Lecture 20 (Nov/7)

HW 4 due in 2 days.

Scribe?

Last time

- Convergence guarantees
- Computational concerns
- Secant method
- Symmetric rank-1 update

Today

- Rank 2 updates
- BFGS
- DFP

Recap from last class

We wanted a modified Newton Method satisfying

(1) B_k symmetric

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

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

(4) $B_k \succ 0$

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

Last time we substituted

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

and derived

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

This is called the Symmetric Rank One update (SR 1).

Big issue: B_{k+1} might not be positive definite! $\Rightarrow$ No descent.

How can we overcome this issue?

Go a rank higher.

Rank-two updates

The idea is two consider

(5b) $B_k - B_{k-1}$ is rank two.

Recall that a symmetric matrix R is rank two if, and only if,

$$R = \alpha u u^T + \beta v v^T.$$

$\alpha, \beta \in \mathbb{R}$
$u, v \in \mathbb{R}^d$

Thus we have many more degrees of freedom! Yet we can still easily update B_k^{-1} due to the Woodbury identity.

BFGS

BFGS is a Quasi-Newton method invented by Broyden, Fletcher, Goldberg and Shanno in 1970.

Independently.

My academic great-grand father.

We can make a guess for u and v based on SRL:

Recall that $w = B_k s_{k+1} - y_{k+1}$

$$u = y_{k+1}$$
$$v = B_k s_{k+1}$$

Due to (3) we have that

$$(B_k + \alpha\, y_{k+1} y_{k+1}^T + \beta\, B_k s_{k+1} s_{k+1}^T B_k)\, s_{k+1} = y_{k+1}.$$

Reordering,

$$B_k s_{k+1} (1 + \beta (s_{k+1}^T B_k s_{k+1})) + y_{k+1} (\alpha\, y_{k+1}^T s_{k+1} - 1) = 0.$$

Thus, the equality holds if

$$1 + \beta\, s_{k+1}^T B_k s_{k+1}^T \Rightarrow \beta = \frac{-1}{s_{k+1}^T B_k s_{k+1}}$$

$$\alpha\, y_{k+1}^T s_{k+1} - 1 = 0 \Rightarrow \alpha = \frac{1}{y_{k+1}^T s_{k+1}},$$

which leads to the update

$$B_{k+1} = B_k + \frac{\overbrace{y_{k+1} y_{k+1}^T}^{\text{PSD}}}{y_{k+1}^T s_{k+1}} - \frac{\overbrace{B_k s_{k+1} s_{k+1}^T B_k}^{\text{PSD}}}{s_{k+1}^T B_k s_{k+1}}$$

Lemma: Assume $B_k \succ 0$ and $y_{k+1}^T s_{k+1} > 0$, then $B_{k+1} \succ 0$.

Proof: HW 5 $\square$

α is always positive if f is strongly-convex since

$$y_{k+1}^T s_{k+1} = (\nabla f(x_k) - \nabla f(x_{k-1}))^T (x_k - x_{k-1})$$
$$\geq \mu \|x_k - x_{k-1}\|^2.$$

But for general we would need to ensure that this holds. Note that if

$$\nabla f(x_k + \alpha p_k)^T p_k \geq \overset{c<1}{c}\, \nabla f(x_k)^T p_k$$

$$\Rightarrow \quad y_{k+1}^T s_{k+1} \geq \alpha \underbrace{(c-1)}_{<0\ (c<1)} \underbrace{\nabla f(x_k)^T p_k}_{<0\ (\text{Descent})}$$

$$\Rightarrow \quad y_{k+1}^T s_{k+1} > 0.$$

This motivates the so-called Wolfe conditions for line search: for some $\eta \in (0,1)$, $c \in (\eta, 1)$

Descent

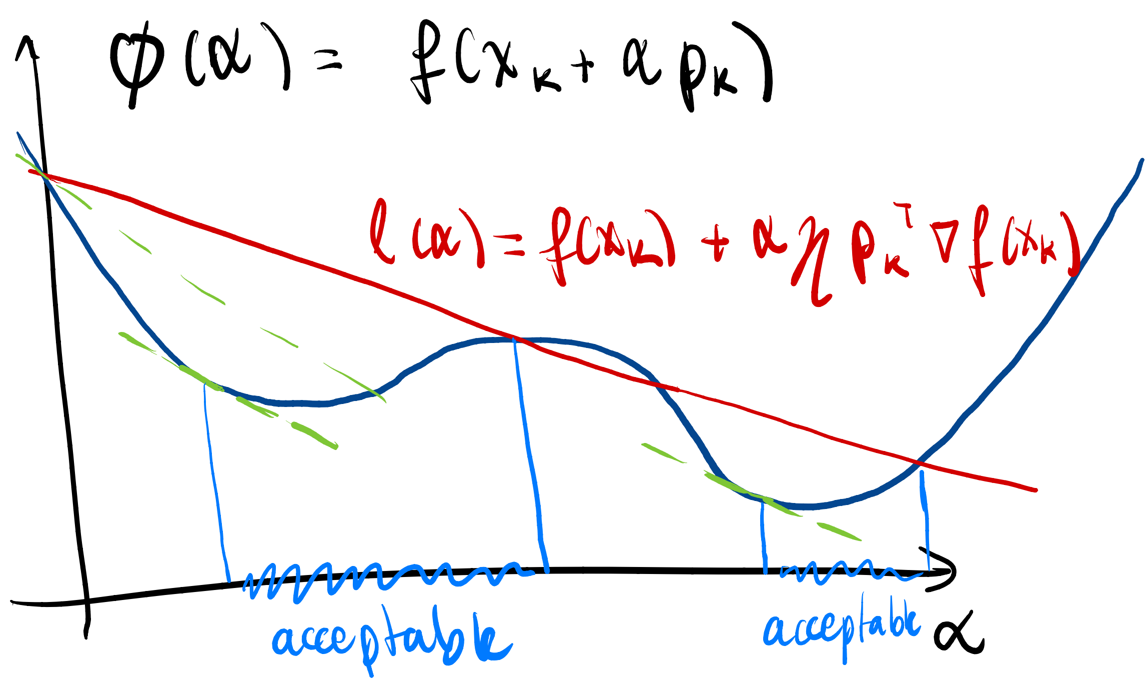
(1) $f(x_k + \alpha p_k) \leq f(x_k) - \eta \alpha p_k^T \nabla f(x_k)$

(2) $\nabla f(x_k + \alpha p_k)^T p_k \geq c \cdot \nabla f(x_k)^T p_k$.

Ensure $B_{k+1} \succ 0$

Lemma: There exist intervals satisfying the Wolfe Conditions for any C^1 function, bounded from below.

Intuition



Details left as exercise.

Davidon-Fletcher-Powell

Our choice of u and v were arbitrary, we could as well pick

$$\left.\begin{aligned} u &= B_k^{-1} y_k \\ v &= s_{k+1} \end{aligned}\right\}$$ Motivated by the update of B_k^{-1}.

Going through some algebra shows that

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

You'll show in HW5 that this update also preserves positive definiteness.

Thus there are infinitely many valid rank two updates.

What are the "best" rank two update?

Philosophy: We want to keep information from B_k, while satisfyin (3).

There are two interpretations.

▷ Relative entropy

The KL-divergence measures the similarity between distributions. We could minimize the KL-divergence of

Normal dist. → $N(0, B_{k+1})$ from $N(0, B_k)$.

This leads to the problem

$$\begin{aligned} \min_X \quad & \operatorname{tr}(B_k^{-1} X) - \log\det(B_k^{-1} X) \\ \text{s.t.} \quad & X \text{ symmetric} \\ & X s_{k+1} = y_{k+1}. \\ & X \succ 0 \end{aligned}$$

This is a constrained convex problem in X and one can show (using KKT conditions) that the minimizers recover BFGS.

The entropy is not symmetric
if we minimize $N(0, B_k)$ from
$N(0, B_{k+1})$, we obtain

$$\min_{X} \ tr(X^{-1} B_k) - \log\det(X^{-1} B_k)$$

s.t. X^{-1} symmetric ($\Leftrightarrow$ X sym)
$X^{-1} y_{k+1} = s_{k+1}$ ($\Leftrightarrow$ $X s_{k+1} = y_{k+1}$)
$X^{-1} \succ 0$ ($\Leftrightarrow$ $X \succ 0$)

This is convex in X^{-1} and the
optimal solution recovers DFP.

▷ Matrix norm

We could try to pick based on
a matrix norm

$$\min_{X} \ \| X - B_k \|$$

Any matrix norm

s.t. X is symmetric
(♡) $X s_{k+1} = y_{k+1}$
$X \succ 0$

Different matrix norms yield

different Quasi-Newton Methods.

If we pick $\|X\| := \|W^{1/2} X W^{1/2}\|_F$
where $W y_{k+1} = s_{k+1}$

Inverse of a matrix solving secant equation.

$\Rightarrow$ The DFP update minimizes (♡).

If we consider updating the inverse.

$$\begin{aligned} \min \quad & \|X^{-1} - B_k^{-1}\| \\ \text{s.t.} \quad & X \text{ sym} \\ & X \succeq 0 \\ & X s_{k+1} = y_{k+1}. \end{aligned}$$

Then this recovers BFGS

$$B^{-1} = (BFGS)^{-1}.$$