Exercise sheet 5: Exam 2024-2025 (adapted)

Optimization for data science, M2 MIAGE ID/ID Apprentissage

January 21, 2026

Exercise 1: Gradient descent

In this exercise, we consider a logistic regression problem of the form

$$\underset{\boldsymbol{w},\boldsymbol{v}\in\mathbb{R}^d}{\text{minimize}}\, f^{log}(\boldsymbol{w}) := \log(1+\exp(-y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w})), \tag{1}$$

where our dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is such that $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ for any $i = 1, \ldots, n$. The function f^{log} is $\mathcal{C}^1$.

a) Write down the iteration of gradient descent applied to problem (1) using a constant stepsize.

b) The function f^{log} is $\mathcal{C}^{1,1}_L$ for some $L > 0$. If the stepsize from question a) is chosen as $\frac{1}{L}$, what guarantee do we have on this iteration?

c) Give another way of choosing the stepsize than using a constant stepsize.

d) Knowing that f^{log} is convex, what convergence rate can we expect for gradient descent on problem (1)? What quantity does this rate apply to?

e) What would the answer to question d) be if the function was strongly convex instead of just convex?

f) The accelerated gradient method (also known as Nesterov's method) has a better convergence rate than gradient descent on convex problems. What is this rate?

g) Explain the main algorithmic idea behind accelerated gradient.

h) In terms of algorithm, what is the difference between accelerated gradient for convex functions and for strongly convex functions?

Exercise 2: Least squares VS smoothed biweight loss

In this exercise, we consider data under the form of a matrix $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix}$ and a vector $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$. Our goal is to find a linear model that best fits the data, i.e. a vector $\boldsymbol{w} \in \mathbb{R}^d$ such that $\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w} \approx y_i$ for $i = 1, \dots, n$.

a) We first consider the classical linear regression task represented by the following optimization problem:

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimize}}\, f^{lin}(\boldsymbol{w}) := \frac{1}{2n} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 = \frac{1}{2n} \sum_{i=1}^{n} (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w} - y_i)^2. \tag{2}$$

i) Justify that the optimal value of problem (2) is nonnegative. Can it be equal to 0?

ii) The objective function f^{lin} of problem (2) is convex. What does convexity imply on the local minima of this problem?

iii) Under additional conditions on the data, we can guarantee that f^{lin} is μ-strongly convex for some $\mu > 0$. What can we say about the set of solutions of problem (2) in that case?

b) The linear regression problem (2) is known to be sensitive to outliers in the data. An alternate formulation, that is more robust to these outliers, relies on the smoothed biweight loss, and gives rise to the problem

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimize}}\, f^{sb}(\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w} - y_i), \quad \text{where} \quad \phi(t) = \frac{t^2}{1+t^2}. \tag{3}$$

The objective function f^{sb} is $\mathcal{C}^2$ and nonconvex.

i) Consider a vector $\bar{\boldsymbol{w}} \in \mathbb{R}^d$ such that $\nabla f^{sb}(\bar{\boldsymbol{w}}) = \mathbf{0}_{\mathbb{R}^d}$. Explain why this vector is not necessarily a local minimum of problem (3).

ii) Suppose that the point $\bar{\boldsymbol{w}}$ from the previous question is not a local minimum of problem (3), and suppose that we run gradient descent starting from $\boldsymbol{w}_0 = \bar{\boldsymbol{w}}$. Justify that the method converges to $\bar{\boldsymbol{w}}$.

iii) To avoid converging towards points like $\bar{\boldsymbol{w}}$, how should $\boldsymbol{w}_0$ be picked in gradient descent?

Exercise 3: Stochastic gradient

In this exercise, we consider an optimization problem of the form

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, f^{sto}(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} f_i^{sto}(\boldsymbol{w}), \tag{4}$$

where f_i^{sto} is $\mathcal{C}^1$ for every $i = 1, \dots, n$. We further assume that every f_i^{sto} depends on the ith example in a dataset of n elements, with $n \gg 100$.

a) Justify that the set of solutions of (4) is identical to that of

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, \frac{1}{2\,n}\sum_{i=1}^{n} f_i^{sto}(\boldsymbol{w}).$$

b) Write down the iteration of stochastic gradient applied to problem (4) using a constant stepsize.

c) Recall the definition of an epoch. How many iterations of stochastic gradient can be performed within a budget of one epoch?

d) We suppose that f^{sto} is nonconvex. Under appropriate assumptions, one can show that after $K \geq 1$ iterations, the convergence rate (in expectation) of stochastic gradient is $\mathcal{O}\left(\frac{1}{K^{1/4}}\right)$.

   i) What is the convergence rate of gradient descent on a nonconvex problem? Is it better or worse than that of stochastic gradient?

   ii) Suppose that we use a budget of $E \geq 1$ epochs. What convergence rate do we obtain for gradient descent?

   iii) Given the same budget of E epochs, what is the convergence rate of stochastic gradient? Is it better or worse than that of gradient descent?

e) Write down the iteration of batch stochastic gradient applied to problem (4) using a constant stepsize.

f) Explain how batch stochastic gradient generalizes both gradient descent and stochastic gradient.

g) Suppose that we run batch stochastic gradient methods with the following batch sizes: $\left\{1, \frac{n}{128}, \frac{n}{2}, n\right\}$.

   i) When we run each method 10 times, we observe a higher variability in the results for batch size 1 than for the others. What property of batch methods does this illustrate?

   ii) We also observe that the method with batch size $\frac{n}{2}$ converges more slowly than the variants with smaller batch size, yet eventually reaches a smaller function value. Explain this behavior.

Exercise 4: Matrix linear regression

In this exercise, we consider data under the form of two matrices $\boldsymbol{X} \in \mathbb{R}^{p \times d_1}$ and $\boldsymbol{Y} \in \mathbb{R}^{p \times d_2}$. In a regression context, one considers the following extension of linear regression

$$\underset{\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \; f^{ml} := \frac{1}{2p} \|\boldsymbol{X}\,\boldsymbol{W} - \boldsymbol{Y}\|_F^2 = \frac{1}{2p} \sum_{\ell=1}^{p} \|\boldsymbol{X}\boldsymbol{w}_\ell - \boldsymbol{y}_\ell\|_2^2 \,, \tag{5}$$

where $\boldsymbol{w}_\ell \in \mathbb{R}^{d_1}$ and $\boldsymbol{y}_\ell \in \mathbb{R}^p$ are the ℓth columns of $\boldsymbol{W}$ and $\boldsymbol{Y}$, respectively.

a) Given $\ell \in \{1, \dots, p\}$, we consider the problem

$$\underset{\boldsymbol{w}_\ell \in \mathbb{R}^{d_1}}{\text{minimize}} \; f_\ell^{ml}(\boldsymbol{w}_\ell) := \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w}_\ell - \boldsymbol{y}_\ell\|_2^2 \,. \tag{6}$$

  i) Justify that problem (5) has a convex objective.

  ii) Write down the iteration of gradient descent for problem (7) using a constant stepsize.

  iii) Can stochastic gradient be applied to (7) ? Justify your answer.

b) Suppose that $\boldsymbol{w}_1^*, \dots, \boldsymbol{w}_p^*$ are minima of $f_1^{ml}, \dots, f_p^{ml}$, respectively. Justify that the matrix $\boldsymbol{W}^* = [\boldsymbol{w}_1^* \;\cdots\; \boldsymbol{w}_p^*]$ is a solution of problem (5).

c) Based on the previous questions, propose an algorithm to solve problem (5).

Solutions

Solutions for Exercise 1

a) The kth iteration of gradient descent for this problem is given by $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha \nabla f^{log}(\boldsymbol{w}_k)$ where $\alpha > 0$ is a constant stepsize.

b) With this stepsize, we are guaranteed that $f^{log}(\boldsymbol{w}_{k+1}) \leq f^{log}(\boldsymbol{w}_k)$ at every iteration (with a strict decrease as long as $\nabla f^{log}(\boldsymbol{w}_k) \neq \mathbf{0}$).

c) One could choose the stepsize *a priori* as a decreasing sequence $\{\alpha_k\}$ converging to 0. Another option would be to choose α_k in an adaptive fashion, that depends on the current iterate $\boldsymbol{w}_k$ and/or f^{log}.

Note: Only one answer was required.

d) After $K \geq 1$ iterations, the iterate $\boldsymbol{w}_K$ of gradient descent satisfies

$$f^{log}(\boldsymbol{w}_K) - \min_{\boldsymbol{w} \in \mathbb{R}^d} f^{log}(\boldsymbol{w}) \;\leq\; \mathcal{O}\left(\frac{1}{K}\right).$$

e) The rate (that applies to the same quantity than in the previous question) would be $\mathcal{O}(t^K)$ with $t \in (0, 1)$.

f) The rate of accelerated gradient on convex problems is $\mathcal{O}\left(\frac{1}{K^2}\right)$.

g) Accelerated gradient consists in combining a gradient step with the previous step taken by the algorithm, the momentum step.

h) The momentum parameter appearing in accelerated gradient must be set differently for convex problems and for strongly convex problems.

Solutions for Exercise 2

a) *Classical linear regression*

i) For every $\boldsymbol{w} \in \mathbb{R}^d$, $f^{lin}(\boldsymbol{w}) \geq 0$ because it is a sum of squares. As a result, $\min_{\boldsymbol{w} \in \mathbb{R}^d} f^{lin}(\boldsymbol{w}) \geq 0$, with equality if and only if it exists $\boldsymbol{w}^*$ such that $\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y} = \mathbf{0}$.

ii) By convexity, all local minima of the problem are global minima.

iii) If f^{lin} is μ-strongly convex, the set of solutions of the problem consists of a single element.

b) *Smoothed biweight loss problem*

i) Since the function is nonconvex, a point with zero gradient is not necessarily a local minimum. It could be a saddle point or a global minimum.

ii) If $\boldsymbol{w}_0 = \bar{\boldsymbol{w}}$, then $\nabla f^{sb}(\boldsymbol{w}_0) = \mathbf{0}$, and thus $\boldsymbol{w}_1 = \boldsymbol{w}_0 - \alpha_0 \nabla f^{sb}(\boldsymbol{w}_0) = \boldsymbol{w}_0$. It follows that all iterates are equal to $\bar{\boldsymbol{w}}$, hence the method converges to $\bar{\boldsymbol{w}}$.

iii) Picking $\boldsymbol{w}_0$ at random in $\mathbb{R}^d$ avoids points with zero gradients that are not local minima.

Solutions for Exercise 3

a) Multiplying the objective function by a positive constant does not change the set of solutions. Indeed,

$$\begin{aligned}\boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f_i^{sto}(\boldsymbol{w}) \quad &\Leftrightarrow \quad \frac{1}{n}\sum_{i=1}^n f_i^{sto}(\boldsymbol{w}^*) \le \frac{1}{n}\sum_{i=1}^n f_i^{sto}(\boldsymbol{w}) \qquad \forall \boldsymbol{w}\in\mathbb{R}^d \\ &\Leftrightarrow \quad \frac{1}{2n}\sum_{i=1}^n f_i^{sto}(\boldsymbol{w}^*) \le \frac{1}{2n}\sum_{i=1}^n f_i^{sto}(\boldsymbol{w}) \qquad \forall \boldsymbol{w}\in\mathbb{R}^d \\ &\Leftrightarrow \quad \boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f_i^{sto}(\boldsymbol{w}).\end{aligned}$$

b) $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha\nabla f_{i_k}(\boldsymbol{w}_k)$, where $\alpha > 0$ and i_k is an index drawn randomly in $\{1,\dots,n\}$.

c) An epoch is a unit of cost corresponding to n accesses to data points in a dataset with n elements. With a budget of one epoch, one can perform n iterations of stochastic gradient.

d) i) The convergence rate of gradient descent is $\mathcal{O}\left(\frac{1}{K^{1/2}}\right)$. It is a better (faster) rate than stochastic gradient in terms of dependency on K, and it is also deterministic, unlike the rate in expectation for stochastic gradient.

ii) Since 1 epoch corresponds to the budget of one gradient descent iteration, we obtain the convergence rate $\mathcal{O}(\frac{1}{E^{1/2}})$.

iii) With a budget of E epochs, we perform nE iterations of stochastic gradient, which yields a convergence rate in $\mathcal{O}\left(\frac{1}{(nE)^{1/4}}\right)$. For large n and small E, the rate for stochastic gradient is better (faster) than the rate for gradient descent.

e) $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\alpha}{|\mathcal{S}_k|}\sum_{i\in\mathcal{S}_k}\nabla f_i(\boldsymbol{w}_k)$, where $\mathcal{S}_k$ is a set of indices drawn randomly with or without replacement in $\{1,\dots,n\}$, and $\alpha > 0$.

f) Taking $|\mathcal{S}_k| = 1$ corresponds to stochastic gradient, while taking $|\mathcal{S}_k| = n$ and drawing without replacement corresponds to gradient descent.

g) i) This observation illustrates that using a batch is a variance reduction technique.

ii) A variant with large batch typically converges to a smaller neighborhood of the solution, hence a smaller function value. However, it typically converges more slowly than variants with smaller batch sizes, because its behavior is close to that of gradient descent.

Solutions for Exercise 4

a) Given $\ell \in \{1,\dots,p\}$, we consider the problem

$$\operatorname*{minimize}_{\boldsymbol{w}_\ell\in\mathbb{R}^{d_1}} f_\ell^{ml}(\boldsymbol{w}_\ell) := \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}_\ell - \boldsymbol{y}_\ell\|_2^2. \tag{7}$$

i) The objective function of problem (5) is a sum of convex functions, which is convex.

ii) Letting $\boldsymbol{w}_\ell^k$ denote the kth iterate of gradient descent applied to problem (7), we have

$$\boldsymbol{w}_\ell^{k+1} = \boldsymbol{w}_\ell^k - \alpha \nabla f_\ell^{ml}(\boldsymbol{w}_\ell^k),$$

where $\alpha > 0$.

iii) Problem (7) is a finite-sum problem, and thus stochastic gradient can be applied to this problem.

b) The objective function of problem (5) can be rewritten as $f(\boldsymbol{W}) = \frac{1}{p}\sum_{\ell=1}^{p} f_\ell^{ml}(\boldsymbol{w}_\ell)$, where $\boldsymbol{w}_\ell$ is the ℓth column of $\boldsymbol{W}$. If $\boldsymbol{w}_1^*, \dots, \boldsymbol{w}_p^*$ are minima of $f_1^{ml}, \dots, f_p^{ml}$, respectively, it follows that $f_\ell^{ml}(\boldsymbol{w}_\ell^*) \le f_\ell^{ml}(\boldsymbol{w}_\ell)$ for any $\boldsymbol{w}_\ell$ and any $\ell = 1, \dots, p$. As a result, one also has $f(\boldsymbol{W}^*) \le f(\boldsymbol{W})$ for every $\boldsymbol{W}$, showing that the matrix $\boldsymbol{W}^* = [\boldsymbol{w}_1^* \ \cdots \ \boldsymbol{w}_p^*]$ is a solution of problem (7).

c) An algorithm for problem (5) could consist in running p variants of gradient descent on all problems of the form (7).