Review

1. $f(\mathbf{x})$ *is called convex if* $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \le \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})$.
2. $f(\mathbf{x})$ *is called strictly convex if* $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})$.
3. $f(\mathbf{x})$ *is called strongly convex with a constant parameter* $\mu > 0$ *if*

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \le \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y}) - \frac{\mu}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2.$$

5. *East to verify that* $f(\mathbf{x})$ *is strongly convex with* $\mu > 0$ *if and only if* $f(\mathbf{x}) - \frac{\mu}{2}\|x\|^2$ *is convex. Strong convexity with* $\mu = 0$ *is convexity.*
6. *It is easy to see that*

$$\textit{strong convexity} \Rightarrow \textit{strict convexity} \Rightarrow \textit{convexity}.$$

Example: ① $f(x) = |x|$ is convex but not strictly convex

② $f(x) = e^x$ is convex but not strongly convex

③ $f(x) = x^2$ is strongly convex

Equivalent Conditions:

① If $\nabla f(x)$ is continuous,

$$\text{convexity} \Leftrightarrow \begin{cases} 1.\ f(\mathbf{x}) \ge f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle, \quad \forall \mathbf{x},\mathbf{y}. \\ 2.\ \langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \ge 0, \quad \forall \mathbf{x},\mathbf{y}. \end{cases}$$

$$\text{Strong convexity} \Leftrightarrow \begin{cases} 1.\ f(\mathbf{x}) \ge f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x},\mathbf{y}. \\ 2.\ \langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \ge \mu\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall \mathbf{x},\mathbf{y}. \end{cases}$$

② If $\nabla^2 f(x)$ is continuous

1) Convexity $\Leftrightarrow \nabla^2 f(x) \geq 0$, $\forall x$

2) Strong convexity $\Leftrightarrow \nabla^2 f(x) \geq \mu I$, $\forall x$, $\mu > 0$

3) Strict convexity $\Leftarrow \nabla^2 f(x) > 0$

$f(x) = x^4$ is strictly convex, $f''(0) = 0$

Optimality Conditions:

Theorem 2.1 (First Order Necessary Conditions). *For a C^1 function (first order derivatives exist and are continuous) $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$, if $\mathbf{x}^*$ is a local minimizer, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Theorem 2.2 (Second Order Necessary Conditions). *For a C^2 function (second order derivatives exist and are continuous) $f(\mathbf{x}) : R^n \longrightarrow \mathbb{R}$, if $\mathbf{x}^*$ is a local minimizer, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) \geq 0$ (Hessian matrix is positive semi-definite).*

Theorem 2.3 (Second Order Sufficient Conditions). *For a C^2 function (second order derivatives exist and are continuous) $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$, if $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) > 0$ (Hessian matrix is positive definite), then $\mathbf{x}^*$ is a strict local minimizer.*

Only strong convexity $\Rightarrow \nabla^2 f(x) > 0$, $\forall x$.

Theorem 2.4. *Assume $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex.*

1. *Any local minimizer is also a global minimizer.*
2. *If $f(\mathbf{x})$ is also continuously differentiable (the same as C^1 functions), then $\mathbf{x}^*$ is a global minimizer if and only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Theorem 2.5. *Assume $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is strongly convex and also continuously differentiable (the same as C^1 functions). Then $f(\mathbf{x})$ has a unique global minimizer $\mathbf{x}^*$, which is the only critical point of the function.*

1) Convex $f(x)$ may not have a minimizer : $f(x) = x$

2) Strictly Convex $f(x)$ may not have a minimizer : $f(x) = e^x$

3) Strong Convex $f(x)$ has a unique minimizer

Singular Values of $A \in \mathbb{R}^{n \times n}$ is denoted by $\sigma_i(A)$

Definition $\sigma_i(A) = \sqrt{\lambda_i(A^T A)} = \sqrt{\lambda_i(AA^T)} \geq 0$

Facts/Theorems :

① If A is real symmetric, $\sigma_i(A) = |\lambda_i(A)|$

② If A is real symmetric and PSD, $\sigma_i(A) = \lambda_i(A)$

③ $\|A\| = \max_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|} = \max_i \sigma_i(A)$

Example : $f(x) = \frac{1}{2} x^T K x - x^T b$

$\nabla f = Kx - b$

$\nabla^2 f = K$

$\Delta x = \frac{1}{n+1}$

$K = \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}_{n \times n}$

$K > 0 \Rightarrow \sigma_i(K) = \lambda_i(K)$

$\lambda_i(K) = 4 \frac{1}{\Delta x^2} \sin^2\left(\frac{\pi}{2} i \Delta x\right)$

$= 4 \frac{1}{\Delta x^2} \sin^2\left(\frac{\pi}{2} i \Delta x\right)$

So we get

① $\|K\| \le \max_i \sigma_i = 4\frac{1}{\Delta x^2}\sin^2\left(\frac{\pi}{2}\frac{n}{n+1}\right) < 4\frac{1}{\Delta x^2}$

② $\lambda_1 < \lambda_2 < \cdots < \lambda_n$

$\Rightarrow \lambda_1 I \le K \le \lambda_n I$ meaning $\begin{cases} \lambda_n I - K \text{ is PSD} \\ K - \lambda_1 I \text{ is PSD} \end{cases}$

③ $\lambda_n = 4\frac{1}{\Delta x^2}\sin^2\left(\frac{\pi}{2}\frac{n}{n+1}\right) < 4\frac{1}{\Delta x^2}$

$\lambda_1 = 4\frac{1}{\Delta x^2}\sin^2\left(\frac{\pi}{2}\Delta x\right)$ $\quad \Delta x = \frac{1}{n+1}$

So $\|\nabla^2 f\| = \|K\| < 4\frac{1}{\Delta x^2}$ implies

$\frac{h^T \nabla^2 f\, h}{h^T h} \le \lambda_n < 4\frac{1}{\Delta x^2}$ $\quad \frac{h^T}{h}$

→ C-F-W minmax principle

$\Rightarrow (y-x)^T \nabla^2 f[\cdot] (y-x) < \frac{4}{\Delta x^2}\|y-x\|^2$

Lemma 2.1 (Descent Lemma)**.** *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L, then*

$$f(\mathbf{y}) \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Remark 2.3. *Notice that there is no assumption on the existence of Hessian. But if assuming $\|\nabla^2 f\| \le L$, then by Theorem 1.4,*

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{y})^T \nabla^2 f(\mathbf{z})(\mathbf{x} - \mathbf{y})$$

which implies $\sigma_i(\nabla^2 f) \le L \Rightarrow |\lambda_i(\nabla^2 f)| \le L \Rightarrow \frac{h^T \nabla^2 f\, h}{h^T h} \le L$

$$f(\mathbf{y}) \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2,$$

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2,.$$

Remark 2.4. *Notice that there is no assumption on convexity. But if assuming strong convexity of $f(\mathbf{x})$, by Theorem 1.1,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

If $\nabla^2 f(x) \geq \mu I$, $\min_i \lambda_i(\nabla^2 f) \geq \mu \Rightarrow \frac{h^T \nabla^2 f h}{h^T h} \geq \mu$

Lemma 2.2 (Sufficient Decrease Lemma)**.** *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L, then the gradient descent method* (2.1) *satisfies*

$$f(\mathbf{x}) - f(\mathbf{x} - \eta \nabla f(\mathbf{x})) \geq \eta(\frac{L}{2} - \eta)\|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x}, \forall \eta > 0.$$

Proof. Lemma 2.1 gives $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$

$$f(\mathbf{x} - \eta \nabla f(\mathbf{x})) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), -\eta \nabla f(\mathbf{x}) \rangle + \frac{L}{2}\|\eta \nabla f(\mathbf{x})\|^2.$$

□

Convergence for $0 < \eta < \frac{2}{L}$

$$f(x_{k+1}) - f(x_k) \leq -\eta\left(1 - \frac{L}{2}\eta\right)\|\nabla f(x_k)\|^2$$

$$\eta \in \left(0, \tfrac{2}{L}\right) \Rightarrow \omega = \eta\left(1 - \tfrac{L}{2}\eta\right) > 0$$

$$f(x_k) - f(x_{k+1}) \geq \omega \|\nabla f(x_k)\|^2$$

① Sum it for $k = 0, 1, 2, \ldots$

$$\sum_{k=0}^{\infty} [f(x_k) - f(x_{k+1})] \geq \omega \sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2$$

② $\{f(x_k)\}$ is a decreasing sequence, thus it is also bounded $(f(x_k) \leq f(x_k) < f(x_0))$

Completeness Theorem of Real numbers

A monotone bounded sequence has a limit.

So $\lim_{k\to\infty} f(x_k)$ exists (doesn't imply $\lim_{k\to\infty} x_k$ exists)

$$\text{LHS} = f(x_0) - \lim_{k\to\infty} f(x_k)$$

$$\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2 \le \frac{1}{\omega}\left[f(x_0) - \lim_{k\to\infty} f(x_k)\right]$$

$g_n = \sum_{k=0}^{n} \|\nabla f(x_k)\|^2$ is ↑ and bounded above

The series $\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2$ converges

$$\Rightarrow \lim_{k\to\infty} \|\nabla f(x_k)\|^2 = 0 \Rightarrow \lim_{k\to\infty} \nabla f(x_k) = 0$$

(doesn't imply $\lim_{k\to\infty} x_k$ exists)

④ Let $g_N = \min_{0\le k\le N} \|\nabla f(x_k)\|$, then

$$f(x_k) - f(x_{k+1}) \ge \omega \|\nabla f(x_k)\|^2$$

$$\Rightarrow \sum_{k=0}^{N} \|\nabla f(x_k)\|^2 \le \frac{1}{\omega}\left[f(x_0) - f(x_{N+1})\right]$$

$$\le \frac{1}{\omega}\left[f(x_0) - f(x_*)\right]$$

$$(N+1)\, g_N^2 \le \sum_{k=0}^{N} \|\nabla f(x_k)\|^2$$

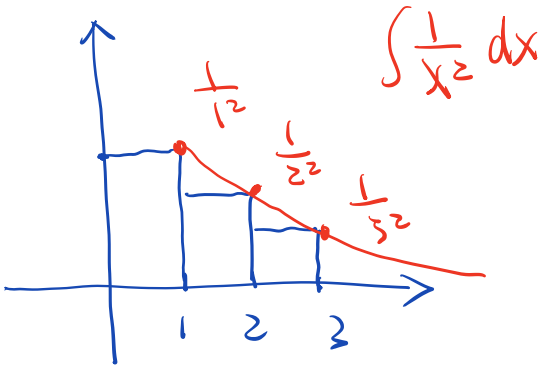
$$\Rightarrow g_N \le \frac{1}{\sqrt{N+1}} \sqrt{\frac{1}{\omega}\left[f(x_0) - f(x_*)\right]}$$

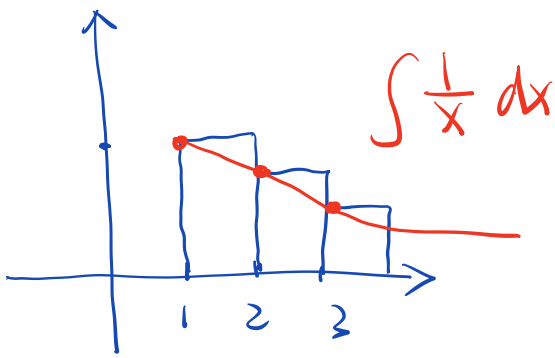
Remark: convergence of $\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2$

does not imply convergence of

$$\sum_{k=0}^{\infty} \|\nabla f(x_k)\|$$

$$\sum_{n=0}^{\infty} \frac{1}{n^2} < +\infty \qquad \sum_{n=0}^{\infty} \frac{1}{n} = +\infty$$





Theorem Assume ∇f is L-continuous.
Assume $f(x) \geq f(x^*)$, $\forall x \in \mathbb{R}^n$

Then for $x_{k+1} = x_k - \eta \nabla f(x_k)$

where $\eta \in (0, \frac{2}{L})$ is a constant:

① $f(x_{k+1}) - f(x_k) \leq -\eta(1 - \frac{L}{2}\eta) \|\nabla f(x_k)\|^2$

② $\lim_{k \to \infty} x_k$ may not exist!

③ $\lim_{k \to \infty} \|\nabla f(x_k)\| = 0$

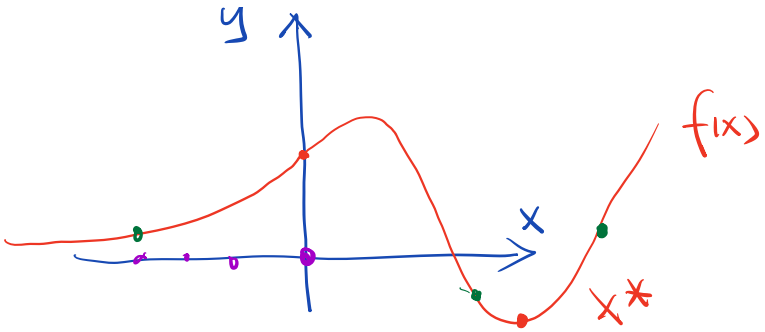
$\omega = \eta(1 - \frac{L}{2}\eta)$

④ $\min_{0 \leq k \leq N} \|\nabla f(x_k)\| \leq \frac{1}{\sqrt{N+1}} \sqrt{\frac{1}{\omega}[f(x_0) - f(x^*)]}$

Example: Let $f(x) = \begin{cases} e^x, & x \leq 0 \\ \text{smooth}, & x > 0 \end{cases}$

such that $\begin{cases} \nabla f \text{ is } L\text{-conitinuous with } L = 1 \\ f(x) \text{ has a global minimizer} \end{cases}$

$\frac{d^2}{dx^2} e^x \leq 1$ for $x \leq 0 \Rightarrow L = 1$ is possible.



Consider Gradient Descent

$$\begin{cases} x_{k+1} = x_k - \eta f'(x_k) \\ x_0 = 0 \\ \eta = 1 \in (0, \frac{2}{L}) \end{cases}$$

Everything satisfies the Theorem, so

① $\lim_{k\to\infty} f(x_k)$ exists

② $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$

This means $\|x_{k+1} - x_k\| < \varepsilon$ for large k

③ But $\{x_k\}$ does not converge!

$x_0 = 0 \Rightarrow x_k < 0, \ k \geq 1$

$$\Rightarrow x_{k+1} = x_k - \eta f'(x_k)$$
$$= x_k - e^{x_k}$$

So $\nabla f(x_k) \to 0 \Rightarrow e^{x_k} \to 0 \Rightarrow x_k \to +\infty$

The sequence $\{x_k\}$ diverges.

because it's not Cauchy.

$$\|x_{k+1} - x_k\| = \|f'(x_k)\| \to 0$$

↳ This is not Cauchy.

Cauchy means $|x_m - x_n|$ is small for any large m & n.