Convex Optimization
Take home exam part II

Antonio Coín Castro

April 12, 2021

Exercise 1. *Prove Jensen's inequality: if f is convex on $\mathbb{R}^d$ and $\sum_1^k \lambda_i = 1$ with $0 \leq \lambda_i \leq 1$, we have for any $x_1, \dots, x_k \in \mathbb{R}^d$*

$$f\left(\sum_{i=1}^k \lambda_i x_i\right) \leq \sum_{i=1}^k \lambda_i f(x_i).$$

Proof. First of all, note that for $k = 1$ there is nothing to prove, since necessarily $\lambda_1 = 1$, and for $k = 2$ the claim is just the definition of a convex function. For $k = 3$, observe that if we write $1 - \lambda_1 = \lambda_2 + \lambda_3$, and provided that $\lambda_2 + \lambda_3 > 0$, we have:

$$\begin{aligned}
\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 &= \lambda_1 x_1 + (\lambda_2 + \lambda_3)\frac{\lambda_2 x_2}{\lambda_2 + \lambda_3} + (\lambda_2 + \lambda_3)\frac{\lambda_3 x_3}{\lambda_2 + \lambda_3} \\
&= \lambda_1 x_1 + (1 - \lambda_1)\mu x_2 + (1 - \lambda_1)(1 - \mu) x_3 \qquad \left[\mu := \lambda_2/(\lambda_2 + \lambda_3)\right] \\
&= \lambda_1 x_1 + (1 - \lambda_1)\left(\mu x_2 + (1 - \mu) x_3\right).
\end{aligned}$$

Note that we can assume without loss of generality that $\lambda_2 + \lambda_3 > 0$, since otherwise $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = 0$, and the claim would be a triviality. Now, observing that $\mu \in [0, 1]$, applying f repeatedly to the preceding expression and using its convexity, we have:

$$\begin{aligned}
f\left(\sum_{i=1}^3 \lambda_i x_i\right) &= f\left(\lambda_1 x_1 + (1 - \lambda_1)\left(\mu x_2 + (1 - \mu) x_3\right)\right) \\
&\leq \lambda_1 f(x_1) + (1 - \lambda_1) f(\mu x_2 + (1 - \mu) x_3) \\
&\leq \lambda_1 f(x_1) + (1 - \lambda_1)\left(\mu f(x_2) + (1 - \mu) f(x_3)\right) \\
&= \sum_{i=1}^3 \lambda_i f(x_i).
\end{aligned}$$

The process carried out above for $k = 3$ gives us the idea on how to proceed by induction on k. Indeed, suppose that the claim holds for k, and consider now $k + 1$ points. Then, it is immediate to see that proceeding as before we can write $1 - \lambda_1 = \sum_2^{k+1} \lambda_i > 0$, so that

$$0 \leq \frac{\lambda_i}{1 - \lambda_1} \leq 1 \quad \text{and} \quad \sum_{i=2}^{k+1} \frac{\lambda_i}{1 - \lambda_1} = 1.$$

Then, it follows that

$$
\begin{aligned}
f\left(\sum_{i=1}^{k+1}\lambda_i x_i\right) &= f\left(\lambda_1 x_1 + (1-\lambda_1)\left(\sum_{i=2}^{k+1}\frac{\lambda_i}{1-\lambda_1}x_i\right)\right) \\
&\leq \lambda_1 f(x_1) + (1-\lambda_1) f\left(\sum_{i=2}^{k+1}\frac{\lambda_i}{1-\lambda_1}x_i\right) \\
&\leq \lambda_1 f(x_1) + (1-\lambda_1)\sum_{i=2}^{k+1}\frac{\lambda_i}{1-\lambda_1}f(x_i) && \text{[Induction hypothesis]} \\
&= \sum_{i=1}^{k+1}\lambda_i f(x_i). && \square
\end{aligned}
$$

Exercise 2. *Rewrite the Lagrange multiplier analysis in the lecture slides under the assumption that from $g(x,y)=0$ we can find instead a function $x=h(y)$ such that $g(h(y),y)=0$.*

Solution. Consider $f, g : \mathbb{R}^2 \to \mathbb{R}$ and the minimization problem

$$\min f(x,y) \quad \text{s.t} \quad g(x,y)=0.$$

We will henceforth only work in the feasible set $C=\{(x,y)\in\mathbb{R}^2 : g(x,y)=0\}$. Now suppose that we can apply the *implicit function theorem* to the first component of g (i.e. g is continuously differentiable and $\partial g/\partial x \neq 0$) to find a continuously differentiable function $x=h(y)$ such that $g(h(y),y)=0$. Thus, we can write

$$f(x,y) = f(h(y),y) = \Psi(y).$$

Now, it follows from the chain rule that Ψ is differentiable, and thus at a minimum y^* with $x^* = h(y^*)$ we have

$$0 = \Psi'(y^*) = \frac{\partial f}{\partial x}(x^*,y^*)h'(y^*) + \frac{\partial f}{\partial y}(x^*,y^*). \tag{1}$$

But since $g(h(y),y)$ is also differentiable and $g(h(y),y)=0$, we have

$$0 = \frac{\partial g}{\partial x}(x^*,y^*)h'(y^*) + \frac{\partial g}{\partial y}(x^*,y^*) \implies h'(y^*) = -\frac{\frac{\partial g}{\partial y}(x^*,y^*)}{\frac{\partial g}{\partial x}(x^*,y^*)}. \tag{2}$$

Putting together (1) and (2) we arrive at

$$\frac{\partial f}{\partial y}(x^*,y^*)\frac{\partial g}{\partial x}(x^*,y^*) - \frac{\partial f}{\partial x}(x^*,y^*)\frac{\partial g}{\partial y}(x^*,y^*) = 0.$$

The above expression can be reinterpreted as $\nabla f \perp \left(-\frac{\partial g}{\partial y}, \frac{\partial g}{\partial x}\right)$ at (x^*,y^*), and since by definition $\left(-\frac{\partial g}{\partial y}, \frac{\partial g}{\partial x}\right) \perp \nabla g$, we have $\nabla f \parallel \nabla g$ at (x^*,y^*), that is, they are proportional:

$$\nabla f(x^*,y^*) = -\lambda^* \nabla g(x^*,y^*) \text{ for some } \lambda^* \neq 0.$$

From this line of reasoning we can conclude that for the *Lagrangian*

$$\mathcal{L}(x,y,\lambda) = f(x,y) + \lambda g(x,y)$$

we have that at a minimum (x^*, y^*) there exists a $\lambda^* \neq 0$ such that

$$\nabla_{x,y} \mathcal{L}(x^*, y^*, \lambda^*) = \nabla f(x^*, y^*) + \lambda^* \nabla g(x^*, y^*) = 0. \tag{3}$$

Thus, a way of solving our original minimization problem is to define its Lagrangian and solve simultaneously (3) and the constraint $g(x, y) = \partial\mathcal{L}/\partial\lambda = 0$. Summing up, we have to solve three equations in three unknowns, namely $\nabla\mathcal{L} = 0$. As we can see, the conclusion remains the same regardless of the component (x or y) of the constraint g for which we solve in the beginning.

Exercise 3. *Using the lecture slides, write down in as much detail as possible the computations needed at each iteration of the Projected Gradient algorithm to solve the constrained Ridge problem.*

Solution. First of all, consider the constrained Ridge problem

$$\min_{w,b} \frac{1}{2} \operatorname{mse}(w, b) = \frac{1}{2n} \sum_{p=1}^{n} (y_p - w^T x_p - b)^2 \quad \text{s.t.} \quad \|w\|_2 \leq \rho.$$

By considering $\tilde{w} = (b, w^T)^T$ and $\tilde{x}_p = (1, x_p^T)^T$ we can reformulate the problem as

$$\min_{\tilde{w} \in C} f(\tilde{w}) = \frac{1}{2n} \|y - \tilde{X}\tilde{w}\|^2,$$

where $C = \mathbb{R} \times \bar{B}(0, \rho)$, $\tilde{X}$ is the $n \times (d+1)$ data matrix with the $\tilde{x}_p^T$ as rows and y is the target vector $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$. Note that C is a closed non-empty convex set, because it is the product of two such sets, and since f is convex and $\mathscr{C}^1$, we can apply the *Projected Gradient* algorithm to iteratively find a solution. This algorithm is a particular case of the Proximal Gradient algorithm in which we consider a sort of indicator function of the constraint set C acting as g. Specifically, it is easy to see that the above problem is equivalent to solving

$$\min_{\tilde{w} \in \mathbb{R}^{d+1}} f(\tilde{w}) + i_C(\tilde{w}), \quad \text{where} \quad i_C(\tilde{w}) = \begin{cases} 0 & \text{if } \tilde{w} \in C, \\ +\infty & \text{if } \tilde{w} \notin C. \end{cases}$$

The original algorithm would require us to compute $\operatorname{prox}_{\lambda i_C}(\tilde{w})$, but it turns out that in this case this is the same as computing $P_C(\tilde{w})$, where P_C is the (Euclidean) projection operator

$$P_C(x) = \arg\min_{y \in C} \|x - y\|_2.$$

The Projected Gradient algorithm to solve our problem is outlined below.

Algorithm 1: Projected Gradient.

Data: $\epsilon, \tilde{w}_0$
k=0
for $k = 1, 2, \ldots$ **do**
  Choose a step size γ_k
  $\tilde{w}_{k+1} = P_C(\tilde{w}_k - \gamma_k \nabla f(\tilde{w}_k))$ ▷ *Gradient update*
  if $\|\tilde{w}_{k+1} - \tilde{w}_k\| \leq \epsilon$ **then** ▷ *Stopping condition*
    return $\tilde{w}_{k+1}$
  end
end

The general idea of this method is to perform the usual unconstrained gradient descent update and then project back onto C to ensure that the constraints are not violated. First of all, we need to choose an initial point $\tilde{w}_0 = (b_0, w_0^T)^T \in C$ (possibly at random) and an application-dependent stopping sensitivity $\epsilon > 0$, which will control how near our successive solutions will have to be in order to stop the algorithm and claim convergence.

Then, after choosing a suitable step size (more on that later) we perform the gradient update and projection at each iteration. In our case, the projection onto the set C is easy to compute, since in the first component we don't do anything, and in the second component we project onto an Euclidean ball (the points already inside the ball are obviously left untouched):

$$P_C(\tilde{w}) = P_C(b, w) = \left(b, \frac{\rho w}{\max\{\rho, \|w\|\}} \right).$$

The only remaining thing to do is to compute the gradient $\nabla f(\tilde{w})$:

$$\nabla f(\tilde{w})^T = \left(\frac{\partial f(b,w)}{\partial b}, \nabla_w f(b,w)^T \right) = -\frac{1}{n} \sum_{p=1}^{n} \tilde{x}_p^T (y_p - w^T x_p - b) = -\frac{1}{n} \left(\mathbf{1}^T (y - \tilde{X}\tilde{w}),\ X^T (y - \tilde{X}\tilde{w}) \right),$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones. All in all, the gradient update step for a point $\tilde{w}_k = (b_k, w_k^T)^T$ amounts to computing

$$\tilde{w}_{k+1} = P_C(\tilde{w}_k - \gamma_k \nabla f(\tilde{w}_k)) = \left(b_k + \frac{\gamma_k}{n} \mathbf{1}^T (y - \tilde{X}\tilde{w}),\ \rho \frac{w_k + \frac{\gamma_k}{n} X^T (y - \tilde{X}\tilde{w})}{\max\left\{ \rho, \|w_k + \frac{\gamma_k}{n} X^T (y - \tilde{X}\tilde{w})\| \right\}} \right)^T.$$

Note that we could have assumed X and y to be centered and work with $b = 0$, simplifying the calculations a bit.

Finally, the choice of a step size γ_k determines the convergence rate of the method. For example[1], if ∇f is $L-$Lipschitz, we know that if we choose a constant step $\gamma_k = \frac{1}{L}$ we can achieve a convergence rate of $\mathcal{O}(1/k)$. As a matter of fact, this is the case with our function, and for instance in the case $b = 0$ it is quite straightforward to derive a Lipschitz constant:

$$\|\nabla f(w_1) - \nabla f(w_2)\|_2 = \|X^T X (w_1 - w_2)\|_2 \leq \|X^T X\| \|w_1 - w_2\|_2 \implies L \leq \|X^T X\|,$$

where $\|\cdot\|$ is any matrix norm compatible with the Euclidean vector norm. As an alternative approach, we could also use cross-validation techniques to find good values of γ_k.

Exercise 4. *We want to apply Lagrangian theory to solve the homogeneous constrained Ridge problem*

$$\arg\min_w \text{mse}(w) = \frac{1}{n} \sum_{p=1}^{n} (y_p - w^T x_p)^2 \quad \text{s.t.} \quad \|w\|_2^2 \leq \rho^2.$$

a) Write down its Lagrangian and the detailed formulation of the KKT conditions at an optimal solution w^* with multiplier λ^*.

[1]Other conditions and suggestions for a step size are presented in Calamai, P. H., & Moré, J. J. (1987). Projected gradient methods for linearly constrained problems. *Mathematical programming*, 39(1), 93-116.

Solution. For convenience, define $f(w) = \text{mse}(w)/2 = \|y - Xw\|^2/2n$ and $g(w) = (\|w\|^2 - \rho^2)/2$, where $y = (y_1, \ldots, y_n)^T$ and X is the data matrix with the x_p^T as rows. Our constrained minimization problem can then equivalently be written as:

$$\min_w f(w) \quad \text{s.t} \quad g(w) \leq 0.$$

First of all, the Lagrangian is

$$\mathcal{L}(w, \lambda) = \frac{1}{2n} \sum_{p=1}^{n} (y_p - w^T x_p)^2 + \frac{\lambda}{2}(\|w\|^2 - \rho^2).$$

Since f and g are convex and $\mathscr{C}^1$, the KKT conditions are necessary and sufficient for an optimal solution. We know that in general they can be expressed as

$$\begin{aligned} \nabla f(w^*) + \lambda^* \nabla g(w^*) &= 0, \\ \lambda^* g(w^*) &= 0, \end{aligned}$$

so in our particular case they are

$$-\frac{1}{n} X^T (y - Xw^*) + \lambda^* w^* = 0, \tag{4}$$

$$\lambda^* (\|w^*\|^2 - \rho^2) = 0. \tag{5}$$

b) Assuming that $\lambda^* > 0$, use the gradient KKT conditions to show that w^* also solves a standard Ridge regression problem for the optimal value λ^* of the regularization parameter.

Solution. Consider the standard Ridge regression problem with regularization parameter $\lambda > 0$:

$$\arg\min_w R(w) = \frac{1}{2n}\|y - Xw\|^2 + \frac{\lambda}{2}\|w\|^2.$$

Since this problem is unconstrained, convex and differentiable, we can find the minimum by setting $\nabla R(w) = 0$ and solving for w:

$$0 = \nabla R(w) = -\frac{1}{n} X^T (y - Xw) + \lambda w. \tag{6}$$

Now we see by looking at the gradient KKT condition (4) that if $\lambda = \lambda^*$, then w^* verifies (6) and thus is the solution of the standard Ridge regression problem with regularization parameter λ^*. Moreover, the complementary slackness condition (5) implies that $\|w^*\|^2 = \rho^2$.

In this case we can even provide the explicit optimal solution:

$$w^* = \frac{1}{n} \left(\frac{1}{n} X^T X + \lambda^* I \right)^{-1} X^T y.$$

Note that $n^{-1} X^T X + \lambda^* I$ is non-singular even though the positive semidefinite matrix $n^{-1} X^T X$ might be singular, because the addition of a positive constant in the diagonal shifts the eigenvalues away from zero.

c) Assuming now that $\lambda^* = 0$, write down the optimal solution in this case and use it to get a lower bound for ρ.

Solution. If $\lambda^* = 0$, condition (4) implies that

$$(X^T X)w^* = X^T y. \tag{7}$$

If $X^T X$ is non-singular (i.e. X has linearly independent columns) we can solve for w^* and write the optimal solution in closed form:

$$w^* = (X^T X)^{-1} X^T y.$$

Now, considering any matrix norm compatible with the Euclidean norm (for example the one induced by it), taking norms in (7) yields

$$\|X^T y\| = \|(X^T X)w^*\| \leq \|X^T X\|\|w^*\|.$$

Finally, since the optimal solution must obey the constraint $\|w^*\|^2 \leq \rho^2$ and we are supposing that $\rho > 0$, we have $\|w^*\| \leq \rho$. Plugging this information in the above expression we arrive at our desired bound for ρ:

$$\|X^T y\| \leq \|X^T X\|\rho \implies \rho \geq \frac{\|X^T y\|}{\|X^T X\|}.$$

Exercise 5. *The primal soft margin L_2 SVC problem with labels $\{+1, -1\}$ is*[2]

$$\arg\min_{w,b,\xi} \frac{1}{2}\|w\|_2^2 + \frac{C}{2}\sum_{p=1}^{n} \xi_p^2 \quad \text{s.t.} \quad \begin{cases} y_p(w^T x_p + b) \geq 1 - \xi_p, \\ 1 \leq p \leq n. \end{cases}$$

Write down its Lagrangian and the corresponding dual problem. Write also the associated KKT conditions and derive the optimal primal solution (w^*, b^*, ξ^*) from the optimal dual solution α^*.

Solution. The Lagrangian for this problem is

$$\begin{aligned} \mathcal{L}(w, b, \xi; \alpha) &= \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{p=1}^{n} \xi_p^2 - \sum_p \alpha_p [y_p(w^T x_p + b) - 1 + \xi_p] \\ &= w^T \left(\frac{1}{2}w - \sum_p \alpha_p y_p x_p\right) + \sum_p \xi_p \left(\frac{C}{2}\xi_p - \alpha_p\right) - b\sum_p \alpha_p y_p + \sum_p \alpha_p, \end{aligned}$$

with $\alpha_p \geq 0$. To get the dual function we solve $\nabla_w \mathcal{L} = 0$, $\frac{\partial \mathcal{L}}{\partial b} = 0$ and $\frac{\partial \mathcal{L}}{\partial \xi_p} = 0$, which yield the KKT stationarity conditions:

$$0 = \nabla_w \mathcal{L} = w - \sum_p \alpha_p y_p x_p \implies w = \sum_p \alpha_p y_p x_p. \tag{8}$$

$$0 = \frac{\partial \mathcal{L}}{\partial b} = \sum_p \alpha_p y_p. \tag{9}$$

$$0 = \frac{\partial \mathcal{L}}{\partial \xi_p} = 2\frac{C}{2}\xi_p - \alpha_p \implies C\xi_p = \alpha_p, \quad 1 \leq p \leq n. \tag{10}$$

[2] We considered $C/2$ instead of C to simplify the calculations.

Substituting back in the Lagrangian we arrive at the dual function

$$\begin{aligned}\Theta(\alpha) &= -\frac{1}{2}\sum_{p,q}\alpha_p\alpha_q y_p y_q x_p^T x_q - \frac{1}{2C}\sum_p \alpha_p^2 - b\overbrace{\sum_p \alpha_p y_p}^{0} + \sum_p \alpha_p \\ &= -\frac{1}{2}\sum_{p,q}\alpha_p\alpha_q\left(y_p y_q x_p^T x_q + \frac{\delta_{pq}}{C}\right) + \sum_p \alpha_p \\ &= -\frac{1}{2}\alpha^T\left(Q + \frac{I}{C}\right)\alpha + \sum_p \alpha_p,\end{aligned}$$

with constraints $\sum_p \alpha_p y_p = 0$ and $\alpha_p \geq 0$. In the above expression, Q is the symmetric matrix given by $Q_{pq} = y_p y_q x_p^T x_q$, I is the identity matrix, $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, and δ_{pq} is Kronecker's delta function, defined to be 1 if $p = q$ and 0 otherwise. We know that strong duality holds in this case, so flipping the sign to get a minimization problem, the dual problem has the following expression:

$$\min_{\alpha\in\mathbb{R}^n}\left\{\frac{1}{2}\alpha^T\left(Q + \frac{I}{C}\right)\alpha - \sum_p \alpha_p\right\} \quad \text{s.t.} \quad \begin{cases}\sum_p \alpha_p y_p = 0, \\ \alpha_p \geq 0, \\ 1 \leq p \leq n.\end{cases}$$

The KKT complementary slackness conditions in this case are

$$\alpha_p^*[y_p((w^*)^T x_p + b^*) - 1 + \xi_p^*] = 0, \quad 1 \leq p \leq n. \tag{11}$$

To derive the optimal primal solution from the dual one, we resort to the KKT conditions. Firstly, from (8) we have

$$w^* = \sum_{p=1}^n \alpha_p^* y_p x_p.$$

Secondly, from (10) we conclude that

$$\xi_p^* = \frac{\alpha_p^*}{C}, \quad 1 \leq p \leq n.$$

Now, observe that in a non-trivial classification setting where we have at least one positive and one negative example, at least one of the α_p^* is non-zero (otherwise we would have $y_p b^* \geq 1$ for all p and every example would be of the same class). If we take any $\alpha_p^* > 0$, the conditions in (11) tell us that

$$\begin{aligned}0 &= y_p((w^*)^T x_p + b^*) - 1 + \xi_p^* \\ &= y_p\left(\sum_{q=1}^n \alpha_q^* y_q x_q^T x_p + b^*\right) - 1 + \frac{\alpha_p^*}{C} \\ &= y_p\sum_{q=1}^n \alpha_q^* y_q\left(x_q^T x_p + \frac{\delta_{pq}}{C}\right) + y_p b^* - 1.\end{aligned}$$

Thus, solving for b^* we get the optimal primal bias term in closed form[3]:

$$b^* = \frac{1}{y_p} - \sum_{q=1}^{n} \alpha_q^* y_q \left(x_q^T x_p + \frac{\delta_{pq}}{C} \right) = y_p - \sum_{q=1}^{n} \alpha_q^* y_q \left(x_q^T x_p + \frac{\delta_{pq}}{C} \right).$$

[3] Although in theory the formula works for any support vector, to increase the stability of the method in practice and compensate for numerical errors it is advised to average the values of b^* obtained for each $\alpha_p^* > 0$.