Self grades are due at 11 PM on March 22, 2024.

1. **Quadratic inequalities**

Consider the set S defined by the following inequalities:

$$(x_1 \geq -x_2 + 1 \text{ and } x_1 \leq 0) \text{ or } (x_1 \leq -x_2 + 1 \text{ and } x_1 \geq 0)\,. \tag{1}$$

To be more precise,

$$S_1 = \{\vec{x} \in \mathbb{R}^2 \mid x_1 \geq -x_2 + 1, x_1 \leq 0\} \tag{2}$$
$$S_2 = \{\vec{x} \in \mathbb{R}^2 \mid x_1 \leq -x_2 + 1, x_1 \geq 0\} \tag{3}$$
$$S = S_1 \cup S_2. \tag{4}$$

(a) Draw the set S. Is it convex?

Solution:

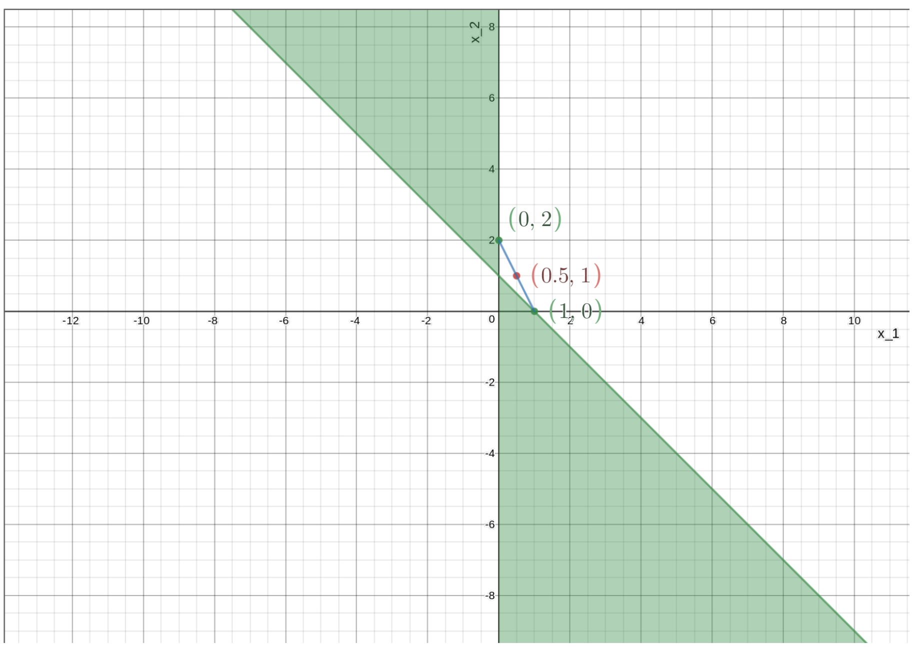

Figure 1: Set S

The set S as shown in Fig. 1 is not convex. We can prove this by counterexample. $(0, 2)$ and $(1, 0)$ both belong to the set, but the midpoint $(1/2, 1)$ does not.

(b) Show that the set S, can be described as a single quadratic inequality of the form $q(\vec{x}) = \vec{x}^\top A\vec{x} + 2\vec{b}^\top \vec{x} + c \leq 0$, for matrix $A = A^\top \in \mathbb{R}^{2\times 2}$, $\vec{b} \in \mathbb{R}^2$ and $c \in \mathbb{R}$ i.e S can be written as $S = \{\vec{x} \in \mathbb{R}^2 \mid q(\vec{x}) \leq 0\}$). Find $A, \vec{b}, c$.

Hint: Can you combine the constraints to make one quadratic constraint?

Solution: Within set S, $x_1 + x_2 - 1 \geq 0$ when $x_1 \leq 0$ and $x_1 + x_2 - 1 \leq 0$ when $x_1 \geq 0$. It follows that $q(\vec{x}) = x_1(x_1 + x_2 - 1) \leq 0$ if and only if it is in the set. Expressing $q(\vec{x})$ in the desired form:

$$q(\vec{x}) = x_1^2 + x_1x_2 - x_1 = \vec{x}^\top A\vec{x} + 2\vec{b}^\top \vec{x} + c$$

where

$$A = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 0 \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} -1/2 \\ 0 \end{bmatrix}, \quad c = 0.$$

(c) Recall the definition of the convex hull of a set $A \subseteq \mathbb{R}^n$ is the set of all convex combinations of points in A, i.e.,

$$\operatorname{conv}(A) = \left\{ \sum_{i=1}^{k} \theta_i \vec{x}_i \;\middle|\; k \in \mathbb{N}, \theta_1, \ldots, \theta_k \geq 0, \sum_{i=1}^{k} \theta_i = 1, \vec{x}_1, \ldots, \vec{x}_k \in A \right\}. \tag{5}$$

What is the convex hull of S?

Solution: The convex hull of the set is the whole space, $\mathbb{R}^2$. To see this note than any point $z = (z_1, z_2) \in \mathbb{R}^2$ can be written as $z = \frac{x+y}{2}$ with $x, y \in S$ as follows:
$x = (2z_1, 1 - 2z_1), y = (0, 2(z_1 + z_2) - 1)$.

(d) We will now consider some convex optimization problems over S_1 that illustrate the role of the constraints in the optimization problem. For each of the following optimization problems find the optimal point, $\vec{x}^*$. Describe the constraints that are active in attaining the optimal value. *Hint: Suppose that there exists a point $\vec{x}$ such that $\nabla f(\vec{x}) = 0$. From the first order characterization of a convex function $\vec{x}$ would be an optimum value for f subject to no constraints. If $\vec{x}$ is not in the constraint set S_1, then the optimum point must be on the boundary of the set, i.e. it satisfies at least one of the constraints defining S_1 with equality.*

i. Minimize $f(\vec{x}) = (x_1 + 1)^2 + (x_2 - 3)^2$ subject to $\vec{x} \in S_1$.
Solution: We first compute the unconstrained optimal value of f. Notice that f is a convex function. Therefore, we can compute its optimal value by computing its gradient and setting it to 0. Doing so, we obtain the optimal value of f to be 0 attained at the point $\vec{x}^* = (-1, 3)$. Now, since $\vec{x}^* \in S_1$, $\vec{x}^*$ is the solution to the constrained optimization problem as well.

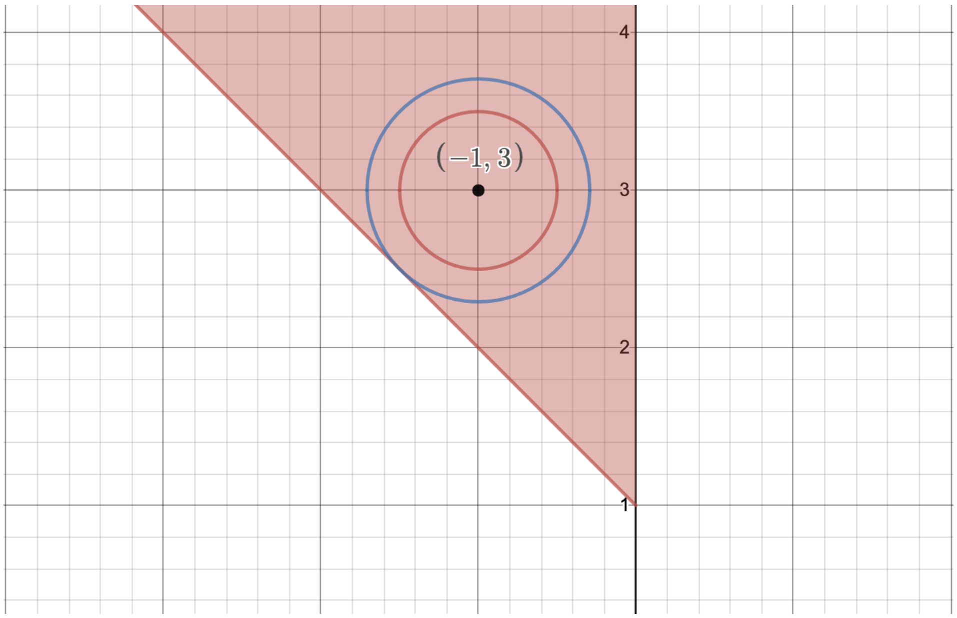

Figure 2: This figure illustrates the position of the optimum, $x^* = (-1, 3)$, and the level sets of the objective function, f, which are concentric circles around x^*.

ii. Minimize $f(\vec{x}) = (x_1 + 2)^2 + (x_2 - 2)^2$ subject to $\vec{x} \in S_1$.

Solution: Proceeding as in the proof for the previous problem, we first find the solution to the unconstrained optimization problem. We get that the unconstrained problem is minimized at the point $\vec{x}_u^* = (-2, 2)$. However, this point is not in the feasible set, S_1. Therefore, the true optimum, $\vec{x}^*$, has one or more constraints active. Now, we will attempt to solve the problem with one active constraint. Suppose the one active constraint is $x_1 \geq -x_2 + 1$. Since this constraint is active, we must try and minimize $f(\vec{x})$ subject to $\vec{x}$ satisfying $x_1 = -x_2 + 1$. Note that any point on this line can be written in the form $(0, 1) + \alpha(-1, 1)$. Now consider the function, $g(\alpha)$:

$$g(\alpha) = f((0, 1) + \alpha(-1, 1)) = (\alpha - 2)^2 + (\alpha - 1)^2.$$

Note that the function, $f(\alpha)$, is convex in α. Therefore, we can minimize $g(\alpha)$ by taking its derivative and setting it to 0. By doing this, we get that $\alpha = 3/2$ is the unique minimizer of $g(\alpha)$. Therefore, the minimizer of f subject to $x_1 = -x_2 + 1$ is the point $(-3/2, 5/2)$, and the function value is 0.5. Similarly, the minimizer of f assuming the second constraint, $x_1 \leq 0$, is active is obtained at the point $(0, 2)$, and the function value at this point is 4, which is higher than the value at $(-3/2, 5/2)$. The final possibility is that both constraints are active. However, the optimal value of f subject to both constraints being active will be greater than the value of f obtained at $(-3/2, 5/2)$ which is in S_1. Therefore, we get that $f(\vec{x})$ is minimized at the point $\vec{x}^* = (-3/2, 5/2)$ subject to $\vec{x} \in S_1$. There is one active constraint at $\vec{x}^*$.

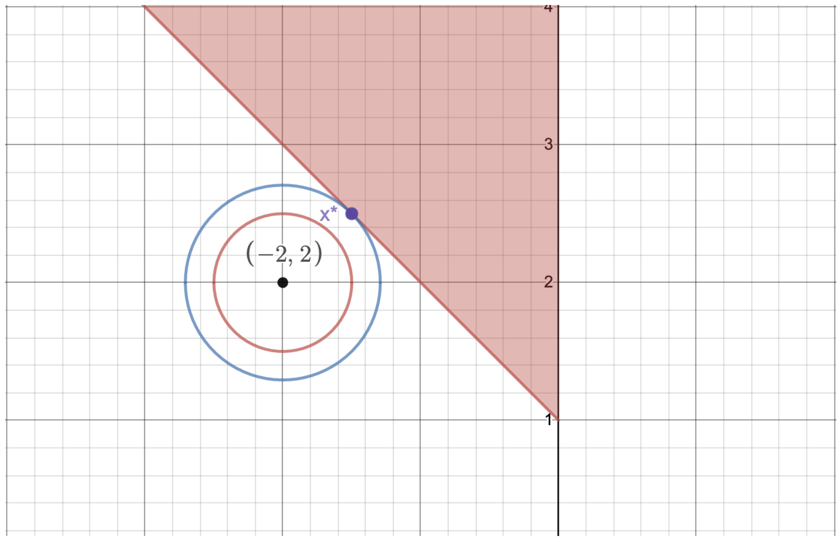

Figure 3: This figure illustrates the position of the optimum, $x^* = (-3/2, 5/2)$, and the level sets of the objective function, f, which are concentric circles around $(-2, 2)$. Note that in this case, the unconstrained optimum does not lie in the set, S_1 and the optimal point lies on the boundary of one of the constraints.

iii. Minimize $f(\vec{x}) = x_1^2 + x_2^2$ subject to $\vec{x} \in S_1$.

Solution: Proceeding as before, we first check the case where 0 constraints are active. However, the unconstrained minimizer of f is $(0, 0)$ which is not in S_1. Now, we check the cases where one of the constraints is active. Assume that the constraint $x_1 \leq 0$ is active. In this case the optimizer is again obtained at the point $(0, 0)$ which is not in S_1. We then consider the case where the constraint $x_1 \geq -x_2 + 1$ is active. As before, we define the function, $g(\alpha)$ as:

$$g(\alpha) = f((0, 1) + \alpha(-1, 1)) = \alpha^2 + (\alpha + 1)^2.$$

By optimizing over α by setting its gradient with respect to α and setting it to 0, we get the optimal setting of α is $-1/2$. However, note that the point $(1/2, 1/2)$ does not belong to S_1 either. Therefore, the only remaining possibility is the possibility that both constraints are active. This can happen solely at the point $(0, 1)$. At this point, the value of the function f is 1, the optimizer $\vec{x}^* = (0, 1)$ and both constraints are active at $\vec{x}^*$.

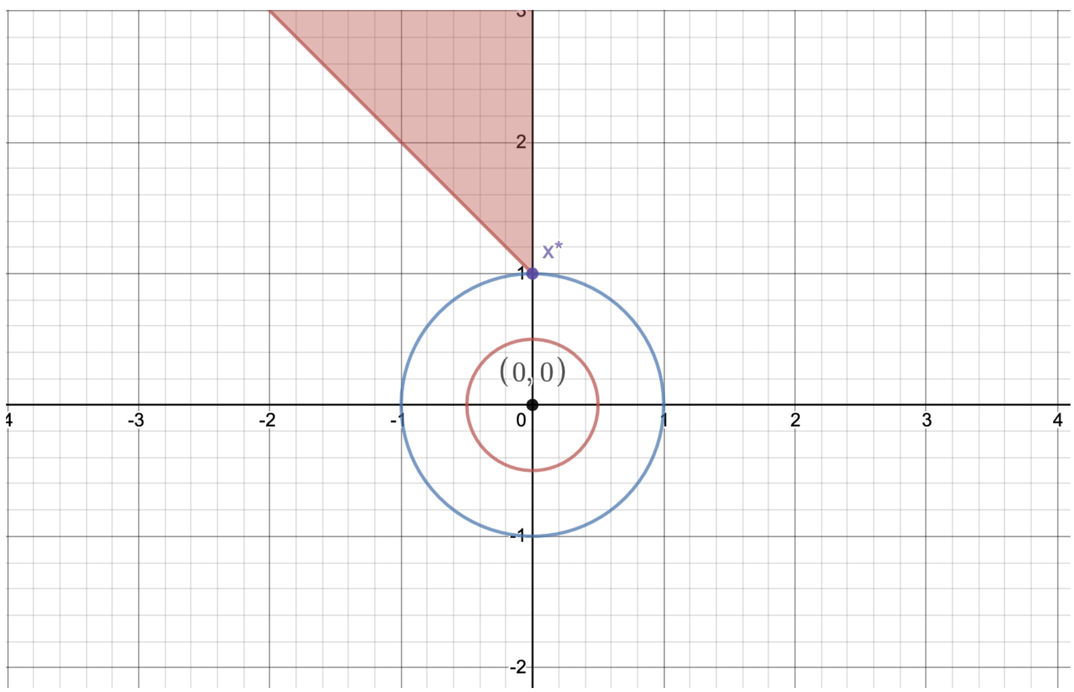

Figure 4: This figure illustrates the position of the optimum, $x^* = (0, 1)$, and the level sets of the objective function, f, which are concentric circles around $(0, 0)$. Note that in this case, the unconstrained optimum does not lie in the set, S_1 and the optimal point lies on the boundary of *both* of the constraints.

2. Direction of Steepest Ascent

For a differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ we want to show that the gradient $\nabla f(\vec{x})$ is the direction of steepest ascent at the point $\vec{x}$.

(a) Let us define the rate of change of the function $f(\vec{x})$ at the point $\vec{x}$ along an arbitrary unit vector $\vec{u}$ as:

$$D_{\vec{u}} f(\vec{x}) = \lim_{h \to 0} \frac{f(\vec{x} + h\vec{u}) - f(\vec{x})}{h}. \tag{6}$$

We call this the directional derivative. Show that the directional derivative can be equivalently expressed as $D_{\vec{u}} f(\vec{x}) = \vec{u}^\top [\nabla f(\vec{x})]$.

HINT: Use Taylor approximation of the function around the point $\vec{x}$ and evaluate it at the point $\vec{x} + h\vec{u}$.

Solution: Using Taylor's theorem we can express the function $f(\vec{x})$ as

$$f(\vec{x} + h\vec{u}) = f(\vec{x}) + [\nabla f(\vec{x})]^\top [h\vec{u}] + o(h). \tag{7}$$

We rearrange the terms, and dividing both sides by h we get

$$\frac{f(\vec{x} + h\vec{u}) - f(\vec{x})}{h} = [\nabla f(\vec{x})]^\top [\vec{u}] + \frac{o(h)}{h}. \tag{8}$$

Now we take the limit of both sides as $h \to 0$; we get

$$\begin{aligned} \lim_{h \to 0} \frac{f(\vec{x} + h\vec{u}) - f(\vec{x})}{h} &= [\nabla f(\vec{x})]^\top [\vec{u}] + \lim_{h \to 0} \left(\frac{o(h)}{h} \right) && (9) \\ &= [\nabla f(\vec{x})]^\top [\vec{u}]. && (10) \end{aligned}$$

Note that $\lim_{h \to 0} \frac{o(h)}{h} = 0$ because $o(h)$ decays faster than h as $h \to 0$.

(b) Show that

$$\frac{\nabla f(\vec{x})}{\|\nabla f(\vec{x})\|_2} = \underset{\|\vec{u}\|_2 = 1}{\operatorname{argmax}} \, \vec{u}^\top [\nabla f(\vec{x})]. \tag{11}$$

Solution: Using Cauchy-Schwarz inequality we can write:

$$\begin{aligned} \vec{u}^\top [\nabla f(\vec{x})] &\le \|\vec{u}\|_2 \|\nabla f(\vec{x})\|_2 && (12) \\ &= \|\nabla f(\vec{x})\|_2, && (13) \end{aligned}$$

so the maximum value that the expression can take is $\|\nabla f(\vec{x})\|_2$. Now it remains to show that this value is attained for the choice $\vec{u} = \frac{\nabla f(\vec{x})}{\|\nabla f(\vec{x})\|_2}$.

$$\begin{aligned} \frac{[\nabla f(\vec{x})]^\top}{\|\nabla f(\vec{x})\|_2} \nabla f(\vec{x}) &= \frac{\|\nabla f(\vec{x})\|_2^2}{\|\nabla f(\vec{x})\|_2} && (14) \\ &= \|\nabla f(\vec{x})\|_2. && (15) \end{aligned}$$

3. Convergence of Gradient Descent for Ridge Regression

Let $A \in \mathbb{R}^{m \times n}$, $\vec{y} \in \mathbb{R}^m$, and $\lambda > 0$. Consider a slight variation of the ridge regression problem where the least squares loss is normalized by the number of data points:

$$\min_{\vec{x} \in \mathbb{R}^n} f_\lambda(\vec{x}) \qquad \text{where} \qquad f_\lambda(\vec{x}) \doteq \frac{1}{2}\left\{\frac{1}{m}\|A\vec{x} - \vec{y}\|_2^2 + \lambda\|\vec{x}\|_2^2\right\}. \tag{16}$$

In this problem, we will examine the behavior of gradient descent (GD) on this problem, and in particular the interplay between the learning rate $\eta > 0$ and regularization parameter $\lambda > 0$ in determining convergence.

(a) Show that the unique solution to the problem in Equation (16) is

$$\vec{x}_\lambda^\star = \left(A^\top A + \lambda m I\right)^{-1} A^\top \vec{y}. \tag{17}$$

Solution: The function $\lambda\|\vec{x}\|_2^2$ in Equation (16) is strictly convex, so f_λ is strictly convex. Thus the problem in Equation (16) has strictly convex objective and convex feasible set $\mathbb{R}^n$, so it has at most one solution. And we can find a solution by setting the gradient to $\vec{0}$:

$$\begin{aligned}
\nabla f_\lambda(\vec{x}_\lambda^\star) &= \frac{1}{m}A^\top(A\vec{x}_\lambda^\star - \vec{y}) + \lambda\vec{x}_\lambda^\star \qquad (18)\\
&= \left(\frac{A^\top A}{m} + \lambda I\right)\vec{x}_\lambda^\star - \frac{1}{m}A^\top\vec{y} \qquad (19)\\
\implies \left(\frac{A^\top A}{m} + \lambda I\right)\vec{x}_\lambda^\star &= \frac{1}{m}A^\top\vec{y} \qquad (20)\\
\implies \left(A^\top A + \lambda m I\right)\vec{x}_\lambda^\star &= A^\top\vec{y} \qquad (21)\\
\implies \vec{x}_\lambda^\star &= \left(A^\top A + \lambda m I\right)^{-1} A^\top\vec{y}. \qquad (22)
\end{aligned}$$

(b) Show that the GD update

$$\vec{x}_{t+1} = \vec{x}_t - \eta\left(\frac{1}{m}A^\top(A\vec{x}_t - \vec{y}) + \lambda\vec{x}_t\right) \tag{23}$$

can be rearranged into the form

$$\vec{x}_{t+1} - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)(\vec{x}_t - \vec{x}_\lambda^\star). \tag{24}$$

Use this to show that

$$\vec{x}_t - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)^t(\vec{x}_0 - \vec{x}_\lambda^\star). \tag{25}$$

for every positive integer t.

Solution: We have

$$\begin{aligned}
\vec{x}_{t+1} &= \vec{x}_t - \eta\left(\frac{1}{m}A^\top(A\vec{x}_t - \vec{y}) + \lambda\vec{x}_t\right) \qquad (26)\\
&= \vec{x}_t - \eta\cdot\frac{A^\top A}{m}\vec{x}_t + \eta\cdot\frac{1}{m}A^\top\vec{y} + \eta\lambda\vec{x}_t \qquad (27)\\
&= \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)\vec{x}_t + \eta\cdot\frac{1}{m}A^\top\vec{y} \qquad (28)
\end{aligned}$$

$$\implies \vec{x}_{t+1} - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)\vec{x}_t + \eta \cdot \left(\frac{A^\top A}{m} + \lambda I\right)\vec{x}_\lambda^\star - \vec{x}_\lambda^\star \tag{29}$$

$$= \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)(\vec{x}_t - \vec{x}_\lambda^\star). \tag{30}$$

Iterating this relation obtains the second equality.

(c) We now discuss the insight that the SVD can give us regarding the convergence of GD. Let $A = U\Sigma V^\top$ be a full SVD of A. Let $\vec{z}_t = V^\top \vec{x}_t$ and $\vec{z}_\lambda^\star = V^\top \vec{x}_\lambda^\star$. Show that

$$\vec{z}_t - \vec{z}_\lambda^\star = \left(I - \eta\left(\frac{\Sigma^\top\Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_\lambda^\star), \tag{31}$$

and, moreover, show that for each $i \in \{1, \ldots, n\}$, we have

$$(\vec{z}_t)_i - (\vec{z}_\lambda^\star)_i = \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t ((\vec{z}_0)_i - (\vec{z}_\lambda^\star)_i) \tag{32}$$

where $\sigma_i\{A\}$ is the i^{th} largest singular value of A. This shows that the *rate of convergence* of $\vec{z}_t$ to $\vec{z}_\lambda^\star$ along the i^{th} component is influenced by the interaction between $\sigma_i\{A\}$, λ, and η, but critically no other singular values. Thus, one considers the V basis to be the "natural" basis for thinking about GD for ridge regression.

Solution: If $A = U\Sigma V^\top$ then

$$A^\top A = V\Sigma^\top U^\top U\Sigma V^\top = V\Sigma^\top\Sigma V^\top. \tag{33}$$

Thus we have

$$\vec{x}_t - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{34}$$

$$= \left(I - \eta\left(\frac{V\Sigma^\top\Sigma V^\top}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{35}$$

$$= \left(I - \eta\left(V\left(\frac{\Sigma^\top\Sigma}{m} + \lambda I\right)V^\top\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{36}$$

$$= \left(I - \eta V\left(\frac{\Sigma^\top\Sigma}{m} + \lambda I\right)V^\top\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{37}$$

$$= \left(V\left(I - \eta\left(\frac{\Sigma^\top\Sigma}{m} + \lambda I\right)\right)V^\top\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star) \tag{38}$$

$$= V\left(I - \eta\left(\frac{\Sigma^\top\Sigma}{m} + \lambda I\right)\right)^t V^\top(\vec{x}_0 - \vec{x}_\lambda^\star) \tag{39}$$

$$\implies V^\top(\vec{x}_t - \vec{x}_\lambda^\star) = \left(I - \eta\left(\frac{\Sigma^\top\Sigma}{m} + \lambda I\right)\right)^t V^\top(\vec{x}_0 - \vec{x}_\lambda^\star) \tag{40}$$

$$\implies \vec{z}_t - \vec{z}_\lambda^\star = \left(I - \eta\left(\frac{\Sigma^\top\Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_\lambda^\star). \tag{41}$$

Now note that the quantity $I - \eta(\frac{\Sigma^\top\Sigma}{m} + \lambda I)$ is a diagonal matrix. Thus we have

$$\vec{z}_t - \vec{z}_\lambda^\star = \left(I - \eta\left(\frac{\Sigma^\top\Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_\lambda^\star) \tag{42}$$

$$\implies (\vec{z}_t - \vec{z}^\star_\lambda)_i = \left(I - \eta\left(\frac{\Sigma^\top\Sigma}{m} + \lambda I\right)\right)^t_i (\vec{z}_0 - \vec{z}^\star_\lambda)_i \tag{43}$$

$$\implies (\vec{z}_t)_i - (\vec{z}^\star_\lambda)_i = \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t ((\vec{z}_0)_i - (\vec{z}^\star_\lambda)_i) \tag{44}$$

(d) Show that $\lim_{t\to\infty} \vec{z}_t = \vec{z}^\star_\lambda$ for all initializations $\vec{x}_0 = V\vec{z}_0$ if and only if

$$\max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1. \tag{45}$$

Use this to show that GD converges for all initializations $\vec{x}_0$ if and only if

$$\eta \in \left(0, \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}\right) \tag{46}$$

where $\sigma_{\max}\{A\} = \sigma_1\{A\}$ is the largest singular value of A.

Solution: We have

$$\lim_{t\to\infty} \vec{z}_t = \vec{z}^\star_\lambda, \qquad \forall \vec{x}_0 \tag{47}$$

$$\iff \lim_{t\to\infty} (\vec{z}_t)_i = (\vec{z}^\star_\lambda)_i, \qquad \forall i \quad \forall \vec{x}_0 \tag{48}$$

$$\iff \lim_{t\to\infty} \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t = 0, \qquad \forall i \tag{49}$$

$$\iff \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1, \qquad \forall i \tag{50}$$

$$\iff \max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1. \tag{51}$$

This proves the first part of the question. The second part of the question follows by noting that

$$\max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1 \tag{52}$$

$$\iff \max_{i\in\{1,\dots,n\}} \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right) < 1 \tag{53}$$

$$\text{and} \quad \min_{i\in\{1,\dots,n\}} \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right) > -1 \tag{54}$$

$$\iff 1 - \eta\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right) < 1 \tag{55}$$

$$\text{and} \quad 1 - \eta\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right) > -1. \tag{56}$$

Now the first equation is always satisfied for $\eta > 0$ and $\lambda > 0$ because $\frac{\sigma_{\min}\{A\}^2}{m} + \lambda > 0$ so $1 - \eta(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda) < 1$. The second equation is satisfied when $\eta < \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}$. Since $\lim_{t\to\infty} \vec{x}_t = \vec{x}^\star_\lambda$ if and only if $\lim_{t\to\infty} \vec{z}_t = \vec{z}^\star_\lambda$, we have that gradient descent converges for all initializations $\vec{x}_0$ if and only if $0 < \eta < \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}$.

(e) **(OPTIONAL)** One way we can derive an "optimal" learning rate $\eta^\star$ is to minimize the largest rate of convergence:

$$\eta^\star \in \operatorname*{argmin}_{\eta\in\mathbb{R}} \max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right|. \tag{57}$$

One important property of $\eta^\star$ is that it makes the rates of convergence $\left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right|$ associated with the largest and smallest singular values of A equal. Use this property to show that

$$\eta^\star = \frac{2m}{\sigma_{\max}\{A\}^2 + \sigma_{\min}\{A\}^2 + 2\lambda m} \tag{58}$$

where $\sigma_{\min}\{A\} = \sigma_n\{A\}$ is the n^{th} largest singular value of A.

NOTE: There are several useful notions of optimal learning rate; this is just one of them.

Solution: We have

$$\left|1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right)\right| = \left|1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)\right| \tag{59}$$

$$1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right) = -\left(1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)\right) \tag{60}$$

$$1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right) = \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right) - 1 \tag{61}$$

$$2 = \eta^\star\left(\frac{\sigma_{\max}\{A\}^2 + \sigma_{\min}\{A\}^2}{m} + 2\lambda\right) \tag{62}$$

$$\eta^\star = \frac{2m}{\sigma_{\max}\{A\}^2 + \sigma_{\min}\{A\}^2 + 2m\lambda}. \tag{63}$$

Here the second equality is the most challenging to derive. It follows from the first inequality by the following reasoning:

- If $1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right)$ and $1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)$ have the same sign, then by the first equality, they must be equal. This means that $\sigma_{\max}\{A\} = \sigma_1\{A\} = \sigma_2\{A\} = \cdots = \sigma_n\{A\} = \sigma_{\min}\{A\}$ and the optimal learning rate $\eta^\star$ sets each rate $1 - \eta^\star\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)$ to 0 simultaneously, ensuring convergence in one step. If both sides are 0 then the second equality holds (because $0 = -0$).
- Otherwise, $1-\eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right)$ and $1-\eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)$ have opposite signs. Since $\sigma_{\max}\{A\} > \sigma_{\min}\{A\}$ (since if they were equal we would be in the first case), we have $1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right) > 1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)$. Thus $1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right)$ must be positive and $1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)$ must be negative. The absolute value of a negative number is its negative, so the second equality follows directly from the first equality.

(f) The attached notebook, `gd_convergence.ipynb`, will examine the computational aspects of GD on ridge regression. Implement the GD and stochastic gradient descent (SGD) functions at the top of the notebook, which are marked with `TODO`s.

(g) Click through the notebook and run the sections $n = 1$, $n = 2$, and $n \gg 2$. Change the values of λ and η and re-run the cells a few times. Write down your observations about how the convergence of GD works under different values of λ and η.

Solution: We know from previous parts that when $\eta \geq \left(\frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}\right)$, gradient descent does not converge. This is also observed in the notebook that when η is increased by too much, gradient descent does not converge. Moreover, when η is too small, it does not converge within the prescribed number of iterations either.

(h) In the sections $n = 1$, $n = 2$, and $n \gg 2$, change the calls to GD to instead call SGD. Write down your observations about how the convergence of SGD works under different values of λ and η. Compare the behavior of GD and SGD.

Solution: It is much harder to tune λ and η to make stochastic gradient descent converge. Due to the randomness, the trajectory is much noisier than gradient descent.

(i) You might have noticed that if we think of convergence in the "last iterate" sense, i.e., $\lim_{T\to\infty} \vec{x}_T = \vec{x}_\lambda^\star$, then *SGD rarely converges*. This is because even if we reach the global optimum, the gradient estimate used by SGD is in general nonzero, and so the iterates end up bouncing around near the optimum. Another different, weaker, notion of convergence under which one might show that SGD actually does converge is convergence "in time average", i.e., $\lim_{T\to\infty} \overline{\vec{x}}_T = \vec{x}_\lambda^\star$ where $\overline{\vec{x}}_T \doteq \frac{1}{T}\sum_{t=1}^{T} \vec{x}_t$. Visualize this by adding the argument `time_avg=True` to each plotting function; the plot will now visualize the sequence of $\overline{\vec{x}}_t$. Re-run the notebook. Write down your observations, especially regarding the stability of SGD and convergence in the last-iterate sense versus the time-average sense.

Solution: When η is set too large, the trajectories don't converge no matter when `time_avg=True` or when `time_avg=False`. On the other hand, when η is not set too large, the trajectories under `time_avg=True` don't bounce around that much near the optimum compared with when `time_avg=False`. Instead, the closer $\vec{x}$ is to the optimum, the more stablized it becomes, which shows that it converges in the limit.

4. Dual of the dual of a linear program

Consider a standard linear program P:

$$\min_{\vec{x}} \quad \vec{c}^\top \vec{x} \tag{64}$$
$$\text{s.t.} \quad A\vec{x} = \vec{b} \tag{65}$$
$$\vec{x} \geq \vec{0}. \tag{66}$$

where $\vec{x}, \vec{c} \in \mathbb{R}^n, \vec{b} \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n}$. $\vec{x} \geq \vec{0}$ means $x_i \geq 0$ for all $i = 1, \ldots, n$.

(a) Formulate the Lagrangian of the problem P, and write the dual problem.

Note: The dual problem should not have the variable $\vec{x}$.

Solution: Denote the dual variable associated with the equality constraint in P by $\vec{\nu} \in \mathbb{R}^m$, and the dual variable associated with the inequality by $\vec{\lambda} \in \mathbb{R}^n$. The Lagrangian of P is given by

$$\mathcal{L}(\vec{x}, \vec{\nu}, \vec{\lambda}) = \vec{c}^\top \vec{x} + \vec{\nu}^\top (A\vec{x} - \vec{b}) + \lambda^\top (-\vec{x}) = -\vec{b}^\top \vec{\nu} + (\vec{c} + A^\top \vec{\nu} - \vec{\lambda})^\top \vec{x}. \tag{67}$$

To get the dual problem, we take the minimum of the Lagrangian with respect to our primal variable $\vec{x}$, i.e.,

$$g(\vec{\nu}) = \min_{\vec{x}} \mathcal{L}(\vec{x}, \vec{\nu}, \vec{\lambda}) = \begin{cases} -\vec{b}^\top \vec{\nu} & \text{if } \vec{c} + A^\top \vec{\nu} - \vec{\lambda} = \vec{0} \\ -\infty & \text{otherwise.} \end{cases} \tag{68}$$

The dual problem is then given by D:

$$\max_{\vec{\nu}, \vec{\lambda}} \quad -\vec{b}^\top \vec{\nu} \tag{69}$$
$$\text{s.t.} \quad \vec{c} + A^\top \vec{\nu} - \vec{\lambda} = \vec{0} \tag{70}$$
$$\vec{\lambda} \geq \vec{0}. \tag{71}$$

We can simplify this further by noting that $\vec{\lambda}$ is constrained to be $\vec{\lambda} \geq \vec{0}$, since it is the dual variable associated with an inequality constraint. Then, $\vec{c} + A^\top \vec{\nu} - \vec{\lambda} = \vec{0} \iff \vec{c} + A^\top \vec{\nu} = \vec{\lambda} \geq \vec{0}$, and we can eliminate $\vec{\lambda}$ to get

$$\max_{\vec{\nu}} \quad -\vec{b}^\top \vec{\nu} \tag{72}$$
$$\text{s.t.} \quad \vec{c} + A^\top \vec{\nu} \geq \vec{0}. \tag{73}$$

(b) Express the dual problem as an equivalent minimization problem. Find the dual of this minimization problem, i.e., the dual of the dual. Compare it to the original linear program formulation.

Solution: Say the optimal value for the dual problem is given by

$$d^* = \max_{\vec{\nu}} \quad -\vec{b}^\top \vec{\nu} \quad \text{s.t.} \quad \vec{c} + A^\top \vec{\nu} \geq \vec{0}. \tag{74}$$

The dual problem can be expressed by an equivalent minimization problem:

$$d^* = -\min_{\vec{\nu}} \quad \vec{b}^\top \vec{\nu} \tag{75}$$
$$\text{s.t.} \quad \vec{c} + A^\top \vec{\nu} \geq 0. \tag{76}$$

We call the minimization problem our new primal problem P'. We can obtain the dual problem of the problem in 74 by dualizing P' while ignoring the negative sign, and then adding it back in at the end. Denote the dual variable associated with the inequality constraint $\vec{c}+A^\top\vec{\nu} \geq 0$ by $\vec{z} \in \mathbb{R}^n$. The Lagrangian for P' is given by

$$\mathcal{L}(\vec{\nu}, \vec{z}) = \vec{b}^\top\vec{\nu} + \vec{z}^\top(-\vec{c} - A^\top\vec{\nu}) = -\vec{c}^\top\vec{z} + (-A\vec{z} + \vec{b})^\top\vec{\nu}. \tag{77}$$

The dual is given by taking the minimum of the Lagrangian over the *primal* variable of the problem we are dualizing, i.e., $\vec{\nu}$.

$$h(\vec{z}) = \min_{\vec{\nu}} \mathcal{L}(\vec{\nu}, \vec{z}) = \begin{cases} -\vec{c}^\top\vec{z} & \text{if } A\vec{z} - \vec{b} = 0 \\ -\infty & \text{otherwise.} \end{cases} \tag{78}$$

The dual problem for P' is then given by

$$\max_{\vec{z}} \quad -\vec{c}^\top\vec{z} \tag{79}$$

$$\text{s.t.} \quad A\vec{z} = \vec{b} \tag{80}$$

$$\vec{z} \geq \vec{0}. \tag{81}$$

Then the dual of the problem obtained in part a) is given by

$$\begin{array}{rc} -\max_{\vec{z}} & -\vec{c}^\top\vec{z} \\ \text{s.t.} & A\vec{z} = \vec{b} \\ & \vec{z} \geq \vec{0} \end{array} \quad = \quad \begin{array}{rc} \min_{\vec{z}} & \vec{c}^\top\vec{z} \\ \text{s.t.} & A\vec{z} = \vec{b} \\ & \vec{z} \geq \vec{0} \end{array} \tag{82}$$

which is the linear problem in (64). This tells us that the dual of the dual of a linear program is the same linear program.

5. Minimizing a Sum of Logarithms

Consider the following problem, which arises in estimation of transition probabilities of a discrete-time Markov chain:

$$p^\star = \max_{\vec{x}\in\mathbb{R}^n} \quad \sum_{i=1}^n \alpha_i \log(x_i) \tag{83}$$

$$\text{s.t.} \quad \vec{x} \geq 0, \quad \vec{1}^\top \vec{x} = c, \tag{84}$$

where $c > 0$ and $\alpha_i > 0$, $i = 1, \ldots, n$. (Recall that if $\vec{x}$ is a vector then by "$\vec{x} \geq 0$" we mean "$x_i \geq 0$ for each i.") We will determine in closed-form a minimizer, and show that the optimal objective value of this problem is

$$p^\star = \alpha \log(c/\alpha) + \sum_{i=1}^n \alpha_i \log(\alpha_i), \tag{85}$$

where $\alpha \doteq \sum_{i=1}^n \alpha_i$. We will show this in a series of steps.

(a) First, express the problem as a minimization problem which has optimal value $p^\star_{\min}$.

Solution: We have

$$\max_{\vec{x}\in\mathbb{R}^n} f_0(\vec{x}) = -\min_{\vec{x}\in\mathbb{R}^n} (-f_0(\vec{x})), \tag{86}$$

so

$$p^\star = -\min_{\vec{x}\in\mathbb{R}^n} \quad \sum_{i=1}^n -\alpha_i \log(x_i) \tag{87}$$

$$\text{s.t.} \quad \vec{x} \geq 0, \quad \vec{1}^\top \vec{x} = c. \tag{88}$$

The minimization problem we now consider is

$$p^\star_{\min} = \min_{\vec{x}\in\mathbb{R}^n} \quad \sum_{i=1}^n -\alpha_i \log(x_i) \tag{89}$$

$$\text{s.t.} \quad \vec{x} \geq 0, \quad \vec{1}^\top \vec{x} = c, \tag{90}$$

so that $p^\star_{\min} = -p^\star$.

(b) In optimization, we often "relax" problems of the form $p^\star_{\min} = \min_{\vec{x}\in\mathcal{X}} f_0(\vec{x})$, i.e., replacing the constraint set $\mathcal{X}$ with a larger constraint set $\mathcal{X}_r$, and instead solving $p^\star_r = \min_{\vec{x}\in\mathcal{X}_r} f_0(\vec{x})$, then showing a connection between $p^\star_{\min}$ and $p^\star_r$. In this problem, a particular relaxation we will use is to replace the equality constraint $\vec{1}^\top \vec{x} = c$ with an inequality constraint $\vec{1}^\top \vec{x} \leq c$.

Show that the relaxed problem has the same optimal value as the original problem, i.e., $p^\star_r = p^\star_{\min}$, and the two problems have the same solutions.

HINT: First argue that $p^\star_r \leq p^\star_{\min}$. Then, suppose for the sake of contradiction that $p^\star_r < p^\star_{\min}$. Let $\vec{x}^r$ be a solution to the relaxed minimization problem which has objective value $p^\star_r$. Consider the vector $\vec{x}$ given by

$$\vec{x} \doteq \begin{bmatrix} c - 1^\top \vec{x}^r + x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{bmatrix}. \tag{91}$$

Show that $\vec{x}$ is feasible for the original problem and has objective value $< p_r^\star$. Argue that this implies $p_{\min}^\star < p_r^\star$ and derive a contradiction. Finally, argue that any solution to the relaxed problem is a solution to the original problem, and vice-versa — you might need to use a construction similar to $\vec{x}$.

Solution: We want to show that the relaxed problem

$$p_r^\star = \min_{\vec{x}\in\mathbb{R}^n} \quad \sum_{i=1}^n -\alpha_i \log(x_i) \tag{92}$$
$$\text{s.t.} \quad \vec{x} \geq 0, \quad \vec{1}^\top \vec{x} \leq c, \tag{93}$$

has the same set of solutions as the original minimization problem. We begin by showing that $p_{\min}^\star = p_r^\star$. Indeed, since the relaxed problem minimizes the same objective function over a larger feasible set, $p_r^\star \leq p_{\min}^\star$. We now show that $p_r^\star \geq p_{\min}^\star$.

Suppose for the sake of contradiction that $\vec{x}^r$ is an optimal solution to the relaxed problem which achieves objective value $p_r^\star < p_{\min}^\star$. If $\vec{x}^r$ were feasible for the original minimization problem, then it would be a better solution than the solutions which achieve $p_{\min}^\star$, which is already a contradiction. Thus, suppose $\vec{x}^r$ is infeasible for the original minimization problem, i.e., $\vec{1}^\top \vec{x}^r < c$. Then consider the following solution vector:

$$\vec{x}^\star \doteq \begin{bmatrix} c - \vec{1}^\top \vec{x}^r + x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{bmatrix}. \tag{94}$$

We claim that this choice of solution vector both fulfills all the constraints of the original problem, and achieves a better optimal value. In both parts, we use a crucial inequality:

$$x_1^\star = \underbrace{(c - \vec{1}^\top \vec{x}^r)}_{>0} + x_1^r > x_1^r. \tag{95}$$

To show that $\vec{x}^\star \geq 0$, we just need to show that the first entry $x_1^\star$ is non-negative. This is given by $x_1^\star > x_1^r \geq 0$.

To show that $\vec{1}^\top \vec{x}^\star = c$, we calculate:

$$\begin{aligned}
\vec{1}^\top \vec{x}^\star &= \sum_{i=1}^n x_i^\star && (96)\\
&= x_1^\star + \sum_{i=2}^n x_i^\star && (97)\\
&= c - \vec{1}^\top \vec{x}^r + x_1^r + \sum_{i=2}^n x_i^\star && (98)\\
&= c - \vec{1}^\top \vec{x}^r + \sum_{i=1}^n x_i^\star && (99)\\
&= c. && (100)
\end{aligned}$$

Finally, we show that the objective value is strictly improved:

$$\sum_{i=1}^n -\alpha_i \log(x_i^\star) = -\alpha_1 \log(x_1^\star) + \sum_{i=2}^n -\alpha_i \log(x_i^\star) \tag{101}$$

$$< -\alpha_1 \log(x_1^r) + \sum_{i=2}^{n} -\alpha_i \log(x_i^\star) \tag{102}$$

$$= -\alpha_1 \log(x_1^r) + \sum_{i=2}^{n} -\alpha_i \log(x_i^r) \tag{103}$$

$$= \sum_{i=1}^{n} -\alpha_i \log(x_i^r). \tag{104}$$

Thus $\vec{x}^r$ could not be a solution to the relaxed problem, a contradiction.

This establishes that $p_r^\star \geq p_{\min}^\star$ and thus $p_r^\star = p_{\min}^\star$. This argument also shows that all solutions for the relaxed problem are feasible for the original problem, and since $p_r^\star = p_{\min}^\star$, they are the same set of solutions.

How did we cook up $\vec{x}^\star$? The main idea is that since the objective function considered each x_i independently, one can come up with a "better" point for any suboptimal point x^r just by moving one of the x_i, in our case x_1. And since it's monotonically decreasing in each x_i, we can make the x_i larger to get the desired result. Finally, also because the objective is monotonically decreasing in each x_i, we should make the x_i as large as possible subject to the constraints; this is how we came up with the fact that $x_1^\star$ needs to be large. The precise value of $x_1^\star$ is just bookkeeping to ensure that the constraint hits equality for $\vec{x}^\star$.

(c) After relaxing the equality constraint to an inequality constraint, form the Lagrangian $\mathcal{L}(\vec{x}, \vec{\lambda}, \mu)$ for the relaxed minimization problem, where λ_i is the dual variable corresponding to the inequality $x_i \geq 0$, and μ is the dual variable corresponding to the inequality constraint $\vec{1}^\top \vec{x} \leq c$.

Solution: The Lagrangian for this problem is

$$\mathcal{L}(\vec{x}, \mu) = \sum_{i=1}^{n} \alpha_i \log(1/x_i) + \sum_{i=1}^{n} \lambda_i(-x_i) + \mu(\vec{1}^\top \vec{x} - c) \tag{105}$$

$$= \sum_{i=1}^{n} (\alpha_i \log(1/x_i) + (\mu - \lambda_i)x_i) - \mu c, \tag{106}$$

(d) Now derive the dual function $g(\vec{\lambda}, \mu)$ for the relaxed minimization problem, and solve the dual problem $d_r^\star = \max\limits_{\substack{\vec{\lambda} \geq \vec{0} \\ \mu \geq 0}} g(\vec{\lambda}, \mu)$. What are the optimal dual variables $\vec{\lambda}^\star, \mu^\star$?

Solution: We have

$$g(\vec{\lambda}, \mu) = \min_{\vec{x}} \mathcal{L}(\vec{x}, \vec{\lambda}, \mu) = -\mu c + \sum_{i=1}^{n} \min_{x_i \geq 0} (\alpha_i \log(1/x_i) + (\mu - \lambda_i)x_i) \tag{107}$$

$$= -\mu c + \sum_{i=1}^{n} \begin{cases} (\alpha_i \log((\mu - \lambda_i)/\alpha_i) + \alpha_i), & \mu - \lambda_i > 0 \\ -\infty, & \mu - \lambda_i \leq 0 \end{cases} \tag{108}$$

$$= \begin{cases} -\mu c + \sum_{i=1}^{n} (\alpha_i \log((\mu - \lambda_i)/\alpha_i) + \alpha_i), & \forall i \colon \mu - \lambda_i > 0 \\ -\infty, & \exists i \colon \mu - \lambda_i \leq 0 \end{cases} \tag{109}$$

The minimum with respect to x_i in the first expression is attained at the unique point $x_i = \alpha_i/(\mu - \lambda_i)$, which we obtain by verifying that the expression is convex with respect to $\vec{x}$ and setting the gradient to 0. The dual is thus $d_r^\star = \max_{\substack{\vec{\lambda} \geq \vec{0} \\ \mu \geq 0}} g(\vec{\lambda}, \mu)$. To solve for the optimal dual variables, we solve for $\vec{\lambda}^\star$ first and then $\mu^\star$. For *every* choice of μ, it is optimal to pick $\vec{\lambda}^\star = \vec{0}$ so as to increase the quantity in the logarithm (because $\vec{\lambda} \geq \vec{0}$). Setting $\vec{\lambda}^\star = \vec{0}$, taking the gradient of $g(\vec{0}, \mu)$ with respect to μ, and setting it to 0, we obtain the optimal

$$\mu^\star = \frac{\sum_{i=1}^n \alpha_i}{c} = \frac{\alpha}{c}. \tag{110}$$

(e) Show that strong duality holds for the relaxed problem, so $p_r^\star = d_r^\star$.

Solution: We want to apply Slater's condition. We can verify that the objective and constraint functions are convex by taking the Hessian of each and verifying that they are positive semidefinite. For a strictly feasible point, we need to find an $\vec{x} \in \mathbb{R}^n$ such that each $x_i > 0$ and $\sum_{i=1}^n x_i < c$. There are many such $\vec{x}$, but one way to find them is to suppose that all x_i are the same, say χ, and find χ such that $n\chi < c$. This is achieved at $\chi = \frac{c}{2n}$, so $\vec{x} = \frac{c}{2n}\vec{1}$. Thus Slater's condition holds and strong duality holds.

The only inequality constraints are affine constraints, and hence by refined Slater's condition, strong duality should hold (no strictly feasible point is necessary since there are no non-affine inequalities).

(f) From the $\vec{\lambda}^\star, \mu^\star$ obtained in the previous part, how do we obtain the optimal primal variable $\vec{x}^\star$? What is the optimal objective function value $p_r^\star$? Finally, what is $p^\star$?

Solution: We obtain the optimal primal solution as

$$x_i^\star = \frac{\alpha_i}{\mu^\star} = \frac{c\alpha_i}{\alpha}, \quad i = 1, \ldots, n. \tag{111}$$

The expression for the optimal objective value follows by substituting this optimal solution back into the objective:

$$\begin{aligned}
p_r^\star &= \sum_{i=1}^n -\alpha_i \log\left(\frac{c\alpha_i}{\alpha}\right) && (112)\\
&= \sum_{i=1}^n -\left(\alpha_i \log\left(\frac{c}{\alpha}\right) + \alpha_i \log(\alpha_i)\right) && (113)\\
&= -\alpha \log\left(\frac{c}{\alpha}\right) - \sum_{i=1}^n \alpha_i \log(\alpha_i) && (114)\\
p^\star &= -p_{\min}^\star && (115)\\
&= -p_r^\star && (116)\\
&= \alpha \log\left(\frac{c}{\alpha}\right) + \sum_{i=1}^n \alpha_i \log(\alpha_i). && (117)
\end{aligned}$$

6. Homework Process

With whom did you work on this homework? List the names and SIDs of your group members.

NOTE: If you didn't work with anyone, you can put "none" as your answer.