The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381K: Large Scale Optimization — Fall 2015

Problem Set Three Solutions

Constantine Caramanis .

Matlab and Computational Assignments *Please provide a printout of the Matlab code you wrote to generate the solutions to the problems below.*

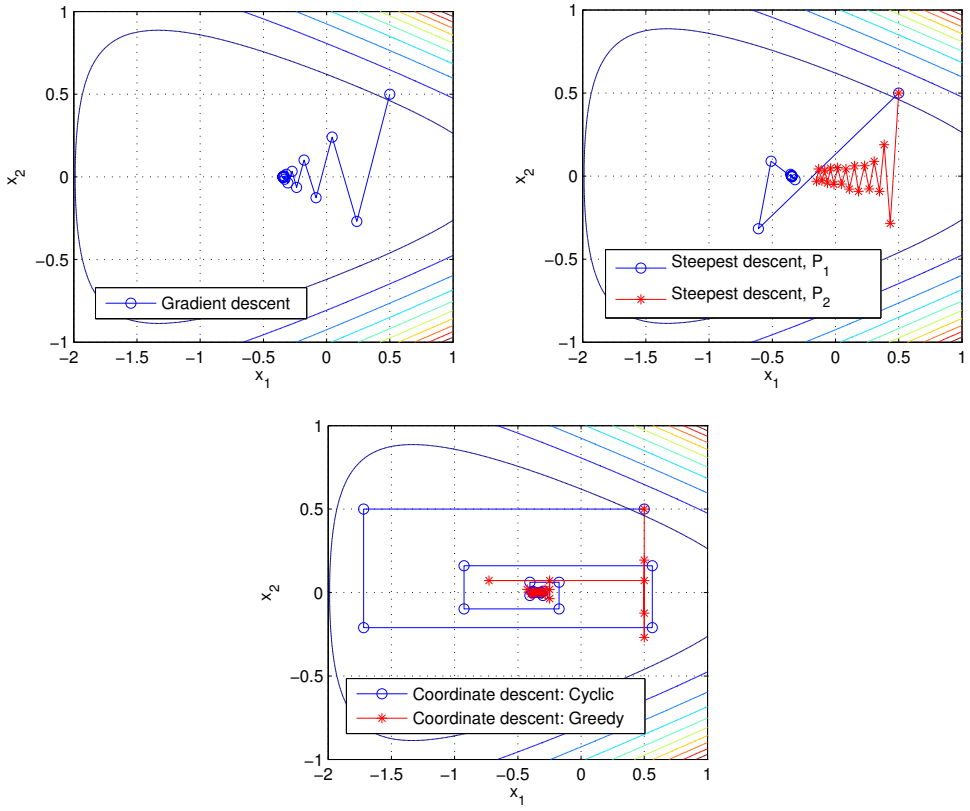
1. *Consider the non-quadratic problem given in Eq. (9.20) in B & V. Implement five flavors of gradient descent algorithms, and provide the convergence plots for all five.*

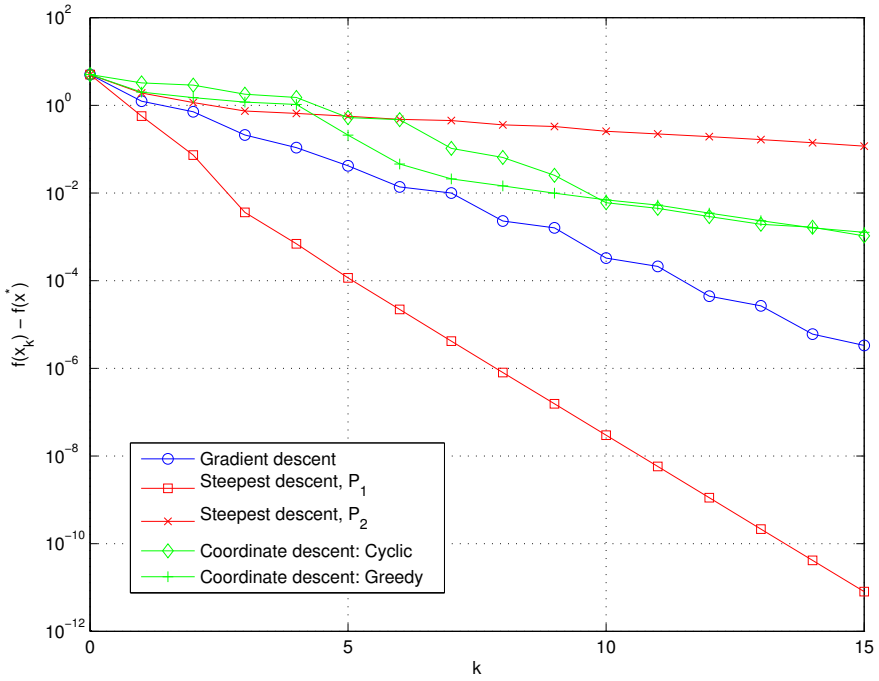
   (a) *Standard gradient descent with backtracking.*

   (b) *Two kinds of Steepest Descent, using the two matrices suggested in the book:*

$$P_1 = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}.$$

   (c) *Cyclic coordinate descent, and greedy coordinate descent, as defined today in class.*
   Solution The following is an example result. The sequences of points can vary depending on the parameters. This example result is obtained using $\alpha = 0.1$, $\beta = 0.7$, $x_0 = (0.5, 0.5)$.





Written Problems

1. **Coordinate Descent**

   (a) *Give an example that shows that coordinate descent may not find the optimum of a convex function. That is, provide a simple function f and a point x such that coordinate descent starting from x will* not *get to the global minimum of f.*

   Solution One of the simplest example is $f(x_1, x_2) = \max\{x_1, x_2\}$. If coordinate descent starts from $(1, 1)$, it will not get to the global minimum $(0, 0)$. The algorithm stops at $(1, 1)$ because each of x_1 and x_2 is at a local minimum.

   (b) *Let $f(x, y) = x^2 + y^2 + 3xy$, where x, y are scalars. Note that f is not convex. Would coordinate descent with exact line search always converge to a stationary point ?*

   Solution Consider the partial derivatives

   $$\frac{\partial f}{\partial x} = 2x + 3y, \quad \frac{\partial f}{\partial y} = 2y + 3x.$$

   We see that $\partial f/\partial x = 0$ for $x = -3y/2$, and $\partial f/\partial y = 0$ for $y = -3x/2$. Since the coordinate descent with exact line search finds the point at which the partial derivative is zero, the update is given by

   $$x_{k+1} = -3y_k/2, \; y_{k+1} = -3x_{k+1}/2.$$

   This shows that each coordinate will increase geometrically with a factor $9/4$. This algorithm does not converge to a stationary point.

2. **Condition Number**. *We saw in class that a fixed step size is able to guarantee linear convergence. The choice of step size we gave in class, however, depended on the function f.*

Show that it is not possible to choose a fixed step size t, that gives convergence for any strongly convex function. That is, for any fixed step size t, show that there exists (by finding one!) a smooth (twice continuously-differentiable) strongly convex function with bounded Hessian, such that a fixed-stepsize gradient algorithm starting from some point x_0, does not converge to the optimal solution.

Solution A strongly convex function with bounded Hessian has a convergence condition that $t < 2/M$ if $\nabla^2 f \preceq MI$. Therefore, the gradient descent with a fixed step size t does not converge if $\nabla^2 f \not\preceq (2/t)I$.

For example, consider $f = (2/t)x^T x$. Note that $\nabla^2 f = (4/t)I \not\preceq (2/t)I$. We get

$$x_{k+1} = x_k - t\nabla f(x_k) = x_k - t \cdot (4/t)x_k = -3x_k,$$

which implies that the algorithm does not converge.

3. **Decreasing Stepsize**.[1] *The previous problem shows that no constant step-size works for every strongly convex function. Consider now, a decreasing step size. Thus, at time k, you use step size $t_k \geq 0$. Show that if this sequence of step sizes satisfies:*

$$\lim_k t_k = 0, \qquad \sum_{k=0}^{\infty} t_k = \infty,$$

then gradient descent converges to the global optimal solution. Hint: Recall that strong convexity implies lower and upper bounds on the Hessian. Each of these bounds in turn gives lower and upper bounds on the value of $f(y)$ with respect to $f(x)$. Use one of these two show that for k large enough,

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2} t_k \|\nabla f(x_k)\|_2^2.$$

Use the other inequality to get (lower) bound on $\|\nabla f(x_k)\|$ in terms of the optimality gap. Then put these together to conclude that gradient descent must converge.

Solution For k large enough, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^\top (x_{k+1} - x_k) + \frac{M}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f(x_k) - t_k \|\nabla f(x_k)\|_2^2 + \frac{M}{2} t_k^2 \|\nabla f(x_k)\|_2^2 \\ &= f(x_k) - \left(1 - \frac{M}{2} t_k\right) t_k \|\nabla f(x_k)\|_2^2 \\ &\leq f(x_k) - \frac{1}{2} t_k \|\nabla f(x_k)\|_2^2 \end{aligned} \tag{1}$$

since the last inequality follows from that t_k is sufficiently small. Let K denote an index such that the inequality satisfies if $k \geq K$. On the other hand, we learned in class that the strong convexity bound gives (Check (4.5) in Lecture Note 4.)

$$\|x_k - x^*\|_2 \leq \frac{2}{m} \|\nabla f(x_k)\|_2^2$$

[1] This problem borrowed from Nati Srebro.

Then we put the above two inequalities together to get

$$f(x_k) - f(x_{k+1}) \geq \frac{mt_k}{4}\|x_k - x^*\|_2, \quad \forall k \geq K.$$

For some $n > K$, it follows that

$$\begin{aligned} f(x_K) - f(x^*) &\geq f(x_K) - f(x_n) \\ &\geq \frac{m}{4} \cdot \sum_{k=K}^{n-1} t_k \|x_k - x^*\|_2 \\ &\geq \frac{m}{4} \cdot \left(\sum_{k=K}^{n-1} t_k\right) \cdot \inf_{K \leq k \leq n-1} \|x_k - x^*\|_2 \end{aligned}$$

According to the above inequality, we must have $\liminf_{k\to\infty} \|x_k - x^*\|_2 = 0$. If there exists $\epsilon > 0$ such that $\liminf_{k\to\infty} \|x_k - x^*\|_2 > \epsilon$, the last line tends to infinity as $n \to \infty$, and this contradicts the inequality.

Now we complete the proof by claiming that $\|x_k - x^*\|_2$ decreases as k grows. We have

$$\begin{aligned} \|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 &= \|x_k - x^*\|_2^2 - \|x_k - t_k \nabla f(x_k) - x^*\|_2^2 \\ &= 2t_k \nabla f(x_k)^\top (x_k - x^*) - t_k^2 \|\nabla f(x_k)\|_2^2 \\ &\geq 2t_k (f(x_k) - f(x^*)) - t_k^2 \frac{2}{t_k} (f(x_k) - f(x_{k+1})) \\ &\geq 2t_k (f(x_{k+1}) - f(x^*)) \\ &> 0 \end{aligned}$$

where the first inequality follows from the convexity $f(x^*) \geq f(x_k) + \nabla f(x_k)^\top (x^* - x_k)$ and (1). Therefore, we get $\lim_{k\to\infty} \|x_k - x^*\|_2 = 0$.

4. **Convex functions**

   (a) *If f_i are convex functions, show that $f(x) := \sup_i f_i(x)$ is also convex.*
   Solution For $\alpha \in [0, 1]$, we have

$$\begin{aligned} f(\alpha x_1 + (1-\alpha)x_2) &= \sup_i f_i(\alpha x_1 + (1-\alpha)x_2) \\ &\leq \sup_i \{\alpha f_i(x_1) + (1-\alpha) f_i(x_2)\} \\ &\leq \alpha \sup_i f_i(x_1) + (1-\alpha) \sup_i f_i(x_2) \\ &= \alpha f(x_1) + (1-\alpha) f(x_2) \end{aligned}$$

   The first inequality follows from that every f_i is convex, and the second follows from that sup is a convex function.

   (b) *Show that the largest eigenvalue of a matrix is a convex function of the matrix (i.e. $\lambda_{\max}(M)$ is a convex function of M). Is the same true for the eigenvalue of largest magnitude?*
   Solution Since the eigenvalue decomposition of a symmetric matrix $M \in \mathbb{S}^n$ is $M = U\Lambda U^T$ where U is a unitary matrix , and Λ is a diagonal matrix with eigenvalues of

M at the diagonal entries. The largest eigenvalue of a symmetric matrix is equal to $\lambda_{\max}(M) = \max_{\|x\|_2=1} x^T M x$. Then we have

$$\begin{aligned}
\lambda_{\max}(\alpha M_1 + (1-\alpha)M_2) &= \max_{\|x\|_2=1} x^\top(\alpha M_1 + (1-\alpha)M_2)x \\
&= \max_{\|x\|_2=1} \{\alpha x^\top M_1 x + (1-\alpha)x^T M_2 x\} \\
&\leq \alpha \cdot \max_{\|x\|_2=1} x^\top M_1 x + (1-\alpha) \cdot \max_{\|x\|_2=1} x^T M_2 x \\
&= \alpha\lambda_{\max}(M_1) + (1-\alpha)\lambda_{\max}(M_2)
\end{aligned}$$

for $\alpha \in [0,1]$, where the inequality follows from that max is a convex function.

The eigenvalue of largest magnitude is not necessarily convex. Consider two symmetric and negative-definite matrices $M_1, M_2 \in \mathbb{S}^n_-$. (A matrix M is negative-definite if $x^T M x < 0$ for any $x \in \mathbb{R}^n$) We see that a convex combination is also negative-definite. Since negative-definite matrices have negative eigenvalues, the eigenvalue of the largest magnitude is the smallest eigenvalue. So we have $\lambda(M) = \min_{\|x\|_2=1} x^T M x$ if $M \in \mathbb{S}^n_-$, and it can be proved that this function is not convex. For $\alpha \in [0,1]$, we have

$$\begin{aligned}
\lambda(\alpha M_1 + (1-\alpha)M_2) &= \min_{\|x\|_2=1} x^\top(\alpha M_1 + (1-\alpha)M_2)x \\
&= \min_{\|x\|_2=1} \{\alpha x^\top M_1 x + (1-\alpha)x^\top M_2 x\} \\
&\geq \alpha \cdot \min_{\|x\|_2=1} x^\top M_1 x + (1-\alpha) \cdot \min_{\|x\|_2=1} x^\top M_2 x \\
&= \alpha\lambda(M_1) + (1-\alpha)\lambda(M_2)
\end{aligned}$$

(c) *Consider a weighted graph with edge weight vector w. Fix two nodes a and b. The* weighted shortest path *from a to b is the path whose sum of edge weights is the minimum, among all paths with one endpoint at a and another at b. Let $f(w)$ be the weight of this path. Show that f is a concave function of w.*

Solution To make a clear proof, let us define some notations at first. Let p denote a path in the graph with $N(p)$ steps, i.e., $p = (p_1, p_2, \ldots, p_{N(p)})$ where each $p_i = (\cdot, \cdot)$ is an edge in the graph, and two consecutive edges p_i and p_{i+1} are connected to the same node. Now we can show that f is concave as follows.

$$\begin{aligned}
f(\alpha w_1 + (1-\alpha)w_2) &= \min_{p: p_1=(a,\cdot), p_{N(p)}=(\cdot,b)} \sum_{i=1}^{N(p)} (\alpha w_1(p_i) + (1-\alpha)w_2(p_i)) \\
&\geq \min_{p: p_1=(a,\cdot), p_{N(p)}=(\cdot,b)} \sum_{i=1}^{N(p)} (\alpha w_1(p_i)) + \min_{p: p_1=(a,\cdot), p_{N(p)}=(\cdot,b)} \sum_{i=1}^{N(p)} ((1-\alpha)w_2(p_i)) \\
&= \alpha f(w_1) + (1-\alpha) f(w_2)
\end{aligned}$$

5. **Convex functions: Jensen's Inequality** *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be any function. It's epigraph is defined as the set:*

$$\text{epi}(f) = \{(x, y) \in \mathbf{R}^{n+1} \,:\, y \geq f(x)\}.$$

(a) *Show that if f is convex, then* epi(f) *is also convex.*

Solution Consider $(x_1, y_1), (x_2, y_2) \in \text{epi}(f)$. Since f is convex, we have

$$\begin{aligned}\alpha y_1 + (1-\alpha)y_2 &\geq \alpha f(x_1) + (1-\alpha)f(x_2) \\ &\geq f(\alpha x_1 + (1-\alpha)x_2)\end{aligned}$$

for $\alpha \in [0,1]$. It follows that $(\alpha x_1 + (1-\alpha)x_2, \alpha y_1 + (1-\alpha)y_2) \in \text{epi}(f)$. epi$(f)$ is convex.

(b) *Prove (the finite version of) Jensen's inequality. Jensen's inequality says that if p is a distribution on $\{x_1, \ldots, x_m\}$ with weights $p_1, \ldots, p_m$, and f is any concave function, then*

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}(X)).$$

Solution (Proof by induction) Jensen's inequality is trivial if $m = 2$ since

$$\mathbb{E}[f(X)] = p_1 f(x_1) + (1-p_1)f(x_2) \leq f(p_1 x_1 + (1-p_1)x_2) = f(\mathbb{E}[X]).$$

for any $p_1 = [0,1]$.

Suppose now that Jensen's inequality satisfies for $(m-1)$ weights. Define $\bar{p} = \sum_{i=1}^{m-1} p_i$. Then we have

$$\begin{aligned}\mathbb{E}[f(X)] &= \sum_{i=1}^{m-1} p_i f(x_i) + p_m f(x_m) \\ &= \bar{p} \cdot \sum_{i=1}^{m-1} \frac{p_i}{\bar{p}} f(x_i) + (1-\bar{p}) f(x_m) \\ &\leq \bar{p} \cdot f\left(\sum_{i=1}^{m-1} \frac{p_i}{\bar{p}} x_i\right) + (1-\bar{p})\, f(x_m) \\ &\leq f\left(\bar{p} \sum_{i=1}^{m-1} \frac{p_i}{\bar{p}} x_i + (1-\bar{p}) x_m\right) \\ &= f\left(\sum_{i=1}^{m} p_i x_i\right) = f(\mathbb{E}[X])\end{aligned}$$

where the first inequality follows from Jensen's inequality for weights $(p_1/\bar{p}, \ldots, p_{m-1}/\bar{p})$ that sum to 1, and the second inequality follows from the concavity of f. Hence Jensen's inequality satisfies for m weights.

6. **Projection** *We have been discussing only unconstrained problems. We will soon consider constraints. One update we will consider has the following form:*

$$x^{(k+1)} = \arg\min_{x \in \mathcal{X}} \left\{ \langle x, \nabla f(x^{(k)}) \rangle + \frac{1}{2t_k} \|x - x^{(k)}\|_2^2 \right\}.$$

Show that the solution is:

$$x^{(k+1)} = \text{Proj}_{\mathcal{X}}(x^{(k)} - t_k \nabla f(x^{(k)})).$$

This is called the Projected Gradient *algorithm.*

Solution

$$\begin{aligned}
x^{(k+1)} &= \mathrm{Proj}_{\mathcal{X}}(x^{(k)} - t_k \nabla f(x^{(k)})) \\
&= \arg\min_{x\in\mathcal{X}} \|x - (x^{(k)} - t_k \nabla f(x^{(k)}))\|_2^2 \\
&= \arg\min_{x\in\mathcal{X}} \|(x - x^{(k)}) + t_k \nabla f(x^{(k)})\|_2^2 \\
&= \arg\min_{x\in\mathcal{X}} \left\{ \|x - x^{(k)}\|_2^2 + 2\langle x - x^{(k)}, t_k \nabla f(x^{(k)})\rangle + \|t_k \nabla f(x^{(k)})\|_2^2 \right\} \\
&= \arg\min_{x\in\mathcal{X}} \left\{ \|x - x^{(k)}\|_2^2 + 2t_k \langle x, \nabla f(x^{(k)})\rangle \right\} \\
&= \arg\min_{x\in\mathcal{X}} \left\{ \frac{1}{2t_k}\|x - x^{(k)}\|_2^2 + \langle x, \nabla f(x^{(k)})\rangle \right\}
\end{aligned}$$

The second last line is obtained by removing constant terms.

7. **Computing Projections** *For the given convex set $\mathcal{X}$, compute the projection of a point z.*

   (a) *$\mathcal{X}$ is a rectangle defined by vectors L and U that satisfy $U_i \geq L_i$. Thus, $\mathcal{X} = \{x : L_i \leq x_i \leq U_i,\ i = 1, \ldots, n\}$.*
   Solution

$$\mathrm{Proj}_{\mathcal{X}} z = (\min(\max(z_1, L_1), U_1), \ldots, \min(\max(z_n, L_n), U_n))$$

   (b) *$\mathcal{X} = \mathbb{R}^n_+$.*
   Solution

$$\mathrm{Proj}_{\mathcal{X}} z = ((z_1)^+, \ldots, (z_n)^+), \quad (z_i)^+ = \max(z_i, 0)$$

   (c) *Euclidean ball: $\{x : \|x\|_2 \leq 1\}$.*
   Solution

$$\mathrm{Proj}_{\mathcal{X}} z = z / \max\{1, \|z\|_2\}$$

   (d) *1-norm ball: $\{x : \sum_i |x_i| \leq 1\}$.*
   Solution For $z \notin \mathcal{X}$, the projection is given by

$$\mathrm{Proj}_{\mathcal{X}} z = \left(\mathrm{sign}(z_1) \cdot (|z_1| - \lambda)^+ \ , \ \ldots \ , \ \mathrm{sign}(z_n) \cdot (|z_n| - \lambda)^+\right) \tag{2}$$

   where $\lambda > 0$ is chosen such that $\|\mathrm{Proj}_{\mathcal{X}} z\|_1 = 1$.
   The proof is given as follows. The original proof can be found in [2, Sec. 4]. Define $\hat{x} \triangleq \mathrm{Proj}_{\mathcal{X}} z$. Since we assume $z \notin \mathcal{X}$, i.e., $\|z\|_1 > 1$, the projection $\hat{x}$ lies on the boundary $\{x : \sum_{i=1}^n |x_i| = 1\}$. We claim that $\hat{x}_i z_i \geq 0$ for every $i \in \{1, \ldots, n\}$. If there exists k such that $\hat{x}_k z_k < 0$, we have $\tilde{x} \triangleq (\hat{x}_1, \ldots, \hat{x}_{k-1}, 0, \hat{x}_{k+1}, \ldots, \hat{x}_n) \in \mathcal{X}$ which satisfies

$$\|z - \hat{x}\|_2^2 = \sum_{i=1}^n (z_i - \hat{x}_i)^2 \geq \sum_{i=1}^n (z_i - \tilde{x}_i)^2 = \|z - \tilde{x}\|_2^2.$$

   This contradicts that $\hat{x}$ is the closest point in $\mathcal{X}$ to z. Since $\hat{x}_i$ should be zero or has the same sign as z_i for every $i \in \{1, \ldots, n\}$, we can only consider the points in $\{x : \sum_{i=1}^n |x_i| = 1,\ x_i z_i \geq 0,\ i = 1, \ldots, n\}$.

By symmetry, the projection can be obtained by

$$\hat{x} = (\text{sign}(z_1)y_1 \ , \ \ldots \ , \ \text{sign}(z_n)y_n)$$

where y is the projection of $|z|$ onto the probability simplex $\{y : \sum_i y_i = 1, y_i \geq 0, i = 1, \ldots, n\}$. Then we borrow the solution of Problem 7(f) to have

$$y = ((|z_1| - \lambda)^+, (|z_2| - \lambda)^+, \ldots, (|z_n| - \lambda)^+)$$

where λ is such that $\|y\|_1 = \sum_i y_i = 1$. Putting the above forms together, we obtain (2).

(e) *Positive semidefinite cone:* $S_+^n = \{M \in S^n : x^\top Mx \geq 0, \ \forall x \in \mathbb{R}^n\}$.
Solution This problem should have assumed that z is symmetric. Then, suppose $z = U\Lambda U^T$ where $U \in \mathbb{S}^n$ is a unitary matrix, i.e, $UU^T = I$, and L is a diagonal matrix with eigenvalues, i.e., $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. We will use the Frobenius norm as the distance between two matrices.

$$\begin{aligned}
\text{Proj}_{\mathcal{X}} z &= \arg\min_{x \in \mathbb{S}_+^n} \|x - z\|_F \\
&= \arg\min_{x \in \mathbb{S}_+^n} \|U^T(x - z)U\|_F \\
&= \arg\min_{x \in \mathbb{S}_+^n} \|U^T xU - \Lambda\|_F \\
&= U(\arg\min_{y \in \mathbb{S}_+^n} \|y - \Lambda\|_F)U^T \\
&= U\Lambda^+ U^T
\end{aligned}$$

where $\Lambda^+ = \text{diag}(\lambda_1^+, \lambda_2^+, \ldots, \lambda_n^+)$. The second last equality is obtained using a change of variables $y = U^T xU$, and the last equality follows from that a matrix $y \in \mathbb{S}_+^n$ must have nonnegative diagonal entries.

(f) *Probability simplex:* $\mathcal{X} = \{x : \sum_i x_i = 1, \ x_i \geq 0, \ i = 1, \ldots, n\}$.
Solution For $z \notin \mathcal{X}$, the projection is given by

$$\text{Proj}_{\mathcal{X}} z = ((z_1 - \lambda)^+, (z_2 - \lambda)^+, \ldots, (z_n - \lambda)^+) \tag{3}$$

where λ is chosen such that $\|\text{Proj}_{\mathcal{X}} z\|_1 = 1$.
The proof is given as follows. The original proof can be found in [1, Sec. 3.3]. Consider the optimization problem

$$\begin{aligned}
\min \quad & \frac{1}{2}\|z - x\|_2^2 \\
\text{s.t.} \quad & \sum_{i=1}^n x_i = 1, \ x_i \geq 0, \ \forall i \in \{1, \ldots, n\}
\end{aligned}$$

The Lagrangian is given by

$$L(x, \lambda, \rho) = \frac{1}{2}\|z - x\|_2^2 + \lambda \left(\sum_{i=1}^n x_i - 1 \right) - \sum_{i=1}^n \rho_i x_i$$

where $\rho_i \geq 0$ for $i \in \{1, \ldots, n\}$. The projection $\hat{x} \triangleq \mathrm{Proj}_{\mathcal{X}} z$ satisfies the KKT conditions

$$\begin{aligned}
&\hat{x}_i - z_i + \lambda - \rho_i = 0, \quad \forall i \in \{1, \ldots, n\}, \\
&\sum_{i=1}^{n} \hat{x}_i = 1, \\
&\rho_i = 0 \text{ or } \hat{x}_i = 0, \quad \forall i \in \{1, \ldots, n\}.
\end{aligned}$$

which result in $\hat{x}_i \in \{z_i - \lambda, 0\}$. $\hat{x}$ should satisfy these conditions for some λ and ρ, but there are many cases.

We now claim that, for any i and j such that $z_i > z_j$, if $\hat{x}_i = 0$ then $\hat{x}_j = 0$. Suppose there exists i and j where $z_i > z_j$ but $\hat{x}_i = 0$ and $\hat{x}_j \neq 0$. Let us define $\tilde{x}$ which has the same components as $\hat{x}$, except for $\tilde{x}_i = \hat{x}_j$ and $\tilde{x}_j = \hat{x}_i$. Then we have $\tilde{x} \in \mathcal{X}$ and

$$\begin{aligned}
\|z - \hat{x}\|_2^2 - \|z - \tilde{x}\|_2^2 &= (z_i - \hat{x}_i)^2 + (z_j - \hat{x}_j)^2 - (z_i - \hat{x}_j)^2 - (z_j - \hat{x}_i)^2 \\
&= z_i + (z_j - \hat{x}_j)^2 - (z_i - \hat{x}_j)^2 - z_j \\
&= 2z_j\hat{x}_j - 2z_i\hat{x}_j \\
&= 2\hat{x}_j(z_j - z_i) > 0.
\end{aligned}$$

This contradicts that $\hat{x}$ is the closest point in $\mathcal{X}$ to z.

Now it follows that the set of indices $I = \{i : \hat{x}_i = z_i - \lambda\}$ must be such that z_i's for $i \in I$ are the $|I|$ largest components of z. Since (3) is the only one satisfying this condition, we obtain (3).

References

[1] Shalev-Shwartz, S., & Singer, Y. (2006). Efficient learning of label ranking by soft projections onto polyhedra. Journal of Machine Learning Research, 7 (July), 1567–1599.

[2] Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the ℓ_1-ball for learning in high dimensions. Proceedings of the 25th International Conference on Machine Learning.