**IE 521: Convex Optimization** **Spring 2017, UIUC**

# Lecture 18: Interior Point Method - Part II – April 3

*Instructor: Niao He* *Scribe: Shuanglong Wang*

*Courtesy warning: These notes do not necessarily cover everything discussed in the class. Please email TA (swang157@illinois.edu) if you find any typos or mistakes.*

In this lecture, we cover the following topics

- Classical Newton Method for Unconstrained Minimization
- (Damped) Newton Method for Self-concordant Functions

**Reference**:

Nesterov, Introductory Lectures on Convex Optimization, 2004, Chapter 1.2.4

## 18.1 Classical Newton Method

Consider the unconstrained minimization

$$\min_{x \in \mathbf{R}^n} f(x)$$

where $f(x)$ is twice continuously differentiable (not necessarily convex) on $\mathbf{R}^n$.

**Newton Method:**

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k), k = 0, 1, 2, ...$$

The direction $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$ is called Newton's direction.

**Interpretation:** The Newton method can be treated as

- *Minimizing quadratic approximation of $f$:* From Taylor's expression, we have:

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

The iteration can be viewed as minimizing the quadratic approximation of $f$:

$$x_{k+1} = \min_x \left\{ f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2} (x - x_k)^T \nabla^2 f(x_k)(x - x_k) \right\}$$

- *Solving linearized optimality condition:* From first-order optimality condition: $\nabla f(x) = 0$. Note that
$$\nabla f(x+h) \approx \nabla f(x) + \nabla^2 f(x)h$$
The iteration can also be viewed as solving linearized optimality condition
$$\nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) = 0$$

**Remark (Newton method vs Gradient Descent)**

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k) \quad \text{(Newton)}$$
$$x_{k+1} = x_k - \gamma_k \nabla f(x_k) \quad \text{(GD)}$$

Firstly, unlike Gradient Descent, Newton's direction is not necessarily a descent direction: for $\nabla f(x) \neq 0$,
$$\nabla f(x)^T d = -\nabla f(x)^T \nabla^2 f(x) \nabla f(x) \not< 0$$
Secondly, Newton method is a second-order method and requires relatively high computational cost comparing to GD.

**Remark (Convergence)**

(i) Newton method can break down if $\nabla^2 f(x)$ is degenerate.

(ii) When $f$ is quadratic and non-degenerate, Newton method converges in one step.

(iii) The method may diverge even for a nice strongly convex function.

(iv) If started close enough to a strict local minimum, the method can converge very fast.

**Example:** consider the convex function $f(x) = \sqrt{1+x^2}$

One can easily compute that $x^* = 0$, and
$$f'(x) = \frac{x}{\sqrt{1+x^2}}, \quad f''(x) = \frac{1}{(1+x^2)^{3/2}}$$
The Newton method works as
$$x_{k+1} = x_k - (1+x_k^2)^{3/2} \frac{x_k}{\sqrt{1+x_k^2}} = -x_k^3$$
Note that

- if $|x_0| < 1$, the method converges extremely fast
- if $|x_0| = 1$, the method oscillates between 1 and -1
- if $|x_0| > 1$, the method diverges.

## 18.2 Classical Analysis

**Theorem 18.1 (Local quadratic convergence of Newton method)** *Assume that*

- *$f$ has a Lipschitz Hessian: $\| \nabla^2 f(x) - \nabla^2 f(y) \|_2 \leq M \| x - y \|_2$ for $M > 0$.*
- *$f$ has a strict local minimum $x^*$: $\nabla^2 f(x^*) \succcurlyeq \mu I$, with $\mu > 0$.*
- *The initial point $x_0$ is close enough to $x^*$: $\| x_0 - x^* \|_2 \leq \frac{\mu}{2M}$*

*Then Newton method is well-defined and converges to $x^*$ at a quadratic rate*

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

Note that for a symmetric matrix $A$, $\| A \|_2 := \sup_{x \neq b} \left\{ \frac{\|Ax\|_2}{\|x\|_2} \right\} = \max_k |\lambda_k(A)|$. We first prove the following simple useful lemma.

**Lemma 18.2** *Suppose $f$ has Lipschitz Hession with constant $M$, then for any $x, y$,*

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\|_2 \leq \frac{M}{2} \|y - x\|_2^2.$$

*Proof:* This is because by basic calculus

$$\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(y - x))(y - x) dt$$

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\|_2 &= \| \int_0^1 \nabla^2 f(x + t(y - x))(y - x) dt - \nabla^2 f(x)(y - x) \|_2 \\
&= \| \int_0^1 \left[ \nabla^2 f(x + t(y - x)) - \nabla^2 f(x) \right] (y - x) dt \|_2 \\
&\leq \int_0^1 Mt \| y - x \|_2^2 \, dt \\
&= \frac{M}{2} \| y - x \|_2^2
\end{aligned}$$

■

Now we are ready to prove the main theorem.

*Proof:* We have

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

The last equality is due to the fact that $\nabla f(x^*) = 0$ by optimality condition. Hence,

$$\begin{aligned} \| x_{k+1} - x^* \|_2 &\leq \| [\nabla^2 f(x_k)]^{-1} \|_2 \| \nabla f(x^*) - \nabla f(x_k) - \nabla^2 f(x_k)(x^* - x_k) \|_2 \\ &\leq \| [\nabla^2 f(x_k)]^{-1} \|_2 \cdot \frac{M}{2} \| x_k - x^* \|_2^2 \end{aligned}$$

The last inequality is due to the previous lemma.
We show now by induction that $\| x_k - x^* \|_2 \leq \frac{\mu}{2M}$.

Assume $\| x_k - x^* \|_2 \leq \frac{\mu}{M}$, then

$$\| \nabla^2 f(x_k) - \nabla f(x^*) \|_2 \leq M \| x_k - x^* \|_2 \leq \frac{\mu}{2}$$

which implies that $-\frac{\mu}{2} I \preccurlyeq \nabla^2 f(x_k) - \nabla^2 f(x^*) \preccurlyeq \frac{\mu}{2} I$. Hence,

$$\nabla^2 f(x_k) \succcurlyeq \nabla^2 f(x^*) - \frac{\mu}{2} \succcurlyeq \frac{\mu}{2} I$$

which implies that $\| [\nabla^2 f(x_k)]^{-1} \|_2 \leq \frac{2}{\mu}$. This leads to

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

and $\| x_{k+1} - x^* \|_2 \leq \frac{\mu}{2M}$, which concludes the proof. ■

Note that the local convergence holds for any unconstrained minimization regardless of convex or not. When $f$ is strongly convex, the analysis is even simpler.

**Theorem 18.3** *Assume that*

- *$f$ has a Lipschitz Hessian* $\| \nabla^2 f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2$
- *$f$ is $\mu$-strongly convex, i,e,* $\nabla^2 f(x) \succcurlyeq \mu I, \forall x$
- *The initial point $x_0$ satisfies* $\| \nabla f(x_0) \|_2 < \frac{2\mu^2}{M}$

*Then the gradient converges to zero quadratically*

$$\| \nabla f(x_{k+1}) \|_2 \leq \frac{M}{2\mu^2} \| \nabla f(x_k) \|_2^2$$

*Proof:* Setting $y = x_{k+1}$, $x = x_k$ in the previous lemma, we have

$$\begin{aligned} \| \nabla f(x_{k+1}) \|_2 &\leq \frac{M}{2} \left\| \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k) \right\|_2^2 \\ &\leq \frac{M}{2} \left\| \left[\nabla^2 f(x_k)\right]^{-1} \right\|_2^2 \cdot \| \nabla f(x_k) \|_2^2 \\ &\leq \frac{M}{2\mu^2} \| \nabla f(x_k) \|_2^2 \end{aligned}$$

■

**Remark (Complexity)** The quadratic convergence implies that:

$$\begin{aligned}\frac{M}{\mu} \parallel x_k - x^* \parallel_2 &\leq \Big[\frac{M}{\mu} \parallel x_{k-1} - x^* \parallel_2 \Big]^2 \\ &\leq \Big[\frac{M}{\mu} \parallel x_{k-2} - x^* \parallel_2 \Big]^4 \\ &\leq ... \\ &\leq \Big[\frac{M}{\mu} \parallel x_0 - x^* \parallel_2 \Big]^{2^k} \\ &\leq (\frac{1}{2})^{2^k}\end{aligned}$$

Hence, $\parallel x_k - x^* \parallel_2 < \frac{\mu}{M} 2^{-2^k}$

To achieve an accuracy $\epsilon$, i.e. $\parallel x_k - x^* \parallel_2 \leq \epsilon$, the number of iterations

$$k \geq \log_2 \log_2 (\frac{M}{\mu\epsilon})$$

**Remark (Affine Invariance)** The Newton method is invariant w.r.t. affine transformation of variables.

Let $A$ be a non-singular matrix consider the function

$$\hat{f}(y) = f(Ay)$$

The Newton step for $f$ and $\hat{f}$ are

$$\begin{aligned}x_{k+1} &= x_k - \Big[\nabla^2 f(x_k)\Big]^{-1} \nabla f(x_k) \\ y_{k+1} &= y_k - \Big[\nabla^2 \hat{f}(y_k)\Big]^{-1} \nabla \hat{f}(y_k)\end{aligned}$$

Let $y_0 = A^{-1} x_0$, then $y_k = A^{-1} x_k$. This can be shown by induction

$$\begin{aligned}y_{k+1} &= y_k - \Big[A^T \nabla^2 f(Ay_k) A\Big]^{-1} \Big[A^T \nabla f(Ay_k)\Big] \\ &= A^{-1} x_k - \Big[A^t \nabla^2 f(x_k) A\Big]^{-1} A^T \nabla f(x_k) \\ &= A^{-1} x_k - A^{-1} \nabla^2 f(x_k)^{-1} \nabla f(x_k) \\ &= A^{-1} x_{k+1}\end{aligned}$$

In other words, Newton's method follow the same trajectory in the 'x-space' and 'y-space'. Hence, the region of quadratic convergence should not depend on the Euclidean metric.

However, in the classical analysis, the assumption and the measure of error, e.g. the Lipschitz continuity of Hessian, depend heavily on the Euclidean metric and is not affine invariant. A natural remedy is to assume self-concordance. Self concordant function are especially well suited for Newton method.