* (Linear) Least Squares:

Given a symmetric PSD matrix A, we want to minimize

$$f(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle$$

$$\nabla f(x) = \underline{Ax - b}$$

$$\nabla^2 f(x) = A \succeq 0 \Rightarrow f \text{ is convex}$$

if $\lambda_{max}(A)$ is the max eval of A, then $f(x)$ is $\lambda_{max}(A)$ - smooth

if A^+ is the Moore-Penrose pseudoinverse of A, then:

$$\forall x : \| \nabla f(x) \|_2 \geq \| \nabla f(A^+ b) \|_2 .$$

→ If the system $Ax = b$ is solvable, then

$$x^* = A^+ b \in \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} f(x) .$$

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \underbrace{\nabla^2 f(x + \gamma(y-x))}_{A} (y-x), y-x \rangle, \quad \gamma \in (0,1) .$$

$$= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle A(y-x), y-x \rangle$$

→ Use this when computing the exact L.s. step size in HW#3.

* Other forms of linear least square problems:

$$\tilde{f}(x) = \frac{1}{2} \| Mx - c \|_2^2 = \frac{1}{2} \langle M^T M x, x \rangle - \langle M^T c, x \rangle + \frac{1}{2} \| c \|_2^2 .$$

$$A = M^T M , \quad b = M^T c$$

$$\underset{x}{\operatorname{argmin}} \tilde{f}(x) = \underset{x}{\operatorname{argmin}} \{ \underbrace{\frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle}_{f(x)} \}$$

$$\tilde{f}(x) - \tilde{f}(x^*) = f(x) - f(x^*) .$$

* <u>Method of Conjugate Gradients:</u>

* Here, we will take A to be symmetric & <u>PD</u>.
* Consider methods of the form:

$$(*) \quad x_k = x_0 - \sum_{i=0}^{k-1} h_{i,k} \nabla f(x_i), \text{ where } h_{i,k} \in \mathbb{R}$$

Both GD and AGD take the form $(*)$.

* As A symm. & PD, it is invertible. $x^* = A^{-1} b = \arg\min_x f(x)$.

x_k defined by $(*)$ is from $x_0 + \text{Lin}\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\}$.

$$\nabla f(x_0) = Ax_0 - b = A(x_0 - x^*), \quad \text{as } b = Ax^*$$

$$\begin{aligned}\nabla f(x_1) &= Ax_1 - b \\ &= A(x_0 - h_{0,1} \nabla f(x_0)) - Ax^* \\ &= A(x_0 - x^*) - h_{0,1} A^2 (x_0 - x^*) \in \text{Lin}\{A(x_0 - x^*), A^2(x_0 - x^*)\}\end{aligned}$$

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

Claim $x_{k+1} \in x_0 + \text{Lin}\{A(x_0 - x^*), \dots, A^{k+1}(x_0 - x^*)\}$

$$x_{k+1} = x_0 - \sum_{i=0}^{k} h_{i,k+1} \nabla f(x_i) = \underbrace{x_0 - \sum_{i=0}^{k-1} h_{i,k+1} \nabla f(x_i)}_{\in x_0 + \text{Lin}\{A(x_0 - x^*), \dots, A^k(x_0 - x^*)\}} - h_{k,k+1} \nabla f(x_k)$$

$$\begin{aligned}\nabla f(x_k) &= A(x_k - x^*) \\ &= A\left(x_0 + \sum_{i=1}^{k} a_i A^i (x_0 - x^*) - x^*\right) \\ &= A(x_0 - x^*) + \sum_{i=1}^{k} a_i A^{i+1} (x_0 - x^*) \\ &\in \text{Lin}\{A(x_0 - x^*), \dots, A^{k+1}(x_0 - x^*)\}\end{aligned}$$

$$\mathcal{K}_k = \mathrm{Lin}\{A(x_0 - x^*), \dots, A^k(x_0 - x^*)\}$$ - Krylov subspace of order k

- Method of Conjugate Gradients:

(CG) $$x_k^{out} = \underset{x \in x_0 + \mathcal{K}_k}{\mathrm{argmin}}\ f(x)$$

* Lemma (1.3.1 in Nesterov's book) For any $k \geq 1$, we have $\mathcal{K}_k = \mathrm{Lin}\{\nabla f(x_0^{out}), \dots, \nabla f(x_{k-1}^{out})\}$.

Proof: By induction on k.

Base case: $k=1$

$$\nabla f(x_0) = A(x_0 - x^*) \Rightarrow \mathcal{K}_1 = \mathrm{Lin}\{A(x_0 - x^*)\} = \mathrm{Lin}\{\nabla f(x_0)\}.$$

Suppose the lemma holds for some $k \geq 1$.

Any point $x_k \in x_0 + \mathcal{K}_k$ can be expressed as:

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

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

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

$$\mathcal{K}_{k+1} = \mathrm{Lin}\{\mathcal{K}_k \cup A^{k+1}(x_0 - x^*)\} = \mathrm{Lin}\{\mathcal{K}_k \cup \nabla f(x_k)\} \ \blacksquare$$

$\Rightarrow$ CG outputs x_k^{out} s.t. $f(x_k^{out}) - f(x^*) \leq \varepsilon$ in at most

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

$L = \lambda_{max}(A)$, $m = \lambda_{min}(A)$.

* <u>Lemma (1.3.2 in Nes'18 book)</u> If x_k^{out} is generated by CG, then $\forall i < k$ $\langle \nabla f(x_k^{out}), \nabla f(x_i^{out}) \rangle = 0$.

<u>Proof</u>: Let $k > i$. Define:

$$\Phi(h_k) = f(\underbrace{x_0 - \sum_{i=0}^{k-1} h_{i,k} \nabla f(x_i^{out})}_{x_k^{out} \in x_0 + \mathcal{K}_k})$$

$$h_k = \underset{h \in \mathbb{R}^d}{\operatorname{argmin}} \Phi(h) \text{ then } x_k^{out} = x_0 - \sum_{i=0}^{k} h_{i,k} \nabla f(x_i^{out})$$
$$= \underset{x \in \mathcal{K}_k}{\operatorname{argmin}} f(x)$$

$$\frac{\partial \Phi(h_k)}{\partial h_{i,k}} = 0 = \langle \nabla f(x_k^{out}), -\nabla f(x_i^{out}) \rangle.$$

* <u>Corollary</u> CG finds $x^* = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} f(x)$ in at most d iterations.

* <u>Corollary</u>. $\forall p \in \mathcal{K}_k$, $\langle \nabla f(x_k^{out}), p \rangle = 0$.