IIIS 2014 Spring: ATCS - Selected Topics in Optimization
Lecture date: Apr. 23rd, 2014
Instructor: Jian Li **Scribe: Yulong Zeng**

This is a draft.

1 Review of Last Lecture

1.1 Steepest descent method

The (unnormalized) gradient of Steepest descent method is defined as

$$\Delta_{sd} = \|\nabla f\|_* \Delta_{nsd}$$

where the normalized gradient

$$\Delta_{nsd} = \arg\min_v \{\nabla f(x)v | \|v\| \leq 1\}$$

and the $\|\cdot\|_*$ denote the dual norm.

For a positive matrix A, we define matrix norm w.r.t. A as

$$\|x\|_A^2 = x^{\mathrm{T}} A x$$

which can be used in the dual norm.

2 Newton's Method

2.1 definition

The gradient of Newton's Method is defined as

$$\Delta x_{nt} = \left(\nabla^2 f(x)\right)^{-1} \cdot \nabla f(x)$$

which actually is a special case of steepest descent if we use matrix norm w.r.t. $\nabla^2 f(x)$, i.e., the Hessian matrix of f at this point.

Notice that Newton's method sometimes may not converge[1].

The intuition of Newton's Method is to approximate the function f by a quadratic function $\hat{f}$, then compute the minimize of $\hat{f}$ to get the new point.

Example 1 (Application on Quadratic Function) *Suppose f is a quadratic function ,*

$$\hat{f}(x+v) = f(x) + \nabla f(x)v + \frac{1}{2} v^{\mathrm{T}} \nabla^2 f(x) v$$

Then

$$\nabla \hat{f}(x+v) = 0 \Longrightarrow \nabla f(x) + \nabla^2 f(x) v = 0,\ v = -\left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$$

Which means that we only need 1 iteration to get optimal.

Remark 2 *Newton's method is a descent method, since the inner product* $< \Delta x_{nt}, \nabla f(x) >= -\nabla f^T(x) \nabla^2 f(x) \nabla f(x)$ *is guaranteed to be negative.*

[1] See `http://en.wikipedia.org/wiki/Newton's_method#Failure_analysis`.

2.2 Affine Inveriance

Consider $f : \mathbb{R}^n \to \mathbb{R}$, and an invertible linear transformation $T \in \mathbb{R}^{n\times n}$.

$$\overline{f}(y) = f(Ty)$$

Where T is a $n \times n$ intertable matrix.

It is easy to verify the following

$$\nabla\overline{f}(y) = T^{\mathrm{T}}\nabla f(x),\ \nabla^2\overline{f}(x) = T^{\mathrm{T}}\nabla^2 f(x)T,$$

and the step size of $\overline{f}$ at point x is

$$\Delta y_{nt} = -\big(T^{\mathrm{T}}\nabla^2 f(x)T\big)^{-1}\cdot\big(T\nabla f(x)\big) = T^{-1}\Delta x_{nt}$$

where $x = Ty$.

Remark 3 *Since computing the term $(\nabla^2 f(x))^{-1}$ is usually a difficult job, an improvement of Newton's Method called Quasi-Newton is proposed.*

3 Convergence Analysis

3.1 Convergence of Newton's method

Based on the following assumptions,

$$mI \preceq \nabla^2 f \preceq MI,\ \|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$$

the implement of Newton's method consists of the following two phases.

- **Phase 1.** $\|\nabla f\| > \eta$, then $f(x^{k+1}) - f(x^k) \le -\gamma$, where $\gamma = 2m^2/L^2$ and $0 < \eta < m^2/L$.
- **Phase 2.** Otherwise,
$$\frac{L}{2m^2}\big\|\nabla f(x_{k+1})\big\| \le \Big(\frac{L}{2m^2}\big\|\nabla f(x_k)\big\|\Big)^2$$

Recall that in last class we mentioned that for strongly convex function f,$f(x)-p^* \le \frac{1}{2m}\|\nabla f(x)\|^2$, which means that $f(x)$ is close to optimal when its gradient is small.

Then the number of iterations is

$$\frac{f(x_0 - p^*)}{\gamma} + \log\log\Big(\frac{\epsilon_0}{\epsilon}\Big)$$

Self-concordance is

$$\frac{f(x_0 - p^*)}{\square} + \log\log\Big(\frac{1}{\epsilon}\Big)$$

Where $\square$ is something irrelevant with γ

Note that the convergency result was for Newton's direction combined with backtrack line search(See notes in last class),not for the plain Newton step written in the board of this class.

Remark 4 *In real application, since phase 2 performs large better than phase 1, one can first use other gradient descent method, and then use Newton's method when $\|\nabla f\|$ is sufficiently small.*

3.2 Convergence of Gradient Descent (Constant Step Size)

Assumption 5 (Lipschitz gradient)

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Suppose that $\epsilon \leq t_k \leq (2-\epsilon)/L$, where L is the coefficient in the Lipschitz assumption, then we have

Lemma 6 (Descent Lemma)

$$f(x_{k+1}) - f(x_k) \leq \nabla f(x_k) t_k\big(-\nabla f(x_k)\big) + \frac{L}{2}\big\|t_k \nabla f(x_k)\big\|^2$$

Proposition 7 *The following four conditions are equivalent.* [2]

1. *Lipschitz gradient.*
2. *Descent Lemma*
3. *Coercivity.*
$$\big(\nabla f(x) - \nabla f(y)\big)^{\mathrm{T}}(x-y) \leq M\|x-y\|^2$$
4. $$\|\nabla^2 f\| \leq L$$

Then we prove the convergence of $\{x_k\}$, with the number of iterations in $O(...)$.

Proof: Using the lemma above,

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \nabla f(x_k) t_k\big(-\nabla f(x_k)\big) + \frac{L}{2}\big\|t_k \nabla f(x_k)\big\|^2 \\ &= \big\|\nabla f(x_k)\big\|^2(-t_k + \frac{L}{2}t_k^2) \\ &\leq \big\|\nabla f(x_k)\big\|^2(-t_k + t_k \cdot \frac{2-\epsilon}{2}) \\ &= -\frac{\epsilon}{2}\cdot t_k\big\|\nabla f(x_k)\big\|^2 \\ &\leq -\frac{\epsilon^2}{2}\big\|\nabla f(x_k)\big\|^2 \end{aligned}$$

By convexity of f,

$$\begin{aligned} f(x_k) - f^* &\leq \langle \nabla f(x_k), x_k - x^*\rangle \\ &\leq \|\nabla f(x_k)\| \cdot \|x_k - x^*\| \\ &\leq \|\nabla f(x_k)\| \cdot R \end{aligned}$$

where R is the radius defined as $R = \max_{f(x)\leq f(x_0)} \|x - x^*\|$.

[2]The lemma is recommended to prove in the order $1 \Longrightarrow 2 \Longrightarrow 3 \Longrightarrow 4 \Longrightarrow 1$, left as homework.

Let $\phi_k = f(x_k) - f^*$, then

$$\phi_k - \phi_{k+1} \geq \frac{\epsilon^2}{2}\left\|\nabla f(x_k)\right\|^2 \geq \frac{\epsilon^2}{2} \cdot \frac{\phi_k^2}{R^2}$$

Therefore

$$\frac{1}{\phi_{k+1}} - \frac{1}{\phi_k} = \frac{\phi_k - \phi k + 1}{\phi_k \phi_{k+1}} \geq \frac{\phi_k - \phi k + 1}{\phi_k^2} \geq \frac{\epsilon^2}{2R^2} \Longrightarrow \frac{1}{\phi_k} \geq \frac{1}{\phi_0} + \frac{k\epsilon^2}{2R^2}$$

where $\phi_0 = f(x_0) - f^* \leq LR$.

So, we get that k is the order in $O(\frac{1}{\delta})$ to satisfy $\phi_k \leq \delta$. □

3.3 Lower Bounds for First Order Methods

Assume that

- f is strongly convex with parameter l, i.e. $\nabla^2 f \succeq lI$.
- The initial point $x_0 = 0$.
- (First order)
$$x_k \in \text{Span}\{x_1, \ldots, x_{k-1}, \nabla f(x_1), \ldots, \nabla f(x_{k-1})\}$$

Then we show that the convergence rate of $f(x) = x^{\mathrm{T}} A x - b^{\mathrm{T}}$ is in $\Omega\big(\sqrt{\kappa}\log\frac{1}{\delta}\big)$, where

$$A = \alpha A_0 + \beta I = \frac{L-l}{4}\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix} + lI,\ b = e_1$$

and dimension $d \to \infty$.

Claim 8 *Suppose $Ax^* = b$ and $x^* = (u_1, \ldots, u_k, \ldots, u_d)$, then*

$$u_i = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i$$

Key observations,

- $x_0 = 0$, and $\nabla f(x_0) = b$.
- $x_1 \in \text{Span}\{b\}$.
- $x_2 \in \text{Span}\{Ab, b\}$.
- $x_3 \in \text{Span}\{A^2 b, Ab, b\}$.

- $\vdots$
- $x_k \in \text{Span}\{A^{k-1}b, \ldots, Ab, b\}$. (Krylov subspace)

Also notice that x_k must be in the form that only the first k terms are non-zero.
Therefore,

$$\|x^* - x_k\|^2 \geq \sum_{j=k+1}^{d} u^{2j} \overset{d\to\infty}{=} u^{2(k+1)}\|x^* - x_0\|^2$$

where $u = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$. Hence

$$\|f(x_k) - f(x^*)\| \geq \frac{l}{2}\|x_k - x^*\|^2 \geq \frac{l}{2}u^{2(k+1)}\|x_0 - x^*\|^2$$

which implies that $k = \Omega\big(\sqrt{\kappa}\log\frac{1}{\delta}\big)$ to satisfy $\|f(x_k) - f(x^*)\| < \delta$.

4 Conjugate Gradient

Goal: minimize

$$x^{\mathrm{T}}Ax - b^{\mathrm{T}}x$$

given that $A \succ 0$.

Suppose that $\mathcal{K} = \text{Span}\{b, Ab, \ldots, A^k b\}$, and $\{v_0, \ldots, v_k\}$ is a basis of $\mathcal{K}$.
Then the question is equivalent to

Finding the best vector $\overline{\alpha}$ in $\mathcal{K}$ that minimizes

$$\Big\|x^* - \sum_i \alpha_i v_i\Big\|_A$$

Observation.
If v_i and v_j are A-orthogonal to each other, i.e., $v_i^{\mathrm{T}}Av_j = 0,\ \forall i \neq j$, then

$$\Big\|x^* - \sum_i \alpha_i v_i\Big\|_A^2 = \sum_i (\alpha_i^2 v_i^{\mathrm{T}}Av_i - 2\alpha_i b^{\mathrm{T}}v_i) + \|x^*\|_A^2$$

(all the cross terms $v_iAv_j (i \neq j)$ disappear)
which implies that $\alpha_i = b^{\mathrm{T}}v_i/\|v_i\|_A^2$.

KEY IDEA, we can construct the basis $\{v_0, \ldots, v_k\}$ which are A-orthogonal to each other efficiently step by step, i.e.,

$$\text{Span}\{v_0, \ldots, v_i\} = \text{Span}\{b, Ab, \ldots, A^i b\} \overset{denote}{=} \mathcal{K}_i$$

Claim 9 *Av_{i-1} is A-orthogonal to $v_0, \ldots, v_{i-3}$.*

One-line proof: for $j \in [i-3]$, $Av_j \in \mathcal{K}_{j+1}$, which is A-orthogonal to v_{i-1}. Hence $v_{i-1}^T AAv_j = 0 = (Av_{i-1})^T Av_j$, which prove the claim. □

Therefore in each iteration, we only need to A-orthogonalize Av_{i-1}, v_{i-1} and v_{i-2} to obtain v_i. Since by construction v_i can write as a linear combination of $Av_{i-1}, v_{i-1}, v_{i-2}$, and by Claim 9 and induction $Av_{i-1}, v_{i-1}, v_{i-2}$ are A-orthogonal to $v_0, \ldots, v_{i-3}$, we have v_i is A-orthogonal to $v_0, \ldots, v_{i-3}$, which proves the Key idea.

References

[1] Cover T M. On determining the irrationality of the mean of a random variable. Ann. Statist, 1973, 1(5): 862-871.