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`

March 26, 2021

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 (see [1–4] for some of my recent research papers in this area); on the contrary please cite this work if you find it useful (arxiv or DOI: 10.5281/zenodo.4638695).

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 is bounded:

$$\exists B > 0 : \ \sup_{\omega} \|\nabla_X L(\omega, X)\|^2 \leq B, \ \forall X \in \mathbb{R}^N. \tag{3}$$

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

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

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

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

   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

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

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

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

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).

Item 3: When ρ_n is constant equal to ρ inequality (6) is equivalent to

$$d_{n+1} - \rho\frac{B}{\mu} \leq (1 - \rho\mu)(d_n - \rho\frac{B}{\mu}).$$

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

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

and by iteration, for any $k \geq 1$:

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

Taking $k \to \infty$ we obtain $\limsup_k \left(d_k - \rho\frac{B}{\mu}\right)_+ = 0$ hence the conclusion (7) for ρ smaller then $1/\mu$ and $\epsilon\mu/B$.

Item 4: For non-constant ρ_n and arbitrary fixed ϵ we obtain from (6)

$$d_{n+1} - \epsilon \leq (1 - \rho_n\mu)(d_n - \epsilon) + \rho_n(\rho_n B - \mu\epsilon).$$

When n is large enough $\rho_n(\rho_n B - \mu\epsilon) \leq 0$ and thus

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

therefore

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

---

[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 $\mathbb{E}[g(U,V)] = \mathbb{E}[\mathbb{E}[g(U,V)|\mathcal{F}_n]] = \mathbb{E}[\int g(v,U)P_V(dv)]$.

Iterating such inequalities we obtain

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

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. $\qquad\square$

Lemma 2. *Let $\mu > 0$ 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 $n \geq 0$:*

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

Proof. Recall that for any $x \in]0, 1[$ we have $\log(1 - x) \leq -x$; then:

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

which concludes the proof. $\qquad\square$

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. Note that in particular it is true when ρ_n is a constant. But in general, if we forget the stochastic part[4], 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 $X(t) = e^{-\alpha t} X(0)$ needs 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[5].

On the other hand, an important example that satisfies (8) is $\rho_n = \frac{c_1}{c_2 + n}$, with $c_1, c_2 > 0$. 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.

References

[1] Imen Ayadi and Gabriel Turinici. Stochastic Runge-Kutta methods and adaptive SGD-G2 stochastic gradient descent, 2020. arxiv:2002.09304, ICPR2020 paper.

---

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

[5]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.

[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; CADL-ICPR 2020 workshop paper.

[4] Gabriel Turinici. X-Ray Sobolev Variational Auto-Encoders, 2020. arxiv:1911.13135.