Notes on Nocedal and Wright's "Numerical Optimization"
Chapter 5 – "Conjugate Gradient Methods"

Lucas N. Ribeiro

Introduction I

This method was proposed in the 1950s by Hestenes and Stiefel for solving large-scale linear systems of equations. Two fundamental approaches: linear and nonlinear conjugate gradient methods

Linear conjugate gradient

- Iterative method for solving $Ax = b$, where A is $n \times n$ symmetric and positive definite.
- minimizes $\phi(x) = 0.5x^{\mathsf{T}}Ax - b^{\mathsf{T}}x$

Conjugate Direction Methods I

Definition – conjugacy

A set of nonzero vectors $\{p_0, \ldots, p_\ell\}$ is said to be *conjugate* with repect to the symmetric positive definite matrix A if

$$p_i^{\mathsf{T}} A p_j = 0, \quad \forall i \neq j.$$

This property lies in the fact that we can minimizes ϕ in n steps by successively minimizing it along the individual directions in a conjugate set. Given a starting point x_0 and a set of conjugate directions, we generate the sequence $\{x_k\}$ by setting

$$x_{k+1} = x_k + \alpha_k p_k$$

where $\alpha_k = -r_k^{\mathsf{T}} p_k / (p_k^{\mathsf{T}} A p_k)$ is the minimizer of ϕ along $x_k + \alpha p_k$.

Conjugate Direction Methods II

Theorem

Theorem 5.1 The sequence generated by $\{x_k\}$ considering the conjugate direction algorithm converges to the solution x^* of the linear system in at most n steps.

Interpretation: if A was diagonal, then the minimizer of ϕ would be found by performing 1D minimizations along the coordinate directions.
Let's consider a variable transformation when A is not diagonal: $\hat{x} = S^{-1}x$ where $S = [p_0, \ldots p_{n-1}]$. Then $S^{\mathsf{T}}AS$ is diagonal (due to conjugacy) and we optimize now

$$\hat{\phi}(S\hat{x}) = 0.5\hat{x}^{\mathsf{T}}(S^{\mathsf{T}}AS)\hat{x} - (S^{\mathsf{T}}b)^{\mathsf{T}}\hat{x}$$

So now we can find the minimizing value of $\hat{\phi}$ by performing n 1D minimizations along the coord. directions.

Conjugate Direction Methods III

Theorem 5.2 – Expanding subspace minimization

Let x_0 be a starting point and that $\{x_k\}$ is generated by the conjugate direction algorithm. Then $r_k^T p_i = 0$ for $i = 0, \ldots, k-1$

This theorem establishes that the current residual r_k is orthogonal to all previous search directions.
Some possible conjugate vectors: eigenvectors of A (but may be expensive to calculate in large problems).

Basic Properties of the CG Method I

Each direction p_k is chosen to be a linear combination of the steepst descent direction $(-r_k)$ and the previous direction p_{k-1}:

$$p_k = -r_k + \beta_k p_{k-1}$$

where β_k is to be determined by the requirement that p_{k-1} and p_k must be conjugate w.r.t. A. From this, we get:

$$\beta_k = \frac{r_k^{\mathsf{T}} A p_{k-1}}{p_{k-1}^{\mathsf{T}} A p_{k-1}} \tag{1}$$

There is another property that establishes that the residuals r_i are mutually orthogonal and that the search direction p_k and residual r_k are contained in the Krylov subspace of degree k for r_0:

$$r_k \in \text{span}\{r_0, Ar_0, \ldots, A^k r_0\}.$$

CG Algorithm

- Given x_0
- Init. $r_0 = Ax_0 - b$, $p_0 = -r_0$, $k = 0$
- While $r_k \neq 0$
  1. Update solution:
     - $\alpha_k = r_k^T r_k / (p_k^T A p_k)$
     - $x_{k+1} = x_k + \alpha_k p_k$
  2. Update residual: $r_{k+1} = r_k + \alpha_k A p_k$
  3. Calculate new conjugate direction:
     - $\beta_{k+1} = r_{k+1}^T r_{k+1} / (r_k^T r_k)$
     - $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$
  4. $k = k + 1$

- Depending on the eigenvalue spread of A, the algorithm may converge even faster than n steps.
- Preconditioners may be applied to improve the eigenvalue distribution of A

Nonlinear CG methods I

- So far, we considered ϕ as a convex quadratic function. It is possible to consider general convex (or even non-linear) functions by adapting the standard CG.
  - Fletcher Reeves method: line search + replace residual by actual gradient
  - Polak-Ribière method
  - Many others...
- Non-linear CG is appealing because each iteration requires only evaluation of the obj. function and its gradient!
- Complicated convergence analysis

Fletcher Reeves CG

- Given x_0
- Init. $p_0 = -\nabla f(x_0)$, $k = 0$
- While $\nabla f_k \neq 0$
  1. Update solution:
     - Line search for α_k
     - $x_{k+1} = x_k + \alpha_k p_k$
  2. Calculate new conjugate direction:
     - $\beta_{k+1} = \nabla f_{k+1}^{\mathsf{T}} \nabla f_{k+1} / (\nabla f_k^{\mathsf{T}} \nabla f_k)$
     - $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$
  3. $k = k + 1$

Polak-Ribière formula: use $\beta_{k+1}^{PR} = \frac{\nabla f_{k+1}^{\mathsf{T}} (\nabla f_{k+1} - \nabla f_k)}{\|\nabla f_k\|^2}$

To ensure strong Wolfe's condition: $\beta_{k+1}^{+} = \max\{0, \beta_{k+1}^{PR}\}$