3. Linear systems: iterative methods

Given $n \in \mathbb{N} \setminus \{0,1\}$, $A \in \mathbb{R}^{n \times n}$ invertible, and $b \in \mathbb{R}^n$, find $x \in \mathbb{R}^n$ such that $Ax = b$.

Gaussian elimination provides the exact solution in exact arithmetic but requires about $\frac{2}{3}n^3$ arithmetic operations, which is prohibitive if n is large. Iterative methods are an alternative to Gaussian elimination that is interesting if n is large.

Contents:

1. Jacobi and Gauss-Seidel methods;
2. gradient methods if A is symmetric positive-definite.

References:

- Y. Saad, Iterative Methods for Sparse Linear Systems, second edition, SIAM, 2003.
- J. Nocedal & S.J. Wright, Numerical Optimization, second edition, Springer New York, 2006.

1. Jacobi and Gauss-Seidel methods

Both methods rely on a splitting decomposition $A = M - N$, where $m_{i,j}$ equals $a_{i,j}$ or 0, and M is easy to invert. The system $Ax = b$ becomes $Mx = Nx + b$ or $x = M^{-1}Nx + M^{-1}b$, which is a fixed-point equation. Given $x_0 \in \mathbb{R}^n$, the fixed-point method iterates $Mx_{i+1} = Nx_i + b$ for all $i \in \mathbb{N}$. Thus, every iteration amounts to solving an "easy" linear system:

- M is the diagonal of A in the Jacobi method;
- M is the lower part of A in the Gauss-Seidel method.

The matrix M is called the preconditioning matrix or preconditioner. Here is why. The simplest way to rewrite $Ax = b$ as a fixed-point equation is arguably $x = (I - A)x + b$. The method that has just been described corresponds to that simple fixed-point method for the system $M^{-1}Ax = M^{-1}b$, which is called the preconditioned system.

Computational cost:

- $Nx_i + b$ requires at most $2n^2 + n$ arithmetic operations;

- solving the "easy" linear system requires n arithmetic operations for the Jacobi method and n^2 for the Gauss-Seidel method.

Convergence: Lipschitz constant of $g : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto M^{-1}Nx + M^{-1}b$
Clearly, $\mathrm{Lip}_{\mathbb{R}^n}(g) = \|M^{-1}N\|$ for every norm on $\mathbb{R}^{n\times n}$ induced by a norm on $\mathbb{R}^n$.
Thus, convergence if $\|M^{-1}N\| < 1$.

Theorem (Saad 2003, Theorem 4.1).
The iteration $x_{i+1} = M^{-1}Nx_i + M^{-1}b$ converges for every $x_0 \in \mathbb{R}^n$ if and only if $\rho(M^{-1}N) < 1$.

For every $A \in \mathbb{R}^{n\times n}$, $\rho(A) := \max_{i\in\{1,\dots,n\}} |\lambda_i(A)|$ is called the spectral radius of A.
A matrix $A \in \mathbb{R}^{n\times n}$ is said to be strictly diagonally dominant if, for all $j \in \{1,\dots,n\}$, $|a_{j,j}| > \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{i,j}|$. A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is said to be positive-definite if, for all $x \in \mathbb{R}^n \setminus \{0\}$, $x^T A x > 0$.

Theorem (Saad 2003, Theorem 4.9).
If $A \in \mathbb{R}^{n\times n}$ is strictly diagonally dominant, then the associated Jacobi and Gauss-Seidel iterations converge for every $x_0 \in \mathbb{R}^n$.

Theorem (Saad 2003, Theorem 4.10).
If $A \in \mathbb{R}^{n\times n}$ is symmetric positive-definite, then the associated Gauss-Seidel iteration converges for every $x_0 \in \mathbb{R}^n$.

Stopping criterion: stop when $\frac{\|b - Ax_i\|}{\|b\|} \leq \varepsilon$ for some $\varepsilon \in (0,\infty)$.

If this inequality is satisfied and $Ax = b$, then, by the perturbation theorem,

$$\frac{\|x - x_i\|}{\|x\|} \leq \kappa(A) \frac{\|b - Ax_i\|}{\|b\|} \leq \kappa(A)\,\varepsilon.$$

2. Gradient methods for symmetric positive-definite A

This section focuses on the case where A is symmetric positive-definite. Then, the function $\phi : \mathbb{R}^n \to \mathbb{R} : x \mapsto \frac{1}{2} x^T A x - b^T x$ is strictly convex and $A^{-1}b$ is its unique global minimizer. Indeed, for all $x \in \mathbb{R}^n$, $\nabla\phi(x) = Ax - b$ and $\nabla^2 \phi(x) = A$. The gradient can be computed as follows: for all $x, v \in \mathbb{R}^n$, $\underbrace{\langle \nabla\phi(x), v \rangle}_{= v^T \nabla\phi(x)} = \phi'(x)v = \lim_{t \to 0} \frac{\phi(x+tv) - \phi(x)}{t}$ and, since

$$\begin{aligned} \phi(x+tv) - \phi(x) &= \frac{1}{2}(t v^T A x + t x^T A v + t^2 v^T A v) - t b^T v \\ &= \frac{t^2}{2} v^T A v + t v^T (Ax - b), \end{aligned}$$

$\lim_{t \to 0} \frac{\phi(x+tv) - \phi(x)}{t} = v^T(Ax - b)$. Thus, for all $x, v \in \mathbb{R}^n$, $v^T \nabla\phi(x) = v^T(Ax - b)$, which implies $\nabla\phi(x) = Ax - b$, as announced.

Therefore, solving the linear system amounts to minimizing ϕ, which can be done by line-search methods. These methods rely on the following concept.

Definition. A descent direction for ϕ at $x \in \mathbb{R}^n$ is a vector $v \in \mathbb{R}^n$ such that $\phi'(x)v < 0$.

Given a descent direction v for ϕ at $x \in \mathbb{R}^n$, there exists $\alpha_* \in (0, \infty)$ such that, for all $\alpha \in (0, \alpha_*]$, $\phi(x + \alpha v) < \phi(x)$. (Proof. Since ϕ is differentiable at x, for $\varepsilon := -\phi'(x)\frac{v}{2\|v\|}$, there exists $\alpha_* \in (0, \infty)$ such that, for all $\alpha \in (0, \alpha_*]$, $\frac{|\phi(x+\alpha v) - \phi'(x)\alpha v - \phi(x)|}{\alpha \|v\|} \leq \varepsilon$, i.e., $|\phi(x+\alpha v) - \phi'(x)\alpha v - \phi(x)| \leq -\frac{1}{2}\alpha\phi'(x)v$, which implies $\phi(x+\alpha v) \leq \phi(x) + \frac{1}{2}\alpha\phi'(x)v < \phi(x)$.) However, since ϕ is quadratic, it is possible to compute $\alpha_* := \operatorname{argmin}_{\alpha \in (0,\infty)} \phi(x+\alpha v)$. Indeed, as seen above, for all $\alpha \in (0, \infty)$, $\phi(x+\alpha v) = \phi(x) + \alpha v^T \nabla\phi(x) + \frac{\alpha^2}{2} v^T A v$ is minimum if and only if $\alpha = -\frac{v^T \nabla\phi(x)}{v^T A v}$. Moving along a descent direction with this optimal step size is called exact line search. If the descent direction is chosen to be the steepest descent direction, given by the negative gradient, then the method is called the steepest descent method or the gradient method.

$i \leftarrow 0$; $r_0 \leftarrow b - Ax_0$;
while $r_i \neq 0$ do
$\alpha_i \leftarrow \dfrac{r_i^T r_i}{r_i^T A r_i}$
$x_{i+1} \leftarrow x_i + \alpha_i r_i$
$r_{i+1} \leftarrow r_i - \alpha_i A r_i$
$i \leftarrow i+1$
end

Steepest descent with exact line search for ϕ

NB: $r_{i+1} = b - Ax_{i+1}$
$= b - A(x_i + \alpha_i r_i)$
$= r_i - \alpha_i A r_i$

Define, for all $x \in \mathbb{R}^n$, $\|x\|_A := \sqrt{x^T A x}$. Then, $\|\cdot\|_A$ is a norm on $\mathbb{R}^n$ and, for all $x \in \mathbb{R}^n$, $\sqrt{\lambda_{min}(A)}\,\|x\|_2 \leq \|x\|_A \leq \sqrt{\lambda_{max}(A)}\,\|x\|_2$.

Theorem (Nocedal & Wright 2006, Theorem 3.3).
For every $x_0 \in \mathbb{R}^n$, steepest descent with exact line search for ϕ generates a sequence $(x_i)_{i \in \mathbb{N}}$ such that, for all $i \in \mathbb{N}$,

$$\|x_i - x\|_A \leq \left(\frac{\kappa_2(A)-1}{\kappa_2(A)+1}\right)^i \|x_0 - x\|_A.$$

Thus, for all $i \in \mathbb{N}$, $\|x_i - x\|_2 \leq \sqrt{\kappa_2(A)} \left(\dfrac{\kappa_2(A)-1}{\kappa_2(A)+1}\right)^i \|x_0 - x\|_2$.

Preconditioning enables to improve convergence. Recall that, for all $P, Q \in \mathbb{R}^{n \times n}$ symmetric positive-definite, the eigenvalues of PQ are real and positive, although PQ is not necessarily symmetric. If it is possible to find a symmetric positive-definite $M \in \mathbb{R}^{n \times n}$ such that M is easy to invert and $\dfrac{\lambda_{max}(M^{-1}A)}{\lambda_{min}(M^{-1}A)}$ is smaller than $\kappa_2(A)$, then the following method can be considered.

$i \leftarrow 0$; $r_0 \leftarrow b - Ax_0$;
while $r_i \neq 0$ do
find $z \in \mathbb{R}^n$ such that $Mz = r_i$
$\alpha_i \leftarrow \dfrac{z^T r_i}{z^T A z}$
$x_{i+1} \leftarrow x_i + \alpha_i z$
$r_{i+1} \leftarrow r_i - \alpha_i A z$
$i \leftarrow i+1$
end

Preconditioned steepest descent with preconditioner M

Theorem

For every $x_0 \in \mathbb{R}^n$, the preconditioned steepest descent with preconditioner M generates a sequence $(x_i)_{i \in \mathbb{N}}$ such that, for all $i \in \mathbb{N}$,

$$\|x_i - x\|_A \leq \left(\frac{\lambda_{max}(M^{-1}A) - \lambda_{min}(M^{-1}A)}{\lambda_{max}(M^{-1}A) + \lambda_{min}(M^{-1}A)} \right)^i \|x_0 - x\|_A .$$

Choosing a descent direction different from the negative gradient can also improve convergence.

Conjugate gradient method

$i \leftarrow 0$; $r_0 \leftarrow b - Ax_0$; $p_0 \leftarrow r_0$;
while $r_i \neq 0$ do
$$\alpha_i \leftarrow \frac{r_i^T r_i}{p_i^T A p_i}$$
$$x_{i+1} \leftarrow x_i + \alpha_i p_i$$
$$r_{i+1} \leftarrow r_i - \alpha_i A p_i$$
$$\beta_{i+1} \leftarrow \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}$$
$$p_{i+1} \leftarrow r_{i+1} + \beta_{i+1} p_i$$
$i \leftarrow i+1$

For all $i \in \mathbb{N}$, $\nabla \phi(x_i)^T p_i = -r_i^T p_i = -\|r_i\|_2^2$. Thus, the step size corresponds to an exact line search.

Theorem (Nocedal & Wright 2006, Theorem 5.3 and (5.36)).

For every $x_0 \in \mathbb{R}^n$, the conjugate gradient method generates at most n iterates, which satisfy

$$\|x_i - x\|_A \leq 2 \left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^i \|x_0 - x\|_A .$$