Week 3, Lecture 2

BFGS Formula for B_k, B_{k+1}:

$$B_{k+1} = B_k + \alpha y_k y_k^T + \beta (B_k s_k)(B_k s_k)^T$$

B_{k+1} satisfies the secant condition:

$$\beta = -\frac{1}{\langle s_k, B_k^T s_k \rangle}, \quad \alpha = \frac{1}{\langle y_k, s_k \rangle}$$

Theorem 4.2 (See Lecture Notes p.30)

<u>This theorem is examinable</u>.

Newton-type methods:

$$\boxed{p_k = -B_k^{-1} \nabla f(x_k)} \quad (1)$$

Avoiding having to compute B_k^{-1} explicitly here is the last step in BFGS. To avoid computing B_k^{-1} explicitly, we use the Sherman–Morrison–Woodbury Formula.

Background:

$$M = B + UV$$

where:

- $M \in \mathbb{R}^{n \times n}$
- $B \in \mathbb{R}^{n \times n}$
- $U \in \mathbb{R}^{n \times k}$
- $V \in \mathbb{R}^{k \times n}$

"Freshman's Dream"

$$\left.\begin{aligned} a &= b + c \\ \Rightarrow a^2 &= b^2 + c^2 \\ \Rightarrow \frac{1}{a} &= \frac{1}{b} + \frac{1}{c} \end{aligned}\right\} \quad \underline{\text{Wrong}}!$$

$$\frac{\sin\cancel{n}\, x}{\cancel{n}} = \text{six} = 6.$$

In the Sherman-Morrison-Woodbury formula, the Freshman's Dream comes true!

$$M^{-1} = B^{-1} + B^{-1}U\left(\mathbb{I}_k + VB^{-1}U\right)^{-1} VB^{-1}.$$

We apply this to BFGS:

$$B_{k+1}^{-1} = B_k^{-1} + \left(1 + \frac{\langle y_k, B_k^{-1} y_k \rangle}{\langle s_k, y_k \rangle}\right) \frac{s_k s_k^T}{\langle s_k, y_k \rangle}$$

$$- \frac{B_k^{-1} y_k s_k^T + s_k y_k^T B_k^{-1}}{\langle s_k, y_k \rangle} \quad (2)$$

The proof of (2) is in the lecture notes but it's heavy going – we leave it out of the course.

Hence, from (2), we compute B_{k+1}^{-1} directly from B_k^{-1}. Thus, we don't have to invert any matrices in ~~the~~ implementing the BFGS algorithm.

This reduces the operation count of BFGS down from $O(n^3)$ down to $O(n^2)$.

One last thing — a simple alternative to BFGS is the Barzilai – Borwein formula. (§ 4.3)

$$\underline{s}_k = \underline{x}_k - \underline{x}_{k-1}$$

$$\underline{y}_k = \nabla f(\underline{x}_k) - \nabla f(\underline{x}_{k-1})$$

Secant condition:

$$\underline{y}_k = B_k \underline{s}_k \quad (3)$$

We solve (3) approximately. In BFGS, (3) is solved approximately in a subspace of $\mathbb{R}^{n \times n}$ spanned by $\underline{y}_k \underline{y}_k^T$ and $(B_k \underline{s}_k)(B_k \underline{s}_k)^T$

Instead, in Barzilai – Borwein, (3) is solved approximately in a 1D subspace:

$$B_k \simeq \frac{1}{\alpha_k} \mathbb{I}_n$$

To fix α_k, we solve the secant condition in the least-squares sense: Hence, we have to minimize

$$\| B_k \underline{s}_k - \underline{y}_k \|_2^2 .$$

or $\| \overbrace{\frac{1}{\alpha_k} I_n}^{B_k} s_k - y_k \|_2^2$.

Hence,

$$\alpha_k = \arg\min_{\alpha > 0} \| \frac{1}{\alpha} s_k - y_k \|_2^2 . \quad (4)$$

There is an analytical solution to (4):
Introduce $\beta = 1/\alpha$, and let

$$\phi(\beta) = \| \beta s_k - y_k \|_2^2$$
$$= \langle \beta s_k - y_k , \beta s_k - y_k \rangle$$
$$= \beta^2 \langle s_k, s_k \rangle - 2\beta \langle s_k, y_k \rangle + \langle y_k, y_k \rangle .$$

Look at $\phi'(\beta)$:

$$\phi'(\beta) = 2\beta \langle s_k, s_k \rangle - 2 \langle s_k, y_k \rangle .$$

$$\phi'(\beta) = 0 \implies \beta = \frac{\langle s_k, y_k \rangle}{\langle s_k, s_k \rangle} .$$

But $\beta = 1/\alpha$. Hence, the α that minimizes (4) is:

$$\alpha = \frac{\| s_k \|_2^2}{\langle s_k, y_k \rangle} .$$

Exam

Update step:

$$\underline{x}_{k+1} = \underline{x}_k - B_k^{-1} \nabla f(\underline{x}_k)$$

$$= \underline{x}_k - \alpha_k \mathbb{I} \nabla f(\underline{x}_k)$$

$$\Rightarrow \boxed{\underline{x}_{k+1} = \underline{x}_k - \alpha_k \underline{\nabla} f(\underline{x}_k)} \quad (5)$$

$B_k = \frac{1}{\alpha_k} \mathbb{I}$.

$B_k^{-1} = \alpha_k \mathbb{I}^{-1}$

$= \alpha_k \mathbb{I}$.

With BB it is not guaranteed that $f(\underline{x}_{k+1}) < f(\underline{x}_k)$.

Summary so far:

- ~~Three techn.~~ Many techniques to compute the search direction p_k:
  - * SD
  - * Newton
  - * Quasi-Newton
    - — Secant Method
    - — BFGS
    - — BB
- We looked at computing the stepsize via:

$$\boxed{\alpha_k = \underset{\alpha > 0}{\operatorname{argmin}}\ f(\underline{x}_k + \alpha p_k) \quad (6)}$$

In the next chapter (Ch. 5) we will look at approximate solutions of (6) which nevertheless guarantee that our iterative method ($\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{p}_k$) converges.

Plan of next few lectures

- Methods for solving (6) approximately
- Convergence proofs next Tuesday
- Finish up convergence proof next Thursday
- Look at Exercises #1 next Thursday

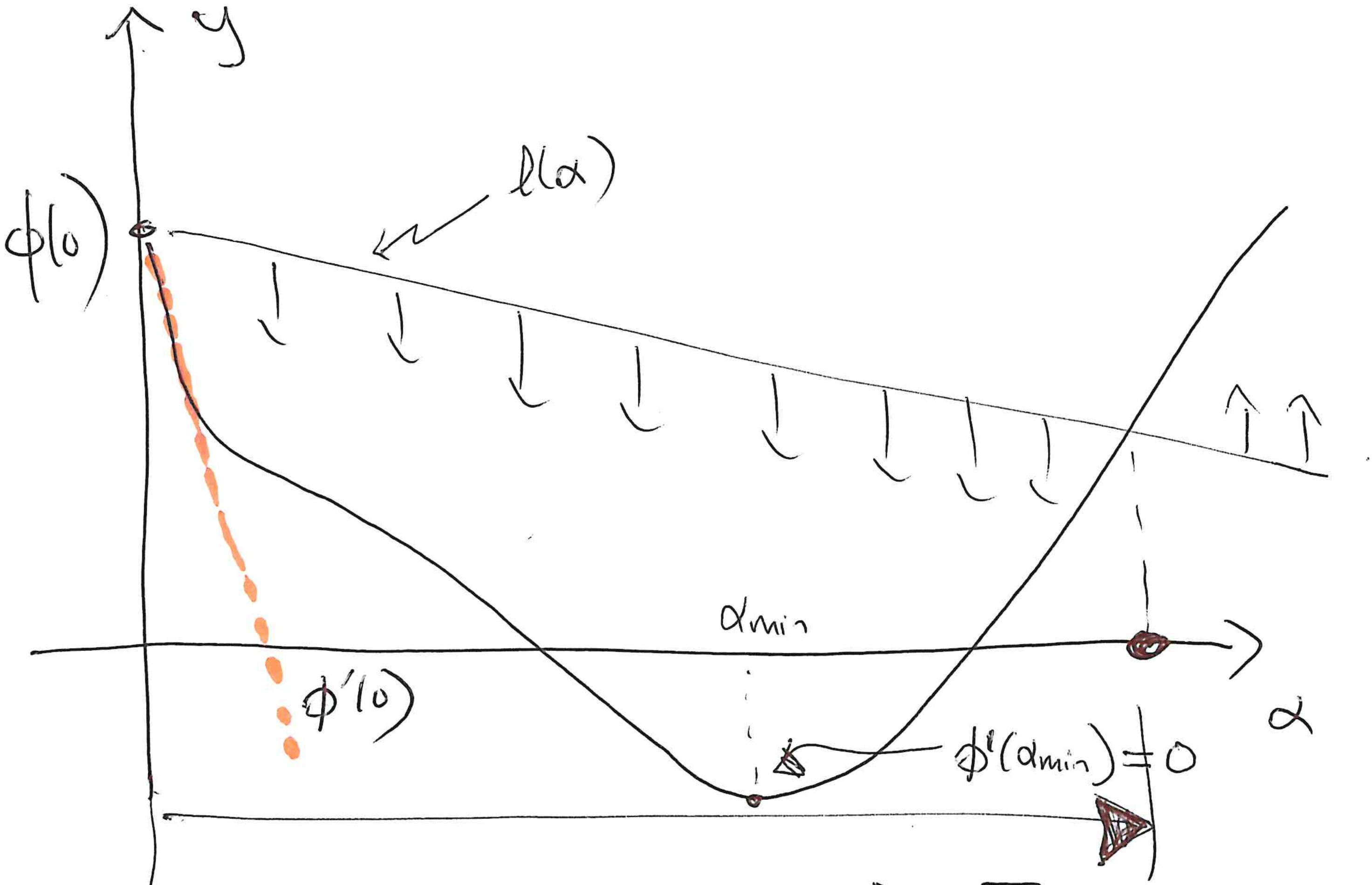
The Strong Wolfe Conditions (SWCs) give a satisfactory approximate solution to (6).

To see the rationale for the SWCs, we look again at the 1D subproblem

$$\phi(\alpha) = f(\underline{x}_k + \alpha \underline{p}_k).$$

The aim is to find an α that gets close to the minimum of ϕ.

To understand the upper bound first, we consider a plot of a typical function $\phi(\alpha)$:



We can't have α too large. To rule out α too large, we construct a line:

$$l(\alpha) = \phi(0) + c_1 \phi'(0)\alpha$$

where $0 < c_1 < 1$.

By searching for an approximate value of α_{min} in a region

$$\phi(\alpha) < l(\alpha),$$

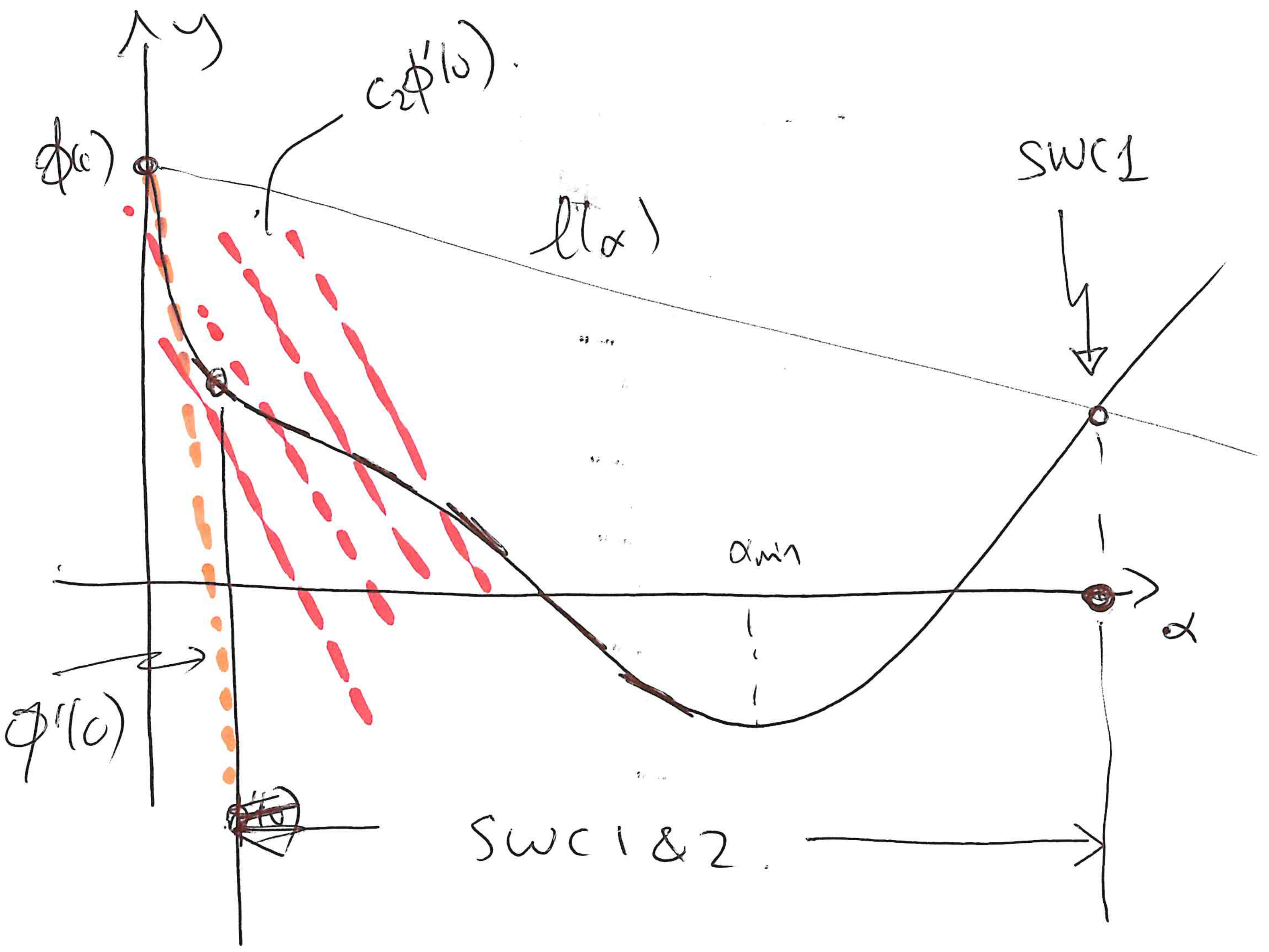
we rule out "α too large".

$$\boxed{\phi(\alpha) < \phi(0) + c_1 \phi'(0)\alpha} \quad \text{SWC 1.}$$

SWC 1 or the <u>Armijo Condition</u>.

By inspection of the plot, we see we need a second condition to rule out "α too small". A desirable value of α will have $\phi'(\alpha)$ small, meaning we are close to the min. Hence, we should avoid $|\phi'(\alpha)|$ being too large. The criterion is thus:

$$|\phi'(\alpha)| < c_2 |\phi'(0)| , \quad c_2 \in (0,1).$$



Take: $0 < c_1 < c_2 < 1$.

Theorem 5.1 Let $\phi(\alpha)$ be a continuously differentiable function which is bounded below, $\phi(\alpha) \geq \phi_{min}$. If $0 < c_1 < c_2 < 1$ then there exists an α satisfying the SWCs. Let $\phi'(0) < 0$.

Week 3, Lecture 3

Proof: Take $0 < c_1 < 1$. Introduce:

$$\Delta(\alpha) = \ell(\alpha) - \phi(\alpha)$$

$$= [\phi(0) + c_1 \alpha \phi'(0)] - \phi(\alpha).$$

$$\phi(\alpha) \geq \phi_m \Rightarrow -\phi(\alpha) \leq -\phi_m.$$

Hence,

$$\Delta(\alpha) \leq [\phi(0) + c_1 \alpha \underbrace{\phi'(0)}_{\text{neg.}}] - \phi_m.$$

Hence, $\boxed{\Delta(\alpha) \rightarrow -\infty \text{ as } \alpha \rightarrow \infty}$ I

Look also at:

$$\frac{d\Delta}{d\alpha} = \frac{dl}{d\alpha} - \phi'(\alpha)$$

$$= c_1 \phi'(0) - \phi'(\alpha).$$

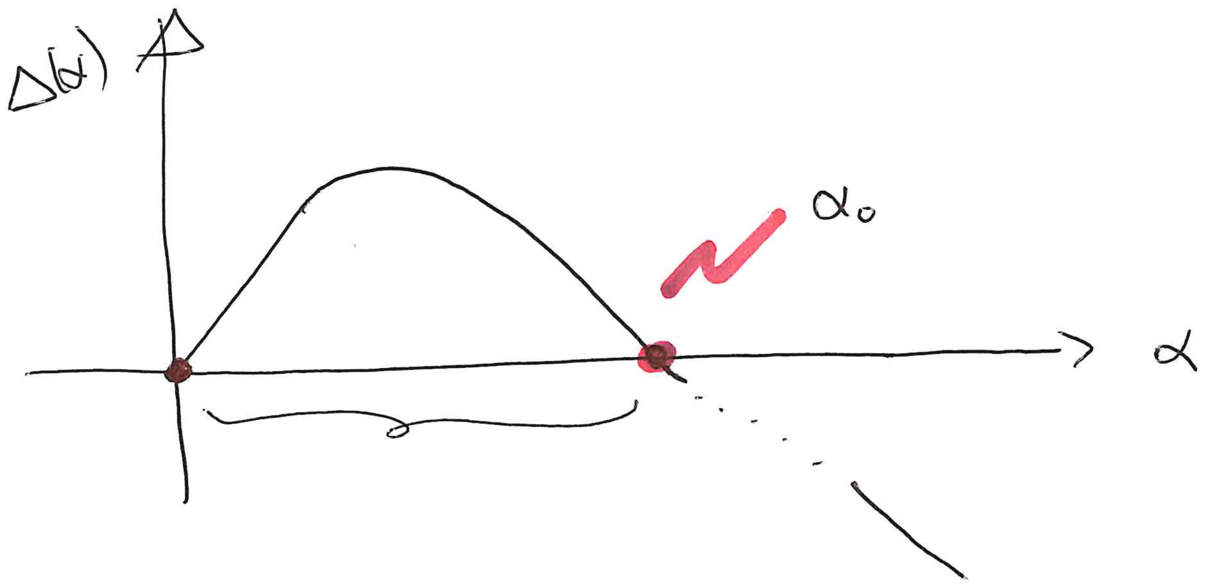
$$\left.\frac{d\Delta}{d\alpha}\right|_{\alpha=0} = c_1 \phi'(0) - \phi'(0)$$

$$= \underbrace{(c_1 - 1)}_{\text{neg}} \underbrace{\phi'(0)}_{\text{neg}}.$$

$$\therefore \boxed{\left.\frac{d\Delta}{d\alpha}\right|_{\alpha=0} > 0} \quad \text{II}.$$

$$\boxed{\Delta(\alpha=0) = l(0) - \phi(0) = 0} \quad \text{III}$$

These observations allow us to make a rough sketch of Δ as a function of α:



By continuity, there exists $\alpha_0 > 0$ such that $\Delta(\alpha_0) = 0$.

Hence, for any $\alpha \in (0, \alpha_0)$, we have $\Delta(\alpha) > 0$. But $\Delta^{(\alpha)} = \ell(\alpha) - \phi(\alpha)$. Hence,

$$\ell(\alpha) > \phi(\alpha) \quad \text{for all} \quad \alpha \in (0, \alpha_0).$$

Hence,

$$\boxed{\phi(\alpha) < \phi(0) + c_1 \phi'(0)\alpha \quad \text{for all} \quad \alpha \in (0, \alpha_0)} \quad \text{SWC 1}.$$

For the second part, we look again at α_0 and we have:

$$\Delta(\alpha_0) = 0 \quad \cancel{\Rightarrow \phi(\alpha_0) = \phi(0)}$$

Hence, $\ell(\alpha_0) - \phi(\alpha_0) = 0$, hence:

$$\phi(\alpha_0) = \phi(0) + \underline{c_1 \alpha_0 \phi'(0)}. \quad (7a)$$

Bring in Taylor's Theorem:

$$\phi(\alpha_0) = \phi(0) + \underline{\alpha_0 \phi'(\beta)}, \quad \beta \in (0, \alpha_0). \quad (7b)$$

Combine (7a) and (7b):

$$\cancel{\alpha_0} \left| c_1 \phi'(0) \right| = \cancel{\alpha_0} \left| \phi'(\beta) \right|, \quad \beta \in (0, \alpha_0).$$

$$\Longrightarrow \quad c_1 \left| \phi'(0) \right| = \left| \phi'(\beta) \right|, \quad \beta \in (0, \alpha_0).$$

$$\left| \phi'(\beta) \right| = c_1 \left| \phi'(0) \right| < c_2 \left| \phi'(0) \right| \quad c_2 > c_1.$$

Hence, SWC 2 is satisfied for $\beta \in (0, \alpha_0)$. $\blacksquare$

<u>Remark</u>: It's good to work with the SWC because we don't have to worry about the signs of derivatives. But a slightly less restrictive set of conditions can also be used to find an approximate value of α_{min}. These are the Wolfe conditions:

$$\phi(\alpha) \leq \phi(0) + c_1 \phi'(0) \alpha$$

$$\phi'(\alpha) > c_2 \phi'(0).$$

A final method to estimate α_{min} is with backtracking line search (§5.3)

It's quick and easy but does not have the same theoretical underpinning as the SWCs.

Choose an initial guess for α_k, call it α.
Fix $\rho \in (0,1)$ and $c \in (0,1)$.

while $f(\underline{x}_k + \alpha \underline{p}_k) > f(\underline{x}_k) + c\alpha \underline{p}_k \cdot \nabla f(\underline{x}_k)$ do

$\quad \alpha \leftarrow \rho\alpha$.

end while