Numerical Optimization
Lecture 14: Nonlinear CG and Constrained Optimization
Sangkyun Lee

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

1 Nonlinear Conjugate Gradient Method

The idea of conjugate gradient can be extended for minimizing a nonlinear objective function $f$ instead of a strictly convex objective $\phi$.

1.1 Fletcher-Reeves Algorithm

Algorithm 1: FR

1. **Input**: $x_0$, $f$, $\nabla f$;
2. $p_0 = -\nabla f(x_0)$, $k = 0$;
3. while $\nabla f(x_k) \neq 0$ do
   4. Compute $\alpha_k$ by line search;
   5. $x_{k+1} = x_k + \alpha_k p_k$;
   6. $\beta_{k+1} = \nabla f(x_{k+1})^T \nabla f(x_{k+1}) / \nabla f(x_k)^T \nabla f(x_k)$;
   7. $p_{k+1} = -\nabla f(x_{k+1}) + \beta_{k+1} p_k$;
   8. $k = k + 1$;
3. end
4. **Output**: $x^* = x_k$.

If $f$ is a strictly convex quadratic function and $\alpha_k$ is computed with an exact line search, then this algorithm reverts to the linear CG algorithm. Algorithm 1 is appealing for solving large-scale problems, since it does not involve any matrix operation is required.

From the update rule of $p_k$, we have

$$\nabla f(x_k)^T p_k = -\|\nabla f(x_k)\|^2 + \beta_k^{FR} \nabla f(x_k)^T p_{k-1}$$

When the line search is exact so that $\alpha_{k-1}$ is a local minimizer of $f$ along $p_{k-1}$, we have $\nabla f(x_k)^T p_{k-1} = 0$ and therefore $p_k$ is a descent direction. Otherwise, the line search need to satisfy the strong Wolfe conditions with $0 < c_1 < c_2 < \frac{1}{2}$ (we skip the proof),

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k)^T p_k$$

$$\|\nabla f(x_k + \alpha_k p_k)^T p_k\| \leq -c_2 \nabla f(x_k)^T p_k$$
1.2 Convergence of FR

**Theorem 14.1.** Suppose that the level set \( \mathcal{L} = \{ x : f(x) \leq f(x_0) \} \) is bounded, and that \( f : \mathbb{R}^n \to \mathbb{R} \) is Lipschitz continuously differentiable in some open neighborhood of \( \mathcal{L} \). Algorithm 1 with a line search satisfying the strong Wolfe conditions with \( 0 < c_1 < c_2 < \frac{1}{2} \) satisfies that

\[
\liminf_{k \to \infty} \| \nabla f(x_k) \|^2_2 = 0.
\]

1.3 Other Nonlinear CG Methods

**Polak-Riviè re (PR)**

\[
\beta_{k+1}^{PR} = \frac{\nabla f(x_{k+1})^T (\nabla f(x_{k+1}) - \nabla f(x_k))}{\nabla f(x_k)^T \nabla f(x_k)}
\]

PR with strong Wolfe LS does not guarantee that \( p_k \) is descent. Choosing \( \beta_{k+1}^+ = \max\{\beta_{k+1}^{PR}, 0\} \) and a simple modification of strong Wolfe LS can give the guarantee back (this algorithm is called PR+). PR and PR+ converge better than FR in practice, but only PR does not have a global convergence result when \( f \) is a general nonlinear function.

On the other hand, under the assumption that \( f \) is strongly convex and LS is exact, then it can be shown for PR that

\[
\lim_{k \to \infty} \| \nabla f(x_k) \|^2_2 = 0.
\]

**Hestenes-Stiefel (HS)**

\[
\beta_{k+1}^{HS} = \frac{\nabla f(x_{k+1})^T (\nabla f(x_{k+1}) - \nabla f(x_k))}{(\nabla f(x_{k+1}) - f(x_k))^T p_k}
\]

HS behaves very similarly to PR.

2 Constrained Optimization

We now consider constrained minimization problems,

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

subject to \( c_i(x) = 0, \; i \in \mathcal{E}, \)

\( c_i(x) \geq 0, \; i \in \mathcal{I}. \)

where \( f \) and \( c_i \)'s are all smooth, real-valued functions on a subset of \( \mathbb{R}^n \), and \( \mathcal{E} \) and \( \mathcal{I} \) are two finite sets of indices for equality and inequality constraints, resp. We define the feasible set \( \Omega \) as follows,

\[
\Omega := \{ x : c_i(x) = 0, \; i \in \mathcal{E}, \; c_i(x) \geq 0, \; i \in \mathcal{I} \}.
\]

2.1 Types of Solutions

Types of minimizers are redefined as follows,

- \( x^* \) is a local minimizer if \( x^* \in \Omega \) and there is a neighborhood \( \mathcal{N} \) of \( x^* \) such that \( f(x^*) \leq f(x) \) for all \( x \in \mathcal{N} \cap \Omega \).
• $x^*$ is a strict local minimizer if $x^* \in \Omega$ and there is neighborhood $N$ of $x^*$ such that $f(x^*) < f(x)$ for all $x \in N \cap \Omega$ with $x \neq x^*$.

• $x^*$ is an isolated local minimizer if $x^* \in \Omega$ and there is a neighborhood $N$ of $x^*$ such that $x^*$ is the only local minimizer in $N \cap \Omega$.

2.2 Active Set

The active set $A(x)$ at any feasible point $x$ is,

$$A(x) := \mathcal{E} \cup \{ i \in \mathcal{I} : c_i(x) = 0 \}.$$  

At a feasible point $x$, an inequality constraint $i \in \mathcal{I}$ is called active if $c_i(x) = 0$, and inactive if $c_i(x) > 0$.

2.3 Examples

Ex. 1

$$\begin{align*}
\text{min } & f(x_1, x_2) = x_1 + x_2 \\
\text{s.t. } & c_1(x_1, x_2) = x_1^2 + x_2^2 - 2 = 0.
\end{align*}$$

Consider $\nabla f(x)$ and $\nabla c_1(x)$ at $(-1, -1)^T$, $(-1, 1)^T$, $(1, -1)^T$, and $(-1, -1)^T$. The solution is obviously $(-1, -1)^T$.

We see that with $\lambda_1^* = -1/2$,

$$\nabla f(x^*) = \lambda_1^* \nabla c_1(x^*).$$

This can be derived as follows. Suppose that a feasible $x$ is not a minimizer and therefore we find a step $s$ to minimize $f$. We require $s$ satisfies the constraint,

$$0 = c_1(x + s) \approx c_1(x) + \nabla c_1(x)^T s = \nabla c_1(x)^T s.$$  

To produce a decrease in $f$, we require that

$$0 > f(x + s) - f(x) \approx \nabla f(x)^T s.$$  

If, on the other hand, there is no such step $s$, then $x$ is likely to be a local minimizer. This can happen if $\nabla f(x)$ and $\nabla c_1(x)$ are parallel.

We introduce the Lagrangian function,

$$\mathcal{L}(x, \lambda_1) = f(x) - \lambda_1 c_1(x),$$

and the above condition can be stated that

$$\nabla_x \mathcal{L}(x^*, \lambda_1^*) = 0.$$  

This condition is necessary, but clearly not sufficient since $x = (1, 1)^T$ with $\lambda_1 = 1/2$ satisfies the condition as well, which is a maximizer. Also, the sign of $\lambda_1^*$ can be changed, by using $c_1(x) = 2 - x_1^2 - x_2^2 = 0$ instead (then $\lambda_1^* = 1/2$).

References