EE364b Homework 3

1. *Minimizing a quadratic.* Consider the subgradient method with constant step size α, used to minimize the quadratic function $f(x) = (1/2)x^T Px + q^T x$, where $P \succ 0$. For which values of α do we have $x^{(k)} \to x^\star$, for any $x^{(1)}$? What value of α gives fastest asymptotic convergence?

   Solution.

   The only subgradient for a quadratic function is the gradient, $\nabla f(x) = Px + q$. Each subgradient method iteration is

   $$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha(Px^{(k)} + q) \\ &= (I - \alpha P)x^{(k)} - \alpha q. \end{aligned}$$

   In general, the kth iterate is

   $$x^{(k)} = (I - \alpha P)^k x^{(0)} - k\alpha q.$$

   This can be viewed as a discrete-time linear dynamical system, and will be stable (and the subgradient method will converge) if and only if the eigenvalues of $I - \alpha P$ are less than 1 in magnitude. Since $P \succ 0$, all the eigenvalues of P are positive. Thus, we require $\lambda_{\max}(P) < 2$ for convergence. The equivalent constraint on α is that

   $$0 < \alpha < \frac{2}{\lambda_{\max}(P)}.$$

   The asymptotic convergence rate is determined by the eigenvalue of $I - \alpha P$ with largest magnitude, *i.e.*, $\max_{i=1,\ldots,n} |1 - \alpha\lambda_i|$, where λ_i are the eigenvalues of P. We can minimize this expression by requiring that $(1 - \alpha\lambda_{\min}) = -(1 - \alpha\lambda_{\max})$, *i.e.*, that

   $$\alpha = \frac{2}{\lambda_{\max} + \lambda_{\min}}.$$

   In other words, the optimal step size is the inverse of the average of the smallest and largest eigenvalues of P.

2. *Step sizes that guarantee moving closer to the optimal set.* Consider the subgradient method iteration $x^+ = x - \alpha g$, where $g \in \partial f(x)$. Show that if $\alpha < 2(f(x) - f^\star)/\|g\|_2^2$ (which is twice Polyak's optimal step size value) we have

   $$\|x^+ - x^\star\|_2 < \|x - x^\star\|_2,$$

   for any optimal point $x^\star$. This implies that $\mathbf{dist}(x^+, X^\star) < \mathbf{dist}(x, X^\star)$. (Methods in which successive iterates move closer to the optimal set are called *Féjer monotone.* Thus, the subgradient method, with Polyak's optimal step size, is Féjer monotone.)

Solution.

For any subgradient g, $g^T(x - x^\star) \le f(x) - f^\star$. Thus, if $\alpha < 2(f(x) - f^\star)/\|g\|^2$,

$$\alpha < \frac{2g^T(x - x^\star)}{\|g\|^2}$$

and

$$\alpha g^T g - 2g^T(x - x^\star) < 0.$$

Because $\alpha > 0$, we also have

$$\alpha^2 g^T g - 2\alpha g^T(x - x^\star) < 0.$$

Now we write

$$\begin{aligned}
\|x - x^\star\|_2^2 + \alpha^2 g^T g - 2\alpha g^T(x - x^\star) &< \|x - x^\star\|_2^2, \\
x^T x - 2x^T x^\star + x^{\star T} x^\star + \alpha^2 g^T g - 2\alpha g^T(x - x^\star) &< \|x - x^\star\|_2^2, \\
(x - \alpha g)^T(x - \alpha g) - 2(x - \alpha g)^T x^\star + x^{\star T} x^\star &< \|x - x^\star\|_2^2, \\
\|x^+ - x^\star\|_2^2 &< \|x - x^\star\|_2^2,
\end{aligned}$$

and

$$\|x^+ - x^\star\|_2 < \|x - x^\star\|_2$$

as required.

3. *A variation on alternating projections.* We consider the problem of finding a point in the intersection $\mathcal{C} \neq \emptyset$ of convex sets $\mathcal{C}_1, \ldots, \mathcal{C}_m$. To do this, we use alternating projections to find a point in the intersection of the two sets

$$\mathcal{C}_1 \times \cdots \times \mathcal{C}_m \subseteq \mathbf{R}^{mn}$$

and

$$\{(z_1, \ldots, z_m) \in \mathbf{R}^{mn} \mid z_1 = \cdots = z_m\} \subseteq \mathbf{R}^{mn}.$$

Show that alternating projections on these two sets is equivalent to the following iteration: project the current point in $\mathbf{R}^n$ onto each convex set, and then average the results. Draw a simple picture to illustrate this.

Solution.

Let P_i be the (Euclidean) projection operator onto $\mathcal{C}_i$. Let $\mathcal{A} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_m$, and let $\mathcal{B} = \{(z_1, \ldots, z_m) \in \mathbf{R}^{mn} \mid z_1 = \cdots = z_m\}$. To project onto $\mathcal{A}$ we use

$$P(x(k+1)) = (P_1(x_1(k)), \ldots, P_m(x_m(k))).$$

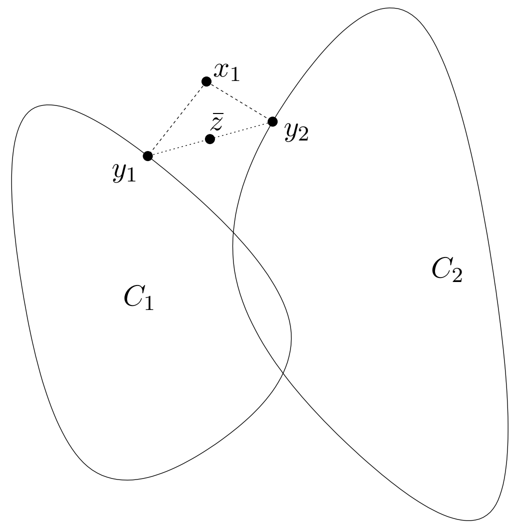


Figure 1: Finding $\bar{z}$, the average of a pair of projections.

The projection of a point $x \in \mathbf{R}^{mn}$ onto a point $z \in \mathcal{B}$, with $z = (z_1, \ldots, z_m)$ is the solution of the optimization problem

$$\begin{array}{ll} \text{minimize} & \|x - z\|^2 \\ \text{subject to} & x_{ij} = \bar{z}_j, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n, \end{array}$$

where $\bar{z} \in \mathbf{R}^n$ and $\bar{z} = z_1 = \cdots = z_m$. This problem is separable in the components $\bar{z}_j$, with minimizer $\bar{z}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$.

Thus, alternating projections with the above convex sets is equivalent to projecting and averaging.

4. *Matrix norm approximation.* We consider the problem of approximating a given matrix $B \in \mathbf{R}^{p\times q}$ as a linear combination of some other given matrices $A_i \in \mathbf{R}^{p\times q}$, $i = 1, \ldots, n$, as measured by the matrix norm (maximum singular value):

$$\text{minimize} \quad \|x_1 A_1 + \cdots + x_n A_n - B\|.$$

   (a) Explain how to find a subgradient of the objective function at x.

   (b) Generate a random instance of the problem with $n = 5$, $p = 3$, $q = 6$. Use `cvx` to find the optimal value $f^\star$ of the problem. Use a subgradient method to solve the problem, starting from $x = 0$. Plot $f - f^\star$ versus iteration. Experiment with several step size sequences.

Solution.

   (a) Let $f(x) = \|A(x)\|$, where $A(x) = x_1 A_1 + \cdots + x_n A_n - B$. We can express f as the pointwise supremum of affine functions of x,

$$f(x) = \|A(x)\| = \sup_{\|u\|_2=1,\ \|v\|_2=1} u^T A(x) v.$$

Each of the functions $f_{u,v}(x) = u^T A(x)v$ is affine in x for fixed u, v, with gradient $\nabla f_{u,v}(x) = (u^T A_1 v, \ldots, u^T A_n v)$.

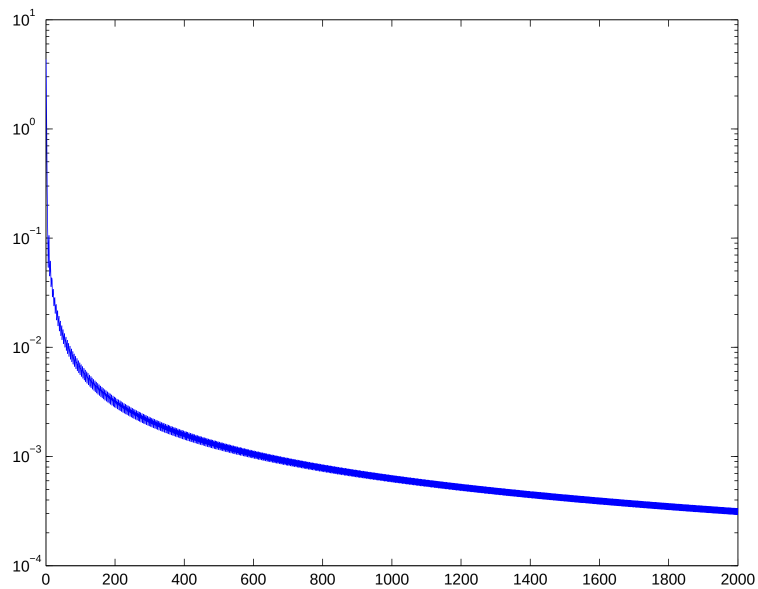
The active functions $f_{u,v}(x)$ are those associated with left and right singular vectors corresponding to the maximum singular value. To find a subgradient, we compute left and right singular unit vectors $\hat{u}$ and $\hat{v}$ associated with $\sigma_{\max}(A(x))$, and take

$$g = \nabla f_{\hat{u},\hat{v}}(x) = (\hat{u}^T A_1 \hat{v}, \ldots, \hat{u}^T A_n \hat{v}).$$

(b) Code appears below.

```
n = 5; p = 10; q = 20;

randn('seed', 124);
A1 = randn()*randn(p,q); A2 = randn()*randn(p,q);
A3 = randn()*randn(p,q); A4 = randn()*randn(p,q);
A5 = randn()*randn(p,q); B = randn(p, q);

cvx_begin
    variable x(5);
    z = x(1)*A1 + x(2)*A2 + x(3)*A3 + x(4)*A4 + x(5)*A5 - B;
    minimize (norm(z))
cvx_end

vs = []; x = zeros(n, 1);
for k = 1:500
    alph = 1/k;
    Z = x(1)*A1 + x(2)*A2 + x(3)*A3 + x(4)*A4 + x(5)*A5 - B;
    [U, E, V] = svd(Z);
    u = U(:,1);
    v = V(:,1);
    g = [u'*A1*v u'*A2*v u'*A3*v u'*A4*v u'*A5*v]';
    x = x - alph*g;
    vs(k) = norm(x(1)*A1 + x(2)*A2 + x(3)*A3 + x(4)*A4 + x(5)*A5 - B);
end

semilogy(vs - cvx_optval);
plot -depsc2 matrix_norm.eps
```

This code generates the plot shown below.



5. *Asynchronous alternating projections.* We consider the problem of finding a sequence of points that approach the intersection $\mathcal{C} \neq \emptyset$ of some convex sets $\mathcal{C}_1, \ldots, \mathcal{C}_m$. In the alternating projections method described in lecture, the current point is projected onto the farthest set. Now consider an algorithm where, at every step, the current point is projected onto any set not containing the point.

   Give a simple example showing that such an algorithm can fail, *i.e.*, we can have $\mathbf{dist}(x^{(k)}, \mathcal{C}) \not\to 0$ as $k \to \infty$.

   Now suppose that an additional hypothesis holds: there is an N such that, for each $i = 1, \ldots, m$, and each k, we have $x^{(j)} \in C_i$ for some $j \in \{k+1, \ldots, k+N\}$. In other words: in each block of N successive iterates of the algorithm, the iterates visit each of the sets. This would occur, for example, if in any block of N successive iterates, we project on each set at least once. As an example, we can cycle through the sets in round-robin fashion, projecting at step k onto C_i with $i = k \,\mathbf{mod}\, m + 1$. (When the point is in the set we are to project onto, nothing happens, of course.)

   When this additional hypothesis holds, we have $\mathbf{dist}(x^{(k)}, \mathcal{C}) \to 0$ as $k \to \infty$. Roughly speaking, this means we can choose projections in any order, provided each set is taken into account every so often.

   We give the general outline of the proof below; you fill in all details. Let's suppose the additional hypothesis holds, and let $x^\star$ be any point in the intersection $\mathcal{C} = \cap_{i=1}^m \mathcal{C}_i$.

   (a) Show that
   $$\|x^{(k+1)} - x^\star\|^2 \leq \|x^{(k)} - x^\star\|^2 - \|x^{(k+1)} - x^{(k)}\|^2.$$
   This shows that $x^{(k+1)}$ is closer to $x^\star$ than $x^{(k)}$ is. Two more conclusions (needed later): The sequence $x^{(k)}$ is bounded, and $\mathbf{dist}(x^{(k)}, \mathcal{C})$ is decreasing.

(b) Iteratively apply the inequality above to show that

$$\sum_{i=1}^{\infty} \|x^{(i+1)} - x^{(i)}\|^2 \le \|x^{(1)} - x^\star\|^2,$$

i.e., the distances of the projections carried out are square-summable. This implies that they converge to zero.

(c) Show that $\mathbf{dist}(x^{(k)}, \mathcal{C}_i) \to 0$ as $k \to \infty$. To do this, let $\epsilon > 0$. Pick M large enough that $\|x^{(k+1)} - x^{(k)}\| \le \epsilon/N$ for all $k \ge M$. Then for all $k \ge M$ we have $\|x^{(k+j)} - x^{(k)}\| \le j\epsilon/N$. Since we are guaranteed (by our hypothesis) that one of $x^{(k+1)}, \ldots, x^{(k+N)}$ is in $\mathcal{C}_i$, we have $\mathbf{dist}(x^{(k)}, \mathcal{C}_i) \le \epsilon$. Since ϵ was arbitrary, we conclude $\mathbf{dist}(x^{(k)}, \mathcal{C}_i) \to 0$ as $k \to \infty$.

(d) It remains to show that $\mathbf{dist}(x^{(k)}, \mathcal{C}) \to 0$ as $k \to \infty$. Since the sequence $x^{(k)}$ is bounded, it has an accumulation point $\tilde{x}$. Let's take a subsequence $k_1 < k_2 < \cdots$ of indices for which $x^{(k_j)}$ converges to $\tilde{x}$ as $j \to \infty$. Since $\mathbf{dist}(x^{(k_j)}, \mathcal{C}_i)$ converges to zero, we conclude that $\mathbf{dist}(\tilde{x}, \mathcal{C}_i) = 0$, and therefore (since $\mathcal{C}_i$ is closed) $\tilde{x} \in \mathcal{C}_i$. So we've shown $\tilde{x} \in \cap_{i=1,\ldots,m} \mathcal{C}_i = \mathcal{C}$.

We just showed that along a subsequence of $x^{(k)}$, the distance to $\mathcal{C}$ converges to zero. But $\mathbf{dist}(x^{(k)}, \mathcal{C})$ is decreasing, so we conclude that $\mathbf{dist}(x^{(k)}, \mathcal{C}) \to 0$.

Solution.

Our example will be in $\mathbf{R}^2$, with

$$\mathcal{C}_1 = \{x \mid x_1 \le 0\}, \quad \mathcal{C}_2 = \{x \mid x_2 - x_1 \le 0\}, \quad \mathcal{C}_3 = \{x \mid x_2 \le -1\},$$

starting from $x^{(1)} = (0, 1)$, and alternating between projections onto $\mathcal{C}_1$ and $\mathcal{C}_2$, always ignoring $\mathcal{C}_3$. It's easy to see (or show) that $x^{(k)} \to 0$, which is not in $\mathcal{C}_3$. In particular, $\mathbf{dist}(x^{(k)}, C) \not\to 0$ as $k \to \infty$.

Now let's prove convergence of the method assuming the additional hypothesis holds. Let $x^\star$ be any point in the intersection $\mathcal{C} = \cap_{i=1}^m \mathcal{C}_i$. We claim that each projection brings the point closer to $x^\star$. Suppose $x^{(k+1)}$ is the projection of $x^{(k)}$ onto $\mathcal{C}_i$ (where $x^{(k)} \notin \mathcal{C}_i$). Then we have

$$(x^{(k)} - x^{(k+1)})^T (x^{(k+1)} - z) \le 0$$

for any $z \in \mathcal{C}_i$, and in particular, for $z = x^\star$. Now we note that

$$\begin{aligned} \|x^{(k)} - x^\star\|^2 &= \|x^{(k)} - x^{(k+1)} + x^{(k+1)} - x^\star\|^2 \\ &= \|x^{(k)} - x^{(k+1)}\|^2 + \|x^{(k+1)} - x^\star\|^2 + 2(x^{(k)} - x^{(k+1)})^T (x^{(k+1)} - x^\star) \\ &\ge \|x^{(k)} - x^{(k+1)}\|^2 + \|x^{(k+1)} - x^\star\|^2, \end{aligned}$$

using our observation above. Thus we have

$$\|x^{(k+1)} - x^\star\|^2 \le \|x^{(k)} - x^\star\|^2 - \|x^{(k+1)} - x^{(k)}\|^2, \tag{1}$$

which shows that $x^{(k+1)}$ is closer to $x^\star$ than $x^{(k)}$ is. It follows that the sequence $x^{(k)}$ is bounded, since it lies in the ball described by $\|z - x^\star\| \le \|x^{(1)} - x^\star\|$. (We're going to use boundedness later.) We also note for future use that $\mathbf{dist}(x^{(k)}, \mathcal{C})$ is decreasing.

Our next step is to show that the distances of the projections carried out go to zero, *i.e.*, we have $\|x^{(k+1)} - x^{(k)}\| \to 0$ as $k \to \infty$. Recursively applying (1) yields

$$\|x^{(k+1)} - x^\star\|^2 \le \|x^{(1)} - x^\star\|^2 - \sum_{i=1}^{k} \|x^{(i+1)} - x^{(i)}\|^2,$$

which implies

$$\sum_{i=1}^{k} \|x^{(i+1)} - x^{(i)}\|^2 \le \|x^{(1)} - x^\star\|^2$$

for all k. This shows that the distances of the projections carried out are square-summable, which implies that they converge to zero.

Let $\epsilon > 0$. Let's pick M large enough that $\|x^{(k+1)} - x^{(k)}\| \le \epsilon/N$ for all $k \ge M$. This implies that for all $k \ge M$ we have $\|x^{(k+j)} - x^{(k)}\| \le j\epsilon/N$. Since we are guaranteed (by our hypothesis) that one of $x^{(k+1)}, \ldots, x^{(k+N)}$ is in $\mathcal{C}_i$, we have

$$\mathbf{dist}(x^{(k)}, \mathcal{C}_i) \le \epsilon.$$

Since ϵ was arbitrary, this shows that $\mathbf{dist}(x^{(k)}, \mathcal{C}_i) \to 0$ as $k \to \infty$.

We're almost done; it remains to show that $\mathbf{dist}(x^{(k)}, \mathcal{C}) \to 0$ as $k \to \infty$. Since the sequence $x^{(k)}$ is bounded, it has an accumulation point $\tilde{x}$. Let's take a subsequence $k_1 < k_2 < \cdots$ of indices for which $x^{(k_j)}$ converges to $\tilde{x}$ as $j \to \infty$. Since $\mathbf{dist}(x^{(k_j)}, \mathcal{C}_i)$ converges to zero, we conclude that $\mathbf{dist}(\tilde{x}, \mathcal{C}_i) = 0$, and therefore (since $\mathcal{C}_i$ is closed) $\tilde{x} \in \mathcal{C}_i$. So we've shown $\tilde{x} \in \cap_{i=1,\ldots,m} \mathcal{C}_i = \mathcal{C}$.

We just showed that along a subsequence of $x^{(k)}$, the distance to $\mathcal{C}$ converges to zero. But $\mathbf{dist}(x^{(k)}, \mathcal{C})$ is decreasing, so we conclude that $\mathbf{dist}(x^{(k)}, \mathcal{C}) \to 0$.

6. *Alternating projections to solve linear equations.* We can solve a set of linear equations expressed as

$$A_i x = b_i, \quad i = 1, \ldots, m,$$

where $A_i \in \mathbf{R}^{m_i \times n}$ are fat and full rank, using alternating projections on the sets

$$\mathcal{C}_i = \{z \mid A_i z = b_i\}, \quad i = 1, \ldots, m.$$

Assuming that the set of equations has solution set $\mathcal{C} \ne \emptyset$, we have $\mathbf{dist}(x^{(k)}, \mathcal{C}) \to 0$ as $k \to \infty$.

In view of exercise (5), the projections can be carried out cyclically, or asynchronously, provided we project onto each set every so often. As an alternative, we could project the current point onto all the sets, and form our next iterate as the average of these projected points, as in exercise (3).

Consider the case $m_i = 1$ and $m = n$ (*i.e.*, we have n scalar equations), and assume there is a unique solution $x^\star$ of the equations. Work out an explicit formula for the update, and show that the error $x^{(k)} - x^\star$ satisfies a (time-varying) linear dynamical system.

Solution.

In the scalar case, each update is a projection on $a_i^T x = b_i$. The update equation is

$$x^{(k+1)} = x^{(k)} - \frac{(a_i^T x^{(k)} - b_i)a_i}{\|a_i\|^2}.$$

Noting that $b_i = a_i^T x^\star$, we have

$$x^{(k+1)} - x^\star = x^{(k)} - x^\star - \frac{a_i^T(x^{(k)} - x^\star)a_i}{\|a_i\|^2}$$

and we can express this as the time-varying linear dynamical system

$$(x^{(k+1)} - x^\star) = \left[I - \frac{a_i a_i^T}{\|a_i\|^2}\right](x^{(k)} - x^\star).$$

If we use the project-and-average scheme for every iteration, we can write the system as

$$(x^{(k+1)} - x^\star) = \left[I - \sum_{i=1}^{n} \frac{a_i a_i^T}{\|a_i\|^2}\right](x^{(k)} - x^\star).$$

In this case we have a time-invariant linear dynamical system.