Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Sixth lecture, May 22, 2023

Practical matters

- **Monday May 29** (next week) is a public holiday
  → **no lecture on May 29!**
- We will have a bonus live-coding lecture on **Tuesday May 30** about total variation regularization in place of the usual exercise session (this material will not be essential to the course).
- The deadline for the fifth exercise sheet will be moved to **Tuesday June 6**. Note that tomorrow's exercise session will happen as planned.

Regularization by truncated iterative methods

Regularization by truncated iterative methods

For simplicity, we will only consider the case when

$$Ax = y \tag{1}$$

is a system of linear equations, i.e., $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, and $y \in \mathbb{R}^m$.

- Iterative methods attempt to solve (1) by finding successive approximations for the solution starting from some initial guess.
- Typically, the computation of such iterations involves multiplications by A and its adjoint, but not explicit computation of inverse operators. (Direct methods, such as *Gaussian elimination*, produce a solution in a finite number of steps.)
- Iterative methods are sometimes the only feasible choice if the problem involves a large number of variables (e.g., in the order of millions), in which case direct methods are prohibitively expensive. Iterations are especially useful if multiplications by A are cheap: for example, if A is sparse or it contains some other structure (e.g., it is a multi-diagonal matrix arising from finite difference or finite element approximation of an elliptic PDE).

Although iterative solvers have not usually been designed for ill-posed equations, they often possess regularizing properties. If the iterations are terminated before “the solution starts to fit to noise”, one often obtains reasonable solutions for inverse problems.

Banach fixed point iteration

Let E be a Banach space and $S \subset E$. Consider a mapping, not necessarily linear, $T\colon E \to E$. We say that S is an *invariant set* for T if $T(S) \subset S$, that is,

$$T(x) \in S \quad \text{for all } x \in S.$$

Moreover, T is a *contraction* on an invariant set S if there exists $0 \leq \kappa < 1$ such that

$$\|T(x) - T(y)\| \leq \kappa \|x - y\| \quad \text{for all } x, y \in S.$$

Finally, a vector $x \in E$ is called a *fixed point* of T if

$$T(x) = x.$$

Theorem (Banach fixed point theorem)

Let E be a Banach space and $S \subset E$ a closed invariant set for the (possibly nonlinear) mapping $T \colon E \to E$. Assume further that T is a contraction in S. Then there exists a unique fixed point $x \in S$ such that $T(x) = x$. Furthermore, this fixed point can be found by the fixed point iteration

$$x = \lim_{k \to \infty} x_k, \quad \textit{where } x_{k+1} = T(x_k),$$

for any $x_0 \in S$.

Proof. Let $T \colon E \to E$ be a mapping, $S \subset E$ a closed invariant set such that $T(S) \subset S$, and let T be a contraction in S,

$$\|T(x) - T(y)\| \leq \kappa \|x - y\| \quad \text{for all } x, y \in S,$$

with $\kappa < 1$. For all $j > 1$, we have

$$\|x_{j+1} - x_j\| = \|T(x_j) - T(x_{j-1})\| \leq \kappa \|x_j - x_{j-1}\|.$$

Inductively, it follows that

$$\|x_{j+1} - x_j\| \leq \kappa^{j-1} \|x_2 - x_1\|.$$

For any $n, k \in \mathbb{N}$, we have

$$\begin{aligned}
\|x_n - x_k\| &\leq \sum_{j=1}^{\max\{n,k\}-\min\{n,k\}} \|x_{\min\{n,k\}+j} - x_{\min\{n,k\}+j-1}\| \\
&\leq \sum_{j=1}^{\max\{n,k\}-\min\{n,k\}} \kappa^{\min\{n,k\}+j-2} \|x_2 - x_1\| \\
&\leq \frac{\kappa^{\min\{n,k\}-1}}{1-\kappa} \|x_2 - x_1\| \xrightarrow{n,k\to\infty} 0,
\end{aligned}$$

where we used the formula for the geometric series. Therefore (x_j) is a Cauchy sequence and thus convergent (since E is a Banach space and thus complete). The limit is in S since S is closed.

Finally, as a contraction, T is (Lipschitz) continuous and we have that

$$x = \lim_{k\to\infty} x_k = \lim_{k\to\infty} T(x_{k-1}) = T\big(\lim_{k\to\infty} x_{k-1}\big) = T(x),$$

as desired. □

Landweber–Fridman iteration

Landweber–Fridman iteration

Instead of considering the original equation

$$Ax = y,$$

let us consider the normal equation

$$A^{\mathrm{T}}Ax = A^{\mathrm{T}}y.$$

Recall that $x \in \mathbb{R}^n$ satisfies the normal equation iff it minimizes the residual

$$\|Ax - y\|.$$

Moreover, there exists a unique element of $\mathbb{R}^n$, given by $x^\dagger := A^\dagger y$, which satisfies the normal equation and $x^\dagger \in \mathrm{Ker}(A)^\perp$ (the minimum norm solution).

Let us define the affine mapping $T\colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$T(x) = x + \beta(A^{\mathrm{T}}y - A^{\mathrm{T}}Ax), \quad \beta \in \mathbb{R}.$$

Note that any solution of the normal equation $A^{\mathrm{T}}Ax = A^{\mathrm{T}}y$ is a fixed point of T.

If β is small enough, then there is only one fixed point of T in $\mathrm{Ker}(A)^{\perp}$, precisely $x^{\dagger}$, and it can be reached by the fixed point iteration if $x_0 = 0$.

Theorem

Let λ_1 be the largest singular value of matrix A and let $0 < \beta < 2/\lambda_1^2$ be fixed. Then the fixed point iteration

$$x_{k+1} = T(x_k), \quad x_0 = 0,$$

converges toward $x^{\dagger}$ as $k \to \infty$.

Proof. Let $S := \mathrm{Ker}(A)^\perp = \mathrm{Ran}(A^\mathrm{T})$. Clearly $T(S) \subset S$ since

$$T(x) = x + A^\mathrm{T}(\beta y - \beta Ax) \in \mathrm{Ran}(A^\mathrm{T})$$

for all $x \in \mathrm{Ran}(A^\mathrm{T})$. Thus S is invariant under T.

Recall that A and its transpose can be written using the SVD of A as

$$Ax = \sum_{j=1}^{p} \lambda_j (v_j^\mathrm{T} x) u_j \quad \text{and} \quad A^\mathrm{T} y = \sum_{j=1}^{p} \lambda_j (u_j^\mathrm{T} y) v_j,$$

where $p = \mathrm{rank}(A)$ and λ_j are the positive singular values of A. The singular vectors $\{v_j\}_{j=1}^p$ and $\{u_j\}_{j=1}^p$ span $S = \mathrm{Ker}(A)^\perp$ and $\mathrm{Ran}(A)$, respectively, and thus

$$x = \sum_{j=1}^{p} (v_j^\mathrm{T} x) v_j \quad \text{for all } x \in S.$$

Let $x, z \in S$. Then $x - z \in S$ and

$$\begin{aligned} T(x) - T(z) &= (x - z) - \beta A^{\mathrm{T}} A(x - z) \\ &= \sum_{j=1}^{p} v_j^{\mathrm{T}}(x - z)v_j - \beta \sum_{j=1}^{p} \lambda_j^2 (v_j^{\mathrm{T}}(x - z))v_j \\ &= \sum_{j=1}^{p} (1 - \beta\lambda_j^2)(v_j^{\mathrm{T}}(x - z))v_j. \end{aligned}$$

Since λ_1 is the largest singular value, it follows that

$$-1 < \beta\lambda_j^2 - 1 \le \beta\lambda_1^2 - 1 < 2 - 1 = 1 \quad \text{for all } j \in \{1, \ldots, p\}.$$

Hence

$$\kappa := \max_{j=1,\ldots,p} |\beta\lambda_j^2 - 1| < 1.$$

In consequence,

$$\begin{aligned}\|T(x) - T(y)\|^2 &\leq \sum_{j=1}^{p} (1 - \beta\lambda_j^2)^2 (v_j^{\mathrm{T}}(x - z))^2 \\ &\leq \kappa^2 \sum_{j=1}^{p} (v_j^{\mathrm{T}}(x - z))^2 = \kappa^2 \|x - z\|^2,\end{aligned}$$

which shows that T is a contraction on S. Since S is a closed invariant set for T, there exists a unique fixed point of T in S.

Finally, recall that $x^\dagger = A^\dagger y$ belongs to $S = \mathrm{Ker}(A)^\perp$ and it satisfies the normal equation. Since $x_0 = 0$ is in S (it is orthogonal to all vectors), the fixed point iteration starting from x_0 converges to $x^\dagger$. □

Regularization properties of Landweber–Fridman

In what follows, we will assume that $0 < \beta < 2/\lambda_1^2$.

In the exercises, it will be shown that the k^{th} iterate of the Landweber–Fridman iteration can be written explicitly as

$$x_k = \sum_{j=1}^{p} \frac{1}{\lambda_j}(1 - (1 - \beta\lambda_j^2)^k)(u_j^{\mathrm{T}} y)v_j, \quad k = 0, 1, \dots .$$

Since we assumed $|1 - \beta\lambda_j^2| < 1$, then

$$(1 - \beta\lambda_j^2)^k \xrightarrow{k\to\infty} 0.$$

This is what one would expect since

$$x^\dagger = \sum_{j=1}^{p} \frac{1}{\lambda_j}(u_j^{\mathrm{T}} y)v_j.$$

While $k \in \mathbb{N}$ is finite, the coefficients appearing in the series representation

$$x_k = \sum_{j=1}^{p} \frac{1}{\lambda_j}(1 - (1 - \beta\lambda_j^2)^k)(u_j^{\mathrm{T}} y) v_j \tag{2}$$

satisfy

$$\begin{aligned} \frac{1}{\lambda_j}(1 - (1 - \beta\lambda_j^2)^k) &= \frac{1}{\lambda_j}\left(1 - \sum_{\ell=0}^{k} \binom{k}{\ell}(-1)^{\ell}\beta^{\ell}\lambda_j^{2\ell}\right) \\ &= \frac{1}{\lambda_j}\sum_{\ell=1}^{k} \binom{k}{\ell}(-1)^{\ell+1}\beta^{\ell}\lambda_j^{2\ell} = \sum_{\ell=1}^{k} \binom{k}{\ell}(-1)^{\ell+1}\beta^{\ell}\lambda_j^{2\ell-1}, \end{aligned}$$

which converges to zero as $\lambda_j \to 0$ (for a fixed k).

In consequence, while k is "small enough", no coefficient of $(u_j^{\mathrm{T}} y) v_j$ in (2) is so large that the component of the measurement noise in the direction u_j is amplified in an uncontrolled manner. (Compare with Tikhonov regularization, where the corresponding coefficients are $\lambda_j/(\lambda_j^2 + \delta)$.)

Discrepancy principle for Landweber–Fridman iteration

Let $y \in \mathbb{R}^m$ be a noisy version of some underlying “exact” data vector $y_0 \in \mathbb{R}^m$, and assume that

$$\|y - y_0\| \approx \varepsilon > 0.$$

The Morozov discrepancy principle for the Landweber–Fridman iteration is analogous to the truncated SVD: choose the smallest $k \geq 0$ such that the residual satisfies

$$\|y - Ax_k\| \leq \varepsilon.$$

Q: *When does an index $k \geq 1$ satisfying $\|y - Ax_k\| \leq \varepsilon$ exist?*
A: When $\varepsilon > \|Py - y\| = \|y - A(A^\dagger y)\| = \|y - Ax^\dagger\|$, where $P = AA^\dagger$ is the orthogonal projection onto $\mathrm{Ran}(A)$ (cf. 3^{rd} exercises) and $x^\dagger = A^\dagger y$ is the minimum norm solution. Since the sequence $(x_k)_{k=0}^\infty$ converges to $x^\dagger$, for any $\varepsilon > \|y - Ax^\dagger\|$, there exists $k = k_\varepsilon \in \mathbb{N}$ such that

$$\|x_k - x^\dagger\| \leq \frac{1}{\|A\|}(\varepsilon - \|y - Ax^\dagger\|).$$

By the reverse triangle inequality

$$\begin{aligned} \|y - Ax_k\| - \|y - Ax^\dagger\| &\leq \|(y - Ax_k) - (y - Ax^\dagger)\| \\ &\leq \|A\|\|x_k - x^\dagger\| \\ &\leq \varepsilon - \|y - Ax^\dagger\|. \end{aligned}$$

From this, we deduce that $\|y - Ax_k\| \leq \varepsilon$ as desired.

Conjugate gradient method

Krylov subspace methods

Krylov subspace methods are iterative solvers for (large scale) matrix equations of the form $Ax = y$, $A \in \mathbb{R}^{n \times n}$. In general terms, the solution vector $x \in \mathbb{R}^n$ is approximated as a linear combination of vectors of the form u, Au, A^2u, $\ldots$, with some given $u \in \mathbb{R}^n$. If multiplication by A is cheap – for example, when A is sparse – Krylov subspace methods can be particularly efficient.

We consider only the most well-known Krylov subspace method, the conjugate gradient method. It is worth mentioning that other methods in this class include, e.g., the generalized minimum residual method (GMRES) and the biconjugate gradient method (BiCG).

Assumptions on A and A-dependent inner product

In what follows, we assume that the system matrix $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite:

$$A^{\mathrm{T}} = A \quad \text{and} \quad u^{\mathrm{T}}Au > 0 \quad \text{for all } u \in \mathbb{R}^n \setminus \{0\}.$$

Note that this implies that A is injective.[†] By the fundamental theorem of linear algebra, A is invertible. Furthermore, the inverse $A^{-1} \in \mathbb{R}^{n\times n}$ is also symmetric and positive definite.

We define

$$\langle u, v\rangle_A := u^{\mathrm{T}}Av \quad \text{and} \quad \|u\|_A := \sqrt{\langle u, u\rangle_A}.$$

Since A was assumed to be symmetric and positive definite, it is straightforward to check that $\langle\cdot,\cdot\rangle_A \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defines an inner product on $\mathbb{R}^n$. In consequence, $\|\cdot\|_A \colon \mathbb{R}^n \to \mathbb{R}$ is a norm.

Finally, we say that non-zero vectors $\{s_0, \dots, s_k\} \subset \mathbb{R}^n$ are A-*conjugate* if

$$\langle s_i, s_j\rangle_A = s_i^{\mathrm{T}}As_j = 0 \quad \text{whenever } i \neq j,$$

i.e., they are orthogonal with respect to the inner product $\langle\cdot,\cdot\rangle_A$.

[†] $Ax = Ay \Rightarrow A(x-y) = 0 \Rightarrow (x-y)^{\mathrm{T}}A(x-y) = 0 \Rightarrow x - y = 0.$

Error, residual, and minimization problem

Let $x_* = A^{-1}y \in \mathbb{R}^n$ denote the unique solution of the equation

$$Ax = y$$

for a given $y \in \mathbb{R}^n$. We define the error and residual corresponding to some approximate solution $x \in \mathbb{R}^n$ by

$$e = x_* - x \quad \text{and} \quad r = y - Ax = Ae.$$

Let $\phi\colon \mathbb{R}^n \to \mathbb{R}$ be the A-dependent quadratic functional

$$\phi(x) = \|e\|_A^2 = e^{\mathrm{T}}Ae = r^{\mathrm{T}}A^{-1}r = \|r\|_{A^{-1}}^2.$$

Since $\|\cdot\|_A$ is a norm, $\phi(x) \geq 0$ for all $x \in \mathbb{R}^n$ and

$$\phi(x) = 0 \quad \Leftrightarrow \quad e = 0 \quad \Leftrightarrow \quad x = x_*.$$

Minimizing ϕ is equivalent to solving $Ax = y$.

The conjugate gradient method is an iterative scheme which, at each step of the iteration, returns $x_{k+1} = \arg\min_{x\in\mathcal{S}_k}\phi(x)$, where

$$\mathcal{S}_k := \{x \in \mathbb{R}^n \mid x = x_0 + c_0 s_0 + \cdots + c_k s_k,\ c_0, \ldots, c_k \in \mathbb{R}\}$$

is a hyperplane determined by a sequence of vectors $s_0, \ldots, s_k \in \mathbb{R}^n$.

Starting from an initial guess $x_0 \in \mathbb{R}^n$, the successive iterates are given by

$$x_{k+1} = x_k + \alpha_k s_k, \quad k = 0, 1, 2, \ldots .$$

Define the *residual* $r_k = y - Ax_k$ corresponding to iterate x_k and let $s_0 = r_0$ be the *initial search direction*. Then the parameters are

$$\alpha_k = \frac{s_k^{\mathrm{T}} r_k}{s_k^{\mathrm{T}} A s_k} \quad \text{for } k \geq 0, \qquad \text{(“step size”)}$$

$$s_k = r_k + \beta_{k-1} s_{k-1},\ \beta_{k-1} = -\frac{s_{k-1}^{\mathrm{T}} A r_k}{s_{k-1}^{\mathrm{T}} A s_{k-1}} \text{ for } k \geq 1. \ \text{(“search direction”)}$$

We proceed to show that the search directions defined by the above recursion are A-conjugate (and thus linearly independent) and the iterates x_{k+1} obtained using this algorithm are minimizers of the functional $\phi(x)$ on the hyperplanes $\mathcal{S}_k$. Note especially that $\mathcal{S}_{n-1} = \mathbb{R}^n$, so an exact solution (up to rounding errors) is achieved in at most n iteration steps.

$\boxed{\textbf{Step 1:} \text{ If } s_0, \ldots, s_k \text{ are } A\text{-conjugate, then } r_{k+1} \perp \operatorname{span}\{s_0, \ldots, s_k\}.}$

Now $x_{k+1} = x_k + \alpha_k s_k = x_{k-1} + \alpha_{k-1} s_{k-1} + \alpha_k s_k = \cdots = x_0 + \sum_{j=0}^{k} \alpha_j s_j$ and $r_{k+1} = y - Ax_{k+1} = y - Ax_0 - \sum_{j=0}^{k} \alpha_j A s_j = r_0 - \sum_{j=0}^{k} \alpha_j A s_j$.
Let $\ell \in \{0, \ldots, k\}$. Then

$$\begin{aligned}
r_{k+1}^{\mathrm{T}} s_\ell &= \left(r_0 - \sum_{j=0}^{k} \alpha_j A s_j \right)^{\mathrm{T}} s_\ell && (A^{\mathrm{T}} = A) \\
&= r_0^{\mathrm{T}} s_\ell - \sum_{j=0}^{k} \alpha_j s_j^{\mathrm{T}} A s_\ell && (s_j^{\mathrm{T}} A s_\ell = 0 \text{ for } j \neq \ell) \\
&= r_0^{\mathrm{T}} s_\ell - \alpha_\ell s_\ell^{\mathrm{T}} A s_\ell && (\alpha_\ell = \tfrac{s_\ell^{\mathrm{T}} r_\ell}{s_\ell^{\mathrm{T}} A s_\ell}) \\
&= r_0^{\mathrm{T}} s_\ell - s_\ell^{\mathrm{T}} r_\ell && (r_\ell = r_0 - \textstyle\sum_{j=0}^{\ell-1} \alpha_j A s_j) \\
&= r_0^{\mathrm{T}} s_\ell - s_\ell^{\mathrm{T}} r_0 + \sum_{j=0}^{\ell-1} \alpha_j s_\ell^{\mathrm{T}} A s_j \\
&= 0,
\end{aligned}$$

as desired.

$\boxed{\textbf{Step 2: } s_0, \ldots, s_k \text{ are } A\text{-conjugate and linearly independent.}}$

By induction with respect to $k \in \mathbb{N}_0$. If $k = 0$, then $\{s_0\}$ is trivially A-conjugate. Suppose that the claim has been proved for some $k \in \mathbb{N}_0$; we show that $s_{k+1}^{\mathrm{T}} A s_j = 0$ for all $j \in \{0, \ldots, k\}$.

Let $j \in \{0, \ldots, k\}$. Then

$$s_{k+1}^{\mathrm{T}} A s_j = (r_{k+1} + \beta_k s_k)^{\mathrm{T}} A s_j = r_{k+1}^{\mathrm{T}} A s_j + \beta_k s_k^{\mathrm{T}} A s_j.$$

If $0 \le j \le k-1$, then the above expression vanishes by the previous slide and the induction hypothesis. Let $j = k$. Then

$$\begin{aligned} s_{k+1}^{\mathrm{T}} A s_k &= r_{k+1}^{\mathrm{T}} A s_k + \beta_k s_k^{\mathrm{T}} A s_k \qquad \left(\beta_k = -\tfrac{s_k^{\mathrm{T}} A r_{k+1}}{s_k^{\mathrm{T}} A s_k}\right) \\ &= 0, \end{aligned}$$

as desired. For the linear dependence, write $c_0 s_0 + \cdots + c_k s_k = 0$ for some undetermined coefficients $c_0, \ldots, c_k \in \mathbb{R}$. For any $\ell \in \{0, \ldots, k\}$, multiplying from the left by $s_\ell^{\mathrm{T}} A$ yields

$$c_0 s_\ell^{\mathrm{T}} A s_0 + \cdots + c_k s_\ell^{\mathrm{T}} A s_k = 0 \Rightarrow c_\ell s_\ell^{\mathrm{T}} A s_\ell = 0 \overset{\substack{x^{\mathrm{T}} A x = 0 \\ \text{iff } x = 0}}{\Rightarrow} c_\ell = 0$$

as desired.

Step 3: $h_* = \underset{h \in \mathbb{R}^{k+1}}{\arg\min} \phi(x_0 + S_k h)$ iff $h_* = (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0$, where $x_0 \in \mathbb{R}^n$, $r_0 = y - Ax_0$, $S_k = [s_0, \ldots, s_k]$, and $s_0, \ldots, s_k \in \mathbb{R}^n$ are lin. independent.

We first verify that the expression $(S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0$ is well-defined by showing that $S_k^{\mathrm{T}} A S_k \in \mathbb{R}^{(k+1)\times(k+1)}$ is invertible. By the positive definiteness of A,

$$S_k^{\mathrm{T}} A S_k z = 0 \quad \Rightarrow z^{\mathrm{T}} S_k^{\mathrm{T}} A S_k z = 0 \quad \Rightarrow \|S_k z\|_A^2 = 0 \quad \Rightarrow S_k z = 0,$$

which means that $z = 0$ since the columns of S_k are linearly independent. Hence $S_k^{\mathrm{T}} A S_k$ is injective, and $(S_k^{\mathrm{T}} A S_k)^{-1}$ exists by the fundamental theorem of linear algebra.

The residual corresponding to $x = x_0 + S_k h$ satisfies

$$r = y - A(x_0 + S_k h) = r_0 - AS_k h,$$

thus (recall that $\phi(x) = r^{\mathrm{T}} A^{-1} r$ for $r = y - Ax$)

$$\begin{aligned} \phi(x_0 + S_k h) &= (r_0 - AS_k h)^{\mathrm{T}} A^{-1} (r_0 - AS_k h) \\ &= r_0^{\mathrm{T}} A^{-1} r_0 - 2 r_0^{\mathrm{T}} S_k h + h^{\mathrm{T}} S_k^{\mathrm{T}} A S_k h. \end{aligned}$$

We obtained

$$\phi(x_0 + S_k h) = r_0^{\mathrm{T}} A^{-1} r_0 - 2 r_0^{\mathrm{T}} S_k h + h^{\mathrm{T}} S_k^{\mathrm{T}} A S_k h.$$

The Hessian of $h \mapsto \phi(x_0 + S_k h)$ is $2 S_k^{\mathrm{T}} A S_k$, which is positive definite since

$$u^{\mathrm{T}} (S_k^{\mathrm{T}} A S_k) u = (S_k u)^{\mathrm{T}} A (S_k u) \geq 0 \quad \text{for all } u \in \mathbb{R}^{k+1},$$

where equality holds iff $S_k u = 0 \Leftrightarrow u = 0$. Hence $h \mapsto \phi(x_0 + S_k h)$ is convex, and we can find its unique minimizer by solving the zero point of its gradient:

$$\begin{aligned} 0 &= \nabla_h \phi(x_0 + S_k h) = 2 S_k^{\mathrm{T}} A S_k h - 2 S_k^{\mathrm{T}} r_0 \\ \Leftrightarrow \quad h &= (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0. \end{aligned}$$

Step 4: Let $x_0 \in \mathbb{R}^n$ be the initial guess and $S_k = [s_0, \ldots, s_k]$, where $s_0, \ldots, s_k \in \mathbb{R}^n$ are the conjugate gradient search directions. The conjugate gradient iterates satisfy $x_{k+1} = \arg\min_{h \in \mathbb{R}^{k+1}} \phi(x_0 + S_k h)$.

Let $a_j = (\alpha_0, \ldots, \alpha_j)^{\mathrm{T}} \in \mathbb{R}^{j+1}$, where $\alpha_i = \frac{s_i^{\mathrm{T}} r_i}{s_i^{\mathrm{T}} A s_i}$ are the line search parameters of the conjugate gradient method. Then

$$x_j = x_0 + \sum_{i=0}^{j-1} \alpha_i s_i = x_0 + S_{j-1} a_{j-1}, \quad j = 1, \ldots, k+1.$$

The residual corresponding to x_j is

$$r_j = y - Ax_j = (y - Ax_0) - AS_{j-1} a_{j-1} = r_0 - AS_{j-1} a_{j-1}$$

and hence

$$s_j^{\mathrm{T}} r_j = s_j^{\mathrm{T}} r_0 - s_j^{\mathrm{T}} A S_{j-1} a_{j-1} = s_j^{\mathrm{T}} r_0 - \underbrace{s_j^{\mathrm{T}} [As_0, \ldots, As_{j-1}]}_{=0} a_{j-1},$$

since $s_j^{\mathrm{T}} A s_i = 0$, $i < j$, due to A-conjugacy. Therefore

$$\alpha_j = \frac{s_j^{\mathrm{T}} r_j}{s_j^{\mathrm{T}} A s_j} = \frac{s_j^{\mathrm{T}} r_0}{s_j^{\mathrm{T}} A s_j}, \quad j = 0, \ldots, k.$$

The line search parameters can be written as

$$\alpha_j = \frac{s_j^{\mathrm{T}} r_j}{s_j^{\mathrm{T}} A s_j} = \frac{s_j^{\mathrm{T}} r_0}{s_j^{\mathrm{T}} A s_j}, \quad j = 0, \dots, k.$$

On the other hand, since $\{s_0, \dots, s_k\}$ are A-conjugate, we have that

$$\begin{aligned}(S_k^{\mathrm{T}} A S_k)^{-1} &= \mathrm{diag}(s_0^{\mathrm{T}} A s_0, \dots, s_k^{\mathrm{T}} A s_k)^{-1} \\ &= \mathrm{diag}\left(\frac{1}{s_0^{\mathrm{T}} A s_0}, \dots, \frac{1}{s_k^{\mathrm{T}} A s_k}\right).\end{aligned}$$

Especially, this means that the minimizer h^* of $\phi(x_0 + S_k h)$ over the hyperplane $\mathcal{S}_k$ is given by

$$h^* = (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0 = \mathrm{diag}\left(\frac{1}{s_0^{\mathrm{T}} A s_0}, \dots, \frac{1}{s_k^{\mathrm{T}} A s_k}\right) \begin{bmatrix} s_0^{\mathrm{T}} r_0 \\ \vdots \\ s_k^{\mathrm{T}} r_0 \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_k \end{bmatrix} = a_k.$$

In consequence, $x_{k+1} = x_0 + S_k a_k = x_0 + S_k h_*$.

Remark. In the conjugate gradient method, the search directions are given by $s_0 = r_0$ and

$$s_k = r_k + \beta_{k-1} s_{k-1}, \quad k \geq 1,$$

where $r_k = y - Ax_k$. Note that $\mathrm{span}\{s_0, \ldots, s_k\} = \mathrm{span}\{r_0, \ldots, r_k\}$.

Especially, the conjugate gradient iterate x_{k+1} satisfies

$$\begin{aligned} x_{k+1} = \underset{x \in x_0 + \mathrm{span}\{s_0, \ldots, s_k\}}{\arg\min} \|x - x_*\|_A^2 &= \underset{x \in x_0 + \mathrm{span}\{r_0, \ldots, r_k\}}{\arg\min} \|x - x_*\|_A^2 \\ &= \underset{x \in x_0 + \mathcal{K}_k}{\arg\min} \|x - x_*\|_A^2, \end{aligned}$$

where the *search space* $\mathcal{K}_k := \mathrm{span}\{r_0, Ar_0, \ldots, A^{k-1} r_0\}$ is precisely the k^{th} Krylov subspace of A with the initial vector $r_0 = y - Ax_0$. Some basic properties of Krylov subspaces:

- $A(\mathcal{K}_k) \subset \mathcal{K}_{k+1}$.
- $\mathcal{K}_{k-1} \subset K_k$ (Krylov subspaces are nested).
- $\dim \mathcal{K}_k \leq k$ (dimension of the k^{th} Krylov subspace is at most k).
- $\dim \mathcal{K}_k \leq \dim \mathcal{K}_{k-1} + 1$ (dimension of the successive Krylov space is at most one higher than that of the former).

The conjugate gradient algorithm is usually presented in slightly different form. Assuming that the iteration has not yet converged at the iterate x_k, we can deduce the following formulae for $\alpha_k = \frac{s_k^{\mathrm{T}} r_k}{s_k^{\mathrm{T}} A s_k}$ and $\beta_k = -\frac{s_k^{\mathrm{T}} A r_{k+1}}{s_k^{\mathrm{T}} A s_k}$.

Simplifying α_k: Since $r_k \perp s_{k-1}$, we have that

$$s_k^{\mathrm{T}} r_k = (r_k + \beta_{k-1} s_{k-1})^{\mathrm{T}} r_k = \|r_k\|^2 \quad \Rightarrow \quad \alpha_k = \frac{\|r_k\|^2}{s_k^{\mathrm{T}} A s_k}. \tag{3}$$

Simplifying β_k: since $r_{k+1} \perp \mathrm{span}\{s_0, \ldots, s_k\} \ni r_k$ and $r_{k+1} = r_k - \alpha_k A s_k$, then

$$\|r_{k+1}\|^2 = r_{k+1}^{\mathrm{T}}(r_k - \alpha_k A s_k) \overset{(3)}{=} -\frac{\|r_k\|^2}{s_k^{\mathrm{T}} A s_k} r_{k+1}^{\mathrm{T}} A s_k = \beta_k \|r_k\|^2$$

and thus

$$\beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}.$$

This leads to the “standard form” of the method.

Pseudocode for the conjugate gradient algorithm

```
Given: symmetric, positive definite system matrix $A \in \mathbb{R}^{n \times n}$,
data $y \in \mathbb{R}^n$.
  1. Choose initial guess $x_0 \in \mathbb{R}^n$.
  2. Set $k = 0$, $r_0 = y - Ax_0$, $s_0 = r_0$;
     Repeat until the chosen stopping rule is satisfied:
        3. $\alpha_k = \|r_k\|^2 / (s_k^{\mathrm{T}} A s_k)$;
        4. $x_{k+1} = x_k + \alpha_k s_k$;
        5. $r_{k+1} = r_k - \alpha_k A s_k$;
        6. $\beta_k = \|r_{k+1}\|^2 / \|r_k\|^2$;
        7. $s_{k+1} = r_{k+1} + \beta_k s_k$;
        8. $k \leftarrow k + 1$;
     end
```

Numerical example

Let us consider minimization with the *steepest descent* directions

$$s_k = -\nabla\phi(x_k) = 2(y - Ax_k), \quad k = 0, 1, \ldots . \tag{4}$$

In general, the convergence of the sequence $\{x_k\}$ toward the global minimizer $x_* = A^{-1}y$ can be fairly slow. We demonstrate this with the following example.

Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Now

$$\phi(x) = x_1^2 + 5x_2^2.$$

We plot the level contours of ϕ and the sequence $\{x_k\}_{k=0}^{5}$ starting from $x_0 = (1, 0.3)^{\mathrm{T}}$. The true solution $x_* = (0, 0)^{\mathrm{T}}$ is marked with a blue cross.

We also illustrate minimization over the hyperplanes $\mathcal{S}_0$ and $\mathcal{S}_1$, i.e., $x_0 + S_0 h_*$ and $x_0 + S_1 h_*$ with $S_0 = [s_0] \in \mathbb{R}^{2\times 1}$ and $S_1 = [s_0, s_1] \in \mathbb{R}^{2\times 2}$, where s_0 and s_1 were computed using the sequential method (4).

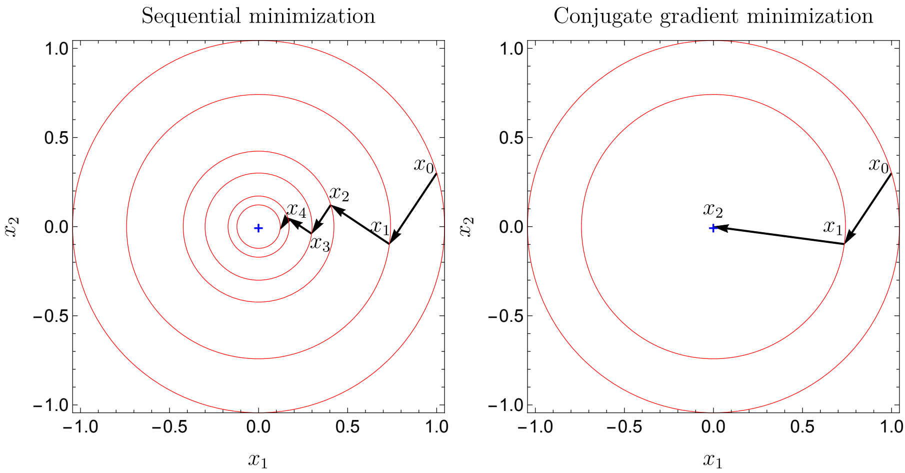


Figure: Left: Minimization using steepest descent search directions $s_k = -\nabla\phi(x_k)$. Right: In the linear case, the conjugate gradient method iteratively finds the optima over the hyperplanes $\mathcal{S}_1$ and $\mathcal{S}_2$. The CG method converges to the actual solution $x_* = (0,0)^{\mathrm{T}}$ (marked with a blue cross) in $n = 2$ iterations (which equals the dimensionality of the ambient space $\mathbb{R}^2$).

Conjugate gradient method for inverse problems

According to the previous construction, if the conjugate gradient method is applied to the equation

$$Ax = y,$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, an exact solution (up to rounding errors) is achieved in at most n iteration steps, i.e., $x_n = x_* = A^{-1}y$. However, the algorithm typically converges satisfactorily much quicker. A (pessimistic) convergence rate is proved in the first exercise of week 4.

With ill-posed problems, one should be more cautious and terminate the iterations well before convergence to avoid fitting the solution to noise. In fact, since the conjugate gradient method often converges very fast, one should be extremely cautious.

Let us consider a general ill-posed matrix equation

$$Ax = y,$$

where $A \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^m$ are given.

- If $m = n$ and there is some available prior information suggesting that A is, at least in theory, positive (semi-)definite, one can apply the conjugate gradient algorithm directly on the original equation.
- More generally, one may still consider the normal equation

  $$A^{\mathrm{T}}Ax = A^{\mathrm{T}}y,$$

  which corresponds to solving the original equation in the sense of least squares.

The system matrix $A^{\mathrm{T}}A = (A^{\mathrm{T}}A)^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ is symmetric and

$$u^{\mathrm{T}}A^{\mathrm{T}}Au = \|Au\|^2 > 0 \quad \text{for all } u \in \mathbb{R}^n \setminus \mathrm{Ker}(A).$$

Thus the conditions of the conjugate gradient algorithm are almost satisfied, and one may look for the solution of the inverse problem by using the conjugate gradient algorithm with A replaced by $A^{\mathrm{T}}A$ and y by $A^{\mathrm{T}}y$.[†]

As a stopping criterion, one may try, e.g., the Morozov principle for the original equation: terminate the iteration when

$$\|y - Ax_k\| \leq \varepsilon$$

for some $\varepsilon > 0$, which measures the amount of noise in y in some sense.

[†] Small remark on implementation: matrix-matrix products are typically far more expensive to compute than matrix-vector products. For example, instead of computing expressions like `residual = A'*y - A'*A*x0` when implementing the conjugate gradient method in MATLAB, one should use parentheses to parse the computation like `residual = A'*y - A'*(A*x0)`. Similarly `residual = A.T@y - A.T@(A@x0)` in Python.

Numerical example: backward heat equation revisited

Let us revisit the backward heat equation:

$$\begin{cases} \partial_t u(x,t) = \partial_x^2 u(x,t) & \text{for } (x,t) \in (0,\pi) \times \mathbb{R}_+, \\ u(0,\cdot) = u(\pi,\cdot) = 0 & \text{on } \mathbb{R}_+, \\ u(\cdot,0) = f & \text{on } (0,\pi), \end{cases}$$

where $f\colon (0,\pi) \to \mathbb{R}$ is the initial heat distribution.

Inverse problem: Reconstruct the initial state f based on noisy measurements of $u(\cdot,T)$ at time $T > 0$.

Let $x_j = jh$, $j = 0,\ldots,100$ with $h = \pi/100$, and denote $U(t) = (U_j(t))_{j=1}^{99}$ and $F = (f(x_j))_{j=1}^{99}$. At time $t = T > 0$, the discretized heat distribution $U := U(T)$ is given by

$$U = AF,$$

where $A = \mathrm{e}^{TB} \in \mathbb{R}^{99\times 99}$ and $B = h^{-2}\mathrm{tridiag}(1,-2,1) \in \mathbb{R}^{99\times 99}$.

As ground truth, we take

$$f(x) = \begin{cases} 1 & \text{if } x \in [1,2], \\ 0 & \text{if } x \in (0,1) \cup (2,\pi). \end{cases}$$

We assume that the simulated data $U = U(T) \in \mathbb{R}^{99}$ at time $T = 0.1$ is contaminated with mean-zero Gaussian noise with standard deviation 0.01, and that the discrepancy between the measured data and the underlying "exact" data equals the square root of the expected value of the squared norm of the noise vector, i.e.,

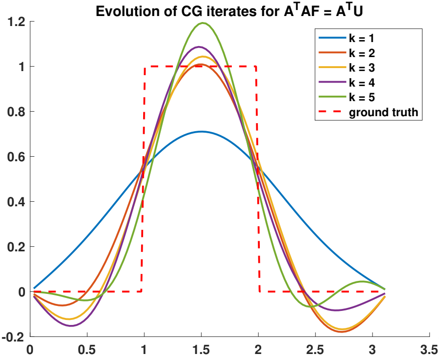
$$\varepsilon = \sqrt{99 \cdot 0.01^2} \approx 0.0995.$$

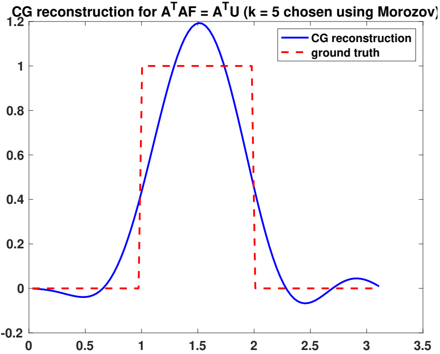
We use the conjugate gradient method to solve the normal equation

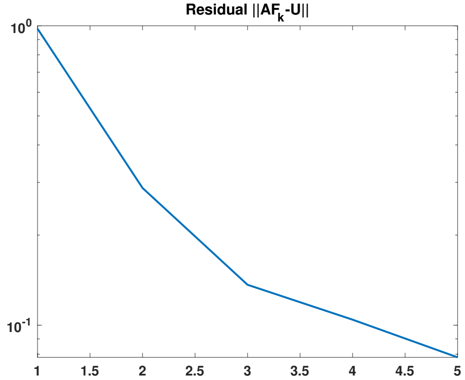
$$A^{\mathrm{T}} A F = A^{\mathrm{T}} U,$$

and terminate the algorithm for the first CG iterate F_k such that

$$\|AF_k - U\| \leq \varepsilon.$$


Evolution of CG iterates for $A^TAF = A^TU$
k = 1
k = 2
k = 3
k = 4
k = 5
ground truth
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
1.5
2
2.5
3
3.5


CG reconstruction for $A^TAF = A^TU$ (k = 5 chosen using Morozov)
CG reconstruction
ground truth
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
1.5
2
2.5
3
3.5


Residual ||AF_k-U||
10^0
10^-1
1
1.5
2
2.5
3
3.5
4
4.5
5

Although we have simply scratched the surface by covering some of the basic ideas surrounding the conjugate gradient scheme and demonstrating how an “early stopping rule” can provide reasonable solutions for inverse problems, the regularizing properties of the conjugate gradient method have been analyzed more explicitly in the literature. A classic textbook specifically about this subject is:

M. Hanke. *Conjugate gradient type methods for ill-posed problems*. Pitman Research Notes in Mathematics Series, 327.