Indian Institute of Technology Kanpur
CS774 Optimization Techniques

Scribe: Akshat Agarwal
Instructor: Purushottam Kar
Date: September 26, 2016

LECTURE
14

Conjugate Gradient and Newton Method

1 Introduction

In the previous lecture, the conjugate gradient (CG) algorithm was introduced, along with the concept of Krylov subspaces and their application to analysis of convergence of the CG method. It was proven that the CG method will converge in d iterations, where d is the dimensionality of the objective function.

Following from the previous lecture, we will first look at obtaining a lower bound on the performance of CG for quadratic optimization, by invoking Chebyshev polynomials.

Then we will look at various formulations of the Newton method (NT) for different problem settings and define the second order approximation whose optimal solution comes from the Newton method update. We will also define linear, super-linear, quadratic and sub-linear rates of convergence and show that under certain constraints, the gradient of objective function in NT has a quadratic rate of convergence.

2 Conjugate Gradient method

The conjugate gradient method is one of the most prominent and fastest methods available for solving sparse systems of linear equations or linear regression problems. These are both quadratic optimization problems, and can be framed in the manner defined below.

Definition 14.1. Objective function for quadratic optimization
Let $\mathbf{x} \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$. Then, we have

$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + b^T\mathbf{x} \tag{1}$$

where $A \succ 0$

We have from the previous lecture,

$$\frac{1}{2}\left\|\mathbf{x}^t - \mathbf{x}^*\right\|_A^2 \leq \frac{1}{2}\left\|\mathbf{x}^*\right\|_A^2 \min_{\substack{deg(q)\leq t \\ |q(0)=1|}} \max_q q(\lambda_i)^2 \tag{2}$$

where $\mathbf{x}^*$ is the optimum solution, $\mathbf{x}^t$ is the value of iterate after t iterations, and λ_i's are the eigenvalues of A

Remark 14.1. A useful approach is to minimize Eqn. 2 over the range $[\lambda_{min}, \lambda_{max}]$ rather than at a finite number of points, because

$$\max_i q^2(\lambda_i) \leq \max_{\lambda \in [\lambda_{min}, \lambda_{max}]} q^2(\lambda) \tag{3}$$

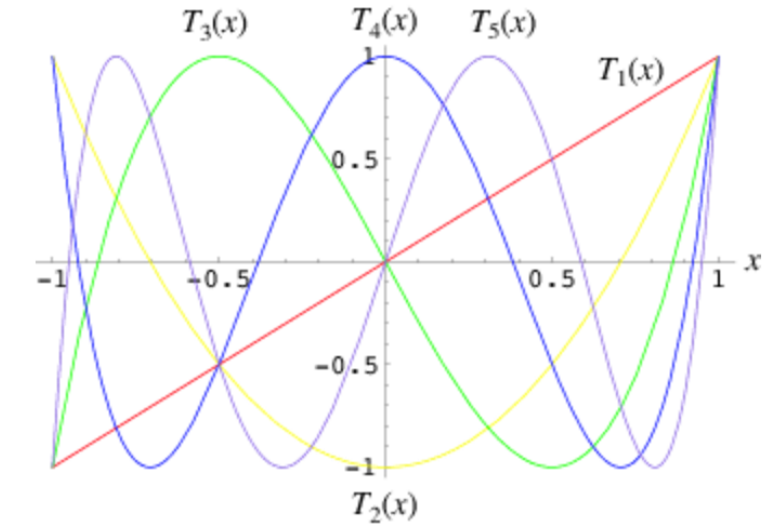

Figure 1: Chebyshev Polynomials for $k = 1, 2, 3, 4, 5$. Image Source: Wolfram Mathworld [1]

Definition 14.2. Chebyshev Polynomials
The Chebyshev polynomial of degree k is

$$T_k(x) = \frac{1}{2}\left[\left(\sqrt{x^2-1}+x\right)^K + \left(x-\sqrt{x^2-1}\right)^K\right] \tag{4}$$

The Chebyshev polynomials have the property that $|T_k(x)| \leq 1$ on the domain $x \in [-1, 1]$, and that $|T_k(x)|$ increases as quickly as possible outside that domain. The first few polynomials have been shown in Fig. 1.

Theorem 14.1. For $f(\mathbf{x})$ as defined in Defn. 14.1, if the CG method is used, we have convergence as $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t$ where κ is the condition number.

Proof. The Chebyshev polynomial family is used to accomplish the minimization of Eqn. 2 over the range $[\lambda_{min}, \lambda_{max}]$, by choosing suitable $q(\lambda)$. It can be showed (Shewchuk (1994), p.55) that the following choice for $q(\lambda)$ minimizes Eqn. 2:

$$q_t(\lambda) = \frac{T_t\left(\frac{\lambda_{max}+\lambda_{min}-2\lambda}{\lambda_{max}-\lambda_{min}}\right)}{T_t\left(\frac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)} \tag{5}$$

Let the argument of T_t in the numerator be $\mathcal{N}(\lambda)$. Note that the denominator is a constant term which does not depend on λ. Then

$$\max_{\lambda} \mathcal{N}(\lambda) = 1 \; at \; \lambda = \lambda_{min}$$
$$\min_{\lambda} \mathcal{N}(\lambda) = -1 \; at \; \lambda = \lambda_{max}$$

[1] `http://mathworld.wolfram.com/ChebyshevPolynomialoftheFirstKind.html`

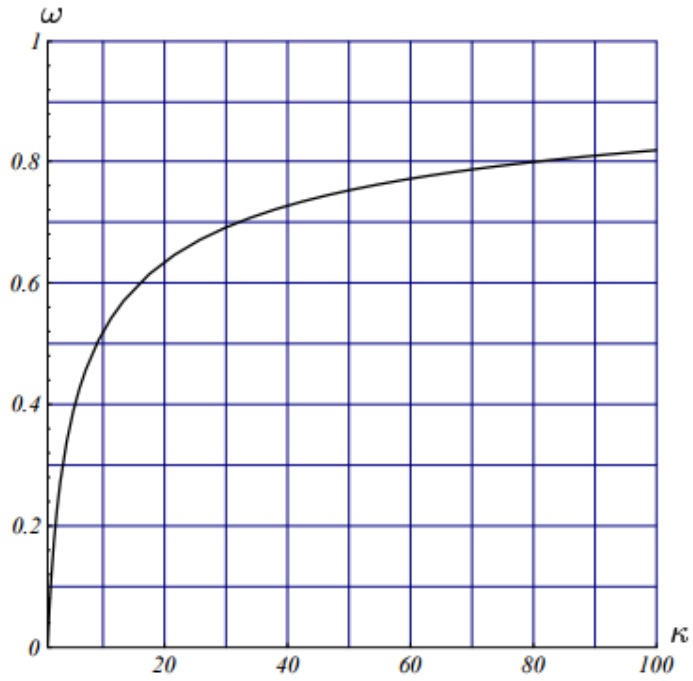

Figure 2: Convergence of Conjugate Gradients (per iteration) as a function of condition number
Source: Shewchuk (1994)

Therefore, $\mathcal{N}(\lambda) \in [-1, 1] \Rightarrow |T_t(\mathcal{N}(\lambda))| \leq 1$ and hence

$$\begin{aligned}(q_t(\lambda))^2 &\leq \left(T_t^{-1}\left(\frac{\lambda_{max}+\lambda_{min}}{\lambda_{max}-\lambda_{min}}\right)\right)^2 \\ &= \left(T_t^{-1}\left(\frac{\kappa+1}{\kappa-1}\right)\right)^2 \\ &= 4\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^t + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t\right]^{-2}\end{aligned} \tag{6}$$

As t increases, the second term inside the square brackets converges to zero, hence the convergence of CG is expressed with the weaker inequality

$$(q_t(\lambda))^2 \leq 4\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \tag{7}$$

Substituting this result into 2, we get

$$\left\|\mathbf{x}^t - \mathbf{x}^*\right\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t \|\mathbf{x}^*\|_A \tag{8}$$

□

Fig. 2 charts the convergence per iteration of Conjugate Gradient method. In practice, CG usually converges faster then Eqn. 8 suggests, because of good eigenvalue distributions and good starting points.

Remark 14.2. Conjugate Gradient should be used for regression or least squares problems, or problems that can be decomposed into a combination of such problems. Projection does not work well with CG, and hence there is no guarantee of its performance in constrained optimization settings.

3 Rates of convergence

The error ϵ_t at time t for an optimization problem can be defined in many ways, like $\left|f(\mathbf{x}^t) - f(\mathbf{x}^*)\right|$ or $\left\|\mathbf{x}^t - \mathbf{x}^*\right\|_2$ or $\left\|\mathbf{x}^t - \mathbf{x}^*\right\|_A$, where $\mathbf{x}^*$ is the minima.

Example 14.1. Example of a sub-linear rate of convergence is $\epsilon_t = \frac{1}{\sqrt{t}}$

Definition 14.3. A problem has a linear rate of convergence if

$$\lim_{t\to\infty} \frac{\epsilon_{t+1}}{\epsilon_t} < 1 \tag{9}$$

For example, $\epsilon_t = e^{-t}$ is a linear rate of convergence. This rate guarantees convergence.

Definition 14.4. A problem has a super-linear rate of convergence if

$$\lim_{t\to\infty} \frac{\epsilon_{t+1}}{\epsilon_t} = 0 \tag{10}$$

For example, $\epsilon_t = e^{-t^2}$ is a super-linear rate of convergence. This rate also guarantees convergence

Definition 14.5. A problem has a quadratic rate of convergence if

$$\lim_{t\to\infty} \frac{\epsilon_{t+1}}{\epsilon_t^2} < \infty \tag{11}$$

For example, $\epsilon_t = e^{-e^t}$ is a quadratic rate of convergence. This rate does NOT guarantee convergence

For a quadratic rate of convergence,

$$\begin{aligned}
\epsilon_{t+1} &\leq c\epsilon_t^2 \\
&\leq c.c^2\epsilon_{t-1}^4 \\
&\leq c.c^2.c^4\epsilon_{t-2}^8 \dots \\
&\leq c^{2^{t+1}}\epsilon_0^{2^{t+1}} \\
&= (c\epsilon_0)^{2^{t+1}}
\end{aligned} \tag{12}$$

If $c\epsilon_0 < 1$ then this quadratic rate guarantees convergence, otherwise not.

Example 14.2. If $\epsilon_0 = \frac{1}{2c}, \epsilon_{t+1} \leq \left(\frac{1}{2}\right)^{2^{t+1}}$

4 Newton's Method

Newton's method (NT) is a method for finding the roots of a real-valued function.

Algorithm 1: Newton Method

Input: $f(\mathbf{x})$
Output: $\mathbf{x}^T$
1: Initialize $\mathbf{x}^0$
2: **for** $t = 0, 1, 2, \ldots, T$ **do**
3: $\quad d^t = \left[\nabla^2 f(\mathbf{x}^t)\right]^{-1} \nabla f(\mathbf{x}^t)$
4: $\quad \mathbf{x}^{t+1} \leftarrow \mathbf{x}^t - \alpha_t d^t$ `//Applying Armijo's rule for` α_t
5: **end for**
6: **return** $\mathbf{x}^T$

Definition 14.6. Consider $g : \mathbb{R} \to \mathbb{R}$. Then a first order approximation to g can be written as $\hat{g}(x) = g(x^t) + g'(x^t)(x - x^t)$. To find the root of $g(x)$, set

$$\begin{aligned} &\hat{g}(x) = 0 \\ \Rightarrow &\frac{-g(x^t)}{g'(x^t)} = x - x^t \\ \Rightarrow &x = x^t - \frac{g(x^t)}{g'(x^t)} \end{aligned} \tag{13}$$

In general, Newton's method will get stuck if not initialized properly.

Definition 14.7. Consider $g : \mathbb{R}^n \to \mathbb{R}^n$. Then a first order approximation to g can be written as $\hat{g}(\mathbf{x}) = g(\mathbf{x}^t) + < J_g(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t >$, where $J_{ij} = \frac{\partial g_i(x)}{\partial \mathbf{x}_j}$. To find the root of $g(\mathbf{x})$, set

$$\begin{aligned} &\hat{g}(\mathbf{x}) = 0 \\ \Rightarrow &\mathbf{x} = \mathbf{x}^t - [J_g(\mathbf{x}^t)]^{-1} g(\mathbf{x}^t) \end{aligned} \tag{14}$$

Here, if the Jacobian is singular, NT gets stuck.

Definition 14.8. Let $f(\mathbf{x})$ be a twice continuously differentiable function, and $g = \nabla f$. Then the root of g will be a minima for f. Let $H(\mathbf{x}) = J_g(\mathbf{x}) = \nabla^2 f(\mathbf{x})$. Then the root of g is at

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \left[\nabla^2 f(\mathbf{x}^t)\right]^{-1} \nabla f(\mathbf{x}^t) \tag{15}$$

This is called the Newton step, for g at $\mathbf{x}$. Positive definiteness of $\nabla^2 f(\mathbf{x}^t)$ implies that $\nabla f(\mathbf{x}^t)^T (\mathbf{x}^{t+1} - \mathbf{x}^t) = -\nabla f(\mathbf{x}^t)^T \left[\nabla^2 f(\mathbf{x}^t)\right]^{-1} \nabla f(\mathbf{x}^t) < 0$, unless $\nabla f(\mathbf{x}^t) = 0$, which means that the Newton step is a descent direction, unless $\mathbf{x}^t$ is optimal. (Boyd and Vandenberghe (2004))

Definition 14.9. The update rule for the Damped Newton Method is:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t \left[\nabla^2 f(\mathbf{x}^t)\right]^{-1} \nabla f(\mathbf{x}^t) \tag{16}$$

Remark 14.3. The update rule for the Newton Method as defined in Defn. 14.8 is the solution of minimizing a second order approximant to f, i.e.

$$\begin{aligned} &\hat{f}(\mathbf{x}) = f(\mathbf{x}^t) + < \nabla f(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t > + \frac{1}{2} < \mathbf{x} - \mathbf{x}^t, H(\mathbf{x}^t)(\mathbf{x} - \mathbf{x}^t) > \\ &\arg\min_{\mathbf{x}} \hat{f}(\mathbf{x}) = \mathbf{x}^t - \left[\nabla^2 f(\mathbf{x}^t)\right]^{-1} \nabla f(\mathbf{x}^t) \end{aligned} \tag{17}$$

Example 14.3. For a least squares problem as in Defn. 14.1,

$$\begin{aligned} \nabla f(\mathbf{x}) &= A\mathbf{x} - \mathbf{b} \\ \nabla^2 f(\mathbf{x}) &= A \\ \Rightarrow \mathbf{x}^{t+1} &= \mathbf{x}^t - A^{-1}(A\mathbf{x}^t - \mathbf{b}) \\ &= A^{-1}\mathbf{b} \end{aligned} \tag{18}$$

Therefore, NT gives one step convergence, but at the cost of calculating the inverse of A^{-1} which is $\mathcal{O}(d^3)$.

Theorem 14.2. Fundamental Theorem of Calculus. For $f : \mathbb{R} \Rightarrow \mathbb{R}$, we have

$$\begin{aligned} f(b) &= f(a) + \int_a^b f'(t)dt \\ &= f(a) + \int_0^1 f'(a + t(b-a))(b-a)dt \end{aligned} \tag{19}$$

Similarly, for $f : \mathbb{R}^n \Rightarrow \mathbb{R}^n$,

$$f(\mathbf{x}) = f(\mathbf{y}) + \int_0^1 J_f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y})dt \tag{20}$$

4.1 NT has quadratic rate of convergence

Assume that

- $f(\mathbf{x})$ is strongly convex, i.e.

$$f(\mathbf{x}) \geq f(\mathbf{y}) + < \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} > + \frac{\alpha}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \tag{21}$$

- $f(\mathbf{x})$ has $\mathcal{L}$-Lipschitz Hessians, i.e.

$$\left\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\right\|_2 \leq \mathcal{L} \|x - y\|_2 \tag{22}$$

  where the 2-norm of a matrix is its spectral norm, i.e. its maximum eigenvalue.

Claim 14.3. $\nabla f(\mathbf{x}^t)$ has a quadratic rate of convergence

Proof.

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \left[\nabla^2 f(\mathbf{x}^t)\right]^{-1} \nabla f(\mathbf{x}^t) \tag{23}$$

$$\begin{aligned} Let\ \Delta^t &= \mathbf{x}^{t+1} - \mathbf{x}^t \\ &= -\left[\nabla^2 f(\mathbf{x}^t)\right]^{-1} \nabla f(\mathbf{x}^t) \\ \Rightarrow &\nabla f(\mathbf{x}^t) = -\nabla^2 f(\mathbf{x}^t)\Delta^t \end{aligned} \tag{24}$$

$$\begin{aligned} \nabla f(\mathbf{x}^{t+1}) &= \nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t) + \nabla f(\mathbf{x}^t) \\ &= \nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t) - \nabla^2 f(\mathbf{x}^t)\Delta^t \\ &= \int_0^1 \nabla^2 f(\mathbf{x}^t + t\Delta^t)\Delta^t dt - \nabla^2 f(\mathbf{x}^t)\Delta^t \end{aligned} \tag{25}$$

$$
\begin{aligned}
\Rightarrow \left\|\nabla f(\mathbf{x}^{t+1})\right\|_2 &= \left\| \int_0^1 (\nabla^2 f(\mathbf{x}^t + t\Delta^t) - \nabla^2 f(\mathbf{x}^t))\Delta^t dt \right\|_2 \\
&\leq \int_0^1 \left\|\nabla^2 f(\mathbf{x}^t + t\Delta^t) - \nabla^2 f(\mathbf{x}^t)\right\|_2 \left\|\Delta^t\right\|_2 dt \\
&\leq \int_0^1 \mathcal{L} \left\|\Delta^t\right\|_2^2 dt \\
&= \frac{\mathcal{L}}{2} \left\|\Delta^t\right\|_2^2 \\
&= \frac{\mathcal{L}}{2} \left\| \left[\nabla^2 f(\mathbf{x}^t)\right]^{-1} \nabla f(\mathbf{x}^t) \right\|_2^2 \\
&\leq \frac{\mathcal{L}}{2} \left\| \left[\nabla^2 f(\mathbf{x}^t)\right]^{-1} \right\|_2^2 \left\|\nabla f(\mathbf{x}^t)\right\|_2^2 \\
\Rightarrow \left\|\nabla f(\mathbf{x}^{t+1})\right\|_2 &\leq \frac{\mathcal{L}}{2\alpha} \left\|\nabla f(\mathbf{x}^t)\right\|_2^2
\end{aligned}
\tag{26}
$$

Hence, proved that $\nabla f(\mathbf{x}^t)$ has a quadratic rate of convergence. □

References

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Jonathan Richard Shewchuk. An introduction to the conjugate gradient method without the agonizing pain, 1994.