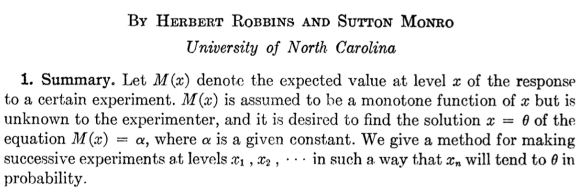
Adding some noise

Introduction to the Stochastic Gradient Descent algorithm.

Samuel Vaiter

2024-02-09

Table of contents



A STOCHASTIC APPROXIMATION METHOD[1]

By Herbert Robbins and Sutton Monro

University of North Carolina

1. Summary. Let $M(x)$ denote the expected value at level x of the response to a certain experiment. $M(x)$ is assumed to be a monotone function of x but is unknown to the experimenter, and it is desired to find the solution $x = \theta$ of the equation $M(x) = \alpha$, where α is a given constant. We give a method for making successive experiments at levels $x_1, x_2, \cdots$ in such a way that x_n will tend to θ in probability.

Figure 1: Fundamental paper of stochastic approximation by Herbert Robbins and Monro (1951)

Stochastic gradient descent (SGD) is the workhorse of modern machine learning. Almost all recent achievements ($\geq$ 2010s) in computer vision, natural language processing are somehow connected to the use of SGD (or one of its multiple generalization). Despite this modern aspect, SGD takes its root in the seminal paper of Herbert Robbins and Monro (1951), more than 70 years ago. The key insight of Robbins and Monro is that **first orders method are robust**. What does it means? The gradient descent does *not* need an exact computation of the gradient, it only need that *in average* the computation is correct up to some "stochastic" errors. We will formalize this idea in this chapter. again inspired by the lecture notes of Bach (2021). A key insight from the seminal paper by Bottou and Bousquet (2007) is that in machine learning, high precision is not required since it is meaningless to optimize below the statistical error. This paper considerably relaunched the interest towards SGD, especially for empirical risk minimization.

Stochastic approximation

Let $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space. We consider the following stochastic approximation problem

$$x^\star \in \operatorname*{argmin}_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_{\xi \sim \mathbb{P}}[F(x; \xi)]. \tag{1}$$

Observe that Equation 1 encompasses two important problem of supervised machine learning: expected risk minimization and empirical risk minimization. Let us recall basic fact on risk minimization. Our goal is given observations[1] (couple of features and responses) $(c_i, r_i) \in C \times R$, for $i = 1, \dots, N$, the goal is to predict a new response $r \in R$ given a new feature $f \in C$. Consider now the test distribution $\mathbb{P}$ on $\Omega = C \times R$ and a loss function $\ell : R \times R \to \mathbb{R}$.

- *Expected risk minimization.* The expected risk (or simply risk) of a predictor $\phi : C \to R$ is the quantity

  $$\mathbb{R}(\phi) = \mathbb{E}_{(c,r)\sim\mathbb{P}}\left[\ell(r, \phi(c))\right] = \int_{C \times R} \ell(r, \phi(c)) d\mathbb{P}(c, r).$$

  This minimization problem fits the framework of Equation 1 with $x = (c, r)$, $F = \ell$ and $\mathbb{P}$ is the joint probability on $F \times R$.
- *Empirical risk minimization.* The practical counterpart of the expected risk is the empirical risk defined as

  $$\hat{\mathbb{R}}(\phi) = \frac{1}{N} \sum_{i=1}^{N} \ell(r_i, f(c_i)).$$

  Again, this minimization problem fits the framework of Equation 1 with $x = (c, r)$, $F = \ell$, but now $\mathbb{P}$ is the empirical distribution based on the test data (c_i, r_i).

Stochastic gradient descent

We suppose that we have access to stochastic estimation $g_{t+1}(x^t)$ of the gradient $\nabla f(x^{(t)})$, i.e.,

$$x^{(t+1)} = x^{(t)} - \eta^{(t)} g^{(t+1)}(x^{(t)}; \xi^{(t)}). \tag{2}$$

The stochastic gradient algorithm is NOT a *descent* method in the sense of of the oracle $g^{(t+1)}(x^{(t)}; \xi^{(t)})$ is not necessarly a descent direction as defined for instance in our lecture on the gradient descent.

Two situations:

- *Stochastic approximation.* Here we have access to an additive noisy version of the "true" gradient:

  $$g^{(t)}(x; \xi) = \nabla f(x) + \xi.$$

[1] Variables c and r are typically denoted x and y in a ML courses, but we want to keep them free for the optimization variable.

- *Learning from samples.* On the contrary, here we have for instance $\mathbb{P}$ that is a uniform choice between $i \in 1, \dots, N$:
$$g^{(t)}(x;\xi) = \nabla f_\xi(x),$$
where $\xi \sim \mathcal{U}(1, \dots, N)$.

The standard assumption when dealing with SGD is the access to an *unbiased* estimator of the gradient.

Unbiasedness of the stochastic gradient estimate

$$\mathbb{E}_\xi \left[g_{t+1}(x^{(t)};\xi) \middle| x^{(t)} \right] = \nabla f(x^{(t)}). \tag{3}$$

Note that it is possible to deal with biased oracles, but this goes beyond the limit of this course.

We are going to study the convergence of SGD *in expectation*, that is the main way to analyze in the machine learning community the stochastic convergence. Nevertheless, the stochastic approximation literature, starting from (H. Robbins and Siegmund 1971), are often concerned with *almost sure* convergence based on martingale tools.

For convex function, we will assume that the noisy gradient $g^{(t)}$ are bounded almost-surely.

Bounded gradient

$$\exists B > 0,\ \forall t \in \mathbb{N}, \quad \|g^{(t+1)}(x^t;\xi)\|_2^2 \leq B^2 \quad \mathbb{P}\text{-almost surely.} \tag{4}$$

The simplest strategy to choose the step size, as for the batch gradient descent, is to fix some $\eta^{(t)} = \eta > 0$. Unfortunately, this method *does not* converge towards a minimizers of Equation 1! It is nevertheless, the most frequent use of SGD in practice. Why? Because, despite the nonconvergence, it is possible to show that the iterations are "close enough" to the true minimizer in the following sense (even if f is not convex).

Theorem 0.1 (Convergence up to a ball of SGD)**.** *Assume that f is L-smooth and Equation 3, Equation 4 are satisfied. For an initial guess $x^{(0)}$, the SGD iterates $x^{(t)}$ with constant step-size $\eta^{(t)} = \eta$ satisfy*
$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \|\nabla f(x^{(t)})\|^2\right] \leq \frac{f(x^{(0)}) - f^\star}{T\eta} + \frac{\eta L B^2}{2},$$
where $f^\star$ is the optimal value of Equation 1.

A consequence is that when T goes to $+\infty$, the expected squared-norm of the gradient of a point uniformly chosen from the trajectory of SGD lies in a ball of radius $\frac{\eta L B^2}{2} \propto \eta$ centered in 0. This is illustrated in the video below

anim_iterates.mp4

The situation is radically different from the batch mode where the asymptotic regime produce a true solution.

Proof of Theorem Theorem 0.1. This is a slight adaptation of the descent lemma from the batch gradient descent proof. We start from

$$f(x^{(t+1)}) \leq f(x^{(t)}) - \eta \langle g^{(t+1)}(x^t; \xi^{(t)}),\, \nabla f(x^{(t)}) \rangle + \eta^2 \frac{L}{2} \| g^{(t+1)}(x^t; \xi^{(t)} \|^2.$$

In the light of Equation 3, Equation 4, we have thus taking the expectation

$$\mathbb{E}[f(x^{(t+1)})] \leq \mathbb{E}[f(x^{(t)})] - \eta \mathbb{E}[\|\nabla f(x^{(t)})\|^2] + \eta^2 \frac{LB^2}{2}.$$

Telescoping the difference between $f(x^{(t+1)})$ and $f(x^{(t)})$, we have

$$\mathbb{E}[f(x^{(T)}) - f(x^{(0)})] \leq \frac{T\eta^2 LB^2}{2} - \eta \sum_{t=0}^{T-1} \|\nabla f(x^{(t)})\|^2.$$

We obtain the claimed result by dividing by $T\eta$ and observing that $f(x^{(T)}) \geq f^\star$. □

Convergence rate in the convex setting

We now dive away from the constant stepsize framework, and will study what happens when $\eta^{(t)}$ varies accross iterations. The following theorem gives the $O(1/\sqrt{t})$ rate of SGD in the convex case.

Theorem 0.2 (Rate of Equation 2 for convex functions)**.** *Assume that f is convex. For an initial guess $x^{(0)}$, suppose that f has a minimizer $x^\star$ satisfying*

$$\|x^\star - x^{(0)}\| \leq D,$$

for some $D > 0$. Assume that Equation 3, Equation 4 are satisfied. Then, the SGD iterates $x^{(t)}$ with step-size $\eta^{(t)} = \frac{D}{B\sqrt{t}}$ satisfy[2]

$$\mathbb{E}\left[f(\bar{x}^{(t)}) - f(x^\star)\right] \leq DB \frac{2 + \log(t)}{2\sqrt{t}},$$

where $\bar{x}^{(t)}$ is the averaged version of $x^{(t)}$, i.e.,

$$\bar{x}^{(t)} = \frac{1}{\sum_{k=1}^t \eta^{(k)}} \sum_{k=1}^t \eta^{(k)} x^{(k-1)}.$$

[2]Note that f is *not* random, but $\bar{x}^{(t)}$ is, so that $\mathbb{E}[\bar{x}^{(t)}]$ is meaningful.

Proof. Let $t \geq 1$. We have the following decomposition

$$\begin{aligned}\|x^{(t+1)} - x^\star\|^2 &= \|x^t - \eta^{(t)} g^{(t+1)}(x^{(t)}; \xi^{(t)}) - x^\star\|^2 \\ &= \|x^{(t)} - x^\star\|^2 - 2\eta^{(t)} \langle g^{(t+1)}(x^{(t)}; \xi^{(t)}),\, x^{(t)} - x^\star \rangle \\ &\quad + (\eta^{(t)})^2 \|g^{(t+1)}(x^{(t)}; \xi^{(t)})\|^2.\end{aligned}$$

Using the linearity of the expectation, we get

$$\begin{aligned}\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] = \mathbb{E}[\|x^{(t)} - x^\star\|^2] &- 2\eta^{(t)} \mathbb{E}[\langle g^{(t+1)}(x^{(t)}; \xi^{(t)}),\, x^{(t)} - x^\star \rangle] \\ &+ (\eta^{(t)})^2 \mathbb{E}[\|g^{(t+1)}(x^{(t)}; \xi^{(t)})\|^2].\end{aligned}$$

We keep the first term as it is, the last term is bounded by

$$(\eta^{(t)})^2 \mathbb{E}[\|g^{(t+1)}(x^{(t)}; \xi^{(t)})\|^2] \leq (\eta^{(t)})^2 B^2 \quad \text{a.s.} \qquad \text{(bounded gradient)}$$

To bound the middle term, we use the total law of conditional expectation

$$\begin{aligned}&\mathbb{E}\left[\langle g^{(t+1)}(x^{(t)}; \xi^{(t)}),\, x^{(t)} - x^\star \rangle\right] \\ =&\, \mathbb{E}\left[\mathbb{E}[\langle g^{(t+1)}(x^{(t)}; \xi^{(t)}),\, x^{(t)} - x^\star \rangle | x^{(t)}]\right] && \text{(total law)} \\ =&\, \mathbb{E}\left[\langle \mathbb{E}[g^{(t+1)}(x^{(t)}; \xi^{(t)}) | x^{(t)}],\, x^{(t)} - x^\star \rangle\right] && \text{(linearity of cond. expectation)} \\ =&\, \mathbb{E}\left[\langle \nabla f(x^{(t)}),\, x^{(t)} - x^\star \rangle\right] && \text{(unbiasedness)} \\ \leq&\, \mathbb{E}\left[f(x^{(t)}) - f(x^\star)\right],\end{aligned}$$

where the last inequality comes from the differential characterization of a convex function. Combining these three terms gives

$$\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] \leq \mathbb{E}[\|x^{(t)} - x^\star\|^2] - 2\eta^{(t)} \mathbb{E}[f(x^{(t)}) - f(x^\star)] + (\eta^{(t)})^2 B^2. \tag{5}$$

Reorganizing the terms, we have

$$\eta^{(t)} \mathbb{E}[f(x^{(t)}) - f(x^\star)] \leq \frac{1}{2}\left(\mathbb{E}[\|x^{(t)} - x^\star\|^2] - \mathbb{E}[\|x^{(t+1)} - x^\star\|^2]\right) + \frac{1}{2}(\eta^{(t)})^2 B^2.$$

We observe that summing over the iterations $t = 0 \dots T-1$ leads to a natural telescoping sum on the first term of the right-hand side:

$$\sum_{t=0}^{T-1} \eta^{(t)} \mathbb{E}[f(x^{(t)}) - f(x^\star)] \leq \frac{1}{2}\left(\mathbb{E}[\|x^{(0)} - x^\star\|^2] - \mathbb{E}[\|x^{(T)} - x^\star\|^2]\right) + \frac{1}{2} B^2 \sum_{t=0}^{T-1} (\eta^{(t)})^2.$$

Since $\mathbb{E}[\|x^{(T)} - x^\star\|^2]$ is nonnegative, we have the cruder bound

$$\sum_{t=0}^{T-1} \eta^{(t)} \mathbb{E}[f(x^{(t)}) - f(x^\star)] \leq \frac{1}{2}\|x^{(0)} - x^\star\|^2 + \frac{1}{2} B^2 \sum_{t=0}^{T-1} (\eta^{(t)})^2.$$

Dividing by $\sum_{t=0}^{T-1} \eta^{(t)}$, we obtain

$$\frac{1}{\sum_{t=0}^{T-1} \eta^{(t)}} \sum_{t=0}^{T-1} \eta^{(t)} \mathbb{E}[f(x^{(t)}) - f(x^\star)] \leq \frac{\|x^{(0)} - x^\star\|^2}{2 \sum_{t=0}^{T-1} \eta^{(t)}} + \frac{B^2 \sum_{t=0}^{T-1} (\eta^{(t)})^2}{2 \sum_{t=0}^{T-1} \eta^{(t)}}. \tag{6}$$

The left-hand side of Equation 6 is an upper-bound of $\mathbb{E}[f(\bar{x}^{(t)}) - f(x^\star)]$. Indeed,

$$\begin{aligned}
\frac{1}{\sum_{t=0}^{T-1} \eta^{(t)}} \sum_{t=0}^{T-1} \eta^{(t)} \mathbb{E}[f(x^{(t)}) - f(x^\star)] &= \mathbb{E}\left[\frac{1}{\sum_{t=0}^{T-1} \eta^{(t)}} \sum_{t=0}^{T-1} \eta^{(t)} (f(x^{(t)}) - f(x^\star)) \right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{\eta^{(t)}}{\sum_{t=0}^{T-1} \eta^{(t)}} f(x^{(t)}) - f(x^\star) \right] \\
&\geq \mathbb{E}\left[f\left(\sum_{t=0}^{T-1} \frac{\eta^{(t)}}{\sum_{t=0}^{T-1} \eta^{(t)}} x^{(t)} \right) - f(x^\star) \right] \\
&= \mathbb{E}[f(\bar{x}^{(t)}) - f(x^\star)],
\end{aligned}$$

where the inequality comes from the Jensen's inequality.

We now look at an upper-bound of the right-hand side. We need two things:

- The sum $\sum_{t=0}^{T-1} \eta^{(t)}$ needs to diverge to $+\infty$ so that the first term goes to 0.
- The sum $\sum_{t=0}^{T-1} \eta^{(t)}$ needs to diverge to $+\infty$ more "quickly" that the sum of squares $\sum_{t=0}^{T-1} (\eta^{(t)})^2$ so that the second term goes also to 0.

A typical example of such a sequence is $\eta^{(t)} = \alpha(t+1)^{-1/2}$ for some $\alpha > 0$. Indeed, we have

$$\frac{\|x^{(0)} - x^\star\|^2}{2 \sum_{t=0}^{T-1} \eta^{(t)}} + \frac{B^2 \sum_{t=0}^{T-1} (\eta^{(t)})^2}{2 \sum_{t=0}^{T-1} \eta^{(t)}} \leq \frac{D^2 + B^2 \alpha^2 \sum_{t=1}^{T} t^{-1}}{2\alpha \sum_{t=1}^{T} t^{-1/2}}.$$

Using the fact that

$$\sum_{t=1}^{T} t^{-1/2} \geq \sum_{t=1}^{T} T^{-1/2} = \frac{T}{\sqrt{T}} = \sqrt{T},$$

we have that

$$\frac{D^2 + B^2 \alpha^2 \sum_{t=1}^{T} t^{-1}}{2\alpha \sum_{t=1}^{T} t^{-1/2}} \leq \frac{1}{2\alpha\sqrt{T}} \left(D^2 + B^2 \alpha^2 \sum_{t=1}^{T} t^{-1} \right).$$

We bound the harmonic serie $\sum_{t=1}^{T} t^{-1}$ by $1 + \log T$, hence

$$\frac{\|x^{(0)} - x^\star\|^2}{2 \sum_{t=0}^{T-1} \eta^{(t)}} + \frac{B^2 \sum_{t=0}^{T-1} (\eta^{(t)})^2}{2 \sum_{t=0}^{T-1} \eta^{(t)}} \leq \frac{1}{2\alpha\sqrt{T}} \left(D^2 + B^2 \alpha^2 (1 + \log T) \right).$$

We obtain the claimed bound by setting $\alpha = D/B$. □

Several remarks needs to be stated:

- As claimed before, SGD *is not* a descent method. Moreover, we proved here the convergence *in expectation* of the *value function* of the *averaged* iterates $x^{(t)}$. Note that *almost-sure* convergence can be proved, but with more involved arguments, see for instance (Sebbouh, Gower, and Defazio 2021) and references therein.
- It is possible to replace the noisy gradient step by a *projected* noisy gradient step, i.e., using the algorithm defined by

$$x^{(t+1)} = \Pi_{B(x^{(0)},D)}\left(x^{(t)} - \eta^{(t)} g^{(t+1)}(x^{(t)};\xi^{(t)})\right).$$

  The obtained results are exactly the same.
- It is also possible to remove the assumption that $x^{(0)}$ lives in a neighborhood of an optimal solution $x^\star$ and requires instead that F is B-smooth. Note that in this case, we do not obtain better results (contrary to the nonstochastic setting).
- The bound can be shown to optimal (Agarwal et al. 2012, Theorem 1) (at least in the Lipschitz setting).

Minibatch Stochastic Gradient Descent

The minibatch idea is to make a tradeoff between the fast rate of the gradient descent but linear depency on N, and the slow rate of SGD but dimension-free dependency: we reduce the variance of the noise, but have a higher running time. Minibatching can also take advantage of parallel architectures such as GPU or SIMD instructions on CPU. A minibatch is built upon B call to the stochastic oracle, and its SGD step associated is written as

$$x^{(t+1)} = x^{(t)} - \frac{\eta^{(t)}}{M}\sum_{m=1}^{M} g^{(t+1)}(x^{(t)};\xi_b^{(t)}) \quad \text{where} \quad (\xi_1^{(t)},\dots,\xi_M^{(t)}) \sim \mathcal{M}(M,N).$$

Observing that $\tilde{g}^{(t+1)}(x,\boldsymbol{\xi}) = \frac{1}{M}\sum_{m=1}^{M} g^{(t+1)}(x;\xi_m)$ is an unbiased estimator of the gradient of f, indeed

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}}\left[\tilde{g}^{(t+1)}(x^{(t)};\boldsymbol{\xi})\middle| x^{(t)}\right] &= \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{M}\sum_{m=1}^{M} g^{(t+1)}(x^{(t)};\xi_m)\middle| x^{(t)}\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\boldsymbol{\xi}}\left[g^{(t+1)}(x^{(t)};\xi_m)\middle| x^{(t)}\right] && \text{by linearity} \\
&= \frac{1}{M}\sum_{m=1}^{M}\nabla f(x^{(t)}) = \nabla f(x^{(t)}) && \text{unbiasedness.}
\end{aligned}$$

The minibatch gradient is also bounded, indeed

$$\begin{aligned}\|\tilde{g}^{(t+1)}(x^{(t)};\boldsymbol{\xi})\|_2^2 &= \|\frac{1}{M}\sum_{m=1}^{M} g^{(t+1)}(x^{(t)};\xi_m)\|_2^2 \\ &\leq \frac{1}{M^2}\sum_{m=1}^{M}\|g^{(t+1)}(x^{(t)};\xi_m)\|_2^2 \\ &\leq \frac{B^2}{M} \quad \mathbb{P}\text{-almost surely.}\end{aligned}$$

We can hence apply Theorem 0.2. A lot of effort has been done in the last year to determine what is the optimal minibatch size, see e.g., Gower et al. (2019), or to derive schedule policy with increasing size of the minibatch.

Agarwal, Alekh, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. 2012. "Information-Theoretic Lower Bounds on the Oracle Complexity of Stochastic Convex Optimization." *IEEE Transactions on Information Theory* 58 (5): 3235–49. https://doi.org/10.1109/TIT.2011.2182178.

Bach, Francis. 2021. "Learning Theory from First Principles." https://www.di.ens.fr/~fbach/ltfp_book.pdf.

Bottou, Léon, and Olivier Bousquet. 2007. "The Tradeoffs of Large Scale Learning." In *Advances in Neural Information Processing Systems*. Vol. 20. Curran Associates, Inc. https://papers.nips.cc/paper/2007/hash/0d3180d672e08b4c5312dcdafdf6ef36-Abstract.html.

Gower, Robert Mansel, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. 2019. "SGD: General Analysis and Improved Rates." In *Proceedings of the 36th International Conference on Machine Learning*, 5200–5209. PMLR. https://proceedings.mlr.press/v97/qian19b.html.

Robbins, Herbert, and Sutton Monro. 1951. "A Stochastic Approximation Method." *Ann. Math. Stat.* 22 (3): 400–407. https://doi.org/10.1214/aoms/1177729586.

Robbins, H., and D. Siegmund. 1971. "A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications." In *Optimizing Methods in Statistics*, edited by Jagdish S. Rustagi, 233–57. Academic Press. https://doi.org/10.1016/B978-0-12-604550-5.50015-8.

Sebbouh, Othmane, Robert M. Gower, and Aaron Defazio. 2021. "Almost Sure Convergence Rates for Stochastic Gradient Descent and Stochastic Heavy Ball." In *Proceedings of Thirty Fourth Conference on Learning Theory*, 3935–71. PMLR. https://proceedings.mlr.press/v134/sebbouh21a.html.