ACM 113 Introduction to Optimization - Problem Set 2

Ling Li, ling@cs.caltech.edu

April 24, 2001

2.1 Let $\lambda_{\min}\left(B_k^{-1}\right)$ denote the smallest eigenvalue of B_k^{-1}. Since B_k is symmetric and positive-definite, we know B_k^{-1} is also symmetric and positive-definite. Thus*

$$-\nabla f_k^T p_k = \nabla f_k B_k^{-1} \nabla f_k \geq \lambda_{\min}\left(B_k^{-1}\right) \|\nabla f_k\|^2 .$$

Also from $\|p_k\| = \left\|B_k^{-1}\nabla f_k\right\| \leq \left\|B_k^{-1}\right\| \|\nabla f_k\|$, and

$$\lambda_{\min}\left(B_k^{-1}\right) = \frac{1}{\lambda_{\max}(B_k)} = \frac{1}{\|B_k\|},$$

we get

$$\cos\theta_k = \frac{-\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|} \geq \frac{\lambda_{\min}\left(B_k^{-1}\right)\|\nabla f_k\|^2}{\left\|B_k^{-1}\right\| \|\nabla f_k\|^2} = \frac{1}{\|B_k\| \left\|B_k^{-1}\right\|} \geq \frac{1}{M}.$$

2.2 From $p_k^N = -\nabla^2 f_k^{-1}\nabla f_k = -\nabla^2 f_k^{-1} B_k B_k^{-1}\nabla f_k = \nabla^2 f_k^{-1} B_k p_k$, we have

$$p_k - p_k^N = -\nabla^2 f_k^{-1}\left(B_k - \nabla^2 f_k\right) p_k. \tag{1}$$

Thus

$$\left\|p_k - p_k^N\right\| = \left\|\nabla^2 f_k^{-1}\left(B_k - \nabla^2 f_k\right) p_k\right\| \leq \left\|\nabla^2 f_k^{-1}\right\| \left\|\left(B_k - \nabla^2 f_k\right) p_k\right\|. \tag{2}$$

Also from (1),

$$\left\|\left(B_k - \nabla^2 f_k\right) p_k\right\| = \left\|\nabla^2 f_k \left(p_k - p_k^N\right)\right\| \leq \left\|\nabla^2 f_k\right\| \left\|p_k - p_k^N\right\|. \tag{3}$$

*An $n \times n$ symmetric positive-definite matrix A with eigenvalues λ_i can be written as $A = U^T \Lambda U$, where $U^T = U^{-1}$, U is an orthogonal (normalized) basis consisting of eigenvectors of A, and

$$\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Let $\lambda_{\min}$ be the smallest eigenvalue of A. For any vector x, we have

$$x^T A x = (Ux)^T \Lambda (Ux) = \sum_{i=1}^n \lambda_i (Ux)_i^2 \geq \lambda_{\min} \sum_{i=1}^n (Ux)_i^2 = \lambda_{\min} (Ux)^T (Ux) = \lambda_{\min} x^T x,$$

where $(Ux)_i$ denotes the i^{th} component of Ux.

Let $\lambda_{\min}^{(k)}$ and $\lambda_{\max}^{(k)}$ denote the smallest and largest eigenvalues of $\nabla^2 f_k$, (thus $\left(\lambda_{\min}^{(k)}\right)^{-1}$ is the largest eigenvalue of $\nabla^2 f_k^{-1}$) and $\lambda_{\min}^*$ and $\lambda_{\max}^*$ denote the smallest and largest eigenvalues of $\nabla^2 f_*$. Since $f \in C^2$ and $x_k \to x^*$, we have for k sufficient large,

$$\lambda_{\min}^{(k)} > \frac{1}{2}\lambda_{\min}^* > 0, \quad \lambda_{\max}^{(k)} < 2\lambda_{\max}^*.$$

Thus

$$\left\|\nabla^2 f_k\right\| < 2\lambda_{\max}^*, \quad \left\|\nabla^2 f_k^{-1}\right\| < 2\left(\lambda_{\min}^*\right)^{-1}. \tag{4}$$

With (2), (3), and (4), we can assert

$$\lim_{k\to\infty} \frac{\left\|\left(B_k - \nabla^2 f_k\right) p_k\right\|}{\|p_k\|} = 0$$

if and only if

$$\lim_{k\to\infty} \frac{\left\|p_k - p_k^N\right\|}{\|p_k\|} = 0.$$

2.3 If we use

$$\alpha_k = \frac{r_k^T r_k}{p_k^T A p_k}, \text{ and } \beta_{k+1} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$$

in the conjugate gradient method, then from [1, Theorem 12.1] (note that r_k here is just $-r_k$ in [1]), we have for $i > j$,

$$r_i^T r_j = 0, \quad r_i^T p_j = 0.$$

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

$$-r_k^T p_k = r_k^T r_k - \beta_k r_k^T p_{k-1} = r_k^T r_k. \tag{5}$$

So for α_k,

$$-\frac{r_k^T p_k}{p_k^T A p_k} = \frac{r_k^T r_k}{p_k^T A p_k}.$$

From $r_{k+1} - r_k = A(x_{k+1} - x_k) = \alpha_k A p_k$, together with (5), we get for β_{k+1},

$$\frac{r_{k+1}^T A p_k}{p_k^T A p_k} = \frac{r_{k+1}^T \alpha_k^{-1}\left(r_{k+1} - r_k\right)}{p_k^T \alpha_k^{-1}\left(r_{k+1} - r_k\right)} = \frac{r_{k+1}^T r_{k+1} - r_{k+1}^T r_k}{r_{k+1}^T p_k - r_k^T p_k} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}.$$

2.4 The sufficient decrease condition says

$$f(x_k + \alpha_k p_k) \le f_k + C_1 \alpha_k p_k^T \nabla f_k. \tag{6}$$

Last line search gives the descent in the objective function is $|f_k - f_{k-1}|$. If such a descent is also expected in this run of line search, we have

$$f_k - f_{k-1} \approx f(x_k + \alpha_k p_k) - f_k \le C_1 \alpha_k p_k^T \nabla f_k.$$

So approximately (notice that $p_k^T \nabla f_k < 0$)

$$\alpha_k \leq \frac{f_{k-1} - f_k}{C_1 p_k^T \nabla f_k}.$$

Thus one heuristic method is to calculate

$$\alpha_k^{(g)} = \frac{f_k - f_{k-1}}{C_1 p_k^T \nabla f_k}$$

and use

$$\alpha_k^{(0)} = \begin{cases} 5, & \alpha_k^{(g)} > 5; \\ 0.2, & \alpha_k^{(g)} < 0.2; \\ \alpha_k^{(g)}, & \text{otherwise.} \end{cases}$$

as the starting step for the line search.

Such guess may have two benefits: first, by estimating the range of α_k, we may reduce the number of iterations within the line search, thus reduce the number of function evaluations; second, this allows sometimes step larger than 1, which may speed the convergence.

For conjugate gradient method with periodical restart, the $\alpha_k^{(0)}$ is set to 1 when restart is invoked.

2.5 For the stopping condition $\|\nabla f(x_k)\| \leq \epsilon$, if the objective function $f(x)$ is changed by multiplying $f(x)$ by some constant M, the previously appropriate value of ϵ may be too strict (too small) for the new problem, since x_k is expected to have half the precision of $f(x_k)$ [1, Section 11.5].

On the other side, the stopping condition $\|\nabla f(x_k)\| \leq \epsilon |f(x_k)|$ may be inappropriate when the minimum of f is 0 or very near to 0.

However, the stopping criterion $\|\nabla f(x_k)\| \leq \epsilon (1 + |f(x_k)|)$ can cope with all the above difficulties. When $|f(x_k)|$ is large, it is like $\|\nabla f(x_k)\| \leq \epsilon |f(x_k)|$; and when $f(x_k)$ is near 0, it resembles $\|\nabla f(x_k)\| \leq \epsilon$.

2.6 Use $\epsilon = 10^{-8}$ in the stopping criterion $\|\nabla f(x_k)\| \leq \epsilon (1 + |f(x_k)|)$, and $C_1 = 0.35$ in the sufficient decrease condition (6). Here are the results of those 3 methods (in all cases, x_k converges to $(1, 1)^T$):

Start point x_0	Newton method	Conjugate gradient	Steepest descent
$(1.2, 1.2)^T$		$f_{52} = 5.3264 \times 10^{-17}$	$f_{17567} = 9.8419 \times 10^{-17}$
with $\alpha_k^{(0)} \equiv 1$	$f_8 = 1.0883 \times 10^{-25}$	$f_{63} = 1.6416 \times 10^{-17}$	$f_{1339} = 1.1200 \times 10^{-16}$
$(-1.2, 1)^T$		$f_{117} = 1.0306 \times 10^{-20}$	$f_{17463} = 9.6359 \times 10^{-17}$
with $\alpha_k^{(0)} \equiv 1$	$f_{21} = 7.6820 \times 10^{-24}$	p_4 is not descent	$f_{524} = 8.2287 \times 10^{-17}$

The results show that for Newton method and conjugate gradient method, it is easier to find the local minimum (which is also the global minimum) if starting from $(1.2, 1.2)^T$. However, for steepest descent, starting from $(-1.2, 1)^T$ seems to be easier.

The step length for Newton method is 1 for almost all the time, and it is sometimes 0.5 or even smaller. The step length for conjugate gradient is less than 0.008 for almost all the time. This is also the case for steepest descent.

In any case, Newton method surpasses other two methods. The second winner is conjugate gradient method. These correspond to the theoretical expectations, since Newton method is with quadratic rate and the other two are with linear rate, and the rate constant of conjugate gradient is less than that of steepest descent.

References

[1] Stephen G. Nash and Ariela Sofer. *Linear and Nonlinear Programming.* McGraw-Hill series in industrial engineering and management science. The McGraw-Hill Companies, Inc, New York, 1996.