Barrier Method

Ryan Tibshirani
Convex Optimization 10-725/36-725

Last time: Newton's method

Consider the problem

$$\min \ f(x)$$

for f convex, twice differentiable, with $\mathrm{dom}(f) = \mathbb{R}^n$. Newton's method: choose initial $x^{(0)} \in \mathbb{R}^n$, repeat

$$x^{(k)} = x^{(k-1)} - t_k \big(\nabla^2 f(x^{(k-1)})\big)^{-1} \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Step sizes t_k chosen by backtracking line search

If ∇f Lipschitz, f strongly convex, $\nabla^2 f$ Lipschitz, then Newton's method has a local convergence rate $O(\log\log(1/\epsilon))$

Downsides:

- Requires solving systems in Hessian $\leftarrow$ quasi-Newton
- Can only handle equality constraints $\leftarrow$ this lecture

Hierarchy of second-order methods

Assuming all problems are convex, you can think of the following hierarchy that we've worked through:

- Quadratic problems are the easiest: closed-form solution
- Equality-constrained quadratic problems are still easy: we use KKT conditions to derive closed-form solution
- Equality-constrained smooth problems are next: use Newton's method to reduce this to a sequence of equality-constrained quadratic problems
- Inequality- and equality-constrained smooth problems are what we cover now: use interior point methods to reduce this to a sequence of equality-constrained smooth problems

Log barrier function

Consider the convex optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, \;\; i = 1, \dots m \\ & Ax = b \end{aligned}$$

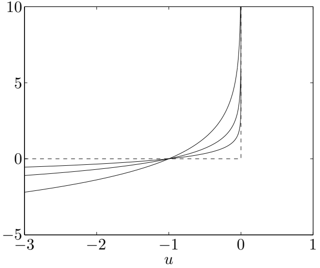
We will assume that f, $h_1, \dots h_m$ are convex, twice differentiable, each with domain $\mathbb{R}^n$. The function

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

is called the log barrier for the above problem. Its domain is the set of strictly feasible points, $\{x : h_i(x) < 0, \; i = 1, \dots m\}$, which we assume is nonempty

Ignoring equality constraints for now, our problem can be written as

$$\min \quad f(x) + \sum_{i=1}^{m} I_{\{h_i(x) \le 0\}}(x)$$



We approximate this representation by adding the log barrier function:

$$\min \quad f(x) - (1/t) \cdot \sum_{i=1}^{m} \log(-h_i(x))$$

where $t > 0$ is a large number

This approximation is more accurate for larger t. But for any value of t, the log barrier approaches ∞ if any $h_i(x) \to 0$

Outline

Today:

- Central path
- Properties and interpretations
- Barrier method
- Convergence analysis
- Feasibility methods

Log barrier calculus

For the log barrier function

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

let us write down its gradient and Hessian, for future reference:

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

and

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

computed using the chain rule

Central path

Consider minimizing our problem, after replacing hard inequalities with barrier term:

$$\begin{array}{ll} \min & tf(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array}$$

(Here we switched placement of t, but its role is the same.) The central path is defined as the solution $x^\star(t)$ as a function of $t > 0$. These solutions are characterized by the KKT conditions:

$$Ax^\star(t) = b, \quad h_i(x^\star(t)) < 0, \quad i = 1, \ldots m$$
$$t\nabla f(x^\star(t)) - \sum_{i=1}^m \frac{1}{h_i(x^\star(t))} \nabla h_i(x^\star(t)) + A^T w = 0$$

for some $w \in \mathbb{R}^m$. As $t \to \infty$, hope is that $x^\star(t) \to x^\star$, solution of our original problem

As an important example, consider the barrier problem for a linear program:

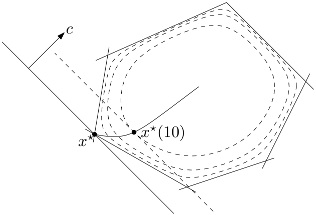
$$\min \quad tc^T x - \sum_{i=1}^{m} \log(e_i - d_i^T x)$$

The barrier function corresponds to polyhedral constraint $Dx \le e$

Stationarity or centrality condition:

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

This means that gradient $\nabla\phi(x^\star(t))$ must be parallel to $-c$, i.e., hyperplane $\{x : c^T x = c^T x^\star(t)\}$ lies tangent to contour of ϕ at $x^\star(t)$



(From B & V page 565)

Dual points from central path

From central points, we can derive feasible dual points for original problem. Given $x^\star(t)$ and corresponding w, we define

$$u_i^\star(t) = -\frac{1}{t h_i(x^\star(t))}, \quad i = 1, \dots m, \quad v^\star(t) = w/t$$

We claim $u^\star(t), v^\star(t)$ are dual feasible for original problem. Why?

- Note that $u_i^\star(t) > 0$ since $h_i(x^\star(t)) < 0$ for all i
- Further, the point $(u^\star(t), v^\star(t))$ lies in domain of Lagrange dual function $g(u, v)$, since by definition

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

  I.e., $x^\star(t)$ minimizes Lagrangian $L(x, u^\star(t), v^\star(t))$ over x, so $g(u^\star(t), v^\star(t)) > -\infty$

This allows us to bound suboptimality of $f(x^\star(t))$, with respect to original problem, via the duality gap. We compute

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

That is, we know that $f(x^\star(t)) - f^\star \le m/t$

This will be very useful as a stopping criterion; it also confirms the fact that $x^\star(t) \to x^\star$ as $t \to \infty$

Interpretation via perturbed KKT conditions

We can think of central path solution $x^\star(t)$ and corresponding dual point $(u^\star(t), v^\star(t))$ as solving the perturbed KKT conditions

$$\nabla f(x^\star(t)) + \sum_{i=1}^m u_i(x^\star(t)) \nabla h_i(x^\star(t)) + A^T v^\star(t) = 0$$
$$u_i^\star(t) \cdot h_i(x^\star) = -1/t, \quad i = 1, \dots m$$
$$h_i(x^\star(t)) \le 0, \quad i = 1, \dots m, \quad Ax^\star(t) = b$$
$$u_i(x^\star(t)) \ge 0, \quad i = 1, \dots m$$

Only difference between these and actual KKT conditions for our original problem is in the second condition: these are replaced by

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

i.e., complementary slackness, in actual KKT conditions

First attempt at an algorithm

Since we have seen that solution $x^\star(t)$ of

$$\begin{array}{ll} \min & tf(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array}$$

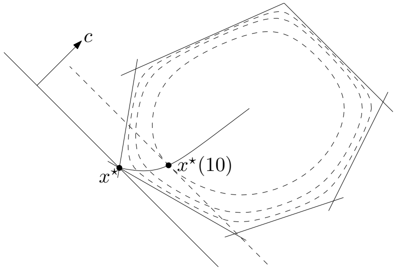
is no more than m/t suboptimal, why don't we simply pick desired accuracy level ϵ, set $t = m/\epsilon$, and then solve above problem using Newton's method?

This is like directly seeking out a point near the end of the central path. Problem is that the required t is often huge, and this causes serious numerical issues in practice. Hence this approach is almost never used

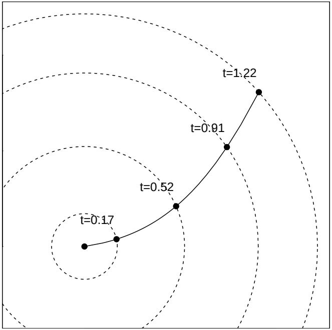
A better approach is to traverse the entire central path, in order to reach the end

Caveat: what does this remind you of?

The central path is closely related to the solution path of statistical optimization problems, defined over a tuning parameter



LP central path



Ridge regression solution path

Solving a statistical problem over a grid of tuning parameter values with warm starts is deeply connected to the central path concept (but this connection is not as developed as it perhaps could be ...)

Barrier method

The barrier method solves a sequence of problems

$$\begin{array}{ll} \min & tf(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array}$$

for increasing values of $t > 0$, until $m/t \le \epsilon$. We start at a value $t = t^{(0)} > 0$, and solve the above problem using Newton's method to produce $x^{(0)} = x^\star(t)$. 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 $x^{(k)} = x^\star(t)$
- Stop if $m/t \le \epsilon$
- Else update $t^{(k+1)} = \mu t$

The first step above is called a centering step (since it brings $x^{(k)}$ onto the central path)

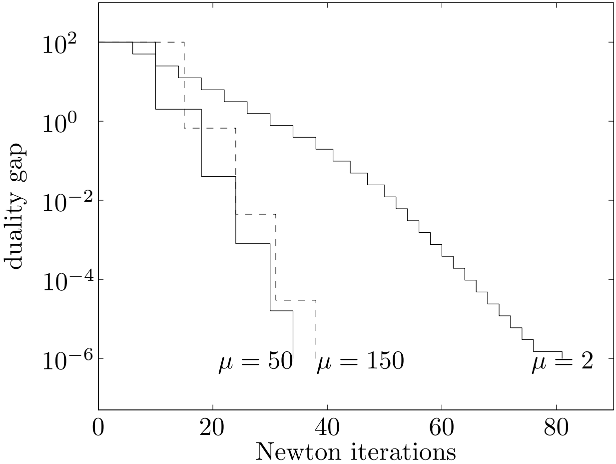
Considerations:

- Choice of μ: if μ is too small, then many outer iterations might be needed; if μ is too big, then Newton's method (each centering step) might take many iterations to converge
- 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 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 μ and $t^{(0)}$ in practice

(However, note that the appropriate range for these parameters is scale dependent)

Example of a small LP in $n = 50$ dimensions, $m = 100$ inequality constraints (from B & V page 571):



Convergence analysis

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

Theorem: The barrier method after k centering steps satisfies

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

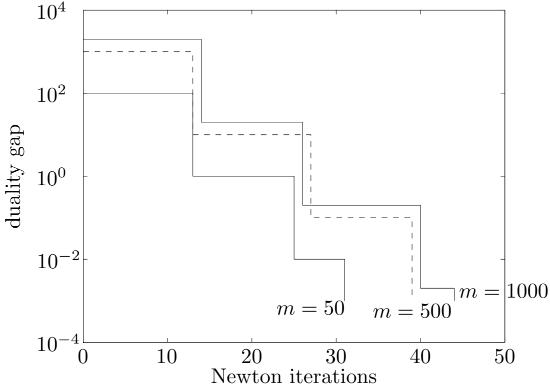
In other words, to reach a desired accuracy level of ϵ, we require

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

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

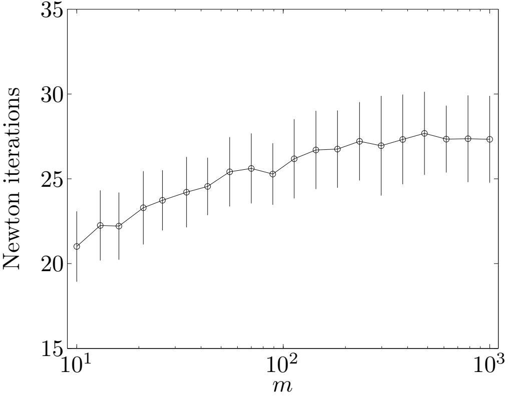
Is it reasonable to assume exact centering? Under mild conditions, Newton's method solves each centering problem to sufficiently high accuracy in nearly a constant number of iterations. (More precise statements can be made under self-concordance)

Example of barrier method progress for an LP with m constraints (from B & V page 575):



Can see roughly linear convergence in each case, and logarithmic scaling with m

Seen differently, the number of Newton steps needed (to decrease initial duality gap by factor of 10^4) grows very slowly with m:



Note that the cost of a single Newton step does depends on m (and moreso, on the problem dimension n)

Feasibility methods

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

This is a point x such that

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

How to find such a feasible x? By solving

$$\begin{aligned} \min_{x,s} \quad & s \\ \text{subject to} \quad & h_i(x) \le s, \quad i = 1, \dots m \\ & Ax = b \end{aligned}$$

The goal is for s to be negative at the solution. This is known as a feasibility method. 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

$$\begin{array}{ll} \min\limits_{x,s} & 1^T s \\ \text{subject to} & h_i(x) \le s_i, \quad i = 1, \dots m \\ & Ax = b, \quad s \ge 0 \end{array}$$

Previously s was the maximum infeasibility across all inequalities. Now each inequality has own infeasibility variable s_i, $i = 1, \dots m$

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

References and further reading

- S. Boyd and L. Vandenberghe (2004), "Convex optimization", Chapter 11
- A. Nemirovski (2004), "Interior point polynomial time methods in convex programming", Chapter 4
- J. Nocedal and S. Wright (2006), "Numerical optimization", Chapters 14 and 19