Lecture 12: Conjugate Gradient Methods

Yudong Chen

Given a symmetric *positive definite* (PD) matrix A, we want to minimize

$$f(x) = \frac{1}{2} x^\top A x - b^\top x.$$

We have $\nabla f(x) = Ax - b$ and $\nabla^2 f(x) = A$. Since $0 \prec A \preccurlyeq \lambda_{\max}(A) I$, f is convex and $\lambda_{\max}(A)$-smooth, and the global minimizer is $\arg\min_x f(x) = x^* = A^{-1} b$.

Example 1. A special case of the above problem is the linear least squares problem

$$f(x) = \frac{1}{2} \|Mx - c\|_2^2 = \frac{1}{2} x^\top \underbrace{M^\top M}_{A} x - (\underbrace{M^\top c}_{b})^\top x + \frac{1}{2} \|c\|_2^2 .$$

Example 2. Minimizing f above is equivalent to solving the linear system

$$Ax = b$$

with symmetric positive definite A. This problem arises in many applications. One example is when $A = \nabla^2 g(z)$ and $b = \nabla g(z)$, so the solution of the linear system is $\left(\nabla^2 g(z)\right)^{-1} \nabla g(z)$, which is the search direction at point z of Newton's method applied to minimizing g. Other examples include A being a covariance matrix or a graph Laplacian matrix.

Question 1. *Why not just compute A^{-1} and use the formula $x^* = A^{-1} b$ to compute the minimizer?*

1 First-order methods and Krylov subspace

(In this section, x_k denotes the iterate of an arbitrary first-order method.)

Consider first order methods for which each iterate x_k lies in the affine subspace

$$x_0 + \text{Lin}\left\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\right\};$$

explicitly,

$$x_k = x_0 - \sum_{i=0}^{k-1} h_{i,k} \nabla f(x_i), \tag{1}$$

where $h_{i,k} \in \mathbb{R}, \forall i, k$. Both GD and AGD take the form (1).

For quadratic f, thanks to the expression $\nabla f(x) = Ax - b = A(x - x^*)$ for the gradient, we have the following.

Lemma 1. *For the quadratic function $f(x) = \frac{1}{2} x^\top A x - b^\top x$ and all $k \geq 0$, we have*

$$x_k \in x_0 + \text{Lin}\left\{A(x_0 - x^*), A^2(x_0 - x^*), \dots, A^k(x_0 - x^*)\right\}$$

Proof. We prove by induction on k. Base case $k = 0$ is trivially true. Suppose

$$x_i - x_0 \in \text{Lin}\left\{A(x_0 - x^*), A^2(x_0 - x^*), \ldots, A^i(x_0 - x^*)\right\}, \qquad \forall i \leq k.$$

It follows that

$$\begin{aligned}\nabla f(x_i) &= A\,(x_i - x^*) \\ &\in \text{Lin}\left\{A(x_0 - x^*), A^2(x_0 - x^*), \ldots, A^{i+1}(x_0 - x^*)\right\}, \qquad \forall i \leq k.\end{aligned}$$

Hence

$$\begin{aligned}x_{k+1} - x_0 &\in \text{Lin}\left\{\nabla f(x_0), \ldots, \nabla f(x_k)\right\} \\ &\subseteq \text{Lin}\left\{A(x_0 - x^*), A^2(x_0 - x^*), \ldots, A^{k+1}(x_0 - x^*)\right\}. \end{aligned} \tag{2}$$

□

Definition 1. The linear subspace

$$\mathcal{K}_k := \text{Lin}\left\{A(x_0 - x^*), A^2(x_0 - x^*), \ldots, A^k(x_0 - x^*)\right\}$$

is called the *Krylov subspace* of order k.

Lemma 1 says all first-order methods in the form (1) satisfy

$$x_k \in x_0 + \mathcal{K}_k, \forall k.$$

2 Conjugate gradient methods

(In this section, x_k denotes the iterate of the CG method specifically.)

The conjugate gradient (CG) method is given by

$$x_k = \arg\min_{x \in x_0 + \mathcal{K}_k} f(x), \qquad k = 1, 2, \ldots$$

By definition, for quadratic f, CG converges at least as fast as any first-order method, including Nesterov's AGD. Therefore, CG inherits the convergence guarantees for AGD: it outputs x_k such that $f(x_k) - f(x^*) \leq \epsilon$ in at most

$$O\left(\min\left\{\sqrt{\frac{L}{\epsilon}}\left\|x_0 - x^*\right\|_2, \sqrt{\frac{L}{m}}\log\frac{L\left\|x_0 - x^*\right\|_2^2}{\epsilon}\right\}\right) \text{ iterations,}$$

where $L = \lambda_{\max}(A)$ and $m = \lambda_{\min}(A) > 0$.

But we can say more.

2.1 Properties of CG

Lemma 2 (Lem 1.3.1 in Nesterov's book)**.** *For any $k \geq 1$, we have*

$$\mathcal{K}_k = \text{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}.$$

Proof. In equation (2) we already established $\text{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\} \subseteq \mathcal{K}_k$. It remains to prove the reverse inclusion.

We use induction on k. Suppose $\text{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\} \supseteq \mathcal{K}_k$. We want to show that $\text{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_k)\} \supseteq \mathcal{K}_{k+1}$.

Note that $x_{k-1} \in x_0 + \mathcal{K}_{k-1}$ can be expressed as

$$x_{k-1} = x_0 + \sum_{i=1}^{k-1} \beta_{i,k-1} A^i (x_0 - x^*).$$

Consider two cases:

- $\nabla f(x_{k-1}) = 0$. Hence

$$\begin{aligned} 0 = \nabla f(x_{k-1}) &= A(x_{k-1} - x^*) \\ &= \underbrace{A(x_0 - x^*) + \sum_{i=1}^{k-2} \beta_{i,k-1} A^{i+1}(x_0 - x^*)}_{\in \mathcal{K}_{k-1}} + \beta_{k-1,k-1} A^k (x_0 - x^*). \end{aligned}$$

  This means $A^k(x_0 - x^*) \in \mathcal{K}_{k-1}$ and $\mathcal{K}_k = \mathcal{K}_{k-1}$. In turn, $A^{k+1}(x_0 - x^*) \in \mathcal{K}_k$ and $\mathcal{K}_{k+1} = \mathcal{K}_k$. We conclude that $\text{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_k)\} \supseteq \mathcal{K}_k = \mathcal{K}_{k+1}$, where the first step follows from induction hypothesis.

- $\nabla f(x_{k-1}) \neq 0$. Then

$$\begin{aligned} \nabla f(x_k) &= A(x_0 - x^*) + \sum_{i=1}^{k} \beta_{i,k} A^{i+1}(x_0 - x^*) \\ &= \underbrace{A(x_0 - x^*) + \sum_{i=1}^{k-1} \beta_{i,k} A^{i+1}(x_0 - x^*)}_{\in \mathcal{K}_k} + \beta_{k,k} A^{k+1}(x_0 - x^*). \end{aligned}$$

  We claim that $\beta_{k,k} \neq 0$. Taking the claim as given, we have

$$\begin{aligned} \mathcal{K}_{k+1} &= \text{Lin}\left\{\mathcal{K}_k \cup A^{k+1}(x_0 - x^*)\right\} \\ &= \text{Lin}\{\mathcal{K}_k \cup \nabla f(x_k)\}. \\ &\subseteq \text{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1}), \nabla f(x_k)\}. \end{aligned}$$

  Proof of claim: If $\beta_{k,k} = 0$, then

$$x_k = x_0 + \sum_{i=1}^{k-1} \beta_{i,k} A^i (x_0 - x^*) \in x_0 + \mathcal{K}_{k-1},$$

  so

$$x_k = \arg\min_{x \in x_0 + \mathcal{K}_k} f(x) = \arg\min_{x \in x_0 + \mathcal{K}_{k-1}} f(x) = x_{k-1}.$$

Note that

$$x_{k-1} - \frac{1}{L}\nabla f(x_{k-1}) \in x_0 + \mathcal{K}_k,$$

hence

$$\begin{aligned} f(x_{k-1}) = f(x_k) &\le f\left(x_{k-1} - \frac{1}{L}\nabla f(x_{k-1})\right) \\ &\le f(x_{k-1}) - \frac{1}{2L}\left\|\nabla f(x_{k-1})\right\|_2^2. \qquad \text{Descent Lemma} \end{aligned}$$

Since $\nabla f(x_{k-1}) \neq 0$, we have $f(x_{k-1}) < f(x_{k-1})$, a contradiction.

□

Lemma 3 (Lem 1.3.2 in Nesterov's book)**.** *For any $0 \le i < k$, we have*

$$\langle \nabla f(x_k), \nabla f(x_i)\rangle = 0.$$

Proof. Define a function $\Phi : \mathbb{R}^k \to \mathbb{R}$ by

$$\Phi(\lambda) = f\Bigg(\underbrace{x_0 - \sum_{i=0}^{k-1} \lambda_i \nabla f(x_i)}_{\in x_0 + \mathcal{K}_k}\Bigg),$$

where $\lambda = (\lambda_0, \lambda_1, \lambda_2, \ldots, \lambda_{k-1})^\top \in \mathbb{R}^k$.

By specification of CG,

$$x_k = \arg\min_{x \in x_0 + \mathcal{K}_k} f(x).$$

This means $x_k = x_0 - \sum_{i=0}^{k-1} \lambda_i^* \nabla f(x_i)$ with

$$\lambda^* = \arg\min_{\lambda \in \mathbb{R}^k} \Phi(\lambda).$$

Therefore, for each i:

$$0 = \frac{\partial \Phi(\lambda)}{\partial \lambda_i}\Big|_{\lambda = \lambda^*} = \langle \nabla f(x_k), -\nabla f(x_i)\rangle.$$

□

Two immediate corollaries:

Corollary 1 (Cor 1.3.1 in Nesterov's book)**.** *CG finds $x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$ in at most d iterations.*

Proof. Lemma 3 says $\nabla f(x_0), \nabla f(x_1), \ldots$ are orthogonal to each other. But in $\mathbb{R}^d$, there cannot be more than d orthogonal non-zero vectors, so we must have $\nabla f(x_d) = 0$ and thus x_d is optimal. □

Corollary 2 (Cor 1.3.2 in Nesterov's book)**.** *$\forall p \in \mathcal{K}_k$, $\langle \nabla f(x_k), p\rangle = 0$.*

Proof. By Lemma 2, $p \in \mathcal{K}_k = \text{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}$. By Lemma 3, any linear combination of $\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}$ is orthogonal to $\nabla f(x_k)$. □

2.2 Why is CG called CG?

Definition 2. Two vectors $p, q \in \mathbb{R}^d$ are said to be conjugate w.r.t. a matrix $A \in \mathbb{R}^{d\times d}$ if $\langle Ap, q\rangle = q^\top Ap = 0$.

We can write the iteration of CG as

$$x_{k+1} = x_k - h_k p_k,$$

where h_k is the stepsize and p_k is the search direction. Later we will show that

$$\forall k \neq i : \quad \langle Ap_k, p_i\rangle = 0.$$

Nocedal-Wright: "Conjugate gradients is a misnomer. It is the search/descent directions that are conjugate, not the gradients."