Tutorial 1: Basics of optimization

Optimization for machine learning, M2 MIAGE ID Apprentissage

September 21, 2023

Exercise 1: Linear least squares

We consider a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, wherein $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for every $i = 1, \dots, n$. We seek a linear model that best fits the data, which we formulate as the following optimization problem:

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

where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$ are given by

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

This problem is among the most classical in data analysis. Its objective function is $\mathcal{C}^2$, and the problem (1) always has at least one solution.

a) Let $\boldsymbol{w}^* \in \mathbb{R}^d$ satisfy $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$ (hence $\boldsymbol{w}^*$ is a solution of the linear system $\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}$). Justify then that $\boldsymbol{w}^*$ is a global minimum of the objective function.

b) The gradient of f at any $\boldsymbol{w} \in \mathbb{R}^d$ is given by $\nabla f(\boldsymbol{w}) = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$. If $\boldsymbol{w}^*$ is a local minimum of f, what is the value of $\nabla f(\boldsymbol{w}^*)$?

c) The Hessian matrix of f at $\boldsymbol{w} \in \mathbb{R}^d$ is given by $\nabla^2 f(\boldsymbol{w}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. Note that it is constant with respect to $\boldsymbol{w}$, and that it only depends on the data matrix $\boldsymbol{X}$.

   i) By construction, we have $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \succeq \boldsymbol{0}$. What property on f does this imply?

   ii) Suppose that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \succeq \mu \boldsymbol{I}_d$ with $\mu > 0$. Given $\boldsymbol{w} \in \mathbb{R}^d$, what can we say about $\nabla^2 f(\boldsymbol{w})$ in that case? What information does this provide about the set of solutions of problem (1)?

Exercise 2: Convex function

Let $q : \mathbb{R}^d \to \mathbb{R}$ be defined as $q(\boldsymbol{w}) = \frac{1}{4}\|\boldsymbol{w}\|^4$. This function is $\mathcal{C}^2$, and for every $\boldsymbol{w} \in \mathbb{R}^d$, we have

$$\nabla q(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 \boldsymbol{w}, \qquad \nabla^2 q(\boldsymbol{w}) = 2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} + \|\boldsymbol{w}\|^2 \boldsymbol{I}_d.$$

a) Using the expression of the Hessian matrix of q, show that the function q is convex. What does it imply on its local minima?

b) Show that the zero vector $\mathbf{0}_{\mathbb{R}^d}$ is a local minimum of q. Does it satisfy the second-order sufficient condition?

c) Given the answer to the previous question, can the function q be strongly convex?

d) Justify that the function has a single global minimum.

Exercise 3: Quasiconvex functions

A function $f : \mathbb{R}^d \to \mathbb{R}$ is called **quasiconvex** if

$$\forall \boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d,\ \forall t \in [0, 1], \quad f(t\boldsymbol{w} + (1-t)\boldsymbol{v}) \leq \max\{f(\boldsymbol{w}), f(\boldsymbol{v})\}. \tag{2}$$

Any convex function is quasiconvex, but the converse is not true.

Let f be a quasiconvex, $\mathcal{C}^2$ function. We consider:

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimize}}\ f(\boldsymbol{w}). \tag{3}$$

a) Write the first- and second-order optimality conditions for problem (3).

b) Since f is quasiconvex, it can be shown that

$$\forall \boldsymbol{w} \in \mathbb{R}^d,\ \forall \boldsymbol{v} \in \mathbb{R}^d, \quad \boldsymbol{v}^{\mathrm{T}} \nabla f(\boldsymbol{w}) = 0 \Rightarrow \boldsymbol{v}^{\mathrm{T}} \nabla^2 f(\boldsymbol{w}) \boldsymbol{v} \geq 0. \tag{4}$$

Let $\boldsymbol{w}^*$ be a first-order stationary point. Justify that $\boldsymbol{w}^*$ is also a second-order stationary point.

Solutions

Solutions for Exercise 1

Underlying goal: Introduce least-squares formulations. Apply the definitions of global minima/solutions and that of convexity.

a) If $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{y}$, then

$$f(\boldsymbol{w}^*) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}\|^2 = \frac{1}{2}\|\mathbf{0}\|^2 = 0.$$

Since f is always nonnegative (definition of a norm), we also have

$$\forall \boldsymbol{w} \in \mathbb{R}^d, f(\boldsymbol{w}) \geq 0 = f(\boldsymbol{w}^*).$$

The latter property corresponds to the definition of a global minimum for f, from which we conclude that $\boldsymbol{w}^*$ is a global minimum of f or, equivalently, a solution of the unconstrained problem (1).

b) The function f is continuously differentiable ($\mathcal{C}^2$, so $\mathcal{C}^1$). If $\boldsymbol{w}^*$ is a local minimum of f, then $\nabla f(\boldsymbol{w}^*) = \mathbf{0}$ per the first-order optimality condition.

i) If $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \succeq \mathbf{0}$, then $\nabla^2 f(\boldsymbol{w}) \succeq \mathbf{0}$ for any $\boldsymbol{w} \in \mathbb{R}^d$. This property is a characterization of convexity for a $\mathcal{C}^2$ function, from which we conclude that f is a convex function.

ii) Similarly to the previous question, the fact that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \succeq \mu \boldsymbol{I}_d$ means that $\nabla^2 f(\boldsymbol{w}) \succeq \mu \boldsymbol{I}_d$ for any $\boldsymbol{w} \in \mathbb{R}^d$. This is again a characterization of strong convexity for $\mathcal{C}^2$ functions, and therefore f is μ-strongly convex. As a result, there exists a unique solution for the optimization problem (or equivalently, f has a unique global minimum).

Solutions for Exercise 2

Goal: Introduce a bit more calculus to get students comfortable with scalar products and matrix-vector products. Give an example of global minimum that does not satisfy the sufficient optimality condition.

a) For any $\boldsymbol{w} \in \mathbb{R}^d$ and any $\boldsymbol{v} \in \mathbb{R}^d$, the linearity of both scalar products and matrix-vector products gives:

$$\begin{aligned}
\boldsymbol{v}^{\mathrm{T}}\nabla^2 q(\boldsymbol{w})\boldsymbol{v} &= \boldsymbol{v}^{\mathrm{T}}(2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} + \|\boldsymbol{w}\|^2\boldsymbol{I}_d)\boldsymbol{v} \\
&= \boldsymbol{v}^{\mathrm{T}}(2\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v} + \|\boldsymbol{w}\|^2\boldsymbol{v}) \\
&= 2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v} + \|\boldsymbol{w}\|^2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} \\
&= 2(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v})^2 + \|\boldsymbol{w}\|^2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} \\
&= 2(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v})^2 + \|\boldsymbol{w}\|^2\|\boldsymbol{v}\|^2 \\
&\geq 0.
\end{aligned}$$

Thus, for any $\boldsymbol{w} \in \mathbb{R}^d$, the Hessian matrix $\nabla^2 q(\boldsymbol{w})$ is positive semidefinite, i.e. $\nabla^2 q(\boldsymbol{w}) \succeq \mathbf{0}$. Consequently, the ($\mathcal{C}^2$) function q is convex, and all its local minima are global.

b) Since the function q is convex, every local minimum is global. Moreover, we have

$$q(\boldsymbol{w}) = \frac{1}{4}\|\boldsymbol{w}\|^4 \geq 0 = q(\mathbf{0}_{\mathbb{R}^d})$$

for any $\boldsymbol{w} \in \mathbb{R}^d$. The zero vector $\mathbf{0}_{\mathbb{R}^d}$ is thus a global minimum of q. If the zero vector were to satisfy the second-order sufficient optimality conditions, we would have $\nabla^2 q(\mathbf{0}_{\mathbb{R}^d}) \succ \mathbf{0}$. However, the expression for $\nabla^2 q$ gives

$$\nabla^2 q(\mathbf{0}_{\mathbb{R}^d}) = \mathbf{0},$$

and the zero matrix is only positive semidefinite (instead of positive definite). As a result, the zero vector does not satisfy the second-order sufficient optimality conditions. *Note: This does not contradict the fact that this vector is a global minimum, as the condition is sufficient but not necessary.*

c) If the function were strongly convex, there would exist $\mu > 0$ such that $\nabla^2 q(\boldsymbol{w}) \succeq \mu \boldsymbol{I}_d \succ \mathbf{0}$ for any $\boldsymbol{w}$, including the zero vector. Since the Hessian is zero at the zero vector, this cannot be true, from which we conclude that q is not strongly convex.

d) For every $\boldsymbol{w} \in \mathbb{R}^d$, we have $q(\boldsymbol{w}) \geq q(\mathbf{0}_{\mathbb{R}^d}) = 0$, hence the zero vector is a global minimum. Moreover, $q(\boldsymbol{w}) = 0$ if and only if $\boldsymbol{w} = \mathbf{0}_{\mathbb{R}^d}$, and thus the zero vector is the only global minimum of q.
Note: Classical argument in this last question, typical first question of an exam.

Solutions for Exercise 3

a) *The result is expected to be known.* The first-order necessary optimality conditions can be stated as follows. If a vector $\boldsymbol{w}^* \in \mathbb{R}^d$ is a local minimum of a $\mathcal{C}^1$ function f, then $\nabla f(\boldsymbol{w}^*) = \mathbf{0}$. The second-order necessary optimality conditions are a stronger characterization. If $\boldsymbol{w}^* \in \mathbb{R}^d$ is a local minimum of f, then

$$\nabla f(\boldsymbol{w}^*) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\boldsymbol{w}^*) \succeq \mathbf{0}.$$

b) Since $\boldsymbol{w}^*$ is a first-order stationary point, it satisfies the first-order necessary conditions, hence $\nabla f(\boldsymbol{w}^*) = \mathbf{0}$ and

$$\forall \boldsymbol{v} \in \mathbb{R}^d, \quad \boldsymbol{v}^{\mathrm{T}} \nabla f(\boldsymbol{w}^*) = \boldsymbol{v}^{\mathrm{T}} \mathbf{0} = 0.$$

The left-hand side of the implication (4) thus holds for $\boldsymbol{w}^*$ and any vector $\boldsymbol{v}$. Thus the right-hand also holds, i.e.

$$\boldsymbol{v}^{\mathrm{T}} \nabla^2 f(\boldsymbol{w}^*) \boldsymbol{v} \geq 0 \ \forall \boldsymbol{v} \in \mathbb{R}^d,$$

which is equivalent to $\nabla^2 f(\boldsymbol{w}^*) \succeq \mathbf{0}$. Therefore, the vector $\boldsymbol{w}^*$ satisfies the second-order necessary optimality conditions, and it is a second-order stationary point.