Exam: Stochastic optimization & generalization in ML
2h

We study the minimization problem

$$\min_{\theta \in \mathbb{R}^d} F(\theta)$$

assuming that F is a differentiable and L-smooth function. To minimize F, we assume that we can run stochastic gradient strategies (SGD), where at each iteration t, we can have access to a random gradient g_t such that

$$\mathbb{E}\left[g_t(\theta_{t-1}) | \mathcal{F}_{t-1}\right] = \nabla F(\theta_{t-1}) \tag{1}$$

where $\mathcal{F}_{t-1}$ is an appropriate filtration such that θ_{t-1} is $\mathcal{F}_{t-1}$-measurable. The iterates of SGD are given by

$$\begin{cases} \theta_0 \in \mathbb{R}^d \\ \theta_{t+1} = \theta_t - \gamma_{t+1} g_{t+1}(\theta_t). \end{cases}$$

1. Describe the two different types of function seen during lectures for which SGD is particularly relevant.
2. For each scenario, precise how the random gradient estimates are constructed, and the associated filtration $\mathcal{F}_t$.
3. Imagine that F is an empirical risk, what type of convergence guarantees does SGD ensure? You will discuss this point depending on the number of iterations done.

We say that F satisfies the Polyak-Łojasiewicz (PL) condition for parameter $\mu > 0$ when for all $\theta \in \mathbb{R}^d$

$$F(\theta) - \inf F \leq \frac{1}{2\mu} \left\| \nabla F(\theta) \right\|_2^2 .$$

4. Show that if F is μ-strongly convex admitting a unique minimizer $\theta^\star$, then F satisfies the Polyak-Łojasiewicz (PL) condition for parameter μ. We say that F is μ-PL.
5. In the context of overparameterized linear regression, consider a dataset $(X_1, Y_1), \ldots,$ (X_n, Y_n), i.i.d. copies of random variables in $\mathbb{R}^d \times \mathbb{R}$, for which the minimization of the least squares criterion reads as

$$\min_{\theta \in \mathbb{R}^d} \underbrace{\frac{1}{2n} \sum_{i=1}^{n} \left(X_i^\top \theta - Y_i \right)^2}_{=: F(\theta)} \quad \text{with also} \quad F(\theta) = \frac{1}{2n} \left\| \mathbb{X}\theta - Y \right\|_2^2$$

when

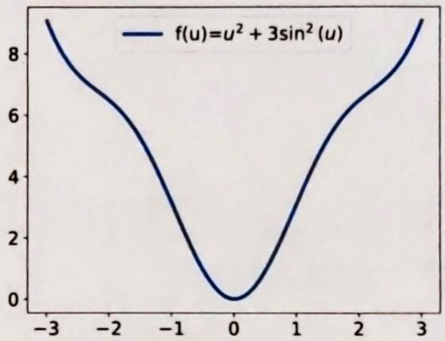
$$\mathbb{X} := \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix} \quad \text{and} \quad Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

We assume that $d > n$, and that $\mathbb{X}\mathbb{X}^\top$ is almost surely invertible.

(a) Show that F is convex but not strongly convex.

(b) Show that F is μ-PL, by precising the value of μ (that should depend on the least singular value $\zeta_n(\mathbb{X})$ of $\mathbb{X}$, and such that $\zeta_n(\mathbb{X}) = \zeta_n(\mathbb{X}^\top)$).

Remark: there are functions satisfying the PL condition while being non-convex, e.g., $f(u) = u^2 + 3\sin^2(u)$ (with $\mu = 1/32$), see figure opposite.



6. Show that when F satisfies the PL property, then

$$\theta^\star \in \operatorname{argmin} F \quad \text{if and only if} \quad \nabla F(\theta^\star) = 0.$$

The goal of the following is to study the convergence of SGD strategies when the function F is L-smooth and μ-PL. To do so, we assume to have access to unbiased gradients as in (1), such that their variance is uniformly bounded by σ^2, i.e., for all $t \geq 1$, for all θ,

$$\mathbb{E}\left[\|g_t(\theta) - \nabla F(\theta)\|_2^2\right] \leq \sigma^2.$$

7. Regarding the SGD iterates, show that for $t \geq 0$,

$$\mathbb{E}\left[F(\theta_{t+1})|\mathcal{F}_t\right] \leq F(\theta_t) - \gamma_{t+1}\left(1 - \frac{L}{2}\gamma_{t+1}\right)\|\nabla F(\theta_t)\|_2^2 + \frac{L}{2}\gamma_{t+1}^2\sigma^2.$$

8. Show that when $\gamma_t \leq 1/L$ for any t, then for $t \geq 0$,

$$\mathbb{E}\left[F(\theta_{t+1})|\mathcal{F}_t\right] \leq F(\theta_t) - \gamma_{t+1}\mu(F(\theta_t) - \inf F) + \frac{L}{2}\gamma_{t+1}^2\sigma^2,$$

and then

$$\mathbb{E}\left[F(\theta_{t+1}) - \inf F\right] \leq (1 - \gamma_{t+1}\mu)\mathbb{E}\left[F(\theta_t) - \inf F\right] + \frac{L}{2}\gamma_{t+1}^2\sigma^2.$$

9. We decide to proceed to T iterations, with a constant step size $\gamma_t = \gamma > 0$ for $t = 1, \ldots, T$.

(a) Provide an upper bound for $\mathbb{E}\left[F(\theta_T) - \inf F\right]$ with respect to $F(\theta_0) - \inf F$.

(b) Discuss the terms in the bound.

(c) When choosing the step size as $\gamma = \ln(T)/(\mu T)$, comment the resulting convergence rate, to be compared to those obtained in class.

Technical reminder:

- for $0 < x < 1$, $\log(1 - x) \leq -x$.