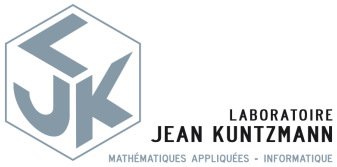

Lecture 5 : Stochastic gradient descent and theoretical properties

So far, we witnessed two first-order algorithms: gradient descent and its Nesterov acceleration. The common points of these algorithms is their access to an oracle of first-order derivatives (i.e., the gradient). In reality, such an oracle can be expensive, and one only has access to a noisy version of the true gradient. Stochastic gradient descent emerge as an alternative for its deterministic counterpart in those cases, and it is the main topic of this lecture.

1 Definition of stochastic gradient descent

Motivation: Adaptation of gradient descent to massive datasets Remind that in machine learning, many optimisation problem admits the following form:

$$\underset{\theta\in\mathbb{R}^d}{\text{Minimize}} \quad \mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell(f(\theta, x_i), y_i)}_{\ell_i(\theta)}, \tag{1}$$

where $\mathcal{D} = (x_i, y_i)_{i=1}^n \in (\mathcal{X}\times\mathcal{Y})^n$ is a dataset, $\ell : \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ is a loss function and $f(\theta,\cdot)$ is a parameterized model (for example, a linear model or a neural network).

If $n \gg 1$ (for example, $n \geq 10^6$), the estimation of gradient of $\mathcal{L}$ is expensive because:

$$\nabla\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla\ell_i(\theta).$$

Different from the update rule of the original gradient descent for minimizning $\mathcal{L}$, which is given by:

$$\theta_{k+1} = \theta_k - \alpha_k\nabla\mathcal{L}(\theta_k), \tag{GD}$$

the stochastic version uses the following update rule:

$$\theta_{k+1} = \theta_k - \alpha_k\nabla\ell_i(\theta_k), \tag{SGD}$$

where $i \in \{1,\dots,n\}$ is a randomly chosen index (usually from $\mathcal{U}(\{1,\dots,n\})$ - the uniform distribution of the set $\{1,\dots,n\}$). With this choise of i_k, even if (SGD) does not take the same direction as (GD) for each realisation of i_k, it does so in average because:

$$\begin{aligned}
\mathbb{E}[\theta_{k+1} \mid \theta_k] &= \theta_k - \alpha_k\mathbb{E}_i[\nabla\ell_i(\theta_k)] \\
&= \theta_k - \alpha_k\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta\ell_i(\theta) \quad (\text{car } i_k \sim \mathcal{U}(\{1,\dots,n\})) \\
&= \theta_k - \alpha_k\nabla\mathcal{L}(\theta_k).
\end{aligned}$$

Moreover, with the same computation cost, the stochastic gradient descent can do $O(n)$ steps, instead of only one (as in the gradient descent case), which makes this method very fast and competitive in practice.

Other forms of stochastic gradient descent More generally, the formula of stochastic gradient descent can be written as:

$$\theta_{k+1} = \theta_k - \alpha_k \underbrace{(\nabla \mathcal{L}(\theta_k) + \epsilon_k)}_{g_k(\theta_k)} \tag{SGD-G}$$

where ϵ_k is a pertubation term inserted to the kth iteration. Below are several popular choices of ϵ_k:

1. In the previous case, we have:

$$\epsilon_k \sim \mathcal{U}(\{\nabla \mathcal{L}(\theta_k) - \nabla \ell_i(\theta_k) \mid i = 1, \dots, n\})$$

2. Gaussian distribution: $\epsilon_k \sim \mathcal{N}(0, \sigma^2)$.

In certain applications (for example, training with differential privacy), ϵ_k is sampled from Gaussian distribution rather than (SGD). As gradient descent, we establish the theoretical guarantees of the stochastic gradient descent in the following:

2 The convergence and the convergence rate of the stochastic gradient descent

In this section, we first introduce the important hypothesis for our theoretical results concerning the stochastic gradient descent defined in (SGD-G). After that, we treat three cases: $\mathcal{L}$ non convex, convex and μ-strongly convex, as we did with gradient descent.

2.1 The hypothesis on the stochastic gradient descent

Assumption 2.1 (Hypothesis on $\mathcal{L}$)**.** *The objective function $\mathcal{L}$ defined in Equation* (1) *is L-smooth.*

Assumption 2.1 can be satisfied when the function $\ell_i(\cdot), i = 1, \dots, n$ are L-smooth, which is usually the case for optimization problems such as linear regression or linear logistics.

Assumption 2.2 (Hypothesis on g_k)**.** *Assume that $g_k(\cdot), k \in \mathbb{N}$ verifies:*

1. $\mathbb{E}[g_k(\theta)] = \nabla \mathcal{L}(\theta)$, $\forall \theta \in \mathbb{R}^d$.

2. $\mathbb{E}[\|g_k(\theta) - \nabla \mathcal{L}(\theta)\|^2] \leq \sigma^2, \forall \theta \in \mathbb{R}^d$.

Several examples that verify Assumption 2.2 will be found in the tutorial session.

Le lemme suivant est la version stochastique du lemme de descente que nous avons vue dans le cours précédent.

Lemma 2.1 (Stochastic version of descent lemma)**.** *Consider $\mathcal{L}$ a function verifying Assumption 2.1 and g_k the stochatic gradient of $\mathcal{L}$ verifying Assumption 2.2, we have:*

$$\mathbb{E}[\mathcal{L}(\theta_{k+1}) \mid \theta_k] \leq \mathcal{L}(\theta_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) \|\nabla \mathcal{L}(\theta_k)\|^2 + \frac{L\alpha_k^2}{2}\sigma^2.$$

Proof. By the property of an L-smooth function, we have:

$$\begin{aligned}\mathcal{L}(\theta_{k+1}) &\leq \mathcal{L}(\theta_k) + \nabla\mathcal{L}(\theta_k)^\top(\theta_{k+1} - \theta_k) + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 \\ &\leq \mathcal{L}(\theta_k) - \alpha_k\langle\nabla\mathcal{L}(\theta_k), g_k(\theta_k)\rangle + \frac{L\alpha_k^2}{2}\|g_k(\theta_k)\|^2.\end{aligned}$$

Taking the expectation of the above inequality, we obtain:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\theta_{k+1}) \mid \theta_k] &\leq \mathcal{L}(\theta_k) - \alpha_k\|\nabla\mathcal{L}(\theta_k)\|^2 + \frac{L\alpha_k^2}{2}\mathbb{E}[\|g_k(\theta_k)\|^2 \mid \theta_k] \\ &\leq \mathcal{L}(\theta_k) - \alpha_k\|\nabla\mathcal{L}(\theta_k)\|^2 + \frac{L\alpha_k^2}{2}\left(\|\nabla\mathcal{L}(\theta_k)\|^2 + \texttt{Var}[g_k(\theta_k) \mid \theta_k]\right) \\ &= \mathcal{L}(\theta_k) - \alpha_k\|\nabla\mathcal{L}(\theta_k)\|^2 + \frac{L\alpha_k^2}{2}\left(\|\nabla\mathcal{L}(\theta_k)\|^2 + \sigma^2\right) \quad \square\end{aligned}$$

2.2 Le cas non convexe

Theorem 2.2 (Stochastic gradient descent for L-smooth function)**.** *Suppose that $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ et $g_k : \mathbb{R}^d \to \mathbb{R}^d$ verifies Assumption 2.1 and Assumption 2.2, respectively. We suppose in addition that $C := \inf_{\theta\in\mathbb{R}^d}\mathcal{L}(\theta) > -\infty$ and $\alpha_k = \alpha < \frac{2}{L}$ is a constant. The iterations generated by* (SGD-G) *verify:*

$$\min_{k=0,\ldots,K-1}\mathbb{E}[\|\nabla\mathcal{L}(\theta_k)\|^2] \leq \frac{\mathcal{L}(\theta_0) - C}{K(\alpha - \frac{L\alpha^2}{2})} + \frac{L\alpha}{1 - \frac{L\alpha}{2}}\sigma^2.$$

Proof. By taking the expectation of the inequality of Lemma 2.1, we obtain:

$$\mathbb{E}[\mathcal{L}(\theta_{k+1})] \leq \mathbb{E}[\mathcal{L}(\theta_k)] - \left(\alpha - \frac{L\alpha^2}{2}\right)\mathbb{E}[\|\nabla\mathcal{L}(\theta_k)\|^2] + \frac{L\alpha^2}{2}\sigma^2$$

By summing this inequality from $k = 0, \ldots, K-1$, we have:

$$\mathbb{E}[\mathcal{L}(\theta_K)] \leq \mathbb{E}[\mathcal{L}(\theta_0)] - \left(\alpha - \frac{L\alpha^2}{2}\right)\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\mathcal{L}(\theta_k)\|^2] + K\frac{L\alpha^2}{2}\sigma^2.$$

This implies:

$$\begin{aligned}\min_{k=0,\ldots,K-1}\mathbb{E}[\|\nabla\mathcal{L}(\theta_k)\|^2] &\leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\mathcal{L}(\theta_k)\|^2] \\ &\leq \frac{1}{K\left(\alpha - \frac{L\alpha^2}{2}\right)}\left(\mathbb{E}[\mathcal{L}(\theta_0)] - \mathbb{E}[\mathcal{L}(\theta_K)] + K\frac{L\alpha^2}{2}\sigma^2\right) \\ &\leq \frac{1}{K\left(\alpha - \frac{L\alpha^2}{2}\right)}\left(\mathbb{E}[\mathcal{L}(\theta_0)] - C + K\frac{L\alpha^2}{2}\sigma^2\right) \quad \square\end{aligned}$$

Remark 2.3. Si nous prenons $\alpha = \frac{1}{L\sqrt{K}}$, nous aurons $\mathbb{E}[\|\nabla\mathcal{L}(\theta_k)\|^2] \leq O\left(\frac{1}{\sqrt{K}}\right)$.

2.3 The strongly convex case

Remind that if $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex, then there eixsts an unique optimal solution $\theta^\star$ and for all θ, we have:

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^\star) \leq \frac{1}{2\mu}\|\nabla\mathcal{L}(\theta)\|^2.$$

Theorem 2.4 (Stochastic gradient descent for μ-strongly convex and L-smooth functions)**.** *Suppose the same hypothesis as in Theorem 2.4. In addition, we suppose that f is μ-strongly convex ($\mu > 0$) and $0 < \alpha \leq \frac{1}{L}$. Then the iterates generated by* (SGD-G) *verify:*

$$\mathbb{E}[\mathcal{L}(\theta_k) - \mathcal{L}^\star] \leq (1 - \mu\alpha)^k \mathbb{E}[\mathcal{L}(\theta_0) - \mathcal{L}^\star] + \frac{L\alpha}{2\mu}\sigma^2, \quad \forall k \in \mathbb{N}.$$

Proof. Denote $\mathcal{L}^\star$ the minimum value of $\mathcal{L}$. By Lemma 2.1, we have :

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\theta_{k+1}) - \mathcal{L}^\star \mid \theta_k] &\leq \mathcal{L}(\theta_k) - \mathcal{L}^\star - \left(\alpha - \frac{L\alpha^2}{2}\right) \|\nabla \mathcal{L}(\theta_k)\|^2 + \frac{L\alpha^2}{2}\sigma^2 \\
&\leq \mathcal{L}(\theta_k) - \mathcal{L}^\star - 2\mu \left(\alpha - \frac{L\alpha^2}{2}\right) (\mathcal{L}(\theta_k) - \mathcal{L}^\star) + \frac{L\alpha^2}{2}\sigma^2 \\
&\leq \left(1 - 2\mu\alpha \left(1 - \frac{L\alpha}{2}\right)\right) (\mathcal{L}(\theta_k) - \mathcal{L}^\star) + \frac{L\alpha^2}{2}\sigma^2 \\
&\leq (1 - \mu\alpha) (\mathcal{L}(\theta_k) - \mathcal{L}^\star) + \frac{L\alpha^2}{2}\sigma^2 \quad (\text{car } \alpha \leq \frac{1}{L})
\end{aligned}$$

By taking the expectation w.r.t. θ_k, we obtain:

$$\mathbb{E}[\mathcal{L}(\theta_{k+1}) - \mathcal{L}^\star] \leq (1 - \mu\alpha)\mathbb{E}[\mathcal{L}(\theta_k) - \mathcal{L}^\star] + \frac{L^2\alpha^2}{2}\sigma^2.$$

By recurrence, we can show that:

$$\mathbb{E}[\mathcal{L}(\theta_k) - \mathcal{L}^\star] \leq (1 - \mu\alpha)^k \mathbb{E}[\mathcal{L}(\theta_0) - \mathcal{L}^\star] + \frac{L\alpha}{2\mu}\sigma^2.$$

□

Remark 2.5. If the number of iterations k is fixed and sufficiently large, we can take:

$$\alpha = \frac{\ln k}{\mu k} \leq \frac{1}{L}.$$

Therefore, the bound Theorem 2.4 becomes:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\theta_k) - \mathcal{L}^\star] &= \left(1 - \frac{\ln k}{k}\right)^k \mathbb{E}[\mathcal{L}(\theta_0) - \mathcal{L}^\star] + \frac{L \ln k}{2k}\sigma^2 \\
&\leq \frac{1}{k}\mathbb{E}[\mathcal{L}(\theta_0) - \mathcal{L}^\star] + \frac{L \ln k}{2k}\sigma^2 \\
&= \tilde{O}\left(\frac{1}{k}\right).
\end{aligned}$$

3 Summary of theoretical guarantees of gradient descent and stochastic gradient descent

Hypothesis	Result	Convergence rate
• $\mathcal{L}$ is L-smooth. • g_k is not biased and of bounded variance. • $\inf_\theta \mathcal{L}(\theta) > 0$.	$\lim_{k\to\infty} \min_{k\in\mathbb{N}} \mathbb{E}[\|\nabla\mathcal{L}(\theta_k)\|] = 0$	$O\left(\sqrt[4]{\frac{1}{k}}\right)$
• $\mathcal{L}$ is L-smooth. • g_k is not biased and of bounded variance. • $\inf_\theta \mathcal{L}(\theta) > 0$. • An optimal solution exists. • $\mathcal{L}$ is convex.	$\lim_{k\to\infty} \mathbb{E}[\mathcal{L}(\theta_k) - \mathcal{L}(\theta^\star)] = 0$	$O\left(\sqrt{\frac{1}{k}}\right)$
• $\mathcal{L}$ is L-smooth. • g_k is not biased and of bounded variance. • $\mathcal{L}$ is μ-strongly convex.	$\lim_{k\to\infty} \mathbb{E}[\mathcal{L}(\theta_k) - \mathcal{L}(\theta^\star)] = 0$	$\tilde{O}\left(\frac{1}{k}\right)$

Table 1: Summary of stochastic gradient descent.

Hypothèses	Descente de gradient	Descente de gradient stochastique
$\mathcal{L}$ is L-smooth	$O\left(\sqrt{\frac{1}{k}}\right)$	$O\left(\sqrt[4]{\frac{1}{k}}\right)$
$\mathcal{L}$ is L-smooth and convex	$O\left(\frac{1}{k}\right)$	$O\left(\sqrt{\frac{1}{k}}\right)$
$\mathcal{L}$ is L-smooth and μ-strongly convex	$O\left(\left(1-\frac{\mu}{L}\right)^k\right)$	$\tilde{O}\left(\frac{1}{k}\right)$

Table 2: Comparison between gradient descent and stochastic gradient descent.