

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 $E: \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}$ is called L -smooth if

$$\forall (x, y) \in \Omega^2, \quad \|\nabla E(x) - \nabla E(y)\| \leq L\|x - y\|, \quad (1)$$

i.e., if $\nabla E: \Omega \rightarrow \mathbb{R}^d$ 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

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

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

Answer I.1. In order to prove the sought-after inequality, we will integrate the inequality defining the L -smoothness of E between x and y . We'd rather integrate a real function (i.e. a function $\mathbb{R} \rightarrow \mathbb{R}$); therefore, define $e: [0, 1] \rightarrow \mathbb{R}, t \mapsto e(t) := E(x + t(y - x))$. 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 $\gamma: \mathbb{R} \rightarrow \Omega, t \mapsto \gamma(t) := x + t(y - x)$, so that $e = E \circ \gamma$.

The so-called “Fundamental Theorem of Calculus” tells us that $e(1) = e(0) + \int_0^1 e'(t) dt$.

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

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

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

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

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

◇

I.2. Convergence of SGD

Notation We are concerned with learning a supervised task on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. The parameter space is $\mathcal{W} \subseteq \mathbb{R}^d$. Let $h: \mathcal{X} \times \mathcal{W} \rightarrow \mathcal{Y}$ be the *prediction function*, and $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ the *loss function*. Denote by $f: \mathcal{W} \times \mathcal{X} \times \mathcal{Y}$ 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 $F: \mathbb{R}^d \rightarrow \mathbb{R}$ is either

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

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.

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.

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

```

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

```

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

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

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 :

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

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

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

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 $\mu_G \geq \mu > 0$ such that, for all $k \in \mathbb{N}$,

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

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

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

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

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

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \quad (6a)$$

$$\leq -(\mu - \frac{1}{2}\alpha_k LM_G)\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 LM \quad (6b)$$

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):

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

◇

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. 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

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}. \quad (7)$$

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 $k \in \mathbb{N}$:

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\bar{\alpha} LM}{2c\mu} + (1 - \bar{\alpha} c\mu)^{k-1} \left(F(w_1) - F_* - \frac{\bar{\alpha} LM}{2c\mu} \right) \quad (8)$$

$$\xrightarrow{k \rightarrow \infty} \frac{\bar{\alpha} LM}{2c\mu}. \quad (9)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Note that $0 \leq 1 - \bar{\alpha}c\mu < 1$, since $1 \geq \frac{L}{c} \geq \bar{\alpha}c\mu > 0$ (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 $k = 1, 2, \dots$ \diamond

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.