白板推導(1∼23) Notes

李笑然 xiaoranli@daum.net

March.2021

Contents

Chapter 1

Comparison of frequentist

1.1 Gaussian distribution

1.1.1 MLE of Gaussian Distribution

Definition 1.1.1. 一次元 *Gaussian distribution*：$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$

θ_{MLE}:

$$\theta_{MLE} = \underset{\theta}{argmax}\, p(x|\theta) = \log \prod_{i=1}^{N} p(x_i|\theta) \tag{1.1}$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) = \sum_{i=1}^{N}\left[\log\frac{1}{2\pi} + \log\frac{1}{\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}\right] \tag{1.2}$$

μ_{MLE}:

$$\mu_{MLE} = \underset{\mu}{argmax} \log p(x|\theta) = \underset{\mu}{argmax} \sum_{i=1}^{N} -\frac{(x_i-\mu)^2}{2\sigma^2} \tag{1.3}$$

$$= \underset{\mu}{argmin} \sum_{i=1}^{N} (x_i-\mu)^2 \tag{1.4}$$

σ^2_{MLE}:

$$\sigma^2_{MLE} = \underset{\sigma}{argmax} \tag{1.5}$$

$$= \underset{\sigma}{argmax} \sum_{i=1}^{N} \left(-\log\sigma - \frac{1}{2\sigma^2}(x_i-\mu)^2\right) \tag{1.6}$$

$\mu Extremum$：Unbiased estimate

$$\frac{\partial}{\partial \mu}\sum_{i=1}^{N}(x_i-\mu)^2=\sum_{i=1}^{N}2\cdot(x_i-\mu)\cdot(-1)=0 \tag{1.7}$$

$$\sum_{i=1}^{N}(x_i-\mu)=\sum_{i=1}^{N}x_i-\sum_{i=1}^{N}\mu=0 \tag{1.8}$$

$$\mu_{MLE}=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{1.9}$$

$\sigma Extremum$：Biased estimate（Unbiased estimate：$\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2$）

$$\frac{\partial}{\partial \sigma}\sum_{i=1}^{N}\left(-\log\sigma-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)=0 \tag{1.10}$$

$$\sum_{i=1}^{N}\left[-\frac{1}{\sigma}-\frac{1}{2}(x_i-\mu)^2\cdot(-2)\sigma^{-3}\right]=0 \tag{1.11}$$

$$\sum_{i=1}^{N}\left[-\frac{1}{\sigma}+(x_i-\mu)^2\cdot\sigma^{-3}\right]=0 \tag{1.12}$$

$$\sum_{i=1}^{N}\left[-\sigma^2+(x_i-\mu)^2\right]=0 \tag{1.13}$$

$$\sum_{i=1}^{N}\sigma^2=\sum_{x_i}^{N}(x_i-\mu)^2 \tag{1.14}$$

$$\begin{aligned}\sigma^2_{MLE}&=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2 &(1.15)\\&=\frac{1}{N}\sum_{i=1}^{N}(x_i^2-2x_i\mu_{MLE}+\mu_{MLE}^2) &(1.16)\\&=\frac{1}{N}\sum_{i=1}^{N}x_i^2-2\mu_{MLE}^2+\mu_{MLE}^2 &(1.17)\\&=\frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu_{MLE}^2 &(1.18)\end{aligned}$$

1.1.2 Biased estimate vs Unbiased estimate

Unbiased estimate：

$$E[\mu_{MLE}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu \tag{1.19}$$

Biased estimate:

$$\begin{aligned}
E[\sigma^2_{MLE}] &= E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2_{MLE}\right] && (1.20)\\
&= E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right) - \left(\mu^2_{MLE} - \mu^2\right)\right] && (1.21)\\
&= E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right] - E\left[\mu^2_{MLE} - \mu^2\right] && (1.22)\\
&= E\left[\frac{1}{N}\sum_{i=1}^{N} (x_i^2 - \mu^2)\right] - E[\mu^2_{MLE}] - E[\mu^2] && (1.23)\\
&= \frac{1}{N}\sum_{i=1}^{N} E[x_i^2 - \mu^2] - E[\mu^2_{MLE}] - \mu^2 && (1.24)\\
&= \frac{1}{N}\sum_{i=1}^{N} (E[x_i^2] - \mu^2) - E[\mu^2_{MLE}] - E^2[\mu_{MLE}] && (1.25)\\
&= \frac{1}{N}\sum_{i=1}^{N} Var[x_i] - Var[\mu_{MLE}] && (1.26)\\
&= \frac{1}{N}\sum_{i=1}^{N} \sigma^2 - Var\left[\frac{1}{N}\sum_{i-1}^{N} x_i\right] && (1.27)\\
&= \sigma^2 - \frac{1}{N^2}\sum_{i=1}^{N} \sigma^2 && (1.28)\\
&= \sigma^2 - \frac{1}{N}\sigma^2 && (1.29)\\
&= \frac{N-1}{N}\sigma^2 && (1.30)
\end{aligned}$$

1.1.3 High-dimensional

Definition 1.1.2.

$$\mathcal{X} \sim N(\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{dim}{2}} \cdot |\Sigma|^{\frac{1}{2}}} \exp\Bigg(-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu) \Bigg) \tag{1.31}$$

Definition 1.1.3. *Mahalanobis Distance:*

$$(x-\mu)^t \Sigma^{-1} (x-\mu) \tag{1.32}$$

Definition 1.1.4. *Euclidean distance:*

$$(x-\mu)^t I (x-\mu) \tag{1.33}$$

Define: Variance matrix∈ Positive definite matrix: $U\Lambda U^t, UU^t = U^t U = I, \Lambda = diag(\lambda_i)$

$$\Sigma = U\Lambda U^t \tag{1.34}$$

$$= (\mu_1, \mu_2, \cdots \mu_{dim}) \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{dim} \end{pmatrix} \begin{pmatrix} \mu_1^t \\ \mu_2^t \\ \vdots \\ \mu_{dim}^t \end{pmatrix} \tag{1.35}$$

$$= (\mu_1\lambda_1, \mu_2\lambda_2, \cdots \mu_{dim}\lambda_{dim}) \begin{pmatrix} \mu_1^t \\ \mu_2^t \\ \vdots \\ \mu_{dim}^t \end{pmatrix} \tag{1.36}$$

$$= \sum_{i=1}^{dim} \mu_i \lambda_i \mu_i^t \tag{1.37}$$

$$\Sigma^{-1} = (U\Lambda U^t)^{-1} = (U^t)^{-1}\Lambda^{-1}U^{-1} = U\Lambda^{-1}U^t \tag{1.38}$$

$$= \sum_{i=1}^{dim} \mu_i \frac{1}{\lambda_i} \mu_i^t \tag{1.39}$$

$$(x-\mu)^t\Sigma^{-1}(x-\mu) = (x-\mu)^t\sum_{i=1}^{dim}\mu_i\frac{1}{\lambda_i}\mu_i^t(x-\mu) \tag{1.40}$$
$$= \sum_{i=1}^{dim}(x-\mu)^t\mu_i\frac{1}{\lambda_i}\mu_i^t(x-\mu) \tag{1.41}$$
$$= \sum_{i=1}^{dim}\frac{((x-\mu)^t\mu_i)^2}{\lambda_i} \tag{1.42}$$

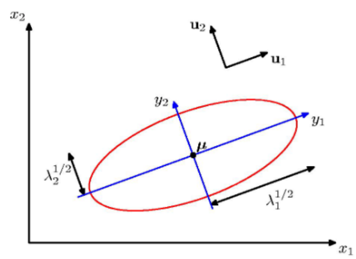

Figure 1.1: 2-dimensional Gaussian distribution

Reference: [1]

1.1.4 Limitations

The complexity: $\Sigma_{dim\times dim} = \frac{dim^2-dim}{2} + dim = O(dim^2)$
Simplification: diagonal matrix(factor analysis) & isotropy(等方性,P-PCA)
Reference: [2]

1.1.5 Conditional and Marginal probability of Mixed Gaussian Distribution

Knowing the joint probability distribution, find the marginal probability distribution and the conditional probability distribution.

Theorem 1.1.1. *If* $\mathcal{X} \sim N(\mu,\Sigma); Y = AX + B$ *then* $\mathcal{Y} \sim N(A\mu + B, A\Sigma A^t)$

[1] https://community.rstudio.com/t/3d-surface-with-a-2d-projection-using-r/17790/2
[2] https://sites.northwestern.edu/msia/2016/12/08/k-means-shouldnt-be-our-only-choice/

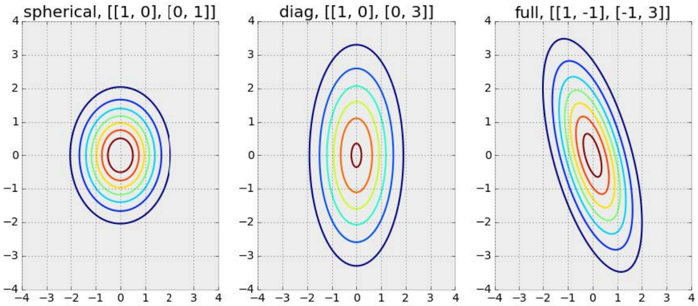

Figure 1.2: Limitations of Gaussian distribution

Proof.

$$E[Y] = E[AX+B] = AE[X] + B = A\mu + B \tag{1.43}$$

$$Var[Y] = Var[AX+B] = A \cdot Var[AX] \cdot A^t = A \cdot \Sigma \cdot A^t \tag{1.44}$$

□

Define: $\mathcal{X} = \begin{pmatrix} X_a \\ X_b \end{pmatrix}; X_a \in \mathbb{R}^{m\times m}, X_b \in \mathbb{R}^{n\times n}; m + n = \mathcal{X}_{dim}; \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}; \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$

Solve for $P(X_a, P(X_b|X_a); P(X_b, P(X_a|X_b)$

Constructive proof: $\neq$ (PRML: Matching method proof)

s.t. $X_a = (I_m, 0)\begin{pmatrix} X_a \\ X_b \end{pmatrix}$

Proof.

$$E[X_a] = (I_m, 0)\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_a \tag{1.45}$$

$$Var[X_a] = (I_m, 0)\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\begin{pmatrix} I_m \\ 0 \end{pmatrix} \tag{1.46}$$

$$= (\Sigma_{aa}, \Sigma_{ab})\begin{pmatrix} I_m \\ 0 \end{pmatrix} = \Sigma_{aa} \tag{1.47}$$

□

Theorem 1.1.2. *Schur complement(シューア補行列) of* Σ_{aa}*:* $\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}$

Define: $X_{b\cdot a} = X_b - \Sigma_{ba}\Sigma_{aa}^{-1}X_a$

Proof.

$$X_{a\cdot b} = (-\Sigma_{ba}\Sigma_{aa}^{-1}, I)\begin{pmatrix} X_a \\ X_b \end{pmatrix} \tag{1.48}$$

$$E[X_{b\cdot a}] = (-\Sigma_{ba}\Sigma_{aa}^{-1}, I)\cdot\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \tag{1.49}$$

$$= \mu_b - \Sigma_{ba}\Sigma_{aa}^{-1}\mu_a \tag{1.50}$$

$$Var[X_{b\cdot a}] = (-\Sigma_{ba}\Sigma_{aa}^{-1}, I)\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\begin{pmatrix} -\Sigma_{aa}^{-1}\Sigma_{ba}^{t} \\ I \end{pmatrix} \tag{1.51}$$

$$= (0, \Sigma_{bb\cdot a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})\begin{pmatrix} -\Sigma_{aa}^{-1}\Sigma_{ba}^{t} \\ I \end{pmatrix} \tag{1.52}$$

$$= \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \tag{1.53}$$

$$X_b = X_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}X_a \tag{1.54}$$

$$E[X_b|X_a] = \mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}X_a \tag{1.55}$$

$$Var[X_b|X_a] = Var[X_{b\cdot a}] = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \tag{1.56}$$

□

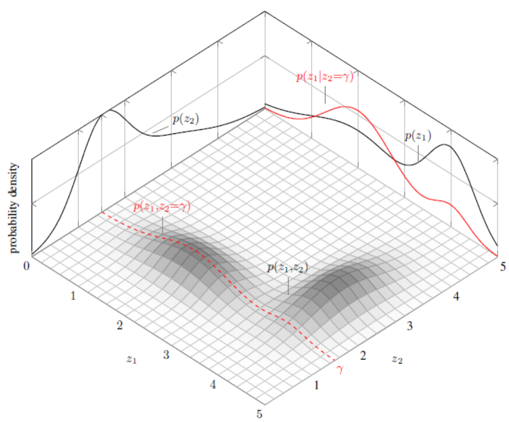

Figure 1.3: Conditional probability Gaussian distribution Link

1.1.6 Joint Probability of Mixed Gaussian Distribution

Know the marginal probability distribution and the conditional probability distribution to find the joint probability distribution.
Define: $P(x) = N(X|\mu, \Lambda^{-1}); P(Y|X) = N(Y|AX+B, L^{-1}); \Lambda^{-1}, L^{-1} \in precisionmatrix = (covariancematrix)^{-1}; Y = AX + B + \epsilon, \epsilon \sim N(0, L^{-1}), \epsilon \perp\!\!\!\perp X$

Solve for $P(Y); P(X|Y)$

Proof. $P(Y)$

$$E[Y] = E[AX + B + \epsilon] = E[AX + B] + E[\epsilon] = A\mu + B \tag{1.57}$$

$$Var[Y] = Var[Ax + B + \epsilon] = Var[AX + B] + Var[\epsilon] = A \cdot \lambda \cdot A^t + L^{-1} \tag{1.58}$$

□

Proof. $P(X, Y)$

$$joint\ probability = \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{bmatrix} \mu \\ A\mu + B \end{bmatrix}, \begin{bmatrix} \Lambda^{-1} & Cov(x, y) \\ Cov(x, y) & A \cdot \lambda \cdot A^t + L^{-1} \end{bmatrix}\right] \tag{1.59}$$

$$\tag{1.60}$$

□

Proof. $Cov(x, y)$

$$\begin{aligned}
Cov(x, y) &= E[(x - E[x]) \cdot (y - E[y])^t] & (1.61)\\
&= E[(x - \mu)(y - A\mu - b)^t] & (1.62)\\
&= E[(x - \mu)(Ax + b + \epsilon - A\mu - b)^t] & (1.63)\\
&= E[(x - \mu)(Ax - A\mu + \epsilon)^t] & (1.64)\\
&= E[(x - \mu)(Ax - A\mu)^t + (x - \mu)\epsilon] & (1.65)\\
&= E[(x - \mu)(Ax - A\mu)^t] + E[(x - \mu)\epsilon] & (1.66)\\
&= E[(x - \mu)(Ax - A\mu)^t] & (1.67)\\
&= E[(x - \mu)(x - \mu)^t \cdot A^t] & (1.68)\\
&= E[(x - \mu)(x - \mu)^t] \cdot A^t & (1.69)\\
&= Var[x] \cdot A^t & (1.70)\\
&= \Lambda^{-1} A^t & (1.71)
\end{aligned}$$

□

Knowing the joint probability distribution, find the conditional probability distribution.

$$joint\ probability = \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{bmatrix} \mu \\ A\mu + B \end{bmatrix}, \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1} A^t \\ \Lambda^{-1} A^t & A \cdot \lambda \cdot A^t + L^{-1} \end{bmatrix}\right] \tag{1.72}$$

$$\tag{1.73}$$

1.2 Exponential Distribution

1.2.1 Introduction

Sufficient statistics, Conjugation, maximum entropy, generalized linear model, probability graph model, variational inference.

Theorem 1.2.1. *Exponential distribution:*
η*: Parameter vector;* $x \in \mathbb{R}^n$*;* $\phi(x)$*: Sufficient statistics(online learning)*

$$p(x|\eta) = h(x) \exp\left(\eta^t \phi(x) - A(\eta)\right) \tag{1.74}$$

Theorem 1.2.2. *Partition function:* z

$$p(x|\theta) = \frac{1}{z}\hat{p}(x|\theta) \tag{1.75}$$

$$\int p(x|\theta)dx = \int \frac{1}{z}\hat{p}(x|\theta)dx \tag{1.76}$$

$$1 = \int \frac{1}{z}\hat{p}(x|\theta)dx \tag{1.77}$$

$$z = \int \hat{p}(x|\theta)dx \tag{1.78}$$

Theorem 1.2.3. *Log partition function:* $A(\eta)$

$$p(x|\eta) = h(x) \cdot \exp\left(\eta^t \phi(x)\right) \cdot \exp\left(A(\eta)\right) \tag{1.79}$$

$$= \frac{1}{\exp\left(A(\eta)\right)} h(x) \cdot \exp\left(\eta^t \phi(x)\right) \tag{1.80}$$

$$= \frac{1}{z}\hat{p}(x|\theta) \tag{1.81}$$

1.2.2 Proof of Gaussian Distribution to Exponential Distribution

Proof.

$$
\begin{aligned}
p(x) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\big[-\frac{(x-\mu)^2}{2\sigma^2}\big] && (1.82)\\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\big[-\frac{1}{2\sigma^2}(x^2-2\mu x+\mu^2)\big] && (1.83)\\
&= \exp\log(2\pi\sigma^2)^{-\frac{1}{2}}\cdot\exp\left[-\frac{1}{2\sigma^2}(x^2(x^2-2\mu x)-\frac{\mu^2}{2\sigma^2}(x^2\right] && (1.84)\\
&= \exp\log(2\pi\sigma^2)^{-\frac{1}{2}}\cdot\exp\left[-\frac{1}{2\sigma^2}(-2\mu,1)\begin{pmatrix}x\\x^2\end{pmatrix}-\frac{\mu^2}{2\sigma^2}\right] && (1.85)\\
&= \exp\left[(\frac{\mu}{\sigma^2},-\frac{1}{2\sigma^2})\begin{pmatrix}x\\x^2\end{pmatrix}-\left(\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\log 2\pi\sigma^2\right)\right) && (1.86)\\
&= \exp\left[(\eta_1,\eta_2)\begin{pmatrix}x\\x^2\end{pmatrix}-\left(-\frac{\eta_1^2}{4\eta_2}+\frac{1}{2}\log(-\frac{\pi}{\eta_2})\right)\right] && (1.87)
\end{aligned}
$$

□

1.2.3 Relationship between $\phi(x)$ and $A(\eta)$

$A'(\eta) = E_{p(x|\eta)}[\eta(x)]$, $A''(\eta) = Var_{p(x|\eta)}[\eta(x)]$, $A(\eta)$ is convex function.

Proof.

$$\begin{aligned}
p(x|\eta) &= h(x)\cdot\exp\big(\eta^t\phi(x)\big)\cdot\exp\big(A(\eta)\big) && (1.88)\\
&= \frac{1}{\exp\big(A(\eta)\big)}h(x)\cdot\exp\big(\eta^t\phi(x)\big) && (1.89)\\
\exp\big(A(\eta)\big) &= \int h(x)\cdot\exp\big(\eta^t\phi(x)\big)dx && (1.90)\\
\exp\big(A(\eta)\big)\dot{A}'(\eta) &= \frac{\partial}{\partial}\Big[\int h(x)\cdot\exp\big(\eta^t\phi(x)\big)dx\Big] && (1.91)\\
&= \int h(x)\cdot\exp\big(\eta^t\phi(x)\big)\cdot\phi(x)dx && (1.92)\\
A'(\eta) &= \frac{\int h(x)\cdot\exp\big(\eta^t\phi(x)\big)\cdot\phi(x)dx}{\exp\big(A(\eta)\big)} && (1.93)\\
&= \int h(x)\cdot\exp\big(\eta^t\phi(x)-A(\eta)\big)\cdot\phi(x)dx && (1.94)\\
&= \int p(x|\eta)\cdot\phi(x)dx && (1.95)\\
&= E_{p(x|\eta)}[\eta(x)] && (1.96)
\end{aligned}$$

□

1.2.4 MLE of Exponential Distribution

Proof.

$$\begin{aligned}
\eta_{MLE} &= \underset{\eta}{argmax}\log\prod_{i=1}^{N}p(x_i|\eta) && (1.97)\\
&= \underset{\eta}{argmax}\sum_{i=1}^{N}\log p(x_i|\eta) && (1.98)\\
&= \underset{\eta}{argmax}\sum_{i=1}^{N}\log\Big[h(x_i)\cdot\exp\big(\eta^t\phi(x_i)-A(\eta)\big)\Big] && (1.99)\\
&= \underset{\eta}{argmax}\sum_{i=1}^{N}\Big[\log h(x_i)\cdot\eta^t\phi(x_i)-A(\eta)\Big] && (1.100)\\
&= \underset{\eta}{argmax}\sum_{i=1}^{N}\Big[\eta^t\phi(x_i)-A(\eta)\Big] && (1.101)
\end{aligned}$$

$$\frac{\partial}{\partial \eta}\sum_{i=1}^{N}\left[\eta^t\phi(x_i)-A(\eta)\right] = \sum_{i=1}^{N}\frac{\partial}{\partial \eta}\left[\eta^t\phi(x_i)-A(\eta)\right] \tag{1.102}$$

$$= \sum_{i=1}^{N}\phi(x_i) - \sum_{i=1}^{N}A'(\eta) \tag{1.103}$$

$$= \sum_{i=1}^{N}\phi(x_i) - NA'(\eta) \tag{1.104}$$

$$\sum_{i=1}^{N}\phi(x_i) - NA'(\eta) = 0 \tag{1.105}$$

$$A'(\eta_{MLE}) = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i) \tag{1.106}$$

□

1.2.5 Uniform Distribution of Maximum Entropy

Theorem 1.2.4.

$$H[p] = -\sum_{x} p(x)\log p(x) \tag{1.107}$$

Define: $\sum_{i=1}^{k} p_i = 1$

Proof. Uniform distribution $\Longleftrightarrow$ Maximum Entropy

$$\mathcal{L}(p,\lambda) = \sum_{i=1}^{k} p_i \log p_i + \lambda(1-\sum_{i=1}^{k} p_i) \tag{1.108}$$

$$\frac{\mathcal{L}}{p_i} = \log p_i + p_i\frac{1}{p_i} - \lambda = 0 \tag{1.109}$$

$$\log p_i + 1 - \lambda = 0 \tag{1.110}$$

$$\exp(\lambda - 1) = \hat{p}_i \tag{1.111}$$

$$\hat{p}_i = constant \tag{1.112}$$

$$\hat{p}_i = \frac{1}{k} \tag{1.113}$$

□

1.2.6 Maximum Entropy to Exponential Distribution

Define: $\sum_{i=1}^{k} p_i = 1; E_p[f(x)] = E_{\hat{p}}[f(x)] = \Delta$

Proof.

$$\mathcal{L}(p, \lambda_0, \lambda_1) = \sum_{i=1}^{k} p(x) \log p(x) + \lambda_0 (1 - \sum_{x} p(x)) + \lambda^t (\Delta - E_p[f(x)]) \tag{1.114}$$

$$\begin{aligned}
\frac{\mathcal{L}}{p(x)} = \sum_{x} \left(\log p(x) + p(x) \cdot \frac{1}{p(x)} \right) - \sum_{x} \lambda_0 - \sum_{x} \lambda^t f(x) = 0 && (1.115) \\
\sum_{x} \left(\log p(x) + 1 - \lambda_0 - \lambda^t f(x) \right) = 0 && (1.116) \\
\log p(x) + 1 - \lambda_0 - \lambda^t f(x) = 0 && (1.117) \\
\lambda^t f(x) + \lambda_0 - 1 = \log p(x) && (1.118) \\
\exp(\lambda^t f(x) + \lambda_0 - 1) = p(x) && (1.119) \\
\exp \left[\lambda^t f(x) - (\lambda_0 - 1) \right] = p(x) && (1.120)
\end{aligned}$$

□

1.3 Linear regression

1.3.1 Two geometric interpretations of Linear Regression

Theorem 1.3.1. *Least squares method*

$$\begin{aligned}
L(w) &= \sum_{i=1}^{N} ||w^t x_i - y_i||^2 && (1.121) \\
&= (W^t X^t - Y^t) && (1.122) \\
&= W^t X^t X W - W^t X^t Y - Y^t X W + Y^t Y && (1.123) \\
&= W^t X^t X W - 2W^t X^t Y + Y^t Y && (1.124)
\end{aligned}$$

Linear regression has analytical solutions:

$$\begin{aligned}
\frac{\partial L(w)}{\partial w} = 2X^t X W - 2X^t Y = 0 && (1.125) \\
X^t X W = X^t Y && (1.126) \\
W = (X^t X)^{-1} X^t Y && (1.127)
\end{aligned}$$

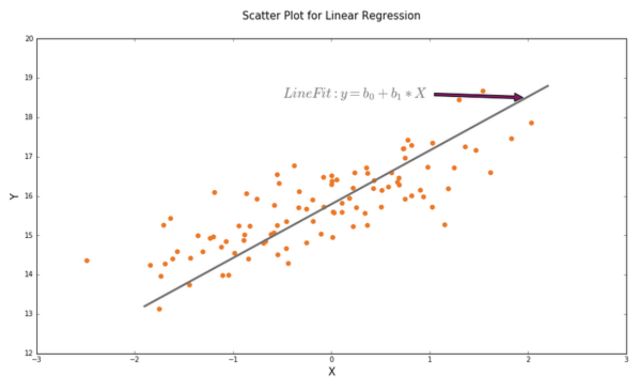

Figure 1.4: geometric interpretations 1th

Proof. geometric interpretations 2nd

$$f(w) = W^t X = X^t \beta \tag{1.128}$$
$$X^t(Y - X\beta) = 0_{dim \times 1} \tag{1.129}$$
$$X^t Y = X^t X \beta \tag{1.130}$$
$$\beta = (X^t X)^{-1} X^t Y \tag{1.131}$$

□

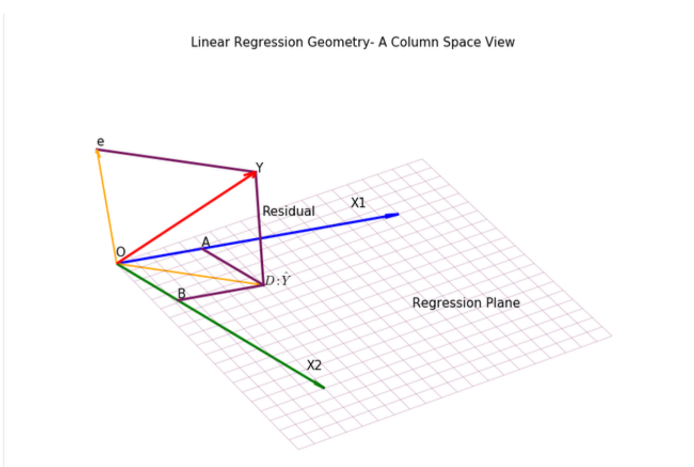

Figure 1.5: geometric interpretations 2nd

Reference: [3]

[3] https://www.datasciencecentral.com/profiles/blogs/linear-regression-geometry

1.3.2 MLE with Gaussian noise for Least Squares Method

Define: $\epsilon \sim N(0,\sigma); y = w^t x + \epsilon; y|x;w \sim N(w^t x, \sigma^2); p(y|x;w) = \frac{1}{\sqrt{2\pi}} exp\Big(-\frac{(y-w^t x)^2}{2\sigma^2}\Big)$

Proof. MLE:log-likelihood

$$\log P(Y|X;w) = \log \prod_{i=1}^{N} P(y_i|x_i;w) \tag{1.132}$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} + \log exp\Big(-\frac{(y_i - w^t x_i)^2}{2\sigma^2}\Big) \tag{1.133}$$

$$= \sum_{i=1}^{N} \Big(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - w^t x_i)^2}{2\sigma^2}\Big) \tag{1.134}$$

$$\hat{w} = \underset{w}{argmax} - \frac{(y_i - w^t x_i)^2}{2\sigma^2} \tag{1.135}$$

$$= \underset{w}{argmin} (y_i - w^t x_i)^2 \tag{1.136}$$

□

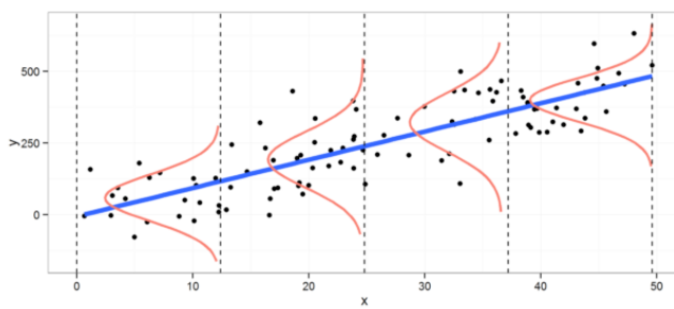

Figure 1.6: MLE with Gaussian noise for least squares method

Reference: [4]

1.3.3 L2 in frequency perspective

Turn the positive semi-definite matrix into a positive definite matrix.
$L1 : Lasso, P(w) = ||w||_1$
$L2 : Ridge(Weight\ decay), P(w) = ||w||_2^2 = w^t w$
Ridge regression's analytical solutions:

[4] `https://suriyadeepan.github.io/2017-01-22-mle-linear-regression/`

Proof. L2: positive semi-definite → positive definite matrix

$$J(w) = \sum_{i=1}^{N} ||w^t x_i - y_i||^2 + \lambda w^t w \tag{1.137}$$
$$= (w^t X^t - Y^t)(Xw - Y) + \lambda w^t w \tag{1.138}$$
$$= w^t X^t X w - w^t X^t Y - Y^t X w + Y^t Y + \lambda w^t w \tag{1.139}$$
$$= w^t X^t X w - 2 w^t X^t Y + Y^t Y + \lambda w^t w \tag{1.140}$$
$$= w^t (X^t X + \lambda I) w - 2 w^t X^t Y + Y^t Y \tag{1.141}$$

$$\frac{\partial J(w)}{\partial w} = 2(X^t X + \lambda I) w - 2X^t Y = 0 \tag{1.142}$$
$$\hat{w} = (X^t X + \lambda I)^{-1} X^t Y \tag{1.143}$$

□

1.3.4 MAP For L2

Proof. Maximum A Posteriori for ridge regression
s.t. $y|x; w \sim N(w^t x, \lambda^2) \to p(y|x) = \frac{1}{\sqrt{2\pi}\sigma} exp\big(-\frac{(y-w^t x)^2}{2\sigma^2}\big)$
$w \sim N(0, \lambda_0^2) \to p(y|x) = \frac{1}{\sqrt{2\pi}\sigma} exp\big(-\frac{||w||^2}{2\sigma_0^2}\big)$

$$\hat{w}_{MAP} = \underset{w}{argmax} \prod_{i=1}^{N} p(w|y) \tag{1.144}$$
$$= \underset{w}{argmax} \log \prod_{i=1}^{N} p(y|w) \cdot p(w) \tag{1.145}$$
$$= \underset{w}{argmax} \sum_{i=1}^{N} \log \left(\frac{1}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{2\pi}\sigma_0} \right) + \log exp\left(-\frac{(y - w^t x)^2}{2\sigma^2} - \frac{||w||^2}{2\sigma_0^2} \right) \tag{1.146}$$
$$= \underset{w}{argmin} \sum_{i=1}^{N} \frac{(y - w^t x)^2}{2\sigma^2} + \frac{||w||^2}{2\sigma_0^2} \tag{1.147}$$
$$= \underset{w}{argmin} \sum_{i=1}^{N} (y - w^t x)^2 + \frac{\sigma^2}{\sigma_0^2} ||w||_2^2 \tag{1.148}$$

□

1.4 Linear Classification

1.4.1 Linear regression to Linear classification

線形分類 = 線形回帰（Linear regression）+ 活性化関数（Activation function）
線形分類の特徴: 線形、グローバル、未処理のデータ

$$
\text{線形回帰}\begin{cases}
\text{非線形}\ (Non-linear)\begin{cases}\text{固有値変換}\ (\text{多項式回帰})\\ \text{活性化関数は非線形}\ (sigmoid)\\ Neural\ Networks, Perceptron\end{cases}\\
\text{非グローバル}\ (Non-global):\ Linear\ spline,\ regression\ Decisiontree\\
\text{処理したデータ}\ (Processed\ data):\ PCA,\ \text{多様体学習}\ (Manifoldlearning)
\end{cases}
$$

1.4.2 Classification Model Tree

$$
\text{線形分類}\begin{cases}
Hard\ output\begin{cases}Perceptron, SVM\\ Fisherlineardiscriminant\end{cases}\\
Soft\ output\begin{cases}Discrimination\ model:\ Logistic\ regression\\ Generative\ model\begin{cases}Gaussian\ Discriminant\ Analysis\\ Naive\ Bayes\end{cases}\end{cases}
\end{cases}
$$

1.4.3 Perceptron

1957 から提出したの分類モデル、Deep Learning の深層 Neural Network は多層 Perceptron です、誤分類されたポイントの数を最小限に抑える (Pocket algorithm: 誤分類を許す)。Optimizer: SGD。

Theorem 1.4.1. *Define:* $\mathcal{X} \in Misclassified\ sample$

$$Loss(w) = \sum_{i=1}^{N} I(y_i w^t x_i < 0) \tag{1.149}$$

$$= \sum_{x_i \in \mathcal{X}} -y_i w^T x_i \tag{1.150}$$

$$\nabla_w Loss = -y_i x_i \tag{1.151}$$

1.4.4 Fisher Linear Discriminant Analysis

次元削減方法（PCA の様に）と見なすことができます。 分類のために、多次元データの次元を 1 次元に減らします。

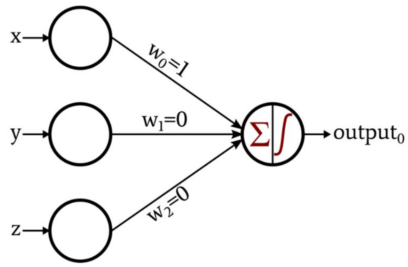

Figure 1.7: Basic Perceptron Neural Network

射影分散 (Covariance) を最大化する，クラスター内の間隔は小さく（high coupling）、クラスター間の間隔は大きくなります（Low aggregation）。

Theorem 1.4.2. *Define:* $x_{c1} \in (x_i|y_i = +1); x_{c2} \in (x_i|y_i = -1)$
$|x_1| = N_1; |x_{c2}| = N_2$
$\mu_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} w^t x_i; \sigma_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(w^t x_i - \mu_1)(w^t x_i - \mu_1)^t$
$\mu_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} w^t x_i; \sigma_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}(w^t x_i - \mu_1)(w^t x_i - \mu_2)^t$
Objective function:

$$\underset{w}{argmax} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1 + \sigma_2} \tag{1.152}$$

Proof.

$$\begin{aligned}
\hat{w} &= \frac{\left(\frac{1}{N_1}\sum_{i=1}^{N_1} w^t x_i - \frac{1}{N_2}\sum_{i=1}^{N_2} w^t x_i\right)^2}{\frac{1}{N_1}\sum_{i=1}^{N_1}\left(w_t x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^t x_j\right)\left(w_t x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^t x_j\right)^t + \sigma_2} && (1.153)\\
&= \frac{\left[w^t\left(\frac{1}{N_1}\sum_{i=1}^{N_1} x_i - \frac{1}{N_2}\sum_{i=1}^{N_2} w^t\right)\right]^2}{\left[\frac{1}{N_1}\sum_{i=1}^{N_1} w^t(x_i - \bar{x}_{c1})(x_i - \bar{x}_{c1})^t w\right] + \sigma_2} && (1.154)\\
&= \frac{\left(w^t(\bar{x}_{c1} - \bar{x}_{c2})\right)^2}{w^t\left[\frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - \bar{x}_{c1})(x_i - \bar{x}_{c1})^t\right]w + \sigma_2} && (1.155)\\
&= \frac{w^t(\bar{x}_{c1} - \bar{x}_{c2})(\bar{x}_{c1} - \bar{x}_{c2})^t w}{w^t\sigma_1 w + w^t\sigma_2 w} && (1.156)\\
&= \frac{w^t(\bar{x}_{c1} - \bar{x}_{c2})(\bar{x}_{c1} - \bar{x}_{c2})^t w}{w^t(\sigma_1 + \sigma_1)w} && (1.157)
\end{aligned}$$

□

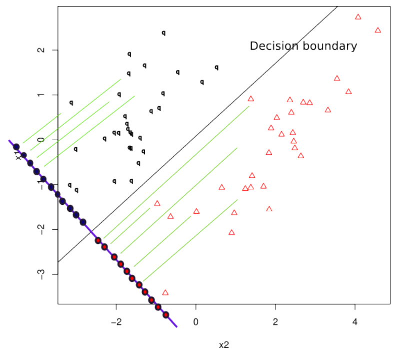

Figure 1.8: High coupling and Low aggregation

1.4.5 Logistic Regression

ただ通常の線形回帰モデルプラスシグモイド (sigmoid) 活性化関数 (activation function) です

$$P(y|x) = P_1^y P_0^{1-y} \tag{1.158}$$
$$\tag{1.159}$$

$$\begin{cases} P_1 = P(y=1|x) & = \sigma(w^t x) = \frac{1}{1+e^{-w^t x}}, y = 1 \\ P_0 = P(y=0|x) & = 1 - \sigma(w^t x) = \frac{e^{-w^t x}}{1+e^{-w^t x}}, y = 0 \end{cases}$$

Theorem 1.4.3. *MLE: - Cross entropy Loss*

$$\hat{w} = \underset{w}{argmax} \log P(y|x) \tag{1.160}$$
$$= \underset{w}{argmax} \log \prod_{i=1}^{N} P(y_i|x_i) \tag{1.161}$$
$$= \underset{w}{argmax} \sum_{i=1}^{N} \log P(y_i|x_i) \tag{1.162}$$
$$= \underset{w}{argmax} \sum_{i=1}^{N} \left(y_i \log \frac{1}{1+exp(-w^t x)} + (1-y_i) \log \frac{1}{1+exp(-w^t x)} \right) \tag{1.163}$$

1.4.6 Gaussian Discriminant Analysis

ただ確率的生成モデル (Generative model) の条件付き確率は、ガウス分布 (μ違う, Σ同じ) として計算されます。

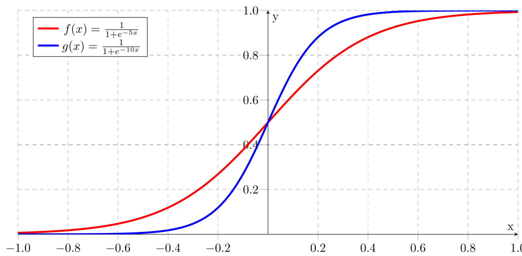

Figure 1.9: Sigmoid function

$$\hat{y} = \underset{y\in\{0,1\}}{argmax}\, P(y|x) = \underset{y}{argmax}\, P(y)\cdot P(x|y) \tag{1.164}$$

Log likelihood :

$$\underset{\mu_1,\mu_2,\Sigma,\phi}{argmax} \log\prod_{i=1}^{N} P(x_i,y_i) = \sum_{i=1}^{N} \log N(\mu_1,\Sigma)^{y_i} + \log N(\mu_2,\Sigma)^{1-y_i} + \log\phi^{y_i}(1-\phi)^{1-y_i} \tag{1.165}$$

$p(y)$ is distributed according to a Bernoulli distribution:

$$y \sim Bernoulli(\phi) \Longleftrightarrow \phi^{y}(1-\phi)^{1-y} \tag{1.166}$$

$p(x|y)$ is distributed according to a multivariate normal distribution:

$$x|y=1 \sim \mathcal{N}(\mu_1,\Sigma) \Longleftrightarrow p(x|y=1) = \frac{1}{(2\pi)^{\frac{n}{2}}\cdot|\Sigma|^{\frac{1}{2}}}\exp\Big(-\frac{1}{2}(x-\mu_1)^t\Sigma^{-1}(x-\mu_1)\Big) \tag{1.167}$$

$$x|y=0 \sim \mathcal{N}(\mu_2,\Sigma) \Longleftrightarrow p(x|y=0) = \frac{1}{(2\pi)^{\frac{n}{2}}\cdot|\Sigma|^{\frac{1}{2}}}\exp\Big(-\frac{1}{2}(x-\mu_2)^t\Sigma^{-1}(x-\mu_2)\Big) \tag{1.168}$$

Maximizing the log-likelihood:

$$\mathcal{L}(\phi,\mu_1,\mu_2,\Sigma) = \log\prod_{i=1}^{N} p(x_i,y_i;\phi,\mu_1,\mu_2,\Sigma) \tag{1.169}$$

$$= \log\prod_{i=1}^{N} p(x_i|y_i;\phi,\mu_1,\mu_2,\Sigma)p(y_i;\phi) \tag{1.170}$$

$$= \sum_{i=1}^{N}\big[\log\mathcal{N}(\mu_1,\Sigma)^{y_i} + \mathcal{N}(\mu_2,\Sigma)^{1-y_i}\log\phi^{y_i}(1-\phi^{1-y_i})\big] \tag{1.171}$$

Proof. ϕ

$$\frac{\partial \mathcal{L}(\phi, \mu_1, \mu_2, \Sigma)}{\phi} = \sum_{i=1}^{N} y_i \frac{1}{\phi} + (1 - y_i)\frac{1}{1-\phi}(-1) \tag{1.172}$$

$$= \sum_{i=1}^{N} y_i \frac{1}{\phi} - (1 - y_i)\frac{1}{1-\phi} \tag{1.173}$$

$$\sum_{i=1}^{N} y_i(1-\phi) - (1-y_i)\phi = 0 \tag{1.174}$$

$$\sum_{i=1}^{N} y_i - y_i\phi - \phi + y_i\phi = 0 \tag{1.175}$$

$$\sum_{i=1}^{N} (y_i - \phi) = 0 \tag{1.176}$$

$$\sum_{i=1}^{N} y_i - N\phi = 0 \tag{1.177}$$

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{N_1}{N} = \hat{\phi} \tag{1.178}$$

□

Proof. μ_1, μ_2

$$\frac{\partial \mathcal{L}(\phi, \mu_1, \mu_2, \Sigma)}{\mu_1} = \sum_{i=1}^{N} y_i \Big(-\frac{1}{2}(x_i - \mu_i)^t \Sigma^{-1}(x_i - \mu_i) \Big) \tag{1.179}$$

$$= -\frac{1}{2}\sum_{i=1}^{N} y_i (x_i^t \Sigma^{-1} - \mu_i^t \Sigma^{-1})(X_i - \mu_1) \tag{1.180}$$

$$= -\frac{1}{2}\sum_{i=1}^{N} y_i (x_i^t \Sigma^{-1} x_i - 2\mu_1^t \Sigma^{-1} x_i + \mu_1^t \Sigma^{-1}\mu_1) \tag{1.181}$$

$$= -\frac{1}{2}\sum_{i=1}^{N} y_i (-2\Sigma^{-1} x_i + 2\Sigma^{-1}\mu_1) \tag{1.182}$$

$$\tag{1.183}$$

$$\sum_{i=1}^{N} y_i(\Sigma^{-1}\mu_1 - \Sigma^{-1}x_i) = 0 \tag{1.184}$$

$$\sum_{i=1}^{N} y_i(\mu_1 - x_i) = 0 \tag{1.185}$$

$$\sum_{i=1}^{N} y_i\mu_1 = \sum_{i=1}^{N} y_i x_i \tag{1.186}$$

$$\frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} y_i x_i}{N_1} = \hat{\mu}_1 \tag{1.187}$$

□

Proof. Σ
Define: $S = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^t$

$$\sum_{i=1}^{N} \log \mathcal{N}(\mu, \Sigma) = \sum_{i=1}^{N} \log \frac{1}{2\pi}^{\frac{n}{2}} + \log|\Sigma|^{-\frac{1}{2}} - \frac{1}{2}(x_i - \mu)^t\Sigma^{-1}(x_i - \mu) \tag{1.188}$$

$$= \sum_{i=1}^{N} C - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i - \mu)^t\Sigma^{-1}(x_i - \mu) \tag{1.189}$$

$$= C - \frac{1}{2}N\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^t\Sigma^{-1}(x_i - \mu) \tag{1.190}$$

$$= -\frac{1}{2}N\log|\Sigma| - \frac{1}{2}N \cdot tr(S \cdot \Sigma^{-1}) + C \tag{1.191}$$

$$\sum_{i=1}^{N} y_i \log \mathcal{N}(\mu_1, \Sigma) + \sum_{i=1}^{N} y_i \log \mathcal{N}(\mu_1, \Sigma) = -\frac{1}{2}N\log|\Sigma| - \frac{1}{2}N_1 tr(S_1\Sigma^{-1} - \frac{1}{2}N_2 tr(S_2\Sigma^{-1}) + C \tag{1.192}$$

$$= -\frac{1}{2}\big(N\log|\Sigma| + N_1 tr(S_1\Sigma^{-1} + N_2 tr(S_2\Sigma^{-1})\big) + C \tag{1.193}$$

$$\frac{\partial\mathcal{L}(\phi, \mu_1, \mu_2, \Sigma)}{\Sigma} = -\frac{1}{2}(N\Sigma^{-1} - N_1S_1\Sigma^{-2} - N_2S_2\Sigma^{-2}) \tag{1.194}$$

$$-\frac{1}{2}(N\Sigma^{-1} - N_1S_1\Sigma^{-2} - N_2S_2\Sigma^{-2}) = 0 \tag{1.195}$$

$$N\Sigma - N_1S_1 - N_2S_2 = 0 \tag{1.196}$$

$$\frac{1}{N}(N_1S_1 + N_2S_2) = \hat{\Sigma} \tag{1.197}$$

□

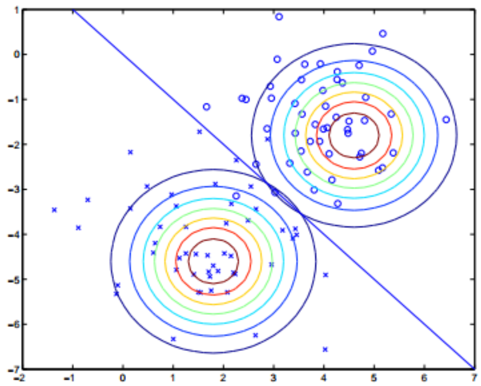

Figure 1.10: Sigmoid function

Reference: [5]

$$\begin{cases} \text{先験的確率}\ (Prioriprobability): \begin{cases} Two-categories:\ y \sim Bernoulli \longrightarrow Binomial \\ Multi-category:\ y \sim Categorial \longrightarrow Multinomial \end{cases} \\ \text{尤度関数}\ (Likelihoodfunction): \begin{cases} x_{Discrete\ variable}:\ x_i \sim Categorial \\ x_{Continuous\ variable}:\ x_i \sim Gaussian \end{cases} \\ \text{事後確率}\ (Posteriorprobability) \end{cases}$$

1.4.7 SVMs

間隔を最大化する分類器。ジオメトリ (geometric) 意義は凸最適化問題 (Convex optimization) に変換されます。元の問題はラグランジュ乗数法 (Lagrange multiplier) によって制約のない問題に変換されます。単純なデータの場合は二次計画法 (Quadratic programming) の問題であり、複雑な問題の場合は二元性 (Duality), カーネル (Kernel) の考え方を使用します。

ハードマージンSVM(hard margin svm) :

$$s.t.\ \left\{(x_i, y_i)\right\}_{i=1}^{N}, x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$$

$$\begin{cases} \underset{w,b}{max}\ \underset{x_i:x_n}{min}\ \frac{1}{\|w\|} y_i(w^t x_i + b) \\ Define:\ y_i(w^t x_i + b) > 0,\ for\ \forall\ i = 1, \ldots, N \end{cases} \Longrightarrow \begin{cases} \underset{w,b}{min}\ \frac{1}{2} w^t w \\ s.t.\ y_i(w^t x_i + b) \geq 1,\ for\ \forall\ i = 1, \ldots, N \end{cases} \tag{1.198}$$

[5] https://tariq-hasan.github.io/concepts/machine-learning-gaussian-discriminant-analysis/

ソフトマージンSVM(soft margin svm) :

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}w^t w + C\sum_{i=1}^{N} max\{0, 1 - y_i(w^t x_i + b)\} \\ s.t.\ y_i(w^t x_i + b) \geq 1 - [1 - y_i(w^t x_i + b)];\ 1 - y_i(w^t x_i + b) \geq 0 \end{cases} \tag{1.199}$$

二元性 (Duality):

弱い双対性 (Weak duality) は $min\,max \geq max\,min$ です；強い双対性 (Strong duality) は $min\,max \geq max\,min$ です。元の問題の目的関数は 2 次であり、その制約は線形であるため、強い双対性です。

$$\begin{cases} \min\limits_{w,b} \max\limits_{\lambda} \frac{1}{2}w^t w + \sum_{i=1}^{N} \lambda_i[1 - y_i(w^t x_i + b)] \\ s.t.\ \lambda_i \geq 0 \end{cases} \implies \begin{cases} \max\limits_{\lambda} \min\limits_{w,b} \frac{1}{2}w^t w + \sum_{i=1}^{N} \lambda_i[1 - y_i(w^t x_i + b)] \\ s.t.\ \lambda_i \geq 0 \end{cases} \tag{1.200}$$

1.4.8 Kernel Method

低次元の非線形分離可能データを高次元の線形分離可能データに変換する場合、低次元ベクトルを計算して高次元ベクトルにマッピングし、内積（SVM）を見つけることは非常に複雑です。 この点で、高次元ベクトルの内積はカーネル法によって直接取得できます。

Theorem 1.4.4. *positive definite kernel:*
$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$;
$\forall x, x' \in \mathcal{X}, \exists K(x, x')$;
If: $\exists \phi : \mathcal{X}, \phi \in \mathcal{H}$;
Then: $Kernel(x, x') =< \phi(x), \phi(x') >$

Proof. $Kernel(x, x') =< \phi(x), \phi(x') > \iff$ Gram matrix(Positive semi-definite matrix) $\square$

1.4.9 Generative Model

Model the sample data itself. $https://en.wikipedia.org/wiki/Generative\_model$

$$\begin{cases} Naive\ Bayes \\ Mixture\ Model:\ GMM \\ Time-series\ Model:\ HMM, Kalman\ Filter, Particle\ Fitter \\ Non-parameter\ Bayesian\ Model:\ GP, DP \\ Mixed\ Memership:\ LDA \\ Factorial\ Model:\ FA, P-PCA, ICA \\ Energy-based\ Model:\ Boltzmann\ Machine \\ VAE \\ GAN \\ Autoregressive\ Model \\ Flow-based\ Model \end{cases}$$

1.5 Dimensionality Reduction

1.5.1 PCA

PCA(Principal components analysis) は元の特徴空間の再構築 (2 つの線形関連変数を 2 つの線形独立変数に変換します)。射影分散 (Covariance) を最大化する，再構成距離を最小化する (元のデータを再構築するためのコスト)，$u_{mapping\ vector}$ は PCA の PC(Principal components) です。

射影分散 (Covariance) 最大化:

$s.t.\ u^t_{mapping\ vector} \cdot u_{mapping\ vector} = 1$

$$argmax \frac{1}{N} \sum_{i=1}^{N} ((x_i - \hat{x})^t u_{mapvect})^2 = u^t_{mapvec} \cdot S \cdot u_{mapvec} \tag{1.201}$$

$$\tag{1.202}$$

Optimizationfunction: ラグランジュ乗数 (Lagrange Multiplier)

$$\mathfrak{L}(u, \lambda) = u^t_{mapvec} \cdot S \cdot u_{mapvec} + \lambda(u^t_{mapvec} \cdot u_{mapvec} - 1) \tag{1.203}$$

再構成距離を最小化:

$s.t.\ u_k^t \cdot u_k = 1$

$$argmax \sum_{i=1}^{N} \left\| \sum_{k=1}^{p} (x_i^t u_k) u_k - \sum_{k=1}^{q} (x_i^t u_k) u_k \right\| = \sum_{k=q+1}^{p} u_k^t \cdot S \cdot u_k \tag{1.204}$$

$$= \sum_{k=q+1}^{p} \lambda_i \tag{1.205}$$

参照:

$Data: \ X \in \mathbb{R}^{n \times p}$

$$Sample\ Mean: \ \hat{x}_{p\times 1} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N} X^t I_N \tag{1.206}$$

$$Sample\ Covariance: \ S_{p\times p} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x})(x_i - \hat{x})^t = \frac{1}{N} X^t H_{centering\ matrix} X \tag{1.207}$$

1.5.2 PCA vs SVD

実はデータセンタリング (Data centralization) 後の SVD 分解と共分散行列 (Covariance matrix) で固有値分解の意味は同じです。だからデータセンタリング後の SVD 分解の方法が共分散行列 (Covariance matrix) の計算しなくでもいい、早いです。

$$S = \frac{1}{N} X^t H X = X^t H^t H X = V \Sigma U^t \cdot U \Sigma V^t = V \Sigma^2 V^t \tag{1.208}$$

1.5.3 P-PCA

P-PCA(Probabilitic PCA) は線形ガウスモデル (Linear Gaussian model) です、ターゲットを隠れ変数に変換し、最尤法 (MLE) でモデルを作成し、EM 使用して取得されます。隠れた変数は特定の分布に従います。

$s.t.\ x(observed\ data) \in \mathbb{R}^p, \ z(latent\ variable) \in \mathbb{R}^q, \ q < p$

$$\begin{cases} z \sim N(0_q, I_q) \\ x = wz + \mu + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 I_p) \end{cases} \implies \begin{cases} Inference: \ P(z|x) \\ Learning(EM): \ w, \mu, \sigma^2 \end{cases}$$

1.5.4 EM For GMM

EM:

$$\theta^{(t+1)} = \underset{\theta}{argmax} \int_z \log P(x,z|\theta) \cdot P(z|x,\theta^{(t)}) dz \tag{1.209}$$

E-step: $p(z|x,\theta^t) \to E_{z|x,\theta^t}[\log p(x,z|\theta)]$;
M-step: $\theta^{t+1} = \underset{\theta}{argmax}\, E_{z|x,\theta^t}[\log p(x,z|\theta)]$

EM For GMM: Define:
X: observed variable;
Z: latent variable(clusterer)
$X|Z = C_k \sim N(X|\mu_k \Sigma_k)$
$\theta = (p_1, ..., p_k, \mu_1, ..., \mu_k, \Sigma_1, ..., \Sigma_k)$
$p(x) = \sum_{k=1}^{K} p_k \cdot N(X|\mu_k \Sigma_k)$
$p(x,z) = p(z) \cdot p(x|z) = p_(z) \cdot N(X|\mu_k \Sigma_k)$
$p(z|x) = \frac{p(x,z)}{p(x)} = \frac{p_(z) \cdot N(X|\mu_k \Sigma_k)}{\sum_{k=1}^{K} p_k \cdot N(X|\mu_k \Sigma_k)}$

Proof. E-step:

$$
\begin{aligned}
Q(\theta,\theta^t) &= \int_z \log p(x,z|\theta) \cdot p(z|x,\theta^t) dz && (1.210)\\
&= \sum_z \log \prod_{i=1}^{N} p(x_i,z_i|\theta) \cdot \prod_{i=1}^{N} p(z_i|x_i,\theta^t) && (1.211)\\
&= \sum_{z_1,...,z_n} \sum_{i=1}^{N} \log p(x_i,z_i|\theta) \cdot \prod_{i=1}^{N} p(z_i|x_i,\theta^t) && (1.212)\\
&= \sum_{z_1,...,z_n} \big[\log p(x_1,z_1|\theta) + \cdots + \log p(x_n,z_n|\theta) \big] \cdot \prod_{i=1}^{N} p(z_i|x_i,\theta^t) && (1.213)
\end{aligned}
$$

$$\because \sum_{z_1,\ldots,z_n} \log p(x_1, z_1|\theta) \prod_{i=1}^{N} p(z_i|x_i, \theta^t) \tag{1.214}$$

$$= \sum_{z_1,\ldots,z_n} \log p(x_1, z_1|\theta) \cdot p(z_1|x_1, \theta^t) \cdot \prod_{i=2}^{N} p(z_i|x_i, \theta^t) \tag{1.215}$$

$$= \sum_{z_1} \log p(x_1, z_1|\theta) \cdot p(z_1|x_1, \theta^t) \cdot \sum_{z_2,\ldots,z_n} \prod_{i=2}^{N} p(z_i|x_i, \theta^t) \tag{1.216}$$

$$= \sum_{z_1} \log p(x_1, z_1|\theta) \cdot p(z_1|x_1, \theta^t) \cdot \sum_{z_2} \prod_{i=2}^{N} p(z_i|x_i, \theta^t) \tag{1.217}$$

$$= \sum_{z_1} \log p(x_1, z_1|\theta) \cdot p(z_1|x_1, \theta^t) \tag{1.218}$$

$$\therefore Q(\theta, \theta^t) = \sum_{i=1} \sum_{z_i} \log p(x_i, z_i|\theta) \cdot p(z_i|x_i, \theta^t) \tag{1.219}$$

$$= \sum_{i=1}^{N} \sum_{z_i}^{N} \log p_{z_i} \cdot N(x_i|\mu_{z_i}, \Sigma_{z_i}) \cdot \frac{p_{z_i} \cdot N(x_i|\mu_{z_i}^t, \Sigma_{z_i}^t)}{\sum_{k=i}^{K} p_k^t \cdot N(x_i|\mu_{z_i}^t, \Sigma_{z_i}^t)} \tag{1.220}$$

$$= \sum_{i=1}^{N} \sum_{z_i} \log \left[p_{z_i} \cdot N(x_i|\mu_{z_i}, \Sigma_{z_i})\right] \cdot p(z_i|x_i, \theta^t) \tag{1.221}$$

$$= \sum_{z_i} \sum_{i=1}^{N} \log \left[p_{z_i} \cdot N(x_i|\mu_{z_i}, \Sigma_{z_i})\right] \cdot p(z_i|x_i, \theta^t) \tag{1.222}$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{N} \log \left[p_k \cdot N(x_i|\mu_k, \Sigma_k)\right] \cdot p(z_i|x_i, \theta^t) \tag{1.223}$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{N} \left[\log p_k + \log N(x_i|\mu_k, \Sigma_k)\right] \cdot p(z_i|x_i, \theta^t) \tag{1.224}$$

$$\tag{1.225}$$

□

Proof. M-step: p^{t+1}

$\theta^{t+1} = \underset{\theta}{argmax}\, Q(\theta, \theta^t)$

$$\mathcal{L}(p, \lambda) = \sum_{k=1}^{K}\sum_{i=1}^{N} \log p_k \cdot p(z_i = c_k|x_i, \theta^t) + \lambda(\sum_{k}^{K} p_k - 1) \tag{1.226}$$

$$= \cdots \tag{1.227}$$

□

1.5.5 Spectral Clustering

Compactness: K-means,GMM
Connectivity: Spectral clustering Model Introduction: Based on weighted undirected graph

Reference: [6]

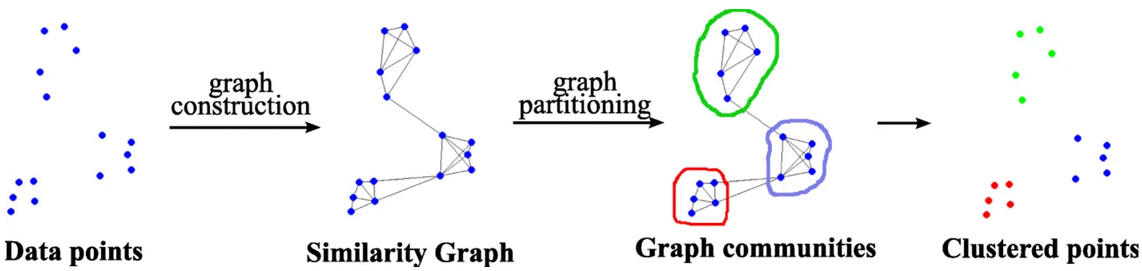

Figure 1.11: K-means vs Spectral clustering

Define: $G = \{V(Vertexset), E(Edgeset)\}$;
$V = \{1, 2, ..., N\} = \mathcal{X}$;
$W(SimilaritymatrixorAffinitymatrix) = [w_{ij}], 1 \leq i, j \leq\ N$;
If: $w_{ij} \in E$; then: $w_{ij} = K(x_i, (x_j)) = \exp\left\{ -\frac{||x_i - x_j||_2^2}{2\sigma^2}\right\}$ else: $w_{ij} = 0$
$cut(v) = cut(C_1, C_2, ..., C_K) = \sum_{k=1}^{K} w(C_k, \bar{C}_k) = \sum_{k=1}^{K} w(C_k, V) - w(C_k, C_k)$;
$V = \cup_{k=1}^{K} C_k; C_i \cap C_j = \phi$
Degree: Normalized cut(V): $d_i = \sum_{j=1}^{N} w_{ij}$
Indicator vector:$Y = (y_1, y_2, ..., y_N^t)_{N\times K}$

$$Y^tY = (y_1, ..., y_N)\begin{pmatrix} y_1^t \\ \vdots \\ y_N^t \end{pmatrix} = \sum_{i=1}^{N} y_i y_i^t = \begin{pmatrix} N_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & N_k \end{pmatrix} = \begin{pmatrix} \sum_{i\in C_1} \cdot 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{i\in C_k} \cdot 1 \end{pmatrix}$$

$$\mathrm{D} = \begin{pmatrix} d_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_N \end{pmatrix} = diag(w \cdot 1_N)$$

[6] https://link.springer.com/article/10.1007/s41109-019-0248-7?shared-article-renderer

$$Y^tWY = (y_1, ..., y_N) \begin{pmatrix} w_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_{NN} \end{pmatrix} \begin{pmatrix} y_1^t \\ \vdots \\ y_N^t \end{pmatrix} = (\sum_{i=1}^N y_i w_{i1}, ..., \sum_{i=1}^N y_i w_{iN}) \begin{pmatrix} y_1^t \\ \vdots \\ y_N^t \end{pmatrix} =$$

$$\sum_{i=1}^N \sum_{j=1}^N y_i y_j w_{ij} = \begin{pmatrix} \sum_{i\in C_1}\sum_{j\in C_1} w_{ij} & \dots & \sum_{i\in C_1}\sum_{j\in C_k} w_{ij} \\ \vdots & \ddots & \vdots \\ \sum_{i\in C_k}\sum_{j\in C_1} w_{ij} & \dots & \sum_{i\in C_k}\sum_{j\in C_k} w_{ij} \end{pmatrix}$$

$$\hat{Y} = \underset{Y}{argmin} \sum_{k=1}^K \frac{w(C_k, V) - w(C_k, C_k)}{\sum_{i\in C_k} d_i} \tag{1.228}$$

$$= tr \begin{pmatrix} \frac{w(C_k,V)-w(C_k,C_k)}{\sum_{i\in C_1} d_i} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{w(C_k,V)-w(C_k,C_k)}{\sum_{i\in C_k} d_i} \end{pmatrix} \cdot \begin{pmatrix} \sum_{i\in C_1} d_i & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{i\in C_k} d_i \end{pmatrix}^{-1} \tag{1.229}$$

$$= tr \begin{pmatrix} w(C_k, V) - w(C_k, C_k) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w(C_k, V) - w(C_k, C_k) \end{pmatrix} \cdot (Y^tDY)^{-1} \tag{1.230}$$

$$= tr \left[\begin{pmatrix} \sum_{i\in C_1} \cdot d_i & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{i\in C_k} \cdot d_i \end{pmatrix} - \begin{pmatrix} w(C_1, C_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w(C_k, C_k) \end{pmatrix} \right] \cdot (Y^tDY)^{-1} \tag{1.231}$$

$$= tr \left[\begin{pmatrix} \sum_{i\in C_1} \cdot d_i & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{i\in C_k} \cdot d_i \end{pmatrix} - \begin{pmatrix} \sum_{i\in C_1}\sum_{j\in C_1} w_{ij} & \dots & \sum_{i\in C_1}\sum_{j\in C_k} w_{ij} \\ \vdots & \ddots & \vdots \\ \sum_{i\in C_k}\sum_{j\in C_1} w_{ij} & \dots & \sum_{i\in C_k}\sum_{j\in C_k} w_{ij} \end{pmatrix} \right] \cdot (Y^tDY)^{-1} \tag{1.232}$$

$$= \underset{Y}{argmin}\, tr\big(Y^t(D-W)Y\dot{(}Y^tDY)^{-1}\big) \tag{1.233}$$

$D - W$: Laplacian matrix

Chapter 2

Bayesian Inference

$$
\begin{cases}
Representation: \begin{cases}
\underset{Bayesian\ network}{Directed\ graph}: \begin{cases}
Single: Naive\ bayes: P(x|y)=\prod_{i=i}^{dim} P(x_i|y=1)\\
Mix: GMM\\
Time: \begin{cases} Markov\ chain\\ Gaussian\ process(Infinite\ dimensionsGD) \end{cases}\\
Dynamic\ model: \begin{cases} HMM: (Discrete)\\ LDS(Gaussian, linear, kalman\ filter)\\ Particle\ filters(nonGaussian, nonLinear) \end{cases}
\end{cases}\\
Undirected\ graph: Markov\ network\\
Gaussian\ graph(Continuous\ variable): \begin{cases} Gaussian\ bayesian\ network\\ Gaussian\ markov\ network \end{cases}
\end{cases}\\
Inferense: \begin{cases}
Accurate: \begin{cases} Variable\ elimination\\ Belief\ propagation(Sum-product\ algorithm)(Tree\ structure)\\ Junction\ Tree\ algorithm(Normal\ graph) \end{cases}\\
Approximate: \begin{cases} Variation\ method(determine)\\ Loop\ belief\ propagation(Ring\ graph)\\ Monte\ Carlo\ \ : Importance\ sampling, MCMC(stochastic) \end{cases}
\end{cases}\\
Learning: \begin{cases}
Structure\ learning\\
Parameter\ learning: \begin{cases} Complete\ data\\ Hidden\ variable: EM \end{cases}
\end{cases}
\end{cases}
$$

2.1 Representation

2.1.1 Introduction

$$\begin{cases} \underset{Bayesian\ network}{Directed\ graph}: \begin{cases} Single: Naive\ bayes: P(x|y) = \prod_{i=i}^{dim} P(x_i|y=1) \\ Mix: GMM \\ Time: \begin{cases} Markov\ chain \\ Gaussian\ process(Infinite\ dimensions GD) \end{cases} \\ Dynamic\ model: \begin{cases} HMM: (Discrete) \\ LDS(Gaussian, linear, kalman\ filter) \\ Particle\ filters(nonGaussian, nonLinear) \end{cases} \end{cases} \\ Undirected\ graph: Markov\ network \\ Gaussian\ graph(Continuous\ variable): \begin{cases} Gaussian\ bayesian\ network \\ Gaussian\ markov\ network \end{cases} \end{cases}$$

2.1.2 Moral Graph

- Directed graph: $p(x) = \prod_x p(x_i|x_{parents})$
- Undirected graph: $p(x) = \frac{1}{z}\prod_{i=1}^{k} \phi_{ci}(x_{ci} = Largest\ group\ set)$

$Moral\ graph: graph(Directed\ tree) \to Undirected\ graph(Undirected\ ring)$

$$P_{directed}(x) = \prod_x P(x_i|x_{parents}) \tag{2.1}$$

$$P_{undirected}(x) = \frac{1}{z}\prod_{i=1} k\phi_{clique_i}(x_{clique_i}) \tag{2.2}$$

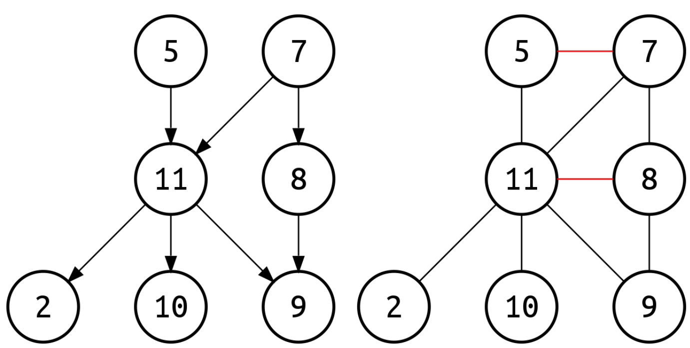

Figure 2.1: Directed graph to Undirected

2.1.3 Factor Graph

リング構造からツリー構造への変換。$Moral\ graph \rightarrow factor\ graph(Undirected\ tree)$

head2head(V structure):$parents(x_i)$ を接続して。これは、因数分解のさらなる分解と見なすことができます。

$$P(x) = \prod_{s \in graph\ node} f_s(x_s) \tag{2.3}$$

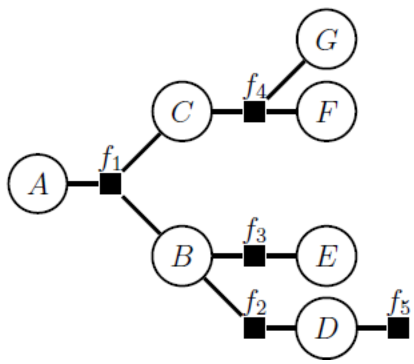

Figure 2.2: Factor graph

$$P_{figure3.1}(x) = f_1(A,B,C) \cdot f_2(B,D) \cdot f_3(B,E) \cdot f_4(C,F,G) \cdot f_5(D) \tag{2.4}$$

2.2 Inference

2.2.1 Introduction

Frequency: optimization problem
Bayesian: integral problem
The task is to find the probability $p(x) = p(x_1, x_2, ..., x_n)$
z: latent variable + parament

- Marginal probability: $p(x_i) = \sum_{x_{i+1}} \sum_{x_{i+2}} \cdots \sum_{x_n}$
- Conditional Probability: $p(\hat{x}|x) = \int_\theta p(\hat{x}, \theta|x) d\theta = \int_\theta p(\hat{x}|\theta) \cdot p(\theta|x) d\theta = E_{\theta|x}[p(\hat{x}|\theta)]$
- MAP Inference: $\hat{z} = \underset{z}{argmax}\, p(z|x) \propto p(z,x)$

$$\begin{cases} Accurate: \begin{cases} Variable\ elimination \\ Belief\ propagation(Sum-product\ algorithm)(Tree\ structure) \\ Junction\ Tree\ algorithm(Normal\ graph) \end{cases} \\ Approximate: \begin{cases} Variation\ method(Deterministicinference) \\ Loop\ belief\ propagation(Ring\ graph) \\ Monte\ Carlo\ \ : Importance\ sampling, MCMC(stochastic) \end{cases} \end{cases}$$

2.2.2 Variable Elimination

Multiplicative Distribution Law：The disadvantage

- Repeated calculation(no stored procedure)
- Ordering is NP-hard

Define: $a, b, c, d \in \{0, 1\}; a \to b \to c \to d$

$$
\begin{aligned}
p(d) &= \sum_{a,b,c} p(a,b,c,d) && (2.5)\\
&= \sum_{a,b,c} p(a) \cdot p(b|a) \cdot p(c|b) p(d|c) && (2.6)\\
&= \sum_{a,b,c} p(a=0) \cdot p(b=0|a=0) \cdot p(c=0|b=0) p(d|c=0) && (2.7)\\
&= \sum_{a,b,c} p(a=1) \cdot p(b=0|a=1) \cdot p(c=0|b=0) p(d|c=0) && (2.8)\\
&= \sum_{a,b,c} p(a=0) \cdot p(b=1|a=0) \cdot p(c=0|b=1) p(d|c=0) && (2.9)\\
&= \sum_{a,b,c} p(a=0) \cdot p(b=0|a=0) \cdot p(c=1|b=0) p(d|c=1) && (2.10)\\
&= \sum_{a,b,c} p(a=1) \cdot p(b=1|a=1) \cdot p(c=0|b=1) p(d|c=0) && (2.11)\\
&= \sum_{a,b,c} p(a=1) \cdot p(b=0|a=1) \cdot p(c=1|b=0) p(d|c=1) && (2.12)\\
&= \sum_{a,b,c} p(a=0) \cdot p(b=1|a=0) \cdot p(c=1|b=1) p(d|c=1) && (2.13)\\
&= \sum_{a,b,c} p(a=1) \cdot p(b=1|a=1) \cdot p(c=1|b=1) p(d|c=1) && (2.14)\\
&= \sum_{b,c} p(c|b) \cdot p(d|c) \sum_{a} p(a) \cdot p(b|a) && (2.15)\\
&= \sum_{b} p(c|b) \cdot \sum_{c} p(d|c) \cdot m_{ab}(b) && (2.16)\\
&= \sum_{c} p(d|c) \cdot m_{bc}(c) && (2.17)\\
&= m_{cd}(d) && (2.18)
\end{aligned}
$$

2.2.3 Belief Propagation(Sum-product)

実は VE ＋ Caching です、木の構造に適しています。Repeated calculation 必要はなく、必要な情報量 $m_{i \to j}$ だけでいいです。

$$m_{j\to i}(x_i) = \sum_{x_j} \varphi_j(x_j) \cdot \varphi_{ij}(x_i, x_j) \prod_{k \in Neighbor(j)-i} m_{k\to j}(x_j) \tag{2.19}$$

$$P(x_i) = \varphi_i(x_i) \cdot \prod_{k \in Neighbor(i)} m_{k\to i}(x_i) \tag{2.20}$$

Algorithm 1 Sequential Implementationg

Require: Get root ,Assume a is root
  for x_i *in Neighbor(Root)* **do**
    Collect $m_{ij}(x_i)$
  end for
  for x_j *in Neighbor(Root)* **do**
    Distribute $m_{ij}(x_j)$
  end for

Algorithm 2 Parellel Implementation

  for x_i *in All_nodes* **do**
    $x_i = Collect\ Neighbor(x_i) \cdot x_i$
    Distribute Neighbor(x_i)
  end for

Reference: [1]

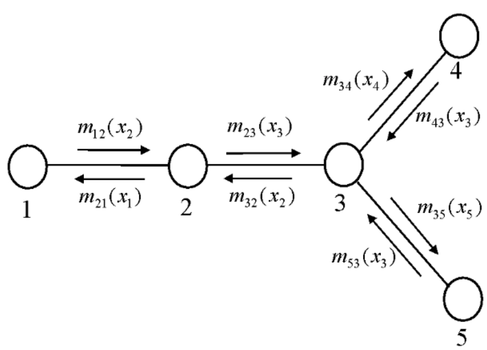

Figure 2.3: Belief propagation

[1] Understanding Belief Propagation and its Generalizations Jonathan S. Yedidia MERL 201 Broadway Cambridge, MA 02139

2.2.4 Max-product

- Belief propagation の改善

- Viterbi の拡張。

$$m_{j\to i} = \max_{x_j} \varphi_j \cdot \varphi_{ij} \prod_{k\in Neighbor(j)-i} m_{k\to j} \tag{2.21}$$

2.3 Variational Inference

2.3.1 VI based Mean field

平均場理論 (Mean field theory) に基づいて、複素確率構造は多くの小さな構造に分割されます。 座標降下法 (Coordinate descent method) を使用して、最大事後推定 (Posterior probability) を解きます。
s.t.
X: observed data
Z: latent variable + parament
(X,Z): complete data

$$\log p(X) = \log p(X,Z) - \log p(Z|X) \tag{2.22}$$

$$= \log \frac{p(X,Z)}{q(Z)} - \log \frac{p(Z|X)}{q(Z)} \tag{2.23}$$

$$\int_Z \log p(X) q(Z) dz = \int_Z q(Z) \cdot \log \frac{p(X,Z)}{q(Z)} dz - \int_Z q(Z) \cdot \log \frac{p(Z|X)}{q(Z)} dz \tag{2.24}$$

$$\log p(X) = ELBO + KL(q||p) \tag{2.25}$$

Define: ELBO $= \mathcal{L}(q)$
$\hat{q}(Z) = \underset{q(Z)}{argmax}\, \mathcal{L}(q) \Rightarrow \hat{q}(Z) \approx p(Z|X)$

$$\mathcal{L}(q) = \int_Z q(Z) \cdot \log p(X,Z) dz - \int_Z q(Z) \cdot \log q(Z) dz \tag{2.26}$$

s.t. Mean field theory: $q(Z) = \prod_{i=1}^{M} q_i(Z_i)$

$$\int_Z q(Z) \cdot \log p(X, Z) dz = \int_Z \prod_{i=1}^{M} q_i(Z_i) \cdot \log p(X, Z) dz_1, ..., dz_M \tag{2.27}$$

$$= \int_{Z_j} q_j(Z_j) \Big(\int_{Z_{\neq j}} \prod_{i \neq j}^{M} q_i(Z_i) \cdot \log p(X, Z) dz_i \Big) dz_j \tag{2.28}$$

$$= \int_{Z_j} q_j(Z_j) \Big(\int_{Z_{\neq j}} \log p(X, Z) \cdot \prod_{i \neq j}^{M} q_i(Z_i) dz_i \Big) dz_j \tag{2.29}$$

$$= \int_{Z_j} q_j(Z_j) \cdot E_{\prod_{i \neq j}^{M} q_i(Z_i)} \big[\log p(X, Z) \big] dz_j \tag{2.30}$$

$$= \int_{Z_j} q_j(Z_j) \cdot \big[\log \hat{p}(X, Z_j) \big] dz_j \tag{2.31}$$

$$\int_Z q(Z) \cdot \log q(Z) dz = \int_Z \prod_{i}^{M} q_i(Z_i) \cdot \sum_{i}^{M} \log q_i(Z_i) dz \tag{2.32}$$

$$= \int_Z \prod_{i}^{M} q_i(Z_i) \cdot \big[\log q_1(Z_1) + \cdots + \log q_M(Z_M) \big] dz \tag{2.33}$$

$$\because \int_Z \prod_{i}^{M} q_i \cdot \log q_1 dz = \int_{Z_{1,\ldots,M}} q_1 \cdots q_M \cdot \log q_1 dz_{1,\ldots,M} \tag{2.34}$$

$$= \int_{Z_1} q_1 \log q_1 dz_1 \cdot \int_{Z_2} q_2 dz_2 \cdots \int_{Z_M} q_M dz_M \tag{2.35}$$

$$= \int_{Z_1} q_1 \log q_1 dz_1 \tag{2.36}$$

$$\therefore \int_Z q(Z) \cdot \log q(Z) dz = \sum_{i=1}^{M} \int_{Z_i} q_i(Z_i) \cdot \log q_i(Z_i) dz_i \tag{2.37}$$

$$= \sum_{i=1}^{M} \int_{Z_j} q_j(Z_j) \cdot \log q_j(Z_j) dz_j + C \tag{2.38}$$

$$\mathcal{L}(q) = \int_Z q(Z) \cdot \log p(X,Z) dz - \int_Z q(Z) \cdot \log q(Z) dz \tag{2.39}$$
$$= \int_{Z_j} q_j(Z_j) \cdot \log \frac{\hat{p}(X,Z_j)}{q_j(Z_j)} dz_j \tag{2.40}$$
$$= -KL(q_j|\hat{p}(X,Z_j)) \tag{2.41}$$

Object function:

$$\hat{q} = \underset{q}{argmin}\, KL(q||p) = \underset{q}{argmin}\, \mathcal{L}(q) \tag{2.42}$$

Coordinate Ascend:

$$\hat{q_1}(Z_1) = \int_{q_2} \cdots \int_{q_M} q_2 \cdots q_M \big[\log p_\theta(x_i,Z)\big] dq_2 \cdots dq_M \tag{2.43}$$
$$\hat{q_2}(Z_2) = \int_{\hat{q}_1} \int_{q_3} \cdots \int_{q_M} \hat{q}_1 q_3 \cdots q_M \big[\log p_\theta(x_i,Z)\big] d\hat{q}_1 q_3 \cdots dq_M \tag{2.44}$$
$$\vdots \tag{2.45}$$
$$\hat{q_M}(Z_M) = \int_{\hat{q}_1} \cdots \int_{\hat{q}_{M-1}} \hat{q}_1 \cdots \hat{q}_{M-1} \big[\log p_\theta(x_i,Z)\big] d\hat{q}_1 \cdots d\hat{q}_{M-1} \tag{2.46}$$
$$\tag{2.47}$$

2.3.2 SGVI (SGVB)

$$ELBO = E_{q_\phi(Z)}\left[\log \frac{p_\theta(x_i,Z)}{q_\phi(Z)}\right] \tag{2.48}$$
$$= E_{q_\phi(Z)}\big[\log p_\theta(x_i,Z) - \log q_\phi(Z)\big] \tag{2.49}$$
$$= \mathcal{L}(q) \tag{2.50}$$

$$\hat{\phi} = \underset{\phi}{argmin}\, \mathcal{L}(q) \tag{2.51}$$
$$\nabla_\phi \mathcal{L}(q) = \nabla_\phi E_{q_\phi(Z)}\big[\log p_\theta(x_i,Z) - \log q_\phi\big] \tag{2.52}$$
$$= \nabla_\phi \int_Z q_\phi \cdot \big[\log p_\theta(x_i,Z) - \log q_\phi\big] dz \tag{2.53}$$
$$= \int_Z \nabla_\phi q_\phi \cdot \big(\log p_\theta(x_i,Z) - \log q_\phi\big) dz + \int_Z q_\phi \nabla_\phi \big[\log p_\theta(x_i,Z) - \log q_\phi\big] dz \tag{2.54}$$

$$
\begin{aligned}
\because \int_Z q_\phi \nabla_\phi \big[\log p_\theta(x_i, Z) - \log q_\phi\big] dz &= -\int_Z q_\phi \nabla_\phi \log q_\theta dz && (2.55)\\
&= -\int_Z q_\phi \cdot \frac{1}{q_\theta} \cdot \nabla_\phi q_\theta dz && (2.56)\\
&= -\int_Z \nabla_\phi q_\theta dz && (2.57)\\
&= -\nabla_\phi \int_Z q_\theta dz && (2.58)\\
&= -\nabla_\phi && (2.59)\\
&= 0 && (2.60)
\end{aligned}
$$

$$
\begin{aligned}
\therefore \hat{\phi} &= \int_Z \nabla_\phi q_\phi \cdot \big(\log p_\theta(x_i, Z) - \log q_\phi\big) dz + 0 && (2.61)\\
&= \int_Z q_\phi \cdot \nabla_\phi \log q_\phi \cdot \big(\log p_\theta(x_i, Z) - \log q_\phi\big) dz && (2.62)\\
&= E_{q_\phi}\Big[\nabla_\phi \log q_\phi \cdot \big(\log p_\theta(x_i, Z) - \log q_\phi\big)\Big] && (2.63)\\
& s.t. Z^L \sim q_\phi(Z), l = 1, 2, ..., L && (2.64)\\
&\approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\phi \log q_\phi(Z^l)(\log p_\phi(x^i, z^i) - \log q_\phi(Z^l)) && (2.65)
\end{aligned}
$$

Reparametrization Trick s.t.
$Z \sim q_\phi(Z|x_i) = \epsilon \sim p(\epsilon)$
$|q_\phi(Z|x_i) \cdot dz| = |p(\epsilon) \cdot d\epsilon|$

$$
\begin{aligned}
\hat{\phi} &= E_{q_\phi}\Big[\nabla_\phi \log q_\phi \cdot \big(\log p_\theta(x_i, Z) - \log q_\phi\big)\Big] && (2.66)\\
&= E_{p(\epsilon)}\Big[\nabla_\phi \log q_\phi \cdot \big(\log p_\theta(x_i, Z) - \log q_\phi(Z|x_i)\big)\Big] && (2.67)\\
& \epsilon \sim p(\epsilon) && (2.68)\\
&= E_{p(\epsilon)}\Big[\nabla_Z \log q_\phi \cdot \big(\log p_\theta(x_i, Z) - \log q_\phi(Z|x_i)\big)\Big] \cdot \nabla_\phi g_\phi(\epsilon^l, x^i) && (2.69)
\end{aligned}
$$

SGVI:

$$
\phi^{t+1} \leftarrow \phi^t + \lambda \cdot \nabla_\phi \mathcal{L}(\phi) \quad (2.70)
$$

2.4 Sampling

2.4.1 Probability distribution sampling

確率 PDF を CDF に変換して、$y^{(i)} \sim U(0,1)$ 一様分布 (uniform distribution) に関連付けます。(PDF は複雑ので PDF から CDF まで難しい)

$$x^{(i)} = cdf^{-1}(y^{(i)}) \tag{2.71}$$

Reference: [2]

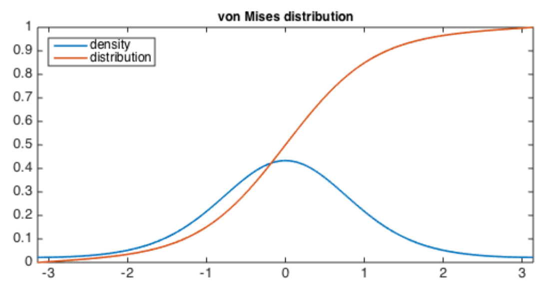

Figure 2.4: PDF を CDF に変換

2.4.2 Rejection sampling

必要な分布 $mq(z)$ は非常に複雑であり、直接サンプリングできないため、単純な分布 $q(z)$(proposed distribution) 提案分布を作成します $(\forall z^{(i)}, mq(z^i)) \geq p(z^{(i)})$。ランダムサンプリングが 2 つの分布の間にある場合は拒否され、真の分布内にある場合は受け入れられます。Acceptance rate が高いほど、サンプリング効率が高くなります。

$$Acceptance\ rate = \frac{p(z^{(i)})}{mq(z^{(i)})} \tag{2.72}$$

Reference: [3]

[2] https://www.chebfun.org/examples/stats/ResamplingRandomVariables.html
[3] https://towardsdatascience.com/monte-carlo-integration-and-sampling-methods-25d5af53e1

Algorithm 3 Rejection sampling

Require: $z^{(i)} \sim q(z)$; $u^{(i)} \sim U(0,1)$
Ensure: $\forall z^{(i)}, mq(z^{i})) \geq p(z^{(i)})$
  if $u \leq Acceptance\ rate$ **then**
    $acceptance : z^{(i)}$
  else
    $rejection : z^{(i)}$
  end if

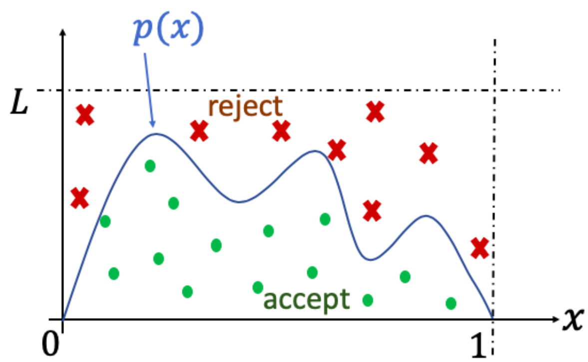

Figure 2.5: Concept of Rejection Sampling

2.4.3 Importance sampling

重要度サンプリングは、確率分布を直接サンプリングするのではなく、確率分布の期待値を直接サンプリングします。

$$E_{p(z)}[f(x)] = \int p(z) \cdot f(z)dz = \int f(z) \cdot \frac{p(z)}{q(z)} \cdot q(z)dz \tag{2.73}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} f(z^{(i)}) \cdot \frac{p(z^{(i)})}{q(z^{(i)})} \tag{2.74}$$

2.4.4 MCMC-MH

Markov chain Monte Carlo はサンプリングに基づくランダム近似法。(別に数値積分)
Markov Chain: Time and state are discrete.
State space: $\{x_1, x_2, ..., x_m\}$ [4]

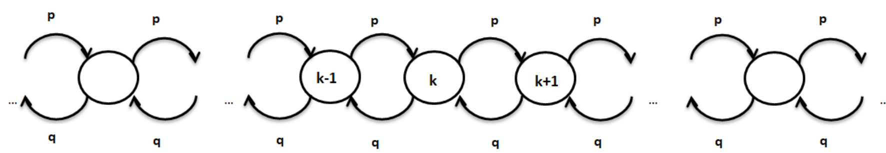

Figure 2.6: Markov Chain

Transition matrix (stochastic matrix): $Q = \begin{pmatrix} Q_{11} & Q_{12} & \cdots & Q_{1m} \\ Q_{21} & Q_{22} & \cdots & Q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{m1} & Q_{m2} & \cdots & Q_{mm} \end{pmatrix}$

$$\because p^{(t+1)} = \left(p^{(t+1)}_{(x_1)} p^{(t+1)}_{(x_2)} \cdots p^{(t+1)}_{(x_m)}\right) \tag{2.75}$$

$$\because p^{(t+1)}_{(x_j)} = \sum_{i}^{K} p^{(t)}_{(x=i)} \cdot Q_{i1} \tag{2.76}$$

$$\therefore p^{(t+1)} = \left(\sum_{i}^{K} p^{(t)}_{(x=i)} \cdot Q_{i1}, \sum_{i}^{K} p^{(t)}_{(x=i)} \cdot Q_{i2}, \cdots, \sum_{i}^{K} p^{(t)}_{(x=i)} \cdot Q_{im}\right)_{1\times m} \tag{2.77}$$

$$= p^{(t)}_{1\times m} \cdot Q \tag{2.78}$$

Detailed Balance: $p(x) \cdot P(x \to x^*) = p(x) \cdot P(x^* \to x)$

[4] https://www.analyticsvidhya.com/blog/2021/02/markov-chain-mathematical-formulation-intuitive-explanation-ap

s.t. $\alpha(x, x^*)$: Acceptance rate

$$\begin{aligned}
p(x) \cdot P(x \to x^*) &= p(x) \cdot P(x^* \to x) && (2.79)\\
p(x) \cdot Q(x \to x^*) \cdot \alpha(x, x^*) &= p(x^*) \cdot Q(x^* \to x) \cdot \alpha(x^*, x) && (2.80)\\
&= p(x) \cdot Q(x \to x^*) \cdot min\left(1, \frac{p(x^*) \cdot Q(x^* \to x)}{p(x) \cdot Q(x \to x^*)}\right) && (2.81)\\
&= min\big(p(x) \cdot Q(x \to x^*), p(x^*) \cdot Q(x^* \to x)\big) && (2.82)\\
&= p(x^*) \cdot Q(x^* \to x) \cdot min\left(1, \frac{p(x) \cdot Q(x \to x^*)}{p(x^*) \cdot Q(x^* \to x)}\right) && (2.83)\\
&= p(x^*) \cdot Q(x^* \to x) \cdot \alpha(x^*, x) && (2.84)
\end{aligned}$$

Algorithm 4 Metropolis Hasting

Require: $u \sim U(0, 1)$; $x^* \sim Q(x|x^{i-1})$
Ensure: $\alpha = min\left(1, \frac{p(x^*) \cdot Q(x^* \to x)}{p(x) \cdot Q(x \to x^*)}\right)$
  if $u \leq \alpha$ **then**
    $x^{(i)} = x^*$
  else
    $x^{(i)} = x^{(i-1)}$
  end if

Stationary Distribution: [5]

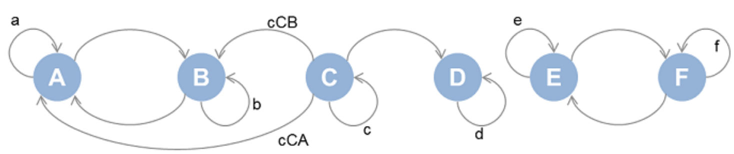

Figure 2.7: Define (positive) transition probabilities between states A through F as shown in the above image.

2.4.5 MCMC-Gibbs

MH の $\alpha = 1$ Define: $z_i \sim p(z_i | z_{1,2,\ldots,i-1,i+1,\ldots,n})$

[5] https://jp.mathworks.com/help/symbolic/markov-chain-analysis-and-stationary-distribution.html?lang=en

$$
\begin{aligned}
&z_1^{t+1} \sim p(z_1|z_2^t, ..., z_n^t) && (2.85)\\
&z_2^{t+1} \sim p(z_2|z_1^t, z_3^t..., z_n^t) && (2.86)\\
&\vdots && (2.87)\\
&z_i^{t+1} \sim p(z_i|z_1^t, ..., z_{i-1}^t, z_{i+1}^t, ..., z_n^t) && (2.88)
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\frac{p(z^*) \cdot Q(z^* \to z)}{p(z) \cdot Q(z \to z^*)} &= \frac{p(z_i^*|p(z_{-i}^*)) \cdot p(z_{-i}^*) \cdot p(z_i|z_{-i}^*)}{p(z_i|z_{-i}^* \cdot p(z_{-i}) \cdot p(z_i^*|z_{-i})} && (2.89)\\
&= \frac{p(z_i^*|p(z_{-i}^*)) \cdot p(z_{-i}^*) \cdot p(z_i^*|z_{-i}^*)}{p(z_i^*|z_{-i}^* \cdot p(z_{-i}^*) \cdot p(z_i^*|z_{-i}^*)} && (2.90)\\
&= 1 && (2.91)
\end{aligned}
$$

□

2.5 Dynamic System（State Space Model）

$$
\begin{cases}
Learning : \lambda_{MLE} = \underset{\lambda}{argmax}\, P(x|\lambda) : Baum\ Welch(EM) \\
Inference : \begin{cases}
Decoding : z = \underset{Z}{argmax}\, P(Z|X) : Viterbi\ algorithm \\
Prob\ of\ evidence : P(X|\theta) : Forward - Backward\ algorithm \\
Filtering : P(z_t|x_1, x_2, ..., x_t) : Forward\ algorithm \\
Smoothing : P(z_t|x_1, x_2, ..., x_T) : Forward - Backward\ algorithm \\
Prediction : \begin{cases} P(z_{t+1}|x_1, x_2, ..., x_t) : Forward\ algorithm \\ P(x_{t+1}|x_1, x_2, ..., x_t) : Forward\ algorithm \end{cases}
\end{cases}
\end{cases}
$$

2.5.1 HMM

Define:
State sequence: $I = i_1, i_2, ..., i_T$
Observation sequence: $O = O_1, O_2, ..., O_T$
State value collection: $Q = \{q_1, q_2, ..., q_N\}$
Collection of observations: $V = \{v_1, v_2, ..., v_N\}$
One model:

- $\lambda = (\pi, A, B)$
- π = Initial probability distribution$\rightarrow \pi = (\pi_1, \pi_2, ..., \pi_N), \sum_{i=1}^N \pi_i = 1$

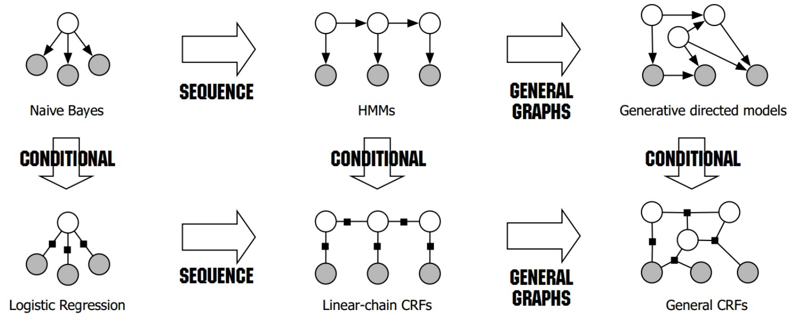

Figure 2.8: PGMs

- A = State transition matrix→ $a_{ij} = P(i_{t+1} = q_j | i_t = q_i)$
- B = Emission matrix→ $b_j(k) = P(o_t = v_k | i_t = q_j)$

Two hypotheses:

- Homogeneous Markov hypothesis→ $P(i_{t+1} | i_1, ..., i_t, o_1, ..., o_t) = P(i_{t+1} | i_t)$
- Observational independence hypothesis→ $P(o_t | i_1, ..., i_t, o_1, ..., o_t) = P(o_t | i_t)$

Three questions

- Evaluation: $P(O|\lambda)$: Forward-Backward
- Learning: $\lambda_{MLE} = \underset{\lambda}{argmax}\, P(O|\lambda)$: Baum Welch(EM)
- Decoding: $\hat{I} = \underset{I}{argmax}\, P(I|O, \lambda)$: Viterbi

Evaluation:

$$P(O|\lambda) = \sum_I P(I, O|\lambda) \tag{2.92}$$
$$= \sum_I P(O|I, \lambda) \cdot P(I|\lambda) \tag{2.93}$$
$$\tag{2.94}$$

$$\because P(O|I,\lambda) = \prod_{t=1}^{T} b_{it}(o_t) \tag{2.95}$$

$$\because P(I|\lambda) = P(i_1, i_2, ..., i_T|\lambda) \tag{2.96}$$

$$= P(i_T|i_1, i_2, ..., i_{T-1}\lambda) \cdot P(i_1, i_2, ..., i_{T-1}\lambda) \tag{2.97}$$

$$= \pi(a_{i1}) \cdot \prod_{t=2}^{T} a_{i_{t-1},i_t} \tag{2.98}$$

$$\therefore P(O|\lambda) = \prod_{t=1}^{T} b_{it}(o_t) \cdot \pi(a_{i1}) \cdot \prod_{t=2}^{T} a_{i_{t-1},i_t} \tag{2.99}$$

$$= \sum_{i_1} \cdots \sum_{i_T} \pi(a_{i1}) \cdot \prod_{t=2}^{T} a_{i_{t-1},i_t} \cdot \prod_{t=1}^{T} b_{it}(o_t) \tag{2.100}$$

Forward Algorithm:
Define: $\alpha_t(i) = P(o_1, ..., o_t, i_t = q_i|\lambda)$
Then $P(O|\lambda) = \sum_{i=1}^{N} P(O, i_t = q_i|\lambda) = \sum_{i}^{N} \alpha_T(i)$

$$\alpha_{t+1}(j) = \sum_{i}^{N} P(o_1, ..., o_t, o_{t+1}, i_{t+1} = q_j, i_t = q_i|\lambda) \tag{2.101}$$

$$= \sum_{i=1}^{N} P(o_{t+1}|o_1, ..., o_t, i_{t+1} = q_j, i_t = q_i, \lambda) \cdot P(o_1, ..., o_t, i_{t+1} = q_j, i_t = q_i|\lambda) \tag{2.102}$$

$$= \sum_{i=1}^{N} P(o_{t+1}|i_{t+1} = q_j) \cdot P(o_1, ..., o_t, i_{t+1} = q_j, i_t = q_i|\lambda) \tag{2.103}$$

$$= \sum_{i=1}^{N} P(o_{t+1}|i_{t+1} = q_j) \cdot P(i_{t+1} = q_j|o_1, ..., o_t, i_t = q_i, \lambda) \cdot P(o_1, ..., o_t, i_t = q_i|\lambda) \tag{2.104}$$

$$= \sum_{i=1}^{N} b_j(o_{t+1}) \cdot a_{ij} \cdot \lambda_t(i) \tag{2.105}$$

Backward Algorithm:
s.t. $\beta_t(i) = P(o_{t+1}, ..., o_T | i_t = q_i, \lambda), \cdots, \beta_1(i) = P(o_2, ..., o_T | i_1 = q_i, \lambda)$

$$\begin{aligned}
P(O|\lambda) &= P(o_1, ...o_T|\lambda) && (2.106)\\
&= \sum_{i=1}^{N} P(o_1, ...o_T, i_1 = q_i) && (2.107)\\
&= \sum_{i=1}^{N} P(o_1, ...o_T|i_1 = q_i) \cdot P(i_1 = q_i) && (2.108)\\
&= \sum_{i=1}^{N} P(o_1|o_2, ...o_T, i_1 = q_i) \cdot P(o_1, ...o_T|i_1 = q_i) \cdot \pi_i && (2.109)\\
&= \sum_{i=1}^{N} P(o_1|i_1 = q_i)\beta_1(i) \cdot \pi_i && (2.110)\\
&= \sum_{i=1}^{N} b_i(o_1)\pi_i\beta_1(i) && (2.111)
\end{aligned}$$

$$\begin{aligned}
\beta_t(i) &= P(o_{t+1}, ..., o_T|i_t = q_i) && (2.112)\\
&= \sum_{j=1}^{N} P(o_{t+1}, ..., o_T, i_{t+1} = q_j|i_t = q_i) && (2.113)\\
&= \sum_{j=1}^{N} P(o_{t+1}, ..., o_T|i_{t+1} = q_j, i_t = q_i) \cdot P(i_{t+1} = q_j|i_t = q_i) && (2.114)\\
&= \sum_{j=1}^{N} P(o_{t+1}, ..., o_T|i_{t+1} = q_j) \cdot a_{ij} && (2.115)\\
&= \sum_{j=1}^{N} P(o_{t+1}|o_{t+2}, ..., o_T, i_{t+1} = q_j) \cdot P(o_{t+2}, ..., o_T|i_{t+1} = q_j) \cdot a_{ij} && (2.116)\\
&= \sum_{j=1}^{N} b_j(o_{t+1}) \cdot a_{ij} \cdot \beta_{t+1}(j) && (2.117)
\end{aligned}$$

Learning: Baum-Welch
EM: $\theta^(t+1) = \underset{\theta}{argmax} \int_Z \log P(X, Z|\theta) \cdot P(Z|X, \theta^(t))dz$

$\lambda = (\pi, A, B)$:

$$\lambda^{t+1} = \underset{\theta}{argmax} \sum_I \log P(O, I|\theta) \cdot P(I|O, \theta^(t)) \tag{2.118}$$

$$= \sum_I \Big[\big(\log \pi_{i1} + \sum_{t=2}^{T} \log a_{i_{t-1}, i_t} + \sum_{t=1}^{T} \log b_{i_t}(0_t)\big) \cdot P(O, I|\lambda^t)\Big] \tag{2.119}$$

π:

$$\pi^{(t+1)} = \underset{\pi}{argmax} \sum_I \big[\log \pi_{i_1} \cdot P(O, I|\lambda^t)\big] \tag{2.120}$$

$$= \underset{\pi}{argmax} \sum_{i_1} \cdots \sum_{i_T} \big[\log \pi_{i_1} \cdot P(O, i_1, ..., i_T|\lambda^t)\big] \tag{2.121}$$

$$= \underset{\pi}{argmax} \sum_{i=1}^{N} \big[\log \pi_i \cdot P(O, i_1 = q_i|\lambda^t)\big] \tag{2.122}$$

$$s.t.\ \sum_{i=1}^{N} \pi_i = 1 \tag{2.123}$$

$$\mathcal{L}(\pi, \eta) = \sum_{i=1}^{N} \log \pi_i P(O, i_1 = q_i|\lambda^{(t)}) + \eta(\sum_{i=1}^{N} \pi_i - 1) \tag{2.124}$$

$$\tag{2.125}$$

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = \frac{1}{\pi_i} P(O, i_1 = q_i|\lambda^{(t)}) + \eta = 0 \tag{2.126}$$

$$\sum_{i}^{N} \big[P(O, i_1 = q_i|\lambda^{(t)}) + \pi_i \eta\big] = 0 \tag{2.127}$$

$$P(O|\lambda^t) + \eta = 0 \tag{2.128}$$

$$\eta = -P(O|\lambda^{(t)}) \tag{2.129}$$

$$\pi_i = \frac{P(O, i_1 = q_i|\lambda^{(t)})}{P(O|\lambda^{(t)}} \tag{2.130}$$

Decoding:

$$\delta_t(i) = \max_{i_1, ... i_{t-1}} P(o_1, ..., o_t, i_1, ..., i_{t-1}, i_t = q_i) \tag{2.131}$$

$$\delta_{t+1}(i) = \max_{i_1, ... i_t} P(o_1, ..., o_{t+1}, i_1, ..., i_t, i_{t+1} = q_j) \tag{2.132}$$

$$= \max_{1 \le i \le N} \delta_t(i) \cdot a_{ij} \cdot b_j(o_{t+1}) \tag{2.133}$$

$$\varphi_{t+1}(j) = \underset{1 \le i \le N}{argmax}\, \delta_t(i) \cdot a_{ij} \tag{2.134}$$

2.5.2 Kalman filter

HMM focus on decoding,Linear Dynamic System and Non-linear Non-Gauss focus on filter
Define:
$P(Z_t|Z_{t-1}) = N(A \cdot Z_{t-1} + B, Q)$
$P(X_t|Z_t) = N(C \cdot Z_t + D, R)$
$P(Z_1) = N(\mu_1, \sigma_1)$
$Z_t = A \cdot Z_{t-1} + B + \epsilon, \epsilon \sim N(0, Q)$
$X_t = C \cdot Z + D + \delta, \delta \sim N(0, R)$
Filter Prob: $P(Z_t|X_1, ..., X_t)$
Step1: Prediction $\rightarrow$ prior
$P(Z_t|X_1, ..., X_{t-1}) = \int_{Z_{t-1}} P(Z_{t-1|X_1,...,X_{t-1}})dz_{t-1}$ Step2: Update $\rightarrow$ posterrior
$P(Z_t|X_1, ..., X_t) \approx P(X_t|Z_t) \cdot P(Z_t|X_1, ..., X_{t-1})$

Proof. Step1: Prediction

$$P(Z_t|X_1, ..., X_{t-1}) = \int_{Z_{t-1}} P(Z_{t-1}, Z_t|X_1, ..., X_{t-1})dz_{t-1} \tag{2.135}$$
$$= \int_{Z_{t-1}} P(Z_t|Z_{t-1}, X_1, ..., X_{t-1}) \cdot P(Z_{t-1}|X_1, ..., X_{t-1})dz_{t-1} \tag{2.136}$$
$$\tag{2.137}$$

□

Proof. Step2: Update

$$P(Z_t|X_1, ..., X_t) = \frac{P(x_1, ..., X_t, Z_t)}{P(X_1, ..., X_t)} \tag{2.138}$$
$$= \frac{1}{C} \cdot P(x_1, ..., X_t, Z_t) \tag{2.139}$$
$$= \frac{1}{C} \cdot P(X_t|X_1, ..., X_{t-1}, Z_t) \cdot P(X_1, ..., X_{t-1}, Z_t) \tag{2.140}$$
$$= \frac{1}{C} \cdot P(X_t|Z_t) \cdot P(Z_t|X_1, ..., X_{t-1}) \cdot P(X_1, ..., X_{t-1}) \tag{2.141}$$
$$= \frac{D}{C} \cdot P(X_t|Z_t) \cdot P(Z_t|X_1, ..., X_{t-1}) \tag{2.142}$$

□

2.5.3 Particle filter - SIS

Non-linear Non-Gauss Dynamic System
Monte Carlo Method: $P(Z|X) \to E_{Z|X}[f(Z)] = \int f(Z) \cdot p(Z)dz$
$\approx \frac{1}{N} \sum_{i=1}^{N} f(Z^{(i)})$
Importance Sampling: $E[f(Z)] = \int f(Z) \cdot p(Z) \cdot dz = \int f(Z) \cdot \frac{p(Z)}{q(Z)} \cdot q(Z)dz$; q(Z): proposal dist.
$\approx \frac{1}{N} \sum_{i=1}^{N} f(Z^{(i)}) \cdot \frac{p(Z^{(i)})}{q(Z^{(i)})}$; $\frac{p(Z^{(i)})}{q(Z^{(i)})}$ is weight $w^{(i)}$
Sequential Importance Sampling:
$P(Z_t|X_{1:t}) \to P(Z_{1:t}|X_{1:t})$
So: $w^{(i)} \approx \frac{P(Z_{1:t}, X_{1:t})}{q(Z_{1:t}|X_{1:t})}$

Proof. $w_t^{(i)} \to w_{(t-1)}^i$

$$P(Z_{1:t}|X_{1:t}) = \frac{P(Z_{1:t}, X_{1:t})}{P(X_{1:t})} \tag{2.143}$$

$$= \frac{1}{C} \cdot P(Z_{1:t}, X_{1:t}) \tag{2.144}$$

$$= \frac{1}{C} \cdot P(X_t|Z_{1:t}, X_{1:t-1}) \cdot P(Z_{1:t}, X_{1:t-1}) \tag{2.145}$$

$$= \frac{1}{C} \cdot P(X_t|Z_t) \cdot P(Z_{1:t}, X_{1:t-1}) \tag{2.146}$$

$$= \frac{1}{C} \cdot P(X_t|Z_t) \cdot P(Z_t|Z_{1:t-1}, X_{1:t-1}) \cdot P(Z_{1:t-1}, X_{1:t-1}) \tag{2.147}$$

$$= \frac{1}{C} \cdot P(X_t|Z_t) \cdot P(Z_t|Z_{t-1}) \cdot P(Z_{1:t-1}, X_{1:t-1}) \tag{2.148}$$

$$= \frac{1}{C} \cdot P(X_t|Z_t) \cdot P(Z_t|Z_{t-1}) \cdot P(Z_{1:t-1}|X_{1:t-1}) \cdot P(X_{1:t-1}) \tag{2.149}$$

$$= \frac{D}{C} \cdot P(X_t|Z_t) \cdot P(Z_t|Z_{t-1}) \cdot P(Z_{1:t-1}|X_{1:t-1}) \tag{2.150}$$

s.t.
$q(Z_{1:t}|X_{1:t}) = q(Z_t|Z_{1:t-1}, X_{1:t}) \cdot q(Z_{1:t-1}|X_{1:t-1})$

$$w_t^{(i)} \approx \frac{P(Z_{1:t}, X_{1:t})}{q(Z_{1:t}|X_{1:t})} \tag{2.151}$$

$$\approx \frac{P(X_t|Z_t) \cdot P(Z_t|Z_{t-1}) \cdot P(Z_{1:t-1}|X_{1:t-1})}{q(Z_t|Z_{1:t-1}, X_{1:t}) \cdot q(Z_{1:t-1}|X_{1:t-1})} \tag{2.152}$$

$$\approx \frac{P(X_t|Z_t) \cdot P(Z_t|Z_{t-1})}{q(Z_t|Z_{1:t-1}, X_{1:t})} \cdot w_{(t-1)}^i \tag{2.153}$$

□

Algorithm 5 Sequential Importance Sampling

Require: $t-1 \rightarrow w_{t-1}^{(i)}$ *is end*
for $i = 1, \cdots, N$ in t **do**
$Z_t^{(i)} \sim q(Z_t|Z_{t-1}, X_{1:t})$
$w_t^{(i)} = w_{t-1}^{(i)}$
end for
Normalized: $w_t^{(i)} = \sum_{i=1}^{N} w_t^i = 1$; (Prob: Weight Degradation)

2.5.4 Particle filter - SIR

Prob: Weight Degradation $\rightarrow$ Resampling
Resampling: $q(Z_t|Z_{1:t-1}, X_{1:t}) = P(Z_t|Z_{t-1})$ (Generate and Test)

$$w_t^{(i)} \approx \frac{P(Z_{1:t}, X_{1:t})}{q(Z_{1:t}|X_{1:t})} \approx \frac{P(X_t|Z_t) \cdot P(Z_t|Z_{t-1})}{P(Z_t|Z_{t-1})} \cdot w_{(t-1)}^i \tag{2.154}$$

Algorithm 6 Sampling Importance Resampling

Require: $t-1 \rightarrow w_{t-1}^{(i)}$ *is end*
for $i = 1, \cdots, N$ in t **do**
$Z_t^{(i)} \sim P(Z_t|Z_{t-1})$
$w_t^{(i)} = \frac{P(X_t|Z_t^{(i)}) \cdot P(Z_t^{(i)}|Z_{t-1}^{(i)})}{P(Z_t^{(i)}|Z_{t-1}^{(i)})} \cdot w_{(t-1)}^i = P(X_t|Z_t^{(i)}) \cdot w_{(t-1)}^i$
end for
Normalized: $w_t^{(i)} = \sum_{i=1}^{N} w_t^i = 1$
Resampling: $w_t^{(i)} = \frac{1}{N}$

2.5.5 CRF

[6] HMM:

$$P(X, Y|\lambda) = \prod_{t=1}^{T} P(x_t, y_t|\lambda) \tag{2.155}$$

$$= \prod_{t=1}^{T} P(y_t|y_{t-1}, \lambda) \cdot P(x_t|y_t, \lambda) \tag{2.156}$$

[6] https://zhuanlan.zhihu.com/p/34736498

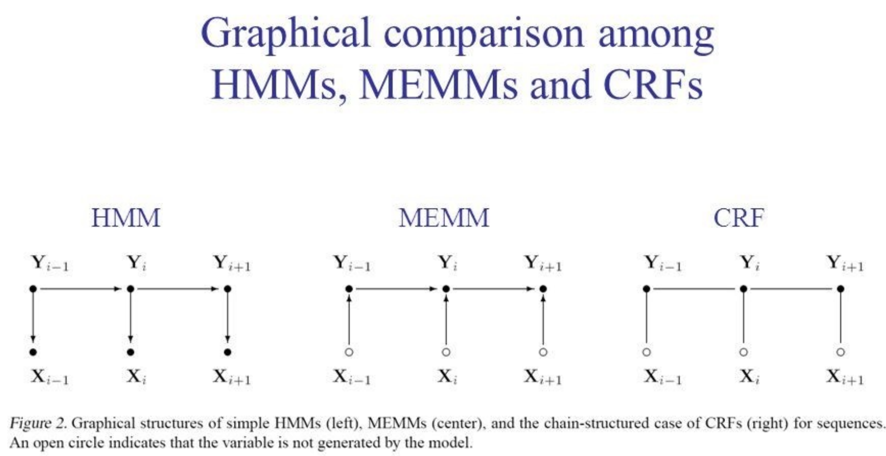

Figure 2.9: HMM → MEMM → CRF

MEMM: (Label Bias Problem - mass score)

$$P(Y|X,\lambda) = \prod_{t=1}^{T} P(y_t|y_{t-1}, x_{1:T}, \lambda) \tag{2.157}$$

無向グラフモデル（MRF）の因数分解の定義：

$$P(X) = \frac{1}{Z}\prod_{i=1}^{K} \psi_i(X_{c_i}) \tag{2.158}$$

$$= \frac{1}{Z}\prod_{i=1}^{K} \exp\big[-E_i(X_{c_i})\big] \tag{2.159}$$

$$= \frac{1}{Z}\exp\sum_{i=1}^{K} F_i(X_{c_i}) \tag{2.160}$$

Then Linear CRF(PDF)：

$$
\begin{aligned}
P(Y|X) &= \frac{1}{Z} \exp \sum_{t=1}^{T} F_t(y_{t-1}, y_t x_{1:T}) && (2.161)\\
&= \frac{1}{Z} \exp \sum_{t=1}^{T} \left(\triangle_{y_t, x_{1:T}} + \triangle_{y_{t-1}, y_t, x_{1:T}} \right) && (2.162)\\
&= \frac{1}{Z} \exp \sum_{t=1}^{T} \left(\sum_{k=1}^{K} \lambda_k I_k(y_{t-1}, y_t, x_{1:T}) + \sum_{l=1}^{L} \eta_l I_l'(y_t, x_{1:T}) \right) && (2.163)\\
& && (2.164)
\end{aligned}
$$

Define:

$$
y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}; \; x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{pmatrix}; \; \lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_K \end{pmatrix}; \; \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_L \end{pmatrix}; \; I = \begin{pmatrix} I_1 \\ I_2 \\ \vdots \\ I_K \end{pmatrix} = I(y_{t-1}, y_t, x);
$$

$$
I' = \begin{pmatrix} I_1' \\ I_2' \\ \vdots \\ I_L' \end{pmatrix} = I'(y_t, x)
$$

$$
\begin{aligned}
P(Y=y|X=x) &= \frac{1}{Z(x,\lambda,\eta)} \exp \sum_{t=1}^{T} \left(\lambda^T \cdot I(y_{t-1}, y_t, x_{1:T}) + \eta^T I'(y_t, x_{1:T}) \right) && (2.165)\\
&= \frac{1}{Z(x,\lambda,\eta)} \exp \left(\lambda^T \cdot \sum_{t=1}^{T} I(y_{t-1}, y_t, x_{1:T}) + \eta^T \sum_{t=1}^{T} I'(y_t, x_{1:T}) \right) && (2.166)
\end{aligned}
$$

Define:

$$
\theta = \begin{pmatrix} \lambda \\ \eta \end{pmatrix}_{K+L}; \; H = \begin{pmatrix} \sum_{t=1}^{T} I \\ \sum_{t=1}^{T} I' \end{pmatrix}_{K+L}
$$

Then:

$$
\begin{aligned}
P(Y=y|X=x) &= \frac{1}{Z(x,\theta)} \exp \theta^T \cdot H(y_t, y_{t-1}, x) && (2.167)\\
&= \frac{1}{Z(x,\theta)} \exp < \theta, H > && (2.168)
\end{aligned}
$$

LEARNING AND INFERENCE:

Learning: parameter estimation
Given training data: $\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ then $\hat{\theta} = argmax \prod_{i=1}^{N} P(y^{(i)}|x^{(i)})$
Inference - marginal problem: $P(y_t|x)$
MAP Inference - decoding: $\hat{y} = \underset{y=y_{1:T}}{argmax}\, P(y|x)$

Proof. Inference - marginal problem(sum product): Given $P(Y = y|X = x) \rightarrow P(y_t = i|x)$

$$P(y|x) = \sum_{y_{1:t-1}} \sum_{y_{t+1:T}} \frac{1}{Z} \prod_{t=1}^{T} \psi_t(y_{t-1}, y_t, x) \tag{2.169}$$

$$= \frac{1}{Z} \sum_{y_{1:t-1}} \psi_1(y_0, y_1, x) \cdot \psi_2(y_1, y_2, x) \cdots \psi_t(y_{t-1}, y_t, x) \cdot \sum_{y_{t+1:T}} \psi_1(y_t, y_{t+1}, x) \cdots \psi_T(y_{T-1}, y_T, x) \tag{2.170}$$

□

Proof. Learning: parameter estimation

$$\hat{\theta} = argmax \prod_{i=1}^{N} P(y^{(i)}|x^{(i)}) \tag{2.171}$$

$$\hat{\lambda}, \hat{\eta} = \underset{\lambda,\eta}{argmax} \prod_{i=1}^{N} P(y^{(i)}|x^{(i)}) \tag{2.172}$$

$$= \underset{\lambda,\eta}{argmax} \frac{1}{Z(x, \lambda, \eta)} \exp \sum_{t=1}^{T} \left(\lambda^T \cdot I(y_{t-1}, y_t, x_{1:T}) + \eta^T I'(y_t, x) \right) \tag{2.173}$$

$$L(\lambda, \eta, x^{(i)}) = \underset{\lambda,\eta}{argmax} \sum_{i=1}^{N} \left[\log Z(x^{(i)}, \lambda, \eta) + \sum_{t=1}^{T} \left(\lambda^T \cdot I(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)}) + \eta^T I'(y_t^{(i)}, x^{(i)}) \right) \right] \tag{2.174}$$

Gradient ascent or $\cdots$:

$$\nabla_\lambda L = \sum_{i=1}^{N}\Big(\sum_{t=1}^{T} I(y_{t-1},y_t,x^{(i)}) - \nabla_\lambda \log Z(x^{(i)},\lambda,\eta)\Big) \tag{2.175}$$

$$= \sum_{i=1}^{N}\Big(\sum_{t=1}^{T} I(y_{t-1},y_t,x^{(i)}) - E\Big[\sum_{t=1}^{T} I(y_{t-1},y_t,x^{(i)})\Big]\Big) \to log - partition\ function \tag{2.176}$$

$$= \sum_{i=1}^{N}\Big(\sum_{t=1}^{T} I(y_{t-1},y_t,x^{(i)}) - \sum_{y} P(y|x^{(i)})\cdot I(y_{t-1},y_t,x^{(i)})\Big) \tag{2.177}$$

$$= \sum_{i=1}^{N}\Big(\sum_{t=1}^{T} I(y_{t-1},y_t,x^{(i)}) - \sum_{t=1}^{T}\Big(\sum_{y} P(y|x^{(i)})\cdot I(y_{t-1},y_t,x^{(i)})\Big)\Big) \tag{2.178}$$

$$= \sum_{i=1}^{N}\Big(\sum_{t=1}^{T} I(y_{t-1},y_t,x^{(i)}) - \sum_{t=1}^{T}\Big(\sum_{y_{1:t-2}}\sum_{y_{t-1:t}}\sum_{y_{t+1:T}} P(y|x^{(i)})\cdot I(y_{t-1},y_t,x^{(i)})\Big)\Big) \tag{2.179}$$

$$= \sum_{i=1}^{N}\Big(\sum_{t=1}^{T} I(y_{t-1},y_t,x^{(i)}) - \sum_{t=1}^{T}\sum_{y_{t-1:t}}\Big(\sum_{y_{1:t-2}}\sum_{y_{t+1:T}} P(y|x^{(i)})\cdot I(y_{t-1},y_t,x^{(i)})\Big)\Big) \tag{2.180}$$

$$= \sum_{i=1}^{N}\Big(\sum_{t=1}^{T} I(y_{t-1},y_t,x^{(i)}) - \sum_{t=1}^{T}\sum_{y_{t-1:t}}\Big(P(y_{t-1},y_t,x^{(i)})\cdot I(y_{t-1},y_t,x^{(i)})\Big)\Big) \tag{2.181}$$

$$= \sum_{i=1}^{N}\Big(\sum_{t=1}^{T} I(y_{t-1},y_t,x^{(i)}) - \sum_{t=1}^{T}\sum_{y_{t-1:t}}\Big(marginal\ problem\cdot I(y_{t-1},y_t,x^{(i)})\Big)\Big) \tag{2.182}$$

□

2.5.6 RBM

Restricted Boltzmann Machine
Define:
$X \in R^{p\times 1} = (h,v)^T$
$h \in R^{m\times 1}; v \in R^{n\times 1}; p = m+n$

Boltzmann Distribution(Gibbs Distribution):

$$P(x) = \frac{1}{Z} \exp\big\{ - E(x)\big\} \tag{2.183}$$
$$P(h,v) = \frac{1}{Z} \exp\big\{ - E(h,v)\big\} \tag{2.184}$$
$$\tag{2.185}$$

$$E(h,v) = -\big(h^T wv + \alpha^T v + \beta^T h\big) \tag{2.186}$$
$$= -\Bigg(\sum_{i=1}^{m}\sum_{j=1}^{n} h_i w_{ij} v_j + \sum_{j=1}^{n} \alpha_j v_j + \sum_{i=1}^{m} \beta_i h_i \Bigg) \tag{2.187}$$

Inference:
$Posterior: P(h|v), P(v|h)$ Define: $h_l \in \{0,1\}$ →Binary RBF

$$P(h|v) = \prod_{l=1}^{m} P(h_l|v) \tag{2.188}$$
$$P(h=1|v) = \frac{P(h_l=1, h_{-l}, v)}{P(h_{-l}, v)} \tag{2.189}$$
$$= \frac{P(h_l=1, h_{-l}, v)}{P(h_l=1, h_{-l}, v) + P(h_l=0, h_{-l}, v} \tag{2.190}$$
$$\tag{2.191}$$

Because:

$$E(h,v) = -\Bigg(\sum_{i=1}^{m}\sum_{j=1}^{n} h_i w_{ij} v_j + \sum_{j=1}^{n} \alpha_j v_j + \sum_{i=1}^{m} \beta_i h_i \Bigg) \tag{2.192}$$
$$= -\Bigg(\sum_{i=1, i\neq j}^{m}\sum_{j=1}^{n} h_i w_{ij} v_j + h_l \sum_{j=1}^{n} w_{ij} v_j + \sum_{j=1}^{n} \alpha_j v_j + \sum_{i=1, i\neq j}^{m} \beta_i h_i + \beta_l h_l \tag{2.193}$$
$$\tag{2.194}$$

Define:
$h_l \sum_{j=1}^{n} w_{ij} v_j + \beta_l h_l = h_l \Bigg(\sum_{j=1}^{n} w_{ij} v_j + \beta_l \Bigg) = h_l \cdot H_l(v)$
Therefore:
$E(h,v) = h_l \cdot H_l(v) + \bar{H}_l(h_{-l}, v)$

Then:

$$P(h=1|v) = \frac{\frac{1}{Z}\exp\left\{H_l(v)+\bar{H}_l(h_{-l},v)\right\}}{\frac{1}{Z}\exp\left\{H_l(v)+\bar{H}_l(h_{-l},v)\right\}+\frac{1}{Z}\exp\left\{\bar{H}_l(h_{-l},v)\right\}} \tag{2.195}$$

$$= \frac{1}{1+\exp\left\{-H_l(v)\right\}} \tag{2.196}$$

$$= \sigma(H_l(v)) \tag{2.197}$$

$$= \sigma\left(\sum_{j=1}^{n} w_{lj}v_j + \beta_l\right) \tag{2.198}$$

Inference:
Marginal: $P(v)$
Define: $W = [w_{ij}]_{m\times n}$: Row vector

$$
\begin{aligned}
P(v) &= \sum_h P(h,v) && (2.199)\\
&= \sum_h \frac{1}{Z}\exp\big\{-E(h,v)\big\} && (2.200)\\
&= \sum_h \frac{1}{Z}\exp\big\{\big(h^T wv + \alpha^T v + \beta^T h\big)\big\} && (2.201)\\
&= \sum_{h_1}\cdots\sum_{h_m}\exp\big\{h^T wv + \alpha^T v + \beta^T h\big\} && (2.202)\\
&= \exp(\alpha^T v)\cdot\sum_{h_1}\cdots\sum_{h_m}\exp\big\{h^T wv + \beta^T h\big\} && (2.203)\\
&= \exp(\alpha^T v)\cdot\sum_{h_1}\cdots\sum_{h_m}\exp\Big\{\sum_{i=1^m}\big(h_i w_i v + \beta_i h_i\big)\Big\} && (2.204)\\
&= \exp(\alpha^T v)\cdot\sum_{h_1}\cdots\sum_{h_m}\exp\big\{h_i w_i v + \beta_i h_i\big\} && (2.205)\\
&= \exp(\alpha^T v)\cdot\sum_{h_1}\exp\big\{h_1 w_1 v + \beta_1 h_1\big\}\cdots\sum_{h_m}\exp\big\{h_m w_m v + \beta_m h_m\big\} && (2.206)\\
&= \exp(\alpha^T v)\cdot\big(1+\exp\big\{w_1 v + \beta_1\big\}\big)\cdots\big(1+\exp\big\{w_m v + \beta_m\big\}\big) && (2.207)\\
&= \exp(\alpha^T v)\cdot\exp\Big\{\log(1+\exp\big\{w_1 v + \beta_1\big\})\Big\}\cdots\exp\Big\{\log(1+\exp\big\{w_m v + \beta_m\big\})\Big\} && (2.208)\\
&= \exp\left(alpha^T v + \sum_{i=1}^m \log\big(1+\exp\big\{w_i v + \beta_i\big\}\big)\right) && (2.209)\\
&= \exp\left(alpha^T v + \sum_{i=1}^m Softplus(w_i v + \beta_i)\right) && (2.210)
\end{aligned}
$$

[7]

2.6 Gaussian Graph

2.6.1 Conditional independence

High-dimensional Gaussian distribution pdf:
Define:
$X_i \sim N(\mu_i, \Sigma_i)$

[7] https://medium.datadriveninvestor.com/an-intuitive-introduction-of-restricted-boltzmann-machine-rbm-14f4382

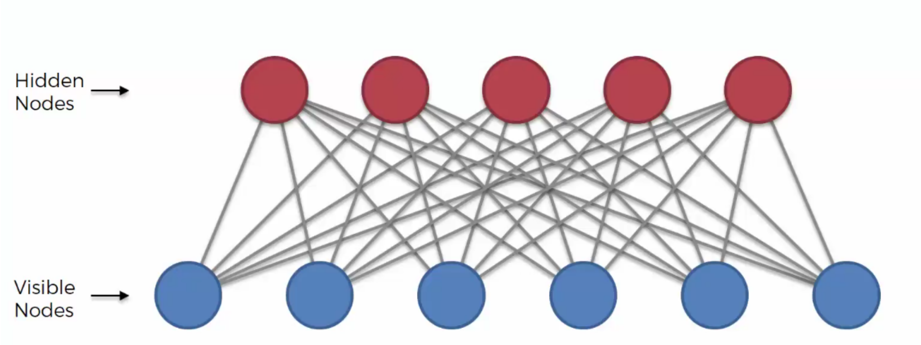

Figure 2.10: Restricted Boltzmann Machine

$X = (x_1, x_2, ..., x_p)^T$

$P(X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$

Local Marginal independent:

$$\Sigma = (\sigma_{ij}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \vdots & & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{1p} \end{pmatrix}_{p\times p} \rightarrow x_i \perp\!\!\!\perp x_j \Leftrightarrow \sigma_{ij} = 0 \tag{2.211}$$

$$\tag{2.212}$$

Local Precision matrix (Information matrix):

$$\Lambda = \Sigma^{-1} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1p} \\ \vdots & & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{1p} \end{pmatrix}_{p\times p} \rightarrow x_i \perp\!\!\!\perp x_j\big|_{-\{x_i,x_j\}} \Leftrightarrow \lambda_{ij} = 0 \tag{2.213}$$

$x = (x_i, -\{x_i\})^T = (x_a, x_b)^T \rightarrow$
Woodbury Formula and Schur Complementary: then

$$\forall x_i, x_i\big|_{-\{x_i\}} \sim N(\Sigma_{i\neq j}\frac{\lambda_{ij}}{\lambda_{ii}}x_j, \lambda_{ii}^{-1}) \tag{2.214}$$

2.6.2 Gaussian Bayesian Network

GBN(global) is based on linear Gaussian model(local, Kalman Filter):
Linear Gaussian model: $P(x) = \prod_{i=1}^{p} P(x_i|x_{i-1})$

$$P(x) = N(x|\mu_x, \Sigma_x) \tag{2.215}$$
$$P(y|x) = N(y|Ax + b, \Sigma_y) \tag{2.216}$$

GBN: $P(x) = \prod_{i=1}^{p} P(x_i|\vec{x}_{pa(i)})$

Define:
$\mu \in R^{p\times 1}$;
$\epsilon \in R^{p\times 1}$;
$S = diag(\sigma_i)$

$$P(x) = N(x|\mu_x, \Sigma_x) \tag{2.217}$$
$$P(\vec{x}_i|x_{pa(i)}) = N(X_i|\vec{\mu}_i + w_i^T\vec{x}_{pa(i)}, \sigma_i^2) \tag{2.218}$$
$$x_i = \mu_i + \sum_{j\in x_{pa(i)}} w_{ij}\cdot(x_j - \mu_j) + \sigma_i\cdot\epsilon_i \tag{2.219}$$

Then:

$$x_i - \mu_i = \sum_{j\in x_{pa(i)}} w_{ij}\cdot(x_j - \mu_j) + \sigma_i\cdot\epsilon_i \tag{2.220}$$
$$x - \mu = w\cdot(x-\mu) + \epsilon\cdot S \tag{2.221}$$
$$(I - w)\cdot(x-\mu) = \epsilon\cdot S \tag{2.222}$$
$$X - \mu = (I-w)^{-1}\epsilon\cdot S \tag{2.223}$$
$$\Sigma = cov(x) = cov(x-\mu) = cov((I-w)^{-1}\epsilon\cdot S) \tag{2.224}$$

2.6.3 Gaussian Markov Network

High-dimensional Gaussian distribution pdf:

$$P(X) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

Factorization of undirected graph model：

$$P(x) = \frac{1}{Z}\prod_{i=1}^{p}\psi_i(x_i)\cdot\prod_{i,j\in X}\psi(x_i,x_j) \tag{2.225}$$
$$= \frac{1}{Z}\prod_{i=1}^{p} node\ potential\cdot\prod_{i,j\in X} edge\ potential \tag{2.226}$$
$$\tag{2.227}$$

Gaussian Markov Network pdf:

Define:
$x \in R^{P\times 1}; \Lambda \in R^{p\times p}$
Λ: precision matrix
$\Lambda\mu$: potential matrix $\rightarrow h \in R^{p\times 1}$

$$
\begin{aligned}
P(x) &\approx \exp\big\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\big\} && (2.228)\\
&= \exp\big\{-\frac{1}{2}(x^T\Lambda-\mu^T\Lambda)(x-\mu)\big\} && (2.229)\\
&= \exp\big\{-\frac{1}{2}(x^T\Lambda x-x^T\Lambda\mu-\mu^T\Lambda x+\mu^T\Lambda\mu)\big\} && (2.230)\\
&= \exp\big\{-\frac{1}{2}(x^T\Lambda x-2\mu^T\Lambda x+\mu^T\Lambda\mu)\big\} && (2.231)\\
&\approx \big\{-\frac{1}{2}x^T\Lambda x+(\Lambda\mu)^T x)\big\} && (2.232)
\end{aligned}
$$

Then:

$$
\begin{aligned}
x_i &= -\frac{1}{2}x_i^2\cdot\lambda_{ii}+h_i x_i \rightarrow \textit{node potential} && (2.233)\\
x_i, x_j &= -\frac{1}{2}x_i^2(\lambda_{ij}x_i x_j+\lambda_{ji}x_j x_i) = -\lambda_{ij}x_i x_j \rightarrow \textit{edge potential} && (2.234)
\end{aligned}
$$

Then: If $\lambda_{ij} == 0$ then $x_i \perp\!\!\!\perp x_j$

2.6.4 Bayesian Linear Regression

Define:
$X \sim R^{N\times p}, Y \sim R^{N\times 1}$
Model:
$f(x) = w^T x = x^T w$; $y = f(x)+\epsilon$; $\epsilon \sim N(0,\sigma^2)$
Inference:

$$
\begin{aligned}
P(w|Data) &= P(w|X,Y) && (2.235)\\
&= \frac{w, Y|X}{P(Y|X)} && (2.236)\\
&= \frac{P(Y|w,X)\cdot P(w)}{\int P(Y|w,X)\cdot P(w)dw} && (2.237)
\end{aligned}
$$

Because:

$$P(Y|w,X) = \prod_{i=1}^{N} P(y_i|w,x_i) \tag{2.238}$$
$$= \prod_{i=1}^{N} N(y_i|w^T x_i, \sigma^2) \rightarrow likelihood \tag{2.239}$$
$$P(w) = N(0,\Sigma_p) \rightarrow prior \tag{2.240}$$

Therefore:

$$P(w|Data) \approx \prod_{i=1}^{N} N(y_i|w^T x_i, \sigma^2) \cdot N(0,\Sigma_p) \tag{2.241}$$
$$Gaussian \approx Gaussian \cdot Gaussian \tag{2.242}$$
$$= \prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\Big\{ -\frac{1}{2\sigma^2}(y_i - w^T x_i)^2 \Big\} \cdot N(0,\Sigma_p) \tag{2.243}$$
$$= \frac{1}{(2\pi)^{\frac{N}{2}} \cdot \sigma^N} \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - w^T x_i)^2 \Big\} \cdot N(0,\Sigma_p) \tag{2.244}$$
$$= \frac{1}{(2\pi)^{\frac{N}{2}} \cdot \sigma^N} \exp\Big\{ -\frac{1}{2}(Y - Xw)^T \sigma^{-2} \cdot I(Y - Xw) \Big\} \cdot N(0,\Sigma_p) \tag{2.245}$$
$$\approx N(Xw, \sigma^{-2} I) \cdot N(0,\Sigma_p) \tag{2.246}$$
$$\approx \exp\Big\{ -\frac{1}{2}(Y - Xw)^T \sigma^{-2} \cdot I(Y - Xw) \Big\} \cdot \exp\Big\{ -\frac{1}{2} w^T \cdot \Sigma_p^{-1} w \Big\} \tag{2.247}$$
$$= \exp\Big\{ -\frac{1}{2\sigma^2}(Y^T - w^T X^T)(Y - Xw) - \frac{1}{2} w^T \Sigma_p^{-1} w \Big\} \tag{2.248}$$
$$= \exp\Big\{ -\frac{1}{2\sigma^2}(Y^T Y - 2Y^T X w + w^T X^T X w) - \frac{1}{2} w^T \Sigma_p^{-1} w \Big\} \tag{2.249}$$
$$\tag{2.250}$$

Because:

$$\exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)) = \exp(-\frac{1}{2}(x^T\Sigma^{-1} - \mu^T\Sigma^{-1})(x-\mu)) \tag{2.251}$$
$$= \exp(-\frac{1}{2}(x^T\Sigma^{-1}x - 2\mu^T\Sigma^{-1}x) + \triangle) \tag{2.252}$$

Then:

$$\begin{aligned}
Quadratic\ term \rightarrow & -\frac{1}{2\sigma^2}\cdot(w^TX^TXw)-\frac{1}{2}w^T\Sigma_p^{-1}w && (2.253)\\
&= -\frac{1}{2}\Big(w^T(\sigma^{-2}x^Tx+\Sigma_p^{-1})w\Big) && (2.254)\\
&= -\frac{1}{2}\Big(w^T\Sigma_w^{-1}w\Big) && (2.255)\\
\Sigma_w &= \sigma^{-2}x^Tx+\Sigma_p^{-1} && (2.256)\\
One\ time\ term \rightarrow & -\frac{1}{2\sigma^2}\cdot(-2)Y^TXw && (2.257)\\
&= \sigma^{-2}Y^TXw && (2.258)\\
&= \mu_w^T\Sigma_w^{-1}w && (2.259)\\
\Sigma_w^{-1}\mu_w &= \sigma^{-2}X^TY && (2.260)\\
\mu_w &= \sigma^{-2}\Sigma_wX^TY && (2.261)
\end{aligned}$$

Prediction:
Given: $x^* \rightarrow y^*$

$$P(y^*|Data,x^*)=\int_w P(y^*|w,Data,x^*)\cdot P(w|Data,x^*)dw \quad (2.262)$$

Because:

$$w\sim N(\mu_w,\Sigma_w)\rightarrow x^{*T}w\sim N(x^{*T}\mu_w,x^{*T}\Sigma_wx^*) \quad (2.263)$$

Then:

$$P(y^*|Data,x^*)=N(x^{*T}\mu_w,x^{*T}\Sigma_wx^*+\sigma^2) \quad (2.264)$$

2.6.5 Gaussian Process Regression

2.7 Learning

2.7.1 Introduction

$$\begin{cases} Structure\ learning \\ Parameter\ learning: \begin{cases} Complete\ data \\ Hidden\ variable: EM \end{cases} \end{cases}$$

2.7.2 Proof of convergence of EM

$$\theta_{MLE} = \underset{\theta}{argmax} \log p(x|\theta) \tag{2.265}$$

$$init(z) \downarrow \tag{2.266}$$

$$\theta^{(t+1)} = \underset{\theta}{argmax} \int_z \log P(x,z|\theta) \cdot P(z|x,\theta^{(t)}) dz \tag{2.267}$$

$$= \underset{\theta}{argmax} E_{z|x\theta^{(t)}}[\log P(x,z|\theta)] \tag{2.268}$$

Proof. $\log p(x|\theta^t) \leq \log p(x|\theta^{t+1})$

$$\log p(x|\theta) = \log p(x,z|\theta) - \log p(z|x,\theta) \tag{2.269}$$

$$\int_z p(z|x,\theta^t) \cdot \log p(x|\theta) dz = \int_z p(z|x,\theta^t) \log p(x,z|\theta) dz - \int_z p(z|x,\theta^t) \log p(z|x,\theta) dz \tag{2.270}$$

$$= Q(\theta,\theta^t) - H(\theta,\theta^t) \tag{2.271}$$

$$\because \theta^{(t+1)} = \underset{\theta}{argmax} E_{z|x\theta^{(t)}}[\log P(x,z|\theta)] \tag{2.272}$$

$$\therefore Q(\theta^{t+1},\theta^t) \geq Q(\theta^t,\theta^t) \tag{2.273}$$

$$\tag{2.274}$$

$$\because H(\theta^{t+1},\theta^t) - H(\theta^t,\theta^t) = \int_z p(z|x,\theta^t) \log p(x,z|\theta^{t+1}) dz - \int_z p(z|x,\theta^t) \log p(z|x,\theta^t) dz \tag{2.275}$$

$$= \int_z p(z|x,\theta^t) \cdot \log \frac{p(z|x,\theta^{t+1})}{p(z|x,\theta^t)} \tag{2.276}$$

$$= -KL(p(z|x,\theta^t||p(z|x,\theta^{t+1}) \tag{2.277}$$

$$\leq 0 \ (or\ Jensen's\ inequality) \tag{2.278}$$

$$\therefore H(\theta^{t+1},\theta^t) \geq H(\theta^t,\theta^t) \tag{2.279}$$

□

Reference: [8]

[8] https://people.duke.edu/ ccc14/sta-663/EMAlgorithm.html

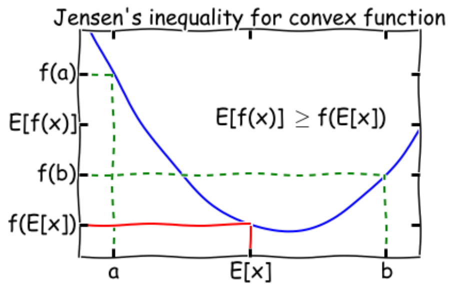

Figure 2.11: Jensen's inequality

2.7.3 ELBO+KL For EM

EM(Expectation maximization)Algorithm は座標昇順法 (Coordinate descent) のような反復更新。実は ELBO(Evidence lower bound) を最大化します。

$$\theta^{(t+1)} = \underset{\theta}{argmax} \int_z \log P(x,z|\theta) \cdot P(z|x,\theta^{(t)})dz \tag{2.280}$$

E-step: $p(z|x,\theta^t) \to E_{z|x,\theta^t}[\log p(x,z|\theta)]$;
M-step: $\theta^{t+1} = \underset{\theta}{argmax}\, E_{z|x,\theta^t}[\log p(x,z|\theta)]$

Proof. $\log p(x|\theta) = ELBO + KL(q||p)$

$$\log p(x|\theta) = \log p(x,z|\theta) - \log p(z|x,\theta) \tag{2.281}$$

$$= \log \frac{p(x,z|\theta)}{q(z)} - \log \frac{p(z|x,\theta)}{q(z)} \tag{2.282}$$

$$\int_z q(z) \cdot \log p(x|\theta)dz = \int q(z) \cdot \log \frac{p(x,z|\theta)}{q(z)}dz - \int q(z) \cdot \log \frac{p(z|x,\theta)}{q(z)}dz \tag{2.283}$$

$$\tag{2.284}$$

maximize evidence lower bound:

$$\hat{\theta} = \underset{\theta}{argmax} \int q(z) \cdot \log \frac{p(x,z|\theta)}{q(z)}dz \tag{2.285}$$

$$= \underset{\theta}{argmax} \int q(z|x,\theta^t) \cdot \log \frac{p(x,z|\theta)}{q(z|x,\theta^t)}dz \tag{2.286}$$

$$= \underset{\theta}{argmax} \int q(z|x,\theta^t) \cdot \log p(x,z|\theta)dz \tag{2.287}$$

□

Reference: [9]

[9]https://people.duke.edu/ ccc14/sta-663/EMAlgorithm.html

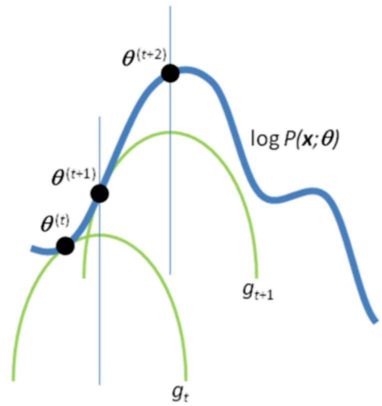

Figure 2.12: Expectation maximization

2.7.4 Jensen's inequality For EM

CS229-Andrew Ng

Proof.

$$\log p(x|\theta) = \log \int_z p(x,z|\theta)dz \tag{2.288}$$
$$= \log \int_z \frac{p(x,z|\theta)}{q(z)} \cdot q(z)dz \tag{2.289}$$
$$= \log E_{q(z)}\left[\frac{p(x,z|\theta)}{q(z)}\right] \tag{2.290}$$
$$\geq E_{q(z)}\left[\log \frac{p(x,z|\theta)}{q(z)}\right] \tag{2.291}$$

If $\frac{p(x,z|\theta)}{q(z)} == C$ then "=";

$$\therefore q(z) = \frac{1}{c}p(x,z|\theta) \tag{2.292}$$
$$1 = \int_z q(z)dz = \int_z \frac{1}{c}p(x,z|\theta)dz \tag{2.293}$$
$$= \frac{1}{c}\int_z p(x,z|\theta)dz \tag{2.294}$$
$$1 = \frac{1}{c}p(x|\theta)dz \tag{2.295}$$
$$c = p(x|\theta) \tag{2.296}$$

$$\therefore q(z) = \frac{1}{p(x|\theta)} p(x, z|\theta) = p(z|x, \theta) \tag{2.297}$$

$$\tag{2.298}$$

$$\therefore \log p(x|\theta) = ELBO + p(z|x,\theta) \tag{2.299}$$

$$= E_{q(z)}\left[\log \frac{p(x,z|\theta)}{q(z)}\right] + p(z|x,\theta) \tag{2.300}$$

□

Reference: [10]

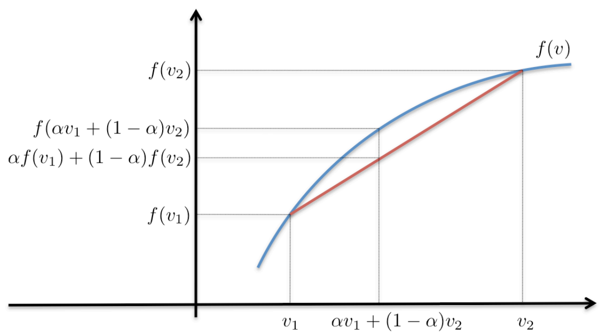

Figure 2.13: Jensen's inequality

[10] `http://willwolf.io/2018/11/11/em-for-lda/`