Convex optimization for machine learning – TD1

Olivier Fercoq
Bruno Costacèque

Novembre 2022

1 Exercises

Exercise 1. (Banach-Picard's fixed point theorem)

Let E be a complete metric space and $f : E \to E$ a **contracting** mapping, that is:

$$\exists \rho \in [0, 1[,\ \forall x, y \in E,\ \mathrm{d}\big(f(x), f(y)\big) \leq \rho \mathrm{d}(x, y).$$

Prove that f admits **one unique** fixed point[1] x^*, and that the sequence defined by:

$$\begin{cases} x_0 \in E \\ x_{n+1} = f(x_n),\ \forall n \in \mathbb{N}. \end{cases}$$

converges to x^* with rate of convergence ρ^n, *i.e.*:

$$\forall n \in \mathbb{N},\ \mathrm{d}(x_n, x^*) \leq \rho^n \mathrm{d}(x_0, x^*).$$

Exercise 2. (Gradient calculus)

1. Calculate the gradient of the following functions. A, M and Q are fixed matrices, and $\mathbf{b}$ is a fixed vector. More precisely, M belongs to $\mathcal{M}_{m,n}(\mathbb{R})$ and Q to $\mathcal{M}_{p,n}(\mathbb{R})$.

   (a)
   $$\begin{array}{rccc} f_1 & : & \mathbb{R}^n & \to & \mathbb{R} \\ & & \mathbf{x} & \mapsto & \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2 \end{array}$$

   (b)
   $$\begin{array}{rccc} f_2 & : & \mathbb{R}^n & \to & \mathbb{R} \\ & & \mathbf{x} & \mapsto & \sum_{i=1}^n \log(1 + \exp(x_i)) \end{array}$$

   (c)
   $$\begin{array}{rccc} f_3 & : & \mathcal{M}_{m,p}(\mathbb{R}) & \to & \mathbb{R} \\ & & P & \mapsto & \frac{1}{2}\|M - PQ\|_F^2 \end{array}$$

2. Let g_1, g_2 and g_3 be functions such that $g_1 : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$, $g_2 : \mathbb{R}^{n_2} \to \mathbb{R}^{n_3}$ and $g_3 : \mathbb{R}^{n_3} \to \mathbb{R}$. Let:

   $$f_4 = g_3 \circ g_2 \circ g_1.$$

   Compute the gradient of f_4 using the Jacobian matrices of g_i for $i \in [\![1, 3]\!]$.

[1] A point such that $f(x^*) = x^*$.

3. Suppose that computing one element of the Jacobian matrices costs C_J and that multiplying two numbers costs C_M. How much does it cost to compute $\nabla f_4(x)$?

Exercise 3. Let f be a $\mathcal{C}^2$ function from $\mathbb{R}^d$ to $\mathbb{R}$. Assume there exist $\mu, L \in \mathbb{R}_+$ such that[2]

$$\mu I_d \preceq \nabla^2 f(\mathbf{x}) \preceq L I_d.$$

1. Let $M \in \mathcal{S}(\mathbb{R}^d)$ be a real symmetric matrix satisfying the previous assumption and note $\|\cdot\|_2$ its operator matrix with respect to the Euclidean norm on $\mathbb{R}^d$.

   (a) Let λ_1 be the greatest eigenvalue of M. Show that

   $$\|M\|_2 = |\lambda_1|.$$

   (b) Prove that:

   $$\mu \leq \|M\|_2 \leq L.$$

2. Show that the fixed point operator $F : \mathbf{x} \mapsto \mathbf{x} - \gamma \nabla f(\mathbf{x})$ is contractant for every γ in $]0, 2/L[$.

3. Show that the gradient method converges linearly[3] to some x^*.

4. How many iterations are necessary to ensure that $\|x_k - x^*\| \leq \varepsilon$?

[2] Where $\preceq$ denotes the partial order known as *Loewner order*: if M, N are positive semi-definite matrices, $M \preceq N$ if $N - M$ is positive semi-definite.

[3] That is:

$$\exists \rho \in]0,1[, \ \lim_{n \to \infty} \big(\mathrm{d}(x_{n+1}, x^*)\big)^{\frac{1}{n}} \leq \rho.$$

2 Solutions

Exercise 1 • Assume first we have proved that the sequence $(x_n)_n$ converges to some limit x^*. Then by continuity of f (which is ρ-Lipschitz by assumption), we have:

$$f(x^*) = f\big(\lim_{n\to\infty} x_n\big) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x^*,$$

and so x^* would be a fixed point. It is unique since if there were another one, say y^*, we would have:

$$\mathrm{d}(x^*, y^*) = \mathrm{d}\big(f(x^*), f(y^*)\big) \leq \rho \mathrm{d}(x^*, y^*) < \mathrm{d}(x^*, y^*),$$

a contradiction.

- Now let us prove that $(x_n)_n$ indeed converges. So far we have no real clue about the value of its limit. But we know E is a complete metric space, and so Cauchy sequences always converge in E to some limit[4]. So we have to prove that (x_n) is a Cauchy sequence. Let $n, m \in \mathbb{N}$. We have:

$$\mathrm{d}(x_{n+m}, x_n) = \mathrm{d}\big(f^{\circ n}(x_m), f^{\circ n}(x_0)\big) \leq \rho^n \mathrm{d}(x_m, x_0).$$

Letting n go to infinity, we get that $(x_n)_n$ is indeed a Cauchy sequence in the complete space E. Using that x^* is a fixed point of f, the same reasoning yields:

$$\forall n \in \mathbb{N},\ \mathrm{d}(x^*, x_n) \leq \rho^n \mathrm{d}(x^*, x_0).$$

Remarque 1. ► *The proof of this theorem is very simple in comparison to its usefulness. For example it is the main argument behind the proof of Cauchy-Lipschitz, another powerful theorem which ensures the existence and the uniqueness of a solution for a very large class of ODEs*[5]*.*

► ***All** the hypotheses of the statement of this theorem are important. For example, if ρ is not strictly lesser than 1, troubles may appear. In particular, it is not enough to ask that*

$$\forall x, y \in E,\ \mathrm{d}\big(f(x), f(y)\big) < \mathrm{d}(x, y),$$

unless E is compact (and metric, so complete). It is a good exercise to try and prove this result.

Exercise 2 1. (a) The function f_1 can be decomposed as $f_1 = N \circ \varphi$, where $N(\mathbf{x}) = \|\mathbf{x}\|_2^2/2$ et $\varphi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$. The latter is a linear map, so its Jacobian matrix is A since:

$$\varphi(\mathbf{x} + \mathbf{h}) = A(\mathbf{x} + \mathbf{h}) = \varphi(\mathbf{x}) + A\mathbf{h}.$$

As for the Euclidean norm function, its differential can be computed directly too:

$$N(\mathbf{x} + \mathbf{h}) = N(\mathbf{x}) + \langle \mathbf{x}, \mathbf{h} \rangle + N(\mathbf{h}),$$

and so $\mathrm{d}N(\mathbf{x}).\mathbf{h} = \langle \mathbf{x}, \mathbf{h} \rangle$, thus its gradient is just $2\mathbf{x}$. Thus the chain rule gives:

$$\mathrm{d}f_1(\mathbf{x}).\mathbf{h} = \mathrm{d}N(A\mathbf{x} + \mathbf{b}).(A\mathbf{h}) = \langle A\mathbf{x} + \mathbf{b}, A\mathbf{h} \rangle = \big\langle A^\top (A\mathbf{x} + \mathbf{b}), \mathbf{h} \big\rangle,$$

which implies that:

$$\nabla f_1(\mathbf{x}) = A^\top (A\mathbf{x} + \mathbf{b}).$$

[4] The original purpose of Cauchy sequences and complete spaces is precisely to show that a given sequence converges, even without knowing a candidate for its limit! It is extremely useful to prove very general and powerful existence theorems, as you will soon see.

[5] Unfortunately this kind of result does not exist for PDEs, which is a huge source of troubles...

(b) The easiest way to explicit this gradient is to compute the partial derivatives of f_2. We find:

$$\forall i \in \mathbb{R}^n, \forall \mathbf{x} \in \mathbb{R}^n,\ \partial_i f_2(\mathbf{x}) = \frac{e^{x_i}}{1+e^{x^i}} = \frac{1}{1+e^{-x_i}} =: \sigma(x_i),$$

so the i-th partial derivative of f_2 is the **logistic function**, which is often used as an activation function for neural networks, especially in classification. Thus the gradient is:

$$\nabla f_2(\mathbf{x}) = \begin{pmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_n) \end{pmatrix}.$$

(c) Formally speaking, this computation looks exactly the same as in the first question, but this time, the argument P is on the left of the parameter matrix Q (while the argument $\mathbf{x}$ was on the right of the parameter matrix A). This time, we will develop the squared norm of $M-(P+H)Q$ divided by two to find directly the differential of f_3 at P taken in $H \in \mathcal{M}_{m,p}(\mathbb{R})$:

$$f_3(P+H) = \frac{1}{2}\|M-PQ\|_F^2 - \langle M-PQ, HQ\rangle_F + \frac{1}{2}\|HQ\|_F^2,$$

and so

$$\mathrm{d}f_3(P).H = -\langle M-PQ, HQ\rangle_F = -\mathrm{Tr}\big((M-PQ)Q^\top H^\top\big) = \big\langle (M-PQ)Q^\top, H\big\rangle_F.$$

Therefore, we get the following gradient (don't forget the minus sign!):

$$\nabla f_3(\mathbf{x}) = -(M-PQ)Q^\top.$$

Remarque 2. *When dealing with this kind of tedious (but important) computation, it is essential to do as many sanity-checks as possible. Here, you could verify that your result is well-defined by looking at the dimensions of the matrices involved: $M-PQ$ belongs to $\mathcal{M}_{m,n}(\mathbb{R})$ and $Q^\top$ belongs to $\mathcal{M}_{n,p}(\mathbb{R})$, so our gradient lives in $\mathcal{M}_{m,p}(\mathbb{R})$, which is consistent with the definition of f_3. Plus its expression bears a lot of similarities with the gradient of f_1!*

2. Applying the chain rule two times, first to $g_3 \circ g_2$ and g_1, and then to g_3 and g_2, we find:

$$\begin{aligned}
\forall \mathbf{x}, \mathbf{h} \in \mathbb{R}^{n_1},\ \mathrm{d}f_4(\mathbf{x}) &= \mathrm{d}(g_3 \circ g_2)(g_1(\mathbf{x})).\big(\mathrm{d}g_1(\mathbf{x}).\mathbf{h}\big) \\
&= \mathrm{d}g_3\big(g_2(g_1(\mathbf{x}))\big).\big(\mathrm{d}g_2(g_1(\mathbf{x})).\mathrm{d}g_1(\mathbf{x}).\mathbf{h}\big) \\
&= \mathrm{d}g_3\big((g_2 \circ g_1)(\mathbf{x})\big).\big(\mathrm{d}g_2(g_1(\mathbf{x})).\mathrm{d}g_1(\mathbf{x}).\mathbf{h}\big) \\
&= \mathrm{d}g_3\big((g_2 \circ g_1)(\mathbf{x})\big).\big(\mathrm{d}g_2(g_1(\mathbf{x})) \circ \mathrm{d}g_1(\mathbf{x})\big).\mathbf{h} \\
&= \big\langle Jg_1(\mathbf{x})^\top Jg_2(g_1(\mathbf{x}))^\top \nabla g_3\big((g_2 \circ g_1)(\mathbf{x})\big), \mathbf{h}\big\rangle,
\end{aligned}$$

and this horrible expressions reads as 'the differential of g_3 computed at point $(g_2 \circ g_1)(\mathbf{x})$ taken[6] in the differential of g_2 at $g_1(\mathbf{x})$ composed with the differential of g_1 at $\mathbf{x}$ in $\mathbf{h}$'. Notice the different organization of the parentheses between the third and the fourth lines! Finally we find:

$$\nabla f_4(\mathbf{x}) = Jg_1(\mathbf{x})^\top Jg_2(g_1(\mathbf{x}))^\top \nabla g_3\big((g_2 \circ g_1)(\mathbf{x})\big).$$

[6] As a linear mapping.

3. This question looks simple but there is a huge subtlety here. First, let's compute the cost $C_{m,n,p}$ of multiplying two matrices $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and $B \in \mathcal{M}_{n,p}(\mathbb{R})$. The (i,j)-th coefficient of AB is the scalar product of the i-th line of A and the j-th column of B, so it requires mnp operations to compute all the coefficients AB and:

$$C_{m,n,p} = mnpC_M.$$

Now, if we compute the gradient of f_4 by multiplying matrices from left to right in the formula of the previous question, we must compute the product of two matrices, then multiply the result with a vector, so the multiplication cost is $C_{n_1,n_2,n_3} + C_{n_1,n_3,1} = (n_1n_2n_3 + n_1n_3)C_M$. But if we multiply from right to left, we **always** manipulate vectors and so we find instead $C_{n_2,n_3,1} + C_{n_1,n_2,1} = (n_2n_3 + n_1n_2)C_M$. This approach is much more efficient than the 'natural' way of multiplying matrices! It accounts a lot for the success of modern deep learning algorithms, even though it looks terribly simple!

Finally, using the backpropagation approach rather than the naive one, we find a total cost equal to $(n_1n_2 + n_2n_3 + n_3)C_J + (n_1n_2 + n_2n_3)C_M$.

Remarque 3. *f_1 is useful for least squares and regression problems while f_2 is useful for logistic regression and binary classification, and f_3 is useful for nonnegative matrix factorization.*

Exercise 3 1. (a) Since M is a real symmetric matrix, it is diagonalizable in an orthonormal basis:

$$\exists P \in \mathbf{O}(d),\ M = P\mathrm{diag}(\lambda_1, \dots, \lambda_d)P^\top =: PDP^\top.$$

Let $\mathbf{x}$ a unit norm vector. We have:

$$\|M\mathbf{x}\|^2 = \|PDP^\top\mathbf{x}\|^2 = \|DP^\top\mathbf{x}\|^2 = \sum_{i=1} \lambda_i^2 x_i \leq \lambda_1^2\|\mathbf{x}\|^2 = \lambda_1^2.$$

Thus $\|M\|_2 \leq |\lambda_1\|$. The equality case is attained by taking $\mathbf{x}$ the unit norm eigenvector associated with λ_1.

(b) Since the eigenvalues of $M - \mu I_d$ are the eigenvalues of M minus μ, and since the eigenvalues of a semi-definite positive matrix are positive, we have that:

$$\mathrm{Sp}(M - \mu I_d) \subseteq \mathbb{R}_+.$$

In particular, all the eigenvalues of M are greater[7] than $\mu \geq 0$, and thus:

$$\mu \leq \lambda_1 = |\lambda_1| = \|M\|_2.$$

The same line of reasoning yields $\|M\|_2 \leq L$, thus proving the desired result.

2. The gradient of $\mathbf{x} \mapsto \mathbf{x} - \gamma\nabla f(\mathbf{x})$ is $F : \mathbf{x} \mapsto I_d - \gamma\nabla^2 f(\mathbf{x})$. We must prove that its operator norm is strictly less than 1 for all $\mathbf{x}$. Since:

$$\forall \mathbf{x} \in \mathbb{R}^d,\ \mathrm{Sp}\big(F(\mathbf{x})\big) = 1 - \gamma\mathrm{Sp}\big(F(\mathbf{x})\big),$$

we know that the greatest eigenvalue λ_1 of $F(\mathbf{x})$ belongs to $[1 - \gamma L, 1 - \gamma\mu]$ by assumption. As a result:

$$\forall \mathbf{x} \in \mathbb{R}^d, -1 < 1 - \gamma L \leq \|F(\mathbf{x})\|_2 \leq 1 - \gamma\mu < 1.$$

[7] And so, the hypothesis on f implies that $\nabla^2 f(\mathbf{x})$ is semi-definite positive for all $\mathbf{x}$, *i.e.* f is convex on $\mathbb{R}^d$.

In particular[8]

$$\rho := \sup_{\mathbf{x}\in\mathbb{R}^d} \|F(\mathbf{x})\|_2 < 1,$$

and so $\mathbf{x} \mapsto \mathbf{x} - \gamma\nabla f(\mathbf{x})$ is contractant.

3. It suffices to apply the Banach-Picard's fixed point theorem to the contracting map $\mathbf{x} \mapsto \mathbf{x} - \gamma\nabla f(\mathbf{x})$.

4. Thanks to the the first exercise, we have the estimate:

$$\forall n \in \mathbb{N},\ \|\mathbf{x}_n - \mathbf{x}^*\| \leq \rho^n \|\mathbf{x}_0 - \mathbf{x}^*\|,$$

where $(x_n)_n$ is defined by the gradient method with starting point $\mathbf{x}_0$. Since:

$$\begin{aligned}\rho^n\|\mathbf{x}_0 - \mathbf{x}^*\| \leq \varepsilon &\iff n\log(\rho) + \log\left(|\mathbf{x}_0 - \mathbf{x}^*\|\right) \leq \log(\varepsilon)\\ &\iff n \geq \frac{1}{\log(\rho)}\left[\log(\varepsilon) - \log\left(\|\mathbf{x}_0 - \mathbf{x}^*\|\right)\right] =: \eta\end{aligned}$$

we need at least $\lceil\eta\rceil = \lfloor\eta\rfloor + 1$ iterations before we get an estimate of $\mathbf{x}^*$ with an error less than ε. Of course this result is not very useful in practice since we do not know $\mathbf{x}^*$. But it underlies the importance of choosing our initial guess $\mathbf{x}_0$ as closely as possible from $\mathbf{x}^*$.

[8] The terms $1-\gamma L$ and $1-\gamma\mu$ are essential, as strict inequalities become large inequalities when taking the supremum, and so we would have lost the very contraction property we have been trying to prove! For example, $1 - n^{-1} < 1$, but $\sup_{n\geq 0}(1 - n^{-1}) = 1$...

Convex optimization for machine learning – TD2

Olivier Fercoq
Bruno Costacèque

Novembre 2022

1 Exercises

Exercise 1. (Proximal operator) Let $f : \mathbb{R}^n \to \mathbb{R} \cup +\infty$ be a convex lower-semicontinuous function such that $\mathrm{dom} f \neq \emptyset$.

1. Recall the definition of the domain of a convex function.

2. It is possible to prove (but we do not ask you to do it) that $\exists \mathbf{x}_0 \in \mathrm{dom} f$ such that $\partial f(\mathbf{x}_0) \neq \emptyset$. Using this information, show that there exists $\alpha \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^n$ such that for all $\mathbf{x}$ we have $f(\mathbf{x}) \geq \alpha + \langle \mathbf{w}, \mathbf{x} \rangle$.

3. Let us fix $\mathbf{x} \in \mathbb{R}^n$. Let us define $g : \mathbf{y} \mapsto f(\mathbf{y}) + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$. Show that g is strongly convex.

4. Show that $\lim\limits_{\|\mathbf{y}\| \to \infty} g(\mathbf{y}) = +\infty$.

5. Show that g has a minimizer and that it is unique. We will denote this minimizer as $\mathrm{prox}_f(\mathbf{x})$. The function $\mathrm{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ is called the *proximal operator* of f.

Exercise 2. Let us denote:

$$\mathrm{prox}_g(\mathbf{y}) = \arg\min_{\mathbf{x} \in \mathcal{X}} \Big\{ g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2 \Big\}$$

the proximal operator of g at $\mathbf{y}$. Fix $\gamma > 0$. Show that if f and g are convex, and f differentiable, then the fixed points of the nonlinear equation

$$\mathbf{x} = \mathrm{prox}_{\gamma g}\big(\mathbf{x} - \gamma \nabla f(\mathbf{x})\big) \tag{1}$$

are the minimizers of the function $F = f + g$.

Exercise 3. (Taylor-Lagrange inequality) The goal of this exercise is to prove Taylor-Lagrange inequality. This is a fundamental inequality for the study of gradient descent and related methods.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz *i.e.* $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|$ for all $\mathbf{x}, \mathbf{y}$.

1. Prove that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\big\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \big\rangle \leq L\|\mathbf{y} - \mathbf{x}\|^2$.

2. Set $\varphi(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$ for all $t \in [0, 1]$. Prove that:

$$f(\mathbf{y}) - f(\mathbf{x}) - \big\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \big\rangle = \varphi(1) - \varphi(0) - \varphi'(0).$$

3. Deduce that:

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \mathrm{d}t$$

4. Using the first question, conclude that

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

2 Solutions

Exercise 1

1. The domain of the function f is the set of points $\mathbf{x} \in \mathbb{R}^n$ such that $f(\mathbf{x}) \neq +\infty$.

2. By definition of the sub-differential, we have for $\mathbf{q}_0 \in \partial f(\mathbf{x}_0)$:

$$\forall \mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) \geq f(\mathbf{x}_0) + \langle \mathbf{q}_0, \mathbf{x} - \mathbf{x}_0 \rangle = \big[f(\mathbf{x}_0) - \langle \mathbf{q}_0, \mathbf{x}_0 \rangle\big] + \langle \mathbf{q}_0, \mathbf{x} \rangle.$$

   Setting $\alpha := f(\mathbf{x}_0) - \langle \mathbf{q}_0, \mathbf{x}_0 \rangle$ and $\mathbf{w} = \mathbf{q}_0$, we obtain the desired result.

3. By definition, a function h is strongly convex if $\mathbf{y} \mapsto h(\mathbf{y}) - \frac{\mu}{2}\|\mathbf{y}\|^2$ is convex for some $\mu \geq 0$. Since

$$g(\mathbf{y}) = \Big[f(\mathbf{y}) - \langle \mathbf{x}, \mathbf{y} \rangle + \frac{1}{2}\|\mathbf{x}\|^2\Big] + \frac{1}{2}\|\mathbf{y}\|^2,$$

   and the function f is convex, as well as the affine function $\mathbf{y} \mapsto -\langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{x}\|^2/2$, it is clear that g is 1-strongly convex.

4. By definition of g, and by using question 2, we have:

$$g(\mathbf{y}) \geq \alpha + \langle \mathbf{w}, \mathbf{y} \rangle + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

   This inequality proves that g converges to $+\infty$ when $\|y\|$ goes to infinity (since $\langle \mathbf{w}, \mathbf{y} \rangle \geq -\|\mathbf{w}\|\|\mathbf{y}\|$ by Cauchy-Schwarz inequality).

5. - **(Existence)** The existence of a minimizer $\mathbf{x}^*$ of the l.s.c. and coercive function g is a consequence of the results seen during the lecture.

   - **(Unicity)** We will prove a general and important result which is useful to keep in mind: a strongly convex function h admits at most one minimizer. Assume the converse. There exist distinct $\mathbf{x}^*, \mathbf{x}^{**}$ such that:

$$h(\mathbf{x}^*) = h(\mathbf{x}^{**}) = \min_{\mathbf{x} \in \mathcal{X}} h(\mathbf{x}).$$

     Let $t \in]0, 1[$ (any t in this interval will do, you can assume $t = 1/2$ will do if you wish[1]). Since h is μ-strongly convex for some $\mu > 0$, we have by the very definition of the convexity of $\mathbf{x} \mapsto h(\mathbf{x}) - \mu\|\mathbf{x}\|^2/2$:

$$\begin{aligned} h\big(t\mathbf{x}^* + (1-t)\mathbf{x}^{**}\big) &\leq t h(\mathbf{x}^*) + (1-t)h(\mathbf{x}^{**}) - \frac{\mu}{2}\Big[\|t\mathbf{x}^* + (1-t)\mathbf{x}^{**}\|^2 - t\|\mathbf{x}^*\|^2 - (1-t)\|\mathbf{x}^{**}\|^2\Big] \\ &= h(\mathbf{x}^*) - \frac{\mu}{2}\Big[\|t\mathbf{x}^* + (1-t)\mathbf{x}^{**}\|^2 - t\|\mathbf{x}^*\|^2 - (1-t)\|\mathbf{x}^{**}\|^2\Big] \\ &< h(\mathbf{x}^*) \\ &\leq h\big(t\mathbf{x}^* + (1-t)\mathbf{x}^{**}\big), \end{aligned}$$

     since the term within the brackets is positive by (strict) convexity of the squared Euclidean norm function. The last inequality stems from the fact $\mathbf{x}^{**}$ is a minimizer of h. But it is absurd! So h can only have one minimizer at most. This result proves that g admits exactly one minimizer.

   This minimizer depends on $\mathbf{x}$ and since it is unique, it defines a function of $\mathbf{x}$: the proximal operator of f at point $\mathbf{x}$.

[1] But it's more annoying to write!

Exercise 2 By assumption, $\mathbf{x} = \mathrm{prox}_{\gamma g}\big(\mathbf{x} - \gamma\nabla f(\mathbf{x})\big)$, so it minimizes the function:

$$\varphi : \mathbf{y} \mapsto \gamma g(\mathbf{y}) + \frac{1}{2}\big\|\mathbf{y} - \big(\mathbf{x} - \gamma\nabla f(\mathbf{x})\big)\big\|^2$$

By Fermat's rule, this implies that $\mathbf{0} \in \partial\varphi(\mathbf{x})$. Let's compute this sub-gradient.

$$\begin{aligned}
\partial\varphi(\mathbf{x}) &= \partial(\gamma g)(\mathbf{x}) + \big\{\mathbf{x} - \big(\mathbf{x} - \gamma\nabla f(\mathbf{x})\big)\big\} \\
&= \partial(\gamma g)(\mathbf{x}) + \big\{\gamma\nabla f(\mathbf{x})\big\} \\
&= \gamma\big(\partial g(\mathbf{x}) + \big\{\nabla f(\mathbf{x})\big\}\big) \\
&= \gamma\partial(f+g)(\mathbf{x}),
\end{aligned}$$

since $\gamma \geq 0$ (this justifies that $\partial(\gamma g)(\mathbf{x}) = \gamma\partial g(\mathbf{x})$). And so

$$\mathbf{0} \in \partial(f+g)(\mathbf{x}).$$

Therefore, the reciprocal of Fermat's rule ensures that $\mathbf{x}$ is a minimizer of $f+g$.

Exercise 3

1. Cauchy-Schwarz inequality and the assumption on f give the desired inequality:

$$\big\langle\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\big\rangle \leq \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|\|\mathbf{y} - \mathbf{x}\| \leq L\|\mathbf{y} - \mathbf{x}\|^2.$$

2. It is clear that, by definition, $\varphi(0) = f(\mathbf{x})$ and $\varphi(1) = f(\mathbf{y})$. We must now compute the derivative of φ at $t = 0$. Notice that $\varphi = f \circ A$, where A is the affine map from $[0,1]$ to $\mathbb{R}^n$ defined by $A(t) = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$. The chain rule gives:

$$\begin{aligned}
\forall h \in \mathbb{R},\ \varphi'(t)h &= \mathrm{d}\varphi(t).h \\
&= \mathrm{d}f\big(A(t)\big).\big(\mathrm{d}A(t).h\big) \\
&= \mathrm{d}f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})).\big(h(\mathbf{y} - \mathbf{x})\big) \\
&= h\mathrm{d}f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})).(\mathbf{y} - \mathbf{x}) \\
&= h\big\langle\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x}\big\rangle,
\end{aligned}$$

   by linearity of the differential. Simplifying by h and taking $t = 0$, we get $\varphi'(0) = \big\langle\nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\big\rangle$ as expected.

3. By the fundamental theorem of calculus applied to φ, we have:

$$\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\mathrm{d}t = \int_0^1 \big\langle\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x}\big\rangle\mathrm{d}t.$$

   Since $\varphi'(0) = \int_0^1 \big\langle\nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\big\rangle$ thanks to the previous question, we get the result.

4. It suffices to use the inequality of the first question in the integral of the previous question to obtain that

$$f(\mathbf{y}) - f(\mathbf{x}) - \big\langle\nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\big\rangle \leq L\|\mathbf{y} - \mathbf{x}\|^2 \int_0^1 \frac{t^2}{t}\mathrm{d}t = \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Convex optimization for machine learning – TD3

Olivier Fercoq
Bruno Costacèque

January 2022

1 Exercises

Exercise 1. (Proximal gradient for logistic regression)

We consider a classification problem defined by observations $(\mathbf{x}_i, y_i)_{1\leq i\leq n}$ where for all i, $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$. The coordinates of $\mathbf{x}_i$ are the explanatory variables of the model. We note $x_{i,j}$ the i-th observation of the j-th explanatory variable, so that $\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,p})$.

We propose the following linear model for the generation of the data. Each observation is supposed to be independent and there exists a vector $\mathbf{w} \in \mathbb{R}^p$ and $w_0 \in \mathbb{R}$ such that for all i, $(y_i, \mathbf{x}_i)$ is a realization of the random variable $(Y, \mathbf{X})$ whose law satisfies

$$\mathbb{P}_{w_0,\mathbf{w}}(Y = 1|\mathbf{X}) = \frac{\exp(\mathbf{X}^\top \mathbf{w} + w_0)}{1 + \exp(\mathbf{X}^\top \mathbf{w} + w_0)} \,.$$

1. In this model, what are the explanatory variables? The observed outcomes? The parameters? Can you think of a real-life situation to which you could apply this model?

2. We set $\sigma := x \mapsto (1 + \exp(-x))^{-1}$; this function is known as the *logistic function*.

   Show that $\forall i \in [\![1, n]\!]$, $\mathbb{P}_{w_0,\mathbf{w}}(Y_i = y_i|\mathbf{x}_i) = \dfrac{1}{1 + \exp(-y_i(\mathbf{x}_i^\top \mathbf{w} + w_0))} = \sigma\big(y_i(\mathbf{x}_i^\top \mathbf{w} + w_0)\big)$.

3. Show that the maximum likelihood estimator is

$$(\widehat{w_0}, \widehat{\mathbf{w}}) = \arg\min_{w_0,\mathbf{w}} \sum_{i=1}^{n} \log\Big(1 + \exp\big(-y_i(\mathbf{x}_i^\top \mathbf{w} + w_0)\big)\Big)$$

4. Denote $f(w_0, \mathbf{w}) = \sum_{i=1}^n \log\Big(1 + \exp\big(-y_i(\mathbf{x}_i^\top \mathbf{w} + w_0)\big)\Big)$. Compute $\nabla f(w_0, \mathbf{w})$.

5. Compute the proximal operator of $\mathbf{x} \mapsto \frac{\lambda}{2}\|\mathbf{x}\|^2$.

6. Write the proximal gradient method for the logistic regression problem with ridge regularizer

$$(\widehat{w_0}^\lambda, \widehat{\mathbf{w}}^\lambda) = \arg\min_{w_0,\mathbf{w}} \sum_{i=1}^{n} \log\Big(1 + \exp\big(-y_i(\mathbf{x}_i^\top \mathbf{w} + w_0)\big)\Big) + \frac{\lambda}{2}\|\mathbf{w}\|^2.$$

7. Compute the Hessian matrix and write Newton's method for the same problem.

Exercise 2. (LASSO)

We consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1 \,.$$

1. Prove that the solution is $\{0\}$ for large λ.

2. For an arbitrary λ, provide the expression of the proximal gradient algorithm, using the step size $\gamma_k := 1/L$.

3. Assume that the initial point is at distance D from a minimizer. How many iterations are needed (at most) to achieve an ε-minimizer?

2 Solutions

Exercise 1 1. The explanatory variables are the $\mathbf{x}_i$, and the observations are the y_i. The parameters are $\mathbf{w}$ and w_0. This model is usually applied to solve binary prediction problems. For example, to predict if a patient will have diabetes ($y = 1$) or not ($y = -1$) according to explanatory variables $\mathbf{x} = (x^1, \cdots, x^p)$, like his weight, his age, his job, *etc.* It could also be used to predict if someone will vote for a certain party or not, according to various social and economic indicators. This model is closer from classification methods than true regression.

2.
- Assume first $y_i = 1$. By multiplying the numerator and the denominator of $\mathbb{P}_{w_0,\mathbf{w}}(Y = y_i|\mathbf{X})$ by $\exp(-(\mathbf{X}^\top\mathbf{w} + w_0))$, we get:

$$\mathbb{P}_{w_0,\mathbf{w}}(Y = y_i|\mathbf{X}) = \frac{\exp(\mathbf{X}^\top\mathbf{w} + w_0)}{1 + \exp(\mathbf{X}^\top\mathbf{w} + w_0)} = \frac{1}{1 + \exp(-(\mathbf{X}^\top\mathbf{w} + w_0))} = \frac{1}{1 + \exp(-y_i(\mathbf{X}^\top\mathbf{w} + w_0))}.$$

- Assume now that $y_i = -1$. Then:

$$\mathbb{P}_{w_0,\mathbf{w}}(Y = y_i|\mathbf{X}) = 1 - \mathbb{P}_{w_0,\mathbf{w}}(Y = 1|\mathbf{X}) = \frac{1}{1 + \exp(\mathbf{X}^\top\mathbf{w} + w_0)} = \frac{1}{1 + \exp(-y_i(\mathbf{X}^\top\mathbf{w} + w_0))}.$$

3. Remember that the likelihood L of a discrete probability distribution μ_θ in the observations $\mathbf{y}_1, \cdots, \mathbf{y}_n$ and with parameter θ is the probability of observing all these observations at the same time, assuming they are independent:

$$L(\theta, \mathbf{y}_1, \cdots, \mathbf{y}_n) = \mathbb{P}(Y_1 = y_1, \cdots, Y_n = y_n) = \prod_{i=1}^{n} \mathbb{P}(Y_1 = y_i) = \prod_{i=1}^{n} \mu_\theta(\{y_i\}),$$

where the Y_i are i.i.d. random vectors with distribution μ_θ. In particular they have the same distribution as Y_1. Furthermore the maximizers of a positive function f are the same as those of $\log \circ f$, so we will focus on the log-likelihood ℓ rather than on the likelihood itself, since it is easier to deal with. Here $\theta = (w_0, \mathbf{w})$ and we want to find $(\widehat{\mathbf{w}}, \widehat{w}_0)$ which are the likeliest parameters of the model, owing to the available data (observations + explanatory variables), that is, those which maximize:

$$\ell(\widehat{\mathbf{w}}, \widehat{w}_0, \mathbf{x}_1, \cdots, \mathbf{x}_n) = -\sum_{i=1}^{n} \log\Big(1 + \exp\big(-y_i(\mathbf{x}_i^\top\widehat{\mathbf{w}} + \widehat{w}_0)\big)\Big).$$

Because of the minus sign, this means finding the minimizers of:

$$f : (w_0, \mathbf{w}) \mapsto \sum_{i=1}^{n} \log\Big(1 + \exp\big(-y_i(\mathbf{x}_i^\top\mathbf{w} + w_0)\big)\Big),$$

and so the maximum likelihood estimator(s) of $(w_0, \mathbf{w})$ are the points $(\widehat{\mathbf{w}}, \widehat{w}_0)$ which minimize this function.

4. It suffices to compute the partial derivatives with respect to the $p + 1$ variables $w_0, w_1, \cdots, w_p$. We find:

$$\nabla f(w_0, \mathbf{w}) = \begin{pmatrix} -\sum_{i=1}^{n} y_i \sigma\big(-y_i(\mathbf{x}_i^\top\mathbf{w} + w_0)\big) \\ -\sum_{i=1}^{n} x_{i1} y_i \sigma\big(-y_i(\mathbf{x}_i^\top\mathbf{w} + w_0)\big) \\ \vdots \\ -\sum_{i=1}^{n} x_{ip} y_i \sigma\big(-y_i(\mathbf{x}_i^\top\mathbf{w} + w_0)\big) \end{pmatrix},$$

5. By definition, computing the proximal operator of $g = \mathbf{x} \mapsto \frac{\lambda}{2}\|\mathbf{x}\|^2$ means to find the minimizer[1] of:

$$h(\mathbf{x}) := \frac{\lambda}{2}\|\mathbf{x}\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2 = \frac{\lambda+1}{2}\|\mathbf{x}\|^2 - \langle \mathbf{x}, \mathbf{y}\rangle + \frac{1}{2}\|\mathbf{y}\|^2.$$

This function is convex (as a sum of such functions) and its gradient is:

$$\nabla h(\mathbf{x}) = (\lambda+1)\mathbf{x} - \mathbf{y}.$$

As a result, its unique minimizer, *i.e.* the proximal operator of g, is

$$\mathrm{prox}_g(\mathbf{y}) = \frac{1}{\lambda+1}\mathbf{y}.$$

6. By definition, the recurrence relation given by the proximal gradient method for the logistic regression problem with given step sequence $(\gamma_k)_k$ is:

$$\begin{aligned}(w_0^{k+1}, \mathbf{w}^{k+1}) &= \mathrm{prox}_{\gamma_k g}\big((w_0^k, \mathbf{w}^k) - \gamma_k \nabla f(w_0^k, \mathbf{w}^k)\big) \\ &= \frac{1}{\lambda+1}\big((w_0^k, \mathbf{w}^k) - \gamma_k \nabla f(w_0^k, \mathbf{w}^k)\big).\end{aligned}$$

This result is consistent with the definition of the ridge regularization: when $\lambda = 0$, the penalization term vanishes and we find the usual gradient descent algorithm. And when $\lambda = +\infty$, the only value of $(\widehat{w_0}, \widehat{\mathbf{w}})$ which solves the logistic regression problem with ridge regularizer is $(\widehat{w_0}, \widehat{\mathbf{w}}) = (\mathbf{0}_{\mathbb{R}^p}, 0)$.

7. We must compute the cross partial derivatives $\partial_{i,j} f$ of the $\mathcal{C}^2$-class function $f : \mathbb{R}^{p+1} \to \mathbb{R}$. We set:

$$z_k := y_k(\mathbf{x}_k^\top \mathbf{w} + w_0).$$

Remember that the y_k are numbers equal to ± 1, so $y_k^2 = 1$. Also notice that $\sigma'(x) = e^{-x}\sigma(x)$, so that:

$$\frac{\mathrm{d}}{\mathrm{d}x}\sigma(-x) = -e^x \sigma(-x)^2.$$

- When $i = j = 1$, we differentiate the first term of the gradient with respect to the first variable. We find:

$$\partial_{1,1} f(w_0, \mathbf{w}) = \sum_{k=1}^{n} e^{z_k}\sigma(-z_k)^2.$$

[1] Which is unique, since g satisfies the hypotheses of the exercise 1 of the previous tutorial: g is convex and lower-semicontinuous with a nonempty domain.

- When $i \in [\![2, p+1]\!]$ and $j = 1$, we differentiate the i-th term of the gradient with respect to the first variable. We find:

$$\forall i \in [\![2, p+1]\!],\ \partial_{i,1} f(\mathbf{x}) = \sum_{k=1}^{n} x_{k,i} e^{z_k} \sigma(-z_k)^2.$$

- For $i, j \in [\![2, p+1]\!]$, we find:

$$\forall i, j \in [\![2, p+1]\!],\ \partial_{i,j} f(\mathbf{x}) = \sum_{k=1}^{n} x_{k,i} x_{k,j} e^{z_k} \sigma(-z_k)^2$$

Once the Hessian matrix is known, we must compute for each k the vector $(\nabla^2 f(\mathbf{x}_k))^{-1} \nabla f(\mathbf{x}_k)$. This means solving the system $\nabla^2 f(\mathbf{x}_k)\mathbf{x} = \nabla f(\mathbf{x}_k)$. In the present case there does not seem to exist a closed expression for the solution, so we have to resort to numerical solvers to find the solution $(\widehat{w_0}^{\lambda}, \widehat{\mathbf{w}}^{\lambda})$ of the logistic regression problem with ridge regularizer by using Newton's algorithm.

Exercise 2 1. We set:

$$f(\lambda, \mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1.$$

It is clear that:

$$\lim_{\lambda \to \infty} f(\lambda, \mathbf{x}) = \frac{1}{2}\|\mathbf{b}\|^2 + \iota_{\{\mathbf{0}\}}(\mathbf{x}) = \begin{cases} \frac{1}{2}\|\mathbf{b}\|^2 & \text{if } \mathbf{x} = \mathbf{0} \\ +\infty & \text{if } \mathbf{x} \neq \mathbf{0} \end{cases}.$$

But $f(\lambda, \mathbf{0}) = \frac{1}{2}\|\mathbf{b}\|_2^2$ for all $\lambda \geq 0$, so intuitively, as soon as λ is large enough, the minimum of $f(\lambda, \cdot)$ should be reached in $\mathbf{0}$, since $f(\lambda, \mathbf{x})$ tends to infinity when λ goes to infinity if $\mathbf{x}$ is not zero. The trouble is that for each fixed λ, the minimizer of $f(\lambda, \cdot)$ depends on λ, so it is not clear how it will behave when λ goes to infinity.

To understand that, we need a uniform control on the minimizer (as a function of λ). It will be provided by the norm of $A^\top b$. First notice that we work in finite dimension, so that all norms are equivalent. In particular, there exists $c > 0$ such that $\|\mathbf{x}\|_1 \geq c\|\mathbf{x}\|_2$. Thus:

$$\begin{aligned} f(\lambda, \mathbf{x}) - f(\lambda, \mathbf{0}) &= \frac{1}{2}\|A\mathbf{x} - b\|_2^2 + \lambda\|\mathbf{x}\|_1 - \frac{1}{2}\|\mathbf{b}\|_2^2 \\ &= \frac{1}{2}\|A\mathbf{x}\|_2^2 - \langle A\mathbf{x}, \mathbf{b}\rangle + \lambda\|\mathbf{x}\|_1 \\ &\geq -\langle A\mathbf{x}, \mathbf{b}\rangle + \lambda\|\mathbf{x}\|_1 \\ &\geq -\langle A\mathbf{x}, \mathbf{b}\rangle + c\lambda\|\mathbf{x}\|_2 \\ &= \|\mathbf{x}\|_2\big(-\frac{\langle A\mathbf{x}, \mathbf{b}\rangle}{\|\mathbf{x}\|_2} + c\lambda\big) \\ &= \|\mathbf{x}\|_2\big(-\frac{\langle \mathbf{x}, A^\top\mathbf{b}\rangle}{\|\mathbf{x}\|_2} + c\lambda\big) \\ &\geq \|\mathbf{x}\|_2\big(-\|A^\top\mathbf{b}\|_2 + c\lambda\big) \end{aligned} \tag{1}$$

and the term between parentheses is positive as soon as $\lambda \geq \|A^\top \mathbf{b}\|_2/c$. For such λ, $\mathbf{0}$ is always a minimizer of $f(\lambda, \cdot)$. It is the unique possible minimizer thanks to equation (1) (if $\mathbf{x}$ is a minimizer of $f(\lambda, \cdot)$ and is not null, then so is its norm and $f(\lambda, \mathbf{x}) - f(\lambda, \mathbf{0}) > 0$. Absurd.

2. Using the notations of the lecture notes, we take $f = \mathbf{x} \mapsto \|A\mathbf{x} - \mathbf{b}\|_2$ and $g = \mathbf{x} \mapsto \|\mathbf{x}\|_1$. It is clear that both functions are convex, and that f is differentiable, with a Lipschitz gradient:

$$\nabla f(\mathbf{x}) = A^\top (A\mathbf{x} - \mathbf{b})$$

(see the first lecture!). In that particular case, we can take the matrix norm of $A^\top A$ as a Lipschitz constant, since:

$$\begin{aligned} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 &= \|A^\top A\mathbf{x} - A^\top A\mathbf{y}\|_2 \\ &\leq \sup_{\mathbf{z} \neq 0} \frac{\|A^\top A\mathbf{z}\|_2}{\|\mathbf{z}\|_2} \|\mathbf{x} - \mathbf{y}\|_2 \\ &= \|A^\top A\|_{2,2} \|\mathbf{x} - \mathbf{y}\|_2. \end{aligned}$$

Here the proximal operator is a bit more complicated to compute than usual, since we must solve the optimization problem:

$$\text{For all } \mathbf{y} \in \mathbb{R}^d, \text{ find } \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} \Big\{ \lambda \|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \Big\}.$$

Notice that the proximal operator will be a function from $\mathbb{R}^n$ to $\mathbb{R}^n$. Notice also that:

$$\lambda \|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 = \sum_{i=1}^n \big(\lambda |x_i\| + \frac{1}{2}(x_i - y_i)^2 \big) = \sum_{i=1}^n f_i(\lambda, x_i),$$

with obvious notations. This means that we have to minimize a sum of functions with separate variables. In that case, it is obvious we just have to minimize each function separately, *i.e.* the solution $\mathbf{x} = (x_1, \ldots, x_n)$ will have coordinates x_i which minimize f_i for each i. So it is enough to compute the proximal operator of $x \mapsto |x|$. This is not hard, but the computation itself is a bit tedious because of the numerous cases to distinguish... Set $g : x \mapsto \lambda|x| + (x - y)^2/2$. We know that x minimizes g iff $0 \in \partial g(x)$. Since:

$$\partial g(x) = \begin{cases} \operatorname{sign}(x)\lambda + x - y & \text{if } x \neq 0 \\ [-\lambda, \lambda] - y & \text{if } x = 0. \end{cases}$$

we must distinguish two cases (I gladly acknowledge that it is not obvious at all just by looking at that expression of the subgradient!):

- **If** $|y| > \lambda$: then 0 cannot belong to $\partial g(0)$, so the minimizer of g must be some $x \neq 0$. If $x \in \mathbb{R}_+^*$, we have that

$$0 \in \partial g(x) \iff x = y - \lambda,$$

which is possible only if $y > \lambda$. Otherwise $y < -\lambda$, and we find for $x \in \mathbb{R}_-^*$:

$$0 \in \partial g(x) \iff x = y + \lambda,$$

so, by reuniting those two cases:

$$x = y + \operatorname{sign}(y)\lambda.$$

- **If** $|y| \leq \lambda$, then $0 \in [-\lambda, \lambda] - y = \partial g(x)$, and 0 is a minimizer of g.

Once more we can reunite all previous cases in one expression, just by factorising by $\operatorname{sign}(y)$:

$$x = \operatorname{sign}(y)\big(|y| - \lambda\big)_+,$$

where $y_+ = \max(y, 0)$ is the *positive part* of y. This function is called *soft thresholding operator* (in one dimension). More generally, the proximal operator of $\mathbf{x} \mapsto \lambda\|\mathbf{x}\|_1$ is just:

$$\operatorname{prox}_{\lambda\|\cdot\|_1}(\mathbf{y}) = \Big(\operatorname{sign}(y_1)\big(|y_1| - \lambda\big)_+, \ldots, \operatorname{sign}(y_n)\big(|y_n| - \lambda\big)_+\Big).$$

Knowing this, it is an easy matter to write the proximal gradient algorithm for $f(\lambda, \cdot)$ (be careful: it is necessary to replace λ by λ/L in the previous expression!).

3. Using the course, we know that:

$$f(\lambda, \mathbf{x}_k) - f(\lambda, \mathbf{x}^*) \leq \frac{L\|\mathbf{x}_k - \mathbf{x}^*\|_2^2}{2k} = \frac{LD^2}{2k},$$

where $L = \|A\|_{2,2}$. By definition, an ε-minimizer of $f(\lambda, \cdot)$ is a point $\mathbf{x}$ such that $f(\lambda, \mathbf{x})$ is at distance at most ε from the minimum $f(\lambda, \mathbf{x}^*)$. So we just need to find k such that:

$$\frac{LD^2}{2k} \leq \varepsilon.$$

With the bound provided by the course, we find that we need at least $k = LD^2/(2\varepsilon)+1$ iterations to be sure that $\mathbf{x}_k$ is an ε-minimizer for the LASSO problem.

Remarque 1.
- *This is a sufficient condition, not a necessary one: we may have already an ε-minimizer with less iterations, hence the need to get bounds as sharp as possible to avoid redundant iterations!*
- *It does not mean that $\mathbf{x}_k$ is actually close from $\mathbf{x}^*$! To determine that, stronger assumptions may be required (for example, strong convexity (see your lecture notes)).*

Convex optimization for machine learning – TD4

Olivier Fercoq
Bruno Costacèque

January 2022

1 Exercises

Exercise 1. (Projected stochastic gradient)
We consider the following optimization problem

$$\min_{\mathbf{x}\in C} \sum_{i=1}^{n} f_i(\mathbf{x}) \tag{1}$$

where $C = [0,1]^d$ and for all i, $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable.

1. We define the convex indicator function of the set C as $\iota_C(x) = \begin{cases} 0 & \text{if } \mathbf{x} \in C \\ +\infty & \text{if } \mathbf{x} \notin C \end{cases}$.
   Show that (1) is equivalent to $\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{n} f_i(\mathbf{x}) + \iota_C(\mathbf{x})$.
2. Compute the proximal operator of ι_C
3. Write the proximal stochastic gradient method for the resolution of (1).

Exercise 2. (Optimisation with explicit constraints)
We consider the following optimization problem

$$\min_{\mathbf{x}\in C} f(\mathbf{x}) \tag{2}$$

where $C \subset \mathbb{R}^d$ is a convex set and $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex.

1. We define the convex indicator function of the set C as
   $$\iota_C(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in C \\ +\infty & \text{if } \mathbf{x} \notin C \end{cases}$$
   Show that (2) is equivalent to
   $$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) + \iota_C(\mathbf{x}) \tag{3}$$
2. Show that for all $\mathbf{x} \in C$, $\partial\iota_C(\mathbf{x}) = \{\mathbf{q} \in \mathbb{R}^n \ : \ \forall \mathbf{y} \in C, \langle \mathbf{q}, \mathbf{y} - \mathbf{x}\rangle \leq 0\}$ and that $\partial\iota_C(\mathbf{x})$ is a cone (it is called the *normal cone* to C at $\mathbf{x}$). Show that for all $\mathbf{x} \notin C$, $\partial\iota_C(\mathbf{x}) = \emptyset$.
3. Show that $\mathbf{x}^*$ is a solution to (3) if and only if $-\nabla f(\mathbf{x}^*) \in \partial\iota_c(\mathbf{x}^*)$.
4. Denote $\mathcal{H}_{\mathbf{w},b} = \{\mathbf{x} \in \mathcal{X} : \langle \mathbf{w}, \mathbf{x}\rangle + b = 0\}$. Compute $\partial\iota_{\mathcal{H}_{\mathbf{w},b}}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$.

   (Hint: Show that we can assume $b = 0$ without loss of generality, and use the fact that $\mathcal{H}_{\mathbf{w},0}$ is a vector space.)
5. Prove that the distance of a point $\mathbf{z}$ to $\mathcal{H}$ is equal to

$$\mathrm{d}(\mathbf{z}, \mathcal{H}_{\mathbf{w},b}) = \min_{\mathbf{x}\in\mathcal{H}_{\mathbf{w},b}} \|\mathbf{x} - \mathbf{z}\|_2 = \frac{|\langle \mathbf{w}, \mathbf{z}\rangle + b|}{\|\mathbf{w}\|_2}\,.$$

2 Solutions

Exercise 1 1. It is obvious that $\mathbf{x} \mapsto \sum_{i=1}^n f_i(\mathbf{x}) + \iota_C(\mathbf{x})$ can admit a minimum only on C (the f_i are differentiable, so in particular they take finite values). On the other hand, if $\mathbf{x} \in C$, this function is the same as $\mathbf{x} \mapsto \sum_{i=1}^n f_i(\mathbf{x})$. Thus the constrained minimization problem is the same as the unconstrained one with the convex indicator function.

2. It is clear that:

$$\mathrm{prox}_{\iota_C}(\mathbf{x}) = \underset{\mathbf{y}\in\mathbb{R}^d}{\mathrm{argmin}}\Big\{\iota_C(\mathbf{y}) + \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2\Big\} = \underset{\mathbf{y}\in C}{\mathrm{argmin}}\ \|\mathbf{x}-\mathbf{y}\|^2.$$

And it is well known that the point minimizing this distance (to the closed convex set $C = [0,1]^d$) is the projection of $\mathbf{x}$ on C :

$$\mathrm{prox}_{\iota_C}(\mathbf{x}) = \mathrm{proj}_C(\mathbf{x}).$$

3. Notice that:

$$\sum_{i=1}^n f_i(\mathbf{x}) = n\mathbb{E}[f_I(\mathbf{x})] = \mathbb{E}[nf_I(\mathbf{x})],$$

where I is a random variable with uniform distribution on $[\![1,n]\!]$. In other words, we choose at random function among the n functions f_i, we evaluate it at $\mathbf{x}$, and thus define a random function. Then (1) is equivalent to minimizing n times the expectation of this random function with respect to $\mathbf{x}$.

In this context, the ξ_k (using the notations of the lecture) will have the uniform distribution $\mathcal{U}_{[\![1,n]\!]}$. So we will apply the stochastic proximal gradient method to $(\mathbf{x},\xi_k) \mapsto nf_{\xi_k}(\mathbf{x}) + \iota_C(\mathbf{x})$. It is just the usual proximal gradient algorithm, but applied to a uniformly chosen function nf_i, with a penalization function given by[1] $g = \iota_C$. Finally we find the following algorithm:

$$\mathbf{x}_{k+1} = \mathrm{proj}_C\big(\mathbf{x}_k - \gamma_k n\nabla f_{\xi_k}(\mathbf{x}_k)\big),$$

where γ_k is a sequence of step sizes.

Exercise 2 1. It is the same line of reasoning as in the first question of the previous exercise.

2. • For each $\mathbf{x} \in \mathbb{R}^d$, we must find all vectors $\mathbf{q} \in \mathbb{R}^d$ such that:

$$\forall \mathbf{y} \in \mathbb{R}^d,\ \iota_C(\mathbf{y}) \geq \iota_C(\mathbf{x}) + \langle \mathbf{q}, \mathbf{y}-\mathbf{x}\rangle. \tag{4}$$

If $\mathbf{x} \in C$, then the only non-trivial case is when $\mathbf{y} \in C$ too. So $\mathbf{q}$ must satisfy:

$$0 \geq 0 + \langle \mathbf{q}, \mathbf{y}-\mathbf{x}\rangle = \langle \mathbf{q}, \mathbf{y}-\mathbf{x}\rangle.$$

This condition is clearly both necessary and sufficient, so:

$$\partial\iota_C(\mathbf{x}) = \{\mathbf{q} \in \mathbb{R}^d \mid \langle \mathbf{q}, \mathbf{y}-\mathbf{x}\rangle \leq 0\}.$$

Furthermore this set is indeed a cone: if $q \in \partial\iota_C(\mathbf{x})$, then it is evident that $\lambda\mathbf{q} \in C$ if $\lambda \geq 0$.

[1] Notice that g is convex (because C is convex) and l.s.c. (because C is closed). Also $\forall\gamma > 0,\ \gamma g = g$!

- And if $\mathbf{x} \notin C$, (4) is obviously impossible since $\iota_C(\mathbf{y}) = 0$ if $\mathbf{y} \in C$, while $\iota_C(\mathbf{y}) = +\infty$. Thus, for $\mathbf{x} \notin C$, $\partial \iota_C(\mathbf{x}) = \emptyset$.

3. A solution $\mathbf{x}^*$ necessarily belongs to C thanks to the first question. Thus the subgradient of ι_C at $\mathbf{x}^*$ exists. And since f is differentiable, Fermat's rule gives that:

$$\mathbf{x}^* \text{is a solution of (3)} \iff \mathbf{0}_{\mathbb{R}^d} \in \partial(f + \iota_C)(\mathbf{x}^*) = \{\nabla f(\mathbf{x}^*)\} + \{\mathbf{q} \in \mathbb{R}^d \mid \langle \mathbf{q}, \mathbf{y} - \mathbf{x} \rangle \leq 0\}.$$

Hence, there is $\mathbf{q} \in \partial \iota_C(\mathbf{x}^*)$ such that:

$$\mathbf{0}_{\mathbb{R}^d} = \nabla f(\mathbf{x}^*) + \mathbf{q},$$

and so $-\nabla f(\mathbf{x}^*) \in \partial \iota_C(\mathbf{x}^*)$.

4. For each $\mathbf{x} \in \mathcal{H}_{\mathbf{w},b}$, we want to find all $\mathbf{q} \in \mathbb{R}^d$ such that:

$$\forall \mathbf{y} \in \mathcal{H}_{\mathbf{w},b},\ \langle \mathbf{q}, \mathbf{y} - \mathbf{x} \rangle \leq 0. \tag{5}$$

Let $\mathbf{x}_0 \in \mathcal{H}_{\mathbf{w},b}$ and notice that:

$$\langle \mathbf{q}, \mathbf{y} - \mathbf{x} \rangle = \big\langle \mathbf{q}, (\mathbf{y} - \mathbf{x}_0) - (\mathbf{x}_0 - \mathbf{x}) \big\rangle.$$

Since $\mathbf{y} - \mathbf{x}_0$ and $\mathbf{x}_0 - \mathbf{x}$ belong to $\mathcal{H}_{\mathbf{w},b}$, we may assume that $\mathbf{x}$ and $\mathbf{y}$ belong to the vector space $\mathcal{H}_{\mathbf{w},0} = \mathrm{Span}(\mathbf{w})^\perp$. But then $\lambda \mathbf{y} \in \mathcal{H}_{\mathbf{w},0}$ too for all $\lambda \in \mathbb{R}$. In particular the inequality (5) reads:

$$\forall \mathbf{y} \in \mathcal{H}_{\mathbf{w},b},\ \forall \lambda \in \mathbb{R},\ \lambda \langle \mathbf{q}, \mathbf{y} \rangle \leq \langle \mathbf{q}, \mathbf{x} \rangle,$$

which is possible only if $\langle \mathbf{q}, \mathbf{y} \rangle = 0$ for all $\mathbf{y} \in \mathcal{H}_{\mathbf{w},0}$. On the other hand, it is easy to see that $\lambda \mathbf{w} \in \partial \iota_{\mathcal{H}_{\mathbf{w},b}}$ for all $\lambda \in \mathbb{R}$. As a result:

$$\partial \iota_{\mathcal{H}_{\mathbf{w},b}} = \partial \iota_{\mathcal{H}_{\mathbf{w},0}} = \mathrm{Span}(\mathbf{w})^{\perp\perp} = \mathrm{Span}(\mathbf{w}).$$

5. Here the function to minimize[2] is $f_{\mathbf{z}} : \mathbf{x} \mapsto \|\mathbf{x} - \mathbf{z}\|_2^2$, and the constraint is that $\mathbf{x}$ must belong to the convex set $\mathcal{H}_{\mathbf{w},b}$. So we can apply question 3 to find the minimizer. We must find a point $\mathbf{x}^*$ such that $-\nabla f(\mathbf{x}^*) = -2(\mathbf{x}^* - \mathbf{z}) \in \partial \iota_C(\mathbf{x}^*) = \mathrm{Span}(\mathbf{w})$. That is, for some $\lambda \in \mathbb{R}$,

$$\mathbf{x}^* = \mathbf{z} - \frac{\lambda}{2}\mathbf{w}.$$

To find λ, remember that $\mathbf{x}^*$ must belong to $C = \mathcal{H}_{\mathbf{w},b}$, so we must have:

$$-b = \langle \mathbf{w}, \mathbf{x}^* \rangle = \langle \mathbf{w}, \mathbf{z} \rangle - \frac{\lambda}{2}\|\mathbf{w}\|_2^2,$$

and so $\lambda = 2(b + \langle \mathbf{w}, \mathbf{x}^* \rangle)/\|\mathbf{w}\|_2^2$. Thus:

$$\mathbf{x}^* = \mathbf{z} - \frac{b + \langle \mathbf{w}, \mathbf{x}^* \rangle}{\|\mathbf{w}\|_2^2}\mathbf{w}.$$

Evaluating $\mathbf{x} \mapsto \|\mathbf{x} - \mathbf{z}\|$ in $\mathbf{x}^*$, we find the desired result.

[2] Notice the square! Without it, this function is not differentiable at $\mathbf{z}$...

Convex optimization for machine learning – TD5

Olivier Fercoq
Bruno Costacèque

January 2022

1 Exercises

Exercise 1. (Support vector machine)

Consider a training set formed by couples $(\mathbf{x}_i, y_i)$ for $i \in \{1, \dots, n\}$ where $\mathbf{x}_i$ is a feature vector in $\mathcal{X}$ and $y_i \in \{-1, +1\}$ for all i. The (affine) hyperplane $\mathcal{H}_{\mathbf{w},b} := \{\mathbf{x} \in \mathbb{R}^d,\ \langle \mathbf{w}, \mathbf{x}\rangle + b = 0\}$ is called *separating* if

$$\forall i, \quad y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) > 0\,.$$

In the sequel, we assume that a separating hyperplane exists. Among all separating hyperplanes, we seek to find the one which maximizes the minimum distance

$$f(\mathbf{w}, b) = \min_{i=1,\dots,n} d(\mathbf{x}_i, \mathcal{H}_{\mathbf{w},b})\,.$$

1. Show that if $(\mathbf{w}, b)$ defines a separating hyperplane, then[1] $f(\mathbf{w}, b) = c(\mathbf{w}, b)/\|\mathbf{w}\|$ where $c(\mathbf{w}, b) = \min_i y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b)$.

Thus, we are interested in solving the problem

$$\max_{\mathbf{w},b} \frac{c(\mathbf{w}, b)}{\|\mathbf{w}\|} \text{ such that } \forall i,\ y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 0\,. \tag{1}$$

Let $(\mathbf{w}^*, b^*)$ be a solution and define

$$\mathbf{v}^* = \frac{\mathbf{w}^*}{c(\mathbf{w}^*, b^*)} \text{ and } \quad a^* = \frac{b^*}{c(\mathbf{w}^*, b^*)}$$

2. Justify that $(\mathbf{w}^*, b^*)$ and $(\mathbf{v}^*, a^*)$ define the same separating hyperplane.

3. Prove that $(\mathbf{v}^*, a^*)$ solves the optimization problem

$$\max_{\mathbf{v},a} \frac{1}{\|\mathbf{v}\|} \text{ such that } \forall i,\ y_i(\langle \mathbf{v}, \mathbf{x}_i\rangle + a) \geq 1\,.$$

4. Deduce that $(\mathbf{v}^*, a^*)$ solves the optimization problem

$$\min_{\mathbf{v},a} \frac{\|\mathbf{v}\|^2}{2} \text{ such that } \forall i,\ 1 - y_i(\langle \mathbf{v}, \mathbf{x}_i\rangle + a) \leq 0\,. \tag{2}$$

What is the purpose of this question? (in comparison to the previous one for example)

5. Write the Lagrangian[2] $L(\mathbf{v}, a; \phi)$.

[1] Remember the previous tutorial where we proved (or not) that:

$$\mathrm{d}(\mathbf{z}, \mathcal{H}_{\mathbf{w},b}) = \min_{\mathbf{x} \in \mathcal{H}_{\mathbf{w},b}} \|\mathbf{x} - \mathbf{z}\| = \frac{|\langle \mathbf{w}, \mathbf{z}\rangle + b|}{\|\mathbf{w}\|}\,.$$

[2] Notice the ';' which separates the primal variables $(\mathbf{v}, a)$ from the dual variable ϕ.

6. Write the KKT conditions.

7. Let $(\mathbf{v}, a; \phi)$ be a saddle point of the Lagrangian. Show that ϕ_i is non-zero only if $y_i(\langle \mathbf{v}, \mathbf{x}_i \rangle + a) = 1$.

The training points $(\mathbf{x}_i, y_i)$ satisfying the above property are the closest to the hyperplane $\mathcal{H}_{\mathbf{v},a}$. The corresponding $\mathbf{x}_i$'s are often called *support vectors*.

8. If one is given a dual solution $\phi^\star$, how to recover a primal solution $(\mathbf{v}^*, a^*)$ from $\phi^\star$?

Define the $n \times n$ matrices $K = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle)_{i,j=1\ldots n}$, $D = \mathrm{diag}(y_1 \ldots y_n)$ and $\mathbf{1}^T = (1, \ldots, 1)$.

9. Prove that the dual problem reduces to

$$\min_{\substack{\phi \geq 0 \\ \langle \mathbf{y}, \phi \rangle = 0}} \frac{1}{2} \langle KD\phi, D\phi \rangle - \langle \mathbf{1}, \phi \rangle .$$

10. Assume that this algorithm has identified a dual solution $\phi^\star$. Write explicitely the classifier as a function of $\phi^\star$.

11. What part of the training data do you need in order to implement the above classifier?

2 Solutions

Exercise 1 1. Using the recalled identity in the footnote, we get that:

$$f(\mathbf{w}, b) = \min_{i=1,\dots,n} \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} = \min_{i=1,\dots,n} \frac{|y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)|}{\|\mathbf{w}\|} = \min_{i=1,\dots,n} y_i \frac{\langle \mathbf{w}, \mathbf{x}_i \rangle + b}{\|\mathbf{w}\|},$$

since we know that all $y_i\langle \mathbf{w}, \mathbf{x}_i \rangle + b$ are positive and $y_i = \pm 1$.

2. We have to prove that $\mathcal{H}_{\mathbf{v}^*,a^*} = \mathcal{H}_{\mathbf{w}^*,b^*}$. But if $\mathbf{x} \in \mathcal{H}_{\mathbf{w}^*,b^*}$, then:

$$\langle \mathbf{v}^*, \mathbf{x} \rangle + a^* = \frac{\langle \mathbf{w}^*, \mathbf{x} \rangle}{c(\mathbf{w}^*, b^*)} + \frac{b^*}{c(\mathbf{w}^*, b^*)} = 0$$

by definition of $\mathcal{H}_{\mathbf{w}^*,b^*}$, and so $\mathcal{H}_{\mathbf{w}^*,b^*} \subset \mathcal{H}_{\mathbf{v}^*,a^*}$. The reciprocal is obvious, so we get the desired result.

3. Since $(\mathbf{w}^*, b^*)$ is a solution of (1), and:

$$\frac{c(\mathbf{w}, b)}{\|\mathbf{w}\|} = \frac{1}{\left\| \frac{\mathbf{w}}{c(\mathbf{w},b)} \right\|},$$

we see that replacing $\mathbf{w}^*$ by $\mathbf{v}^* = \mathbf{w}^*/c(\mathbf{w}^*, b^*)$ and b^* by $a^* = b^*/c(\mathbf{w}^*, b^*)$ yields that $(\mathbf{v}^*, a^*)$ satisfies:

$$\frac{1}{\|\mathbf{v}^*\|} = \max_{\mathbf{v},a} \frac{1}{\|\mathbf{v}\|},$$

with the constraint that

$$\forall j,\ y_j(\langle \mathbf{v}^*, \mathbf{x}_j \rangle + a^*) = \frac{1}{c(\mathbf{w}^*, b^*)} y_j(\langle \mathbf{w}^*, \mathbf{x}_j \rangle + b^*) \geq \frac{1}{c(\mathbf{w}^*, b^*)} \min_i\ y_j(\langle \mathbf{w}^*, \mathbf{x}_j \rangle + b^*) = 1.$$

4. It suffices to notice that $h : x \mapsto 1/(2x^2)$ is a non-increasing function which turns $1/\|\mathbf{v}\|$ in $\|\mathbf{v}\|^2/2$, and the maximization problem into a minimization one:

$$\min_{\mathbf{v},a} \frac{\|\mathbf{v}\|^2}{2} = h\Big(\max_{\mathbf{v},a} \frac{1}{\|\mathbf{v}\|}\Big)$$

The constraint does not change. The main purpose of this question is to replace the non-differentiable function $\mathbf{v} \mapsto 1/\|\mathbf{v}\|$ by $\mathbf{v} \mapsto \|\mathbf{v}\|^2/2$, which will be much more pleasant to work with. Notice also that normalizing by $c(\mathbf{w}^*, b^*)$ allows us to replace the strict inequalities $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 0$ by $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + a) \geq 1$.

5. With the notations of the lecture, the function f is $\mathbf{v} \mapsto \|\mathbf{v}\|^2/2$ while g is:

$$\mathbf{x} \mapsto \big(g_1(\mathbf{x}), \dots, g_n(\mathbf{x})\big), \text{ where } g_i(\mathbf{x}) = 1 - y_i(\langle \mathbf{v}, \mathbf{x}_i \rangle + a),$$

and the associated Laplacian equals:

$$L(\mathbf{v}, a; \phi) = \frac{1}{2}\|\mathbf{v}\|^2 + \langle \phi, g(\mathbf{x}) \rangle - \iota_{\mathbb{R}^n_+}(\phi).$$

6. There are four conditions which together are known as the KKT conditions and must be satisfied by $(\mathbf{v}^*, a^*, \phi^*)$ to be a saddle point of the Lagrangian. First we give them as they are formulated in the lecture notes (see theorem 6.3.2. page 35), and then we rewrite them in a more useful way. Let $\mathbf{y} := (y_1, \dots, y_n)$. Remember that in the course x denotes the primal variable(s). Here $x = (\mathbf{v}, a)$. Also remember that for fixed ϕ such that $\phi_i \leq 0$ for all i (and so $\iota_{\mathbb{R}^n_-}(\phi) = 0$), we know that $L(\cdot, \cdot; \phi)$ is differentiable, so we can use the usual gradient instead of the subdifferential.

   - *Primal feasibility:* $g(\mathbf{v}^*, a^*) \preceq \mathbf{0}_{\mathbf{R}^n}$, that is, $\forall i \in [\![1, n]\!],\ g_i(\mathbf{v}^*, a^*) \leq 0$.
   - *Dual feasibility:* $\phi^* \succeq \mathbf{0}_{\mathbb{R}^n}$, that is, $\forall i \in [\![1, n]\!],\ \phi_i^* \leq 0$.
   - *Complementary slackness:* $\langle \phi^*, g(\mathbf{v}^*, a^*) \rangle = 0$.
   - *First order primal feasibility:*

$$\nabla_{(\mathbf{v},a)} L(\mathbf{v}^*, a^*; \phi^*) = \mathbf{0}_{\mathbb{R}^{n+1}}.$$

   In the sequel, it will be more convenient to work with those expressions, which are exactly the previous conditions, but formulated differently:

   - *Primal feasibility:* $\forall i \in [\![1, n]\!],\ y_i(\langle \mathbf{v}^*, \mathbf{x}_i \rangle + a^*) \geq 1$.
   - *Dual feasibility:* $\forall i \in [\![1, n]\!],\ \phi_i^* \geq 0$.
   - *Complementary slackness:* $\forall i \in [\![1, n]\!],\ g_i(\mathbf{v}^*) = 1 - y_i(\langle \mathbf{v}^*, \mathbf{x}_i \rangle + a^*) = 0$ or $\phi_i^* = 0$.
   - *First order primal feasibility:*

$$\nabla_{\mathbf{v}} L(\mathbf{v}^*, a^*; \phi^*) = \mathbf{v}^* - \sum_{i=1}^{n} y_i \phi_i^* \mathbf{x}_i^* = \mathbf{0}_{\mathbb{R}^n},$$

$$\partial_a L(\mathbf{v}^*, a^*; \phi^*) = -\langle \mathbf{y}, \phi^* \rangle = 0.$$

   The notation $\nabla_{(\mathbf{v},a)} L(\mathbf{v}, a; \phi)$ denotes the gradient of the Lagrangian with respect to $\mathbf{v}$ and a. It outputs a vector of $\mathbb{R}^{n+1}$ which formally reads as:

$$\nabla_{(\mathbf{v},a)} L(\mathbf{v}, a; \phi) = \begin{pmatrix} \nabla_{\mathbf{v}} L(\mathbf{v}, a, \phi) \\ \partial_a L(\mathbf{v}, a) \end{pmatrix}$$

   The reformulated third condition is a consequence of the original complementary slackness condition, the original primal feasibility condition, as well as of the dual feasibility condition: we know that $\langle \phi^*, g(\mathbf{v}^*, a^*) \rangle = 0$, that $\forall i \in [\![1, n]\!],\ g_i(\mathbf{v}^*, a^*) \leq 0$ and $\phi_i \leq 0$. So we get that a sum of nonnegative terms vanishes. Thus all terms must be equal to zero, which gives the reformulated complementary slackness condition.

7. This is a direct consequence of the complementary slackness condition: $\phi^* \neq 0$ so

$$\forall i \in [\![1, n]\!],\ y_i(\langle \mathbf{v}^*, \mathbf{x}_i \rangle + a^*) = 1.$$

8. Let us explain what this question means exactly: assume that you want to solve a complicated optimization problem. Maybe the dual problem is easier, so you manage to find its solution ϕ^*. Now you want to deduce from that $\mathbf{x}^* = (\mathbf{v}^*, a^*)$ the solution of your original problem, the primal problem. But you know that $(\mathbf{v}^*, a^*; \phi^*)$ is a saddle point of your Lagrangian. So KKT conditions will help you express the unknown $(\mathbf{v}^*, a^*)$ in terms of what you know, that is, ϕ^*.

   The first equation of the first order primal feasibility condition yields:

$$\mathbf{v}^* = \sum_{i=1}^{n} y_i \phi_i^* \mathbf{x}_i^*$$

(and the right-hand side does not depend on $\mathbf{v}^*$ nor a^*!). And using the previous question and the fact that $y_i^2 = 1$, we can recover a^* as well:

$$\forall i \in [\![1, n]\!] \text{ such that } \phi_i^* \neq 0,\ a^* = y_i - \langle \mathbf{v}^*, \mathbf{x}_i \rangle,$$

If this condition is never satisfied, that is, if all $\phi_i^* = 0$, then $\mathbf{v}^* = \mathbf{0}_{\mathbb{R}^n}$ and by the primal feasibility condition we have:

$$\forall i \in [\![1, n]\!],\ y_i a^* \geq 1,$$

which is possible only if every $y_i = 1$ or $y_i = -1$ for all i. Thus as long the two classes of the classification problem have at least one element each, we can find one index i_0 such that $\phi_{i_0} \neq 0$. As a result, using the previous expression of $\mathbf{v}^*$, we can take:

$$a^* = y_{i_0} - \langle \mathbf{v}^*, \mathbf{x}_{i_0} \rangle = y_{i_0} - \sum_{i=1}^{n} y_i \phi_i^* \langle \mathbf{x}_i, \mathbf{x}_{i_0}^* \rangle.$$

9. The dual problem is found by keeping the dual variables feasible[3] and minimizing the Lagrangian with respect to the primal variables. As before, the first-order conditions are:

$$\nabla_v L(\mathbf{v}^*, a^*; \phi) = \mathbf{v}^* - \sum_{i=1}^{n} y_i \phi_i \mathbf{x}_i^* = \mathbf{0}_{\mathbb{R}^n},$$

$$\partial_a L(\mathbf{v}^*, a^*; \phi) = -\langle \mathbf{y}, \phi \rangle = 0.$$

Substituting this expression of $\mathbf{v}^*$ in the definition of L, we get:

$$\begin{aligned}
L(\mathbf{v}^*, a^*; \phi) &= \frac{1}{2}\|\mathbf{v}^*\|^2 + \langle \phi, g(\mathbf{x}) \rangle + \iota_{\mathbb{R}_+^n}(\phi) \\
&= \frac{1}{2}\langle \mathbf{v}^*, \mathbf{v}^* \rangle + \langle \phi, g(\mathbf{x}) \rangle \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \phi_i \phi_j^* \langle \mathbf{x}_i^*, \mathbf{x}_j^* \rangle + \langle \phi, g(\mathbf{x}) \rangle \\
&= \frac{1}{2}\langle KD\phi, D\phi \rangle + \sum_{i=1}^{n} \phi_i \big(1 - y_i(\langle \mathbf{v}^*, \mathbf{x}_i \rangle + a^*)\big) \\
&= \frac{1}{2}\langle KD\phi, D\phi \rangle + \langle \phi, \mathbf{1} \rangle - \sum_{i=1}^{n} \phi_i y_i (\langle \mathbf{v}^*, \mathbf{x}_i \rangle + a^*) \\
&= \frac{1}{2}\langle KD\phi, D\phi \rangle + \langle \phi, \mathbf{1} \rangle - \sum_{i=1}^{n} \phi_i y_i \langle \mathbf{v}^*, \mathbf{x}_i \rangle - a^* \langle \phi, \mathbf{y} \rangle \\
&= \frac{1}{2}\langle KD\phi, D\phi \rangle + \langle \phi, \mathbf{1} \rangle - \langle KD\phi, D\phi \rangle \\
&= -\frac{1}{2}\langle KD\phi, D\phi \rangle + \langle \phi, \mathbf{1} \rangle,
\end{aligned}$$

and so the dual problem consists in maximizing $-\frac{1}{2}\langle KD\phi, D\phi \rangle + \langle \phi, \mathbf{1} \rangle$ under the constraints that $\phi_i \geq 0$ for all i and $\langle \phi, \mathbf{y} \rangle = 0$.

[3] That is, $\forall i \in [\![1, n]\!], \phi_i \geq 0$.

10. The classifier takes the following form, where a^* has been defined in the previous questions:

$$C : \mathbf{x} \mapsto \text{sign}(\langle \mathbf{v}^*, \mathbf{x} \rangle + a^*).$$

11. The only necessary data is the support vectors, that is the vectors $\mathbf{x}_i$ such that $\phi_i > 0$.