Homework 10: Stochastic gradient descent

Hand in: 06.01.2026 (Tuesday)
Please follow the submission instructions from the webpage of the course.

Correction: tutorial session on 08.01.2026 (Thursday)

These exercises involve some knowledge in probability theory. Please have a look at the formulas given at the end of this document to help you solving the exercises.

Exercise 1: The randomized Kaczmarz method (12 points + 2 bonus points)

In this exercise, we consider the following linear regression problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - a_i^\top x\right)^2 . \tag{1}$$

We assume that the model perfectly fits the data, i.e., there exists $x^\star$ such that $y_i = a_i^\top x^\star$ for all $i = 1, \dots, N$.

Furthermore, we assume that the vectors a_i are such that, for some constants L and $\mu > 0$:

$$\frac{1}{N}\sum_{i=1}^{N} a_i a_i^\top \succcurlyeq \mu I_n, \quad \text{and} \quad \text{for } i = 1, \dots, N, \quad \|a_i\|^2 \le L. \tag{2}$$

We perform the stochastic incremental gradient method on the problem (1) with a fixed step-size $\alpha = 1/L$.

1. Show that the objective function in (1) is L-smooth and μ-strongly convex.
2. Express the update rule of the stochastic incremental gradient method for solving (1), i.e. express x_{k+1} as a function of x_k, and the randomly selected index i_k at iteration k.
3. Show that the following equation holds:
$$x_{k+1} - x^\star = M_{i_k}(x_k - x^\star),$$
where M_i is a matrix that you need to determine, and i_k is the randomly selected index at iteration k.
4. Show the following equality for expected value of the squared norm of the error:
$$\mathbb{E}\left[\|x_{k+1} - x^\star\|^2\right] = \mathbb{E}\left[(x_k - x^\star)^\top P (x_k - x^\star)\right],$$
for a positive semi-definite matrix $P \succcurlyeq 0$ that you need to determine.
5. Show that $P \preccurlyeq (1 - \frac{\mu}{L}) I_n$.
6. Conclude that for this problem, the stochastic incremental gradient method converges linearly in expectation. More precisely, show that the following inequality holds:
$$\mathbb{E}\left[\|x_k - x^\star\|^2\right] \le \left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^\star\|^2 .$$

Bonus question (2 points) :

Which rate would you obtain if you had used the full-batch gradient descent method instead, for the same step-size $\alpha = 1/L$?

Compare the number of iterations needed for the two methods to reach a given accuracy $\varepsilon > 0$ in expected value, i.e., such that $\mathbb{E}\left[\|x_k - x^\star\|^2\right] \le \varepsilon \|x_0 - x^\star\|^2$. Also compare the number of vector-vector products needed for both methods to reach the same accuracy.

Exercise 2: Stochastic linear regression (10 points)

In this exercise, we consider the following stochastic linear regression problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\mathbb{E}\left[(y - a^\top x)^2\right], \tag{3}$$

where the random variable $(a, y) \in \mathbb{R}^n \times \mathbb{R}$ are generated according to the following model:

$$y = a^\top x^\star + e, \tag{4}$$

$$a \sim \mathcal{N}(0, \sigma_a^2 I_n), \quad e \sim \mathcal{N}(0, \sigma_e^2), \tag{5}$$

where $x^\star \in \mathbb{R}^n$ is the unknown parameter vector we want to estimate, and $\sigma_a, \sigma_e > 0$ are known constants. Also note that the random variables a and e are independent.

We perform the stochastic gradient descent method on the problem (3), where at each iteration k, we sample a new independent realization of the random variable (a_k, y_k) that follows the same distribution as (a, y). Like before, we choose a constant step-size $\alpha > 0$.

1. What is the solution of the optimization problem (3)?
2. Express x_{k+1} as a function of x_k, a_k, $x^\star$ and e_k.
3. Express the update rule for the expected squared error, i.e. express $\mathbb{E}\left[\|x_{k+1} - x^\star\|^2\right]$ as a function of $\mathbb{E}\left[\|x_k - x^\star\|^2\right]$.
4. Assume that the step-size is chosen such that $\alpha < \frac{2}{(n+2)\sigma_a^2}$. What is the limit of $\mathbb{E}\left[\|x_k - x^\star\|^2\right]$ as k goes to infinity?

   <u>Hint:</u> Find a fixed point of the update rule derived in the previous question, and subtract it from the equation.
5. How should one choose the step-size α to achieve $\lim_{k\to+\infty} \mathbb{E}\left[\|x_k - x^\star\|^2\right] \le \varepsilon$?

Exercise 3: Gradient descent on a random direction (10 points)

In this exercise, we aim to minimize an L-smooth function $f(\cdot)$ by computing, at each iteration, only the directional derivative along a random direction.

More precisely, at each iteration, we pick a random direction $p_k \sim \mathcal{N}(0, I_n)$, and compute the directional derivative of f at the point x_k along the direction p_k, i.e.:

$$\beta_k := \lim_{\varepsilon\to 0} \frac{f(x_k + \varepsilon p_k) - f(x_k)}{\varepsilon}. \tag{6}$$

Then, we update the variable x_k as follows:

$$x_{k+1} = x_k - \frac{\beta_k}{L\,\|p_k\|^2} p_k. \tag{7}$$

The goal of this exercise is to analyze the convergence of this method.

1. Express β_k as a function of $\nabla f(x_k)$ and p_k.
2. Show the following inequality:

$$f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\left(\frac{\beta_k}{\|p_k\|}\right)^2. \tag{8}$$

3. Show the following inequality for the expected value of $f(x_{k+1})$:

$$\mathbb{E}\left[f(x_{k+1})\right] \leq \mathbb{E}\left[f(x_k)\right] - \frac{1}{2Ln}\mathbb{E}\left[\|\nabla f(x_k)\|^2\right]. \tag{9}$$

4. Assume that f is μ-strongly convex and denote by $x^\star$ its unique minimizer. Show the following inequality:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^\star)\right] \leq \left(1 - \frac{\mu}{Ln}\right)\mathbb{E}\left[f(x_k) - f(x^\star)\right]. \tag{10}$$

Hint: Prove that for a μ-strongly convex function, the following inequality holds for all x:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f(x^\star)).$$

5. Conclude the following inequality:

$$\mathbb{E}\left[\|x_k - x^\star\|^2\right] \leq \frac{L}{\mu}\left(1 - \frac{\mu}{Ln}\right)^k \|x_0 - x^\star\|^2. \tag{11}$$

Programming tasks (4 bonus points)

Open the jupyter notebook `programming_exercise5.ipynb`, and fill in the missing parts of the code.

If you are struggling with downloading Jupyter notebook, you can also use it online via

`https://jupyter.org/try-jupyter/lab`.

A couple of probability formulas you might find useful

Let $r \in \mathbb{R}^n$ be a random variable, following a Gaussian distribution with zero mean and covariance matrix $\sigma^2 I_n$, i.e., $r \sim \mathcal{N}(0, \sigma^2 I_n)$, then:

$$\mathbb{E}\left[r\right] = 0, \quad \mathbb{E}\left[rr^\top\right] = \sigma^2 I_n, \qquad \mathbb{E}\left[\|r\|^2\right] = n\sigma^2,$$

$$\mathbb{E}\left[\|r\|^2 rr^\top\right] = \sigma^4(n+2)I_n, \qquad \mathbb{E}\left[\frac{rr^\top}{\|r\|^2}\right] = \frac{1}{n}I_n.$$

Furthermore, if r and s are independent random variables, then for any functions $\phi : \mathbb{R}^n \to \mathbb{R}$ and $\psi : \mathbb{R}^m \to \mathbb{R}$:

$$\mathbb{E}\left[\phi(r)\psi(s)\right] = \mathbb{E}\left[\phi(r)\right]\mathbb{E}\left[\psi(s)\right].$$