10-725/36-725: Convex Optimization **Fall 2018**

Lecture 16: October 22

Lecturer: Ryan Tibshirani *Scribes: Nic Dalmasso, Alan Mishler, Benjamin LeRoy*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

16.1 Equality-constrained Newton's Method

Considering the following problem

$$\min_x f(x) \quad \text{subject to} \quad Ax = b,$$

We might think at the following in order to find a solution:

- *Eliminating Equality Constraints*
  We set $x = Fy + x_0$, where F spans the null space of A, and $Ax_0 = b$. We then solve for y. The problem with this is that in doing so we might lose the structure of the Hessian of the original problem. Let's say the original problem Hessian is diagonal, by operating this transformation we have that:

  $$g(y) = f(FY + x_0) \implies \nabla^2 g(y) = F^T \nabla^2 f(Fy + x_0) F$$

  Generally when multiplying by F the structure of the original Hessian is lost - i.e. if the original Hessian was diagonal or sparse, the Hessian of the transformation will not anymore.

- *Deriving the Dual*
  We can use the Lagrange dual function, which is $-f^*(-A^T v) - b^T v$ and, with strong duality holding by refined Slater's condition, we can maybe express $x^\star$ in terms of the $v^\star$. Although, if A had some type of structure, we would lose it when considering the dual problem.

In this case we would like to apply Newton Method, but we should take into account that the new point proposed is feasible. In equality-constrained Newtons method, we start with $x^{(0)}$ such that $Ax^{(0)} = b$. Then we repeat the updates $x^+ = x + tv$ such that x^+ is again a feasible point, which we guarantee by selecting v in the following way:

$$v = \arg\min_{Az=0} \nabla f(x)^T (z - x) + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x),$$

where again we are approximate the function with a quadratic, as in the unconstrained Newton method. The new point x^+ is still feasible as

$$Ax^+ = Ax + tAv = b + 0 = b.$$

In comparison of the unconstrained Newton's method, this is a slightly bigger system to solve. Considering that v is the solution of minimizing a quadratic subject to equality constraint we have that:

$$\nabla^2 f(x) v = -\nabla f(x) \qquad \text{Uncostrained Newton's Method}$$

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix} \qquad \text{Equality-Constrained Newton's Method}$$

For some w.

16.1.1 Quasi-Newton Methods

Oftentimes it might be expensive to calculate or to store the Hessian matrix and approximate it might be preferable. Quasi-Newton methods aim to approximate the Hessian with a matrix $H \succ 0$, making the update $x^+ = x - tH^{-1}\nabla f(x)$. These methods do not converge at a quadratic rate like the Newton methods, but still enjoy a faster-than-linear convergence rate, denoted *superlinear* rate. The advantage for these methods is that they only take $O(n^2)$ operations per single iteration, which improves on the $O(n^3)$ operations the Newton method takes.

16.2 Barrier Method

Barrier method is an interior point method, category of which we will also explore the primal-dual method. The usefulness of barrier method comes into a place when considering the following hierarchy:

- *Quadratic Problems*: very easy to solve, we have a closed-form solution;
- *Equality-Constrained Quadratic Problems*: still easy, and we use KKT to derive closed-form solutions;
- *Equality-Constrained Smooth Problems*: this is where Newton's Method is useful, as it reduces these problems to a sequence of equality constrained quadratic problems, for which we can calculate the closed-form solutions to;
- *Inequality and Equality-Constrained Smooth Problems*: currently do not know how to solve, this is where interior point methods are useful.

The intuition behind the barrier method is that we are going to push the inequality constraints in the problem in a smooth way, reducing the problem to an equality-constrained problem.
Consider now the following minimization problem:

$$\min_x f(x)$$
$$\text{subject to } h_i(x) \le 0 \;\; i = 1, ..., m$$
$$Ax = b,$$

Where we assume f, h_1,...,h_m to be convex and twice differentiable functions, with the domain in $\mathbb{R}^n$. We define the log-barrier function as:

$$\phi(x) = -\sum_{i=1}^{m} \log(-h_i(x)).$$

The domain of this function is defined on the set of the x such that they are strictly feasible - i.e. they do not attain the equality in the inequality constrained. If we assume that this set is non-empty, which is going to be the case, we automatically also obtain strong duality by Slater's conditions.
The rationale behind the log-barrier function is that, when we ignore equality constraints, the problem above can be written by incorporating the inequality with the identity function, which in turn can be approximated by the log-barrier functions as follows

$$\min_x f(x) + \sum_{i=1}^{m} I_{\{h_i(x) \leq o\}}(x) \approx \min_x f(x) + \frac{1}{t}\phi(x) = \min_x f(x) - \frac{1}{t}\sum_{i=1}^{m} \log(-h_i(x)),$$

for $t > 0$ being a large number. This approximation is more accurate the larger the t, with also the log-barrier approaching ∞ if any $h_i(x) \to 0$.
Calculating the log-barrier function gradient and Hessian can be done using the chain rule and product rule respectively, resulting in:

$$\begin{aligned}
\nabla\phi(x) &= -\sum_{i=1}^{m} \frac{1}{h_i(x)} \nabla h_i(x), \\
\nabla^2\phi(x) &= \sum_{i=1}^{m} \frac{1}{h_i(x)^2} \nabla h_i(x) \nabla h_i(x)^T - \sum_{i=1}^{m} \frac{1}{h_i(x)} \nabla^2 h_i(x).
\end{aligned}$$

16.3 Central Path

We can formally articulate the barrier problem as follows:

$$\begin{aligned}
&\min_x t f(x) + \phi(x) \\
&\text{subject to } Ax = b
\end{aligned}$$

For $t > 0$. The central path is the solution $x^\star(t)$ for a given t. As $t \to \infty$, the log barrier function approaches the indicator function as described above, so we hope that the central path will approach the solution $x^\star$ of the original problem.

There are at least two reasons why we don't just solve for some large value t and hope the solution is close to $x^\star$:

1. As t gets large, Newton's method may get numerically unstable.
2. It takes Newton's method a long time to enter the pure convergence state if we do this. Instead, it's better to take a sequence of increasing t values and use each solution $X^\star(t)$ as a warm start for the next t. If this is done carefully, Newton's method can be kept in the pure convergence phase.

16.3.1 Barrier problem for a Linear Program

In a LP, the barrier problem is

$$\min_x c^T x - \sum_{i=1}^{m} \log(e_i - d_i^T x)$$

with the polyhedral constraint $Dx \leq e$. The contours of the log barrier function approximate this feasible set better and better from the inside as $t \to \infty$.

Taking the gradient setting it equal to 0, we have the gradient optimality condition:

$$0 = tc - \sum_{i=1}^{m} \frac{1}{e_i - e_i^T x^\star(t)} d_i$$

In other words, the gradient of the barrier function at the optimum, $\nabla\phi(x^\star(t))$ must be parallel to $-c$.

16.3.2 Bounding duality gap for original problem

The Lagrangian for the original problem is

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + v^T(Ax - b) \tag{16.1}$$

We can use the KKT conditions for the central path to identify feasible dual points for this Lagrangian. The KKT conditions for the central path are

$$t\nabla f(x^\star(t)) - \sum_{i=1}^{m} \frac{1}{h_i(x^\star(t))} \nabla h_i(x^\star(t)) + A^T w = 0 \tag{16.2}$$

$$Ax^\star(t) = b \tag{16.3}$$

$$h_i(x^\star(t)) < 0, i = 1, \dots m \tag{16.4}$$

For a given $x^\star(t), w$ that satisfy these conditions, let

$$u_i^\star(t) = -\frac{1}{th_i(x^\star(t))}, i = 1 \dots m \tag{16.5}$$

$$v^\star(t) = \frac{w}{t} \tag{16.6}$$

Since $h_i(x^\star(t) < 0$, due to the strict feasibility requirement, we know that $u_i(t) > 0$. Additionally, notice that $\nabla_x L(x^\star(t), u^\star(t), v^\star(t) = 0$, so $(u^\star(t), v^\star(t))$ is in the domain of (16.1) *and* $x^\star(t)$ minimizes $L(x, u^\star(t), v^\star(t))$ over x, so $g(u^\star(t), v^\star(t)) > -\infty$.

Therefore, $u^\star(t), v^\star(t)$ are feasible for the original problem.

We now have

$$g(u^\star(t), v^\star(t)) = f(x^\star(t)) + \sum_{i=1}^{m} u_i^\star(t) h_i(x^\star(t)) + v^\star(t)^T(Ax^\star(t) - b) \tag{16.7}$$

$$= f(x^\star(t)) - m/t \tag{16.8}$$

$$\implies f(x^\star(t)) - f^\star \leq f(x^\star(t)) - g(u^\star(t), v^\star(t)) \leq m/t \tag{16.9}$$

This means that $x^\star(t) \to x^\star$ as $t \to \infty$, as we had hoped. This also provide a useful stopping criterion.

16.4 Another interpretation of the central path

Above, we derived the central path from the barrier problem. Another way to understand the central path is as the solution to the following perturbed KKT conditions:

$$\nabla f(x) + \sum_{i=1}^{m} u_i \nabla h_i(x) + A^T v = 0 \tag{16.10}$$

$$u_i h_i(x) = -1/t, i = 1 \dots m \tag{16.11}$$

$$h_i(x) \leq 0, i = 1, \dots m, Ax = b \tag{16.12}$$

$$u_i \geq 0, i = 1, \dots m \tag{16.13}$$

i.e., we have replaced the usual complementary slackness equations $u_i h_i(x) = 0$ with $u_i h_i(x) = -1/t$. As $t \to \infty$, of course, the perturbed complementary slackness condition will approach the original condition.

16.5 Barrier Method

The Barrier method iteratively (as $t \to \infty$) solves a sequence of problems

$$\min_x \; tf(x) + \phi(x) \quad \text{s.t. } Ax = b \tag{16.14}$$

Allegorically, for a $t^{(0)} > 0$, $\mu > 1$ (t multiplier):

1. For $k \neq 0$, define $t^{(k)} = \mu t^{(k-1)}$
2. Solve Barrier problem (eq 16.14) and get $x^\star(t^{(k)})$ using Newton's method, initialized with $x^\star(t^{(k-1)})$ - for $k \neq 0$ (soft start)

16.5.1 Global Parameters

Although the barrier method is decently robust in practice to choices of μ and $t^{(0)}$,

1. $\mu \downarrow$: potentially lots off iterations to get to global convergence, $\mu \uparrow$: warm starts may not keep in quadratic convergence phase for Newton step
2. $t^{(0)} \downarrow$: potentially lots off iterations to get to global convergence, $t^{(0)} \uparrow$: first step may take a while to get into the quadratic convergence phase

16.5.2 Convergence

Relative to the central path approach of the Barrier method we get convergence of the form:

$$f(x^{(k)} - f^\star \leq \frac{m}{\mu^k t^{(0)}} \quad \text{or} \quad k \geq \frac{\log(\frac{m}{t^{(0)}\epsilon})}{\log(\mu)},$$

and noting that if we are in the quadratic convergence phase the number of steps for numerical convergence for the Newton method is 5 or 6, making this of the correct order if the warm starts keep the descents in the quadratic convergence phase.

16.5.3 Newton Iterations

The above comment about the Newton Method converging in at most 5-6 steps requires us to be in the quadratic convergence phase. This can be formalized with the following statement:

Suppose

- The function $tf + \phi$ is *self-concordant*
- our original problem has bounded sublevel sets

The solving Newton with approximate accuracy gets the total number of iterations $O(log(\frac{m}{t^{(0)}\epsilon}))$.

LP, QP and QCQPs are have $tf + \phi$ functions that are self-concordant (aka above fact is useful). (only requires f, h_i linear or quadratic).

16.6 Feasibility Methods

To leverage some of the above analysis, specifically using the dual formulation in Section 16.3.2, we needed to have a *strictly feasible* point (for Slater's condition). This is very important for $x^{(0)}$ estimation. To find a strictly feasible point we could apply the Barrier to

$$\begin{aligned} &\min_{x,s} \; s \\ &\text{s.t.} \;\; h_i(x) \le s i = 1, ..., n \\ &\qquad\; Ax = b \; , \end{aligned}$$

and terminate when $s \le 0$ (x would then be an optimal point). Deviations of this would also allow us to see which $h_i(x) \le 0$, if any cause the solution be be unsolvable but looking at

$$\begin{aligned} &\min_{x,s} \; 1^t s \\ &\text{s.t.} \;\; h_i(x) \le s_i i = 1, ..., n \\ &\qquad\; Ax = b, \quad s \ge 0 \; , \end{aligned}$$

which again, could be solved with the Barrier method.

16.7 History and Motivation for Interior-Point Methods

See lecture notes (*Primal-dual interior point methods*, slide 14) for some historical bullet points.