Lecture 23

Scribe?

Last time

- L-BFGS
- Conjugate gradient method

Today

- CG continued
- Convergence Guarantees
- Nonlinear least squares

Recall from last class

Lemma 7: Let x_0 and $s_1, \ldots, s_k$ be any vectors. Consider x_{k+1} given by $(*)$, then $\nabla f(x_{k+1})$ is orthogonal (in the standard sense) to

$$\text{span}\{s_1, \ldots, s_k\}.$$

Proof: Equivalently

$$y^* \in \underset{y \in \mathbb{R}^k}{\arg\min}\ f(x_0 + Sy)$$

By 1^{st}-order optimality conditions:

$$S^T \nabla f(\overbrace{x_0 + Sy^*}^{x_{k+1}}) = 0$$

$\Rightarrow$ $\nabla f(x_{k+1})$ is orthogonal to $\mathrm{span}\{s_1, \dots, s_k\}$.

$\square$

Thanks to separability:

Lemma ⋈: Supose that x_{k+1} is given by (★) and s_{k+1} is A-conjugate to each s_i. Then,

$$x_{k+2} \in \underset{\text{s.t. } x = x_{k+1} + \mathrm{span}\{s_{k+1}\}}{\mathrm{argmin}} f(x)$$

is also a solution of

$$x_{k+2} \in \underset{\text{s.t. } x = x_0 + \mathrm{span}\{s_1, \dots, s_{i+1}\}}{\mathrm{argmin}} f(x).$$

$\dashv$

CG Method

Input: $x_0 \in \mathbb{R}^d$, $s_0 = r_0 = b - Ax_0$

Update $i \leq d$:

$$\alpha_i = \underset{\alpha}{\mathrm{argmin}}\, f(x_i + \alpha s_i) \qquad \alpha_i = \frac{s_i^T (b - A x_i)}{\langle s_i, s_i \rangle_A}$$

$$x_{i+1} = x_i + \alpha_i s_i$$

$$r_{i+1} = -\nabla f(x_{i+1}) = b - A x_{i+1}$$

$$s_{i+1} = r_{i+1} - \sum P^A_{s_i}(r_{i+1}) \overset{\text{Lemma below}}{=} r_{i+1} - P^A_{s_i}(r_{i+1})$$

Gram-Schmidt

Theorem: The conjugate gradient method has

1. $\text{span}\{r_1, \dots, r_k\} = \text{span}\{s_0, \dots, s_k\}$.
2. x_{k+1} is given by $(\star)$.

Proof

1. Gram-Schmidt + Lemma ⚡ for independence.
2. Given by Lemma ⛝.

$\square$

CG simplifies a lot:

Lemma: For $j < i$, $\langle r_{i+1}, s_j \rangle_A = 0$.

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

The Theorem ensures that

x_{i+1} minimizes f over $x_0 + L$.

$\Rightarrow$ By Lemma ⚡, $-\nabla f(x_{i+1}) = r_{i+1}$ is orthogonal to L.

$\Rightarrow r_{i+1}^T r_j = 0 \quad \forall j \leq i$

Expanding

$$\begin{aligned}
\langle r_{i+1}, s_j \rangle_A &= r_{i+1}^T A s_j. \\
&= \frac{1}{\alpha_i} r_{i+1}^T A (x_{j+1} - x_j) \\
&= \frac{1}{\alpha_i} r_{i+1}^T \left((b - A x_j) - (b - A x_{j+1}) \right)
\end{aligned}$$

$$= \frac{1}{\alpha_i}\left(\underbrace{r_{i+1}^T r_j}_{=0} - \underbrace{r_{i+1}^T r_{j+1}}_{=0}\right)$$

$\leftarrow$ By Lemma 3

$\square$

This ensures that we don't need to make a lot of unnecessary matrix-vector multiplies.

Convergence guarantees

Recall that with GD we had

$$f(x_k) - \min f \le \left(1 - \frac{1}{K(A)}\right)^k (f(x_0) - \min f)$$

where $K(A) = \dfrac{\lambda_{max}(A)}{\lambda_{min}(A)} = \dfrac{L}{\mu}$

You proved in a HW.

AGD achieved a faster convergence rate with $\sqrt{K(A)}$ instead of cond(A).

CG does just as well (it's optimal)

Theorem: The iterates of CG satisfy

$$f(x_k) - \min f \le \left(\frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}\right)^k (f(x_0) - \min f)$$

$$\le \left(1 - \frac{1}{\sqrt{K(A)}}\right)^k (f(x_0) - \min f).$$

$\dashv$

We are not going to prove this result, as the proof involves some matrix analysis and uses Chebychev polynomials.

Remarks:

- The convergence is way better if $\kappa(A) \approx 1$. A natural idea is to precondition:

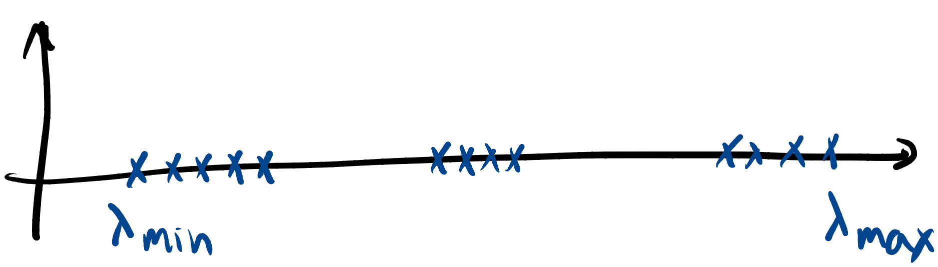
Invertible

$$P A P^T y = P b$$

$\Rightarrow$ $x = P^T y$ is a solution of $Ax = b$.

Active research area: How to come up with good preconditioners?

- For linear systems CG is often preferred over AGD. One reason is it offers faster convergence when eigenvalues are clustered, e.g.,



▷ How about asymmetric A?

GMRES is a popular algorithm that updates

$$x_{k+1} = \arg\min \tfrac{1}{2}\|Ax - b\|^2 \quad \text{s.t. } x \in x_0 + L_k$$

Computed via Arnoldi.

Krylov subspace.

where $L_{k+1} = \text{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^k r_0\}$.

By the Caley-Hamilton Theorem $A^{-1}b \in L_n$.

This was invented by Saad and Schultz in 1986.

Krylov subspace methods (CG, GMRES, ...) are one of the Top 10 algorithms of the past century (according to SIAM).

▷ How about extensions for general f?

They exist, but the guarantees and performance are not as strong; see Chapter 5.2 of Nocedal & Wright.

Non linear least squares

Assume we have a mapping

residual
$r : \mathbb{R}^d \to \mathbb{R}^n$ and our goal is to minimize

$$\min_x f(x) = \frac{1}{2}\|r(x)\|_2^2$$

From HW4, we have

$$\nabla f(x) = \nabla r(x)^T r(x)$$

$$\nabla^2 f(x) = \underbrace{\nabla r(x)^T \nabla r(x)}_{\text{cheap to compute}} + \underbrace{\sum_{i=1}^{n} \nabla^2 r_i(x)\, r_i(x)}_{\text{expensive}}$$

small near a solution

Our goal is to find a first-order critical point.

Two of the most popular first-order methods are

- Gauss-Newton
- Levenberg-Marquart Method.

Gauss-Newton method

Similarly to Newton pick a direction via

Well-defined if $B_k \succ 0$

$$p_k = \arg\min_p \nabla f(x_k)^T p + \frac{1}{2} p^T \overbrace{\nabla r(x_k)^T \nabla r(x_k)}^{B_k} p$$

$$\Updownarrow$$

$$\nabla r(x_k)^T \nabla r(x_k) p_k = -\nabla r(x_k) r(x_k)$$

$$\Updownarrow$$

$$p_k = \arg\min \frac{1}{2} \| \underbrace{r(x_k) + \nabla r(x_k) p}_{\text{linearization of } r(x_k+p)} \|_2^2$$

We will not show it, but if

$$\mu I \preceq B_k \preceq L I$$

Then, this method combined with appropriate step sizes has descent and globally converges.

When x_k is close to $x^* \Rightarrow B_k$ is close to $\nabla^2 f(x_k)$ and the method has superlinear convergence.

(Chapter 10 of Nocedal & Wright).

Question: What can we do when B_k is not positive definite?

Levenberg - Marquardt Method

Idea: Bypass the lack of unique solutions by adding a norm constraint:

$$(\because)\ p_{k+1} = \operatorname{argmin} \frac{1}{2} \| r(x_k) + \nabla r(x_k)^T p \|_2^2$$
$$\text{s.t.} \quad \|p\| \leq \Delta_k.$$

Trust-region

This prevents the need to pick α_k, but forces to pick Δ_k.

Q: How do we pick Δ_k?

Q: How do we solve $(\because)$?

We will cover trust-region after the break.