AMATH 515: Homework 3
Sid Meka

1. Convexity and composition rules. Suppose that f and g are twice continuously differentiable functions from $\mathbb{R}$ to $\mathbb{R}$, with $h = f \circ g$ their composition, defined by $h(x) = f(g(x))$.

   (a) If f and g are convex, show it is possible for h to be nonconvex (give a counter example). Give additional conditions that ensure the composition is convex.
   Here is our counterexample:
   Let $f(x) = -x$ and $g(x) = x^2$. Then, $f(g(x)) = -x^2$, which is not convex as the second derivative is entirely negative. Suppose f is nondecreasing. Then, since g is convex, we have

   $$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$$

   Therefore, since f is nondecreasing and convex, we have that:

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

   Thus, we have shown our appropriate counterexample to f and g being convex while h is nonconvex.

   (b) If f is convex and g is concave, what are additional conditions that guarantee h is convex?
   The additional conditions we require are for $f(x)$ to be nondecreasing and $g(x)$ to be nonincreasing. We now provide a proof of convexity under these conditions:

   Proof.

   $$\begin{aligned} g(\lambda x + (1-\lambda)y) &\geq \lambda g(x) + (1-\lambda)g(y) && \text{(Convexity of } g\text{)} \\ f(g(\lambda x + (1-\lambda)y)) &\leq f(\lambda g(x) + (1-\lambda)g(y)) \\ f(\lambda g(x) + (1-\lambda)g(y)) &\leq \lambda f(g(x)) + (1-\lambda)f(g(y)) && \text{(Applying Jensen's Inequality)} \\ f(g(\lambda x + (1-\lambda)y)) &\leq \lambda f(g(x)) + (1-\lambda)f(g(y)) && \text{(Combining inequalities)} \end{aligned}$$

   Thus, we have shown under these conditions $h(x) = f(g(x))$ is convex. □

   From the proof, we see that the additional conditions for $f(x)$ and $g(x)$ are sufficient to see $h(x)$ as convex.

   (c) Show that if $f : \mathbb{R}^m \to \mathbb{R}$ is convex and $g : \mathbb{R}^n \to \mathbb{R}^m$ is affine, then h is convex.

   Proof. Since g is affine, we can write $g(x) = Ax + b$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$. Observe that we now have that:

   $$\begin{aligned} g(\lambda x + (1-\lambda)y) &= A(\lambda x + (1-\lambda)y) + b \\ &= \lambda(Ax + b) + (1-\lambda)(Ay + b) \\ &= \lambda g(x) + (1-\lambda)g(y) \end{aligned}$$

   Thus,

   $$\begin{aligned} f(g(\lambda x + (1-\lambda)y)) &= f(\lambda g(x) + (1-\lambda)g(y)) \\ &\leq \lambda f(g(x)) + (1-\lambda)f(g(y)) \end{aligned}$$

   This proves $f \circ g$ is convex. □

(d) Show that the following functions are convex:

i. Logistic regression objective: $\sum_{i=1}^n \log(1 + \exp(a_i^T x)) - b^T Ax$
Let's call $\ell(x) = \sum_{i=1}^n \log(1 + \exp(a_i^T x)) - b^T Ax$, or the logistic regression objective function with respect to x. $\nabla^2 \ell(x) = \sum_{i=1}^n \frac{\exp(a_i^T x)}{(1+\exp(a_i^T x))^2} a_i a_i^T$. Note that $b^T Ax$ is affine meaning that $\nabla^2 b^T Ax = 0$. Additionally, note that $\nabla^2 \ell(x) \succeq 0$, so $\ell(x)$ is convex.

ii. Poisson regression objective: $\sum_{i=1}^n \exp(a_i^T x) - b^T Ax$.
Let's call $\mathfrak{p}(x) = \sum_{i=1}^n \exp(a_i^T x) - b^T Ax$, or the Poisson regression objective function with respect to x. $\nabla^2 \mathfrak{p}(x) = \exp(a_i^T x) a_i a_i^T$. Note that $b^T Ax$ is affine meaning that $\nabla^2 b^T Ax = 0$. Additionally, note that $\nabla^2 \mathfrak{p}(x) \succeq 0$, so $\mathfrak{p}(x)$ is convex.

2. Suppose f is a continuously differentiable function from $\mathbb{R}^n$ to $\mathbb{R}$. Show that the iterations of steepest descent

$$x^{k+1} = x^k - \alpha \nabla f(x^k)$$

can be equivalently obtained by solving the problem

$$x^{k+1} = \underset{x \in \mathbb{R}^n}{\arg\min}\, f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2s} \|x - x^k\|^2,$$

what is the relationship between s and α?

We aim to find the relationship between s and α. Note that from the remark $m_k(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2s}\|x - x^k\|^2$. We will use this $m_k(x)$.

$$\begin{aligned}
x^{k+1} &= \underset{x \in \mathbb{R}^n}{\arg\min}\, f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2s}\|x - x^k\|^2 && \text{(Given)} \\
x^{k+1} &= \underset{x \in \mathbb{R}^n}{\arg\min}\, f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2s}(x - x^k)^T (x - x^k) && \text{(Expanding } \|x - x^k\|^2\text{)} \\
m_k(x) &= f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2s}\|x - x^k\|^2 && \text{(Using } m_k(x)\text{)} \\
m_k(x) &= f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2s}(x - x^k)^T (x - x^k) && \text{(Expanding } \|x - x^k\|^2\text{)} \\
\nabla m_k(x) &= \nabla f(x^k) + \frac{1}{s}(x - x^k) && \text{(Gradient of } m_k(x) \text{ with respect to } x\text{)} \\
\nabla f(x^k) &+ \frac{1}{s}(x - x^k) = 0 && \text{(Setting } \nabla m_k(x) = 0 \text{ to find the minimum)} \\
x^{k+1} &= x^k - s\nabla f(x^k) && \text{(Multiplying by } s\text{)}
\end{aligned}$$

We thus have $x^{k+1} = x^k - s\nabla f(x^k)$ and $x^{k+1} = x^k - \alpha \nabla f(x^k)$. We see that s and α are equal. Thus, $s = \alpha$. Note that $s \neq 0$ as s is our step size and if it were 0, there would be no step size.

3. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth, and let $x^\star$ be a minimizer with value $f(x^\star) = f^\star$.

(a) Show that for any $x \in \mathbb{R}^n$, we have

$$f(x) - f^\star \geq \frac{1}{2L}\|\nabla f(x)\|^2.$$

Proof. We can use the definition of L smoothness to write $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|^2$ for all $x \in \mathbb{R}^n$ and for all $y \in \mathbb{R}^n$. If we set $y = x - \frac{1}{L}\nabla f(x)$, and we additionally substitute this y in the L smoothness inequality: $f\left(x - \frac{1}{L}\nabla f(x)\right) \leq f(x) + \nabla f(x)^T \left(-\frac{1}{L}\nabla f(x)\right) + \frac{L}{2}\left\|-\frac{1}{L}\nabla f(x)\right\|^2$. We have $\nabla f(x)^T \left(-\frac{1}{L}\nabla f(x)\right) = -\frac{1}{L}\|\nabla f(x)\|^2$ due to a direct consequence of the property that the dot product of a vector with itself is its squared norm, and we use this in combination with the distributive property of scalar multiplication in inner products. Furthermore, $\left\|-\frac{1}{L}\nabla f(x)\right\|^2 = \frac{1}{L^2}\|\nabla f(x)\|^2$ because scaling a vector, which is $\nabla f(x)$ here, by a scalar, which is $-\frac{1}{L}$ here, scales its squared norm by the square of that scalar. Hence, we have $f\left(x - \frac{1}{L}f(x)\right) \leq f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{L}{2}\cdot\frac{1}{L^2}\|\nabla f(x)\|^2$. Additionally,

$$\begin{aligned} f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{L}{2}\cdot\frac{1}{L^2}\|\nabla f(x)\|^2 = \\ f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(x)\|^2 = \\ f(x) - \frac{2}{2L}\|\nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(x)\|^2 = \\ f(x) - \frac{1}{2L}\|\nabla f(x)\|^2 \end{aligned}$$

Thus, we will proceed with $f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{L}{2}\cdot\frac{1}{L^2}\|\nabla f(x)\|^2 = f(x) - \frac{1}{2L}\|\nabla f(x)\|^2$.

Since $x^\star$ is a minimizer, $f(x^\star) \leq f\left(x - \frac{1}{L}\nabla f(x)\right)$. Combining with what we have found before and using $f^\star = f(x^\star)$, we have

$$f^\star \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|^2.$$

This means

$$f^\star - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|^2.$$

Negating each side and flipping the inequality sign,

$$f(x) - f^\star \geq \frac{1}{2L}\|\nabla f(x)\|^2.$$

Thus we have shown for any $x \in \mathbb{R}^n$, it follows that $f(x) - f^\star \geq \frac{1}{2L}\|\nabla f(x)\|^2$. □

(b) Show that for any $x, y \in \mathbb{R}^n$ we have

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Hint: apply part (a) to the functions, $h_x(z) = f(z) - \nabla f(x)^T z$ *and* $h_y(z) = f(z) - \nabla f(y)^T z$.

We define:

$$h_x(z) = f(z) - \nabla f(x)^T z$$
$$h_y(z) = f(z) - \nabla f(y)^T z$$

For each of these functions, we note that they inherit convexity and smoothness from f, and their gradients are given by:

$$\nabla h_x(z) = \nabla f(z) - \nabla f(x)$$
$$\nabla h_y(z) = \nabla f(z) - \nabla f(y)$$

Applying the inequality $f(z) - f^\star \geq \frac{1}{2L}\|\nabla f(z)\|^2$, where $f(z^\star) = f^\star$ to h_x at y and to h_y at x, we get:

$$h_x(y) - h_x^\star \geq \frac{1}{2L}\|\nabla h_x(y)\|^2$$

$$h_y(x) - h_y^\star \geq \frac{1}{2L}\|\nabla h_y(x)\|^2$$

Since $h_x^\star = f^\star - \nabla f(x)^T x$ and $h_y^\star = f^\star - \nabla f(y)^T y$, we substitute them:

$$f(y) - \nabla f(x)^T y - (f^\star - \nabla f(x)^T x) \geq \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2$$

$$f(x) - \nabla f(y)^T x - (f^\star - \nabla f(y)^T y) \geq \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2$$

Rearranging both inequalities, we obtain:

$$f(y) - f^\star - \nabla f(x)^T (y - x) \geq \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2$$

$$f(x) - f^\star - \nabla f(y)^T (x - y) \geq \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2$$

Adding the two inequalities:

$$(f(y) - f^\star - \nabla f(x)^T (y - x)) + (f(x) - f^\star - \nabla f(y)^T (x - y)) \geq \frac{2}{2L} \|\nabla f(y) - \nabla f(x)\|^2$$

We can state this more concisely:

$$f(x) + f(y) - 2f^\star - \nabla f(x)^T (y - x) - \nabla f(y)^T (x - y)) \geq \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2$$

Since $x^\star$ is a minimizer of f and $y^\star$ is a minimizer of f, we have $f^\star \leq f(x)$ and $f^\star \leq f(y)$. Adding these two inequalities, we get: $2f^\star \leq f(x)+f(y)$. This can also be stated as $f(x)+f(y)-2f^\star \geq 0$. Since $f(x) + f(y) - 2f^\star \geq 0$, we obtain:

$$-\nabla f(x)^T (y - x) - \nabla f(y)^T (x - y) \geq \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2$$

This means:

$$\nabla f(x)^T (x - y) - \nabla f(y)^T (x - y) \geq \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2$$

Furthermore:

$$(\nabla f(x)^T - \nabla f(y)^T)(x - y) \geq \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2$$

This means:

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2$$

Because $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ are both chosen arbitrarily, we have shown that for any $x, y \in \mathbb{R}^n$, we have $(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2$.

4. Let A be an $N \times d$ matrix with $N < d$ and $\text{rank}(A) = N$, and consider the least-squares problem

$$\min_x f(x) := \frac{1}{N} \|Ax - b\|^2.$$

(a) Characterize the solution space of the system $Ax = b$. You will want to consider some of the "four fundamental subspaces", the column space of A, null space of A,row space of A, and null space of A^T.

Since A is an $N \times d$ matrix with $\text{rank}(A) = N$, we know:

i. The column space of A, denoted $\text{range}(A)$, is $\mathbb{R}^n$.

ii. The null space of A, denoted $\text{null}(A)$, has dimension $d - N$ because A maps from $\mathbb{R}^d$ to $\mathbb{R}^N$, and we thus have: $\dim(\text{null}(A)) = d - N$.

iii. The row space of A is $\mathbb{R}^N$ since A has full row rank.

iv. The null space of A^T, denoted null(A^T), has dimension $d - N$ because A^T is a $d \times N$ matrix of rank N, so its null space has dimension $d - N$.

The equation $Ax = b$ has a solution if and only if $b \in \text{range}(A)$. Since A has full row rank N, the system $Ax = b$ is consistent for all $b \in \mathbb{R}^N$. Furthermore, the general solution to $Ax = b$ is given by: $x = x_p + x_h$, where x_p is a particular solution, or in other words, one specific solution to $Ax = b$ and x_h is a homogeneous solution that satisfies $Ax_h = 0$, so it belongs to null(A). Thus, the solution space of $Ax = b$ is the affine space: $x_p + \text{null}(A)$. Furthermore, there will always be more than one solution because the null space of A is nontrivial.

(b) Write down the Lipschitz constant for the gradient of f in terms of A.

The function $f(x)$ is given by: $f(x) = \frac{1}{N}\|Ax - b\|^2$. Computing its gradient: $\nabla f(x) = \frac{2}{N} A^T(Ax - b)$. The Hessian is: $\nabla^2 f(x) = \frac{2}{N} A^T A$. The Lipschitz constant L of the graient is the largest eigenvalue of $\nabla^2 f(x)$, which is: $L = \frac{2}{N}\lambda_{\max}(A^T A)$. Since A has rank N, the nonzero eigenvalues of $A^T A$ are the squares of the singular values $\sigma_i(A)$, so we can write: $L = \frac{2}{N}\sigma^2_{\max}(A)$.

(c) If you run the steepest-descent algorithm on this problem with $x^{(0)} = 0$ and with the appropriate choice of steplength, how many iterations are required to find a solution that satisfies $\frac{1}{N}\|Ax - b\|^2 \le \epsilon$, according to the convergence theory for gradient descent applied to smooth (not strongly) convex functions?

Since $f(x)$ is a convex quadratic function, gradient descent with optimal step size $\alpha = \frac{1}{L}$ converges at a rate given by $f(x_k) - f^\star \le \frac{L}{2k}\|x_0 - x^\star\|^2$. This ensures: $\frac{1}{N}\|Ax_k - b\|^2 \le \epsilon$ after k iterations, where $k \ge \frac{L}{2\epsilon}\|x_0 - x^\star\|^2$. Note that $\frac{L}{2\epsilon}\|x_0 - x^\star\|^2 = \frac{\sigma^2_{\max}(A)}{N\epsilon}\|x_0 - x^\star\|^2$ because $L = \frac{2}{N}\sigma^2_{\max}(A)$, which we have found previously, so in other words we have $\frac{1}{N}\|Ax_k - b\|^2 \le \epsilon$ after k iterations, where $k \ge \frac{\sigma^2_{\max}(A)}{N\epsilon}\|x_0 - x^\star\|^2$.

(d) Show that every iterate x_k of steepest descent lies in the row space of A, or in other words, that $x_k \in \text{range}(A^T)$ for every k. Explain why this implies that steepest-descent will converge to the minimum norm solution $x_{\min}$ defined as

$$x_{\min} = \arg\min_x \|x\|^2 \quad \text{s.t.} \quad Ax = b.$$

This is one case of the implicit regularization effect of gradient descent.

We initialize $x^{(0)} = 0$ and use the gradient descent update: $x_{k+1} = x_k - \alpha \nabla f(x_k)$. Since the gradient $\nabla f(x_k)$ is always in range(A^T) and $x_0 = 0$, we see that all iterates satisfy: $x_k \in \text{range}(A^T)$ for all k. This means gradient descent never moves outside range(A^T). The minimum norm solution, which we will define as $x_{♫}$ is characterized by: $x_{♫} = A^T(AA^T)^{-1}b$. Gradient descent converges to this solution because it iterates in range(A^T), where the unique solution is $x_{♫}$. Note that here instead of converging to any solution in $x_p + \text{null}(A)$, the gradient descent selects the one with the smallest norm. Furthermore, all iterates lie in range(A^T) leading to convergence to the minimum norm solution.

Steepest descent applied to this ”underdetermined least squares” problem actually enjoys even faster convergence than one would expect given that it is not strongly convex. Indeed, we can actually ignore the zero eigenvalues of the Hessian and only consider the nonzero ones in the convergence rate in this case to obtain a *linear* convergence rate (think about why!). *Okay, I have thought about why.*

5. Consider the regularized problem

$$\min_x f_\mu(x) := \frac{1}{N}\|Ax - b\|^2 + \mu\|x\|^2.$$

for some $\mu > 0$.

(a) Express the minimizer x_μ of this problem in closed form.
The objective function is:

$$f_\mu(x) = \frac{1}{N}\|Ax - b\|^2 + \mu\|x\|^2$$

To find the minimizer x_μ, we set the gradient $\nabla f_\mu(x)$ to zero:

$$\begin{aligned}
\nabla f_\mu(x) &= 0 && \text{(Setting } \nabla f_\mu(x) \text{ to 0)}\\
\nabla f_\mu(x) &= \frac{2}{N}A^T(Ax - b) + 2\mu x && \text{(Expressing } \nabla f_\mu(x)\text{)}\\
\frac{2}{N}A^T(Ax - b) + 2\mu x &= 0\\
\frac{2}{N}A^TAx - \frac{2}{N}A^Tb + 2\mu x &= 0\\
\frac{2}{N}A^TAx + 2\mu x &= \frac{2}{N}A^Tb\\
\frac{1}{N}A^TAx + \mu x &= \frac{1}{N}A^Tb\\
\left(\frac{1}{N}A^TA + \mu I\right)x &= \frac{1}{N}A^Tb && \text{(Factoring out } x\text{)}
\end{aligned}$$

Since A^TA is positive semidefinite and μI is positive definite as it is defined that $\mu > 0$, the matrix $\frac{1}{N}A^TA + \mu I$ is invertible. Thus, the closed form solution is:

$$x_\mu = \left(\frac{1}{N}A^TA + \mu I\right)^{-1}\frac{1}{N}A^Tb$$

(b) If you applied steepest-descent to f_μ with $x^{(0)} = 0$, how many iterations are required to find a solution satisfying $f_\mu(x^k) - f_\mu(x_\mu) \le \epsilon$?
The Hessian of $f_\mu(x)$ is $\nabla^2 f_\mu(x) = \frac{2}{N}A^TA + 2\mu I$. Since $\mu > 0$, the matrix is Positive Definite, hence strongly convex. Now, we can compute the Lipschitz constant, denoted by L, and the strongly convex constant, denoted by m, from the eigenvalues of the Hessian Matrix. This will give us:

$$L = \lambda_{\max}\left(\frac{2}{N}A^TA + 2\mu I\right) \quad \text{and} \quad m = \lambda_{\min}\left(\frac{2}{N}A^TA + 2\mu I\right),$$

where $\lambda_{\max}$ is the largest eigenvalue of $\nabla^2 f_\mu(x)$ and $\lambda_{\min}$ is the smallest eigenvalue of $\nabla^2 f_\mu(x)$. For a m strongly convex and L smooth function, gradient descent with optimal step size converges following this:

$$f_\mu(x_k) - f_\mu(x_\mu) \le \left(1 - \frac{m}{L}\right)^k (f_\mu(x_0) - f_\mu(x_\mu)) \qquad \text{(Let's Name this Convergence Equation)}$$

To ensure $f_\mu(x_k) - f_\mu(x_\mu) \le \epsilon$, we require

$$k \ge \frac{\ln\left(\frac{f_\mu(x_0) - f_\mu(x_\mu)}{\epsilon}\right)}{\ln\left(1 - \frac{m}{L}\right)}.$$

(c) Suppose $\hat{x}$ satisfies $f_\mu(\hat{x}) - f_\mu(x_\mu) \le \epsilon$. Find a tight upper bound on $f_0(\hat{x})$.
We are given a point $\hat{x}$ satisfying: $f_\mu(\hat{x}) - f_\mu(x_\mu) \le \epsilon$. We want to find a tight upper bound on $f_0(\hat{x})$. Note that $f_0(\hat{x}) = \frac{1}{N}\|A\hat{x} - b\|^2$, which corresponds to the unregularized objective function, or when $\mu = 0$. The regularized function is: $f_\mu(x) = \frac{1}{N}\|Ax - b\|^2 + \mu\|x\|^2$. Since x_μ is the minimizer, the function difference satisfies: $f_\mu(\hat{x}) - f_\mu(x_\mu) = \left(\frac{1}{N}\|A\hat{x} - b\|^2 + \mu\|\hat{x}\|^2\right) -$

$\left(\frac{1}{N}\|Ax_\mu - b\|^2 + \mu\|x_\mu\|^2\right)$. Rearrange to isolate $f_0(\hat{x})$, which means we will have $f_0(\hat{x}) = \frac{1}{N}\|A\hat{x} - b\|^2 = f_\mu(\hat{x}) - \mu\|\hat{x}\|^2$. Using the fact that $f_\mu(\hat{x}) - f_\mu(x_\mu) \leq \epsilon$, we substitute: $\frac{1}{N}\|A\hat{x} - b\|^2 \leq f_\mu(x_\mu)+\epsilon-\mu\|\hat{x}\|^2$. Since $f_\mu(x_\mu)$ is the minimum value of f_μ, we can upper bound it in terms of x_μ, which means we will have: $f_\mu(x_\mu) = \frac{1}{N}\|Ax_\mu - b\|^2+\mu\|x_\mu\|^2$. Thus, it follows that: $\frac{1}{N}\|A\hat{x}-b\|^2 \leq \left(\frac{1}{N}\|Ax_\mu - b\|^2 + \mu\|x_\mu\|^2\right) + \epsilon - \mu\|\hat{x}\|^2$. And we don't really need the parentheses for grouping so: $\frac{1}{N}\|A\hat{x}-b\|^2 \leq \frac{1}{N}\|Ax_\mu - b\|^2+\mu\|x_\mu\|^2+\epsilon-\mu\|\hat{x}\|^2$. We observe that $\frac{1}{N}\|Ax_\mu - b\|^2 = f_0(x_\mu)$. Thus, we substitute this into our bound: $f_0(\hat{x}) \leq f_0(x_\mu)+\mu\|x_\mu\|^2+\epsilon-\mu\|\hat{x}\|^2$. Since $\|\hat{x}\|^2 \geq 0$ and $\mu > 0$, it follows that $-\mu\|\hat{x}\|^2 \leq 0$. Thus, $f_0(\hat{x}) \leq f_0(x_\mu) + \mu\|x_\mu\|^2 + \epsilon - \mu\|\hat{x}\|^2 \leq f_0(x_\mu) + \mu\|x_\mu\|^2 + \epsilon$. Or more simply, we can say: $f_0(\hat{x}) \leq f_0(x_\mu) + \mu\|x_\mu\|^2 + \epsilon$. Thus, the tight upper bound that appropriately takes into consideration μ and approximation error is: $f_0(\hat{x}) \leq f_0(x_\mu)+\mu\|x_\mu\|^2+\epsilon$.

6. Consider the function

$$f(x) = \frac{1}{2}x^T Ax - b^T x$$

where A is symmetric and positive definite, and let $x^\star$ denote the minimizer. Show that if the initial point $x^{(0)}$ is chosen such that $x^{(0)} - x^\star$ is parallel to an eigenvector of A, then the steepest descent with exact line-search will find the minimizer in one step.

We have that $f(x) = \frac{1}{2}x^T Ax - b^T x$, and from this we can say, $\nabla f(x) = Ax - b$ and $Ax^\star = b$ as the minimizer $x^\star$ is found by setting $\nabla f(x) = 0$. Iteratively, we can write $x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$. We can rewrite this as $\nabla f(x^{(k)}) = Ax^{(k)} - b$. Furthermore, we can rewrite this statement using $x^\star$: $\nabla f(x^{(k)}) = A(x^{(k)} - x^\star)$ as $x^\star$ is our minimizer.

Now, we move onto the fact that for the purpose of this proof, we will set the initial point $x^{(0)}$ such that $x^{(0)} - x^\star$ is parallel to an eigenvector of A in order to show that such a scenario will have the steepest descent with exact line-search will find the minimizer in one step.

Proof. We can write $Av = \lambda v$ such that λ represents eigenvalue and v is our associated eigenvector. Since $x^{(0)} - x^\star = cv$ for some scalar c, we get that $A^{(0)}(x^{(0)} - x^\star) = cAv$. Because we have $Av = \lambda v$, we can say $A^{(0)}(x^{(0)} - x^\star) = c\lambda v$. Remember that $\nabla f(x^{(k)}) = A^{(0)}(x^{(0)} - x^\star)$. Thus, we can say that $\nabla f(x^{(k)}) = cAv$ or more desirably, $\nabla f(x^{(k)}) = c\lambda v$. We will have α_0 be our initial step size because it is determined by exact line search, which minimizes $f(x)$ in the initial stages along the steepest descent direction, and it ensures that each iteration moves as optimally as possible in the negative gradient direction.

To determine α_0, recall that the exact line search step size is chosen to minimize $f(x)$ along the search direction $-\nabla f(x^{(0)})$. This means that α_0 satisfies:

$$\alpha_0 = \arg\min_{\alpha} \ f(x^{(0)} - \alpha \nabla f(x^{(0)}))$$

Substituting the expression for $f(x)$:

$$f(x^{(0)} - \alpha \nabla f(x^{(0)})) = \frac{1}{2}(x^{(0)} - \alpha \nabla f(x^{(0)}))^T A(x^{(0)} - \alpha \nabla f(x^{(0)})) - b^T(x^{(0)} - \alpha \nabla f(x^{(0)}))$$

Since $\nabla f(x^{(0)})) = A(x^{(0)} - x^\star)$, define $d^{(0)} = -\nabla f(x^{(0)})$, so that $x^{(0)} - \alpha \nabla f(x^{(0)}) = x^{(0)} + ad^{(0)}$.

Plugging this into the function:

$$f(x^{(0)} + \alpha d^{(0)}) = \frac{1}{2}(x^{(0)} + \alpha d^{(0)})^T A(x^{(0)} + \alpha d^{(0)}) - b^T(x^{(0)} + \alpha d^{(0)})$$

Expanding, we get:

$$f(x^{(0)} + \alpha d^{(0)}) = \frac{1}{2}x^{(0)T} Ax^{(0)} + \alpha x^{(0)T} Ad^{(0)} + \frac{\alpha^2}{2} d^{(0)T} Ad^{(0)} - b^T x^{(0)} - \alpha b^T d^{(0)}$$

Since $Ax^\star = b$, we rewrite:

$$f(x^{(0)} + \alpha d^{(0)}) = f(x^{(0)}) + \alpha x^{(0)T} A d^{(0)} + \frac{\alpha^2}{2} d^{(0)T} A d^{(0)} - \alpha b^T d^{(0)}$$

Since $d^{(0)} = -A(x^{(0)} - x^\star)$, we have:

$$x^{(0)T} A d^{(0)} = -(x^{(0)} - x^\star)^T A^2 (x^{(0)} - x^\star)$$

Additionally, from $d^{(0)} = -A(x^{(0)} - x^\star)$, it follows:

$$b^T d^{(0)} = -(x^{(0)} - x^\star)^T A b = -(x^{(0)} - x^\star)^T A^2 (x^{(0)} - x^\star)$$

Thus, the function to minimize reduces to:

$$f(\alpha) = f(x^{(0)}) - \alpha (x^{(0)} - x^\star)^T A^2 (x^{(0)} - x^\star) + \frac{\alpha^2}{2} (x^{(0)} - x^\star)^T A^3 (x^{(0)} - x^\star)$$

We now differentiate $f(\alpha)$ and find $\frac{d}{d\alpha} f(\alpha)$, so we take the derivative with respect to α:

$$\frac{d}{d\alpha} f(\alpha) = -(x^{(0)} - x^\star)^T A^2 (x^{(0)} - x^\star) + \alpha (x^{(0)} - x^\star)^T A^3 (x^{(0)} - x^\star)$$

We set $\frac{d}{d\alpha} f(\alpha)$ to 0 in order to minimize and accordingly call α as α_0 because we are minimizing. So:

$$0 = -(x^{(0)} - x^\star)^T A^2 (x^{(0)} - x^\star) + \alpha_0 (x^{(0)} - x^\star)^T A^3 (x^{(0)} - x^\star)$$

$$\alpha_0 (x^{(0)} - x^\star)^T A^3 (x^{(0)} - x^\star) = (x^{(0)} - x^\star)^T A^2 (x^{(0)} - x^\star)$$

$$\alpha_0 = \frac{(x^{(0)} - x^\star)^T A^2 (x^{(0)} - x^\star)}{(x^{(0)} - x^\star)^T A^3 (x^{(0)} - x^\star)}$$

We will now introduce the use of eigenvector. Since $x^{(0)} - x^\star = cv$ for some scalar c and some eigenvector v for A with eigenvalue λ, we have the following: $Av = \lambda v$, $A^2 v = \lambda^2 v$, and $A^3 v = \lambda^3 v$. Thus, $\alpha_0 = \frac{(x^{(0)} - x^\star)^T A^2 (x^{(0)} - x^\star)}{(x^{(0)} - x^\star)^T A^3 (x^{(0)} - x^\star)}$ can be written as

$$\alpha_0 = \frac{c^2 \lambda^2 v^T v}{c^2 \lambda^3 v^T v}$$

Simplifying:

$$\alpha_0 = \frac{1}{\lambda}$$

We have $\alpha_0 = \frac{1}{\lambda}$, where λ is the eigenvalue corresponding to the eigenvector v of A. This confirms that in one step, the steepest descent method will reach $x^\star$, proving the claim. □

From our proof, we have that if the initial point $x^{(0)}$ is chosen such that $x^{(0)} - x^\star$ is parallel to an eigenvector of A, then the steepest descent with exact line-search will find the minimizer in one step.