Numerical Methods II

Convergence of Gradient and Conjugate Gradient Methods

Marcel Oliver

1 Notation

We apply the Gradient and the Conjugate Gradient methods to the problem of finding the minimum of the quadratic functional

$$\Phi(\boldsymbol{x}) = \tfrac{1}{2}\,\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b}\,, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite. Recall that both methods are *descent methods*. They construct a minimizing sequence

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\,\boldsymbol{d}_k \tag{2}$$

where $\boldsymbol{d}_k$ is a descent direction, i.e. $\boldsymbol{d}_k^T\boldsymbol{\nabla}\Phi(x_k) < 0$, so that

$$\Phi(\boldsymbol{x}_k + \alpha\,\boldsymbol{d}_k) < \Phi(\boldsymbol{x}_k) \tag{3}$$

for small positive values of α. For quadratic functionals (1), the value of α that minimizes Φ along the line through $\boldsymbol{x}_k$ in the direction $\boldsymbol{d}_k$ is easily found to be

$$\alpha = \frac{\boldsymbol{d}_k^T\boldsymbol{r}_k}{\boldsymbol{d}_k^T A\boldsymbol{d}_k}\,. \tag{4}$$

We denote the error in the k-th step by

$$\boldsymbol{e}_k = \boldsymbol{x} - \boldsymbol{x}_k\,, \tag{5}$$

$\boldsymbol{x}$ being the location of the true minimum, and define the *residual*

$$\boldsymbol{r}_k = A\boldsymbol{e}_k = b - A\boldsymbol{x}_k\,. \tag{6}$$

The last equality is true since the location of the minimum is the solution of $A\boldsymbol{x} = \boldsymbol{b}$. See the notes from Numerical Methods I for details.

Finally, we introduce the A-norm

$$\|\boldsymbol{x}\|_A^2 \equiv \boldsymbol{x}^T A\boldsymbol{x}\,. \tag{7}$$

2 Convergence of the Gradient Method

In the gradient method we always walk down the direction of steepest descent, i.e.

$$\boldsymbol{d}_k = -\boldsymbol{\nabla}\Phi(\boldsymbol{x}_k) = \boldsymbol{r}_k\,. \tag{8}$$

A direct computation of the error norm shows that

$$\begin{aligned}
\|\boldsymbol{e}_{k+1}\|_A^2 &= \|\boldsymbol{x} - (\boldsymbol{x}_k + \alpha_k\,\boldsymbol{d}_k)\|_A^2 \\
&= \|\boldsymbol{e}_k - \alpha_k\,\boldsymbol{d}_k\|_A^2 \\
&= \|\boldsymbol{e}_k\|_A^2 - 2\,\alpha_k\,\boldsymbol{e}_k^T A\boldsymbol{d}_k + \alpha_k^2\,\|\boldsymbol{d}_k\|_A^2 \\
&= \|\boldsymbol{e}_k\|_A^2 - 2\,\frac{\boldsymbol{r}_k^T\boldsymbol{r}_k}{\boldsymbol{r}_k^T A\boldsymbol{r}_k}\,\boldsymbol{e}_k^T A\boldsymbol{r}_k + \frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{(\boldsymbol{r}_k^T A\boldsymbol{r}_k)^2}\,\boldsymbol{r}_k^T A\boldsymbol{r}_k \\
&= \|\boldsymbol{e}_k\|_A^2\left(1 - \frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{\boldsymbol{r}_k^T A\boldsymbol{r}_k\,\boldsymbol{r}_k^T A^{-1}\boldsymbol{r}_k}\right).
\end{aligned} \tag{9}$$

To proceed further, we need the following result.

Lemma 1 (Kantorovich inequality)**.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite, and let $0 < \lambda_{\min} < \lambda_{\max}$ denote its smallest and largest eigenvalue, respectively. Then*

$$\min_{\boldsymbol{y}\neq 0}\frac{(\boldsymbol{y}^T\boldsymbol{y})^2}{\boldsymbol{y}^T A\boldsymbol{y}\,\boldsymbol{y}^T A^{-1}\boldsymbol{y}} = \frac{4\,\lambda_{\min}\,\lambda_{\max}}{(\lambda_{\min}+\lambda_{\max})^2}\,. \tag{10}$$

Applying the Kantorovich inequality to the right side of (9) and noting that

$$1 - \frac{4\,\lambda_{\min}\,\lambda_{\max}}{(\lambda_{\min}+\lambda_{\max})^2} = \frac{(\lambda_{\min}-\lambda_{\max})^2}{(\lambda_{\min}+\lambda_{\max})^2}\,, \tag{11}$$

we find that

$$\|\boldsymbol{e}_{k+1}\|_A \le \frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\,\|\boldsymbol{e}_k\|_A \equiv \frac{\kappa-1}{\kappa+1}\,\|\boldsymbol{e}_k\|_A\,, \tag{12}$$

where $\kappa = \lambda_{\max}/\lambda_{\min}$ is the so-called spectral condition number of A.

Proof of the Kantorovich inequality. The Kantorovich inequality (10) is invariant under rescaling of $\boldsymbol{y}$. It is therefore sufficient to prove it for arbitrary unit vectors. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. We can write

$$\boldsymbol{y} = \sum_{i=1}^{n} y_i\,\boldsymbol{v}_i \tag{13}$$

so that

$$\|\boldsymbol{y}\|_2^2 = \sum_{i=1}^{n} y_i^2 = 1\,, \tag{14}$$

$$\boldsymbol{y}^T A \boldsymbol{y} = \sum_{i=1}^{n} y_i^2\,\lambda_i\,, \tag{15}$$

$$\boldsymbol{y}^T A^{-1} \boldsymbol{y} = \sum_{i=1}^{n} y_i^2\,\lambda_i^{-1}\,. \tag{16}$$

Thus, proving the Kantorovich inequality reduces to solving the constraint optimization problem *find the maximum of*

$$g(\boldsymbol{y}) = \boldsymbol{y}^T A \boldsymbol{y}\, \boldsymbol{y}^T A^{-1} \boldsymbol{y} \tag{17}$$

under the constraint

$$h(\boldsymbol{y}) = \boldsymbol{y}^T \boldsymbol{y} - 1 = 0\,. \tag{18}$$

Such maximum must necessarily satisfy

$$\boldsymbol{\nabla} g(\boldsymbol{y}) = \mu\, \boldsymbol{\nabla} h(\boldsymbol{y})\,, \tag{19}$$

where μ is the Lagrange multiplier. By direct computation,

$$\frac{\partial g}{\partial y_i} = 2\,y_i\,\lambda_i\,\boldsymbol{y}^T A^{-1} \boldsymbol{y} + 2\,y_i\,\lambda_i^{-1}\,\boldsymbol{y}^T A \boldsymbol{y}\,, \tag{20}$$

$$\frac{\partial h}{\partial y_i} = 2\,y_i\,, \tag{21}$$

so that (19) reads

$$y_i\,\lambda_i\,\boldsymbol{y}^T A^{-1} \boldsymbol{y} + y_i\,\lambda_i^{-1}\,\boldsymbol{y}^T A \boldsymbol{y} = \mu\, y_i\,, \tag{22}$$

or

$$\left(\lambda_i^2\,\boldsymbol{y}^T A^{-1} \boldsymbol{y} - \lambda_i\,\mu + \boldsymbol{y}^T A \boldsymbol{y}\right) \frac{y_i}{\lambda_i} = 0\,. \tag{23}$$

The expression in parenthesis is a quadratic equation in λ_i, and can only be zero for at most two distinct eigenvalues. Therefore, there can be at most two non-zero components y_i and y_j. (If an eigenvalue occurs with multiplicity larger than one, $g(\boldsymbol{y})$ depends only on the norm of the projection of $\boldsymbol{y}$ onto the corresponding eigenspace, so the statement remains valid.) Dividing (22) by y_i and equating the expressions for indices i and j, we obtain

$$\lambda_i \left(\frac{y_i^2}{\lambda_i} + \frac{y_j^2}{\lambda_j}\right) + \frac{y_i^2\,\lambda_i + y_j^2\,\lambda_j}{\lambda_i} = \lambda_j \left(\frac{y_i^2}{\lambda_i} + \frac{y_j^2}{\lambda_j}\right) + \frac{y_i^2\,\lambda_i + y_j^2\,\lambda_j}{\lambda_j}\,, \tag{24}$$

or

$$(y_j^2 - y_i^2)\left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} - 2\right) = (y_j^2 - y_i^2)\,\frac{(\lambda_i - \lambda_j)^2}{\lambda_i \lambda_j}\,. \tag{25}$$

Therefore, $y_i^2 = y_j^2 = \frac{1}{2}$ unless $\lambda_i = \lambda_j$ and, by direct substitution into (17), candidates for the maximum value are

$$\frac{1}{4}\,(\lambda_i + \lambda_j)\left(\frac{1}{\lambda_i} + \frac{1}{\lambda_j}\right) = \frac{1}{4}\left(\frac{\lambda_i}{\lambda_j} + 1\right)\left(\frac{\lambda_j}{\lambda_i} + 1\right) = \frac{(\lambda_i + \lambda_j)^2}{4\,\lambda_i\,\lambda_j}\,. \tag{26}$$

The expression in the middle shows that it is increasing in the ratio λ_i/λ_j when $\lambda_i > \lambda_j$; we thus take $\lambda_i = \lambda_{\max}$ and $\lambda_j = \lambda_{\min}$ to complete the proof. □

3 Convergence of the CG Method

Recall that one of the key properties of the Conjugate Gradient method is that each new iterate is optimal with respect to *all* descent directions from the so-called Krylov subspace

$$\begin{aligned} V_k &= \mathrm{Span}\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{k-1}\} \\ &= \mathrm{Span}\{A^0\boldsymbol{r}_1, \ldots, A^{k-2}\boldsymbol{r}_1\}\,. \end{aligned} \tag{27}$$

For details, see the handout on the derivation of the CG method. According to (6), we can also write

$$V_k = \mathrm{Span}\{A^1\boldsymbol{e}_1, \ldots, A^{k-1}\boldsymbol{e}_1\}\,. \tag{28}$$

Recall also the recursion for the residual,

$$\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \alpha_k\,A\boldsymbol{d}_k \equiv \boldsymbol{r}_k + A\boldsymbol{w}_{k+1} \tag{29}$$

where $\boldsymbol{w}_{k+1}$ is some vector from V_{k+1}. By definition of $\boldsymbol{e}_k$, we can also write

$$\boldsymbol{e}_{k+1} = \boldsymbol{e}_k + \boldsymbol{w}_{k+1}\,. \tag{30}$$

We conclude inductively that $\boldsymbol{e}_k - \boldsymbol{e}_1 \in V_k$ for $k = 2, 3, \ldots$, and we can write

$$\begin{aligned} \boldsymbol{e}_k &= \boldsymbol{e}_1 + \sum_{i=1}^{k-1} \gamma_i\,A^i\,\boldsymbol{e}_1 \\ &\equiv \phi_k(A)\,\boldsymbol{e}_1 \end{aligned} \tag{31}$$

where

$$\phi_k(x) = \sum_{i=0}^{k-1} \gamma_i\,x^i \tag{32}$$

is some polynomial of degree $k-1$ with $\phi_k(0) = 1$.

Since

$$\begin{aligned}\|\boldsymbol{e}_k\|_A^2 &= (\boldsymbol{x}-\boldsymbol{x}_k)^T A(\boldsymbol{x}-\boldsymbol{x}_k)\\ &= \boldsymbol{x}^T A\boldsymbol{x} - 2\,\boldsymbol{x}_k^T A\boldsymbol{x} + \boldsymbol{x}_k^T A\boldsymbol{x}_k\\ &= \boldsymbol{x}^T\boldsymbol{b} - 2\,\boldsymbol{x}_k^T\boldsymbol{b} + \boldsymbol{x}_k^T A\boldsymbol{x}_k\\ &= 2\,\Phi(\boldsymbol{x}_k) + \boldsymbol{x}^T\boldsymbol{b}\,,\end{aligned} \tag{33}$$

last term on the right being independent of $\boldsymbol{x}_k$, optimality of $\boldsymbol{x}_k$ with respect to a certain subspace is equivalent to optimality of the A-norm of $\boldsymbol{e}_k$ with respect to the same subspace. In other words, CG is constructed in such a way that the polynomial ϕ_k which appears in (31) is the polynomial that minimizes the A norm of $\boldsymbol{e}_k$ among all polynomials or the same or lesser degree.

As in the proof of the Kantorovich inequality we express $\boldsymbol{e}_1$ in terms of the orthonormal eigenvectors of A, i.e.

$$\boldsymbol{e}_1 = \sum_{j=1}^{n} y_j\,\boldsymbol{v}_j\,, \tag{34}$$

so that

$$\boldsymbol{e}_k = \sum_{j=1}^{n} y_j\,\phi_k(\lambda_j)\,\boldsymbol{v}_j\,, \tag{35}$$

$$\|\boldsymbol{e}_k\|_A^2 = \sum_{j=1}^{n} y_j^2\,\phi_k^2(\lambda_j)\,\lambda_j\,. \tag{36}$$

Let P_k denote the vector space of polynomials of degree less or equal to k. The optimality condition of CG can therefore be expressed as

$$\begin{aligned}\|\boldsymbol{e}_k\|_A^2 &= \min_{\substack{\phi\in P_{k-1}\\ \phi(0)=1}} \sum_{j=1}^{n} y_j^2\,\phi^2(\lambda_j)\,\lambda_j\\ &\le \min_{\substack{\phi\in P_{k-1}\\ \phi(0)=1}} \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \phi^2(\lambda)\sum_{j=1}^{n} y_j^2\,\lambda_j\\ &= \min_{\substack{\phi\in P_{k-1}\\ \phi(0)=1}} \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \phi^2(\lambda)\,\|\boldsymbol{e}_1\|_A^2\,.\end{aligned} \tag{37}$$

Let T_k denote the Chebychev polynomial of degree k. Then the following is true.

Lemma 2. *If* $0 < \lambda_{\min} < \lambda_{\max}$, *then*

$$\min_{\substack{\phi\in P_k\\ \phi(0)=1}} \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} |\phi(\lambda)| = T_k\left(\frac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)^{-1} \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k, \tag{38}$$

where $\kappa = \lambda_{\max}/\lambda_{\min}$.

Inserting this result into (37), we find that

$$\|\boldsymbol{e}_k^{\mathrm{cg}}\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k-1}\|\boldsymbol{e}_1\|_A\,, \tag{39}$$

whereas for the gradient method, from equation (12),

$$\|\boldsymbol{e}_k^{\mathrm{grad}}\|_A \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{k-1}\|\boldsymbol{e}_1\|_A\,. \tag{40}$$

This shows that for poorly conditioned matrices the rate of convergence of the conjugate gradient method is much better than that of the gradient method.

Proof of Lemma 2. Recall that the Chebychev polynomials of order k on the interval $[-1,1]$ can be written in the form

$$T_k(x) = \cos(k\,\arccos x)\,. \tag{41}$$

It is clear that $|T_k| \leq 1$ on this interval. Moreover, $T_k(1) = 1$, $T_k(-1) = (-1)^k$, and T_k has $k-1$ distinct extrema on $(-1,1)$, alternating between 1 and -1.

To realize the minimum on the left of (38), we need a polynomial that is uniformly small on the interval $[\lambda_{\min}, \lambda_{\max}]$. Since T_k is of uniform size on $[-1,1]$, we are tempted to use it as a template for constructing a candidate minimizing polynomial by remapping the interval $[\lambda_{\min}, \lambda_{\max}]$ onto $[-1,1]$ by a linear affine change of variable ℓ. Setting

$$\phi(x) = \frac{T_k(\ell(x))}{T_k(\ell(0))} \tag{42}$$

will then also satisfy $\phi(0) = 1$. The ansatz $\ell(x) = mx + c$ together with the requirement that $\ell(\lambda_{\min}) = 1$ and $\ell(\lambda_{\max}) = -1$ gives

$$\ell(x) = \frac{\lambda_{\max} + \lambda_{\min} - 2\lambda}{\lambda_{\max} - \lambda_{\min}}\,. \tag{43}$$

Therefore,

$$\max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} |\phi(\lambda)| = \frac{1}{|T_k(\ell(0))|}\max_{x\in[-1,1]}|T_k(x)| = T_k\left(\frac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)^{-1} \tag{44}$$

To complete the proof of the equality in (38), we need to show that no other polynomial of degree k can yield a bound lower than (44). Assume the contrary, and let ψ denote a polynomial satisfying $\psi(0) = 1$ and

$$\max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} |\psi(\lambda)| < \max_{\lambda\in[\lambda_{\min},\lambda_{\max}]} |\phi(\lambda)|\,. \tag{45}$$

Then $\psi(0) - \phi(0) = 0$, and the graph of ψ must intersect the graph of ϕ exactly k times in the interval $[\lambda_{\min}, \lambda_{\max}]$. This means that $\phi - \psi$ is a polynomial of degree k with $k+1$ zeroes, and must therefore be identically zero.

The proof of the inequality in (38) is a homework exercise. □