

24.7 Steepest Descent and Conjugate Gradients

Section 2 solved $\mathbf{Ax} = \mathbf{b}$ by elimination: about $\frac{2}{3}n^3$ operations, an exact answer after finitely many of them, and a factorization that fills in most of the zeros the matrix had. For a large sparse matrix that is the wrong bargain. This section builds two iterations that touch $\mathbf{A}$ only through the product $\mathbf{x} \mapsto \mathbf{Ax}$, for the case where $\mathbf{A}$ is positive definite, and counts what each of them buys. The first costs about κ steps per decimal digit of accuracy, the second about $\sqrt{\kappa}$. Where that square root comes from is the point of the section.

Throughout, $\mathbf{A} \in M_n(\mathbb{R})$ is symmetric and positive definite, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$ is the exact solution. Iteration indices are superscripts in parentheses. We work over $\mathbb{R}$ because the function this section minimizes is built from $\mathbf{b}^\top \mathbf{x}$, which over $\mathbb{C}$ is not a real number; everything below survives over $\mathbb{C}$ with $\mathbf{b}^* \mathbf{x}$ replaced by its real part, and nothing else changes.

24.7.1 A linear system is a minimization problem

Positive definiteness is a statement about the number $\mathbf{x}^\top \mathbf{Ax}$ (Definition 13.1.1), so it invites us to look for a *quantity* that the solution makes small, rather than an equation the solution satisfies. There is one, and it is the simplest possible candidate: a quadratic minus a linear term.

Solving $\mathbf{Ax} = \mathbf{b}$ is the same as sitting at the bottom of a bowl.

Definition 24.7.1. The Energy Inner Product and the Energy Norm

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite. The **energy inner product** is

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} := \mathbf{y}^\top \mathbf{Ax} \quad (\mathbf{x}, \mathbf{y} \in \mathbb{R}^n),$$

and the **energy norm** is $\|\mathbf{x}\|_{\mathbf{A}} := \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}^{1/2} = (\mathbf{x}^\top \mathbf{Ax})^{1/2}$. Two vectors are **A-orthogonal**, or **conjugate**, when $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = 0$.

This is the inner product of Chapter 13 §04, written there as $\mathbf{y}^* \mathbf{Ax}$. It really is one: it is bilinear because matrix multiplication is; it is symmetric because the 1×1 matrix $\mathbf{y}^\top \mathbf{Ax}$ equals its own transpose $\mathbf{x}^\top \mathbf{Ay}$; and $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} > 0$ for $\mathbf{x} \neq \mathbf{0}$ is clause (P2) of Definition 13.1.1 in its strict form. So $\|\cdot\|_{\mathbf{A}}$ is the induced norm of Definition 11.1.3, and every fact about norms and orthogonality from Chapter 11 applies verbatim. That is the reason for introducing it: it lets us reuse the Pythagorean theorem in a geometry bent by $\mathbf{A}$.

The word *energy* comes from the springs of Section 9, where $\frac{1}{2}\mathbf{x}^\top \mathbf{Kx}$ is the potential energy stored in a displacement $\mathbf{x}$.

Proposition 24.7.1. The Solution Minimizes the Energy Functional

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite, $\mathbf{b} \in \mathbb{R}^n$, and define

$$\phi(\mathbf{x}) := \frac{1}{2} \mathbf{x}^\top \mathbf{Ax} - \mathbf{b}^\top \mathbf{x} \quad (\mathbf{x} \in \mathbb{R}^n).$$

Then, with $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$,

$$\phi(\mathbf{x}) - \phi(\mathbf{x}^*) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{A}}^2 \quad \text{for every } \mathbf{x} \in \mathbb{R}^n.$$

In particular $\mathbf{x}^*$ is the **unique** minimizer of ϕ .

Proof. $\mathbf{A}$ is invertible, since $\mathbf{A} \succ \mathbf{O}$ has only positive eigenvalues (Theorem 13.1.5, (a) $\Rightarrow$ (b)), so $\mathbf{x}^*$ is defined and $\mathbf{A}\mathbf{x}^* = \mathbf{b}$. Expand, using $\mathbf{A}^\top = \mathbf{A}$:

$$\begin{aligned} \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{A}}^2 &= \frac{1}{2}\mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbf{x}^\top \mathbf{A} \mathbf{x}^* + \frac{1}{2}(\mathbf{x}^*)^\top \mathbf{A} \mathbf{x}^* \\ &= \frac{1}{2}\mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbf{b}^\top \mathbf{x} + \frac{1}{2}\mathbf{b}^\top \mathbf{x}^*, \end{aligned}$$

where the second line uses $\mathbf{A}\mathbf{x}^* = \mathbf{b}$ twice and $\mathbf{x}^\top \mathbf{b} = \mathbf{b}^\top \mathbf{x}$. The first two terms are $\phi(\mathbf{x})$. Putting $\mathbf{x} = \mathbf{x}^*$ gives $0 = \phi(\mathbf{x}^*) + \frac{1}{2}\mathbf{b}^\top \mathbf{x}^*$, so the leftover constant is $-\phi(\mathbf{x}^*)$, which is the identity. Since $\|\cdot\|_{\mathbf{A}}$ is a norm, the right-hand side is ≥ 0 with equality only at $\mathbf{x} = \mathbf{x}^*$. This proves the proposition. ■

So every iteration below may be read twice: as an attempt to solve a linear system, and as an attempt to walk downhill on a bowl whose shape is $\mathbf{A}$. The identity also tells us which error to measure. The natural yardstick is not $\|\mathbf{x} - \mathbf{x}^*\|_2$ but $\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{A}}$, because that is the one the algorithms actually make small.

Two quantities recur. The **error** is $\boldsymbol{\varepsilon} = \mathbf{x} - \mathbf{x}^*$ and the **residual** is $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$, and they are related by

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}^* - \mathbf{A}\mathbf{x} = -\mathbf{A}\boldsymbol{\varepsilon}. \quad (24.7.1)$$

The residual is computable and the error is not; Eq. 24.7.1 is the exchange rate between them.

24.7.2 Steepest descent

Fix $\mathbf{x}$ and a direction $\mathbf{d} \neq \mathbf{0}$, and ask what happens along the line $\mathbf{x} + t\mathbf{d}$. Expanding ϕ and using $\mathbf{A}^\top = \mathbf{A}$,

$$\phi(\mathbf{x} + t\mathbf{d}) = \phi(\mathbf{x}) - t\mathbf{d}^\top \mathbf{r} + \frac{1}{2}t^2 \mathbf{d}^\top \mathbf{A} \mathbf{d}, \quad (24.7.2)$$

where $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$. Since $\mathbf{d}^\top \mathbf{A} \mathbf{d} > 0$, the right-hand side is a parabola in t opening upwards, so it has one minimum, at

$$t = \frac{\mathbf{d}^\top \mathbf{r}}{\mathbf{d}^\top \mathbf{A} \mathbf{d}}, \quad \text{with decrease } \frac{(\mathbf{d}^\top \mathbf{r})^2}{2\mathbf{d}^\top \mathbf{A} \mathbf{d}}. \quad (24.7.3)$$

Choosing t this way is an **exact line search**; it costs one product $\mathbf{A}\mathbf{d}$.

Which $\mathbf{d}$ should we take? Read Eq. 24.7.2 for small t : the decrease is $t\mathbf{d}^\top \mathbf{r}$ to first order, and among unit vectors $\mathbf{d}$ the quantity $\mathbf{d}^\top \mathbf{r}$ is largest exactly at $\mathbf{d} = \mathbf{r}/\|\mathbf{r}\|_2$, by Cauchy–Schwarz (Theorem 11.1.5) with its equality case. So the residual is the direction in which ϕ falls fastest at the starting point, measured in the ordinary Euclidean norm.

Warning. “Steepest” is a statement about the first instant, not about the step. The direction that minimizes ϕ over the whole line is not $\mathbf{r}$ but $\mathbf{A}^{-1}\mathbf{r}$: with $\mathbf{d} = \mathbf{A}^{-1}\mathbf{r}$, Eq. 24.7.3 gives $t = 1$ and $\mathbf{x} + \mathbf{d} = \mathbf{x} + \mathbf{A}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{x}^*$, the exact answer in one step. That direction is unavailable, because producing it means solving the very system we are trying to solve. Every method in this section is an attempt to approximate $\mathbf{A}^{-1}\mathbf{r}$ cheaply.

Definition 24.7.2. Steepest Descent

Let $\mathbf{A} \succ \mathbf{O}$ and $\mathbf{b} \in \mathbb{R}^n$, and let $\mathbf{x}^{(0)} \in \mathbb{R}^n$. **Steepest descent** is the iteration: for $k = 0, 1, 2, \dots$, set $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$; if $\mathbf{r}^{(k)} = \mathbf{0}$, stop, since $\mathbf{x}^{(k)} = \mathbf{x}^*$; otherwise put

$$t_k = \frac{(\mathbf{r}^{(k)})^\top \mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^\top \mathbf{A} \mathbf{r}^{(k)}}, \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k \mathbf{r}^{(k)}.$$

The denominator is positive because $\mathbf{r}^{(k)} \neq \mathbf{0}$ and $\mathbf{A} \succ \mathbf{O}$, so t_k is defined, and $t_k > 0$. One step costs one matrix–vector product.

To see how fast this is we need one inequality, and it is a statement about weighted averages, not about matrices at all.

Lemma 24.7.2. Kantorovich’s Inequality

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite with eigenvalues $\lambda_1 \geq \dots \geq \lambda_n > 0$. Then for every $\mathbf{y} \in \mathbb{R}^n$,

$$(\mathbf{y}^\top \mathbf{A} \mathbf{y})(\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{y}) \leq \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1 \lambda_n} (\mathbf{y}^\top \mathbf{y})^2.$$

Idea. In an orthonormal eigenbasis both factors become weighted averages of the eigenvalues, one of λ and one of $1/\lambda$, with the same weights. The obstacle is that $1/\lambda$ is not linear, so the second average is not determined by the first. It is, however, *bounded* by it: on the interval $[\lambda_n, \lambda_1]$ the curve $1/\lambda$ lies below the straight line joining its endpoints. Replacing the curve by that line makes the second average a linear function of the first, and the product becomes a scalar quadratic that we maximize by hand.

Proof. Write $a = \lambda_1$ and $c = \lambda_n$. For every λ with $c \leq \lambda \leq a$,

$$\frac{a + c - \lambda}{ac} - \frac{1}{\lambda} = \frac{\lambda(a + c - \lambda) - ac}{ac\lambda} = \frac{-(\lambda - a)(\lambda - c)}{ac\lambda} \geq 0, \quad (*)$$

because $\lambda - a \leq 0$, $\lambda - c \geq 0$, and $ac\lambda > 0$.

We may assume $\mathbf{y} \neq \mathbf{0}$, the case $\mathbf{y} = \mathbf{0}$ being an equality $0 = 0$, and by homogeneity (both sides scale by s^4 under $\mathbf{y} \mapsto s\mathbf{y}$) we may assume $\mathbf{y}^\top \mathbf{y} = 1$. By the spectral theorem (Corollary 12.5.2) there is an orthonormal basis $\mathbf{q}_1, \dots, \mathbf{q}_n$ of $\mathbb{R}^n$ with $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$. Let $u_i = \mathbf{q}_i^\top \mathbf{y}$ be the coordinates of $\mathbf{y}$ in that basis, and put $w_i = u_i^2$. Then $\sum_i w_i = \|\mathbf{y}\|_2^2 = 1$

by Parseval's identity (Theorem 11.9.2), every $w_i \geq 0$, and

$$\mathbf{y}^\top \mathbf{A} \mathbf{y} = \sum_i \lambda_i w_i =: \mu, \quad \mathbf{y}^\top \mathbf{A}^{-1} \mathbf{y} = \sum_i \frac{w_i}{\lambda_i},$$

the second because $\mathbf{A}^{-1} \mathbf{q}_i = \lambda_i^{-1} \mathbf{q}_i$. Since the w_i are weights summing to 1 and every $\lambda_i \in [c, a]$, the number μ lies in $[c, a]$. Averaging (*) with the weights w_i ,

$$\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{y} \leq \sum_i w_i \frac{a + c - \lambda_i}{ac} = \frac{a + c - \mu}{ac}.$$

Hence $(\mathbf{y}^\top \mathbf{A} \mathbf{y})(\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{y}) \leq \mu(a + c - \mu)/(ac)$. Finally, for any real μ ,

$$\mu(a + c - \mu) = \frac{(a + c)^2}{4} - \left(\frac{a + c}{2} - \mu \right)^2 \leq \frac{(a + c)^2}{4},$$

and dividing by ac gives the bound. This proves the lemma. ■

Remark. Equality is attainable: take $\mathbf{y} = (\mathbf{q}_1 + \mathbf{q}_n)/\sqrt{2}$, so that $w_1 = w_n = \frac{1}{2}$ and every other weight is 0. Then (*) is an equality at each λ_i carrying weight, since those are a and c , and $\mu = (a + c)/2$ makes the last step an equality too. Only the weights $w_i = u_i^2$ enter, so $(\mathbf{q}_1 - \mathbf{q}_n)/\sqrt{2}$ serves just as well, and so does any nonzero multiple of either. [Example \(The bound is attained\)](#) does exactly this, up to sign and scale.

We also record the condition number in the form we shall use it. For $\mathbf{A} \succ \mathbf{O}$, Lemma 17.3.5 gives $\|\mathbf{A}\|_2 = \lambda_1$, since all eigenvalues are positive; the eigenvalues of $\mathbf{A}^{-1}$ are the λ_i^{-1} , so the same lemma gives $\|\mathbf{A}^{-1}\|_2 = \lambda_n^{-1}$. Hence, by Definition 16.8.1,

$$\kappa := \kappa_2(\mathbf{A}) = \frac{\lambda_1}{\lambda_n} \geq 1. \quad (24.7.4)$$

Theorem 24.7.3. Convergence of Steepest Descent

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite with $\kappa = \kappa_2(\mathbf{A})$, let $\mathbf{b} \in \mathbb{R}^n$, and let $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots$ be the steepest descent iterates. Then for every k for which $\mathbf{x}^{(k)}$ is defined,

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\|_{\mathbf{A}} \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^k \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_{\mathbf{A}}.$$

Idea. One step at a time. Write the new energy-norm error in terms of the old one; the exact line search makes the difference a single explicit fraction, whose numerator and denominator are exactly the three quantities Kantorovich's inequality compares. What comes out is $1 - 4\lambda_1\lambda_n/(\lambda_1 + \lambda_n)^2$, and that number is the square of $(\kappa - 1)/(\kappa + 1)$.

Proof. It is enough to prove the estimate for one step and then iterate. Fix k and write $\mathbf{x} = \mathbf{x}^{(k)}$, $\mathbf{r} = \mathbf{r}^{(k)}$, $t = t_k$, and $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^{(k)} := \mathbf{x}^{(k)} - \mathbf{x}^*$, and assume $\mathbf{r} \neq \mathbf{0}$. By Eq. 24.7.1, $\mathbf{r} = -\mathbf{A}\boldsymbol{\varepsilon}$, so

$$\|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 = \boldsymbol{\varepsilon}^\top \mathbf{A} \boldsymbol{\varepsilon} = (\mathbf{A}^{-1}\mathbf{r})^\top \mathbf{A} (\mathbf{A}^{-1}\mathbf{r}) = \mathbf{r}^\top \mathbf{A}^{-1} \mathbf{r},$$

using $\mathbf{A}^\top = \mathbf{A}$ and $(\mathbf{A}^{-1})^\top = \mathbf{A}^{-1}$. The new error is $\boldsymbol{\varepsilon} + t\mathbf{r}$, so

$$\begin{aligned} \|\boldsymbol{\varepsilon} + t\mathbf{r}\|_{\mathbf{A}}^2 &= \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 + 2t\mathbf{r}^\top \mathbf{A} \boldsymbol{\varepsilon} + t^2 \mathbf{r}^\top \mathbf{A} \mathbf{r} \\ &= \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 - 2t\mathbf{r}^\top \mathbf{r} + t^2 \mathbf{r}^\top \mathbf{A} \mathbf{r}, \end{aligned}$$

the second line because $\mathbf{A}\boldsymbol{\varepsilon} = -\mathbf{r}$. Substituting $t = (\mathbf{r}^\top \mathbf{r})/(\mathbf{r}^\top \mathbf{A} \mathbf{r})$ collapses the last two terms:

$$\|\boldsymbol{\varepsilon} + t\mathbf{r}\|_{\mathbf{A}}^2 = \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 - \frac{(\mathbf{r}^\top \mathbf{r})^2}{\mathbf{r}^\top \mathbf{A} \mathbf{r}} = \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 \left(1 - \frac{(\mathbf{r}^\top \mathbf{r})^2}{(\mathbf{r}^\top \mathbf{A} \mathbf{r})(\mathbf{r}^\top \mathbf{A}^{-1} \mathbf{r})}\right),$$

where the last step divides and multiplies by $\|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 = \mathbf{r}^\top \mathbf{A}^{-1} \mathbf{r}$, which is nonzero because $\mathbf{r} \neq \mathbf{0}$. By Lemma 24.7.2 applied to $\mathbf{y} = \mathbf{r}$,

$$\frac{(\mathbf{r}^\top \mathbf{r})^2}{(\mathbf{r}^\top \mathbf{A} \mathbf{r})(\mathbf{r}^\top \mathbf{A}^{-1} \mathbf{r})} \geq \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2},$$

so the bracket is at most

$$1 - \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2} = \frac{(\lambda_1 - \lambda_n)^2}{(\lambda_1 + \lambda_n)^2} = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2,$$

the last equality on dividing numerator and denominator by λ_n^2 and using Eq. 24.7.4. Since $\boldsymbol{\varepsilon} + t\mathbf{r} = \mathbf{x}^{(k+1)} - \mathbf{x}^* = \boldsymbol{\varepsilon}^{(k+1)}$, taking square roots gives $\|\boldsymbol{\varepsilon}^{(k+1)}\|_{\mathbf{A}} \leq \frac{\kappa-1}{\kappa+1} \|\boldsymbol{\varepsilon}^{(k)}\|_{\mathbf{A}}$, and an induction on k proves the theorem. ■

Example. The bound is attained

Let $\mathbf{A} = \text{diag}(9, 1)$, $\mathbf{b} = (9, 1)$, and $\mathbf{x}^{(0)} = (2, -8)$. Run three steps of steepest descent and compare with Theorem 24.7.3.

Solution. Here $\mathbf{x}^* = (1, 1)$ and $\kappa = 9$, so the bound's factor is $\frac{4}{5}$. The first residual is

$$\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)} = (9, 1) - (18, -8) = (-9, 9),$$

so $\mathbf{r}^\top \mathbf{r} = 162$ and $\mathbf{r}^\top \mathbf{A} \mathbf{r} = 9 \cdot 81 + 81 = 810$, giving $t_0 = \frac{1}{5}$. Then

$$\mathbf{x}^{(1)} = (2, -8) + \frac{1}{5}(-9, 9) = \left(\frac{1}{5}, -\frac{31}{5}\right), \quad \mathbf{r}^{(1)} = \left(\frac{36}{5}, \frac{36}{5}\right).$$

The two components of $\mathbf{r}^{(1)}$ are again equal in absolute value, so $t_1 = \frac{1}{5}$ as well, and $\mathbf{x}^{(2)} = \left(\frac{41}{25}, -\frac{119}{25}\right)$, $\mathbf{x}^{(3)} = \left(\frac{61}{125}, -\frac{451}{125}\right)$. The squared energy-norm errors are

$$90, \quad \frac{288}{5}, \quad \frac{4608}{125}, \quad \frac{73728}{3125},$$

and each is exactly $\frac{16}{25}$ times the one before. So the ratio is $\frac{4}{5}$ at every step: the bound of [Theorem 24.7.3](#) holds with equality here, and no better constant is possible.

The picture behind the example is a staircase. On an elongated bowl the steepest direction points across the valley rather than along it, so the iterates cross and re-cross. The residuals $\mathbf{r}^{(0)}$ and $\mathbf{r}^{(2)}$ in [Example \(The bound is attained\)](#) are parallel: the method keeps searching in directions it has already used.

Quick check. Show that consecutive steepest descent residuals are always orthogonal, $(\mathbf{r}^{(k+1)})^\top \mathbf{r}^{(k)} = 0$. What does that say about the shape of the path?

Solution. $\mathbf{r}^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)} - t_k \mathbf{A}\mathbf{r}^{(k)} = \mathbf{r}^{(k)} - t_k \mathbf{A}\mathbf{r}^{(k)}$, so

$$(\mathbf{r}^{(k+1)})^\top \mathbf{r}^{(k)} = \mathbf{r}^\top \mathbf{r} - t_k \mathbf{r}^\top \mathbf{A}\mathbf{r} = 0$$

by the choice of t_k , writing $\mathbf{r} = \mathbf{r}^{(k)}$. So each step turns a right angle: the path is a staircase. In two dimensions that forces every second direction to be parallel, which is why [Example \(The bound is attained\)](#) never escapes its worst case.

24.7.3 Conjugate directions

The defect is now visible: each line search spoils the minimality achieved along the previous direction. We want directions $\mathbf{p}^{(0)}, \mathbf{p}^{(1)}, \dots$ for which a search along $\mathbf{p}^{(k)}$ leaves the earlier work intact, so that after k steps the iterate is optimal over the whole k -dimensional affine piece explored so far.

Read the requirement off the goal. Being optimal over $\mathbf{x}^{(0)} + \text{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k-1)})$ means, by [Proposition 24.7.1](#), that the error is $\mathbf{A}$ -orthogonal to that span; and one step changes the error by $t_k \mathbf{p}^{(k)}$. So the condition that preserves the earlier orthogonality is that $\mathbf{p}^{(k)}$ itself be $\mathbf{A}$ -orthogonal to $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k-1)}$. That is forced, not guessed. Which conjugate directions? The cheapest answer: take the residual and remove only what $\mathbf{A}$ -orthogonality forbids.

Definition 24.7.3. The Conjugate Gradient Iteration

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{x}^{(0)} \in \mathbb{R}^n$. Set

$$\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}, \quad \mathbf{p}^{(0)} = \mathbf{r}^{(0)}.$$

For $k = 0, 1, 2, \dots$: if $\mathbf{r}^{(k)} = \mathbf{0}$, stop. Otherwise put

$$\alpha_k = \frac{(\mathbf{r}^{(k)})^\top \mathbf{r}^{(k)}}{(\mathbf{p}^{(k)})^\top \mathbf{A}\mathbf{p}^{(k)}}, \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)},$$

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}, \quad \beta_k = \frac{(\mathbf{r}^{(k+1)})^\top \mathbf{r}^{(k+1)}}{(\mathbf{r}^{(k)})^\top \mathbf{r}^{(k)}},$$

and $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{p}^{(k)}$. This is the **conjugate gradient** iteration, or **CG**.

A nonexample, by dropping one clause. Keep everything about the recursion and only weaken $\mathbf{A} \succ \mathbf{O}$ to “symmetric and invertible”. Take

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \mathbf{b} = (1, 1), \quad \mathbf{x}^{(0)} = \mathbf{0}.$$

Then $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = (1, 1)$ and $(\mathbf{p}^{(0)})^\top \mathbf{A} \mathbf{p}^{(0)} = 1 - 1 = 0$: the very first α_0 divides by zero, and the iteration does not start. What failed is exactly clause (P2) of Definition 13.1.1 in its strict form, which is what guarantees $\mathbf{p}^\top \mathbf{A} \mathbf{p} > 0$ for $\mathbf{p} \neq \mathbf{0}$. Symmetry alone is not enough, and nor is invertibility; both Proposition 24.7.1 (the bowl has a bottom) and Theorem 24.7.4 (the search directions are usable) need definiteness.

Three remarks on the shape of the recursion. The update of $\mathbf{r}$ is not a fresh evaluation of $\mathbf{b} - \mathbf{A} \mathbf{x}$ but a consequence of the update of $\mathbf{x}$; the two agree, since $\mathbf{b} - \mathbf{A} \mathbf{x}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{A} \mathbf{p}^{(k)}$. Only one product $\mathbf{A} \mathbf{p}^{(k)}$ is formed per step, serving both α_k and $\mathbf{r}^{(k+1)}$. And the new direction corrects the residual by a multiple of the **single** previous direction, not of all of them — the surprise, and the reason CG is short.

Theorem 24.7.4. Conjugate Gradients: Orthogonality and Optimality

Let $\mathbf{A} \succ \mathbf{O}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x}^{(0)} \in \mathbb{R}^n$, and write $\mathcal{K}_k = \mathcal{K}_k(\mathbf{A}, \mathbf{r}^{(0)})$ for the Krylov subspaces of Definition 24.6.1, with $\mathcal{K}_0 = \{\mathbf{0}\}$. Suppose the iteration of Definition 24.7.3 has produced nonzero $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}$. Then:

- (a) $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)}$ are nonzero and pairwise $\mathbf{A}$ -orthogonal, and every α_j and β_j with $j \leq k$ is defined, with $\alpha_j > 0$;
- (b) $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}$ are pairwise orthogonal, and $(\mathbf{r}^{(k+1)})^\top \mathbf{p}^{(j)} = 0$ for every $j \leq k$;
- (c) $\text{span}(\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}) = \text{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)}) = \mathcal{K}_{k+1}$, a space of dimension $k+1$;
- (d) $\mathbf{x}^{(k+1)}$ is the **unique** minimizer of $\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{A}}$ over the affine set $\mathbf{x}^{(0)} + \mathcal{K}_{k+1}$.

Idea. Everything is one induction, and the two coefficients are exactly what the induction needs. Step k has to deliver two orthogonality relations: the new residual against the old directions, and the new direction against the old directions in the $\mathbf{A}$ -inner product. The first is forced by α_k , the second by β_k , and each forces it only against the *latest* predecessor; the older ones come free, because $\mathbf{A} \mathbf{p}^{(j)}$ is a combination of $\mathbf{r}^{(j)}$ and $\mathbf{r}^{(j+1)}$, and the new residual is already orthogonal to both. Optimality (d) is then the Pythagorean theorem, applied in the energy inner product.

Proof. Parts (a), (b), and (c) are proved together by induction on k . Throughout, $\mathbf{A}^\top = \mathbf{A}$ is used silently to move $\mathbf{A}$ across a dot product, and $\mathbf{A} \mathcal{K}_j \subseteq \mathcal{K}_{j+1}$ directly from Definition 24.6.1.

Base $k = 0$. Here $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} \neq \mathbf{0}$, so $(\mathbf{p}^{(0)})^\top \mathbf{A} \mathbf{p}^{(0)} > 0$ and $\alpha_0 = \|\mathbf{r}^{(0)}\|_2^2 / (\mathbf{p}^{(0)})^\top \mathbf{A} \mathbf{p}^{(0)}$ is defined and positive; that is (a). Part (c) is $\text{span}(\mathbf{r}^{(0)}) = \text{span}(\mathbf{p}^{(0)}) = \mathcal{K}_1$, which is Definition 24.6.1, and the dimension is 1 because $\mathbf{r}^{(0)} \neq \mathbf{0}$. For (b), a single vector is vacuously an orthogonal list, and

$$(\mathbf{r}^{(1)})^\top \mathbf{p}^{(0)} = (\mathbf{r}^{(0)})^\top \mathbf{p}^{(0)} - \alpha_0 (\mathbf{p}^{(0)})^\top \mathbf{A} \mathbf{p}^{(0)} = \|\mathbf{r}^{(0)}\|_2^2 - \|\mathbf{r}^{(0)}\|_2^2 = 0,$$

using $\mathbf{r}^{(1)} = \mathbf{r}^{(0)} - \alpha_0 \mathbf{A} \mathbf{p}^{(0)}$, $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$, and the definition of α_0 .

Inductive step. Assume (a), (b), (c) for k , and assume $\mathbf{r}^{(k+1)} \neq \mathbf{0}$, so that β_k and $\mathbf{p}^{(k+1)}$ are formed. We prove (a), (b), (c) for $k+1$.

Step 1: the residuals stay orthogonal. By (b), $\mathbf{r}^{(k+1)}$ is orthogonal to $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)}$, and by (c) that list spans the same space as $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}$. Hence $\mathbf{r}^{(k+1)} \perp \mathbf{r}^{(j)}$ for every $j \leq k$, and with (b) this makes $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k+1)}$ pairwise orthogonal.

Step 2: (c) for $k+1$. From $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}$ with $\mathbf{r}^{(k)}, \mathbf{p}^{(k)} \in \mathcal{K}_{k+1}$, we get $\mathbf{r}^{(k+1)} \in \mathcal{K}_{k+2}$, and then $\mathbf{p}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{p}^{(k)} \in \mathcal{K}_{k+2}$ as well. Moreover

$$\text{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k+1)}) = \text{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)}, \mathbf{r}^{(k+1)}) = \text{span}(\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k+1)}),$$

the first equality because $\mathbf{p}^{(k+1)}$ differs from $\mathbf{r}^{(k+1)}$ by a multiple of $\mathbf{p}^{(k)}$, the second by (c). By Step 1 the vectors $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k+1)}$ are nonzero and pairwise orthogonal, hence independent (Theorem 11.1.4), so this common span has dimension $k+2$. It sits inside $\mathcal{K}_{k+2}$, which is spanned by $k+2$ vectors and so has dimension at most $k+2$; equal dimensions force equal spaces.

Step 3: $\mathbf{p}^{(k+1)} \neq \mathbf{0}$, and α_{k+1} is defined. By (b), $(\mathbf{r}^{(k+1)})^\top \mathbf{p}^{(k)} = 0$, so

$$(\mathbf{p}^{(k+1)})^\top \mathbf{r}^{(k+1)} = \|\mathbf{r}^{(k+1)}\|_2^2 + \beta_k (\mathbf{p}^{(k)})^\top \mathbf{r}^{(k+1)} = \|\mathbf{r}^{(k+1)}\|_2^2 > 0. \quad (\dagger)$$

Hence $\mathbf{p}^{(k+1)} \neq \mathbf{0}$, so $(\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(k+1)} > 0$ and α_{k+1} is defined and positive.

Step 4: (a) for $k+1$. We must show $(\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(j)} = 0$ for $j \leq k$. For $j = k$, use $\mathbf{A}\mathbf{p}^{(k)} = (\mathbf{r}^{(k)} - \mathbf{r}^{(k+1)})/\alpha_k$:

$$(\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(k)} = (\mathbf{r}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(k)} + \beta_k (\mathbf{p}^{(k)})^\top \mathbf{A}\mathbf{p}^{(k)}.$$

By Step 1 the first term is $(0 - \|\mathbf{r}^{(k+1)}\|_2^2)/\alpha_k$, and $(\mathbf{p}^{(k)})^\top \mathbf{A}\mathbf{p}^{(k)} = \|\mathbf{r}^{(k)}\|_2^2/\alpha_k$ by the definition of α_k . The two cancel exactly when $\beta_k = \|\mathbf{r}^{(k+1)}\|_2^2/\|\mathbf{r}^{(k)}\|_2^2$, which is how β_k was defined. For $j < k$, the term $\beta_k (\mathbf{p}^{(k)})^\top \mathbf{A}\mathbf{p}^{(j)}$ vanishes by (a), and

$$\mathbf{A}\mathbf{p}^{(j)} = \frac{\mathbf{r}^{(j)} - \mathbf{r}^{(j+1)}}{\alpha_j} \in \text{span}(\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}),$$

so $(\mathbf{r}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(j)} = 0$ by Step 1.

Step 5: (b) for $k+1$. The first half is Step 1. For the second, $\mathbf{r}^{(k+2)} = \mathbf{r}^{(k+1)} - \alpha_{k+1} \mathbf{A}\mathbf{p}^{(k+1)}$, so for $j = k+1$,

$$(\mathbf{r}^{(k+2)})^\top \mathbf{p}^{(k+1)} = (\mathbf{r}^{(k+1)})^\top \mathbf{p}^{(k+1)} - \alpha_{k+1} (\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(k+1)} = 0$$

by $(\dagger)$ and the definition of α_{k+1} ; and for $j \leq k$,

$$(\mathbf{r}^{(k+2)})^\top \mathbf{p}^{(j)} = (\mathbf{r}^{(k+1)})^\top \mathbf{p}^{(j)} - \alpha_{k+1} (\mathbf{p}^{(k+1)})^\top \mathbf{A}\mathbf{p}^{(j)} = 0 - 0 = 0$$

by (b) and Step 4. This completes the induction.

Proof of (d). By construction $\mathbf{x}^{(k+1)} = \mathbf{x}^{(0)} + \sum_{j \leq k} \alpha_j \mathbf{p}^{(j)}$, which lies in $\mathbf{x}^{(0)} + \mathcal{K}_{k+1}$ by (c). Let $\boldsymbol{\varepsilon} = \mathbf{x}^{(k+1)} - \mathbf{x}^*$. By Eq. 24.7.1, $\mathbf{A}\boldsymbol{\varepsilon} = -\mathbf{r}^{(k+1)}$, so for every $\mathbf{w} \in \mathcal{K}_{k+1} = \text{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k)})$,

$$\langle \boldsymbol{\varepsilon}, \mathbf{w} \rangle_{\mathbf{A}} = \mathbf{w}^\top \mathbf{A}\boldsymbol{\varepsilon} = -\mathbf{w}^\top \mathbf{r}^{(k+1)} = 0$$

by (b). Now take any $\mathbf{x} = \mathbf{x}^{(0)} + \mathbf{z}$ with $\mathbf{z} \in \mathcal{K}_{k+1}$, and write $\mathbf{x} - \mathbf{x}^* = \boldsymbol{\varepsilon} + \mathbf{w}$ with $\mathbf{w} = \mathbf{z} - \sum_{j \leq k} \alpha_j \mathbf{p}^{(j)} \in \mathcal{K}_{k+1}$. By the Pythagorean theorem (Theorem 11.1.3) in the energy inner product,

$$\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{A}}^2 = \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2 + \|\mathbf{w}\|_{\mathbf{A}}^2 \geq \|\boldsymbol{\varepsilon}\|_{\mathbf{A}}^2,$$

with equality only for $\mathbf{w} = \mathbf{0}$, that is only for $\mathbf{x} = \mathbf{x}^{(k+1)}$. This proves the theorem. ■

Corollary 24.7.5. Finite Termination

In exact arithmetic the conjugate gradient iteration reaches $\mathbf{r}^{(m)} = \mathbf{0}$, and hence $\mathbf{x}^{(m)} = \mathbf{x}^*$, for some $m \leq n$. If $\mathbf{A}$ has only d distinct eigenvalues, then $m \leq d$.

Proof. Suppose $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k)}$ are all nonzero. By Theorem 24.7.4 (c), $\dim \mathcal{K}_{k+1} = k + 1$. Since $\mathcal{K}_{k+1} \subseteq \mathbb{R}^n$, this forces $k + 1 \leq n$; so the residuals cannot all be nonzero for $k = n$, and $\mathbf{r}^{(m)} = \mathbf{0}$ for some $m \leq n$. Then $\mathbf{A}\mathbf{x}^{(m)} = \mathbf{b}$, so $\mathbf{x}^{(m)} = \mathbf{x}^*$.

For the refinement, let $\mu_1, \dots, \mu_d$ be the distinct eigenvalues and put $q(x) = (x - \mu_1) \cdots (x - \mu_d)$. By Corollary 12.5.2, $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ with $\mathbf{Q}$ orthogonal and $\mathbf{D}$ diagonal carrying the eigenvalues, so $q(\mathbf{A}) = \mathbf{Q}q(\mathbf{D})\mathbf{Q}^\top = \mathbf{O}$, every diagonal entry of $q(\mathbf{D})$ being q at an eigenvalue. Hence $\mathbf{A}^d \mathbf{r}^{(0)}$ is a combination of $\mathbf{r}^{(0)}, \mathbf{A}\mathbf{r}^{(0)}, \dots, \mathbf{A}^{d-1}\mathbf{r}^{(0)}$, so $\mathcal{K}_{d+1} = \mathcal{K}_d$ and $\dim \mathcal{K}_k \leq d$ for every k . If $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(d)}$ were all nonzero we would get $\dim \mathcal{K}_{d+1} = d + 1 > d$, a contradiction. So $m \leq d$. ■

Example. Two distinct eigenvalues, two steps

Let $\mathbf{A} = \mathbf{I}_3 + \mathbf{J}$, where $\mathbf{J}$ is the all-ones matrix, let $\mathbf{b} = \mathbf{e}_1$ and $\mathbf{x}^{(0)} = \mathbf{0}$. Run CG.

Solution. $\mathbf{A}$ has rows $(2, 1, 1), (1, 2, 1), (1, 1, 2)$. Since $\mathbf{J}\mathbf{1} = 3\mathbf{1}$ and $\mathbf{J}$ kills $\mathbf{1}^\perp$, the eigenvalues of $\mathbf{A}$ are 4 (once) and 1 (twice): $d = 2$. Also $\mathbf{A}^{-1} = \mathbf{I} - \frac{1}{4}\mathbf{J}$, so $\mathbf{x}^* = \mathbf{e}_1 - \frac{1}{4}\mathbf{1} = (\frac{3}{4}, -\frac{1}{4}, -\frac{1}{4})$.

Step 0. $\mathbf{r}^{(0)} = \mathbf{p}^{(0)} = (1, 0, 0)$, $(\mathbf{r}^{(0)})^\top \mathbf{r}^{(0)} = 1$, $\mathbf{A}\mathbf{p}^{(0)} = (2, 1, 1)$, $(\mathbf{p}^{(0)})^\top \mathbf{A}\mathbf{p}^{(0)} = 2$, so $\alpha_0 = \frac{1}{2}$ and

$$\mathbf{x}^{(1)} = (\tfrac{1}{2}, 0, 0), \quad \mathbf{r}^{(1)} = (1, 0, 0) - \tfrac{1}{2}(2, 1, 1) = (0, -\tfrac{1}{2}, -\tfrac{1}{2}).$$

Then $\beta_0 = \frac{1}{2}/1 = \frac{1}{2}$ and $\mathbf{p}^{(1)} = (0, -\frac{1}{2}, -\frac{1}{2}) + \frac{1}{2}(1, 0, 0) = (\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2})$.

Step 1. The rows of $\mathbf{A}$ applied to $\mathbf{p}^{(1)}$ give $1 - \frac{1}{2} - \frac{1}{2} = 0$, then $\frac{1}{2} - 1 - \frac{1}{2} = -1$, then $\frac{1}{2} - \frac{1}{2} - 1 = -1$; so $\mathbf{A}\mathbf{p}^{(1)} = (0, -1, -1)$ and $(\mathbf{p}^{(1)})^\top \mathbf{A}\mathbf{p}^{(1)} = 1$. With $(\mathbf{r}^{(1)})^\top \mathbf{r}^{(1)} = \frac{1}{2}$ this gives $\alpha_1 = \frac{1}{2}$ and

$$\mathbf{x}^{(2)} = (\tfrac{1}{2}, 0, 0) + \tfrac{1}{2}(\tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}) = (\tfrac{3}{4}, -\tfrac{1}{4}, -\tfrac{1}{4}) = \mathbf{x}^*,$$

with $\mathbf{r}^{(2)} = (0, -\frac{1}{2}, -\frac{1}{2}) - \frac{1}{2}(0, -1, -1) = \mathbf{0}$. Two steps, as Corollary 24.7.5 predicts, for a 3×3 system. Checks: $(\mathbf{r}^{(1)})^\top \mathbf{r}^{(0)} = 0$ and $(\mathbf{p}^{(1)})^\top \mathbf{A}\mathbf{p}^{(0)} = (\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}) \cdot (2, 1, 1) = 1 - \frac{1}{2} - \frac{1}{2} = 0$.

Warning. **Finite termination is a theorem about exact arithmetic, and it is not how CG is used.** In the floating-point model of Section 1 the computed residuals lose their orthogonality, as the Lanczos vectors of Section 6 do. The two algorithms are closely related: started from the same $\mathbf{A}$ and the same vector they build bases of the same Krylov spaces, and the tridiagonal matrix of Theorem 24.6.3 can be assembled from the CG coefficients — which are *not* the α_j and β_j of Theorem 24.6.3, those being the entries of that tridiagonal matrix, but different numbers. **That correspondence, due to Lanczos and to Hestenes and Stiefel, and the rounding behavior it carries across, due to Paige and Greenbaum, are described here and not proved; nothing in this section depends on either.** A run of CG on a matrix of order 10^6 is stopped after a few hundred steps, and the theorem that matters is [Theorem 24.7.7](#) below.

24.7.4 Chebyshev polynomials

[Theorem 24.7.4](#) (d) has an equivalent form that turns the whole question of speed into a question about polynomials, and that question has a classical answer. We need one family of polynomials and two facts about it.

The Chebyshev polynomials stay within 1 on $[-1, 1]$ and then grow geometrically outside it.

Definition 24.7.4. Chebyshev Polynomials

The **Chebyshev polynomials** $T_0, T_1, T_2, \dots \in \mathbb{R}[x]$ are defined by

$$T_0 = 1, \quad T_1 = x, \quad T_{k+1} = 2xT_k - T_{k-1} \quad (k \geq 1).$$

The name comes from Chapter 11. Its table at the end of §10 lists the **monic** orthogonal polynomial sequence of Definition 11.10.1 for the weight $(1-t^2)^{-1/2}$ on $(-1, 1)$, and the T_k above are its members rescaled — which is why the recurrence here is not the one of Theorem 11.10.3, that theorem being stated for the monic normalization, where no factor 2 appears. The book has not proved that identification and nothing below uses it. Here the recurrence *is* the definition, and we prove what we need from it directly. The first few are

$$T_2 = 2x^2 - 1, \quad T_3 = 4x^3 - 3x, \quad T_4 = 8x^4 - 8x^2 + 1.$$

We shall need the degrees, and the recurrence supplies them by induction on k . For $k = 0$ and $k = 1$, $\deg T_k = k$, with leading coefficient 1 in both cases. Let $k \geq 1$ and suppose $\deg T_{k-1} = k-1$ and $\deg T_k = k$ with leading coefficient c_k . Then $2xT_k$ has degree $k+1$ with leading coefficient $2c_k$, while $\deg T_{k-1} = k-1 < k+1$, so no cancellation is possible in $T_{k+1} = 2xT_k - T_{k-1}$ and $\deg T_{k+1} = k+1$ with leading coefficient $2c_k$. Hence $\deg T_k = k$ for every k , and since $c_1 = 1$ the doubling gives $c_k = 2^{k-1}$ for $k \geq 1$.

The one trick in the subject is a change of variable. Every complex number x can be written as $\frac{1}{2}(w + w^{-1})$ for some $w \neq 0$: the equation $w^2 - 2xw + 1 = 0$ has two roots in $\mathbb{C}$, neither of them 0 since their product is 1. In that variable the recurrence becomes trivial.

Lemma 24.7.6. Chebyshev Polynomials: Size Inside and Growth Outside

For every $k \geq 0$ and every $w \in \mathbb{C} \setminus \{0\}$,

$$T_k\left(\frac{w + w^{-1}}{2}\right) = \frac{w^k + w^{-k}}{2}. \quad (24.7.5)$$

Consequently:

- (a) $|T_k(x)| \leq 1$ for every real $x \in [-1, 1]$, and $T_k(1) = 1$;
- (b) for every real $z > 1$,

$$T_k(z) = \frac{1}{2} \left((z + \sqrt{z^2 - 1})^k + (z - \sqrt{z^2 - 1})^k \right) \geq \frac{1}{2} (z + \sqrt{z^2 - 1})^k,$$

and, separately, $T_k(z) \geq 1$.

Idea. Under $x = \frac{1}{2}(w + w^{-1})$ the recurrence of [Definition 24.7.4](#) says nothing more than $w^{k+1} + w^{-(k+1)} = (w + w^{-1})(w^k + w^{-k}) - (w^{k-1} + w^{-(k-1)})$, which is true because the cross terms cancel. The two facts are then two choices of w : on $[-1, 1]$ take $|w| = 1$, so that w^k and w^{-k} are again on the unit circle; outside, take w real and greater than 1, so that w^k grows geometrically.

Proof. The identity [Eq. 24.7.5](#). Induct on k . For $k = 0$ both sides are 1, and for $k = 1$ both are $\frac{1}{2}(w + w^{-1})$. Assume the identity for $k - 1$ and k , with $k \geq 1$, and write $x = \frac{1}{2}(w + w^{-1})$. Then

$$\begin{aligned} T_{k+1}(x) &= 2x T_k(x) - T_{k-1}(x) \\ &= (w + w^{-1}) \frac{w^k + w^{-k}}{2} - \frac{w^{k-1} + w^{-(k-1)}}{2} \\ &= \frac{w^{k+1} + w^{-(k+1)} + w^{k-1} + w^{-(k+1)} - w^{k-1} - w^{-(k-1)}}{2}, \end{aligned}$$

which is $\frac{1}{2}(w^{k+1} + w^{-(k+1)})$, as required.

(a). Let $x \in [-1, 1]$ and put $w = x + i\sqrt{1 - x^2}$. Then $|w|^2 = x^2 + (1 - x^2) = 1$, so $w \neq 0$ and $w^{-1} = \bar{w}$ ([Theorem 0.6.2](#)), whence $\frac{1}{2}(w + w^{-1}) = \frac{1}{2}(w + \bar{w}) = x$. By [Eq. 24.7.5](#) and $|u + v| \leq |u| + |v|$ ([Theorem 0.6.4](#)),

$$|T_k(x)| = \frac{1}{2} |w^k + w^{-k}| \leq \frac{1}{2} (|w|^k + |w|^{-k}) = 1.$$

For $x = 1$ take $w = 1$, giving $T_k(1) = \frac{1}{2}(1 + 1) = 1$.

(b). Let $z > 1$ and put $w = z + \sqrt{z^2 - 1}$, a real number > 1 . Then

$$w(z - \sqrt{z^2 - 1}) = z^2 - (z^2 - 1) = 1,$$

so $w^{-1} = z - \sqrt{z^2 - 1}$ and $\frac{1}{2}(w + w^{-1}) = z$. Now [Eq. 24.7.5](#) gives the displayed formula. Both w^k and w^{-k} are positive, so dropping the second gives the displayed inequality. For the separate bound, $\frac{1}{2}(s + s^{-1}) \geq 1$ for every $s > 0$, because $s + s^{-1} - 2 = (s^{1/2} - s^{-1/2})^2 \geq 0$; applying this with $s = w^k > 0$ gives $T_k(z) = \frac{1}{2}(w^k + w^{-k}) \geq 1$. This proves the lemma. ■

Warning. **Part (a) is false the moment x leaves $[-1, 1]$, and spectacularly so.** $T_{10}(1.1)$ is already about 42.2, and $T_{50}(1.1)$ exceeds 2×10^9 . That explosion is not a defect; it is the whole resource. The polynomial in the next theorem is small where the eigenvalues live and large at 0, and the ratio of the two is what CG converts into speed.

Remark. That the Chebyshev polynomials are also **extremal** — among monic polynomials of degree k , the scaled $T_k/2^{k-1}$ has the smallest maximum modulus on $[-1, 1]$ — is classical, is not proved here, and is not needed. What [Theorem 24.7.7](#) uses is only that T_k , rescaled, is an admissible competitor in the minimum of [Eq. 24.7.6](#): small on the interval where the eigenvalues lie, and large at 0.

24.7.5 The square root

Theorem 24.7.7. Convergence of Conjugate Gradients

Let $\mathbf{A} \in M_n(\mathbb{R})$ be symmetric positive definite with $\kappa = \kappa_2(\mathbf{A})$, let $\mathbf{b} \in \mathbb{R}^n$, and let $\mathbf{x}^{(k)}$ be the conjugate gradient iterates from $\mathbf{x}^{(0)}$, with $\mathbf{x}^{(k)} := \mathbf{x}^*$ once the iteration has terminated. Then for every $k \geq 0$,

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\|_{\mathbf{A}} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_{\mathbf{A}}.$$

Idea. Three moves. ① [Theorem 24.7.4](#) says the CG error is the smallest among all errors reachable from the Krylov space; writing out what “reachable” means turns that into: the smallest value of $\|q(\mathbf{A})\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}}$ over polynomials q of degree $\leq k$ with $q(0) = 1$. ② In an eigenbasis that quantity is at most $\max_i |q(\lambda_i)|$ times the initial error, so we need one polynomial that is 1 at 0 and small on $[\lambda_n, \lambda_1]$. ③ Chebyshev’s polynomial, rescaled so that the interval becomes $[-1, 1]$, is that polynomial, and [Lemma 24.7.6](#) measures it.

Proof. If $\kappa = 1$ then $\lambda_1 = \lambda_n$, so $\mathbf{A} = \lambda_1 \mathbf{I}$ by the spectral theorem, and $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$ gives $\alpha_0 = 1/\lambda_1$ and $\mathbf{x}^{(1)} = \mathbf{x}^*$; the bound then reads $0 \leq 0$ for $k \geq 1$ and $\|\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}} \leq 2\|\boldsymbol{\varepsilon}^{(0)}\|_{\mathbf{A}}$ for $k = 0$, both true. So assume $\kappa > 1$. We may also assume the iteration has not terminated before step k , since once it has, the left-hand side is 0.

Step 1: the error is a polynomial in $\mathbf{A}$ applied to the initial error. Write $\boldsymbol{\varepsilon}^{(0)} = \mathbf{x}^{(0)} - \mathbf{x}^*$, so $\mathbf{r}^{(0)} = -\mathbf{A}\boldsymbol{\varepsilon}^{(0)}$ by [Eq. 24.7.1](#). A vector of $\mathbf{x}^{(0)} + \mathcal{K}_k$ has the form $\mathbf{x} = \mathbf{x}^{(0)} + \sum_{j=0}^{k-1} c_j \mathbf{A}^j \mathbf{r}^{(0)}$, and then

$$\mathbf{x} - \mathbf{x}^* = \boldsymbol{\varepsilon}^{(0)} - \sum_{j=0}^{k-1} c_j \mathbf{A}^{j+1} \boldsymbol{\varepsilon}^{(0)} = q(\mathbf{A}) \boldsymbol{\varepsilon}^{(0)}, \quad q(x) = 1 - \sum_{j=0}^{k-1} c_j x^{j+1}.$$

As the c_j range over $\mathbb{R}^k$, q ranges over **all** polynomials of degree at most k with $q(0) = 1$. So [Theorem 24.7.4](#) (d) says

$$\|\varepsilon^{(k)}\|_{\mathbf{A}} = \min\{\|q(\mathbf{A})\varepsilon^{(0)}\|_{\mathbf{A}} : \deg q \leq k, q(0) = 1\}. \quad (24.7.6)$$

Step 2: one eigenbasis estimate. By Corollary 12.5.2 take an orthonormal basis $\mathbf{q}_1, \dots, \mathbf{q}_n$ with $\mathbf{A}\mathbf{q}_i = \lambda_i\mathbf{q}_i$, and write $\varepsilon^{(0)} = \sum_i u_i\mathbf{q}_i$. Then $q(\mathbf{A})\mathbf{q}_i = q(\lambda_i)\mathbf{q}_i$, so

$$\|q(\mathbf{A})\varepsilon^{(0)}\|_{\mathbf{A}}^2 = \sum_i \lambda_i q(\lambda_i)^2 u_i^2 \leq \left(\max_i |q(\lambda_i)|\right)^2 \sum_i \lambda_i u_i^2,$$

and $\sum_i \lambda_i u_i^2 = \|\varepsilon^{(0)}\|_{\mathbf{A}}^2$. Every λ_i lies in $[\lambda_n, \lambda_1]$, so with Eq. 24.7.6, for any admissible q ,

$$\|\varepsilon^{(k)}\|_{\mathbf{A}} \leq \left(\max_{\lambda_n \leq \lambda \leq \lambda_1} |q(\lambda)|\right) \|\varepsilon^{(0)}\|_{\mathbf{A}}. \quad (24.7.7)$$

Step 3: the polynomial. Put

$$z = \frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n} = \frac{\kappa + 1}{\kappa - 1} > 1, \quad q(\lambda) = \frac{1}{T_k(z)} T_k\left(\frac{\lambda_1 + \lambda_n - 2\lambda}{\lambda_1 - \lambda_n}\right).$$

The inner expression is an affine function of λ carrying λ_n to 1 and λ_1 to -1 , so it maps $[\lambda_n, \lambda_1]$ into $[-1, 1]$; and it sends $\lambda = 0$ to z . By Lemma 24.7.6 (b), $T_k(z) \geq 1 > 0$, so q is a well-defined polynomial of degree at most k , and $q(0) = T_k(z)/T_k(z) = 1$. By Lemma 24.7.6 (a),

$$\max_{\lambda_n \leq \lambda \leq \lambda_1} |q(\lambda)| \leq \frac{1}{T_k(z)}.$$

Step 4: measuring $T_k(z)$. Since $\kappa > 1$,

$$z^2 - 1 = \frac{(\kappa + 1)^2 - (\kappa - 1)^2}{(\kappa - 1)^2} = \frac{4\kappa}{(\kappa - 1)^2}, \quad \sqrt{z^2 - 1} = \frac{2\sqrt{\kappa}}{\kappa - 1},$$

so, using $\kappa - 1 = (\sqrt{\kappa} - 1)(\sqrt{\kappa} + 1)$,

$$z + \sqrt{z^2 - 1} = \frac{\kappa + 1 + 2\sqrt{\kappa}}{\kappa - 1} = \frac{(\sqrt{\kappa} + 1)^2}{(\sqrt{\kappa} - 1)(\sqrt{\kappa} + 1)} = \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}.$$

By Lemma 24.7.6 (b), $T_k(z) \geq \frac{1}{2}((\sqrt{\kappa} + 1)/(\sqrt{\kappa} - 1))^k$, so $1/T_k(z) \leq 2((\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1))^k$. Feeding this into Eq. 24.7.7 proves the theorem. ■

The square root is the whole gain. Take $\kappa = 100$. Steepest descent contracts by at most $99/101 \approx 0.9802$ per step, so reducing the energy-norm error by 10^{-6} needs about 691 steps; conjugate gradients contracts by $9/11 \approx 0.8182$, and $2(9/11)^k \leq 10^{-6}$ already at $k = 73$. Both cost one product with $\mathbf{A}$ per step, so the ratio of the counts is the ratio of the work.

Quick check. Why does the factor 2 in Theorem 24.7.7 cost so little, and why is it there at all?

Solution. It costs little because it is a constant against a geometric factor: absorbing it changes the step count by $\log 2 / \log((\sqrt{\kappa} + 1)/(\sqrt{\kappa} - 1))$, about 3.5 steps out of 73 when $\kappa = 100$. It is there because the lower bound for $T_k(z)$ in [Lemma 24.7.6](#) (b) throws away the term w^{-k} ; the surviving $\frac{1}{2}$ is exactly this 2.

Remark. The bound uses only the two ends of the spectrum, and it is often pessimistic. [Eq. 24.7.7](#) holds for *every* admissible polynomial, so if the eigenvalues fall into a few tight clusters one may take q small near each cluster and do far better; [Corollary 24.7.5](#) is the extreme case. Conjugate gradients sees the whole spectrum, not just its ends.

24.7.6 Two footnotes: Chebyshev iteration and preconditioning

The polynomial in Step 3 depends only on λ_1 and λ_n , so one can *fix* it in advance and update the iterates by the recurrence of [Definition 24.7.4](#). This is the **Chebyshev iteration**. Three things are said about it: that it attains the bound of [Theorem 24.7.7](#) rather than beating it; that it needs estimates of λ_1 and λ_n which CG does not; and that in exchange it forms no inner products, which matters when the work is split across many processors. **None of the three is proved here.** The method and its analysis are due to Golub and Varga, and nothing in this section depends on them.

Since the cost is governed by $\sqrt{\kappa}$, it pays to replace $\mathbf{Ax} = \mathbf{b}$ by an equivalent system of smaller condition number. Choosing a symmetric positive definite $\mathbf{M}$ close to $\mathbf{A}$, and writing $\mathbf{M} = \mathbf{CC}^\top$, one applies CG to

$$(\mathbf{C}^{-1}\mathbf{AC}^{-\top})\mathbf{y} = \mathbf{C}^{-1}\mathbf{b}, \quad \mathbf{x} = \mathbf{C}^{-\top}\mathbf{y},$$

whose matrix is symmetric positive definite ([Proposition 13.1.2](#)) and whose eigenvalues are the generalized eigenvalues of the pencil $\mathbf{A} - \lambda\mathbf{M}$ of [Definition 13.4.1](#), as [Exercise C3](#) below asks you to check. This is **preconditioning**, and it is a subject rather than a theorem: which $\mathbf{M}$ to choose depends on where $\mathbf{A}$ came from, and no theorem of linear algebra answers that. We state the mechanism and stop.

24.7.7 Exercises

24.7.7.1 A. Check your understanding

Exercise. A1

- Define the energy inner product of a symmetric positive definite $\mathbf{A}$, and say which clause of [Definition 13.1.1](#) makes it an inner product.
- State [Proposition 24.7.1](#), and explain in one sentence why it makes $\|\cdot\|_{\mathbf{A}}$ the natural yardstick for the error.
- State the optimality property of the k -th conjugate gradient iterate, naming the set over which it is optimal.
- Determine whether the following statement is correct, and justify your answer: in exact arithmetic, conjugate gradients applied to a 1000×1000 positive definite matrix with 5 distinct eigenvalues terminates after at most 5 steps.

- (e) Write down the two Chebyshev facts of [Lemma 24.7.6](#), and say which of the two supplies the $\sqrt{\kappa}$ in [Theorem 24.7.7](#).

Solution.

- (a) $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{y}^{\top} \mathbf{A} \mathbf{x}$. Symmetry and bilinearity hold for any symmetric $\mathbf{A}$; what makes it an inner product is $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = \mathbf{x}^{\top} \mathbf{A} \mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$, which is clause (P2) of Definition 13.1.1 in its strict form.
- (b) $\phi(\mathbf{x}) - \phi(\mathbf{x}^{\star}) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}^{\star}\|_{\mathbf{A}}^2$, where $\phi(\mathbf{x}) = \frac{1}{2} \mathbf{x}^{\top} \mathbf{A} \mathbf{x} - \mathbf{b}^{\top} \mathbf{x}$ and $\mathbf{x}^{\star} = \mathbf{A}^{-1} \mathbf{b}$. So the energy norm of the error is the excess of ϕ over its minimum, up to the factor $\frac{1}{2}$: it is the quantity every descent method is built to decrease.
- (c) By [Theorem 24.7.4](#) (d), $\mathbf{x}^{(k)}$ is the unique minimizer of $\|\mathbf{x} - \mathbf{x}^{\star}\|_{\mathbf{A}}$ over the affine set $\mathbf{x}^{(0)} + \mathcal{K}_k(\mathbf{A}, \mathbf{r}^{(0)})$.
- (d) Correct. By [Corollary 24.7.5](#) the number of steps is at most the number d of distinct eigenvalues, here 5, and the order 1000 plays no part. The proof is that $q(\mathbf{A}) = \mathbf{0}$ for the degree-5 polynomial with the distinct eigenvalues as roots, so the Krylov spaces stop growing at dimension 5.
- (e) $|T_k| \leq 1$ on $[-1, 1]$, and $T_k(z) \geq \frac{1}{2}(z + \sqrt{z^2 - 1})^k$ for $z > 1$. The second supplies the square root: at $z = (\kappa + 1)/(\kappa - 1)$ the number $z + \sqrt{z^2 - 1}$ equals $(\sqrt{\kappa} + 1)/(\sqrt{\kappa} - 1)$, and the square root enters through $\sqrt{z^2 - 1} = 2\sqrt{\kappa}/(\kappa - 1)$.

24.7.7.2 B. Practice

Exercise. B1: Steepest descent by hand

Let $\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$, $\mathbf{b} = (2, 1)$, and $\mathbf{x}^{(0)} = (0, 0)$.

- (a) Compute $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ by steepest descent.
- (b) Compute $\|\mathbf{x}^{(k)} - \mathbf{x}^{\star}\|_{\mathbf{A}}^2$ for $k = 0, 1, 2$ and compare the ratios with the bound of [Theorem 24.7.3](#).

Solution. Here $\mathbf{x}^{\star} = (1, 1)$ and $\kappa = 2$, so the bound's factor is $(\kappa - 1)/(\kappa + 1) = \frac{1}{3}$.

- (a) $\mathbf{r}^{(0)} = (2, 1)$, $\mathbf{r}^{\top} \mathbf{r} = 5$, $\mathbf{r}^{\top} \mathbf{A} \mathbf{r} = 2 \cdot 4 + 1 = 9$, so $t_0 = \frac{5}{9}$ and $\mathbf{x}^{(1)} = (\frac{10}{9}, \frac{5}{9})$. Then $\mathbf{r}^{(1)} = (2, 1) - \mathbf{A} \mathbf{x}^{(1)} = (2 - \frac{20}{9}, 1 - \frac{5}{9}) = (-\frac{2}{9}, \frac{4}{9})$. Now $\mathbf{r}^{\top} \mathbf{r} = \frac{20}{81}$ and $\mathbf{r}^{\top} \mathbf{A} \mathbf{r} = 2 \cdot \frac{4}{81} + \frac{16}{81} = \frac{24}{81}$, so $t_1 = \frac{20}{24} = \frac{5}{6}$ and

$$\mathbf{x}^{(2)} = \left(\frac{10}{9} - \frac{5}{6} \cdot \frac{2}{9}, \frac{5}{9} + \frac{5}{6} \cdot \frac{4}{9} \right) = \left(\frac{25}{27}, \frac{25}{27} \right).$$

- (b) The errors are $(-1, -1)$, $(\frac{1}{9}, -\frac{4}{9})$, $(-\frac{2}{27}, -\frac{2}{27})$, with squared energy norms

$$3, \quad 2 \cdot \frac{1}{81} + \frac{16}{81} = \frac{18}{81} = \frac{2}{9}, \quad 3 \cdot \frac{4}{729} = \frac{4}{243}.$$

The ratios are $\frac{2}{27}$ and $\frac{2}{27}$, both well below the bound's $(\frac{1}{3})^2 = \frac{1}{9}$. The bound is not attained here because the initial residual does not split its weight equally between the two eigenvector directions, which is what equality in [Lemma 24.7.2](#) demands.

Exercise. B2: Conjugate gradients by hand

Let $\mathbf{A} = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$, $\mathbf{b} = (1, 2)$, and $\mathbf{x}^{(0)} = (0, 0)$. Run conjugate gradients to termination, and verify at each stage that the residuals are orthogonal and the directions $\mathbf{A}$ -orthogonal.

Solution. $\det \mathbf{A} = 11$, so $\mathbf{x}^* = \frac{1}{11}(3 \cdot 1 - 1 \cdot 2, -1 \cdot 1 + 4 \cdot 2) = (\frac{1}{11}, \frac{7}{11})$.

Step 0. $\mathbf{r}^{(0)} = \mathbf{p}^{(0)} = (1, 2)$; $(\mathbf{r}^{(0)})^\top \mathbf{r}^{(0)} = 5$; $\mathbf{A}\mathbf{p}^{(0)} = (6, 7)$; $(\mathbf{p}^{(0)})^\top \mathbf{A}\mathbf{p}^{(0)} = 6 + 14 = 20$. So $\alpha_0 = \frac{1}{4}$, $\mathbf{x}^{(1)} = (\frac{1}{4}, \frac{1}{2})$ and

$$\mathbf{r}^{(1)} = (1, 2) - \frac{1}{4}(6, 7) = (-\frac{1}{2}, \frac{1}{4}).$$

Check: $(\mathbf{r}^{(1)})^\top \mathbf{r}^{(0)} = -\frac{1}{2} + \frac{1}{2} = 0$. Then $(\mathbf{r}^{(1)})^\top \mathbf{r}^{(1)} = \frac{1}{4} + \frac{1}{16} = \frac{5}{16}$, $\beta_0 = \frac{5/16}{5} = \frac{1}{16}$ and

$$\mathbf{p}^{(1)} = (-\frac{1}{2}, \frac{1}{4}) + \frac{1}{16}(1, 2) = (-\frac{7}{16}, \frac{3}{8}).$$

Check: $(\mathbf{p}^{(1)})^\top \mathbf{A}\mathbf{p}^{(0)} = (-\frac{7}{16})6 + \frac{3}{8} \cdot 7 = -\frac{42}{16} + \frac{42}{16} = 0$.

Step 1. $\mathbf{A}\mathbf{p}^{(1)} = (4(-\frac{7}{16}) + \frac{3}{8}, -\frac{7}{16} + 3 \cdot \frac{3}{8}) = (-\frac{11}{8}, \frac{11}{16})$, so

$$(\mathbf{p}^{(1)})^\top \mathbf{A}\mathbf{p}^{(1)} = \frac{77}{128} + \frac{33}{128} = \frac{55}{64}, \quad \alpha_1 = \frac{5/16}{55/64} = \frac{4}{11}.$$

Then

$$\mathbf{x}^{(2)} = (\frac{1}{4}, \frac{1}{2}) + \frac{4}{11}(-\frac{7}{16}, \frac{3}{8}) = (\frac{11}{44} - \frac{7}{44}, \frac{11}{22} + \frac{3}{22}) = (\frac{1}{11}, \frac{7}{11}) = \mathbf{x}^*,$$

and

$$\mathbf{r}^{(2)} = (-\frac{1}{2}, \frac{1}{4}) - \frac{4}{11}(-\frac{11}{8}, \frac{11}{16}) = (-\frac{1}{2} + \frac{1}{2}, \frac{1}{4} - \frac{1}{4}) = \mathbf{0}.$$

Two steps for a 2×2 system, which is what [Corollary 24.7.5](#) allows.

Exercise. B3: Chebyshev by recurrence

- (a) Compute T_5 from [Definition 24.7.4](#).
- (b) Evaluate $T_5(1)$, $T_5(-1)$, and $T_5(0)$, and check them against [Lemma 24.7.6](#) (a).
- (c) Evaluate $T_5(2)$ from the polynomial, and check it against the closed formula of [Lemma 24.7.6](#) (b).

Solution.

- (a) $T_2 = 2x^2 - 1$, $T_3 = 2x(2x^2 - 1) - x = 4x^3 - 3x$, $T_4 = 2x(4x^3 - 3x) - (2x^2 - 1) = 8x^4 - 8x^2 + 1$, and

$$T_5 = 2x(8x^4 - 8x^2 + 1) - (4x^3 - 3x) = 16x^5 - 20x^3 + 5x.$$

- (b) $T_5(1) = 16 - 20 + 5 = 1$, $T_5(-1) = -1$, $T_5(0) = 0$. All three have absolute value at most 1, as (a) of the lemma requires, and the first is the equality $T_k(1) = 1$.
- (c) $T_5(2) = 16 \cdot 32 - 20 \cdot 8 + 10 = 512 - 160 + 10 = 362$. By the closed formula with $z = 2$,

$w = 2 + \sqrt{3}$ and $w^{-1} = 2 - \sqrt{3}$, so

$$T_5(2) = \frac{1}{2}((2 + \sqrt{3})^5 + (2 - \sqrt{3})^5).$$

Expanding, $(2 \pm \sqrt{3})^5 = 362 \pm 209\sqrt{3}$, so the half-sum is 362. The two agree, and the growth is visible: T_5 is at most 1 on $[-1, 1]$ and already 362 at 2.

24.7.7.3 C. Going deeper

Exercise. C1: Sharpness of Kantorovich

Let $a \geq c > 0$ and let $\mathbf{A} = \text{diag}(a, c)$.

- Prove that for $\mathbf{y} = (1, 1)$ the inequality of Lemma 24.7.2 is an equality.
- Hence deduce that the constant $(\kappa - 1)/(\kappa + 1)$ in Theorem 24.7.3 cannot be replaced by any smaller one.

Solution.

- $\mathbf{y}^\top \mathbf{A} \mathbf{y} = a + c$, $\mathbf{y}^\top \mathbf{A}^{-1} \mathbf{y} = a^{-1} + c^{-1} = (a + c)/(ac)$, and $\mathbf{y}^\top \mathbf{y} = 2$. So the left side is $(a + c)^2/(ac)$ and the right side is $((a + c)^2/(4ac)) \cdot 4 = (a + c)^2/(ac)$. They agree.
- Take $\mathbf{b} = (a, c)$, so that $\mathbf{x}^* = (1, 1)$, and $\mathbf{x}^{(0)} = (1 - \frac{1}{a}, 1 - \frac{1}{c})$, so that $\boldsymbol{\varepsilon}^{(0)} = (-\frac{1}{a}, -\frac{1}{c})$. Then $\mathbf{r}^{(0)} = -\mathbf{A}\boldsymbol{\varepsilon}^{(0)} = (1, 1)$, which is the vector of part (a). By the proof of Theorem 24.7.3 the one-step ratio is

$$1 - \frac{(\mathbf{r}^\top \mathbf{r})^2}{(\mathbf{r}^\top \mathbf{A} \mathbf{r})(\mathbf{r}^\top \mathbf{A}^{-1} \mathbf{r})} = 1 - \frac{4ac}{(a + c)^2} = \left(\frac{\kappa - 1}{\kappa + 1} \right)^2,$$

so the first step attains the bound exactly. Moreover $\mathbf{r}^{(1)} = \mathbf{r}^{(0)} - t_0 \mathbf{A} \mathbf{r}^{(0)} = (1 - t_0 a, 1 - t_0 c)$ with $t_0 = 2/(a + c)$, giving $\mathbf{r}^{(1)} = ((c - a)/(a + c))(1, -1)$; its two components again have equal absolute value, so part (a) applies to it up to sign and scale. That is enough: replacing $\mathbf{y}$ by $s\mathbf{y}$ multiplies both sides of Lemma 24.7.2 by s^4 , and changing the sign of a coordinate changes neither side, since $\mathbf{A}$ and $\mathbf{A}^{-1}$ are diagonal. So every step attains the bound. Hence no constant smaller than $(\kappa - 1)/(\kappa + 1)$ can work. (Example (The bound is attained) is the case $a = 9$, $c = 1$, rescaled.)

Exercise. C2: Finite termination, from polynomials

Give a second proof of the second statement of Corollary 24.7.5 — that d distinct eigenvalues force termination in at most d steps — using Eq. 24.7.6 instead of a dimension count. *Hint: consider $q(x) = \prod_{i=1}^d (1 - x/\mu_i)$.*

Solution. Let $\mu_1, \dots, \mu_d$ be the distinct eigenvalues of $\mathbf{A}$; all are positive, so $q(x) = \prod_{i=1}^d (1 - x/\mu_i)$ is defined, has degree d , and satisfies $q(0) = 1$. It is therefore admissible in Eq. 24.7.6 with $k = d$. By Step 2 of the proof of Theorem 24.7.7,

$$\|\boldsymbol{\varepsilon}^{(d)}\|_{\mathbf{A}}^2 \leq \sum_i \lambda_i q(\lambda_i)^2 u_i^2 = 0,$$

since every eigenvalue λ_i of $\mathbf{A}$ equals some μ_j and so is a root of q . Hence $\boldsymbol{\varepsilon}^{(d)} = \mathbf{0}$ and $\mathbf{x}^{(d)} = \mathbf{x}^*$. (Eq. 24.7.6 was stated for iterates before termination; if the iteration stops earlier the conclusion is immediate.)

Exercise. C3: What preconditioning does to the spectrum

Let $\mathbf{A}, \mathbf{M} \in M_n(\mathbb{R})$ be symmetric positive definite and write $\mathbf{M} = \mathbf{C}\mathbf{C}^\top$ with $\mathbf{C}$ invertible, which is possible by Theorem 13.2.3.

- (a) Prove that $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}$ is symmetric positive definite.
- (b) Prove that $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}$ and $\mathbf{M}^{-1}\mathbf{A}$ are similar, and deduce that their eigenvalues coincide.
- (c) Hence show that those eigenvalues are exactly the generalized eigenvalues of the pencil $\mathbf{A} - \lambda\mathbf{M}$ in the sense of Definition 13.4.1, and say what Theorem 24.7.7 then predicts for preconditioned conjugate gradients.

Solution.

- (a) It is symmetric because $(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top})^\top = \mathbf{C}^{-1}\mathbf{A}^\top\mathbf{C}^{-\top} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}$. It is positive definite by Proposition 13.1.2 (b) applied with $\mathbf{S} = \mathbf{C}^{-\top}$, which is invertible.
- (b) Put $\mathbf{P} = \mathbf{C}^{-\top}$, which is invertible. Then

$$\mathbf{P}(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top})\mathbf{P}^{-1} = \mathbf{C}^{-\top}\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}\mathbf{C}^\top = \mathbf{C}^{-\top}\mathbf{C}^{-1}\mathbf{A} = (\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{A} = \mathbf{M}^{-1}\mathbf{A},$$

using $(\mathbf{C}\mathbf{C}^\top)^{-1} = \mathbf{C}^{-\top}\mathbf{C}^{-1}$. So the two matrices are similar, and similar matrices have the same characteristic polynomial (Theorem 7.10.3), hence the same eigenvalues with multiplicity.

- (c) By Definition 13.4.1, λ is a generalized eigenvalue of the pencil $\mathbf{A} - \lambda\mathbf{M}$ exactly when $\mathbf{A}\mathbf{v} = \lambda\mathbf{M}\mathbf{v}$ for some $\mathbf{v} \neq \mathbf{0}$, that is when $\mathbf{M}^{-1}\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$; so the generalized eigenvalues are the eigenvalues of $\mathbf{M}^{-1}\mathbf{A}$, which by (b) are those of $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top}$. Running conjugate gradients on $(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-\top})\mathbf{y} = \mathbf{C}^{-1}\mathbf{b}$ therefore converges at the rate of Theorem 24.7.7 with κ replaced by the ratio of the largest to the smallest generalized eigenvalue of $\mathbf{A}$ relative to $\mathbf{M}$. Choosing $\mathbf{M}$ close to $\mathbf{A}$ makes that ratio close to 1; choosing $\mathbf{M} = \mathbf{A}$ makes it exactly 1, at the cost of having to solve with $\mathbf{A}$.