Linear Algebra

X : vector of random variable $X_1, X_2, \cdots, X_n$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

C_X : covariance matrix of X

$$C_X = \begin{bmatrix} Var(X_1) & Cov(X_2, X_1) & \cdots & Cov(X_n, X_1) \\ Cov(X_1, X_2) & Var(X_2) & \cdots & Cov(X_n, X_2) \\ \vdots & \vdots & & \vdots \\ Cov(X_1, X_n) & Cov(X_2, X_n) & \cdots & Var(X_n) \end{bmatrix}$$

$$= E[(X - E(X))(X - E(X))^T]$$

$Y = AX + b$

$$\begin{aligned} C_Y &= E[(Y - E(Y))(Y - E(Y))^T] \\ &= E[(AX+b - E(AX+b))(AX+b - E(AX+b))^T] \\ &= E[(AX - E(AX))(AX - E(AX))^T] \\ &= E[A(X - E(X))(X - E(X))^T A^T] \quad \Big\} \ (AB)^T = B^T A^T \\ &= A E[(X - E(X))(X - E(X))^T] A^T \\ &= A C_X A^T \end{aligned}$$

$f_Y(y)$: probability density function

$$Y = G(X) \quad , \quad X = G^{-1}(Y) = H(Y)$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} H_1(Y) \\ H_2(Y) \\ \vdots \\ H_n(Y) \end{bmatrix}$$

$$J = \det \begin{pmatrix} \partial H_1 / \partial y_1 & \partial H_1 / \partial y_2 & \cdots & \partial H_1 / \partial y_n \\ \partial H_2 / \partial y_1 & \partial H_2 / \partial y_2 & \cdots & \partial H_2 / \partial y_n \\ \vdots & \vdots & & \vdots \\ \partial H_n / \partial y_1 & \partial H_n / \partial y_2 & \cdots & \partial H_n / \partial y_n \end{pmatrix}$$

$$f_Y(y) = f_X(H(y)) \, |J|$$

$$Y = AX + b = G(X)$$

$$X = A^{-1}(Y - b) = H(Y)$$

$$J = \det(A^{-1}) = \frac{1}{\det(A)}$$

$$f_Y(y) = f_X(H(y)) \, |J|$$

$$= f_X(A^{-1}(y-b)) \left| \frac{1}{\det(A)} \right|$$

Probability density function of multivariate Gaussian distribution

$$Z = [Z_1, Z_2, \cdots, Z_n]^T, \quad Z_i \sim N(0, 1), \quad Z_i \text{ i.i.d.}$$

$$f_Z(z) = \prod_{i=1}^{n} f_{Z_i}(z_i)$$

$$= \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} z_i^2\right)$$

$$= \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{1}{2} z^T z\right)$$

Probability density function of multivariate normal distribution

$$X = AZ + m \ , \quad Z \sim (0, I)$$

$$E(X) = m$$

$$C_X = A C_Z A^T = AA^T$$

$$\det(C_X) = \det(AA^T) = \det(A)\det(A^T) = (\det(A))^2$$

$$\sqrt{\det(C_X)} = |\det(A)|$$

$$
\begin{aligned}
f_X(x) &= f_Z(A^{-1}(x-m)) \left| \frac{1}{\det(A)} \right| \\
&= \frac{1}{\sqrt{2\pi}^n} \left| \frac{1}{\det(A)} \right| \exp\left(-\frac{1}{2}(A^{-1}(x-m))^T A^{-1}(x-m)\right) \\
&= \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{\det(C_X)}} \exp\left(-\frac{1}{2}(x-m)^T (A^{-1})^T A^{-1}(x-m)\right) \\
&= \frac{1}{\sqrt{2\pi}^n \sqrt{\det(C_X)}} \exp\left(-\frac{1}{2}(x-m)^T (A^T A)^{-1}(x-m)\right) \\
&= \frac{1}{\sqrt{2\pi}^n \sqrt{\det(C_X)}} \exp\left(-\frac{1}{2}(x-m)^T C_X^{-1}(x-m)\right)
\end{aligned}
$$

MLE

$$-\log p_\theta(y|x) = -\frac{1}{2}(f_\theta(x)-y)^T \Sigma_\theta(x)^{-1}(f_\theta(x)-y) - \frac{1}{2}\log|\Sigma_\theta(x)| - \cancel{\frac{n}{2}\log 2\pi}$$

$$\left(\begin{array}{l} N(f_\theta(x), \Sigma_\theta(x)) \\ \Sigma_\theta(x) = I \end{array} \right. \qquad = -\frac{1}{2}(f_\theta(x)-y)^T(f_\theta(x)-y)$$

$$= -\frac{1}{2}\| f_\theta(x) - y \|^2$$

Analyze error

$$L(\theta, x, y) = -\frac{1}{2}\|f_\theta(x) - y\|^2$$

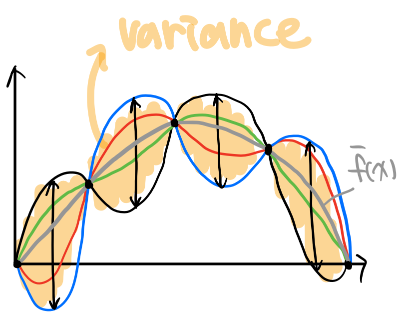
$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

$$p(D) = \prod_{i=1}^{n} p(x_i)\, p(y_i | x_i)$$

$$E_{D\sim p(D)}[\|f_\theta(x) - y\|^2] = E_{D\sim p(D)}[\|f_D(x) - f(x)\|^2]$$
$$= \sum p(D)\, \|f_D(x) - f(x)\|^2$$

let $\bar{f}(x) = E_{D\sim p(D)}[f_D(x)]$

$$E_{D\sim p(D)}[\|f_D(x) - f(x)\|^2] = E_{D\sim p(D)}[\|f_D(x) - \bar{f}(x) + \bar{f}(x) - f(x)\|^2]$$
$$= E_{D\sim p(D)}[\|(f_D(x) - \bar{f}(x)) + (\bar{f}(x) - f(x))\|^2]$$
$$= E_{D\sim p(D)}[\|f_D(x) - \bar{f}(x)\|^2] + E_{D\sim p(D)}[\|\bar{f}(x) - f(x)\|^2]$$
$$+ \cancel{E_{D\sim p(D)}[2(f_D(x) - \bar{f}(x))^T(\bar{f}(x) - f(x))]}$$
$$= \underbrace{E_{D\sim p(D)}[\|f_D(x) - \bar{f}(x)\|^2]}_{\text{Variance}} + \underbrace{E_{D\sim p(D)}[\|\bar{f}(x) - f(x)\|^2]}_{\text{Bias}^2}$$



Bias² related with model capacity not data

variance ↑ : overfitting

bias ↑ : underfitting

Regularization

Bayesian perspective : add prior on parameters

Given D, what is most likely θ ?

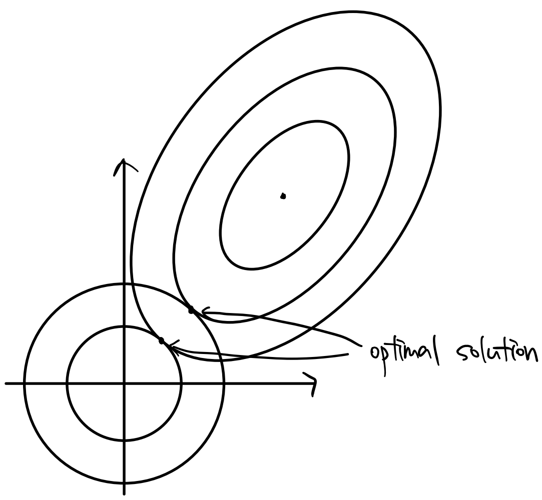
$$p(\theta \mid D) = \frac{p(\theta, D)}{p(D)} \propto p(\theta, D) = p(D \mid \theta)\, \underbrace{p(\theta)}_{\text{prior}}$$

$$p(D \mid \theta) = \prod_{i=1}^{N} p_\theta(y_i \mid x_i)\, p(x_i)$$

chapter 2 - given θ, maximize probability of D

$\Rightarrow$ New loss function : $-\sum \log p_\theta(y_i \mid x_i) - \log p(\theta)$

small number → higher probability



if $\theta \sim N(0, \sigma^2)$ $\qquad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\theta^2}{2\sigma^2}\right)$

$$\log p(\theta) = \sum_i -\frac{\theta_i^2}{2\sigma^2} - \cancel{\frac{1}{2}\log 2\pi\sigma^2}$$

$$= -\lambda \|\theta\|^2 \quad \left(\lambda = \frac{1}{2\sigma^2}, \text{ hyperparameter}\right)$$

→ L2 loss
L2 regularization
weight decay

if $\theta \sim \text{Laplace}(0, b)$ $\qquad \frac{1}{2b} \exp\left(-\frac{|\theta|}{b}\right)$

$$\log p(\theta) = \sum_i -\frac{|\theta_i|}{b} - \cancel{\log 2b}$$

$$= -\lambda |\theta|$$

→ L1 loss
L1 regularization

MLE (Maximum Likelihood Estimation) : $p(D \mid \theta)$
MAP (Maximum A Posterior) : $p(\theta \mid D)$