IE 521 Convex Optimization

Lecture 17: Interior Point Method

Newton's Method

Niao He

16th April 2019

Outline

Path Following Scheme

Newton's Method

- Classical Newton's Method
- Affine Invariance
- Newton vs GD
- Classical Analysis
- Newton's Decrement
- Newton's Method for Self-Concordant Functions
- Damped Newton Method

Path Following Scheme

$$\min_x \quad f(x)$$
$$\text{s.t.} \quad x \in X := \{x : g_i(x) \le 0, i = 1, \ldots, m\} \qquad \text{(P)}$$

Barrier Method: solve a series of unconstrained problems

$$x^*(t) := \operatorname*{argmin}_x \{t \cdot f(x) + F(x)\} \quad (t > 0) \qquad (P_t)$$

Barrier Function:

- $F : \text{int}(X) \to \mathbb{R}$ and $F(x) \to +\infty$ as $x \to \partial(X)$
- F is twice continuously differentiable and convex
- F is *non-degenerate*, i.e. $\nabla^2 F(x) \succ 0, \forall x \in \text{int}(X)$

Central Path:

$$x^*(t) \in \text{int}(X) \longrightarrow x^*, \text{as } t \longrightarrow \infty$$

Path Following Scheme

Newton's Method

Classical Newton's Method

Affine Invariance

Newton vs GD

Classical Analysis

Newton's Decrement

Newton's Method for Self-Concordant Functions

Damped Newton Method

Path Following Scheme

Question: Need to specify

1. the barrier function $F(x)$?
   - Self-concordant barriers, e.g..

$$F(x) = -\sum_{i=1}^{m} \log(-g_i(x))$$

2. the method to solve unconstrained problems (P_t)?
   - Newton's method

3. the policy to update the penalty parameter t?

Classical Newton's Method

Assume $f(x)$ is twice continuously differentiable on $\mathbb{R}^n$.

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

Newton's Method:

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

- Newton's method can break down if $f(x)$ is degenerate.
- $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$ is called Newton's's direction.
- Newton's direction is not necessarily a descent direction.

Classical Newton's Method: Interpretation

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

- $(\star) \iff$ *minimizing quadratic approximation of f*
  - Recall Taylor expansion of $f(x)$:

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

  - $(\star)$ is the solution to the quadratic approximation

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

Remark. When f is quadratic and non-degenerate, Newton's method converges in one step.

Classical Newton's Method: Interpretation

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

- $(\star) \iff$ *solving linearized optimality condition*
  - From first-order optimality condition: $\nabla f(x) = 0$
  - Taylor expansion:

$$\nabla f(x+h) \approx \nabla f(x) + \nabla^2 f(x) h$$

  - $(\star)$ is the solution to the linear system:

$$\nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) = 0$$

Affine Invariance of Newton's Method

Newton's method is invariant w.r.t. affine transformation.

- Let A be non-singular and consider the function

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

- The Newton steps for f and $\hat{f}$ are

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

$$\begin{aligned} y_{k+1} &= y_k - \left[\nabla^2 \hat{f}(y_k)\right]^{-1} \nabla \hat{f}(y_k) \\ &= y_k - A^{-1}\left[\nabla^2 f(Ay_k)\right]^{-1} \nabla f(Ay_k) \end{aligned}$$

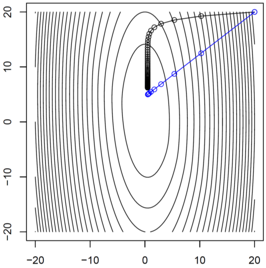
- If $y_0 = A^{-1}x_0$, then $y_k = A^{-1}x_k$.
- Newton's method follows the same trajectory in the 'x-space' and 'y-space'.

Newton's Method vs Gradient Descent

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

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k) \qquad \text{(GD)}$$

- Affine vs non-affine invariant
- Second-order vs. first-order
- Expensive vs. cheap iteration
- Local vs. global convergence



Newton vs. GD

Figure from Tibshirani lecture notes

Illustration: Convergence of Newton's Method

Consider the nonconvex function:

$$f(x) = x^3$$

Q. Does Newton's method converge? How fast?

$$f'(x) = 3x^2, \quad f''(x) = 6x$$

$$x_{k+1} = x_k - (6x_k)^{-1} \cdot 3x_k^2 = \frac{1}{2} \cdot x_k \qquad \text{(Newton)}$$

- Converges in a linear rate to a stationary point

Illustration: Convergence of Newton's Method

Consider the strictly convex function:

$$f(x) = \sqrt{1+x^2}$$

Q. Does Newton's method converge? How fast?

$$f'(x) = \frac{x}{\sqrt{1+x^2}}, \quad f''(x) = \frac{1}{(1+x^2)^{3/2}}$$

$$x_{k+1} = x_k - (1+x_k^2)^{3/2}\frac{x_k}{\sqrt{1+x_k^2}} = -x_k^3 \qquad \text{(Newton)}$$

- if $|x_0| < 1$, converges in a cubic rate (extremely fast)
- if $|x_0| = 1$, oscillates between 1 and -1
- if $|x_0| > 1$, diverges

Local Quadratic Convergence

Theorem. Assume that

- f has a Lipschitz Hessian: for some $M > 0$,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq M\|x - y\|_2$$

- f has a strict local minimum x^*: for some $\mu > 0$,

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

- The initial point x_0 is close enough to x^*:

$$\|x_0 - x^*\|_2 \leq \frac{\mu}{2M}$$

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

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

Local Quadratic Convergence

Corollary. It follows that

$$
\begin{aligned}
\frac{M}{\mu}\|x_k - x^*\|_2 &\leq \left[\frac{M}{\mu}\|x_{k-1} - x^*\|_2\right]^2 \\
&\leq \ldots \\
&\leq \left[\frac{M}{\mu}\|x_0 - x^*\|_2\right]^{2^k} \\
&\leq (\frac{1}{2})^{2^k}
\end{aligned}
$$

- The number of iterations to achieve an accuracy ϵ, i.e. $\|x_k - x^*\|_2 \leq \epsilon$, is at most

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

- The above results hold true for any unconstrained minimization regardless of convexity.

Lemma on Hessian Lipschitzness

Lemma. Assume f has a Lipschitz Hessian 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.

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

Proof of Local Convergence

First, we have

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

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

We can show by induction that

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

This concludes the proof.

Local Convergence for Strongly Convex Functions

Theorem. Assume that

- f has a Lipschitz Hessian with constant $M > 0$;
- f is μ-strongly convex: $\nabla^2 f(x) \succeq \mu I, \forall x$;
- The initial point x_0 satisfies $\|\nabla f(x_0)\|_2 \leq \frac{2\mu^2}{M}$.

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. This is because

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

Issue with Affine Invariance

- Recall that Newton's method is invariant w.r.t. affine transformations.
- 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.

Newton's Decrement

Definition. *Newton decrement* is defined as :

$$\lambda_f(x) = \sqrt{\nabla f(x)[\nabla^2 f(x)]^{-1} \nabla f(x)}$$

- Relates to the decrease of the second order Taylor expansion after a Newton step:

$$f(x) - \min_h \left\{ f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) h \right\} = \frac{1}{2} \lambda_f^2(x)$$

- Can be viewed as an approximate bound of the suboptimality gap $f(x) - f^*$.
- Newton decrement is also *affine-invariant*.

Newton's Decrement vs Local Norms

$$\lambda_f(x) = \sqrt{\nabla f(x)[\nabla^2 f(x)]^{-1}\nabla f(x)}$$

- Equals to the local norm of Newton's direction $d(x)$:

$$\|d(x)\|_x = \|-\nabla^2 f(x)^{-1}\nabla f(x)\|_x = \lambda_f(x)$$

- Equals to the conjugate local norm of $\nabla f(x)$:

$$\|\nabla f(x)\|_{x,*} = \|[\nabla^2 f(x)]^{-1/2}\nabla f(x)\|_2 = \lambda_f(x)$$

Newton's Decrement and Self-concordance

Recall that standard self-concordant functions f has nice properties inside the Dikin ellipsoid: $\forall y : \|y - x\|_x = \gamma < 1$,

(1) $y \in \mathrm{dom}(f)$

(2) $(1-r)^2 \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \frac{1}{(1-r)^2} \nabla^2 f(x)$

(3) $\frac{\gamma^2}{1+\gamma} \leq \langle \nabla f(y) - \nabla f(x), y - x \rangle \leq \frac{\gamma^2}{1-\gamma}$

(4) $\omega(\gamma) \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \omega_*(\gamma)$, where $\omega(\gamma) = \gamma - \ln(1+\gamma)$, $\omega_*(\gamma) = -\gamma - \ln(1-\gamma)$.

Proposition.

- If $\lambda_f(x) < 1$, the point $x_+ = x - d(x) \in dom(f)$
- If x^* is a minimizer of f, then $\lambda_f(x^*) = 0$
- If $\lambda_f(x_0) < 1$ for some $x_0 \in dom(f)$, then f has a unique minimizer.

Newton's Decrement and Self-concordance

Proposition. If $\lambda_f(x_0) < 1$ for some $x_0 \in dom(f)$, then f has a unique minimizer.

Proof.
Suffice to show the level set $\{y : f(y) \le f(x_0)\}$ is bounded.

$$\begin{aligned} f(y) &\ge f(x_0) + \langle \nabla f(x_0), y - x_0 \rangle + \omega(\|y - x_0\|_{x_0}) \\ &\ge f(x_0) - \|\nabla f(x_0)\|_{x_0,*} \cdot \|y - x_0\|_{x_0} + \omega(\|y - x_0\|_{x_0}) \\ &= f(x_0) - \lambda_f(x_0) \cdot \|y - x_0\|_{x_0} + \omega(\|y - x_0\|_{x_0}) \end{aligned}$$

$$\text{Hence, } f(y) \le f(x_0) \Longrightarrow \frac{\omega(\|y - x_0\|_{x_0})}{\|y - x_0\|_{x_0}} \le \lambda_f(x_0) < 1$$

Note the function $\phi(t) = \frac{\omega(t)}{t} = 1 - \frac{1}{t}\ln(1+t)$ is strictly increasing in $t \ge 0$. Hence, $\|y - x_0\|_{x_0} \le t^*$ for some t^*.

Example: Newton's Decrement

Consider the self-concordant function

$$f(x) = \epsilon x - \ln(x)$$

with $\text{dom}(f) := \{x : x > 0\}$.

$$\lambda_f(x) = \sqrt{\left(\epsilon - \frac{1}{x}\right)\left(\frac{1}{x^2}\right)^{-1}\left(\epsilon - \frac{1}{x}\right)} = |1 - \epsilon x|$$

- When $\epsilon \leq 0$, $\lambda_f(x) \geq 1$, and the function is unbounded below and there does not exist a minimizer.
- When $\epsilon > 0$, $\lambda_f(x) < 1$, for $x \in (0, \frac{2}{\epsilon})$, there exists a unique minimizer $x^* = \frac{1}{\epsilon}$.

Affine-invariant Metrics

Newton method: initialize $x_0 \in \mathrm{dom}(f)$ and update via

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

Accuracy Measure:

- Function gap: $f(x_k) - f(x^*)$
- Newton's decrement: $\lambda_f(x_k) = \|\nabla f(x_k)\|_{x_k,*}$
- Local distance to the minimizer: $\|x_k - x^*\|_{x_k}$
- Distance to the minimizer: $\|x_k - x^*\|_{x^*}$

Remark. Indeed, all of these measures are independent of Euclidean metric and equivalent locally.

Affine-invariant Metrics

Proposition. (Nesterov, 2004) When $\lambda_f(x) < 1$, we have

1. $f(x) - f(x^*) \leq \omega_*(\lambda_f(x)) \leq \frac{\lambda_f(x)^2}{2(1-\lambda_f(x))^2}$
2. $\|x - x^*\|_x \leq \frac{\lambda_f(x)}{1-\lambda_f(x)}$
3. $\|x - x^*\|_{x^*} \leq \frac{\lambda_f(x)}{1-\lambda_f(x)}$

We will focus mainly on the convergence in terms of $\lambda_f(x)$.

Local Convergence of Self-concordant Functions

Theorem. If $x_k \in \mathrm{dom}(f)$ and $\lambda_k < 1$, then $x_{k+1} \in \mathrm{dom}(f)$ and

$$\lambda_{k+1} \leq \left(\frac{\lambda_k}{1-\lambda_k}\right)^2.$$

Remark. Let λ^* be such that $\frac{\lambda^*}{(1-\lambda^*)^2} = 1$.

- If $\lambda_k < \lambda^*$, $\lambda_{k+1} < \lambda_k$.
- Region of quadratic convergence is

$$\lambda_f(x) \leq \lambda^* = \frac{3-\sqrt{5}}{2} \approx 0.38.$$

- Still might diverge if not started with a point with $\lambda_f(x)$ small enough.

Q. How to ensure global convergence?

Proof of Local Convergence

- Note $\|x_{k+1} - x_k\|_{x_k} = \lambda_f(x_k) = \lambda_k < 1$, so $x_{k+1} \in \mathrm{dom}(f)$.
- It holds that

$$\nabla^2 f(x_{k+1}) \succeq (1-\lambda_k)^2 \nabla^2 f(x_k)$$

$$\lambda_{k+1} \leq \frac{1}{1-\lambda_k} \sqrt{\nabla f(x_{k+1})^T \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_{k+1})}$$

- Note

$$\nabla f(x_{k+1}) = \nabla f(x_{k+1}) - \nabla f(x_k) - [\nabla^2 f(x_k)](x_{k+1} - x_k)$$
$$= \underbrace{\left[\int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x_k) dt\right]}_{G}(x_{k+1} - x_k)$$

- Hence,

$$\lambda_{k+1} \leq \frac{1}{1-\lambda_k} \sqrt{(x_{k+1} - x_k)^T G^T [\nabla^2 f(x_k)]^{-1} G (x_{k+1} - x_k)}$$

Proof of Local Convergence (Cont'd)

- Further,

$$\lambda_{k+1} \leq \frac{\lambda_k}{1-\lambda_k} \| \underbrace{[\nabla^2 f(x_k)]^{-1/2} G [\nabla^2 f(x_k)]^{-1/2}}_{H} \|_2$$

- Note that

$$G \succeq \nabla^2 f(x_k) \int_0^1 \left[(1 - t\lambda_k)^2 - 1 \right] dt = \left(\frac{\lambda_k^2}{3} - \lambda_k \right) \nabla^2 f(x_k)$$

$$G \preceq \nabla^2 f(x_k) \int_0^1 \left[\frac{1}{(1 - t\lambda_k)^2} - 1 \right] dt = \frac{\lambda_k}{1-\lambda_k} \nabla^2 f(x_k)$$

- This implies that

$$\|H\|_2 \leq \max \left\{ \lambda_k - \frac{\lambda_k^2}{3}, \frac{\lambda_k}{1-\lambda_k} \right\} = \frac{\lambda_k}{1-\lambda_k}.$$

Damped Newton Method

Damped Newton method: initialize $x_0 \in \mathrm{dom}(f)$ and update via

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

Remark. Damped Newton procedure is always well-defined:

$$\|x_{k+1} - x_k\|_{x_k} = \frac{\lambda_f(x_k)}{1+\lambda_f(x_k)} < 1 \Rightarrow x_{k+1} \in W_1^0(x_k) \subseteq \mathrm{dom}(f).$$

Global Convergence of Damped Newton

Theorem. The damped Newton method satisfies that

1. (Descent phase) $\forall k \geq 0$,
$$f(x_{k+1}) \leq f(x_k) - \omega(\lambda_f(x_k)).$$
2. (Quadratic convergence phase) If $\lambda_k(x_k) < \frac{1}{4}$, then
$$\lambda_f(x_{k+1}) \leq 2[\lambda_f(x_k)]^2.$$

Proof.

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \omega_*(\|x_{k+1} - x_k\|_{x_k}) \\ &= f(x_k) - \frac{\lambda_f(x_k)}{1 + \lambda_f(x_k)} + \omega_*\Big(\frac{\lambda_f(x_k)}{1 + \lambda_f(x_k)}\Big) \end{aligned}$$

where $\omega_*(t) = -t - \ln(1 - t)$.

Iteration Complexity of Damped Newton Method

Remark.

- *Damped Newton stage:* when $\lambda_f(x_k) \geq \beta \in (0, 1/4)$

$$f(x_{k+1}) \leq f(x_k) - \omega(\beta) \Rightarrow N_1 \leq \frac{f(x_0) - f(x^*)}{\omega(\beta)}.$$

- *Damped/Basic Newton stage:* when $\lambda_f(x_k) < \beta$

$$\lambda_f(x_{k+1}) \leq 2\Big[\lambda_f(x_k)\Big]^2 \Rightarrow N_2 \leq O(1)\log_2\log_2(\frac{1}{\epsilon}).$$

The total complexity to find a solution with $\lambda_f(x) \leq \epsilon$:

$$O(1)[f(x_0) - f^* + \log\log(\frac{1}{\epsilon})]$$

References

- Nesterov (2004), Introductory Lectures on Convex Optimization, Chapter 4.1.4-5
- Nemirovski (2004), Interior Point Polynomial Time Methods in Convex Programming, Chapter 1