CS292F Convex Optimization: Concepts, Algorithms and Analysis **Spring 2020**

Lecture 12: May 6

Lecturer: Yu-Xiang Wang *Scribes: Zhuowei Cheng*

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

12.1 Additional Notes on Newton's method

Let's look at $f(x^{(k+1)}) \leq f(x^{(k)}) - \gamma$, for some $\gamma > 0$.

For non-self-concordant analysis. $K_0 = poly(L/m)$, For self concordant analysis $k_0 = O(1)$

12.1.1 Non self-concordant analysis

$x^+ = x + v = - \bigtriangledown^2 f(x)^{-1} \bigtriangledown f(x)$

$$\begin{aligned} f(x^+) &\leq f(x) + < \bigtriangledown f(x), tv > + \frac{Lt^2}{2}||v||^2 (L - smoothness) \\ &\leq f(x) - t \bigtriangledown f(x)^T \bigtriangledown^2 f(x)^{-1} \bigtriangledown f(x) + \frac{Lt^2}{2}|| \bigtriangledown^2 f(x)^{-1} \bigtriangledown f(x)||^2 \\ &= f(x) - t \bigtriangledown f(x)^T \bigtriangledown^2 f(x)^{-1} \bigtriangledown f(x) + \frac{Lt^2}{2}|| \bigtriangledown^2 f(x)^{-1} \bigtriangledown f(x)||^2 \end{aligned} \tag{12.1}$$

Let $\lambda^2(x) = \bigtriangledown f(x)^T \bigtriangledown^2 f(x)^{-1} \bigtriangledown f(x)$, use $|| \bigtriangledown^2 f(x)^{-1/2}||_{op} \leq \sqrt{\frac{1}{m}}$

$$\begin{aligned} f(x^+) &\leq f(x) - t\lambda^2(x) + \frac{Lt^2}{2m}|| \bigtriangledown^2 f(x)^{-1/2} \bigtriangledown f(x)||_2^2 \\ &= f(x) - t(1 - \frac{Lt}{2m})\lambda^2(x) \end{aligned} \tag{12.2}$$

t sufficiently small In the first phase, $k \leq k_0$, $|| \bigtriangledown f(x)||_2 \geq \eta_0$

Next step, involves lower bounding $\lambda^2(x)$ using η_0.

First, choose $t \leq \frac{m}{2}$

$$f(x) - t(1 - \frac{Lt}{2m})\lambda^2(x) \leq f(x) - \frac{t}{2}|| \bigtriangledown f(x)||^2_{\bigtriangledown^2 f^{-1}(x)} \tag{12.3}$$

Notice that, $|| \bigtriangledown f(x)||^2 \bigtriangledown^2 f^{-1}(x) \geq \frac{1}{L}|| \bigtriangledown f(x)||^2 = \frac{\eta^2}{L} = \frac{\eta^2}{L}$ where L is from $mI \leq \bigtriangledown^2 f(x) \leq LI$

with this fixed choice of t, we get $\gamma = \frac{m\eta^2}{2L^2}$. Back

12.1.2 self-concordant analysis

Definition 12.1. *f(x) is self-concordant if* $|f^{'''}| \leq 2f^{''}(x)^{\frac{3}{2}}$. *f(x) is R-self-concordant if* $|f^{'''}| \leq 2Rf^{''}(x)^{\frac{3}{2}}$. *When* $x \in R^d$, *this needs to apply to every direction.*

Property of self-concordance: if f(x) is self-concordant,

$$(1 - R||v||_H)^2 \bigtriangledown^2 f(x) \leq \bigtriangledown^2 f(x+v) \leq \frac{1}{1 - R||v||_{\bigtriangledown^2 f(x)}}^2 \bigtriangledown^2 f(x) \tag{12.4}$$

Let say, $\exists \xi \in [x, x+tv]$

$$\begin{aligned} f(x+tv) &= f(x) + < \bigtriangledown f(x), tv > + t^2 v^T \bigtriangledown^2 f(\xi) v \\ &= f(x) - t\lambda^2(x) + t^2 \bigtriangledown f(x)^T (\bigtriangledown^2 f(x))^{-1} \bigtriangledown^2 f(\xi) (\bigtriangledown f(x))^{-1} \bigtriangledown f(x) \\ &\leq f(x) - t\lambda^2(x) + t^2 \bigtriangledown f(x)^T (\bigtriangledown^2 f(x))^{-1} \frac{\bigtriangledown^2 f(x)}{(1 - t||v||_{\bigtriangledown^2 f(x)})^2} (\bigtriangledown^2 f(x))^{-1} \bigtriangledown f(x) \\ &\leq f(x) - t\lambda^2(x) + \frac{t^2}{(1 - t\lambda(x))^2} \lambda^2(x) \end{aligned} \tag{12.5}$$

Let $t \leq min\{\frac{1}{x\lambda(x)}, \frac{1}{8}\}$,

$$\begin{aligned} f(x) - t\lambda^2(x) + \frac{t^2}{1 - t\lambda(x)} \lambda^2(x) &\leq f(x) - t\lambda^2(x)(1 - \frac{t}{(1/2)^2})(t \leq \frac{1}{2\lambda(x)}) \\ &\leq f(x) - \frac{t}{2}\lambda^2(x) \leq f(x) - \frac{1}{16}\lambda^2(x)(t \leq \frac{1}{8}) \end{aligned} \tag{12.6}$$

12.2 Interior Point Methods

12.2.1 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

12.2.2 Log barrier function

Consider the convex optimization problem

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

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

$$min_x \;\; f(x) + \sum_{i=1}^{m} I_{\{h_i(x) \leq 0\}}(x) \tag{12.7}$$

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

$$min_x \;\; f(x) - \frac{1}{t} \sum_{i=1}^{m} log(-h_i(x)) \tag{12.8}$$

where $t > 0$ is a large number. This approximation is more accurate for larger t.

We assume that f, h_1, ..., h_m are convex, twice differentiable, each with domain R^n. The function

$$\phi(x) = -\sum_{i=1}^{m} log(-h_i(x)) \tag{12.9}$$

is called the log barrier for the above problem. Its domain is the set of strictly feasible points, $x : h_i(x) < 0, i = 1, ..., m$, which we assume is nonempty.(Note this implies strong duality holds)

For the log barrier function, we have for its gradient:

$$\bigtriangledown \phi(x) = -\sum_{i=1}^{m} \frac{1}{h_i(x)} \bigtriangledown^2 h_i(x) \tag{12.10}$$

and for its Hessian:

$$\bigtriangledown^2 \phi(x) = \sum_{i=1}^{m} \frac{1}{h_i(x)^2} \bigtriangledown h_i(x) \bigtriangledown h_i(x)^T - \sum_{i=1}^{m} \frac{1}{h_i(x)} \bigtriangledown^2 h_i(x) \tag{12.11}$$

12.2.3 Central Path

Consider barrier problem:

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

In linear and quadratic cases, $tf(x) + \phi(x)$ is self-concordant. The central path is defined as the solution $x^*(t)$ as a function of $t > 0$. Hope is that, as $t \longrightarrow \infty$, we will have $x^*(t) \longrightarrow x^*$. However, start with very large t can be more difficult. We can start with moderate t to solve the easier problem and then update t.

An special case is the barrier problem for a linear program:

$$min_x \ \ tc^T x + \sum_{i=1}^{m} log(e_i - d_i^T x) \tag{12.12}$$

Gradient optimality condition:

$$0 = tc + \sum_{i=1}^{m} \frac{1}{e_i - d_i^T x^*(t)} d_i \tag{12.13}$$

If we are close to the boundary, then the gradient is pushing you away. The gradient $\bigtriangledown\phi(x^*(t))$ must be parallel to $-c$.

Central path is characterized by its KKT conditions:

$$\begin{aligned} t \bigtriangledown f(x^*(t)) - \sum_{i=1}^{m} \frac{1}{h_i(x^*(t))} \bigtriangledown h_i(x^*(t)) + A^T x = 0 \\ Ax^*(t) = b, h_i(x^*(t)) < 0, i = 1, ..., m \end{aligned} \tag{12.14}$$

for some $x \in R^m$. Usually we don't care about dual variable at all when solving the primal. But sometimes having access to the feasible dual variable is useful. Suppose we can construct a feasible dual variable. By calculating the dual objective function and taking the difference we can get the duality gap. Use the duality gap to see if we are close the the optimal solution.

In the case, we can derive feasible dual points. Define

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

We claim $u_i^*(t)$, $v^*(t)$ are dual feasible for original problem, whose Lagrangian is

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

First we can easily see that $u_i^*(t) > 0$. Next, substitute with $u_i^*(t)$ and $v^*(t)$

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

For $x = x^*$

$$L(x, u, v) = f^*(x) + \sum_{i=1}^{m} \frac{1}{t} + \frac{w^T}{t}(Ax^* - b)$$

we get back to the minimizer of the Lagrange at u and v take the gradient of $min_x f(x) - \sum_{i=1}^{m} \frac{h_i(x)}{th_i^*(x)} + \frac{w^T(Ax-b)}{t}$ and assign to 0. u^* and v^* satisfy the stationarity condition.

Assigning $\bigtriangledown L(x, u^*, v^*) = 0$ give the optimal solution on the central path. $g(u^*(t), v^(t)*) > -\infty$ Thus u^* and v^* are feasible.

This allows us to bound suboptimality of $f(x^*(t))$, via the duality gap.

$$g(u^*(t), v^(t)*) = f(x^*(t)) - \frac{m}{t}$$

This confirms that $t \longrightarrow \infty$, we will have $x^*(t) \longrightarrow x^*$

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

$$\begin{aligned} \bigtriangledown f(x) + \sum_{i=1}^{m} u_i \bigtriangledown h_i(x) + A^T v = 0 \\ u_i h_i(x) = -\frac{1}{t}, t = 1, ...m \\ Ax = b, h_i(x^*(t)) \leq 0, i = 1, ..., m \\ u_i \geq 0, i = 1, ...m \end{aligned} \tag{12.15}$$

12.2.4 Barrier Method

The barrier method solves a sequence of problems

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

for increasing values of $t > 0$, until duality gap satisfies $\frac{m}{t} < \epsilon$.

In practice, the performance of the barrier method is often quite robust to the choice of μ and $t^{(0)}$.

12.2.5 Convergence analysis

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

Theorem 12.2. *The barrier method after k centering steps satisfies*

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

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

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

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

If we assume we are already in the quadratic convergence phase, then we can calculate how quickly it converges.

This can be formalized under self-concordant. 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 $O(log(\frac{m}{t^{(0)}\epsilon}))$. $tf + \phi = tf - \sum_{i=1}^{m} log(-h_i)$ is self-concordant when f, h_i are all linear or quadratic. So this covers LPs, QPs, QCQPs.

12.2.6 Feasiblity methods

We have assumed that we have a strictly feasible point for the first centering step, i.e., for computint $x^{(0)} = x^*$, solution of barrier problem at $t = t^{(0)}$. This is x such that

$$h_i(x) < 0, i = 1, ...m, Ax = b$$

How to find such a feasible x?

Choose any x such that Ax=b, substitute that into $h_i(x)$, then $s_i = h_i(x_0)$.

Solve

$$\begin{array}{c} min_{x,s} \quad s \\ \text{subject to} \quad h_i(x) \leq s, i = 1, ..., m \\ \text{Ax=b} \end{array}$$

approximately with barrier method. The goal is for s to be negative at the solution. We don't need to solve it to high accuracy. Once we find a feasible (x,s) with $s < 0$, we can terminate early.

An alternative is to solve the problem

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

The advantage is that when the original system is infeasible, the solution of the above problem is informative.

Revisiting the perturbed KKT condition, we can view it as a nonlinear system of equations.

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

where

$$h(x) = \begin{pmatrix} h_1(x) \\ ... \\ h_m(x) \end{pmatrix}, Dh(x) = \begin{bmatrix} \bigtriangledown h_1(x)^T \\ ... \\ \bigtriangledown h_m(x)^T \end{bmatrix}$$

Newton's method, is to exactly solve the nonlinear system F(y)=0. Approximating $F(y + \Delta y) \approx F(y) + DF(y)\Delta y$ leads to

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

Apply to r(x,u,v)=0. First approach is to use $u_i = -\frac{1}{(th_i(x))}$ and eliminate u, we get

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

Thus the Newton update(Δx, Δv) is determined by

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

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

Another way is to directly apply Newton root-finding update, without eliminating u.

$$\begin{aligned} r_{dual} &= \bigtriangledown f(x) + Dh(x)^T u + A^T v \\ r_{cent} &= -diag(u)h(x) - \frac{1}{t}t \\ r_{prim} &= Ax - b \end{aligned} \tag{12.16}$$

Now root-finding update $\Delta y = (\Delta x, \Delta u, \Delta v)$ is given by

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

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

The two methods are different algorithm that traces different sequences of solutions. The latter method gave us the primal-dual interior point method and the first version recovers the barrier method. One problem of the second approach is that the dual iterates are not necessarily feasible for the original dual problem.

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

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

It's an interesting heuristic but it's not a valid upper bound for the suboptimality. If the dual solution out of the method converge to the actual solution, then the surrogate duality gap should go to 0.

The primal-dual interior-point method is as follows:

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

References

[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"