10-725/36-725: Convex Optimization **Fall 2015**

Lecture 15: October 12

Lecturer: Ryan Tibshirani *Scribes: Xuanchong Li, Wanchao Liang*

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

15.1 Last Lecture Review: Newton's Method

15.1.1 Equality-Constrained Newton's Method

Consider a equality-constrained problem,

$$\min_x f(x)$$

subject to

$$Ax = b$$

There are several options to handle the equality constrains.

- **Eliminating the equality constrains**: Express x as $x = Fy + x_0$ where F spans null space of A and $Ax_0 = b$
- **Deriving the dual**: Check the Lagrange dual function $-f^*(-A^T v) - b^T v$ and wether the strong duality holds. If so, we might be able to express $x^\star$ in terms of $v^\star$
- **Equality-constrained Newton**: Take the Newton update only in the direction of feasible set. This is the most straightforward option.

15.1.2 Equality-constrained Newton in Details

In equality-constrained Newton, we start from a feasible point x_0 such that $Ax_0 = b$. Then we do the update.

$$x^+ = x + tv$$

where

$$v = \underset{Az=0}{\arg\min} \bigtriangledown f(x)^T (z - x) + \frac{1}{2}(z - x)^T \bigtriangledown^2 f(x)(z - x)$$

The direction of v ensures that after each update, the point is still feasible.
To get v, we need to solve a quadratic programming with equality constraints. We can have closed form solution by writing down the KKT condition.

$$\begin{bmatrix} \bigtriangledown^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} - \bigtriangledown f(x) \\ 0 \end{bmatrix} \tag{15.1}$$

Then we can solve the linear system easily.

15.1.2.1 Newton's Method Summary

Consider the problem

$$\min_x f(x)$$

where f is convex, twice differentiable, with $dom(f) = R^n$. Newton's method starts from an initial $x^0 \in R^n$, repeat the update

$$x^k = x^{k-1} - t_k(\bigtriangledown^2 f(x^{k-1}))^{-1} \bigtriangledown f(x^{k-1}), k = 1, 2, 3, ...$$

The step size t_k can be chosen by backtracking line search.
If we assume $\bigtriangledown f$ Lipschitz, f strongly convex, $\bigtriangledown^2 f$ Lipschitz, then Newton's method has a convergence rate of $O(\log\log(\epsilon))$
Newton's method has two downsides.

- It requires compute Hessian. It might be expensive We can use quasi-Newton to approximate Hessian.
- It can only handle equality constraints. The barrier method can be applied to inequality constraints.

15.2 Hierarchy of Second-order method

Considering the convex problems, we have the hierarchy of problems in terms of difficulty.

- **Quadratic problem**: It is easiest since we can get closed form solution by setting the derivative to zero.
- **Quadratic problem with equality constrains**: It is still easy since we can write down the KKT conditions to derive the closed form solution.
- **Smooth problem with equality constrains**: Use Newton's method to handle equality constrains
- **Smooth problem with equality and inequality constrains**: these problems can be handled by interior-point method. It can be reduced to a sequence of equality-constrained smooth problems. Barrier method replaces the inequality constrains by smooth function in the criterion.

15.3 Log Barrier Function

Consider the convex optimization problem.

$$\min_x f(x)$$

subject to

$$h_i(x) \leq 0, i = 1, ..., m$$

$$Ax = b$$

Assume the criterion f and inequality constrains h_i are convex and twice differentiable, with domain R^n. We define the $\phi(x)$ as

$$\phi(x) = -\sum_{i=1}^{m} \log(-h_i(x)) \tag{15.2}$$

Its domain is the set of strictly feasible points.

If we ignore the equality constrains, the problem can be written as this.

$$\min_x f(x) + \sum_{i=1}^{m} I_{\{h_i(x)\leq 0\}}(x)$$

It can ben approximated by the barrier function:

$$\min_x f(x) - \frac{1}{t}\sum_{i=1}^{m} \log(-h_i(x)) \tag{15.3}$$

Here the t is a large number which controls how accurate the approximation is (shown in Figure 15.1). When t is larger, it is more similar as the indicator function. The barrier method is about solving the approximated problem without any inequality constrains.

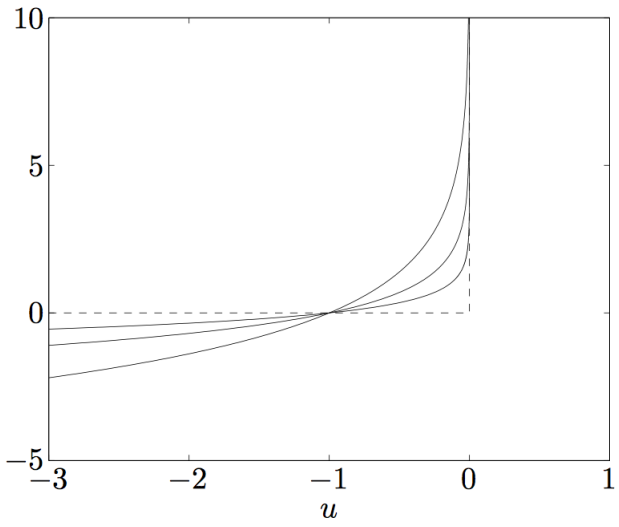


Figure 15.1: Approximation of indicator function

15.4 Log Barrier Calculus

Here we derive the first and second order derivative of log barrier function.

$$\phi(x) = -\sum_{i=1}^{m} \log(-h_i(x)) \tag{15.4}$$

$$\bigtriangledown\phi(x) = -\sum_{i=1}^{m} \frac{1}{h_i(x)} \bigtriangledown h_i(x) \tag{15.5}$$

$$\bigtriangledown^2\phi(x) = \sum_{i=1}^{m} \frac{1}{h_i(x)^2} \bigtriangledown h_i(x) \bigtriangledown h_i(x)^T - \sum_{i=1}^{m} \frac{1}{h_i(x)} \bigtriangledown^2 h_i(x) \tag{15.6}$$

15.5 Central Path

Replacing the inequality constrains with the log barrier function, we are solving the following equality-constrained problem rather than the original problem.

$$\min_x tf(x) + \phi(x)$$

subject to

$$Ax = b$$

For simplicity, the t is moved from log barrier function to the criterion.
Then we can write the solution of the above problem as a function of t. It is called the central path.

$$Ax^\star(t) = b, h_i(x^\star(t)) < 0, i = 1, ..., m$$

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

As we push t to infinity, the $x^\star(t)$ will converge to $x^\star$, which is the solution of the original problem. This is the motivation of barrier method.

15.6 Example: Linear Programming

The barrier problem for linear programming is

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

The original inequality constrain is $Dx \leq e$ The stationarity condition here is also called centrality condition. It is as follows.

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

It shows the gradient $\bigtriangledown\phi(x^\star(t))$ must be parallel to $-c$. It means at each iteration, the hyperplane $\{x : c^T x = c^T x^\star(t)\}$ lies tangent to the contour of ϕ at $x^\star(t)$ (shown in Figure 15.2)

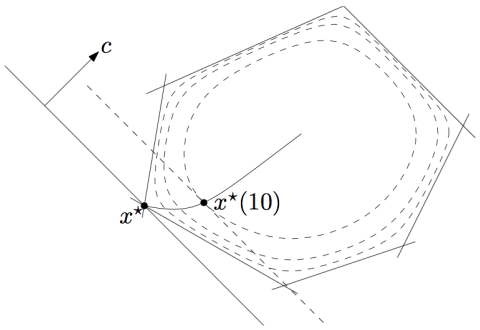


Figure 15.2: Central Path for Linear Programming

15.7 Dual Points From Central Path

We know the point $x^\star, w$ on the central path satisfy the following conditions

$$Ax^\star(t) = b, h_i(x^\star(t)) < 0, i = 1, ..., m$$

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

Divide t on the both sides of centrality condition.

$$\bigtriangledown f(x^\star(t)) - \sum_{i=1}^{m} \frac{1}{th_i(x^\star(t))} \bigtriangledown h_t(x^\star(t)) + A^T \frac{w}{t} = 0$$

Naturally, we can define $u^\star, v^\star$ as the dual points for original problem.

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

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

It is feasible since $h_i(x^\star(t)) < 0 \implies u_i^\star(t) > 0$. By its definition, $x^\star$ minimize the Lagrangian $L(x, u^\star(t), v^\star(t))$ over x. So $(u^\star(t), v^\star(t))$ lies in the domain of Lagrange dual function $g(u, v)$
Then we can write the $g(u, v)$

$$\begin{aligned} g(u^\star(t), v^\star(t)) &= f(x^\star(t)) + \sum_{i=1}^{m} u_i^\star(t) h_i(x^\star(t)) + v^\star(t)^T (Ax^\star(t) - b) \\ &= f(x^\star(t)) - \frac{m}{t} \end{aligned}$$

From this, we have the bound of duality gap

$$f(x^\star(t)) - f^\star \leq \frac{m}{t}$$

It can be the stopping criterion for barrier method. Also, it confirms that $\lim_{t\to\infty} x^\star(t) = x$

15.8 Interpretation via Perturbed KKT Conditions

Barrier method can be interpreted by another perspective: *perturbed KKT conditions*. So far, we can see the central path solution $x^\star(t)$ and its dual point $(u^\star(t), v^\star(t))$ satisfy the following *perturbed KKT conditions*:

- Stationarity:
$$\bigtriangledown f(x^\star(t)) + \sum_{i=1}^{m} u_i(x^\star(t)) \bigtriangledown h_i(x^\star(t)) + A^T v^\star(t) = 0$$
- Complementary slackness:
$$u_i^\star(t) h_i(x^\star(t)) = \frac{-1}{t}, i = 1, ..., m$$
- Primal feasibility:
$$h_i(x^\star(t)) \leq 0, i = 1, ..., m, Ax^\star(t) = b$$

- Dual feasibility:

$$u_i(x^\star(t)) \geq 0, i = 1, ..., m$$

Note that the only difference between these and the actual KKT conditions is the complementary slackness. In the actual KKT condition, it is

$$u_i^\star(t) h_i^\star(x^\star(t)) = 0, i = 1, ..., m$$

15.9 First attempt at an Algorithm

Why we need to repeatedly solve the log barrier problem for t be bigger and bigger, why is this necessary at all? Since we have seen the solution $x^\star(t)$ of

$$\min_x \quad f(x) + \phi(x)$$

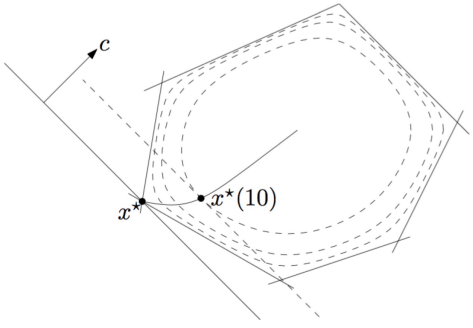
subject to

$$Ax = b$$

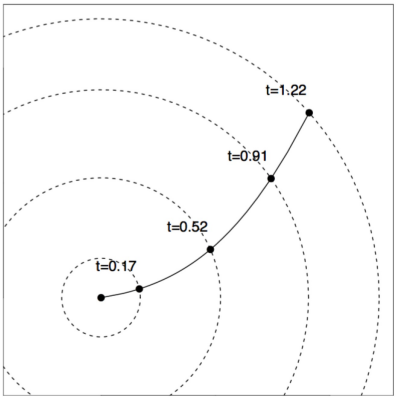
is no more than m/t suboptimal (we proved this from the KKT condition). Why don't we simply pick the desired accuracy level ϵ, set $t = m/\epsilon$, then solve the problem once using Newton's method and declare victory?

That is to say, we directly focus on the end of the central path. It seems a good idea in principle, but works poorly in practice because it incurs serious numerical problems of doing this. It requires t to be extremely large in order for this to be true, and because you are not starting anywhere close to optimality, and even not where to start, it costs serious issues.

A much better approach is to traverse the entire central path. The reason that it's a good approach intuitively is that when t is reasonably small, the objective is fairly smooth. Newton's method will converge pretty quickly. Then we make t larger, and solve a less smooth problem. But we will start at the solution of the previous value of t, we are already in a pretty good spot, which also leads quick convergence (we call it "warm start").



LP central path



Ridge regression solution path

From the left figure above, it can be known that we are staying fairly close to the central path, that will be helpful to reach the optimal point, rather than picking out the extreme point to begin with. Also, the

central path is closely related to the solution path of statistical optimization problems with warm starts, defined over a tuning parameter.

15.10 Barrier Method

In this section, we will present a formal algorithm. The barrier method basically solves problems repeatedly, makes t larger in each step with uses warm starts, and use Newton's method to solve the problem.

To state it formally, the barrier method solves a sequence of problems

$$\min_x \quad tf(x) + \phi(x)$$

subject to

$$Ax = b$$

We pick some initial value of $t = t^{(0)} > 0$, and solve the above problem using Newton's method to produce $x^{(0)} = x^*(t)$. Then for a barrier parameter $\mu > 1$, we repeat, for $k = 1, 2, 3, \ldots$

- Solve the barrier problem at $t = t^{(k)}$, using Newton's method initialized at $x^{(k-1)}$, to produce $x^{(k)} = x^*(t)$
- stop if the duality gap $m/t \leq \epsilon$
- Else update $t^{(k+1)} = \mu t$ (increase t by the barrier parameter)

The first step above is called a centering step (since it brings $x^{(k)}$ onto/close the central path). The idea of the barrier method is to always remain the quadratic convergence space. Although this algorithm seems simple enough, below is some considerations while implementing this method.

- The choice of μ (the barrier parameter): if μ is too small, then many outer iterations might be needed; if μ is too big, then there are many inner iterations, the Newton's method (each centering step) might take many iterations to converge.
- The choice of $t^{(0)}$: if $t^{(0)}$ is too small, then many outer iterations might be needed; if $t^{(0)}$ is too big, then the very first Newton's solve (first centering step) might require many iterations to compute $x^{(0)}$

Fortunately, the performance of the barrier method is often quite robust to the choice of the μ and $t^{(0)}$ in practice (However, note that the appropriate range for these parameters is scale dependent).

We can see an example of a small LP illustrating the effects of duality gap by the choices of μ in Fig. 15.3.

15.11 Convergence Analysis

Assume that we solve the centering steps exactly. The following result is immediate

Theorem 15.1 *The barrier method after k centering steps satisfies*

$$f(x^{(k)}) - f* \leq \frac{m}{\mu^k t^{(0)}}$$

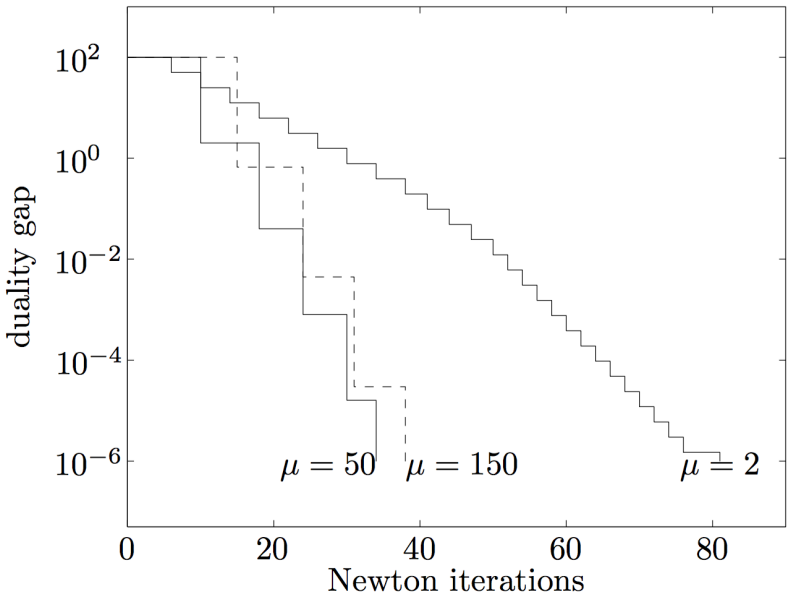


Figure 15.3: Example of a small LP with $n = 50$ dimensions, $m = 100$ inequalities (from B&V)

In other words, to reach a desired accuracy level of ϵ, we require

$$\frac{\log(m/t^{(0)}\epsilon)}{\log \mu} + 1$$

centering steps with the barrier method (plus the initial centering step).

Also, it is not reasonable to assume every Newton's method step gives us exact centering. But under mild conditions, Newton's method solves each centering problem to sufficiently high accuracy in nearly a constant number of iterations. (More precise statements can be made under self-concordance).

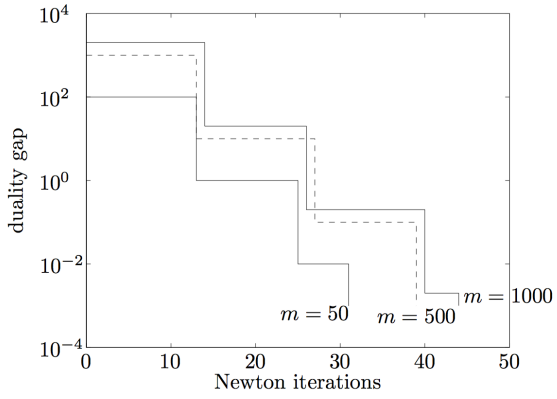


Figure 15.4: Example of a small LP in n = 50 dimensions, m = 100 inequalities (from B&V)

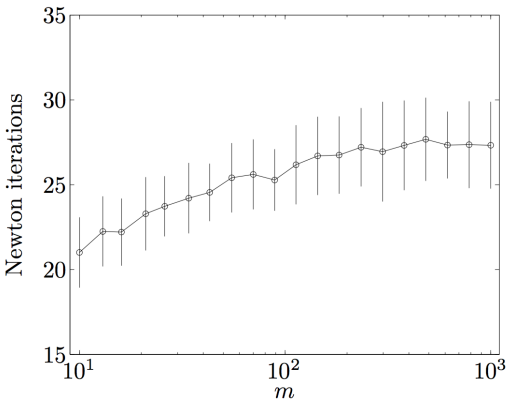


Figure 15.5: Example of barrier method for LP with m constraints (from B&V)

We can see from the above Fig.15.4 that in each case, it achieves roughly linear convergence, and logarithm scaling with m. Figure.15.5 shows that the number of Newton steps needed (to decrease the initial duality gap by factor of 10^4) grows very slowly with m.

15.12 Feasibility Methods

In the previous algorithm, we have implicitly assumed that we have a strictly feasible point for the first centering step, i.e., for computing $x^{(0)} = x^*$, solution of barrier problem at $t = t^{(0)}$

This is a point x such that

$$h_i(x) < 0, \quad i = 1, \ldots, m, \quad Ax = b$$

How to find such a feasible x? By solving

$$\min_{x,s} \quad s$$

subject to

$$\begin{aligned} &h_i(x) \leq s, \quad i = 1, \ldots, m \\ &Ax = b \end{aligned}$$

The goal is to find a strictly feasible point, and for s to be negative at the solution (then we quit). This is known as a feasibility method, and solving this problem is called running a feasibility program (phase 1 method in B&V). We can apply the barrier method to the above problem, since it is easy to find a strictly feasible starting point

Note that we do not need to solve this problem to high accuracy. Once we find a feasible (x, s) with $s < 0$, we can terminate early.

An alternative is to solve the problem

$$\min_{x,s} \quad 1^T s$$

subject to

$$\begin{aligned} &h_i(x) \leq s_i, \quad i = 1, \ldots, m \\ &Ax = b, \quad s \geq 0 \end{aligned}$$

Previously s was the maximum infeasibility across all inequalities. Now each inequality has own infeasibility variable $s_i, i = 1, \ldots, m$. So instead of having a global parameter for every constraint, we have a local parameter for every constraint.

One advantage: when the original system is infeasible, the solution of the above problem will be informative. The nonzero entries of s will tell us which of the constraints cannot be satisfied. The disadvantage is that it may be a slight more complicated to solve.