4 Linear Algebra

- Review

$$A = n\left[\begin{array}{c} m \\ \; \end{array}\right]$$

A: $n \times m$ matrix

$A = [a_1 \ a_2 \cdots \ a_m]$ $\quad a_i \in \mathbb{R}^n$

def

① range(A) = column space of A $= \{x = \sum_{i=1}^{m} \alpha_i a_i : \alpha_i \in \mathbb{R}\} \in \mathbb{R}^n$

② rank(A) = dim (range(A))

③ null(A) $= \{x \in \mathbb{R}^m : Ax = 0\}$

④ $W^\perp :=$ orthogonal complement of set $W \in \mathbb{R}^m$

$= \{v \in \mathbb{R}^m : v^T w = 0 \quad \forall w \in W\}$

⑤ W is subspace of $\mathbb{R}^n$ if

i) $0 \in W$

ii) $w_1, w_2 \in W \Rightarrow w_1 + w_2 \in W$

iii) $w \in W \Rightarrow cw \in W \ (\forall c \in \mathbb{R})$

thms

① rank(A) + dim(null(A)) = n

② $\mathbb{R}^m = W \oplus W^\perp$ (if W is subspace of $\mathbb{R}^m$)

③ $(\text{row}(A))^\perp = \text{null}(A)$

$(\text{col}(A))^\perp = \text{null}(A^T)$

- # Norm

definition $\|\cdot\|$ is a norm on $\mathbb{R}^n$ if $\|x\| \in [0, \infty)$ $\forall x \in \mathbb{R}^n$,

① $\|x+y\| \le \|x\| + \|y\|$ $\quad \forall x, y \in \mathbb{R}^n$

② $\|\alpha x\| = |\alpha| \, \|x\|$ $\quad \forall \alpha \in \mathbb{R}, x \in \mathbb{R}^n$

③ $\|x\| = 0 \iff x = 0 \in \mathbb{R}^n$

Famous Norms

① L_2 norm (= Euclidean norm)

$$\|x\|_2 = (x_1^2 + \cdots + x_n^2)^{\frac{1}{2}}$$

② L_p norm

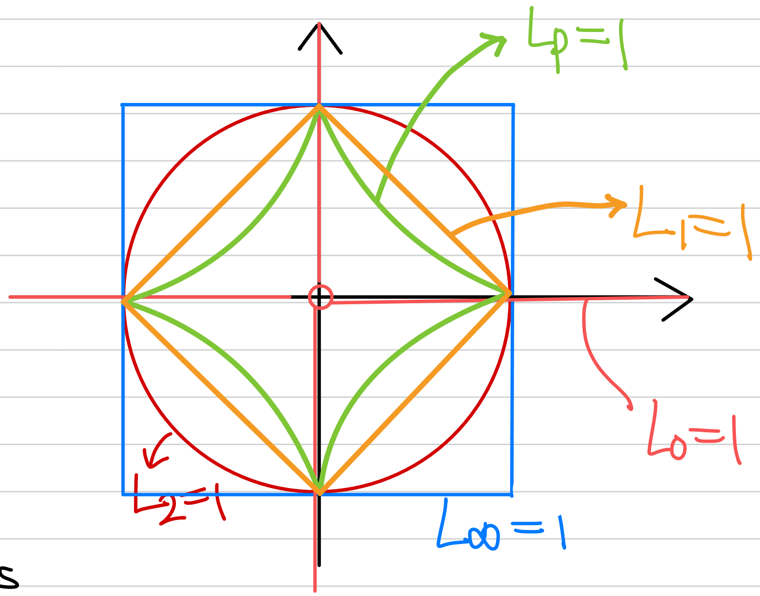
$$\|x\|_p = (x_1^p + \cdots + x_n^p)^{\frac{1}{p}}$$

③ L_∞ norm

$$\|x\|_\infty = \max_i |x_i|$$

④ L_0 norm

$$\|x\|_0 = \# \text{ of nonzero elements}$$



ex L_1 norm is less sensitive to outliers than L_2 norm

• Orthogonal

<u>definition</u> $U \in \mathbb{R}^{n\times n}$ is orthogonal (= unitary) if transpose of it is inverse of it

$\Leftrightarrow U^TU = I \Leftrightarrow UU^T = I$

$\Leftrightarrow$ column vectors of U forms a orthonormal basis of $\mathbb{R}^n$

• Singular Vector Decomposition

$A \in \mathbb{R}^{m\times n}$ $\quad A = U\Sigma V^T$ $\quad$ ($u_1, \cdots, u_m$: orthonormal basis of $\mathbb{R}^m$

$v_1, \cdots, v_n$: orthonormal basis of $\mathbb{R}^n$)

<u>def</u>

$$\begin{array}{c} \\ m \end{array}\overset{n}{\begin{bmatrix} A \end{bmatrix}} = \begin{array}{c} \\ m \end{array}\overset{m}{\begin{bmatrix} U \end{bmatrix}} \begin{array}{c} \\ m \end{array}\overset{n}{\begin{bmatrix} \Sigma \end{bmatrix}} \begin{array}{c} \\ n \end{array}\overset{n}{\begin{bmatrix} V^T \end{bmatrix}}$$

$$= [u_1 | u_2 \cdots | u_m] \, \Sigma \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \; 0 \end{bmatrix}$$

$$= u_1\sigma_1 v_1^T + \cdots + u_r\sigma_r v_r^T$$

$$= \begin{array}{c} \\ m \end{array}\overset{r}{[u_1 | \cdots | u_r]} \; \begin{array}{c} \\ r \end{array}\overset{r}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}$$

$$\Rightarrow A^{-1} = V\Sigma^{-1}U^T$$

- **Pseudo - inverse**

definition

$$A^+ = V_c \Sigma_c^{-1} U_c \qquad \Sigma_c^{-1} = \begin{bmatrix} 1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_r \end{bmatrix}$$

$* \left(U_c U_c^T \neq id \qquad V_c V_c^T \neq id \right.$

$\left. U_c^T U_c = id \qquad V_c^T V_c = id \right)$

property

minimum achiever of $\|A\theta - B\|^2$: $A = U \Sigma V^T$

then $\theta = A^+ B \rightarrow$ minimum achiever

• Soft Prediction

def
- hard prediction : $f_\theta(x) = y \in \mathcal{Y}$
- soft prediction : $f_\theta(x) = (\ \ ) \in$ distribution of y $\notin \mathcal{Y}$ $f_\theta : \mathcal{X} \to \tilde{\mathcal{Y}}$

• Loss function

definition $\ell : \tilde{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ $\ell(f_\theta(x^{(i)}), y) \to$ real #

: how well $\tilde{y}$ approximates

ex> quadratic loss (= Mean Square Error)

$$\ell(\tilde{y}, y) = \|\tilde{y} - y\|_2^2$$

- **Expected Risk** (= True risk, population risk)

$\underline{\text{def}}$ $R[f] = \mathbb{E}[\ell(f(X), Y)]$

→ minimize $R[f]$

- **Bayes Risk** (= optimal risk)

$\underline{\text{def}}$ $R^* = \inf_f R[f]$

→ infimum over all possible functions

optimal achieving f^* : Bayes predictor

$R^* = R[f^*]$

- **Empirical Risk Minimization**

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(x^{(i)}), y^{(i)})$$

Actually, Training is Optimizing

→ minimizing $\mathcal{L}(\theta)$: Gradient Descent

- **Some Math Knowledge**

def
① directional derivative $= \nabla f(x)\cdot v = \langle \nabla f(x), v\rangle = \nabla f(x)^T v$
② convexity $\Leftrightarrow f''(x) \geq 0$ in $\mathbb{R}^d \Leftrightarrow \nabla^2 f \succeq 0$
③ f is L-Lipschitz if $f(x_1) - f(x_2) \leq L\|x_1 - x_2\|$
④ f is L-smooth if ∇f is L-Lipschitz.
⑤ f is μ-strongly convex if $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2$

thms
① f: twice diff L-smooth $\Leftrightarrow \nabla^2 f \preceq LI$
② $f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2$
③ f: minimizer x_0 then
$\frac{1}{2L}\|\nabla f(z)\|^2 \leq f(z) - f(x_0) \leq \frac{L}{2}\|z - x_0\|^2$
④ f: μ-strongly-convex $\Rightarrow \nabla^2 f \succeq \mu I \Leftrightarrow \nabla^2 f - \mu I \succeq 0$
⑤ f: convex $\Rightarrow (\nabla f(x) - \nabla f(y))^T(x-y) \geq 0$
⑥ f: μ-strongly convex $\Rightarrow (\nabla f(x) - \nabla f(y))^T(x-y) \geq \mu\|x-y\|^2$

- **Co-Coercivity** : name of inequality

f: convex, L-smooth
$\Rightarrow (\nabla f(x) - f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$

- **Polyak - Lojasiewicz (PL)**

if f is μ-strongly convex, then f satisfies PL condition

$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f(x_{opt}))$

• Gradient Descent

$f: \mathbb{R}^d \to \mathbb{R}$, L-smooth, $x^{(0)}$: pick randomly

$x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)})$ α: learning rate

If f has global maximum, $\exists \alpha > 0$ s.t Gradient Descent converges

• Gradient Descent + convex + L-smooth

$$f(x^T) - f(x_{opt}) \le \frac{1}{T}\left(\frac{1}{2\alpha}\|x^{(0)} - x_{opt}\|^2\right) = O\left(\frac{1}{T}\right)$$

• Gradient Descent + convex + μ-strongly convex + L-smooth

Lemma f: μ-strongly convex & L-smooth

$\Rightarrow h(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex & $(L-\mu)$-smooth

$$f(x^T) - f(x_{opt}) \le \frac{L}{2} c^T \|x^{(0)} - x_{opt}\|^2 = O(c^T)$$

$c < 1$: exponentially fast convergence

conclusion

① GD + convex	$O(\frac{1}{T})$
② GD + strong convex	$O(e^{-cT})$
③ GD + convex	$\exists$ algorithm $O(\frac{1}{T^2})$

- **Operator**

$T: \mathbb{R}^d \to \mathbb{R}^d$

def
1. T is nonexpansive if $\|Tx - Ty\| \le \|x - y\|$
2. T is contractive if $\|Tx - Ty\| \le L\|x - y\|$ $(L<1)$
3. T is θ- averaged if $T = (1-\theta)I + \theta S$, $I = id$, S : nonexpansive
4. x is fixed point of T if $Tx = x$

- **Fixed point iterator** (Picard iteration)

$x^{(0)}$: starting point, T is θ - averaged, x_f : fixed point

$$\|x^{(k+1)} - x^{(k)}\| \le \frac{1}{k+1} \cdot \frac{\theta}{1-\theta} \|x^{(0)} - x_f\|^2$$

pf by $\|(1-\theta)x + \theta y\|^2 = (1-\theta)\|x\|^2 + \theta\|y\|^2 - \theta(1-\theta)\|x-y\|^2$

$\Longrightarrow$ converging rate : $O(\frac{1}{k})$

ex> Gradient descent is fixed point iteration of θ- averaged operator

$T = I - \alpha \nabla f$

$S = I - \frac{2}{L}\nabla f$ $\quad \theta = \frac{\alpha L}{2}$

+) if f is μ-strongly convex,

$\|Tx - Ty\| \le L\|x - y\|$, $L<1$

$\therefore T$ is contractive operator

- # Nesterov's accelerated gradient method (AGM)

x_0: random initialization | f: convex, L-smooth

$y_0 : x_0$

① $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

② $y_{k+1} = x_{k+1} + \frac{k-1}{k+2}(x_{k+1} - x_k)$

* Equivalent formula

① $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

② $z_{k+1} = z_k - \frac{k+1}{2L}\nabla f(y_k)$

③ $y_{k+1} = \left(1 - \frac{2}{k+2}\right)x_{k+1} + \frac{2}{k+2} z_{k+1}$

& $x_0 = y_0 = z_0$

* Converging rate

global minimizer x^*

$$f(x_k) - f(x^*) \leq \frac{2L}{k^2}\|x_0 - x^*\|^2$$

* First order Method

any iterate algorithm that selects x_{k+1} in the set

$\{x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), \cdots, \nabla f(x_k)\}\}$

$\forall k \leq \frac{d-1}{2}$, $\exists f$, $\forall$ 1st order method, $f(x) - f(x^*) \geq \frac{L\|x_0 - x^*\|^2}{32(k+1)}$

(d: dimension)

• Momentum

$V_0 = 0$

$V_{k+1} = \beta V_k + \alpha \nabla f(x_k)$

$x_{k+1} = x_k - V_{k+1}$

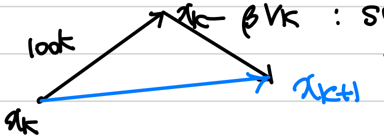
$(\beta = 0 \ : \ GD)$

• Nesterov oscillation

$V_{k+1} = \beta V_k + \alpha \nabla f(x_k - \beta V_k)$

$x_{k+1} = x_k - V_{k+1}$

$x_k - \beta V_k$: sudo-step



$\nabla f(x_k - \beta V_k)$

gd : $O\left(\frac{1}{k^2}\right)$

gd+momentum : $O\left(\frac{1}{k}\right)$

- **Stochastic gradient descent**

1st SGD : minimize $f(x)$

$$\min f(\theta) = \frac{1}{N}\sum_{i=1}^{N} f_i(\theta)$$

① SGD : i_k is randomly picked from $\{1, 2, \cdots, N\}$

$$\theta_{k+1} = \theta_k - \alpha \cdot \nabla f_{i_k}(\theta)$$

sample gradients

② minibatch : $I_k \subseteq \{1, 2, \cdots, N\}$, $|I_k| = B$

$$\theta_{k+1} = \theta_k - \alpha \cdot \boxed{\frac{1}{b}\sum_{i \in I_k} \nabla f_i(\theta)}$$

g_k : stochastic gradient

③ $\mathbb{E}[f(\bar{\theta}_k) - f(\theta^*)] \le \frac{\sqrt{L^2+\sigma^2}\,\|\theta_0 - \theta^*\|^2}{\sqrt{k+1}}$ when $\alpha = \frac{\|\theta_0 - \theta^*\|^2}{\sqrt{L^2+\sigma^2}\,\sqrt{k+1}}$, $\bar{\theta}_k = \frac{1}{k+1}\sum_{k=0}^{K} \theta_k$,

: SGD rate

$\therefore O\left(\frac{1}{\sqrt{k}}\right)$

$f(\theta)$: L-Lipschitz, continuous, convex

θ_0 : starting point K : total iteration count

- **Regularization** purpose: prevent from overfitting

$r(\theta) = \sum_{i=1}^{d} \theta_i^2$

$\sum_{i=1}^{N} \ell(f_\theta(x^i), y^i) + \lambda\|\theta\|^2$

Ridge Regression : ① linear + ② L_2 penalty ($\lambda\|\theta\|^2$)

ridge regression

goal: minimizing $\|X\theta - Y\|^2 + \lambda\|\theta\|^2$

if X : full rank $\Rightarrow$ minimum achiever is $\theta^* = (X^TX + \lambda I)^{-1} X^TY$

- f: G-Lipschitz, continuous, convex

x^* : minimizer of $f(x) + \frac{\mu}{2\|x\|^2}$

$\mathbb{E}_k[g_k] = \nabla f(x_k)$, $Var_k(g_k) \le \sigma^2$

$\alpha_k = \frac{1}{\mu(k+1)}$, $\bar{x}_k = \frac{1}{k+1} \sum_{i=0}^{k} x_i$

$\longrightarrow f(\bar{x}_k) - f(x^*) \le \frac{2(G^2+\sigma^2)}{\mu} \cdot \frac{1+\log(k+1)}{k+1}$

$\therefore$ convergence rate is $O(\frac{\log k}{k})$ which is **smaller than SGD** $O(\frac{1}{\sqrt{k}})$

- comparison

	GD	SGD
convex	$O(\frac{1}{k})$	$O(\frac{1}{\sqrt{k}})$
strongly convex	$O(c^k)$ $(c<1)$	$O(\frac{\log k}{k}) \xrightarrow[\text{can become}]{} O(\frac{1}{k})$

- **RMSProp**

$S_t = r S_{t-1} + (1-r)(\nabla \mathcal{L}(x_t))^2$ → element wise!

: vector

$\begin{pmatrix} 4 \\ 2 \\ 1 \end{pmatrix}^2 = \begin{pmatrix} 16 \\ 4 \\ 1 \end{pmatrix}$

$$x_{t+1} = x_t - \frac{\alpha}{\sqrt{S_t + \varepsilon}} \nabla \mathcal{L}(x_t)$$

elementwise $\begin{pmatrix} 4 \\ 2 \\ 1 \end{pmatrix} / \begin{pmatrix} 2 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 1/4 \end{pmatrix}$

- **Adam (Adaptive moment Estimation)**

$$V_t = \beta_1 V_{t-1} + (1-\beta_1) \nabla \mathcal{L}(x_t)$$

$$S_t = \beta_2 S_{t-1} + (1-\beta_2)(\nabla \mathcal{L}(x_t))^2$$

$$\hat{V}_t = \frac{V_t}{1-\beta_1} \qquad \hat{S}_t = \frac{S_t}{1-\beta_2}$$

$$x_{t+1} = x_t - \frac{\alpha}{\sqrt{S_t + \varepsilon}} \hat{V}_t$$

<u>thm</u> Adam converges to $\forall$ convex fcn

$\Rightarrow$ it is not proved

- SAM (Sharpness Aware Minimization)

  finding flat minima

  $$\min_{x} \left\{ \max_{\varepsilon : \|\varepsilon\| \le \rho} \mathcal{L}(x+\varepsilon) \right\}$$