Lecture 12. Convergence rate of gradient descent.

Definition (L-Lipschitz) $\forall x, y.\ \|f(x)-f(y)\| \le L\|x-y\|$.

Example. $f(x)=w^Tx$ is $\|w\|$-Lipschitz. (using l_2-norm)

$$|w^T(x-y)| \le \|w\|\cdot\|x-y\| \quad \text{by Cauchy-Schwarz.}$$

Example. $f(x)=Qx$ $(\mathbb{R}^n\to\mathbb{R}^n)$ is $\lambda_{max}(Q)$-Lipschitz if $Q\succeq 0$.

$$\|f(x)-f(y)\| = \|Q(x-y)\| = \left((x-y)^TQ^2(x-y)\right)^{1/2}$$

$$\le \left(\lambda_{max}(Q^2)\|x-y\|^2\right)^{1/2} = \lambda_{max}(Q)\|x-y\|.$$

Recall that we hope $f'(x)$ or $\nabla f(x)$ not change rapidly. so:

Definition. (L-smoothness). Suppose $f:\mathbb{R}^n\to\mathbb{R}$ is differentiable.

Then f is L-smooth if ∇f is L-Lipschitz. ie. $\forall x, y\in \text{dom} f$

$$\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|.$$

Example. $f(x)=x^TQx$ with $Q\succeq 0$ is $2\lambda_{max}(Q)$-smooth.

Recall that $f'(x)=\lim_{y\to x}\frac{f(y)-f(x)}{y-x}$. f is L-Lipschitz $\Rightarrow |f'|\le L$.

$f:\mathbb{R}\to\mathbb{R}$ is twice differentiable and L-smooth $\Rightarrow |f''(x)|\le L$.

Lemma. A twice continuous differentiable $f:\mathbb{R}^n\to\mathbb{R}$ is L-smooth iff

$-LI\preceq\nabla^2 f(x)\preceq LI$. $\forall x\in\text{dom} f$ (or equivalently. $|\lambda(\nabla^2 f(x))|\le L$).

Proof. "$\Leftarrow$". Assume $-LI \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \text{dom } f$.

Fix $x, y \in \text{dom } f$. If $f: \mathbb{R} \to \mathbb{R}$. then $f'(x) - f'(y) = f''(z)(x-y)$

for some z. by the mean value theorem. However, no such theorem

for vector-value function $\nabla f(x)$ if $f: \mathbb{R}^n \to \mathbb{R}$.

Let $\varphi(t) = (\nabla f(y) - \nabla f(x))^T \nabla f(x + t(y-x))$. $\varphi: [0,1] \to \mathbb{R}$.

$\varphi(1) - \varphi(0) = (\nabla f(y) - \nabla f(x))^T (\nabla f(y) - \nabla f(x)) = \| \nabla f(y) - \nabla f(x) \|^2$.

$= \varphi'(t)$ for some $t \in (0,1)$. by the MVT.

$\varphi'(t) = (\nabla f(y) - \nabla f(x))^T \nabla^2 f(x + t(y-x))(y-x)$.

$\leq \| \nabla f(y) - \nabla f(x) \| \cdot \| \nabla^2 f(x + t(y-x)) \cdot (y-x) \|$ by the CSI.

$\Rightarrow \| \nabla f(y) - \nabla f(x) \| \leq \| \nabla^2 f(x + t(y-x)) \cdot (y-x) \| \leq L \| y - x \|$.

"$\Leftarrow$". Assume f is L-smooth. Fix $x, v \in \mathbb{R}^n$. $\psi(t) = \nabla f(x + tv)^T v$

$\Rightarrow | \psi(t) - \psi(0) | \leq \| \nabla f(x+tv) - \nabla f(x) \| \cdot \| v \| \leq t L \| v \|^2$

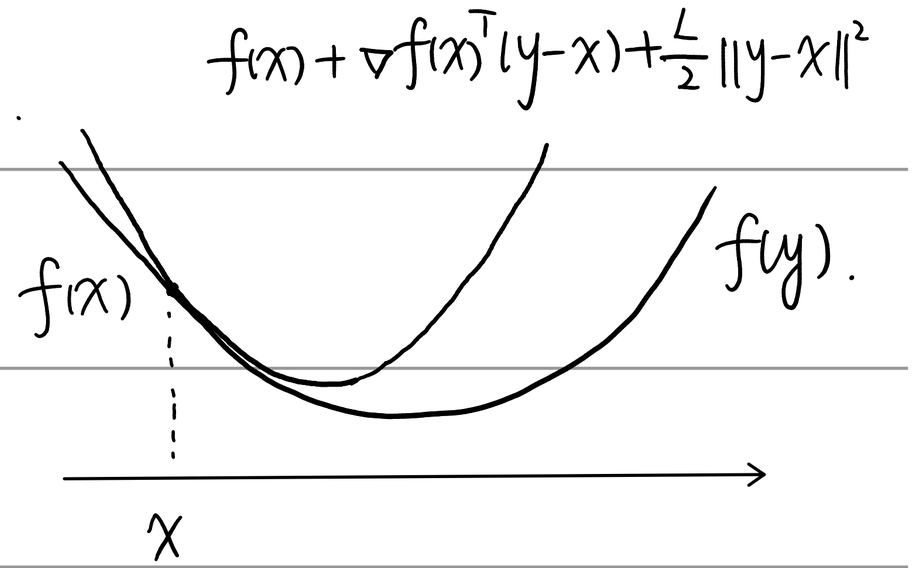
$\Rightarrow \left| \frac{\psi(t) - \psi(0)}{t} \right| \leq L \| v \|^2$ taking the limit and using the chain rule.

$| \psi'(0) | = | v^T \nabla^2 f(x) v | \leq L \| v \|^2 \Rightarrow -LI \preceq \nabla^2 f(x) \preceq LI$. □

In particular. if f is convex. then f is L-smooth iff $\nabla^2 f(x) \preceq LI$.

Lemma. If f is L-smooth. then $f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \| y - x \|^2$.

Proof. Fix x, y. Let $g(\theta) = f(x+\theta(y-x))$.



$g'(\theta) = \nabla f(x+\theta(y-x))^T (y-x)$.

$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(\theta)\, d\theta$

Since $\nabla f(x)^T(y-x) = g'(0)$. $f(y) - f(x) - \nabla f(x)^T(y-x) = \int_0^1 g'(\theta) - g'(0)\, d\theta$.

Note that $|g'(\theta) - g'(0)| = |(\nabla f(x+\theta(y-x)) - \nabla f(x))^T (y-x)|$

$\le \|\nabla f(x+\theta(y-x)) - \nabla f(x)\| \cdot \|y-x\|$ by Cauchy-Schwarz

$\le \theta \cdot L \|y-x\|^2$ by L-smoothness.

So $f(y) - f(x) - \nabla f(x)^T(y-x) \le \int_0^1 \theta L \|y-x\|^2 d\theta = \frac{L}{2}\|y-x\|^2$. □

Remark. Do not need convexity or concavity. It holds for general cases.

Now we can analyze the convergence of gradient descent method.

Easy part: we hope $f(x_{k+1}) < f(x_k)$. $f(x_{k+1}) = f(x_k - \nabla f(x_k)\cdot t)$.

So $f(x_{k+1}) \le f(x_k) - \nabla f(x_k)^T \cdot t \cdot \nabla f(x_k) + \frac{L}{2}\|t \nabla f(x_k)\|^2$ if L-smooth.

$= f(x_k) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x_k)\|^2 < f(x_k)$ if $t < 2/L$.

Difficult part: bound the convergence rate. further require $t \le 1/L$.

Then $f(x_{k+1}) \le f(x_k) - \frac{t}{2}\|\nabla f(x_k)\|^2$ need to bound $\|\nabla f(x_k)\|^2$.

$\|x_{k+1}\|^2 = \|x_k\|^2 + \|\nabla f(x_k)\|^2 t^2 - 2t \nabla f(x_k)^T x_k$. $\nabla f(x_k)^T x_k$ nothing to know

$$\|x_{k+1}-x^*\|^2 = \|x_k-x^*\|^2 + t^2\|\nabla f(x_k)\|^2 - 2t\nabla f(x_k)^T(x_k-x^*).$$

$$\leq \|x_k-x^*\|^2 + t^2\|\nabla f(x_k)\|^2 - 2t(f(x_k)-f(x^*))$$

$$\Rightarrow \|\nabla f(x_k)\|^2 \geq \frac{1}{t^2}(\|x_{k+1}-x^*\|^2 - \|x_k-x^*\|^2) + \frac{2}{t}(f(x_k)-f(x^*))$$

$$\Rightarrow \frac{t}{2}\|\nabla f(x_k)\|^2 \geq \frac{1}{2t}(\|x_{k+1}-x^*\|^2 - \|x_k-x^*\|^2) + f(x_k)-f^*$$

$$\Rightarrow f(x_{k+1}) \leq f(x_k) - \frac{1}{2t}(\|x_{k+1}-x^*\|^2 - \|x_k-x^*\|^2) - f(x_k) + f^*$$

$$= f^* - \frac{1}{2t}(\|x_{k+1}-x^*\|^2 - \|x_k-x^*\|^2)$$

$$\Rightarrow \sum_{k=0}^{T-1} f(x_{k+1}) - f^* \leq \frac{1}{2t}\sum_{k=0}^{T-1}(\|x_k-x^*\|^2 - \|x_{k+1}-x^*\|^2) \leq \frac{1}{2t}\|x_0-x^*\|^2.$$

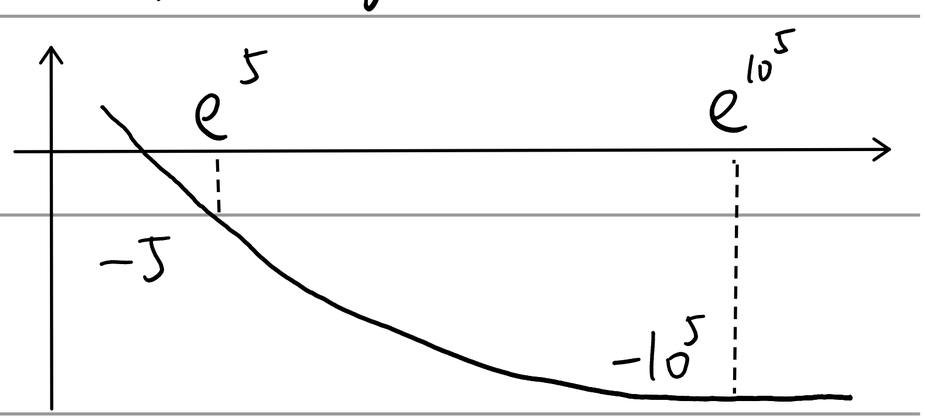
$$\Rightarrow f(x_T)-f^* \leq \frac{\|x_0-x^*\|^2}{2tT}. \quad \text{since } f(x_k) > f(x_{k+1}) \geq f^*. \qquad \square$$

Theorem. Suppose f is convex and L-smooth. Let $0 < t \leq 1/L$. Then

$$f(x_T)-f^* \leq \frac{\|x_0-x^*\|^2}{2tT}. \quad \text{where} \quad x_{k+1} = x_k - t\nabla f(x_k)$$

Remark. convergence rate is $O(1/T)$; $O(1/\varepsilon)$ steps to get ε-approximation.

Consider function $f(x) = \begin{cases} -\log x & x < e^{10^5} \\ -10^5 & x \geq e^{10^5} \end{cases}$



$f'(x) = -\frac{1}{x}$. f is 1-smooth if $x \geq 1$.

$$x_{k+1} = x_k + t/x_k < x_k + 1/x_k \Rightarrow x_{k+1}^2 < x_k^2 + 2 + 1/x_k^2 < x_k^2 + 3.$$

To get $f(x_T) < f^* + \varepsilon = -10^5 + \varepsilon$. $x_T > e^{10^5-\varepsilon} \Rightarrow T > \frac{1}{3}(e^{10^5-\varepsilon})^2$.

Good example: $f(x) = x^2$. $x_{k+1} = x_k(1-2t)$. $f(x_T) = x_0^2(1-2t)^{2T}$.

Definition (m-strong convexity). f is strongly convex with $m > 0$. or m-strongly convex. if $g(x) = f(x) - \frac{m}{2}\|x\|^2$ is convex.

Remark. $\|x\|^2$ is not significant. $\forall y \in \mathbb{R}^n$ fixed. $g(x)$ convex iff $f(x) - \frac{m}{2}\|x-y\|^2 = f(x) - \frac{m}{2}\|x\|^2 + m y^T x - \frac{m}{2}\|y\|^2$ convex.

Proposition. Suppose f is a twice continuously differentiable function. Then f is m-strongly convex iff $\nabla^2 f(x) \succeq mI$. i.e. $\lambda_{min}(\nabla^2 f(x)) \geq m$.

Proof. Apply second-order condition for convexity to $g(x)$. □.

Lemma. Suppose f is a differentiable function. Then f is m-strongly convex iff $\forall x, y.\ f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|y-x\|^2$.

Proof. $g(x) = f(x) - \frac{m}{2}\|x\|^2$ convex $\iff g(y) \geq g(x) + \nabla g(x)^T (y-x)$.

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

$\iff f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}(\|y\|^2 + \|x\|^2 - 2x^T y)$ □.

Remark. L-smoothness $\Rightarrow f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2$.

Example. $-\log x$. x^4 are not strongly convex.

$f(x) = w^T x$ is not strongly convex.

$f(x) = x^T Q x$ is $2\lambda_{min}(Q)$-strongly convex. if $Q > 0$.

Consider $f(x) = \frac{m}{2} x^2$. $m > 0$. $x_{k+1} = x_k(1 - mt) = (1-mt)^{k+1} x_0$.

Theorem. If f is m-strongly convex and L-smooth. Fix $t \le 1/L$ and let $x^* = \operatorname{argmin} f$. Suppose $\{x_k\}$ is given by the gradient descent method. Then $\|x_T - x^*\|^2 \le (1-mt)^T \|x_0 - x^*\|^2$.

Remark. m-strong convexity $\Rightarrow$ strict convexity. so x^* is unique.

$$f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|y-x\|^2 > f(x) + \nabla f(x)^T (y-x).$$

Remark It further gives $f(x_T) - f^* \le \frac{L}{2}(1-mt)^T \|x_0 - x^*\|^2$ since

$f(x_T) \le f(x^*) + \nabla f(x^*)^T (x_T - x^*) + \frac{L}{2}\|x_T - x^*\|^2$ by L-smoothness.