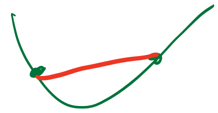
Convexity



1. $f(\mathbf{x})$ *is called convex if* $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})$.

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*

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

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,

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

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

$A \geq B \iff A - B \geq 0$

② If $\nabla^2 f(x)$ is continuous

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

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

3) strict convexity $\Leftarrow \nabla^2 f(x) \succ 0$

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

but not strongly convex

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 $\Rightarrow \nabla^2 f(x) \succ 0$, $\forall x$.

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 : $f(x) = x^2$

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

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

$\hookrightarrow$ eigenvalue of (A^TA)

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 $\|A\| = \max\limits_{x \in \mathbb{R}^n} \dfrac{\|Ax\|}{\|x\|} = \max\limits_i \sigma_i(A)$ spectral norm of A

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

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

$\nabla f = Kx - b$

$\nabla^2 f = K$

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

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

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

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

(K is the discrete Laplacian matrix see my MA 615 notes)

So we get

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

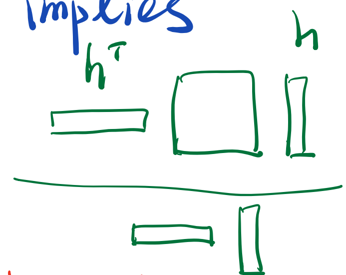
② $\lambda_1 < \lambda_2 < \cdots < \lambda_n$

$\Rightarrow \lambda_1 I \leq K \leq \lambda_n I$ 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



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

$\longrightarrow$ C-F-W minmax principle

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

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

$\|\nabla f(x) - \nabla f(y)\| \leq L \|x-y\|$

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

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

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

$\|K\| \leq L$

which implies

$-L \leq \lambda_i(K) \leq L$

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

$-L \leq \frac{h^T K h}{h^T h} \leq L$

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

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

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

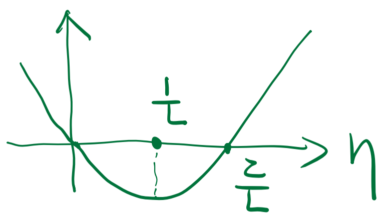
$\square$

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

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

I. To have $f(x_{k+1}) < f(x_k)$, we need

$$-\eta\left(1-\frac{L}{2}\eta\right) = \frac{L}{2}\left(\eta^2 - \frac{2}{L}\eta\right) = \frac{L}{2}\left(\eta - \frac{1}{L}\right)^2 - \frac{L}{2} < 0$$



$\Rightarrow 0 < \eta < \frac{2}{L}$

$\boxed{\text{Stability}}$

① GD $x_{k+1} = x_k - \eta \nabla f(x_k)$ with $\eta > 0$ is

numerically stable if $\eta < \frac{2}{L}$

$f(x_*) \le f(x_{k+1}) < f(x_k)$ if global minimum $f(x_*)$ exists.

② In practice it's hard to have exact L.

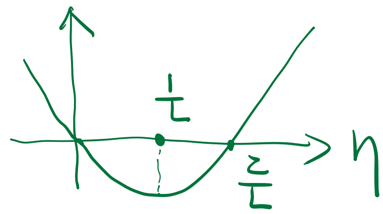
Assume ∇f is L-continuous, then

GD is stable for any $\eta \in (0, \frac{2}{L})$ with unknown L

$\Rightarrow$ GD is stable for small enough η.

II. "Best" constant step size is to minimize $-\eta(1-\frac{L}{2}\eta)$

$\eta = \frac{1}{L}$



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

$$\leq -\frac{L}{2}\|\nabla f(x_k)\|^2$$

$\min_{\eta} f(x_k - \eta \nabla f(x_k))$

"Best" only in the sense of minimizing $-\eta(1-\frac{L}{2}\eta)$

III. Convergence of Constant Step Size $\eta \in (0, \frac{2}{L})$

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

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

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

① Sum it for $k = 0, 1, 2, \dots$

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

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

Completeness Theorem of Real numbers

- monotone bounded sequence as a limit.

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

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

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

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

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

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

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

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

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

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

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

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

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

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 contant:

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

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

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

③ $\min_{0 \leq k \leq N} \|\nabla f(x_k)\| \leq \frac{1}{\sqrt{N+1}} \sqrt{\frac{1}{\omega}[f(x_0) - f(x_*)]}$

III. Convergence of $\{x_k\}$ for convex $f(x)$

Theorem If ∇f is L-continuous and $f(x)$ is convex:

(a) $f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y)$

(b) $\|\nabla f(x) - \nabla f(y)\|^2 \leq L\langle \nabla f(x) - \nabla f(y), x - y \rangle$

Theorem

Assume ∇f is L-continuous.

Assume $f(x) \geq f(x_*)$, $\forall x \in \mathbb{R}^n$

Assume $f(x)$ is convex.

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

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

$$f(x_k) - f(x_*) \leq \frac{1}{\frac{1}{f(x_0)-f(x_*)} + k\omega \frac{1}{\|x_0 - x_*\|^2}} < \frac{\|x_0 - x_*\|^2}{k\omega}$$

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

Remark: $f(x_k) - f(x_*) < \frac{\|x_0 - x_*\|^2}{\omega} \cdot \frac{1}{k}$

gives convergence rate $O(\frac{1}{k})$, under the assumptions of only convexity and L-continuity.