AMA 505: Optimization Methods

Subject Lecturer: Ting Kei Pong

Lecture 10
Unconstrained Optimization
Conjugate gradient method

Conjugate gradient method

In this lecture, we focus on

$$\operatorname*{Minimize}_{x\in\mathbb{R}^n}\ f(x) := \frac{1}{2}x^TAx - b^Tx \tag{1}$$

for a given $A \succ 0$ and $b \in \mathbb{R}^n$.

- The function f is convex because $\nabla^2 f(x) = A \succ 0$.
- First-order optimality condition is given by $Ax = b$. Since $A \succ 0$ is invertible, f has a unique minimizer given by $A^{-1}b$.
- (1) arises when truncated Newton's methods are adopted.

$A^{-1}b$?

- Direct method for computing $A^{-1}b$ such as Gaussian elimination / Cholesky factorization + back substitution takes $O(n^3)$ flops.
- Use iterative methods with low per-iteration cost?
- Recall: Steepest descent with exact line search on f:

> **Steepest descent with exact line search**
> Start at $x^0 \in \mathbb{R}^n$. For each $k = 0, 1, 2, \ldots$,
> - ⋆ Set $d^k = b - Ax^k$.
> - ⋆ Pick α_k so that
>
> $$\alpha_k \in \operatorname{Arg\,min}\{f(x^k + \alpha d^k) : \ \alpha \geq 0\}. \tag{2}$$
>
> - ⋆ Set $x^{k+1} = x^k + \alpha_k d^k$.

Flops per iteration is $O(n^2)$, but can take lots of iterations.

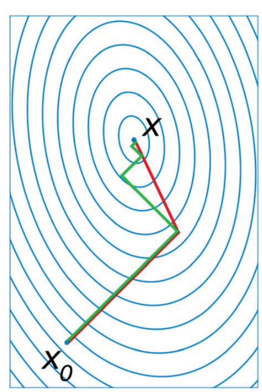
Note that the exact line search is well defined as long as $d^k \neq 0$, because $\psi(\alpha) := f(x^k + \alpha d^k)$ is a quadratic with leading coefficient $\frac{1}{2}{d^k}^T A d^k$. An explicit formula for α_k was discussed in Tutorial 1.

n cube : n 立方

Conjugate gradient method

We introduce the conjugate gradient method:

- Flops per iteration is $O(n^2)$;
- It converges in at most n steps;
- It keeps track of $O(1)$ vectors of dimension n per iteration.



- The red line shows the progress made by conjugate gradient method: it converges in 2 iterations.
- The green line shows the progress made by steepest descent with exact line search: it shows the signature zig-zag behavior.

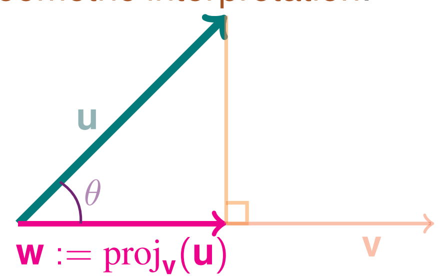
Idea: Modify the steepest descent direction to fit the (ellipse) geometry.

Review

Definition: (Projection) Let $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n \backslash \{0\}$. The projection of u onto v is defined as

$$\mathrm{proj}_v(u) := \frac{u^T v}{\|v\|_2^2} v.$$

Geometric interpretation:



1. Length of w equals

$$\|u\|_2 \cos\theta = \|u\|_2 \frac{u^T v}{\|u\|_2 \|v\|_2} = \frac{u^T v}{\|v\|_2}.$$

2. Unit vector along w is $\frac{v}{\|v\|_2}$.

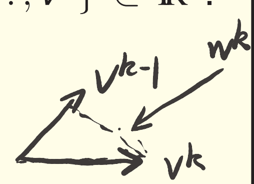
Thus,

$$\mathrm{proj}_v(u) = \frac{u^T v}{\|v\|_2} \cdot \frac{v}{\|v\|_2} = \frac{u^T v}{\|v\|_2^2} v.$$

Gram-Schmidt process

Theorem 10.1: (Gram-Schmidt process)
Given a set of linearly independent vectors $\{v^0, \ldots, v^k\} \subset \mathbb{R}^n$. Set $w^0 = v^0$ and for each $j = 1, \ldots, k$

$$w^k = v^k - \sum_{j=0}^{k-1} \frac{{v^k}^T w^j}{\|w^j\|_2^2} w^j$$



Then $w^i \neq 0$ for all i. Moreover, ${w^i}^T w^j = 0$ whenever $i \neq j$, and for each $i = 0, 1, \ldots, k$, it holds that

$$\mathrm{Span}\{v^0, \ldots, v^i\} = \mathrm{Span}\{w^0, \ldots, w^i\}.$$

Remark:

- If $\{v^0, \ldots, v^{k-1}\}$ is linearly independent but $\{v^0, \ldots, v^k\}$ is not, then $w^k = 0$. Indeed, in this case, $v^k = \sum_{i=0}^{k-1} \alpha_i w^i$ for some α_i. Multiplying both sides by $(w^j)^T$ for each $j = 0, \ldots, k-1$, we have $\alpha_i = {v^k}^T w^i / \|w^i\|_2^2$. Hence, $w^k = 0$.

seem as inner product $\langle v, w\rangle = v^T \cdot A \cdot w$

(Generalized) Gram-Schmidt process

Theorem 10.2: ((Generalized) Gram-Schmidt process)
Given $A \in \mathbb{R}^{n\times n}$ with $A \succ 0$ and a set of linearly independent vectors $\{v^0, \dots, v^k\} \subset \mathbb{R}^n$. Set $w^0 = v^0$ and for each $j = 1, \dots, k$

$$w^k = v^k - \sum_{j=0}^{k-1} \frac{{v^k}^T A w^j}{{w^j}^T A w^j} w^j$$

$\Rightarrow$ that means, $w^i \perp w^j$ in $\langle\cdot,\cdot\rangle$ inner product def

Then $w^i \neq 0$ for all i. Moreover, ${w^i}^T A w^j = 0$ whenever $i \neq j$, and for each $i = 0, 1, \dots, k$, it holds that

$$\mathrm{Span}\{v^0, \dots, v^i\} = \mathrm{Span}\{w^0, \dots, w^i\}.$$

Remark:

- If $\{v^0, \dots, v^{k-1}\}$ is linearly independent but $\{v^0, \dots, v^k\}$ is not, then $w^k = 0$. Indeed, in this case, $v^k = \sum_{i=0}^{k-1} \alpha_i w^i$ for some α_i. Multiplying both sides by $(Aw^j)^T$ for each $j = 0, \dots, k-1$, we have $\alpha_i = {v^k}^T A w^i / ({w^i}^T A w^i)$. Hence, $w^k = 0$.

$\mathrm{Min}\ f(x) = \frac{1}{2} x^T A x - bx$

Conjugate gradient method: Conceptual version

Conjugate gradient method: Conceptual version

see note

Start at $x^0 \in \mathbb{R}^n$ and $d^0 = -\nabla f(x^0) = b - Ax^0$.

For each $k = 0, 1, 2, \dots$,

- If $d^k = 0$, terminate.
- Pick α_k so that

$$\alpha_k \in \mathrm{Arg\,min}\{f(x^k + \alpha d^k) : \ \alpha \geq 0\}.$$

- Set $x^{k+1} = x^k + \alpha_k d^k$ and

$$d^{k+1} = -\nabla f(x^{k+1}) - \sum_{j=0}^{k} \frac{[-\nabla f(x^{k+1})]^T A d^j}{{d^j}^T A d^j} d^j$$

To do: upper to n orthogonal $\{d^k\}_{k=1}^n$.

- Prove the correctness: i.e., when $d^k = 0$, what happens?
- The update of d^{k+1} requires access to $O(k)$ vectors and k grows! — further simplify it?

Conjugate gradient method: Conceptual version cont.

Theorem 10.3:
Let $A \succ 0$ and $x^0 \in \mathbb{R}^n$. Set $d^0 = -\nabla f(x^0)$. For $k = 0, 1, \ldots,$ suppose that $d^0, \ldots, d^k \neq 0$, where for each $i = 0, \ldots, k-1$,

$$d^{i+1} = -\nabla f(x^{i+1}) - \sum_{j=0}^{i} \frac{[-\nabla f(x^{i+1})]^T A d^j}{{d^j}^T A d^j} d^j,$$

with $x^{i+1} = x^i + \alpha_i d^i$ and α_i coming from exact line search. Then $[\nabla f(x^j)]^T \nabla f(x^{k+1}) = 0$ and ${d^j}^T \nabla f(x^{k+1}) = 0$ for $j < k+1$.

Proof: The proof is by induction. Let $k = 0$ and $d^0 \neq 0$. Then $d^0 = -\nabla f(x^0)$, and ${d^0}^T \nabla f(x^1) = 0$ holds because of exact line search.

$\nabla f(x^0)^T \cdot \nabla f(x^1) = 0$

Conjugate gradient method: Conceptual version cont.

Proof of Theorem 10.3 cont.: Now, suppose that the claim is true for $k = m$, i.e., if $d^0, \ldots, d^m \neq 0$, then $[\nabla f(x^j)]^T \nabla f(x^{m+1}) = 0$ and ${d^j}^T \nabla f(x^{m+1}) = 0$ for $j < m+1$. We prove the claim for $k = m+1$.

Suppose that $d^0, \ldots, d^{m+1} \neq 0$. Then for each $j < m+1$, we have

$$\begin{aligned} {d^j}^T \nabla f(x^{m+2}) &= {d^j}^T [A(x^{m+1} + \alpha_{m+1} d^{m+1}) - b] \leftarrow \text{2, } \nabla f(x) \text{ def} \\ &= {d^j}^T (Ax^{m+1} - b) + \alpha_{m+1} {d^j}^T A d^{m+1} \leftarrow \text{GCS} \\ &= {d^j}^T (Ax^{m+1} - b) = {d^j}^T \nabla f(x^{m+1}) = 0. \end{aligned}$$

induction

Next, ${d^{m+1}}^T \nabla f(x^{m+2}) = 0$ follows from exact line search.

Finally, we also have $[\nabla f(x^j)]^T \nabla f(x^{m+2}) = 0$ for $j < m+2$ because

$$\text{span}\{d^0, \ldots, d^{m+1}\} = \text{span}\{\nabla f(x^0), \ldots, \nabla f(x^{m+1})\}.$$

This completes the induction argument.

${d^j}^T A d^m = 0 \Rightarrow {d^j}^T \cdot \nabla f(x^m) = 0 \Rightarrow \nabla f(x^j)^T \cdot \nabla f(x^m) = 0.$

Benchmark

Conjugate gradient method: Conceptual version cont.

- Rewrite d^{k+1}: Note that $\nabla f(x^{j+1}) - \overset{?}{\nabla} f(x^j) = \alpha_j A d^j$ for $j \le k$, and $\alpha_j > 0$. (Observe that if $\alpha_j = 0$, then $\nabla f(x^j) = d^j = 0$.) Thus

$$
\begin{aligned}
d^{k+1} &= -\nabla f(x^{k+1}) - \sum_{j=0}^{k} \frac{[-\nabla f(x^{k+1})]^T A d^j}{{d^j}^T A d^j} d^j \\
&= -\nabla f(x^{k+1}) + \sum_{j=0}^{k} \frac{\nabla f(x^{k+1})^T [\nabla f(x^{j+1}) - \nabla f(x^j)]}{{d^j}^T [\nabla f(x^{j+1}) - \nabla f(x^j)]} d^j \\
&= -\nabla f(x^{k+1}) + \frac{\nabla f(x^{k+1})^T [\nabla f(x^{k+1}) - \nabla f(x^k)]}{{d^k}^T [\nabla f(x^{k+1}) - \nabla f(x^k)]} d^k \\
&= -\nabla f(x^{k+1}) - \frac{\nabla f(x^{k+1})^T \nabla f(x^{k+1})}{{d^k}^T \nabla f(x^k)} d^k
\end{aligned}
$$

Conjugate gradient method: Conceptual version cont.

- Thus, we have $d^{k+1} = -\nabla f(x^{k+1}) - \frac{\nabla f(x^{k+1})^T \nabla f(x^{k+1})}{{d^k}^T \nabla f(x^k)} d^k$. On the other hand, using Theorem 10.3, we have for $k \ge 1$ that

$$
\begin{aligned}
[\nabla f(x^k)]^T d^k &= [\nabla f(x^k)]^T \left[-\nabla f(x^k) + \sum_{j=0}^{k-1} \frac{[\nabla f(x^k)]^T A d^j}{{d^j}^T A d^j} d^j \right] \\
&= -[\nabla f(x^k)]^T \nabla f(x^k) = -\|\nabla f(x^k)\|_2^2.
\end{aligned}
$$

The same formula also holds for $k = 0$. Consequently,

$$
\boxed{d^{k+1} = -\nabla f(x^{k+1}) + \frac{\|\nabla f(x^{k+1})\|_2^2}{\|\nabla f(x^k)\|_2^2} d^k.}
$$

Conjugate gradient method: Formal version

Conjugate gradient method: Formal version

Start at $x^0 \in \mathbb{R}^n$ and $d^0 = -\nabla f(x^0) = b - Ax^0$.

For each $k = 0, 1, 2, \ldots$,

- If $d^k = 0$, terminate.
- Pick α_k so that
$$\alpha_k \in \operatorname{Arg\,min}\{f(x^k + \alpha d^k) : \ \alpha \geq 0\}.$$
- Set $x^{k+1} = x^k + \alpha_k d^k$ and
$$d^{k+1} = -\nabla f(x^{k+1}) + \frac{\|\nabla f(x^{k+1})\|_2^2}{\|\nabla f(x^k)\|_2^2} d^k$$

Remark:

- Proof of correctness: when $d^{k+1} = 0$ but $d^k \neq 0$, then $\nabla f(x^{k+1})$ is a multiple of d^k. Since ${d^k}^T \nabla f(x^{k+1}) = 0$ according to Theorem 10.3, we conclude that $\nabla f(x^{k+1}) = 0$.
- The update of d^{k+1} requires access to only 3 vectors!

Finding α_k

$\alpha_k = \arg\min f(x^k + \alpha d^k)$

We next derive a formula for α_k.

- Recall that $f(x) = \frac{1}{2}x^TAx - b^Tx$.
- Hence, $\psi(\alpha) := f(x^k + \alpha d^k)$ is a quadratic with leading coefficient $\frac{1}{2}{d^k}^TAd^k > 0$.
- To find α_k, we compute
$$\frac{d}{d\alpha} f(x^k + \alpha d^k) = {d^k}^T \nabla f(x^k + \alpha d^k) = {d^k}^T(Ax^k + \alpha Ad^k - b).$$
Set the above to zero and solve for α, we obtain that
$$\alpha_k = \frac{{d^k}^T(b - Ax^k)}{{d^k}^TAd^k} = \frac{[-\nabla f(x^k)]^T d^k}{{d^k}^TAd^k} = \frac{\|\nabla f(x^k)\|_2^2}{{d^k}^TAd^k}.$$

See also Tutorial 1.

Conjugate gradient method: Actual version

Conjugate gradient method: Actual version

Start at $x^0 \in \mathbb{R}^n$ and $r^0 = d^0 = b - Ax^0$.

For each $k = 0, 1, 2, \ldots,$

- If $\|r^k\|$ (or, less commonly, $\|d^k\|$) is below a tolerance, terminate.
- (Exact line search) Compute

$$\alpha_k = \frac{{r^k}^T r^k}{{d^k}^T A d^k}, \quad x^{k+1} = x^k + \alpha_k d^k, \quad r^{k+1} = r^k - \alpha_k A d^k.$$

- (Update d^{k+1}) Compute

$$\alpha \quad \beta_k = \frac{{r^{k+1}}^T r^{k+1}}{{r^k}^T r^k}, \quad d^{k+1} = r^{k+1} + \beta_k d^k.$$

Remark: β_k not saved

- One matrix-vector multiplication per iteration if Ad^k is saved.
- Keeping track of four vectors, x^k, r^k, d^k and the Ad^k saved.

Convergence rate

Conjugate gradient method must terminate in at most n iterations: because $\{d^0, \ldots, d^{n-1}\}$ must be a basis of $\mathbb{R}^n$, if $d^k \neq 0$ for each k. Then necessarily, $d^n = 0$.

In practice, conjugate gradient method can converge much more quickly. We state the following result without proof. See Theorem 5.5 of Ref 2.

Theorem 10.4: (Luenberger)
Consider the conjugate gradient method for minimizing $f(x) = \frac{1}{2}x^T Ax - b^T x$ for some $b \in \mathbb{R}^n$ and $A \succ 0$. Let $\{x^k\}$ be the sequence generated and let x^* be the minimizer of f. If A has eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, then

$$(x^{k+1} - x^*)^T A(x^{k+1} - x^*) \leq \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right)^2 (x^0 - x^*)^T A(x^0 - x^*)$$

$\hookrightarrow \geq \lambda_1 \cdot \|x^{k+1} - x^*\|_2^2$

so if $\lambda_{n-k} - \lambda_1 = 0$, we have $x^{k+1} = x^*$.

Example

Example: Consider $\underset{x\in\mathbb{R}^n}{\text{Minimize}}\, f(x)$ with

$$f(x) = \frac{1}{2}x^T \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix} x - x_1 - 2x_2.$$

Perform two iterations of conjugate gradient method, starting with $x^0 = (2, 1)$. Write down x^1 and x^2.

Remark: Note that the matrix $\begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}$ is positive definite.

Solution: We have

$$r^0 = d^0 = b - Ax^0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -8 \\ -3 \end{bmatrix}.$$

Example cont.

Solution cont.: We have

$$\alpha_0 = \frac{{r^0}^T r^0}{{d^0}^T A d^0} = \frac{73}{331}, \;\; x^1 = x^0 + \alpha_0 d^0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \frac{73}{331} \begin{bmatrix} -8 \\ -3 \end{bmatrix} = \begin{bmatrix} 0.2356 \\ 0.3384 \end{bmatrix}.$$

We can then compute the next residual.

$$r^1 = r^0 - \alpha_0 A d^0 = \begin{bmatrix} -8 \\ -3 \end{bmatrix} - \frac{73}{331} \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} -8 \\ -3 \end{bmatrix} = \begin{bmatrix} -0.2810 \\ 0.7492 \end{bmatrix}.$$

Next, we compute β_0:

$$\beta_0 = \frac{\|r^1\|_2^2}{\|r^0\|_2^2} = \frac{(-0.2810)^2 + (0.7492)^2}{(-8)^2 + (-3)^2} = 0.0088.$$

Example cont.

Solution cont.: Next, we compute d^1:

$$d^1 = r^1 + \beta_0 d^0 = \begin{bmatrix} -0.2810 \\ 0.7492 \end{bmatrix} + 0.0088 \begin{bmatrix} -8 \\ -3 \end{bmatrix} = \begin{bmatrix} -0.3511 \\ 0.7299 \end{bmatrix}.$$

Finally, we have

$$\alpha_1 = \frac{{r^1}^T r^1}{{d^1}^T A d^1} = \frac{(-0.2810)^2 + (0.7492)^2}{\begin{bmatrix} -0.3511 & 0.7299 \end{bmatrix} \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} -0.3511 \\ 0.7299 \end{bmatrix}} = 0.4122,$$

and

$$x^2 = x^1 + \alpha_1 d^1 = \begin{bmatrix} 0.2356 \\ 0.3384 \end{bmatrix} + 0.4122 \begin{bmatrix} -0.3511 \\ 0.7299 \end{bmatrix} = \begin{bmatrix} 0.0909 \\ 0.6364 \end{bmatrix}.$$

No exam

Nonlinear conjugate gradient method

There are extensions of conjugate gradient method for minimizing a general $f \in C^1(\mathbb{R}^n)$.

Nonlinear conjugate gradient method: Conceptual version

Start at $x^0 \in \mathbb{R}^n$ and $d^0 = -\nabla f(x^0)$.

For each $k = 0, 1, 2, \ldots$,

- If $\|d^k\|_2$ is small, terminate.
- Pick α_k judiciously. Set $x^{k+1} = x^k + \alpha_k d^k$ and

$$d^{k+1} = -\nabla f(x^{k+1}) + \frac{\|\nabla f(x^{k+1})\|_2^2}{\|\nabla f(x^k)\|_2^2} d^k$$

Remark:

- If α_k is not chosen carefully, d^k can fail to be a descent direction at x^k.
- The above choice of d^k is due to Fletcher and Reeves.

Nonlinear conjugate gradient method cont.

Remark cont.:

- If α_k is chosen according to exact line search and if d^k is a descent direction, then ${d^k}^T \nabla f(x^{k+1}) = 0$. Then d^{k+1} is a descent direction at x^{k+1} if x^{k+1} is nonstationary, because:

$$ {d^{k+1}}^T \nabla f(x^{k+1}) = -\|\nabla f(x^{k+1})\|_2^2 < 0. $$

However, exact line search can be hard to perform for general f.

- In practice, one chooses α_k to satisfy strong Wolfe conditions:

Strong Wolfe conditions:
Let $0 < c_1 < c_2 < \frac{1}{2}$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. Find $\alpha > 0$ so that

$$ f(x + \alpha d) \leq f(x) + \alpha c_1 [\nabla f(x)]^T d, $$
$$ |[\nabla f(x + \alpha d)]^T d| \leq c_2 |[\nabla f(x)]^T d|. $$

The nonvoidness can be proved similarly as Theorem 3.3.