Conjugate Grad Methods.

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

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

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 $(\nabla^2 g(z))^{-1}\nabla g(z)$, 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?*

Cost effictive. $A: n\times n$. $\simeq O(n^3)$ time

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

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

explicitly,

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

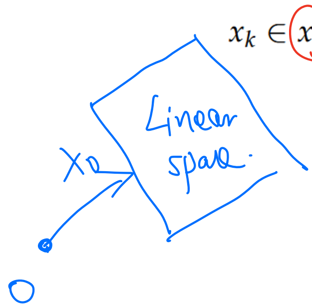
$Ax^* = b.$

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 $\boxed{\nabla f(x)} = Ax - b = \boxed{A(x - x^*)}$ for the gradient, we have the following.

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

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



Pf: By induction. $k=0$. √.

Suppose, $k=i$ $\quad X_i \in X_0 + \operatorname{Lin}\{A(X_0 - X^*), \dots, A^i(X_0 - X^*)\}.$

$\nabla f(x_i) = A(x_i - x^*) \in \mathrm{Lin}\{A(x_0-x^*), \cdots, A^{i+1}(x_0-x^*)\}.$

$x_{i+1} - x_0 \in \mathrm{Lin}\{\nabla f(x_0), \cdots, \nabla f(x_i)\}$
$\subseteq \mathrm{Lin}\{A(x_0-x^*), \cdots, A^{i+1}(x_0-x^*)\}.$ $\square$.

Definition 1. The linear subspace

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

is called the *Krylov subspace* of order k.

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

(for quadratic functions)

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

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 $f(x_k) - f(x^*) \le \epsilon$ in at most

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

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 $k \ge 1$, we have*

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

Pf: During pf of Lemma 1, we've shown $\mathrm{Lin}\{\nabla f(x_0), \cdots, \nabla f(x_{k-1})\} \subseteq \mathcal{K}_k$.
Let's show the reverse. Also by induction.

Suppose $\mathcal{K}_k \subseteq \mathrm{Lin}\{\nabla f(x_0), \cdots \nabla f(x_{k-1})\}$.

We use lemma I, $x_{k-1} \in x_0 + \mathcal{K}_{k-1}$

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

① $\nabla f(x_{k-1}) = 0$.

Then $0 = \nabla f(x_{k-1}) = A(x_{k-1} - x^*) = A(x_{k-1} - x_0 + x_0 - x^*)$

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

$$= \underbrace{A(x_0 - x^*) + \sum_{i=1}^{k-2} \beta_{i,k-1} A^{i+1} (x_0 - x^*)}_{\in \mathcal{K}_{k-1}} + \beta_{k-1,k-1} A^k (x_0 - x^*)$$

Hence, $A^k(x_0 - x^*) \in \mathcal{K}_{k-1} \Rightarrow \mathcal{K}_k = \mathcal{K}_{k-1}$.

In turn, $\mathcal{K}_{k+1} = \mathcal{K}_k$.

$$\therefore \mathcal{K}_{k+1} = \mathcal{K}_k \subseteq \mathrm{Lin}\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\} \subseteq \mathrm{Lin}\{\nabla f(x_0), \dots, \nabla f(x_k)\}.$$

② $\nabla f(x_{k-1}) \neq 0$.

$$\nabla f(x_k) = A(x_k - x^*) = A(x_k - x_0 + x_0 - x^*)$$

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

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

$\Rightarrow A^{k+1}(x_0 - x^*) \sim \nabla f(x_k)$

Claim: $\beta_{k,k} \neq 0$.

Take this claim, $\mathcal{K}_{k+1} = \{\mathcal{K}_k \cup A^{k+1}(x_0 - x^*)\}$

$$= \{\mathcal{K}_k \cup \nabla f(x_k)\}.$$

$$\subset \{\nabla f(x_0), \dots, \nabla f(x_k)\}.$$

Pf that $\beta_{k,k} \neq 0$: prove.

S.f.c., $\beta_{k,k} = 0$.

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

$$\therefore x_k = \operatorname*{argmin}_{x \in x_0 + \mathcal{K}_k} f(x) = \operatorname*{argmin}_{x \in x_0 + \mathcal{K}_{k-1}} f(x) = x_{k-1}.$$

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

Then
$$f(x_{k-1}) = f(x_k) \leq f\left(x_{k-1} - \frac{1}{L}\nabla f(x_{k-1})\right) \leq f(x_{k-1}) + \left\langle \nabla f(x_{k-1}), -\frac{1}{L}\nabla f(x_{k-1}) \right\rangle + \frac{L}{2}\left\| -\frac{1}{L}\nabla f(x_{k-1})\right\|_2^2$$
$$= -\frac{1}{L}\|\nabla f(x_{k-1})\|_2^2$$

$$\Rightarrow \|\nabla f(x_{k-1})\|^2 \leq 0 \quad \Rightarrow \nabla f(x_{k-1}) = 0. \quad \text{Contradicts!}$$

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

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

pf: Define $\Phi(\lambda) = f(\underbrace{x_0 - \sum_{i=0}^{k-1} \lambda_i \nabla f(x_i)}_{\in x_0 + \mathcal{K}_k})$ $\quad \lambda = (\lambda_0, \cdots, \lambda_{k-1})^T \in \mathbb{R}^k.$

CG: $x_k = \operatorname*{argmin}_{x \in x_0 + \mathcal{K}_k} f(x).$ $\quad$ Lemma 3, $\mathcal{K}_k = \mathrm{Lin}\{\nabla f(x_0), \cdots, \nabla f(x_{k-1})\}.$

Note that $\inf_{x \in x_0 + \mathcal{K}_k} f(x) = \inf_{\lambda \in \mathbb{R}^k} f(x_0 - \sum_{i=0}^{k-1} \lambda_i \nabla f(x_i))$

Hence $x_k = x_0 - \sum_{i=0}^{k-1} \lambda_i^* \nabla f(x_i)$ where $\lambda^* = (\lambda_0^*, \cdots, \lambda_{k-1}^*) = \operatorname*{argmin}_{\lambda \in \mathbb{R}^k} f(x_0 - \sum_{i=0}^{k-1} \lambda_i \nabla f(x_i))$
$$= \operatorname*{argmin}_{\lambda \in \mathbb{R}^k} \Phi(\lambda)$$

$\forall i,\quad \frac{\partial}{\partial \lambda_i} \Phi(\lambda^*) = 0. \quad \Phi(\lambda^*) = f(x_k)$

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

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* $\boxed{\textit{at most } d \textit{ iterations.}}$

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