AI – Machine Learning

Artificial Intelligence
Research Group

Variational Inference

- Markov Chain Monte Carlo (MCMC)
- Variational Inference

$$\log\left(p(X)\right) = \log(p(X,Z)) - \log(p(Z|X)) \qquad Z\text{: Latent Variable}$$

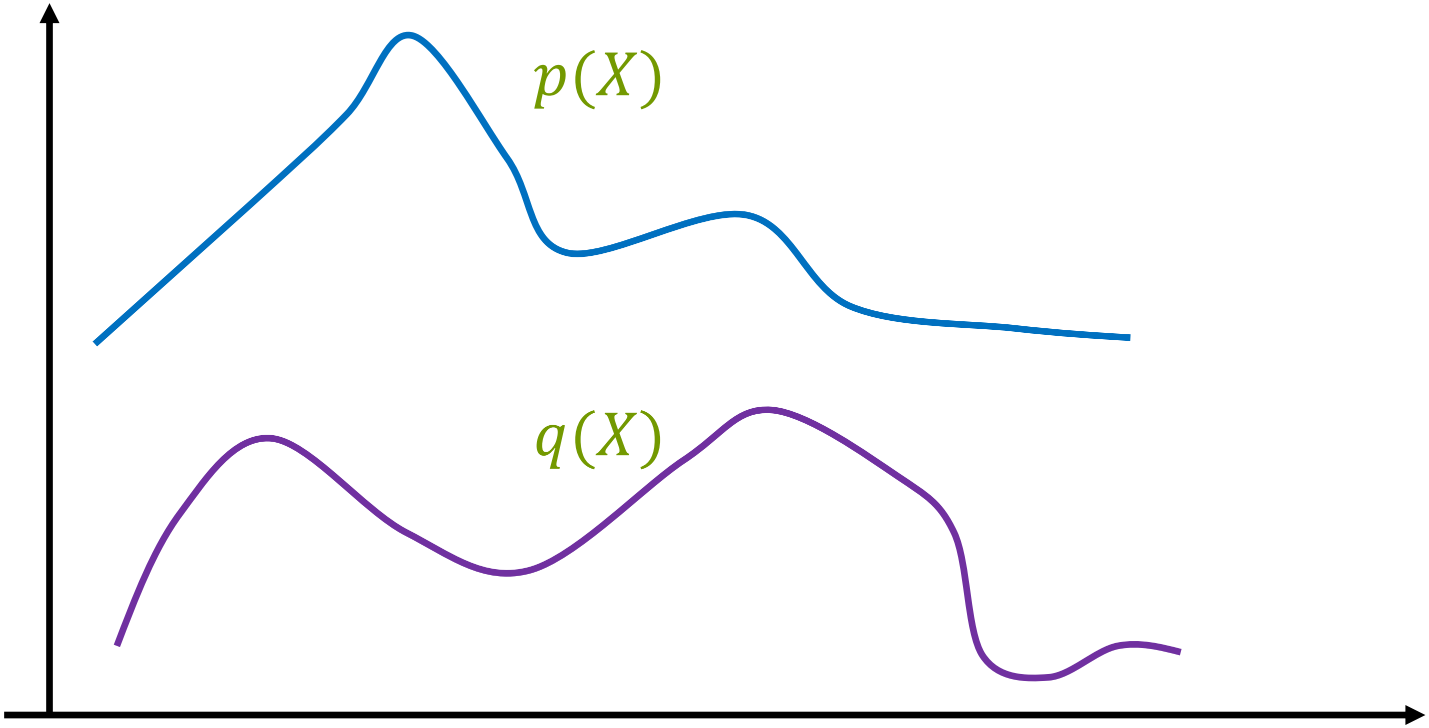
$$= \log\left(\frac{p(X,Z)}{q(Z)}\right) - \log\left(\frac{p(Z|X)}{q(Z)}\right)$$

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

$$\int_Z \log\left(p(X)\right) q(Z)dZ = \log\left(p(X)\right)\int_Z q(Z)dZ = \log(p(X))$$

$$\underbrace{\int_Z \log\left(p(X,Z)\right) q(Z)dZ - \int_Z \log\left(q(Z)\right) q(Z)dZ}_{\mathcal{L}(q)\text{: Evidence Lower Bound (EBOL)}} - \underbrace{\int_Z \log\left(\frac{p(Z|X)}{q(Z)}\right) q(Z)dZ}_{\mathrm{KL}(q(Z)\,\|\,p(Z|X))}$$

KL-divergence



$$\mathrm{KL}(p,q)=\int_X \log\left(\frac{p(X)}{q(X)}\right)p(X)\,dX$$

Evidence Lover Bound

$$\log(p(X)) = \log \int_Z p(X,Z)dZ$$

$$= \log \int_Z p(X,Z)\frac{q(Z)}{q(Z)}dZ$$

$$= \log \int_Z \frac{p(X,Z)}{q(Z)}q(Z)dZ$$

$$= \log\left(E_q\left[\frac{p(X,Z)}{q(Z)}\right]\right)$$

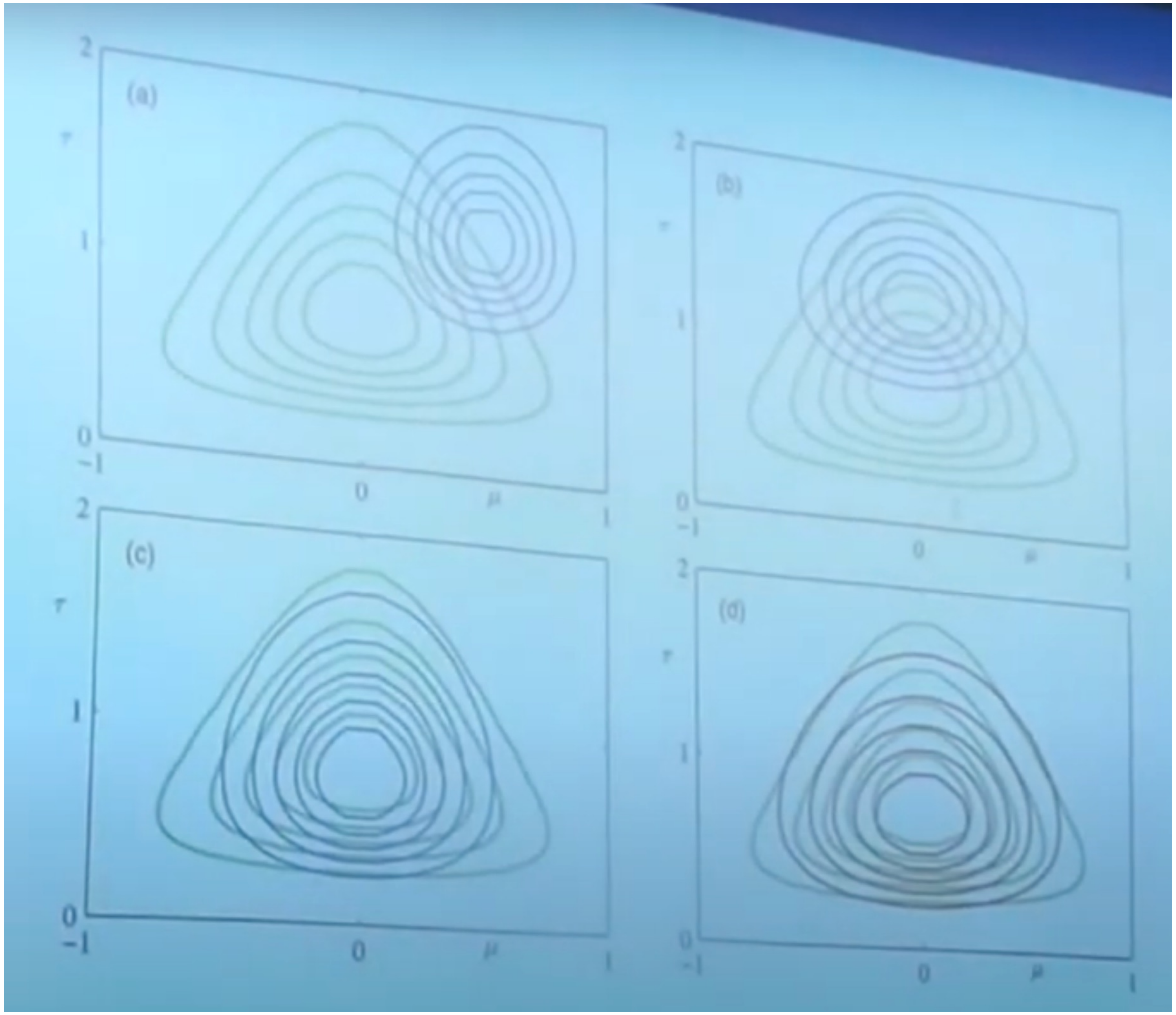
$$\geq E_q\left[\log\left(\frac{p(X,Z)}{q(Z)}\right)\right]$$ Using Jensen's inequality

$$= E_q[\log(p(X,Z))] - E_q[\log(q(Z))]$$

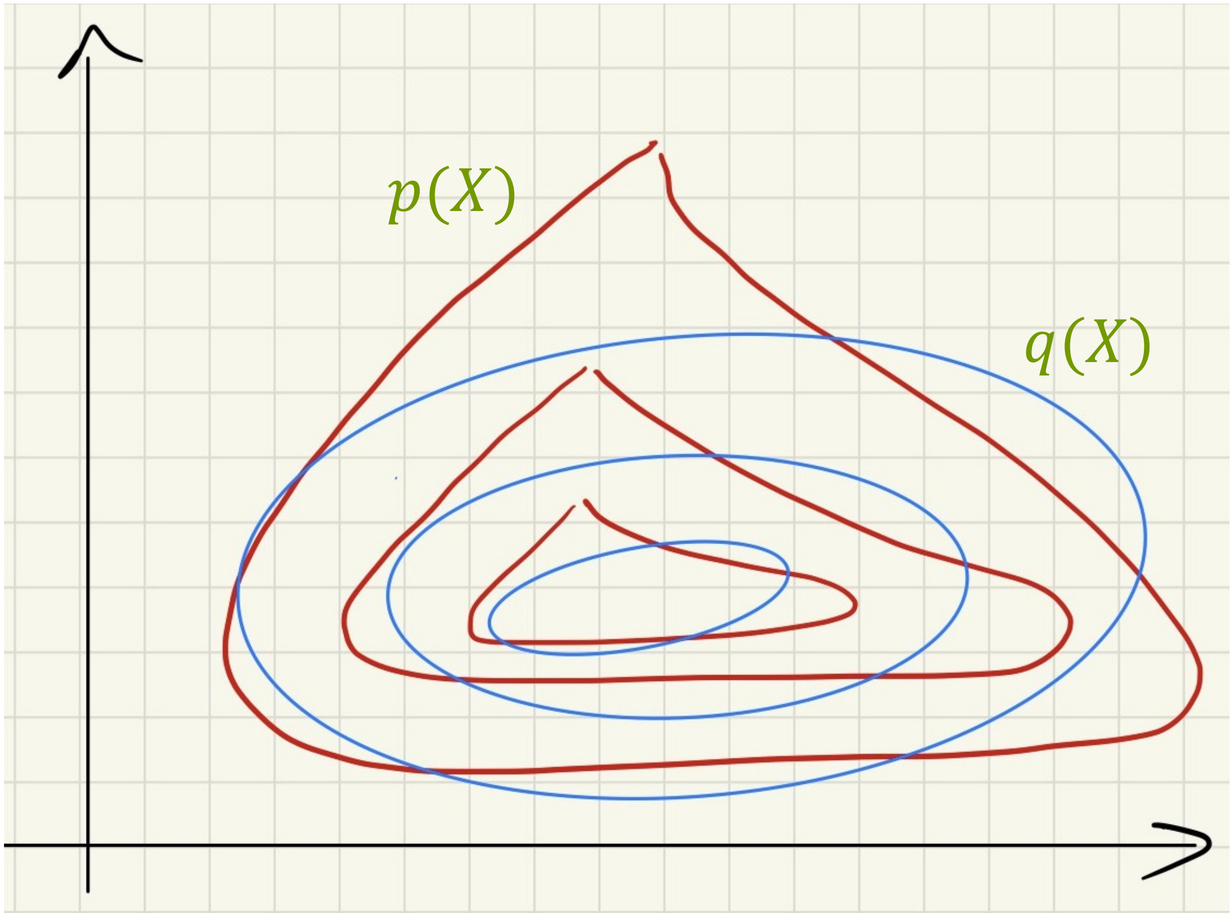
$$\triangleq \mathcal{L}(q)$$

$$\log\big(p(X)\big) - \mathcal{L}(q) = \mathrm{KL}(q \parallel p)$$

How to Choose distribution $q(X)$



How to Choose distribution $q(X)$



How to Choose distribution $q(X)$

$q(Z) \qquad Z = \{Z_1, \cdots, Z_m\}$

$$p(Z) \neq p_1(Z_1)p_2(Z_2)\cdots p_m(Z_m)$$

$$q(Z) = q_1(Z_1)q_2(Z_2)\cdots q_m(Z_m) = \prod_{i=1}^{m} q_i(Z_i)$$

$$\mathcal{L}(q) = \int_Z \log\big(p(X,Z)\big)q(Z)dZ - \int_Z \log\big(q(Z)\big)\, q(Z)dZ$$

$$= \underbrace{\int_Z \log\big(p(X,Z)\big)\prod_{i=1}^{m} q_i(Z_i)\, dZ}_{\textbf{Part (1)}} - \underbrace{\int_Z \log\left(\prod_{i=1}^{m} q_i(Z_i)\right)\prod_{i=1}^{m} q_i(Z_i)\, dZ}_{\textbf{Part (2)}}$$

Part (1) and (2)

$$\int_Z \log(p(X,Z)) \prod_{i=1}^{m} q_i(Z_i)\, dZ$$

$$= \int_{Z_1} \int_{Z_2} \cdots \int_{Z_m} \prod_{i=1}^{m} q_i(Z_i) \log(p(X,Z)) dZ_1 dZ_1 \cdots dZ_m$$

$$= \int_{Z_j} q_j(Z_j) \left(\int_{Z_{i \neq j}} \prod_{i \neq j}^{m} q_i(Z_i) \log(p(X,Z)) dZ_{i \neq j} \right) dZ_j$$

Part (1) $= \int_{Z_j} q_j(Z_j) E_{i \neq j}[\log(P(X,Z))] dZ_j$

Part (2) $= \int_Z \prod_{i=1}^{M} q_i(Z_i) \sum_{i=1}^{\mathrm{M}} \log(q_i(Z_i))\, dZ$

$$= \sum_{i=1}^{M} \int_{Z_i} q_i(Z_i) \log(q_i(Z_i))\, dZ_i$$

Part (2)

$$\int_{x_1}\int_{x_2}[f(x_1)+f(x_2)]\,p(x_1,x_2)dx_1dx_2$$

$$=\int_{x_1}\int_{x_2}f(x_1)\,p(x_1,x_2)dx_1dx_2+\int_{x_1}\int_{x_2}f(x_2)\,p(x_1,x_2)dx_1dx_2$$

$$=\int_{x_1}f(x_1)\left(\int_{x_2}p(x_1,x_2)\,dx_2\right)dx_1+\int_{x_2}f(x_2)\left(\int_{x_1}p(x_1,x_2)\,dx_1\right)dx_2$$

$$=\int_{x_1}f(x_1)p(x_1)dx_1+\int_{x_2}f(x_2)p(x_2)dx_2$$

Put All Together

For Particular $q_i(Z_i)$ Part (2) can be:

$$\text{Part (2)} = \int_{Z_j} q_j(Z_j)\log(q_j(Z_j))dZ_j + \text{const}$$

$$\mathcal{L}(q) = \text{Part(1)} - \text{Part(2)}$$

$$\mathcal{L}(q) = \int_{Z_j} q_j(Z_j)E_{i\neq j}[\log(P(X,Z))]dZ_j - \int_{Z_j} q_j(Z_j)\log(q_j(Z_j))dZ_j + \text{const}$$

$$\mathcal{L}\left(\tilde{p}_j(X,Z_j)\right) = E_{i\neq j}[\log(P(X,Z))]$$

$$\mathcal{L}(q_j) = \int_{Z_j} q_j(Z_j)\log\left[\frac{\tilde{p}_j(X,Z_j)}{q_j(Z_j)}\right] + \text{const}$$

$$\mathcal{L}(q_j^*(Z_j)) = E_{i\neq j}[\log(p(X,Z))]$$

Example

Let $X = \{x_1, \dots, x_n\}$:

$$p(X|\mu,\tau) = \prod_{i=1}^{n} \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-\tau}{2}(x_i-\mu)^2\right)$$

$$= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(\frac{-\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$p(\mu|\tau) = \mathcal{N}(\mu_0, (\lambda_0\tau)^{-1}) \propto \exp\left(\frac{-\lambda_0\tau}{2}(\mu-\mu_0)^2\right)$$

$$p(\tau) = \mathrm{Gamma}(\tau|a_0,b_0) \propto \tau^{a_0-1}\exp^{-b_0\tau}$$

Complete data-likelihood is:

$$p(X,\mu,\tau) = p(X|\mu,\tau)p(\mu|\tau)p(\tau)$$

Example

$$p(X,\mu,\tau) \propto p(X|\mu,\tau)p(\mu|\tau)p(\tau) = \mathcal{N}(\mu_n,(\lambda_n\tau)^{-1})\mathrm{Gamma}(\tau|a_n,b_n)$$

where:

$$\mu_n = \frac{\lambda_0\mu_0 + n\bar{x}}{\lambda_0 + n}$$

$$\lambda_n = \lambda_0 + n$$

$$a_n = a_0 + n/2$$

$$b_n = b_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\lambda_0 n(\bar{x}-\mu_0)^2}{2(\lambda_0+n)}$$

However, for demo purpose, we assume

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau)$$

$$\log\left(q_\mu^*(\mu)\right) = \int_\tau \log\left(p(\mu,\tau|X)\right)q_\tau(\tau)d(\tau)$$

Example

$$\log\left(q_\mu^*(\mu)\right) = \mathrm{E}_{q_\tau}[\log(p(\mu,\tau|X))]$$

$$= \mathrm{E}_{q_\tau}[\log(p(X|\mu,\tau)) + \log(p(\mu|\tau))] + \text{const}$$

$$= \mathrm{E}_{q_\tau}\left[\frac{-\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{-\lambda_0\tau}{2}(\mu-\mu_0)^2\right] + \text{const}$$

$$= -\frac{\mathrm{E}_{q_\tau}[\tau]}{2}\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const}$$

$$\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2 = (n+\lambda_0)\left(\mu - \frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}\right)^2 + \text{const}$$

$$\log\left(q_\mu^*(\mu)\right) = \mathcal{N}\left(\frac{n\bar{x}+\lambda_0\mu_0}{n+\lambda_0}, \mathrm{E}_{q_\tau}[\tau](n+\lambda_0)\right)$$

$$\log\left(q_\tau^*(\tau)\right) = \int_\mu \log\left(p(\mu,\tau|X)\right) q_\mu(\mu) d(\mu)$$

Variational Inference

$q(Z) \approx \quad p(Z|X)$

X: Data

Z: Parameters

$$q(Z) = \prod_{i=1}^{m} q_i(Z_i) \qquad Z = \{Z_1, \cdots Z_m\}$$

$$\log\left(q_j^*(Z_j)\right) = E_{i \neq j}[\log(p(X, Z))]$$

$p(Z)$ $\quad p(Z_i|Z_{-i}) \quad Z_{-i} = \{Z_1, \cdots, Z_{i-1}, Z_{i+1}, \cdots Z_m\}$

Exponential family distribution

$q(Z_i|\lambda)$

Exponential Family Distribution

$$f(x|\mu,\sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \text{ for } -\infty < x < \infty$$

$$p(x|\eta) = \underbrace{h(x)}_{x \text{ only}} \exp(\underbrace{T(x)^\intercal \eta}_{\textit{Sufficient statistics}} - \underbrace{A(\eta)}_{\eta \text{ only}})$$

$$\underset{\eta}{\operatorname{argmax}}[\log p(X|\eta)] \qquad X = \{x_1, \cdots, x_n\}$$

$$\underset{\eta}{\operatorname{argmax}}\left[\log \prod_{i=1}^{n} p(x_i|\eta)\right]$$

$$\underset{\eta}{\operatorname{argmax}}\left[\log\left\{\left(\prod_{i=1}^{n} h(x_i)\right)\exp\left(\left(\sum_{i=1}^{n} T(x_i)\right)^\intercal \eta - nA(\eta)\right)\right\}\right]$$

Exponential Family Distribution

$$\underset{\eta}{\operatorname{argmax}} \left[\log \left\{ \left(\prod_{i=1}^{n} h(x_i) \right) \exp\left(\left(\sum_{i=1}^{n} T(x_i) \right)^{\intercal} \eta - nA(\eta) \right) \right\} \right]$$

$$\underbrace{\underset{\eta}{\operatorname{argmax}} \left[\left(\sum_{i=1}^{n} T(x_i) \right)^{\intercal} \eta - nA(\eta) \right]}_{\mathcal{L}(\eta)}$$

$$\frac{\partial \mathcal{L}(\eta)}{\partial \eta} = \sum_{i=1}^{n} T(x_i) - nA'(\eta) = 0$$

$$A'(\eta) = \sum_{i=1}^{n} \frac{T(x_i)}{n}$$

Exponential Family Distribution

$$p(x|\eta) = h(x)\exp(T(x)^{\intercal}\eta - A(\eta))$$

Distribution	**Ball**	**Bucket**
Multinomial	n	$p_1, \cdots, p_k$
Binomial	n	p
Categorical	1	$p_1, \cdots, p_k$
Bernoulli	1	p

Gaussian Distribution

$$p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$p(x|\eta) = h(x)\exp(T(x)^\intercal \eta - A(\eta))$$

$$\begin{aligned}
p(x|\mu,\sigma^2) &= \exp\left(-\frac{x^2 - 2x\mu - \mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \\
&= \exp\left(\underbrace{\begin{bmatrix} x \\ x^2 \end{bmatrix}^\intercal}_{T(x)^\intercal} \underbrace{\begin{bmatrix} \frac{\mu}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{bmatrix}}_{\eta} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)}_{A(\eta)}\right)
\end{aligned}$$

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{bmatrix}$$

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2} \end{bmatrix}$$

Gaussian Distribution

$$p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\hat{\mu}_{\text{MLE}} = \underset{\mu}{\text{argmax}} \left[\sum_{i=1}^{n} \log p(x_i|\mu,\sigma^2)\right] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma^2}_{\text{MLE}} = \underset{\sigma^2}{\text{argmax}} \left[\sum_{i=1}^{n} \log p(x_i|\mu,\sigma^2)\right] = \frac{1}{n}\sum_{i=1}^{n} (x_i-\mu)^2$$

$$p(x|\eta) = \exp\left(\begin{bmatrix} x \\ x^2 \end{bmatrix}^{\intercal} \begin{bmatrix} \frac{\mu}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{bmatrix} - \left(\frac{-\eta_1^2}{4\eta_2}\right) - \frac{1}{2}\log(-2\eta_2) + \frac{1}{2}\log(2\pi)\right)$$

$$A'(\eta) = \sum_{i=1}^{n} \frac{T(x_i)}{n}$$

Gaussian Distribution

$$p(x|\eta) = \exp\left(\begin{bmatrix} x \\ x^2 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \frac{\mu}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{bmatrix} - \left(\frac{-\eta_1^2}{4\eta_2}\right) - \frac{1}{2}\log(-2\eta_2) + \frac{1}{2}\log(2\pi)\right)$$

$$A'(\eta) = \sum_{i=1}^{n} \frac{T(x_i)}{n}$$

$$\begin{bmatrix} \frac{\partial A(\eta)}{\partial \eta_1} \\ \frac{\partial A(\eta)}{\partial \eta_2} \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{n} x_i}{n} \\ \frac{\sum_{i=1}^{n} x_i^2}{n} \end{bmatrix}$$

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{-1}{2\eta_2} \end{bmatrix}$$

$$\frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} = \mu^2 + \sigma^2$$

Any questions?

AI Research Group
Fudan University