Indian Institute of Technology Kanpur
CS774 Optimization Techniques

Scribe: Debojyoti Dey
Instructor: Purushottam Kar
Date: September 22, 2016

LECTURE
13

Conjugate Gradient Descent

1 Introduction

In the last lecture we have seen the construction of Conjugate Gradient Descent method for quadratic optimization problem. In this lecture, we will present an analysis of the algorithm. We will prove optimality of the step lengths in all iterations. We will express a solution of the algorithm as a point in Krylov subspace. We will finally show the convergence property of CG method as a minmax problem in Krylov subspace.

2 A quick recap

In quadratic optimization problem, we need to minimize the objective function as follows,

$$\min_{x \in \mathcal{X}} f(x) = \min_{x \in \mathcal{X}} \left\{ \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x} \right\} \qquad A \in \mathbb{R}^{d\times d} \text{ and } \mathbf{b} \in \mathbb{R}^d$$

In Conjugate Gradient Descent, we are driven by the following motivation while taking a gradient step: "Never explore the already explored gradient!". The directions of update in this method are pairwise conjugate wrt A. We obtain the subsequent orthogonal directions by Gram Schimdt orthogonalization process, modified by the fact that the inner products used are induced by A.

Algorithm 1: Conjugate Gradient Descent Method

```
Input: $A \succ 0, \mathbf{b}$
$\mathbf{x}^0 \leftarrow \mathbf{0}$
for $t = 0, 1, \ldots, T-1$ do
    $\mathbf{r}^t \leftarrow \mathbf{b} - A\mathbf{x}^t$
    $\mathbf{p}^t \leftarrow \mathbf{r}^t - \sum_{i<t} \frac{\langle \mathbf{r}^t, A\mathbf{p}^i \rangle}{\langle \mathbf{p}^i, A\mathbf{p}^i \rangle}\mathbf{p}^i$        //Gram Schimdt step
    $\alpha_t \leftarrow \frac{\langle \mathbf{p}^t, \mathbf{b} \rangle}{\langle \mathbf{p}^t, A\mathbf{p}^t \rangle}$        //optimal step length
    $\mathbf{x}^{t+1} \leftarrow \mathbf{x}^t + \alpha_t \mathbf{p}^t$
end for
return $\mathbf{x}^T$
```

We claim that the step lengths α^t used in algorithm 1 are optimal, individually and together. In the next section, we prove our claim to be true.

3 Analysis of CG Algorithm

We will discuss three properties of Conjugate Gradient method. The claims are as follows:

Claim 13.1. Step length α_t is conditionally optimal that is,

$$\alpha_t = \arg\min_{\alpha\in\mathbb{R}} f(\mathbf{x}^{t+1}|\mathbf{x}^t, \mathbf{p}^t) = \arg\min_{\alpha\in\mathbb{R}} f(\mathbf{x}^t + \alpha\mathbf{p}^t)$$

Proof. $\mathbf{x}^{t+1} \leftarrow \mathbf{x}^t + \alpha_t\mathbf{p}^t$
We assume that $\mathbf{x}^{t+1}$ attains the lowest residual at step t.
For $\mathbf{x}^{t+1}$ to be the optimal, $\mathbf{p}^t$ has to be orthogonal to $\mathbf{r}^{t+1} = \mathbf{b} - A\mathbf{x}^{t+1}$ with respect to A. Otherwise there will exist a negative component of the residual along $\mathbf{p}^t$ which can still be reduced by moving in the direction and thus contradicting our hypothesis. Therefore,

$$\begin{aligned}
&\left\langle \mathbf{p}^t, \mathbf{b} - A(\mathbf{x}^t + \alpha_t\mathbf{p}^t)\right\rangle = 0 \\
&\Rightarrow \left\langle \mathbf{p}^t, \mathbf{b}\right\rangle = \left\langle \mathbf{p}^t, A\mathbf{x}^t\right\rangle + \alpha_t\left\langle \mathbf{p}^t, A\mathbf{p}^t\right\rangle \\
&\Rightarrow \left\langle \mathbf{p}^t, \mathbf{b}\right\rangle = 0 + \alpha_t\left\langle \mathbf{p}^t, A\mathbf{p}^t\right\rangle \qquad \text{as } \mathbf{x}^t \in \mathit{span}\left\{\mathbf{p}^0, \mathbf{p}^1, \cdots, \mathbf{p}^{t-1}\right\} \\
&\Rightarrow \alpha_t = \frac{\left\langle \mathbf{p}^t, \mathbf{b}\right\rangle}{\left\langle \mathbf{p}^t, A\mathbf{p}^t\right\rangle} \qquad (1)
\end{aligned}$$

Hence, the value of α_t used in algorithm 1 is optimal. □

Claim 13.2. Step lengths $\{\alpha_i\}_{i<t}$ are optimal which means,

$$\mathbf{x}^t = \arg\min_{\mathbf{x}\in \mathit{span}\{\mathbf{p}^0,,\mathbf{p}^1,\cdots,\mathbf{p}^{t-1}\}} f(\mathbf{x})$$

Proof. We need to prove that there exists no point other than $\mathbf{x}^t$ in the space generated by $\left(\mathbf{p}^0,,\mathbf{p}^1,\cdots,\mathbf{p}^{t-1}\right)$ which can further reduce the value of $f(\mathbf{x})$. From optimality condition of constrained optimization we already know that, $\mathbf{x}^\star \in \mathcal{X}$ is a local optima only if,

$$\forall \mathbf{x} \in \mathcal{X}, \langle \nabla f(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star\rangle \geq 0$$

where $\mathcal{X}$ is the constraint space. In our case the constraint space is the entire t dimensional subspace given by $\left\{p^i\right\}_{i<t}$. Thus for $\mathbf{x}^t$ to be optimal there should be no point $\mathbf{x}$ in the current subspace such that $f(\mathbf{x}) < f(\mathbf{x}^t)$. We show the fact in the following.

$$\begin{aligned}
&\left\langle \nabla f(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t\right\rangle = 0 \qquad \forall \mathbf{x} \in \mathit{span}\left\{\mathbf{p}^0,,\mathbf{p}^1,\cdots,\mathbf{p}^{t-1}\right\} \\
&\Leftarrow \left\langle \nabla f(\mathbf{x}^t), \mathbf{p}^i\right\rangle = 0 \qquad \forall i < t \qquad (2)
\end{aligned}$$

Next we prove equation 2.

$$\begin{aligned}
\left\langle \nabla f(\mathbf{x}^t), \mathbf{p}^i\right\rangle &= \left\langle b - A\mathbf{x}^t, \mathbf{p}^i\right\rangle \\
&= \left\langle b, \mathbf{p}^i\right\rangle - \left\langle \sum_{j<t} A\alpha_j\mathbf{p}^j, \mathbf{p}^i\right\rangle \\
&= \left\langle b, \mathbf{p}^i\right\rangle - \left\langle A\alpha_i\mathbf{p}^i, \mathbf{p}^i\right\rangle \\
&= \left\langle b, \mathbf{p}^i\right\rangle - \left\langle \mathbf{p}^i, A\mathbf{p}^i\right\rangle \frac{\left\langle \mathbf{p}^i, \mathbf{b}\right\rangle}{\left\langle \mathbf{p}^i, A\mathbf{p}^i\right\rangle} \qquad \text{using } \alpha_i \text{ from equation 1} \\
&= 0
\end{aligned}$$

Hence $\mathbf{x}^t$ is proved to be the optimal solution of $f(x)$ in the subspace it belongs to. □

Claim 13.3. $\left\langle \mathbf{p}^t, A\left(\mathbf{x}^{t+1} - \mathbf{x}^\star\right)\right\rangle = 0$, $\forall t$

Proof. Left as an exercise □

4 Krylov Subspaces

Given a $(d \times d)$ matrix A and d-dimensional vector $\mathbf{b}$, Krylov subspace of order t is defined to be Wikipedia (2016b) the linear subspace spanned by the images of b under the first t powers of A(starting from $A^0 = I$), that is,

$$\mathcal{K}_t = span\left\{\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \cdots, A^{t-1}\mathbf{b}\right\} \qquad \text{with } \mathcal{K}_0 = \{\mathbf{0}\}$$

Following are the properties that hold true,

$$\begin{aligned} &\mathcal{K}_i = \mathcal{K}_{i-1} \implies \mathcal{K}_j = \mathcal{K}_{i-1} && \forall j \geq i-1 \\ &\mathcal{K}_{d+1} = \mathcal{K}_d \end{aligned}$$

5 Characteristic Polynomial

We consider a square matrix $A \in \mathbb{R}^{d \times d}$. The characteristic polynomial of A is defined as follows:

$$\begin{aligned} p(x) &= det(A - \lambda I) \\ &= x^d + a_1 x^{d-1} + \cdots + a_d && \forall i, a_i \in \mathbb{R} \end{aligned}$$

The zeros of the polynomial are the eigenvalues of A.

6 Cayley-Hamilton Theorem

In linear algebra, Cayley-Hamilton theorem states Wikipedia (2016a) that every square matrix over a commutative ring(such as the real and complex field) satisfies its own characteristic equation. For the characteristic equation of the matrix A in the last section we get,

$$\begin{aligned} &p(A) = 0 \\ &\Rightarrow A^d + a_1 A^{d-1} + \cdots + a_d I = 0 \\ &\Rightarrow \frac{A^{d-1}b + a_1 A^{d-2} + \cdots + a_{d-1}b}{-a_d} = A^{-1}b \quad \text{(by multiplication with } A^{-1}b \text{ and re-arranging)} \end{aligned}$$

Corollary 13.4. For quadratic optimization problem, we have the closed form solution $\mathbf{x}^\star = A^{-1}b$. Therefore, we can say that $\mathbf{x}^\star \in \mathcal{K}_d$

Claim 13.5. $span\left\{\mathbf{p}^0, \mathbf{p}^1, \cdots, \mathbf{p}^{t-1}\right\} = \mathcal{K}_{t-1}$

Proof. Left as an exercise. □

Corollary 13.6. In the CG algorithm 1 we can see that $\mathbf{x}_t$ is a linear combination of $\mathbf{p}^0, \mathbf{p}^1, \cdots, \mathbf{p}^{t-1}$. Therefore, from the previous claim $\mathbf{x}^t \in \mathcal{K}_{t-1}$

Corollary 13.7. Using the above facts, it can be said that CG converges in d iterations.

7 Analysis of Convergence

In this section we will derive a formulation of convergence of CG method minimizing the quadratic function,

$$f(x) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x} \tag{3}$$

The optimal solution of the problem is given by $x^\star = A^{-1}b$. Using the value of $x^\star$ in equation 3 we get,

$$\begin{aligned} f(x^\star) &= \frac{1}{2}\mathbf{b}^T A^{-1}\mathbf{b} - \mathbf{b}^T A\mathbf{b} \\ &= -\frac{1}{2}\mathbf{b}^T A^{-1}\mathbf{b} = -\frac{1}{2}\left\|A^{-1}\mathbf{b}\right\|_A^2 = \frac{1}{2}\left\|\mathbf{x}^\star\right\|_A^2 \end{aligned}$$

Lemma 13.8 (Shewchuk (1994)). Potential function $f(\mathbf{x}^t) - f(\mathbf{x}^\star) = -\frac{1}{2}\left\|\mathbf{x}^t - \mathbf{x}^\star\right\|_A^2$

Proof. Let, $\mathbf{e}_t = \mathbf{x}^t - \mathbf{x}^\star$

$$\begin{aligned} f(\mathbf{x}^t) = f(\mathbf{x}^\star + \mathbf{e}_t) &= \frac{1}{2}\mathbf{x}^{\star T} A\mathbf{x}^\star + \frac{1}{2}\mathbf{e}_t^T A\mathbf{e}_t^T + \mathbf{e}_t^T A\mathbf{x}^\star - \mathbf{b}^T\mathbf{x}^\star - \mathbf{b}^T\mathbf{e}_t \\ &= \left(\frac{1}{2}\mathbf{x}^{\star T} A\mathbf{x}^\star - \mathbf{b}^T\mathbf{x}^\star\right) + \frac{1}{2}\mathbf{e}_t^T A\mathbf{e}_t^T + \mathbf{e}_t^T b - \mathbf{b}^T\mathbf{e}_t \\ &= f(x^\star) + \frac{1}{2}\left\|\mathbf{x}^t - \mathbf{x}^\star\right\|_A^2 \end{aligned}$$

$\therefore f(\mathbf{x}^t) - f(\mathbf{x}^\star) = \frac{1}{2}\left\|\mathbf{x}^t - \mathbf{x}^\star\right\|_A^2$ □

Using our knowledge from previous sections,

$$\mathbf{x}^t \in \mathcal{K}_{t-1} \iff \mathbf{x}^t = p(A)\mathbf{b} \qquad \text{where } deg(p) < t$$

We get the following,

$$\begin{aligned} f(\mathbf{x}^t) - f(\mathbf{x}^\star) = \frac{1}{2}\left\|\mathbf{x}^t - \mathbf{x}^\star\right\|_A^2 &= \min_{deg(p)<t} \frac{1}{2}\left\|p(A)\mathbf{b} - A^{-1}\mathbf{b}\right\|_A^2 \\ &= \min_{deg(p)<t} \frac{1}{2}\left\|(p(A) - A^{-1})\mathbf{b}\right\|_A^2 \end{aligned} \tag{4}$$

Let Q is a $(d \times d)$ matrix with the singular vectors of A as its columns. The corresponding eigenvalues of A are $\lambda_1, \lambda_2, \cdots, \lambda_d$. Then,

$$A = Q\Lambda Q^T \Rightarrow p(A) = Qp(\Lambda)Q^T$$

where $Q^T Q = I$ and $p(\Lambda) = diag(p(\lambda_1), p(\lambda_2), \cdots, p(\lambda_d))$. Putting the values in equation 4 we get,

$$
\begin{aligned}
f(\mathbf{x}^t) - f(\mathbf{x}^\star) &= \min_{deg(p)<t} \frac{1}{2} \left\| (p(\Lambda) - \Lambda^{-1})\mathbf{c} \right\|_\Lambda^2 && \text{where } Q^T\mathbf{b} = \mathbf{c} \\
&= \min_{deg(p)<t} \frac{1}{2} \sum_{i=1}^{d} (p(\lambda_i) - \lambda_i^{-1})^2 \mathbf{c}_i^2 \lambda_i \\
&= \min_{deg(p)<t} \frac{1}{2} \sum_{i=1}^{d} \frac{1}{\lambda_i} (\lambda_i p(\lambda_i) - 1)^2 \mathbf{c}_i^2 \\
&= \min_{\substack{deg(q)\le t \\ |q(0)|=1}} \frac{1}{2} \sum_{i=1}^{d} \frac{\mathbf{c}_i^2}{\lambda_i} (q(\lambda_i))^2 \\
&\le \left(\frac{1}{2} \sum_{i=1}^{d} \frac{\mathbf{c}_i^2}{\lambda_i} \right) \min_{\substack{deg(q)\le t \\ |q(0)|=1}} \max_i (q(\lambda_i))^2
\end{aligned}
\tag{5}
$$

Remark 13.1. Show that $\frac{1}{2}\sum_{i=1}^{d} \frac{\mathbf{c}_i^2}{\lambda_i} = \frac{1}{2}\left\| x^\star \right\|_A^2$.

Therefore, the inequality 5 can be written as:

$$
f(\mathbf{x}^t) - f(\mathbf{x}^\star) = \frac{1}{2} \left\| \mathbf{x}^t - \mathbf{x}^\star \right\|_A^2 \le \frac{1}{2} \left\| x^\star \right\|_A^2 \min_{\substack{deg(q)\le t \\ |q(0)|=1}} \max_i (q(\lambda_i))^2 \tag{6}
$$

Remark 13.2. If there are only m distinct eigenvalues of A, Conjugate Gradient method converges in m steps.

Remark 13.3. If the eigenvalues of A are clustered in k groups centered at $l_1, l_2, \cdots, l_k$ respectively such that,

$$
\forall \lambda_i, \exists j \in [k] \text{ so that } |\lambda_i - l_j| \le \epsilon \qquad \text{where } \epsilon > 0
$$

then CG converges in k steps.

8 Conclusion

In this chapter, we analysed the Conjugate Gradient Descent algorithm. We verified the optimality of step lengths. We derived an upper bound of convergence in terms of characteristic polynomial of A. In the next lecture we will mimic Chebyshev's polynomial as the characteristic polynomial and show its upper bound in terms of condition number $\mathcal{K}$ of matrix A.

References

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

Wikipedia. Cayleyhamilton theorem — wikipedia, the free encyclopedia, 2016a. URL `https://en.wikipedia.org/w/index.php?title=Cayley%E2%80%93Hamilton_theorem&oldid=743445932`. [Online; accessed 9-October-2016].

Wikipedia. Krylov subspace — wikipedia, the free encyclopedia, 2016b. URL https://en.wikipedia.org/w/index.php?title=Krylov_subspace&oldid=727610996. [Online; accessed 30-June-2016].