M2A 2023-2024
M2 Statistique & Learning

Exercises 3: Stochastic gradient descent

Exercise 1.

The goal of this exercise is to show the following theorem.

Theorem 1. *Assume that F is μ-strongly convex, and that the stochastic (sub-)gradients g_t used are almost surely bounded , i.e., $\|g_t(\theta)\| \leq b$ for any $\theta \in B$ with B a Euclidean ball of radius r. Furthermore, assume that any $\theta^\star \in \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$ belongs to B. If $(\theta_t)_t$ are the projected SGD iterates with steps $\gamma_t = \frac{2}{\mu(t+1)}$, i.e., for all $t \geq 0$,*

$$\theta_{t+1} = \mathrm{proj}_B\left(\theta_t - \gamma_{t+1} g_{t+1}(\theta_t)\right)$$

we have

$$\mathbb{E}F\Big(\frac{2}{t(t+1)}\sum_{s=1}^{t} s\theta_{s-1}\Big) - F(\theta^\star) \leq \frac{2b^2}{\mu(t+1)}.$$

This theorem establishes a convergence rate of projected SGD for a particular averaging.

We denote by $\mathcal{F}_t$ the minimal $\sigma-$field that makes the first t stochastic gradients measurable. In particular for online optimization with data coming in a streaming fashion, $F(\theta) = \mathbb{E}\left[\ell(Y, f_\theta(X))\right]$ and $\mathcal{F}_t = \sigma((X_1,Y_1),\dots,(X_t,Y_t))$, with $(X_1,Y_1),\dots,(X_t,Y_t)$ the data collected so far. For stochastic algorithms used to minimize an empirical risk function $F = \frac{1}{n}\sum_{i=1}^n f_i$, $\mathcal{F}_t = \sigma(i_1,\dots,i_t)$ with $i_1,\dots,i_t$ the random indices uniformly drawn in $\{1,\dots,n\}$. In both settings, we assume to have access to unbiased gradients, meaning that for all $t \geq 0$,

$$\mathbb{E}\left[g_{t+1}(\theta_t)|\mathcal{F}_t\right] = \nabla F(\theta_t).$$

1. *Show that*

$$\mathbb{E}[\|\theta_t - \theta^\star\|^2|\mathcal{F}_{t-1}] \leq \|\theta_{t-1} - \theta^\star\|^2 + \gamma_t^2 b^2 - 2\gamma_t\langle \nabla F(\theta_{t-1}), \theta_{t-1} - \theta^\star\rangle.$$

2. *With the choice of steps described in the theorem, deduce that*

$$\mathbb{E}F(\theta_{t-1}) - F(\theta^\star) \leq \frac{\mu(t-1)}{4}\mathbb{E}\|\theta_{t-1} - \theta^\star\|^2 - \frac{\mu(t+1)}{4}\mathbb{E}\|\theta_t - \theta^\star\|^2 + \frac{b^2}{\mu(t+1)}.$$

3. *Show that*

$$\sum_{s=1}^{t} s\mathbb{E}(F(\theta_{s-1}) - F(\theta^\star)) \leq \frac{b^2}{\mu}t,$$

   and conclude.

Exercise 2 (Kaczmarz : a random projection algorithm)**.** *We want to solve the linear system $\Phi w = y$, where $\Phi = (\phi(x_1),\dots,\phi(x_n))^\top = (\phi_1,\dots,\phi_n)^\top \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$. In what follows we will note $S \subset \mathbb{R}^d$ the set of solutions of the equation $\Phi w = y$, and we assume that $S \neq \emptyset$.*

1. *For all $1 \leq i \leq n$, we note $H_i \subset \mathbb{R}^d$ the hyperplane defined by*

$$H_i := \{w \in \mathbb{R}^d \mid \langle \phi_i, w\rangle = y_i\}.$$

   Check that S is the intersection of all the H_i's.

We decide to solve our equation $\Phi w = y$ with the following strategy: starting from some $w^0 \in \mathbb{R}^d$, we chose randomly an hyperplane H_i, and *project* w^0 onto H_i, which gives us a new point w^1. We repeat this as many times as needed, projecting each time our current point onto a sampled hyperplane. This method is called *Kaczmarz method*, and is an example of a so-called alternated projection algorithm.

In what follows, we will assume that $w^0 = 0$, and that at every iteration, each hyperplane H_i can be sampled with probability $p_i = \frac{\|\phi_i\|}{\sum_j \|\phi_j\|}$.

2. *Make a drawing illustrating how the algorithm works.*

3. *Write the closed form formula relating w^{t+1} with w^t. You'll need for this the formula for projecting a point onto an hyperplane:*

$$(\forall w \in \mathbb{R}^d) \quad \mathrm{proj}_{H_i}(w) = w - \frac{\langle \phi_i, w\rangle - y_i}{\|\phi_i\|^2}\phi_i.$$

4. *Show that, for all $w^* \in S$ and $t \in \mathbb{N}$, we have*

$$w^{t+1} - w^* = (I - C_{i_t})(w^t - w^*), \qquad \textit{avec} \quad C_{i_t} = \frac{\phi_{i_t}\phi_{i_t}^\top}{\|\phi_{i_t}\phi_{i_t}^\top\|}.$$

5. *Deduce that:*

$$\|w^{t+1} - w^*\|^2 \leq \langle (I - C_{i_t})(w^t - w^*), w^t - w^*\rangle.$$

6. *Compute $C := \mathbb{E}\left[C_i\right]$. You must pay attention to what means $\mathbb{E}$ here, in particular with respect to which law we sample $i \in \{1, \dots, n\}$.*

7. *Verify that C is a semidefinite positive symmetric matrix, such that $\mathrm{Im}\, C = \mathrm{span}\,(\phi_i)_i$, and $\|C\| \leq 1$.*

8. *Check that for all $t \in \mathbb{N}$, $w^t \in \mathrm{Ker}\, C^\perp$.*

9. *Show that, for all $t \in \mathbb{N}$, and for $w^* \in S \cap \mathrm{Ker}\, C^\perp$ (we assume it exists) we have*

$$\mathbb{E}\left[\|w^{t+1} - w^*\|^2\right] \leq \theta\, \mathbb{E}\left[\|w^t - w^*\|^2\right],$$

where $\theta = 1 - \sigma_{min}(C)$.

10. *Explain why $\theta \in [0, 1[$. What can be said of the convergence rate for this algorithm?*

Exercise 3 (Kaczmarz is a particular case of SGD)**.** *We keep here the same context as in Exercise 2. Let $F(w) = \frac{1}{2n}\|\Phi w - y\|^2$.*

1. *Show that we can write F under the form $\frac{1}{n}\sum\limits_{i=1}^{n} f_i$, such that*

   - *if we apply SGD to this sum*
   - *if we use importance sampling to sample the f_i's*
   - *if we use an appropriate stepsize*

   then we obtain exactly the Kaczmarz method.

2. *This link between Kaczmarz and SGD being made, what does the class say about the asymptotic behavior of $\mathbb{E}\left[\|w^t - w^*\|^2\right]$? Compare and discuss with your answer at the end of Exercise 2.*

3. *What is the value of the stepsize λ_k for the SGD algorithm here?*

4. *Compute $\mathcal{L}$, the expected smoothness constant of f with respect to the importance sampling. Express λ_k in terms of $\mathcal{L}$.*