Stochastic Gradient Descent

Optimization for Machine Learning — Exercise #4

Monday 22nd May, 2023

Part I: Theory

A nice reference on Stochastic Gradient Descent is [1]. We will see some highlights from its §4.

I.1. Useful inequalities

Exercise I.1 (Inequality of L-smooth functions)**.** Recall that, given $L > 0$, a function $E\colon \Omega \subset \mathbb{R}^d \to \mathbb{R}$ is called L-smooth if

$$\forall (x, y) \in \Omega^2, \quad \|\nabla E(x) - \nabla E(y)\| \leqslant L\|x - y\|, \tag{1}$$

i.e., if $\nabla E\colon \Omega \to \mathbb{R}^d$ is L-Lipschitz.

Show that if E is L-smooth and Ω is convex (i.e. $\forall (x, y) \in \Omega^2$, $\forall t \in [0, 1]$, $x + t(y - x) \in \Omega$), then

$$\forall (x, y) \in \Omega^2, \quad E(y) \leqslant E(x) + \langle \nabla E(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 \tag{2}$$

Hint: Express $E(y)$ with the integral formula $E(y) = E(x) + \int_0^1 \frac{\partial}{\partial t} E(x + t(y - x))\,\mathrm{d}t$.

Answer I.1. In order to prove the sought-after result, we will integrate the inequality defining the L-smoothness of E between x and y. We would rather integrate a real function (i.e. a function $\mathbb{R} \to \mathbb{R}$); therefore, define $e\colon [0, 1] \to \mathbb{R}$, $t \mapsto e(t) := E(x + t(y - x))$. The function e is well defined since Ω is convex. Notice that $e(0) = E(x)$ and $e(1) = E(y)$.

We can decompose e with $\gamma\colon \mathbb{R} \to \Omega$, $t \mapsto \gamma(t) := x + t(y - x)$, so that $e = E \circ \gamma$.

The so-called "Fundamental Theorem of Calculus" tells us that $e(1) = e(0) + \int_0^1 e'(t)\,\mathrm{d}t$.

We compute $e'(t)$ thanks to the chain rule for Jacobians $\frac{\mathrm{d}(E\circ\gamma)(t)}{\mathrm{d}t} = J_{(E\circ\gamma)}(t) = J_E(\gamma(t))J_\gamma(t)$:

$$e'(t) = \frac{\mathrm{d}(E \circ \gamma)(t)}{\mathrm{d}t} = (\nabla_{\gamma(t)} E(\gamma(t)))^\top \frac{\mathrm{d}\gamma(t)}{\mathrm{d}t} = \langle \nabla_{\gamma(t)} E(\gamma(t)), \gamma'(t)\rangle = \langle \nabla_{\gamma(t)} E(\gamma(t)), y - x\rangle,$$

so that $e(1) = e(0) + \int_0^1 \langle \nabla E(x + t(y-x)), y - x\rangle \,\mathrm{d}t$. Therefore,

$$
\begin{aligned}
E(y) &= E(x) + \int_0^1 \langle \nabla E(x + t(y-x)), y - x\rangle \,\mathrm{d}t \\
&= E(x) + \langle \nabla E(x), y - x\rangle + \int_0^1 \langle \nabla E(x + t(y-x)) - \nabla E(x), y - x\rangle \,\mathrm{d}t \\
\implies E(y) &\leqslant E(x) + \langle \nabla E(x), y - x\rangle + \int_0^1 \|\nabla E(x + t(y-x)) - \nabla E(x)\| \|y - x\| \,\mathrm{d}t \quad (3) \\
&\leqslant E(x) + \langle \nabla E(x), y - x\rangle + \int_0^1 L\|x + t(y-x) - x\| \|y - x\| \,\mathrm{d}t \quad (4) \\
&\leqslant E(x) + \langle \nabla E(x), y - x\rangle + L\|y - x\|^2 \int_0^1 t \,\mathrm{d}t \\
&\leqslant E(x) + \langle \nabla E(x), y - x\rangle + \frac{L}{2}\|y - x\|^2,
\end{aligned}
$$

where (3) comes from the Cauchy-Schwarz inequality ($|\langle u, v\rangle| \leqslant \|u\|\|v\|$), and (4) from the L-smoothness assumption on E.

$\diamondsuit$

I.2. Convergence of SGD

This section follows the beginning of [1, Section 4]. The goal is to prove the convergence rate of stochastic gradient descent under the assumptions of L-smoothness and c-strong convexity for the functional, and with fixed step size. This is [1, Theorem 4.6]. First, we give the setup.

Problem setup Let $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$. We are concerned with learning a supervised task on $\mathcal{X} \times \mathcal{Y}$. The parameter space is $\mathcal{W} \subseteq \mathbb{R}^d$. Let $h\colon \mathcal{X} \times \mathcal{W} \to \mathcal{Y}$ be the *prediction function*, and $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ the *loss function*. Denote by $f\colon \mathcal{W} \times \mathcal{X} \times \mathcal{Y}$ the composition of ℓ and h.

With ξ a random variable selecting samples from $\mathcal{X} \times \mathcal{Y}$, the **expected risk** R can be written as $R(w) = \mathbb{E}_\xi[f(w;\xi)]$. The **empirical risk** R_n can be obtained when ξ takes n realizations $\{\xi(i)\}_{i\in[n]}$ corresponding to n training samples $\{(x_i, y_i)\}_{i\in[n]}$. Denoting $f_i(w) := f(w, \xi(i))$, one has $R_n(w) = \frac{1}{n}\sum_{i=1}^n f_i(w)$.

The objective function $F\colon \mathbb{R}^d \to \mathbb{R}$ is either

$$
F(w) = \begin{cases} R(w) \\ \text{or} \\ R_n(w) \end{cases} \tag{1}
$$

We assume to be able to compute the realization of a random variable ξ_k. Given an iteration w_k and the realization of ξ_k, we assume to be able to compute a stochastic vector $g(w_k, \xi_k) \in \mathbb{R}^d$ (the stochastic gradient).

Because of the stochastic nature of $g(w_k, \xi_k)$, we are not assured to decrease the objective function at every step. But, we can carry an expectation analysis and show that, in expectation (over ξ_k), we do make progress in the minimization problem.

Algorithm 1 Stochastic Gradient Descent algorithm [1, Algorithm 4.1].

Choose an initial iterate w_1
for $k = 1, 2, \ldots$ **do**
    Generate a realization of the random variable ξ_k
    Compute a stochastic vector $g(w_k, \xi_k)$
    Choose a step size $\alpha_k > 0$
    Set the new iterate as $w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$.
end for

We first require the objective F to be L-smooth.

Assumption 1 (F is L-smooth). There exists a constant $L > 0$ such that F is L-smooth.

Exercise I.2 (Descent update with L-smooth function, [1, Lemma 4.2])**.** Under Assumption 1, show that the iterates of SGD (Algorithm 1) satisfy the following inequality for all $k \in \mathbb{N}$:

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leqslant -\alpha_k \langle \nabla F(w_k), \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \rangle + \frac{\alpha_k^2 L}{2} \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2], \qquad (2)$$

where $\mathbb{E}_{\xi_k}[X]$ denotes the expectation of a random variable X with respect to ξ_k given w_k. In this formalism, we assume to have run SGD k times, and we would like to analyse our gain *in expectation* for the next step. Therefore, $F(w_{k+1})$ depends on ξ_k (it is a random variable). Assume that all the $\{\xi_k\}_{k \in \mathbb{N}}$ are jointly independent.

Answer I.2. Let $k \in \mathbb{N}$. We write the L-smoothness inequality (2) for F at iteration k:

$$F(w_{k+1}) - F(w_k) \leqslant \langle \nabla F(w_k), w_{k+1} - w_k \rangle + \frac{L}{2} \|w_{k+1} - w_k\|^2.$$

With the update $w_{k+1} = w_k - \alpha_k g(w_k, \xi_k)$, one has $w_{k+1} - w_k = -\alpha_k g(w_k, \xi_k)$ and

$$F(w_{k+1}) - F(w_k) \leqslant -\alpha_k \langle \nabla F(w_k), g(w_k, \xi_k) \rangle + \frac{\alpha_k^2 L}{2} \|g(w_k, \xi_k)\|^2.$$

To obtain (2), take the expectation with respect to ξ_k, and notice that $F(w_k)$ does not depend on ξ_k: only $F(w_{k+1})$ and $g(w_k, \xi_k)$ do. $\Diamond$

We need some more assumptions on the stochastic estimation $g(w_k, \xi_k)$ in order to control $\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|^2]$. More specifically, we require bounding the first and second moment of $g(w_k, \xi_k)$ like so:

Assumption 2. 1. There exist scalars $\mu_G \geqslant \mu > 0$ such that, for all $k \in \mathbb{N}$,

$$\langle \nabla F(w_k), \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \rangle \geqslant \mu \|\nabla F(w_k)\|_2^2 \qquad (3)$$

$$\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\| \leqslant \mu_G \|\nabla F(w_k)\|_2 \qquad (4)$$

2. The second moment of g is bounded: there exist $M, M_G \geqslant 0$, such that

$$\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \leqslant M + M_G \|\nabla F(w_k)\|_2^2 \tag{5}$$

Exercise I.3 ([1, Lemma 4.4])**.** Show that, under Assumptions 1 and 2, the iterates of SGD in Algorithm 1 satisfy the following inequalities for all $k \in \mathbb{N}$:

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leqslant -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] &\text{(6a)}\\ &\leqslant -(\mu - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M &\text{(6b)}\end{aligned}$$

Answer I.3. We simply use the lower-bound (3) on $\mathbb{E}[\|g(w_k, \xi_k)\|]$ and the upper-bound (5) on $\|\mathbb{E}[g(w_k, \xi_k)]\|^2$ in order to upper-bound (2):

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leqslant -\alpha_k \langle \nabla F(w_k), \mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\rangle + \frac{\alpha_k^2 L}{2}\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2]\\ &\leqslant -\alpha_k \langle \nabla F(w_k), \mu\|\nabla F(w_k)\|\rangle + \frac{\alpha_k^2 L}{2}(M + M_G\|\nabla F(w_k)\|^2)\\ &\leqslant \alpha_k(\frac{\alpha_k L M_G}{2} - \mu)\|\nabla F(w_k)\|^2 + \frac{\alpha_k^2 L M}{2}\end{aligned}$$

$\Diamond$

A last assumption is helpful to make: strong convexity of the objective function F. This will allow to obtain a linear rate of convergence to a neighbourhood of a solution, if the step size is not too big.

Assumption 3 (F is c-strongly convex). There exists a constant $c > 0$ such that F is c-strongly convex.

Exercise I.4 ([1, Theorem 4.6])**.** Under Assumptions 1 to 3, suppose that Algorithm 1 is run with a fixed step size $\alpha_k =: \bar{\alpha}$ for all $k \in \mathbb{N}$, satisfying

$$0 < \bar{\alpha} \leqslant \frac{\mu}{L M_G}. \tag{7}$$

Denote $F_* := \min_w F(w)$ (exists and is unique by Assumption 3). Show that the expected optimality gap satisfies the following inequality for all $k \in \mathbb{N}$:

$$\begin{aligned}\mathbb{E}[F(w_k) - F_*] &\leqslant \frac{\bar{\alpha} L M}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1}\left(F(w_1) - F_* - \frac{\bar{\alpha} L M}{2c\mu}\right) &\text{(8)}\\ &\xrightarrow[k\to\infty]{} \frac{\bar{\alpha} L M}{2c\mu}. &\text{(9)}\end{aligned}$$

Answer I.4. We will combine the previous bound (6) with the inequality coming from the c-strong convexity of F. Indeed, if F is c-strongly convex, then

$$\forall w \in \Omega, \quad 2c(F(w) - F_*) \leqslant \|\nabla F(w)\|^2,$$

where $F_* = \min_{w\in\Omega} F(w)$ (exists and is unique by strong-convexity assumption).

At one time-step $k \in \mathbb{N}^*$, the previous bound (6) gives

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leqslant -\bar{\alpha}(\mu - \frac{\bar{\alpha}LM_G}{2})\|\nabla F(w_k)\|^2 + \frac{\bar{\alpha}^2 LM}{2},$$

which combined with $-\|\nabla F(w_k)\|^2 \leqslant -2c(F(w_k) - F_*)$ further gives

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leqslant -2c\bar{\alpha}(\mu - \frac{\bar{\alpha}LM_G}{2})(F(w_k) - F_*) + \frac{\bar{\alpha}^2 LM}{2}.$$

Now, since $0 < \bar{\alpha} \leqslant \frac{\mu}{LM_G}$, $\mu - \frac{\bar{\alpha}LM_G}{2} \geqslant \frac{\mu}{2}$ and

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leqslant -\bar{\alpha}c\mu(F(w_k) - F_*) + \frac{\bar{\alpha}^2 LM}{2}.$$

Subtracting F_* from both sides and factorizing by $F(w_k) - F_*$ leads to

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F_* \leqslant (1 - \bar{\alpha}c\mu)(F(w_k) - F_*) + \frac{\bar{\alpha}^2 LM}{2}.$$

Taking the total expected value (with respect to $\xi_1 \otimes \cdots \otimes \xi_k$, denoted by $\mathbb{E}[\cdot]$) gives

$$\mathbb{E}[F(w_{k+1}) - F_*] \leqslant (1 - \bar{\alpha}c\mu)\mathbb{E}[F(w_k) - F_*] + \frac{\bar{\alpha}^2 LM}{2}. \tag{10}$$

This is an arithmetico-geometric sequence $u_{k+1} = au_k + b$, with ratio $a = 1 - \bar{\alpha}c\mu$ and with fixed-point $\frac{b}{1-a} = \frac{\bar{\alpha}^2 LM}{2}(\bar{\alpha}c\mu)^{-1} = \frac{\bar{\alpha}LM}{2c\mu}$, so that the geometric sequence $v_k := u_k - \frac{b}{1-a}$ is

$$\mathbb{E}[F(w_{k+1}) - F_*] - \frac{\bar{\alpha}LM}{2c\mu} \leqslant (1 - \bar{\alpha}c\mu)\left(\mathbb{E}[F(w_k) - F_*] - \frac{\bar{\alpha}LM}{2c\mu}\right) \tag{11}$$

Note that $0 \leqslant 1 - \bar{\alpha}c\mu < 1$, since $1 \geqslant \frac{L}{c} \geqslant \bar{\alpha}c\mu > 0$ (we have a *contraction*: at each time-step k, the value of the sequence shrinks).

The desired inequality (8) is obtained by applying (11) inductively for $k = 1, 2, \ldots$. ◊

References

[1] Léon Bottou, Frank E. Curtis and Jorge Nocedal. 'Optimization Methods for Large-Scale Machine Learning'. 2016. DOI: 10.1137/16M1080173. arXiv: 1606.04838.