Classical Numerical Analysis, Chapter 07

Abner J. Salgado and Steven M. Wise

asalgad1@utk.edu swise1@.utk.edu
University of Tennessee

Chapter 07, Part 1 of 2
Variational and Krylov Subspace Methods

Solving Via Optimization

In this chapter, we introduce gradient–type methods for solving the system of equations $\mathrm{A}\mathbf{x} = \mathbf{f}$, where $\mathrm{A} \in \mathbb{C}^{n\times n}$ is Hermitian positive definite (HPD). These are iterative methods that include the steepest descent and conjugate gradient methods. We will take advantage of the fact that solving this equation is equivalent to minimizing the quadratic function

$$E_{\mathrm{A}}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^{\mathrm{H}}\mathrm{A}\mathbf{z} - \Re\left(\mathbf{z}^{\mathrm{H}}\mathbf{f}\right)$$

over $\mathbb{C}^n$, and will utilize some simple ideas from the theory of convex optimization.

Basic Facts about HPD Matrices

Theorem (properties of HPD matrices)

Suppose that $\mathrm{A} \in \mathbb{C}^{n\times n}$ *is Hermitian positive definite (HPD). Then the following are true.*

1. *The expression*

$$(\mathbf{x}, \mathbf{y})_{\mathrm{A}} = (\mathrm{A}\mathbf{x}, \mathbf{y})_2 = \mathbf{y}^{\mathrm{H}}\mathrm{A}\mathbf{x}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n,$$

   defines an inner product on $\mathbb{C}^n$.

2. *The object* $\|\mathbf{x}\|_{\mathrm{A}} = \sqrt{\mathbf{x}^{\mathrm{H}}\mathrm{A}\mathbf{x}}$, *where* $\mathbf{x} \in \mathbb{C}^n$ *defines a norm on* $\mathbb{C}^n$.

3. *Let the eigenvalues of* A *be ordered so that* $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. *Then*

$$\sqrt{\lambda_1}\,\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_{\mathrm{A}} \leq \sqrt{\lambda_n}\,\|\mathbf{x}\|_2\,,$$

   for any $\mathbf{x} \in \mathbb{C}^n$.

4. *Let* $\mathbf{f} \in \mathbb{C}^n$ *be given. Then* $\mathbf{x} = \mathrm{A}^{-1}\mathbf{f}$ *if and only if* $\mathbf{x}$ *minimizes the quadratic function* $E_{\mathrm{A}} : \mathbb{C}^n \to \mathbb{R}$ *defined by*

$$E_{\mathrm{A}}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^{\mathrm{H}}\mathrm{A}\mathbf{z} - \Re\left(\mathbf{z}^{\mathrm{H}}\mathbf{f}\right).$$

Proof of (4).

(4: $\Longrightarrow$): Suppose that $\mathbf{x} = \mathrm{A}^{-1}\mathbf{f}$. Let $\mathbf{y} \in \mathbb{C}^n$ be arbitrary and consider

$$\begin{aligned} E_{\mathrm{A}}(\mathbf{x}+\mathbf{y}) &= \frac{1}{2}(\mathbf{x}+\mathbf{y})^{\mathrm{H}}\mathrm{A}(\mathbf{x}+\mathbf{y}) - \Re\left(\mathbf{x}^{\mathrm{H}}\mathbf{f}\right) - \Re\left(\mathbf{y}^{\mathrm{H}}\mathbf{f}\right) \\ &= \frac{1}{2}\mathbf{x}^{\mathrm{H}}\mathrm{A}\mathbf{x} + \frac{1}{2}\mathbf{y}^{\mathrm{H}}\mathrm{A}\mathbf{y} + \Re\left(\mathbf{y}^{\mathrm{H}}(\mathrm{A}\mathbf{x}-\mathbf{f})\right) - \Re\left(\mathbf{x}^{\mathrm{H}}\mathbf{f}\right) \\ &= E_{\mathrm{A}}(\mathbf{x}) + \frac{1}{2}\mathbf{y}^{\mathrm{H}}\mathrm{A}\mathbf{y} \\ &\geq E_{\mathrm{A}}(\mathbf{x}), \end{aligned}$$

where, in the last step, we used that A is HPD.

An alternate approach is to establish the following:

$$E_{\mathrm{A}}(\mathbf{z}) = \frac{1}{2}\left\|\mathbf{z} - \mathrm{A}^{-1}\mathbf{f}\right\|_{\mathrm{A}}^2 - \frac{1}{2}\mathbf{f}^{\mathrm{H}}\mathrm{A}^{-1}\mathbf{f}, \tag{1}$$

for arbitrary $\mathbf{z} \in \mathbb{C}^n$. The right hand side is clearly strictly convex (as a function of $\mathbf{z}$), and has the minimizer $\mathbf{z} = \mathbf{x} = \mathrm{A}^{-1}\mathbf{f}$. Since the right and left hand sides must have the same minimizer, we are done.

Proof of (4), Cont.

(4: $\Longleftarrow$): Now suppose that $\mathbf{x}$ minimizes the function E_{A}, and let $\mathbf{u} \in \mathbb{C}^n$ be an arbitrary unit vector. Now define $g(s,t) = E_{\mathrm{A}}(\mathbf{x} + \alpha\mathbf{u})$, where $\alpha = s + \mathrm{i}t$, $s, t \in \mathbb{R}$. Then

$$
\begin{aligned}
g(s,t) &= \frac{1}{2}\left(\mathbf{x}+\alpha\mathbf{u}\right)^{\mathsf{H}}\mathrm{A}\left(\mathbf{x}+\alpha\mathbf{u}\right) - \Re\left(\left(\mathbf{x}+\alpha\mathbf{u}\right)^{\mathsf{H}}\mathbf{f}\right) \\
&= \frac{1}{2}\mathbf{x}^{\mathsf{H}}\mathrm{A}\mathbf{x} + \Re\left(\bar{\alpha}\mathbf{u}^{\mathsf{H}}\mathrm{A}\mathbf{x}\right) + \frac{|\alpha|^2}{2}\mathbf{u}^{\mathsf{H}}\mathrm{A}\mathbf{u} - \Re\left(\mathbf{x}^{\mathsf{H}}\mathbf{f}\right) - \Re\left(\bar{\alpha}\mathbf{u}^{\mathsf{H}}\mathbf{f}\right) \\
&= E_{\mathrm{A}}(\mathbf{x}) + \Re\left(\bar{\alpha}\mathbf{u}^{\mathsf{H}}\left(\mathrm{A}\mathbf{x}-\mathbf{f}\right)\right) + \frac{|\alpha|^2}{2}\mathbf{u}^{\mathsf{H}}\mathrm{A}\mathbf{u} \\
&= E_{\mathrm{A}}(\mathbf{x}) + \Re\left(\bar{\alpha}\right)\Re\left(\mathbf{u}^{\mathsf{H}}\left(\mathrm{A}\mathbf{x}-\mathbf{f}\right)\right) - \Im\left(\bar{\alpha}\right)\Im\left(\mathbf{u}^{\mathsf{H}}\left(\mathrm{A}\mathbf{x}-\mathbf{f}\right)\right) \\
&\quad + \frac{s^2+t^2}{2}\mathbf{u}^{\mathsf{H}}\mathrm{A}\mathbf{u} \\
&= E_{\mathrm{A}}(\mathbf{x}) + s\Re\left(\mathbf{u}^{\mathsf{H}}\left(\mathrm{A}\mathbf{x}-\mathbf{f}\right)\right) + t\Im\left(\mathbf{u}^{\mathsf{H}}\left(\mathrm{A}\mathbf{x}-\mathbf{f}\right)\right) + \frac{s^2+t^2}{2}\mathbf{u}^{\mathsf{H}}\mathrm{A}\mathbf{u}.
\end{aligned}
$$

Proof of (4), Cont.

Clearly g is a strictly convex, quadratic function on $\mathbb{R}^2$. Moreover, g is minimized at $(s,t) = (0,0)$. Hence,

$$0 = \frac{\partial g}{\partial s}(0,0) = \Re\left(\mathbf{u}^{\mathsf{H}}(\mathrm{A}\mathbf{x} - \mathbf{f})\right)$$

and

$$0 = \frac{\partial g}{\partial t}(0,0) = \Im\left(\mathbf{u}^{\mathsf{H}}(\mathrm{A}\mathbf{x} - \mathbf{f})\right)$$

hold for any vector $\mathbf{u}$. It follows then that $\mathrm{A}\mathbf{x} = \mathbf{f}$. □

Proposition (inner products and HPD matrices)

Suppose that $(\cdot\,,\cdot) : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ *is an inner product. There exists a unique HPD matrix* $\mathrm{A} \in \mathbb{C}^{n\times n}$ *such that*

$$(\mathbf{x}, \mathbf{y}) = (\mathrm{A}\mathbf{x}, \mathbf{y})_2 = (\mathbf{x}, \mathbf{y})_{\mathrm{A}}\,, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{C}^n.$$

Proof.

Let, for $j = 1, \ldots, n$, $\hat{\mathbf{e}}_j$ denote the canonical unit basis vectors of $\mathbb{C}^n$. Define the matrix $\mathrm{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ via

$$a_{i,j} = (\hat{\mathbf{e}}_j, \hat{\mathbf{e}}_i)\,.$$

The reader should show that A has the desired properties. □

Definition (A–conjugate)

Suppose that $(\cdot\,,\cdot) : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ is an inner product and $A \in \mathbb{C}^{n\times n}$ is its associated HPD matrix. We say that $B \in \mathbb{C}^{n\times n}$ is **self–adjoint** with respect to this inner product iff

$$(\mathbf{x}, B\mathbf{y}) = (\mathbf{x}, B\mathbf{y})_A = (B\mathbf{x}, \mathbf{y})_A = (B\mathbf{x}, \mathbf{y}), \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{C}^n.$$

We say that B is **self–adjoint positive definite** with respect to the inner product iff B is self–adjoint and satisfies

$$(\mathbf{x}, B\mathbf{x}) = (\mathbf{x}, B\mathbf{x})_A > 0, \quad \forall\, \mathbf{x} \in \mathbb{C}^n_\star.$$

We say that a set $S \subset \mathbb{C}^n$ of non–zero vectors is called A–**orthogonal** (or A–**conjugate**) iff whenever $\mathbf{x}, \mathbf{y} \in S$, and $\mathbf{x} \neq \mathbf{y}$, then

$$(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y})_A = 0.$$

We say that $S \subset \mathbb{C}^n$ is A–**orthonormal** iff S is A–orthogonal and

$$\|\mathbf{x}\|_A = 1, \quad \forall\, \mathbf{x} \in S.$$

Proposition (square root)

Suppose that $A \in \mathbb{C}^{n\times n}$ *is HPD. Then* A *is invertible and* A^{-1} *is HPD. Furthermore, there exists a unique HPD matrix* $B \in \mathbb{C}^{n\times n}$ *with the property that* $BB = A$.

Proof.

Since A is HPD, it is unitarily diagonalizable, that is, there are a unitary matrix $U \in \mathbb{C}^{n\times n}$ and a diagonal matrix with positive diagonal entries $D = \mathbf{diag}[\lambda_1, \ldots, \lambda_n]$ such that

$$A = UDU^H.$$

Set

$$B = UD^{1/2}U^H \quad \text{and} \quad D^{1/2} = \mathbf{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\right).$$

Then B has the desired properties. Furthermore, it is easy to see that

$$A^{-1} = UD^{-1}U^H, \quad D^{-1} = \mathbf{diag}\left(\lambda_1^{-1}, \ldots, \lambda_n^{-1}\right).$$

We leave it to the reader to check the details. □

Proposition (product of HPD matrices)

Suppose that $\mathrm{A}, \mathrm{B} \in \mathbb{C}^{n\times n}$ *are HPD matrices. Then the product* BA *is self–adjoint and positive definite with respect to* $(\,\cdot\,,\,\cdot\,)_{\mathrm{A}}$*, i.e.,*

$$(\mathbf{x}, \mathrm{BA}\mathbf{y})_{\mathrm{A}} = (\mathrm{BA}\mathbf{x}, \mathbf{y})_{\mathrm{A}}\,, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{C}^n$$

and

$$(\mathrm{BA}\mathbf{x}, \mathbf{x})_{\mathrm{A}} > 0, \quad \forall\, \mathbf{x} \in \mathbb{C}^n_\star.$$

Proposition (similarity)

Suppose that $A \in \mathbb{C}^{n\times n}$ *is HPD and* $B \in \mathbb{C}^{n\times n}$ *is self–adjoint with respect to* $(\cdot,\cdot)_A$, *i.e.,*

$$(\mathbf{x}, B\mathbf{y})_A = (B\mathbf{x}, \mathbf{y})_A\,, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{C}^n.$$

Then, B *is similar to a matrix that is Hermitian, that is, a matrix that is self–adjoint with respect to the Euclidean inner product* $(\,\cdot\,,\,\cdot\,)_2$.

Proof.

Since $B \in \mathbb{C}^{n\times n}$ is self–adjoint with respect to $(\cdot,\cdot)_A$, then the reader should confirm that

$$B^H A = AB.$$

Since A is HPD, there is an invertible matrix $L \in \mathbb{C}^{n\times n}$ such that $A = LL^H$. Define

$$C = L^H B L^{-H}.$$

Proof, Cont.

Then

$$\begin{aligned} C^H &= L^{-1}B^H L \\ &= L^{-1}B^H L L^H L^{-H} \\ &= L^{-1}B^H A L^{-H} \\ &= L^{-1}ABL^{-H} \\ &= L^{-1}LL^H BL^{-H} \\ &= L^H BL^{-H} \\ &= C. \end{aligned}$$

Since C is similar to B, the result follows. □

Theorem (spectral decomposition)

Suppose that $A \in \mathbb{C}^{n\times n}$ *is HPD and* $B \in \mathbb{C}^{n\times n}$ *is self–adjoint with respect to* $(\cdot\,,\cdot)_A$. *Then all of the eigenvalues of* B *are real and there is an* A*–orthonormal basis of* $\mathbb{C}^n$ *consisting of eigenvectors of* B.

Proof.

Applying the Euclidean spectral decomposition theorem to the Hermitian matrix

$$C = L^H B L^{-H},$$

where the invertible matrix L is taken from the Cholesky–type decomposition $A = LL^H$ of A, there are a unitary matrix $U \in \mathbb{C}^{n\times n}$ and a diagonal matrix with real entries $D = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$ such that

$$L^H B L^{-H} = C = UDU^{-1}.$$

Hence, B is similar to a diagonal matrix with real entries:

$$B = \left(L^{-H}U\right) D \left(L^{-H}U\right)^{-1}.$$

Proof, Cont.

Moreover, setting $M = L^{-H}U$, we see that

$$BM = MD,$$

which implies that the columns of the invertible matrix M are eigenvectors of B. It only remains to check that the columns of M form an A–orthonormal set. This is left to the reader as an exercise. □

Corollary (eigenvalues)

Suppose that $A \in \mathbb{C}^{n\times n}$ *is HPD and* $B \in \mathbb{C}^{n\times n}$ *is self–adjoint and positive definite with respect to* $(\cdot\,,\cdot)_A$. *Then the eigenvalues of* B *are all real and positive.*

Gradient Descent Methods

Definition (gradient descent)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ is HPD, $\mathbf{f} \in \mathbb{C}^n$, and define the quadratic function $E_{\mathsf{A}} : \mathbb{C}^n \to \mathbb{R}$ via

$$E_{\mathsf{A}}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^{\mathsf{H}}\mathsf{A}\mathbf{z} - \Re\left(\mathbf{z}^{\mathsf{H}}\mathbf{f}\right).$$

A **gradient descent method** is a two–layer iterative scheme to approximate $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$. Starting from an arbitrary initial guess $\mathbf{x}_0$, the iterations proceed as

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_k \mathbf{d}_{k-1}, \quad k = 1, 2, 3, \ldots,$$

where $\mathbf{d}_{k-1} \in \mathbb{C}^n$ is the $(k-1)$–st **search direction**, supplied by the algorithm, and $\alpha_k \in \mathbb{C}$ is the **step size** given by the condition

$$\alpha_k = \underset{\alpha\in\mathbb{C}}{\operatorname{argmin}}\, E_{\mathsf{A}}\left(\mathbf{x}_{k-1} + \alpha\mathbf{d}_{k-1}\right),$$

which is called a **line search**.

Theorem (gradient descent, real case)

Suppose that $\mathsf{A} \in \mathbb{R}^{n\times n}$ *is SPD,* $\mathbf{f} \in \mathbb{R}^n$*, and define the quadratic function* $E_{\mathsf{A}} : \mathbb{R}^n \to \mathbb{R}$

$$E_{\mathsf{A}}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^{\mathsf{T}}\mathsf{A}\mathbf{z} - \mathbf{z}^{\mathsf{T}}\mathbf{f}.$$

Suppose that the search direction $\mathbf{d}_{k-1} \in \mathbb{R}^n_\star$ *and previous iterate* $\mathbf{x}_{k-1} \in \mathbb{R}^n$ *in a gradient descent method are given, and define* $\mathbf{r}_{k-1} = \mathbf{f} - \mathsf{A}\mathbf{x}_{k-1}$*. Then the step size can be computed exactly via the formula*

$$\alpha_k = \operatorname*{argmin}_{\alpha\in\mathbb{R}} E_{\mathsf{A}}\left(\mathbf{x}_{k-1} + \alpha\mathbf{d}_{k-1}\right) = \frac{\mathbf{d}_{k-1}^{\mathsf{T}}\mathbf{r}_{k-1}}{\mathbf{d}_{k-1}^{\mathsf{T}}\mathsf{A}\mathbf{d}_{k-1}}.$$

In other words, a gradient descent method is well defined once a non–trivial search direction is specified.

Proof.

Consider the quadratic

$$g(\alpha) = E_{\mathsf{A}}\left(\mathbf{x}_{k-1} + \alpha\mathbf{d}_{k-1}\right) = E_{\mathsf{A}}\left(\mathbf{x}_{k-1}\right) - \alpha\mathbf{d}_{k-1}^{\mathsf{T}}\mathbf{r}_{k-1} + \alpha^2\frac{1}{2}\mathbf{d}_{k-1}^{\mathsf{T}}\mathsf{A}\mathbf{d}_{k-1}.$$

Proof, Cont.

A calculation with the first derivative locates the extremum:

$$0 = g'(\alpha_k) = -\mathbf{d}_{k-1}^{\mathsf{T}}\mathbf{r}_{k-1} + \alpha_k \mathbf{d}_{k-1}^{\mathsf{T}} A\mathbf{d}_{k-1},$$

which implies

$$\alpha_k = \frac{\mathbf{d}_{k-1}^{\mathsf{T}}\mathbf{r}_{k-1}}{\mathbf{d}_{k-1}^{\mathsf{T}} A\mathbf{d}_{k-1}}.$$

The second derivative indicates that this is a minimum:

$$g''(\alpha) = \mathbf{d}_{k-1}^{\mathsf{T}} A\mathbf{d}_{k-1} > 0,$$

provided $\mathbf{d}_{k-1} \neq \mathbf{0}$. □

Theorem (gradient descent, complex case)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n$*, and define the quadratic function* $E_{\mathsf{A}} : \mathbb{C}^n \to \mathbb{R}$

$$E_{\mathsf{A}}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^{\mathsf{H}}\mathsf{A}\mathbf{z} - \Re\left(\mathbf{z}^{\mathsf{H}}\mathbf{f}\right).$$

Suppose that the search direction $\mathbf{d}_{k-1} \in \mathbb{C}^n_\star$ *and previous iterate* $\mathbf{x}_{k-1} \in \mathbb{C}^n$ *in a gradient descent method are given, and define* $\mathbf{r}_{k-1} = \mathbf{f} - \mathsf{A}\mathbf{x}_{k-1}$*. Then the step size can be computed exactly via the formula*

$$\alpha_k = \operatorname*{argmin}_{\alpha\in\mathbb{C}} E_{\mathsf{A}}\left(\mathbf{x}_{k-1} + \alpha\mathbf{d}_{k-1}\right) = \frac{\mathbf{d}_{k-1}^{\mathsf{H}}\mathbf{r}_{k-1}}{\mathbf{d}_{k-1}^{\mathsf{H}}\mathsf{A}\mathbf{d}_{k-1}}.$$

Proof.

Let $\alpha = s + \mathrm{i}t$, where $s, t \in \mathbb{R}$. Define the function

$$g(s,t) = E_{\mathsf{A}}\left(\mathbf{x}_{k-1} + \alpha\mathbf{d}_{k-1}\right).$$

Proof, Cont.

Then,

$$\begin{aligned} g(s,t) &= E_{\mathrm{A}}\left(\mathbf{x}_{k-1} + \alpha \mathbf{d}_{k-1}\right) \\ &= E_{\mathrm{A}}\left(\mathbf{x}_{k-1}\right) - \Re\left(\bar{\alpha}\mathbf{d}_{k-1}^{\mathrm{H}}\mathbf{r}_{k-1}\right) + \frac{|\alpha|^2}{2}\mathbf{d}_{k-1}^{\mathrm{H}}\mathrm{A}\mathbf{d}_{k-1} \\ &= E_{\mathrm{A}}\left(\mathbf{x}_{k-1}\right) - s\Re\left(\mathbf{d}_{k-1}^{\mathrm{H}}\mathbf{r}_{k-1}\right) - t\Im\left(\mathbf{d}_{k-1}^{\mathrm{H}}\mathbf{r}_{k-1}\right) + \frac{s^2+t^2}{2}\mathbf{d}_{k-1}^{\mathrm{H}}\mathrm{A}\mathbf{d}_{k-1}. \end{aligned}$$

This is a strictly convex quadratic function of two variables. Setting the first derivatives equal to zero, we find

$$\begin{aligned} 0 = \frac{\partial g}{\partial s}(s_k, t_k) &= -\Re\left(\mathbf{d}_{k-1}^{\mathrm{H}}\mathbf{r}_{k-1}\right) + s_k\mathbf{d}_{k-1}^{\mathrm{H}}\mathrm{A}\mathbf{d}_{k-1}, \\ 0 = \frac{\partial g}{\partial t}(s_k, t_k) &= -\Im\left(\mathbf{d}_{k-1}^{\mathrm{H}}\mathbf{r}_{k-1}\right) + t_k\mathbf{d}_{k-1}^{\mathrm{H}}\mathrm{A}\mathbf{d}_{k-1}, \end{aligned}$$

which implies that

$$\alpha_k = s_k + \mathrm{i}t_k = \frac{\mathbf{d}_{k-1}^{\mathrm{H}}\mathbf{r}_{k-1}}{\mathbf{d}_{k-1}^{\mathrm{H}}\mathrm{A}\mathbf{d}_{k-1}}.$$

□

The Steepest Descent Method

Definition (steepest descent)

Suppose that $A \in \mathbb{C}^{n\times n}$ is HPD, and $\mathbf{f} \in \mathbb{C}^n$. The **steepest descent method** is a gradient descent method for which the search direction $\mathbf{d}_{k-1}$ is defined to be the residual, i.e.,

$$\mathbf{d}_{k-1} = \mathbf{r}_{k-1} = \mathbf{f} - A\mathbf{x}_{k-1},$$

so that the step size is precisely

$$\alpha_k = \frac{\mathbf{r}_{k-1}^H \mathbf{r}_{k-1}}{\mathbf{r}_{k-1}^H A \mathbf{r}_{k-1}}.$$

If $B \in \mathbb{C}^{n\times n}$ is an HPD matrix, the B–**preconditioned steepest descent** method is a gradient descent method with search direction

$$\mathbf{d}_{k-1} = B^{-1}\mathbf{r}_{k-1}, \tag{2}$$

so that the step size is precisely

$$\alpha_k = \frac{\mathbf{r}_{k-1}^H B^{-1} \mathbf{r}_{k-1}}{\mathbf{r}_{k-1}^H B^{-1} A B^{-1} \mathbf{r}_{k-1}}.$$

What's the Deal with Preconditioning?

The idea with preconditioning is that the *preconditioner*, B, should be like A, but easier to invert. In fact, if we choose $B = A$, which is not practical, we would converge in a single iteration, because (2) would yield the error vector as the search direction. One way of realizing preconditioned steepest descent, theoretically, is to observe that it is just normal steepest descent applied to solve the equation

$$B^{-1}A\mathbf{x} = B^{-1}\mathbf{f}. \tag{3}$$

Proposition (orthogonality)

Suppose that $A \in \mathbb{C}^{n\times n}$ *is HPD, and* $\mathbf{f} \in \mathbb{C}^n$. *Suppose that* $\{\mathbf{x}_k\}_{k=1}^{\infty}$ *is computed using the steepest descent method with the starting vector* $\mathbf{x}_0$. *Then the sequence of residual vectors* $\{\mathbf{r}_k\}_{k=1}^{\infty}$, $\mathbf{r}_k = \mathbf{f} - A\mathbf{x}_k$, *has the property that*

$$(\mathbf{r}_k, \mathbf{r}_{k+1})_2 = \mathbf{r}_{k+1}^H \mathbf{r}_k = 0,$$

for $k = 0, 1, 2, \ldots$.

Proof.

The proof is an exercise. □

Remark (orthogonality)

The last proposition shows that the steepest descent method has the property that its next search direction is always orthogonal to the last. We will see that this can lead to some bad outcomes.

Theorem (error equation)

Suppose that $\mathrm{A} \in \mathbb{C}^{n \times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n$*, and* $\mathbf{x} = \mathrm{A}^{-1}\mathbf{f}$*. Suppose that* $\{\mathbf{x}_k\}_{k=1}^{\infty}$ *is computed using the steepest descent method with the starting value* $\mathbf{x}_0 \in \mathbb{C}^n$*. Then the error* $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k$ *satisfies*

$$\|\mathbf{e}_{k+1}\|_{\mathrm{A}}^2 = \gamma_k \|\mathbf{e}_k\|_{\mathrm{A}}^2 ,$$

where

$$\gamma_k = 1 - \frac{\left(\mathbf{r}_k^{\mathrm{H}}\mathbf{r}_k\right)^2}{\left(\mathbf{r}_k^{\mathrm{H}}\mathrm{A}\mathbf{r}_k\right)\left(\mathbf{r}_k^{\mathrm{H}}\mathrm{A}^{-1}\mathbf{r}_k\right)}.$$

Proof.

Suppose that $\mathbf{x} = \mathrm{A}^{-1}\mathbf{f}$. Then, for any $\mathbf{z} \in \mathbb{C}^n$,

$$E_{\mathrm{A}}(\mathbf{z}) = E_{\mathrm{A}}(\mathbf{x}) + \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_{\mathrm{A}}^2 , \tag{4}$$

where, as usual,

$$E_{\mathrm{A}}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^{\mathrm{H}}\mathrm{A}\mathbf{z} - \Re\left(\mathbf{z}^{\mathrm{H}}\mathbf{f}\right).$$

Proof, Cont.

Since $\mathbf{r}_k = \mathbf{f} - \mathrm{A}\mathbf{x}_k$, and, for the standard steepest descent method,

$$\alpha_{k+1} = \frac{\mathbf{r}_k^{\mathrm{H}}\mathbf{r}_k}{\mathbf{r}_k^{\mathrm{H}}\mathrm{A}\mathbf{r}_k},$$

a brief calculation shows that

$$E_{\mathrm{A}}(\mathbf{x}_{k+1}) = E_{\mathrm{A}}(\mathbf{x}_k + \alpha_{k+1}\mathbf{r}_k) = E_{\mathrm{A}}(\mathbf{x}_k) - \frac{1}{2}\frac{\left(\mathbf{r}_k^{\mathrm{H}}\mathbf{r}_k\right)^2}{\mathbf{r}_k^{\mathrm{H}}\mathrm{A}\mathbf{r}_k}. \tag{5}$$

Combining equations (4) and (5) we get

$$\|\mathbf{e}_{k+1}\|_{\mathrm{A}}^2 = \|\mathbf{e}_k\|_{\mathrm{A}}^2 - \frac{\left(\mathbf{r}_k^{\mathrm{H}}\mathbf{r}_k\right)^2}{\mathbf{r}_k^{\mathrm{H}}\mathrm{A}\mathbf{r}_k}. \tag{6}$$

Since $\mathbf{r}_k = \mathrm{A}\mathbf{e}_k$, we have

$$\|\mathbf{e}_k\|_{\mathrm{A}}^2 = \mathbf{r}_k^{\mathrm{H}}\mathrm{A}^{-1}\mathbf{r}_k. \tag{7}$$

Combining (6) and (7), we get the desired result. □

Lemma (Kantorovich inequality)

Let the matrix $\mathrm{A} \in \mathbb{C}^{n\times n}$ *be HPD with spectrum* $\sigma(\mathrm{A}) = \{\lambda_i\}_{i=1}^n$, *with* $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, *and spectral condition number*

$$\kappa = \kappa_2(\mathrm{A}) = \frac{\lambda_n}{\lambda_1}.$$

Then, for any $\mathbf{x} \in \mathbb{C}_\star^n$,

$$\frac{\left(\mathbf{x}^{\mathrm{H}}\mathrm{A}\mathbf{x}\right)\left(\mathbf{x}^{\mathrm{H}}\mathrm{A}^{-1}\mathbf{x}\right)}{\left(\mathbf{x}^{\mathrm{H}}\mathbf{x}\right)^2} \leq \frac{1}{4}\left(\sqrt{\kappa} + \sqrt{\kappa^{-1}}\right)^2.$$

Proof.

See the book. □

Theorem (convergence)

Suppose that $\mathrm{A} \in \mathbb{C}^{n \times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n$*, and* $\mathbf{x} = \mathrm{A}^{-1}\mathbf{f}$*. Suppose that* $\{\mathbf{x}_k\}_{k=1}^{\infty}$ *is computed using the steepest descent method with the starting value* $\mathbf{x}_0 \in \mathbb{C}^n$*. Then the following estimate holds*

$$\|\mathbf{e}_k\|_{\mathrm{A}} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{e}_0\|_{\mathrm{A}},$$

where $\kappa = \kappa_2(\mathrm{A})$*.*

Proof.

Using the Kantorovich inequality and the previous theorem to show that

$$\gamma_k \leq 1 - \frac{4}{\left(\kappa^{-1/2} + \kappa^{1/2}\right)^2} = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2,$$

which implies the result. □

Example

Let us show, by means of an example, that the convergence rate for the steepest descent method, presented in the previous result, is sharp. Let $A = \mathbf{diag}[a, b] \in \mathbb{R}^{2\times 2}$, $a > b > 0$, $\mathbf{f} = \mathbf{0}$, and $\mathbf{x} = \mathbf{0}$. We will prove the following: if

$$\mathbf{x}_i = c_i\,[b, sa]^\intercal \in V_1 = \langle [b, sa]^\intercal \rangle,$$

for some $c_i \in \mathbb{R}_\star$, where $s = \pm 1$, then

$$\mathbf{x}_{i+1} = c_{i+1}\,[b, -sa]^\intercal \in V_2 = \langle [b, -sa]^\intercal \rangle,$$

for some coefficient $c_{i+1} \in \mathbb{R}_\star$. In general, $\mathbf{x}_k \in V_1 \cup V_2$. First $\mathbf{r}_i = -c_i\,[ab, sab]^\intercal$ and $A\mathbf{r}_i = -c_i\left[a^2 b, sab^2\right]^\intercal$. Then

$$\alpha_{i+1} = \frac{\mathbf{r}_i^\intercal \mathbf{r}_i}{\mathbf{r}_i^\intercal A \mathbf{r}_i} = \frac{2c_i^2 a^2 b^2}{c_i^2\left(a^3 b^2 + a^2 b^3\right)} = \frac{2}{a+b}\,.$$

Example (Cont.)

Thus,

$$
\begin{aligned}
\mathbf{x}_{i+1} &= \mathbf{x}_i + \alpha_i \mathbf{r}_i \\
&= \begin{bmatrix} c_i b - \frac{2c_i}{a+b} ab \\ c_i sa - \frac{2c_i}{a+b} sab \end{bmatrix} \\
&= \frac{c_i}{a+b} \begin{bmatrix} b(a+b) - 2ab \\ sa(a+b) - 2sab \end{bmatrix} \\
&= \frac{c_i}{a+b} \begin{bmatrix} b^2 - ab \\ sa^2 - sab \end{bmatrix} \\
&= \frac{c_i(b-a)}{a+b} \begin{bmatrix} b \\ -sa \end{bmatrix}
\end{aligned}
$$

Example (Cont.)

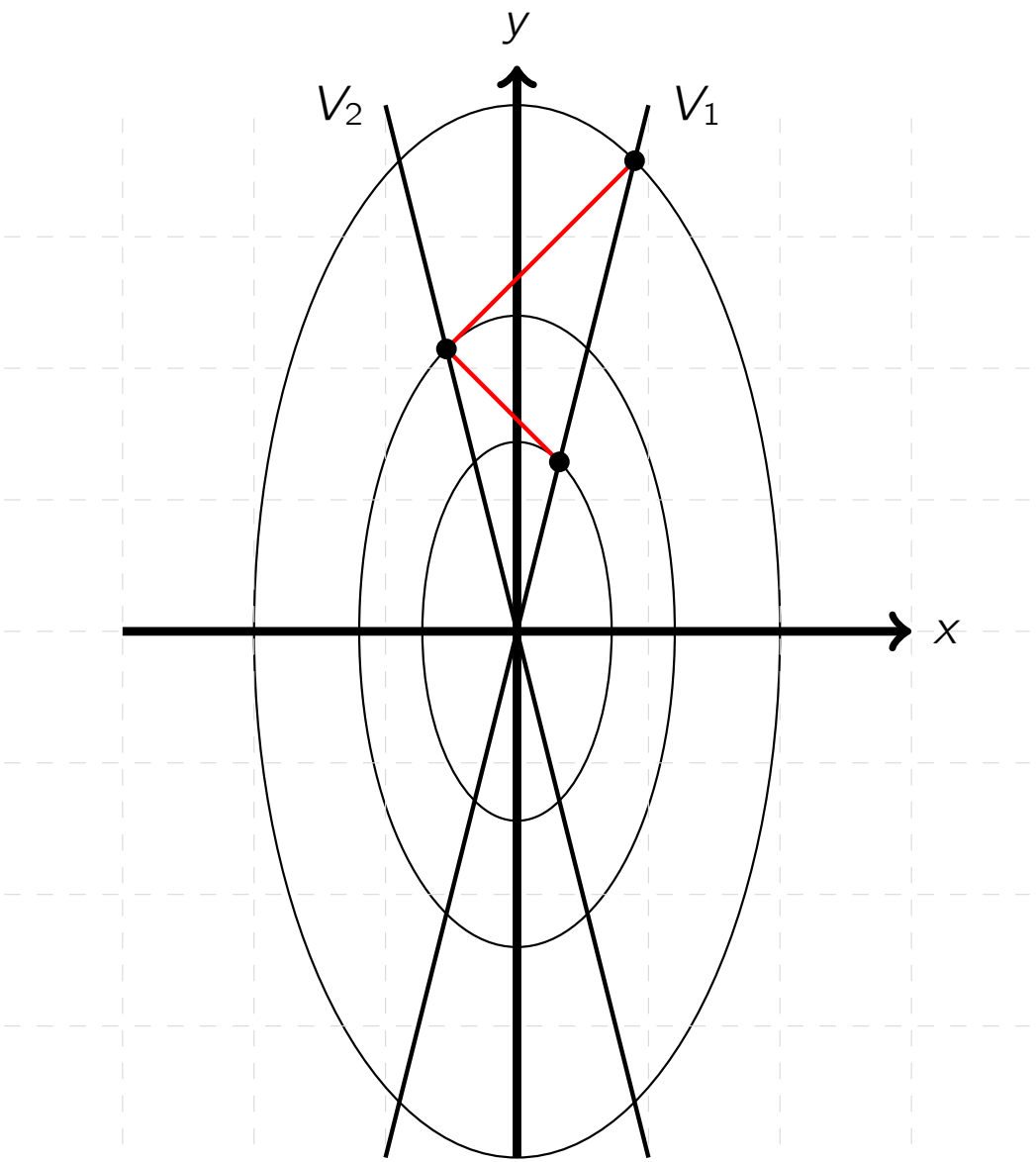
Hence

$$c_{i+1} = \frac{c_i(b-a)}{a+b} = -c_i \frac{\kappa - 1}{\kappa + 1},$$

where $\kappa = \frac{a}{b}$ is precisely the 2–norm condition number of the matrix A. From this, one can see that the convergence rate is the worst possible, according to our convergence theory, because

$$c_k = (-1)^k c_0 \left(\frac{\kappa - 1}{\kappa + 1} \right)^k.$$

Suppose $a = 16$ and $b = 4$



Deterioration of Convergence Rate for Ill-Conditioned Problems

For the steepest descent method, with large spectral condition number κ, we observe that

$$\frac{\kappa - 1}{\kappa + 1} \approx 1 - \frac{2}{\kappa}.$$

In other words, the convergence rate deteriorates as $\kappa \to \infty$, that is, as the system matrix becomes increasingly ill conditioned. We will see, when we study numerical methods for certain types of differential equations, how a family of matrices can become more and more ill conditioned as the size of the matrices increases.

Theorem (error equation)

Suppose that $\mathsf{A}, \mathsf{B} \in \mathbb{C}^{n\times n}$ *are HPD,* $\mathbf{f} \in \mathbb{C}^n$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. Suppose that* $\{\mathbf{x}_k\}_{k=1}^{\infty}$ *is computed using the* B*–preconditioned steepest descent method with the starting value* $\mathbf{x}_0 \in \mathbb{C}^n$*. Then the error* $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k$ *satisfies*

$$\|\mathbf{e}_{k+1}\|_{\mathsf{A}}^2 = \beta_k \|\mathbf{e}_k\|_{\mathsf{A}}^2 ,$$

where

$$\beta_k = 1 - \frac{\left(\mathbf{d}_k^{\mathsf{H}}\mathbf{r}_k\right)^2}{\left(\mathbf{d}_k^{\mathsf{H}}\mathsf{A}\mathbf{d}_k\right)\left(\mathbf{r}_k^{\mathsf{H}}\mathsf{A}^{-1}\mathbf{r}_k\right)} \quad \textit{and} \quad \mathbf{r}_k = \mathsf{B}\mathbf{d}_k .$$

Setting $\mathsf{B} = \mathsf{L}^{\mathsf{H}}\mathsf{L}$*, for some invertible matrix* $\mathsf{L} \in \mathbb{C}^{n\times n}$*,*

$$\mathbf{g}_k = \mathsf{L}\mathbf{d}_k \quad \textit{and} \quad \mathsf{C} = \mathsf{L}^{-\mathsf{H}}\mathsf{A}\mathsf{L}^{-1},$$

the error equation may be expressed using

$$\beta_k = 1 - \frac{\left(\mathbf{g}_k^{\mathsf{H}}\mathbf{g}_k\right)^2}{\left(\mathbf{g}_k^{\mathsf{H}}\mathsf{C}\mathbf{g}_k\right)\left(\mathbf{g}_k^{\mathsf{H}}\mathsf{C}^{-1}\mathbf{g}_k\right)}.$$

Theorem (convergence)

Suppose that $\mathsf{A}, \mathsf{B} \in \mathbb{C}^{n\times n}$ *are HPD,* $\mathbf{f} \in \mathbb{C}^n$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. Suppose that* $\{\mathbf{x}_k\}_{k=1}^{\infty}$ *is computed using the* B*–preconditioned steepest descent method with the starting value* $\mathbf{x}_0 \in \mathbb{C}^n$*. Suppose that* B *has the Cholesky–type factorization* $\mathsf{B} = \mathsf{L}^{\mathsf{H}}\mathsf{L}$*, where* $\mathsf{L} \in \mathbb{C}^{n\times n}$ *is invertible, and define* $\mathsf{C} = \mathsf{L}^{-\mathsf{H}}\mathsf{A}\mathsf{L}^{-1}$*. Then the following error estimate holds:*

$$\|\mathbf{e}_k\|_{\mathsf{A}} \leq \left(\frac{\kappa_{\mathsf{C}} - 1}{\kappa_{\mathsf{C}} + 1}\right)^k \|\mathbf{e}_0\|_{\mathsf{A}},$$

where

$$\kappa_{\mathsf{C}} = \kappa_2(\mathsf{C}) = \frac{\mu_n}{\mu_1}, \quad \sigma(\mathsf{C}) = \{\mu_i\}_{i=1}^{n}, \quad 0 < \mu_1 \leq \cdots \leq \mu_n.$$

Proof.

Observe that $\mathsf{C} = \mathsf{L}^{-\mathsf{H}}\mathsf{A}\mathsf{L}^{-1}$ is HPD. The result now follows by applying the Kantorovich inequality to estimate the size β_k from the previous theorem. □

Generalized Condition Number

Consider the preconditioned system $\mathrm{B}^{-1}\mathrm{A}\mathbf{x} = \mathrm{B}^{-1}\mathbf{f}$, and observe that $\mathrm{B}^{-1}\mathrm{A}$ is self–adjoint and positive definite with respect to $(\cdot\,,\cdot)_{\mathrm{A}}$, since B^{-1} is HPD. It therefore has positive real eigenvalues. Furthermore, one will find that $\mathrm{B}^{-1}\mathrm{A}$ is similar to $\mathrm{C} = \mathrm{L}^{-\mathrm{H}}\mathrm{A}\mathrm{L}^{-1}$. In particular,

$$\mathrm{L}\left(\mathrm{B}^{-1}\mathrm{A}\right)\mathrm{L}^{-1} = \mathrm{L}^{-\mathrm{H}}\mathrm{A}\mathrm{L}^{-1} = \mathrm{C}.$$

Therefore the eigenvalues of C and the preconditioned coefficient matrix $\mathrm{B}^{-1}\mathrm{A}$ are the same. Therefore, we will often write the result of the last theorem in the following way:

$$\|\mathbf{e}_k\|_{\mathrm{A}} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{e}_0\|_{\mathrm{A}}\,,$$

where

$$\kappa = \kappa_{\mathrm{B}^{-1}\mathrm{A}} = \frac{\mu_n}{\mu_1}, \quad \sigma(\mathrm{B}^{-1}\mathrm{A}) = \{\mu_i\}_{i=1}^n\,, \quad 0 < \mu_1 \leq \cdots \leq \mu_n.$$

The number $\kappa_{\mathrm{B}^{-1}\mathrm{A}} \geq 1$ is called the *condition number* of the preconditioned coefficient matrix $\mathrm{B}^{-1}\mathrm{A}$.

The Effect of Preconditioning

The result of this last theorem is quite important. It shows that, if we select the preconditioner B in a careful way, it is possible to dramatically improve the convergence rate. In particular, it is possible, theoretically, to find B such that κ_C is nearly one. The closer κ_C is to one, the faster is the convergence rate.