Convexity

- **Def** $f(x)$ is convex if $\forall x, y \in \mathbb{R}^n$, $\forall \lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda) y) \le \lambda f(x) + (1-\lambda) f(y)$$

- **Def** $f(x)$ is strictly convex if $\forall x, y \in \mathbb{R}^n$, $\forall \lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda) y) < \lambda f(x) + (1-\lambda) f(y)$$

- **Def** $f(x)$ is strongly convex if $\forall x, y \in \mathbb{R}^n$, $\forall \lambda \in (0,1)$

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

μ is a constant (strong convexity const.)

- **Fact** $f(x)$ is μ-strongly convex $\Leftrightarrow$ $f(x) - \frac{\mu}{2}\|x\|^2$ is convex

- Strong Convexity $\Rightarrow$ Strict Convexity $\Rightarrow$ 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,

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

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

$A \succeq B \iff A - B \succeq 0$

② If $\nabla^2 f(x)$ is continuous

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

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

3) Strict Convexity $\Leftarrow \nabla^2 f(x) \succ 0$

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

but not strongly convex

Taylor Theorem $f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$

$z = x + \theta(y-x)$

$h = y - x$, $\nabla^2 f(x) \succeq \mu I \Rightarrow \dfrac{h^T \nabla^2 f h}{h^T h} \geq \mu$

Courant-Fischer-Weyl Minmax Principle:

$A \in \mathbb{R}^{n\times n}$ is real symmetric with eigenvalues

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, then

$$\lambda_n \leq \frac{x^T A x}{x^T x} \leq \lambda_1 \quad , \quad \forall x \in \mathbb{R}^n$$

<u>Def</u> (Lipschitz Continuity)

$f(x)$ is Lipschitz continuous if

$\exists\, L > 0, \ \forall x, y \in \mathbb{R}^n, \ |f(x) - f(y)| \leq L \|x - y\|$.

Theorem: $\|\nabla^2 f(x)\| \leq L \ , \forall x$

$\Rightarrow \nabla f(x)$ is Lipschitz continuous

③

<u>Descent Lemma</u> Assume $\nabla f(x)$ is L-Lipschitz

① $f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|x-y\|^2$

② $f(y) \geq f(x) + \nabla f(x)^T (y-x) - \frac{L}{2}\|x-y\|^2$

Taylor Theorem $f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$

$z = x + \theta(y-x)$

Example: Assume $\|\nabla^2 f(x)\| \leq L, \ \forall x$

$$\Rightarrow |\lambda_i(\nabla^2 f)| = \sigma_i(\nabla^2 f) \leq \sigma_1 \leq L, \ \forall x$$

$$\Rightarrow \left|\frac{h^T \nabla^2 f(z) h}{h^T h}\right| \leq \max\{|\lambda_1|, |\lambda_n|\} = \sigma_1 \leq L$$

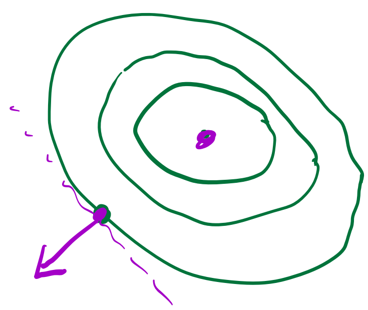
$$-\frac{1}{2}L\|y-x\|^2 \leq \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x) \leq \frac{1}{2}L\|y-x\|^2$$

$\boxed{\text{Sufficient Decrease Lemma}}$

Assume $\nabla f(x)$ is L-Lipschitz. Then $\forall x \in \mathbb{R}^n, \ \forall \eta > 0$

$$f(x) - f(x - \eta\nabla f(x)) \geq \eta\left(1 - \frac{L}{2}\eta\right)\|\nabla f(x)\|^2$$

Gradient Descent $x_{k+1} = x_k - \eta\nabla f(x_k), \ \eta > 0$



Contour lines of

$f(x, y) = (x-1)^2 + (y-2)^2$

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

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

$$= f(x) - \eta\left(1 - \frac{L}{2}\eta\right)\|\nabla f(x)\|^2.$$

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: $f(x) = x^2$

Convergence of Gradient Descent

I. $f(x)$ has Lipschitz Continuous $\nabla f(x)$

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

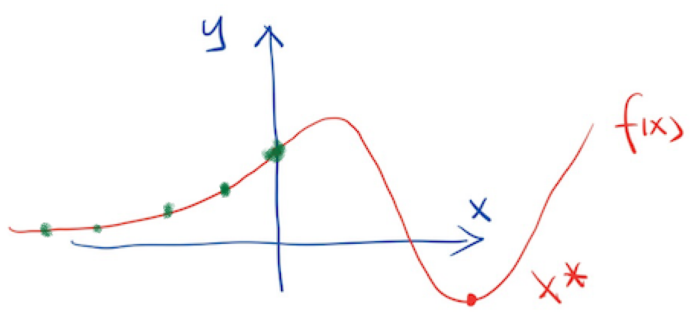
③ $\min_{0 \leq k \leq N} \|\nabla f(x_k)\| \leq \frac{1}{\sqrt{N+1}} \sqrt{\frac{1}{\omega}[f(x_0) - f(x_*)]}$

Example 2.2. *We construct an example for which the gradient descent method produces almost $\|\nabla f(\mathbf{x}_k)\| = \frac{1}{k}$. Consider the following function*

$$f(x) = \begin{cases} e^x, & x \leq 0 \\ g(x), & x > 0 \end{cases},$$

where we pick a function $g(x)$ such that

1. *$f(x)$ is very smooth;*
2. *$|f''(x)| \leq 1$ for any x, which implies $f'(x)$ is L-continuous with $L = 1$;*
3. *$f(x)$ has a global minimizer x_*.*



So a stable step size can be chosen as any positive $\eta < 2$. We consider the following gradient descent iteration with $\eta = 1$:

$$\begin{cases} x_{k+1} = x_k - f'(x_k) \\ x_0 = 0 \end{cases}.$$

Notice that all iterates x_k stays non-positive, it can also be written as

$$x_{k+1} = x_k - e^{x_k}, \quad x_0 = 0.$$

One can easily implement this on MATLAB to verify that numerically we have $|f'(x_k)| \approx \frac{1}{k}$ for this iteration.

Remark: this example shows than $\lim_{k\to 0} \|\nabla f(x_k)\| = 0$ does not imply convergence to critical point.

II. Convergence of GD for convex $f(x)$

L-cont. $\nabla f(x)$

<u>Theorem</u> 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 constant:

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

Remark 2.8. *We obtain convergence rate $\mathcal{O}(\frac{1}{k})$, assuming only convexity of the cost function and Lipschitz-continuity of its gradient. We cannot expect convergence of $\mathbf{x}_k$ to $\mathbf{x}_*$ because a convex function may have multiple global minimizers, e.g., $f(\mathbf{x}) \equiv 0$.*

III. Convergence of GD for $\begin{cases} \text{strongly convex } f(x) \\ L\text{-Cont } \nabla f(x) \end{cases}$

Theorem 2.12 (Global linear rate of gradient descent)**.** *Assume $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is strongly convex with $\mu > 0$ and $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L. Then $f(\mathbf{x})$ has a unique global minimizer: $f(\mathbf{x}) \geq f(\mathbf{x}_*), \forall \mathbf{x}$. The gradient descent method (2.1) with a constant step size $\eta \in (0, \frac{2}{L+\mu}]$ satisfies*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\|^2 \leq \left(1 - \frac{2\eta\mu L}{L+\mu}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|^2.$$

In particular, if $\eta = \frac{2}{L+\mu}$, then we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \left(\frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|,$$

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_*) \leq \frac{L}{2}\left(\frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}_*\|.$$

Remark: $\mu I \leq \nabla^2 f(x) \leq L I, \ \forall x$

$\Rightarrow \begin{cases} f(x) \text{ is strongly convex with } \mu > 0 \\ \nabla f(x) \text{ is } L\text{-Cont. with } L \end{cases}$