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

EE381K: Large Scale Optimization — Fall 2015

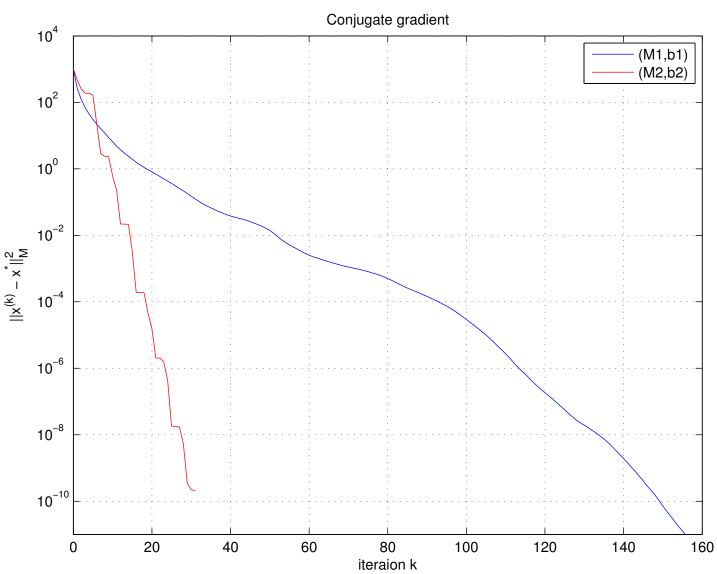
PROBLEM SET FIVE SOLUTIONS

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. **Conjugate Gradient Algorithm.** *Recall the linear conjugate gradient algorithm. Download the file* `http://users.ece.utexas.edu/~cmcaram/EE381V_2012F/ConjugateGradient.mat`*. There you will find matrices and vectors defining two equations: $M_1x = b_1$, and $M_2x = b_2$. The solution, x^*, is there as well, although this is easy to find since both M_1 and M_2 are invertible. Use conjugate gradient to solve these two linear systems, and plot the error, $\log(\|x^{(k)} - x^*\|^2_{M_i})$ vs. iteration k for both.*

   Solution The following figure shows the squared error versus number of iterations. The initial point is $x_0 = 0$, and the stopping criterion is $\|r_k\|_2 = \|Mx_k - b\|_2 < 10^{-4}$.



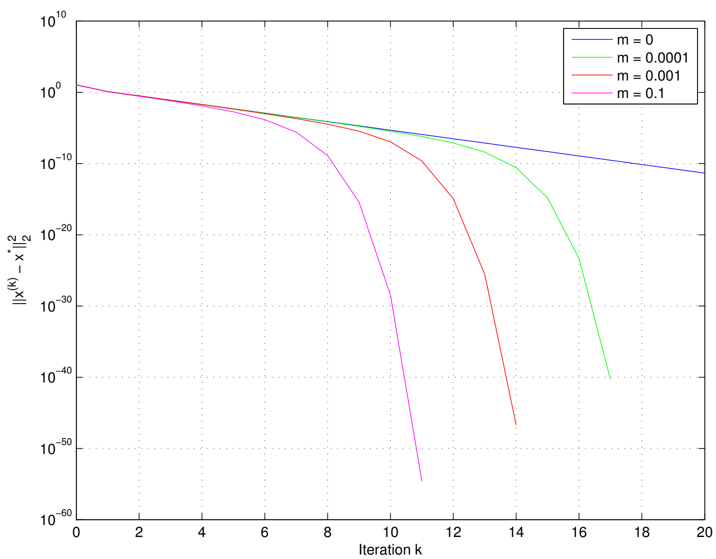
2. **Newton's Method**. *This problem will demonstrate the two convergence behaviors of Newton's method, damped and quadratic, by matlab simulation.*

   Consider $f_m(x) = \|x\|^3 + \frac{m}{2}\|x\|^2$ *for* $m \in \{0, 0.0001, 0.001, 0.1\}$ *and* $x \in \mathbb{R}^5$.

   (a) *For each* m*, implement Newton's method on* $f_m(x)$ *and provide the convergence plots, i.e* $\log(\|x^{(k)} - x^*\|^2)$ *vs. iteration* k*. Use the constant step size* $t = 1$.

   (b) *Using the condition for quadratic convergence, explain how and why your result changes according to* m.

   Solution The following figure shows the squared error versus number of Newton steps.



   We see that the convergence gets faster as m increases. The reason is that the function becomes more strongly convex as m grows, while its Hessian is Lipschitz with the same L. Since we always take a full Newton step in this problem, we can tell that x is in the region of quadratic convergence provided that $\|\nabla f(x)\| < m^2/L$. This explains that Newton iteration begins quadratic convergence within smaller number of steps for greater m.

3. **Central Path**. *Consider the linear optimization problem:*

$$
\begin{aligned}
\text{min :} \quad & 2x_1 + 4x_2 + x_3 + x_4 \\
\text{s.t. :} \quad & x_1 + 3x_2 + x_4 \leq 4 \\
& 2x_1 + x_2 \leq 3 \\
& x_2 + 4x_3 + x_4 \leq 3 \\
& x_i \geq 0, \quad i = 1, 2, 3, 4.
\end{aligned}
$$

   (a) *Find a function* F *that is a self-concordant-barrier function, such that the closure of its domain is equal to the feasible set of the problem. (Recall that* $-\log(a^\top x - b)$ *is a self-concordant-barrier function, as you show in an exercise below.)*

(b) *Find the analytic center* x_F^* *using Newton's method. You can initialize at any point in the domain (e.g.,* $(1/2, 1/2, 1/2, 1/2)$ *or any other point you like).*

$$x_F^* = \arg\min_{x \in \mathrm{dom} F} F(x).$$

(c) *Now you will generate the central path:*

$$x^*(t) = \arg\min_{x \in \mathrm{dom} F} f(t; x),$$

where recall:

$$f(t; x) = tc^\top x + F(x).$$

For $t = 0$*, the solution, and first point of the central path, is the analytic center. At each iteration, you will compute* $t_{k+1} = t_k(1 + \alpha)$*. Experiment with different values of* α*. If* α *is too small, progress may not be that fast as* t *will grow slowly. If* α *is too big, we might move outside the region of quadratic convergence, and although* t *will grow more quickly, each individual step of the central path will take longer to compute.*

(d) *Plot the error,* $\log(\|x^{(k)} - x^*\|)$ *as a function of number of iterations, for different values of* α*.*

Note: Chapter 11 in Boyd & Vandenberghe has much information about central path and barrier methods, although the chapter also contains a lot of information, definitions and ideas we have not yet discussed.

Solution There are many self-concordant-barrier functions we can use. The central path will be different depending on the function. A typical example is the log-barrier function, given by

$$F(x) = \sum_i \log(b_i - a_i^\top x),$$

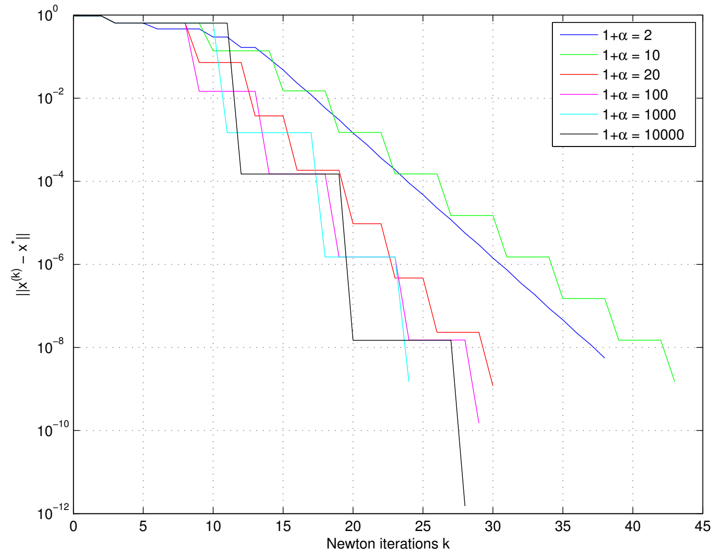
where a_i and b_i are picked such that each linear constraint is written as $a_i^\top x \le b_i$. For this problem, they are given by

$$\begin{aligned}
&a_1^\top = [1, 3, 0, 1],\ b_1 = 4, \quad a_2^\top = [2, 1, 0, 0],\ b_2 = 3, \quad a_3^\top = [0, 1, 4, 1],\ b_3 = 3,\\
&a_4^\top = [-1, 0, 0, 0],\ b_4 = 0, \quad a_5^\top = [0, -1, 0, 0],\ b_5 = 0, \quad a_6^\top = [0, 0, -1, 0],\ b_6 = 0,\\
&a_7^\top = [0, 0, 0, -1],\ b_7 = 0
\end{aligned}$$

In this example solution, the backtracking line search with $\alpha = 0.01$ and $\beta = 0.5$ are chosen. (Note that α in the backtracking line search is a different notation from α in the central path.) The stopping criterion for Newton's method is $\lambda(x)^2/2 < 10^{-5}$.

Using the log-barrier function, the obtained analytic center is at $(0.5547, 0.3096, 0.2541, 0.6572)$. The figure below shows the error as a functions of the number of inner Newton steps for different values of α in the central path. Each convergence plot looks like a staircase, as shown in Figure 11.4 of Boyd & Vandenberghe, where each stair represents a single outer iteration.

When α in the central path is small, it takes more outer iterations to converge, but each outer iteration takes fewer Newton steps. On the other hand, when α is large, the number of required outer iterations is small, while it takes more Newton steps for each outer iteration. As we see in the figure, the number of total Newton steps does not change much as we change the value of α, if α is large enough (more than about 20).

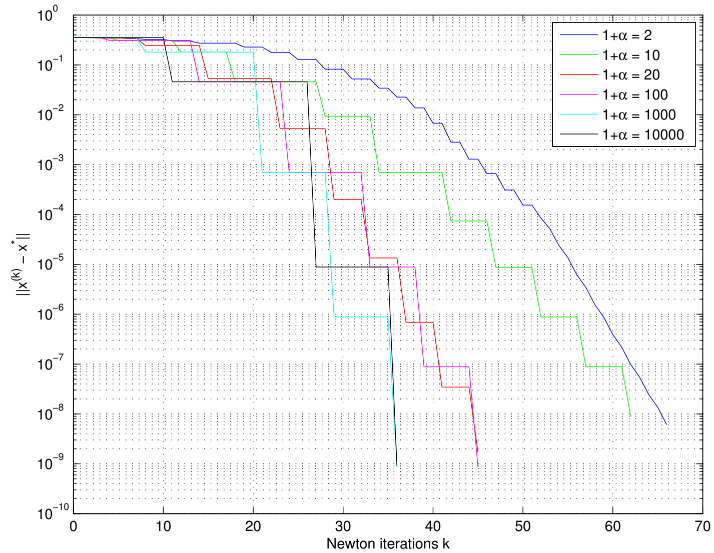


4. *Do the same for the (slightly larger) LP contained in* ***http://users.ece.utexas.edu/~cmcaram/EE381V_2012F/LP_centralpath.mat****. In that file you will find specified: c, A, and b, thus defining the problem:*

$$\begin{aligned} \text{min}: \quad & c^\top x \\ \text{s.t.}: \quad & Ax \leq b \\ & x \geq 0. \end{aligned}$$

Note that you can use CVX to quickly solve both this LP and the previous problem, in order to have the solution.

The following figure shows the convergence result for this large LP where all the parameters are chosen the same as the previous problem. We can also see in this figure that the number of required Newton iterations does not depend on the value of α critically, provided that α is large enough.

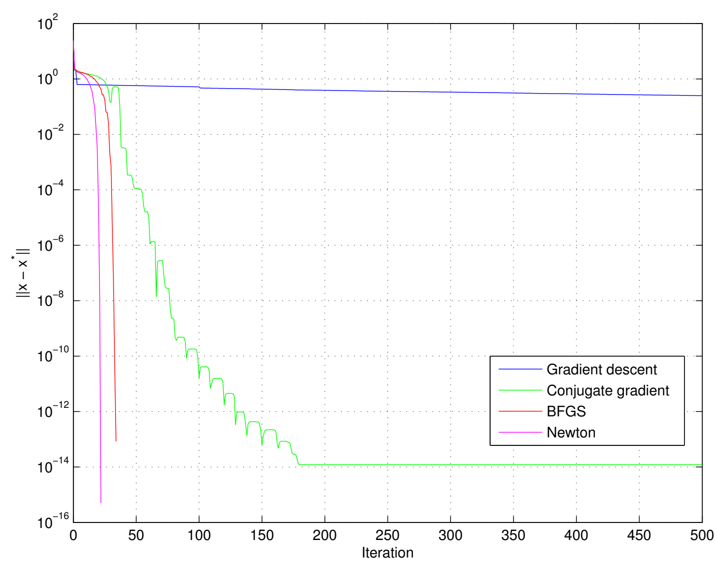


5. **Gradient, Conjugate Gradient, Newton and BFGS**. *Consider the function (called the Rosenbrock function)*

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

This function is not convex, however it not hard to see that it has a unique minimizer $x^* = (1, 1)$, *and that in a neighborhood of this point, the Hessian is positive definite. Initializing at* $x_{\text{init}} = (-1.2, 1)$, *implement (a) gradient descent, (b) (non-linear) Conjugate Gradient, (c) BFGS, and (d) Newton, using a back-tracking line search for all three. Plot the error in each as a function of the iteration.*

Solution Newton's method outperforms the other three, and BFGS follows up. The third is the conjugate gradient, and the gradient descent is the worst.



Written Problems

1. *We proved quadratic convergence of Newton's method (locally), using the assumption that f is strongly convex, smooth, and with L-Lipschitz Hessian. Locally (so, only after BTLS is taking full steps) derive a rate of convergence in the case where the Hessian is α-Hölder continuous. (Recall that a function F is called α-Hölder continuous if $\|F(x) - F(y)\| \leq H\|x-y\|^\alpha$ for all x, y.)*

   Solution As we proved quadratic convergence using L-Lipschitz Hessian, we can get

$$
\begin{aligned}
\|\nabla f(x^+)\| &= \|\nabla f(x + t\Delta x_{\mathrm{nt}})\| \\
&= \left\| \int_0^1 (\nabla^2 f(x + t\Delta x_{\mathrm{nt}}) - \nabla^2 f(x))\Delta x_{\mathrm{nt}} dt \right\| \\
&\leq \int_0^1 \left\|\nabla^2 f(x + t\Delta x_{\mathrm{nt}}) - \nabla^2 f(x)\right\| \|\Delta x_{\mathrm{nt}}\| \, dt \\
&\leq \int_0^1 H \|t\Delta x_{\mathrm{nt}}\|^\alpha \|\Delta x_{\mathrm{nt}}\| \, dt \\
&= \int_0^1 H \|\Delta x_{\mathrm{nt}}\|^{1+\alpha} t^\alpha dt \\
&= \frac{H}{1+\alpha} \|\Delta x_{\mathrm{nt}}\|^{1+\alpha} \\
&\leq \frac{H}{1+\alpha} \left\|(\nabla^2 f(x))^{-1} \nabla f(x)\right\|^{1+\alpha} \\
&\leq \frac{H}{1+\alpha} \left\|(\nabla^2 f(x))^{-1}\right\|^{1+\alpha} \|\nabla f(x)\|^{1+\alpha} \\
&\leq \frac{H}{(1+\alpha)m^{1+\alpha}} \|\nabla f(x)\|^{1+\alpha}
\end{aligned}
$$

   Let $C \triangleq \sqrt[\alpha]{\frac{H}{(1+\alpha)m^{1+\alpha}}}$. Then it follows that

$$
\left(C\|\nabla f(x^+)\|\right) \leq \left(C\|\nabla f(x)\|\right)^{1+\alpha}
$$

   which leads to

$$
\left(C\|\nabla f(x^{(K+k)})\|\right) \leq \left(C\|\nabla f(x^{(K)})\|\right)^{(1+\alpha)^k} \leq \left(\frac{1}{2}\right)^{(1+\alpha)^k}
$$

   where K is sufficiently large to ensure that BTLS takes a full step and $\|\nabla f(x^{(K)})\| \leq 1/2C$. Then, the suboptimality is bounded by

$$
f(x^{(K+k)}) - f(x^*) \leq \frac{1}{2m}\|\nabla f(x^{(K+k)})\|_2^2 \leq \frac{1}{2mC^2}\left(\frac{1}{4}\right)^{(1+\alpha)^k}
$$

   This shows that the local convergence is at least of order $1+\alpha$.

2. **Self Concordant Barriers.** *Recall that a self-concordant function f is called a self-concordant barrier if in addition to the properties for self-concordance, it satisfies the property that $\lambda_f(x)^2$ is uniformly bounded by some constant ν. That is, f is called a ν-self-concordant-barrier if*

$$\lambda_f(x)^2 = \|\nabla f(x)\|^2_{\nabla^2 f(x)^{-1}} \le \nu, \qquad x \in \mathrm{dom}(f).$$

(a) *Explain why we need this property (in connection to minimizing a linear function over a convex set).*

Solution Using a ν-self-concordant barrier function, we can obtain the central path by increasing t linearly and also keeping inner Newton steps in the region of quadratic convergence. Then the central path can be found within a small number of iterations. Let us consider the central path

$$x^*(t) = \arg\min_{x \in \mathrm{dom} F} f(t; x)$$

where f is defined as

$$f(t; x) = tc^\top x + F(x).$$

Suppose F is a ν-self-concordant-barrier, i.e., $\lambda_F(x)^2 = \|\nabla F(x)\|^2_{\nabla^2 F(x)^{-1}} \le \nu$ for $x \in \mathrm{dom}(F)$. Also, assume that we increase t linearly, i.e., $t^+ = (1+\alpha)t$.

When an inner Newton iteration for some t has terminated at x, we expect

$$\nabla f(t; x) = tc + \nabla F(x) = 0.$$

Then, at the beginning of the next Newton iteration with $t^+ = (1+\alpha)t$, we have

$$\begin{aligned}
\lambda_f(t^+; x)^2 &= \|\nabla f((1+\alpha)t; x)\|^2_{\nabla^2 f(t;x)^{-1}} \\
&= \|\nabla f((1+\alpha)t; x)\|^2_{\nabla^2 F(x)^{-1}} \\
&= \|\alpha t c\|^2_{\nabla^2 F(x)^{-1}} \\
&= \|\alpha \nabla F(x)\|^2_{\nabla^2 F(x)^{-1}} \\
&\le \alpha^2 \nu
\end{aligned}$$

Provided that we have picked small enough α, the Newton decrement is also small enough to ensure that the next Newton iteration begins in the region of quadratic convergence.

(b) *For A a positive semidefinite matrix, consider the concave quadratic $\phi(x) = c + b^\top x - \frac{1}{2}x^\top Ax$. Show that $f(x) = -\ln\phi(x)$ is a ν-self-concordant barrier function, with $\nu = 1$.*

Solution The gradient and Hessian of $f(x) = -\ln\phi(x)$ are given by

$$\begin{aligned}
\nabla f(x) &= -\frac{1}{\phi(x)}\nabla\phi(x) \\
\nabla^2 f(x) &= -\frac{1}{\phi(x)}\nabla^2\phi(x) + \frac{1}{\phi^2(x)}\nabla\phi(x)\nabla\phi(x)^\top = \frac{1}{\phi(x)}A + \nabla f(x)\nabla f(x)^\top
\end{aligned}$$

Since we know that $\|\cdot\|_{\nabla^2 f(x)^{-1}}$ is the dual norm of $\|\cdot\|_{\nabla^2 f(x)}$, which can be written as

$$\|\nabla f(x)\|_{\nabla^2 f(x)^{-1}} = \sup_{y:\|y\|_{\nabla^2 f(x)}=1} \nabla f(x)^\top y,$$

it is sufficient to prove $\nabla f(x)^\top y \leq 1$ for any y such that $\|y\|_{\nabla^2 f(x)} = 1$. We have

$$\begin{aligned}
\|y\|^2_{\nabla^2 f(x)} &= \frac{1}{\phi(x)} y^\top A y + y^\top \nabla f(x) \nabla f(x)^\top y \\
&= \frac{1}{\phi(x)} y^\top A y + (\nabla f(x)^\top y)^2 \\
&= 1
\end{aligned}$$

Since A is a positive semidefenite matrix, i.e., $y^\top A y \geq 0$ for every y, and $\phi(x) > 0$ for $x \in \mathrm{dom}(f)$, we have $\frac{1}{\phi(x)} y^\top A y \geq 0$. It follows that $\nabla f(x)^\top y \leq 1$, which completes the proof.

3. **Proof of the Conjugate Gradient Method, parts I & II.** These are written up in the lecture notes on Conjugate Gradient.