MIT 6.854/18.415: Advanced Algorithms Spring 2016

Lecture 16 – April 4, 2016

Prof. Ankur Moitra *Scribes: David Vargas, Alex Markovits, Cenk Baykal*

1 Last Time (Gradient Descent)

Gradient descent is a greedy algorithm to minimize a convex function. We showed that, when the function is β-smooth and α-strongly convex, gradient descent converges to the correct answer at an exponential rate.

2 Today: Interior Point Methods

Let's introduce Interior Point Methods, a new way of solving convex optimization problems, following the work of Karmarker, Nestirov, and Nemirovskii [1] [2].

Recall the following optimization problem:

$$\text{minimize } c^T x \text{ such that } x \in K \text{ where } K \text{ is convex}$$

For the ellipsoid method we turned this into a feasibility problem where the objective function becomes a constraint.

Our new approach is to consider a new optimization problem ($\star$) and turn it into the above:

$$(\star) \text{ minimize } tc^T x + F(x) \text{ where t is a scalar and } F(x) \to \infty \text{ as } x \to \text{boundary}(K)$$

$F(x)$ is a "barrier" function that acts like a force field preventing us from choosing a point outside of K. Starting at $t = 0$ we will only care about the barrier function and would choose x to be the "analytic center" of K. As we increase t we will care more about the linear objective function and get closer to the right answer.

Now let's consider the solution to our new optimization problem:

$$x^\star(t) = \text{ argmin } (\star)$$

and let's work with the following set:

$$\{x^\star(t) \mid 0 \le t \le \infty\}.$$

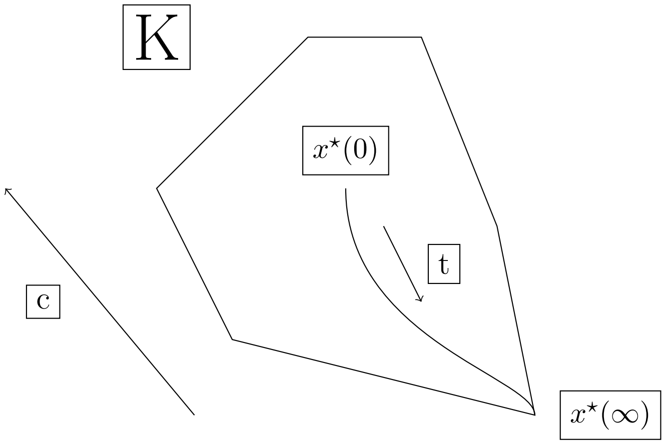
This set off solutions under varying values of t is called the "central path". It interpolates smoothly from $x^\star(0)$, the analytic center of K, to the optimal solution of the original problem, which should be on the boundary of K.

Example:

$$\text{Let } K = \{x \mid Ax \leq b\} \text{ , then we can choose } F(x) = -\sum_{i=1}^{m} \log(b_i - (Ax)_i)$$

Clearly as Ax approaches b the term inside the logarithm will go towards 0, making the logarithm explode towards negative infinity, making the entire term $-\sum_{i=1}^{m} \log(b_i - (Ax)_i)$ explode to infinity. This is a so-called "log barrier".

Here is a visualization:



3 Main Ingredients for Interior Point Methods

1. Newton's Method:

   If we are close to $x^\star(t)$, then Newton's method gives a simple iterative scheme which converges quadratically to it.

2. From $x^\star(t)$, we can consider some $t' > t$ and converge to $x^\star(t')$ quickly since the optima won't be too far apart.

3. There are fast methods to find $x^\star(0)$, the analytic center. We will not cover this in class.

4 Traditional Analysis of Newton's Method

Setup:

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^2. Then from Taylor's Theorem

$$f(x+h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) + o(\|h\|^2)$$

The basic idea of Newton's Method is to throw away the rightmost term of the Taylor expansion, and we find the h value we need to minimize the quadratic approximation of the function at x. We claim that:

$$\text{argmin}_h \; h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) h = -(\nabla^2 f(x))^{-1} \, \nabla f(x)$$

When we differentiate $h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) h$ with respect to h, set it equal to 0, and solve for h, we get $-(\nabla^2 f(x))^{-1} \, \nabla f(x)$. We can check that this actually gives a minimum since $f(x)$ is convex (and thus $\nabla^2 f(x)$ has non-negative eigenvalues).

To run Newton's method, we update:

$$x_{k+1} = x_k - \Big[\nabla^2 f(x_k)\Big]^{-1} \nabla f(x_k).$$

We present the following theorem to formally argue that this procedure quickly converges to the minimum of $f(x)$ as long as x_0 isn not too far away.

Theorem 1. *(Local quadratic convergence of Newton's Method) Suppose f satisfies the following 3 conditions:*

1. $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le R \|x - y\|_2$

2. $\nabla^2 f(x^*) \succeq \mu I$

3. $\|x_0 - x^*\|_2 \le \frac{\mu}{2R}$

Then: $\|x_{k+1} - x^*\|_2 \le \frac{R}{\mu} \|x_k - x^*\|_2^2$

The first condition is a condition on the smoothness of f, upper bounding the operator norm of the difference of two Hessians by the norm of the difference of the two vectors themselves. The second condition lower bounds all of the eigenvalues of the Hessian at x^*, where this value of x is a local optima of f. Our final condition simply upper bounds the distance between our initial and optimal values of x. Using all of these conditions we could prove that the distance between our updated value of x_{k+1} from the optimal is converging quadratically.

Proof. We write:

$$\nabla f(x + h) - \nabla f(x) = \int_0^1 \nabla^2 f(x + sh) h \, ds$$

For our purposes, x will be the optimal value x^* and $x^* + h$ is x_k. Because x^* is optimal, the gradient of f at this value is zero and we have the following relation:

$$\nabla f(x_k) = \int_0^1 \nabla^2 f(x^* + s(x_k - x^*))(x_k - x^*) ds$$

Now if we rewrite our update rule again but subtract x^* from both sides we have

$$x_{k+1} - x^* = x_k - x^* - \Big[\nabla^2 f(x_k)\Big]^{-1} \nabla f(x_k)$$

We will now make two substitutions, using the integral formula for the gradient of f at x_k and rewriting $(x_k - x^*) = (\nabla^2 f(x_k))^{-1}(\nabla^2 f(x_k))(x_k - x^*)$:

$$x_{k+1} - x^* = \Big[\nabla^2 f(x_k)\Big]^{-1}\Bigg[\int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + s(x_k - x^*))\right)(x_k - x^*)ds\Bigg]$$

We can now use the first condition in Theorem 1 (where y in this case is $x^* + s(x_k - x^*)$) to upper bound the norm of the integrand by:

$$\| \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + s(x_k - x^*))\right)\Big(x_k - x^*\Big)\|_2 \leq R||(1-s)x_k - (1-s)x^*||_2 = (1-s)R||x_k - x^*||_2.$$

So when we integrate with respect to s from 0 to 1 (applying the triangle inequality), and plug back into our relation for $x_{k+1} - x^*$, we get:

$$||x_{k+1} - x^*||_2 \leq \|\Big[\nabla^2 f(x_k)\Big]^{-1}\|_2\Big(\frac{R}{2}||x_k - x^*||_2^2\Big)$$

To finish this proof we just need to show the following claim:

Claim 2. $||\Big[\nabla^2 f(x_k)\Big]^{-1}||_2 \leq \frac{2}{\mu}$

To show this we will start with this simple identity:

$$\nabla^2 f(x_k) = \nabla^2 f(x^*) - \Big(\nabla^2 f(x^*) - \nabla^2 f(x_k)\Big).$$

Accordingly,

$$\lambda_{min}(\nabla^2 f(x_k)) \geq \lambda_{min}(\nabla^2 f(x^*)) - \lambda_{max}\Big(\nabla^2 f(x^*) - \nabla^2 f(x_k)\Big).$$

From the second condition of Theorem 1 we know that

$$\lambda_{min}(\nabla^2 f(x^*)) \geq \mu.$$

Additionally, applying the first and third conditions gives:

$$\lambda_{max}(\nabla^2 f(x^*) - \nabla^2 f(x_k)) \leq R||x_k - x^*||$$

$$\lambda_{max}(\nabla^2 f(x^*) - \nabla^2 f(x_k)) \leq \frac{\mu}{2}$$

This allows us to finish the proof for Claim 2:

$$\lambda_{min}(\nabla^2 f(x_k)) \geq \frac{\mu}{2} \implies ||\Big[\nabla^2 f(x_k)\Big]^{-1}||_2 \leq \frac{2}{\mu}$$

Which when substituted into the relation we had before claim 2 gives us the quadratic convergence that we wanted to show:

$$||x_{k+1} - x^*|| \leq \frac{R}{\mu}||x_k - x^*||_2^2$$

□

5 Affine Invariant Newton's Method

The conditions and our proof of Theorem 1 are not affine invariant, but Newton's method itself is. To show this, imagine applying the transformation $y = Ax$ (A is invertible) and optimizing the objective function $\phi(y) = f(A^{-1}y) = f(x)$. The gradient is the following

$$\nabla\phi(y) = (A^{-1})^T \nabla f(A^{-1}y)$$

and the Hessian satisfies

$$\nabla^2\phi(y) = (A^{-1})^T \nabla f(A^{-1}y) A^{-1}.$$

Accordingly, Newton's method will apply the step:

$$\Big[\nabla^2\phi(y)\Big]^{-1} \nabla\phi(y) = A\Big(\nabla^2 f(A^{-1}y)\Big)^{-1} \nabla f(A^{-1}y),$$

which is equivalent to minimizing the objective function in the x space, and then transforming it by A. So, Newton's method is invariant to any invertible transformation A.

Interior point methods will rely on a more general affine invariant analysis of Newton's method, which we don't proof for simplicity, but describe below. We begin by defining:

Definition 3. *(Newton Decrement) The Newton Decrement, $\lambda_f(x)$, is defined as:*

$$\lambda_f(x) := \sqrt{\nabla f(x)^T (\nabla^2 f(x))^{-1} \nabla f(x)},$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the function that we wish to minimize.

Note that by definition, the Newton Decrement is affine invariant. Intuitively, the Newton Decrement quantifies the radius of fast convergence for Newton's Method and hence serves as a proxy for getting to the optimal point. We note that the the Newton Decrement at a locally optimal point is equal to 0.

In light of this definition, we present the affine invariant guarantees of Newton's Method.

Theorem 4. *(Affine invariant local quadratic convergence of Newton's Method) Let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy*

$$\forall h \in \mathbb{R}^n \quad \nabla^3 f(x)[h,h,h] \le 2(h^T \nabla^2 f(x) h)^{\frac{3}{2}} \tag{1}$$

for all $x \in K$. Then, it follows that:

$$\lambda_f(x)\big(x - (\nabla f(x))^{-1} \nabla f(x)\big) \le \left(\frac{\lambda_f(x)}{1 - \lambda_f(x)}\right)^2 \tag{2}$$

for $\{\, x \,|\, \lambda_f(x) < 1\} \equiv$ "Newton decrement ball."

Proof. We refer the reader to pages 15-18 of Prof. Nisheeth Vishnoi's lecture notes [4]. □

We remark that condition (1) above is called the *self-concordance* property of function f.

We conclude this section by remarking that Theorem 4 comprises the first main ingredient for Interior Point methods: understanding how well Newton's Method performs in an affine invariant way if we are close to $x^*(t)$.

6 Analysis of Interior Point Methods

We now proceed to Step 2: if we have an optimal point for t, $x^*(t)$, by how much can we increase t and still remain within the radius of fast convergence (Newton decrement ball) for? We note that this natural question arises from the fact that we start off from the analytic center $x^*(0)$ and seek to reach the optimal solution, $x^*(\infty)$, through the central path by iteratively incrementing t.

Define $F_{t'}(x)$ as follows:

$$F_{t'}(x) := t'c^T x + F(x).$$

Given the minimizer for the old objective value, $x^*(t)$, we want to understand how large can we set $t' > t$ so that we remain in the Newton decrement ball, i.e. so that:

$$\lambda_{F_{t'}}(x^*(t)) < \frac{1}{4}.$$

We first simplify the expression for $\lambda_{F_{t'}}(x^*(t))$ by noting that

$$\begin{aligned}\nabla F_{t'}(x^*(t)) &= \nabla F_t(x^*(t)) + (t'-t)c \\ &= (t'-t)c\end{aligned}$$

where the last equality follows from the fact that $F_t(x^*(t)) = 0$ by definition of $x^*(t)$ as the minimizer of the old objective function $F_t(x)$. Thus, we have

$$\begin{aligned}\lambda_{F_{t'}}(x^*(t)) &= (t'-t)\sqrt{c^T(\nabla^2 F_{t'}(x^*(t)))^{-1}c} \\ &= (t'-t)\sqrt{c^T(\nabla^2 F(x^*(t)))^{-1}c} \qquad (3)\end{aligned}$$

Now, in addition to the self-concordance property (1), we need one more condition on the function to complete the analysis:

$$\nabla^2 F(x) \succeq \frac{1}{\nu}\nabla F(x)\big(\nabla F(x)\big)^T. \qquad (4)$$

We collect one more fact in addition to this condition:

$$\begin{aligned}\nabla F_t(x^*(t)) &= 0 \\ &= tc + \nabla F(x^*(t))\end{aligned}$$

which implies that

$$c = \frac{-\nabla F(x^*(t))}{t}.$$

Combining the above result with the previous condition (4), we simplify (3) to yield:

$$\lambda_{F_{t'}}(x^*(t)) \le \frac{(t'-t)}{t}\sqrt{\nu} \qquad (5)$$

which is exactly what we were looking for. Now, using (5), we can generate an appropriate value for t' such that $\lambda_{F_{t'}}(x^*(t)) < \frac{1}{4}$. Namely, note that $\frac{(t'-t)}{t}\sqrt{\nu} < \frac{1}{4}$ if and only if

$$t' \le t\big(1 + \frac{1}{4\sqrt{\nu}}\big)$$

which is precisely what we wanted to show.

In light of this result pertaining a bound on how large we can *boost* the value of t at each iteration while remaining within the Newton decrement ball, we can now assert a bound on the number of iterations required in order to obtain an ϵ-approximate solution.

Theorem 5. *(Bound on the number of iterations for an ϵ-approximate solution) $\forall \epsilon \in \mathbb{R}_+$, an ϵ-approximate solution to the original optimization problem can be obtained after k iterations, where k is:*

$$k = \mathcal{O}\big(\sqrt{\nu} \log \frac{\nu}{t_0 \epsilon}\big).$$

Proof Sketch. It's not too hard to check that a value of $t = \frac{\nu}{\epsilon}$ is sufficiently high to obtain an ϵ-approximate solution for our convex optimization problem [4]. With a multiplicative increase of $(1 + \frac{1}{4\sqrt{\nu}})$ at each iteration, it would take

$$k = \mathcal{O}\big(\sqrt{\nu} \log \frac{\nu}{t_0 \epsilon}\big)$$

iterations to boost from a starting value t_0 to $\frac{\nu}{\epsilon}$, which yields the result. Additional work is required to show that we can find some starting point which minimizes F_{t_0} for some $t_0 > 0$. □

To conclude, we reiterate that conditions 1 and 4 were required to hold in order to establish this result, i.e.,

$$(\text{a}) \forall h \in \mathbb{R}^n \quad \nabla^3 f(x)[h,h,h] \leq 2(h^T \nabla^2 f(x) h)^{\frac{3}{2}} \tag{1}$$

$$(\text{b}) \nabla^2 F(x) \succeq \frac{1}{\nu} \nabla F(x) \big(\nabla F(x)\big)^T \tag{4}$$

We call a barrier function that satisfies the two conditions above a *ν-self-concordant barrier.*

References

[1] Karmarkar, N. (1984). A New Polynomial Time Algorithm for Linear Programming, Combinatorica, Vol 4, nr. 4, p. 373 - 395

[2] Nesterov, Y. and Nemirovskii, A. (1995). Interior-Point Polynomial Algorithms in Convex Programming. Society for Industrial and Applied Mathematics.

[3] Bubeck, S. ORF523. Class Lecture, ”Interior Point Methods.” Princeton University, Princeton, New Jersey, Feb. 14, 2013 [Online]. Available: https://blogs.princeton.edu/imabandit/2013/02/14/orf523-interior-point-methods

[4] Vishnoi, N. Fundamentals of Convex Optimization. Class Lecture, ”Newton's Method and the Interior Point Method.” École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, Oct. 31, 2014 [Online]. Available: http://tcs.epfl.ch/files/content/sites/tcs/files/Lec3-Fall14-Web.pdf