The convergence of the Stochastic Gradient Descent (SGD) : a self-contained proof

Gabriel TURINICI
CEREMADE - CNRS
Université Paris Dauphine - PSL Research University
Gabriel.Turinici@dauphine.fr

November 14, 2023

Abstract

We give here a proof of the convergence of the Stochastic Gradient Descent (SGD) in a self-contained manner.

1 Introduction

The Stochastic Gradient Descent (SGD) or other algorithms derived from it are used extensively in Deep Learning, a branch of Machine Learning; but the proof of convergence is not always easy to find. The goal of this paper is to adapt various proofs from the literature in a simple format. **In particular no claim of originality is made and this is rather a pedagogic work** (see [1–4] for some of my recent research papers in this area); please cite this presentation if you find it useful.

This proof can be used in any domain where a self-contained presentation is needed.

2 Recall of the general framework

Suppose $(\Omega, F, \mathbb{P})$ is a probability space, $L : \Omega \times \mathbb{R}^N \to \mathbb{R}$ a function depending on a random argument ω and a parameter X (second argument) to be optimized. Denote

$$\mathcal{L}(X) = \mathbb{E}_\omega[L(\omega, X)]. \tag{1}$$

The goal of the SGD is to find a minimum of $\mathcal{L}$. It operates iteratively by taking at iteration n:

- a (deterministic) "learning rate" ρ_n (schedule fixed *a priori*)
- a random $\omega_n \in \Omega$ independent of any other previous random variables is drawn (following the law $\mathbb{P}$)
- and updating by the formula

$$X_{n+1} = X_n - \rho_n \nabla_X L(\omega_n, X_n). \tag{2}$$

3 Hypothesis on L and $\mathcal{L}$

In order to prove the convergence we need some hypothesis that are detailed below

1. The gradient of L satisfies the following bound:

$$\exists C_0, C_1 > 0 : \ \mathbb{E}_\omega \left[\|\nabla_X L(\omega, X)\|^2 \right] \leq C_0 + C_1 \|X\|^2, \ \forall X \in \mathbb{R}^N. \tag{3}$$

2. $\mathcal{L}$ is strongly convex:

$$\exists \mu > 0 : \ \mathcal{L}(Y) \geq \mathcal{L}(X) + \langle \nabla \mathcal{L}(X), Y - X \rangle + \frac{\mu}{2} \|X - Y\|^2, \ \forall X, Y \in \mathbb{R}^N. \tag{4}$$

   Note that for $\mu = 0$ this is just the usual convexity, i.e. the function is above its tangent. For general μ this tells that the function is even above a parabola centered in any X. For regular functions this means that the Hessian $D^2\mathcal{L}$ of $\mathcal{L}$ satisfies $D^2\mathcal{L} \geq \mu \cdot I_N$[1].

4 A convergence result and its proof

We fill prove the following

Theorem 1. *Suppose that each $L(\omega, \cdot)$ is differentiable (a.e. $\omega \in \Omega$)[2] and that $\mathcal{L}$ satisfies the hypothesis* (3) *and* (4)*. Then*

1. *the function $\mathcal{L}$ has an unique minimum X_*;*

2. *For any $n \geq 0$ denote*

$$d_n = \mathbb{E} \left[\|X_n - X_*\|^2 \right]. \tag{5}$$

   Then there exist constants $c_0, c_1 > 0$ such that

$$d_{n+1} \leq (1 - \rho_n \mu + \rho_n^2 c_1) d_n + \rho_n^2 c_0. \tag{6}$$

3. *For any $\epsilon > 0$ there exists a $\rho_\epsilon > 0$ such that if $\rho_n = \rho < \rho_\epsilon$ then*

$$\limsup_{n \to \infty} \mathbb{E} \left[\|X_{n+1} - X_*\|^2 \right] \leq \epsilon. \tag{7}$$

4. *Take ρ_n a sequence such that:*

$$\rho_n \to 0 \text{ and } \sum_{n \geq 1} \rho_n = \infty. \tag{8}$$

   Then $d_n \to 0$, that is $\lim_{n \to \infty} X_n = X_*$, where the convergence is the L^2 convergence of random variables.

[1] Here I_N is the $N \times N$ identity matrix.

[2] This requirement can be largely weakened. For instance in the case of ReLU activation, which corresponds to the positive part $x \mapsto x_+$, one can employ any suitable sub-gradient of the x_+ function and in particular take at the non-regular point $x = 0$ any value between 0 and 1.

Proof. **Item 1**: The existence and uniqueness of the optimum is guaranteed by the assumptions of strong convexity and smoothness of $\mathcal{L}$.
Item 2: We have

$$\begin{aligned}&\mathbb{E}\left[\|X_{n+1}-X_*\|^2\right]=\mathbb{E}\left[\|X_n-X_*-\rho_n\nabla_x L(\omega_n,X_n)\|^2\right]\\&=\mathbb{E}\left[\|X_n-X_*\|^2\right]+\rho_n^2\mathbb{E}\left[\|\nabla_x L(\omega_n,X_n)\|^2\right]-2\rho_n\mathbb{E}\left[\langle X_n-X_*,\nabla_X L(\omega_n,X_n)\rangle\right].\end{aligned} \tag{9}$$

First we remark that[3]

$$\mathbb{E}\left[\langle X_n-X_*,\nabla_x L(\omega_n,X_n)\rangle\right]=\mathbb{E}\left[\langle X_n-X_*,\nabla\mathcal{L}(X_n)\rangle\right].$$

But at its turn

$$\begin{aligned}&\mathbb{E}\left[\langle X_n-X_*,\nabla\mathcal{L}(X_n)\rangle\right]\geq\mathbb{E}\left[\mathcal{L}(X_n)-\mathcal{L}(X_*)+\frac{\mu}{2}\|X_n-X_*\|^2\right]\\&\geq\frac{\mu}{2}\mathbb{E}[\|X_n-X_*\|^2],\end{aligned} \tag{10}$$

the last inequality being guaranteed by the fact that X_* is the minimum. Putting together all relations proved so far one obtains the relation (6) (we have used hypothesis (3) to bound the term $\mathbb{E}\|\nabla_x L(\omega_n,X_n)\|^2$ by $c_0+d_nc_1$).
Item 3: When ρ_n is constant equal to ρ inequality (6) is equivalent to

$$d_{n+1}-\frac{\rho c_0}{\mu-\rho c_1}\leq(1-\rho\mu+\rho^2c_1)\left(d_n-\frac{\rho c_0}{\mu-\rho c_1}\right).$$

Since the function $x\mapsto x_+$ (the positive part) is increasing we obtain for $\rho<\min(1/\mu,\mu/2c_1)$:

$$\left(d_{n+1}-\frac{\rho c_0}{\mu-\rho c_1}\right)_+\leq\left(1-\frac{\rho\mu}{2}\right)\left(d_n-\frac{\rho c_0}{\mu-\rho c_1}\right)_+,$$

and by iteration, for any $k\geq1$:

$$\left(d_{n+k}-\frac{\rho c_0}{\mu-\rho c_1}\right)_+\leq\left(1-\frac{\rho\mu}{2}\right)^k\left(d_n-\frac{\rho c_0}{\mu-\rho c_1}\right)_+.$$

Taking $k\to\infty$ we obtain $\limsup_k\left(d_k-\frac{\rho c_0}{\mu-\rho c_1}\right)_+=0$ hence the conclusion (7) for ρ smaller than $\rho_\epsilon:=\min\{1/\mu,\mu/2c_1,\epsilon\mu/(c_0-\epsilon c_1)\}$.
Item 4: For non-constant ρ_n and arbitrary fixed ϵ we obtain from (6)

$$d_{n+1}-\epsilon\leq\left(1-\frac{\rho_n\mu}{2}\right)(d_n-\epsilon)+\rho_n(c_0\rho_n-\mu\epsilon/2+(\rho_nc_1-\mu/2)d_n).$$

When n is large enough the last term in the right hand side is negative and thus

$$d_{n+1}-\epsilon\leq\left(1-\frac{\rho_n\mu}{2}\right)(d_n-\epsilon),$$

[3]The formal justification is as follows: denote by $\mathcal{F}_n$ the sigma algebra generated by X_1, ..., X_n, ω_1, ..., ω_{n-1}. In particular ω_n is independent of $\mathcal{F}_n$. Recall now that for any random variables U measurable with respect to $\mathcal{F}_n$ and V independent of $\mathcal{F}_n$: $\mathbb{E}[g(U,V)|\mathcal{F}_n]=\int g(v,U)P_V(dv)$ and in particular $\mathbb{E}[g(U,V)]=\mathbb{E}[\mathbb{E}[g(U,V)|\mathcal{F}_n]]=\mathbb{E}[\int g(v,U)P_V(dv)]$.

therefore

$$(d_{n+k} - \epsilon)_+ \leq \left(1 - \frac{\rho_n \mu}{2}\right)(d_n - \epsilon)_+ .$$

Iterating such inequalities we obtain

$$(d_{n+k} - \epsilon)_+ \leq \prod_{\ell=n}^{n+k-1} \left(1 - \frac{\rho_\ell \mu}{2}\right)(d_n - \epsilon)_+ .$$

From the Lemma 2 we obtain $\lim_{k\to\infty} (d_k - \epsilon)_+ = 0$ and since this is true for any ϵ the conclusion follows. □

Lemma 2. *Let $\xi > 0$ and ρ_n a sequence of positive real numbers such that $\rho_n \to 0$ and $\sum_{n\geq 1} \rho_n = \infty$. Then for any $n \geq 0$:*

$$\lim_{k\to\infty} \prod_{\ell=n}^{n+k} (1 - \rho_\ell \xi) = 0. \tag{11}$$

Proof. Since $\rho_n \to 0$, $\rho_\ell \xi < 1$ for ℓ large enough. To keep things simple we suppose this is true starting from n. Recall that for any $x \in]0, 1[$ we have $\log(1 - x) \leq -x$; then:

$$0 \leq \prod_{\ell=n}^{n+k} (1 - \rho_\ell \xi) = e^{\sum_{\ell=n}^{n+k} \log(1-\rho_\ell \xi)} \leq e^{\sum_{\ell=n}^{n+k} (-\rho_\ell \xi)} \overset{k\to\infty}{\longrightarrow} e^{-\infty} = 0, \tag{12}$$

which concludes the proof. □

5 Concluding remarks

We make here some remarks concerning the hypothesis and the use in Neural Networks.

First, consider the hypothesis $\sum_n \rho_n = \infty$; at first it may seem strange but this is not really so[4] . Note that in particular it is true when ρ_n is a constant. But in general, if we forget the stochastic part[5], one can interpret the SGD as following some continuous time dynamics of the type $X'(t) = -\nabla \mathcal{L}(X)$; for the simple quadratic function $\mathcal{L}(X) = \alpha \|X\|^2/2$ the dynamics is $X'(t) = -\alpha X(t)$ with solution $X(t) = e^{-\alpha t} X(0)$ needing an infinite 'time' t to converge to the minimum $X_* = 0_N$. Or here $\sum_n \rho_n$ is the discrete version of the time and thus it is not a surprise to need infinite time to obtain X_* with infinite precision. On the other hand if a finite precision is needed one can just take a constant time step as indicated in the theorem[6].

Note that an important example that satisfies (8) is $\rho_n = \frac{c_3}{c_4 + n}$, with $c_3, c_4 > 0$. In general giving a functional form for ρ_n is termed 'choosing a decay rate', but it may not be clear what the best decay rate is in general.

Finally, concerning the hypothesis (3) and (4) both can be considerably weakened, but at the cost of a longer proof.

[4] One may show on a simple counter-example $L(\omega, X) = (X - \omega)^2/2$ with ω a standard normal that $\sum_n \rho_n < \infty$ will lead to a non-null limit variance; to do so, use a second order version of lemma 2 and the formula $Var(X_{n+1}) = (1 - \rho_n)^2 Var(X_n) + \rho_n^2$ true in this case.

[5] This can be made precise when the stochastic part is added, see [1].

[6] but in this case one may spend a too long time to wait for the convergence to this small neighborhood to arrive see [1] for some ways to accelerate the convergence.

Acknowledgements

A special thanks to Stefania Anita for helpful discussions concerning this work and in particular for suggesting the present form of the hypothesis (3).

References

[1] Imen Ayadi and Gabriel Turinici. Stochastic Runge-Kutta methods and adaptive SGD-G2 stochastic gradient descent, 2020. arxiv:2002.09304, Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 8220-8227, doi: 10.1109/ICPR48806.2021.9412831.

[2] Gabriel Turinici. Stochastic learning control of inhomogeneous quantum ensembles. *Phys. Rev. A*, 100:053403, Nov 2019.

[3] Gabriel Turinici. Convergence Dynamics of Generative Adversarial Networks: The Dual Metric Flows 2020. arXiv:2012.10410; In: Del Bimbo, A., et al. Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science, vol 12661. Springer, Cham. `https://doi.org/10.1007/978-3-030-68763-2_47`

[4] Gabriel Turinici. Radon–Sobolev Variational Auto-Encoders. *Neural Networks*, 141:294-305, 2021.