Computational inverse problems

Nuutti Hyvönen and Juha-Pekka Puska

nuutti.hyvonen@aalto.fi, juha-pekka.puska@aalto.fi

Fifth lecture, March 15, 2021.

2.4.3 Krylov subspace methods

Krylov subspace methods

The Krylov subspace methods are iterative solvers for (large) matrix equations of the form $Ax = y$, $A \in \mathbb{R}^{n \times n}$. Loosely speaking, such methods try to approximate the solution vector $x \in \mathbb{R}^n$ as a linear combination of vectors of the type u, Au, A^2u etc., with some given $u \in \mathbb{R}^n$. If multiplication by A is cheap — e.g., if A is sparse —, the Krylov subspace methods are especially efficient.

On this course, we only consider the most well-known Krylov subspace method, the conjugate gradient method. Other methods of this class include, e.g., the generalized minimal residual method (GMRES), and the biconjugate gradient method (BiCG).

The regularizing properties of the conjugate gradient method can be analyzed explicitly; see, e.g., the monograph

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

However, here we content ourselves with introducing the basic ideas behind the conjugate gradient scheme and demonstrating numerically how application of an 'early stopping rule' provides reasonable solutions for inverse problems.

Assumptions on A and a related inner product

We assume that the system matrix $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, i.e.,

$$A^{\mathrm{T}} = A \qquad \text{and} \qquad u^{\mathrm{T}} A u > 0 \quad \text{for } u \neq 0.$$

In particular, this means that the square matrix A is injective, and consequently invertible due to the fundamental theorem of linear algebra. It is easy to see that the inverse $A^{-1} \in \mathbb{R}^{n \times n}$ is also symmetric and positive definite.

We define an A-dependent inner product and the corresponding norm via

$$\langle u, v \rangle_A = u^{\mathrm{T}} A v \qquad \text{and} \qquad \|u\|_A = \langle u, u \rangle_A^{1/2}.$$

It follows from the assumptions on A that $\langle \cdot, \cdot \rangle_A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ really is an inner product on $\mathbb{R}^n$, and consequently $\| \cdot \|_A : \mathbb{R}^n \to \mathbb{R}$ is a norm.

The error, the residual and a minimization problem

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

$$Ax = y$$

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

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

Let $\phi : \mathbb{R}^n \to \mathbb{R}$ 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, $\phi(x)$ is non-negative and equals zero if and only if

$$e = 0 \qquad \Longleftrightarrow \qquad x = x_*.$$

Hence, minimizing ϕ is equivalent to solving the original equation.

Minimizing ϕ in a given direction

Evaluating ϕ would require the knowledge of x_* or, equivalently, that of A^{-1}; since our ultimate goal is to approximate the solution x_* iteratively, assuming it known is not a feasible option.

Fortunately, if we have some initial guess $x_0 \in \mathbb{R}^n$ and some search direction $0 \neq s_0 \in \mathbb{R}^n$, we can find the minimizer of ϕ over the line

$$\mathcal{S}_0 = \{x \in \mathbb{R}^n \mid x = x_0 + \alpha s_0,\ \alpha \in \mathbb{R}\}$$

without knowing x_*.

Lemma. *The function*

$$\alpha \mapsto \phi(x_0 + \alpha s_0), \quad \mathbb{R} \to \mathbb{R},$$

attains its minimum at

$$\alpha = \alpha_0 := \frac{s_0^{\mathrm{T}} r_0}{\|s_0\|_A^2} = \frac{s_0^{\mathrm{T}} r_0}{s_0^{\mathrm{T}} A s_0},$$

where r_0 is the residual corresponding to the initial guess:

$$r_0 = y - A x_0.$$

Proof. The residual corresponding to $x = x_0 + \alpha s_0$ is

$$r \;=\; y - Ax \;=\; y - Ax_0 - \alpha A s_0 \;=\; r_0 - \alpha A s_0.$$

In consequence,

$$\begin{aligned} \phi(x) &= r^{\mathrm{T}} A^{-1} r \\ &= (r_0 - \alpha A s_0)^{\mathrm{T}} A^{-1} (r_0 - \alpha A s_0) \\ &= \alpha^2 s_0^{\mathrm{T}} A s_0 - 2\alpha s_0^{\mathrm{T}} r_0 + r_0^{\mathrm{T}} A^{-1} r_0, \end{aligned}$$

which, as a function of α, is a parabola that opens upwards, because $s_0^{\mathrm{T}} A s_0 > 0$. Hence, its minimum is at the unique zero of the derivative with respect to α, i.e., at $\alpha = \alpha_0$. □

About the choice of the search directions

Given a sequence of (non-zero) search directions $\{s_k\} \subset \mathbb{R}^n$, we can thus produce a sequence of approximate solutions by first choosing x_0 and then finding iteratively the minimizer of ϕ on the line passing through x_k in the direction s_k as follows:

$$x_{k+1} = x_k + \alpha_k s_k, \qquad \text{with } \alpha_k = \frac{s_k^{\mathrm{T}} r_k}{s_k^{\mathrm{T}} A s_k}, \qquad k = 0, 1, \ldots,$$

where r_k is the residual corresponding to the kth iterate, i.e.,

$$r_k = y - A x_k.$$

Notice that $\{\phi(x_k)\}$ is a decreasing sequence of real numbers because $\phi(x_{k+1})$ is always smaller than — or as small as — $\phi(x_k)$.

However, an efficient choice of the search directions $\{s_k\}$ is a subtle issue.

Probably, one of the first ideas that comes to mind is to choose

$$s_k \;=\; -\nabla\phi(x_k) \;=\; 2(y - Ax_k), \qquad k = 0, 1, \ldots,$$

because it gives the direction of the *steepest descent*. However, this does not in general provide a sequence $\{x_k\}$ that converges fast towards the global minimizer $x_* = A^{-1}y$, as demonstrated by the following example:

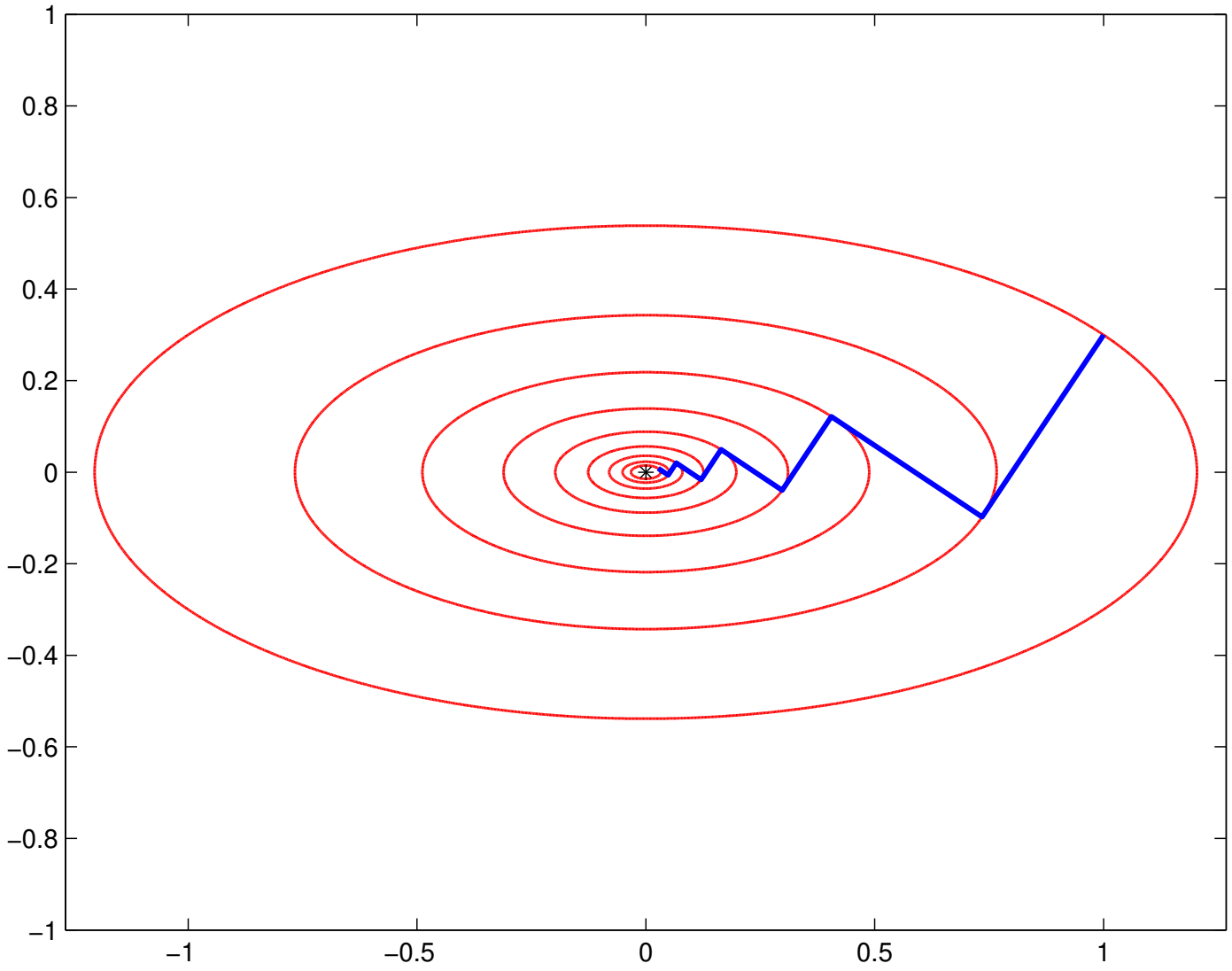
Let

$$A \;=\; \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix} \qquad \text{and} \qquad y \;=\; \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which means, in particular, that

$$\phi(x) \;=\; \phi(x^{(1)}, x^{(2)}) \;=\; (x^{(1)})^2 + 5(x^{(2)})^2.$$

The following image shows level contours of ϕ and the sequence $\{x_k\}_{k=0}^{9}$ starting from $x_0 = (1, 0.3)^{\mathrm{T}}$. The actual solution $x_* = (0, 0)^{\mathrm{T}}$ is marked with an asterisk.


1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1
−0.5
0
0.5
1

Minimizing ϕ over a hyperplane

Let $\{s_0, \ldots, s_k\}$ be a set of linearly independent search direction. Next, we consider finding the minimizer of ϕ on the hyperplane

$$\mathcal{S}_k = \{x \in \mathbb{R}^n \mid x = x_0 + S_k h,\ h \in \mathbb{R}^{k+1}\},$$

where $x_0 \in \mathbb{R}^n$ is the initial guess and $S_k = [s_0, \ldots, s_k] \in \mathbb{R}^{n \times (k+1)}$.

Lemma. *The function*

$$h \mapsto \phi(x_0 + S_k h), \quad \mathbb{R}^{k+1} \to \mathbb{R},$$

attains its minimum at

$$h = h_* = (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0,$$

where $r_0 = y - Ax_0$ is the residual corresponding to the initial guess.

Proof. Let us first prove that $S_k^{\mathrm{T}} A S_k \in \mathbb{R}^{(k+1)\times(k+1)}$ is invertible: Due to the positive definiteness of A, we have

$$S_k^{\mathrm{T}} A S_k z = 0 \quad \Longrightarrow \quad z^{\mathrm{T}} S_k^{\mathrm{T}} A S_k z = 0 \quad \Longrightarrow \quad S_k z = 0,$$

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

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

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

and thus

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

In particular, the coefficient matrix $S_k^{\mathrm{T}} A S_k$ of the quadratic term of $\phi(x_0 + S_k h)$ in h is positive definite:

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

where the equality holds if and only if $S_k u = 0$, i.e., $u = 0$. Thus, the basics of quadratic programming tell us that the unique zero of the gradient of $\phi(x_0 + S_k h)$ with respect to h, i.e.,

$$h_* \; = \; (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0,$$

is the unique minimizer of $\phi(x_0 + S_k h)$ over $h \in \mathbb{R}^{k+1}$. $\square$

A-conjugate search directions

Since finding the minimizer of ϕ over the hyperplane

$$\mathcal{S}_k = \{x \in \mathbb{R}^n \mid x = x_0 + S_k h,\ h \in \mathbb{R}^{k+1}\}$$

involves inverting a $(k+1) \times (k+1)$ matrix, such an approach is not necessarily very attractive.

On the other hand, as demonstrated by the numerical example above, minimizing ϕ sequentially in the directions $s_0, \ldots, s_k$ does not, in general, result in as good approximate solution as doing the minimization over the whole hyperplane $\mathcal{S}_k$ at once. (Clearly, the first two search directions of the numerical example were linearly independent, and thus minimization over the hyperplane $\mathcal{S}_2$, i.e., the whole $\mathbb{R}^2$, would have given the global minimizer $x_* = (0, 0)^{\mathrm{T}}$.)

However, the sequential minimization *does* produce the minimizer over $\mathcal{S}_k$ if the search directions $\{s_0, \ldots, s_k\}$ are chosen in a clever way.

We say that non-zero vectors $\{s_0, \ldots, s_k\} \subset \mathbb{R}^n$ are A-conjugate if

$$\langle s_i, s_j \rangle_A \;=\; s_i^{\mathrm{T}} A s_j \;=\; 0$$

for $i \neq j$. In other words, the vectors $\{s_0, \ldots, s_k\}$ are A-conjugate if they are orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle_A$.

The A-conjugacy condition can be expressed neatly with the help of the matrix $S_k = [s_0, \ldots, s_k] \in \mathbb{R}^{n \times (k+1)}$:

$$S_k^{\mathrm{T}} A S_k \;=\; \begin{bmatrix} s_0^{\mathrm{T}} \\ \vdots \\ s_k^{\mathrm{T}} \end{bmatrix} [As_0, \ldots, As_k] \;=\; \mathrm{diag}(d_0, d_1 \ldots, d_k) \in \mathbb{R}^{(k+1)\times(k+1)},$$

where $d_j = s_j^{\mathrm{T}} A s_j > 0$, $j = 0, \ldots, k$, due to the positive definiteness of the matrix A.

The following theorem demonstrates that it is useful to choose the search directions to be A-conjugate.

Theorem. *Let $x_0 \in \mathbb{R}^n$ be an initial guess and assume that the vectors $\{s_0, \ldots, s_k\} \subset \mathbb{R}^n$ are non-zero and A-conjugate. Then, the sequential minimizer of ϕ over these directions, i.e., $x_{k+1} \in \mathbb{R}^n$ obtained by the iteration*

$$x_{j+1} = x_j + \alpha_j s_j, \qquad \text{with } \alpha_j = \frac{s_j^{\mathrm{T}} r_j}{s_j^{\mathrm{T}} A s_j}, \qquad j = 0, \ldots, k,$$

is the minimizer of ϕ on the hyperplane

$$\mathcal{S}_k = \{x \in \mathbb{R}^n \mid x = x_0 + S_k h,\ h \in \mathbb{R}^{k+1}\}.$$

To put it short,

$$x_{k+1} = x_0 + S_k h_* = x_0 + S_k (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0.$$

Proof. Let $a_j = (\alpha_0, \ldots, \alpha_j)^{\mathrm{T}} \in \mathbb{R}^{j+1}$. With this notation we have

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

Moreover the residual corresponding to x_j is

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

In particular,

$$s_j^{\mathrm{T}} r_j \;=\; s_j^{\mathrm{T}} r_0 - s_j^{\mathrm{T}} AS_{j-1}a_{j-1} \;=\; s_j^{\mathrm{T}} r_0 + s_j^{\mathrm{T}} [As_0, \ldots, As_{j-1}]a_{j-1},$$

where the last term vanishes since s_j is A-conjugate to $\{s_0, \ldots, s_{j-1}\}$.

Hence,

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

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

$$\begin{aligned} (S_k^{\mathrm{T}} A S_k)^{-1} &= \big(\mathrm{diag}(s_0^{\mathrm{T}} A s_0, \ldots, s_k^{\mathrm{T}} A s_k)\big)^{-1} \\ &= \mathrm{diag}(1/(s_0^{\mathrm{T}} A s_0), \ldots, 1/(s_k^{\mathrm{T}} A s_k)), \end{aligned}$$

which means that

$$h_* \;=\; (S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0 \;=\; (S_k^{\mathrm{T}} A S_k)^{-1} \begin{bmatrix} s_0^{\mathrm{T}} r_0 \\ \vdots \\ s_k^{\mathrm{T}} r_0 \end{bmatrix} \;=\; \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_k \end{bmatrix}.$$

Consequently, $a_k = h_*$ and

$$x_{k+1} \;=\; x_0 + S_k a_k \;=\; x_0 + S_k h_*. \qquad \square$$

A useful corollary about the residuals

If the search directions are chosen to be A-conjugate, we have also extra information about the residuals:

Corollary. *If the non-zero search directions* $\{s_j\}_{j=0}^{k} \subset \mathbb{R}^n$ *are* A*-conjugate, then the residual* $r_{k+1} = y - Ax_{k+1}$ *satisfies*

$$r_{k+1} \perp \operatorname{span}\{s_0, \ldots, s_k\},$$

where the orthogonality is in the sense of the standard inner product.

Proof. Since $x_{k+1} = x_0 + S_k h_*$, it holds that

$$r_{k+1} \;=\; (y - Ax_0) - AS_k h_* \;=\; r_0 - AS_k h_*.$$

In consequence,

$$[r_{k+1}^{\mathrm{T}} s_0, \ldots, r_{k+1}^{\mathrm{T}} s_k] \;=\; r_{k+1}^{\mathrm{T}} S_k \;=\; r_0^{\mathrm{T}} S_k - h_*^{\mathrm{T}} S_k^{\mathrm{T}} A S_k \;=\; 0$$

because $h_*^{\mathrm{T}} = ((S_k^{\mathrm{T}} A S_k)^{-1} S_k^{\mathrm{T}} r_0)^{\mathrm{T}} = r_0^{\mathrm{T}} S_k (S_k^{\mathrm{T}} A S_k)^{-1}$.

How to construct A-conjugate search directions?

There are many ways to construct a set of A-conjugate search directions. If one chooses to use Krylov subspaces the result is the conjugate gradient algorithm:

Definition: *The kth Krylov subspace of A with the initial vector $r_0 = y - Ax_0$ is defined as*

$$\mathcal{K}_k = \mathcal{K}(A, r_0) = \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}, \qquad k = 1, 2, \ldots.$$

Note, in particular, that $A(\mathcal{K}_k) \subset \mathcal{K}_{k+1}$.

Take also note that $\mathcal{K}_{k-1} \subset \mathcal{K}_k$, where the dimension of the latter is *at most* k, and it is *at most* one higher than that of the former. (For example, if r_0 is an eigenvector of A, then the vectors spanning $\mathcal{K}_k$ are scalar multiples of each other, which means that $\dim(\mathcal{K}_k) = 1$ for all $k \geq 1$. Fortunately, it turns out that this is not a hindrance.)

The logic of the conjugate gradient algorithm

Let us construct a sequence of A-conjugate search directions inductively. The leading idea is that, given a set of A-conjugate search direction, we can either find a new A-conjugate direction or the previous iterate is already the global minimizer x_*, i.e., the unique solution of $Ax = y$.

1. Choose an initial guess $x_0 \in \mathbb{R}^n$.

2. If $r_0 = y - Ax_0 = 0$, we have found the solution $x_* = x_0$. Otherwise, set $s_0 = r_0$ (, which is, by the way, the steepest descent direction). Note, in particular, that the set of a single search direction $\{s_0\}$ is trivially A-conjugate and

$$\mathcal{K}_1 \;=\; \mathrm{span}\{s_0\} \;=\; \mathrm{span}\{r_0\}.$$

3. Suppose that we have non-zero and A-conjugate search directions $\{s_j\}_{j=0}^{k-1}$, $k \geq 1$, such that

$$\mathcal{K}_m = \text{span}\{s_0, \ldots, s_{m-1}\} = \text{span}\{r_0, \ldots, r_{m-1}\}, \qquad m = 1, \ldots, k, \tag{9}$$

where $r_j = y - Ax_j$, $j = 0, \ldots, k-1$, are the residuals corresponding to the iterates $\{x_j\}_{j=0}^{k-1}$ of the sequential minimization algorithm.

If $r_k = 0$, the algorithm has converged to $x_* = x_k$. Otherwise, we try to choose another A-conjugate and non-zero search direction $s_k \in \mathbb{R}^n$ so that (9) remains valid if k is replaced by $k+1$.

Assume thus that $r_k \neq 0$. Since

$$r_k \;=\; y - Ax_k \;=\; y - A(x_{k-1} + \alpha_{k-1}s_{k-1}) \;=\; r_{k-1} - \alpha_{k-1}As_{k-1}$$

and r_{k-1} and s_{k-1} belong by assumption to $\mathcal{K}_k$, the new residual r_k belongs to $\mathcal{K}_{k+1}$. Since r_k is orthogonal to $\{s_0, \ldots, s_{k-1}\}$, which span $\mathcal{K}_k$ and belong to $\mathcal{K}_{k+1}$, we must have

$$\mathcal{K}_{k+1} \;=\; \mathrm{span}\{s_0, \ldots, s_{k-1}, r_k\} \;=\; \mathrm{span}\{r_0, \ldots, r_{k-1}, r_k\}.$$

Let us try to find the new search direction s_k in the form

$$s_k \;=\; r_k + \beta_{k-1}s_{k-1}, \qquad \beta_{k-1} \in \mathbb{R}.$$

Note that this kind of vector belongs to $\mathcal{K}_{k+1}$ and, furthermore,

$$\mathcal{K}_{k+1} \;=\; \mathrm{span}\{s_0, \ldots, s_{k-1}, r_k\} \;=\; \mathrm{span}\{s_0, \ldots, s_{k-1}, s_k\}.$$

Consequently, all we have to worry about is the A-conjugacy condition:

We want to choose $\beta_{k-1} \in \mathbb{R}^k$ so that

$$\begin{aligned} s_j^{\mathrm{T}} A s_k &= s_j^{\mathrm{T}} A r_k + \beta_{k-1} s_j^{\mathrm{T}} A s_{k-1} \\ &= (A s_j)^{\mathrm{T}} r_k + \beta_{k-1} s_j^{\mathrm{T}} A s_{k-1} = 0 \end{aligned} \tag{10}$$

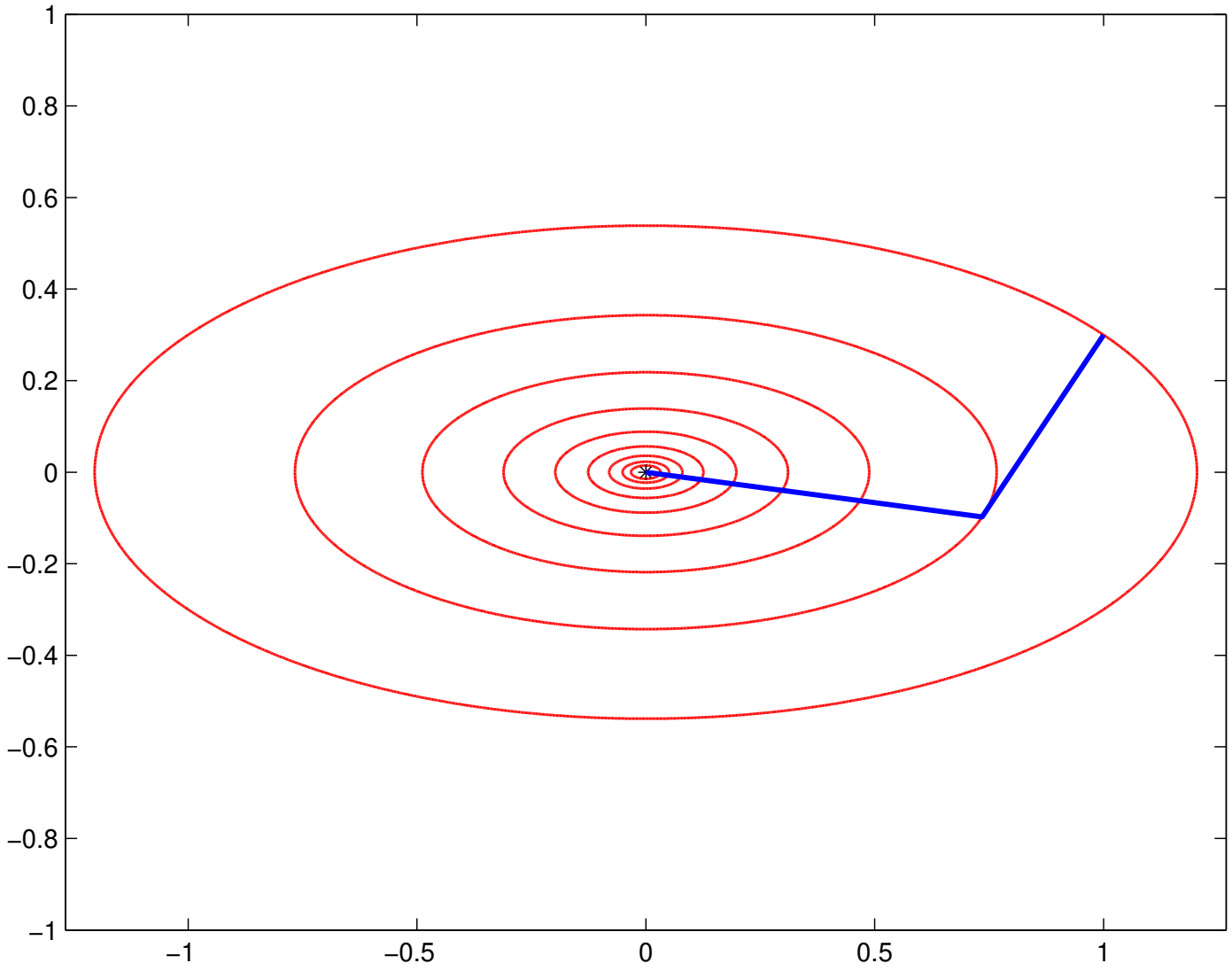
for $j = 0, \ldots, k-1$. Because $\{s_0, \ldots, s_{k-2}\} \subset \mathcal{K}_{k-1}$, we have

$$\{A s_0, \ldots, A s_{k-2}\} \subset \mathcal{K}_k = \operatorname{span}\{s_0, \ldots, s_{k-1}\},$$

and thus the vectors $\{A s_0, \ldots, A s_{k-2}\}$ are orthogonal to r_k. Hence, the A-conjugacy of $\{s_0, \ldots, s_{k-1}\}$ yields that only the last of the equations (10) is non-trivial.

Solving this equation for β_{k-1} results in the needed update rule

$$s_k = r_k + \beta_{k-1} s_{k-1}, \qquad \beta_{k-1} = -\frac{s_{k-1}^{\mathrm{T}} A r_k}{s_{k-1}^{\mathrm{T}} A s_{k-1}}.$$


1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1
−0.5
0
0.5
1