Exercise: Minibatch SGD

Let $n, d \geq 1$ be integers, and consider the optimization problem

$$\min_{\theta \in \mathbb{R}^d} F(\theta), \quad \text{where} \quad F(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta), \quad \forall \theta \in \mathbb{R}^d.$$

In this exercise, we will study a generalization of GD and SGD, called *mini-batch SGD*. The idea is simple:

- Gradient Descent proceeds by computing a full gradient $\frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta_k)$ at each step.
- In contrast, SGD proceeds by only computing *one* gradient $\nabla f_{i_k}(\theta_k)$ at each step, for some index i_k picked uniformly at random in $\{1, \ldots, n\}$.
- Mini-batch SGD is a compromise between the two. For some integer $m \in \{1, \ldots, n\}$, it selects uniformly at random (u.a.r.) a subset $v_{k+1} \subseteq \{1, \ldots, n\}$ of cardinality m at each iteration, and computes the partial gradient $\frac{1}{m} \sum_{i \in v_{k+1}} \nabla f_i(\theta_k)$.

The pseudo-code of this algorithm is given below

Algorithm 1: Mini-batch SGD

Start from $\theta_0 \in \mathbb{R}^d$.

Until termination condition, iterate

Draw a subset $v_{k+1} \subseteq \{1, \ldots, n\}$ of cardinality m u.a.r., independent of the past

$$g_{k+1} = \frac{1}{m} \sum_{i \in v_{k+1}} \nabla f_i(\theta_k)$$

$$\theta_{k+1} = \theta_k - \gamma_{k+1} g_{k+1}.$$

Notation. For any subset $v \subseteq \{1, \ldots, n\}$ of cardinality m, we write for ease

$$f_v = \frac{1}{m} \sum_{i \in v} f_i.$$

In particular, if $m = n$ and $v = \{1, \ldots, n\}$, then $f_v = F$. We also define the noise of a subgradient as

$$\sigma^2 = \mathbb{E}_{v \sim \mathcal{D}}[\|\nabla f_v(\theta^*)\|^2]$$

Here, $\mathcal{D}$ denotes the uniform distribution over subsets of $\{1, \ldots, n\}$ whose cardinality is m. The goal of the exercise is to prove a convergence rate for mini-batch SGD, given in the theorem below.

Theorem 1. *Assume that*

1. *F is μ-strongly convex*
2. *$\mathbb{E}_{\mathcal{D}} \|\nabla f_v(\theta) - \nabla f_v(\theta^*)\|^2 \leq 2L(F(\theta) - F(\theta^*))$ for some constant $L > 0$ (we say that F is "L-smooth in expectation with respect to $\mathcal{D}$")*
3. *The noise σ^2 is finite.*

Take a precision $\varepsilon > 0$, and choose a step size $\gamma = \min\left\{\frac{1}{2L}, \frac{\varepsilon\mu}{4\sigma^2}\right\}$. Then we have

$$\mathbb{E}\|\theta_k - \theta^*\|^2 \leq \varepsilon \quad \text{as soon as} \quad k \geq \max\left\{\frac{2L}{\mu}, \frac{4\sigma^2}{\varepsilon\mu^2}\right\} \log\left(\frac{2\|\theta_0 - \theta^*\|^2}{\varepsilon}\right).$$

Questions

1. Using Assumption 2, justify that $\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(\theta)\|^2\right] \leq 4L(F(\theta) - F(\theta^*)) + 2\sigma^2$.
   (One can use the inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$.
2. Justify that
   $$\|\theta_{k+1} - \theta^*\|^2 = \|\theta_k - \theta^*\|^2 - 2\gamma\left\langle \theta_k - \theta^*, \nabla f_{v_{k+1}}(\theta_k)\right\rangle + \gamma^2\left\|\nabla f_{v_{k+1}}(\theta_k)\right\|^2.$$
3. Writing $\mathbb{E}_k$ for the expectation conditional on θ_k, prove that:
   $$\mathbb{E}_k\left\|\theta_{k+1} - \theta^*\right\|^2 \leq \big(1 - \gamma\mu\big)\left\|\theta_k - \theta^*\right\|^2 - 2\gamma\big[F(\theta_k) - F(\theta^*)\big] + \gamma^2\mathbb{E}_k\left\|\nabla f_{v_{k+1}}(\theta_k)\right\|^2.$$
   (One can use the strong convexity of F: $\forall x, y \in \mathbb{R}^d : F(x) - F(y) \geq \langle \nabla F(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2$).
4. Deduce that, if $\gamma \leq \frac{1}{2L}$, we can take the total expectation and write
   $$\mathbb{E}\left\|\theta_{k+1} - \theta^*\right\|^2 \leq \big(1 - \gamma\mu\big)\mathbb{E}\left\|\theta_k - \theta^*\right\|^2 + 2\gamma^2\sigma^2.$$
5. Deduce that $\mathbb{E}\left\|\theta_k - \theta^*\right\|^2 \leq \big(1 - \gamma\mu\big)^k\left\|\theta_0 - \theta^*\right\|^2 + \frac{2\gamma\sigma^2}{\mu}$.
6. How should one choose k to obtain $\mathbb{E}\left\|\theta_k - \theta^*\right\|^2 \leq \varepsilon$?
7. For any $m \in \{1, \ldots, n\}$, we denote by $\mathcal{D}_m$ the uniform distribution over subsets of $\{1, \ldots, n\}$ whose cardinality is m, and we define the quantity $\sigma_m^2 = \mathbb{E}_{v\sim\mathcal{D}_m}\left\|\nabla f_v(\theta^*)\right\|^2$. Justify that
   $$\sigma_m^2 = \frac{1}{m^2}\sum_{i,j=1}^{n}\Big\langle \nabla f_i(\theta^*), \nabla f_j(\theta^*)\Big\rangle\, \mathbb{P}_{v\sim\mathcal{D}_m}\big(i \in v \text{ and } j \in v\big)$$
8. Justify that $\sigma_n^2 = 0$ and that, for any integers $i, j \in \{1, \ldots, n\}$, we have
   $$\mathbb{P}_{v\sim\mathcal{D}_m}\big(i \in v \text{ and } j \in v\big) = \begin{cases} \frac{m}{n} & \text{if } i = j \\ \frac{m(m-1)}{n(n-1)} & \text{if } i \neq j. \end{cases}$$
9. Deduce that $\sigma_m^2 = \dfrac{\sigma_1^2}{n-1}\Big(\dfrac{n}{m} - 1\Big)$.
10. What is the runtime of the mini-batch SGD algorithm with mini-batches of size m to reach precision ε, assuming that computing one gradient takes 1 second?
11. Prove that the optimal batch size m^* allowing one to reach precision ε in the shortest possible runtime is given by
    $$m^* = \frac{2n\sigma_1^2}{2\sigma_1^2 + L(n-1)\varepsilon\mu},$$
    assuming for simplicity that this quantity is an integer.
12. Discuss the behavior of m^* when $\varepsilon \to 0$ or $\varepsilon \to 1$, assuming n to be "very large".

    Remark 1. *Unfortunately, the quantities σ_1^2, L, and μ are unknown to the practitioner, so m^* cannot be evaluated prior to running the algorithm. Still, this result highlights a trade-off between choosing a small batch size, which allows us to reach low precision (high ε) faster, versus choosing a large batch size, which yields a higher precision $\varepsilon \to 0$ faster.*