STEEPEST DESCENT AND CONJUGATE GRADIENT METHODS

In this exposition, we consider two related iterative methods for the solution of linear systems of equations: the *steepest descent method* and the *conjugate gradients method*. Whereas the classical iterative methods are nowadays treated from the perspective of fixpoint theory, the steepest descent method and conjugate gradient method can approached from several quite different perspectives.

The perspective that we choose in this exposition is the minimization of energy functionals. We translate the linear systems of equations into an analytical problem: minimize the error in some norm. Computational methods to approximate the minimizer of such an energy functional then lead to the approximate solution of linear systems of equations.

1. Energy Functionals

Here and below, we let $\langle \cdot, \cdot \rangle$ be a scalar product on $\mathbb{R}^n$. Consider a matrix $A \in \mathbb{R}^{n \times n}$ that is symmetric positive definite with respect to that scalar product, that is,

$$\forall x, y \in \mathbb{R}^n : \langle Ax, y \rangle = \langle Ay, x \rangle,$$
$$\forall x \in \mathbb{R}^n : \langle Ax, x \rangle > 0.$$

Moreover we let $b \in \mathbb{R}^n$ be a right-hand side vector. We are interested in finding the solution of the linear system of equations

$$Ax = b.$$

We recast this problem of linear algebra into a problem of analysis.

We introduce the energy functional

$$\mathcal{J}(x) = \frac{1}{2}\langle Ax, x \rangle + \langle b, x \rangle.$$

The derivative of $\mathcal{J}$ in x is a linear mapping $\mathrm{D}_x\mathcal{J} : \mathbb{R}^n \to \mathbb{R}^n$ that is given by

$$\begin{aligned} \mathrm{D}_x\mathcal{J}(y) &= \frac{1}{2}\langle Ay, x \rangle + \frac{1}{2}\langle Ax, y \rangle + \langle b, y \rangle \\ &= \frac{1}{2}\langle Ax, y \rangle + \frac{1}{2}\langle Ax, y \rangle + \langle b, y \rangle = \langle Ax - b, y \rangle. \end{aligned}$$

We conclude that

$$\mathrm{D}_x\mathcal{J} = 0 \quad \Longleftrightarrow \quad Ax = b.$$

This characterizes the solution of $Ax^* = b$ as a critical point of the *energy functional* $\mathcal{J}$. Since the solution x^* is unique, we that $\mathcal{J}$ has one and only one critical point.

The symmetric positive definite matrix A induces a scalar product and a norm, called *energy scalar product* and *energy norm*, respectively. They are given by

$$\langle x, y\rangle_A := \langle Ax, y\rangle, \quad \|x\|_A := (\langle Ax, x\rangle)^{\frac{1}{2}} .$$

Thus we can write

$$\mathcal{J}(x) = \frac{1}{2}\|x\|_A^2 - \langle b, x\rangle.$$

We observe

$$\begin{aligned}
\mathcal{J}(x^* + y) &= \frac{1}{2}\langle A(x^* + y), x^* + y\rangle + \langle b, x^* + y\rangle \\
&= \frac{1}{2}\langle Ax^*, x^*\rangle + \frac{1}{2}\langle Ax^*, y\rangle + \frac{1}{2}\langle Ay, y\rangle + \langle b, x^* + y\rangle \\
&= \frac{1}{2}\langle Ax^*, x^*\rangle + \langle Ax^*, y\rangle + \frac{1}{2}\langle Ay, y\rangle + \langle b, x^* + y\rangle \\
&= \frac{1}{2}\langle Ax^*, x^*\rangle + \frac{1}{2}\langle Ay, y\rangle + \langle b, x^*\rangle \\
&= \mathcal{J}(x^*) + \frac{1}{2}\|y\|_A^2.
\end{aligned}$$

Hence the solution of $Ax^* = b$ is a minimizer of the energy functional $\mathcal{J}$.

Remark 1
The energy functional $\mathcal{J}$ has the nice property that we can compute it for any $x \in \mathbb{R}^n$ with the knowledge of the matrix A and the vector b.

However, another functional leads to the same analytical problem and is of theoretical interest as well. Consider the error functional

$$\mathcal{E}(x) := \frac{1}{2}\|x - x^*\|_A^2.$$

To compare $\mathcal{J}$ and $\mathcal{E}$, we compute that

$$\begin{aligned}
\mathcal{E}(x) &= \frac{1}{2}\|x\|_A^2 + \langle x^*, x\rangle_A + \frac{1}{2}\|x^*\|_A^2 \\
&= \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + \frac{1}{2}\langle Ax^*, x^*\rangle \\
&= \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + \frac{1}{2}\langle Ax^*, x^*\rangle \\
&= \mathcal{J}(x) + \frac{1}{2}\langle Ax^*, x^*\rangle.
\end{aligned}$$

This means that the energy $\mathcal{J}$ differs from the energy norm of the error only by a constant (that depends on A and b). In particular, minimizing $\mathcal{J}$ over some subset $V \subseteq \mathbb{R}^n$ is equivalent to minimizing the error $x - x^*$ in energy norm over V.

2. Line Search for Energy functionals

Starting from some initial approximation $x^{(0)}$ to the true solution x^*, we develop a sequence of approximate solutions $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$ by trying to "push down" the energy $\mathcal{J}(x^{(k)})$ at each iterate:

$$\mathcal{J}(x^{(0)}) \geq \mathcal{J}(x^{(1)}) \geq \mathcal{J}(x^{(2)}) \geq \ldots$$

Given the approximation $x^{(k)}$, how do we get the approximation $x^{(k+1)}$ with less energy? One possibility is *line search*.

Conside a non-zero vector $d \in \mathbb{R}^n$. Given any point $x \in \mathbb{R}^n$, we want to find $\alpha \in \mathbb{R}$ such that

$$\mathcal{L}_{x,d}(\alpha) := \mathcal{J}(x + \alpha d)$$

is minimal among all choices of α. In other words, given a starting point x, we want to walk along the line through x in direction d such that $\mathcal{J}$ is minimized over that line.

To find the minimum, we calculate the derivative of $\mathcal{J}$ along that line, that is, the derivative of $\mathcal{L}_{x,d}(\alpha)$ in the parameter α. We get

$$\begin{aligned}\mathcal{L}_{x,d}(\alpha) &= \frac{1}{2}\langle A(x+\alpha d), x+\alpha d\rangle - \langle b, x+\alpha d\rangle. \\ &= \frac{1}{2}\langle Ax, x\rangle + \alpha\langle Ax, d\rangle + \alpha^2\frac{1}{2}\langle Ad, d\rangle - \langle b, x\rangle - \alpha\langle b, d\rangle.\end{aligned}$$

Thus,

$$\partial_\alpha \mathcal{L}_{x,d}(\alpha) = \langle Ax, d\rangle + \alpha\langle Ad, d\rangle - \langle b, d\rangle = \langle Ax - b, d\rangle + \alpha\langle Ad, d\rangle.$$

and

$$\partial_\alpha^2 \mathcal{J}(x + \alpha d) = \langle Ad, d\rangle > 0.$$

Writing $r = b - Ax$, we thus get

$$\partial_\alpha \mathcal{L}_{x,d}(\alpha) = 0 \quad \Longleftrightarrow \quad \alpha = \frac{\langle r, d\rangle}{\langle Ad, d\rangle}.$$

Since the second derivative is positive, this is a minimizer of $\mathcal{J}$ along the line through x in direction d. Consequently, for any given $x, d \in \mathbb{R}^n$, the functional $\mathcal{J}$ is minimized at the point

$$x + \frac{\langle r, d\rangle}{\langle Ad, d\rangle} d.$$

We found out how to minimize the value of $\mathcal{J}$ over some given line in $\mathbb{R}^n$.

Suppose that we have a sequence of search directions $d^{(0)}, d^{(1)}, d^{(2)}, \ldots$, then, for an algorithm, start with some initial guess $x^{(0)}$ and recursively obtain $x^{(k+1)}$ from $x^{(k)}$ by minimizing the energy $\mathcal{J}$ over the line through $x^{(k)}$ in direction $d^{(k)}$. That is, we compute recursively

$$x^{(k+1)} = x^{(k)} + \frac{\langle b - Ax^{(k)}, d^{(k)}\rangle}{\langle Ad^{(k)}, d^{(k)}\rangle} d^{(k)}.$$

At this point we recall the residuals

$$r^{(k)} = b - Ax^{(k)}$$

of each iteration. We can thus write more compactly

$$x^{(k+1)} = x^{(k)} + \frac{\langle r^{(k)}, d^{(k)}\rangle}{\langle Ad^{(k)}, d^{(k)}\rangle} d^{(k)}.$$

We still need to agree on a sequence of search directions.

3. Steepest Descent

We compute the search directions one at a time for each given $x^{(k)}$. Since we want to minimize $\mathcal{J}$, the direction of steepest descent of $\mathcal{J}$ seems to be a good choice.

Recall that the gradient of $\mathcal{J}$ is given by

$$\mathrm{D}_x\mathcal{J} = \langle Ax - b, \cdot \rangle$$

and points into the direction of steepest *ascent*. Hence the residual $r = b - Ax$ at the point x points in the direction of the steepest descent of the functional $\mathcal{J}$.

This leads to Steepest Descent algorithm. Given an approximate solution $x^{(k)}$, we compute the residual $r^{(k)} = b - Ax^{(k)}$ and find $x^{(k+1)}$ by minimizing $\mathcal{J}$ along the line through $x^{(k)}$ in direction $r^{(k)}$. We can do that as long as the residual is not sufficiently small.

Given $A \in \mathbb{R}^{n\times n}$ symmetric positive definite and $b \in \mathbb{R}^n$, for any starting value $x^{(0)} \in \mathbb{R}^n$ we define recursively

$$r^{(k)} = b - Ax^{(k)}, \quad \alpha^{(k)} = \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle Ar^{(k)}, r^{(k)} \rangle}, \quad x^{(k+1)} = x^{(k)} + \alpha^{(k)} r^{(k)}.$$

A very basic pseudocode for the algorithm that one typically implements is the following:

$r = b - Ax$
WHILE $\|r\| > \epsilon$
  $\alpha = \dfrac{\langle r, r \rangle}{\langle Ar, r \rangle}$
  $x = x + \alpha r$
  $r = b - Ax$
END WHILE

Remark 2
This is a good opportunity to demonstrate a few techniques on how to save computations (at the expense) of memory in such iterative methods. Assuming exact arithmetics, we have an equivalent pseudocode given by

$r = b - Ax$
$p = Ar$
Calculate $\langle r, r \rangle$, $\langle p, r \rangle$
WHILE $\langle r, r \rangle > \epsilon$
  $\alpha = \dfrac{\langle r, r \rangle}{\langle p, r \rangle}$
  $x = x + \alpha r$
  $r = r - \alpha p$
  $p = Ar$
  Calculate $\langle r, r \rangle$, $\langle p, r \rangle$
END WHILE

Here, we assume that $\langle r, r\rangle$ and $\langle p, r\rangle$ are saved in two scalar variables. This pseudocode uses only one matrix-vector multiplication. Furthermore, if the CSR (compressed sparse rows) format is used to save A, then the vector updates and scalar products in the main loop can be processed within one (parallelizable) FOR loop.

Remark 3
The resulting algorithm can be interpreted as a non-stationary Richardson method, that is, a Richardson method where the parameter θ changes from iteration to iteration. In this specific example, at every iteration θ is chosen such that the resulting error has minimal energy norm.

4. Convergence of Gradient Method

We can prove convergence rates for the Steepest Descent algorithm by different means. One possibility is to compare it with the Richardson method.

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with smallest and largest eigenvalue $0 < \lambda_{\min} \leq \lambda_{\max}$. We define the condition number of A as

$$\kappa(A) := \frac{\lambda_{\max}}{\lambda_{\min}}.$$

We first prove a result on the convergence of the Richardson iteration.

Lemma 1
Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite with smallest and largest eigenvalues $0 < \lambda_{\min} \leq \lambda_{\max}$, respectively. Letting $\theta = \frac{1}{2}(\lambda_{\min} + \lambda_{\max})$, we have

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

Proof. We prove slightly more general result. Let $k \in \mathbb{N}_0$. With θ as defined above, we prove that

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

Let $A = VDV^T$ be the spectral decomposition of A. Write $x = V\xi$ and $x^\star = V\xi^\star$ with $\xi, \xi^\star \in \mathbb{R}^n$. We then have

$$\left(\mathrm{Id} - \theta A\right)\left(x^* - x\right) = V\left(\mathrm{Id} - \theta D\right)\left(\xi^* - \xi\right)$$

Hence

$$\begin{aligned}
\left\|\left(\mathrm{Id} - \theta A\right)\left(x^* - x\right)\right\|_{A^k}^2 &= \left\langle V(\mathrm{Id} - \theta D)\left(\xi^* - \xi\right), VD^kV^T \cdot V(\mathrm{Id} - \theta D)\left(\xi^* - \xi\right)\right\rangle \\
&= \left\langle(\mathrm{Id} - \theta D)\left(\xi^* - \xi\right), D^k(\mathrm{Id} - \theta D)\left(\xi^* - \xi\right)\right\rangle \\
&= \|\mathrm{Id} - \theta D\|_\infty \langle \xi^* - \xi, \xi^* - \xi\rangle_{D^k} \\
&= \|\mathrm{Id} - \theta D\|_\infty \|x^* - x\|_{A^k}^2.
\end{aligned}$$

Since the function $\lambda \mapsto |1-\theta\lambda|$ has no local minimum in the open interval $(\lambda_{\min}, \lambda_{\max})$, we observe

$$\max_{1\leq i\leq n} |1-\theta D_{ii}| = \max\left\{|1-\theta\lambda_{\min}|, |1-\theta\lambda_{\max}|\right\}.$$

The last expression, as a function in θ assumes its maximum when

$$|1-\theta\lambda_{\min}| = |1-\theta\lambda_{\max}|$$

Since $\lambda_{\min} \leq \lambda_{\max}$, this is equivalent to

$$\theta\lambda_{\min} - 1 = 1 - \theta\lambda_{\max} \quad \Longleftrightarrow \quad \theta = \frac{2}{\lambda_{\min}+\lambda_{\max}}.$$

In particular, we have

$$\max\left\{|1-\theta\lambda_{\min}|, |1-\theta\lambda_{\max}|\right\} = \frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}} = \frac{\kappa(A)-1}{\kappa(A)+1}.$$

This completes the proof. □

Lemma 2
For the iterates $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$ produced by the steepest descent algorithm we have

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

Proof. By the minimization property of the steepest descent algorithm and the contraction property of the Richardson iteration with optimal parameter we have

$$\begin{aligned}\left\|x^* - x^{(k+1)}\right\|_A &\leq \left\|x^* - x^{(k)} + \alpha^{(k)} r^{(k)}\right\|_A \leq \left\|x^* - x^{(k)} + \theta r^{(k)}\right\|_A \\ &\leq \left\|x^* - x^{(k)} + \theta A\left(x^* - x^{(k)}\right)\right\|_A \leq \frac{\kappa(A)-1}{\kappa(A)+1}\left\|x^* - x^{(k)}\right\|_A.\end{aligned}$$

This completes the proof. □

Remark 4
We get the same convergence rate as with the Richardson method. However, we do not need to know any explicit bounds on the eigenvalues but do need that the matrix is symmetric positive definite.

The Richardson iteration can be interpreted as a line search with fixed stepsize. By contrast, the steepest descent method can be interpreted as a Richardson-like method with variable step size.

5. Preconditioned Gradient Method

Let $P \in \mathbb{R}^{n\times n}$ be a preconditioner for the symmetric positive definite system $Ax = b$. The system $P^{-1}Ax = P^{-1}b$ may have a matrix with better condition number but generally will not be symmetric positive definite.

Suppose that $P^{-1} = GG^T$. We may consider the system

$$G^T AGy = G^T b, \quad x = Gy.$$

It then follows that

$$s^{(k)} = G^T b - G^T AGy^{(k)}, \quad \gamma^{(k)} = \frac{\langle s^{(k)}, s^{(k)} \rangle}{\langle G^T AGs^{(k)}, s^{(k)} \rangle}, \quad y^{(k+1)} = y^{(k)} + \gamma^{(k)} s^{(k)}.$$

We have approximate solutions and residual relations

$$x^{(k)} = Gy^{(k)}, \quad s^{(k)} = G^T r^{(k)}, \quad r^{(k)} := b - Ax^{(k)}.$$

Hence

$$x^{(k+1)} = x^{(k)} + \gamma^{(k)} GG^T r^{(k)}$$

and

$$\gamma^{(k)} = \frac{\langle G^T r^{(k)}, G^T r^{(k)} \rangle}{\langle G^T AGG^T r^{(k)}, G^T r^{(k)} \rangle} = \frac{\langle r^{(k)}, GG^T r^{(k)} \rangle}{\langle AGG^T r^{(k)}, GG^T r^{(k)} \rangle} = \frac{\langle r^{(k)}, P^{-1} r^{(k)} \rangle}{\langle AP^{-1} r^{(k)}, P^{-1} r^{(k)} \rangle}.$$

Consequently the recursion can be written

$$r^{(k)} = b - Ax^{(k)}, \quad z^{(k)} = P^{-1} r^{(k)}, \quad \gamma^{(k)} = \frac{\langle r^{(k)}, z^{(k)} \rangle}{\langle Az^{(k)}, z^{(k)} \rangle}, \quad x^{(k+1)} = x^{(k)} + \gamma^{(k)} z^{(k)}.$$

A very basic pseudocode for the preconditioned steepest descent reads

$r = b - Ax$
$z = P^{-1} r$
WHILE $\langle r, z \rangle > \epsilon$
  $\gamma = \dfrac{\langle r, z \rangle}{\langle Az, z \rangle}$
  $x = x + \gamma z$
  $r = b - Ax$
  $z = P^{-1} r$
END WHILE

Remark 5
We can again reduce the amount of computational work at the expense of further auxiliary variables

$r = b - Ax$
$z = P^{-1} r$
$p = Az$
Calculate $\langle r, z \rangle$, $\langle p, z \rangle$
WHILE $\langle r, z \rangle > \epsilon$
  $\gamma = \dfrac{\langle r, z \rangle}{\langle p, z \rangle}$
  $x = x + \gamma z$
  $r = r - \gamma p$
  $z = P^{-1} r$
  $p = Az$
  Calculate $\langle r, z \rangle$, $\langle p, z \rangle$
END WHILE

Note that $\langle r, z \rangle = \langle r, P^{-1} r \rangle$.

Remark 6
If $P^{-1} = \mathrm{Id}$, then we get the same calculations as for the original steepest descent method. This obviously happens if $G = \mathrm{Id}$.

It also happens whenever G is orthogonal matrix, $P^{-1} = GG^T = \mathrm{Id}$. Now, if A has the spectral decomposition $A = QDQ^T$, then using the symmetric preconditioning with $G = Q$ gives an Preconditioned Steepest Descent Algorithm with the same vectors and numbers as the original Steepest Descent Algorithm. However, the Steepest Descent Algorithm applied to the preconditioned system G^TAG has residuals and errors that are different but with the same norm.

We conclude that the convergence behavior of steepest descent depends only on the eigenvalue distribution of A. In particular, to understand the convergence behavior, it is sufficient to consider examples with A diagonal.