\mu}$ then we can achieve the minimax-optimal rate $O(n^{-1})$.

It is worthwhile highlighting a major difference between SGD and deterministic gradient descent: even if $\delta_0 = 0$ (i.e. $\theta_0 = \theta^*$ almost surely), the algorithm will not converge instantly, owing to the noise in the gradients of power σ^2.

Finally, we can see that the (possibly random) choice of θ_0 is not important as there is an exponential decay in from of δ_0 (when $\alpha < 1$).

Under similar hypotheses (but with $\alpha > 1/2$), we can show that the convergence rates (up to a logarithmic factor) are in fact almost sure[2]. Indeed:

Theorem 3.4 (Godichon-Baggioni (2021))**.** *Assume that f_n verify a* weak growth condition*, i.e. there is $\nu > \frac{1}{\alpha} - 1$ and C_ν such that*

$$\forall \theta \in \mathcal{H},\ \mathbb{E}\left[\left\|\nabla f_n(\theta)\right\|^{2+2\nu}\right] \leq C\left(1 + \left\|\theta - \theta^*\right\|^{2+2\nu}\right) \tag{17}$$

and that f is twice continuously-differentiable in a neighborhood of θ^* and $\nabla^2 f(\theta^*) \geq \Sigma > 0$. Then, for any $\alpha \in (1/2, 1)$,

$$\left\|\theta_n - \theta^*\right\|^2 = O\left(\frac{\log n}{n^\alpha}\right) \quad a.s.$$

4 Proof of Theorem 3.3

In this section, we present the complete proof of Theorem 3.3. The proof presents the main techniques used to obtain a non-asymptotic bounds, and is the basis for the bounds in the other results from Bach and Moulines (2013).

The most important part of the proof is Lemma 4.1 and its proof, as it presents the approach to bounding δ_n as a function of δ_{n-1} using the algorithmic recursion. The rest of the proof is then the analysis of the behavior of δ_n, and can be skipped at first reading.

Lemma 4.1. *Under Assumptions 2.1, 2.5, 3.1 and 3.2 and with $\gamma_n = Cn^{-\alpha}$,*

$$\delta_n \leq (1 - 2\mu\gamma_n + 2L^2\gamma_n^2)\gamma_{n-1} + 2\sigma^2\gamma_n^2. \tag{18}$$

Proof. To begin, notice that thanks to the gradient recursion:

$$\begin{aligned}
\left\|\theta_n - \theta^*\right\|^2 &= \left\|\theta_{n-1} - \gamma_n \nabla f_n(\theta_{n-1}) - \theta^*\right\|^2 & (19)\\
&= \left\|\theta_{n-1} - \theta^*\right\|^2 + \gamma_n^2\left\|\nabla f_n(\theta_{n-1})\right\|^2 - 2\langle\theta_{n-1} - \theta^*, \nabla f_n(\theta_{n-1})\rangle & (20)
\end{aligned}$$

In order to separate the randomness stemming from f_n and that stemming from θ_{n-1}, we take the expectation conditionally to $\mathcal{F}_{n-1}$:

$$\mathbb{E}\left[\left\|\theta_n - \theta^*\right\|^2 |\mathcal{F}_{n-1}\right] = \left\|\theta_{n-1} - \theta^*\right\|^2 + \gamma_n^2\mathbb{E}\left[\left\|\nabla f_n(\theta_{n-1})\right\|^2 |\mathcal{F}_{n-1}\right] - 2\langle\theta_{n-1} - \theta^*, \mathbb{E}\left[\nabla f_n(\theta_{n-1})|\mathcal{F}_{n-1}\right]\rangle \tag{21}$$

Using that f is μ-strongly convex, we have $\langle\theta_{n-1} - \theta^*, \nabla f(\theta_{n-1}) - \nabla f(\theta^*)\rangle \geq \mu\left\|\theta_{n-1} - \theta^*\right\|^2$. Thus, since $\nabla f(\theta^*) = 0$,

$$-2\langle\theta_{n-1} - \theta^*, \mathbb{E}\left[\nabla f_n(\theta_{n-1})|\mathcal{F}_{n-1}\right]\rangle \leq -2\mu\left\|\theta_{n-1} - \theta^*\right\|^2 \tag{22}$$

[2] We can show that convergence is almost sure with much weaker hypotheses (see Godichon-Baggioni).

Furthermore, by Cauchy-Schwarz:

$$\mathbb{E}\left[\|\nabla f(\theta_{n-1})\|^2 |\mathcal{F}_{n-1}\right] \leq 2\mathbb{E}\left[\|\nabla f(\theta_{n-1}) - \nabla f(\theta^*)\|^2 |\mathcal{F}_{n-1}\right] + 2\mathbb{E}\left[\|\nabla f(\theta^*)\|^2 |\mathcal{F}_{n-1}\right] \tag{23}$$

Then applying Assumption 2.5 and Assumption 3.2, since θ_{n-1} is $\mathcal{F}_{n-1}$-measurable:

$$\mathbb{E}\left[\|\nabla f(\theta_{n-1})\|^2 |\mathcal{F}_{n-1}\right] \leq 2L^2 \|\theta_{n-1} - \theta^*\|^2 + 2\sigma^2. \tag{24}$$

Taking the expectation and plugging in the two above inequalities, we obtain:

$$\delta_n \leq \underbrace{\left(1 + 2\gamma_n^2 L^2 - 2\mu\gamma_n\right)}_{A_n} \delta_{n-1} + 2\sigma^2\gamma_n^2. \tag{25}$$

□

Remark 4.2. *Note that the upper bound in the lemma an equality for quadratic functions. Indeed, let* $f(\theta) = \frac{1}{2}a^2x^2 + bx + c$. *Then, the smoothness inequality is exact, i.e.* $\nabla f(x_{n-1}) - \nabla f(x^*) = a(x_{n-1} - x^*)$.

In order to be able to iterate the recursion in in Eq. (25) and obtain an upper-bound on δ_n, we first ensure that $A_n \geq 0,\ \forall n$ (this is the case since $\mu \leq L$, $1 - An \leq 2L\gamma_n(1 - L\gamma_n) \leq \frac{1}{2}$).
We can then deduce (by recursion) that:

$$\delta_n \leq \delta_0 \underbrace{\prod_{k=1}^{n} A_k}_{T_n} + 2\sigma^2 \underbrace{\sum_{i=1}^{n} \gamma_i \prod_{k=i+1}^{n} A_k}_{S_n}. \tag{26}$$

Note that we have decomposed the upper bound into two components:

- a transient phase which depends on δ_0.
- a stationary phase which is proportional to the noise power σ^2.

In the rest of the proof we bound S_n and T_n, for a generic decreasing step-size sequence γ_n, such that $\gamma_n \xrightarrow[n\to\infty]{} 0$, and then for $\gamma_n = Cn^{-\alpha}$.

Stationary term S_n Let $n_0 = \inf\{k \in \mathbb{N} \mid \gamma_n \leq \frac{\mu}{2L^2}\}$. If $k \geq n_0$, then $A_n \leq 1 - \mu\gamma_n$. In any case we of course have, $A_n \leq 1 + 2\gamma_n^2 L^2$.
We can split S_n at $k = n_0$, which yields (ignoring the $2\sigma^2$ factor in B_k):

$$S_n = \underbrace{\sum_{k=1}^{n_0} \gamma_k^2 \prod_{i=k+1}^{n} A_i}_{S_{n,1}} + \underbrace{\sum_{k=n_0+1}^{n} \gamma_k^2 \prod_{i=k+1}^{n} A_i}_{S_{n,2}} \tag{27}$$

While all the terms in $S_{n,2}$ can be upper-bounded by $1 - \mu\gamma_i$, we further split $S_{n,1}$: Note that

$$S_{n,1} = \left(\prod_{i=n_0+1}^{n} A_i\right) \sum_{k=1}^{n_0} \gamma_k^2 \prod_{i=k+1}^{n_0} A_i. \tag{28}$$

So,

$$S_n \leq \underbrace{\left(\prod_{i=n_0+1}^{n} 1 - \mu\gamma_i\right)}_{S_n^{(1)}} \underbrace{\sum_{k=1}^{n_0} \gamma_k^2 \prod_{i=k+1}^{n_0} (1 + 2L^2\gamma_i^2)}_{S_n^{(2)}} + \underbrace{\sum_{k=n_0+1}^{n} \gamma_k^2 \prod_{i=k+1}^{n} (1 - \mu\gamma_i)}_{S_n^{(3)}} \tag{29}$$

Bounding $S_n^{(1)}$ Using that $1 + x \leq \exp(x)$ (by convexity), we have:

$$S_n^{(1)} \leq \exp\left(-\mu \sum_{i=n_0}^{n} \gamma_i\right) \tag{30}$$

It is tempting to upperbound the other terms in the same way. In fact, this is very suboptimal because we'd have a sum of exponentials. Rather we finely upper-bound $S_n^{(2)}$ and S_n^3, using the following result.

We use the following telescoping argument:

Lemma 4.3 (**?**)**.** *Let* (u_k) *be a non-negative sequence and* $a \neq 0$*. Then,*

$$\sum_{k=1}^{n} u_k \prod_{i=k+1}^{n} (1 + a u_i) = \frac{1}{a}\left(\prod_{i=1}^{n}(1 + a u_i) - 1\right). \tag{31}$$

with the convention that an empty product is equal to 1*.*

See Section 7.2 for the proof.

Bounding $S_n^{(2)}$

$$\sum_{k=1}^{n_0} \gamma_k^2 \prod_{i=k+1}^{n_0} (1 + 2L^2\gamma_i^2) \leq \frac{1}{2L^2} \prod_{i=1}^{n_0}(1 - 2L^2\gamma_i^2) \leq \frac{1}{2L^2} \exp\left(2L^2 \sum_{i=1}^{n_0} \gamma_i^2\right) \tag{32}$$

Bounding $S_n^{(3)}$ First, we bound

$$\sum_{k=m+1}^{n} \prod_{i=k+1}^{n} (1 - \mu\gamma_i)\gamma_k^2, \tag{33}$$

where $1 \leq m \leq n$.

In order to apply the lemma (note the order of the γ_k terms), we use that the sequence (γ_k) is decreasing, i.e.

$$\sum_{k=m+1}^{n} \prod_{i=k+1}^{n} (1 - \mu\gamma_i)\gamma_k^2 \leq \gamma_m \sum_{k=m+1}^{n} \prod_{i=k+1}^{n} (1 - \mu\gamma_i)\gamma_k \tag{34}$$

which the lemma finally upper-bounds by γ_m/μ.

The remainder term is upper-bounded as:

$$\sum_{k=1}^{n} \gamma_k^2 \prod_{i=k+1}^{n} (1 - \mu\gamma_i) \leq \prod_{i=m+1}^{n} (1 - \mu\gamma_i) \sum_{k=1}^{m} \gamma_k^2 \leq \exp\left(-\mu \sum_{i=m+1}^{n} \gamma_i\right) \sum_{i=1}^{m} \gamma_k^2 \tag{35}$$

where we added terms between 1 and n_0.

Thus, finally:

$$S_n^{(3)} \leq \frac{\gamma_m}{\mu} + \exp\left(-\mu \sum_{i=m+1}^{n} \gamma_i\right) \sum_{i=1}^{m} \gamma_k^2. \tag{36}$$

And:

$$S_n \leq \frac{\gamma_m}{\mu} + \exp\left(-\mu \sum_{i=m+1}^{n} \gamma_i\right) \sum_{i=1}^{m} \gamma_k^2 + \frac{1}{2L^2} \exp\left(-\mu \sum_{i=1}^{n} \gamma_i\right) \exp\left(4L^2 \sum_{i=1}^{n_0} \gamma_i^2\right). \tag{37}$$

Transient term Once again using the exponential convexity bound, we obtain:

$$T_n \leq \exp\left(-2\mu \sum_{k=1}^{n} \gamma_k\right) \exp\left(2L^2 \sum_{k=1}^{n} \gamma_k^2\right) \tag{38}$$

Notice that under the classical hypotheses for the Robbins-Monro algorithm (i.e. $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$), we can see that $T_n \to 0$. The analysis more delicate for S_n. In fact, we'll show that both γ_n and γ_n^2 can have divergent series and SGD still converges in L^2.

Specialization for $\gamma_n = Cn^{-\alpha}$ We need to estimate the difference between $\sum_{k=1}^{n} \gamma_k$ and $\sum_{k=1}^{n} \gamma_k^2$, from above and below (since there are different signs).

We use the following series-integral comparison lemma:

Lemma 4.4 (**?**)**.** *Recall the notation:*

$$\varphi_\beta(n) = \begin{cases} \frac{n^\beta - 1}{\beta}, & \beta \neq 0 \\ \log(n), & \beta = 0. \end{cases} \tag{39}$$

Then for any $0 \leq m < n$,

$$\forall \beta \leq 1,\ \varphi_\beta(n) - \varphi_\beta(m+1) \leq \sum_{k=m+1}^{n} k^{\beta-1} \leq \varphi_\beta(n) - \varphi_\beta(m) \tag{40}$$

If furthermore, $0 \leq \beta \leq 1$,

$$\frac{1}{2}\left(\varphi_\beta(n) - \varphi_\beta(m)\right) \leq \sum_{k=m+1}^{n} k^{\beta-1} \leq \varphi_\beta(n) - \varphi_\beta(m) \tag{41}$$

See Section 7.3 for a proof.

To continue the proof on Theorem 3.3, let us specialize the above result to sums of γ_k. We focus on the case when $0 < \alpha < 1$ for clarity and handle the edge cases at the end of the proof.

Thus, for $0 \leq m < n$, since $0 > \beta - 1 = -\alpha > -1$, $0 < \beta < 1$, we can apply Eq. (41):

$$\forall\ 0 < \alpha < 1,\ \frac{C}{2}\left(\varphi_{1-\alpha}(n) - \varphi_{1-\alpha}(m)\right) \leq \sum_{k=m+1}^{n} \gamma_k \leq C\left(\varphi_{1-\alpha}(n) - \varphi_{1-\alpha}(m)\right). \tag{42}$$

where γ_k depends on C and α.

For $\beta - 1 = -2\alpha$, we have $1 > \beta > -1$, we apply Eq. (40) but we need to enforce $m \geq 1$ because $\varphi_\beta(0)$ is not always defined. In particular,

$$\forall\ 0 < \alpha < 1,\ C^2\varphi_{1-2\alpha}(n) - \varphi_{1-2\alpha}(m+1) \leq \sum_{k=m+1}^{n} \gamma_k^2 \leq C^2\left(\varphi_{1-2\alpha}(n) - \varphi_{1-2\alpha}(m)\right). \tag{43}$$

Without any limits on m we can use the following inequality:

$$\forall\ 0 < \alpha < 1,\ C^2\varphi_{1-2\alpha}(n) - \varphi_{1-2\alpha}(m+1) \leq \sum_{k=m+1}^{n} \gamma_k^2 \leq C^2\left(1 + \varphi_{1-2\alpha}(n)\right) \tag{44}$$

Furthermore since if $0 < \beta < 1$,

$$\varphi_\beta(n) - \varphi_\beta(n/2) = \frac{n^\beta}{\beta}\left(1 - \frac{1}{2^\beta}\right) \geq \frac{n^\beta}{2\beta} \tag{45}$$

We now have all the necessary tools to show that, with $m = n/2$ and taking $n_0 = n$:

$$S_n \leq \frac{C2^\alpha}{n^\alpha \mu} + C^2\left[1 + \varphi_{1-2\alpha}(n)\right] \exp\left(-\frac{C\mu}{4} n^{1-\alpha}\right) \tag{46}$$

$$+ \frac{1}{2L^2} \exp\left(-\frac{\mu C}{2} \varphi_{1-\alpha}(n)\right) \exp\left(4L^2C^2\left(1 + \varphi_{1-2\alpha}(n)\right)\right). \tag{47}$$

Similarly,

$$T_n \leq \exp\left(-2\mu\sum_{k=1}^{n}\gamma_k\right)\exp\left(2L^2\sum_{k=1}^{n}\gamma_k^2\right) \leq \exp\left(-\frac{\mu C}{2}\varphi_{1-\alpha}(n)\right)\exp\left(4L^2C^2\left(1+\varphi_{1-2\alpha}(n)\right)\right) \tag{48}$$

And so, recalling that $\delta_n \leq \delta_0 T_n + 2\sigma^2 S_n$,

$$\delta_n \leq \exp\left(-\frac{\mu C}{2}\varphi_{1-\alpha}(n)\right)\exp\left(4L^2C^2\left(1+\varphi_{1-2\alpha}(n)\right)\right)\left(\frac{\sigma^2}{L^2}+\delta_0\right) \tag{49}$$

$$+\frac{4\sigma^2 C}{n^\alpha \mu}+2\sigma^2C^2\left[1+\varphi_{1-2\alpha}(n)\right]\exp\left(-\frac{C\mu}{4}n^{1-\alpha}\right) \tag{50}$$

Furthermore, using $\varphi_{1-\alpha}(n) > n^{1-\alpha}$ and $2C^2\sigma^2 \leq 4C^2L^2\left(\frac{\sigma^2}{L^2}+\delta_0\right)$, and applying $\exp(x) \geq x$:

$$\delta_n \leq 2\exp\left(-\frac{\mu C}{4}n^{1-\alpha}\right)\exp\left(4L^2C^2\left(1+\varphi_{1-2\alpha}(n)\right)\right)\left(\frac{\sigma^2}{L^2}+\delta_0\right)+\frac{4\sigma^2C}{n^\alpha\mu} \tag{51}$$

Case: $\alpha = 1$ This is case proceeds in the same way, by directly upperbounding $S_n^{(3)}$ with the exponential bound.

This marks the end of the proof of Theorem 3.3. Note that we were not able to - following the approach in the paper - recover the same constants, perhaps due to our approach of doing series-integral comparisons. For the discussion, we keep their constants. □

5 Effect of strong convexity and averaging

In this section, we present and discuss different results from Bach and Moulines (2013) in order to demonstrate the effects of strong convexity of the objective function and of the averaging procedure on non-asymptotic rates. We compare these results with asymptotic rates from Godichon-Baggioni (2021) (obtained using different hypotheses).

5.1 Effect of strong convexity

Recall that in Theorem 3.3, we obtained rates of order $O(n^{-\alpha})$ for any $\alpha \in (0,1]$. Replacing Assumption 3.1 by Assumption 5.1 below (i.e. losing strong convexity), we can hope for at best $O(n^{-\frac{1}{2}})$, which is the minimax-optimal oracle rate.

Assumption 5.1. *The function f acheives its global minimum at $\theta^* \in \mathcal{H}$ (θ^* is not necessarily unique though).*

Theorem 5.2 (SGD no strong convexity)**.** *Under Assumptions 2.1, 2.7, 3.2 and 5.1, for any $\alpha \in [1/2, 1]$, we have:*

$$\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \leq \frac{1}{C}\left(\delta_0+\frac{\sigma^2}{L^2}\right)\exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right)\frac{1+4L^{3/2}C^{3/2}}{\min\{\varphi_{1-\alpha}(n),\varphi_{\alpha/2}(n)\}} \tag{52}$$

From Theorem 5.2, we can see that lack of strong convexity reduces the range of α values for which we have convergence to $(\frac{1}{2},1]$[3].

Furthermore, we fail to reach the minimax-optimal rate of $O(n^{-1/2})$ (it is easy to see that the fastest rate we can have is $O(n^{-1/3})$). In contrast, we achieved the minimax-optimal rate for strongly convex functions for $\alpha = 1$.

In Section 6.1, we illustrate these observations.

[3] The case $\alpha = 1/2$ is excluded for simplicity as it involves a convergence criterion on C, compared to L.

5.2 Effect of averaging

Recall that the averaging procedure consists in considering $\bar{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i$ instead of θ_n.

What can we expect from averaging ?

- In the strongly convex setting, we will not see better results for $\alpha = 1$ since we already have minimax optimal rates. However, we would like to obtain the rate (somewhat) independently of the choice of α, so as to be able to tune the learning rate for the transient phase (see Section 6).
- In the non-strongly convex setting, Theorem 5.2 did not acheive the optimal rates. Averaging could bridge the gap (we will see that this is not quite the case).

Averaging with strong-convexity

We only state a summary of Theorem 4 from Bach and Moulines (2013):

Theorem 5.3. *Under Assumptions 2.1, 2.7, 3.1 and 3.2 and assuming that f_n have Lipschitz Hessians (with a universal Lipschitz constant) and that there exists $\infty > \tau \geq 0$ such that*

$$\mathbb{E}\left[\|\nabla f_n(\theta^*)\|^4 \,|\mathcal{F}_{n-1}\right] \leq \tau^2 \tag{53}$$

then for any $\alpha \in \left[\frac{1}{2}, 1\right]$:

$$\mathbb{E}\left[\left\|\bar{\theta}_n - \theta^*\right\|^2\right] = O(n^{-1}) \tag{54}$$

The underlying assumption in the previous theorem is that Hessian of f be Lipschitz. Another way of thinking about this condition, is how well f is approximated by a quadratic function. Indeed, in the limit case where the Hessian is constant (i.e. Lipschitz constant 0), then the function is quadratic.

Averaging without strong-convexity

Similarly, Theorem 7 (Bach and Moulines, 2013) yields:

Theorem 5.4. *Under Assumptions 2.1, 2.7 and 5.1 and assuming that we have $D, B > 0$ such that*

$$\forall \theta \in \mathcal{H},\ \forall n \leq 0,\ \|\theta\| \leq D \implies \|\nabla f_n(\theta)\| \leq B \tag{55}$$

for any $\alpha \in [0, 1]$:

$$\mathbb{E}\left[f(\bar{\theta}_n) - f(\theta^*)\right] = \begin{cases} O(n^{-\alpha}), & \alpha < \frac{1}{2} \\ \tilde{O}(n^{-1/2}), & \alpha = \frac{1}{2} \\ O(n^{1-\alpha}), & \alpha > \frac{1}{2} \end{cases} \tag{56}$$

where $\tilde{O}$ denotes the presence of a logarithmic factor.

Without strong convexity, $\alpha = \frac{1}{2}$ acheives the minimax-optimal rate (up to a logarithmic factor), allowing for assuming bounded gradients.

5.3 Choice of α

Given a stochastic optimisation problem, the main lever a practitioner has is the learning rate, that is, the choice of C and α. In the discussions around the results presented in these notes, we remarked how both can be chosen a function of the problem's specifications (strong convexity of not, smoothness constants, ...). However, these require knowledge of the model or the underlying distribution (for example, the data-generating distribution in machine learning), which is in general not the case.

Setting the choice of C aside (see Bach and Moulines (2013) for a heuristic), a robust choice of α is $\frac{1}{2}$ with averaging. Indeed, Theorems 5.3 and 5.4 show that with $\alpha = \frac{1}{2}$ and averaging, minimax optimal rates are obtained. In other words, the choice $\alpha = \frac{1}{2}$ is adaptive to the difficulty of the problem.

5.4 Asymptotic rates

Theorem 5.5 (Godichon-Baggioni (2021))**.** *Under the hypotheses of Theorem 3.4, and assuming*[4] *that the Hessian of f is Lipschitz in a neighborhood of θ^*, then for any $\alpha \in (1/2, 1)$, for any $\delta > 0$,*

$$\left\|\bar{\theta}_n - \theta^*\right\|^2 = o\left(\frac{(\log n)^{1+\delta}}{n}\right) \quad a.s. \tag{57}$$

We can observe the same phenomenon as in the non-asymptotic analysis: fast rates are obtained for any $\alpha \in (1/2, 1)$.

In fact, we can prove that (assuming that $\theta \mapsto \mathbb{E}\left[\nabla f_n(\theta)\nabla f_n(\theta)^T\right]$ is continuous at θ^*)

$$\sqrt{n}\left(\bar{\theta}_n - \theta^*\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, H^{-1}\Sigma H^{-}1) \tag{58}$$

where $H = \nabla^2 f(\theta^*)$ and $\Sigma = \mathbb{E}\left[\nabla f_n(\theta^*)\nabla f_n(\theta^*)^T\right]$.

This matches the asymptotic efficiency result we introduced in Theorem 1.1. In plain language, an online estimator (with averaging) is as efficient as the M-estimator.

6 Numerical experiments

In this section, we reproduce synthetic experiments from Bach and Moulines (2013) which illustrate the results in these notes.

6.1 Impact of strong-convexity

Consider the twice continuously-differentiable function f defined by (illustrated in Figure 1 for $q = 2$ and $q = 4$):

$$\begin{aligned} f : \mathbb{R} &\longrightarrow \mathbb{R} \\ \theta &\longmapsto \begin{cases} |\theta|^q, & \theta \in [-1, 1] \\ q(\theta - 1) + 1, & \theta > 1 \\ -q(\theta + 1) + 1, & \theta > 1 \end{cases} \end{aligned}$$

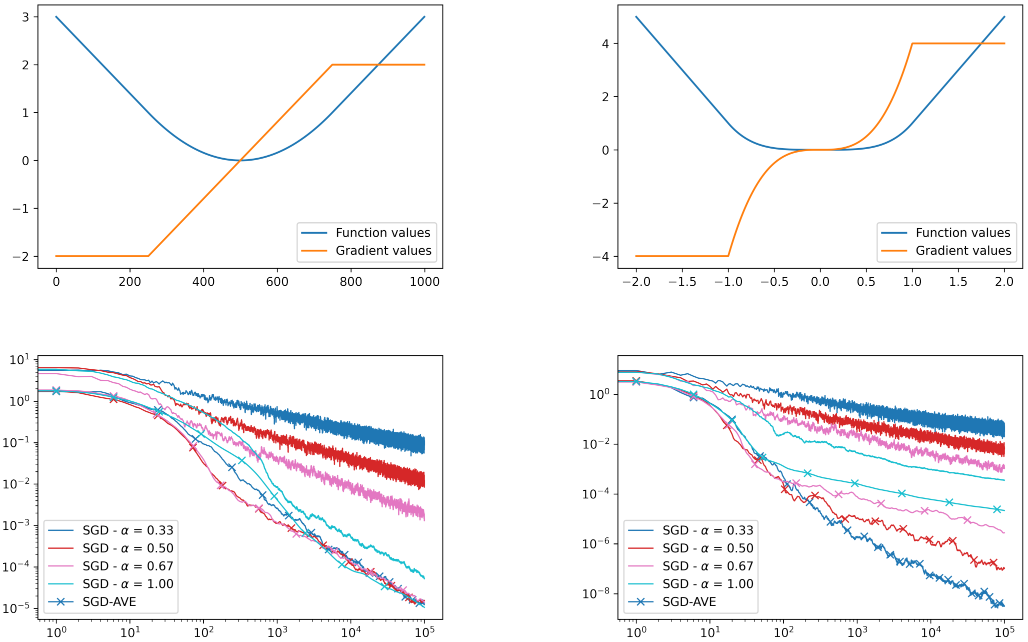
We consider the stochastic approximation where we observe gradients corrupted with Gaussian noise (with standard deviation 4). Repeating $N = 100$ SGD and averaged SGD runs over $n = 10000$ samples, for different values of α, we show how α and the presence of acceleration determine the rate of convergence. Results are summarized in Figure 1.

6.2 Robustness to constants

We remarked after presenting Theorem 3.3 that when $\alpha = 1$, the choice of C is key, and that SGD is not robust to a poor choice of C (recall: it is chosen with respect to μ).

We can illustrate this (and show that this is not the case for $\alpha = \frac{1}{2}$) by comparing performance for different values of C.

[4] This condition is not minimal (see Godichon-Baggioni (2021)).



(a) $f(\theta) = |\theta|^2$.

(b) $f(\theta) = |\theta|^4$.

Figure 1: SGD and averaged SGD (SGD-AVE) with $C = 1$ and varying α. **Top:** function and gradients. Notice that for $q = 2$, f is locally strongly convex, which is not the case for $q = 4$. **Bottom:** average error $f - f(\theta^*)$ over 100 runs. Notice that when we do not have strong convexity (right), averaging does not yield fast rates, while we can see that all values of α give an optimal rate $O(n^{-1})$ rate (as does $\alpha = 1$).

7 Solutions to exercises

7.1 Relating value and iterate convergences

The key idea is to use express $f_n(\theta_{n-1}) - f_n(\theta^*)$ using their gradients, then apply the smoothness hypothesis:

$$f_n(\theta_{n-1}) - f_n(\theta^*) = \int_0^1 \langle \nabla f_n\left(t\theta_{n-1} + (1-t)\theta^*\right), \theta_{n-1} - \theta^* \rangle dt \tag{59}$$

And since $\mathbb{E}\left[\nabla f_n(\theta^*)|\mathcal{F}_{n-1}\right]$:

$$\begin{aligned}
\mathbb{E}\left[f_n(\theta_{n-1}) - f_n(\theta^*)|\mathcal{F}_{n-1}\right] &= \int_0^1 \mathbb{E}\left[\langle \nabla f_n\left(t\theta_{n-1} + (1-t)\theta^*\right), \theta_{n-1} - \theta^* \rangle |\mathcal{F}_{n-1}\right] dt && (60)\\
&\leq \int_0^1 \mathbb{E}\left[\|\nabla f_n\left(t\theta_{n-1} + (1-t)\theta^*\right)\| \|\theta_{n-1} - \theta^*\| |\mathcal{F}_{n-1}\right] dt && (61)\\
&= \int_0^1 \mathbb{E}\left[\|\nabla f_n\left(t\theta_{n-1} + (1-t)\theta^*\right)\| |\mathcal{F}_{n-1}\right] dt \|\theta_{n-1} - \theta^*\| && (62)\\
&\leq \int_0^1 \left(\mathbb{E}\left[\|\nabla f_n\left(t\theta_{n-1} + (1-t)\theta^*\right)\|^2 |\mathcal{F}_{n-1}\right]\right)^{1/2} dt \|\theta_{n-1} - \theta^*\| && (63)\\
&\leq L \|\theta_{n-1} - \theta^*\|^2 \int_0^1 t dt = \frac{L}{2} \|\theta_{n-1} - \theta^*\|^2 && (64)
\end{aligned}$$

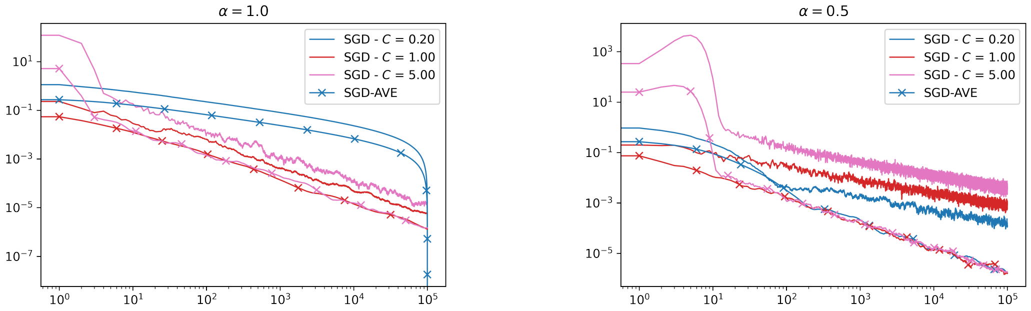


Figure 2: SGD and averaged SGD (SGD-AVE) with $\alpha = 1.0$ (left) and $\alpha = 0.5$ (right), with varying C. **Left** ($\alpha = 1.0$)**:** when C is too large, there are instabilities at the start of the run. If C is too small, then the rate is well under n^{-1}, until the end of the run. On the other hand, $\alpha = 0.5$ **(right),** except the instability at the start of the optimization, all runs have the predicted rates ($n^{-1/2}$ for SGD and n^{-1} for averaged SGD) irrespective of C (i.e. the procedure est robust to knowledge of μ.)

Finally, taking the expectation yields:

$$\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \leq \mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \tag{65}$$

7.2 Proof of Lemma 4.3

Proof.

$$\sum_{k=1}^{n} u_k \prod_{i=k+1}^{n} (1 + au_i) = \sum_{k=1}^{n} \frac{1}{a} \left(\prod_{i=k}^{n} (1 + au_i) - \prod_{i=k+1}^{n} (1 + au_i) \right) \tag{66}$$

$$= \frac{1}{a} \left(\prod_{i=1}^{n} (1 + au_i) - 1 \right) \tag{67}$$

□

7.3 Proof of Lemma 4.4

For completeness, we prove the lemma[5]:

Proof.

Let $\beta \leq 1$. $f : x \mapsto x^{\beta - 1}$ is decreasing on $\mathbb{R}^+$ so if $n \geq 1$

$$\int_{n}^{n+1} f \leq f(n) \leq \int_{n-1}^{n} f \tag{68}$$

so for $1 \geq m \geq n$,

$$\varphi_\beta(n+1) - \varphi_\beta(m+1) = \int_{m+1}^{n+1} f \leq \sum_{k=m+1}^{n} f(k) \leq \int_{m}^{n} f = \varphi_\beta(n) - \varphi_\beta(m). \tag{69}$$

We then obtain Eq. (40) by using the fact that φ_β is increasing, whatever β.

In order to obtain Eq. (41), we can notice it is implied by Eq. (69) if and only if the map $\phi : x \mapsto \varphi_\beta(x+1) - \varphi_\beta(x)/2$ is non-decreasing on $[1, +\infty[$. Studying the variations of ϕ we can see that this is equivalent to

$$x \geq \frac{1}{\frac{1}{2}^{\frac{1}{\beta - 1}} - 1}. \tag{70}$$

[5] Note that the proof we provide is elementary and stronger arguments are possible. In fact, the authors do not place a lower bound on β which makes us believe that a stronger result is possible.

The right-hand term is less than 1 for all $0 \leq \beta \leq 1$ so the condition is verified on $\mathbb{N}^*$. □

Bibliography

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in Neural Information Processing Systems*, 2013.

Antoine Godichon-Baggioni. Notes de cours: Optimisation stochastique, 2021. URL http://godichon.perso.math.cnrs.fr/Cours_algo_sto.pdf.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. *Technical report*, 1988.

Boris Polyak. New method of stochastic approximation type. *Automation and Remote Control*, 1990, 1990.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2011.

R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 2020.

Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, 2009.

A.W. Van Der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, 1998.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511809682.

Julien Mairal and Jean-Philippe Vert. Course: Kernel methods for machine learning, 2021. URL http://members.cbio.mines-paristech.fr/~jvert/svn/kernelcourse/course/2021mva/index.html.