Numerical Optimization
Lecture 12: Conjugate Gradient Method

Sangkyun Lee

The content is from Nocedal and Wright (2006). Topics marked with ** are optional.

1 Linear CG as a Coordinate Descent in a Transformed Space

Consider a minimization of strictly convex quadratic function,

$$\min_{x \in \mathbb{R}^n} \phi(x) := \frac{1}{2} x^T A x - b^T x$$

where $A$ is symmetric positive definite.

If $A$ is diagonal, then the contour of $\phi$ is coordinate-aligned. In this case, successive minimization along coordinate directions $e_1, e_2, \ldots, e_n$ finds the minimizer. When $A$ is not diagonal, $\phi$ is typically not aligned with the standard coordinates, and coordinate-descent may not even converge.

Conjugate gradient method recovers the property of coordinate descent, but in a transformed space. Define a matrix $S$ as

$$S = [p_0 \ p_1 \ \cdots \ p_{n-1}] \in \mathbb{R}^{n \times n}$$

where $\{p_0, p_1, \ldots, p_{n-1}\}$ is the set of conjugate directions w.r.t. $A$. Defining a new variable $\hat{x}$ by

$$\hat{x} = S^{-1} x,$$

gives us a new function

$$\hat{\phi}(\hat{x}) := \phi(S\hat{x}) = \frac{1}{2} \hat{x}^T (S^T A S) \hat{x} - (S^T b)^T \hat{x}.$$

By conjugacy property, the matrix $(S^T A S)$ is diagonal, and therefore we can find the minimizer of $\hat{\phi}$ by performing $n$ one-dimensional minimizations along the standard coordinate directions. Since

$$e_i = S^{-1} p_i,$$

the coordinate descent applied to $\hat{\phi}$ is equivalent to the conjugate gradient approach applied to $\phi$.

2 Expanding Subspace Minimization

Another interesting property of linear CG in analogy to coordinate descent is that after $k$ steps, the function $\phi$ is minimized on the subspace spanned by $\{p_0, p_1, \ldots, p_{k-1}\}$.

From now on we will use a relation that is easily verifiable,

$$r_{k+1} = r_k + \alpha_k A p_k. \quad (12.1)$$
Theorem 12.1. Let \( x_0 \in \mathbb{R}^n \) be a starting point and suppose that \( \{x_k\} \) is generated by

\[
x_{i+1} = x_i + \alpha_i p_i, \quad \alpha_i = -\frac{r_i^T p_i}{p_i^T A p_i}, \quad r_i = Ax_i - b,
\]

for conjugate directions \( p_0, p_1, \ldots, p_{k-1} \). Then,

\[
r_k^T p_i = 0, \quad \text{for } i = 0, 1, \ldots, k - 1,
\]

and \( x_k \) is the minimizer of \( \phi(x) \) over the set

\[
\{x : x = x_0 + \text{span}\{p_0, p_1, \ldots, p_{k-1}\}\}.
\]

Proof. We first show a point \( \tilde{x}_0 \) for all \( i \in S \). From the fact that \( x_0 \) is the minimizer of \( \phi(x) \), so that \( r_0 \) is orthogonal to all previous directions, \( \{p_0, p_1, \ldots, p_{k-1}\} \), which is an important property to be used later.

To show that \( x_k \) satisfies (12.1), we use induction. For the case \( k = 1 \), we have

\[
r_k = r_{k-1} + \alpha_{k-1} A p_{k-1},
\]

so that

\[
p_{k-1}^T r_k = p_{k-1}^T r_{k-1} + \alpha_{k-1} p_{k-1}^T A p_{k-1} = 0
\]

by the definition of \( \alpha_{k-1} \). For the other vectors \( p_i, i = 0, 1, \ldots, k - 2 \), we have

\[
p_i^T r_k = p_i^T r_{k-1} + \alpha_{k-1} p_i^T A p_{k-1} = 0
\]

due the the induction hypothesis and the conjugacy of vectors. This shows that \( r_k^T p_i = 0 \) for \( i = 0, 1, \ldots, k - 1 \), completing the proof.

This theorem tells that the current residual \( r_k \) is orthogonal to all previous search directions, which is an important property to be used later.

So far, the linear CG method has been described with any set of conjugate directions \( \{p_0, p_1, \ldots, p_{n-1}\} \). For example, the set of eigenvectors of \( A \) can serve the purpose, although is is not appealing when \( n \) is large since computing all eigenvectors requires \( O(n^3) \) computation and \( O(n^2) \) storage.

We will discuss next time how to generate conjugate directions iteratively with \( O(n^2) \) operations and without having to store all previous vectors.

References