Lemma 2 Suppose $\bar{x}$ minimizes f over $x^0 + \text{span}(s^1 \dots s^k)$ and s^{k+1} is A-conjugate to each s^i. Then

$$\hat{x} = \underset{x = \bar{x} + \alpha s^{k+1}}{\text{argmin}} f(x)$$ is the minimizer of f over $\text{span}(s^1, \dots, s^{k+1}) + x_0$.

Proof. Essential same as previous seperation we have seen. $\square$

This gives the Conjugate Gradient Method

Given $x_0 \in \mathbb{R}^d$, $s^0 = r^0 = b - Ax^0 = -\nabla f(x^0)$.

Iterate:

$$\alpha_i = \underset{\alpha}{\text{argmin}} f(x^i + \alpha s^i) \leftarrow \alpha_i = \frac{s^{iT}(b - Ax^i)}{\langle s^i, s^i \rangle_A}.$$

$$x^{i+1} = x^i + \alpha_i s^i$$

$$r^{i+1} = -\nabla f(x^{i+1}) = b - Ax^{i+1}$$

0 for all $j \neq i$

$$s^{i+1} = r^{i+1} - \sum_{j=1}^{i} \frac{\langle r^{i+1}, s^j \rangle_A}{\langle s^j, s^j \rangle_A} s^j. = r^{i+1} - \frac{\langle r^{i+1}, s^i \rangle_A}{\langle s^i, s^i \rangle_A} s^i.$$

Theorem The Conjugate Gradient Method has

1. $\text{span}(r^0 \dots r^k) = \text{span}(s^0, \dots, s^k)$ has dimension $k+1$.
2. x^{k+1} minimizes f over $x^0 + \text{span}(r^0, \dots, r^k)$.

Proof. 1. is Gram-Schmidt using Lemma 1 for independence.

2. is what Lemma 2 tells us. $\square$

Claim. For $j<i$, $\langle r^{i+1}, s^j \rangle_A = 0$.

Proof. Let $L = \text{span}(r^0, \dots, r^i) = \text{span}(s^0, \dots, s^i)$

Theorem ensures x^{i+1} minimizes f over $x^0 + L$.

By Lemma 1, $-\nabla f(x^{i+1}) = r^{i+1}$ is orthogonal to L.

$$\Rightarrow \quad r^{i+1\,T} r^j = 0 \quad \forall j \le i$$

Then
$$\begin{aligned}
\langle r^{i+1}, s^j \rangle_A &= r^{i+1\,T} A s^j \\
&= r^{i+1\,T} A (x^{j+1} - x^j) \\
&= r^{i+1\,T} \left((b - Ax^j) - (b - Ax^{j+1}) \right) \\
&= r^{i+1\,T} (r^j - r^{j+1}) \\
&= \underset{0}{\underset{\shortparallel}{r^{i+1\,T} r^j}} - \underset{0}{\underset{\shortparallel}{r^{i+1\,T} r^{j+1}}}
\end{aligned}$$

if $j<i$. $\square$

4. Convergence Rates

Recall Gradient Descent, we had $\delta_k = f(x_k) - f(x^*)$

$$\delta_k \le \left(1 - \frac{1}{\text{Cond}(A)} \right)^k \delta_0$$

where $\text{Cond}(A) = \dfrac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \left(= \dfrac{L}{\mu} \right)$

Conjugate Gradient does better (achieves optimal rate)

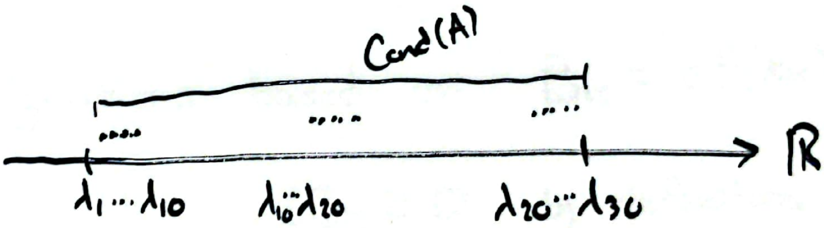
Theorem CGM has

$$\delta_K \le \left(\frac{\sqrt{\mathrm{Cond}(A)} - 1}{\sqrt{\mathrm{Cond}(A)} + 1} \right)^K \delta_0$$

$$\le \left(1 - \frac{1}{\sqrt{\mathrm{Cond}(A)}} \right)^K \delta_0 .$$

Proof. Spectral analysis of A. □

Faster guarantees for special arrangements of $\lambda_1, \dots, \lambda_n$



GMRES, Krylov Subspace find x_K minimizing over

$$K_K = \{ b, Ab, A^2 b, \dots, A^{K-1} b \}$$

(Caley-Hamilton Theorem $\Rightarrow A^{-1} b \in K_n$)

Conjugate Gradient (More Generally)

$\min f(x)$, $x_{K+1} = x_K - \alpha_K s_K$

$s_{K+1} = -\nabla f(x_{K+1}) + \beta_K s_K$

$$\frac{\nabla f(x_K)^T (\nabla f(x_K) - \nabla f(x_{K-1}))}{s_K^T (\nabla f(x_K) - \nabla f(x_{K-1}))}$$

Nonlinear Least Squares

$F(x) = 0 : \mathbb{R}^d \to \mathbb{R}^d$

$$\min \; f(x) = \frac{1}{2}\|F(x)\|_2^2$$

Recall HW4. $\nabla f(x) = \nabla F(x)^T F(x)$

$$\nabla^2 f(x) = \underbrace{\nabla F(x)^T \nabla F(x)}_{\text{pretty nice}} + \underbrace{\sum_{i=1}^{d} \nabla^2 F_i(x) F_i(x)}_{\text{terrible}}$$

In general, f is nonconvex, best we can do

$$\nabla f(x) = \nabla F(x)^T F(x) = 0$$

Two Algorithms based on $B_k = \nabla F(x_k)^T \nabla F(x_k)$.

($B_k \succeq 0$ by definition

Lets assume $B_k \succ 0$, ∇F be ind col).

($B_k \to \nabla^2 f(x^*)$ if $x_k \to x^*$)

1. Gauss-Newton Method
2. Levenberg-Marquardt Method (Trust-Region).

Each step of Gauss-Newton solves

$$p_k = \operatorname{argmin} \; \nabla f(x_k)^T p + \tfrac{1}{2} p^T B_k p$$

$$= \operatorname{argmin} \; \underline{F(x_k)^T \nabla F(x_k)} p + \tfrac{1}{2} p \, \underline{\nabla F(x_k)^T \nabla F(x_k)} p.$$

Solved by $\nabla F(x_k)^T \nabla F(x_k) \, p = - \nabla F(x_k)^T F(x_k)$.

$(B_k p = - \nabla f(x_k))$

$$= \operatorname{argmin} \; \tfrac{1}{2} \| \underbrace{F(x_k) + \nabla F(x_k) p}_{\text{Linearized } F(x_k + p)} \|_2^2$$

At each step we solve linear least squares problem.

(Conjugate Gradient Method good here)

If B_k has eigenvalues uniformly bounded away from 0 and ∞, then we have good descent. Old results $O(1/\sqrt{T})$ convergence.

If $x_k \to x^*$, then $B_k \to \nabla^2 f(x^*)$, then p_k "$\to$" Newton Step "$\to$" $\nabla^2 f(x^*)^{-1} \cdot \nabla f(x_k)$

(Superlinear rate)

Levenberg-Marquardt

$$s_k = \operatorname{argmin} \underbrace{\tfrac{1}{2} \| F(x_k) + \nabla F(x_k) s \|_2^2}_{m_k(s)} \qquad (1)$$
$$\text{s.t. } \|s_k\|_2 \leq \delta_k$$

Well-defined always, always descent (for small δ_k)

How to pick δ_k?

How to solve (1)?

(Trust Regions after break)