Week 2, Lectures 2-3

Online lecture on Tuesday A.M. for the foreseeable future

Plan for today:

- Finish up looking at the quadratic model problem (Ch. 2)
- Line search Methods (Ch. 3)

$$f(\underline{x}) = c + \langle \underline{a}, \underline{x} \rangle + \frac{1}{2} \langle \underline{x}, B\underline{x} \rangle$$

When $B \in \mathbb{R}^{n \times n}$ is no longer positive-definite:

We still assume that B is symmetric.

Theorem: f attains a min. if and only if B is positive-semi-definite and $\underline{a}$ is in the range of B. If B is positive-semi-definite (PSD), then every $\underline{p}$ satisfying $B\underline{p} = -\underline{a}$ is a global minimizer of f.

Proof: Assume that B is P.S.D. and that $\underline{a}$ is in the range of B. Given this assumption, there exists an $\underline{x} \in \mathbb{R}^n$ such that:

$$B\underline{x} = -\underline{a}.$$

For any $\underline{w} \in \mathbb{R}^n$, consider:

$$f(\underline{x}+\underline{w}) = c + \langle \underline{a}, \underline{x}+\underline{w} \rangle + \frac{1}{2} \langle \underline{x}+\underline{w}, B(\underline{x}+\underline{w}) \rangle$$

B symmetric

$$= c + \langle \underline{a}, \underline{x} \rangle + \langle \underline{a}, \underline{w} \rangle + \frac{1}{2} \langle \underline{x}, B\underline{x} \rangle + 1 \langle \underline{x}, B\underline{w} \rangle + \frac{1}{2} \langle \underline{w}, B\underline{w} \rangle$$

$$= \underline{c + \langle \underline{a}, \underline{x} \rangle + \frac{1}{2} \langle \underline{x}, B\underline{x} \rangle} + \underline{\left[\langle \underline{a}, \underline{w} \rangle + \langle \underline{x}, B\underline{w} \rangle \right]} + \frac{1}{2} \langle \underline{w}, B\underline{w} \rangle$$

$$= f(\underline{x}) + \overbrace{\langle -B\underline{x}, \underline{w} \rangle + \langle \underline{x}, B\underline{w} \rangle}^{=0} + \frac{1}{2} \langle \underline{w}, B\underline{w} \rangle$$

$$\Rightarrow f(\underline{x}+\underline{w}) = f(\underline{x}) + \overbrace{\frac{1}{2} \langle \underline{w}, B\underline{w} \rangle}^{\geq 0}.$$

Since B is P.S.D.,

$$f(\underline{x}+\underline{w}) \geq f(\underline{x}) \quad \forall\, \underline{w} \in \mathbb{R}^n.$$

Hence, $\underline{x}$ is a global minimizer.

For the other way around, suppose that f ~~has~~ has a minimizer ($\underline{x}$, say).

By the first-order optimality condition,

$$\nabla f(\underline{x}) = 0 \implies B\underline{x} = -\underline{a}.$$

Hence, $\underline{a}$ is in the range of B.

By the second-order optimality,

$$\left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right|_{\underline{x}} \text{ is P.S.D.}$$

But for the Quadratic Model Problem, this is just B, hence B is P.S.D.

<u>Theorem</u>: f has a unique minimizer if and only if B is strictly P.D.

<u>Proof</u>: Assume that B is P.D. Then, B is invertible, so $\underline{a}$ is in the range of B, so let $\underline{x}$ solve $B\underline{x} = -\underline{a}$.

Consider:

$$f(\underline{x}+\underline{w}) = f(\underline{x}) + \overset{0}{\wedge} \tfrac{1}{2}\langle \underline{w}, B\underline{w}\rangle$$

But B is P.D. so $\langle \underline{w}, B\underline{w}\rangle > 0 \ \forall \underline{w} \neq 0$.

Hence $f(\underline{x}+\underline{w}) > f(\underline{x}) \ \forall \ \underline{w} \ (\neq 0) \in \mathbb{R}^n$.

So $\underline{x}$ is the unique global minimizer.

For the other way around, suppose f has a unique global minimizer (call it $\underline{x}$). We use a proof by contradiction: assume that B is not positive-definite.

THERE IS A GAP HERE — FILLED IN LATER.

Then, we can find a non-zero vector $\underline{w}$ such that $B\underline{w} = 0$. Then,

$$f(\underline{x}+\underline{w}) = f(\underline{x}) + \cancel{\tfrac{1}{2}\langle \underline{w}, B\underline{w} \rangle}$$

$$= f(\underline{x}).$$

Hence, $\underline{x}+\underline{w}$ is also a minimizer. This contradicts uniqueness. Hence, the only way to have a unique global minimizer is if B is strictly P.D.

Take-home message

in case of the model problem

Continuous optimization is a nice application of Calculus and Linear algebra. In the next chapters we will attempt to approximate a general O.P. with a quadratic problem, which can be solved using Linear Algebra.

Chapter 3 — Line Search Methods

Notation for the O.P. :

$$\underline{x}_* = \arg\min_{\underline{x} \in \mathbb{R}^n} f(\underline{x})$$

Line Search Methods are <u>iterative</u>.

We start off with an initial guess for $\underline{x}_*$ (call it $\underline{x}_0$). We make a sequence of improved guesses $\underline{x}_k$, such that:

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

Typically, $\underline{s}_k$ depends on $\nabla f(\underline{x}_k)$ and is broken up into a magnitude and a direction:

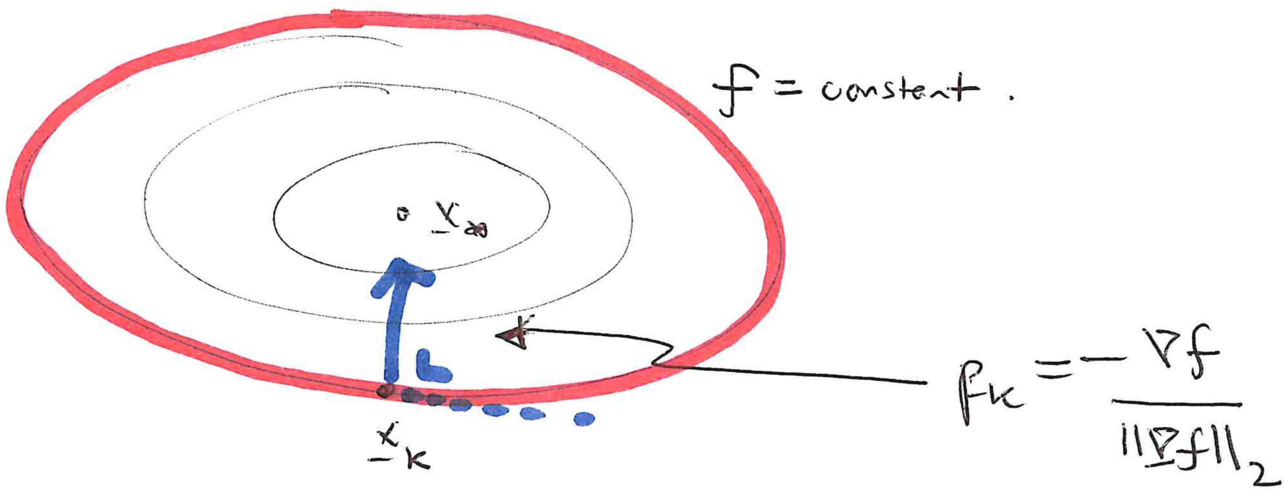
$$\underline{s}_k = \alpha_k p_k ,$$

where typically p_k is a unit vector.

When p_k is a unit vector, α_k can be found by solving a 1D OP:

$$\boxed{\alpha_k = \arg\min_{\alpha > 0} f(\underline{x}_k + \alpha p_k) \quad (*)}$$

§ 3.2 Steepest Descent Method



Look at:

$$f(\underline{x}_k + \alpha \underline{p}) = f(\underline{x}_k) + \overbrace{\alpha p_i \frac{\partial f}{\partial x_i}(\underline{x}_k)}^{\text{DOMINANT}}$$
$$+ \frac{1}{2}\alpha^2 p_i p_j \frac{\partial^2 f}{\partial x_i \partial x_j}(\underline{x}_k + t\underline{p}),$$

$t \in (0, \alpha)$.

To reduce f as much as possible in one iteration ($\underline{x}_k \to \underline{x}_k + \alpha \underline{p}$), we need to make the dominant term as negative as possible. To do this, we simply take:

$$\underline{p} = -\frac{\nabla f(\underline{x}_k)}{\|\nabla f(\underline{x}_k)\|_2}.$$

Then,

$$f(\underline{x}_k + \alpha \underline{p}) = f(\underline{x}_k) - \alpha \left.\frac{\nabla f \cdot \nabla f}{\|\nabla f\|}\right|_{\underline{x}_k} + O(\alpha^2)$$

$$= f(\underline{x}_k) - \alpha \|\nabla f\|_2 \Big|_{\underline{x}_k} + O(\alpha^2).$$

Thus, the ~~cha~~ reduction in f is maximized.

$$\underline{p}_k = -\frac{\nabla f(\underline{x}_k)}{\|\nabla f(\underline{x}_k)\|_2} \qquad (1)$$

is the direction of Steepest Descent.

$\boxed{\text{Pseudocode in notes (p.23)}}$

Another choice for the search direction is the Newton Method (§3.3)

Taylor approximation:

$$f(\underline{x}_k + \underline{p}) \simeq \underbrace{f(\underline{x}_k)}_{= c} + p_i \underbrace{\frac{\partial f}{\partial x_i}(\underline{x}_k)}_{a_i} + \frac{1}{2} p_i p_j \underbrace{\frac{\partial^2 f}{\partial x_i \partial x_j}(\underline{x}_k)}_{B_{ij}}$$

This is the Quadratic Model Problem:

$$m_k(p) = c + \langle a, p \rangle + \frac{1}{2} \langle p, Bp \rangle$$

Minimize $m_k(p)$. If B is invertible, then $p = -B^{-1}\underline{a}$.

This gives the descent direction in the Newton method. Restoring k, we have:

$$p_k^N = - B^{-1}(\underline{x}_k) \nabla f(\underline{x}_k) \qquad (2)$$

$$\boxed{\text{Pseudocode in notes on p. 24}}$$

Equation (2) is the <u>Newton descent direction</u>

For the Newton method to yield a reduction in f from $\underline{x}_k$ to $\underline{x}_k + p_k^N$, we require $B(\underline{x}_k)$ to be positive-definite. <u>Proof</u>:

$$f(\underline{x}_k + t p_k^N) = f(\underline{x}_k) + t \sum_{i=1}^{n} (p_k^N)_i \frac{\partial f}{\partial x_i}(\underline{x}_k) + O(t^2)$$

Hence,

$$f(\underline{x}_n + t p_n^N) = f(\underline{x}_n)$$

$$- t \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(B^{-1})_{ij} \frac{\partial f}{\partial x_j} \right]_{\underline{x}_k} \frac{\partial f}{\partial x_i} + O(t^2)$$

Reason: $p_n^N = - B^{-1} \underline{\nabla} f$.

$$(p_n^N)_i = - \sum_{j=1}^{n} (B^{-1})_{ij} \frac{\partial f}{\partial x_j}$$

(Matrix multiplication).

Hence,

$$f(\underline{x}_n + t p_n^N) = f(\underline{x}_n) \underbrace{- t \underbrace{\langle \nabla f(\underline{x}_n), B^{-1} \underline{\nabla} f \rangle}_{> 0}}_{\text{NEG.}} + O(t^2).$$

Hence,

$$f(\underline{x}_n + t p_n^N) < f(\underline{x}_n),$$

for t suff. small.

Lecture 3

Advantages of using P_n^N :

- Step-length is provided, no need to solve the sub-problem (*)
- Simple criterion for method to work ($P(x_n)$ is pos.-definite)
- Quadratic convergence

We illustrate the idea of quadratic convergence here for a 1D problem where we wish to solve:

$$x_* = \arg\min_{x \in \mathbb{R}} f(x) .$$

By first-order optimality, $f'(x_*) = 0$.

$$\underline{x_{k+1}} = \underline{x_k} - \frac{f'(x_k)}{f''(x_k)} . \quad (3)$$

Error:

$$\epsilon_k = x_* - x_k \qquad\qquad x_k = x_* - \epsilon_k$$

$$\epsilon_{k+1} = x_* - x_{k+1} . \qquad x_{k+1} = x_* - \epsilon_{k+1}$$

Sub in to (3):

$$\underline{\cancel{x_*} - \epsilon_{k+1}} = \underline{\cancel{x_*} - \epsilon_k} - \frac{f'(\overbrace{x_* - \epsilon_k}^{x_k})}{f''(x_* - \epsilon_k)}$$

$$\Rightarrow \epsilon_{k+1} = \epsilon_k + \frac{f'(x_* - \epsilon_k)}{f''(x_* - \epsilon_k)}$$

$$\Rightarrow \epsilon_{k+1} = \epsilon_k + \frac{\left[\cancel{f'(x_*)} - f''(x_*)\epsilon_k + \frac{1}{2} f'''(x_*)\epsilon_k^2 + \cdots\right]}{f''(x_*) - f'''(x_*)\epsilon_k + \cdots}$$

$$\Rightarrow \epsilon_{k+1} = \epsilon_k - \frac{\cancel{f''(x_*)}\epsilon_k\left[1 - \frac{1}{2}\frac{f'''(x_*)}{f''(x_*)}\epsilon_k + \cdots\right]}{\cancel{f''(x_*)}\left[1 - \frac{f'''(x_*)}{f''(x_*)}\epsilon_k + \cdots\right]}$$

BINOMIAL THEOREM: $(1+z)^p = 1 + pz + \frac{p(p-1)}{2} z^2 + \cdots$

$$\epsilon_{k+1} = \epsilon_k - \epsilon_k\left[1 - \frac{1}{2}\frac{f'''(x_*)}{f''(x_*)}\epsilon_k + \cdots\right]\left[1 + \frac{f'''(x_*)}{f''(x_*)}\epsilon_k - \cdots\right]$$

$$\Rightarrow \epsilon_{k+1} = \epsilon_k - \epsilon_k\left[1 + \frac{f'''(x_*)}{f''(x_*)}\epsilon_k - \frac{1}{2}\frac{f'''(x_*)}{f''(x_*)}\epsilon_k + O(\epsilon_k^2)\right]$$

$$\Rightarrow \epsilon_{k+1} = \epsilon_k - \epsilon_k\left[1 + \frac{1}{2}\frac{f'''(x_*)}{f''(x_*)}\epsilon_k + O(\epsilon_k^2)\right]$$

Hence,

$$\epsilon_{k+1} = \cancel{\epsilon_k} - \cancel{\epsilon_k} - \frac{1}{2} \epsilon_k^2 \frac{f'''(x_*)}{f''(x_*)} + O(\epsilon_k^3)$$

Hence,

$$\epsilon_{k+1} = -\frac{1}{2} \epsilon_k^2 \frac{f'''(\overbrace{x_\otimes}^{x_k + \epsilon_k})}{f''(x_\otimes)} + O(\epsilon_k^3).$$

<u>OR</u>

$$\epsilon_{k+1} = -\frac{1}{2} \epsilon_k^2 \frac{f'''(x_k)}{f''(x_*)} + O(\epsilon_k^3).$$

Drawbacks:

- Computation of Hessian at each iteration
- Requires inversion of the Hessian at each iteration ($O(n^3)$)

<u>Amelioration</u> — approximate the Hessian matrix using the SECANT METHOD (§3.4)

To see how the Secant Method works, we go back to the 1D problem, and we look at:

$$f'(\underline{x_k + \delta x}) \simeq \underline{f'(x_k)} + \underline{f''(x_k)}\,\delta x .$$

Take: $\delta x = x_{k+1} - x_k$. Hence, this equation becomes:

$$\underbrace{f'(x_{k+1}) - f'(x_k)}_{y_k} \simeq f''(x_k)\,[\underbrace{x_{k+1} - x_k}_{s_k}] .$$

Approximate Hessian:

$$f''(x_k) \simeq y_k / s_k .$$

Equivalent n-dimensional analogy:

$$\underbrace{\nabla f(\underline{x}_{k+1}) - \nabla f(\underline{x}_k)}_{\underline{y}_k} \simeq B(\underline{x}_k)\,[\underbrace{\underline{x}_{k+1} - \underline{x}_k}_{\underline{s}_k}] .$$

Or $\underline{y}_k = \underbrace{B(\underline{x}_k)}_{B_{k+1}}\,\underline{s}_k .$

Pseudocode:

Choose $\underline{x}_0$ sufficiently close to $\underline{x}_*$

Choose B_0.

for $k = 0, 1, 2, \ldots$

Compute the descent direction by solving $B_k p_k = -\nabla f(x_k)$

Choose the stepsize α_k.

Write $\underline{s}_k = \alpha_k p_k$

Set $\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{s}_k$

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

Update approx Hessian for next iteration by solving $\boxed{\underline{y}_k = B_{k+1} \underline{s}_k}$

end for

Chapter 4 — BFGS method.

Problem — we need to solve for B_{k+1} in the equation

$$\underline{y}_k = B_{k+1} \underline{s}_k \quad (4)$$

1D: $f''(x_k) \simeq y_k / s_k$.

B_{k+1} — $n \times n$ matrix $\Rightarrow$ $n \times n$ unknowns.

n equations in the secant approximation (4)

We solve for B_{k+1} in an approximate sense using the <u>BFGS method</u>. We build an approximation of B_{k+1} out of y_k and $B_k s_k$.

We look at the outer product of y_k with itself:

$$[y_k y_k^T]_{ij} = (y_k)_i (y_k)_j \quad \left\{ \begin{array}{l} \underline{n} \times 1 \\ 1 \times \underline{n} \\ n \times n. \end{array} \right. \quad \left(\quad \right)$$

The same outer product for $B_k s_k$.

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

where α and β are TBC.

Theorem 4.1 gives us the values for α and β:

EXAM

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

<u>Proof</u>: Our approximation of B_{k+1} is:

$$B_{k+1} = B_k + \alpha \underline{y}_k \underline{y}_k^T + \beta (B_k \underline{s}_k)(B_k \underline{s}_k)^T \quad (5)$$

We require this to satisfy:

$$\underline{y}_k = B_{k+1} \underline{s}_k . \quad (6)$$

Sub (5) into (6):

$$\underline{y}_k = \left[B_k + \alpha \underline{y}_k \underline{y}_k^T + \beta (B_k \underline{s}_k) \underline{\underline{(B_k \underline{s}_k)^T}} \right] \underline{s}_k$$

$$= B_k \underline{s}_k + \alpha \underline{y}_k (\underline{y}_k^T \underline{s}_k) + \beta B_k \underline{s}_k (\underline{s}_k^T B_k^T \underline{s}_k)$$

$$\Rightarrow \underline{y}_k = B_k \underline{s}_k + \alpha \underline{y}_k \langle \underline{y}_k , \underline{s}_k \rangle + \beta B_k \underline{s}_k \langle \underline{s}_k , B_k^T \underline{s}_k \rangle$$

Re-arrange:

$$0 = \underline{y}_k \left[-1 + \alpha \langle \underline{y}_k, \underline{s}_k \rangle \right]$$
$$+ \quad B_k \underline{s}_k \left[1 + \beta \langle \underline{s}_k, B_k^T \underline{s}_k \rangle \right]$$

In general, $\underline{y}_k$ and $B_k \underline{s}_k$ are linearly independent, so we require the square brackets to be zero:

$$-1 + \alpha \langle \underline{y}_k, \underline{s}_k \rangle = 0$$
$$\Longrightarrow \quad \alpha = \frac{1}{\langle \underline{y}_k, \underline{s}_k \rangle}$$

Also,

$$1 + \beta \langle \underline{s}_k, B_k^T \underline{s}_k \rangle = 0$$

$$\Longrightarrow \beta = \frac{-1}{\langle \underline{s}_k, B_k^T \underline{s}_k \rangle} \qquad \blacksquare$$