Dr H. Fawzi

Mathematical Tripos Part II: Michaelmas Term 2023

Numerical Analysis – Lecture 18

4.1 Steepest descent and conjugate gradient methods

For solving $A\boldsymbol{x} = \boldsymbol{b}$ with a symmetric positive definite matrix A, we consider iterative methods based on an optimization formulation. Consider the convex quadratic function

$$F(\boldsymbol{x}) := \tfrac{1}{2}\langle \boldsymbol{x}, A\boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle \tag{4.5}$$

where $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{u}^T\boldsymbol{v}$ is the Euclidean inner product. Note that the global minimizer of F is $\boldsymbol{x}^* = A^{-1}\boldsymbol{b}$. Indeed

$$F(\boldsymbol{x}^* + \boldsymbol{h}) - F(\boldsymbol{x}^*) = \langle \boldsymbol{h}, A\boldsymbol{x}^* - \boldsymbol{b}\rangle + \tfrac{1}{2}\langle \boldsymbol{h}, A\boldsymbol{h}\rangle \geq 0$$

for any $\boldsymbol{h}$. Observe that F can also be written as

$$F(\boldsymbol{x}) = \tfrac{1}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|_A^2 + \text{constant}$$

where $\|\boldsymbol{y}\|_A := \langle \boldsymbol{y}, A\boldsymbol{y}\rangle^{1/2} = \sqrt{\boldsymbol{y}^T A\boldsymbol{y}}$ is the A-norm of A. (The constant in the above formulation is a term that does not depend on $\boldsymbol{x}$, so it is irrelevant for the purpose of minimizing F, the constant is $\frac{1}{2}\boldsymbol{b}^T A^{-1}\boldsymbol{b}$.)

Gradient/Steepest descent The *gradient descent* method for minimizing F has iterates

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla F(\boldsymbol{x}^{(k)})$$

where $\nabla F(\boldsymbol{x}^{(k)})$ is the gradient of F at $\boldsymbol{x}^{(k)}$, and $\alpha_k > 0$ is the step size. For our quadratic function, it is easy to verify that

$$\nabla F(\boldsymbol{x}^{(k)}) = A\boldsymbol{x}^{(k)} - \boldsymbol{b} = -\boldsymbol{r}^{(k)}$$

where $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ is the residual. There are multiple ways to choose the step size α_k:

- *Constant step-size* $\alpha_k = \alpha$. In this case the iteration takes the form

  $$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha(A\boldsymbol{x}^{(k)} - \boldsymbol{b}) = (I - \alpha A)\boldsymbol{x}^{(k)} + \alpha\boldsymbol{b}$$

  which is nothing but a Jacobi-like iteration with $D = \alpha^{-1}I$ (we say Jacobi-like because the diagonal of A is not necessarily equal to $\alpha^{-1}I$). We know from previous lectures that the method converges iff

  $$\rho(I - \alpha A) < 1 \iff |1 - \alpha\lambda_i| < 1 \;\; \forall \lambda_i \text{ eigenvalues of } A \iff 0 < \alpha < 2/\rho(A).$$

  For example, assume the eigenvalues of A are all in $[l, L]$ where $0 < l < L$. Then one can choose $\alpha = 1/L$, and in this case the convergence rate is given by $\rho(I - \frac{1}{L}A) = 1 - l/L$, i.e., the error $\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|$ decays like $(1 - l/L)^k$. The quantity $L/l \geq 1$ is known as the condition number of A. We see that, as the condition number grows, the convergence rate becomes worse and worse.

- *Exact line search*. Another way to choose the step size α_k is using line search. Here α_k is chosen so that it achieves the smallest possible value of F along the search direction, i.e., $\alpha_k = \arg\min_\alpha F(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{d}^{(k)})$ where $\boldsymbol{d}^{(k)}$ is the search direction, equal to the negative gradient. Because our function is quadratic, one can get a closed form expression for the optimal α.

  Lemma 4.20 *Let F be the function defined in (4.5). Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x}$ be the residual and let $\boldsymbol{d} \in \mathbb{R}^n$ be a search direction. Then*

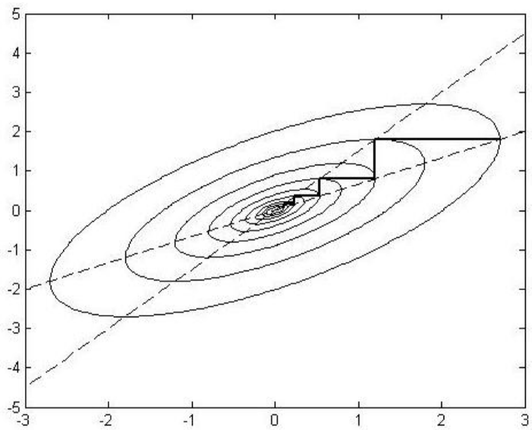
  $$\arg\min_\alpha F(\boldsymbol{x} + \alpha\boldsymbol{d}) = \frac{\langle \boldsymbol{r}, \boldsymbol{d}\rangle}{\langle \boldsymbol{d}, A\boldsymbol{d}\rangle}. \tag{4.6}$$

Proof. The function $F(\boldsymbol{x} + \alpha \boldsymbol{d}) = F(\boldsymbol{x}) - \alpha\langle \boldsymbol{r}, \boldsymbol{d}\rangle + \alpha^2/2\langle \boldsymbol{d}, A\boldsymbol{d}\rangle$ is quadratic in the single variable α. The minimum is attained at α s.t. $-\langle \boldsymbol{r}, \boldsymbol{d}\rangle + \alpha\langle \boldsymbol{d}, A\boldsymbol{d}\rangle = 0$ which gives the desired formula. $\square$

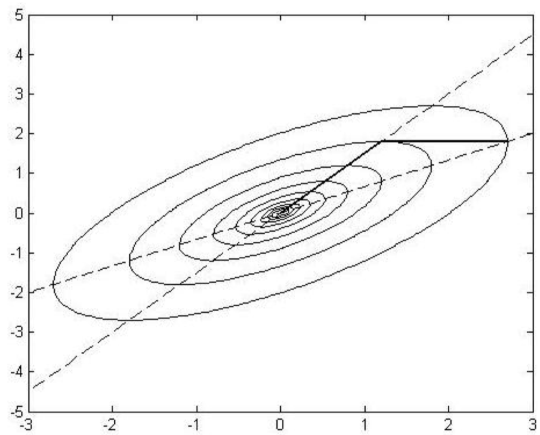
The gradient descent method with exact line search thus takes the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \frac{\|\boldsymbol{r}^{(k)}\|_2^2}{\|\boldsymbol{r}^{(k)}\|_A^2}\boldsymbol{r}^{(k)},$$

where we used the fact that the gradient direction is $\boldsymbol{d} = -\nabla F(\boldsymbol{x}^{(k)}) = \boldsymbol{r}^{(k)}$. It can be shown that the speed of convergence of the gradient descent with exact line search is, like with the constant step size, $\approx (1 - l/L)^k$ where $0 < l < L$ are the smallest and largest eigenvalues of A. The figure below (left) shows an example of the gradient descent method with exact line search applied to a two-dimensional quadratic function F. Note the zig-zag behaviour of the iterates.



(a) Worst case scenario of steepest descent



(b) Conjugate gradient method applied to the same problem as in (a)

Conjugate directions Let's revisit equation (4.6) for a general direction $\boldsymbol{d}$ (i.e., not necessarily equal to the negative gradient). Assume $\boldsymbol{x} = \boldsymbol{x}^{(k)}$, and let $\boldsymbol{e}^{(k)} = \boldsymbol{x}^* - \boldsymbol{x}^{(k)}$ be the error and $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)} = A\boldsymbol{e}^{(k)}$ be the residual. Then we can write $\langle \boldsymbol{r}^{(k)}, \boldsymbol{d}\rangle = \langle \boldsymbol{e}^{(k)}, \boldsymbol{d}\rangle_A$, and so for a general search direction $\boldsymbol{d}$ with an exact line search, the iterate takes the form $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \frac{\langle \boldsymbol{e}^{(k)}, \boldsymbol{d}\rangle_A}{\langle \boldsymbol{d}, \boldsymbol{d}\rangle_A}\boldsymbol{d}$. By substracting $\boldsymbol{x}^*$, the iterates in terms of the error $\boldsymbol{e}^{(k+1)}$ are given by:

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{e}^{(k)} - \frac{\langle \boldsymbol{e}^{(k)}, \boldsymbol{d}\rangle_A}{\langle \boldsymbol{d}, \boldsymbol{d}\rangle_A}\boldsymbol{d}. \tag{4.7}$$

Geometrically, this means that $\boldsymbol{e}^{(k+1)}$ is the projection of $\boldsymbol{e}^{(k)}$ onto the hyperplane that is A-orthogonal to $\boldsymbol{d}$, i.e., we have

$$\langle \boldsymbol{e}^{(k+1)}, \boldsymbol{d}\rangle_A = 0 \tag{4.8}$$

Definition 4.21 (Conjugate directions) The vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ are *conjugate* with respect to a symmetric positive definite matrix A if they are nonzero and A-orthogonal: $\langle \boldsymbol{u}, \boldsymbol{v}\rangle_A := \langle \boldsymbol{u}, A\boldsymbol{v}\rangle = 0$.

The observation above allows us to prove the following important result.

Theorem 4.22 *Let $\boldsymbol{d}^{(0)}, \boldsymbol{d}^{(1)}, \ldots, \boldsymbol{d}^{(n-1)}$ be n nonzero pairwise conjugate directions, and consider the sequence of iterates*

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}, \qquad \alpha_k = \frac{\langle \boldsymbol{r}^{(k)}, \boldsymbol{d}^{(k)}\rangle}{\langle \boldsymbol{d}^{(k)}, A\boldsymbol{d}^{(k)}\rangle}.$$

Let $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ be the residual. Then for each $k = 1, \ldots, n$, $\boldsymbol{r}^{(k)}$ is orthogonal to $\operatorname{span}\{\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(k-1)}\}$. In particular $\boldsymbol{r}^{(n)} = 0$.

Proof. Since $\boldsymbol{r}^{(k)} = A\boldsymbol{e}^{(k)}$, it suffices to show that $\boldsymbol{e}^{(k)}$ is A-orthogonal to $\operatorname{span}\{\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(k-1)}\}$. The proof is by induction on k. For $k = 0$ there is nothing to prove. Assume the statement is true for $k \geq 0$, and consider the equation (4.7) (with $\boldsymbol{d} = \boldsymbol{d}^{(k)}$). From the induction hypothesis, and the fact that the $\boldsymbol{d}^{(i)}$ are pairwise conjugate directions, we see that $\boldsymbol{e}^{(k+1)}$ is A-orthogonal to $\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(k-1)}$. Furthermore, we have already seen in (4.8) that $\langle \boldsymbol{e}^{(k+1)}, \boldsymbol{d}^{(k)} \rangle_A = 0$. Thus this shows that $\boldsymbol{e}^{(k+1)}$ is A-orthogonal to $\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(k)}$ as desired. $\square$

So, if a sequence $(\boldsymbol{d}^{(k)})$ of conjugate directions is at hands, we have an iterative procedure with good approximation properties. In the conjugate gradient method, the (A-orthogonal) basis of conjugate directions is constructed by A-orthogonalization of the sequence of gradients of F at the $\boldsymbol{x}^{(k)}$; or equivalently the sequence of residuals $\{\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{r}^{(k)}\}$.