Exercises on Optimization for Machine Learning

M2 IASD 2025-2026

Version 2.0 - October 15, 2025*

Dauphine | PSL
UNIVERSITÉ PARIS

1. Basics of optimization

Exercise 1.1: Support vector machines

Given a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, finding a binary linear classifier for the data using support vector machines amounts to considering the problem

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \frac{1}{n}\sum_{i=1}^n \max\left\{1 - y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}, 0\right\} + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2, \tag{1}$$

where $\lambda \geq 0$.

Show that the objective of (1) is a convex function for any $\lambda \geq 0$, and that it is λ-strongly convex when $\lambda > 0$.

Exercise 1.2: Least-squares problems

Let $\boldsymbol{x} \in \mathbb{R}^d$ such that $\|\boldsymbol{x}\|_2 \neq 0$ and $\boldsymbol{y} \in \mathbb{R}^d$. The goal of the exercise is to compute a linear model that best fits $\boldsymbol{x}$ to $\boldsymbol{y}$, and to show the interest of using overparameterization in that setting.

a) Consider first the one-dimensional problem

$$\underset{w\in\mathbb{R}}{\text{minimize}} \frac{1}{2}\|w\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{2}$$

Is problem (2) convex? Is its optimal value equal to 0?

b) Consider now the matrix problem

$$\underset{\boldsymbol{W}\in\mathbb{R}^{d\times d}}{\text{minimize}} \frac{1}{2}\|\boldsymbol{W}\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{3}$$

Reformulate the problem so as to show that the objective function is convex in the sense defined in class, i.e. for vector-valued functions.

*Thanks to the students who spotted typos in earlier versions!

c) Show that

$$\underset{\boldsymbol{W}\in\mathbb{R}^{d\times d}}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{W}\boldsymbol{x}-\boldsymbol{y}\|_2^2 = 0$$

and that the problem does not have a unique global minimum.

Exercise 1.3: Strong convexity

Let $f:\mathbb{R}^d\to\mathbb{R}$ be $\mathcal{C}^1$ and μ-strongly convex, and denote by $\boldsymbol{w}^*$ the minimum of f.

a) For any $\boldsymbol{w}\in\mathbb{R}^d$, justify that the function

$$\varphi_{\boldsymbol{w}}:\boldsymbol{z}\ \longmapsto\ f(\boldsymbol{w})+\nabla f(\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{z}-\boldsymbol{w})+\frac{\mu}{2}\|\boldsymbol{z}-\boldsymbol{w}\|^2$$

is $\mathcal{C}^1$ and strongly convex.

b) Using $\nabla\varphi_{\boldsymbol{w}}(\boldsymbol{z})=\nabla f(\boldsymbol{w})+\mu(\boldsymbol{z}-\boldsymbol{w})$ for any $\boldsymbol{z}\in\mathbb{R}^d$, compute $\min_{\boldsymbol{z}}\varphi_{\boldsymbol{w}}(\boldsymbol{z})$ and $\operatorname{argmin}_{\boldsymbol{z}}\varphi_{\boldsymbol{w}}(\boldsymbol{z})$.

c) Conclude from the previous questions that

$$\|\nabla f(\boldsymbol{w})\|_2^2\quad\geq\quad 2\mu\,\left(f(\boldsymbol{w})-f(\boldsymbol{w}^*)\right).$$

Exercise 1.4: Co-coercivity

Let $f:\mathbb{R}^d\to\mathbb{R}$ be $\mathcal{C}_L^{1,1}$ and convex. Suppose that $\boldsymbol{w}^*\in\operatorname{argmin}_{\boldsymbol{w}} f(\boldsymbol{w})$ and let $f^*=f(\boldsymbol{w}^*)$.

a) Let $\boldsymbol{w}\in\mathbb{R}^d$. Show that $f(\boldsymbol{w})-f(\boldsymbol{w}^*)\geq\frac{1}{2L}\|\nabla f(\boldsymbol{w})\|_2^2$.

b) Let $(\boldsymbol{w},\boldsymbol{v})\in(\mathbb{R}^d)^2$. Show that

$$\left(\nabla f(\boldsymbol{v})-\nabla f(\boldsymbol{w})\right)^{\mathrm{T}}(\boldsymbol{v}-\boldsymbol{w})\geq\frac{1}{L}\|\nabla f(\boldsymbol{v})-\nabla f(\boldsymbol{w})\|_2^2.$$

Consider $\boldsymbol{z}\mapsto f(\boldsymbol{z})-\nabla f(\boldsymbol{v})^{\mathrm{T}}\boldsymbol{z}$ and $\boldsymbol{z}\mapsto f(\boldsymbol{z})-\nabla f(\boldsymbol{w})^{\mathrm{T}}\boldsymbol{z}$.

2. Derivatives and differentiation

Exercise 2.1: A simple function (Adapted from 2019-2020 exam)

Consider the one-dimensional function $f:\mathbb{R}\to\mathbb{R}$ defined by

$$f(w)=\sqrt{\sin(w)}+\sin(w). \tag{4}$$

a) Write a directed acylic graph representing the computation of $f(w)$.

b) Write down the instructions for computing $f'(w)$ in forward mode. Use it to evaluate $f'(\frac{\pi}{2})$.

c) Write down the instructions for computing $f'(w)$ in backward mode. Use it to evaluate $f'(\frac{\pi}{2})$.

Exercise 2.2: Sigmoid-type functions (Adapted from 2020-2021 exam)

Given $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in [0,1]$, we consider the two functions $f, g : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(\boldsymbol{w}) = y \ln\left(\sigma(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w})\right) \quad \text{and} \quad g(\boldsymbol{w}) = (1-y) \ln\left(1-\sigma(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w})\right), \tag{5}$$

where $\sigma(t) = \frac{1}{1+\mathrm{e}^{-t}}$. Recall that $\sigma'(t) = \sigma(t)\,(1-\sigma(t))$.

a) Write a computational graph representing the calculation of $f = f(\boldsymbol{w})$ by considering dot product, $\sigma(\cdot)$, logarithm and product as elementary operations.

b) Write down the chain rule for computing $\frac{\partial f}{\partial w}$ in forward mode.

c) Write down the chain rule for computing $\frac{\partial f}{\partial w}$ in backward mode.

d) Write a computational graph representing the calculation of $h = f(\boldsymbol{w}) + g(\boldsymbol{w})$ based on the graph from question a).

e) Write down the chain rule for computing $\frac{\partial h}{\partial \boldsymbol{w}}$ in backward mode.

3. Gradient descent

Exercise 3.1: A strongly convex problem

Consider the minimization problem

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w}) := \frac{1}{2}\|\boldsymbol{w}\|^2. \tag{6}$$

For any $\boldsymbol{w} \in \mathbb{R}^d$, we have $\nabla f(\boldsymbol{w}) = \boldsymbol{w}$ and $\nabla^2 f(\boldsymbol{w}) = \boldsymbol{I}_d$.

a) Justify that $f \in \mathcal{C}_1^{1,1}(\mathbb{R}^d)$ and that f is 1-strongly convex.

b) Adapt the general convergence rates of gradient descent and accelerated gradient on strongly convex problems to this particular case. What do the results suggest about the problem?

c) Write down the first iteration of gradient descent and accelerated gradient for the problem at hand using a stepsize of 1: is the result in agreement with the convergence rates?

Exercise 3.2: Convergence rates

Let $f : \mathbb{R}^d \to \mathbb{R}$. Suppose that $f \in \mathcal{C}_L^{1,1}(\mathbb{R}^d)$ and that f is μ-strongly convex. Suppose that we apply gradient descent with a constant stepsize $\alpha_k = \frac{1}{\mu+L}$ to this problem.

a) Using the descent property stated in class, justify that one has

$$f(\boldsymbol{w}_{k+1}) \leq f(\boldsymbol{w}_k) - \frac{2\mu + L}{2(\mu+L)^2}\|\nabla f(\boldsymbol{w}_k)\|_2^2 \tag{7}$$

holds at every iteration of gradient descent.

b) Using the previous question, show that gradient descent achieves a linear convergence rate with this stepsize. Are additional assumptions on μ and/or L needed?

c) Gradient descent with stepsize $\frac{1}{L}$ achieves a linear rate of convergence in $\left(1-\frac{\mu}{L}\right)^K$ on this problem. Compare this rate with that obtained in question b).

d) Accelerated gradient (with the appropriate parameterization for strongly convex functions) achieves a linear rate of convergence in $\left(1-\sqrt{\frac{\mu}{L}}\right)^K$ on this problem. Compare this rate with that obtained in question b).

Exercise 3.3: Saddle points

Consider the two-dimensional function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(\boldsymbol{w}) := w_1^2\left(w_1^2-2\right)+w_2^4.$$

a) For every $\boldsymbol{w} \in \mathbb{R}^2$, we have $\nabla f(\boldsymbol{w}) = \begin{bmatrix} 4w_1^3-4w_1 \\ 4w_2^3 \end{bmatrix}$. Find all first-order stationary points of f.

b) The formula for the Hessian matrix gives

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \nabla^2 f(\boldsymbol{w}) = \begin{bmatrix} 12w_1^2-4 & 0 \\ 0 & 12w_2^2 \end{bmatrix}.$$

Evaluate the Hessian at the points found in the previous question. What can you conclude about the nature of those points?

c) Without using derivatives, show that the critical points are either global minima or strict saddle points.

Exercise 3.4: Gradient descent with line search

Consider the minimization of a nonconvex, $\mathcal{C}^1$ function $f : \mathbb{R}^d \to \mathbb{R}$, assuming that f is bounded below by $f_{\text{low}} \in \mathbb{R}$. Suppose that we use gradient descent with a stepsize chosen through backtracking line search. More precisely, given $\theta \in (0,1)$ and $c \in (0,1/2)$, we select the stepsize at iteration k as the largest value in $\{\theta^j\}_{j\in\mathbb{N}}$ such that

$$f(\boldsymbol{w}_k - \theta^j \nabla f(\boldsymbol{w}_k)) < f(\boldsymbol{w}_k) - c\theta^j \|\nabla f(\boldsymbol{w}_k)\|_2^2. \tag{8}$$

a) Suppose that $\|\nabla f(\boldsymbol{w}_k)\|_2 > 0$ at iteration k. Show then that the line search terminates within a fixed number of iterations (independent of $\boldsymbol{w}_k$) and give a bound on the value of the stepsize α_k.

b) Show that the convergence rate of gradient descent in this setting is $\mathcal{O}(\frac{1}{\sqrt{K}})$, as in the fixed-stepsize case seen in class. Give the precise values for the constants.

c) What is the maximum number of backtracking steps at each iteration? How does this number illustrate the additional cost of line-search techniques?

Exercise 3.5: Hölder continuity (Adapted from 2021-2022 exam)

In this exercise, we consider a general, unconstrained optimization problem of the form

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{w}), \tag{9}$$

where $f \in \mathcal{C}^1$. Rather than assuming that the gradient is Lipschitz continuous (as in the lectures), we will assume a more general property called *Hölder continuity*. More precisely, for any $\nu \in (0, 1]$ and $L > 0$, we write $f \in \mathcal{C}_L^{1,\nu}$ if $f \in \mathcal{C}^1$ and

$$\forall (\boldsymbol{v}, \boldsymbol{w}) \in (\mathbb{R}^d)^2, \qquad \|\nabla f(\boldsymbol{v}) - \nabla f(\boldsymbol{w})\| \leq L\|\boldsymbol{v} - \boldsymbol{w}\|^\nu. \tag{10}$$

We say that ∇f is (L, ν)-Hölder continuous.

For any $f \in \mathcal{C}_L^{1,\nu}$, we can show that

$$\forall (\boldsymbol{v}, \boldsymbol{w}) \in (\mathbb{R}^d)^2, \qquad f(\boldsymbol{v}) \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^T(\boldsymbol{v} - \boldsymbol{w}) + \frac{L}{1+\nu}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2. \tag{11}$$

Moreover, since $f \in \mathcal{C}^1$, we can apply gradient descent to problem (9): the kth iteration of this method is

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k). \tag{12}$$

Part 1 We first study the case $\nu = 1$ (i.e. $f \in \mathcal{C}_L^{1,1}$), corresponding to the gradient being L-Lipschitz continuous.

a) Recalling that $f \in \mathcal{C}_L^{1,1}$ for this question, justify the choice $\alpha_k = \frac{1}{L}$ for every k.

b) Give the convergence rate result for gradient descent on problem (9), assuming that the function f is nonconvex. What quantity does this rate apply to?

c) Give the convergence rate result for gradient descent on problem (9), assuming that the function f is convex. What quantity does this rate apply to?

d) Assuming that f is convex, name a method with a better convergence rate than gradient descent, and give the associated rate.

Part 2 We now consider the more general setting $\nu \in (0, 1]$ (i.e. $f \in \mathcal{C}_L^{1,\nu}$) and f nonconvex.

a) Based on (11) and your answer to Part 1, justify the choice

$$\alpha_k = \left[\frac{\|\nabla f(w_k)\|^{1-\nu}}{L}\right]^{\frac{1}{\nu}} \tag{13}$$

for the stepsize used in the iteration (12).

b) Using the sequence $\{\alpha_k\}_k$ from the previous question, it is possible to show a convergence rate in $\mathcal{O}\left(\frac{1}{K^{\frac{\nu}{1+\nu}}}\right)$ for gradient descent applied to $f \in \mathcal{C}_L^{1,\nu}$ nonconvex and bounded below. Compare this bound with the results seen in class for the case $\nu = 1$.

c) In addition to $f \in \mathcal{C}_L^{1,\nu}$, suppose that f is twice continuously differentiable.

   i) Without any convexity assumption on f, are we guaranteed to converge towards a local minimum?

   ii) What guarantee seen in class can we provide about the limit points of gradient descent?

4. Nonsmooth and regularized optimization

Exercise 4.1: Robust linear regression (Adapted from 2024-2025 exam)

We consider a linear regression problem of the form

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \;\; f(\boldsymbol{w}) := \frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_1 = \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-y_i|, \tag{14}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1 \;\cdots\; \boldsymbol{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y}\in\mathbb{R}^n$ form the problem data.

a) Justify that the objective function of (14) is convex.

b) Let $\boldsymbol{w}^*$ and $\bar{\boldsymbol{w}}$ be two solutions of (14). Justify that the vector $\frac{\boldsymbol{w}^*+\bar{\boldsymbol{w}}}{2}$ is also a solution of the problem.

c) Due to the use of the ℓ_1 norm, the function f is not necessarily differentiable.

   i) Give a condition on the data $(\boldsymbol{X},\boldsymbol{y})$ under which f is actually continuously differentiable.

   ii) Assuming that the condition from the previous question does not hold, give a point in $\mathbb{R}^d$ at which the function is *not* differentiable.

d) Given any real $t\in\mathbb{R}$, we let $\mathrm{sgn}(t) = 1$ if $t>0$, $\mathrm{sgn}(t)=0$ if $t=0$ and $\mathrm{sgn}(t)=-1$ if $t<0$.

   i) For any $\boldsymbol{w}\in\mathbb{R}^d$ and $i\in\{1,\dots,n\}$, show that the vector $\mathrm{sgn}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-y_i)\boldsymbol{x}_i$ is a subgradient for the function $\boldsymbol{w}\mapsto \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-y_i$ at $\boldsymbol{w}$.

   ii) Consequently, justify that the vector

$$\boldsymbol{g}(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\mathrm{sgn}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}), \quad \text{where} \quad \mathrm{sgn}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}) = \begin{bmatrix}\mathrm{sgn}(\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{w}-y_1)\\ \vdots \\ \mathrm{sgn}(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{w}-y_n)\end{bmatrix} \tag{15}$$

   is a subgradient for f at $\boldsymbol{w}$.

   iii) Under the assumptions of question 3b), let $\boldsymbol{w}_0$ be the point considered in that question. Give another example of a subgradient for f at $\boldsymbol{w}_0$ other than $\boldsymbol{g}(\boldsymbol{w}_0)$.

e) We now consider solving (14) using a subgradient method.

   i) Write down the iteration of a subgradient method using a constant stepsize and the vector $\boldsymbol{g}(\boldsymbol{x})$ defined by (15).

   ii) Do you expect the function values corresponding to the method's iterates to be monotonically decreasing? Justify your answer.

   iii) What kind of guarantees seen in class can we provide about that method for a sufficiently small stepsize?

Exercise 4.2: Proximal operator (Adapted from 2022-2023 exam)

In this exercise, we revisit the proximal operator and the proximal gradient method on a specific problem. Given a vector $\boldsymbol{w} \in \mathbb{R}^d$, we consider

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \ \|\boldsymbol{w}\|_1 + \frac{1}{2\alpha}\|\boldsymbol{w}-\boldsymbol{z}\|_2^2, \tag{16}$$

where $\|\boldsymbol{w}\|_1 = \sum_{i=1}^d |[\boldsymbol{w}]_i|$, $\|\boldsymbol{w}\|_2^2 = \sum_{i=1}^d [\boldsymbol{w}]_i^2$ and $\alpha > 0$.

a) Explain why the objective function of (16) cannot be optimized by gradient-type techniques.

b) Justify that problem (16) and

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \ \alpha\|\boldsymbol{w}\|_1 + \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{z}\|_2^2 \tag{17}$$

have the same solution set (argmin). Since both functions are strongly convex, what can be said about this solution set?

c) Write down an optimality condition for problem (16).

d) In this question, we view problem (16) as computing a proximal operator.

i) Using the definition of the proximal operator, write down the solution of problem (16) as the value of a proximal operator of a certain function.

ii) By repeatedly solving instances of problem (16) using the last solution found as $\boldsymbol{z}$, what algorithm do we obtain?

e) In this question, we change perspective and view problem (17) in a composite form, where $\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{z}\|_2^2$ is a data-fitting term and $\alpha\|\boldsymbol{w}\|_1$ is a regularization term.

i) What is the purpose of such a regularization term? Why is it computationally worth using?

ii) Write down an iteration of proximal gradient applied to this problem with $\boldsymbol{w}_k = \boldsymbol{z}$ and stepsize α. What do you observe then? Is it to be expected?

Exercise 4.3: Regularization and sparsity (Adapted from 2024-2025 exam)

In this exercise, we consider an optimization problem with elastic net regularization of the form

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \, f(\boldsymbol{w}) + \lambda\left(\|\boldsymbol{w}\|_1 + \frac{\mu}{2}\|\boldsymbol{w}\|_2^2\right), \tag{18}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm, $\lambda > 0$ and $\mu > 0$.

a) Recall the purpose of the regularization terms $\|\cdot\|_1$ and $\|\cdot\|_2^2$ in optimization.

b) In this section, we consider the function $\Omega : \boldsymbol{w} \mapsto \lambda\|\boldsymbol{w}\|_1 + \frac{\lambda\mu}{2}\|\boldsymbol{w}\|_2^2$.

i) Recall the definition of the proximal operator associated with Ω, denoted by $\operatorname{prox}_\Omega(\cdot)$. Justify that this particular "prox" can be considered as a function from $\mathbb{R}^d$ to $\mathbb{R}^d$.

ii) Using properties of the proximal operator, one obtains

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \qquad \mathrm{prox}_{\Omega}(\boldsymbol{w}) = \frac{1}{\lambda\mu + 1}\mathrm{prox}_{\lambda\|\cdot\|_1}(\boldsymbol{w}). \tag{19}$$

where $\mathrm{prox}_{\lambda\|\cdot\|_1}(\cdot)$ corresponds to the proximal operator for (a multiple of) the ℓ_1 norm. Justify then that $\mathrm{prox}_{\Omega}(\boldsymbol{w})$ can be computed in closed form.

c) Suppose for this question that f is a continuously differentiable function.

i) Write down an iteration of the proximal gradient algorithm applied to problem (16) using the prox_{Ω} operator.

ii) Using the answer from the previous question, justify that the use of elastic net regularization is likely to produce sparse iterates that decay in norm for large values of the regularization parameter λ.