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

EE381K: Large Scale Optimization — Fall 2015

Problem Set Four Solutions

Constantine Caramanis

Written Problems

1. **Gradient descent and non-convexity**

   Consider the gradient descent algorithm with fixed step size η for the function $f(x) = x'Qx$, where Q is symmetric but not positive semidefinite. (i.e., Q has some negative eigenvalues). Exactly describe the set of initial points from which gradient descent, with any *positive step size, will diverge. What happens at the other points, if the step size is small enough ?*

   Solution Consider a symmetric matrix $Q \in \mathbb{S}^n$. Its eigenvalue decomposition is given by $Q = U\Lambda U^T$ where U is a unitary matrix, i.e., $U^T U = I$, and Λ is a diagonal matrix with eigenvalues at the diagonal, i.e., $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

   The gradient descent step is then given by

   $$x_{k+1} = x_k - \eta \nabla f(x_k) = x_k - 2\eta Q x_k = (I - 2\eta Q)x_k = U(I - 2\eta\Lambda)U^T x_k.$$

   Let us define $\hat{x}_k \triangleq U^T x_k$. We have

   $$\hat{x}_{k+1} = (I - 2\eta\Lambda)\hat{x}_k,$$

   and it follows that

   $$\hat{x}_{k,i} = (1 - 2\eta\lambda_i)^k \hat{x}_{0,i}, \ i \in \{1, \ldots, n\}$$

   where $\hat{x}_{k,i}$ is the ith component of $\hat{x}_k$. For i such that λ_i is negative, $\hat{x}_{k,i}$ will diverge with any positive step size unless $\hat{x}_{0,i} = 0$. Thus, the set of initial points from which gradient descent with any $\eta > 0$ will diverge is given by

   $$\{x_0 = U\hat{x}_0 : \exists i, \hat{x}_{0,i} \neq 0, \lambda_i < 0\}$$

   Starting from the other points, i.e., $x_0 \in \{x_0 = U\hat{x}_0 : \forall i, \hat{x}_{0,i} = 0 \text{ or } \lambda_i \geq 0\}$, the sequence may converge to zero, converge to other points, or oscillate. If $\hat{x}_{0,i} \neq 0$ for the only indices i such that $\lambda_i > 0$, the component will converge to zero for a sufficiently small step size because $|1 - 2\eta\lambda_i| < 1$. Then $x_k = U\hat{x}_k$ will also converge to zero. If $\hat{x}_{0,i} \neq 0$ for some i such that $\lambda_i = 0$, the magnitude of the component will never change because $|1 - 2\eta\lambda_i| = 1$. Therefore, $x_k = U\hat{x}_k$ will not coverage to zero. It will converge to another point or oscillate.

2. **Jacobi Method**

   Recall that coordinate descent (with exact line search) involves minimizing over one coordinate at a time, keeping the other coordinates fixed. The Jacobi method involves, in a sense, doing all minimizations simultaneously. In particular, given a point x, define the vector $\bar{x}$, in *which the value at every coordinate i is determined by the corresponding* individual *coordinate descent update*

$$\bar{x}_i \;:=\; \arg\min_{\psi} f(x_1, \ldots, x_{i-1}, \psi, x_{i+1}, \ldots, x_n)$$

Thus, potentially, every coordinate of $\bar{x}$ could be different from that of x.

The Jacobi method is defined by the iteration

$$x_+ \;=\; x + \alpha(\bar{x} - x)$$

Prove that, for a convex continuously differentiable f, and a step size $\alpha = 1/n$ where n is the number of coordinates, the next iterate of the Jacobi method produces a lower function value than x, provided x does not already minimize the function.
(Hint: express x_+ as a convex combination of n points.)

Solution Define $\bar{x}^{(i)} \triangleq (x_1, \ldots, x_{i-1}, \bar{x}_i, x_{i+1}, \ldots, x_n)$. Then x_+ can be written as

$$x_+ = x + \frac{1}{n}(\bar{x} - x) = x + \frac{1}{n}\sum_{i=1}^{n}(\bar{x}^{(i)} - x) = \frac{1}{n}\sum_{i=1}^{n}\bar{x}^{(i)}.$$

Since f is convex, we obtain

$$f(x_+) = f\left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}^{(i)}\right) \le \frac{1}{n}\sum_{i=1}^{n} f(\bar{x}^{(i)}) < \frac{1}{n}\sum_{i=1}^{n} f(x) = f(x)$$

where the second inequality follows by the definition of $\bar{x}_i$.

3. **Step size in Newton**

Consider the use of Newton's method with constant step size t to minimize the function $\|x\|^3$.

(a) *For what values of t do we obtain global convergence to the minimum (i.e. $x^* = 0$) ? What happens for the other values of t ?*

Solution For $x \in \mathbb{R}^n$, the gradient and the Hessian of the function $f(x) = \|x\|^3$ is given by

$$\nabla f(x) = 3\|x\|x, \quad \nabla^2 f(x) = 3\|x\|I + \frac{3}{\|x\|}xx^\top$$

The eigenvalue decomposition of the Hessian can be written as

$$\begin{aligned}\nabla^2 f(x) &= U\Lambda U^\top \\ &= \left[\begin{array}{c|c|c|c} \dfrac{x}{\|x\|} & \hat{x}_1 & \cdots & \hat{x}_{n-1} \end{array}\right] \cdot \begin{bmatrix} 6\|x\| & 0 & \cdots & 0 \\ 0 & 3\|x\| & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 3\|x\| \end{bmatrix} \cdot \left[\begin{array}{c} x^\top/\|x\| \\ \hline \hat{x}_1^\top \\ \hline \vdots \\ \hline \hat{x}_{n-1}^\top \end{array}\right]\end{aligned}$$

where $\hat{x}_1, \ldots, \hat{x}_{n-1}$ and $x/\|x\|$ form the unitary matrix U, and the eigenvalues are all $3\|x\|$ except for the one equal to $6\|x\|$.

The Newton update is then given by

$$\begin{aligned}
x^+ &= x - t(\nabla^2\|x\|^3)^{-1}\nabla\|x\|^3 \\
&= x - t(U\Lambda^{-1}U^T)\cdot 3\|x\|x \\
&= x - 3\|x\|tU\Lambda^{-1}U^T x \\
&= x - 3\|x\|t\frac{x}{6\|x\|} \\
&= x\left(1 - \frac{t}{2}\right)
\end{aligned}$$

The sequence of x converges to zero, which is the global minimum, if $|1 - t/2| < 1$. Therefore, the step size should be $0 < t < 4$.

(b) *For the values of t for which it does converge, why is the convergence not quadratic ?*

Solution It is because the function $f(x) = \|x\|^3$ is too flat for x close to zero. The eigenvalue is proportional to the norm $\|x\|$, so there does not exists $m > 0$ satisfying $\nabla^2 f \succeq mI$. This demonstrates that the function is not strongly convex, and thus we cannot obtain quadratic convergence using the Newton's method.

4. **Composite functions**

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$ be a convex function, $\phi : \mathbb{R} \to \mathbb{R}$ be both convex and increasing, and define $g(x) = \phi(f(x))$. Assume that both f and g are twice differentiable everywhere. Note that this means g too is convex.

(a) *Consider an initial point $x^{(0)}$, and run two gradient descent with exact line search iterations: one on f, and the other one on g, with this same initial point. Show that the entire sequence of iterates will then be the same.*

Solution The two gradients $\nabla f(x)$ and $\nabla g(x) = \nabla\phi(f(x)) = \phi'(f(x))\nabla f(x)$ have the same direction. Since the exact line search finds the minimum point on the line with the opposite direction of the gradient, it finds the same point when the gradient directions are identical, no matter how large the magnitudes are.

(b) *Is the same true for the Newton methods with exact line search ? Prove your answer, or provide a simple counter-example.* (Hint: use the matrix inversion lemma of Appendix C.4.3 in the text)

Solution The gradient and the Hessian of the composite function are given by

$$\begin{aligned}
\nabla g(x) &= \phi'(f(x))\nabla f(x), \\
\nabla^2 g(x) &= \phi'(f(x))\nabla^2 f(x) + \phi''(f(x))\nabla f(x)\nabla f(x)^\top
\end{aligned}$$

Let us define

$$A \triangleq \phi'(f(x))\nabla^2 f(x),\ B \triangleq \sqrt{\phi''(f(x))}\nabla f(x)$$

Then we have

$$\begin{aligned}
\left(\nabla^2 g(x)\right)^{-1} \nabla g(x) \frac{\sqrt{\phi''(f(x))}}{\phi'(f(x))} &= (A + BB^\top)^{-1} B \\
&= (A^{-1} - A^{-1}B(1 + B^\top A^{-1}B)^{-1}B^\top A^{-1})B \\
&= A^{-1}B - A^{-1}BB^\top A^{-1}B(1 + B^\top A^{-1}B)^{-1} \\
&= A^{-1}B(1 - B^\top A^{-1}B(1 + B^\top A^{-1}B)^{-1}) \\
&= \gamma \frac{\sqrt{\phi''(f(x))}}{\phi'(f(x))} \left(\nabla^2 f(x)\right)^{-1} \nabla f(x)
\end{aligned}$$

where $\gamma = (1 - B^\top A^{-1}B(1 + B^\top A^{-1}B)^{-1})$. The second equality follows from the matrix inversion lemma, and the third equality follows from that $(1 + B^\top A^{-1}B)^{-1}$ is a scalar. We can see that the two Newton directions are the same, i.e., $\left(\nabla^2 g(x)\right)^{-1} \nabla g(x) = c\left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$ for some $c > 0$. Since we use the exact line search, the two sequences of iterates will be the same.