ELE 538: Large-Scale Optimization
Homework 1

Zachary Hervieux-Moore

Wednesday 28$^{\text{th}}$ February, 2018

Exercise 1: Strong convexity: Suppose that f is differentiable. Show that the following two statements are equivalent.

i)

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

ii)

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

Answer:

(i) $\Rightarrow$ (ii):

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

$$f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

Now adding the norm to both sides,

$$\begin{aligned} f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\ \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \mu\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \end{aligned}$$

Now we apply (i) to the LHS of the inequality to get

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \mu\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

Rearranging yields

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

(ii) $\Rightarrow$ (i):
We denote the new function $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ and notice that

$$\begin{aligned}
\langle \nabla g(\boldsymbol{x}) - \nabla g(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle &= \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \mu \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{x} - \boldsymbol{y} \rangle \\
&= \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle - \mu \|\boldsymbol{x} - \boldsymbol{y}\|_2^2
\end{aligned}$$

By (ii) we then have that

$$\langle \nabla g(\boldsymbol{x}) - \nabla g(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0$$

Therefore, $g(\boldsymbol{x})$ is convex. Using another notion of convexity,

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \langle \nabla g(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

Which leads to

$$\begin{aligned}
f(\boldsymbol{y}) - \frac{\mu}{2}\|\boldsymbol{y}\|_2^2 \geq f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2 + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle - \mu \langle \boldsymbol{x}, \boldsymbol{y} - \boldsymbol{x} \rangle \\
f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2}(\|\boldsymbol{x}\|_2^2 - 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \|\boldsymbol{y}\|_2^2) \\
f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}
\end{aligned}$$

Exercise 2: Subgradients: For each of the following convex functions, explain how to calculate a subgradient at a given $\boldsymbol{x} = (x_1, \dots, x_n)$.

a) $f(\boldsymbol{x}) = \max_{i=1,\dots,m} |\boldsymbol{a}_i^T \boldsymbol{x} + b_i|$

b) $f(\boldsymbol{x}) = \sup_{0 \le t \le 1} p(t)$, where $p(t) = x_1 + x_2 t + \cdots + x_n t^{n-1}$

c) $f(\boldsymbol{x}) = x_{[1]} + \cdots + x_{[k]}$, where $x_{[i]}$ denotes the i^{th} largest element of the vector $\boldsymbol{x}$

d) $f(\boldsymbol{x}) = \sup_{\boldsymbol{A}\boldsymbol{y} \preceq \boldsymbol{b}} \boldsymbol{y}^T \boldsymbol{x}$. (You can assume that the polyhedron defined by $\boldsymbol{A}\boldsymbol{y} \preceq \boldsymbol{b}$ is bounded, where "$\preceq$" denotes component-wise inequality).

Answer:

a) A subgradient is any $\boldsymbol{g} = \text{sign}(\boldsymbol{a}_i^T \boldsymbol{x} + b_i)\ \boldsymbol{a}_i$ such that $f(\boldsymbol{x}) = |\boldsymbol{a}_i^T \boldsymbol{x} + b_i|$ for some $i \in \{1, \dots, m\}$. Suppose $\boldsymbol{a}_k$ and b_k achieves the max for $f(\boldsymbol{x})$. Then we check whether or not $\boldsymbol{g}$ a subgradient. We need to check the following condition:

$$f(\boldsymbol{z}) \ge f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{z} - \boldsymbol{x})$$
$$|\boldsymbol{a}_k^T \boldsymbol{z} + b_k| \ge |\boldsymbol{a}_k^T \boldsymbol{x} + b_k| + \text{sign}(\boldsymbol{a}_k^T \boldsymbol{x} + b_i)\boldsymbol{a}_k^T(\boldsymbol{z} - \boldsymbol{x})$$

From here, it is trivial to check the 4 different cases on the signs of the absolute value terms to verify that this is always true.

b) First we rewrite the problem as

$$f(x) = \sup_{0 \le t \le 1} \boldsymbol{x}^T \boldsymbol{a}_t$$

where $\boldsymbol{a}_t^T = [1\ t \cdots\ t^{n-1}]$. By picking the subgradient to be $\boldsymbol{g} = \boldsymbol{a}_s$ where s is the argument that maximizes the supremum, we have

$$f(\boldsymbol{z}) \ge f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{z} - \boldsymbol{x})$$
$$f(\boldsymbol{z}) \ge f(\boldsymbol{x}) + \boldsymbol{a}_s^T \boldsymbol{z} - \boldsymbol{a}_s^T \boldsymbol{x}$$
$$f(\boldsymbol{z}) \ge \boldsymbol{a}_s^T \boldsymbol{z}$$

Where the last line is true because s does not necessarily achieve the supremum at point $\boldsymbol{z}$. Thus picking $g = \boldsymbol{a}_s$ is infact a subgradient.

So we simply need to find $\boldsymbol{a}_s$. By writing down the derivative of the polynomial $p'(t)$ as the characterstic polynomial of a companion matrix, we can efficiently find its eigenvalues using QR decomposition methods. This gives us the critical points of $p(t)$ which we check along with the boundary points $t = 0, 1$. As it is a polynomial on a compact set, one of these points will attain the maximum.

c) Pick the subgradient to be $\boldsymbol{g} = \sum_{i=1}^{k} e_{[i]}$. Where $e_{[i]}$ is the canonical basis vector corresponding to the i^{th} largest entry. Then we have

$$
\begin{aligned}
f(\boldsymbol{z}) &\geq f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{z} - \boldsymbol{x}) \\
\sum_{i=1}^{k} z_{[i]} &\geq \sum_{i=1}^{k} x_{[i]} + \sum_{i=1}^{k} e_{[i]}^T(\boldsymbol{z} - \boldsymbol{x}) \\
\sum_{i=1}^{k} z_{[i]} &\geq \sum_{i=1}^{k} x_{[i]} + \sum_{i=1}^{k} (z_{[i]} - x_{[i]}) \\
0 &\geq 0
\end{aligned}
$$

Thus we do indeed have $g = \partial f$.

d) To find a subgradient for $f(\boldsymbol{x})$, first some the LP $\sup_{\boldsymbol{Ay} \preceq \boldsymbol{b}} \boldsymbol{y}^T \boldsymbol{x}$, since it is bounded, the max is achieved. Denote the solutions by $\boldsymbol{y}^*$ and pick the subgradient to be $\boldsymbol{g} = \boldsymbol{y}^*$. Now we verify that this is a gradient.

$$
\begin{aligned}
f(\boldsymbol{z}) &\geq f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{z} - \boldsymbol{x}) \\
\sup_{\boldsymbol{Aw} \preceq \boldsymbol{b}} \boldsymbol{w}^T \boldsymbol{z} &\geq \sup_{\boldsymbol{Ay} \preceq \boldsymbol{b}} \boldsymbol{y}^T \boldsymbol{x} + \boldsymbol{y}^*(\boldsymbol{z} - \boldsymbol{x}) \\
\sup_{\boldsymbol{Aw} \preceq \boldsymbol{b}} \boldsymbol{w}^T \boldsymbol{z} &\geq \boldsymbol{y^*}^T \boldsymbol{x} + \boldsymbol{y^*}^T \boldsymbol{z} - \boldsymbol{y^*}^T \boldsymbol{x} \\
\sup_{\boldsymbol{Aw} \preceq \boldsymbol{b}} \boldsymbol{w}^T \boldsymbol{z} &\geq \boldsymbol{y^*}^T \boldsymbol{z}
\end{aligned}
$$

Which is true because $\boldsymbol{y}^*$ satisfies the constraints and so is feasible to the problem on the LHS.

Exercise 3: A convex function that is not subdifferentiable: Verify that the following function $f : \mathbb{R}_+ \to \mathbb{R}$ is convex, but not subdifferentiable at $x = 0$:

$$f(x) = \begin{cases} 1, & x = 0 \\ 0, & x > 0 \end{cases}$$

with $\text{domain}(f) = \mathbb{R}_+$

Answer:

First, let us prove f is convex. If $x, y \in (0, \infty)$ and $\lambda \in [0, 1]$, then we have that $f(\lambda x + (1-\lambda)y) = 0$ and $\lambda f(x)_(1-\lambda)f(y) = 0$, so it is trivially convex on $(0, \infty)$. Now suppose $x = 0$ and $y \in (0, \infty)$, then we have

$$f(\lambda x + (1-\lambda)y) = f((1-\lambda)y) = 0 \leq \lambda f(x) + (1-\lambda)f(y) = \lambda$$

Thus, $f(x)$ is convex. We will prove that it is not subdifferentiable at $x = 0$ by contradiction. Suppose that there was a subgradient $g \in \mathbb{R}$. Then we would have

$$\begin{aligned} f(z) &\geq f(x) + g(z - x) \quad \forall z \\ f(z) &\geq 1 + gz \quad \forall z \end{aligned}$$

But $f(z) = 0$ for all $z \in (0, \infty)$ so

$$\begin{aligned} 0 &\geq 1 + gz \quad \forall z \\ -gz &\geq 1 \quad \forall z \end{aligned}$$

Note that since $z > 0$, then we must have $g < 0$. Thus picking $z = \frac{1}{-2g}$ which belongs to the interval $(0, \infty)$, leads to a contradition.

Exercise 4: Matrix norm approximation: We consider the problem of approximating a given matrix $\boldsymbol{B} \in \mathbb{R}^{p\times q}$ as a linear combination of some other given matrices $\boldsymbol{A}_i \in \mathbb{R}^{p\times q}$, $i = 1, \dots, n$ as measured by the matrix norm (maximum singular value):

$$\min \|x_1\boldsymbol{A}_1 + \cdots + x_n\boldsymbol{A}_n - \boldsymbol{B}\|_2$$

a) Explain how to find a subgradient of the objective function at $\boldsymbol{x}$

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

Answer:

a) We note the the matrix norm of a matrix C is equivalent to solving the following problem

$$\sup_{\|y\|_2\leq 1} y^T C^T C y$$

Or put more succinctly, the square root of the largest eigenvalue of C^TC. In out case, $C = x_1\boldsymbol{A}_1 + \cdots + x_n\boldsymbol{A}_n - \boldsymbol{B}$. Thus, C^TC for us is

$$C^TC = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i A_i^T A_j x_j - \sum_{i=1}^{n}(x_i A_i^T B + x_i B^T A_i) + B^T B$$

Suppose $\boldsymbol{y}$ is the eigenvector corresponding to the largest singular value, then we have our objective being equal to

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i \boldsymbol{y}^T A_i^T A_j \boldsymbol{y} x_j - \sum_{i=1}^{n}(x_i \boldsymbol{y}^T A_i^T B\boldsymbol{y} + x_i \boldsymbol{y}^T B^T A_i \boldsymbol{y}) + \boldsymbol{y}^T B^T B\boldsymbol{y}$$

Thus, a subgradient of $f(\boldsymbol{x})$ can be obtained by taking the gradient of the above

$$\boldsymbol{g}_i = \boldsymbol{y}^T \left(2x_i A_i^T A_i + \sum_{i \neq j} (x_j A_i^T A_j + x_j A_j^T A_i) - A_i^T B - B^T A_i \right) \boldsymbol{y}$$

b) The code is appended below but the following three figures show the convergence for difference step size schemes.

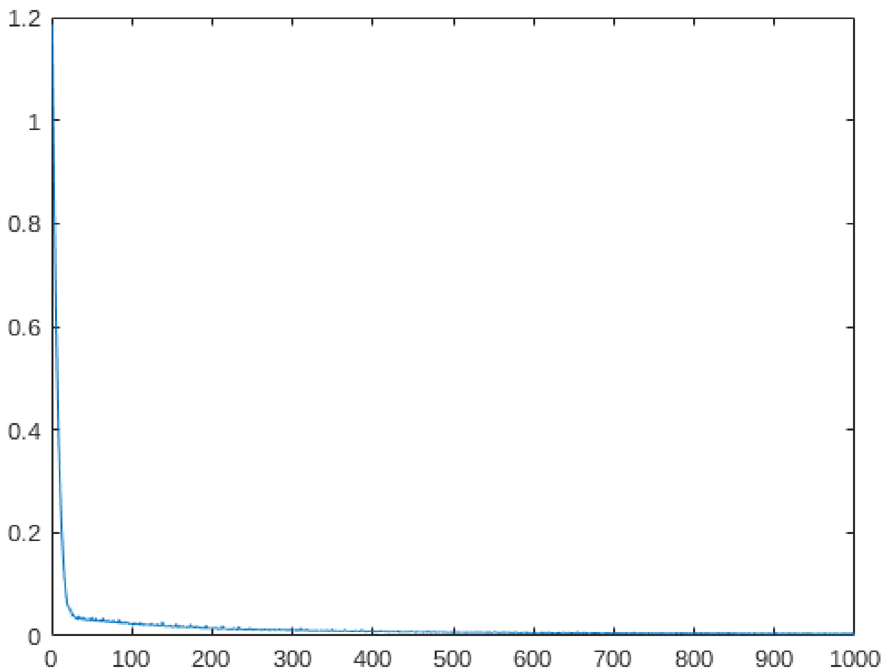

Figure 1: Difference between $f(x_t)$ and f^{opt} at each iteration using a Polyak step size.

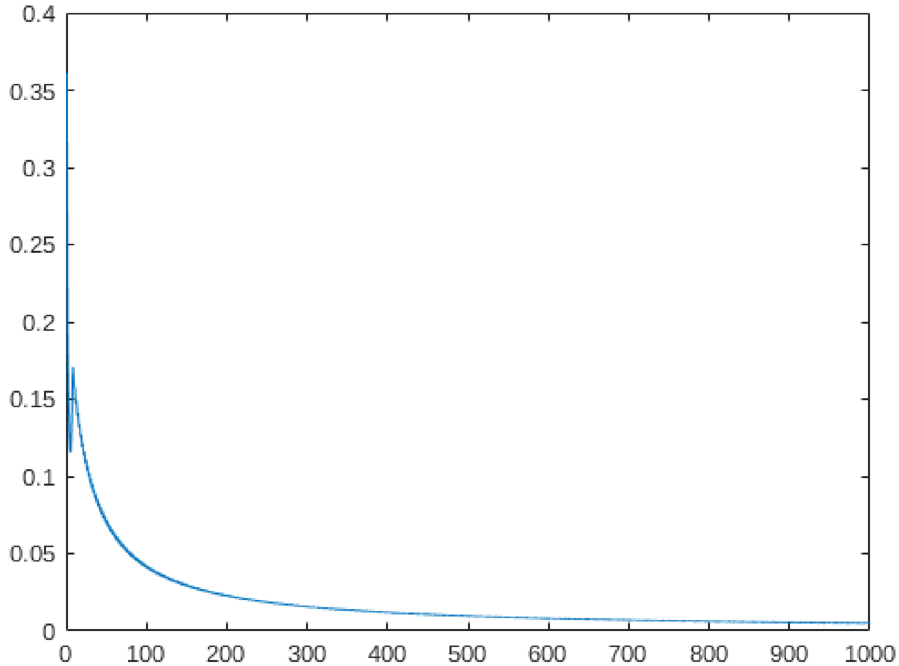

Figure 2: Difference between $f(x_t)$ and f^{opt} at each iteration using a $1/t$ step size. Note the slower convergence.

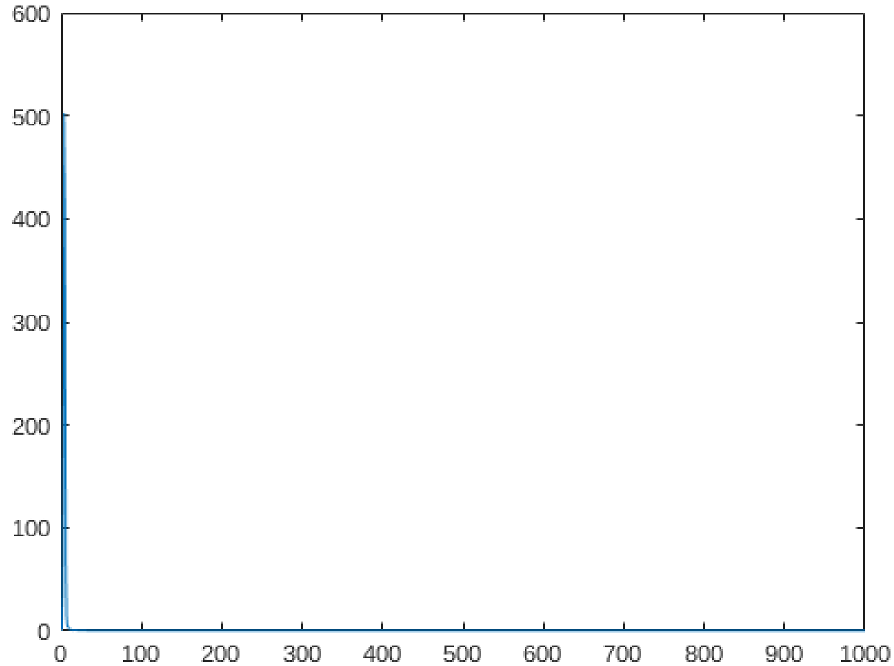

Figure 3: Difference between $f(x_t)$ and f^{opt} at each iteration using a $1/t^2$ step size. Note that divergence at the beggining when the step size is too big. The convergence is also much slower than the other two but the scale of the y-axis is too large to tell.

Code:

```
clear;
clc;

n = 5; p = 3; q = 6;
A = randn(p,q,n);
B = randn(p,q);

cvx_begin
    variable x(n)
    minimize( norm(A(:,:,1)*x(1)+A(:,:,2)*x(2)+A(:,:,3)*x(3)+A
        (:,:,4)*x(4)+A(:,:,5)*x(5) - B, 2) )
cvx_end


f_opt = cvx_optval;
```

```
% Subgradient method, no need to project as we are in R^n
T = 1000;
y = zeros(T,5);
g = zeros(T,5);
f = zeros(T,1);
tmp = zeros(6,6);

for t = 1:T
    inner_mat = A(:,:,1)*y(t,1)+A(:,:,2)*y(t,2)+A(:,:,3)*y(t
        ,3)+A(:,:,4)*y(t,4)+A(:,:,5)*y(t,5) - B;
    f(t) = norm(inner_mat, 2);
    [U,S,V] = svds(inner_mat, 1);

    % Calculate subgradient
    for i = 1:n
        tmp = 2*y(t,i)*A(:,:,i)'*A(:,:,i);

        for j = 1:n
            if j ~= i
                tmp = tmp + y(t,j)*A(:,:,i)'*A(:,:,j) + y(t,j)
                    *A(:,:,j)'*A(:,:,i);
            end
        end

        tmp = tmp - A(:,:,i)'*B - B'*A(:,:,i);
        g(t,i) = V'*tmp*V;
    end

    eta = (f(t)-f_opt)/(norm(g(t,:),2)^2);
    %eta = 1/t^2;
    %eta = 1/(t+30);
    y(t+1,:) = y(t,:) - eta*g(t,:);
end

plot(f-f_opt)
f(T)-f_opt
```

Exercise 5: Step sizes that guarantee moving closer to the optimal set: Consider the subgradient method iteration $\boldsymbol{x}^+ = \boldsymbol{x} - \eta \boldsymbol{g}$, where $\boldsymbol{g} \in \partial f(\boldsymbol{x})$. Let f^* be the optimal objective value. Show that if $\eta < \frac{2(f(\boldsymbol{x})-f^*)}{\|\boldsymbol{g}\|_2^2}$ (which is twice Polyak's optimal step size value) we have

$$\|\boldsymbol{x}^+ - \boldsymbol{x}^*\|_2 < \|\boldsymbol{x} - \boldsymbol{x}^*\|_2$$

for any optimal point $\boldsymbol{x}^*$.

Remark: Methods in which successive iterates move closer to the optimal set are called *Fejer monotone*. Thus, the subgradient method, with Polyak's optimal step size, is *Fejer monotone*.

Answer:

The proof follows immediately using the majorizing function presented by Lemma 4.1 in the notes. We have by the lemma:

$$\|\boldsymbol{x}^+ - \boldsymbol{x}^*\|_2^2 \leq \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 - 2\eta(f(\boldsymbol{x}) - f^*) + \eta^2 \|\boldsymbol{g}\|_2^2$$

If we wish to have the result that $\|\boldsymbol{x}^+ - \boldsymbol{x}^*\|_2 < \|\boldsymbol{x} - \boldsymbol{x}^*\|_2$, then we must have that

$$-2\eta(f(\boldsymbol{x}) - f^*) + \eta^2 \|\boldsymbol{g}\|_2^2 < 0$$

Simple rearranging of this yields

$$\eta < \frac{2(f(\boldsymbol{x}) - f^*)}{\|\boldsymbol{g}\|_2^2}$$

Note that dividing by η is justified as if $\eta = 0$ then we are at the optimal point.

Exercise 6: Gradient of squared distance (bonus): Define the Euclidean projection onto a closed convex set $\mathcal{C}$ as

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}\in\mathcal{C}} \|\boldsymbol{z} - \boldsymbol{x}\|_2$$

and let

$$\mathrm{dist}_{\mathcal{C}}(\boldsymbol{x}) := \|\boldsymbol{x} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})\|_2$$

Show that the gradient of the squared distance $f(\boldsymbol{x}) := \frac{1}{2}\mathrm{dist}^2_{\mathcal{C}}(\boldsymbol{x})$ is

$$\nabla f(\boldsymbol{x}) = \boldsymbol{x} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$$

Here, you can assume (without proof) that $f(\boldsymbol{x})$ is convex.

Answer:

For conciseness, we use the the shorthand notation $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) = \mathcal{P}_{\boldsymbol{x}}$. To show that the $\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}$ is the gradient of $f(\boldsymbol{x})$ we must show

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq (\boldsymbol{y} - \boldsymbol{x})^T(\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}})$$

Plugging in the definitions into the above

$$\|\boldsymbol{y} - \mathcal{P}_{\boldsymbol{y}}\|_2^2 - \|\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}\|_2^2 \geq 2(\boldsymbol{y} - \boldsymbol{x})^T(\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}})$$

Now we introduce some terms to the RHS and simplify

$$\begin{aligned}
\|\boldsymbol{y} - \mathcal{P}_{\boldsymbol{y}}\|_2^2 - \|\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}\|_2^2 &\geq 2(\boldsymbol{y} - \mathcal{P}_{\boldsymbol{y}} + \mathcal{P}_{\boldsymbol{y}} - \boldsymbol{x})^T(\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}) \\
\|\boldsymbol{y} - \mathcal{P}_{\boldsymbol{y}}\|_2^2 - \|\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}\|_2^2 &\geq 2(\boldsymbol{y} - \mathcal{P}_{\boldsymbol{y}})^T(\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}) + 2(\mathcal{P}_{\boldsymbol{y}} - \boldsymbol{x})^T(\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}) \\
\|\boldsymbol{y} - \mathcal{P}_{\boldsymbol{y}}\|_2^2 + \|\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}\|_2^2 - 2(\boldsymbol{y} - \mathcal{P}_{\boldsymbol{y}})^T(\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}) &\geq 2(\mathcal{P}_{\boldsymbol{y}} - \boldsymbol{x})^T(\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}) + 2\|\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}\|_2^2 \\
\|\boldsymbol{y} - \mathcal{P}_{\boldsymbol{y}} - \boldsymbol{x} + \mathcal{P}_{\boldsymbol{x}}\|_2^2 &\geq 2(\mathcal{P}_{\boldsymbol{y}} - \boldsymbol{x})^T(\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}) + 2\|\boldsymbol{x} - \mathcal{P}_{\boldsymbol{x}}\|_2^2
\end{aligned}$$

Now, we expand the norm on the LHS differently

$$\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + 2(\boldsymbol{y}-\boldsymbol{x})^T(\mathcal{P}_{\boldsymbol{x}}-\mathcal{P}_{\boldsymbol{y}}) + \|\mathcal{P}_{\boldsymbol{x}}-\mathcal{P}_{\boldsymbol{y}}\|_2^2 \geq 2(\mathcal{P}_{\boldsymbol{y}}-\boldsymbol{x})^T(\boldsymbol{x}-\mathcal{P}_{\boldsymbol{x}}) + 2\|\boldsymbol{x}-\mathcal{P}_{\boldsymbol{x}}\|_2^2$$

Using the non-expansiveness property of projections, we can upper bound the LHS

$$2\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + 2(\boldsymbol{y}-\boldsymbol{x})^T(\mathcal{P}_{\boldsymbol{x}}-\mathcal{P}_{\boldsymbol{y}}) \geq 2(\mathcal{P}_{\boldsymbol{y}}-\boldsymbol{x})^T(\boldsymbol{x}-\mathcal{P}_{\boldsymbol{x}}) + 2\|\boldsymbol{x}-\mathcal{P}_{\boldsymbol{x}}\|_2^2$$

Simplifying this expression

$$\begin{aligned}
\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + (\boldsymbol{y}-\boldsymbol{x})^T(\mathcal{P}_{\boldsymbol{x}}-\mathcal{P}_{\boldsymbol{y}}) &\geq (\mathcal{P}_{\boldsymbol{y}}-\mathcal{P}_{\boldsymbol{x}})^T(\boldsymbol{x}-\mathcal{P}_{\boldsymbol{x}}) \\
\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 &\geq (\mathcal{P}_{\boldsymbol{y}}-\mathcal{P}_{\boldsymbol{x}})^T(\boldsymbol{x}-\mathcal{P}_{\boldsymbol{x}}+\boldsymbol{y}-\boldsymbol{x}) \\
\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 &\geq (\mathcal{P}_{\boldsymbol{y}}-\mathcal{P}_{\boldsymbol{x}})^T(\boldsymbol{y}-\mathcal{P}_{\boldsymbol{x}})
\end{aligned}$$

Introducing $\mathcal{P}_{\boldsymbol{y}}-\mathcal{P}_{\boldsymbol{y}}$ to the inner product of the RHS yields

$$\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \geq \|\mathcal{P}_{\boldsymbol{y}}-\mathcal{P}_{\boldsymbol{x}}\|_2^2 + (\mathcal{P}_{\boldsymbol{y}}-\mathcal{P}_{\boldsymbol{x}})^T(\boldsymbol{y}-\mathcal{P}_{\boldsymbol{y}})$$

Now, we note that the second term in the RHS $(\mathcal{P}_{\boldsymbol{y}}-\mathcal{P}_{\boldsymbol{x}})^T(\boldsymbol{y}-\mathcal{P}_{\boldsymbol{y}}) \geq 0$ by the convexity of $\mathcal{C}$. To see this, note that $\boldsymbol{y}-\mathcal{P}_{\boldsymbol{y}}$ is perpendicular to the tangent at $\mathcal{P}_{\boldsymbol{y}}$ and that $\mathcal{P}_{\boldsymbol{x}}$ and $\mathcal{P}_{\boldsymbol{y}}$ are contained inside $\mathcal{C}$ and so these two vectors must have a positive inner product. Thus, we can lower bound the LHS by dropping this term

$$\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \geq \|\mathcal{P}_{\boldsymbol{y}}-\mathcal{P}_{\boldsymbol{x}}\|_2^2$$

Which is our non-expansiveness property that we know is true. Thus, our original inequality must be true and we conclude that

$$\nabla f(\boldsymbol{x}) = \boldsymbol{x} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$$