Notation: for the sake of clarity, we will denote vectors as $\mathbf{x}$ and their components as x_i so that bold font identifies vectors and regular font indentifies real numbers.

Exercise 1.

Let $\mathcal{C}$ denote the set considered at each point in the exercise.

- The set $\mathcal{C}$ is the epigraph of the function $f : x \in \mathbb{R} \to \frac{1}{x}$. We have that $\operatorname{dom} f = \mathbb{R}_+^*$ is convex and $f''(x) = \frac{2}{x^3} \geq 0$ for all $x \in \mathbb{R}_+^*$. Thus, $\boxed{\mathcal{C} \text{ is convex}}$.
- $\boxed{\text{The set } \mathcal{C} \text{ is not necessarily convex}}$. Indeed, let $e_x := (x, 0, \ldots, 0) \in \mathbb{R}^n$ for any $x \in \mathbb{R}$. Let $S := \{e_{-1}, e_1\}$ and $T := \{e_0\}$. We get that $e_{-1}, e_1 \in \mathcal{C}$ but $\frac{1}{2}e_{-1} + \frac{1}{2}e_1 = e_0 \notin \mathcal{C}$.
- We have that $x \in \mathcal{C}$ if and only if for all $s_2 \in S_2$, there exists $s_1 \in S_1$ such that $x = s_1 - s_2$. Thus, $\mathcal{C} = \bigcap_{s_2 \in S_2} S_1 - s_2$. Since $S_1 - s_2$ is convex for all $s_2 \in S_2$ and the intersection of convex sets remains convex, we get that $\boxed{\text{the set } \mathcal{C} \text{ is convex}}$.
- Take $x_1, x_2 \in \mathcal{C}$. Thus, there exist $y_1, y_2 \in S_2$ such that $x_1 + y_1 \in S_1$ and $x_2 + y_2 \in S_1$. For any $\alpha \in [0, 1]$, we have that $\alpha y_1 + (1 - \alpha) y_2 \in S_2$ by convexity of S_2. Moreover, notice that

$$\alpha x_1 + (1 - \alpha) x_2 + \alpha y_1 + (1 - \alpha) y_2 = \alpha (x_1 + y_1) + (1 - \alpha) (x_2 + y_2). \tag{1}$$

  By convexity of S_1, the right-hand side of (1) is in S_1 and so $\alpha x_1 + (1 - \alpha) x_2 \in \mathcal{C}$. Thus $\boxed{\mathcal{C} \text{ is convex}}$.

Exercise 2.

- We compute the Hessian matrix $\mathbf{H}_f$. We have

$$\mathbf{H}_f = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

  We have that $\mathbf{H}_f \not\succeq 0$ and $\mathbf{H}_f \not\preceq 0$. Therefore f is not concave nor convex. Also, let $\alpha \in \mathbb{R}$. The set $S_\alpha := \left\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 x_2 \leq \alpha\right\}$ is not covex nor concave. Therefore f is not quasiconvex nor quasiconcave.

- The Hessian matrix

$$\mathbf{H}_f = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

  is such that $\mathbf{x}^t \mathbf{H}_f \mathbf{x} > 0$ for all $x \in \mathbb{R}_{++}^n$. Since $\mathbb{R}_{++}^n$ is convex, we deduce that f is convex (so it is also quasiconvex). By looking at $-\mathbf{H}_f$ we see that f is not concave. Moreover, for any $\alpha \in \mathbb{R}$, the set $S'_\alpha := \left\{(x_1, x_2) \in \mathbb{R}_{++}^2 \mid -x_1 x_2 \leq \alpha\right\}$ is convex for all α. Therefore f is quasiconcave.

- We have that

$$\mathbf{H}_f(x_1, x_2) = \begin{pmatrix} 0 & -\frac{1}{x_2^2} \\ -\frac{1}{x_2^2} & 2\frac{x_1}{x_2^3} \end{pmatrix}$$

  and thus we compute

$$\begin{pmatrix} y_1 & y_2 \end{pmatrix} \begin{pmatrix} 0 & -\frac{1}{x_2^2} \\ -\frac{1}{x_2^2} & 2\frac{x_1}{x_2^3} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = -2\frac{y_1 y_2}{x_2^2} + 2\frac{x_1 y_2^2}{x_2^3}. \tag{2}$$

  For instance, the left-hand side of (2) is positive for $y_1 = y_2 = 1$, $x_2 = 1$ and x_1 large enough. On the other hand, it is negative for $y_1 = y_2 = 1$, $x_2 = 1$ and $0 < x_1 < \varepsilon$ for small enough. This entails that $\mathbf{H}_f \not\succeq 0$ and $\mathbf{H}_f \not\preceq 0$. Therefore f is not convex nor concave. On the other hand, the set $S_\alpha = \left\{(x_1, x_2) \in \mathbb{R}_{++}^n \mid x_1/x_2 \leq \alpha\right\} = \{(x_1, x_2) \mid x_1 \leq \alpha x_2\}$ is a half-plane and so it is convex. Thus f is quasiconvex. We show in the same way that f is quasiconcave.

- Notice that $\mathbb{S}^n_{++}$ is convex. Given this, it suffices to show that the function $g : t \in \mathbb{R} \mapsto \mathrm{Tr}\,(A+tB)$, is convex for all $A \in \mathbb{S}^n_{++}$ and $B \in \mathbb{S}^n$ at $t = 0$. Notice that, since $A \in \mathbb{S}^n_{++}$, A is invertible and $A^{-1}B$ is symetric. Moreover, there exists Q such that $A^{-1}B = Q\Lambda Q^{-1}$, where $\Lambda = \mathrm{Diag}\,(\lambda_1, \ldots, \lambda_n)$. We also have that $\left(I + tA^{-1}B\right) = Q\,(I + t\Lambda)\,Q^{-1}$ thus, for t small enough, $I + t\Lambda$ is invertible. Also, we remark that, there exists $\widetilde{Q}$ such that $A = \widetilde{Q}\widetilde{\Lambda}\widetilde{Q}^{-1}$ with $\widetilde{\Lambda} = \mathrm{Diag}\left(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n\right)$ with $\widetilde{\lambda}_i > 0$ for all $1 \leq i \leq n$. We thus have that

$$(A+tB)^{-1} = \left(I + tA^{-1}B\right) A^{-1}Q\,(I+t\Lambda)^{-1}\,Q^{-1}\widetilde{Q}^{-1}\widetilde{\Lambda}\widetilde{Q}.$$

Thus, $\mathrm{Tr}\left((A+tB)^{-1}\right) = \sum_{i=1}^n \frac{1}{\widetilde{\lambda}_i(1+t\lambda_i)}$. We finally get $g''(t) = \sum_{i=1}^n \frac{2\lambda_i^2}{\widetilde{\lambda}_i(1+t\lambda_i)^3} > 0$ for t small enough. We finally get that f is convex.

Exercise 3.

- For any $s \in \mathbb{R}^n$, let $h(\mathbf{x}) := \mathbf{s}^t\mathbf{x} - \|\mathbf{x}\|_2^2$. By taking $\nabla h(\mathbf{x}) = s - 2x$, we get $\nabla h(\mathbf{x}_*) = 0$ iff $x_* = \frac{1}{2}\mathbf{s}$. Since $\mathbf{H}_h \preceq 0$, our choice of x_* yields a maximum of h. Therefore, $\|\mathbf{s}\|_2^{2,*} = \frac{1}{2}\|\mathbf{s}\|_2^2 - \frac{1}{4}\|\mathbf{s}\|_2^2 = \frac{1}{4}\|\mathbf{s}\|_2^2$.

  For the ℓ_1 norm, we observe the following. Let $s \in \mathbb{R}$. If $s > 0$, then we have that $\sup_{x\in\mathbb{R}} sx - |x| = +\infty$ (by taking $x \to +\infty$ if $s > 0$ and $x \to -\infty$ if $s < 0$). On the other hand, if $|s| \leq 1$, we have $sx - |x| \leq 0$, for all choices of $x \in \mathbb{R}$. By chosing $x = 0$, we get that $sx - |x| = 0$. Now, let $\mathbf{s} \in \mathbb{R}^n$. If there exists i_0 such that $|s_{i_0}| > 1$ (that is, $\|\mathbf{s}\|_\infty > 1$), we deduce that, by choosing $x_i = 0$ for all $i \neq i_0$ and by letting x_{i_0} to infinity to with the appropriate sign, we obtain $\sup_{\mathbf{x}\in\mathbb{R}^n} \mathbf{s}^t\mathbf{x} - \|\mathbf{x}\|_1 = +\infty$. On the other hand, if $\|\mathbf{s}\|_\infty \leq 1$, we get that

$$\sup_{\mathbf{x}\in\mathbb{R}^n} \mathbf{s}^t\mathbf{x} - \|\mathbf{x}\|_1 = \sup_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^n (s_ix_i - |x_i|) \leq \sum_{i=1}^n \sup_{x\in\mathbb{R}} (s_ix - |x|) \leq 0.$$

  But such upperbound can be obtained by choosing $\mathbf{x} = \mathbf{0}$.

  All in all we have that $\|\cdot\|_1^* = \mathrm{i}_{\{\|\cdot\|_\infty \leq 1\}}(\cdot)$

- Let $g * h$ denote the infimal convolution of g and h. For any $\mathbf{s} \in \mathbb{R}^n$ we have:

$$\begin{aligned}
(g*h)^*(s) &= \sup_{\mathbf{x}} \mathbf{s}^t\mathbf{x} - \min_{\mathbf{u}+\mathbf{v}=\mathbf{x}} (g(\mathbf{u}) + h(\mathbf{v})) \\
&= \sup_{\mathbf{x}} \mathbf{s}^t\mathbf{x} + \max_{\mathbf{u}+\mathbf{v}=\mathbf{x}} (-g(\mathbf{u}) - h(\mathbf{v})) \\
&= \sup_{\mathbf{x}} \max_{\mathbf{u}+\mathbf{v}=\mathbf{x}} \left(\mathbf{s}^t\mathbf{x} - g(\mathbf{u}) - h(\mathbf{v})\right) \\
&= \sup_{\mathbf{x}} \max_{\mathbf{u}+\mathbf{v}=\mathbf{x}} \left(\mathbf{s}^t\mathbf{u} - g(\mathbf{u})\right) + \left(\mathbf{s}^t\mathbf{v} - h(\mathbf{v})\right) \\
&= \sup_{\mathbf{u},\mathbf{v}} \left(\mathbf{s}^t\mathbf{u} - g(\mathbf{u}) + \mathbf{s}^t\mathbf{v} - h(\mathbf{v})\right) \\
&= \sup_{\mathbf{u}} \left(\mathbf{s}^t\mathbf{u} - g(\mathbf{u})\right) + \sup_{\mathbf{u}} \left(\mathbf{s}^t\mathbf{v} - h(\mathbf{v})\right).
\end{aligned}$$

  Thus, $(g*h)^*(\mathbf{s}) = g^*(\mathbf{s}) + h^*(\mathbf{s})$.

  Now, we compute

$$\min_{\mathbf{u}+\mathbf{v}=\mathbf{x}} (g(\mathbf{u}) + h(\mathbf{v})) = \min_{\mathbf{u}+\mathbf{v}=\mathbf{x}} \left(\|\mathbf{u}\|_1 + \frac{1}{2\alpha}\|\mathbf{v}\|_2^2\right) = \min_{\mathbf{u}} \left(\|\mathbf{u}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}-\mathbf{u}\|_2^2\right).$$

  Let $h(\mathbf{u}) := \min_{\mathbf{u}} \left(\|\mathbf{u}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}-\mathbf{u}\|_2^2\right)$. We have:

$$\min_{\mathbf{u}} h(\mathbf{u}) = \min_{\mathbf{u}} \left(\|\mathbf{u}\|_1 + \frac{1}{2\alpha}\|\mathbf{u}\|_2^2\right) = \min_{\mathbf{u}} \left(\sum_{i=1}^n \left(|u_i| + \frac{1}{2\alpha}(x_i - u_i)^2\right)\right) = \sum_{i=1}^n \min_u |u| + \frac{1}{2\alpha}(x_i - u)^2.$$

Let $h_i(u) := |u| + \frac{1}{2\alpha}(x_i - u)^2$. We compute the minimum of h_i on $\mathbb{R} \setminus \{0\}$. We have that for anu $u \in \mathbb{R} \setminus \{0\}$, we get $h_i'(u) = \operatorname{sign}(u) - \frac{1}{\alpha}(x_i - u)$. Also $h_i''(u) = \frac{1}{\alpha} > 0$. We first compute the minimum of h_i on $\mathbb{R}_-^*$. We have:

$$h_i'(u) = 0 \iff -1 + \frac{1}{\alpha}(x_i - u) = 0 \iff u = x_i + \alpha.$$

Which requires $x_i + \alpha < 0$, i.e. $x_i < -\alpha$.

We now compute the minimum of h_i on $\mathbb{R}_+^*$. We have:

$$h_i'(u) = 0 \iff 1 - \frac{1}{\alpha}(x_i - u) = 0 \iff u = x_i - \alpha.$$

Which requires $x_i > \alpha$. We thus have the existence of a unique minimum of h_i on $\mathbb{R} \setminus \{0\}$ if either $x_i > \alpha$ or $x_i < -\alpha$. If such conditions are met, we deduce that this is also a global minimum on $\mathbb{R}$ by the continuity of h_i.

Otherwise, if we have $-\alpha < x_i < \alpha$, we get $h_i'(u) < 0$ for all $u < 0$ and $h_i'(u) > 0$ for all $u > 0$. we deduce that, in such case, the minimum of h_i is attained at 0. All in all, we have that

$$m_i := \min_u h_i(u) = \begin{cases} -x_i - \frac{1}{2}\alpha & \text{if} \quad x_i < -\alpha, \\ \frac{x_i^2}{2\alpha} & \text{if} \quad -\alpha \leq x_i \leq \alpha, \\ x_i - \frac{1}{\alpha} & \text{if} \quad x_i > \alpha. \end{cases}$$

And thus:

$$f(\mathbf{x}) = \min_{\mathbf{u}} h(\mathbf{u}) = \sum_i^n m_i.$$

Finally we notice that since f is continuous and $\operatorname{dom}(f) = \mathbb{R}^n$ is closed, f is a closed function. Moreover, since all the m_is are convex, f is also convex. Since f is closed and convex, we have $f = f^{**}$.

- Let $f : x \in \mathbb{R} \mapsto \log(1 + \exp(x))$ and let $s \in \mathbb{R}$, we define $h : \mathbb{R} \to \mathbb{R}$ by $h(x) = sx - \log(1 + \exp(x))$. We have $h'(x) = s - \frac{\exp(x)}{1+\exp(x)}$ and $h''(x) = -\frac{\exp(x)}{(1+\exp(x))^2} < 0$. So h is concave.
  - For $s < 0$, we have $\lim_{x\to+\infty} h(x) = +\infty$.
  - If $s = 0$, we have $h(x) < 0$ and $\lim_{x\to-\infty} h(x) = 0$.
  - If $0 < s < 1$, we have

    $$h'(x) = 0 \iff s - \frac{\exp(x)}{1 + \exp(x)} = 0 \iff x = \log\left(\frac{s}{1-s}\right).$$

    By the concavity of h, we have that

    $$\sup_{x\in\mathbb{R}} h(x) = s\log\left(\frac{s}{1-s}\right) - \log\left(1 + \frac{s}{1-s}\right) = s\log(s) + (1-s)\log(1-s).$$

  - If $s = 1$, we have $h(x) = \log\left(\frac{\exp(x)}{1+\exp(x)}\right) < 0$ and $\lim_{x\to+\infty} \log\left(\frac{\exp(x)}{1+\exp(x)}\right) = 0$.
  - If $s > 1$, we have $\lim_{x\to+\infty} h(x) = +\infty$.

  All in all, we have that

$$f^*(s) = \begin{cases} +\infty & \text{if} \quad s < 0, \\ 0 & \text{if} \quad s = 0, \\ s\log(s) + (1-s)\log(1-s) & \text{if } 0 < s < 1, \\ 0 & \text{if} \quad s = 1, \\ +\infty & \text{if} \quad s > 1. \end{cases}$$

- The Lagrangian of LASSO is

$$\mathcal{L}(\mathbf{x},\mathbf{y},\lambda) = \frac{1}{2}\|\mathbf{y}-\mathbf{b}\|_2^2 + \alpha\|x\|_1 + \lambda^t(A\mathbf{x}-\mathbf{y}).$$

We exploit the fact that $(\alpha f)^*(\mathbf{s}) = \alpha f^*(\mathbf{s}/\alpha)$ for all $\mathbf{s}\in\mathbb{R}^n$ and $\alpha\neq 0$ to deduce that $(\alpha\|\cdot\|_1)^* = \alpha\,\mathrm{i}_{\{\|\cdot\|_\infty\leq 1\}}\left(\frac{\cdot}{\alpha}\right) = \mathrm{i}_{\{\|\cdot\|_\infty\leq 1\}}\left(\frac{\cdot}{\alpha}\right)$. Similarly, we have $\left(\frac{1}{2}\|\cdot\|_2^2\right)^* = \frac{1}{2}\|\cdot\|_2^2$. Finally we notice that $(f(\cdot-\mathbf{c}))^* = f^*(\cdot)+(\cdot)^t\mathbf{c}$. Therefore, we have that $\left(\frac{1}{2}\|\cdot-\mathbf{c}\|_2^2\right)^*(\lambda) = \frac{1}{2}\|\lambda\|_2^2+\lambda^t\mathbf{c}$. We finally get that the (Fenchel) dual for LASSO is

$$\begin{aligned}
&\max_{\lambda\in\mathbb{R}^m} -\left(\frac{1}{2}\|\cdot\|_2^2\right)^*(\lambda) - (\alpha\|\cdot\|_1)^*(-A^t\lambda)\\
=&\max_{\lambda} -\frac{1}{2}\|\lambda\|_2^2 + \frac{1}{2}\|\mathbf{b}\|_2^2 - \mathrm{i}_{\{\|\cdot\|_\infty\leq 1\}}\left(-\frac{A^t\lambda}{\alpha}\right)\\
=&\max_{\lambda} -\frac{1}{2}\|\lambda+\mathbf{b}\|_2^2 + \frac{1}{2}\|\mathbf{b}\|_2^2 - \mathrm{i}_{\{\|\cdot\|_\infty\leq 1\}}\left(-\frac{A^t\lambda}{\alpha}\right)\\
=&\max_{\lambda} -\frac{1}{2}\|\lambda-\mathbf{b}\|_2^2 + \frac{1}{2}\|\mathbf{b}\|_2^2 - \mathrm{i}_{\{\|\cdot\|_\infty\leq 1\}}\left(\frac{A^t\lambda}{\alpha}\right).
\end{aligned}$$

Where the last equality is given by the change of variables $\lambda\mapsto-\lambda$. We have the desired result.

- To keep notation consistent, we denote by $\mathbf{w}^{(i)}$ what is denoted w_i in the homework text. The j-th coordinate of $\mathbf{w}^{(i)}$ will be denoted $\mathbf{w}_j^{(i)}$. It is clear that the original problem and the formulation with the intermediate constraints are equivalent. The Lagrangian of the intermediate constrained problem is

$$\mathcal{L}\left(\lambda,\nu,\mathbf{w},\mathbf{w}^{(1)},\ldots,\mathbf{w}^{(m)}\right) = \frac{1}{m}\sum_{i=1}^m h_i\left(\mathbf{w}^{(i)}\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^m \nu_i^t\left(\mathbf{w}^{(i)}-\mathbf{w}\right).$$

We want to compute

$$\begin{aligned}
g(\nu) &= \inf_{\substack{\mathbf{w}\in\mathbb{R}^n\\ \mathbf{w}^{(1)},\ldots,\mathbf{w}^{(m)}\in\mathbb{R}^n}} \frac{1}{m}\sum_{i=1}^m h_i\left(\mathbf{w}^{(i)}\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^m \nu_i^t\left(\mathbf{w}^{(i)}-\mathbf{w}\right)\\
&= \inf_{\substack{\mathbf{w}\in\mathbb{R}^n\\ \mathbf{w}^{(1)},\ldots,\mathbf{w}^{(m)}\in\mathbb{R}^n}} \frac{1}{m}\sum_{i=1}^m \left(h_i\left(\mathbf{w}^{(i)}\right) - \left(m\nu_i^t\right)\mathbf{w}^{(i)}\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 - \sum_{i=1}^m \nu_i^t\mathbf{w}\\
&= -\frac{1}{m}\sum_{i=1}^m \sup_{\mathbf{w}^{(i)}\in\mathbb{R}^n}\left(\left(m\nu_i^t\right)\mathbf{w}^{(i)} - h_i\left(\mathbf{w}^{(i)}\right)\right) - \sup_{\mathbf{w}\in\mathbb{R}^n} -\frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^m \nu_i^t\mathbf{w}.
\end{aligned}$$

We have that

$$\sup_{\mathbf{w}^{(i)}\in\mathbb{R}^n}\left(\left(m\nu_i^t\right)\mathbf{w}^{(i)} - h_i\left(\mathbf{w}^{(i)}\right)\right) = h_i^*(m\nu_i)$$

Let $f:\mathbf{w}\in\mathbb{R}^n\mapsto -\frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^m \nu_i\mathbf{w}$. Notice that f is concave. We have

$$\nabla f(\mathbf{w}) = -\lambda\mathbf{w} + \sum_{i=1}^m \nu_i = 0 \iff \mathbf{w} = \frac{1}{\lambda}\sum_{i=1}^m \nu_i.$$

Finally:

$$\sup_{\mathbf{w}\in\mathbb{R}^n}\sum_{i=1}^m \nu_i\mathbf{w} - \frac{\lambda}{2}\|\mathbf{w}\|_2^2 = \frac{1}{\lambda}\sum_{1\leq i,j\leq m}\nu_i^t\nu_j - \frac{\lambda}{2}\left\|\frac{1}{\lambda}\sum_{i=1}^m\nu_i\right\|_2^2 = \frac{1}{2\lambda}\left\|\sum_{i=1}^m\nu_i\right\|_2^2.$$

All in all, we have:

$$g(\nu) = -\frac{1}{m}\sum_{i=1}^{m} h_i^*(m\nu_i) - \frac{1}{2\lambda}\left\|\sum_{i=1}^{m}\nu_i\right\|_2^2.$$

Now, let $h_i : \mathbf{w} \in \mathbb{R}^n \mapsto \log(1+\exp(-y_i\mathbf{x}_i^t\mathbf{w}))$ and denote $\ell : x \in \mathbb{R} \mapsto \log(1+\exp(x))$. We notice that $\mathbf{w} \in \mathbb{R}^n \mapsto \exp(-y_i\mathbf{x}_i^t\mathbf{w})$ is concave and log is convex and non-decreasing. Thus, h_i is convex. So, for any given $\mathbf{s} \in \mathbb{R}^n$, we get that the function $\varphi : \mathbf{w} \in \mathbb{R}^n \mapsto \mathbf{s}^t\mathbf{w} - h_i(\mathbf{w})$ is concave. Set $x = -y_i\mathbf{x}_i\mathbf{w}$ so that $\mathbf{w} = -\frac{y_i\mathbf{x}_i}{\|x_i\|_2^2}x$. For $s = -y_i\frac{\mathbf{s}^t\mathbf{x}_i}{\|\mathbf{x}_i\|_2^2}$ we have:

$$\sup_{x\in\mathbb{R}} -y_i\frac{\mathbf{s}^t\mathbf{x}_i}{\|\mathbf{x}_i\|_2^2}x - \log\left(1+\exp\left(-y_i\mathbf{x}_i^t\cdot\left(-\frac{y_i\mathbf{x}_i}{\|\mathbf{x}_i\|_2^2}x\right)\right)\right) =$$

$$= \sup_{x\in\mathbb{R}} sx - \log(1+\exp(x)) = \ell^*(s) = \ell^*\left(-y_i\frac{\mathbf{s}^t\mathbf{x}_i}{\|\mathbf{x}_i\|_2^2}\right).$$

All in all, we have that the Lagrangian dual to the logistic regression problem is

$$g(\nu) = -\frac{1}{m}\sum_{i=1}^{m}\ell^*\left(-y_i m\frac{\nu_i^t\mathbf{x}_i}{\|\mathbf{x}_i\|_2^2}\right) - \frac{1}{2\lambda}\left\|\sum_{i=1}^{m}\nu_i\right\|_2^2.$$

- We set th domain of this problem to be $\mathbb{S}^n$. For any $X \in \mathbb{S}^n$, let $(\lambda_i(X))_{i=1}^n$ be its eigenvalues enumerated with multiplicity. We claim that for any $X \in \mathbb{S}^n$, $\max_{Y\succeq 0} -\operatorname{Tr}(XY)$ is finite iff $Y \succeq 0$. We have

$$\max_{Y\succeq 0} -\operatorname{Tr}(XY) = \max_{Y\succeq 0} -\sum_{i=1}^{n}\lambda_i(X)\lambda_i(Y) = \begin{cases} -\infty & \text{if } X \not\succeq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed,if we have $X \not\succeq 0$, it has a strictly negative eigenvalue λ_{i_0}. Then, by taking $Y = \operatorname{Diag}(0,\dots,0,y,0,\dots,0)$ with y at position i_0, we get $\operatorname{Tr}(XY) = -y\lambda_{i_0}$, which tends to $-\infty$ for $y \to +\infty$. If $X \succeq 0$, then $-\operatorname{Tr}(XY) \leq 0$, so taking $Y = 0$ yields the desired result. Therefore, the problem in the statement is equivalent to

$$\min_{X\in\mathbb{S}^n}\max_{\substack{\lambda\in\mathbb{R}^m \\ S\succeq 0}} \operatorname{Tr}(A_0X) + \sum_{i=1}^{m}\lambda_i(\operatorname{Tr}(A_iX) - b_i) - \operatorname{Tr}(SX).$$

So, we can write the Lagrangian of the problem as

$$\mathcal{L}(X,\lambda,S) = \operatorname{Tr}(A_0X) + \sum_{i=1}^{m}\lambda_i(\operatorname{Tr}(A_iX) - b_i) - \operatorname{Tr}(SX).$$

We want to compute for $S \succeq 0$:

$$\begin{aligned} g(\lambda,S) &= \operatorname{Tr}(A_0X) + \sum_{i=1}^{m}\lambda_i(\operatorname{Tr}(A_iX) - b_i) - \operatorname{Tr}(SX) \\ &= \operatorname{Tr}\left(\left(A_0 + \sum_{i=1}^{m}\lambda_iA_i - S\right)X\right) - \lambda^t\mathbf{b} \\ &= \begin{cases} -\lambda^t\mathbf{b} & \text{if } A_0 + \sum_{i=1}^{m}\lambda_iA_i - S = 0, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Notice that for S_* to be optimal, we need to have $S_* = A_0\sum_{i=1}^{m}\lambda_iA_i$. If there exists an optimal X_*, then we have $-\operatorname{Tr}(S_*X_*) \leq 0$ since $S_* \in \mathbb{S}^n$ and $\operatorname{Tr}(A_iX_*) - b_i = 0$ for all i. This

gives us primal feasibility. We also have $S \succeq 0$ by construction, which yields dual feasibility. Since $S_* = A_0 + \sum_{i=1}^m \lambda_i A_i$, we have

$$\frac{\partial \operatorname{Tr}(A_0 X)}{\partial X} + \sum_{i=1}^m \frac{\partial (\operatorname{Tr}(A_i X) - b_i)}{\partial X} - \frac{\partial \operatorname{Tr}(S_* X)}{\partial X} = A_0^t + \sum_{i=1}^m \lambda_i A_i^t - S_*^t = 0,$$

which yields the first order condition. Since our original problem is convex, having strong duality is equivalent to the missing KKT condition, which is complementary slackness, i.e. $-\operatorname{Tr}(S_* X_*) = 0$.

Exercise 4.

- Let p^* be the optimal value for the original formulation of the problem and let q^* be optimal value for the reformulated version. Since a feasible solution of the original problem automatically gives a feasible solution to the reformulated problem (by setting $s_i = \max\{0, 1 - y_i \mathbf{x}_i^t \mathbf{w}\}$), we have $p^* \leq q^*$. Since minimizing the objective function of the reformulated problem yields optimal $s_i^* = \max\{0, 1 - y_i \mathbf{x}_i^t \mathbf{w}^*\}$, we get that $q^* \leq p^*$. Therefore the two formulations are equivalent.

  The Lagrangian of the problem is the following:

$$\mathcal{L}(\mathbf{w}, \mathbf{s}, \lambda, \lambda') = \frac{1}{2}\|\mathbf{w}\|_2^2 + \alpha \sum_{i=1}^m s_i - \sum_{i=1}^m \lambda_i \left(\mathbf{w}^t \mathbf{x}_i y_i + s_i - 1\right) - \sum_{i=1}^m \lambda_i' s_i.$$

  Notice that the Lagrangian is convex in $\mathbf{w}$ and $\mathbf{s}$. Thus, to compute the Lagrange dual, we compute

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^m \lambda_i \mathbf{x}_i y_i = 0 \Longleftrightarrow \mathbf{w} = X\lambda.$$

  We also have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{s}}\Big|_j = \alpha - \lambda_j - \lambda_j' = 0.$$

  In order to achieve optimality we need $\lambda_i \geq 0$ and $\lambda_i' \geq 0$, this implies that $\lambda_j \leq \alpha$. Therefore, plugging this into $\mathcal{L}$, we get

$$\begin{aligned}
\mathcal{L}(\mathbf{w}^*, \mathbf{s}^*, \lambda, \lambda') &= \max_{\substack{0 \leq \lambda \leq \alpha \\ 0 \leq \lambda' \leq \alpha}} \left(\frac{1}{2}\|X\lambda\|_2^2 - \sum_{i=1}^m \lambda_i \left(\lambda^t X^t\right) \mathbf{x}_i y_i^2 + \sum_{i=1}^m \underbrace{(\alpha - \lambda_i - \lambda_i')}_{=0} + \sum_{i=1}^m \lambda_i \right) \\
&= \max_{0 \leq \lambda \leq \alpha} \left(\frac{1}{2}\|X\lambda\|_2^2 - \lambda^t X^t X \lambda + \sum_{i=1}^m \lambda_i \right) \\
&= \max_{0 \leq \lambda \leq \alpha} D(\lambda).
\end{aligned}$$

  We note in passing that we have the natural estimate for $\mathbf{w} = X\lambda$.

- From now on, we will refer to the Projected Gradient Ascent algorithm as PGA and to the Randomized Coordinate Ascent algorithm as RCA.

  We first prove that one iteration of (RCA) optimizes one dimension of λ at a time. We first notice that we keep thoughout the algorithm the invariant $\mathbf{w}^{(k)} = X\lambda^{(k)}$, required for optimality of the primal variable. One iteration step consists in performing a gradient ascent step with respect to the randomly chosen variable λ_{i_k}. We have

$$\frac{\partial D}{\partial \lambda_{i_k}}(\lambda) = 1 - y_{i_k} \mathbf{x}_{i_k}^t \left(\sum_{j=1}^m y_j \mathbf{x}_j \lambda_j \right).$$

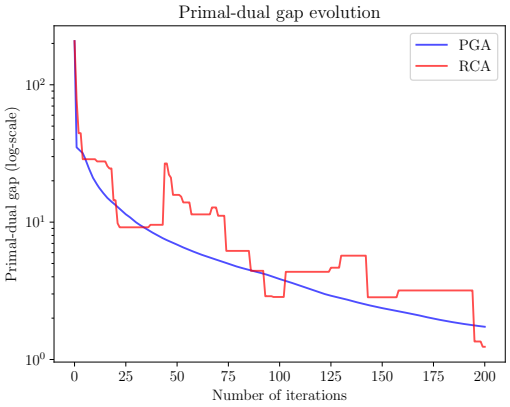
As for the step size, we choose a Newton-like step of $\left|\frac{\partial D}{\partial \lambda_{i_k}}(\lambda)\right| = \frac{1}{\|\mathbf{x}_{i_k}\|_2^2}$. The corresponding gradient ascent step is

$$\lambda_{i_k}^{(k+1)} = \lambda_{i_k}^{(k)} - \frac{1 - y_{i_k}\mathbf{x}_{i_k}^t \mathbf{w}^{(k)}}{\|\mathbf{x}_{i_k}\|_2^2}.$$

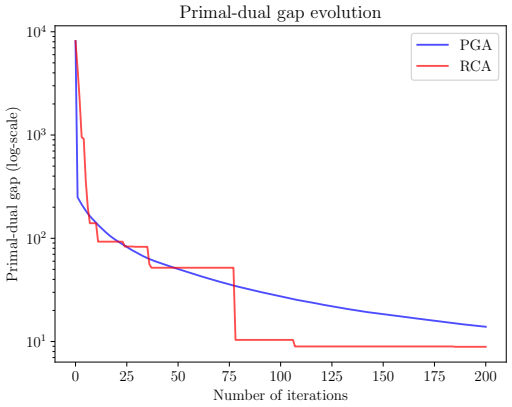
We project this onto $[0, \alpha]$ to ensure feasibility.

We also keep our invariant $\mathbf{w}^{(k+1)} = X\lambda^{(k+1)}$. We only need to update the i_k-th dimension to get $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + y_{i_k}\mathbf{x}_{i_k}\left(\lambda_{i_k}^{(k+1)} - \lambda_{i_k}\right)$. We have the desired result.

Now, let us look at the performances of these algorithms. One can look at and reproduce our same simulations at `https://github.com/vincenzo-politelli/CVX`. We tested the algorithms on the LIBSVM datasets "sonar" (208 samples and 60 features). and "mushrooms" (8124 samples and 112 features). For a value of $\alpha = 1$ and a maximum number of iterations of 200. We look at the primal-dual gap agains the number of iterations and we obtain the following.



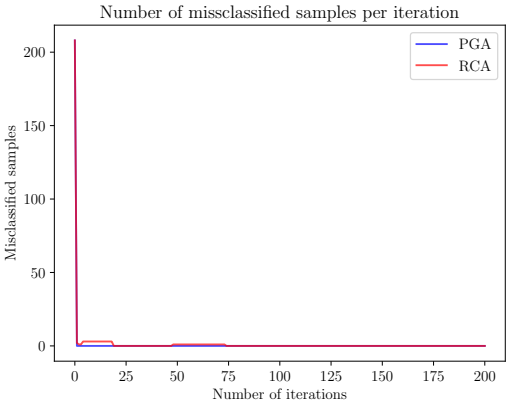
(a) Data for "sonar" dataset with $\alpha = 1$ and a maximum number of iterations of 200



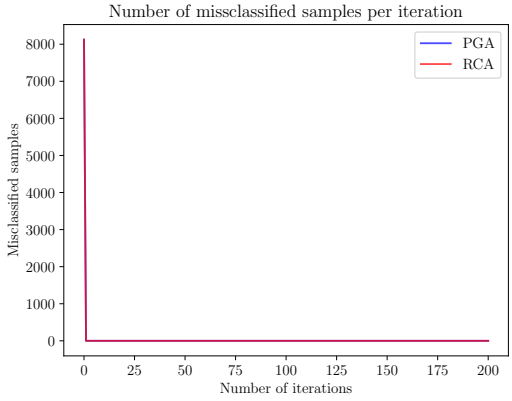
(b) Data for "mushrooms" dataset with $\alpha = 1$ and a maximum number of iterations of 200

We can see that the quality of the classification is comparable and that the primal-dual gap drops very quickly. In the case of the "sonar" dataset, for PGA, the final primal-dual gap is of 1.7312; while in the case of RCA such value amounts to 1.2373. In the case of the "mushrooms" dataset, the primal-dual gap for PGA is 13.9067, while for RCA, such value is 8.8974. Thus, overall, RCA seems to be slightly more precise. We can see that the fact that in RCA we do not take into account all the information about the gradient makes the convergence less smooth but still precise. On the other hand, RCA is much faster than PGA. Ineed, in the case of the "sonar" dataset, we PGA takes around 0.0164 s to run, while RCA takes 0.0063 s. In the case of the "mushrooms" dataset, PGA takes 81.541 s to complete, while RCA takes only 0.1542 s. We evince that the fact that RCA does not compute the whole gradient makes it much faster (but still precise enough).

As for the misclassifications, we get that the two algorithms perform very well, the two dataset yielding 0 misclassified samples only after the first few iterations.



(a) Data for "sonar" dataset with $\alpha = 1$ and a maximum number of iterations of 200



(b) Data for "mushrooms" dataset with $\alpha = 1$ and a maximum number of iterations of 200

- Consider the LP problem

$$\begin{aligned} \min_{\substack{\mathbf{w},\mathbf{v}\in\mathbb{R}^n \\ \mathbf{v}\in\mathbb{R}^m}} \quad & \sum_{i=1}^{n} u_i + \alpha \sum_{i=1}^{m} v_i \\ \text{subject to} \quad & \mathbf{u} \geq \mathbf{w} \text{ and } \mathbf{u} \geq -\mathbf{w}, \\ & \mathbf{v} \geq \mathbf{0}_n \text{ and } \mathbf{v} \geq \mathbf{1}_n - X^t\mathbf{w}. \end{aligned}$$

  Let p^* be the optimal solution to the original problem and let q^* bet the optimal solution to the above reformulation. Clearly any feasible solution to the original pr/oblem is a feasible solution of the reformulation. Thus $p^* \leq q^*$. Moreover, since $w_i \geq |w_i|$ and $v_i \geq \max\{0, 1 - y_i\mathbf{x}_i^t\mathbf{w}\}$, minimizing with respect to $\mathbf{w}, \mathbf{u}, \mathbf{v}$ requires to choose $u_i^* = |w_i^*|$ and $v_i = \max\{0, 1 - y_i\mathbf{x}_i^t\mathbf{w}^*\}$. Thus, $q^* \leq p^*$ and the two problems are equivalent.

  One can find on `https://github.com/vincenzo-politelli/CVX` the code that we use to solve this linear optimization problem. We have different values for the vector $\mathbf{w}$ (as expected) and the method is pretty fast, yielding an execution time of 0.01830 s for the "sonar" dataset and of 0.2853 s for the "mushrooms" dataset. Thus, it seems to be just slightly less efficient than RCA. In this case as well, the final number of misclassified samples is 0.

- We make the preliminary remark that feasible solutions to the above LP problem aways exist. Indeed if suffices to take $\mathbf{w} = \mathbf{1}_n$, $\mathbf{u} = \mathbf{0}_n$ and $v_i > \max\{0, 1 - y_i\mathbf{x}_i^t\mathbf{w}\}$ for all i. Let $\mathbf{x}$ denote the concatenation $(\mathbf{w}, \mathbf{u}, \mathbf{v})$ and let A be a matrix such that $A\mathbf{x} \leq 0$ iff the linear constraints in the LP formulation above are satisfied. Let also $\mathbf{c} := (\underbrace{0, \dots, 0}_{n\text{-times}}, \underbrace{1, \dots, 1}_{n\text{-times}}, \underbrace{\alpha, \dots, \alpha}_{m\text{-times}})$. Let $n' := 2n + m$. Then, we have the equivalent reformulation of the above LP problem.

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{c}^t\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} \leq 0 \end{aligned}$$

  Let A_i be the i-th row of A and let m' be the number of rows of A. As seen in the course, we consider the family of functions with logarithmic barrier which are parametrized by $t > 0$:

$$\varphi_t : \mathbf{x} \in \mathbb{R}^{n'} \mapsto t\mathbf{c}^t\mathbf{x} - \sum_{i=1}^{m'} \log\left(A_i^t\mathbf{x}\right).$$

  As seen in the course, we propose the following interior-point scheme (Algorithm **??**).

Algorithm 1: Interior-point method algorithm

1 **??**

Input: A strictly feasible solution $\mathbf{x}$, $t = t_0 > 0$, $\varepsilon > 0$ and $\mu = 1 + 1/\sqrt{m'}$.

2 **while** True **do**

3 **minimize** $\varphi(t)$ by Newton's method to get $\mathbf{x}^*(t)$;

4 $\mathbf{x} \leftarrow \mathbf{x}^*(t)$;

5 **quit** if $m'/t < \varepsilon$;

6 $t \leftarrow \mu t$;

By the self-concordance of φ_t, the analysis provided in the course yields that Newton's method takes at most $m'(\mu - 1 - \log(\mu))$ iterations to converge. We recall that one step of Newton's methods is of the form $\mathbf{x}^+ = \mathbf{x} - \nabla^2\varphi_t(\mathbf{x})^{-1}\nabla\varphi_t(\mathbf{x})$, and requires $\mathcal{O}\left((n')^3\right)$ time. So, for $\mu = 1 + 1/\sqrt{m'}$, one iteration of our interior-point method algorithm takes time $m'\left(\frac{1}{\sqrt{m'}} - \log\left(1 + \frac{1}{\sqrt{m'}}\right)\right)\mathcal{O}\left((n')^3\right) = \mathcal{O}\left(\sqrt{m'}\,(n')^3\right)$.