Lecture 2

Conjugate Gradient Method

New Directions in Mathematics

Adérito Araújo (alma@mat.uc.pt)
March 22, 2024

Gradient Method

- For SPD matrices, solving $Ax = b$ is equivalent to finding the minimizer $x \in \mathbb{R}^n$ of the quadratic form

$$\phi(y) = \frac{1}{2}y^T Ay - y^T b = (y, y)_A - (y, b)$$

- Two phases: (i) choosing a descent direction (the residual); (ii) picking up a point of local minimum for ϕ along that direction
- For a given direction $p^{(k)}$, the value of α_k was obtained such that $\phi(x^{(k)} + \alpha p^{(k)})$ is minimized

$$\alpha_k = \frac{{p^{(k)}}^T r^{(k)}}{{p^{(k)}}^T A p^{(k)}} = \frac{(p^{(k)}, r^{(k)})}{(p^{(k)}, p^{(k)})_A} \qquad (4)$$

- For the gradient method $p^{(k)} = r^{(k)}$

Richardson and Gradient Methods

- Richardson Method ($P = I$)

$$
\begin{aligned}
&x^{(0)} \text{ given; } r^{(0)} = b - Ax^{(0)}\\
&\textbf{for } k = 0, 1, \ldots\\
&\qquad \text{solve } Iz^{(k)} = r^{(k)}\\
&\qquad \text{compute } \alpha_k\\
&\qquad x^{(k+1)} = x^{(k)} + \alpha_k z^{(k)}\\
&\qquad r^{(k+1)} = r^{(k)} - \alpha_k A z^{(k)}\\
&\textbf{until convergence}
\end{aligned}
$$

- Gradient Method

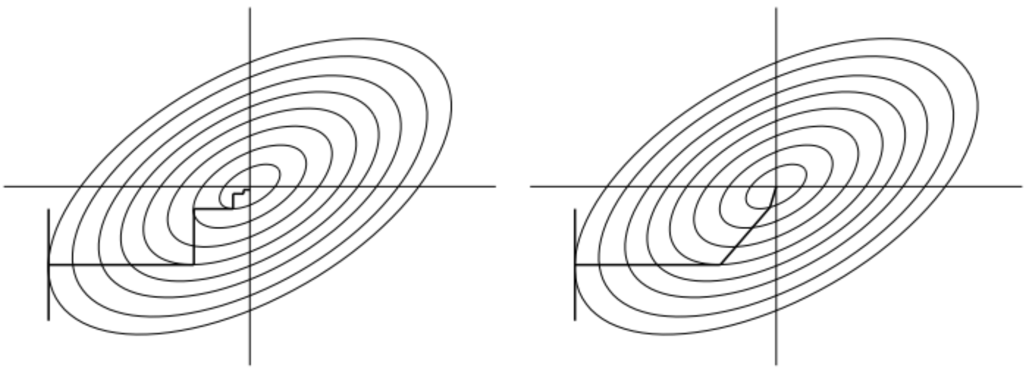
$$
\begin{aligned}
&x^{(0)} \text{ given; } r^{(0)} = b - Ax^{(0)}\\
&\textbf{for } k = 0, 1, \ldots\\
&\qquad \alpha_k = \frac{{r^{(k)}}^T r^{(k)}}{{r^{(k)}}^T A r^{(k)}}\\
&\qquad x^{(k+1)} = x^{(k)} + \alpha_k r^{(k)}\\
&\qquad r^{(k+1)} = r^{(k)} - \alpha_k A r^{(k)}\\
&\textbf{until convergence}
\end{aligned}
$$

- Exercise 2.1: Prove that, for $p^{(k)} = r^{(k)}$

$$(p^{(k)}, r^{(k+1)}) = {p^{(k)}}^T r^{(k+1)} = 0 \quad \Leftrightarrow \quad p^{(k)} \perp r^{(k+1)},$$

i.e., the new residual becomes orthogonal to the search direction

Improve Steepest Descent Method



- For the gradient method

$$\|e^{(k+1)}\|_A \leqslant \frac{K_2(A) - 1}{K_2(A) + 1} \|e^{(k)}\|_A$$

- Goal: Improve the convergence, minimizing $\|e^{(k)}\|_A$ at each step

General Direction Method

- Exercise 2.2: Prove that, for α_k given by (4)

$$\phi(x^{(k)} + \alpha_k p^{(k)}) = \phi(x^{(k)}) - \frac{1}{2}\frac{(p^{(k)}, r^{(k)})^2}{(p^{(k)}, p^{(k)})_A}.$$

- Find search direction $p^{(k)}$ such that $(p^{(k)}, r^{(k)}) \neq 0$
- General Direction Method

  $x^{(0)}$ given; $r^{(0)} = b - Ax^{(0)}$; $p^{(0)} = r^{(0)}$
  for $k = 0, 1, \dots$

  $$\alpha_k = \frac{{r^{(k)}}^T r^{(k)}}{{r^{(k)}}^T A r^{(k)}}$$
  $$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$$
  $$r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)}$$
  choose $p^{(k+1)}$ such that $(p^{(k+1)}, r^{(k+1)}) \neq 0$

  until convergence
- For $x^{(0)} = 0$, since $x^{(k)} \in \langle p^{(0)}, p^{(1)}, \dots, p^{(k-1)} \rangle$ then

$$\langle x^{(1)}, x^{(2)}, \dots, x^{(k)} \rangle = \langle p^{(0)}, p^{(1)}, \dots, p^{(k-1)} \rangle$$

Conjugate Direction Method

- Goal: Find search direction $p^{(k+1)}$ that provides a faster convergence
- Let $p^{(0)} = r^{(0)}$. Search for directions $p^{(k+1)}$ in such way that

$$(p^{(j)}, p^{(k+1)})_A = 0, \quad j = 0, 1, \dots, k, \qquad (5)$$

  i.e., the directions are A-conjugate (or A-orthogonal)
- Exercise 2.3: Prove that, if $p^{(j)} \neq 0$, $j = 0, 1, \dots, k-1$, are A-conjugate, then:
  1. the directions $\{p^{(0)}, p^{(1)}, \dots, p^{(k-1)}\}$ are linearly independent;
  2. the algorithm converges in at most n steps.

Conjugate Gradient (CG) Method

- Let $p^{(0)} = r^{(0)}$. Search for directions of the form

$$p^{(k+1)} = r^{(k+1)} - \beta_k p^{(k)}, \quad k = 0, 1, \ldots$$

where $\beta_k \in \mathbb{R}$ must be determined in such way that (5) holds, i.e., the directions are A-conjugate

- Exercise 2.4: Prove that, for $j = k$,

$$\beta_k = \frac{(p^{(k)}, r^{(k+1)})_A}{(p^{(k)}, p^{(k)})_A}, \quad k = 0, 1, \ldots$$

and, by induction, using the above β_k, that

$$(p^{(j)}, p^{(k+1)})_A = 0, \quad j = 0, 1, \ldots, k-1.$$

Conjugate Gradient (CG) Method

- Algorithm: Conjugate Gradient Method

$x^{(0)}$ given; $r^{(0)} = b - Ax^{(0)}$; $p^{(0)} = r^{(0)}$
for $k = 0, 1, \ldots$

$$\alpha_k = \frac{(p^{(k)}, r^{(k)})}{(p^{(k)}, p^{(k)})_A} \quad \text{\% step lenght}$$

$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)} \quad \text{\% update solution}$$

$$r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)} \quad \text{\% update residual}$$

$$\beta_k = \frac{(p^{(k)}, r^{(k+1)})_A}{(p^{(k)}, p^{(k)})_A} \quad \text{\% improvement this step}$$

$$p^{(k+1)} = r^{(k+1)} - \beta_k p^{(k)} \quad \text{\% search direction}$$

until convergence

- Exercise 2.5: Show that the algorithm requires only one matrix-vector product $Ap^{(k)}$ per iteration.

Conjugate Gradient (CG) Method

- Algorithm: Conjugate Gradient Method

  $x^{(0)} = 0$; $r^{(0)} = b$; $p^{(0)} = r^{(0)}$
  for $k = 0, 1, \ldots$

  $$\alpha_k = \frac{(p^{(k)}, r^{(k)})}{(p^{(k)}, p^{(k)})_A} \qquad \text{\% step lenght}$$

  $$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)} \qquad \text{\% update solution}$$

  $$r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)} \qquad \text{\% update residual}$$

  $$\beta_k = \frac{(p^{(k)}, r^{(k+1)})_A}{(p^{(k)}, p^{(k)})_A} \qquad \text{\% improvement this step}$$

  $$p^{(k+1)} = r^{(k+1)} - \beta_k p^{(k)} \qquad \text{\% search direction}$$

  until convergence

- Exercise 2.6: Show that

$$\alpha_k = \frac{\|r^{(k)}\|^2}{\|p^{(k)}\|_A^2} \quad \text{and} \quad \beta_k = -\frac{\|r^{(k+1)}\|^2}{\|r^{(k)}\|^2}.$$

Krylov Subspace

Krylov Subspace

- Krylov Subspace: $\mathcal{K}_k = \mathcal{K}_k(A; b) = \langle b, Ab, \ldots, A^{k-1}b \rangle$
- CG for $Ax = b$, $A \in \mathbb{R}^{n \times n}$ SPD, $x^{(0)} = 0$, $p^{(0)} = r^{(0)} = b$

Theorem 2.1

As long as $r^{(k-1)} \neq 0$ (CG not yet converged), the algorithm proceeds without divisions by zero and

$$\begin{aligned}\mathcal{K}_k &= \langle x^{(1)}, x^{(2)}, \ldots, x^{(k)} \rangle = \langle p^{(0)}, p^{(1)}, \ldots, p^{(k-1)} \rangle \\ &= \langle r^{(0)}, r^{(1)}, \ldots, r^{(k-1)} \rangle = \langle b, Ab, \ldots, A^{k-1}b \rangle.\end{aligned}$$

- Exercise 2.7: Prove that the residuals are orthogonal,

$$r^{(k)^T} r^{(j)} = 0, \quad j < k,$$

and the search directions are A-conjugate (or A-orthogonal),

$$p^{(k)^T} A p^{(j)} = 0, \quad j < k.$$

Convergence Result

Theorem 2.2: Monotonic convergence

If the iteration has not yet converged then $x^{(k)}$ is the only point in $\mathcal{K}_k$ that minimizes $\|e^{(k)}\|_A$. The convergence is monotonic,

$$\|e^{(k)}\|_A \leqslant \|e^{(k-1)}\|_A,$$

and $\|e^{(k)}\|_A = 0$ is achieved for some $k \leqslant n$.

- Proof: For any other point $y = x^{(k)} - \Delta y \in \mathcal{K}_k$, the error is

$$\begin{aligned}\|e\|_A^2 &= (e^{(k)} + \Delta y)^T A (e^{(k)} + \Delta y) \\ &= (e^{(k)})^T A e^{(k)} + (\Delta y)^T A (\Delta y) + 2(e^{(k)})^T A (\Delta y)\end{aligned}$$

But $(e^{(k)})^T A(\Delta y) = (r^{(k)})^T(\Delta y) = 0$ since $r^{(k)} \perp \mathcal{K}_k$, so Δy minimizes $\|e\|_A^2$. Since $A \in$ SPD, the monotonic convergence follow from $\mathcal{K}_k \subseteq \mathcal{K}_{k+1}$, and $\mathcal{K}_k \subseteq \mathbb{R}^n$ unless converged. $\square$

Optimization in CG

- CG can be interpreted as a minimization algorithm
- We know it minimizes $\|e\|_A$, but this cannot be evaluated
- CG minimizes the quadratic function $\phi(y) = \frac{1}{2}y^T Ay - y^T b$:

$$\begin{aligned}\|e^{(k)}\|_A &= (e^{(k)})^T Ae^{(k)} = (x - x^{(k)})^T A(x - x^{(k)}) \\ &= (x^{(k)})^T Ax^{(k)} - 2(x^{(k)})^T Ax + x^T Ax \\ &= (x^{(k)})^T Ax^{(k)} - 2(x^{(k)})^T x^T b \\ &= 2\phi(x^{(k)}) + \text{ constant}\end{aligned}$$

- At each step α_k is choosen to minimizes $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$
- Conjugated search directions $p^{(k)}$ give minimization over $\mathcal{K}_k$

Polynomial Approximation by CG

- $P_k = \{p: \ p \text{ is a polynomial of degree} \leqslant k, \ p(0) = 1\}$
- Find $p_k \in P_k$ such that

$$\|p_k(A)e^{(0)}\|_A = \text{minimum}. \tag{6}$$

Theorem 2.3

If the CG iteration has not yet converged, the problem (6) has a unique solution $p_k \in P_k$ and the iterate $x^{(k)}$ has error $e^{(k)} = p_k(A)e^{(0)}$ for this same polynomial p_k. Moreover

$$\frac{\|e^{(k)}\|_A}{\|e^{(0)}\|_A} = \inf_{p \in P_k} \frac{\|p(A)e^{(0)}\|_A}{\|e^{(0)}\|_A} \leqslant \inf_{p \in P_k} \max_{\lambda \in \Lambda(A)} |p(\lambda)|.$$

- Proof: It is clear that $x^{(k)} = q_{k-1}(A)b = q_{k-1}(A)Ax$ with q_{k-1} of degree $k-1$, Then $e^{(k)} = p_k(A)e^{(0)}$ with $p_k \in P_k$. The equality follows from Theorem 2.2; for the inequality, expand in eigenvectors of A and conclude the result. □

Rate of Convergence

- Exercise 2.8: Prove that, if A has only k distinct eigenvalues, the the CG method converges in at most k steps.

Theorem 2.4: Rate of convergence

The error $e^{(k)}$ at the k-th iteration (with $k < n$) is orthogonal to $p^{(j)}$, $j = 0, \dots, k-1$, and

$$\|e^{(k)}\|_A \leqslant \frac{2c^k}{1+c^{2k}}\|e^{(0)}\|_A \leqslant 2c^k\|e^{(0)}\|_A, \qquad \text{with } c = \frac{\sqrt{K_2(A)}-1}{\sqrt{K_2(A)}+1}.$$

- Note that
$$\frac{\sqrt{K_2(A)}-1}{\sqrt{K_2(A)}+1} \sim 1 - \frac{2}{\sqrt{K_2(A)}},$$
and the convergence to a specified tolerance can be expected in $\mathcal{O}(\sqrt{K_2(A)})$ iterations

Some Remarks

- CG was proposed by Hestenes and Stiefel in 1952 as a direct method
- For systems with matrices of large size, CG is usually employed as an iterative method
- The dependence of the error reduction factor on the condition number of the matrix is more favourable when compared with the steepest descent method
- We have derived only an upper bound for the error; the convergence may be faster

Preconditioned Conjugate Gradient (PCG) Method

- If P is SPD (preconditioning matrix)

$$P^{-1/2}AP^{-1/2}y = P^{-1/2}b, \quad y = P^{1/2}x$$

- Not explicitly require the computation of $P^{1/2}$ or $P^{-1/2}$
- Algorithm: Preconditioned Conjugate Gradient Method

$x^{(0)}$ and P given; $r^{(0)} = b - Ax^{(0)}$; $z^{(0)} = P^{-1}r^{(0)}$; $p^{(0)} = r^{(0)}$
for $k = 0, 1, \ldots$

$$\alpha_k = \frac{{p^{(k)}}^T r^{(k)}}{{p^{(k)}}^T A p^{(k)}} \quad \text{\% step lenght}$$
$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)} \quad \text{\% update solution}$$
$$r^{(k+1)} = r^{(k)} - \alpha_k A p^{(k)} \quad \text{\% update residual}$$
$$Pz^{(k+1)} = r^{(k+1)} \quad \text{\% update residual}$$
$$\beta_k = \frac{(Ap^{(k)})^T z^{(k+1)}}{(Ap^{(k)})^T p^{(k)}} \quad \text{\% improvement this step}$$
$$p^{(k+1)} = z^{(k+1)} - \beta_k p^{(k)} \quad \text{\% search direction}$$

until convergence

Homework Exercises

- Exercise 2.9: Let $A \in \mathbb{R}^{805\times 805}$ matrix with eigenvalues 1.00, 1.01, 1.02, ..., 8.98. 8.99, 9.00 and also 10, 12, 16, 24. How many steps CG must take to be sure of reducing of $\|e^{(0)}\|_A$ by a factor 10^6?

- Exercise 2.10: The CG is applied to a SPD matrix A with results $\|e^{(0)}\|_A = 1$, $\|e^{(10)}\|_A = 2 \times 2^{-10}$. Based solely on this data, what bound can you give for $K_2(A)$ and $\|e^{(20)}\|_A$?

- Exercise 2.11: Let $A \in \mathbb{R}^{100\times 100}$ tridiagonal SPD matrix with $1, 2, \ldots, 100$ on the diagonal and 1 on the sub/super-diagonals, and set $b = (1, 1, \ldots, 1)^T$. Write a program that takes 100 steps of CG and the steepest descent (SD) iterations to approximately solve $Ax = b$. Produce a plot with four curves: the computed residual $\|r^{(k)}\|_2$ for CG, the actual residual $\|b - Ax^{(k)}\|$ for CG, the residual $\|r^{(k)}\|_2$ for SD, and the estimate $2c^k$ of Theorem 2.4. Comment on the results.