Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning: a short recap

Henri Alam, Maxime Haddouche, Redwan Mekrami

Abstract

This document consists in a short presentation of the main results of Moulines and Bach [2011]. We will also compare their results and remark with the SGD described in the OCO lecture.

1 Framework

Moulines and Bach [2011] presents several results for the classical *Stochastic Gradient Descent* (sometimes called the *Robbins-Monro algorithm*) and a simple modification of it where iterates are averaged (*Polyak-Ruppert averaging*) when optimising on a Hilbert space $\mathcal{H}$. More precisely, we do not restrict ourselves to a convex $\mathcal{K} \subset \mathcal{H}$

More concretely we describe first Robbins-Monro algorithm in this case :

Algorithm 1: Stochastic Gradient descent (Robbins-Monro algorithm) onto a Hilbert space $\mathcal{H}$.

Parameters : Epoch T, step-sizes (γ_n)
Initialisation : Initial point $\theta_0 \in \mathcal{K}$
for *each iteration* n *in* $1..T$ **do**
  Sample $\hat{\nabla} f(\theta_n)$
  Update
$$\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})$$
end
Return θ_T

Where (f_n) is a sequence of convex differentiable randoù functions whom their gradient are unbiased estimates of the gradient of a certain function f we wish to minimize through the epochs. We denote by f_n' the gradient of f_n which is a random variable centered in f'.

The Polyak-Rupper (PR) averaging consists in returning $\bar{\theta}_T = \frac{1}{T+1}\sum_{n=0}^{T} \theta_n$ instead of θ_n. This averaging is known to ensure more stability to the final output.

Here, we focus on a specific form of the steps γ_n: we set two real constants $\alpha, C \in [0,1]\times]0;+\infty[$ and we have:

$$\forall n, \gamma_n := Cn^{-\alpha}$$

Our goal is then to provide explicit non-asymptotic bounds to measure the efficiency of those algorithms.

2 Main results

A classical assumption in online optimisation is that the sequence (f_n) is assume to be μ-strongly convex, Moulines and Bach [2011] provided two bounds for SGD (with or without PR averaging) exploiting this assumption, but also two bounds not involving this somewhat-restrictive hypothesis.

For the sake of clarity, we provide in Appendix A, a set of 9 hypotheses we will need to state rigorously our theorems.

2.1 Theorems with the strong convexity assumption

SGD

Before stating our first theorem (see proof in the appendix), we introduce the following family of functions: $\varphi_\beta : \mathbb{R}_+\backslash\{0\} \to \mathbb{R}$ given by:

$$\varphi_\beta(t) = \begin{cases} \frac{t^\beta - 1}{\beta} & \text{if } \beta \neq 0 \\ \log t & \text{if } \beta = 0 \end{cases}$$

The function $\beta \mapsto \varphi_\beta(t)$ is continuous for all $t > 0$. Moreover, for $\beta > 0, \varphi_\beta(t) < \frac{t^\beta}{\beta}$, while for $\beta < 0$, we have $\varphi_\beta(t) < \frac{1}{-\beta}$ (both with asymptotic equality when t is large).

Theorem 1 (Stochastic gradient descent, strong convexity)**.** *Assume **(H1,H2,H3,H4)**. Denote by $\delta_n = \mathbb{E}\left\|\theta_n - \theta^*\right\|^2$, where $\theta_n \in \mathcal{H}$ is the n -th iterate of the recursion in Eq. (1), with $\gamma_n = Cn^{-\alpha}$. We have, for $\alpha \in [0,1]$:*

$$\delta_n \leqslant \begin{cases} 2\exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right)\exp\left(-\frac{\mu C}{4}n^{1-\alpha}\right)\left(\delta_0 + \frac{\sigma^2}{L^2}\right) + \frac{4C\sigma^2}{\mu n^\alpha}, & \text{if } 0 \leqslant \alpha < 1 \\ \frac{\exp\left(2L^2C^2\right)}{n^{\mu C}}\left(\delta_0 + \frac{\sigma^2}{L^2}\right) + 2\sigma^2C^2\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}}, & \text{if } \alpha = 1 \end{cases}$$

Forgetting initial conditions. Bounds depend on the initial condition $\delta_0 = \mathbb{E}\left[\left\|\theta_0 - \theta^*\right\|^2\right]$ and the variance σ^2 of the noise term. The initial condition is forgotten sub-exponentially fast for $\alpha \in (0,1)$, but not for $\alpha = 1$. For $\alpha < 1$, the asymptotic term in the bound is $\frac{4C\sigma^2}{\mu n^\alpha}$

Behavior for $\alpha = 1$. For $\alpha = 1$, we have $\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}} \leqslant \frac{1}{\mu C/2-1}\frac{1}{n}$ if $C\mu > 2$, $\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}} = \frac{\log n}{n}$ if $C\mu = 2$ and $\frac{\varphi_{\mu C/2-1}(n)}{n^{\mu C/2}} \leqslant \frac{1}{1-\mu C/2}\frac{1}{n^{\mu C/2}}$ if $C\mu > 2$. Therefore, for $\alpha = 1$, the choice of C is critical: too small C leads to convergence at arbitrarily small rate of the form $n^{-\mu C/2}$, while too large C leads to explosion due to the initial condition.

Setting C too large. There is a potentially catastrophic term when C is chosen too large, i.e., $\exp\left(4L^2C^2\varphi_{1-2\alpha}(n)\right)$, which leads to an increasing bound when n is small. If $\alpha < 1$ this catastrophic term is in front of a sub-exponentially decaying factor, so its effect is mitigated once the term in $n^{1-\alpha}$ takes over $\varphi_{1-2\alpha}(n)$. Moreover, the asymptotic term is not involved in it.

PR averaging

Theorem 2 (Averaging, strong convexity)**.** *Assume **(H1, H2', H3, H4, H7, H8)**. Then, for $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$ and $\alpha \in (0,1)$, we have:*

$$\begin{aligned}\left(\mathbb{E}\|\bar{\theta}_n - \theta^*\|^2\right)^{1/2} \leqslant{}& \frac{\left[\operatorname{tr} f''(\theta^*)^{-1}\Sigma f''(\theta^*)^{-1}\right]^{1/2}}{\sqrt{n}} + \frac{6\sigma}{\mu C^{1/2}}\frac{1}{n^{1-\alpha/2}} + \frac{MC\tau^2}{2\mu^{3/2}}\left(1 + (\mu C)^{1/2}\right)\frac{\varphi_{1-\alpha}(n)}{n} \\ &+ \frac{4LC^{1/2}}{\mu}\frac{\varphi_{1-\alpha}(n)^{1/2}}{n} + \frac{8A}{n\mu^{1/2}}\left(\frac{1}{C} + L\right)\left(\delta_0 + \frac{\sigma^2}{L^2}\right)^{1/2} \\ &+ \frac{5MC^{1/2}\tau}{2n\mu}A\exp\left(24L^4C^4\right)\left(\delta_0 + \frac{\mu\mathbb{E}\left[\|\theta_0 - \theta^*\|^4\right]}{20C\tau^2} + 2\tau^2C^3\mu + 8\tau^2C^2\right)^{1/2}\end{aligned}$$

where A is a constant that depends only on μ, C, L and α

Forgetting initial condition. There is no sub-exponential forgetting of initial conditions, but rather a decay at rate $O(n^{-2})$. This is a known problem which may slow down the convergence, a common practice being to start averaging after a certain number of iterations. Moreover, the constant A may be large when LC is large, thus the catastrophic terms are more problematic than for stochastic gradient descent, because they do not appear in front of sub-exponentially decaying terms. This suggests to take CL small.

2.2 Theorems without the strong convexity assumption

SGD

Theorem 3 (Stochastic gradient descent, no strong convexity)**.** *Assume (H1,H2',H4,H9). Then, if $\gamma_n = Cn^{-\alpha}$, for $\alpha \in [1/2, 1]$, we have:*

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

When $\alpha = 1/2$, the bound goes to zero only when $LC < 1/4$, at rates which can be arbitrarily slow. For $\alpha \in (1/2, 2/3)$, we get convergence at rate $O\left(n^{-\alpha/2}\right)$, while for $\alpha \in (2/3, 1)$, we get a convergence rate of $O\left(n^{\alpha-1}\right)$. For $\alpha = 1$, the upper bound is of order $O\left((\log n)^{-1}\right)$, which may be very slow (but still convergent). The rate of convergence changes at $\alpha = 2/3$, where we get our best rate $O\left(n^{-1/3}\right)$,.

PR averaging

Theorem 4 (averaging, no strong convexity)**.** *Assume (H1,H2',H4,H9). Then, if $\gamma_n = Cn^{-\alpha}$, for $\alpha \in [1/2, 1]$, we have*

$$\mathbb{E}\left[f\left(\bar{\theta}_n\right) - f\left(\theta^*\right)\right] \leqslant \frac{1}{C}\left(\delta_0 + \frac{\sigma^2}{L^2}\right) \frac{\exp\left(2L^2C^2\varphi_{1-2\alpha}(n)\right)}{n^{1-\alpha}} \left[1 + (2LC)^{1+\frac{1}{\alpha}}\right] + \frac{\sigma^2 C}{2n}\varphi_{1-\alpha}(n)$$

If $\alpha = 1/2$, then we only have convergence under $LC < 1/4$ (as in Theorem 3), with potentially slow rate, while for $\alpha > 1/2$, we have a rate of $O\left(n^{-\alpha}\right)$, with otherwise similar behavior than for the strongly convex case with no bounded gradients. Here, averaging has allowed the rate to go from $O\left(\max\left\{n^{\alpha-1}, n^{-\alpha/2}\right\}\right)$ to $O\left(n^{-\alpha}\right)$

3 Experiments.

Our experimental setup deals with linear SVM. Our function f is the following, for any (a_i, b_i) in the dataset MNIST:

$$f(x) = \frac{1}{n}\sum_{1}^{n} \text{hinge}(b_i x^T a_i) + \frac{\lambda}{2}||x||^2$$

Where $hinge(x) = \max(0, 1 - x)$.
As recalled in [Tsianos and Rabbat, 2012, Eq. 4], f is λ-strongly convex. Our theorems of interest are then theorem 1, theorem 2.

We implement SGD without PR averaging on the next subsection.

3.1 SGD in practice

We chose $\lambda = 1$.

We chose to study SGD with different step-sizes $\gamma_n = Cn^{-\alpha}$. We chose $C \in \{1/20, 1/2, 2\}$ and $\alpha \in \{1/2, 1\}$. The results are presented below:

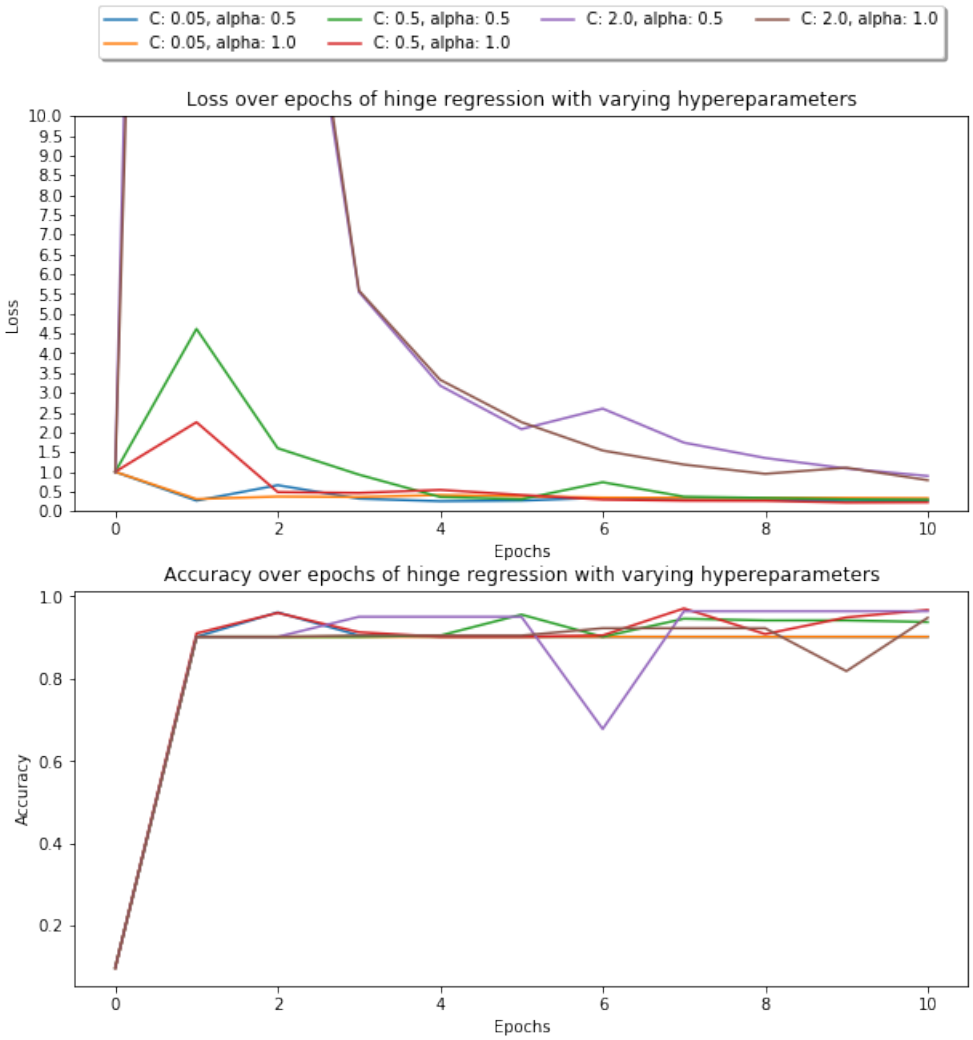

Figure 1: Evolution of the loss and the accuracy on the MNIST test dataset for basic SGD.

We can compare what has been remarked under theorem 1. When $\alpha = 1$, we said that "too small C leads to convergence at arbitrarily small rate of the form $n^{-\mu C/2}$, while too large C leads to explosion due to the initial condition." We can see this on the accuracy plot: at 10 epochs when $\alpha = 1$, the curve $C = 0.05$ is below the curve $C = 2$, the last one being below the curve $C = 0.5$,.

We also said that "if $\alpha < 1$, this catastrophic term is in front of a sub-exponentially decaying factor, so its effect is mitigated once the term in $n^{1-\alpha}$ takes over $\varphi_{1-2\alpha}(n)$. Moreover, the asymptotic term is not involved in it." We can see that for $\alpha = 1/2$ the curves $C = 0.5$ and $C = 2.0$ provides the same accuracy after ten epochs: the influence of C is indeed attenuated after several epochs.

We now deal with SGD with PR averaging.

3.2 SGD with Polyak-Ruppert averaging in practice

We kept $\alpha = 1$ and $C \in \{1/20, 1/2, 2\}$ and $\alpha \in \{1/2, 1\}$. We obtained the following curves:

The first striking remark is that PR averaging is providing in practice far more stability. We lost the erratic behavior of the purple curve compare to fig. 1. We then notice that we precised that taking LC small was suggested. We can see in practice on the loss curve that if $C = 1/2$ or $C = 2$, then we have a peak of divergence at epoch 1. This peak is not visible for $C = 0.05$. One can see that the accuracy are quite similar except for $/alpha = C = 0.5$, we can only say that we need to find

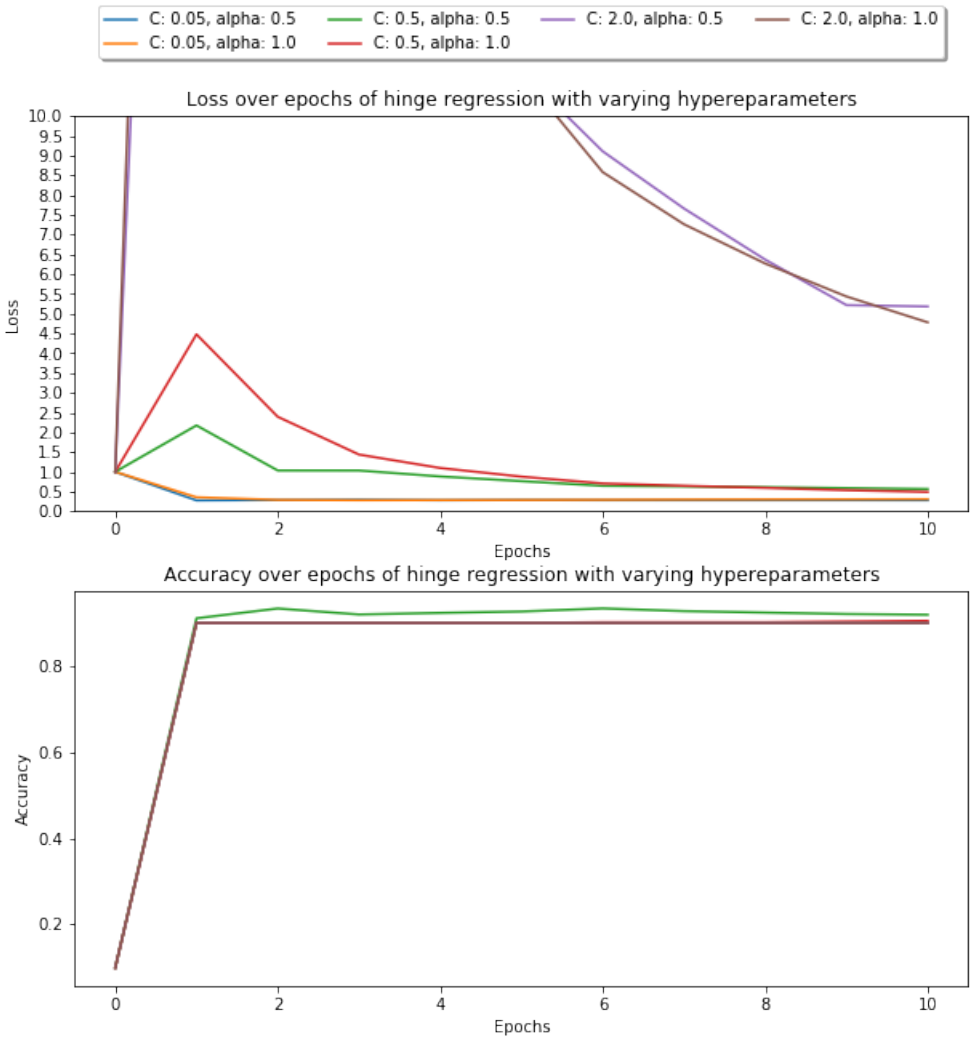

Figure 2: Evolution of the loss and the accuracy on the MNIST test dataset for SGD with PR averaging.

balanced parameters to maximise our accuracy. One can still not say much on how to find those optimal parameters.

References

E. Moulines and F. R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in neural information processing systems*, pages 451–459, 2011.

K. I. Tsianos and M. G. Rabbat. Distributed strongly convex optimization. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 593–600. IEEE, 2012.

A Hypotheses needed for the main theorems

(H1) Let $(\mathcal{F}_n)_n \geqslant 0$ be an increasing family of σ -fields. θ_0 is $\mathcal{F}_0$ -measurable, and for each $\theta \in \mathcal{H}$, the random variable $f'_n(\theta)$ is square-integrable, $\mathcal{F}_n$ -measurable and

$$\forall \theta \in \mathcal{H}, \quad \forall n \geqslant 1, \quad \mathbb{E}\left(f'_n(\theta) \mid \mathcal{F}_{n-1}\right) = f'(\theta), \text{ w.p.1.}$$

(H2) For each $n \geqslant 1$, the function f_n is almost surely convex, differentiable, and:

$$\forall n \geqslant 1, \forall \theta_1, \theta_2 \in \mathcal{H}, \quad \mathbb{E}\left(\left\|f'_n(\theta_1) - f'_n(\theta_2)\right\|^2 \mid \mathcal{F}_{n-1}\right) \leqslant L^2 \left\|\theta_1 - \theta_2\right\|^2, \quad \text{w.p. 1.}$$

(H'2) For each $n \geqslant 1$, the function f_n is almost surely convex, differentiable with Lipschitz-continuous gradient f'_n, with constant L, that is:

$$\forall n \geqslant 1, \forall \theta_1, \theta_2 \in \mathcal{H}, \quad \left\|f'_n(\theta_1) - f'_n(\theta_2)\right\| \leqslant L \left\|\theta_1 - \theta_2\right\|, \quad \text{w.p. 1.}$$

(H3) The function f is μ-strongly convex with respect to the norm of $\mathcal{H}$.

(H4) If we assume **(H3)**, then f admits a minimum at a unique vector θ^*. Our assumption is the following : there exists $\sigma^2 > 0$ such that for all $n \geqslant 1$, w.p.1 :

$$\mathbb{E}\left(\left\|f'_n(\theta^*)\right\|^2 \mid \mathcal{F}_{n-1}\right) \leqslant \sigma^2$$

(H7) For each $n \geqslant 1$, the function f_n is almost surely twice differentiable with Lipschitz-continuous Hessian operator f''_n, with Lipschitz constant M. That is, for all $\theta_1, \theta_2 \in \mathcal{H}$ and for all $n \geqslant 1$, $\left\|f''_n(\theta_1) - f''_n(\theta_2)\right\| \leqslant M \left\|\theta_1 - \theta_2\right\|$, where $\|\cdot\|$ is the operator norm.

Note that $(\mathbf{H}7)$ needs only to be satisfied for $\theta_2 = \theta^*$. For least-square regression, we have $M = 0$, while for logistic regression, we have $M = R^3/4$.

(H8) There exists $\tau \in \mathbb{R}_+$, such that for each $n \geqslant 1, \mathbb{E}\left(\left\|f'_n(\theta^*)\right\|^4 \mid \mathcal{F}_{n-1}\right) \leqslant \tau^4$ almost surely. Moreover, there exists a nonnegative self-adjoint operator Σ such that for all $n, \mathbb{E}\left(f'_n(\theta^*) \otimes f'_n(\theta^*) \mid \mathcal{F}_{n-1}\right) \leqslant \Sigma$ almost-surely.

(H9) The function f attains its global minimum at a certain $\theta^* \in \mathcal{H}$ (which may not be unique).