Lecture 11: CS395T Numerical Optimization for Graphics and AI — Conjugate Gradient Methods (Nonlinear)

Qixing Huang
The University of Texas at Austin
huangqx@cs.utexas.edu

1 Disclaimer

This note is adapted from


2 Introduction

In this section, we discuss nonlinear variants of the conjugate gradient, which have proved to be quite successful in practice.

2.1 Fletcher-Reeves method

The FR method (denoted as CG-FR) is based on a simple modification of the linear version of CG:

- Given $x_0$;
- Evaluate $f_0 = f(x_0), \nabla f_0 = \nabla f(x_0)$;
- Set $p_0 \leftarrow -\nabla f(x_0), k \leftarrow 0$;
- while $\nabla f_k \neq 0$
  - Compute $\alpha_k$ and set $x_{k+1} = x_k + \alpha_k p_k$;
  - Evaluate $\nabla f_{k+1}$;
  - $\beta_{k+1}^{FR} \leftarrow \frac{\nabla f_{k+1}^T \nabla f_{k+1}}{\nabla f_k^T \nabla f_k}$;
  - $p_{k+1} \leftarrow -\nabla f_{k+1} + \beta_{k+1}^{FR} p_k$;
  - $k \leftarrow k + 1$;
- end (while)
Note that in CG-FR, line search is used instead of the explicit formula for $\alpha_k$ in the linear case. So to make a global convergence argument, we have to be careful about the step-size $\alpha_k$. In fact, the angle between the search direction $p_k$ of the gradient $\nabla f_k$ may even be bigger than $90^\circ$.

In fact, we have

$$\nabla f_k^T p_k = -\|\nabla f_k\|^2 + \beta_k^{FR} \nabla f_k^T p_{k-1}.$$ If the line search is exact, so that $\alpha_{k-1}$ is, we have $\nabla f_k^T p_{k-1} = 0$. In this case, we have $\nabla f_k^T p_k < 0$, so that $p_k$ is indeed a descent direction. If the line search is inexact, then $\beta_k^{FR} \nabla f_k^T p_{k-1} > \|\nabla f_k\|^2$, then $p_k$ may not be a descent direction. Fortunately, we can avoid this situation by requiring the step length $\alpha_k$ to satisfy the strong Wolfe conditions, which we restate here:

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

(1)

$$|\nabla f(x_k + \alpha_k p_k)^T p_k| \leq -c_2 \nabla f_k^T p_k,$$

(2)

where $0 < c_1 < c_2 < \frac{1}{2}$. We will show that (2) ensures $p_k$ is a descent direction.

**Lemma 2.1.** Suppose that the algorithm is implemented with a step length $\alpha_k$ that satisfies the strong Wolfe conditions (4) with $0 < c_2 < \frac{1}{2}$. Then the method generates descent directions $p_k$ that satisfy the following inequalities:

$$-\frac{1}{1 - c_2} \leq \frac{\nabla f_k^T p_k}{\|\nabla f_k\|^2} \leq \frac{2c_2 - 1}{1 - c_2},$$

for all $k = 0, 1, \ldots$

(3)

**Proof.** We prove this by induction. When $k = 0$, (3) is obvious, since

$$\frac{\nabla f_k^T p_k}{\|\nabla f_k\|^2} = -1.$$

We prove (3) by induction. Suppose it is true for all integers that are small than $k$, now consider

$$\frac{\nabla f_{k+1}^T p_{k+1}}{\|\nabla f_{k+1}\|^2} = \frac{\nabla f_{k+1}^T (-\nabla f_{k+1} + \beta_{k+1}^{FR} p_k)}{\|\nabla f_{k+1}\|^2} = -1 + \frac{\beta_{k+1}^{FR} \nabla f_{k+1}^T p_k}{\|\nabla f_{k+1}\|^2}$$

$$= -1 + \frac{\nabla f_{k+1}^T p_k}{\|\nabla f_{k+1}\|^2}.$$ Since

$$|f_{k+1}^T p_k| \leq c_2 \|f_k^T p_k\|,$$

and

$$-\frac{1}{1 - c_2} \leq \frac{\nabla f_k^T p_k}{\|\nabla f_k\|^2} \leq \frac{2c_2 - 1}{1 - c_2}.$$ It follows that

$$-\frac{1}{1 - c_2} \leq \frac{\nabla f_k^T p_k}{\|\nabla f_k\|^2} \leq -1 + c_2 \frac{1}{1 - c_2} = \frac{2c_2 - 1}{1 - c_2}.$$ 

Remark 2.1. The Lemma above can also be used to explain a weakness of the CG-FR method. We will argue that if the method generates a bad direction and a tiny step, then the next direction and next step are also likely to be poor. Let $\theta_k$ be the angle between $p_k$ and the steepest descent direction $-\nabla f_k$, defined by

$$\cos(\theta_k) = -\frac{\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|}.$$ Suppose that $p_k$ is a poor search direction, in the sense that it makes an angle of nearly $90^\circ$ with $-\nabla f_k$, that is, $\cos(\theta_k) \approx 0$. Note that

$$\frac{1 - 2c_2}{1 - c_2} \|\nabla f_k\| \leq \cos(\theta_k) \leq \frac{1}{1 - c_2} \|\nabla f_k\| \|p_k\|,$$

for all $k = 0, 1, \ldots$
From these inequalities, we deduce that \( \cos(\theta_k) \approx 0 \) if and only if
\[
\|\nabla f_k\| \ll \|p_k\|.
\]
Since \( p_k \) is almost orthogonal to the gradient, it is likely that the step from \( x_k \) to \( x_{k+1} \) is tiny, that is, \( x_{k+1} \approx x_k \). If so, we have \( \nabla f_{k+1} \approx \nabla f_k \), and therefore
\[
\beta_{k+1} = 1,
\]
by the definition of \( \beta_{k+1} \). Note that \( p_{k+1} = -\nabla f_{k+1} + \beta_{k+1}p_k \), \( \nabla f_{k+1} \approx \nabla f_k \) and \( \|\nabla f_k\| \ll \|p_k\| \), we conclude that
\[
p_{k+1} \approx p_k.
\]
In other words, it is better to restart CG-FR when the angle between \( p_k \) and \( \nabla f_k \) becomes close to \( 90^\circ \).

### 2.2 Global Convergence

For the purposes of this section, we make the following (non-restrictive) assumptions on the objective function.

1. The levelset \( \mathcal{L} := \{x|f(x) \leq f(x_0)\} \) is bounded;
2. In some open neighborhood \( \mathcal{N} \) of \( \mathcal{L} \), the objective function \( f \) is Lipschitz continuously differentiable.

Now comes to the global convergence of CG-FR.

**Theorem 2.1.** Suppose that assumptions hold, and that CG-FR is implemented with a line search that satisfies the strong Wolfe conditions, with \( 0 < c_1 < c_2 < \frac{1}{2} \). Then
\[
\liminf_{k \to \infty} \|\nabla f_k\| = 0.
\]

**Proof.** The proof is by contradiction. Suppose there exists a \( \gamma > 0 \) such that
\[
\|\nabla f_k\| \geq \gamma,
\]
for all \( k \) sufficiently large.

First of all, the strong Wolfe condition implies that
\[
\sum_{k=0}^{\infty} \cos^2(\theta_k)\|\nabla f_k\|^2 < \infty.
\]
Note that
\[
\cos^2(\theta_k) = \left( \frac{\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|} \right)^2 = \left( \frac{\nabla f_k^T p_k}{\|\nabla f_k\|} \right)^2 \left( \frac{\|\nabla f_k\|}{\|p_k\|} \right)^2 \geq \left( \frac{1 - 2c_2}{1 - c_2} \right)^2 \left( \frac{\|\nabla f_k\|}{\|p_k\|} \right)^2.
\]
It turns out
\[
\sum_{k=0}^{\infty} \frac{\|\nabla f_k\|^4}{\|p_k\|^2} < \infty.
\]
Since \( \|\nabla f_k\| \geq \gamma \), it follows that
\[
\sum_{k=0}^{\infty} \frac{1}{\|p_k\|^2} < \infty.
\]
Now we derive an upper bound on \( \|p_k\| \). First of all,
\[
\|p_k\|^2 = \| - \nabla f_k + \beta_k p_{k-1} \|^2 \leq \|\nabla f_k\|^2 + 2\beta_k \|\nabla f_k^T p_{k-1}\| + \beta_k^2 \|p_{k-1}\|^2.
\]
Using the Wolfe condition, we have
\[
|\nabla f_k^T p_{k-1}| \leq c_2 |\nabla f_{k-1}^T p_{k-1}| \leq \frac{c_2}{1 - c_2} \|\nabla f_{k-1}\|^2.
\]

It follows that
\[
\|p_k\|^2 \leq \|\nabla f_k\|^2 + \frac{2c_2}{1 - c_2} \beta_k \|\nabla f_{k-1}\|^2 + \beta_k^2 \|p_{k-1}\|^2
\leq \frac{1 + c_2}{1 - c_2} \|\nabla f_k\|^2 + \beta_k^2 \|p_{k-1}\|^2.
\]

Applying the recursion, we have
\[
\|p_k\|^2 \leq \frac{1 + c_2}{1 - c_2} \|\nabla f_k\|^2 + \frac{\|\nabla f_k\|^4}{\|\nabla f_{k-1}\|^4} \|p_{k-1}\|^2
\leq \frac{1 + c_2}{1 - c_2} \|\nabla f_k\|^2 + \frac{\|\nabla f_k\|^4}{\|\nabla f_{k-1}\|^4} \|p_{k-2}\|^2
\leq \frac{1 + c_2}{1 - c_2} \|\nabla f_k\|^2 \sum_{j=0}^{k} \frac{1}{\|\nabla f_j\|^2}
\leq \frac{1 + c_2}{1 - c_2} (k + 1) \frac{\pi^4}{7^2}.
\]

This means
\[
\sum_{k=0}^{\infty} \frac{1}{\|p_k\|^2} = O(\sum_{k=1}^{\infty} \frac{1}{k}) = \infty,
\]
leading to a contradiction.

\[\Box\]

**Remark 2.2.** In general, if we can show that there exist constants $c_4, c_5 > 0$ such that
\[
cos(\theta_k) \geq c_4 \frac{\|\nabla f_k\|}{\|p_k\|}, \quad \frac{\|\nabla f_k\|}{\|p_k\|} \geq c_5 > 0, \quad k = 1, 2, \ldots,
\]
then
\[
\lim_{k \to \infty} \|\nabla f_k\| = 0.
\]