Lecture 19 (Nov/2)

Scribe?

Last time

- Exam results.
- Modified Newton
- 3 variants

Today

- Convergence guarantees
- Computational concerns
- Secant method
- Quasi-Newton Methods

Convergence Guarantees

When $\nabla^2 f(x_k) \succeq \varepsilon I$, all the variants yield $B_k = \nabla^2 f(x_k)$. Thus, the template reduces to Newton's method.

So local quadratic convergence still holds (under strong convexity).

For global convergence we need a Descent Lemma.

Lemma: Suppose ∇f is L-Lipschitz and $x_{k+1} \leftarrow x_k - \alpha_k B_k^{-1} \nabla f(x_k)$ with $B_k \succ 0$. Then,

$$f(x_{k+1}) \leq f(x_k) - \left(\frac{\alpha}{\lambda_{max}(B_k)} - \frac{L\alpha^2}{2\lambda_{min}^2(B_k)} \right) \|\nabla f(x_k)\|^2.$$

Proof: HW 5.

This recovers the GD result when $B_k = I$.

Backtracking works.
Using this Lemma we derive

$$f(x_{k+1}) \le f(x_k) - c\,\|\nabla f(x_k)\|^2$$
$$\le f(x_0) - c\sum_{i=0}^{k} \|\nabla f(x_i)\|^2$$

which leads to the following result

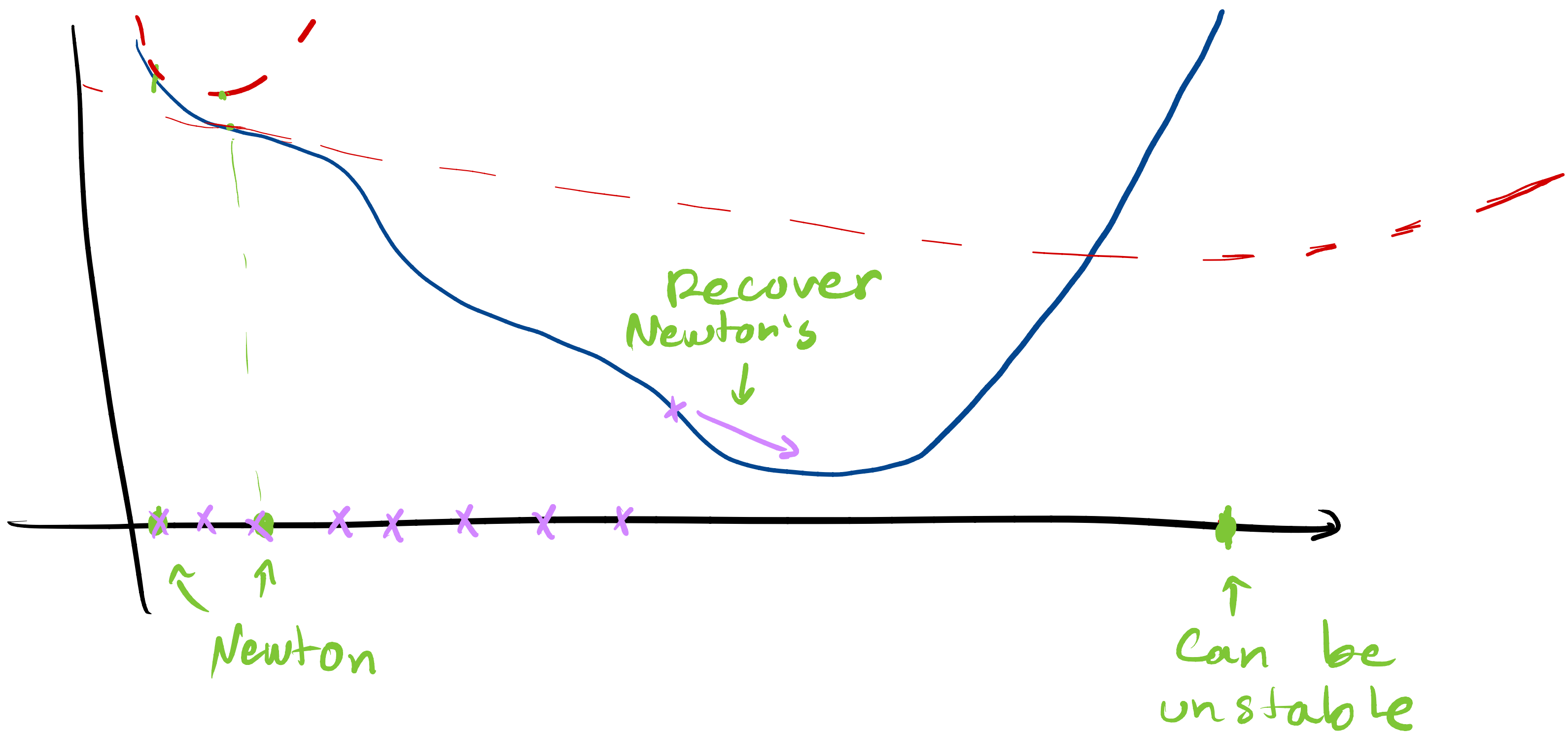
Theorem: If $f: \mathbb{R}^d \to \mathbb{R}$ has L-Lipschitz gradient and $\min f > 0$, and B_k has eigenvalues bounded away from 0 and ∞, then there exists a constant M s.t.

$$\min_{i \le k} \|\nabla f(x_i)\| \le \frac{M}{\sqrt{k}}$$

Proof: HW 5.

Intuition

Modified Newton converges globally, slowly, but if we approach a "strong" local minimum ($\nabla^2 f(x^*) \succeq \varepsilon I$) then, it recovers Newton's fast quadratic convergence.



Computational concerns (Again)

We still need to compute $\nabla^2 f(x)$, which consumes $O(d^3)$ when done directly.

We might also have bad conditioning. If $\nabla f(x_k)$ is singular

$\Rightarrow$ B_k has eigenvalues $O(\varepsilon)$

we have to be careful about ε.

Bad conditioning does appear in practice. For example when considering high degree polynomial systems:

In HW 4 we have

$$F(x) = \begin{pmatrix} (A - \lambda I)x \\ x^T x - 1 \end{pmatrix} = 0$$

Then $\|F(x)\|^2$ has degree 4.

As another simple example:

$$f(x,y) = x^2 + y^4$$

$$\Rightarrow \quad \nabla f(x,y) = \begin{bmatrix} 2x \\ 4y^3 \end{bmatrix} \quad \text{and} \quad \nabla^2 f(x,y) = \begin{bmatrix} 2 & \\ & 12y^2 \end{bmatrix}$$

As $y \to 0$, $\nabla^2 f(x,y) \to \begin{bmatrix} 2 & \\ & 0 \end{bmatrix}$.

Idea: Generalize the secant method

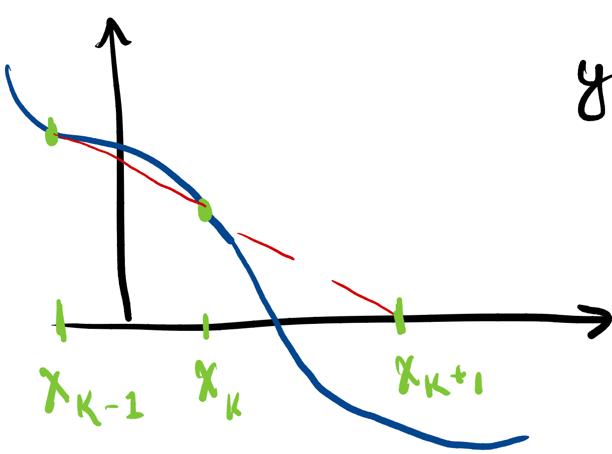
Just to remind you, the secant method finds a root of $F: \mathbb{R} \to \mathbb{R}$ by approximating

$$\nabla F(x_k) \approx \frac{F(x_k) - F(x_{k-1})}{\underbrace{x_k - x_{k-1}}_{B_k}}$$

and updating

$$x_{k+1} = x_k - \frac{F(x_k)}{B_k}.$$

It is locally superlinear yet not quadratic.



It avoids computing the Jacobian / Hessian.

Goal: Get these two features for $\mathbb{R}^d$. (Hopefully at a cost of $O(d^2)$).

No inverses.

The model build by the secant method "preserves" first order information at x_k and x_{k-1}:

$$m_k(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}\left(\frac{f'(x_k) - f'(x_{k-1})}{x_k - x_{k-1}}\right)(x - x_k)^2$$

Notice that both

$$m_k'(x_k) = f'(x_k) \quad \text{and} \quad m_k'(x_{k-1}) = f'(x_{k-1}).$$

Inspired by this, we want a method that satisfies

(1) B_k symmetric This is not super important.

(2) $m_k(x_k) = f(x_k)$, $\nabla m_k(x_k) = \nabla f(x_k)$

(3) $\nabla m_k(x_{k-1}) = \nabla f(x_k)$ ← Capture curvature

(4) $B_k \succ 0$

(5) Udpating and inverting B_k is cheap. $O(d^2)$

By (2) we have that

$$m_k(x) = f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T B_k (x - x_k).$$

Then, taking derivatives

$$\nabla m_k(x_k) = \nabla f(x_k) + B_k(x_{k-1} - x_k) \overset{(3)}{=} \nabla f(x_{k-1})$$

$$\Rightarrow B_k(\underbrace{x_{k-1} - x_k}_{s_k}) = \underbrace{\nabla f(x_k) - \nabla f(x_{k-1})}_{y_k}.$$

This gives $\frac{d(d+1)}{2}$ variables and d constraints (lots of solutions).

To satisfy (5) we need cheap updates B_k^{-1} from B_{k-1}^{-1}:

(5a) $B_k - B_{k-1}$ is rank one.

We leverage an important result.

<u>Lemma</u> (Sherman-Morrison) For any invertable $u, v \in \mathbb{R}^d$. If $v^T A^{-1} u \neq 1$, then $(A + uv^T)$ is invertable and

$$(A + uv^T)^{-1} = A^{-1} - \frac{(A^{-1}u)(A^{-T}v)^T}{1 + v^T A^{-1} u}.$$

$\dashv$

Proof: HW 5 (Woodbury Identity) $\square$

You'll prove a more general formula for rank r updates.

Update formulas for Quasi-Newton.

We assume (1)-(4), and (5a), then

$$B_k - B_{k-1} = \alpha w w^T$$

symmetry

$w \in \mathbb{R}^d$

$\alpha \in \mathbb{R}$

Because of (3)

$$B_{k+1} s_{k+1} = y_{k+1}$$

Let's consider two cases

Case 1 $\quad B_k s_{k+1} = y_{k+1} \Rightarrow w = 0.$ ✓

Case 2 $\quad B_k s_{k+1} \neq y_{k+1}$

$$\Rightarrow (B_k + \alpha w w^T) s_{k+1} = y_{k+1}$$

$$\Rightarrow -(\alpha w^T s_{k+1}) w = B_k s_{k+1} - y_{k+1}$$

$\alpha w^T s_{k+1} \neq 0$ $\Rightarrow \quad w = \beta (B_k s_{k+1} - y_k)$

Therefore,

$$B_{k+1} = B_k + \underbrace{\beta^2 \alpha}_{\gamma} (B_k s_{k+1} - y_k)(B_k s_{k+1} - y_k)^T$$

Then we obtain

$$B_k s_{k+1} + \gamma (B_k s_{k+1} - y_{k+1})(B_k s_{k+1} - y_{k+1}) s_{k+1} = y_{k+1}$$

$$\Rightarrow \left(1 + \gamma (B_k s_{k+1} - y_{k+1})^T s_{k+1}\right) B_k s_{k+1}$$

$$- \left(\underbrace{1 + \gamma (B_k s_{k+1} - y_{k+1})^T s_{k+1}}_{\text{need to make this zero}}\right) y_{k+1} = 0$$

$$\Rightarrow \quad \gamma = - \frac{1}{(B_k s_{k+1} - y_{k+1})^T s_{k+1}}$$

Thus

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

This is called the Symmetric Rank One update (SR1).

Big issue here: B_{k+1} might not be positive definite!