Final Exam Solutions

You may use any books, notes, or pdf/text files on your laptop (tablets, cell phones and other devices are prohibited). **You are not permitted to connect to the internet or communicate with anyone in anyway during the exam. Please disconnect wireless, bluetooth etc. now. Any (even seemingly minor) violations will result in a failing grade for the course.**
All sub-problems have roughly equal weight. Some are easy. Others, not so much.

Total	/ 100
Q1	/ 20
Q2	/ 15
Q3	/ 25
Q4	/ 15
Q5	/ 25

1- **Short questions (20 pts). (2.5 points (a) and (b), (5) points (c)-(d))**
Specify whether the functions in parts (a)-(d) are strictly convex, convex, or nonconvex, and give a justification (brief justification in parts a and b)

(a) $f(x) = \frac{x}{1+\frac{1}{x}}$ for $x > 0$.

(b)

$$f(x) = \begin{cases} 0 & |x| \leq 1 \\ |x-1| & \text{otherwise} \end{cases}$$

(c) $-\text{trace}(\boldsymbol{X}) \cdot \text{logdet}\left(\frac{\boldsymbol{X}}{\text{trace}(\boldsymbol{X})}\right)$ on $\{\boldsymbol{X} \mid \boldsymbol{X} \succ \boldsymbol{0}\}$.

(d)

$$f(\boldsymbol{x}) = \sup_{t \in [a,b]} \log\left(\frac{p(t)}{q(t)}\right) \quad \text{where} \quad p(t) = e^{x_1 \sin(t)} + e^{x_2 \sin(2t)} + \ldots + e^{x_n \sin(nt)}$$
$$\text{and} \quad q(t) = e^{x_1 \sin(t) + x_2 \sin(2t) + \ldots x_n \sin(nt)}$$

Here a and b are real constants, with $a < b$.

(e) What is the proximal mapping associated with the regularizer

$$\mathcal{R}(\boldsymbol{z}) = \sum_{i=1}^{n} w_i |z_i|,$$

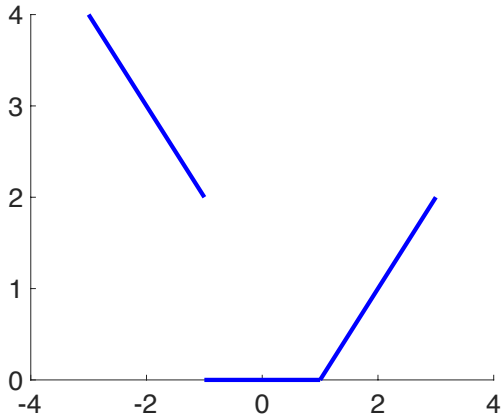
where $w_i > 0$ are known non-negative weights.

Solution

(a) This function is strictly convex as

$$f''(x) = \frac{2}{(x+1)^3} > 0.$$

(b) Nonconvex, see drawing below.



(c) Strictly convex. Note that -logdet($\boldsymbol{X}$) is convex. Hence its perspective function $f(\boldsymbol{X}, t) = -t \cdot \text{logdet}\left(\frac{\boldsymbol{X}}{t}\right)$ is convex over $(\boldsymbol{X}, t)$. Now compose the latter with a linear function of the form $\boldsymbol{X} \to (\boldsymbol{X}, \text{trace}(\boldsymbol{X}))$.

(d) Strictly convex. Note that $f(\boldsymbol{x})$ can be simplified into the form

$$f(\boldsymbol{x}) = \sup_{t \in [a,b]} \log\left(\frac{p(t)}{q(t)}\right) = \sup_{t \in [a,b]} \left(\log\left(\sum_{k=1}^{n} e^{x_k \sin(kt)}\right) - \left(\sum_{k=1}^{n} x_k \sin(kt)\right)\right)$$

Now note that for a fixed t the first term is convex (as log-sum-exp is convex) and the second term is linear. So there different is convex for a fixed t. Supremum of convex functions preserves convexity so f is convex.

(e) The proximal mapping is defined as

$$\mathcal{S}_{\mathcal{R}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z} \in \mathbb{R}^n} \quad \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{x}\|_{\ell_2}^2 + \mathcal{R}(\boldsymbol{z}).$$

For the weighted ℓ_1 norm defined above we have

$$\mathcal{S}_{\mathcal{R}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z} \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \left(\frac{1}{2} |z_i - x_i|^2 + w_i |z_i|\right),$$

Since the loss above is separable we have

$$[\mathcal{S}_{\mathcal{R}}(\boldsymbol{x})]_i = \arg\min_{z_i \in \mathbb{R}} \frac{1}{2} |z_i - x_i|^2 + w_i |z_i|$$

However, in class we learned that this value is given by soft thresholding at level w_i. Therefore,

$$[\mathcal{S}_{\mathcal{R}}(\boldsymbol{x})]_i = \arg\min_{z_i \in \mathbb{R}} \frac{1}{2} |z_i - x_i|^2 + w_i |z_i| = \mathcal{S}_{w_i}(x_i),$$

where

$$\mathcal{S}_{\mu}(x) = \begin{cases} x - \mu & \text{if } x \geq \mu \\ 0 & \text{if } |x| \leq \mu \\ x + \mu & \text{if } x \leq -\mu \end{cases}$$

Thus

$$\mathcal{S}_{\mathcal{R}}(\boldsymbol{x}) = \begin{bmatrix} \mathcal{S}_{w_1}(x_1) \\ \mathcal{S}_{w_2}(x_2) \\ \vdots \\ \mathcal{S}_{w_n}(x_n) \end{bmatrix}.$$

■

2- **SDP. (15 points)**
We wish to solve the following optimization problem

$$\text{minimize} \quad \frac{(\boldsymbol{c}^T\boldsymbol{x})^2}{\boldsymbol{d}^T\boldsymbol{x}} \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x}+\boldsymbol{b} \geq 0,$$

where we assume that $\boldsymbol{d}^T\boldsymbol{x} > 0$ whenever $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b} \geq \boldsymbol{0}$.

(a) Is this optimization problem convex? If yes, justify why. If no, cast it into an equivalent convex optimization problem. (5 points)

(b) How would you solve this problem using an SDP solver. Justify your answer. (10 points)

Solution

(a) Yes, it is convex. Note that the function $f(z,t) = z^2/t$ is convex over (z,t) with $t > 0$. Thus, its pre-composition with a linear function of the form $\boldsymbol{x} \to (\boldsymbol{c}^T\boldsymbol{x}, \boldsymbol{d}^T\boldsymbol{x})$ remains convex.

(b) Using the epigraph trick we can rewrite the above problem in the form

$$\begin{aligned} &\text{minimize} && t \\ &\text{subject to} && \boldsymbol{A}\boldsymbol{x}+\boldsymbol{b} \geq 0, \\ & && \frac{(\boldsymbol{c}^T\boldsymbol{x})^2}{\boldsymbol{d}^T\boldsymbol{x}} \leq t. \end{aligned}$$

Using the Shur trick from class the latter is equivalent to

$$\begin{aligned} &\text{minimize} && t \\ &\text{subject to} && \begin{bmatrix} \mathbf{diag}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}) & 0 & 0 \\ 0 & t & \boldsymbol{c}^T\boldsymbol{x} \\ 0 & \boldsymbol{c}^T\boldsymbol{x} & \boldsymbol{d}^T\boldsymbol{x}. \end{bmatrix} \succeq \boldsymbol{0}. \end{aligned}$$

■

3- **Its all dual. (25 pts)**
Consider the following optimization problem

$$\text{minimize} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad g(\boldsymbol{x}) \leq 0.$$

where

$$f(\boldsymbol{x}) = f(x_1, x_2) = e^{-x_1} \quad \text{and} \quad g(\boldsymbol{x}) = g(x_1, x_2) = \frac{x_1^2}{x_2} \quad \text{with domain} \quad \mathcal{X} = \{(x_1, x_2) | x_2 > 0\}.$$

Please note that the optimization variables are x_1 and x_2.

(a) Is the function g convex over $\mathcal{X}$?

(b) What is the primal optimal value?

(c) Derive the dual optimization problem. Note that $\mathcal{X}$ is an implicit constraint. When creating the dual do not make it explicit.

(d) What is the optimal value of the dual optimization problem. Does strong duality hold? Why or why not?

(e) Consider the perturbed problem

$$\text{minimize} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad g(\boldsymbol{x}) \leq u,$$

with $u > 0$. What is the duality gap?

Solution

(a) Yes, quadratic over linear is convex when $x_2 > 0$. If you would like a proof note that

$$\nabla^2 g(\boldsymbol{x}) = \begin{bmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{bmatrix} = \frac{2}{x_2^3} \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix} \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix}^T \geq 0.$$

(b) Since $x_1^2/x_2 \leq 0$ and $x_2 > 0$ then all feasible solutions are (x_1, x_2) with $x_1 = 0$ and $x_2 > 0$. The objective only depends on x_1 and so the optimal value is simply $e^0 = 1$.

(c) The Lagrangian is given by

$$\mathcal{L}(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}) = \begin{cases} +\infty & \text{if } x_2 \leq 0 \\ e^{-x_1} + \lambda \frac{x_1^2}{x_2} & \text{if } x_2 > 0 \end{cases}$$

The Lagrangian obtains it's minimum value of 0 when $x_2 = x_1^3$ and $x_1 \to \infty$. Thus,

$$g(\lambda) = 0.$$

Therefore, the dual problem is given by

$$\max_{\lambda \geq 0} \quad g(\lambda) = \max_{\lambda \geq 0} \quad 0.$$

(d) The primal optimal value is equal to 1 and dual optimal value is equal to zero. Therefore, strong duality does not hold. We expect this to happen as Slater's condition does not hold in this case.

(e) A short answer that would have been sufficient is that Slater's condition holds and therefore we automatically have strong duality with zero duality gap.

A longer answer is that in this case the primal optimization problem is of the form

$$\text{minimize} \quad e^{-x_1} \quad \text{subject to} \quad x_1^2 \leq x_2 u.$$

Now note that for $x_1 = t$ with $t > 0$ and $x_2 = t^2/u$ the objective value is equal to e^{-t} and the constraint is satisfied. By making t arbitrary large the primal objective can be made arbitrary small so that the primal optimal value is equal to zero.

The Lagrangian is given by

$$\mathcal{L}(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda \left(g(\boldsymbol{x}) - u\right) = \begin{cases} +\infty & \text{if } x_2 \leq 0 \\ e^{-x_1} + \lambda \left(\frac{x_1^2}{x_2} - u \right) & \text{if } x_2 > 0 \end{cases}$$

The Lagrangian converges to $-\lambda u$ when $x_2 = x_1^3$ and $x_1 \to \infty$. Thus,

$$g(\lambda) = -\lambda u.$$

So dual problem is

$$\max_{\lambda \geq 0} \quad g(\lambda) = \max_{\lambda \geq 0} \quad -\lambda u.$$

Therefore, the dual optima value is equal to 0. Thus, strong duality does indeed hold and the duality gap is equal to 0.

■

4- **Fixed sum of distances from k given points. (15 pts)**
Fix a positive real number d and fix k distinct points $(u_1, v_1), (u_2, v_2), \ldots, (u_k, v_k)$ in $\mathbb{R}^2$. The generalized k-ellipse with foci (u_i, v_i) and radius d is the following curve in the plane:

$$\left\{(x, y) \in \mathbb{R}^2 : \sum_{i=1}^{k} \sqrt{(x - u_i)^2 + (y - v_i)^2} = d\right\}$$

Figure 1 below depicts such generalized ellipses.

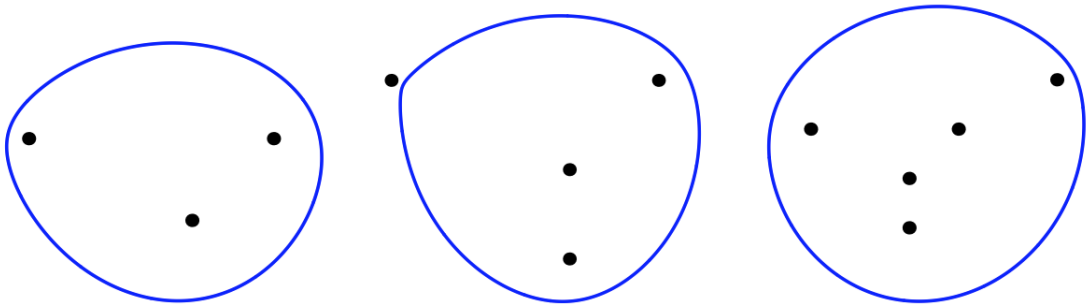


Figure 1: A 3-ellipse, a 4-ellipse, and a 5-ellipse, each with its foci.

(a) Is the set ϵ_k defined below convex? Justify your answer. (5 points)

$$\mathcal{E}_k := \left\{(x, y) \in \mathbb{R}^2 : \sum_{i=1}^{k} \sqrt{(x - u_i)^2 + (y - v_i)^2} \leq d\right\}$$

(b) How would you find a point at the intersection of two such generalized ellipses using an SDP solver. Justify your answer. (10 points)

Solution

(a) Yes, it is convex. Consider the vector $\boldsymbol{z} = \begin{bmatrix} x \\ y \end{bmatrix}$ and vectors $\boldsymbol{p}_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix}$. Then the set $\mathcal{E}^{(k)}$ can be rewritten in the form

$$\mathcal{E}_k := \left\{\boldsymbol{z} \in \mathbb{R}^2 : f(\boldsymbol{z}) \leq d\right\} \quad \text{with} \quad f(\boldsymbol{z}) = \sum_{i=1}^{k} \|\boldsymbol{z} - \boldsymbol{p}_i\|_{\ell_2}.$$

Note that $f(\boldsymbol{z})$ is a sum of convex functions and hence convex. Therefore, its sub-level sets including $\mathcal{E}_k$ is also convex.

(b) Let us consider the two k-ellipses

$$\mathcal{E}_k := \left\{ \boldsymbol{z} \in \mathbb{R}^2 : f(\boldsymbol{z}) \leq d \right\} \quad \text{with} \quad f(\boldsymbol{z}) = \sum_{i=1}^{k} \|\boldsymbol{z} - \boldsymbol{p}_i\|_{\ell_2} .$$

and

$$\tilde{\mathcal{E}}_k := \left\{ \boldsymbol{z} \in \mathbb{R}^2 : \tilde{f}(\boldsymbol{z}) \leq \tilde{d} \right\} \quad \text{with} \quad \tilde{f}(\boldsymbol{z}) = \sum_{i=1}^{k} \|\boldsymbol{z} - \tilde{\boldsymbol{p}}_i\|_{\ell_2} .$$

To find a point at the intersection (or show non-exists) it suffices to solve the following optimization problem

$$\min_{\boldsymbol{z} \in \mathbb{R}^2} \quad 0 \quad \text{subject to} \quad f(\boldsymbol{z}) \leq d \quad \text{and} \quad \tilde{f}(\boldsymbol{z}) \leq \tilde{d}.$$

The optimal value will be zero if a point exists at the intersection and $+\infty$ if such a point does not exist. To be able to solve this via SDP we only need to cast the constraints $f(\boldsymbol{z}) \leq d$ and $\tilde{f}(\boldsymbol{z}) \leq \tilde{d}$ as an LMI.

To this aim note that using the epigraph trick we have

$$\sum_{i=1}^{k} \|\boldsymbol{z} - \boldsymbol{p}_i\|_{\ell_2} \leq d \quad \Leftrightarrow \quad \begin{cases} \|\boldsymbol{z} - \boldsymbol{p}_i\|_{\ell_2} \leq t_i & \text{for } i = 1, 2, \ldots, k \\ \sum_{i=1}^{k} t_i \leq d. \end{cases}$$

Using Shur's trick the latter can be written in the form

$$\sum_{i=1}^{k} t_i \leq d \quad \text{and} \quad \begin{bmatrix} t_i \boldsymbol{I}_{2\times 2} & \boldsymbol{z} - \boldsymbol{p}_i \\ (\boldsymbol{z} - \boldsymbol{p}_i)^T & t_i \end{bmatrix} \succeq \boldsymbol{0} \quad \text{for } i = 1, 2, \ldots, k.$$

Putting everything together we arrive at

$$\begin{aligned} \min_{\boldsymbol{z} \in \mathbb{R}^2, \boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^k} \quad & 0 \\ \text{subject to} \quad & \boldsymbol{1}^T \boldsymbol{t} \leq d \quad \text{and} \quad \boldsymbol{1}^T \boldsymbol{s} \leq \tilde{d}, \\ & \begin{bmatrix} t_i \boldsymbol{I}_{2\times 2} & \boldsymbol{z} - \boldsymbol{p}_i \\ (\boldsymbol{z} - \boldsymbol{p}_i)^T & t_i \end{bmatrix} \succeq \boldsymbol{0} \quad \text{for } i = 1, 2, \ldots, k \\ & \begin{bmatrix} s_i \boldsymbol{I}_{2\times 2} & \boldsymbol{z} - \tilde{\boldsymbol{p}}_i \\ (\boldsymbol{z} - \tilde{\boldsymbol{p}}_i)^T & s_i \end{bmatrix} \succeq \boldsymbol{0} \quad \text{for } i = 1, 2, \ldots, k \end{aligned}$$

which is an SDP.

■

5- **Structured parameter estimation from nonlinear measurements. (25 pts)** We wish to recover an unknown signal $\boldsymbol{x}$ from nonlinear measurements of the form

$$y_r = |\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle|^2 + w_r.$$

Here, $\boldsymbol{a}_r \in \mathbb{R}^n$ are known measurement vectors, y_r are the measurements and w_r are the noise in the measurements. We assume that the noise w_r are distributed i.i.d. according to a Poisson noise model. Specifically the measurements are generated i.i.d. according to the following distribution

$$\text{prob}(y_r = k) = \frac{e^{-\mu_r} \mu_r^k}{k!} \quad \text{with} \quad \mu_r = \left(\langle \boldsymbol{a}_r, \boldsymbol{x} \rangle\right)^2,$$

with $\boldsymbol{x} \in \mathbb{R}^n$ denoting the unknown signal.

(a) Write down the objective function $\mathcal{L}(\boldsymbol{z})$ with variable $\boldsymbol{z}$ that you would minimize to find the signal $\boldsymbol{x}$ (Note that the objective function does not have to be convex). (5 points)

(b) In practice, due to various practical considerations (e.g. that taking measurements are time consuming) we do not have a lot of measurements in the sense that $m << n$ so that just minimizing the objective function in part (a) may not be enough as there may be many global optima for the objective in part (a). Luckily however we have a lot of side information that we can utilize. In particular, in our problem we know that an orthonormal rotation of the signal vector $\boldsymbol{x}$ is sparse. Specifically, $\boldsymbol{U}^T \boldsymbol{x}$ has only a few non-zero entries with $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ a known orthonormal matrix (i.e. obeying $\boldsymbol{U}^T \boldsymbol{U} = \boldsymbol{I}$). Based on our side information we wish to solve one of the two following optimization problems. (Below $\tau, \lambda \geq 0$ are known tuning parameters)

$$\min_{\boldsymbol{z} \in \mathbb{R}^p} \quad \mathcal{L}(\boldsymbol{z}) \quad \text{subject to} \quad \left\| \boldsymbol{U}^T \boldsymbol{z} \right\|_{\ell_1} \leq \tau, \tag{0.1}$$

$$\min_{\boldsymbol{z} \in \mathbb{R}^p} \quad \mathcal{L}(\boldsymbol{z}) + \lambda \left\| \boldsymbol{U}^T \boldsymbol{z} \right\|_{\ell_1}. \tag{0.2}$$

What are the iterative updates when solving (0.1) via the Frank-Wolfe algorithm? (5 points)

(c) What are the iterative updates when solving (0.2) via ISTA? (Hint: use the case of $\boldsymbol{U} = \boldsymbol{I}$ as a starting point) (5 points)

(d) What changes would you make to the iterations of part (c) to make the algorithm converge faster. Please write down the corresponding iterations. (5 points)

(e) How would you solve (0.2) using a stochastic algorithm (i.e. a stochastic version of part (c)). Please write down the iterations. (5 points)

Solution

(a) We need to minimize the negative log-likelihood. Following page 354 of the book we arrive at the objective

$$\mathcal{L}(\boldsymbol{z}) = \sum_{r=1}^{m} |\langle \boldsymbol{a}_r, \boldsymbol{z}\rangle|^2 - y_r \log\left(|\langle \boldsymbol{a}_r, \boldsymbol{z}\rangle|^2\right)$$

(b) Note that

$$\nabla\mathcal{L}(\boldsymbol{z}) = 2\left(\sum_{r=1}^{m}\left(1 - \frac{y_r}{|\langle \boldsymbol{a}_r, \boldsymbol{z}\rangle|^2}\right)\boldsymbol{a}_r\boldsymbol{a}_r^T\right)\boldsymbol{z}.$$

Also define

$$\mathcal{K} = \left\{\boldsymbol{z} \in \mathbb{R}^n \;\middle|\; \left\|\boldsymbol{U}^T\boldsymbol{z}\right\|_{\ell_1} \le \tau\right\}.$$

Note that the Frank-Wolfe iterations are given by

$$\begin{aligned}\boldsymbol{u}_t &\in \arg\min_{\boldsymbol{u}\in\mathcal{K}} \; \boldsymbol{u}^T\nabla\mathcal{L}(\boldsymbol{z}_t),\\ \boldsymbol{z}_{t+1} &= \frac{t-1}{t+1}\boldsymbol{z}_t + \frac{2}{t+1}\boldsymbol{u}_t.\end{aligned}$$

To calculate $\boldsymbol{u}_t$ note that

$$\begin{aligned}\boldsymbol{u}_t \in \arg\min_{\boldsymbol{u}\in\mathcal{K}} \; \boldsymbol{u}^T\nabla f(\boldsymbol{z}_t) &= \arg\min_{\boldsymbol{u}\in\mathcal{K}} \; \boldsymbol{u}^T\nabla\mathcal{L}(\boldsymbol{z}_t)\\ &= \arg\min_{\boldsymbol{u}:\|\boldsymbol{U}^T\boldsymbol{u}\|_{\ell_1}\le\tau} \; \boldsymbol{u}^T\nabla\mathcal{L}(\boldsymbol{z}_t)\\ &= \arg\min_{\boldsymbol{u}:\|\boldsymbol{U}^T\boldsymbol{u}\|_{\ell_1}\le\tau} \; \boldsymbol{u}^T\boldsymbol{U}\boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{z}_t)\\ &= \arg\min_{\boldsymbol{u}:\|\boldsymbol{U}^T\boldsymbol{u}\|_{\ell_1}\le\tau} \; \langle\boldsymbol{U}^T\boldsymbol{u}, \boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{z}_t)\rangle\\ &= \boldsymbol{U}\cdot\arg\min_{\boldsymbol{w}:\|\boldsymbol{w}\|_{\ell_1}\le\tau} \; \langle\boldsymbol{w}, \boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{z}_t)\rangle\\ &= -\tau\,\mathrm{sgn}([\boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{z}_t)]_i)(\boldsymbol{U}\boldsymbol{e}_{i_{\max}})\\ &= -\tau\,\mathrm{sgn}([\boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{z}_t)]_i)\boldsymbol{U}_{i_{\max}}\end{aligned}$$

where

$$i_{\max} = \arg\max_i \; \left|[\boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{z}_t)]_i\right|.$$

Thus the iterations are equal to

$$\begin{aligned}\boldsymbol{u}_t &= -\tau\,\mathrm{sgn}([\boldsymbol{U}^T\nabla\mathcal{L}(\boldsymbol{z}_t)]_i)\boldsymbol{U}_{i_{\max}},\\ \boldsymbol{z}_{t+1} &= \frac{t-1}{t+1}\boldsymbol{z}_t + \frac{2}{t+1}\boldsymbol{u}_t.\end{aligned}$$

(c) We rewrite (0.2) as

$$\min_{\boldsymbol{z}\in\mathbb{R}^n} \frac{1}{\lambda}\mathcal{L}(\boldsymbol{z}) + \mathcal{R}(\boldsymbol{z}) \quad \text{with} \quad \mathcal{R}(\boldsymbol{z}) = \left\|\boldsymbol{U}^T\boldsymbol{z}\right\|_{\ell_1}.$$

The iterations of ISTA are of the form

$$\boldsymbol{z}_{t+1} = \text{prox}_{\mu\mathcal{R}}\left(\boldsymbol{z}_t - \mu\nabla\mathcal{L}(\boldsymbol{z}_t)\right),$$

where

$$\text{prox}_{\mu\mathcal{R}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \frac{1}{2}\left\|\boldsymbol{z}-\boldsymbol{x}\right\|_{\ell_2}^2 + \mu\mathcal{R}(\boldsymbol{z}).$$

Note for $\mathcal{R}(\boldsymbol{z}) = \left\|\boldsymbol{U}^T\boldsymbol{z}\right\|_{\ell_1}$ we have

$$\begin{aligned}
\text{prox}_{\mu\mathcal{R}}(\boldsymbol{x}) =& \arg\min_{\boldsymbol{z}} \frac{1}{2}\left\|\boldsymbol{z}-\boldsymbol{x}\right\|_{\ell_2}^2 + \mu\left\|\boldsymbol{U}^T\boldsymbol{z}\right\|_{\ell_1}, \\
=& \arg\min_{\boldsymbol{z}} \frac{1}{2}\left\|\boldsymbol{U}^T\left(\boldsymbol{z}-\boldsymbol{x}\right)\right\|_{\ell_2}^2 + \mu\left\|\boldsymbol{U}^T\boldsymbol{z}\right\|_{\ell_1}, \\
=& \arg\min_{\boldsymbol{z}} \frac{1}{2}\left\|\boldsymbol{U}^T\boldsymbol{z}-\boldsymbol{U}^T\boldsymbol{x}\right\|_{\ell_2}^2 + \mu\left\|\boldsymbol{U}^T\boldsymbol{z}\right\|_{\ell_1}, \\
=& \boldsymbol{U}\cdot\arg\min_{\boldsymbol{w}} \frac{1}{2}\left\|\boldsymbol{w}-\boldsymbol{U}^T\boldsymbol{x}\right\|_{\ell_2}^2 + \mu\left\|\boldsymbol{w}\right\|_{\ell_1}, \\
=& \boldsymbol{U}\cdot\mathcal{S}_\mu(\boldsymbol{U}^T\boldsymbol{x}),
\end{aligned}$$

where $\mathcal{S}_\mu$ is the soft-thresholding operator and the prox of $\mu\left\|\cdot\right\|_{\ell_1}$ as discussed in class. Thus the ISTA iterations are

$$\boldsymbol{z}_{t+1} = \boldsymbol{U}\cdot\mathcal{S}_\mu\left(\boldsymbol{U}^T\left(\boldsymbol{z}_t - \frac{\mu}{\lambda}\nabla\mathcal{L}(\boldsymbol{z}_t)\right)\right).$$

(d) In this case we would run FISTA instead of ISTA. The iterations of FISTA are of the form

$$\begin{aligned}
\boldsymbol{v}_t =& \boldsymbol{z}_t - \beta\left(\boldsymbol{z}_t - \boldsymbol{z}_{t-1}\right) \\
\boldsymbol{z}_{t+1} =& \text{prox}_{\mu\mathcal{R}}\left(\boldsymbol{v}_t - \frac{\mu}{\lambda}\nabla\mathcal{L}(\boldsymbol{v}_t)\right).
\end{aligned}$$

More specifically, it takes the form

$$\begin{aligned}
\boldsymbol{v}_t =& \boldsymbol{z}_t - \beta\left(\boldsymbol{z}_t - \boldsymbol{z}_{t-1}\right), \\
\boldsymbol{z}_{t+1} =& \boldsymbol{U}\cdot\mathcal{S}_\mu\left(\boldsymbol{U}^T\left(\boldsymbol{v}_t - \frac{\mu}{\lambda}\nabla\mathcal{L}(\boldsymbol{v}_t)\right)\right).
\end{aligned}$$

(e) Note that the objective is of the form

$$\sum_{r=1}^{m}\left(|\langle \boldsymbol{a}_r, \boldsymbol{z}\rangle|^2 - y_r \log\left(|\langle \boldsymbol{a}_r, \boldsymbol{z}\rangle|^2\right) + \frac{\lambda}{m}\left\|\boldsymbol{U}^T\boldsymbol{z}\right\|_{\ell_1}\right)$$

Thus we pick a random $\phi_t \in \{1, 2, \ldots, m\}$ and run the iterations

$$\boldsymbol{z}_{t+1} = \boldsymbol{U} \cdot \mathcal{S}_{\mu}\left(\boldsymbol{U}^T\left(\boldsymbol{z}_t - 2\frac{\mu m}{\lambda}\left(1 - \frac{y_r}{\left|\langle \boldsymbol{a}_{\phi_t}, \boldsymbol{z}\rangle\right|^2}\right)(\boldsymbol{a}_{\phi_t}^T\boldsymbol{z}_t)\boldsymbol{a}_{\phi_t}\right)\right).$$

■