Numerical Optimization

Lecture Notes #12
Conjugate Gradient Methods — Nonlinear CG

Fall 2022

Outline

1. Nonlinear Conjugate Gradient Methods
   - New Ideas... Fletcher-Reeves, etc...
   - Practical Considerations
   - Convergence

2. Projects
   - Separate Handouts, etc...

Nonlinear Conjugate Gradient Methods

We now turn our attention to making the CG methods useful for optimization problems (the non-linear situation).

The **Fletcher-Reeves** (`CG-FR`, published in 1964) extension requires two modifications to the `CG` algorithm:

1: The computation of the step length α_k is replaced by a line-search which minimizes the non-linear objective $f(\cdot)$ along the search direction $\bar{\mathbf{p}}_k$.

2: The instances of the residual $\bar{\mathbf{r}}$ (which are just $\nabla\Phi(\cdot)$ for the quadratic objective in standard `CG`) are replaced by the gradient of the non-linear objective $\nabla f(\cdot)$.

Fletcher, R., and Reeves, C. M. "Function minimization by conjugate gradients." *The computer journal*, 7, no. 2 (1964), 149-154.

The Fletcher-Reeves FR-CG Algorithm

Algorithm: Fletcher-Reeves

Given $\bar{\mathbf{x}}_0$:

Evaluate $f_0 = f(\bar{\mathbf{x}}_0)$, $\nabla f_0 = \nabla f(\bar{\mathbf{x}}_0)$.

Set $\bar{\mathbf{p}}_0 = -\nabla f_0$, $k = 0$

while ($\|\nabla f_k\| > 0$, ...)

$$
\begin{aligned}
\alpha_k &= \texttt{linesearch(...)} \\
\bar{\mathbf{x}}_{k+1} &= \bar{\mathbf{x}}_k + \alpha_k \bar{\mathbf{p}}_k \\
\nabla f_{k+1} &= \texttt{Evaluate } \nabla f(\bar{\mathbf{x}}_{k+1}) \\
\beta_{k+1}^{\mathrm{FR}} &= \frac{\nabla f_{k+1}^T \nabla f_{k+1}}{\nabla f_k^T \nabla f_k}, \qquad \texttt{Save numerator!} \\
\bar{\mathbf{p}}_{k+1} &= -\nabla f_{k+1} + \beta_{k+1}^{\mathrm{FR}} \bar{\mathbf{p}}_k \\
k &= k+1
\end{aligned}
$$

end-while

Comments: The Fletcher-Reeves FR−CG Algorithm

Sanity check: If $f(\bar{\mathbf{x}})$ is a strongly convex quadratic, and α_k the exact minimizer, then FR−CG reduces to linear CG.

Each iteration requires evaluation of the objective function (for the line-search), and the gradient of the objective. — No Hessian evaluation, nor matrix operations are required. **Good** for large non-linear optimization problems.

If we require that α_k satisfies the **strong Wolfe conditions**

$$\begin{array}{rcl} f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k) & \leq & f(\bar{\mathbf{x}}_k) + c_1 \alpha \bar{\mathbf{p}}_k^T \nabla f_k \\ |\bar{\mathbf{p}}_k^T \nabla f(\bar{\mathbf{x}}_k + \alpha \bar{\mathbf{p}}_k)| & \leq & c_2 |\bar{\mathbf{p}}_k^T \nabla f_k| \end{array}$$

where $0 < c_1 < c_2 < \frac{1}{2}$, then FR−CG converges globally.

Variants: The Polak-Ribière (PR-CG) Method

The following modification to FR-CG was suggested by Polak-Ribière

$$\beta_{k+1}^{\text{FR}} = \frac{\nabla f_{k+1}^T \nabla f_{k+1}}{\nabla f_k^T \nabla f_k} \quad \to \quad \boldsymbol{\beta_{k+1}^{\text{PR}} = \frac{\nabla f_{k+1}^T (\nabla f_{k+1} - \nabla f_k)}{\nabla f_k^T \nabla f_k}}$$

when f is a strongly convex quadratic, and the line search is exact, the gradients are orthogonal and $\beta_{k+1}^{\text{FR}} = \beta_{k+1}^{\text{PR}}$.

On general non-linear objectives, an inexact line-searches PR-CG tends to be **more robust** and **more efficient** than FR-CG.

Polak, Elijah, and Gerard Ribiere. "Note sur la convergence de méthodes de directions conjugués." *Revue française d'informatique et de recherche opérationnelle.* Série rouge 3, no. 16 (1969): 35-43.

Variants: The Polak-Ribière (PR−CG) Method

One problem: The strong Wolfe conditions **do not** guarantee that $\bar{\mathbf{p}}_k$ is always descent direction for PR−CG. In order to fix this, β is defined to be

$$\beta^{+}_{k+1} = \max(\beta^{\mathsf{PR}}_{k+1}, 0)$$

the resulting algorithm is known as PR+.

There are a number of other choices for β in the literature, but they are not (in general) more efficient than Polak-Ribière PR−CG/PR+.

Practical Considerations

If the line-search uses quadratic (or cubic) interpolation along the search direction $\bar{\mathbf{p}}_k$, then if/when $f(\cdot)$ is a strictly convex quadratic, the step lengths α_k will be the exact 1D-minimizers $\Rightarrow$ the non-linear algorithm reduces to linear `CG`. [THIS IS HIGHLY DESIRABLE!]

Restarting: `CG` gets its favorable convergence properties from the conjugacy of the search directions **near** the optimum. If we start "far" from the optimum, the algorithm does not necessarily gain anything from maintaining this conjugacy.

Therefore, we should periodically restart the algorithm, by setting $\beta = 0$ (*i.e.* taking a steepest-descent step).

The n-step convergence is only guaranteed when we start with a steepest-descent step, and the model is quadratic. Hence a restart close to $\bar{\mathbf{x}}^*$ will (approximately) guarantee n-step convergence.

Practical Considerations: Restarting Conditions

Restarting conditions: The most common condition is based on the fact that for the strictly quadratic objective, the residuals are orthogonal. Hence, when two consecutive residuals are "far" from orthogonal

$$\frac{\nabla f_k^T \nabla f_{k-1}}{\nabla f_k^T \nabla f_k} \geq \nu \sim 0.1$$

a restart is triggered.

The formula

$$\beta_{k+1}^{+} = \max(\beta_{k+1}^{\mathrm{PR}}, 0)$$

in PR+ can be viewed as a restart-condition. This is not practical since these "restarts" are very infrequent — in practice $\beta_{k+1}^{\mathrm{PR}}$ is positive most of the time.

Nonlinear CG: Global Convergence

Linear CG: Global convergence properties well understood, and optimal.

Nonlinear CG: Convergence properties not so well understood, except in special cases. The behavior is sometimes surprising and bizarre!

We look at some results, under the following non-restrictive assumptions

Assumptions:

(i) The level set $\mathcal{L} = \{\bar{\mathbf{x}} \in \mathbb{R}^n : f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}}_0)\}$ is bounded.

(ii) In some neighborhood $\mathcal{N}$ of $\mathcal{L}$, the objective function f is Lipschitz continuously differentiable, *i.e.* there exists a constant $L > 0$ such that

$$\|\nabla f(\bar{\mathbf{x}}) - \nabla f(\bar{\mathbf{y}})\| \leq L\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\|, \quad \forall \bar{\mathbf{x}}, \bar{\mathbf{y}} \in \mathcal{N}$$

Global Convergence: FR-CG

Theorem

Suppose that the assumptions hold, and that FR-CG is implemented with a line search which satisfies the strong Wolfe conditions, with $0 < c_1 < c_2 < \frac{1}{2}$*. Then*

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

This does not say that the limit of the sequence of gradients $\{\nabla f_k\}$ is zero; but it does tell us that at least the sequence is not bounded away from zero.

If, however, we restart the algorithm every n steps, we get n-step quadratic convergence:

$$\|\bar{\mathbf{x}}_{k+n} - \bar{\mathbf{x}}^*\| = \mathcal{O}(\|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}^*\|^2).$$

Global Convergence: PR−CG

In practice PR−CG performs better than FR−CG, but we cannot prove a theorem like the one for FR−CG on the previous slide.

The following surprising result **can** be shown:

Theorem

Consider the Polak-Ribiere PR−CG method with an ideal line search. There exists a twice continuously differentiable objective function $f : \mathbb{R}^3 \to \mathbb{R}$ *and a starting point* $\bar{\mathbf{x}}_0 \in \mathbb{R}^3$ *such that the sequence of gradients* $\{\|\nabla f_k\|\}$ *is* **bounded away from zero**.

The modification (PR+)

$$\beta_{k+1}^{+} = \max(\beta_{k+1}^{\text{PR}}, 0)$$

fixes this strange behavior, and it is possible to show global convergence for PR+.

Separate Handouts, etc...

INITECH

T.P.S REPORT

COVER SHEET

Prepared By: ____________ Date: ________

System: ________ Program Language: ____ Platform: ____ OS: ____

Unit Code: ________ Customer: ____________________

Unit Code Tested: ______________________________

Due Date: ________ Approved By: ____________________

Test Date: ________ Tested By: ____________________

Total Run Time: ________ Total Error Count: ________________

Error Reference: ______________________________

Errors Logged: ________ Log Location: ____________________

Passed: ________ Moved to Production: ________________

Comments: ______________________________

____

____

____

CONFIDENTIAL