Lecture 21

HW 4 due Today, Expect HW 5 soon.

Scribe?

Last time

- Rank 2 updates
- BFGS
- DFP

Today

- Convergence guarantees for BFGS.
- Proof

Quasi-Newton methods convergence guarantees

Analyzing the iterates directly is hard instead we show that the trajectories of the iterates are similar to those of other algorithms.

The guarantees we are about to see are weak, they only apply to strongly convex functions.

In practice, Quasi-Newtod method work well for most functions.

We will see two results:

Theorem: (Linear convergence) Let B_0 be a positive definite matrix and let $f:\mathbb{R}^d \to \mathbb{R}$ be a C^2 function, such that

As good as GD

$$\mu I \preceq \nabla f^2(x) \preceq L I.$$

μ-strongly convex

L-smooth

Then, the iterates of BFGS converge linearly $x_k \to x^*$.

⊣

Theorem: (Local superlinear convergence) Let B_0 be a PD matrix and $f:\mathbb{R}^d \to \mathbb{R}$ C^2, with a minimizer x^* with $\nabla^2 f(x^*) \succ 0$ and for all x, y near x^* we have

$$\| \nabla^2 f(x) - \nabla^2 f(y) \| \leq G \| x - y \|,$$

Then, if x_0 starts close enough to x^* we have that the iterates of BFGS converges superlinearly $x_k \to x^*$.

⊣

Today we focus on proving a weaker version of the first Theorem (we shall only prove convergence).

See Nocedal & Wright Section 6.4 for additional proofs.

The argument is based on the following result:

Theorem (   ) Let $f: \mathbb{R}^d \to \mathbb{R}$ be L-smooth. Consider an update

$$x_{k+1} \leftarrow x_k + \alpha_k p_k$$

where p_k is descent direction an α_k satisfies the Wolfe Conditions. Then,

$$\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x_k)\|^2 < \infty.$$

angle between p_k and $-\nabla f(x_k)$

Proof: By the second Wolfe cond.

$$(c-1)\nabla f(x_k)^T p_k \leq (\nabla f(x_{k+1}) - \nabla f(x_k))^T p_k$$

L-Lips

$x_{k+1} - x_k = \alpha_k p_k$

$$\le \alpha_k L \|p_k\|^2$$

Then,

$$\alpha_k \ge \frac{(c-1)}{L} \frac{\nabla f(x_k)^T p_k}{\|p_k\|^2}$$

Using the first Wolfe Condition

$$f(x_{k+1}) \le f(x_k) - \eta \frac{(1-c)}{L} \frac{(\nabla f(x_k)^T p_k)^2}{\|p_k\|^2}$$

$\cos\theta = \frac{\langle u, v\rangle}{\|u\|\|v\|}$

$$= f(x_k) - \eta \frac{(1-c)}{L} \cos^2\theta_k \|\nabla f(x_k)\|^2$$

Recursing

$$\le f(x_0) - \eta \frac{(1-c)}{L} \sum_{j=0}^{k} \cos^2\theta_j \|\nabla f(x_j)\|^2$$

Reordering

$$\sum^{k} \cos^2\theta_j \|\nabla f(x_j)\|^2 \le \frac{L}{\eta(1-c)} \left(f(x_0) - \min f\right)$$

Letting $k \uparrow \infty$, yields the result.

$\square$

Idea: If we show that $\cos^2\theta_k \geq \delta > 0$

$$\Rightarrow \liminf \|\nabla f(x_k)\|^2 \to 0.$$

This enough to have convergence for strongly convex functions since

Descent $\rightarrow$ $f(x_{k+1}) - \min f \leq f(x_k) - \min f$

$$\frac{\mu}{2}\left(f(x_{k+1}) - \min f\right) \leq \|\nabla f(x_{k+1})\|^2;$$

Error bound from midterm.

and $\frac{\mu}{2}\|x_k - x^*\|^2 \leq f(x_k) - \min f.$

Quadratic growth.

Proof: We focus on showing

$$\cos\theta_k^2 \geq \delta > 0.$$

where

$$\theta_k = \text{angle}\left(B_k^{-1}\nabla f(x_k), -\nabla f(x_k)\right)$$

Note that

$$s_{k+1} = -\alpha_k B_k^{-1} \nabla f(x_k)$$

Then

$$\text{angle}(s_{k+1}, B_k s_{k+1}) = \theta_k .$$

We will prove a boung using the relative entropy. Define

$$\psi(B) = \underbrace{tr(B)}_{\Sigma \lambda_i} - \underbrace{\log(\det(B))}_{\log(\Pi \lambda_i)}$$

One can show $\psi(B) > 0$ for $B \succ 0$.

Let's show that if $\cos \theta_k^2 \to 0$
$\Rightarrow \psi(B_k) < 0$ for large k.

Facts: ← Check!

$$tr(B_{k+1}) = tr(B_k) - \overbrace{\frac{\|B_k s_{k+1}\|^2}{s_{k+1}^T B_k s_{k+1}}}^{q_k / \cos^2\theta_k} + \overbrace{\frac{\|y_{k+1}\|^2}{y_{k+1}^T s_{k+1}}}^{L_k}$$

$$\det(B_{k+1}) = \det(B_k) \frac{y_{k+1}^T s_{k+1}}{s_{k+1}^T B_k s_{k+1}}$$

$$\frac{z_k^T G_k z_k}{\|z_k\|^2} \quad z_k = G_k^{1/2} s_{k+1}$$

$$\frac{s_{k+1}^T G_k s_{k+1}}{\|s_{k+1}\|^2} \qquad \frac{s_{k+1} G_k^2 s_{k+1}}{s_{k+1} G_k s_{k+1}}$$

We define

$$\mu_k = \frac{y_{k+1}^T s_{k+1}}{\|s_{k+1}\|^2}, \quad L_k = \frac{y_{k+1} y_{k+1}}{y_{k+1}^T s_{k+1}}$$

Then, $\mu \leq \mu_k$ and $L_k \leq L$. ($\star$)

Further define

$$q_k = \frac{s_{k+1} B_k s_{k+1}}{\|s_{k+1}\|^2}.$$

Then,

$$\det(B_{k+1}) = \det(B_k) \frac{\mu_k}{q_k}$$

and

$$\frac{\|B_k s_{k+1}\|^2}{s_{k+1}^T B_k s_{k+1}^T} = \frac{\|B_k s_{k+1}\|^2 \|s_{k+1}\|^2}{(s_{k+1}^T B_k s_{k+1})^2} \frac{(s_{k+1}^T B_k s_{k+1})}{\|s_{k+1}\|^2}$$

$$= \frac{q_k}{\cos^2 \theta_k}.$$

Thus,

$$\Psi(B_{k+1}) = tr(B_k) + L_k - \frac{q_k}{\cos^2\theta_k}$$

$$- \ln(\det B_k) - \ln q_k + \ln \mu_k$$

$$= \Psi(B_k) + (L_k - \ln \mu_k - 1)$$

$$+ \underbrace{\left[1 - \frac{q_k}{\cos^2\theta_k} + \ln \frac{q_k}{\cos^2\theta_k}\right]}_{1 - t + \ln(t) \leq 0 \quad \forall t>0} + \ln \cos^2\theta$$

$$\leq \Psi(B_k) + \overbrace{L - \ln \mu - 1}^{c}$$

$$+ \ln \cos^2\theta_k$$

$$\leq \Psi(B_0) + c(k+1) + \sum_{j=0}^{k} \ln \cos^2\theta_j$$

Assume seeking contradiction that $\cos^2\theta_j \to 0 \Rightarrow \ln \cos^2\theta_j \to -\infty$.

Let $k_0 > 0$ s.t $\forall j > k_0 \; \ln \cos^2\theta_j < -2c$.

Thus,

$$0 \leq \psi(B_k) \leq \psi(B_0) + c(k+1) + \sum_{j=0}^{k_0} \ln \cos^2\theta_j - \sum_{j=k_0+1}^{k} 2c$$

$$= \psi(B_0) + \sum_{j}^{k_0} \ln \cos^2\theta_j + 2ck_1 + c - ck.$$

$$< 0$$

For large k.

$\psi\downarrow$ $\square$

Let's prove $(\star)$, note that

$$y_{k+1} = \nabla f(x_{k-1}) - \nabla f(x_k)$$

$$= \int_0^1 \nabla^2 f(x_k + t(x_{k-1} - x_k))\,(x_k - x_{k-1})\,dt$$

$$= \underbrace{\left[\int_0^1 \nabla^2 f(x_k + t(x_{k-1} - x_k))\,dt\right]}_{G_k} s_{k+1}$$

Since G_k is an integral of Hessians

$\Rightarrow \quad \mu I \preceq G_k \preceq L I$

which implies $(\star)$.