Classical Numerical Analysis, Chapter 07

Abner J. Salgado and Steven M. Wise

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

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

The Conjugate Gradient Method

Krylov Subspaces

We have arrived at the so–called conjugate gradient (CG) method. Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be Hermitian and positive definite (HPD) and $\mathbf{f} \in \mathbb{C}^n$. Recall that we are interested in finding a solution to

$$\mathsf{A}\mathbf{x} = \mathbf{f},$$

or, equivalently,

$$\mathbf{x} = \operatorname*{argmin}_{\mathbf{z}\in\mathbb{C}^n} E_{\mathsf{A}}(\mathbf{z}), \qquad E_{\mathsf{A}}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^{\mathsf{H}}\mathsf{A}\mathbf{z} - \Re\left(\mathbf{z}^{\mathsf{H}}\mathbf{f}\right).$$

To solve $\mathsf{A}\mathbf{x} = \mathbf{f}$ we will, instead, try to solve the minimization problem in a clever way, namely, by doing it over an increasing sequence of subspaces of $\mathbb{C}^n$.

Definition (Krylov subspace)

Given $\mathsf{A} \in \mathbb{C}^{n\times n}$ and $\mathbf{0} \neq \mathbf{q} \in \mathbb{C}^n$, the **Krylov subspace** of degree m is

$$\mathcal{K}_m(\mathsf{A}, \mathbf{q}) = \operatorname{span}\left\{\mathsf{A}^k\mathbf{q} \,\middle|\, k = 0, \ldots, m-1\right\}.$$

The Conjugate Gradient Method

Definition

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n_\star$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. The* **zero–start conjugate gradient method** *is an iterative scheme for producing a sequence of approximations* $\{\mathbf{x}_k\}_{k=1}^\infty$ *from the starting point* $\mathbf{x}_0 = \mathbf{0} \in \mathbb{C}^n$ *according to the following prescription: setting* $\mathcal{K}_k = \mathcal{K}_k(\mathsf{A}, \mathbf{f})$*, the k–th iterate is obtained by*

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathrm{z}\in\mathcal{K}_k} E_{\mathsf{A}}(\mathbf{z}). \tag{1}$$

Notice that, by construction, $\mathcal{K}_m(\mathsf{A}, \mathbf{q}) \subseteq \mathcal{K}_{m+1}(\mathsf{A}, \mathbf{q})$. Thus, we are minimizing over an increasing family of nested subspaces of $\mathbb{C}^n$.

Later on, we will mention the important case of starting the conjugate gradient method with a non–zero starting vector $\mathbf{x}_0$. In many cases, the user may have some insight on how to start the conjugate gradient method so as to obtain faster convergence.

Definition (Galerkin approximation)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ is HPD, $\mathbf{f} \in \mathbb{C}^n$, $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$, and W is a subspace of $\mathbb{C}^n$. The vector $\mathbf{x}_W \in W$ is called the **Galerkin approximation** of $\mathbf{x}$ in W iff

$$(\mathsf{A}\mathbf{x}_W, \mathbf{w})_2 = (\mathbf{f}, \mathbf{w})_2, \quad \forall\, \mathbf{w} \in W. \tag{2}$$

So, given $W \leq \mathbb{C}$ and A HPD, how do we find $\mathbf{x}_W$? Does a solution exist?

Theorem (existence and uniqueness)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n$, $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$, *and* W *is a subspace of* $\mathbb{C}^n$. *The Galerkin approximation* $\mathbf{x}_W \in W$ *exists and is unique.*

Proof.

Let $B = \{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ be an A–orthonormal basis for W, i.e.,

$$(\mathbf{w}_i, \mathbf{w}_j)_{\mathsf{A}} = (\mathsf{A}\mathbf{w}_i, \mathbf{w}_j)_2 = \delta_{i,j},$$

for all $1 \le i, j \le k \le n$. (How do we know such a basis exists?) Then (2) holds iff

$$(\mathsf{A}\mathbf{x}_W, \mathbf{w}_i)_2 = (\mathbf{f}, \mathbf{w}_i)_2, \quad i = 1, \ldots, k. \tag{3}$$

Since B is a basis, there are unique constants $c_1, \ldots, c_k \in \mathbb{C}$ such that $\mathbf{x}_W = \sum_{j=1}^k c_j \mathbf{w}_j$. Plugging this into (3) we get the trivial diagonal system

$$c_i = \sum_{j=1}^k c_j (\mathbf{w}_j, \mathbf{w}_i)_{\mathsf{A}} = (\mathbf{f}, \mathbf{w}_i)_2, \quad i = 1, \ldots, k. \tag{4}$$

This proves existence and uniqueness. □

Proposition (properties of Galerkin approximations)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n$, $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$, W *is a subspace of* $\mathbb{C}^n$, *and* $\mathbf{x}_W$ *is the Galerkin approximation to* $\mathbf{x}$ *in* W*.*

1. *The residual is orthogonal to* W*. That is, if* $\mathbf{r} = \mathbf{f} - \mathsf{A}\mathbf{x}_W$, *we have*

$$(\mathbf{r}, \mathbf{w})_2 = 0, \quad \forall\, \mathbf{w} \in W.$$

2. *Galerkin orthogonality: Define the error* $\mathbf{e} = \mathbf{x} - \mathbf{x}_W$*. Then we have*

$$(\mathsf{A}\mathbf{e}, \mathbf{w})_2 = 0, \quad \forall\, \mathbf{w} \in W.$$

3. *Optimality:*

$$(\mathsf{A}\mathbf{e}, \mathbf{e})_2 \le (\mathsf{A}(\mathbf{x} - \mathbf{w}), \mathbf{x} - \mathbf{w})_2, \quad \forall\, \mathbf{w} \in W.$$

Proof of properties of Galerkin approximations.

We only prove optimality. The other properties are trivial. Let $\mathbf{w} \in W$ be arbitrary. Using Galerkin orthogonality, and the Cauchy–Schwarz inequality for the A–norm,

$$\begin{aligned} \|\mathbf{e}\|_A^2 &= (A\mathbf{e}, \mathbf{e})_2 \\ &= (A\mathbf{e}, \mathbf{x} - \mathbf{x}_W)_2 \\ &= (A\mathbf{e}, \mathbf{x} - \mathbf{x}_W)_2 + (A\mathbf{e}, \mathbf{x}_W - \mathbf{w})_2 \\ &= (A\mathbf{e}, \mathbf{x} - \mathbf{x}_W + \mathbf{x}_W - \mathbf{w})_2 \\ &= (A\mathbf{e}, \mathbf{x} - \mathbf{w})_2 \\ &\le \|\mathbf{e}\|_A \|\mathbf{x} - \mathbf{w}\|_A . \end{aligned}$$

If $\|\mathbf{e}\|_A = 0$, the result is trivial. If $\|\mathbf{e}\|_A > 0$,

$$\|\mathbf{e}\|_A \le \|\mathbf{x} - \mathbf{w}\|_A ,$$

as we claimed. □

Theorem (characterization of Galerkin approximations)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n$*, and* W *is a subspace of* $\mathbb{C}^n$*. Then, the following are equivalent:*

1. *The vector* $\mathbf{x}_W \in W$ *is a minimizer of* E_{A} *over* W*:*

$$\mathbf{x}_W = \operatorname*{argmin}_{z\in W} E_{\mathsf{A}}(\mathbf{z})$$

2. *The vector* $\mathbf{x}_W \in W$ *is a Galerkin approximation of* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*:*

$$(\mathsf{A}\mathbf{x}_W, \mathbf{w})_2 = (\mathbf{f}, \mathbf{w})_2, \quad \forall\, \mathbf{w} \in W.$$

Proof.

Exercise. □

Theorem (convergence)

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be HPD, $\mathbf{f} \in \mathbb{C}^n_\star$, and $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$. Suppose that $\{\mathbf{x}_k\}_{k=1}^{\infty}$ is the sequence generated by the zero-start CG algorithm. Then, there is an $m_\star \in \{1, \ldots, n\}$ for which

$$\mathbf{x}_k \neq \mathbf{x}, \quad 1 \leq k \leq m_\star - 1, \quad \mathbf{x}_k = \mathbf{x}, \quad k \geq m_\star,$$

and $\dim \mathcal{K}_k(\mathsf{A}, \mathbf{f}) = k$, for $k = 1, \ldots, m_\star$.

Proof.

Let $\mathcal{K}_m = \mathcal{K}_m(\mathsf{A}, \mathbf{f})$. Notice that $\dim \mathcal{K}_m \leq m$, so that if we show that equality actually holds, all the statements will follow.

We will proceed by induction. Set $m = 1$ and notice that, since $\mathbf{f} \neq \mathbf{0}$,

$$\mathcal{K}_1 = \operatorname{span}\{\mathbf{f}\} \quad \Longrightarrow \quad \dim \mathcal{K}_1 = 1.$$

Assume now that, for all $m = 1, \ldots, k$, with $k < n - 1$, we have $\dim \mathcal{K}_m = m$ and $\mathbf{x}_m \neq \mathbf{x}$. Therefore, the residual $\mathbf{r}_k = \mathbf{f} - \mathsf{A}\mathbf{x}_k \neq \mathbf{0}$ and $\mathbf{r}_k \in \mathcal{K}_{k+1}$.

Proof, Cont.

Notice that, using the characterization of Galerkin approximations given in a previous theorem, we have that $\mathbf{x}_k \in \mathcal{K}_k$ is the Galerkin approximation of $\mathbf{x}$ in in the subspace $\mathcal{K}_k$. Thus, the residual $\mathbf{r}_k$ must be orthogonal to $\mathcal{K}_k$, i.e.,

$$(\mathbf{r}_k, \mathbf{w})_2 = 0, \quad \forall \, \mathbf{w} \in \mathcal{K}_k.$$

In other words we have shown that $\mathbf{0} \neq \mathbf{r}_k \in \mathcal{K}_{k+1} \setminus \mathcal{K}_k$. This is only possible if $\dim \mathcal{K}_{k+1} > \dim \mathcal{K}_k$. In other words $\dim \mathcal{K}_{k+1} = k + 1$ and the result follows.

Finally, $\mathbf{x}$, the true solution must be in one of the Krylov subspaces, i.e., there is some $m_\star \leq n$, such that $\mathbf{x} \in \mathcal{K}_{m_\star}$.

□

Theorem (equivalence)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n_\star$, *and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$. *The sequence generated by the zero–start conjugate gradient method,* $\{\mathbf{x}_k\}_{k=1}^{m_\star}$, *is the same sequence as that generated by the following recursive algorithm: given* $\mathbf{x}_0 = \mathbf{0}$, *define* $\mathbf{r}_0 = \mathbf{f} - \mathsf{A}\mathbf{x}_0 = \mathbf{f}$ *and* $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{f}$. *For* $k = 0, \ldots, m_\star - 1$ *compute:*

1. *Update the iterate:*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_{k+1}\mathbf{p}_k, \quad \lambda_{k+1} = \frac{(\mathbf{r}_k, \mathbf{p}_k)_2}{(\mathsf{A}\mathbf{p}_k, \mathbf{p}_k)_2}.$$

2. *Update the residual:*

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \lambda_{k+1}\mathsf{A}\mathbf{p}_k.$$

3. *Update the search direction:*

$$\mathbf{p}_{k+1} = \mathbf{r}_{k+1} - \mu_{k+1}\mathbf{p}_k, \quad \mu_{k+1} = \frac{(\mathsf{A}\mathbf{r}_{k+1}, \mathbf{p}_k)_2}{(\mathsf{A}\mathbf{p}_k, \mathbf{p}_k)_2}.$$

4. *If* $k = m_\star - 1$, *stop. Otherwise, index* k *and go to step 1.*

Theorem (equivalence, Cont.)

Clearly,

$$\mathbf{r}_{m_\star} = \mathbf{0} = \mathbf{p}_{m_\star},$$

but it is guaranteed that

$$\mathbf{0} \neq \mathbf{r}_k \in \mathcal{K}_{k+1} \setminus \mathcal{K}_k, \quad \mathbf{0} \neq \mathbf{p}_k \in \mathcal{K}_{k+1} \setminus \mathcal{K}_k, \quad k = 0, \ldots, m_\star - 1,$$

and the following orthogonalities hold:

$$(\mathbf{r}_j, \mathbf{r}_i)_2 = \left(\mathbf{p}_j, \mathbf{p}_i\right)_{\mathsf{A}} = 0,$$

for all $0 \leq j < i \leq m_\star - 1$.

Proof.

The proof is a non-intuitive induction nightmare hellscape. See the book. □

A Couple of Corollaries

Corollary

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}_\star^n$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. The sequence generated by the zero–start conjugate gradient method,* $\{\mathbf{x}_k\}_{k=1}^{m_\star}$ *has the following property: for all* $k \in \{1, \ldots, m_\star\}$*,*

$$\mathbf{x}_k \in \mathcal{K}_k \setminus \mathcal{K}_{k-1},$$

which implies that

$$\langle \mathbf{x}_1, \ldots, \mathbf{x}_k \rangle = \mathcal{K}_k.$$

Corollary

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}_\star^n$ *and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. If the zero–start conjugate gradient (CG) algorithm is employed to produce the approximation sequence* $\{\mathbf{x}_j\}_{j=1}^{m_\star}$*, then, for all* $1 \le i \le m_\star$

$$\mathcal{K}_i(\mathsf{A}, \mathbf{f}) = \left\langle \mathbf{f}, \mathsf{A}\mathbf{f}, \ldots, \mathsf{A}^{i-1}\mathbf{f} \right\rangle = \langle \mathbf{x}_1, \ldots, \mathbf{x}_i \rangle = \langle \mathbf{p}_0, \ldots, \mathbf{p}_{i-1} \rangle = \langle \mathbf{r}_0, \ldots, \mathbf{r}_{i-1} \rangle .$$

Corollary (equivalent formulation)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n_\star$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. The sequence generated by the zero–start conjugate gradient method,* $\{\mathbf{x}_k\}_{k=1}^{m_\star}$*, is the same sequence as that generated by the following recursive algorithm: given* $\mathbf{x}_0 = \mathbf{0}$*, define* $\mathbf{r}_0 = \mathbf{f} - \mathsf{A}\mathbf{x}_0 = \mathbf{f}$ *and* $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{f}$*. For* $0 \leq k \leq m_\star - 1$ *compute:*

1. *Update the iterate:*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_{k+1}\mathbf{p}_k, \quad \lambda_{k+1} = \frac{(\mathbf{r}_k, \mathbf{r}_k)_2}{(\mathsf{A}\mathbf{p}_k, \mathbf{p}_k)_2}.$$

2. *Update the residual:*

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \lambda_{k+1}\mathsf{A}\mathbf{p}_k.$$

3. *Update the search direction:*

$$\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_{k+1}\mathbf{p}_k, \quad \beta_{k+1} = \frac{(\mathbf{r}_{k+1}, \mathbf{r}_{k+1})_2}{(\mathbf{r}_k, \mathbf{r}_k)_2}.$$

4. *If* $k = m_\star - 1$*, stop. Otherwise, index* k *and go to step 1.*

Corollary (equivalent formulation, Cont.)

It follows that

$$\mathbf{r}_{m_\star} = \mathbf{0} = \mathbf{p}_{m_\star},$$

but, it is guaranteed that

$$\mathbf{0} \neq \mathbf{r}_k \in \mathcal{K}_{k+1} \setminus \mathcal{K}_k, \quad \mathbf{0} \neq \mathbf{p}_k \in \mathcal{K}_{k+1} \setminus \mathcal{K}_k, \quad k = 0, \ldots, m_\star - 1.$$

The following orthogonalities hold:

$$(\mathbf{r}_j, \mathbf{r}_i)_2 = \left(\mathbf{p}_j, \mathbf{p}_i\right)_\mathsf{A} = 0,$$

for all $0 \leq j < i \leq m_\star - 1$.

Theorem (minimization)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n_\star$ *is given, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. Let, for some* $m \in \{1, \ldots, n\}$*,* $\{\mathbf{x}_i\}_{i=0}^m$*, denote any sequence of vectors with* $\mathbf{x}_0 = \mathbf{0}$ *— with associated residual vectors* $\mathbf{r}_j = \mathbf{f} - \mathsf{A}\mathbf{x}_j$ *— that satisfies*

$$\mathcal{K}_j = \mathcal{K}_j(\mathbf{f}, \mathsf{A}) = \langle \mathbf{f}, \mathsf{A}\mathbf{f}, \ldots, \mathsf{A}^{j-1}\mathbf{f} \rangle = \langle \mathbf{x}_1, \ldots, \mathbf{x}_j \rangle = \langle \mathbf{r}_0, \ldots, \mathbf{r}_{j-1} \rangle, \quad \mathbf{r}_{j-1} \neq \mathbf{0},$$

for all $j = 1, \ldots, m$*, with the orthogonality relations*

$$\mathbf{r}_k^{\mathsf{H}}\mathbf{r}_i = 0,$$

for all $0 \leq k < i \leq m$*. Then the* j*–th iterate* $\mathbf{x}_j$ *is the unique vector in* $\mathcal{K}_j$ *that minimizes the error function* $\phi(\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\mathsf{A}}$*. Furthermore,* ϕ *is monotonically decreasing:*

$$\|\mathbf{e}_j\|_{\mathsf{A}} = \|\mathbf{x} - \mathbf{x}_j\|_{\mathsf{A}} = \phi(\mathbf{x}_j) \leq \phi(\mathbf{x}_{j-1}) = \|\mathbf{x} - \mathbf{x}_{j-1}\|_{\mathsf{A}} = \|\mathbf{e}_{j-1}\|_{\mathsf{A}}.$$

Proof of minimization.

Let $\mathbf{z} \in \mathcal{K}_j$ be arbitrary, and define $\mathbf{w} = \mathbf{x}_j - \mathbf{z} \in \mathcal{K}_j$. Then

$$\begin{aligned}\phi^2(\mathbf{z}) &= \|\mathbf{x} - \mathbf{z}\|_{\mathsf{A}}^2 \\ &= \|\mathbf{x} - \mathbf{x}_j + \mathbf{x}_j - \mathbf{z}\|_{\mathsf{A}}^2 \\ &= \|\mathbf{e}_j + \mathbf{w}\|_{\mathsf{A}}^2 \\ &= (\mathbf{e}_j + \mathbf{w})^{\mathsf{H}}\mathsf{A}(\mathbf{e}_j + \mathbf{w}) \\ &= \|\mathbf{e}_j\|_{\mathsf{A}}^2 + 2\Re\left(\mathbf{w}^{\mathsf{H}}\mathsf{A}\mathbf{e}_j\right) + \|\mathbf{w}\|_{\mathsf{A}}^2 \\ &= \|\mathbf{e}_j\|_{\mathsf{A}}^2 + 2\Re\left(\mathbf{w}^{\mathsf{H}}\mathbf{r}_j\right) + \|\mathbf{w}\|_{\mathsf{A}}^2.\end{aligned}$$

Since $\mathbf{w} \in \mathcal{K}_j = \langle \mathbf{r}_0, \ldots, \mathbf{r}_{j-1} \rangle$, there exist unique $\alpha_0, \ldots, \alpha_{j-1} \in \mathbb{C}$ such that

$$\mathbf{w} = \sum_{i=0}^{j-1} = \alpha_i \mathbf{r}_i.$$

Proof, Cont.

Using the orthogonality $\mathbf{r}_i{}^{\mathsf{H}}\mathbf{r}_j = 0$, for all $0 \le i < j$, it is clear that $\mathbf{w}^{\mathsf{H}}\mathbf{r}_j = 0$. Hence

$$\phi^2(\mathbf{z}) = \|\mathbf{e}_j\|_{\mathsf{A}}^2 + \|\mathbf{w}\|_{\mathsf{A}}^2 \ \ge \|\mathbf{e}_j\|_{\mathsf{A}}^2 \ ,$$

with equality in the last relation if and only if $\mathbf{w} = \mathbf{0}$, or equivalently, if and only if $\mathbf{z} = \mathbf{x}_j$. Hence $\mathbf{x}_j$ is the unique minimizer of ϕ over $\mathcal{K}_j$.

Since we have the space containment $\mathcal{K}_{j-1}(\mathbf{f}, \mathsf{A}) \subseteq \mathcal{K}_j(\mathbf{f}, \mathsf{A})$ we must have

$$\begin{aligned}
\|\mathbf{e}_j\|_{\mathsf{A}} &= \phi(\mathbf{x}_j) \\
&= \inf\{\phi(\mathbf{z}) \mid \mathbf{z} \in \mathcal{K}_j(\mathbf{f}, \mathsf{A})\} \\
&\le \inf\{\phi(\mathbf{z}) \mid \mathbf{z} \in \mathcal{K}_{j-1}(\mathbf{f}, \mathsf{A})\} \\
&= \phi(\mathbf{x}_{j-1}) \\
&= \|\mathbf{e}_{j-1}\|_{\mathsf{A}} \, .
\end{aligned}$$

□

Theorem (polynomial bound)

Suppose the zero–start conjugate gradient algorithm is applied to solve $\mathsf{A}\mathbf{x} = \mathbf{f}$ *where* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD and* $\mathbf{f} \in \mathbb{C}^n_\star$*. Then, if the iteration has not already converged (*$\mathbf{r}_{i-1} \neq \mathbf{0}$*), there is a unique polynomial*

$$p_i \in \mathbb{P}_{i,\star} = \{p \in \mathbb{P}_i \mid p(0) = 1\}$$

that minimizes $\|p(\mathsf{A})\mathbf{e}_0\|_\mathsf{A}$ *over all* $p \in \mathbb{P}_{i,\star}$*. The iterate* $\mathbf{x}_i$ *has the error* $\mathbf{e}_i = p_i(\mathsf{A})\mathbf{e}_0$ *and, consequently,*

$$\frac{\|\mathbf{e}_i\|_\mathsf{A}}{\|\mathbf{e}_0\|_\mathsf{A}} \leq \inf_{p\in\mathbb{P}_{i,\star}} \max_{\lambda\in\sigma(\mathsf{A})} |p(\lambda)| .$$

Proof.

Recall that

$$E_\mathsf{A}(\mathbf{z}) = \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_\mathsf{A}^2 - \frac{1}{2}\mathbf{f}^\mathsf{H}\mathsf{A}^{-1}\mathbf{f},$$

where $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$ is the exact solution.

Proof, Cont.

Therefore, by definition of the zero–start CG algorithm, we have

$$\mathbf{x}_i = \operatorname*{argmin}_{z\in\mathcal{K}_i} \|\mathbf{x}-\mathbf{z}\|_{\mathsf{A}} , \quad \min_{z\in\mathcal{K}_i} \|\mathbf{x}-\mathbf{z}\|_{\mathsf{A}} = \|\mathbf{e}_i\|_{\mathsf{A}} ,$$

and $\mathbf{x}_i \in \mathcal{K}_i$ is uniquely determined. Now, observe that $\mathbf{e}_0 = \mathbf{x}$, since $\mathbf{x}_0 = \mathbf{0}$, and, consequently, $\mathbf{r}_0 = \mathbf{f}$. Thus, for any $\mathbf{z} \in \mathcal{K}_i$, there are constants $c_j \in \mathbb{C}$, $1 \le j \le i$, such that $\mathbf{z} = \sum_{j=1}^{i}(-c_j)\mathsf{A}^{j-1}\mathbf{f}$. Consequently,

$$\mathbf{x}-\mathbf{z} = \mathbf{x} + \sum_{j=1}^{i} c_j \mathsf{A}^{j-1}\mathbf{f} = \mathbf{e}_0 + \sum_{j=1}^{i} c_j \mathsf{A}^{j-1}\mathbf{r}_0 = \mathbf{e}_0 + \sum_{j=1}^{i} c_j \mathsf{A}^{j}\mathbf{e}_0 = p(\mathsf{A})\mathbf{e}_0,$$

where $p(t) = 1 + \sum_{j=1}^{i} c_j t^j \in \mathbb{P}_{i,\star}$. It follows, then, that the minimization problem above is equivalent to

$$p_i = \operatorname*{argmin}_{p\in\mathbb{P}_{i,\star}} \|p(\mathsf{A})\mathbf{e}_0\|_{\mathsf{A}} , \quad \min_{p\in\mathbb{P}_{i,\star}} \|p(\mathsf{A})\mathbf{e}_0\|_{\mathsf{A}} = \|\mathbf{e}_i\|_{\mathsf{A}} ,$$

and $p_i \in \mathbb{P}_{i,\star}$ is, of course, uniquely determined.

Proof, Cont.

It therefore follows that

$$\left\|\mathbf{e}_i\right\|_{\mathsf{A}} = \inf_{p\in\mathbb{P}_{i,\star}} \left\|p(\mathsf{A})\mathbf{e}_0\right\|_{\mathsf{A}} \leq \inf_{p\in\mathbb{P}_{i,\star}} \left\|p(\mathsf{A})\right\|_{\mathsf{A}} \left\|\mathbf{e}_0\right\|_{\mathsf{A}},$$

and, consequently,

$$\frac{\left\|\mathbf{e}_i\right\|_{\mathsf{A}}}{\left\|\mathbf{e}_0\right\|_{\mathsf{A}}} \leq \inf_{p\in\mathbb{P}_{i,\star}} \left\|p(\mathsf{A})\right\|_{\mathsf{A}}. \tag{5}$$

To finish up, suppose that $\mathbf{z} \in \mathbb{C}^n_\star$. Let $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ be an orthonormal basis of eigenvectors of A. Set $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\} \subset (0, \infty)$, with $\mathsf{A}\mathbf{w}_j = \lambda_j\mathbf{w}_j$, for $j = 1, \ldots, n$. There exist constants $\alpha_j \in \mathbb{C}$, $j = 1, \ldots, n$, such that $\mathbf{z} = \sum_{j=1}^n \alpha_j\mathbf{w}_j$, and

$$\left\|\mathbf{z}\right\|_{\mathsf{A}}^2 = \mathbf{z}^{\mathsf{H}}\mathsf{A}\mathbf{z} = \sum_{j=1}^n |\alpha_j|^2\lambda_j.$$

Furthermore,

$$\left\|p(\mathsf{A})\mathbf{z}\right\|_{\mathsf{A}}^2 = \mathbf{z}^{\mathsf{H}}p(\mathsf{A})^{\mathsf{H}}\mathsf{A}p(\mathsf{A})\mathbf{z} = \sum_{j=1}^n |\alpha_j|^2\lambda_j|p(\lambda_j)|^2.$$

Proof, Cont.

As a consequence,

$$\frac{\|p(\mathsf{A})\mathbf{z}\|_{\mathsf{A}}^2}{\|\mathbf{z}\|_{\mathsf{A}}^2} = \frac{\sum_{j=1}^n |\alpha_j|^2 \lambda_j |p(\lambda_j)|^2}{\sum_{j=1}^n |\alpha_j|^2 \lambda_j} \leq \max_{\lambda \in \sigma(\mathsf{A})} |p(\lambda)|^2,$$

which implies that

$$\|p(\mathsf{A})\|_{\mathsf{A}} \leq \max_{\lambda \in \sigma(\mathsf{A})} |p(\lambda)| \, .$$

Putting this estimate together with that in (5), we have

$$\frac{\|\mathbf{e}_i\|_{\mathsf{A}}}{\|\mathbf{e}_0\|_{\mathsf{A}}} \leq \inf_{p \in \mathbb{P}_{i,\star}} \max_{\lambda \in \sigma(\mathsf{A})} |p(\lambda)| \, ,$$

the desired result. □

Theorem (convergence)

Suppose the zero–start conjugate gradient algorithm is applied to solve $\mathsf{A}\mathbf{x} = \mathbf{f}$ *where* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD and* $\mathbf{f} \in \mathbb{C}^n_\star$*. If* A *has only* k *distinct eigenvalues,* $k < n$*, then the algorithm converges in at most* k *steps.*

Proof.

Let $\sigma(\mathsf{A}) = \{\lambda_j\}_{j=1}^k$ denote the set of k distinct eigenvalues of A. From the last theorem, for $i = 1, \ldots, k$,

$$\frac{\|\mathbf{e}_i\|_{\mathsf{A}}}{\|\mathbf{e}_0\|_{\mathsf{A}}} \leq \max_{\lambda\in\sigma(\mathsf{A})} |q_i(\lambda)| \,,$$

for any polynomial q_i of degree at most i, with the property that $q_i(0) = 1$. Let us define, for any $i = 1, \ldots, k$,

$$q_i(x) = \prod_{j=1}^{i} \left(1 - \frac{x}{\lambda_j}\right) .$$

Proof, Cont.

Clearly, $q_i(0) = 1$, and for all $j = 1, \ldots, i$

$$q_i(\lambda_j) = 0.$$

But, on the other hand, if $j = i + 1, \ldots, k$,

$$q_i(\lambda_j) \neq 0.$$

Now, once $i = k$,

$$q_i(\lambda) = 0, \quad \text{for all} \quad \lambda \in \sigma(\mathsf{A}).$$

Thus, $\|\mathbf{e}_k\|_{\mathsf{A}} = 0$. Of course, convergence could happen at an earlier stage if, by chance, $\mathbf{x} \in \mathcal{K}_i$, for some $i < k$. □

Theorem (convergence of CG)

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ *be HPD and* $\mathbf{f} \in \mathbb{C}^n_\star$. *The error for the zero–start conjugate gradient method satisfies*

$$\|\mathbf{e}_k\|_{\mathsf{A}} \leq 2\left(\frac{\sqrt{\kappa_2(\mathsf{A})}-1}{\sqrt{\kappa_2(\mathsf{A})}+1}\right)^k \|\mathbf{e}_0\|_{\mathsf{A}}.$$

Proof.

Suppose that $\sigma(\mathsf{A}) = \{\lambda_1, \ldots \lambda_n\}$, $0 < \lambda_1 \leq \cdots \leq \lambda_n$. It follows that

$$\|\mathbf{e}_k\|_{\mathsf{A}} \leq \max_{\lambda\in[\lambda_1,\lambda_n]} |q_k(\lambda)| \|\mathbf{e}_0\|_{\mathsf{A}},$$

for any polynomial q_k of degree at most k, such that $q_k(0) = 1$. Since this polynomial is arbitrary, we may choose it to minimize the right hand side of this expression. It turns out that shifted and rescaled versions of the classical Chebyshev polynomials minimize this choice.

Proof, Cont.

Namely, we set

$$q_k(t) = \frac{1}{T_k(1/\rho)} T_k\left(\frac{1}{\rho}\left(1 - \frac{2}{\lambda_1 + \lambda_n} t\right)\right), \quad \rho = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}$$

and

$$T_k(t) = \begin{cases} \cos(k \arccos(t)), & |t| \le 1, \\ \cosh(k \cosh^{-1}(t)), & |t| > 1. \end{cases}$$

With this choice, we obtain the bound

$$\max_{\lambda \in [\lambda_1, \lambda_n]} |q_k(\lambda)| \le \frac{1}{T_k(1/\rho)}.$$

Now, since $\rho < 1$ we set $\sigma = \cosh^{-1}(1/\rho)$ to see that

$$T_k(1/\rho) = \frac{1}{2}(e^{k\sigma} + e^{-k\sigma}) \ge \frac{1}{2} e^{k\sigma}.$$

Proof, Cont.

Using that $\sigma = \cosh^{-1}(1/\rho) = \ln\left(1/\rho + \sqrt{1/\rho^2 - 1}\right)$, we get

$$e^{k\sigma} = \left[\frac{1}{\rho}\left(1 + \sqrt{1-\rho^2}\right)\right]^k .$$

And, using that $\rho = \frac{\kappa_2(\mathsf{A})-1}{\kappa_2(\mathsf{A})+1}$, we then obtain

$$\frac{1}{T_k(1/\rho)} \leq 2e^{-k\sigma} = 2\left(\frac{\sqrt{\kappa_2(\mathsf{A})}-1}{\sqrt{\kappa_2(\mathsf{A})}+1}\right)^k ,$$

and the result follows. □

Non-Zero Starting Vectors

Definition (standard CG)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n_\star$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. The* **(standard) conjugate gradient method** *is an iterative scheme for producing a sequence of approximations* $\{\mathbf{x}_k\}_{k=1}^\infty$ *from the starting point* $\mathbf{x}_0 \in \mathbb{C}^n$ *according to the following recursive formula: setting* $\mathcal{K}_k = \mathcal{K}_k(\mathsf{A}, \mathbf{r}_0)$*, where* $\mathbf{r}_0 = \mathbf{f} - \mathsf{A}\mathbf{x}_0$*, the k–th iterate is obtained by*

$$\mathbf{x}_k = \mathbf{x}_k' + \mathbf{x}_0, \quad \mathbf{x}_k' = \operatorname*{argmin}_{\mathbf{z}\in\mathcal{K}_k} E_{\mathsf{A}}(\mathbf{z} + \mathbf{x}_0). \tag{6}$$

Proposition (equivalence)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD,* $\mathbf{f} \in \mathbb{C}^n_\star$, $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$, $\mathbf{x}_0 \in \mathbb{C}^n$, *and set*

$$\mathbf{x}' = \mathbf{x} - \mathbf{x}_0, \quad \mathbf{r}_0 = \mathbf{f} - \mathsf{A}\mathbf{x}_0 .$$

The sequence $\{\mathbf{x}'_k\}_{k=1}^\infty$ *generated by the zero–start conjugate gradient algorithm to approximate the solution to* $\mathsf{A}\mathbf{x}' = \mathbf{r}_0$ *is equivalent to the sequence* $\{\mathbf{x}_k\}_{k=1}^\infty$ *generated by the (standard) conjugate gradient algorithm with the starting vector* $\mathbf{x}_0$ *to approximate the solution to* $\mathsf{A}\mathbf{x} = \mathbf{f}$ *in the sense that*

$$\mathbf{x}_k = \mathbf{x}_0 + \mathbf{x}'_k \quad \textit{and} \quad \mathbf{x} - \mathbf{x}_k = \mathbf{x}' - \mathbf{x}'_k .$$

Furthermore, as long as $\mathbf{x}_0 \neq \mathbf{x}$, *there is an integer* $m_\star \in \{1, \ldots, n\}$ *such that*

$$\mathbf{x}_k \neq \mathbf{x}, \quad k = 1, \ldots, m_\star - 1, \qquad \mathbf{x}_k = \mathbf{x}, \quad k \geq m_\star .$$

Advantage of Non-Zero Starts

The last result states that approximating $\mathsf{A}\mathbf{x} = \mathbf{f}$ using the standard conjugate gradient algorithm is exactly equivalent to approximating $\mathsf{A}\mathbf{x}' = \mathbf{r}_0$ using the zero–start conjugate gradient algorithm. In short, their convergence properties are the same. Of course, often it is advantageous to use a nonzero starting vector, especially in the case where one already has a good approximation to the exact solution of an equation of interest.

Preconditioned Conjugate Gradient

Proposition (some useful facts)

Suppose that $\mathsf{A}, \mathsf{B} \in \mathbb{C}^{n\times n}$ *are HPD,* $\mathbf{f} \in \mathbb{C}^n_\star$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f} \in \mathbb{C}^n_\star$*. Let* $\mathsf{B} = \mathsf{L}^{\mathsf{H}}\mathsf{L}$ *be a Cholesky–type factorization for* B*, where* $\mathsf{L} \in \mathbb{C}^{n\times n}$ *is invertible. Define*

$$\mathsf{C} = \mathsf{L}^{-\mathsf{H}}\mathsf{A}\mathsf{L}^{-1}.$$

Then, C *is HPD and* $\mathsf{B}^{-1}\mathsf{A}$ *is similar to* C*. Consequently,*

$$\sigma(\mathsf{B}^{-1}\mathsf{A}) = \sigma(\mathsf{C}) \subset (0, \infty).$$

Furthermore, the following problems are equivalent:

1. *Find* $\mathbf{x} \in \mathbb{C}^n$ *such that*
$$\mathsf{A}\mathbf{x} = \mathbf{f}.$$
2. *Find* $\mathbf{x} \in \mathbb{C}^n$ *such that*
$$\mathsf{B}^{-1}\mathsf{A}\mathbf{x} = \mathsf{B}^{-1}\mathbf{f}.$$
3. *Find* $\mathbf{x} \in \mathbb{C}^n$ *such that*
$$\mathsf{L}\mathbf{x} = \mathbf{y}, \quad \mathsf{C}\mathbf{y} = \mathbf{q},$$
where $\mathbf{q} = \mathsf{L}^{-\mathsf{H}}\mathbf{f}$*.*

The equation $\mathsf{C}\mathbf{y} = \mathbf{q}$ *is called the preconditioned system.*

Definition

Suppose that $\mathsf{A}, \mathsf{B} \in \mathbb{C}^{n\times n}$ are HPD, $\mathbf{f} \in \mathbb{C}^n_\star$, and $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$. Assume that C, L, $\mathbf{q}$, and $\mathbf{y}$ are as defined in Proposition 3.1. The **zero–start** B**–preconditioned conjugate gradient (PCG) method** is an iterative scheme for producing the sequence $\{\mathbf{y}_k\}_{k=1}^\infty$ by applying the standard zero–start conjugate gradient method to approximate the solution to the preconditioned system, $\mathsf{C}\mathbf{y} = \mathbf{q}$. The sequence of approximations for the solution of interest, $\mathbf{x}$, denoted $\{\mathbf{x}_k\}_{k=1}^\infty$, is defined by $\mathbf{x}_k = \mathsf{L}^{-1}\mathbf{y}_k$.

Proposition (equivalence I)

Suppose that $\mathsf{A}, \mathsf{B} \in \mathbb{C}^{n\times n}$ *are HPD,* $\mathbf{f} \in \mathbb{C}^n_\star$*, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. Assume that* C*,* L*,* $\mathbf{q}$*, and* $\mathbf{y}$ *are as defined in Proposition 3.1. The sequence,* $\{\mathbf{y}_k\}_{k=1}^\infty$*, generated by the zero–start* B*–PCG method is the same sequence as that generated by the following recursive algorithm: given* $\mathbf{y}_0 = \mathbf{0}$*, define* $\mathbf{s}_0 = \mathbf{q} - \mathsf{C}\mathbf{y}_0 = \mathbf{q}$ *and* $\mathbf{d}_0 = \mathbf{s}_0 = \mathbf{q}$*. For* $k \geq 0$ *compute:*

1. *Update the iterate:*

$$\mathbf{y}_{k+1} = \mathbf{y}_k + \theta_{k+1}\mathbf{d}_k, \quad \theta_{k+1} = \frac{(\mathbf{s}_k, \mathbf{s}_k)_2}{(\mathsf{C}\mathbf{d}_k, \mathbf{d}_k)_2}.$$

2. *Update the residual:*

$$\mathbf{s}_{k+1} = \mathbf{s}_k - \theta_{k+1}\mathsf{C}\mathbf{d}_k.$$

3. *Update the search direction:*

$$\mathbf{d}_{k+1} = \mathbf{s}_{k+1} + \nu_{k+1}\mathbf{d}_k, \quad \nu_{k+1} = \frac{(\mathbf{s}_{k+1}, \mathbf{s}_{k+1})_2}{(\mathbf{s}_k, \mathbf{s}_k)_2}.$$

4. *If* $\mathbf{s}_{k+1} = \mathbf{0}$ *stop. Otherwise, index* k *and goto step 1.*

Proposition (equivalence I, Cont.)

Define $\mathcal{M}_k = \mathcal{K}_k(\mathsf{C}, \mathbf{q})$*. Then, there is an integer* $m_\star^{\mathsf{C}} \in \{1, \ldots, n\}$ *such that*

$$\mathbf{s}_{m_\star^{\mathsf{C}}} = \mathbf{0} = \mathbf{d}_{m_\star^{\mathsf{C}}},$$

and

$$\mathbf{0} \neq \mathbf{s}_k \in \mathcal{M}_{k+1} \setminus \mathcal{M}_k, \quad \mathbf{0} \neq \mathbf{d}_k \in \mathcal{M}_{k+1} \setminus \mathcal{M}_k, \quad k = 0, \cdots, m_\star^{\mathsf{C}} - 1.$$

Furthermore, the following orthogonalities hold:

$$(\mathbf{s}_j, \mathbf{s}_i)_2 = (\mathbf{d}_j, \mathbf{d}_i)_{\mathsf{C}} = 0,$$

for all $0 \le j < i \le m_\star^{\mathsf{C}} - 1$*. Finally, the following convergence estimate holds:*

$$\|\mathbf{y} - \mathbf{y}_k\|_{\mathsf{C}} \le 2 \left(\frac{\sqrt{\kappa_2(\mathsf{C})} - 1}{\sqrt{\kappa_2(\mathsf{C})} + 1} \right)^k \|\mathbf{y} - \mathbf{y}_0\|_{\mathsf{C}}.$$

Corollary (equivalence II)

With the same assumptions and notation as in the last proposition, define, for $0 \leq k \leq m_\star^{\mathrm{C}}$,

$$\mathbf{x}_k = \mathrm{L}^{-1}\mathbf{y}_k, \quad \mathbf{r}_k = \mathrm{L}^{\mathrm{H}}\mathbf{s}_k, \quad \mathbf{p}_k = \mathrm{L}^{-1}\mathbf{d}_k.$$

These vectors may be generated directly by the following recursive algorithm: $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{f}$ *and* $\mathbf{p}_0 = \mathrm{B}^{-1}\mathbf{f}$. *For* $0 \leq k \leq m_\star^{\mathrm{C}} - 1$:

1. *Update the iterate:*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \theta_{k+1}\mathbf{p}_k, \quad \theta_{k+1} = \frac{(\mathrm{B}^{-1}\mathbf{r}_k, \mathbf{r}_k)_2}{(\mathrm{A}\mathbf{p}_k, \mathbf{p}_k)_2}.$$

2. *Update the residual:*

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \theta_{k+1}\mathrm{A}\mathbf{p}_k.$$

3. *Update the search direction:*

$$\mathbf{p}_{k+1} = \mathrm{B}^{-1}\mathbf{r}_{k+1} + \nu_k\mathbf{p}_k, \quad \nu_{k+1} = \frac{(\mathrm{B}^{-1}\mathbf{r}_{k+1}, \mathbf{r}_{k+1})_2}{(\mathrm{B}^{-1}\mathbf{r}_k, \mathbf{r}_k)_2}.$$

4. *If* $k = m_\star^{\mathrm{C}} - 1$, *stop. Otherwise, index* k *and goto step 1.*

Corollary (equivalence II, Cont.)

The following error estimate is valid:

$$\|\mathbf{x}-\mathbf{x}_k\|_{\mathsf{A}} \leq 2\left(\frac{\sqrt{\kappa_2(\mathsf{C})}-1}{\sqrt{\kappa_2(\mathsf{C})}+1}\right)^k \|\mathbf{x}-\mathbf{x}_0\|_{\mathsf{A}}.$$

Choosing Preconditioners

Given the HPD matrix $\mathsf{A} \in \mathbb{C}^{n\times n}$, how do we choose, practically, the HPD preconditioner $\mathsf{B} \in \mathbb{C}$?

This is a very hard problem and is treated in Math 673.

In general, there is no one right answer to the question. One must carefully examine the class of matrices from which A comes.

For example, for the common tridiagonal stiffness matrix, Gil Strang's favorite matrix, there is a certain class of preconditioners that work well. But these preconditioners might not work well at all for a matrix of another class.