Numerical methods for Partial Differential Equations

Adriano Festa
Politecnico of Turin, Italy

Undergraduate Lecture at KSU
2022

Table of Contents

1. The model of elastic string or heated beam
2. The model of elastic membrane or heated plate
3. Solution of large linear algebraic systems
4. Models of temporal evolution
5. Time-advancing schemes
6. Convection-diffusion and transport problems
7. Conservation Laws - Introduction to Finite Volumes

When dealing numerically with differential models one is often lead to solve very large linear systems with sparse matrices. We will write such a system as

$$\mathbf{A}\mathbf{x} = \mathbf{b} \,, \tag{56}$$

where $\mathbf{A}$ is a square matrix of order n, non-singular and sparse, $\mathbf{b}$ the column vector of sources and $\mathbf{x}$ the column vector of unknowns.

We will denote by $\overline{\mathbf{x}}$ the *vector that solves* this system of equations.

For our purposes it is useful to introduce the notion of *residual vector* (residual) of equation (56) relative to a vector $\mathbf{x}$:

$$\mathbf{r}(\mathbf{x}) = \mathbf{b} - \mathbf{A}\mathbf{x} \,. \tag{57}$$

Note

$$\overline{\mathbf{x}} \text{ solves } (56) \qquad \Longleftrightarrow \qquad \mathbf{r}(\overline{\mathbf{x}}) = \mathbf{0} \,.$$

Solution methods

- **Direct methods**: they lead to solving the system by a finite number of algebraic operations (assuming ideally an exact arithmetic) through suitable matrix transformations (*factorizations*)
  Classical examples are the **Gaussian elimination method**, or the **Cholesky method** for symmetric positive-definite matrices, which rely, respectively, on the following factorizations:
  $$\mathbf{PA} = \mathbf{LU} \ , \qquad \mathbf{A} = \mathbf{C}\mathbf{C}^T \ .$$
  Using a direct method might have a prohibitive cost, both in terms of memory storage, given that the matrices $\mathbf{L}$ and $\mathbf{U}$ usually have much bigger number of non-zero entries than $\mathbf{A}$ (a phenomenon called *fill-in*), and also in terms of number of operations required to get to the solution.
- **Iterative methods**: they generate, starting from a tentative vector $\mathbf{x}^0$, a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ that converges to $\overline{\mathbf{x}}$.
  (Classical methods are Jacobi's, Gauss-Seidel's, relaxation methods, ...).

If $\mathbf{A}$ is a symmetric positive-definite matrix, then the solution $\overline{\mathbf{x}}$ of the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

can be seen as the *unique minimum point* of the functional

$$J : \mathbb{R}^n \to \mathbb{R}, \qquad J(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}\ ,$$

i.e., one has precisely

$$J(\overline{\mathbf{x}}) = \min_{\mathbf{x}\in\mathbb{R}^n} J(\mathbf{x})\ .$$

Then, the idea is building a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ such that

$$J(\overline{\mathbf{x}}) \leq \cdots < J(\mathbf{x}^{k+1}) < J(\mathbf{x}^k) < J(\mathbf{x}^{k-1}) < \cdots < J(\mathbf{x}^0)\ ,$$

with

$$\overline{\mathbf{x}} = \lim_{k\to\infty} \mathbf{x}^k\ .$$

Properties of the functional J

Let $\mathbf{x}$ be any element in $\mathbb{R}^n$ and let $\delta\mathbf{x}$ denote an arbitrary incremento given to $\delta\mathbf{x}$. Then,

$$\begin{aligned} J(\mathbf{x}+\delta\mathbf{x}) &= \tfrac{1}{2}(\mathbf{x}+\delta\mathbf{x})^T\mathbf{A}(\mathbf{x}+\delta\mathbf{x}) - (\mathbf{x}+\delta\mathbf{x})^T\mathbf{b} \\ &= \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b} \;+\; (\mathbf{A}\mathbf{x}-\mathbf{b})^T\delta\mathbf{x} \;+\; \tfrac{1}{2}\delta\mathbf{x}^T\mathbf{A}\delta\mathbf{x} \end{aligned}$$

or, equivalently,

$$J(\mathbf{x}+\delta\mathbf{x}) \;=\; J(\mathbf{x}) \;-\; \mathbf{r}(\mathbf{x})^T\delta\mathbf{x} \;+\; \tfrac{1}{2}\delta\mathbf{x}^T\mathbf{A}\delta\mathbf{x}\,.$$

Comparing this to the Taylor expansion of the functional J in $\mathbf{x}$ for the increment $\delta\mathbf{x}$

$$J(\mathbf{x}+\delta\mathbf{x}) \;=\; J(\mathbf{x}) \;+\; \nabla J(\mathbf{x})^T\delta\mathbf{x} \;+\; \tfrac{1}{2}\delta\mathbf{x}^T\mathbf{H}J(\mathbf{x})\delta\mathbf{x} + \cdots\,,$$

we deduce that

$$\nabla J(\mathbf{x}) = -\mathbf{r}(\mathbf{x}) \qquad \text{and} \qquad \mathbf{H}J(\mathbf{x}) = \mathbf{A}\,.$$

The absence of terms higher than degree two indicates that J is a *quadratic* functional (an upper-concave parabola if $n=1$, an elliptic paraboloid if $n=2$, and so on).

The relation

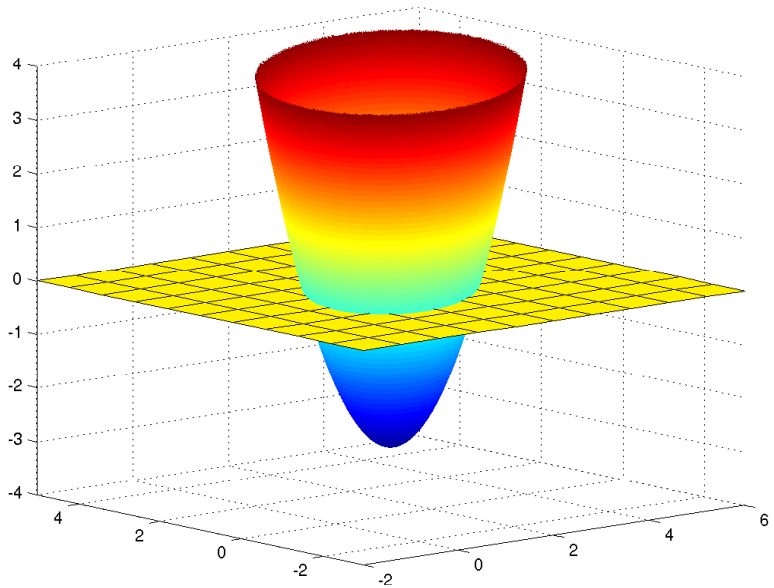
$$J(\mathbf{x}+\delta\mathbf{x}) \;=\; J(\mathbf{x}) \;-\; \mathbf{r}(\mathbf{x})^T\delta\mathbf{x} \;+\; \tfrac{1}{2}\delta\mathbf{x}^T\mathbf{A}\delta\mathbf{x}$$

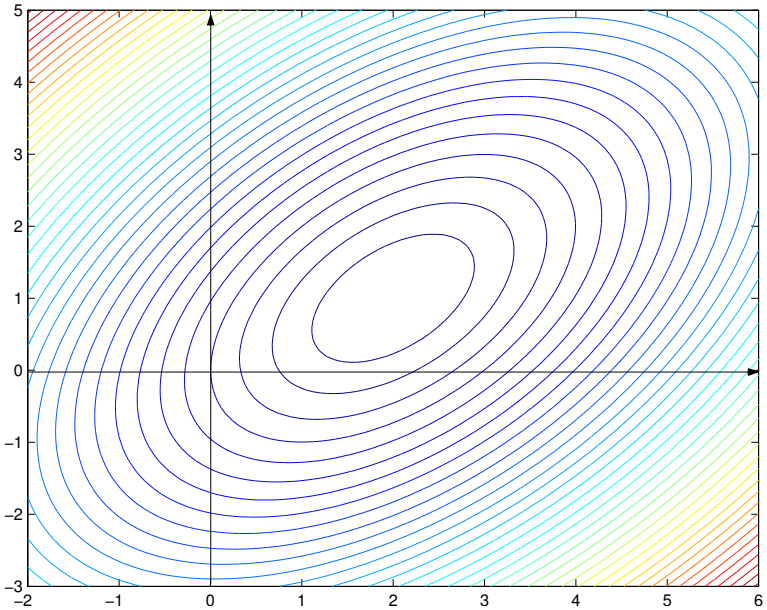
with $\mathbf{x} = \overline{\mathbf{x}}$ yields, for any $\delta\mathbf{x} \neq \mathbf{0}$,

$$J(\overline{\mathbf{x}}+\delta\mathbf{x}) \;=\; J(\overline{\mathbf{x}}) \;+\; \tfrac{1}{2}\delta\mathbf{x}^T\mathbf{A}\delta\mathbf{x} \;>\; J(\overline{\mathbf{x}})\,,$$

since $\mathbf{A}$ is definite positive. This confirm the already anticipated property

$$J(\overline{\mathbf{x}}) = \min_{\mathbf{x}\in\mathbb{R}^n} J(\mathbf{x})\,.$$





Mechanical interpretation

Consider the algebraic system

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$$

produced by the finite-element discretization of the elastic membrane problem. The corresponding functional $J : \mathbb{R}^N \to \mathbb{R}$ is given by

$$J(\boldsymbol{v}) = \tfrac{1}{2}\boldsymbol{v}^T \mathbf{A}\boldsymbol{v} \; - \; \boldsymbol{v}^T \boldsymbol{f} \; .$$

Let $v_h = \sum_{j=1}^N v_j \varphi_j \in V_h$ be the discrete admissible displacement associated with the vector $\boldsymbol{v} = (v_j)_{1 \le j \le N}$. Then, we have

$$\boldsymbol{v}^T \mathbf{A}\boldsymbol{v} = \int_\Omega \mu \|\nabla v_h\|^2 \, d\boldsymbol{x} \qquad \text{and} \qquad \boldsymbol{v}^T \boldsymbol{f} = \int_\Omega f \, v_h \, d\boldsymbol{x} \; ,$$

hence,

$$J(\boldsymbol{v}) = \tfrac{1}{2} \int_\Omega \mu \|\nabla v_h\|^2 \, d\boldsymbol{x} - \int_\Omega f \, v_h \, d\boldsymbol{x} = \mathcal{E}(v_h) \; ,$$

where we heve defined, on the space V of all admissible displacements, the functional

$$\mathcal{E} : V \to \mathbb{R} \, , \qquad \mathcal{E}(v) = \tfrac{1}{2} \int_\Omega \mu \|\nabla v\|^2 \, d\boldsymbol{x} - \int_\Omega f \, v \, d\boldsymbol{x} \; .$$

The latter represents the *total energy* of configuration v, given by the sum of the elastic strain $\frac{1}{2} \int_\Omega \mu \|\nabla v\|^2 \, d\boldsymbol{x}$ associated to the deformation with respect to the rest position, and the potential energy $- \int_\Omega f \, v \, d\boldsymbol{x}$ relative to the external force.

The minimum condition

$$J(\boldsymbol{u}) = \min_{\boldsymbol{v} \in \mathbb{R}^N} J(\boldsymbol{v}) \ ,$$

can be translated into

$$\mathcal{E}(u_h) = \min_{v_h \in V_h} \mathcal{E}(v_h) \ , \tag{58}$$

It tells that the solution $u_h \in V_h$ to the discrete variational problem is the configuration that *minimises the membrane's total energy among all admissible discrete configurations.*

In a similar fashion, one can prove that the solution $u \in V$ of the exact variational problem is characterised by

$$\mathcal{E}(u) = \min_{v \in V} \mathcal{E}(v) \ ,$$

which expresses the **Minimum Principle of the Total Energy**: in an equilibrium state the membrane's configuration *minimises the total energy of the system.*

This Principle is equivalent to the **Principle of Virtual Work**.

Descent direction and descend stepsize

A descent method is defined by a recursive relation of the sort

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{p}^k \ , \qquad k = 0, 1, 2, \dots \ ,$$

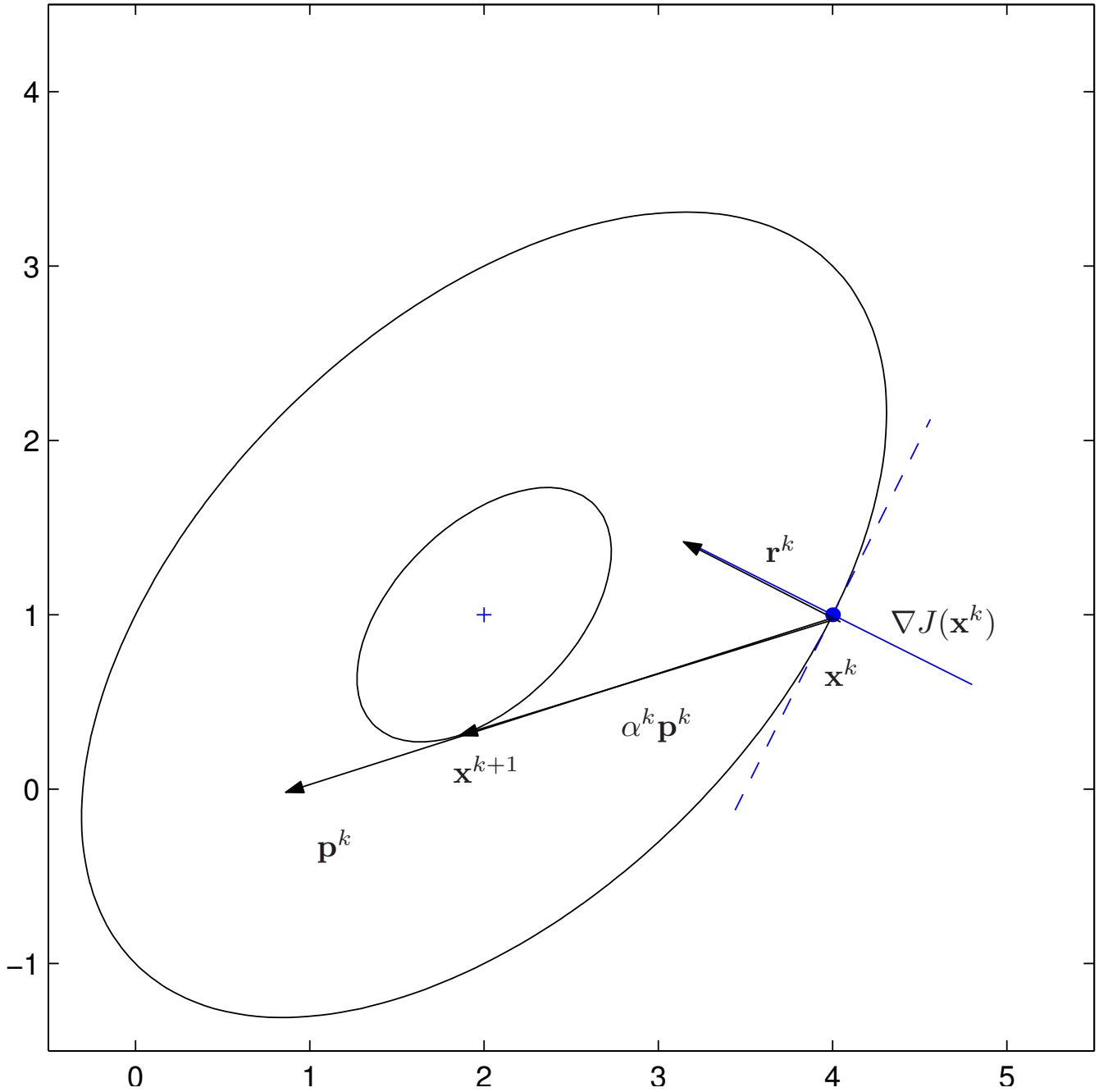
where the vector $\mathbf{p}^k$ is the **descent vector**, whereas the scalar α^k is the **descent step size**.

A vector $\mathbf{p}^k$ is an **admissible** descent vector if

$$\frac{\partial J}{\partial p^k}(\mathbf{x}^k) = \nabla J(\mathbf{x}^k)^T \mathbf{p}^k < 0 \ ,$$

that is, if the functional decreases as one moves very little along $\mathbf{p}^k$. Setting $\mathbf{r}^k = \mathbf{r}(\mathbf{x}^k) = -\nabla J(\mathbf{x}^k)$, the previous relationship becomes

$$(\mathbf{r}^k)^T \mathbf{p}^k > 0 \ .$$


4
3
2
1
0
−1
0
1
2
3
4
5
$\mathbf{r}^k$
$\nabla J(\mathbf{x}^k)$
$\mathbf{x}^k$
$\alpha^k\mathbf{p}^k$
$\mathbf{x}^{k+1}$
$\mathbf{p}^k$

Line-search

Once a descent vector $\mathbf{p}^k$ has been fixed, the step size α^k is uniquely determined by the requirement that the functional J has the maximum decrease in that direction.
Setting

$$\varphi(\alpha) = J(\mathbf{x}^k + \alpha \mathbf{p}^k) \ ,$$

the step size α^k will be such that

$$\varphi(\alpha^k) = \min_{\alpha \in \mathbb{R}} \varphi(\alpha) \ .$$

We have

$$\varphi(\alpha) = J(\mathbf{x}^k) - \alpha (\mathbf{r}^k)^T \mathbf{p}^k + \frac{1}{2} \alpha^2 (\mathbf{p}^k)^T \mathbf{A} \mathbf{p}^k \ ,$$

proving that $\varphi(\alpha)$ is a concave parabola, since $(\mathbf{p}^k)^T \mathbf{A} \mathbf{p}^k > 0$.

Hence, the minumum point is characterized by the relation

$$\varphi'(\alpha^k) = 0 \ , \qquad \text{which gives} \qquad \alpha^k = \frac{(\mathbf{r}^k)^T \mathbf{p}^k}{(\mathbf{p}^k)^T \mathbf{A} \mathbf{p}^k} \ .$$

The new residual is then computed as follows:

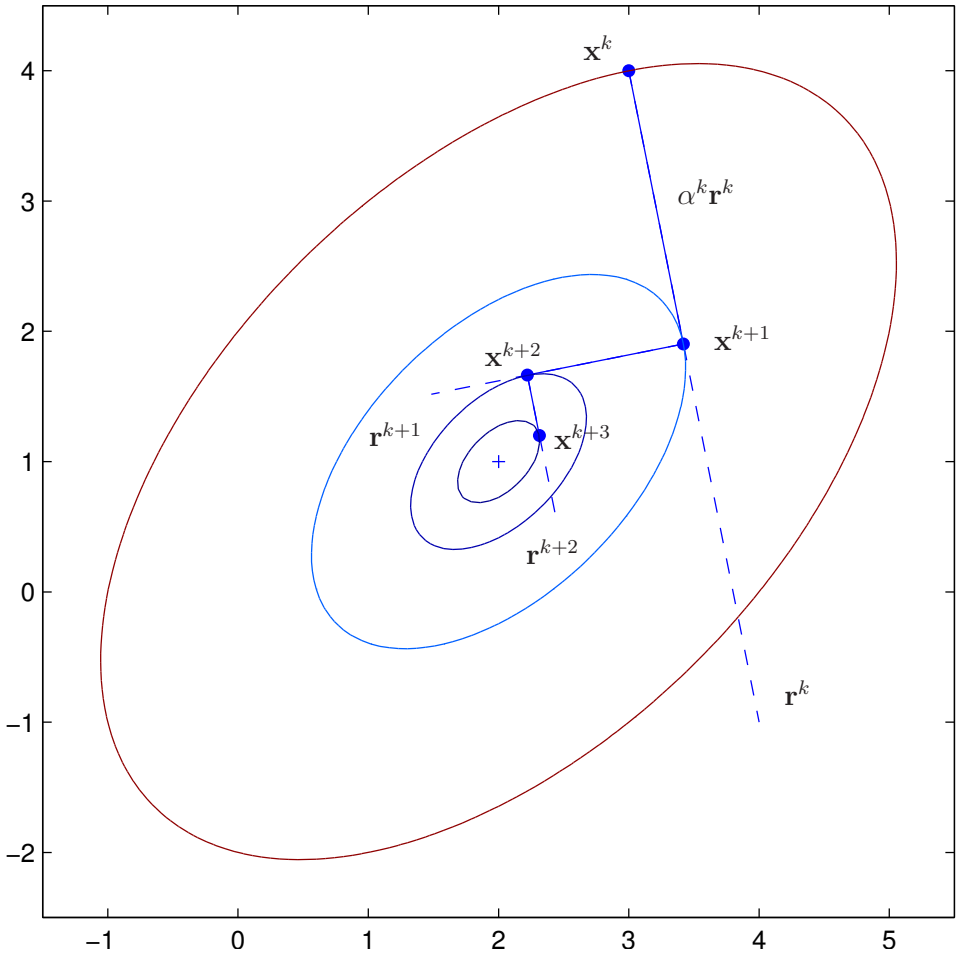
$$\mathbf{r}^{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}^{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}^k - \alpha^k \mathbf{A}\mathbf{p}^k = \mathbf{r}^k - \alpha^k \mathbf{A}\mathbf{p}^k \ .$$

The gradient method

Recalling the admissibility condition $(\mathbf{r}^k)^T \mathbf{p}^k > 0$, a natural choice consists in choosing

$$\mathbf{p}^k = \mathbf{r}^k \quad (= -\nabla J(\mathbf{x}^k)) \,,$$

meaning we move in the direction of local **steepest descent**.



The algorithm of the gradient goes like this:

$\mathbf{x}^0$ arbitrary

$\mathbf{r}^0 = \mathbf{b} - \mathbf{A}\mathbf{x}^0$

For $k = 0, 1, \ldots$ until convergence

$$\mathbf{z}^k = \mathbf{A}\mathbf{r}^k$$

$$\alpha^k = \frac{(\mathbf{r}^k)^T \mathbf{r}^k}{(\mathbf{r}^k)^T \mathbf{z}^k}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{r}^k$$

$$\mathbf{r}^{k+1} = \mathbf{r}^k - \alpha^k \mathbf{z}^k$$

Note that at each iteration we perform just *one matrix-vector multiplication* (to compute $\mathbf{z}^k$).
This is the most expensive part of the algorithm, the other steps being just operations on vectors.

The conjugate gradient method (CG-method)

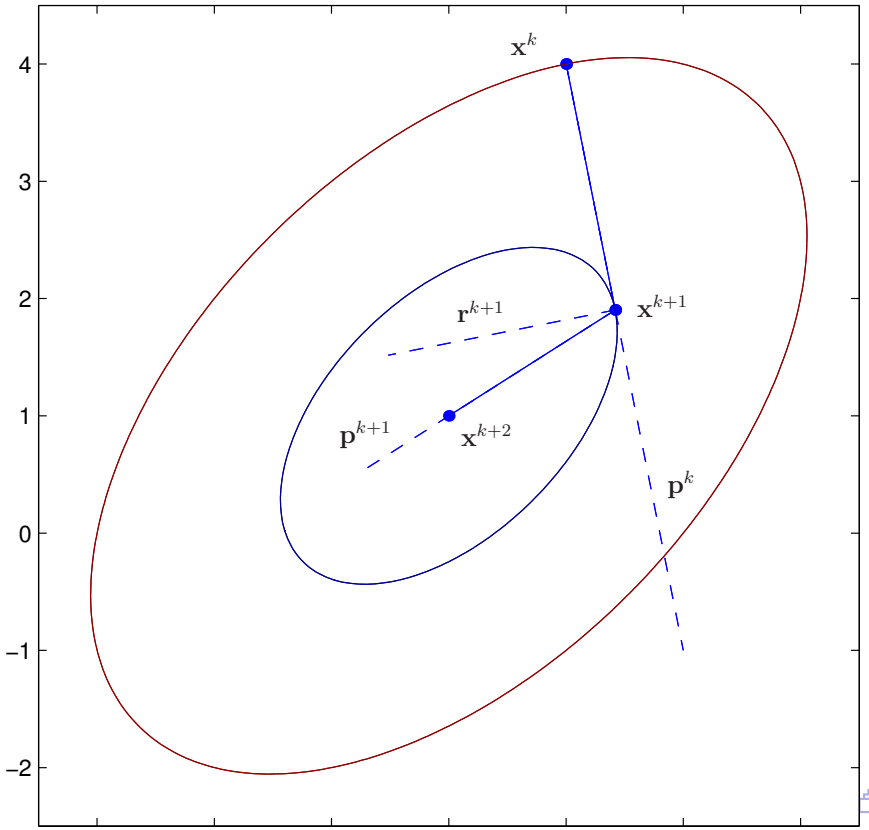
A faster descent is obtained by choosing the new descent direction $\mathbf{p}^{k+1}$ in the form

$$\mathbf{p}^{k+1} = \mathbf{r}^{k+1} + \beta \mathbf{p}^k$$

where $\beta = \beta^{k+1}$ is determined so to satisfy

$$(\mathbf{p}^k)^T \mathbf{A} \mathbf{p}^{k+1} = 0 \, .$$

This means that two successive descent directions $\mathbf{p}^k$ and $\mathbf{p}^{k+1}$ are $\mathbf{A}$-*orthogonal* (or, said equivalently, $\mathbf{A}$-*conjugate*).



The conjugate gradient algorithm is as follows:

$$\mathbf{x}^0 \text{ arbitrary}$$
$$\mathbf{r}^0 = \mathbf{b} - \mathbf{A}\mathbf{x}^0$$
$$\mathbf{p}^0 = \mathbf{r}^0$$
For $k = 0, 1, \ldots$ until convergence
$$\mathbf{z}^k = \mathbf{A}\mathbf{p}^k$$
$$\alpha^k = \frac{(\mathbf{r}^k)^T \mathbf{p}^k}{(\mathbf{p}^k)^T \mathbf{z}^k}$$
$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{p}^k$$
$$\mathbf{r}^{k+1} = \mathbf{r}^k - \alpha^k \mathbf{z}^k$$
$$\beta^{k+1} = \frac{(\mathbf{r}^{k+1})^T \mathbf{r}^{k+1}}{(\mathbf{r}^k)^T \mathbf{r}^k}$$
$$\mathbf{p}^{k+1} = \mathbf{r}^{k+1} + \beta^{k+1} \mathbf{p}^k$$

Note that, as before, the algorithm demands just *one matrix-vector multiplication* for each iteration.

Convergence - Rates of convergence

Let us measure the error $\mathbf{x}^k - \overline{\mathbf{x}}$ in the so-called *energy norm* $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A}\mathbf{x}}$.

One shows that for the **gradient method** one has

$$\|\mathbf{x}^k - \overline{\mathbf{x}}\|_{\mathbf{A}} \leq \left(\frac{\mathrm{cond}_2(\mathbf{A}) - 1}{\mathrm{cond}_2(\mathbf{A}) + 1} \right)^k \|\mathbf{x}^0 - \overline{\mathbf{x}}\|_{\mathbf{A}} \,, \qquad k \geq 0 \,,$$

whereas for the **conjugate gradient method** one has

$$\|\mathbf{x}^k - \overline{\mathbf{x}}\|_{\mathbf{A}} \leq 2 \left(\frac{\sqrt{\mathrm{cond}_2(\mathbf{A})} - 1}{\sqrt{\mathrm{cond}_2(\mathbf{A})} + 1} \right)^k \|\mathbf{x}^0 - \overline{\mathbf{x}}\|_{\mathbf{A}} \,, \qquad k \geq 0 \,.$$

This implies that to obtain a relative error reduction proportional to some $\varepsilon > 0$, i.e., to obtain

$$\frac{\|\mathbf{x}^k - \overline{\mathbf{x}}\|_{\mathbf{A}}}{\|\mathbf{x}^0 - \overline{\mathbf{x}}\|_{\mathbf{A}}} \simeq \varepsilon$$

the number k of needed iterations will satisfy

$$k \simeq \tfrac{1}{2} |\log \varepsilon| \, \mathrm{cond}_2(\mathbf{A}) \qquad \textbf{(gradient method)} \,,$$

$$k \simeq \tfrac{1}{2} |\log \varepsilon| \, \sqrt{\mathrm{cond}_2(\mathbf{A})} \qquad \textbf{(conjugate gradient method)} \,.$$

Preconditioning

In order to furtherly reduce the number of iterations to achieve a given accuracy, a popular strategy consists in applying the Conjugate Gradient method to the equivalent algebraic system

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{x} = \mathbf{P}^{-1}\mathbf{b}\ ,$$

where $\mathbf{P}$ is a so-called **preconditioning matrix** for $\mathbf{A}$, namely it satisfies the conditions

- $\mathrm{cond}_2(\mathbf{P}^{-1}\mathbf{A}) \ll \mathrm{cond}_2(\mathbf{A})$;
- the entries of $\mathbf{P}$ are computable in a cheap way;
- a matrix-product multiplication $\mathbf{z} = \mathbf{P}^{-1}\mathbf{y}$, or equivalently the solution of a linear system

  $$\mathbf{P}\mathbf{z} = \mathbf{y}\ ,$$

  can be accomplished at a comparable cost to that of computing $\mathbf{z} = \mathbf{A}\mathbf{y}$.

By such a strategy, we obtain a **preconditioned conjugate gradient method**.

A technique that often leads to a rather effective preconditioning matrix is to construct an *incomplete factorisation* of $\mathbf{A}$.

Example

For a uniform triangulation of a square by means of $2N^2$ rectangular triangles with edge length $h = N^{-1}$, one has:

Method	Iter vs condition number	Iter vs h or N
Gradient	$\mathcal{O}(\mathrm{cond}_2(\boldsymbol{A}))$	$\mathcal{O}(h^{-2}) = \mathcal{O}(N^2)$
Conjugate Gradient	$\mathcal{O}(\sqrt{\mathrm{cond}_2(\boldsymbol{A})})$	$\mathcal{O}(h^{-1}) = \mathcal{O}(N)$
Preconditioned Conjugate Gradient	$\mathcal{O}(\sqrt{\mathrm{cond}_2(\boldsymbol{P}^{-1}\boldsymbol{A})})$	$\mathcal{O}(h^{-1/2}) = \mathcal{O}(\sqrt{N})$

Here, denoting by

$$\mathbf{A} = \mathbf{C}\mathbf{C}^T$$

the *Cholesky factorization* of the symmetric positive-definite matrix $\mathbf{A}$, one uses as a preconditioner some

$$\mathbf{P} = \boldsymbol{\mathcal{C}}\boldsymbol{\mathcal{C}}^T \,, \qquad \text{with} \qquad \boldsymbol{\mathcal{C}} \simeq \mathbf{C} \,.$$