Stochastic Gradient Descent

Optimization for Machine Learning — Exercise #4

Monday 22$^{\text{nd}}$ 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 $E(y) = E(x) + \int_0^1 \frac{\partial}{\partial t} E(x + t(y - x))\, \mathrm{d}t$.

I.2. Convergence of SGD

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).

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

```
Choose an initial iterate w_1
for k = 1, 2, ... do
    Generate a realization of the random variable ξ_k
    Compute a stochastic vector g(w_k, ξ_k)
    Choose a step size α_k > 0
    Set the new iterate as w_{k+1} ← w_k − α_k g(w_k, ξ_k).
end for
```

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.

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 $k \in \mathbb{N}$:

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

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.

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 $k \in \mathbb{N}$,

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

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

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

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

$$\leqslant -(\mu - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M \tag{6b}$$

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

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

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 $k \in \mathbb{N}$:

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

$$\xrightarrow[k\to\infty]{} \frac{\bar{\alpha} L M}{2c\mu}. \tag{9}$$

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`.