Interior Point Method

Yu-Xiang Wang
CS292F

(Based on Ryan Tibshirani's 10-725)

Last time: Newton's method

Consider the problem

$$\min_x \ f(x)$$

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

An important variant is equality-constrained Newton: start with $x^{(0)}$ such that $Ax^{(0)} = b$. Then we repeat the updates

$$x^+ = x + tv, \quad \text{where}$$

$$v = \underset{Az=0}{\operatorname{argmin}} \ \nabla f(x)^T (z - x) + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x)$$

which keep x^+ in feasible set $Ax = b$

Here v is characterized by KKT system

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

for some w. Hence Newton direction v is again given by solving a linear system in the Hessian (albeit a bigger one)

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-constrained and equality-constrained smooth problems are what we cover now: use interior-point methods to reduce this to a sequence of equality-constrained problems

Log barrier function

Consider the convex optimization problem

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

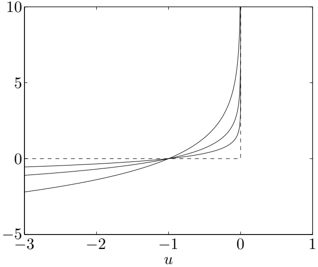
We will assume that f, $h_1, \dots h_m$ are convex, twice differentiable, each with domain $\mathbb{R}^n$. The function

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

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. (Note this implies strong duality holds)

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

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



We can approximate the sum of indicators by the log barrier:

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

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

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

we have for its gradient:

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

and for its Hessian:

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

Central path

Consider barrier problem:

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

The central path is defined as the solution $x^\star(t)$ as a function of $t > 0$

- Hope is that, as $t \to \infty$, we will have $x^\star(t) \to x^\star$, solution to our original problem
- Why don't we just set t to be some huge value, and solve the above problem? Directly seek solution at end of central path?
- Problem is that this is seriously inefficient in practice
- Much more efficient to traverse the central path, as we will see

An important special case: barrier problem for a linear program:

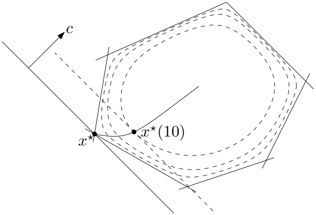
$$\min_{x} \ tc^T x - \sum_{i=1}^{m} \log(e_i - d_i^T x)$$

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

Gradient optimality 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)

KKT conditions and duality

Central path is characterized by its KKT conditions:

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

for some $w \in \mathbb{R}^m$. But we don't really care about dual variable for barrier problem ...

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

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

We claim $u^\star(t), v^\star(t)$ are dual feasible for original problem, whose Lagrangian is

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

Why?

- Note that $u_i^\star(t) > 0$ since $h_i(x^\star(t)) < 0$ for all $i = 1, \ldots, m$
- 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 $g(u^\star(t), v^\star(t)) > -\infty$

Duality gap

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 hope that $x^\star(t) \to x^\star$ as $t \to \infty$

Perturbed KKT conditions

We can now reinterpret central path $(x^\star(t), u^\star(t), v^\star(t))$ as solving the perturbed KKT conditions:

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

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

$$u_i \cdot h_i(x) = 0, \quad i = 1, \ldots m$$

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

Barrier method

The barrier method solves a sequence of problems

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

for increasing values of $t > 0$, until duality gap satisfies $m/t \leq \epsilon$

We fix $t^{(0)} > 0$, $\mu > 1$. We use Newton to compute $x^{(0)} = x^\star(t)$, a solution to barrier problem at $t = t^{(0)}$. For $k = 1, 2, 3, \ldots$

- Solve the barrier problem at $t = t^{(k)}$, using Newton initialized at $x^{(k-1)}$, to yield $x^{(k)} = x^\star(t)$
- Stop if $m/t \leq \epsilon$, 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)

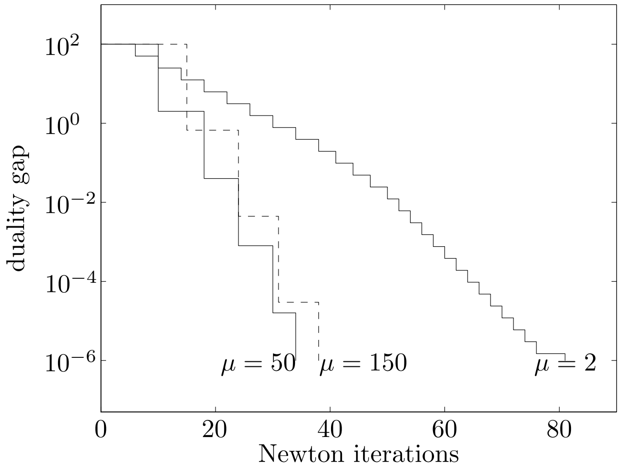
Considerations:

- 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
- 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 solve (first centering step) might require many iterations

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

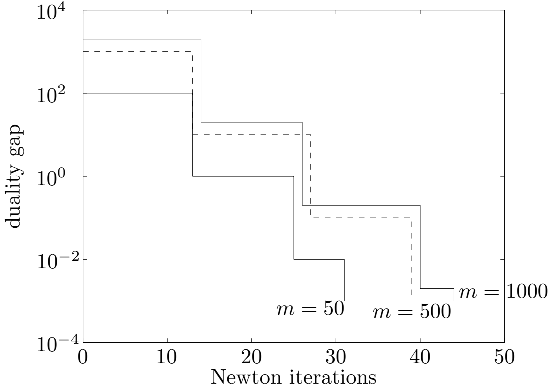
$$f(x^{(k)}) - f^\star \le \frac{m}{\mu^k t^{(0)}}$$

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

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

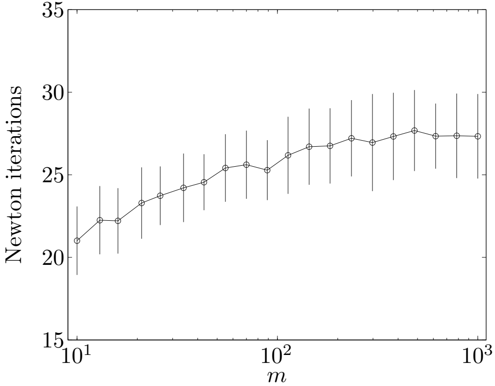
centering steps with the barrier method (plus initial centering step)

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 cost of a single Newton step does depends on m (and on problem dimension n)

How many Newton iterations?

Informally, due to careful central path traversal, in each centering step, Newton is already in quadratic convergence phase, so takes nearly constant number of iterations

This can be formalized under self-concordance. Suppose:

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

Then we can terminate each Newton solve at appropriate accuracy, and the total number of Newton iterations is still $O(\log(m/(t^{(0)}\epsilon))$ (where constants do not depend on problem-specific conditioning). See Chapter 11.5 of B & V

Importantly, $tf + \phi = tf - \sum_{i=1}^m \log(-h_i)$ is self-concordant when f, h_i are all linear or quadratic. So this covers LPs, QPs, QCQPs

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 x such that

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

How to find such a feasible x? By solving

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

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

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

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

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

Perturbed KKT conditions

Barrier method iterates $(x^\star(t), u^\star(t), v^\star(t))$ can be motivated as solving the perturbed KKT conditions:

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

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

$$u_i \cdot h_i(x) = 0, \quad i = 1, \ldots m$$

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

Perturbed KKT as nonlinear system

Can view this as a nonlinear system of equations, written as

$$r(x,u,v) = \begin{pmatrix} \nabla f(x) + Dh(x)^T u + A^T v \\ -\mathrm{diag}(u) h(x) - (1/t)1 \\ Ax - b \end{pmatrix} = 0$$

where

$$h(x) = \begin{pmatrix} h_1(x) \\ \dots \\ h_m(x) \end{pmatrix}, \quad Dh(x) = \begin{bmatrix} \nabla h_1(x)^T \\ \dots \\ \nabla h_m(x)^T \end{bmatrix}$$

Newton's method, recall, is generally a root-finder for a nonlinear system $F(y) = 0$. Approximating $F(y + \Delta y) \approx F(y) + DF(y)\Delta y$ leads to

$$\Delta y = -(DF(y))^{-1} F(y)$$

What happens if we apply this to $r(x,u,v) = 0$ above?

Newton on perturbed KKT, v1

Approach 1: from middle equation (relaxed comp slackness), note that $u_i = -1/(th_i(x))$, $i = 1, \ldots m$. So after eliminating u, we get

$$r(x, v) = \begin{pmatrix} \nabla f(x) + \sum_{i=1}^m (-\frac{1}{th_i(x)}) \nabla h_i(x) + A^T v \\ Ax - b \end{pmatrix} = 0$$

Thus the Newton root-finding update $(\Delta x, \Delta v)$ is determined by

$$\begin{bmatrix} H_{\text{bar}}(x) & A^T \\ A & 0 \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta v \end{pmatrix} = -r(x, v)$$

where $H_{\text{bar}}(x) = \nabla^2 f(x) + \sum_{i=1}^m \frac{1}{th_i(x)^2} \nabla h_i(x) \nabla h_i(x)^T + \sum_{i=1}^m (-\frac{1}{th_i(x)}) \nabla^2 h_i(x)$

This is just the KKT system solved by one iteration of Newton's method for minimizing the barrier problem

Newton on perturbed KKT, v2

Approach 2: directly apply Newton root-finding update, without eliminating u. Introduce notation

$$\begin{aligned} r_{\text{dual}} &= \nabla f(x) + Dh(x)^T u + A^T v \\ r_{\text{cent}} &= -\text{diag}(u)h(x) - (1/t)t \\ r_{\text{prim}} &= Ax - b \end{aligned}$$

called the dual, central, and primal residuals at $y = (x, u, v)$. Now root-finding update $\Delta y = (\Delta x, \Delta u, \Delta v)$ is given by

$$\begin{bmatrix} H_{\text{pd}}(x) & Dh(x)^T & A^T \\ -\text{diag}(u)Dh(x) & -\text{diag}(h(x)) & 0 \\ A & 0 & 0 \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta u \\ \Delta v \end{pmatrix} = - \begin{pmatrix} r_{\text{dual}} \\ r_{\text{cent}} \\ r_{\text{prim}} \end{pmatrix}$$

where $H_{\text{pd}}(x) = \nabla^2 f(x) + \sum_{i=1}^m u_i \nabla^2 h_i(x)$

Some notes:

- In v2, update directions for the primal and dual variables are inexorably linked together
- Also, v2 and v1 leads to different (nonequivalent) updates
- As we saw, one iteration of v1 is equivalent to inner iteration in the barrier method
- And v2 defines a new method called primal-dual interior-point method, that we will flesh out shortly
- One complication: in v2, the dual iterates are not necessarily feasible for the original dual problem ...

Surrogate duality gap

For barrier method, we have simple duality gap: m/t, since we set $u_i = -1/(th_i(x))$, $i = 1, \ldots m$ and saw this was dual feasible

For primal-dual interior-point method, we can construct surrogate duality gap:

$$\eta = -h(x)^T u = -\sum_{i=1}^{m} u_i h_i(x)$$

This would be a bonafide duality gap if we had feasible points, i.e., $r_{\text{prim}} = 0$ and $r_{\text{dual}} = 0$, but we don't, so it's not

What value of parameter t does this correspond to in perturbed KKT conditions? This is $t = m/\eta$

Primal-dual interior-point method

Putting it all together, we now have our primal-dual interior-point method. Start with $x^{(0)}$ such that $h_i(x^{(0)}) < 0$, $i = 1, \ldots, m$, and $u^{(0)} > 0$, $v^{(0)}$. Define $\eta^{(0)} = -h(x^{(0)})^T u^{(0)}$. We fix $\mu > 1$, repeat for $k = 1, 2, 3 \ldots$

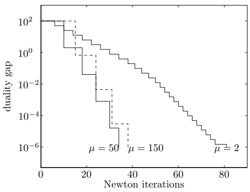
- Define $t = \mu m / \eta^{(k-1)}$
- Compute primal-dual update direction Δy
- Use backtracking to determine step size s
- Update $y^{(k)} = y^{(k-1)} + s \cdot \Delta y$
- Compute $\eta^{(k)} = -h(x^{(k)})^T u^{(k)}$
- Stop if $\eta^{(k)} \leq \epsilon$ and $(\|r_{\text{prim}}\|_2^2 + \|r_{\text{dual}}\|_2^2)^{1/2} \leq \epsilon$

Note that we stop based on surrogate duality gap, and approximate feasibility. (Line search maintains $h_i(x) < 0$, $u_i > 0$, $i = 1, \ldots, m$)

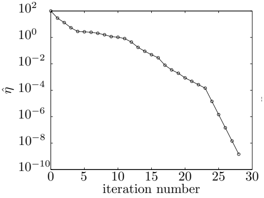
Example: barrier versus primal-dual

Example from B & V 11.3.2 and 11.7.4: standard LP with $n = 50$ variables and $m = 100$ equality constraints

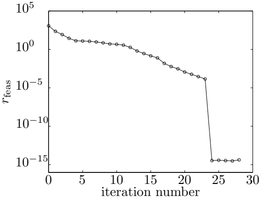
Barrier method uses various values of μ, primal-dual method uses $\mu = 10$. Both use $\alpha = 0.01$, $\beta = 0.5$



Barrier duality gap



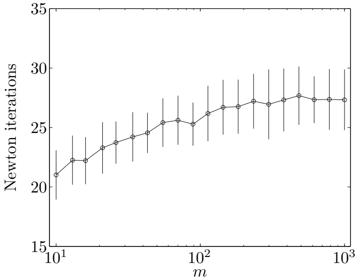
Primal-dual surrogate duality gap



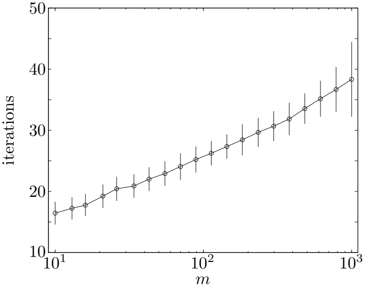
Primal-dual feasibility gap, where $r_{\mathrm{feas}} = (\|r_{\mathrm{prim}}\|_2^2 + \|r_{\mathrm{dual}}\|_2^2)^{1/2}$

Can see that primal-dual is faster to converge to high accuracy

Now a sequence of problems with $n = 2m$, and n growing. Barrier method uses $\mu = 100$, runs two outer loops (decreases duality gap by 10^4); primal-dual method uses $\mu = 10$, stops when surrogate duality gap and feasibility gap are at most 10^{-8}



Barrier method



Primal-dual method

Primal-dual method requires only slightly more iterations, despite the fact that it is producing much higher accuracy solutions

Some history

- Dantzig (1940s): the simplex method, still today is one of the most well-known/well-studied algorithms for LPs
- Klee and Minty (1972): pathological LP with n variables and $2n$ constraints, simplex method takes 2^n iterations to solve
- Khachiyan (1979): polynomial-time algorithm for LPs, based on ellipsoid method of Nemirovski and Yudin (1976). Strong in theory, weak in practice
- Karmarkar (1984): interior-point polynomial-time method for LPs. Fairly efficient (US Patent 4,744,026, expired in 2006)
- Renegar (1988): Newton-based interior-point algorithm for LP. Best known complexity ... until Lee and Sidford (2014)
- Spielman and Teng (2001): Smoothed Analysis of Simplex method: Why is it efficient in practice? (Goedel Prize, 2004)
- Modern state-of-the-art LP solvers typically use both simplex and interior-point methods

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
- S. Wright (1997), "Primal-dual interior-point methods," Chapters 5 and 6
- J. Renegar (2001), "A mathematical view of interior-point methods"