Recap:

- Convex functions and their minimizers
- Quadratic Model Problem (QMP)

} Tuesday online lecture

Plan for today:

- Finish looking at QMP
- Look at Linesearch Methods (Chapter 3)

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

- $\underline{x}$ is the variable in $\mathbb{R}^n$
- $\underline{a}$ is a constant vector in $\mathbb{R}^n$
- $B$ is an $n \times n$ matrix with real entries.

When $B$ is symmetric positive-definite:

$$\nabla f = \underline{a} + B\underline{x}$$

$$\nabla f = 0 \Rightarrow \underline{x} = \underline{x}_*, \text{ where } B\underline{x}_* = -\underline{a}.$$

Theorem: $f$ attains a minimum if and only if $B$ is positive semi-definite and $\underline{a}$ is in the range of $B$.

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

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

Take $\underline{w} \in \mathbb{R}^n$ arbitrary:

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

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

Since $B$ is symmetric, the two underlined terms are the same, so we have: ($B^T = B$)

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

$$= f(\underline{x}) + \underbrace{\langle \underline{a}, \underline{w} \rangle + \langle B\underline{x}, \underline{w} \rangle}_{-\underline{a} = B\underline{x}} + \tfrac{1}{2}\langle \underline{w}, B\underline{w} \rangle$$

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

$\therefore f(\underline{x}+\underline{w}) = f(\underline{x}) + \frac{1}{2}\langle \underline{w}, B\underline{w} \rangle \geq f(\underline{x}) \quad \forall \underline{w} \in \mathbb{R}^n$

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

Conversely, we suppose that $f$ has a minimizer (at $\underline{x}$, say).

By first-order optimality,

$$\nabla f(\underline{x}) = 0$$

But $\nabla f = \underline{a} + B\underline{x}$ for the Q.M.P.

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

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

By 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 Q.M.P., the Hessian matrix is just $B$.

Hence, $B$ is P.S.D. ∎

Theorem: $f$ has a unique minimizer if and only if $B$ is symmetric positive definite.

Assume first that $B$ is positive definite. Hence, $B$ is invertible, so $\underline{a}$ is in the range of $B$, so there exists an $\underline{x}$ such that $B\underline{x} = -\underline{a}$ ($\underline{x} = -B^{-1}\underline{a}$).

Look at:

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

Hence:

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

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

Conversely, assume that $f$ has a unique global minimizer, call it $\underline{x}$. Assume for contradiction that $B$ is not positive-definite. Then, there exists $\underline{w} \neq 0$ such that $B\underline{w} = -\lambda \underline{w}$, $\lambda \geq 0$.

By the previous calculation,

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

$$= f(\underline{x}) - \frac{1}{2}\lambda \langle \underline{w}, \underline{w} \rangle$$

$$= f(\underline{x}) - \frac{1}{2}\lambda \|\underline{w}\|_2^2$$

$$\Rightarrow f(\underline{x} + \underline{w}) \leq f(\underline{x})$$

This contradicts the fact that $\underline{x}^*$ is a unique global minimizer. To resolve the contradiction, it must be true that $B$ is positive-definite. ∎

Why do we spend so much time on the QMP?

Answer: Because it is a model for all the other problems in continuous optimization.

We move on to looking at algorithms to find the minimizer numerically, starting with Linesearch Methods (Ch. 3).

Notation:

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

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

Linesearch methods are iterative: we start with an initial guess for $\underline{x}_*$ $(=\underline{x}_0)$. We make a sequence of improved guesses $\underline{x}_k$ such that:

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

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

$$\underline{s}_k = \alpha_k \underline{p}_k$$

where $\underline{p}_k$ is typically a unit vector, and $\alpha_k$ is obtained by solving a 1D OP:

$$\alpha_k = \arg\min_{\alpha>0} f(\underline{x}_k + \alpha \underline{p}_k)$$

## § 3.2 Steepest Descent Method.

Look at (Taylor's Remainder Theorem, Summation Convention)

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

DOMINANT, for $\alpha$ small

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

To reduce $f$ as much as possible in one iteration, we need to make the highlighted term as

we need to make the highlighted term as negative as possible (with $\|p\|_2 = 1$):

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

Hence:

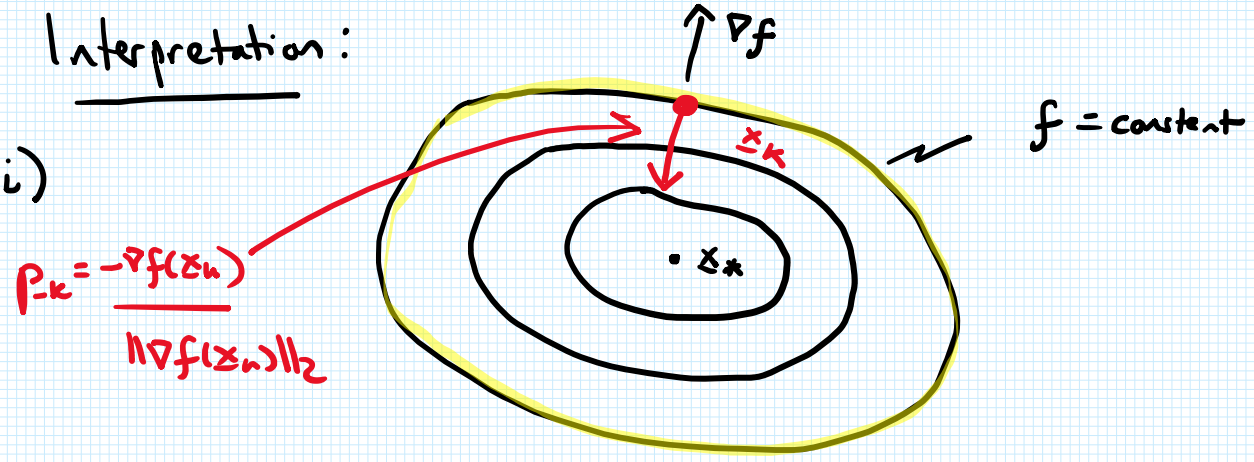
$$f(x_k + \alpha p) = f(x_k) - \alpha \frac{\nabla f(x_k) \cdot \nabla f(x_k)}{\|\nabla f(x_k)\|_2} + O(\alpha^2)$$

$$= f(x_k) - \alpha \|\nabla f(x_k)\|_2 + O(\alpha^2)$$

The direction in (1) is the direction of steepest descent.

Interpretation:

i)




ii) Croix Rousse, Lyon. 2017 Direction of steepest ascent, never got lost!

iii) Stochastic Gradient Descent.

iv) ODEs – Euler Method

Gradient Descent is to Optimization what Euler is to ODEs

Another choice for Search Direction

Newton (§ 3.3)

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

($\underline{p}$: a and unit vector bundled into one)

This is the QMP:

At each step in the iteration, we minimize $m_k(\underline{p})$:

$$m_k(\underline{p}) = c + \langle \underline{a}, \underline{p} \rangle + \frac{1}{2} \langle \underline{p}, B\underline{p} \rangle \text{ , where:}$$

$$c = f(\underline{x}_k)$$

$$\underline{a} = \nabla f(\underline{x}_k) \quad \Rightarrow \quad B_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(\underline{x}_k) .$$

$$\nabla_p m_k(\underline{p}) = 0 \Rightarrow \underline{p} = -B^{-1}\underline{a} ,$$

provided $B$ is invertible. This gives us the Newton step:

$$\boxed{\underline{p}_k^N = - B^{-1}(\underline{x}_k) \nabla f(\underline{x}_k)} \quad (2)$$

To reduce the cost function in going from one iteration to the next, we require $B(\underline{x}_k)$ to be positive-definite.

$$f(\underline{x}_k + t\,\underline{p}_k^N) = f(\underline{x}_k)$$

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

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

$$\Rightarrow f(x_k + t p_k^N)$$

$$= f(x_k) + t \sum_{i=1}^{n} \left[ -B^{-1}(x_k) \nabla f(x_k) \right]_i \frac{\partial f}{\partial x_i}(x_k) + O(t^2)$$

$$= f(x_k) - t \sum_{i=1}^{n} \sum_{j=1}^{n} B^{-1}_{ij}(x_k) \frac{\partial f}{\partial x_j}(x_k) \frac{\partial f}{\partial x_i}(x_k) + O(t^2)$$

$$= f(x_k) - t \underbrace{\left\langle \nabla f(x_k), B^{-1}(x_k) \nabla f(x_k) \right\rangle}_{< 0} + O(t^2)$$

Hence
$f(x_k + t p_k^N) < f(x_k)$ for $t$ sufficiently small.

We characterize the Newton Method

- Step length is provided, no need to compute $\alpha_k$ separately
- Simple criterion for success — $B(x_k)$ should be positive-definite
- Quadratic convergence

We illustrate the idea of quadratic convergence for the 1D case:

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

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

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

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

$$\epsilon_k = x_* - x_k$$

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

Sub into (3):

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

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

$$= \epsilon_k + \frac{\left[\overbrace{f'(x_*)}^{=0} - f''(x_*)\epsilon_k + \frac{1}{2}f'''(x_*)\epsilon_k^2 + \cdots\right]}{\left[f''(x_*) - f'''(x_*)\epsilon_k + \cdots\right]}$$

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

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

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

Hence:

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

$$= \epsilon_k - \epsilon_k - \epsilon_k^2 \frac{f'''(x_k)}{f''(x_k)} + \frac{1}{2}\epsilon_k^2 \frac{f'''(x_k)}{f''(x_k)} + \dots$$

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

$$= \frac{}{}$$

E.g. $\epsilon_k = O(10^{-2})$

then $\epsilon_{k+1} = O(10^{-4})$

and $\epsilon_{k+1} = O(10^{-8})$

Hence, quadratic convergence is shown. ∎

Drawbacks:

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

To speed things up, we use versions of the Secant Method.

Idea in 1D:

$$f'(x_k + \delta x) \simeq f'(x_k) + f''(x_k)\,\delta x$$

Take $\delta x = x_{k+1} - x_k$

$$\Rightarrow f'(x_{k+1}) \simeq f'(x_k) + \overbrace{f''(x_k)}^{\text{Hessian}}(x_{k+1} - x_k)$$

$$\Rightarrow f'(x_{k+1}) \simeq f'(x_k) + f''(x_k)(x_{k+1} - x_k)$$

$$\Rightarrow \underbrace{f'(x_{k+1}) - f'(x_k)}_{y_k} \simeq f''(x_k) \underbrace{(x_{k+1} - x_k)}_{s_k}$$

Approximation of Hessian:

$$f''(x_k) \simeq y_k / s_k \quad . \qquad (4a)$$

Equivalent in $n$ dimensions:

$$\underbrace{\nabla f(x_{k+1}) - \nabla f(x_k)}_{\underline{y}_k} \simeq \underbrace{B(\underline{x}_k)}_{B_{k+1}} \underbrace{(\underline{x}_{k+1} - \underline{x}_k)}_{\underline{s}_k} \qquad (4b)$$

**Problem**

$f''(x_k) \cdot s_k = y_k$, solvable for $f''(x_k)$ } (4a) is ok

(4b): $n$ linear equations, $n^2$ unknowns.

Equation (4b) is under-determined, for $n > 1$.

BFGS addresses this problem.

We build an approximation for $B_{k+1}$ out of $B_k$, $\underline{y}_k$, $B_k \underline{s}_k$.

$B_{k+1} \simeq$ Lin. Combination of $B_k$, $y_k$, $B_k s_k$

We look at the outer product of $y_k$ with itself:

$$\left[ y_k y_k^T \right]_{ij} = (y_k)_i (y_k)_j$$

$n \times 1$, $1 \times n$, $(n \times 1) \times (1 \times n) \to n^2$

$$B_{k+1} = B_k + \alpha y_k y_k^T + \beta (B_k s_k)(B_k s_k)^T \quad (5)$$

where $\alpha$ and $\beta$ are TBC.

Theorem 4.1 gives us the values of $\alpha$ and $\beta$:

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

EXAM

Proof: We must choose $\alpha$ and $\beta$ such that that

$$y_k = B_{k+1} s_k \quad (6)$$

i.e. we must choose $\alpha$ and $\beta$ such that $B_{k+1}$ satisfies the secant condition. This fixes $\alpha$ and $\beta$ through direct calculation.

We sub the BFGS approximation (5) into (6):

$$y_k = \left[ B_k + \alpha y_k y_k^T + \beta (B_k s_k)(B_k s_k)^T \right] s_k$$

$$= B_k s_k + \alpha y_k y_k^T s_k + \beta B_k s_k (B_k s_k)^T s_k$$

$$= B_k s_k + \alpha y_k (y_k^T s_k) + \beta B_k s_k (s_k^T B_k^T s_k)$$

$$\Rightarrow y_k = B_k s_k + \alpha y_k \langle y_k, s_k \rangle + \beta B_k s_k \langle s_k, B_k^T s_k \rangle$$

By linear independence:

$$1 = \alpha \langle y_k, s_k \rangle \Rightarrow \alpha = \frac{1}{\langle y_k, s_k \rangle}$$

$$0 = 1 + \beta \langle s_k, B_k^T s_k \rangle \Rightarrow \beta = \frac{-1}{\langle s_k, B_k^T s_k \rangle}$$ ∎

Where are we going?

- Determination of $\alpha$ and $\beta$ in BFGS ✓
- Properties of BFGS algorithm (Tuesday, Online)
- Next Thursday: We show that calculating the Hessian matrix using BFGS is $O(n^2)$.