Tutorial 5 : Stochastic gradient descent and theoretical properties

Exercise 1 (Optimization problems satisfy the hypothesis)**.** Prove that the stochastic gradient descent of the following cases satisfy our hypothesis on g_k, i.e., $\mathbb{E}[g_k] = \nabla\mathcal{L}$ and $\texttt{Var}[g_k] \le \sigma^2$ for some fixed constant $\sigma > 0$.

1. Gradient pertubed by Gaussian noise, i.e., $g_k(\theta) = \nabla\mathcal{L}(\theta) + \epsilon_k$ and $\epsilon_k \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

2. We consider the optimization problem of a linear regression problem of the form:

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\log(1 + \exp(-y_i x_i^\top \theta))}_{\ell_i(\theta)}. \tag{1}$$

   and $g_i(x) = \nabla\ell_i(x), i \sim \mathcal{U}(\{1,\dots,n\})$ (i is sampled uniformly from the set $\{1,\dots,n\}$).

Solution of Exercise 1. We consider two cases one-by-one:

1. We have trivially that: $\mathbb{E}[g_k] = \nabla\mathcal{L}(\theta_k)$ and $\mathrm{Var}[g_k] = d\sigma^2$.

2. The expectation condition is clearly satisfied. We just need to control the variance condition. Indeed, since:

$$\nabla\ell_i(\theta) = -\frac{y_i x_i}{1 + \exp(y_i x_i^\top \theta)},$$

   which is bounded by $\max_i |y_i|\|x_i\|$. Therefore, the variance is always bounded.

□

Exercise 2 (Reduce the variant in stochastic gradient descent)**.** We saw that the variance of $\sigma^2 = \texttt{Var}[\|g_k\|^2]$ plays an important role in the analysis of stochastic gradient descent. The smaller is σ, the better the bound is. In reality, practitioners employ many techniques to reduce σ. Minibatching is one of them. Indeed, consider an optimization problem of the form:

$$\underset{\theta\in\mathbb{R}^d}{\text{Minimize}} \quad \mathcal{L}(\theta) = \sum_{i=1}^{n} \ell_i(\theta).$$

Instead of taking $g_k = \nabla\ell_i(\theta_k)$, minibatching takes:

$$\tilde{g}_k^2 = \frac{1}{|S|}\sum_{i\in S} \nabla\ell_i(\theta_k).$$

where $S \subseteq \{1,\dots,n\}$ is uniformly sampled from $\{1,\dots,n\}$ with replacement. Prove that:

$$\texttt{Var}[\|\tilde{g}_k\|^2] = \frac{1}{|S|}\texttt{Var}[\|g_k\|^2].$$

Solution of Exercise 2. Since elements of S are uniformly sampled from $\{1,\dots,n\}$ with replacement, we have:

$$\begin{aligned}\mathrm{Var}[\|\tilde{g}_k\|^2] &= \frac{1}{|S|^2}\mathrm{Var}[\|\sum_{i\in S}\nabla\ell_i(\theta_k)\|^2] \\ &= \frac{1}{|S|^2}\sum_{i\in S}\mathrm{Var}[\|\nabla\ell_i(\theta_k)\|] \\ &= \frac{1}{|S|}\mathrm{Var}[\|g_k\|^2]\end{aligned}$$

□

Exercise 3 (Stochastic gradient descent with convex and L-smooth functions)**.** Consider the stochastic gradient descent when optimizing a convex and L-smooth function f. Assume that the stochastic gradient g_k satisfy the hypothesis as in the lecture. We want to prove that if the step size $\alpha_k = \alpha < \frac{1}{2L}$, then:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{\|x_0 - x^\star\|^2}{T\alpha} + \sigma^2\alpha.$$

where $\bar{x}^T = \frac{1}{T}\sum_{i=0}^{T-1} x_i$. Follow these steps:

1. Prove that:

$$\mathbb{E}[\|x_{k+1} - x^\star\|^2|x_k] = \|x_k - x^\star\|^2 - 2\alpha\nabla f(x_k)^\top(x_k - x^\star) + \alpha^2(\|\nabla f(x_k)\|^2 + \sigma^2)$$

2. Prove that:

$$\mathbb{E}[\|x_{k+1} - x^\star\|^2|x_k] \leq \|x_k - x^\star\|^2 + 2\alpha(1 - L\alpha)(f(x_k) - f(x^\star)) + \alpha^2\sigma^2$$

3. Conclude.

4. Optimizing α to get the best bound for the timestep K.

Solution of Exercise 3. Answers for each question are given as follows:

1. We have:

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^\star\|^2 \mid x_k] &= \mathbb{E}[\||x_k - x^\star - \alpha g_k\|^2] \\ &= \|x_k - x^\star\|^2 - 2\mathbb{E}[\alpha\langle x_k - x^\star, g_k\rangle] + \alpha^2\mathbb{E}[\|x_k\|^2] \\ &= \|x_k - x^\star\|^2 - 2\alpha\langle x_k - x^\star, \nabla f(x_k)\rangle + \alpha^2(\|\nabla f(x_k)\|^2 + \sigma^2).\end{aligned}$$

2. By convexity, we have:

$$f(y) \leq f(x) + \nabla f(x)^\top(y - x) \implies \nabla f(x_k)^\top(x^\star - x) \leq f(x^\star) - f(x_k).$$

   By L-smoothness, we have:

$$f(x_k) - f(x^\star) \geq f(x_k) - f(x_k - \frac{1}{L}\nabla f(x_k)) \geq \frac{1}{2L}\|\nabla f(x_k)\|^2.$$

   Plugging these two inequalities into question 1 gives the result.

3. From the second questions, we can conclude that:

$$\mathbb{E}[\|x_{k+1} - x^\star\|^2] \leq \mathbb{E}[\|x_k - x^\star\|^2] + \underbrace{2\alpha(1 - L\alpha)}_{\leq \alpha} \mathbb{E}[f(x_k) - f(x^\star)] + 2\alpha^2\sigma^2.$$

   Telescoping, we get:

$$\mathbb{E}[\sum_{i=1}^{T}(f(x_i) - f^\star)] \leq \frac{1}{\alpha}\|x_0 - x^\star\|^2 + \alpha\sigma^2.$$

   We finish the proof by remarking that:

$$\sum_{i=1}^{T}(f(x_i) - f^\star) \leq T(f(\bar{x}_T) - f^\star).$$

4. Choose $\alpha = \frac{1}{\sqrt{T}}$.

□

Exercise 4 (Choice of step-size)**.** Consider the same setting as the previous exercise. However, this time, we will choose α_k differently and $\alpha_k < \frac{1}{2L}, \forall k \in \mathbb{N}$. In that case, the previous result becomes:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{\|x_0 - x^\star\|^2}{\sum_i \alpha_i} + \sigma^2 \frac{\sum_i \alpha_k^2}{\sum_i \alpha_i}.$$

where $\bar{x}^T = \frac{\sum_{i=0}^{T-1} x_i \alpha_i}{\sum_i \alpha_i}$ (weighted average).

1. Using the same schema, reprove the above theoretical guarantee.
2. The Robbins-Monro step-sizes are sequences of $\{\alpha_k\}_{k\in\mathbb{N}}$ satisfies:
   (a) $\sum_k \alpha_k^2 < \infty$.
   (b) $\sum_k \alpha_k$ is not bounded.

   Prove that this choice allows $\mathbb{E}[f(\bar{x}^T) - \inf f] \to 0$.
3. For which value of β that $\alpha_k = k^{-\beta}$ satisfies Robbins Monro criteria.
4. Which values of β yields the best asymptotic convergence rate.

Solution of Exercise 4. Answers for each question are given below:

1. Using the same argument, we obtain:

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^\star\|^2] &\leq \mathbb{E}[\|x_k - x^\star\|^2] + \underbrace{2\alpha_k(1 - L\alpha_k)}_{\leq \alpha} \mathbb{E}[f(x_k) - f(x^\star)] + 2\alpha_k^2\sigma^2 \\ &= \mathbb{E}[\|x_k - x^\star\|^2] + \alpha_k\mathbb{E}[f(x_k) - f(x^\star)] + 2\alpha_k^2\sigma^2\end{aligned}$$

   Telescoping and re-arranging the term yield the desired inequality.

2. With the Robbins-Monro step-sizes, we have:

$$\lim_{k\to\infty} \mathbb{E}[f(\bar{x}^k) - f^\star] \leq \lim_{k\to\infty} \frac{\|x_0 - x^\star\|^2}{\sum_i \alpha_i} + \lim_{k\to\infty} \sigma^2 \frac{\sum_i \alpha_i^2}{\sum_i \alpha_i} = 0.$$

3. We need $\beta > 1/2$ so that $\sum_k \alpha_k^2 < \infty$ and $\beta \leq 1$ so that $\sum_k \alpha_k = \infty$.

4. Note that for $\alpha = k^{-\beta}, \beta \in (0,1)$, we have:

$$\sum_{k=1}^{T} \alpha_k = O(T^{1-\beta}).$$

Therefore, the best value is $\beta \to 1/2$. Note that at $\beta = 1/2$, the rate becomes:

$$\frac{\ln T}{T^{1/2}}$$

$\square$

Exercise 5 (Subgaussian distributions)**.** In mathematical analysis of stochastic optimization algorithms, many works assume that the noise ϵ_k belongs to certain families of distribution. Subgaussian, a generalization of Gaussian distribution is among the most popular. In this exercise, we take a look at this family. Prove the following equivalent definition of a subgaussian distribution: Let X be a random variable:

1. There exists $K_1 > 0$ such that:

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{t^2}{K_1^2}\right), \forall t \geq 0.$$

2. There exists $K_2 > 0$ such that:

$$\mathbb{E}[|X|^p]^{\frac{1}{p}} \leq K_2\sqrt{p}, \forall p > 1.$$

3. There exists $K_3 > 0$ such that:

$$\mathbb{E}\exp\left(\frac{X^2}{K_3}\right) \leq 2.$$

Solution of Exercise 5. We prove that 1) implies 2) implies 3) implies 1).

- 1) $\implies$ 2): WLOG, we consider $K_2 = 1$ and we have:

$$\begin{aligned}
\mathbb{E}[|X|^p] &= \int_0^\infty P(|X|^p > t)dt \\
&= \int_0^\infty P(|X| > t^{1/p})dt \\
&= \int_0^\infty P(|X| > t)pt^{p-1}dt \\
&\leq \int_0^\infty 2pt^{p-1}\exp(-t^2)dt \\
&= \int_0^\infty pt^{p/2-1}\exp(-t)dt = p\Gamma(p/2) \leq 3p(p/2)^{p/2}.
\end{aligned}$$

- 2) $\implies$ 3): We have:

$$\begin{aligned}
\mathbb{E}[\exp(X^2/K_3)] &= \mathbb{E}\left[1 + \sum_{k\geq 1}\frac{X^{2k}}{k!K_3^k}\right] \\
&\leq \mathbb{E}\left[1 + \sum_{k\geq 1}\frac{K_2^{2k}(2k)^k}{k!K_3^k}\right] \\
&\leq \mathbb{E}\left[1 + \sum_{k\geq 1}\left(\frac{2kK_2^2}{K_3k/e}\right)^k\right]
\end{aligned}$$

Picking K_3 large enough will do the job.

- 3) $\implies$ 1): This is the Markov inequality.

$\square$