Convex Optimization: Project #2

Due on Wednesday 22rd, 2023 at 23:59pm

Jacopo Peroni

Problem 1

By definition, we have that

$$h(z) = \inf_{t\in\mathbb{R}} \delta|t| + \frac{1}{2}(z-t)^2$$

Let's define the function

$$\tilde{h}_z(t) = \delta|t| + \frac{1}{2}(z-t)^2$$

This function, if restricted to $\mathbb{R}_+$ or $\mathbb{R}^-$ is differentiable and convex because is the sum of an affine and a convex function, so there exists a minimum.
In $\mathbb{R}_+$

$$\frac{d}{dt}\tilde{h}_z(t) = \delta - (z-t) = \delta - z + t = 0 \Longrightarrow t = z - \delta$$

That is an admissible solution if $z > \delta$, in this case, the minimum is

$$\tilde{h}_z(z-\delta) = \delta(z-\delta) + \frac{1}{2}(z-z+\delta)^2 = \delta z - \frac{1}{2}\delta^2$$

Otherwise ($z < \delta$) the minimum is obtained at the boundary. At $+\infty$ the function clearly explodes at infinity, so the minimum is obtained in $t = 0$

$$\tilde{h}_z(0) = \frac{1}{2}z^2$$

In $\mathbb{R}^-$

$$\frac{d}{dt}\tilde{h}_z(t) = -\delta - (z-t) = -\delta - z + t = 0 \Longrightarrow t = z + \delta$$

That is an admissible solution if $z < -\delta$, in this case, the minimum is

$$\tilde{h}_z(z+\delta) = \delta(z+\delta) + \frac{1}{2}(z+z+\delta)^2 = \delta z + \delta^2 + 2z^2 + 2z\delta + \frac{1}{2}\delta^2 = 3z\delta + \frac{3}{2}\delta^2 + 2z^2$$

Otherwise ($z > -\delta$) the minimum is obtained at the boundary. At $-\infty$ the function clearly explodes at infinity, so the minimum is obtained in $t = 0$

$$\tilde{h}_z(0) = \frac{1}{2}z^2$$

Putting the things together $h(z)$ is defined in this way

$$h(z) = \begin{cases} \frac{1}{2}z^2 & |z| \le \delta \\ \min\left(\delta z - \frac{1}{2}\delta^2, 3z\delta + \frac{3}{2}\delta^2 + 2z^2\right) & |z| > \delta \end{cases}$$

Thus

$$h(z) = \begin{cases} \frac{1}{2}z^2 & |z| \le \delta \\ \delta z - \frac{1}{2}\delta^2 & |z| > \delta \end{cases} = L_\delta(z)$$

To show the equivalence between the two problems it's useful to notice that

$$\inf_{t\in\mathbb{R}} f(t) + g(z-t) = \inf_{(u,v)\in\mathbb{R}\times\mathbb{R};\ u+v=z} f(u) + g(v) = \inf_{t\in\mathbb{R}} f(z-t) + g(t)$$

Proven this we can show the fundamental result on $L_\delta(z)$

$$L_\delta\left(\omega^T x_i + b - y_i\right) = \inf_{t\in\mathbb{R}} \delta|t| + \frac{1}{2}\left(\omega^T x_i + b - y_i - t\right)^2 = \inf_{t\in\mathbb{R}} \delta\left|\omega^T x_i + b - y_i - t\right| + \frac{1}{2}t^2$$

Thus

$$\underset{\omega\in\mathbb{R}^m, b\in\mathbb{R}}{\text{minimize}} \sum_{i=1}^{m} L_\delta\left(\omega^T x_i + b - y_i\right) + \frac{\rho}{2}\|\omega\|_2^2 =$$

$$= \underset{\omega\in\mathbb{R}^m, b\in\mathbb{R}}{\text{minimize}} \sum_{i=1}^{m} \inf_{t_i\in\mathbb{R}} \delta\left|\omega^T x_i + b - y_i - t_i\right| + \frac{1}{2}t_i^2 + \frac{\rho}{2}\|\omega\|_2^2 =$$

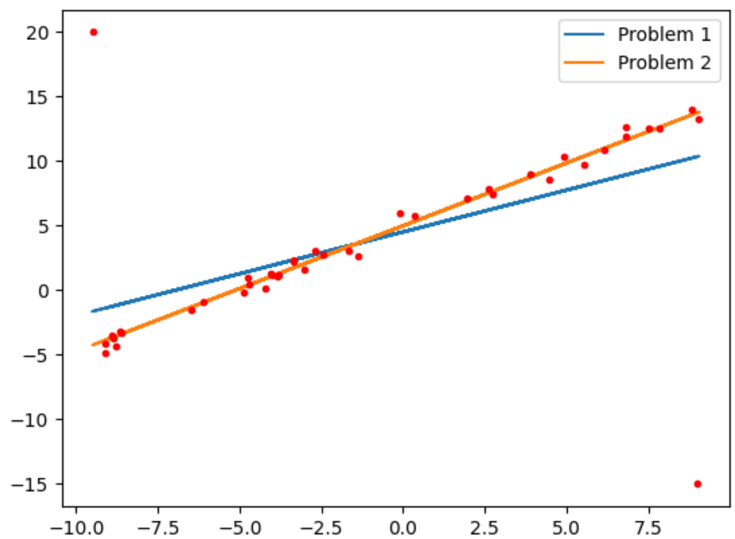
$$= \underset{\omega\in\mathbb{R}^m, b\in\mathbb{R}, t\in\mathbb{R}^m}{\text{minimize}} \frac{1}{2}\|t\|_2^2 + \sum_{i=1}^{m} \delta\left|\omega^T x_i + b - y_i - t_i\right| + \frac{\rho}{2}\|\omega\|_2^2$$

Adding two epigraphical variables for each index i, where we use two of them because the term $\omega^T x_i + b - y_i - t_i$ is minimized in its absolute value, thus the problem becomes

$$\begin{array}{ll} \underset{\omega\in\mathbb{R}^m, b\in\mathbb{R}, t\in\mathbb{R}^m, r^+\in\mathbb{R}_+^m, r^-\in\mathbb{R}_+^m}{\text{minimize}} & \frac{1}{2}\|t\|_2^2 + \sum_{i=1}^{m} \delta\mathbf{1}^T(r^+ + r^-) + \frac{\rho}{2}\|\omega\|_2^2 \\ \text{subject to} & \begin{array}{ll} \omega^T x_i + b - y_i - t_i \leq r_i^+ & \forall i \in \{1,\ldots,m\} \\ y_i - \omega^T x_i - b + t_i \leq r_i^- & \forall i \in \{1,\ldots,m\} \end{array} \end{array}$$

Problem 2

In the following picture, we can see how the two regressions, solutions of the two problems, react differently to outliers.
As we can see in the first problem, given $\delta = +\infty$ the function $L_\delta(z) = (1/2)z^2$, so it is more sensible to outliers.
Our data set has two outliers at the extremes, which leads the linear regression to have a different derivative if compared to the one given by problem 2.
Clearly, in problem 2, given the linear dependence of the points in the minimized function, the sensibility to outliers is less than in problem 1, as we can see from the picture, where the linear regression is closer to the one we would expect.



Problem 3

$$\begin{array}{ll} \underset{\omega\in\mathbb{R}^m, b\in\mathbb{R}, t\in\mathbb{R}^m, r^+\in\mathbb{R}^m_+, r^-\in\mathbb{R}^m_+}{\text{minimize}} & \frac{1}{2}\|t\|_2^2 + \sum_{i=1}^{m} \delta\mathbf{1}^T(r^+ + r^-) + \frac{\rho}{2}\|\omega\|_2^2 \\ \text{subject to} & \omega^T x_i + b - y_i - t_i - r_i^+ \leq 0 \quad \forall i \in \{1,\dots,m\} \\ & y_i - \omega^T x_i - b + t_i - r_i^- \leq 0 \quad \forall i \in \{1,\dots,m\} \end{array}$$

Subproblem 3.1

The Lagrangian of this problem is a map like this

$$L : \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^m_+ \times \mathbb{R}^m_+ \times \mathbb{R}^m_+ \times \mathbb{R}^m_+ \longrightarrow \mathbb{R}$$

Defined like follows

$$L(\omega, b, t, r^+, r^-, \lambda^+, \lambda^-) =$$

$$\frac{1}{2}\|t\|_2^2 + \delta\mathbf{1}^T\left(r^+ + r^-\right) + \frac{\rho}{2}\|\omega\|_2^2 + \sum_{i=1}^{m} \lambda_i^+ \left(\omega^T\phi\left(x_i\right) + b - y_i - t_i - r_i^+\right) +$$

$$+ \sum_{i=1}^{m} \lambda_i^- \left(y_i - \omega^T\phi\left(x_i\right) - b + t_i - r_i^-\right)$$

Subproblem 3.2

To state the dual problem we need to define the function $g(\lambda^+, \lambda^-)$ as

$$g\left(\lambda^+, \lambda^-\right) = \inf_{\omega\in\mathbb{R}^m, b\in\mathbb{R}, t\in\mathbb{R}^m, r^+\in\mathbb{R}^m_+, r^-\in\mathbb{R}^m_+} L(\omega, b, t, r^+, r^-, \lambda^+, \lambda^-)$$

To define better the function g let's rewrite L grouping the coefficients of the different variables together

$$L\left(\omega, b, t, r^+, r^-, \lambda^+, \lambda^-\right) =$$

$$= \omega^T\left(\frac{\rho}{2}\omega + \sum_{i=1}^{m}\left(\lambda_i^+ - \lambda_i^-\right)\phi\left(x_i\right)\right) + b\sum_{i=1}^{m}\left(\lambda_i^+ - \lambda_i^-\right) + t^T\left(\frac{1}{2}t + \lambda^+ - \lambda^-\right) +$$

$$+ {r^+}^T\left(\delta\mathbf{1} - \lambda^+\right) + {r^-}^T\left(\delta\mathbf{1} - \lambda^-\right) + y^T\left(\lambda^- - \lambda^+\right)$$

Using the change of variables $\beta_i = \lambda_i^- - \lambda_i^+ \ \forall i \in \{1,\dots,m\}$, we get

$$L\left(\omega, b, t, r^+, r^-, \lambda^+, \lambda^-, \beta\right) =$$

$$= \omega^T\left(\frac{\rho}{2}\omega - \sum_{i=1}^{m}\beta_i\phi\left(x_i\right)\right) - b\sum_{i=1}^{m}\beta_i + t^T\left(\frac{1}{2}t - \beta\right) + {r^+}^T\left(\delta\mathbf{1} - \lambda^+\right) + {r^-}^T\left(\delta\mathbf{1} - \lambda^-\right) + y^T\beta$$

Given that r^+ and r^- are in $\mathbb{R}^m_+$ we have that

$$\inf_{r^\pm \in \mathbb{R}^m_+} r^{\pm^T}\left(\delta \mathbf{1} - \lambda^\pm\right) = \begin{cases} 0 & \lambda^\pm \leq \delta \mathbf{1} \\ -\infty & \text{otherwise} \end{cases}$$

To rephrase this condition in terms of β

$$\beta = \lambda^- - \lambda^+$$

$$\lambda^- \geq 0 \Longrightarrow \beta \geq -\lambda^+ \geq -\delta \mathbf{1}$$

$$\lambda^+ \geq 0 \Longrightarrow \beta \leq \lambda^- \leq \delta \mathbf{1}$$

Thus we obtain that

$$-\delta \leq \beta_i \leq \delta \ \forall i \in \{1, \ldots, m\}$$

Given that b is unconstrained the problem is feasible only if

$$\sum_{i=1}^m \beta_i = 0$$

The condition of minimality with respect to ω is computed like this

$$\nabla_\omega L\left(\omega, b, t, r^+, r^-, \lambda^+, \lambda^-\right) = \rho\omega - \sum_{i=1}^m \beta_i \phi\left(x_i\right) = 0 \Longrightarrow \omega = \frac{1}{\rho} \sum_{i=1}^m \beta_i \phi\left(x_i\right)$$

Analogously for what concerns t

$$\nabla_t L\left(\omega, b, t, r^+, r^-, \lambda^+, \lambda^-\right) = t + \beta = 0 \Longrightarrow t = -\beta$$

Putting these results together in the definition of the Lagrangian we get

$$g(\beta) = \frac{1}{\rho} \sum_{i'=1}^m \beta_{i'} \phi\left(x_{i'}\right)^T \left(\frac{1}{2} \sum_{i=1}^m \beta_i \phi\left(x_i\right) - \sum_{i=1}^m \beta_i \phi\left(x_i\right)\right) - \beta^T \left(-\frac{1}{2}\beta + \beta\right) + y^T \beta =$$

$$= \sum_{i=1}^m y_i \beta_i - \frac{1}{2} \sum_{i=1}^m \beta_i^2 - \frac{1}{2\rho} \sum_{i=1}^m \sum_{i'=1}^m \beta_i \phi\left(x_i\right)^T \phi\left(x_{i'}\right) \beta_{i'}$$

Thus the dual problem is

$$\begin{aligned} \underset{\beta \in \mathbb{R}^m}{\text{maximize}} \quad & -\frac{1}{2} \sum_{i=1}^m \beta_i^2 - \frac{1}{2\rho} \sum_{i=1}^m \sum_{i'=1}^m \beta_i \phi\left(x_i\right)^T \phi\left(x_{i'}\right) \beta_{i'} + \sum_{i=1}^m y_i \beta_i \\ \text{subject to} \quad & \textstyle\sum_{i=1}^m \beta_i = 0 \\ & -\delta \leq \beta_i \leq \delta \quad \forall i \in \{1, \ldots, m\} \end{aligned}$$

Subproblem 3.3

If $(\omega^\star, b^\star, t^\star, r^{+^\star}, r^{-^\star}, \beta^\star)$ is optimal they have to follow KKT conditions, in particular the fourth one.
Rewriting the results of the calculations of the previous subproblem, we get

$$0 = \nabla_\omega L\left(\omega^\star, b^\star, t^\star, r^{+^\star}, r^{-^\star}, \beta^\star\right) = \rho\omega^\star - \sum_{i=1}^{m} \beta_i^\star \phi\left(x_i\right) \Longrightarrow$$

$$\Longrightarrow \omega_j^\star = \frac{1}{\rho}\sum_{i=1}^{m} \beta_i^\star \phi_j\left(x_i\right) \;\; \forall j \in \{1, \ldots, N\}$$

And

$$0 = \nabla_t L\left(\omega^\star, b^\star, t^\star, r^{+^\star}, r^{-^\star}, \beta^\star\right) = t^\star + \beta^\star \Longrightarrow t_i^\star = -\beta_i^\star \; \forall i \in \{1, \ldots, m\}$$

Subproblem 3.4

If we build the lagrangian for the dual problem we get

$$\tilde{L}\left(\beta, \tilde{\lambda}^+, \tilde{\lambda}^-, \mu\right) = -\frac{1}{2}\sum_{i=1}^{m} \beta_i^2 - \frac{1}{2\rho}\sum_{i=1}^{m}\sum_{i'=1}^{m} \beta_i \phi\left(x_i\right)^T \phi\left(x_{i'}\right)\beta_{i'} + \sum_{i=1}^{m} y_i\beta_i + \sum_{i=1}^{m} \tilde{\lambda}_i^+\left(\beta_i - \delta\right) +$$

$$-\sum_{i=1}^{m} \tilde{\lambda}_i^-\left(\beta_i + \delta\right) + \mu\sum_{i=1}^{m} \beta_i$$

If $(\beta^\star, \tilde{\lambda}^{+^\star}, \tilde{\lambda}^{-^\star}, \mu^\star)$ is optimal for this lagrangian, they have to follow KKT conditions, and so for the fourth KKT condition we get

$$0 = \nabla_\beta \tilde{L}\left(\beta^\star, \tilde{\lambda}^{+^\star}, \tilde{\lambda}^{-^\star}, \mu^\star\right) = -\beta^\star - \frac{1}{\rho}\kappa\beta^\star + y + \tilde{\lambda}^{+^\star} - \tilde{\lambda}^{-^\star} + \mu^\star\mathbf{1} \Longrightarrow$$

$$\Longrightarrow -\mu^\star = y - \beta^\star - \frac{1}{\rho}\kappa\beta^\star + \tilde{\lambda}^{+^\star} - \tilde{\lambda}^{-^\star}$$

Where κ is the kernel matrix.
Considering the third KKT condition we get $\forall i \in \{i, \ldots, m\}$

$$\tilde{\lambda}_i^{+^*}\left(\beta_i^\star - \delta\right) = 0$$

$$\tilde{\lambda}_i^{-^*}\left(\beta_i^\star + \delta\right) = 0$$

Thus for the $i \in \{i, \ldots, m\}$ such that $\beta_i^\star \in (-\delta, \delta)$, we get

$$\tilde{\lambda}_i^{+^*} = 0$$

$$\tilde{\lambda}_i^{-^*} = 0$$

So the condition on μ becomes

$$-\mu^\star = y - \beta^\star - \frac{1}{\rho}\kappa\beta^\star$$

From the definition of the inequality constraints in the primal problem, we can see that Slater's condition holds (we can fix $t = 0$, $\omega = \mathbf{1}$ and $r^+ > \max(y - x, 0)$ and $r^- > \max(y - x, 0)$), so strong duality holds. Thus the bi-dual problem is the primal problem. Given this we can notice that μ corresponds to the primal variable b because it's the only variable that is multiplying $\sum_{i=1}^m \beta_i$, so finally we get

$$b^\star = y - \beta^\star - \frac{1}{\rho}\kappa\beta^\star$$

Subproblem 3.5

Let's define the matrix

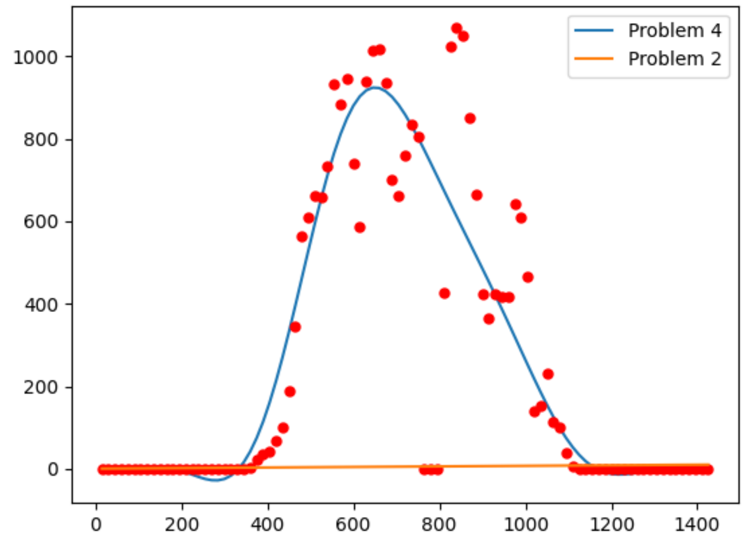
$$K = \begin{bmatrix} \phi(x_1) & \dots & \phi(x_m) \end{bmatrix}$$

We get that $\kappa = K^T K$
Consider $v \in \mathbb{R}^{m \times 1}$

$$v^T \kappa v = v^T K^T K v = (Kv)^T Kv = w^T w \geq 0$$

This proves that κ is positive semidefinite.

Subproblem 3.6



As we can clearly see, problem 2 gives us back a linear regression that is fixed at almost zero because most of the points have a 0 y value.
On the other hand, problem 4 exploits the power of kernel regression by adding a non-linearity, such that in the high-dimensional feature space there is a linear separator that can be estimated by regression.