Section 02. Simple Iterative Solvers

Linear Algebraic Solvers

A fundamental problem in scientific computing:

Given a large sparse matrix $A \in \mathbb{R}^{N \times N}$ and $\vec{f} \in \mathbb{R}^N$, how to solve $A\vec{u} = \vec{f}$?

Why is it so difficult?

- Large number of unknowns
- Sometimes ill-conditioned
- PDE-system with several different physical variables
- Linear solution is usually the bottleneck in simulation
- Linear solution is difficult to scale (optimality in terms of cost and parallel scalability)

Goals:

accuracy, convergence, robustness, efficiency, applicability, scalability, optimality, reliability, user-friendliness, cost-effectiveness, ...

Iterative Solvers

Iterative methods

- A very long history: Newton, Euler, Guass, ...
- Stationary iterative methods, Krylov subspace methods, domain decomposition, multigrid, ...
- Steepest descent method, Newton's method, power method, ...

Preconditioning methods

- A relatively long history: Turing 1948, Meijerink, van der Vorst 1977 (IC preconditioner) ...
- ILU, Sparse Approximate Inverse, Diagonal, SGS, SSOR, ...
- Domain decomposition, RAS, multigrid, nonlinear preconditioning, ...
- Problem dependent, usually requires at least the coefficient matrix A

Smoothers / local relaxation methods

- Employed by multigrid, or more generally, subspace correction methods
- Usually just simple iterative methods, sometimes not even convergent methods

KSM's as iterative methods or accelerators; linear stationary method as preconditioners or smoothers.

Linear Stationary Iterative Solvers

Abstract problem setting:

Let V be a finite-dimensional linear vector space, $\mathcal{A} : V \mapsto V$ be a non-singular linear operator, and $f \in V$. Find a $u \in V$, such that

$$\mathcal{A}u = f. \tag{10}$$

Algorithm (Stationary iterative method $u^{\text{new}} = ITER(u^{\text{old}})$)

1. Form the residual: $r = f - \mathcal{A}u^{\text{old}}$
2. Solve or approximate the error equation: $\mathcal{A}e = r$ by $\hat{e} = \mathcal{B}r$
3. Correct the previous iterative solution: $u^{\text{new}} = u^{\text{old}} + \hat{e}$

That is to say, the new iteration is obtained by computing

$$u^{\text{new}} = u^{\text{old}} + \mathcal{B}(f - \mathcal{A}u^{\text{old}}), \tag{11}$$

where $\mathcal{B}$ is called the iterator. Apparently, $\mathcal{B} = \mathcal{A}^{-1}$ for nonsingular operator $\mathcal{A}$ also defines an iterator, which yields a direct method. We wish to construct $\mathcal{B} \approx \mathcal{A}^{-1}$, but easier to compute.

Some Simple Examples

Consider the linear system $A\vec{u} = \vec{f}$. Assume the coefficient matrix $A \in \mathbb{R}^{N\times N}$ can be partitioned as $A = L + D + U$, where the three matrices $L, D, U \in \mathbb{R}^{N\times N}$.

Example (Richardson method)

The simplest iterative method for solving $A\vec{u} = \vec{f}$ might be the Richardson method

$$\vec{u}^{\text{new}} = \vec{u}^{\text{old}} + \omega\Big(\vec{f} - A\vec{u}^{\text{old}}\Big).$$

We can choose an optimal weight ω to improve performance of this method.

Example (Weighted Jacobi method)

The weighted or damped Jacobi method can be written as

$$\vec{u}^{\text{new}} = \vec{u}^{\text{old}} + \omega D^{-1}(\vec{f} - A\vec{u}^{\text{old}}).$$

This method solves one equation for one variable at a time, simultaneously.

More Simple Examples

Example (Gauss–Seidel method)

The Gauss–Seidel (G-S) method can be written as

$$\vec{u}^{\,\text{new}} = \vec{u}^{\,\text{old}} + (D+L)^{-1}(\vec{f} - A\vec{u}^{\,\text{old}}).$$

Thus we have

$$\vec{u}^{\,\text{new}} = \vec{u}^{\,\text{old}} + D^{-1}\Big(\vec{f} - L\vec{u}^{\,\text{new}} - (D+U)\vec{u}^{\,\text{old}}\Big).$$

Example (Successive over-relaxation method)

The successive over-relaxation (SOR) method can be written as

$$(D+\omega L)\vec{u}^{\,\text{new}} = \omega\vec{f} - \Big(\omega U + (\omega-1)D\Big)\vec{u}^{\,\text{old}}.$$

The weight ω is usually in $(1,2)$. This is in fact the extrapolation of $\vec{u}^{\,\text{old}}$ and $\vec{u}^{\,\text{new}}$ obtained in the G-S method. If $\omega = 1$, then it reduces to the G-S method.

Smoothing Effect of Richardson Method

Error components: Assume that

$$A\vec{\xi}^k = \lambda_k \vec{\xi}^k, \quad k = 1, \ldots, N,$$

where $0 < \lambda_1 \leq \cdots \leq \lambda_N$. $\{\vec{\xi}^k\}_{k=1}^N$ forms a basis of $\mathbb{R}^N$. We can then write

$$\vec{u} - \vec{u}^{(m)} = \sum_{k=1}^N \alpha_k^{(m)} \vec{\xi}^k.$$

Error propagation equation of the Richardson method ($\omega = \frac{1}{\lambda_N}$):

$$\begin{aligned}
\vec{u} - \vec{u}^{(m)} &= (I - \omega A)(\vec{u} - \vec{u}^{(m-1)}) = \cdots = (I - \omega A)^m (\vec{u} - \vec{u}^{(0)}). \\
&\Longrightarrow \sum_{k=1}^N \alpha_k^{(m)} \vec{\xi}^k = (I - \omega A)^m \sum_{k=1}^N \alpha_k^{(0)} \vec{\xi}^k = \sum_{k=1}^N \alpha_k^{(0)} (1 - \omega\lambda_k)^m \vec{\xi}^k \\
&\Longrightarrow \alpha_k^{(m)} = (1 - \omega\lambda_k)^m \alpha_k^{(0)} = \left(1 - \frac{\lambda_k}{\lambda_N}\right)^m \alpha_k^{(0)}, \quad k = 1, \ldots, N
\end{aligned}$$

We observe that the Richardson method converges very fast for high-frequency error components (large k) but very slow for low-frequency components (small k).

Smoothing Effect of Jacobi Method

Use the weighted Jacobi method with $\omega = 2/3$ to solve the problem $A\vec{u} = \vec{0}$:

- If the initial guess just equal to the eigenvector $\vec{\xi}^{\max}$ of $\lambda_{\max}$, the convergence is fast.
- If begin with a different initial guess $\vec{\xi}^{\min}$, the convergence becomes slow.

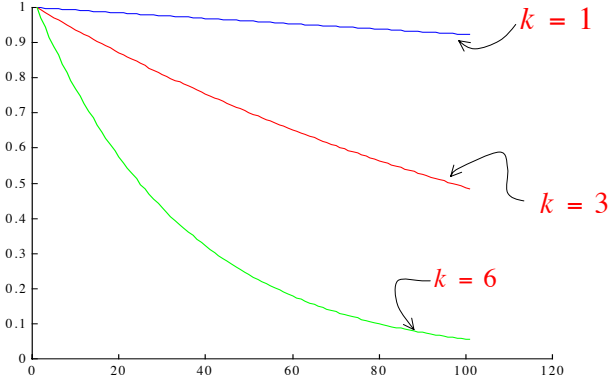

Figure: Error decay in $\|\cdot\|_\infty$-norm for weighted Jacobi (Richardson) method with initial guess $\vec{\xi}^{k}$.

Inner Product and Symmetry

Inner product: $(u, v) := \int_\Omega uv\, dx$ and $(u, v) := \sum_{i=1}^N u_i v_i$, if $V = \mathbb{R}^N$.

Symmetry: Define the adjoint operator (transpose) of the linear operator $\mathcal{A}$ as $\mathcal{A}^T : V \mapsto V$, such that

$$(\mathcal{A}^T u, v) := (u, \mathcal{A}v), \quad \forall u, v \in V.$$

A linear operator $\mathcal{A}$ on V is symmetric if and only if

$$(\mathcal{A}u, v) = (u, \mathcal{A}v), \quad \forall u, v \in \text{domain}(\mathcal{A}) \subseteq V.$$

If $\mathcal{A}$ is densely defined and $\mathcal{A}^T = \mathcal{A}$, then $\mathcal{A}$ is a self-adjoint operator.

Null space and range: Denote the null space and the range of $\mathcal{A}$ as

$$\begin{aligned} \mathcal{N}(\mathcal{A}) = \text{null}(\mathcal{A}) &:= \{v \in V \,:\, \mathcal{A}v = 0\}, \\ \mathcal{R}(\mathcal{A}) = \text{range}(\mathcal{A}) &:= \{u = \mathcal{A}v \,:\, v \in V\}. \end{aligned}$$

The null space is also called the kernel space and the range is called the image space. We have

$$\text{null}(\mathcal{A}^T)^\perp = \overline{\text{range}(\mathcal{A})} \quad \text{and} \quad \text{null}(\mathcal{A}^T) = \text{range}(\mathcal{A})^\perp.$$

Spectrum and Condition Number

Spectrum: The set of eigenvalues of $\mathcal{A}$ is called the spectrum, denoted as $\sigma(\mathcal{A})$. The spectrum of any bounded symmetric $\mathcal{A}$ is real, i.e., all eigenvalues are real, although a symmetric operator may have no eigenvalues.

Spectral radius: $\rho(\mathcal{A}) := \sup\{|\lambda| : \lambda \in \sigma(\mathcal{A})\}$. We have that

$$\lambda_{\min}(\mathcal{A}) = \min_{v \in V\setminus\{0\}} \frac{(\mathcal{A}v, v)}{\|v\|^2} \quad \text{and} \quad \lambda_{\max}(\mathcal{A}) = \max_{v \in V\setminus\{0\}} \frac{(\mathcal{A}v, v)}{\|v\|^2}.$$

Symmetric positive definite: $\mathcal{A}$ is called an SPD if and only if $\mathcal{A}$ is symmetric and $(\mathcal{A}v, v) > 0$, for any $v \in V\setminus\{0\}$. Since $\mathcal{A}$ is SPD, all of its eigenvalues are positive real numbers. This will be the main problem class for this lecture.

Spectral condition number: $\kappa(\mathcal{A}) := \lambda_{\max}(\mathcal{A})/\lambda_{\min}(\mathcal{A})$.

- For isomorphic mapping $\mathcal{A} \in \mathscr{L}(V;V)$, $\kappa(\mathcal{A}) := \|\mathcal{A}\|_{\mathscr{L}(V;V)}\|\mathcal{A}^{-1}\|_{\mathscr{L}(V;V)}$.
- For indefinite problems, we can define $\kappa(\mathcal{A}) := \frac{\sup_{\lambda\in\sigma(\mathcal{A})}|\lambda|}{\inf_{\lambda\in\sigma(\mathcal{A})}|\lambda|}$.
- All these definitions are consistent for symmetric positive definite problems.

Condition Number Analysis

Lemma (Estimation of condition number)

If μ_0 and μ_1 are positive constants satisfying

$$\mu_0(\mathcal{A}u, u) \leq (\mathcal{B}^{-1}u, u) \leq \mu_1(\mathcal{A}u, u), \quad \forall u \in V, \tag{12}$$

then the condition number

$$\kappa(\mathcal{B}\mathcal{A}) \leq \mu_1/\mu_0.$$

Lemma (Some equivalent conditions)

If $\mathcal{A}$ and $\mathcal{B}$ are SPD operators on V, then we have the inequalities (12) are equivalent to

$$\mu_0(\mathcal{B}u, u) \leq (\mathcal{A}^{-1}u, u) \leq \mu_1(\mathcal{B}u, u), \quad \forall u \in V, \tag{13}$$

or

$$\mu_1^{-1}(\mathcal{A}u, u) \leq (\mathcal{A}\mathcal{B}\mathcal{A}u, u) \leq \mu_0^{-1}(\mathcal{A}u, u), \quad \forall u \in V, \tag{14}$$

or

$$\mu_1^{-1}(\mathcal{B}u, u) \leq (\mathcal{B}\mathcal{A}\mathcal{B}u, u) \leq \mu_0^{-1}(\mathcal{B}u, u), \quad \forall u \in V. \tag{15}$$

An Alternative Inner Product

$\mathcal{A}$-inner product: If $\mathcal{A}$ is an SPD operator, it induces a new inner product (and a norm as well):

$$(u, v)_{\mathcal{A}} := (\mathcal{A}u, v) \qquad \forall u, v \in V.$$

Symmetry with respect to $\mathcal{A}$-inner product:
For any bounded linear operator $\mathcal{B} : V \mapsto V$, we can define two transposes with respect to the inner products $(\cdot, \cdot)$ and $(\cdot, \cdot)_{\mathcal{A}}$, respectively; namely,

$$(\mathcal{B}^T u, v) = (u, \mathcal{B}v), \qquad (\mathcal{B}^* u, v)_{\mathcal{A}} = (u, \mathcal{B}v)_{\mathcal{A}}.$$

Relations between these two symmetries: By the above definitions, it is easy to show that

- $\mathcal{B}^* = \mathcal{A}^{-1}\mathcal{B}^T\mathcal{A}$
- $(\mathcal{B}\mathcal{A})^* = \mathcal{B}^T\mathcal{A}$
- If $\mathcal{B}^T = \mathcal{B}$, we do not necessarily have $(\mathcal{B}\mathcal{A})^T = \mathcal{B}\mathcal{A}$; but ...
- If $\mathcal{B}^T = \mathcal{B}$, we have a key identity $(\mathcal{B}\mathcal{A})^* = \mathcal{B}^T\mathcal{A} = \mathcal{B}\mathcal{A}$.

Lemma (Spectral radius of self-adjoint operators)

If $\mathcal{B}^T = \mathcal{B}$, then $\rho(\mathcal{B}) = \|\mathcal{B}\|$. Similarly, if $\mathcal{B}^* = \mathcal{B}$, then $\rho(\mathcal{B}) = \|\mathcal{B}\|_{\mathcal{A}}$.

Convergence of Iterative Methods

For the linear stationary iterative method, it is easy to see that

$$u - u^{(m)} = (\mathcal{I} - \mathcal{B}\mathcal{A})\big(u - u^{(m-1)}\big) = \cdots = (\mathcal{I} - \mathcal{B}\mathcal{A})^m\big(u - u^{(0)}\big) = \mathcal{E}^m\big(u - u^{(0)}\big),$$

where $\mathcal{I} : V \mapsto V$ is the identity operator and the operator $\mathcal{E} := \mathcal{I} - \mathcal{B}\mathcal{A}$ is called the error propagation operator (or error reduction operator or iterative operator).

Lemma (Spectral radius and $\mathcal{A}$-norm)

If $\mathcal{A}$ is SPD and $\mathcal{B}$ is symmetric, then $\rho(\mathcal{I} - \mathcal{B}\mathcal{A}) = \|\mathcal{I} - \mathcal{B}\mathcal{A}\|_{\mathcal{A}}$.

Theorem (Convergence of linear stationary method)

The iteration converges for any initial guess if the spectral radius $\rho(\mathcal{I} - \mathcal{B}\mathcal{A}) < 1$, which is equivalent to $\lim_{m\to+\infty}(\mathcal{I} - \mathcal{B}\mathcal{A})^m = 0$. $\Longrightarrow$ The converse direction is also true.

- If $\mathcal{A}$ and $\mathcal{B}$ are both SPD, the eigenvalues of $\mathcal{B}\mathcal{A}$ are real and the spectral radius satisfies that $\rho(\mathcal{I} - \mathcal{B}\mathcal{A}) = \max\big(\lambda_{\max}(\mathcal{B}\mathcal{A}) - 1,\, 1 - \lambda_{\min}(\mathcal{B}\mathcal{A})\big)$. How eigenvalues distribute?
- It is important to note that the spectral radius of $\mathcal{E}$ only reflects the asymptotic convergence behavior of the iterative method.

Simple Example: Spectral Radius and Convergence Behavior

Suppose we have an iterative method with an error propagation matrix

$$E := \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 \\ 0 & \cdots & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{N\times N}$$

and the initial error is $\vec{e}^{(0)} := \vec{u} - \vec{u}^{(0)} = (0, \ldots, 0, 1)^T \in \mathbb{R}^N$. Notice that $\rho(E) \equiv 0$ in this example. However, if applying this error propagation matrix to form a sequence of approximations, we will find the convergence is actually very slow for a large N. In fact,

$$\|\vec{e}^{(0)}\|_2 = \|\vec{e}^{(1)}\|_2 = \cdots = \|\vec{e}^{(N-1)}\|_2 = 1 \qquad \text{and} \qquad \|\vec{e}^{(N)}\|_2 = 0.$$

Analyzing the spectral radius of the iterative matrix alone will not provide much useful information about the speed of an iterative method.

Stationary Methods as Preconditioners

Remark (Another equivalent condition)

If $\mathcal{A}$ and $\mathcal{B}$ are symmetric positive definite operators on a finite-dimensional space V, $\alpha > 0$ and $0 < \delta < 1$, then it is easy to verify the following two conditions are equivalent:

$$-\alpha(\mathcal{A}u, u) \leq (\mathcal{A}(\mathcal{I} - \mathcal{B}\mathcal{A})u, u) \leq \delta(\mathcal{A}u, u), \quad \forall u \in V \tag{16}$$

and

$$(1+\alpha)^{-1}(\mathcal{A}u, u) \leq (\mathcal{B}^{-1}u, u) \leq (1-\delta)^{-1}(\mathcal{A}u, u), \quad \forall u \in V. \tag{17}$$

Let $\mathcal{B}$ be a symmetric iterator for the SPD operator $\mathcal{A}$. An iterative method is convergent if

$$\rho := \rho(\mathcal{I} - \mathcal{B}\mathcal{A}) = \|\mathcal{I} - \mathcal{B}\mathcal{A}\|_{\mathcal{A}} < 1.$$

The method is not only converging but also a contraction, i.e.,

$$\|u - u^{(m)}\|_{\mathcal{A}} \leq \rho^m \|u - u^{(0)}\|_{\mathcal{A}} \to 0 \quad \text{as } m \to +\infty.$$

Furthermore, by definition, we have

$$\Big((\mathcal{A} - 2\mathcal{A}\mathcal{B}\mathcal{A} + \mathcal{A}\mathcal{B}\mathcal{A}\mathcal{B}\mathcal{A})u, u\Big) \leq \rho^2 (u, u)_{\mathcal{A}}, \quad \forall u \in V.$$

Gradient Descent Method

If $\mathcal{A}$ is SPD, the linear system (first-order optimality) is equivalent to a quadratic min problem.

Problem (Quadratic minimization problem)

Let $\mathcal{A} : V \mapsto V$ be an SPD operator. Consider the following convex minimization problem:

$$\min_{u \in V} \mathcal{F}(u) := \frac{1}{2}(\mathcal{A}u, u) - (f, u). \tag{18}$$

Suppose we have an initial approximation u^{old} and construct a new approximation

$$u^{\text{new}} = u^{\text{old}} + \alpha\, p$$

with a given nonzero search direction $p \in V$ and a stepsize α. In order to find the "best possible" stepsize, we can solve a one-dimensional problem (i.e., the exact line-search method):

$$\min_{\alpha \in \mathbb{R}} \mathcal{F}(\alpha) := \frac{1}{2}\left(u^{\text{old}} + \alpha p, u^{\text{old}} + \alpha p\right)_{\mathcal{A}} - (f, u^{\text{old}} + \alpha p).$$

Gradient Descent and Richardson

By simple calculation, we obtain the following quadratic form

$$\mathcal{F}(\alpha) := \frac{1}{2}\alpha^2(\mathcal{A}p, p) - \alpha(f - \mathcal{A}u^{\text{old}}, p) + \frac{1}{2}\big(\mathcal{A}u^{\text{old}}, u^{\text{old}}\big) - (f, u^{\text{old}}),$$

and the optimal stepsize is

$$\alpha_{\text{opt}} = \frac{(f - \mathcal{A}u^{\text{old}}, p)}{(\mathcal{A}p, p)} = \frac{(r^{\text{old}}, p)}{(\mathcal{A}p, p)}, \quad \text{with } r^{\text{old}} = f - \mathcal{A}u^{\text{old}}. \tag{19}$$

"Best" search direction: $p := -\nabla\mathcal{F}(u^{\text{old}}) = r^{\text{old}} \quad \Longrightarrow \quad$ Steepest descent direction

Remark (Steepest descent and Richardson methods)

If A is a SPD matrix, then $A\vec{u} = \vec{f}$ is equivalent to the quadratic minimization problem

$$\underset{\vec{u}\in\mathbb{R}^N}{\operatorname{argmin}} \frac{1}{2}\vec{u}^T A\vec{u} - \vec{f}^T\vec{u}.$$

The search direction in the Richardson method is exactly the same as the steepest decent method.

Convergence of Gradient Descent Method

Theorem (Convergence rate of gradient descent method)

If we apply the exact line-search using the stepsize

$$\alpha_m := \frac{\left(r^{(m)}, r^{(m)}\right)}{\left(r^{(m)}, r^{(m)}\right)_{\mathcal{A}}},$$

then the convergence rate of the SD method satisfies that

$$\left\|u - u^{(m)}\right\|_{\mathcal{A}} \leq \left(\frac{\kappa(\mathcal{A}) - 1}{\kappa(\mathcal{A}) + 1}\right)^m \left\|u - u^{(0)}\right\|_{\mathcal{A}}. \tag{20}$$

- The method is simple, robust, and easy to implement;
- It is cheap to compute at each iteration;
- But it is too slow in practice;
- We will discuss a few improvements in this section.

Symmetrized Iterative Methods

Algorithm (Symmetrized iterative method $u^{\text{new}} = SITER(u^{\text{old}})$)

1. $u^{(m+\frac{1}{2})} = u^{(m)} + \mathcal{B}\big(f - \mathcal{A}u^{(m)}\big)$
2. $u^{(m+1)} = u^{(m+\frac{1}{2})} + \mathcal{B}^T\big(f - \mathcal{A}u^{(m+\frac{1}{2})}\big)$

In turn, we obtain a symmetrized iterative method

$$u - u^{(m+1)} = (\mathcal{I} - \mathcal{B}^T\mathcal{A})(\mathcal{I} - \mathcal{B}\mathcal{A})(u - u^{(m)}) = (\mathcal{I} - \mathcal{B}\mathcal{A})^*(\mathcal{I} - \mathcal{B}\mathcal{A})(u - u^{(m)}).$$

If this new method satisfies the relation

$$u - u^{(m+1)} = (\mathcal{I} - \overline{\mathcal{B}}\mathcal{A})(u - u^{(m)}),$$

then it has a symmetric iteration operator

$$\overline{\mathcal{B}} := \mathcal{B}^T + \mathcal{B} - \mathcal{B}^T\mathcal{A}\mathcal{B} = \mathcal{B}^T(\mathcal{B}^{-T} + \mathcal{B}^{-1} - \mathcal{A})\mathcal{B} =: \mathcal{B}^T\mathcal{K}\mathcal{B}.$$

Lemma (Error decay property)

For any $v \in V$, we have $\|v\|_{\mathcal{A}}^2 - \|(\mathcal{I} - \mathcal{B}\mathcal{A})v\|_{\mathcal{A}}^2 = \big(\overline{\mathcal{B}}\mathcal{A}v, v\big)_{\mathcal{A}}$.

Effect of Symmetrization

We notice that $\overline{\mathcal{B}}^T = \overline{\mathcal{B}}$ and $(\mathcal{I} - \overline{\mathcal{B}}\mathcal{A})^* = \mathcal{I} - \overline{\mathcal{B}}\mathcal{A}$. Furthermore, the above lemma shows that

$$\big((\mathcal{I} - \overline{\mathcal{B}}\mathcal{A})v, v\big)_{\mathcal{A}} = \|(\mathcal{I} - \mathcal{B}\mathcal{A})v\|_{\mathcal{A}}^2, \quad \forall v \in V.$$

Since $\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}$ is self-adjoint w.r.t. $(\cdot,\cdot)_{\mathcal{A}}$, we have $\|\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}\|_{\mathcal{A}} = \rho(\mathcal{I} - \overline{\mathcal{B}}\mathcal{A})$. Hence,

$$\|\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}\|_{\mathcal{A}} = \sup_{\|v\|_{\mathcal{A}}=1} \big((\mathcal{I} - \overline{\mathcal{B}}\mathcal{A})v, v\big)_{\mathcal{A}} = \sup_{\|v\|_{\mathcal{A}}=1} \|(\mathcal{I} - \mathcal{B}\mathcal{A})v\|_{\mathcal{A}}^2 = \|\mathcal{I} - \mathcal{B}\mathcal{A}\|_{\mathcal{A}}^2. \tag{21}$$

This immediately gives the following relations:

$$\rho(\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}) = \|\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}\|_{\mathcal{A}} = \|\mathcal{I} - \mathcal{B}\mathcal{A}\|_{\mathcal{A}}^2 \geq \rho(\mathcal{I} - \mathcal{B}\mathcal{A})^2.$$

Hence, if the symmetrized method converges, then the original method also converges; but the opposite direction might not be true. Furthermore, we have obtained the identity:

$$\|\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}\|_{\mathcal{A}} = \rho(\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}) = \sup_{v \in V \setminus \{0\}} \frac{\big((\mathcal{I} - \overline{\mathcal{B}}\mathcal{A})v, v\big)_{\mathcal{A}}}{\|v\|_{\mathcal{A}}^2}.$$

An Observation

In the Error Decay Lemma, we have already seen that

$$\left\|(\mathcal{I} - \mathcal{B}\mathcal{A})v\right\|_{\mathcal{A}}^2 = \left\|v\right\|_{\mathcal{A}}^2 - \left(\overline{\mathcal{B}}\mathcal{A}v, \mathcal{A}v\right).$$

Contraction property: $\left\|\mathcal{I} - \mathcal{B}\mathcal{A}\right\|_{\mathcal{A}} < 1$ if and only if $\overline{\mathcal{B}}$ is SPD.

Example (If $\overline{\mathcal{B}}$ is not SPD, $\mathcal{B}$ might still converge)

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad I - BA = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix},$$

then we have

$$\|I - BA\|_A = 2 > 1, \quad \overline{B} = \begin{bmatrix} 1 & 0 \\ 0 & -3 \end{bmatrix}, \quad \text{and} \quad I - \overline{B}A = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}.$$

Hence $\rho(I - BA) = 0 < 4 = \rho(I - \overline{B}A)$. So the iterator B converges but $\overline{B}$ does not.

Convergence of Symmetrized Method

Theorem (Convergence of Symmetrized Algorithm)

The symmetrized iteration SITER is convergent if and only if $\overline{\mathcal{B}}$ is SPD.

Proof.

First of all, we notice that

$$\mathcal{I} - \overline{\mathcal{B}}\mathcal{A} = (\mathcal{I} - \mathcal{B}^T\mathcal{A})(\mathcal{I} - \mathcal{B}\mathcal{A}) = \mathcal{A}^{-\frac{1}{2}}(\mathcal{I} - \mathcal{A}^{\frac{1}{2}}\mathcal{B}^T\mathcal{A}^{\frac{1}{2}})(\mathcal{I} - \mathcal{A}^{\frac{1}{2}}\mathcal{B}\mathcal{A}^{\frac{1}{2}})\mathcal{A}^{\frac{1}{2}},$$

which has the same spectrum as $(\mathcal{I} - \mathcal{A}^{\frac{1}{2}}\mathcal{B}^T\mathcal{A}^{\frac{1}{2}})(\mathcal{I} - \mathcal{A}^{\frac{1}{2}}\mathcal{B}\mathcal{A}^{\frac{1}{2}})$. So all eigenvalues of $\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}$ are non-negative. Hence, we have $\lambda \leq 1$ for all $\lambda \in \sigma(\overline{\mathcal{B}}\mathcal{A})$. SITER converges if and only if $\rho(\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}) < 1$. Because

$$\sigma(\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}) = \{1 - \lambda \,:\, \lambda \in \sigma(\overline{\mathcal{B}}\mathcal{A})\},$$

SITER converges if and only if $\sigma(\overline{\mathcal{B}}\mathcal{A}) \subseteq (0, 2)$. Therefore, the convergence is equivalent to $\sigma(\overline{\mathcal{B}}\mathcal{A}) \subseteq (0, 1]$, i.e., $\overline{\mathcal{B}}\mathcal{A}$ is SPD w.r.t. $(\cdot, \cdot)_{\mathcal{A}}$. □

Convergence Rate of Iterative Methods

$\bar{\mathcal{B}}$ is SPD $\iff$ $\bar{\mathcal{B}}$ defines a convergent method $\implies$ $\mathcal{B}$ defines a convergent method

Theorem (Convergence rate)

If $\overline{\mathcal{B}}$ is SPD, the convergence rate of the stationary iterative method (or its symmetrization) is

$$\|\mathcal{I} - \mathcal{B}\mathcal{A}\|_{\mathcal{A}}^2 = \|\mathcal{I} - \overline{\mathcal{B}}\mathcal{A}\|_{\mathcal{A}} = 1 - \frac{1}{c_1}, \quad \text{with } c_1 := \sup_{\|v\|_{\mathcal{A}}=1} (\overline{\mathcal{B}}^{-1}v, v).$$

Sketch of proof:

Since $((\mathcal{I} - \overline{\mathcal{B}}\mathcal{A})v, v)_{\mathcal{A}} = \|v\|_{\mathcal{A}}^2 - (\overline{\mathcal{B}}\mathcal{A}v, v)_{\mathcal{A}}$, we have

$$\|\mathcal{I} - \mathcal{B}\mathcal{A}\|_{\mathcal{A}}^2 = 1 - \inf_{\|v\|_{\mathcal{A}}=1} (\overline{\mathcal{B}}\mathcal{A}v, v)_{\mathcal{A}} = 1 - \lambda_{\min}(\overline{\mathcal{B}}\mathcal{A}) = 1 - \frac{1}{c_1},$$

where

$$c_1 = \lambda_{\max}((\overline{\mathcal{B}}\mathcal{A})^{-1}) = \sup_{\|v\|_{\mathcal{A}}=1} \left((\overline{\mathcal{B}}\mathcal{A})^{-1}v, v\right)_{\mathcal{A}} = \sup_{\|v\|_{\mathcal{A}}=1} \left(\overline{\mathcal{B}}^{-1}v, v\right).$$

Symmetric Positive Semidefinite Problems

Consider a more general linear system:

- $\mathcal{A} : V \mapsto V$ is a symmetric and positive semidefinite operator.
- V is a finite dimensional Hilbert space with inner product $(\cdot, \cdot)$.
- $\mathcal{Q} : V \mapsto \mathcal{R}(\mathcal{A})$ is the orthogonal projection under the inner product $(\cdot, \cdot)$.

Consider the general iterative method (ITER)

$$u^{\text{new}} = u^{\text{old}} + \mathcal{B}(f - \mathcal{A}u^{\text{old}}),$$

where $\mathcal{B}$ is a linear operator from V to V and it might be singular.

Let $|v|_{\mathcal{A}} := (v, v)_{\mathcal{A}}^{1/2}$ for any $v \in V$. Similar to the SPD case, we have

$$\begin{aligned} |v|_{\mathcal{A}}^2 - |(\mathcal{I} - \mathcal{B}\mathcal{A})v|_{\mathcal{A}}^2 &= (\overline{\mathcal{B}}\mathcal{A}v, v)_{\mathcal{A}} = \left((\mathcal{B} + \mathcal{B}^T - \mathcal{B}^T\mathcal{A}\mathcal{B})\mathcal{A}v, v\right)_{\mathcal{A}} \\ &= 2(\mathcal{B}\mathcal{A}v, v)_{\mathcal{A}} - (\mathcal{B}\mathcal{A}v, \mathcal{B}\mathcal{A}v)_{\mathcal{A}}, \quad \forall v \in V. \end{aligned}$$

Convergence Analysis for SPSD Problems

Theorem (Convergence for Semidefinite Problems)

The iterative algorithm ITER converges if $\mathcal{Q}\,\overline{\mathcal{B}}\,\mathcal{Q}$ is SPD on $\mathcal{R}(\mathcal{A})$.

The above convergence condition is equivalent to that there exists a positive constant α such that

$$2(\mathcal{B}\mathcal{A}v, v)_{\mathcal{A}} - (\mathcal{B}\mathcal{A}v, \mathcal{B}\mathcal{A}v)_{\mathcal{A}} \geq \alpha(v, v)_{\mathcal{A}}, \quad v \in V.$$

Corollary (Convergence conditions)

Suppose the following two assumptions hold:

(A1) There exists $\omega \in (0, 2)$ such that $(\mathcal{B}\mathcal{A}v, \mathcal{B}\mathcal{A}v)_{\mathcal{A}} \leq \omega(\mathcal{B}\mathcal{A}v, v)_{\mathcal{A}}, \ \ \forall v \in V$;

(A2) There exists $\beta > 0$ such that $(\mathcal{B}\mathcal{A}v, \mathcal{B}\mathcal{A}v)_{\mathcal{A}} \geq \beta(v, v)_{\mathcal{A}}, \ \ \forall v \in V$.

Then the iterative method ITER converges.

Convergence Rate for SPSD Problems

Under the assumptions (A1) and (A2), we find that

$$|v|_{\mathcal{A}}^2 - |(\mathcal{I} - \mathcal{B}\mathcal{A})v|_{\mathcal{A}}^2 = 2(\mathcal{B}\mathcal{A}v, v)_{\mathcal{A}} - (\mathcal{B}\mathcal{A}v, \mathcal{B}\mathcal{A}v)_{\mathcal{A}}$$
$$\geq \Big(\frac{2}{\omega} - 1\Big)(\mathcal{B}\mathcal{A}v, \mathcal{B}\mathcal{A}v)_{\mathcal{A}} \geq \frac{\beta(2-\omega)}{\omega}|v|_{\mathcal{A}}^2.$$

This immediately implies that $|\mathcal{I} - \mathcal{B}\mathcal{A}|_{\mathcal{A}}^2 \leq 1 - \frac{\beta(2-\omega)}{\omega} < 1$.

Theorem (Convergence rate for Semidefinite Problems)

Under the assumptions (A1) and (A2), the iterative method ITER satisfies that

$$|\mathcal{I} - \mathcal{B}\mathcal{A}|_{\mathcal{A}}^2 = 1 - \frac{1}{c_1}, \quad \text{with } c_1 := \sup_{v \in \mathcal{R}(\mathcal{A}), |v|_{\mathcal{A}}=1} \inf_{c \in \mathcal{N}(\mathcal{A})} (v + c, v + c)_{(\mathcal{Q}\overline{\mathcal{B}}\mathcal{Q})^\dagger}.$$

The Moore–Penros inverse is denoted by $\mathcal{B}^\dagger$. In case $\mathcal{N}(\mathcal{B}) = \mathcal{N}(\mathcal{A})$ and $\mathcal{R}(\mathcal{B}) = \mathcal{R}(\mathcal{A})$, then $\mathcal{B}^\dagger : V \mapsto V$ is a zero extension of $\mathcal{B}^{-1} : \mathcal{R}(\mathcal{A}) \mapsto \mathcal{R}(\mathcal{A})$, i.e. $\mathcal{B}^\dagger c = 0,\ \forall c \in \mathcal{N}(\mathcal{A})$ and $\mathcal{B}^\dagger v = B^{-1}v,\ \forall v \in \mathcal{R}(\mathcal{A})$.

GS for SPSD Problems

Suppose GS can be applied to the given SPSD problem, i.e. $B = (D+L)^{-1}$ and $\overline{B} = B + B^T - B^T AB$. It is easy to check that

$$\overline{B}^{-1} = (D+L)D^{-1}(D+L^T) =: S, \quad \text{on } \mathcal{R}(A).$$

In the previous theorem, we have

$$c_1 := \sup_{v\in\mathcal{R}(A)} \frac{\big((\mathcal{Q}\overline{B}\mathcal{Q})^\dagger v, v\big)}{(v,v)_A} = \sup_{v\in\mathcal{R}(A)} \frac{\big((\mathcal{Q}S^{-1}\mathcal{Q})^\dagger v, v\big)}{(v,v)_A}.$$

Let $w := (\mathcal{Q}S^{-1}\mathcal{Q})^\dagger v$. Then $w \in \mathcal{R}(A)$ and $w = S(v + c(v)),$ with $c(v) := S^{-1}w - v$.
By the definition, we can find $c(v) \in \mathcal{N}(A)$ and

$$c_1 = \sup_{v\in\mathcal{R}(A)} \frac{\big(S(v+c(v)), v+c(v)\big)}{(v,v)_A}.$$

Let $\xi := \arg\inf_{c\in\mathcal{N}(A)} \big(S(v+c), v+c\big)$. Then ξ uniquely satisfies that $\big(S(v+\xi), c\big) = 0$, for all $c \in \mathcal{N}(A)$. Apparently, $\xi = c(v)$ also satisfies this equation. So $\xi = c(v)$ and

$$c_1 = \sup_{v\in\mathcal{R}(A)} \frac{\big(S(v+c(v)), v+c(v)\big)}{(v,v)_A} = \sup_{v\in\mathcal{R}(A)} \inf_{c\in\mathcal{N}(A)} \frac{\big(S(v+c), v+c\big)}{(v,v)_A}.$$