Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains

Aymeric Dieuleveut  
Ecole Polytechnique Fédérale de Lausanne  
aymeric.dieuleveut@epfl.ch

Alain Durmus  
Ecole Normale Supérieure de Cachan  
durmus@cmla.ens-cachan.fr

Francis Bach  
Ecole Normale Supérieure de Paris, Inria  
francis.bach@ens.fr

Abstract

We consider the minimization of an objective function given access to unbiased estimates of its gradient through stochastic gradient descent (SGD) with constant step-size. While the detailed analysis was only performed for quadratic functions, we provide an explicit asymptotic expansion of the moments of the averaged SGD iterates that outlines the dependence on initial conditions, the effect of noise and the step-size, as well as the lack of convergence in the general (non-quadratic) case. For this analysis, we bring tools from Markov chain theory into the analysis of stochastic gradient. We then show that Richardson-Romberg extrapolation may be used to get closer to the global optimum and we show empirical improvements of the new extrapolation scheme.

1 Introduction

We consider the minimization of an objective function given access to unbiased estimates of the function gradients. This key methodological problem has raised interest in different communities: in large-scale machine learning [4, 17, 18], optimization [10, 11], and stochastic approximation [6, 13, 16]. The most widely used algorithms are stochastic gradient descent (SGD), a.k.a. Robbins-Monro algorithm [15], and some of its modifications based on averaging of the iterates [13, 14, 19].

While the choice of the step-size may be done robustly in the deterministic case [see, e.g., 3], this remains a traditional theoretical and practical issue in the stochastic case. Indeed, early work suggested to use step-size decaying with the number \( k \) of iterations as \( O(1/k) \) [15], but it appeared to be non-robust to ill-conditioning and slower decays such as \( O(1/\sqrt{k}) \) together with averaging lead to both good practical and theoretical performance [1].

We consider in this paper constant step-size SGD, which is often used in practice. Although the algorithm is not converging in general to the global optimum of the objective function, constant step-sizes come with benefits: (a) there is single parameter value to set as opposed to the several choices of parameters to deal with decaying step-sizes, e.g., as \( 1/(\Box k + \Delta)^\gamma \); the initial conditions are forgotten exponentially fast for well-conditioned (e.g., strongly convex) problems [8, 9], and the performance, although not optimal, is sufficient in practice (in a machine learning set-up, being only 0.1% away from the optimal prediction often does not matter).

The main goals of this paper are (a) to gain a complete understanding of the properties of constant-step-size SGD in the strongly convex case, and (b) to propose provable improvements to get closer to the optimum when precision matters or in high-dimensional settings. We consider the iterates of the SGD recursion on \( \mathbb{R}^d \) defined starting from \( \theta_0 \in \mathbb{R}^d \), for \( k \geq 0 \), and a step-size \( \gamma > 0 \) by

\[
\theta_k^{(\gamma)} = \theta_k^{(\gamma)} - \gamma \left[ f' \left( \theta_k^{(\gamma)} \right) + \varepsilon_{k+1}^{(\gamma)} \right],
\]

where \( f \) is the objective function to minimize (in machine learning the generalization performance), \( \varepsilon_{k+1}^{(\gamma)} \) the zero-mean statistically independent noise (in machine learning, obtained from a single i.i.d. observation of a data point). Following Bach and Moulines [2], we leverage the property that the sequence of iterates \( (\theta_k^{(\gamma)})_{k \geq 0} \) is an homogeneous Markov chain.

This interpretation allows us to capture the general behavior of the algorithm. In the strongly convex case, this Markov chain converges exponentially fast to the unique stationary distribution \( \pi \), highlighting the facts that
Let $f : \mathbb{R}^d \to \mathbb{R}$ be an objective function, satisfying the following assumptions: we assume the function $f$ to be strongly convex with strong convexity constant $\mu$, namely $f - f(\theta) \geq \frac{\mu}{2} \| \cdot \|^2$ is convex; and that $f$ is five times continuously differentiable with uniformly second to fifth bounded derivatives. Especially $f$ is $L$-smooth: $\forall \theta \in \mathbb{R}^d$, the largest eigenvalue of $f''(\theta)$ is less than $L$.
Regarding the sequence of random functions \((\varepsilon_k)_{k \geq 1}\) (in Eq. 1)), we assume that there exists a filtration \((\mathcal{F}_k)_{k \geq 0}\) (i.e., for all \(k \in \mathbb{N}\), \(\mathcal{F}_k \subset \mathcal{F}_{k+1}\)) on some probability space \((\Omega, \mathcal{F}, P)\) such that for any \(k \in \mathbb{N}\), for any \(\theta \in \mathbb{R}^d\), \(\varepsilon_{k+1}(\theta)\) is an \(\mathcal{F}_{k+1}\)-measurable random variable and \(\mathbb{E}[\varepsilon_{k+1}(\theta) | \mathcal{F}_k] = 0\). In addition, \((\varepsilon_k)_{k \in \mathbb{N}}\), are independent and identically distributed (i.i.d.) random fields. Finally, \(\varepsilon_k(\theta_*)\) admits bounded moments up to the order 4: \(\mathbb{E}[\varepsilon_k(\theta_*)^4] < \infty\).

We observe a noisy gradient \(f_{k+1}'(\theta)_k = f'_{k+1}(\theta) + \varepsilon_{k+1}(\theta)\) which is an unbiased estimator of \(f'\). We assume that for any \(k \in \mathbb{N}^*\), \(f_k'\) is almost surely \(L\)-coercive [22]: that is, for any \(\eta, \theta \in \mathbb{R}^d\), \(L(f_k' (\theta) - f_k'(\eta), \theta - \eta) \geq \|f_k'(\theta) - f_k'(\eta)\|^2\).

**Example: learning from i.i.d. observations.** Our main motivation comes from machine learning: namely, we consider sets \(X, Y\), and a convex loss function \(\ell : X \times Y \times \mathbb{R}^d \to \mathbb{R}\). The objective function is the generalization error \(f_\delta(\theta) = \mathbb{E}_{X,Y}[\ell(X, Y, \theta)]\). Given \(n \in \mathbb{N}\) i.i.d. observations \((x_k, y_k)_{k \in [1:n]}\), for any \(k \in [1:n]\), we define
\[
f_k(\cdot) = \ell(x_k, y_k, \cdot)
\]
the loss with respect to observation \(k\). SGD then corresponds to following gradient of the loss on a single independent observation \((x_k, y_k)_{k \geq 1}\) at each step. Assumption on the noise is then satisfied with \(\mathcal{F}_k := \sigma((x_j, y_j)_{1 \leq j \leq k})\) in least-squares regression, \(X = \mathbb{R}^d, Y = \mathbb{R}\), and the loss function \(\ell(X, Y, \theta) = ((X, Y - \theta)^2, \) and the function \(f_k\) is then quadratic. In logistic regression, \(\ell(X, Y, \theta) = \log(1 + \exp(-Y(X, \theta)))\).

### 2.1. Limit distribution.
A first step is to prove the existence of a unique stationary distribution \(\pi_\gamma\) for the Markov chain \((\theta_k(\gamma))_{k \geq 0}\) and convergence to \(\pi_\gamma\). A fundamental tool in Markov chain theory is the *Markov kernel* [7], which is the equivalent for continuous spaces of the *transition matrix* in finite state spaces.

**Markov kernel.** We denote \(R_\gamma\) the Markov kernel associated with the SGD iterates \((\theta_k^{(\gamma)})_{k \geq 0}\). For all starting points \(\theta \in \mathbb{R}^d\), for \(k \in \mathbb{N}\), \(\delta_0 R_k^{\gamma}\) is the distribution of \(\theta_k^{(\gamma)}\) starting at \(\theta\). Moreover, for a function \(\varphi : \mathbb{R}^d \to \mathbb{R}\) we define \(R_k^{\gamma}\varphi : \mathbb{R}^d \to \mathbb{R}\) such that \(R_k^{\gamma}\varphi : \theta \mapsto \int_{\mathbb{R}^d} \varphi(\eta) \delta_0 R_k^{\gamma}(d\eta)\).

To show that \((\theta_k^{(\gamma)})_{k \geq 0}\) admits a unique stationary distribution \(\pi_\gamma\) and describe the convergence of \((\delta_0 R_k^{\gamma})_{k \geq 0}\) to \(\pi_\gamma\), we use the Wasserstein distance \(W_2\) [21].

**Theorem 1.** For any step size \(\gamma < L^{-1}\), the Markov chain \((\theta_k^{(\gamma)})_{k \geq 0}\), defined by the recursion (1), admits a unique stationary distribution \(\pi_\gamma\), with finite second order moment. In addition for all \(\theta \in \mathbb{R}^d, k \in \mathbb{N}\):
\[
W_2^2(\delta_0 R_k^{\gamma}, \pi_\gamma) \leq (1 - 2\mu\gamma(1 - \gamma L))k \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 d\pi_\gamma(\vartheta).
\]

To prove the existence of the limit, one shows that for any \(\theta \in \mathbb{R}^d\), \((\delta_0 R_k^{\gamma})_{k \geq 0}\) is a Cauchy sequence in a particular Polish space. The existence of this limit allows to analyze the behavior of the chain under the limit distribution and the position of the limit point. Moreover, the speed of convergence allows so show the existence of solutions to the Poisson equation, that characterize the speed of convergence of the bias and variance term.

### 2.2. Behavior under the limit distribution, expansion of \(\theta^{(\gamma)}\) around \(\theta_*\), as \(\gamma \to 0\).
When \(f\) is a quadratic function, \(\theta_\gamma = \theta_*\). Indeed, since \(\pi_\gamma\) is invariant for \((\theta_k^{(\gamma)})_{k \geq 0}\), if \(\theta_0^{(\gamma)}\) is distributed according to \(\pi_\gamma\), then so is \(\theta_1^{(\gamma)}\). Thus as \(\theta_1^{(\gamma)} = \theta_0^{(\gamma)} - \gamma f'_{0}^{(\gamma)} + \gamma \varepsilon_1^{(\gamma)}\) taking expectations on both sides leads to \(\int_{\mathbb{R}^d} f'(\vartheta) d\pi_\gamma(\vartheta) = 0\). For a quadratic function, the gradient is linear: \(\int_{\mathbb{R}^d} f'(\vartheta) d\pi_\gamma(\vartheta) = f'(\theta_\gamma) = 0\), thus \(\theta_\gamma = \theta_*\). This highlights the crucial fact that for a quadratic function, the mean under the limit distribution is the optimal point, which explains why averaged least-mean squares algorithm does not saturate with constant step size [2]. If \(f\) is not quadratic but regular enough, we have the following first order development:

**Theorem 2.** For \(\gamma \to 0\), we have \(\theta_\gamma = \theta_* + \gamma \Delta + O(\gamma^2)\).

Combining Theorems 1 and 2, we get that for \(\gamma\) small enough and all \(k \geq 1\), \(\mathbb{E}(\bar{\theta}_k^{(\gamma)} - \theta_*) = \frac{A(\theta_0, \gamma)}{k} + \gamma \Delta + O(\gamma^2) + O(e^{-k\mu\gamma})\). This expansion in the step size \(\gamma\) shows that a Richardson-Romberg extrapolation can be used to have better estimates of \(\theta_\gamma\). Consider the average iterates \((\bar{\theta}_k^{(\gamma)})_{k \geq 0}\) and \((\bar{\theta}_k^{(\gamma)})_{k \geq 0}\) associated with SGD with step size \(2\gamma\) and \(\gamma\) respectively. Then
\[
\mathbb{E}(2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)} - \theta_*) = \frac{2A(\theta_0, \gamma) - A(\theta_0, 2\gamma)}{k} + O(\gamma^2) + O(e^{-k\mu\gamma}),
\]
and therefore this very simple trick improves the convergence by a factor of $\gamma$. In practice, while the unaveraged gradient iterate $\tilde{\theta}_k^{(\gamma)}$ saturates rapidly, $\bar{\theta}_k^{(\gamma)}$ may already perform well enough to avoid saturation on real data-sets [2], and Richardson-Romberg extrapolated iterate $2\bar{\theta}_k^{(\gamma)} - \bar{\theta}_k^{(2\gamma)}$ rarely reaches saturation.

### 2.3. Expansion for a given $\gamma > 0$ when $k$ tends to $+\infty$

We also show that the convergence of $\tilde{\theta}_k^{(\gamma)}$ to $\bar{\theta}_\gamma$, when $k \to \infty$, and the decay of $\mathbb{E}[[\tilde{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma]^2]$ to 0 can be described very precisely. The expected squared distance decomposes as a sum of a bias term, that scales as $k^{-2}$ and depends on the starting point $\theta_0$, plus a variance term, that scales as $k^{-1}$ and only depends on the asymptotic distribution $\pi_\gamma$, plus linearly decaying residual terms. The asymptotic bias and variance can be easily expressed as moments of solutions to several Poisson equations.

**Poisson equation.** For any (locally-) Lipschitz function $\varphi : \mathbb{R}^d \to \mathbb{R}^q$, there exists a function $\psi_\gamma : \mathbb{R}^d \to \mathbb{R}^q$ that satisfies $\pi_\gamma(\psi_\gamma) = 0$, and $\pi_\gamma(I - R_\gamma)\psi_\gamma = \varphi$. We call it Poisson solution associated with $\varphi$. Let $\psi_\gamma$, $\chi_\gamma$ and $\chi_\gamma^2$ be the Poisson solutions associated respectively to $\theta \mapsto \theta - \theta_\gamma$, $\theta \mapsto \|\psi_\gamma(\theta)\|^2$ and $\theta \mapsto \|\psi_\gamma(\theta) - \varphi(\theta)\|^2$.

**Theorem 3 (Convergence of the Markov chain).** Let $\gamma \in (0, 1/(2L))$. For any starting point $\theta_0 \in \mathbb{R}^d$, with $\rho := (1 - \gamma_\mu)^{1/2}$:

$$
\mathbb{E}\left[\left\|\tilde{\theta}_k^{(\gamma)} - \bar{\theta}_\gamma\right\|^2\right] = \frac{\mathbb{E}_{\theta \sim \pi_\gamma} \left[\left\|\psi_\gamma(\theta)\|^2 - \left\|\psi_\gamma(\theta) - \varphi(\theta)\right\|^2\right]_k}{k} + \frac{\left\|\psi_\gamma(\theta_0)\right\|^2 + \chi_\gamma(\theta_0) - \chi_\gamma^2(\theta_0)}{k^2} + O(\rho^k).
$$

When $f_2$ is a quadratic function, it is possible, for any $\gamma > 0$, to compute $\psi_\gamma$ and $\chi_\gamma^{1/2}$ explicitly; we exactly recover the result of Défossez and Bach [5].

### 2.4. Experiments.

For the sake of illustration, we perform experiments on simulated data, for logistic regression, with $n = 10^7$ observations, for $d = 10$ (Fig. 2). We consider SGD with constant step-sizes $1/R^2$, $1/2R^2$ (and $1/4R^2$) with (plain lines) or without (dashed lines) averaging, with $R^2$ the smoothness constant. Without averaging, the chain saturates with an error proportional to $\gamma$. We consider Richardson-Romberg iterates (red), which saturate at a much lower level, and performs much better than decaying step sizes (as $1/\sqrt{n}$) on the first iterations, as it forgets the initial conditions faster. We also propose an estimator that uses 3 different step sizes to perform a higher order interpolation. More precisely, we compute $\tilde{\theta}_k^{(\gamma)} := \frac{8}{3} \tilde{\theta}_k^{(\gamma)} - 2\bar{\theta}_k^{(2\gamma)} + \frac{1}{3} \bar{\theta}_k^{(4\gamma)}$. With such an estimator, the first 2 terms in the expansion vanish, scaling as $\gamma$ and $\gamma^2$.

---

**References**


