Lecture 22

Scribe?

HW 5 out today

Last time

- Convergence guarantees for BFGS.
- Proof

Today

- L-BFGS
- Conjugate gradient method

L-BFGS

With BFGS, we solved multiple problems

+ We have descent.
+ Local superlinear convergence
+ Only have to compute $\nabla f(x_k)$ per iter.
- However, we have a storage cost of $O(d^2)$. Thus it only works up to $d = 10^4 \sim 10^5$ (on personal computer).

To tackle higher sizes we can forget far away iterates.

In HW 5 you'll show that BFGS updates

$$B_k^{-1} = B_0^{-1} + \alpha_1 w_1 w_1^T + \dots + \alpha_{2k} w_{2k} w_{2k}^T$$

Instead of keeping all $2k$ vectors we can just keep that last m where $m \sim 2-30$ (usually).

This leads to $O(dm)$ memory when $B_0^{-1} = I$.

or any other simple to apply linear map.

Because

$$B_k^{-1} \nabla f(x_k) = \nabla f(x_k) + \sum_{j=k-m}^{k} \alpha_j (w_j^T \nabla f(x_k)) w_j.$$

Conjugate gradient

Gauss looking for planets

Today we go back to least squares

$$\min_x \underbrace{\tfrac{1}{2} x^T A x - b^T x}_{f(x) :=}$$

where $A \succ 0$ and $b \in \mathbb{R}^d$.

Optimality conditions say this is

equivalent to solving $Ax = b$.

We needed this for Newton.

What would happened if I knew $A = V \Lambda V^T$?

spectral decomposition

I could simplify the problem

change of basis

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} x^T A x - bx = \min_{y \in \mathbb{R}^d} \frac{1}{2} (Vy)^T A (Vy) - b^T V y$$

$$= \min_{y \in \mathbb{R}^d} \frac{1}{2} y^T \underbrace{\cancel{V^T V}}_{I} \Lambda \underbrace{\cancel{V^T V}}_{I} y - b^T V y$$

$$= \min_{y} \frac{1}{2} y^T \Lambda y - b^T V y$$

This is a separable problem

$$= \min \sum y_i^2 \lambda_i - b^T v_i y_i$$

$\Rightarrow$ Minimized at $y_i^* = \operatorname{argmin} \; y_i^2 \lambda_i - b^T v_i y_i$

$$= \frac{v_i^T b}{\lambda_i}$$

Minimized at $x^* = \sum^d v_i y_i^* = \sum^d \frac{v_i v_i^T b}{\lambda_i}$.

Computing the spectral decomposition is too expensive.

Question: Is there a cheaper way to obtain a separable problem?

Conjugate vectors

Any $A \succ 0$ defines an inner product

$$\langle x, y\rangle_A = x^T A y$$

$\langle x, x\rangle \geq 0 \ \forall x$
$\langle x, x\rangle = 0 \Leftrightarrow x = 0$
$\langle \cdot, z\rangle$ is linear
$\langle x, y\rangle = \langle y, x\rangle$
Real

Def: 1. Two vectors x & y A-conjugate if

$$\langle x, y\rangle_A = 0.$$

2. Given a linear subspace $L \subseteq \mathbb{R}^d$

$$L_A^\perp = \{ y \mid \langle x, y\rangle_A = 0 \quad \forall x \in L \}.$$

3 The projection of x onto y w.r.t $\langle \cdot, \cdot\rangle_A$ is

$$P_y^A(x) = \frac{\langle x, y\rangle_A}{\langle y, y\rangle_A} y.$$

Lemma: Let $s_1, \ldots, s_k$ be A-conjugate pairwise. Then, they are linearly independent and for all $x \in \text{span}\{s_i\}_{i=1}^k$

$$x = \sum_{i=1}^{K} P^A_{s_i}(x).$$

Note that if $s_1, \dots, s_d$ are A-conjugate then

$$S = \begin{pmatrix} s_1 & \cdots & s_d \end{pmatrix}$$

yields

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} x^T A x - bx \overset{\text{change of basis}}{=} \min_{y \in \mathbb{R}^d} \frac{1}{2} (Sy)^T A (Sy) - b^T S y$$

$$= \min \sum_{i,j} y_i y_j s_i^T A s_j - \sum_i b^T s_i y_i$$

Decomposable

$$= \min \sum_i y_i^2 s_i^T A s_i - b^T s_i y_i$$

$$\Rightarrow \quad y_i^* = \frac{b^T s_i}{s_i^T A s_i} \qquad \Rightarrow \quad x^* = \sum_{i=1}^{d} \frac{s_i s_i^T b}{s_i^T A s_i}.$$

Question: How do we obtain conjugate vectors?

Gram-Schmidt Orthogonalization

Input: $A \succ 0$, and linearly independent $x_1, \dots, x_k$
Output: $s_1, \dots, s_k$ A-conjugates s.t.
$$\text{span}\{s_i\} = \text{span}\{x_i\}.$$

- $s_1 = x_1$
- Recursively update
$$s_{i+1} = x_{i+1} - \sum_i P^A_{s_i}(x_{i+1})$$

Check:

- $\text{span}\{s_j\} = \text{span}\{x_j\}$
- $\langle s^{i+1}, s^j \rangle = 0 \quad \forall j < i+1$
- $s^{i+1} \neq 0$.

This algorithm is nice but when $k = d$, we have that we have to do d steps each with complexity $O(d^2) \Rightarrow O(d^3)$ complexity.

Matrix multiplication to compute $\langle x, s_i \rangle_A$.

Question: Can we find a good approximation of the solution of $Ax = b$ without doing $O(d^3)$ work?

Conjugate Gradient Method

Idea: Construct the basis $s_1, \dots, s_k$ using the residuals

$$r_k = b - Ax_k = \nabla f(x_k).$$

Then select

$$(\star) \quad x_{k+1} = \operatorname{argmin} f(x) \quad \text{s.t. } x \in x_0 + \operatorname{span}\{s_1, \dots, s_k\}.$$

Let us see two supporting results

Lemma 3: Let x_0 and $s_1, \dots, x_k$ be any vectors. Consider x_{k+1} given by $(\star)$, then $\nabla f(x_{k+1})$ is orthogonal (in the standard sense) to

$$\operatorname{span}\{s_1, \dots, s_k\}.$$

Proof: Equivalently

$$y^* \in \operatorname*{argmin}_{y \in \mathbb{R}^k} f(x_0 + Sy)$$

By 1st-order optimality conditions:

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

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

Thanks to separability:

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

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

is also a solution of

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

CG Method

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

Update $i \leq d$:

$$\alpha_i = \arg\min_{\alpha} f(x_i + \alpha s_i) \qquad \alpha_i = \frac{s_i^T(b - Ax_i)}{\langle s_i, s_i \rangle_A}$$

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

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

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

Gram-schmidt