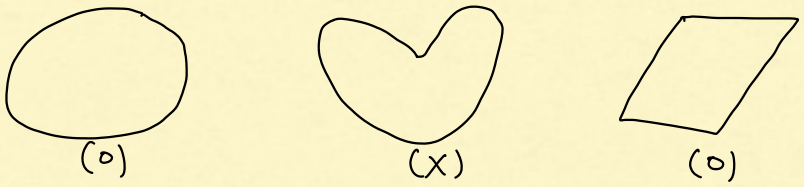
Opt Lec 1.

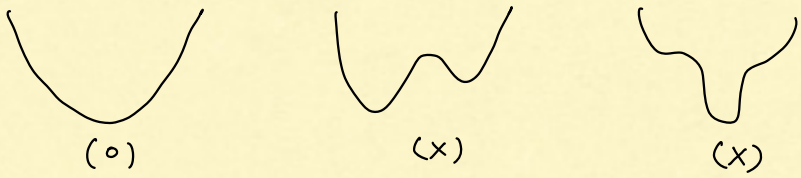
- Convexity

1. Convex set(C) : When $x \in C, y \in C$, then $\alpha x + (1-\alpha) y \in C \quad \forall \alpha \in [0,1]$



(O) (X) (O)

2. Convex functions (f) : $\alpha f(x_1) + (1-\alpha) f(x_2) \geq f(\alpha x_1 + (1-\alpha) x_2) \quad \forall x_1, x_2, \forall \alpha \in [0,1]$



(O) (X) (X)

$\Rightarrow$ { linear functions ($a^T x$), affine functions ($a^T x + b$), norm
exponential functions ($e^{\alpha x}$), negative log functions ($-\log \beta x$) }

☆ f is a convex function (①) iff its epigraph is a convex set (②)
(where epigraph is $\{(x,t) \mid f(x) \leq t\}$)

pf) ①⇒② $f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \leq \lambda t_1 + (1-\lambda) t_2$ ($\because \begin{matrix} f(x_1) \leq t_1 \\ f(x_2) \leq t_2 \end{matrix}$)

②⇒① $f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ ($\because \{(x_1, f(x_1)), (x_2, f(x_2))\} \subset epi(f)$)

☆ Convexity is preserved with addition and max operator between convex functions and non-negative scalar multiplication is okay, too.

3. Norms (p) : non-negative-valued scalar functions on a given vector space V.

satisfying the following conditions ($p : V \to [0, +\infty)$)

① $p(x) + p(y) \geq p(x+y)$ for all $x, y \in V$ (triangular inequality)

② $p(ax) = |a| p(x)$ (homogeneous or scalable)

③ $p(x) = 0$ if and only if $x = 0$

(semi-norm satisfies only ① and ②)

* Frequently used norms

→ 0-norm : $\|x\|_0 = \#$ of nonzero element

→ 1-norm : $\|x\|_1 = \sum_{i=1}^{d} |x_i|$

→ 2-norm : $\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$

→ max-norm : $\|x\|_\infty = \max_i |x_i|$

$\Rightarrow$ p-norm : $\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}$

★ Every norm is convex

pf) $\alpha p(x) + (1-\alpha) p(y) = p(\alpha x) + p((1-\alpha)y) \geq p(\alpha x + (1-\alpha)y)$

$\therefore \alpha p(x) + (1-\alpha) p(y) \geq p(\alpha x + (1-\alpha)y)$

4. Jensen's inequality.

$\Rightarrow$ Let f is convex, $x_1, \cdots, x_m \in dom(f)$, $\lambda_1, \cdots, \lambda_m \in R_+$ s.t. $\sum_{i=1}^{m} \lambda_i = 1$

Then, $f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \leq \sum_{i=1}^{m} \lambda_i f(x_i)$ $\left(\Leftrightarrow f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]\right)$

pf) Mathematical Induction

Step 1. Show $m=2$ is correct

$\Rightarrow f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2)$ ($\because f$ is convex)

Step 2. Show $m=k+1$ is correct assuming $m=k$ is correct

$$\Rightarrow f\left(\sum_{i=1}^{k+1} \lambda_i x_i\right) = f\left((1-\lambda_{k+1}) \sum_{i=1}^{k} \frac{\lambda_i x_i}{1-\lambda_{k+1}} + \lambda_{k+1} x_{k+1}\right)$$

$$\leq (1-\lambda_{k+1}) f\left(\sum_{i=1}^{k} \frac{\lambda_i x_i}{1-\lambda_{k+1}}\right) + \lambda_{k+1} f(x_{k+1})$$

$$\leq \cancel{(1-\lambda_{k+1})} \sum_{i=1}^{k} \frac{\lambda_i}{\cancel{1-\lambda_{k+1}}} f(x_i) + \lambda_{k+1} f(x_{k+1})$$

$$= \sum_{i=1}^{k+1} \lambda_i f(x_i)$$

$\therefore f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \leq \sum_{i=1}^{m} \lambda_i f(x_i)$

• Young's inequality

When $1<p<\infty$, $a>0, b>0$, then $ab \le \frac{p-1}{p} a^{\frac{p}{p-1}} + \frac{1}{p} b^p$

(When $p=2$, it turns to Cauchy-Schwarz inequality)

pf) $ab = \exp(\log a + \log b)$

$= \exp\left(\frac{p-1}{p} \cdot \frac{p}{p-1} \log a + \frac{1}{p} p \cdot \log b\right)$

$\le \frac{p-1}{p} \exp\left(\frac{p}{p-1} \log a\right) + \frac{1}{p} \exp(p \log b)$ ($\because$ convexity)

$= \frac{p-1}{p} \cdot a^{\frac{p}{p-1}} + \frac{1}{p} b^p$

• Hölder's inequality

When $x, y \in V$, then $x^T y \le \|x\|_q \|y\|_p$ where $\frac{1}{p} + \frac{1}{q} = 1$

pf) $\sum_{i=1}^{\infty} \frac{x_i}{\|x\|_q} \cdot \frac{y_i}{\|y\|_p} \le \sum_{i=1}^{\infty} \frac{|x_i|}{\|x\|_q} \cdot \frac{|y_i|}{\|y\|_p}$

$\le \sum_{i=1}^{\infty} \left(\frac{1}{q} \frac{|x_i|^q}{\|x\|_q^q} + \frac{1}{p} \cdot \frac{|y_i|^p}{\|y\|_p^p} \right)$ ($\because$ Young's inequality)

$= \frac{1}{q} + \frac{1}{p} = 1$

$\Rightarrow \sum_{i=1}^{\infty} x_i \cdot y_i \le \|x\|_q \cdot \|y\|_p$

• Minkowski's inequality

When $1<p<\infty$, $x, y \in V$, then $\|x+y\|_p \le \|x\|_p + \|y\|_p$

pf) $(\|x+y\|_p)^p = \sum_i |x_i + y_i|^p$

$= \sum_i |x_i + y_i| \cdot |x_i + y_i|^{p-1}$

$= \sum_i (|x_i| + |y_i|) |x_i + y_i|^{p-1}$

$= \sum_i \left(|x_i| \cdot |x_i + y_i|^{p-1} + |y_i| \cdot |x_i + y_i|^{p-1} \right)$

$\le \left\{ \left(\sum_i |x_i|^p \right)^{\frac{1}{p}} + \left(\sum_i |y_i|^p \right)^{\frac{1}{p}} \right\} \left(\sum_i |x_i + y_i|^{(p-1)\cdot\frac{p}{p-1}} \right)^{\frac{p-1}{p}}$ ($\because$ Hölder's inequality)

$= (\|x\|_p + \|y\|_p) \cdot (\|x+y\|_p)^{p-1}$

$\therefore \|x+y\|_p \le \|x\|_p + \|y\|_p$

Opt Lec 2.

- Convex optimization

1. Convex optimization problem

$$\Rightarrow \min \underbrace{f(x)}_{\text{convex function}} \quad \text{s.t.} \quad \underbrace{x \in C}_{\text{convex set}}$$

2. Convergence rate ($g(t)$)

$$\Rightarrow \frac{f(x_t) - f(x^*)}{g(t)} \leq \text{constant}$$

ex1) $g(t) = \frac{1}{t} \rightarrow f(x_t) - f(x^*) \leq \varepsilon$ when $t \geq \frac{1}{\varepsilon}$

ex2) $g(t) = \frac{1}{\sqrt{t}} \rightarrow f(x_t) - f(x^*) \leq \varepsilon$ when $t \geq \frac{1}{\varepsilon^2}$

ex3) $g(t) = e^{-t} \rightarrow f(x_t) - f(x^*) \leq \varepsilon$ when $t \geq -\ln \varepsilon$

3. Differentiable

① f is differentiable at x_0

$\Rightarrow$ there exists a linear map $J: \mathbb{R}^m \rightarrow \mathbb{R}^n$ such that

$$\lim_{h \rightarrow 0} \frac{\| f(x_0 + h) - f(x_0) - J(h) \|}{\| h \|} = 0$$

② f is differentiable

$$\Rightarrow \nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_d} \right)^T \text{ exists at every point } x \in \text{dom}(f)$$

③ f is twice differentiable

$$\Rightarrow \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d} \end{bmatrix} \text{ exists at every point } x \in \text{dom}(f)$$

4. Tangent hyperplane of f at x_0 ($g(x)$)

$\Rightarrow g(x) = f(x_0) + \nabla f(x_0)^T (x - x_0)$

5. Local minima(x): there exist $\varepsilon > 0$ s.t. $f(x) \le f(y)$ $\forall y$ s.t. $\|y - x\| \le \varepsilon$

6. Critical point(x): $\nabla f(x) = 0$

(When f is convex, a critical point is a global minima)

★ 1st order characterization of convexity

$\Rightarrow$ Let $\mathrm{dom}(f)$ is open, f is differentiable

Then, f is convex (①) iff $\mathrm{dom}(f)$ is convex & $f(y) \ge f(x) + \nabla f(x)^T (y - x)$ $\forall x, y \in \mathrm{dom}(f)$ (②)

(if f is non-differentiable, $\nabla f(x) \approx g(x) \in \partial f(x)$)

(f is lower-bounded by its tangent hyperplane for every point in domain)

pf) ① $\Rightarrow$ ② $\lambda f(y) + (1-\lambda) f(x) \ge f(\lambda y + (1-\lambda)x) = f(\lambda(y-x) + x)$

$\Rightarrow \lambda f(y) \ge \lambda f(x) + f(x + \lambda(y-x)) - f(x)$

$\Rightarrow f(y) \ge f(x) + \dfrac{f(x + \lambda(y-x)) - f(x)}{\lambda(y-x)} \cdot (y-x)$

$\Rightarrow f(y) \ge f(x) + \nabla f(x)^T (y-x)$ when $\lambda \to 0$

② $\Rightarrow$ ① Let $z = (1-\lambda)x + \lambda y$

$f(x) \ge f(z) + \nabla f(z)^T (x - z)$ $\times (1-\lambda)$

$+) \; f(y) \ge f(z) + \nabla f(z)^T (y - z)$ $\times \lambda$

$(1-\lambda) f(x) + \lambda f(y) \ge f(z) + \nabla f(z)^T ((1-\lambda)x + \lambda y - z)$

$\Rightarrow (1-\lambda) f(x) + \lambda f(y) \ge f((1-\lambda)x + \lambda y)$

★ 2nd order characterization of convexity

⇒ Let $\mathrm{dom}(f)$ is open, f is twice differentiable

Then, f is convex (①) iff $\mathrm{dom}(f)$ is convex & $\nabla^2 f(x)$ is positive semidefinite $\forall x \in \mathrm{dom}(f)$ (②)

(A is symmetric & positive semidefinite : $x^TAx = x^TPDP^Tx = y^TDy \geq 0 \quad \forall x \in \mathbb{R}^d$

⇒ eigenvalues are all non-negative

diagonalizable with orthogonal matrix (P) → $P^T = P^{-1}$

pf) ① ⇒ ② $g(y) := f(y) - f(x) - \nabla f(x)^T(y-x)$ ⟶ convex + affine = convex

⇒ $\nabla g(y) = \nabla f(y) - \nabla f(x)$ ⟶ $\nabla g(x) = 0$ ⇒ (x is a global minimizer of g ; $\nabla^2 g(x)$ is positive semi definite ($\because$ 2nd order necessity condition for a minimizer))

⇒ $\nabla^2 g(y) = \nabla^2 f(y)$ ⟶ $\nabla^2 g(x) = \nabla^2 f(x)$

⇒ $\nabla^2 f(x)$ is positive semi definite $\forall x$ since x was arbitrary

② ⇒ ① $f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x + t(y-x))(y-x)$ for some $t \in [0,1]$

($\geq 0 \because \nabla^2 f$ is p.s.d.)

⇒ $f(y) \geq f(x) + \nabla f(x)^T(y-x)$

★ Taylor theorem

⇒ $f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(c)(y-x)$

where $c = x + (y-x)t$, $t \in [0,1]$

($\frac{1}{2}\alpha \|y-x\|^2 \leq \frac{1}{2}(y-x)^T \nabla^2 f(c)(y-x) \leq \frac{1}{2}\beta \|y-x\|^2$)

pf) $f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T M (y-x)$

Let $g(t) = -f(y) + f(t) + \nabla f(t)^T(y-t) + \frac{1}{2}(y-t)^T M (y-t)$

⇒ $g(x) = g(y) = 0$

⇒ $\exists t \in [0,1]$ s.t. $c = x + t(y-x)$, $\nabla g(c) = 0$ ($\because$ Rolle's theorem)

⇒ $M = \nabla^2 f(c)$

✡ Constrained minimization

⇒ Let f is convex, χ is a convex set

Then, x^* is a minimizer of f over χ ① iff $\nabla f(x^*)^T (y - x^*) \geq 0$ ②

($f(x^*) \leq f(y) \quad \forall y \in \chi$)

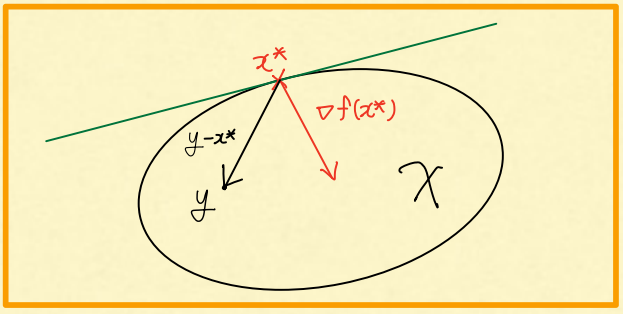
pf.) ① ⇒ ②

Case 1. x^* is inside χ

$\Rightarrow \nabla f(x^*) = 0$

Case 2. x^* is on the boundary of χ

$\Rightarrow \nabla f(x^*)(x - x^*) \geq 0$



② ⇒ ① $f(y) \geq f(x^*) + \underbrace{\nabla f(x^*)(y - x^*)}_{(\geq 0)} \geq f(x^*)$

Opt Lec 3.

- Gradient descent

1. Gradient descent algorithm.

$\Rightarrow x_{t+1} = x_t - \gamma \nabla f(x_t)$

(γ could be a function of $t \Rightarrow \gamma_t$)

2. Convergence rate (t)

$\Rightarrow$ find t such that $f(x_t) - f(x^*) \le \varepsilon$

3. L-Lipschitz continuous (f)

$\Rightarrow \|f(x) - f(y)\| \le L \|x - y\| \quad \forall x, y$

(When f is differentiable, then $\|\nabla f(x)\| \le L \ \forall x$)

4. β-smooth (f) $\not\Longrightarrow$ convex

$\Rightarrow \|\nabla f(x) - \nabla f(y)\| \le \beta \|x - y\|$

($\Rightarrow f(x) - f(y) \ge \nabla f(x)^T (x - y) - \frac{\beta}{2} \|x - y\|^2 \quad \forall x, y$)

(When f is twice differentiable, then $\|\nabla^2 f(x)\| \le \beta \ \forall x$)

pf

★ $\nabla^2 f(x) \preceq \beta I \Rightarrow$ eigenvalues of Hessian are less than equal to β

$\|\nabla f(x) - \nabla f(y)\| \le \beta \|x - y\|$

$\Rightarrow (\nabla f(x) - \nabla f(y))^T (x - y) \le \|\nabla f(x) - \nabla f(y)\| \|x - y\| \le \beta \|x - y\|^2$

Let $g(t) = f(x + (y - x)t)$

$\Rightarrow g'(t) - g'(0) = \left(\nabla f(x + (y - x)t) - \nabla f(x)\right)^T (y - x) \le \frac{1}{t} \beta \|(y - x)t\|^2 = \beta t \|y - x\|^2$

$\Rightarrow f(y) = g(1) = g(0) + \int_0^1 g'(t)\, dt$

$\le f(x) + \int_0^1 g'(0) + \beta t \|y - x\|^2 \, dt = f(x) + \int_0^1 \nabla f(x)^T (y - x) + \beta t \|y - x\|^2 dt$

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

5. α-strongly convex (f) ⟶ convex

$\Rightarrow f(x) - f(y) \le \nabla f(x)^T (x-y) - \frac{\alpha}{2}\|x-y\|^2 \quad \forall x, y$

pf

$(\Rightarrow \|\nabla f(x) - \nabla f(y)\| \ge \alpha \|x-y\|)$

(When f is twice differentiable, then $\|\nabla^2 f(x)\| \ge \alpha \ \forall x$)

★ $\nabla^2 f(x) \succeq \alpha I \Rightarrow$ eigenvalues of Hessian are greater than equal to α

$$f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{\alpha}{2}\|y-x\|^2$$
$$+) \ f(x) \ge f(y) + \nabla f(y)^T (x-y) + \frac{\alpha}{2}\|x-y\|^2$$
$$(\nabla f(x) - \nabla f(y))^T (x-y) \ge \alpha \|x-y\|^2 \Rightarrow \|\nabla f(x) - \nabla f(y)\| \ge \alpha \|x-y\|$$

★ $\alpha I \preceq \nabla^2 f(x) \preceq \beta I$ if f is α-strongly convex and β-smooth

pf)

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

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

$\Rightarrow \alpha\|x-y\|^2 \le (\nabla f(x) - \nabla f(y))^T (x-y) \le \beta \|x-y\|^2$

Let $x = y + ht$

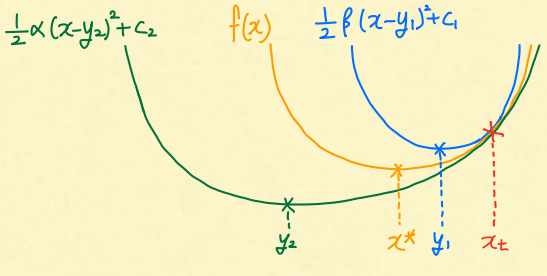
$\Rightarrow \alpha \cdot \|ht\|^2 \le (\nabla f(y+ht) - \nabla f(y))^T \cdot ht \le \beta \cdot \|ht\|^2$

$\Rightarrow \alpha \cdot \|h\|^2 \le \left(\frac{\nabla f(y+ht) - \nabla f(y)}{t}\right)^T h \le \beta \cdot \|h\|^2$

$\Rightarrow h^T \alpha I h \le h^T \cdot \nabla^2 f(y) \cdot h \le h^T \beta I h$ as $t \to 0$

⟶ $\nabla^2 f(y) - \alpha I$ is positive semi-definite

⟶ $\beta I - \nabla^2 f(y)$ is positive semi-definite



★ $y_2 \le x^* \le y_1$

$(y_1 = x_t - \frac{1}{\beta}\nabla f(x_t))$

$(y_2 = x_t - \frac{1}{\alpha}\nabla f(x_t))$

6. Polyak-Łojasiewicz inequality

$\Rightarrow f(x) - f(x^*) \le \frac{1}{2\alpha}\|\nabla f(x)\|^2 \quad \forall x \in \chi$

(α-strongly convex function satisfies.)

pf) $f(x) - f(x^*) \le \nabla f(x)^T (x - x^*) - \frac{\alpha}{2}\|x - x^*\|^2$

$\le \|\nabla f(x)\|\,\|x - x^*\| - \frac{\alpha}{2}\|x - x^*\|^2$ ($\because$ Hölder's inequality)

$= -\frac{\alpha}{2}\left(\|x - x^*\| - \frac{1}{\alpha}\|\nabla f(x)\|\right)^2 + \frac{1}{2\alpha}\|\nabla f(x)\|^2$

$\le \frac{1}{2\alpha}\|\nabla f(x_t)\|^2$

7. Linear convergence (Sequence of $f(x_0), f(x_1), \cdots$)

$\Rightarrow$ there exists a constant $c \in (0,1)$ such that

$$\lim_{t\to\infty} \frac{f(x_{t+1}) - f(x^*)}{f(x_t) - f(x^*)} = c$$

(When $\frac{f(x_{t+1}) - f(x^*)}{f(x_t) - f(x^*)} \le c \;\; \forall t$, then $f(x_t) - f(x^*) \le c^t (f(x_0) - f(x^*))$)

✡ Convergence rate for gradient descent

* Vanilla analysis

$\Rightarrow f(x_t) - f(x^*) \le \nabla f(x_t)^T (x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^T (x_t - x^*)$

$\le \frac{1}{2\gamma}\left(\|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$

$= \frac{\gamma}{2}\|\nabla f(x_t)\|^2 + \frac{1}{2\gamma}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$

$\Rightarrow \sum_{t=0}^{T-1}\left(f(x_t) - f(x^*)\right) \le \sum_{t=0}^{T-1}\frac{\gamma}{2}\|\nabla f(x_t)\|^2 + \frac{1}{2\gamma}\left(\underbrace{\|x_0 - x^*\|^2}_{(:= R)} - \|x_T - x^*\|^2\right)$

① L-Lipschitz continuous

- $f(\bar{x}) - f(x^*) \le \frac{1}{T}\sum_{t=0}^{T-1} f(x_t) - f(x^*)$ ($\because$ Jensen's inequality)

$\le \frac{\gamma}{2} \cdot L^2 + \frac{1}{2\gamma T}R^2$ ($\because$ Vanilla analysis)

$\le \frac{RL}{\sqrt{T}}$ ($\because \gamma^* = \frac{R}{L\sqrt{T}}$) ← minimization

$\therefore O\left(\frac{1}{\epsilon^2}\right)$ steps

② β-smooth

- $f(x_t) - f(x_{t+1}) \geq \nabla f(x_t)^T (x_t - x_{t+1}) - \frac{\beta}{2}\|x_t - x_{t+1}\|^2$ ($\because \beta$-smooth)

$$= \left(\gamma - \frac{\beta\gamma^2}{2}\right)\|\nabla f(x_t)\|^2$$

(non-increasing) $\geq \frac{1}{2\beta}\|\nabla f(x_t)\|^2$ ($\because \gamma^* = \frac{1}{\beta}$) ↘ maximization

(Since f is non-increasing, set γ to maximize the amount of decrease)

- $\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \sum_{t=0}^{T-1} \frac{1}{2\beta}\|\nabla f(x_t)\|^2 + \frac{\beta}{2}R^2$ ($\because$ Vanilla analysis)

$$\leq \sum_{t=0}^{T-1}\left(f(x_t) - f(x_{t+1})\right) + \frac{\beta}{2}R^2$$

$$= f(x_0) - f(x_T) + \frac{\beta}{2}R^2$$

- $f(x_T) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f(x^*)) \leq \frac{\beta}{2T}R^2$ ($\because$ non-increasing)

$\therefore O(\frac{1}{\varepsilon})$ steps

③ α-strongly convex & β-smooth

- $\nabla f(x_t)^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\alpha}{2}\|x_t - x^*\|^2$ ($\because \alpha$-strongly convex)
- $\nabla f(x_t)^T(x_t - x^*) \leq \frac{\gamma}{2}\|\nabla f(x_t)\|^2 + \frac{1}{2\gamma}(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$ ($\because$ vanilla analysis)

$\therefore f(x_t) - f(x^*) - \frac{\gamma}{2}\|\nabla f(x_t)\|^2 \leq -\frac{\alpha}{2}\|x_t - x^*\|^2 + \frac{1}{2\gamma}(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$

- RHS $\geq$ LHS $\geq f(x_t) - f(x_{t+1}) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2$ ($\because$ non-increasing f)

$$\geq \frac{1}{2\beta}\|\nabla f(x_t)\|^2 - \frac{1}{2\beta}\|\nabla f(x_t)\|^2 = 0$$ ($\because \beta$-smooth)

$\therefore \|x_{t+1} - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)\|x_t - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^{t+1}\|x_0 - x^*\|^2$

- $f(x_T) - f(x^*) \leq \nabla f(x^*)^T(x_T - x^*) + \frac{\beta}{2}\|x_T - x^*\|^2 \leq \frac{\beta}{2}\left(1 - \frac{\alpha}{\beta}\right)^T R^2$

$\therefore O(\log\frac{1}{\varepsilon})$ steps

- $f(x_t) - f(x_{t+1}) \geq \frac{1}{2\beta}\|\nabla f(x_t)\|^2$ ($\because \beta$-smooth)

$$\geq \frac{1}{2\beta} \cdot 2\alpha\left(f(x_t) - f(x^*)\right)$$ ($\because$ Polyak Łojasiewicz inequality)

$\Rightarrow f(x_{t+1}) - f(x^*) \leq \left(1 - \frac{\alpha}{\beta}\right)(f(x_t) - f(x^*))$ ($\because$ subtract $f(x^*)$ on both side)

$\Rightarrow f(x_T) - f(x^*) \leq \left(1 - \frac{\alpha}{\beta}\right)^T(f(x_0) - f(x^*))$

Opt Lec 4.

- Projected Gradient descent

1. Constrained Optimization problem ⟶ Convex constrained problem
   (when f is a convex function, $\mathcal{X}$ is a closed convex set)

   ⇒ minimize $f(x)$ s.t. $x \in \mathcal{X}$.

   (① Projected Gradient Descent
   ② Transform it into unconstrained version)

2. Projected Gradient Descent

   ⇒ $x_{t+1} = \Pi_{\mathcal{X}}(x_t - \gamma \nabla f(x_t)) = \arg\min_{x \in \mathcal{X}} \| x - (x_t - \gamma \nabla f(x)) \|^2$

3. Property of projection.

   ① $(x - \Pi_{\mathcal{X}}(y))^T (y - \Pi_{\mathcal{X}}(y)) \le 0$ where $x \in \mathcal{X}$, $y \in \mathbb{R}^d$

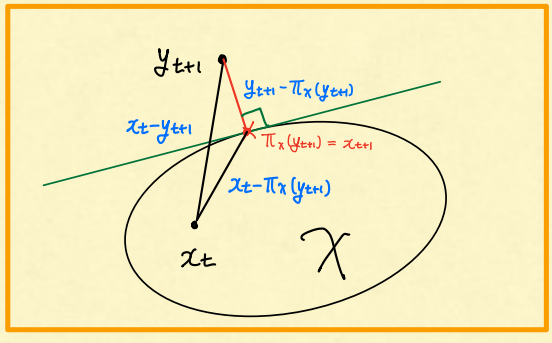
   pf) Let $g(x) = \|x - y\|^2$ ⟶ convex

   ⇒ $\Pi_{\mathcal{X}}(y)$ is a minimizer of $g(x)$

   ⇒ $\nabla g(\Pi_{\mathcal{X}}(y))^T (x - \Pi_{\mathcal{X}}(y)) \ge 0$

   ⇒ $2 (\Pi_{\mathcal{X}}(y) - y)^T (x - \Pi_{\mathcal{X}}(y)) \ge 0$

   ∴ $(x - \Pi_{\mathcal{X}}(y))^T (y - \Pi_{\mathcal{X}}(y)) \le 0$



   ② $\|x - \Pi_{\mathcal{X}}(y)\|^2 + \|y - \Pi_{\mathcal{X}}(y)\|^2 \le \|x - y\|^2$

   pf) $(x - \Pi_{\mathcal{X}}(y))^T (y - \Pi_{\mathcal{X}}(y)) = \frac{1}{2}\left(\|x - \Pi_{\mathcal{X}}(y)\|^2 + \|y - \Pi_{\mathcal{X}}(y)\|^2 - \|x - y\|^2\right) \le 0$

   ∴ $\|x - \Pi_{\mathcal{X}}(y)\|^2 + \|y - \Pi_{\mathcal{X}}(y)\|^2 \le \|x - y\|^2$

✡ Convergence rate for projected gradient descent

* vanilla analysis

$$\Rightarrow f(x_t)-f(x^*) \le \nabla f(x_t)^T(x_t-x^*) = \frac{1}{\gamma}(x_t-y_{t+1})^T(x_t-x^*)$$

$$\le \frac{1}{2\gamma}\left(\|x_t-y_{t+1}\|^2+\|x_t-x^*\|^2-\|y_{t+1}-x^*\|^2\right)$$

$$= \frac{\gamma}{2}\cdot\|\nabla f(x_t)\|^2+\frac{1}{2\gamma}\left(\|x_t-x^*\|^2-\|x^*-y_{t+1}\|^2\right)$$

$$\le \frac{\gamma}{2}\|\nabla f(x_t)\|^2+\frac{1}{2\gamma}\left(\|x_t-x^*\|^2-\|x^*-x_{t+1}\|^2-\|y_{t+1}-x_{t+1}\|^2\right) \quad (\because \text{2nd property})$$

$$\Rightarrow \sum_{t=0}^{T-1}\left(f(x_t)-f(x^*)\right) \le \sum_{t=0}^{T-1}\frac{\gamma}{2}\|\nabla f(x_t)\|^2+\frac{1}{2\gamma}\left(\|x_0-x^*\|^2-\|x_T-x^*\|^2-\sum_{t=0}^{T-1}\|y_{t+1}-x_{t+1}\|^2\right)$$

$(:= R^2)$

① L-Lipschitz continuous

- $f(\bar{x})-f(x^*) \le \frac{1}{T}\sum_{t=0}^{T-1}\left(f(x_t)-f(x^*)\right)$ $(\because$ Jensen's inequality$)$

$$\le \frac{\gamma}{2}\cdot L^2+\frac{1}{2\gamma T}R^2 \quad (\because \text{Vanilla analysis})$$

$$\le \frac{RL}{\sqrt{T}} \quad \left(\because \gamma^* = \frac{R}{L\sqrt{T}}\right) \quad \text{minimization}$$

$\therefore O\left(\frac{1}{\varepsilon^2}\right)$ steps

② β-smooth

- $f(x_t)-f(x_{t+1}) \ge \nabla f(x_t)^T(x_t-x_{t+1})-\frac{\beta}{2}\|x_t-x_{t+1}\|^2$

$$= \beta(x_t-y_{t+1})^T(x_t-x_{t+1})-\frac{\beta}{2}\|x_t-x_{t+1}\|^2 \quad (\gamma := \tfrac{1}{\beta})$$

$$= \frac{\beta}{2}\left(\|x_t-y_{t+1}\|^2+\|x_t-x_{t+1}\|^2-\|y_{t+1}-x_{t+1}\|^2\right)-\frac{\beta}{2}\|x_t-x_{t+1}\|^2$$

$$= \frac{\beta}{2}\left(\|x_t-y_{t+1}\|^2-\|y_{t+1}-x_{t+1}\|^2\right)$$

$$= \frac{1}{2\beta}\|\nabla f(x_t)\|^2-\frac{\beta}{2}\|y_{t+1}-x_{t+1}\|^2 \quad (\text{yet, can not guarantee non-increasing})$$

$$\ge \frac{1}{2\beta}\|\nabla f(x_t)\|^2-\frac{\beta}{2}\|y_{t+1}-x_t\|^2+\frac{\beta}{2}\|x_t-x_{t+1}\|^2 \quad (\because \text{2nd property})$$

$$= \frac{1}{2\beta}\|\nabla f(x_t)\|^2-\frac{1}{2\beta}\|\nabla f(x_t)\|^2+\frac{\beta}{2}\|x_t-x_{t+1}\|^2$$

$$(\text{non-increasing}) \quad = \frac{\beta}{2}\|x_t-x_{t+1}\|^2 \ge 0$$

- $\sum_{t=0}^{T-1}\left(f(x_t)-f(x^*)\right) \le \sum_{t=0}^{T-1}\frac{1}{2\beta}\|\nabla f(x)\|^2+\frac{\beta}{2}\left(R^2-\sum_{t=0}^{T-1}\|y_{t+1}-x_{t+1}\|^2\right)$ $(\because$ vanilla analysis$)$

$$\le \sum_{t=0}^{T-1}\left(f(x_t)-f(x_{t+1})\right)+\frac{\beta}{2}R^2$$

$$= f(x_0)-f(x_T)+\frac{\beta}{2}R^2$$

- $f(x_T)-f(x^*) \le \frac{1}{T}\sum_{t=1}^{T}\left(f(x_t)-f(x^*)\right) \le \frac{\beta}{2T}R^2$ $(\because$ non-increasing$)$

$\therefore O(\frac{1}{\varepsilon})$ steps

③ α-strongly convex & β-smooth

- $\nabla f(x_t)^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\alpha}{2}\|x_t - x^*\|^2$ ($\because$ α-strongly convex)
- $\nabla f(x_t)^T(x_t - x^*) \leq \frac{\gamma}{2}\|\nabla f(x_t)\|^2 + \frac{1}{2\gamma}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 - \|y_{t+1} - x_{t+1}\|^2\right)$

$\therefore f(x_t) - f(x^*) - \frac{\gamma}{2}\|\nabla f(x_t)\|^2 \leq -\frac{\alpha}{2}\|x_t - x^*\|^2 + \frac{1}{2\gamma}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 - \|y_{t+1} - x_{t+1}\|^2\right)$

- RHS $\geq$ LHS $\geq f(x_t) - f(x_{t+1}) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2$ ($\because$ non-increasing f)

$$\geq \frac{1}{2\beta}\|\nabla f(x_t)\|^2 - \frac{\beta}{2}\|y_{t+1} - x_{t+1}\|^2 - \frac{1}{2\beta}\|\nabla f(x_t)\|^2$$

$$= -\frac{\beta}{2}\|y_{t+1} - x_{t+1}\|^2$$

$\therefore \|x_{t+1} - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)\|x_t - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^{t+1}\|x_0 - x^*\|^2$

- $f(x_T) - f(x^*) \leq \nabla f(x^*)^T(x_T - x^*) + \frac{\beta}{2}\|x_T - x^*\|^2$

$$\leq \|\nabla f(x^*)\| \cdot \|x_T - x^*\| + \frac{\beta}{2}\|x_T - x^*\|^2 \quad (\nabla f(x^*) \text{ may not be } 0)$$

$$\leq \|\nabla f(x^*)\| \cdot \left(1 - \frac{\alpha}{\beta}\right)^{\frac{T}{2}} R + \frac{\beta}{2}\left(1 - \frac{\alpha}{\beta}\right)^T R^2$$

$\therefore O(\log \frac{1}{\varepsilon})$ (but slower than gradient descent)

- $f(x_T) - f(x^*) \leq \left(1 - \frac{\alpha}{\beta}\right)^T (f(x_0) - f(x^*))$ no longer holds

($\because f(x_t) - f(x_{t+1}) \geq \frac{1}{2\beta}\|\nabla f(x_t)\|^2$ no longer holds)

Opt Lec 5.

- Lagrange Multiplier Theory

1. Inequality constrained problem

⇒ minimize $f(x)$ → convex function

s.t. $h_i(x) = 0$ for $1 \le i \le m$ → affine function

$g_j(x) \le 0$ for $1 \le j \le r$ → convex function

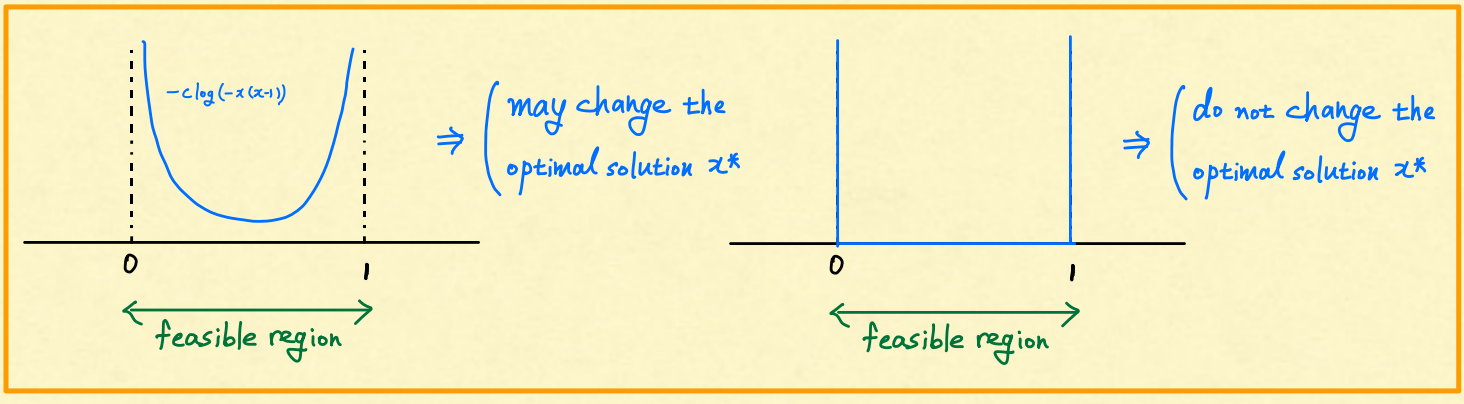
2. Barrier method

⇒ add a high cost to infeasibility or approaching boundary from the interior

(transform constrained problem to unconstrained version)

ex1) min $f(x)$ s.t. $x(x-1) \le 0$ → min $f(x) - c \log(-x(x-1))$

penalty function



3. Lagrange Dual Problem

$\Rightarrow \max_{\mu,\lambda} L(\mu,\lambda)$ → Lagrange dual function

s.t. $\lambda_j \geq 0$ for $1 \leq j \leq r$

where $L(\mu,\lambda) := \min_x \Lambda(x,\mu,\lambda) = \min_x f(x) + \sum_{i=1}^{m} \mu_i h_i(x) + \sum_{j=1}^{r} \lambda_j g_j(x)$

(Lagrange function: $\Lambda(x,\mu,\lambda)$; Lagrange Multiplier: μ_i, λ_j)

$\left(\sum_{i=1}^{m} \mu_i h_i(x) + \sum_{j=1}^{r} \lambda_j g_j(x)\right.$ is a penalty function$\left.\right)$

$\Rightarrow$ updates

$$\mu_i \longleftarrow \mu_i + \gamma \cdot \frac{\partial \Lambda(x^*(\mu,\lambda),\mu,\lambda)}{\partial \mu_i} = \mu_i + \gamma h_i(x)$$

$$\lambda_j \longleftarrow \lambda_j + \gamma \cdot \frac{\partial \Lambda(x^*(\mu,\lambda),\mu,\lambda)}{\partial \lambda_j} = \lambda_j + \gamma g_j(x)$$

✡ $x^*(\mu,\lambda)$ may not be in X, but $x^*(\mu^*,\lambda^*)$ is in X.

✡ Lagrange dual function is concave.

pf) $L(\alpha\mu^{(1)} + (1-\alpha)\mu^{(2)}, \alpha\lambda^{(1)} + (1-\alpha)\lambda^{(2)})$

$$= \min_x \left(f(x) + \sum_{i=1}^{m} \left(\alpha\mu_i^{(1)} + (1-\alpha)\mu_i^{(2)}\right) h_i(x) + \sum_{j=1}^{r} \left(\alpha\lambda_j^{(1)} + (1-\alpha)\lambda_j^{(2)}\right) g_j(x) \right)$$

$$\geq \alpha \left(\min_x f(x) + \sum_{i=1}^{m} \mu_i^{(1)} h_i(x) + \sum_{j=1}^{r} \lambda_j^{(1)} g_j(x) \right) + (1-\alpha) \left(\min_x f(x) + \sum_{i=1}^{m} \mu_i^{(2)} h_i(x) + \sum_{j=1}^{r} \lambda_j^{(2)} g_j(x) \right)$$

✡ Important remarks on optimal solution

① $\nabla_x L(\mu^*,\lambda^*) = 0$

② $h_i(x^*(\mu^*,\lambda^*)) = 0 \quad \forall i$

③ $\lambda_j^* g_j(x^*(\mu^*,\lambda^*)) = 0 \quad \forall j$

- $g_j(x) < 0 \Rightarrow \lambda_j^* = 0$
- $g_j(x) > 0 \Rightarrow \lambda_j^* = \infty$
- $g_j(x) = 0 \Rightarrow \lambda_j^* \geq 0$

✡ Weight decay can be expressed as Lagrange dual problem

pf) $x^* = \arg\min_x f(x)$ s.t. $\|x\| \leq c$

$\Leftrightarrow x^* = \arg\min_x f(x) + \lambda(\|x\| - c)$ s.t. $\lambda \geq 0$

$= \arg\min_x f(x) + \lambda\|x\|$ s.t. $\lambda \geq 0$

(∴ For all c, there is a corresponding value of λ.)

- Proximal Gradient

1. Proximal gradient descent

⇒ decompose the original objective function : $f(x) = g(x) + h(x)$

(g is a nice function and h is a simple additional term)

⇒ $x_{t+1} = \arg\min_y \left(\frac{1}{2\gamma} \| y - (x_t - \gamma \nabla g(x_t)) \|^2 + h(y) \right)$

$:= prox_{h,\gamma}(x_t - \gamma \nabla g(x_t))$

(Generalized Gradient $(G_{h,\gamma}(x_t)) = \frac{1}{\gamma}(x_t - x_{t+1}) = \frac{1}{\gamma}\left(x_t - prox_{h,\gamma}(x_t - \gamma \nabla g(x_t))\right)$

($x_{t+1} = prox_{h,\gamma}(x_t - \gamma \nabla g(x_t)) \xrightarrow{\text{differentiation}} \nabla h(x_{t+1}) = G_{h,\gamma}(x_t) - \nabla g(x_t)$)

☆ G.D. and P.G.D can be generalized to proximal gradient descent.

① Gradient descent ⟶ $h(x) = 0$

② Projected gradient descent ⟶ $h(x) = 0$ (for $x \in \chi$) or $h(x) = \infty$ (for $x \notin \chi$)

✡ Convergence rate for proximal gradient descent

* β - smooth

- $g(x_{t+1}) \le g(x_t) + \nabla g(x_t)^T (x_{t+1} - x_t) + \frac{\beta}{2} \|x_{t+1} - x_t\|^2$ ($\because \beta$-smooth)

$= g(x_t) + \nabla g(x_t)^T (x_{t+1} - x_t) + \frac{\beta}{2} \| -\gamma G_{h,\gamma}(x_t) \|^2$

$= g(x_t) + \nabla g(x_t)^T (x_{t+1} - x_t) + \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2$ ($\because \gamma := \frac{1}{\beta}$)

- $f(x_{t+1}) = g(x_{t+1}) + h(x_{t+1})$

$\le g(x_t) + \nabla g(x_t)^T (x_{t+1} - x_t) + \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2 + h(x_{t+1})$

($\because g$ is convex) $\le g(z) + \nabla g(x_t)^T (x_t - z) + \nabla g(x_t)^T (x_{t+1} - x_t) + \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2 + h(x_{t+1})$

($\because h$ is convex) $\le g(z) + \nabla g(x_t)^T (x_{t+1} - z) + \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2 + h(z) + \nabla h(x_{t+1})^T (x_{t+1} - z)$

$= f(z) + \nabla g(x_t)^T (x_{t+1} - z) + \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2 + (G_{h,\gamma}(x_t) - \nabla g(x_t))^T (x_{t+1} - z)$

$= f(z) + G_{h,\gamma}(x_t)^T (x_{t+1} - z) + \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2$

$= f(z) + G_{h,\gamma}(x_t)^T (x_t - z + x_{t+1} - x_t) + \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2$

$= f(z) + G_{h,\gamma}(x_t)^T (x_t - z) - \frac{1}{\beta} \| G_{h,\gamma}(x_t) \|^2 + \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2$

$= f(z) + G_{h,\gamma}(x_t)^T (x_t - z) - \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2$

$z = x_t$

$\therefore f(x_{t+1}) \le f(x_t) - \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2$ (non-increasing)

$z = x^*$

- $f(x_{t+1}) \le f(x^*) + G_{h,\gamma}(x_t)^T (x_t - x^*) - \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2$

$= f(x^*) + \beta \cdot (x_t - x_{t+1})^T (x_t - x^*) - \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2$

$= f(x^*) + \frac{\beta}{2} \left(\|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \right) - \frac{1}{2\beta} \| G_{h,\gamma}(x_t) \|^2$

$= f(x^*) + \frac{\beta}{2} \left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \right)$

- $f(x_T) - f(x^*) \le \frac{1}{T} \sum_{t=0}^{T-1} \left(f(x_{t+1}) - f(x^*) \right) \le \frac{\beta}{2T} \left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2 \right) \le \frac{\beta}{2T} R^2$

$\therefore O\left(\frac{1}{\varepsilon}\right)$ steps

- subgradient

1. subgradient (g)

$\Rightarrow f(y) \geq f(x) + g^T(y-x) \quad \forall y \in X.$

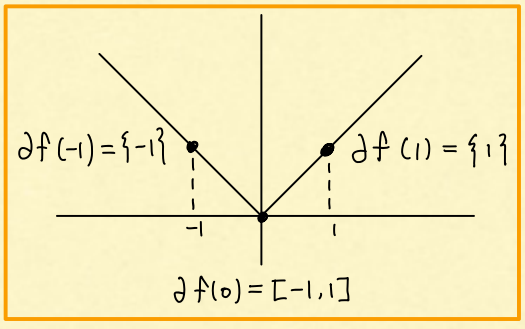
($\partial f(x) :=$ the set of subgradients of f at x)

$\Rightarrow \forall x \in dom(f), \forall g \in \partial f(x), \|g\| \leq L \iff \forall x \in dom(f), |f(x)-f(y)| \leq L\|x-y\|$
(Lipchitz-continuous)

2. subgradient descent

$\Rightarrow x_{t+1} = x_t - \gamma_t g_t \quad$ where $g_t \in \partial f(x_t)$

(projected subgradient descent is also possible)



✡ Convergence rate for subgradient descent

* Lipchitz-continuous

- $f(x_t) - f(x^*) \leq g_t^T(x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^T(x_t - x^*)$

$= \frac{1}{2\gamma}(\|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$

$= \frac{\gamma}{2}\|g_t\|^2 + \frac{1}{2\gamma}(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$

- $f(\bar{x}) - f(x^*) \leq \frac{1}{T}\sum_{t=0}^{T-1}(f(x_t) - f(x^*))$ ($\because$ Jensen's inequality)

$= \frac{\gamma}{2T}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\gamma T}(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2)$

$\leq \frac{\gamma}{2}L^2 + \frac{1}{2\gamma T}R^2$ ($\because \|g_t\| \leq L$)

$\leq \frac{RL}{2\sqrt{T}} + \frac{RL}{2\sqrt{T}} = \frac{RL}{\sqrt{T}}$ ($\because \gamma^* = \frac{R}{L\sqrt{T}}$)

$\therefore O(\frac{1}{\epsilon^2})$ steps

- Mirror descent

1. Dual

⇒ dual space (V^*): a vector space consisting of all linear functionals on original vector space with a vector space structure of pointwise addition & scalar multiplication

($\forall x, y \in V$, $f \in V^*$ s.t. $f(c_1x + c_2y) = c_1 f(x) + c_2 f(y)$)

↳ (definition of linear functional)

(if $V = \mathbb{R}^d$, then $V^* = \mathbb{R}^d$)

⇒ dual norm ($\|\cdot\|_*$): $\|g\|_* = \|g\|_q = \sup_{x \in V} g^T x$ s.t. $\|x\|_p \le 1$ $(x \in V, g \in V^*)$

☆ $\frac{1}{p} + \frac{1}{q} = 1$ where p is a primal norm and q is a dual norm)

pf) $g^T x \le \|g\|_q \|x\|_p$ where $\frac{1}{p} + \frac{1}{q} = 1$ (∵ Hölder's inequality)

$\Rightarrow g^T x \le \|g\|_q$ (∵ $\|x\|_p \le 1$)

☆ L-Lipchitz continuous : $\forall x \in X$, $g \in \partial f(x)$, $\|g\|_* \le L$ ($\|\cdot\|_1$ converge faster than $\|\cdot\|_2$)

☆ β-smooth : $\forall x, y \in X$, $\|\nabla f(x) - \nabla f(y)\|_* \le \beta \|x - y\|$

2. Bregman divergence ($D_\phi(x, y)$) : $\phi(x) - \phi(y) - \nabla\phi(y)^T(x - y)$

($\nabla_x D_\phi(x, y) = \nabla\phi(x) - \nabla\phi(y)$)

☆ $(\nabla\phi(x) - \nabla\phi(y))^T(x - z) = D_\phi(x, y) + D_\phi(z, x) - D_\phi(z, y)$

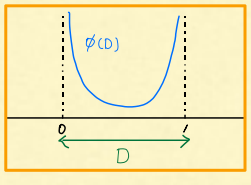
pf) $D_\phi(x, y) + D_\phi(z, x) - D_\phi(z, y)$

$= \phi(x) - \phi(y) - \nabla\phi(y)^T(x - y) + \phi(z) - \phi(x) - \nabla\phi(x)^T(z - x) - \phi(z) + \phi(y) + \nabla\phi(y)^T(z - y)$

$= -\nabla\phi(y)^T(x - y) - \nabla\phi(x)^T(z - x) + \nabla\phi(y)^T(z - y) = (\nabla\phi(x) - \nabla\phi(y))^T(x - z)$

3. Mirror map (ϕ): a function ($\phi : D \to \mathbb{R}^n$) where $X \subset \overline{D}$ & $D (\subset \mathbb{R}^n)$ is a convex, open set as follows

↳ closed set of D



① ϕ is strictly convex and differentiable

② The gradient of ϕ takes all possible values ⟶ $\nabla\phi(D) = \mathbb{R}^n$

③ The gradient of ϕ diverges on the boundary of D ⟶ $\lim_{x \to \partial D} \|\nabla\phi(D)\| = +\infty$

4. Mirror descent

$\Rightarrow \; x_{t+1} = \underset{x \in \mathcal{X} \cap D}{\operatorname{argmin}} \; \gamma_t \nabla f(x_t)^T x + D_\phi(x, x_t)$

① x_t is mapped to $\nabla \phi(x_t)$ ⟶ Mirror map (primal → dual)

② compute $\nabla \phi(x_t) - \gamma \nabla f(x_t)$ ⟶ Gradient descent in dual space

③ Find y_{t+1} s.t. $\nabla \phi(y_{t+1}) = \nabla \phi(x_t) - \gamma \nabla f(x_t)$ ⟶ Inverse mirror map (dual → primal)

④ Projection $x_{t+1} = \Pi_{\mathcal{X}}^{\phi}(y_{t+1}) = \underset{x \in \mathcal{X}}{\operatorname{argmin}} D_\phi(x, y_{t+1})$ ⟶ Projection in primal space

★ Relationship with proximal gradient descent

pf) $x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} D_\phi(x, y_{t+1}) = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left(\phi(x) - \phi(y_{t+1}) - \nabla \phi(y_{t+1})^T (x - y_{t+1}) \right)$

$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left(\phi(x) - \nabla \phi(y_{t+1})^T x \right) = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left(\phi(x) - (\nabla \phi(x_t) - \gamma \nabla f(x_t))^T x \right)$

$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left(\phi(x) - \phi(x_t) - \nabla \phi(x_t)^T (x - x_t) + \gamma \nabla f(x_t)^T x \right)$

$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left(\underbrace{\gamma \nabla f(x_t)^T x}_{①} + \underbrace{D_\phi(x, x_t)}_{②} \right)$

(① : local minimization inducing to move opposite direction of gradient
② : additional term inducing not to move far from the previous point)

★ Well used mirror map : $\phi(x_t) = \sum_{i=1}^{d} x_t^{(i)} \log x_t^{(i)}$ where $x_t^{(i)}$ is the ith element of x_t

① $\nabla \phi(x_t) = \begin{bmatrix} \vdots \\ \log x_t^{(i)} + 1 \\ \vdots \end{bmatrix}$

② $\nabla \phi(x_t) - \gamma \nabla f(x_t) = \begin{bmatrix} \vdots \\ \log y_{t+1}^{(i)} + 1 \\ \vdots \end{bmatrix} = \nabla \phi(y_{t+1})$

③ $y_{t+1}^{(i)} = x_t^{(i)} \exp\left(-\gamma \cdot \frac{\partial f(x_t)}{\partial x_t^{(i)}} \right)$

④ $x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} D_\phi(x, y_{t+1}) = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left(\phi(x) - \nabla \phi(y_{t+1})^T x \right)$

$= \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left(\sum_{i=1}^{d} x^{(i)} \log \frac{x^{(i)}}{y_{t+1}^{(i)}} - \|x\|_1 \right) = \frac{y_{t+1}}{\|y_{t+1}\|_1}$ where $\mathcal{X} = \{ x \in \mathbb{R}^d \mid \|x\|_1 = 1, x^{(i)} \geq 0 \}$

(differentiation)

Convergence rate for mirror descent

* f be L-Lipchitz continuous & ϕ be α-strongly convex

- $\nabla_{x_{t+1}} D(x_{t+1}, y_{t+1})^T (x_{t+1} - x) \le 0$ ($\because x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left(D_\phi(x, y_{t+1}) \right)$ and D is convex)

  $\Rightarrow \nabla \phi(x_{t+1})^T (x_{t+1} - x) \le \nabla \phi(y_{t+1})^T (x_{t+1} - x) \le \nabla \phi(y_{t+1})^T (x_{t+1} - y_{t+1} + y_{t+1} - x)$

  $\Rightarrow -\nabla \phi(x_{t+1})^T (x - x_{t+1}) - \nabla \phi(y_{t+1})^T (x_{t+1} - y_{t+1}) \le -\nabla \phi(y_{t+1})^T (x - y_{t+1})$

  $\therefore D_\phi(x, x_{t+1}) + D_\phi(x_{t+1}, y_{t+1}) \le D_\phi(x, y_{t+1})$

- $D_\phi(x_t, y_{t+1}) - D_\phi(x_{t+1}, y_{t+1}) = \phi(x_t) - \phi(x_{t+1}) - \nabla \phi(y_{t+1})^T (x_t - x_{t+1})$

  ($\because \alpha$-strongly convex) $\le \nabla \phi(x_t)^T (x_t - x_{t+1}) - \frac{\alpha}{2} \|x_t - x_{t+1}\|^2 - \nabla \phi(y_{t+1})^T (x_t - x_{t+1})$

  $= \left(\nabla \phi(x_t) - \nabla \phi(y_{t+1}) \right)^T (x_t - x_{t+1}) - \frac{\alpha}{2} \|x_t - x_{t+1}\|^2$

  $= \gamma g_t^T (x_t - x_{t+1}) - \frac{\alpha}{2} \|x_t - x_{t+1}\|^2$

  $\le \gamma L \|x_t - x_{t+1}\| - \frac{\alpha}{2} \|x_t - x_{t+1}\|^2$

  $\le -\frac{\alpha}{2} \left(\|x_t - x_{t+1}\| - \frac{\gamma L}{\alpha} \right)^2 + \frac{\alpha}{2} \cdot \frac{\gamma^2 L^2}{\alpha^2} \le \frac{\gamma^2 L^2}{2\alpha}$

- $f(x_t) - f(x) \le g_t^T (x_t - x)$ where $g_t \in \partial f(x_t)$

  $\le \frac{1}{\gamma} \left(\nabla \phi(x_t) - \nabla \phi(y_{t+1}) \right)^T (x_t - x)$

  $= \frac{1}{\gamma} \left(D_\phi(x_t, y_{t+1}) + D_\phi(x, x_t) - D_\phi(x, y_{t+1}) \right)$

  $\le \frac{1}{\gamma} \left(D_\phi(x_t, y_{t+1}) + D_\phi(x, x_t) - D_\phi(x, x_{t+1}) - D_\phi(x_{t+1}, y_{t+1}) \right)$

  $\Rightarrow f(x_t) - f(x^*) \le \frac{1}{\gamma} \left(D_\phi(x_t, y_{t+1}) - D_\phi(x_{t+1}, y_{t+1}) + D_\phi(x^*, x_t) - D_\phi(x^*, x_{t+1}) \right)$

  $\le \frac{\gamma L^2}{2\alpha} + \frac{1}{\gamma} \left(D_\phi(x^*, x_t) - D_\phi(x^*, x_{t+1}) \right)$

- $f(\bar{x}) - f(x^*) \le \frac{1}{T} \sum_{t=1}^{T} \left(f(x_t) - f(x^*) \right) \le \frac{\gamma L^2}{2\alpha} + \frac{1}{\gamma T} \left(D_\phi(x^*, x_1) - D_\phi(x^*, x_{T+1}) \right)$

  $\le \frac{\gamma L^2}{2\alpha} + \frac{R^2}{\gamma T}$ ($R^2 := \sup_{x \in \mathcal{X}} \left(\phi(x) - \phi(x_1) \right)$)

  $\le RL \sqrt{\frac{2}{\alpha T}}$ ($\because \gamma^* = \frac{R}{L} \sqrt{\frac{2\alpha}{T}}$)

$\therefore O\left(\frac{1}{\epsilon^2}\right)$ steps

Opt Lec 6.

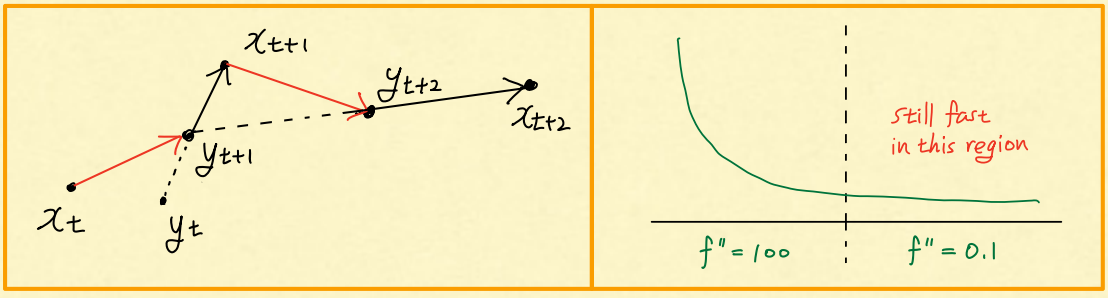
- Nesterov's accelerated gradient descent

1. Nesterov's accelerated gradient descent

$$\Rightarrow \begin{cases} y_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t) \\ x_{t+1} = y_{t+1} + \left(\frac{\sqrt{k}-1}{\sqrt{k}+1}\right)(y_{t+1} - y_t) \end{cases} \quad \text{where } k = \frac{\beta}{\alpha} \quad (1 \le k \le \infty)$$

$k=1 \Rightarrow 0$, $k=\infty \Rightarrow 1$

- Not informative
- Getting harder



2. key properties

Let f is α-strongly convex and β-smooth

$$\begin{cases} \phi_1(x) = f(x_1) + \frac{\alpha}{2}\|x - x_1\|^2 \\ \phi_{s+1}(x) = \left(1 - \frac{1}{\sqrt{k}}\right)\phi_s(x) + \frac{1}{\sqrt{k}}\left(f(x_s) + \nabla f(x_s)^T(x - x_s) + \frac{\alpha}{2}\|x - x_s\|^2\right) \end{cases}$$

$\Rightarrow \phi_s(x) = \text{affine function} + \frac{\alpha}{2}\|x - v_s\|^2$

① $\phi_{s+1}(x) - f(x) \le \left(1 - \frac{1}{\sqrt{k}}\right)^s (\phi_1(x) - f(x))$

pf) $\phi_{s+1}(x) - f(x) = \left(1 - \frac{1}{\sqrt{k}}\right)(\phi_s(x) - f(x)) + \frac{1}{\sqrt{k}}\left(f(x_s) + \nabla f(x_s)^T(x - x_s) + \frac{\alpha}{2}\|x - x_s\|^2 - f(x)\right)$

$\le \left(1 - \frac{1}{\sqrt{k}}\right)(\phi_s(x) - f(x))$ ($\because \alpha$-strongly convex)

$\le \left(1 - \frac{1}{\sqrt{k}}\right)^s (\phi_1(x) - f(x))$

$= \left(1 - \frac{1}{\sqrt{k}}\right)^s \left(f(x_1) + \frac{\alpha}{2}\|x - x_1\|^2 - f(x)\right)$

$\le \left(1 - \frac{1}{\sqrt{k}}\right)^s \left(\nabla f(x)^T(x_1 - x) + \frac{\alpha+\beta}{2}\|x - x_1\|^2\right)$ ($\because \beta$-smooth)

$\therefore \phi_{s+1}(x) - f(x) \le \left(1 - \frac{1}{\sqrt{k}}\right)^s (\phi_1(x) - f(x))$

② $f(y_s) \le \min_x \phi_s(x)$

- $\phi_s(x) = \min_x \left(\phi_s(x)\right) + \frac{\alpha}{2}\|x - v_s\|^2 = \phi_s^* + \frac{\alpha}{2}\|x - v_s\|^2$

- $\phi_{s+1}(x) = \left(1 - \frac{1}{\sqrt{k}}\right)\phi_s(x) + \frac{1}{\sqrt{k}}\left(f(x_s) + \nabla f(x_s)^T(x - x_s) + \frac{\alpha}{2}\|x - x_s\|^2\right)$

$$= \min_x \left(\phi_{s+1}(x)\right) + \frac{\alpha}{2}\|x - v_{s+1}\|^2 = \phi_{s+1}^* + \frac{\alpha}{2}\|x - v_{s+1}\|^2$$

$$\Rightarrow v_{s+1} = -\frac{1}{\alpha}\left(\left(1 - \frac{1}{\sqrt{k}}\right)\cdot\frac{\alpha}{2}\cdot(-2v_s) + \frac{1}{\sqrt{k}}\cdot\nabla f(x_s) + \frac{1}{\sqrt{k}}\cdot\frac{\alpha}{2}(-2x_s)\right)$$

$$= \left(1 - \frac{1}{\sqrt{k}}\right)\cdot v_s + \frac{1}{\sqrt{k}}x_s - \frac{1}{\alpha\sqrt{k}}\nabla f(x_s)$$

$$\Rightarrow \|x_s - v_{s+1}\|^2 = \left\|\left(1 - \frac{1}{\sqrt{k}}\right)(x_s - v_s) + \frac{1}{\alpha\sqrt{k}}\nabla f(x_s)\right\|^2$$

$$= \left(1 - \frac{1}{\sqrt{k}}\right)^2\|x_s - v_s\|^2 + \frac{1}{\alpha^2 k}\|\nabla f(x_s)\|^2 + 2\left(1 - \frac{1}{\sqrt{k}}\right)\cdot\frac{1}{\alpha\sqrt{k}}\nabla f(x_s)^T(x_s - v_s)$$

- $v_{s+1} - x_{s+1} = \left(1 - \frac{1}{\sqrt{k}}\right)v_s + \frac{1}{\sqrt{k}}x_s - \frac{1}{\alpha\sqrt{k}}\nabla f(x_s) - x_{s+1}$

$$= \left(1 - \frac{1}{\sqrt{k}}\right)\left(x_s + \sqrt{k}(x_s - y_s)\right) + \frac{1}{\sqrt{k}}x_s - \frac{1}{\alpha\sqrt{k}}\nabla f(x_s) - x_{s+1} \quad (\because \text{mathematical induction})$$

$$= \sqrt{k}(x_s - y_s) + y_s - \frac{\sqrt{k}}{\beta}\nabla f(x_s) - x_{s+1}$$

$$= \sqrt{k}\,y_{s+1} - (\sqrt{k} - 1)y_s - x_{s+1} \quad (\because y_{s+1} = x_s - \tfrac{1}{\beta}\nabla f(x_s))$$

$$= \sqrt{k}\,y_{s+1} - (\sqrt{k} - 1)y_s - y_{s+1} - \frac{\sqrt{k} - 1}{\sqrt{k} + 1}(y_{s+1} - y_s)$$

$$= (\sqrt{k} - 1)(y_{s+1} - y_s)\frac{\sqrt{k}}{\sqrt{k} + 1} = \sqrt{k}(x_{s+1} - y_{s+1})$$

- $\phi_{s+1}(x_s) = \left(1 - \frac{1}{\sqrt{k}}\right)\phi_s(x_s) + \frac{1}{\sqrt{k}}f(x_s)$

$$= \left(1 - \frac{1}{\sqrt{k}}\right)\phi_s^* + \frac{\alpha}{2}\left(1 - \frac{1}{\sqrt{k}}\right)\|x_s - v_s\|^2 + \frac{1}{\sqrt{k}}f(x_s) = \phi_{s+1}^* + \frac{\alpha}{2}\|x_s - v_{s+1}\|^2$$

$$\Rightarrow \phi_{s+1}^* = \left(1 - \frac{1}{\sqrt{k}}\right)\phi_s^* + \frac{\alpha}{2\sqrt{k}}\left(1 - \frac{1}{\sqrt{k}}\right)\|x_s - v_s\|^2 + \frac{1}{\sqrt{k}}f(x_s) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2 - \frac{1}{\sqrt{k}}\left(1 - \frac{1}{\sqrt{k}}\right)\nabla f(x_s)^T(x_s - v_s)$$

$$= \left(1 - \frac{1}{\sqrt{k}}\right)\phi_s^* + \frac{\alpha\sqrt{k}}{2}\left(1 - \frac{1}{\sqrt{k}}\right)\|x_s - y_s\|^2 + \frac{1}{\sqrt{k}}f(x_s) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2 + \left(1 - \frac{1}{\sqrt{k}}\right)\nabla f(x_t)^T(x_s - y_s)$$

- $f(y_{s+1}) \le f(x_s) - \frac{1}{2\beta}\|\nabla f(x_s)\|^2 \quad (\because \beta\text{-smooth},\ \gamma := \frac{1}{\beta})$

$$= \left(1 - \frac{1}{\sqrt{k}}\right)f(x_s) + \frac{1}{\sqrt{k}}f(x_s) - \frac{1}{2\beta}\|\nabla f(x_s)\|^2$$

$$= \left(1 - \frac{1}{\sqrt{k}}\right)\left(f(x_s) - f(y_s)\right) + \left(1 - \frac{1}{\sqrt{k}}\right)f(y_s) + \frac{1}{\sqrt{k}}f(x_s) - \frac{1}{2\beta}\|\nabla f(x_s)\|^2$$

$$\le \left(1 - \frac{1}{\sqrt{k}}\right)\nabla f(x_s)^T(x_s - y_s) + \left(1 - \frac{1}{\sqrt{k}}\right)\phi_s^* + \frac{1}{\sqrt{k}}f(x_s) - \frac{1}{2\beta}\|\nabla f(x_s)\|^2 \quad (\because \text{mathematical induction})$$

$$\le \phi_{s+1}^* \quad \left(\because \frac{\alpha\sqrt{k}}{2}\left(1 - \frac{1}{\sqrt{k}}\right)\|x_s - y_s\|^2 \ge 0\right)$$

$\therefore f(y_{s+1}) \le \phi_{s+1}^* = \min_x \left(\phi_{s+1}(x)\right)$

Convergence rate for Nesterov's accelerated gradient descent

* α-strongly convex & β-smooth

- $\phi_s(x^*) - f(x^*) \le \left(1 - \frac{1}{\sqrt{k}}\right)^{s-1} \left(\nabla f(x^*)^T (x_1 - x^*) + \frac{\alpha+\beta}{2} \|x^* - x_1\|^2\right)$

$$= \frac{\alpha+\beta}{2} \|x^* - x_1\|^2 \cdot \left(1 - \frac{1}{\sqrt{k}}\right)^{s-1}$$

$$\le \frac{\alpha+\beta}{2} \|x^* - x_1\|^2 \cdot \exp\left(-\frac{s-1}{\sqrt{k}}\right)$$

- $f(y_t) - f(x^*) \le \phi_t^* - f(x^*) \le \phi_t(x^*) - f(x^*) \le \frac{\alpha+\beta}{2} \|x_1 - x^*\|^2 \cdot \exp\left(-\frac{t-1}{\sqrt{k}}\right)$

$\therefore O\left(\log \frac{1}{\epsilon}\right)$ steps

Opt Lec 7.

- Stochastic Gradient Descent I

1. Stochastic oracle (g)

   ⇒ fixed input point $x \longrightarrow g(x)$ s.t. $E[g(x)] \in \partial f(x)$

   ⇒ random input point $x \longrightarrow g(x)$ s.t. $E[g(x)|x] \in \partial f(x)$

2. Stochastic Gradient Descent

   ⇒ $x_{t+1} = x_t - \gamma g_t$ where $g_t = g(x_t)$ and $E[g(x_t)] \in \partial f(x_t)$

☆ Assumptions for simple analysis

① f is α-strongly convex and β-smooth

$\left(\alpha \|x-y\| \le \|\nabla f(x) - \nabla f(y)\| \le \beta \|x-y\|\right)$

② $E\left[\|g_t - \nabla f(x_t)\|^2\right] = E\left[\|g_t\|^2\right] - \|\nabla f(x)\|^2 \le \sigma^2 \quad \forall t$

Convergence rate for Stochastic Gradient Descent

* α-strongly convex & β-smooth

- $E[f(x_t) - f(x_{t+1}) \mid x_t] \geq E[\nabla f(x_t)^T (x_t - x_{t+1}) - \frac{\beta}{2}\|x_t - x_{t+1}\|^2 \mid x_t]$ ($\because \beta$-smooth)

$$= E[\gamma \nabla f(x_t)^T g_t - \frac{\beta\gamma^2}{2}\|g_t\|^2 \mid x_t]$$

$$= \gamma \|\nabla f(x_t)\|^2 - \frac{\beta\gamma^2}{2} E[\|g_t\|^2 \mid x_t] \quad (\because E[g_t \mid x_t] = \nabla f(x_t))$$

$$\geq \gamma\left(1 - \frac{\beta\gamma}{2}\right)\|\nabla f(x_t)\|^2 - \frac{\beta\gamma^2}{2}\sigma^2 \quad (\because \text{assumption})$$

- $E[f(x_{t+1}) - f(x^*) \mid x_t] \leq E[f(x_t) - f(x^*) \mid x_t] - \gamma\left(1 - \frac{\beta\gamma}{2}\right)\|\nabla f(x_t)\|^2 + \frac{\beta\gamma^2}{2}\sigma^2$

$$\leq f(x_t) - f(x^*) - \gamma\left(1 - \frac{\beta\gamma}{2}\right) \cdot 2\alpha \cdot (f(x_t) - f(x^*)) + \frac{\beta\gamma^2}{2}\sigma^2$$

($\because \alpha$-strongly convex $\Rightarrow$ Polyak Lojasiewicz inequality)

$$\leq \left(1 - \alpha\gamma(2 - \beta\gamma)\right)\left(f(x_t) - f(x^*)\right) + \frac{\beta\gamma^2}{2}\sigma^2$$

- $E[f(x_T) - f(x^*)] \leq \left(1 - \alpha\gamma(2-\beta\gamma)\right)^T \left(f(x_0) - f(x^*)\right) + \frac{\beta\gamma^2}{2}\sigma^2 \left(\sum_{t=0}^{T-1} \left(1 - \alpha\gamma(2-\beta\gamma)\right)^t\right)$

$$\leq \left(1 - \alpha\gamma(2-\beta\gamma)\right)^T \left(f(x_0) - f(x^*)\right) + \frac{\beta\gamma^2}{2}\sigma^2 \cdot \frac{1 - \left(1 - \alpha\gamma(2-\beta\gamma)\right)^T}{\alpha\gamma(2-\beta\gamma)}$$

(To boost convergence speed, we can increase batch size ($\sigma^2 \downarrow$) or decrease learning rate ($\gamma \downarrow$))

- When $\gamma = \frac{1}{\beta}$

$$\Rightarrow E[f(x_T) - f(x^*)] \leq \left(1 - \frac{\alpha}{\beta}\right)^T \left(f(x_0) - f(x^*)\right) + \frac{\sigma^2}{2} \cdot \frac{1 - \left(1 - \frac{\alpha}{\beta}\right)^T}{\alpha}$$

$$\leq \left(1 - \frac{\alpha}{\beta}\right)^T \left(f(x_0) - f(x^*)\right) + \frac{\sigma^2}{2\alpha}$$

$\therefore O(\log \frac{1}{\varepsilon})$ steps

- When $\gamma = \frac{1}{\sqrt{T}}$

$$\Rightarrow E[f(x_T) - f(x^*)] \leq \left(1 - \frac{\alpha}{\sqrt{T}}\left(2 - \frac{\beta}{\sqrt{T}}\right)\right)^T \left(f(x_0) - f(x^*)\right) + \frac{\beta}{2}\sigma^2 \frac{1 - \left(1 - \frac{\alpha}{\sqrt{T}}\left(2 - \frac{\beta}{\sqrt{T}}\right)\right)^T}{\frac{\alpha}{\sqrt{T}}\left(2 - \frac{\beta}{\sqrt{T}}\right)}$$

$$\leq \exp\left(-2\alpha\sqrt{T} + \alpha\beta\right)\left(f(x_0) - f(x^*)\right) + \frac{\beta}{2T}\sigma^2 \cdot \frac{1}{\frac{\alpha}{\sqrt{T}}\left(2 - \frac{\beta}{\sqrt{T}}\right)}$$

$$\leq \exp\left(-2\alpha\sqrt{T} + \alpha\beta\right)\left(f(x_0) - f(x^*)\right) + \frac{\beta}{2}\sigma^2 \cdot \frac{1}{\alpha(2\sqrt{T} - \beta)}$$

$\therefore O(\frac{1}{\varepsilon^2})$ steps

$\Rightarrow O(\exp(-\sqrt{T}))$ $\Rightarrow O((\log\frac{1}{\varepsilon})^2)$ steps

$\Rightarrow O(\frac{1}{\sqrt{T}})$ $\Rightarrow O(\frac{1}{\varepsilon^2})$ steps (dominant)