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

- Section 5 of *Numerical Optimization* by Jorge Nocedal and Stephen J. Wright. Springer series in operations research and financial engineering. Springer, New York, NY, 2. ed. edition, (2006)

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 $\boldsymbol{x}_0$;
- Evaluate $f_0 = f(\boldsymbol{x}_0)$, $\nabla f_0 = \nabla f(\boldsymbol{x}_0)$;
- Set $\boldsymbol{p}_0 \leftarrow -\nabla f(\boldsymbol{x}_0)$, $k \leftarrow 0$;
- while $\nabla f_k \neq 0$
- Compute α_k and set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$;
- Evaluate ∇f_{k+1};
- $\beta_{k+1}^{FR} \leftarrow \frac{\nabla f_{k+1}^T \nabla f_{k+1}}{\nabla f_k^T \nabla f_k}$;
- $\boldsymbol{p}_{k+1} \leftarrow -\nabla f_{k+1} + \beta_{k+1}^{FR} \boldsymbol{p}_k$;
- $k \leftarrow k+1$;
- end (while)

Note that in CG-FR, line search is used instead of the explicit formula for α_k in the linear case. So to make a global convergence argument, we have to be careful about the step-size α_k. In fact, the angle between the search direction $\boldsymbol{p}_k$ of the gradient ∇f_k may even be bigger than $90°$.

In fact, we have

$$\nabla f_k^T \boldsymbol{p}_k = -\|\nabla f_k\|^2 + \beta_k^{FR} \nabla f_k^T \boldsymbol{p}_{k-1}.$$

If the line search is exact, so that α_{k-1} is a local minimizer of f along the direction $\boldsymbol{p}_{k-1}$, we have $\nabla f_k^T \boldsymbol{p}_{k-1} = 0$. In this case, we have $\nabla f_k^T \boldsymbol{p}_k < 0$, so that $\boldsymbol{p}_k$ is indeed a descent direction. If the line search is inexact, then $\beta_k^{FR} \nabla f_k^T \boldsymbol{p}_{k-1} > \|\nabla f_k\|^2$, then $\boldsymbol{p}_k$ may not be a descent direction. Fortunately, we can avoid this situation by requiring the step length α_k to satisfy the strong Wolfe conditions, which we restate here:

$$f(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) \le f(\boldsymbol{x}_k) + c_1 \alpha_k \nabla f_k^T \boldsymbol{p}_k, \tag{1}$$

$$|\nabla f(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k)^T \boldsymbol{p}_k| \le -c_2 \nabla f_k^T \boldsymbol{p}_k, \tag{2}$$

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

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

$$-\frac{1}{1-c_2} \le \frac{\nabla f_k^T \boldsymbol{p}_k}{\|\nabla f_k\|^2} \le \frac{2c_2 - 1}{1 - c_2}, \qquad \text{for all } k = 0, 1, \ldots \tag{3}$$

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

$$\frac{\nabla f_k^T \boldsymbol{p}_k}{\|\nabla f_k\|^2} = -1.$$

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

$$\begin{aligned}\frac{\nabla f_{k+1}^T \boldsymbol{p}_{k+1}}{\|\nabla f_{k+1}\|^2} &= \frac{\nabla f_{k+1}^T(-\nabla f_{k+1} + \beta_{k+1}^{FR} \boldsymbol{p}_k)}{\|\nabla f_{k+1}\|^2} = -1 + \beta_{k+1}^{FR} \frac{\nabla f_{k+1}^T \boldsymbol{p}_k}{\|\nabla f_{k+1}\|^2} \\ &= -1 + \frac{\nabla f_{k+1}^T \boldsymbol{p}_k}{\|\nabla f_k\|^2}.\end{aligned}$$

Since

$$|f_{k+1}^T \boldsymbol{p}_k| \le c_2 |f_k^T \boldsymbol{p}_k|,$$

and

$$-\frac{1}{1-c_2} \le \frac{\nabla f_k^T \boldsymbol{p}_k}{\|\nabla f_k\|^2} \le \frac{2c_2 - 1}{1 - c_2}.$$

It follows that

$$-\frac{1}{1-c_2} \le \frac{\nabla f_k^T \boldsymbol{p}_k}{\|\nabla f_k\|^2} \le -1 + c_2 \frac{1}{1-c_2} = \frac{2c_2 - 1}{1 - c_2}.$$

□

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 θ_k be the angle between $\boldsymbol{p}_k$ and the steepest descent direction $-\nabla f_k$, defined by*

$$\cos(\theta_k) = -\frac{\nabla f_k^T \boldsymbol{p}_k}{\|\nabla f_k\| \|\boldsymbol{p}_k\|}.$$

Suppose that $\boldsymbol{p}_k$ is a poor search direction, in the sense that it makes an angle of nearly $90°$ with $-\nabla f_k$, that is, $\cos(\theta_k) \approx 0$. Note that

$$\frac{1-2c_2}{1-c_2} \frac{\|\nabla f_k\|}{\|\boldsymbol{p}_k\|} \le \cos(\theta_k) \le \frac{1}{1-c_2} \frac{\|\nabla f_k\|}{\|\boldsymbol{p}_k\|}, \qquad \text{for all } k = 0, 1, \ldots$$

From these inequalities, we deduce that $\cos(\theta_k) \approx 0$ *if and only if*

$$\|\nabla f_k\| << \|\boldsymbol{p}_k\|.$$

Since $\boldsymbol{p}_k$ *is almost orthogonal to the gradient, it is likely that the step from* $\boldsymbol{x}_k$ *to* $\boldsymbol{x}_{k+1}$ *is tiny, that is,* $\boldsymbol{x}_{k+1} \approx \boldsymbol{x}_k$*. If so, we have* $\nabla f_{k+1} \approx \nabla f_k$*, and therefore*

$$\beta_{k+1} = 1,$$

by the definition of β_{k+1}*. Note that* $\boldsymbol{p}_{k+1} = -\nabla f_{k+1} + \beta_{k+1}\boldsymbol{p}_k$, $\nabla f_{k+1} \approx \nabla f_k$ *and* $\|\nabla f_k\| << \|\boldsymbol{p}_k\|$*, we conclude that*

$$\boldsymbol{p}_{k+1} \approx \boldsymbol{p}_k.$$

In other words, it is better to restart CG-FR when the angle between $\boldsymbol{p}_k$ *and* ∇f_k *becomes close to* $90°$*.*

2.2 Global Convergence

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

1. The levelset $\mathcal{L} := \{\boldsymbol{x} | f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$ 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

$$\cos^2(\theta_k) = (\frac{\nabla f_k^T \boldsymbol{p}_k}{\|\nabla f_k\|\|\boldsymbol{p}_k\|})^2 = (\frac{\nabla f_k^T \boldsymbol{p}_k}{\|\nabla f_k\|^2})^2 \left(\frac{\|\nabla f_k\|}{\|\boldsymbol{p}_k\|}\right)^2 \geq (\frac{1-2c_2}{1-c_2})^2 \left(\frac{\|\nabla f_k\|}{\|\boldsymbol{p}_k\|}\right)^2.$$

It turns out

$$\sum_{k=0}^{\infty} \frac{\|\nabla f_k\|^4}{\|\boldsymbol{p}_k\|^2} < \infty.$$

Since $\|\nabla f_k\| \geq \gamma$, it follows that

$$\sum_{k=0}^{\infty} \frac{1}{\|\boldsymbol{p}_k\|^2} < \infty.$$

Now we derive an upper bound on $\|\boldsymbol{p}_k\|$. First of all,

$$\|\boldsymbol{p}_k\|^2 = \| - \nabla f_k + \beta_k \boldsymbol{p}_{k-1}\|^2 \leq \|\nabla f_k\|^2 + 2\beta_k |\nabla f_k^T \boldsymbol{p}_{k-1}| + \beta_k^2 \|\boldsymbol{p}_{k-1}\|^2.$$

Using the Wolfe condition, we have

$$|\nabla f_k^T \boldsymbol{p}_{k-1}| \leq c_2 |\nabla f_{k-1}^T \boldsymbol{p}_{k-1}| \leq \frac{c_2}{1-c_2} \|\nabla f_{k-1}\|^2.$$

It follows that

$$\begin{aligned}
\|\boldsymbol{p}_k\|^2 &\leq \|\nabla f_k\|^2 + \frac{2c_2}{1-c_2} \beta_k \|\nabla f_{k-1}\|^2 + \beta_k^2 \|\boldsymbol{p}_{k-1}\|^2 \\
&\leq \frac{1+c_2}{1-c_2} \|\nabla f_k\|^2 + \beta_k^2 \|\boldsymbol{p}_{k-1}\|^2.
\end{aligned}$$

Applying the recursion, we have

$$\begin{aligned}
\|\boldsymbol{p}_k\|^2 &\leq \frac{1+c_2}{1-c_2} \|\nabla f_k\|^2 + \frac{\|\nabla f_k\|^4}{\|\nabla f_{k-1}\|^4} \|\boldsymbol{p}_{k-1}\|^2 \\
&\leq \frac{1+c_2}{1-c_2} (\|\nabla f_k\|^2 + \frac{\|\nabla f_k\|^4}{\|\nabla f_{k-1}\|^2}) + \frac{\|\nabla f_k\|^4}{\|\nabla f_{k-2}\|^4} \|\boldsymbol{p}_{k-2}\|^2 \\
&\leq \frac{1+c_2}{1-c_2} \|\nabla f_k\|^4 \sum_{j=0}^{k} \frac{1}{\|\nabla f_j\|^2} \\
&\leq \frac{1+c_2}{1-c_2} (k+1) \frac{\overline{\gamma}^4}{\gamma^2}.
\end{aligned}$$

This means

$$\sum_{k=0}^{\infty} \frac{1}{\|\boldsymbol{p}_k\|^2} = O(\sum_{k=1}^{\infty} \frac{1}{k}) = \infty,$$

leading to a contradiction. □

Remark 2.2. *In general, if we can show that there exist constants $c_4, c_5 > 0$ such that*

$$\cos(\theta_k) \geq c_4 \frac{\|\nabla f_k\|}{\|\boldsymbol{p}_k\|}, \qquad \frac{\|\nabla f_k\|}{\|\boldsymbol{p}_k\|} \geq c_5 > 0, \qquad k = 1, 2, \ldots,$$

then

$$\lim_{k \to \infty} \|\nabla f_k\| = 0.$$