« Alireza Gargoori Motlegh _ 98102176 »

« Homework 0 _ Reinforcement Learning »

Estimator's Variance

1

$$\underset{\theta}{Bias}(w) = \underset{\theta}{\mathbb{E}}[w] - \theta \quad , \quad \underset{\theta}{Var}(w) = \underset{\theta}{\mathbb{E}}\left[(w - \underset{\theta}{\mathbb{E}}[w])^2\right]$$

$$MSE(w,\theta) = \underset{\theta}{\mathbb{E}}\left[(w-\theta)^2\right] = \underset{\theta}{\mathbb{E}}\left[(w - \underset{\theta}{\mathbb{E}}[w] + \underset{\theta}{\mathbb{E}}[w] - \theta)^2\right] =$$

$$\underset{\theta}{\mathbb{E}}\left[(w - \underset{\theta}{\mathbb{E}}[w])^2 + 2(w - \underset{\theta}{\mathbb{E}}[w])(\underset{\theta}{\mathbb{E}}[w] - \theta) + (\underset{\theta}{\mathbb{E}}[w] - \theta)^2\right] =$$

$$\boxed{\underset{\theta}{\mathbb{E}}\left[(w - \underset{\theta}{\mathbb{E}}[w])^2\right]} + 2\boxed{\underset{\theta}{\mathbb{E}}\left[(w - \underset{\theta}{E}[w])(\underset{\theta}{E}[w] - \theta)\right]} + \boxed{\underset{\theta}{\mathbb{E}}\left[(\underset{\theta}{\mathbb{E}}[w] - \theta)^2\right]} =$$

$\underset{\theta}{Var}(w)$ ★ ★★

$$\star \quad \underset{\theta}{\mathbb{E}}\left[(w - \underset{\theta}{E}[w])(\underset{\theta}{\mathbb{E}}[w] - \theta) \,|\, \theta\right] = (\underset{\theta}{\mathbb{E}}[w] - \theta)\,\underset{\theta}{\mathbb{E}}\left[w - \underset{\theta}{\mathbb{E}}[w]\right] =$$

$$(\underset{\theta}{\mathbb{E}}[w] - \theta)\Big(\underset{\theta}{\mathbb{E}}[w] - \underbrace{\underset{\theta}{\mathbb{E}}\big(\underset{\theta}{\mathbb{E}}[w]\big)}_{= \underset{\theta}{\mathbb{E}}[w]}\Big) = (\underset{\theta}{\mathbb{E}}[w] - \theta) \times 0 = 0$$

$$\star\star \quad \mathop{\mathbb{E}}_{\theta}\left[\left(\mathop{\mathbb{E}}_{\theta}[w]-\theta\right)^2\right] = \mathbb{E}\left[\left(\mathop{\mathbb{E}}_{\theta}[w]-\theta\right)^2 \mid \theta\right] = \left(\mathop{\mathbb{E}}_{\theta}[w]-\theta\right)^2 = \left(\mathop{Bias}_{\theta}(w)\right)^2$$

$$\xrightarrow[\star\star]{\star} \quad \boxed{MSE(w,\theta) = \mathop{\mathbb{E}}_{\theta}\left[(w-\theta)^2\right] = \mathop{Var}_{\theta}(w) + \left(\mathop{Bias}_{\theta}(w)\right)^2}$$

So, to have a low error (low MSE), we must have a low variance (good precision) & low bias (good accuracy), which is quite not possible because of bias-variance tradeoff.

2 $$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \quad , \quad X_i \overset{iid}{\sim} \mathbb{P}(X) \quad , \quad \theta = \mathbb{E}[f(X)]$$

$$\mathop{Bias}_{\theta}\left(\hat{\theta}_n\right) = \mathop{\mathbb{E}}_{\theta}\left[\hat{\theta}_n\right] - \theta = \mathop{\mathbb{E}}_{\theta}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right] - \theta \overset{\text{linearity}}{=}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta}\left[f(X_i)\right] - \theta \overset{iid}{=} \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\theta}\left[f(X)\right] - \theta =$$

$$\frac{1}{n} \times (n\theta) - \theta = \theta - \theta = 0 \implies \boxed{\hat{\theta}_n \text{ is an unbiased estimator for } \theta.}$$

3 $f_1(x) = 1 + \left(\frac{x}{2}\right)^2$, $f_2(x) = \left(\frac{x}{4}\right)^{10}$, $X \sim$ Uniform $[0, 4]$

$$\mathbb{E}_p[f_1(X)] = \int_{-\infty}^{\infty} f_1(x)\, f_X(x)\, dx = \frac{1}{4}\int_0^4 \left(1 + \left(\frac{x}{2}\right)^2\right) dx = \frac{7}{3} \approx 2.33$$

$$\mathbb{E}_p[f_2(X)] = \int_{-\infty}^{\infty} f_2(x)\, f_X(x)\, dx = \frac{1}{4}\int_0^4 \left(\frac{x}{4}\right)^{10} dx = \frac{1}{11} \approx 0.091$$

The code I wrote produced: $\hat{\theta}_1 = 2.144$, $\hat{\theta}_2 = 0.047$

$\text{Bias}_{\theta_1}(\hat{\theta}_1) = -0.19$ $\quad$ $\text{Bias}_{\theta_2}(\hat{\theta}_2) = -0.044$

$\text{Var}_{\theta_1}(\hat{\theta}_1) = 1.003$ $\quad$ $\text{Var}_{\theta_2}(\hat{\theta}_2) = 0.017$

As we can see, the bias and variance of this estimator is greater (in magnitude) for $f_1(x)$ than $f_2(x)$. The reason is that $f_1(x)$ is greater than $f_2(x)$ in magnitude and hence, its estimation has a larger bias & variance; since Monte Carlo estimator's variance is propotional to the magnitude of the estimating function.

4 $$\hat{\theta} = \mathop{\mathbb{E}}_{q}\left[\frac{p(a)}{q(a)} * f(a)\right] = \sum_a q(a)\frac{p(a)}{q(a)} f(a) = \sum_a p(a) f(a) = \mathop{\mathbb{E}}_{p}[f(a)]$$

$$\Longrightarrow \boxed{\text{Bias}(\hat{\theta}) = \underbrace{\mathop{\mathbb{E}}_{q}\left[\frac{p(a)}{q(a)} * f(a)\right]}_{\hat{\theta}} - \underbrace{\mathop{\mathbb{E}}_{p}[f(a)]}_{\theta} = 0 \quad : \text{unbiassedness}}$$

($q(a) = 0$ must imply $p(a) f(a) = 0$)

$$\hat{\theta}_q = \frac{1}{n}\sum_{i=1}^{n}\frac{p(X_i) f(X_i)}{q(X_i)} \quad , \quad \hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

$$\mathop{\text{Var}}_{q}(\hat{\theta}_q) = \mathop{\mathbb{E}}_{q}\left[\left(\frac{p(x)}{q(x)} f(x)\right)^2\right] - \left(\mathop{\mathbb{E}}_{q}\left[\frac{p(a)}{q(a)} * f(a)\right]\right)^2 \Longrightarrow$$

$$\boxed{\mathop{\text{Var}}_{q}(\hat{\theta}_q) = \sum_a q(a)\left(\frac{p(a)}{q(a)} f(a)\right)^2 - \theta^2}$$

$$\mathop{\text{Var}}_{p}(\hat{\theta}_n) = \mathop{\mathbb{E}}_{p}[f^2(x)] - \left(\mathop{\mathbb{E}}_{p}[f(x)]\right)^2 = \sum_a p(a) f^2(a) - \theta^2 \Longrightarrow$$

$$\boxed{\mathop{\text{Var}}_{p}(\hat{\theta}_n) = \sum_a p(a) f^2(a) - \theta^2}$$

So, the reduction in the variance is:

$$\underset{p}{Var}\left(\hat{\theta}_n\right) - \underset{q}{Var}\left(\hat{\theta}_q\right) = \sum_x f^2(x)p(x) - \sum_x f^2(x)\frac{p^2(x)}{q(x)} = \sum_x f^2(x)\left(1 - \frac{p(x)}{q(x)}\right)p(x)$$

$$\Rightarrow \boxed{\text{Reduced Var} = \sum_x f^2(x)\left(1 - \frac{p(x)}{q(x)}\right)p(x)}$$

5. $$\text{minimize} \quad \sum_x f^2(x)\left(1 - \frac{p(x)}{q(x)}\right)p(x)$$

$$\text{s.t.} \quad \sum_x q(x) = 1$$

$$\rightarrow \mathcal{L}(q,\lambda) = \sum f^2(x)\left(1 - \frac{p(x)}{q(x)}\right)p(x) + \lambda\left(\sum_x q(x) - 1\right)$$

$$\frac{\partial \mathcal{L}(q,\lambda)}{\partial q} = 0 \Rightarrow f^2(x)\frac{p^2(x)}{q^2(x)} + \lambda = 0 \Rightarrow -\lambda q^2(x) = f^2(x)p^2(x)$$

$$\rightarrow q^2(x) = \frac{-1}{\lambda} f^2(x)p^2(x) \xrightarrow{q(x)\geq 0} q^*(x) = \frac{1}{|\lambda|}|f(x)|p(x) \rightarrow$$

$$\boxed{q^*(x) \propto |f(x)|p(x)} \quad , \quad \lambda = \sum_x |f(x)|p(x)$$

We prove no other q has a lower variance then q^*:

$$q^*(x) = \frac{1}{\lambda} |f(x)| p(x) \quad , \quad \lambda = \sum_x |f(x)| p(x)$$

$$\underset{q^*}{\text{Var}} = \sum_x \frac{f^2(x) p^2(x)}{q^*(x)} - \theta^2 = \lambda \sum_x \frac{f^2(x) p^2(x)}{|f(x)| p(x)} - \theta^2$$

$$= \lambda \sum_x |f(x)| p(x) - \theta^2 = \left[\sum_x |f(x)| p(x) \right]^2 - \theta^2$$

$$= \left[\sum_x \frac{|f(x)| p(x)}{q(x)} q(x) \right]^2 - \theta^2 \leq \sum_x \frac{f^2(x) p^2(x)}{q^2(x)} q(x) - \theta^2$$

Cauchy Schwarz w.r.t. $q(x)$ (also $\sum_x q^2 \leq \sum_x q = 1$)

$$= \sum_x \frac{f^2(x) p^2(x)}{q(x)} - \theta^2 = \underset{q}{\text{Var}}$$

6 $\quad p = \mathbb{P}(X > 5) \quad , \quad X \sim N(0,1) \quad , \quad Q(x) = \begin{cases} e^{-(x-5)} & x \geq 5 \\ 0 & \text{o.w.} \end{cases}$

$$\underset{p}{\mathbb{E}}[f(x)] = \underset{p}{\mathbb{E}}[I_{x>5}] = \int_{-\infty}^{\infty} I_{x>5} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int_5^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx = 2.87 \times 10^{-7}$$

$$\underset{Q}{\mathbb{E}}\left[\frac{P(x)}{Q(x)} I_{x>5}\right] = \int_5^{\infty} e^{-(x-5)} \frac{\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}}{e^{-(x-5)}} dx = \underset{P}{\mathbb{E}}\left[I_{x>5}\right] = 2.87 \times 10^{-7}$$

$$\underset{P}{Var}\left(I_{x>5}\right) = \underset{P}{\mathbb{E}}\left[I^2_{x>5}\right] - \left(\underset{P}{\mathbb{E}}\left[I_{x>5}\right]\right)^2 = \underset{P}{\mathbb{E}}\left[I_{x>5}\right] - \underset{P}{\mathbb{E}}^2\left[I_{x>5}\right] =$$

$$2.87 \times 10^{-7} - 8.24 \times 10^{-14} \sim 10^{-7}$$

$$\underset{Q}{Var}\left(\frac{P(x)}{Q(x)} I_{x>5}\right) = \underset{Q}{\mathbb{E}}\left[\left(\frac{P(x)}{Q(x)} I_{x>5}\right)^2\right] - \left(\underset{Q}{\mathbb{E}}\left[\frac{P(x)}{Q(x)} I_{x>5}\right]\right)^2 =$$

$$\int_{-\infty}^{\infty} Q(x) \frac{P^2(x)}{Q^2(x)} I^2_{x>5} dx - 8.24 \times 10^{-14} =$$

$$\int_5^{\infty} \frac{\frac{1}{2\pi} e^{-x^2}}{e^{-(x-5)}} dx - 8.24 \times 10^{-14} = 2.40 \times 10^{-13} - 8.24 \times 10^{-14} \sim 10^{-13}$$

As we can see, variance of the new estimator (importance sampling estimator) is about 10^{-13}, while the simple Monte Carlo estimator's variance is about 10^{-7}; So the variance is reduced significantly.

Markov Chain

1 $$\mathbb{P}(X_n=s_n, X_{n-1}=s_{n-1}, \dots, X_0=s_0) = \mathbb{P}(X_0=s_0) \times \mathbb{P}(X_1=s_1 \mid X_0=s_0) \times$$

$$\mathbb{P}(X_2=s_2 \mid X_1=s_1, X_0=s_0) \times \dots \times \mathbb{P}(X_n=s_n \mid X_{n-1}=s_{n-1}, X_{n-2}=s_{n-2}, \dots, X_0=s_0)$$

markov property
$$\overset{*}{=} \mathbb{P}(X_0=s_0) \times \mathbb{P}(X_1=s_1 \mid X_0=s_0) \times \mathbb{P}(X_2=s_2 \mid X_1=s_1) \times \dots \times \mathbb{P}(X_n=s_n \mid X_{n-1}=s_{n-1})$$

$$= \mathbb{P}(X_0=s_0) \times \underbrace{P \times P \times \dots \times P}_{n} = P^{(0)} P^{(n)}$$

2 $$\mathbb{P}(X_{t_1+t_2} \mid X_{F_1}, \dots, X_{F_n}) \overset{\text{Bayes}}{=} \frac{\mathbb{P}(X_{t_1+t_2}, X_{F_1}, X_{F_2}, \dots, X_{F_n})}{\mathbb{P}(X_{F_1}, X_{F_2}, \dots, X_{F_n})} =$$

$$\frac{P^{(0)} \times P^{(\min(F))} \times P^{(\max(F)-\min(F))} \times P^{(t_1+t_2)-\max(F)}}{P^{(0)} \times P^{(\min(F))} \times P^{(\max(F)-\min(F))}} = P^{(t_1+t_2)-\max(F)}$$

$$\Longrightarrow \boxed{\mathbb{P}(X_{t_1+t_2} \mid X_{F_1}, \dots, X_{F_n}) = \mathbb{P}(X_{t_1+t_2} \mid X_{\max(F)})}$$

3 $$P^{(n+m)} = \left[p^{(n+m)} \right]_{ij} \quad \text{where} \quad p_{ij}^{(n+m)} = \mathbb{P}\left(X_{m+n} = j \mid X_0 = i \right)$$

$$p_{ij}^{(n+m)} = \mathbb{P}\left(X_{n+m} = j \mid X_0 = i \right) \overset{*}{=} \sum_k \mathbb{P}\left(X_{m+n} = j, X_m = k \mid X_0 = i \right)$$

* finite n-states

$$= \sum_k \mathbb{P}\left(X_{m+n} = j \mid X_m = k, X_0 = i \right) \mathbb{P}\left(X_m = k \mid X_0 = i \right)$$

Chain rule

$$= \sum_k \mathbb{P}\left(X_{m+n} = j \mid X_m = k \right) \mathbb{P}\left(X_m = k \mid X_0 = i \right)$$

Markov property

$$= \sum_k p_{kj}^{(n)} p_{ik}^{(m)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)} \implies \boxed{P^{(m+n)} = P^{(m)} P^{(n)}}$$

4 $$P^{(n)} = P^n = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n \end{bmatrix}^n = \begin{bmatrix} p_1^{(n)} & 0 & \cdots & 0 \\ 0 & p_2^{(n)} & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n^{(n)} \end{bmatrix}$$

$$P^{(0)^T} = \underbrace{\begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}}_{W} \begin{pmatrix} p_1^{(0)} \\ p_2^{(0)} \\ \vdots \\ p_n^{(0)} \end{pmatrix} = \begin{bmatrix} w_1 p_1^{(0)} \\ w_2 p_2^{(0)} \\ \vdots \\ w_n p_n^{(0)} \end{bmatrix}$$

$W :=$ weight matrix

$$\text{trace}(W) = \sum_{i=1}^{n} w_i = 1$$

$$\pi_{ss} = \lim_{n\to\infty} P^{(0)} P^{(n)} = \begin{bmatrix} w_1 P_1^{(0)} \\ w_2 P_2^{(0)} \\ \vdots \\ w_n P_n^{(0)} \end{bmatrix}^T \begin{bmatrix} P_1^{(n)} & 0 & \cdots & 0 \\ 0 & P_2^{(n)} & 0 \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & P_n^{(n)} \end{bmatrix} = \begin{bmatrix} w_1 P_1^{(0)} P_1^{(n)} \\ w_2 P_2^{(0)} P_2^{(n)} \\ \vdots \\ w_n P_n^{(0)} P_n^{(n)} \end{bmatrix}$$

$$\Longrightarrow \quad \Pi_{ss} = \begin{bmatrix} w_1 \pi_1 \\ w_2 \pi_2 \\ \vdots \\ w_n \pi_n \end{bmatrix} \quad , \quad \sum_{i=1}^{n} w_i = 1$$

So any convex combination of individual steady-states could be the steady-state for the combination of n independent markov chains; hence, the Π_{ss} distribution is not unique.

The long run behaviour of this markov chain would be determined by the coefficients of w; which indicate the importance (weight) of each individual markov chain. Therefore, $\Pi_{ss} \sim \Pi_{\text{argmax}(w)}$

5

$$\pi_1(j) = \lim_{n\to\infty} P^{(n)}_{i,j} \quad \forall i,j$$

- $$\pi_1(j) = \lim_{n\to\infty} P^{(n)}_{i,j} \overset{*3}{=} \lim_{n\to\infty} \sum_k P^{(n-1)}_{i,k} P_{k,j} = \sum_k \lim_{n\to\infty} P^{(n-1)}_{i,k} P_{k,j}$$

$$= \sum_k \pi_1(k) P_{k,j} \Longrightarrow \boxed{\pi_1(j) = \sum_k \pi_1(k) P_{k,j}, \ \forall j}$$

which is equivalent to the stationary state : $\pi = \pi P$

$\Longrightarrow$ The limiting distribution is always a stationary distribution.

- $$\pi_1(j) = \sum_k \pi_1(k) P_{k,j} \Longrightarrow \pi^{(n)}_2(j) = \sum_k \pi^{(0)}_2(k) P^{(n)}_{k,j}$$

$$\longrightarrow \lim_{n\to\infty} \pi^{(n)}_2(j) = \lim_{n\to\infty} \sum_k \pi^{(0)}_2(k) P^{(n)}_{k,j} = \sum_k \underbrace{\lim_{n\to\infty} P^{(n)}_{k,j}}_{=\pi_2(j)} \pi^{(0)}_2(k)$$

$$\overset{\text{definition}}{=} \sum_k \pi_2(j) \pi^{(0)}_2(k) = \pi_2(j) \underbrace{\sum_k \pi^{(0)}_2(k)}_{=1} = \pi_2(j) \Longrightarrow \boxed{\pi^{(n)}_2(j) = \pi_2(j)}$$

$\longrightarrow$ $\pi^{(0)}$ has no effect in limiting distribution on long run behaviour.

Information Theory

1 $$H(X) = -\sum_{x} P(x) \log P(x) = -\mathop{\mathbb{E}}_{X}[\log P(X)]$$

$$= \mathop{\mathbb{E}}_{X}[-\log P(X)] = \mathop{\mathbb{E}}_{X}\left[\log \frac{1}{P(X)}\right] \geq 0 \quad \text{since} \quad \log \frac{1}{P(X)} \geq 0$$

for $0 \leq P(X) \leq 1$

- equality holds iff $\log \frac{1}{P(X)} = 0$ with probability 1 ; therefore X must be deterministic.

2 Since we want the maximum uncertainty ($\equiv$ maximum entropy),

X must have a uniform distribution.

$X \sim \text{Uniform}\left(\frac{1}{M}\right)$, where M is the number of outcomes of X R.V. :

$$\Rightarrow H(X) = -\sum_{x} P(x) \log P(X) = -\sum_{x} \frac{1}{M} \log \frac{1}{M} = -M \times \frac{1}{M} \times \log \frac{1}{M} = \log M$$

$$\Rightarrow \boxed{H(X) = \log M}$$

3 $$H(X|Y) = -\sum_{x,y} \mathbb{P}(x,y) \log \mathbb{P}(x|y)$$

- $$H(X,Y) = -\sum_{x,y} \mathbb{P}(x,y) \log \mathbb{P}(x,y) = -\sum_{x,y} \mathbb{P}(x,y) \log \left(\mathbb{P}(x|y)\mathbb{P}(y) \right) =$$

$$-\sum_{x,y} \mathbb{P}(x,y) \left(\log \mathbb{P}(x|y) + \log \mathbb{P}(y) \right) = -\sum_{x,y} \mathbb{P}(x,y) \log \mathbb{P}(x|y) -$$

$$\sum_{x,y} \mathbb{P}(x,y) \log \mathbb{P}(y) = H(X|Y) - \sum_{y} \log \mathbb{P}(y) \underbrace{\sum_{x} \mathbb{P}(x,y)}_{\mathbb{P}(y)} =$$

$$H(X|Y) - \sum_{y} \mathbb{P}(y) \log \mathbb{P}(y) = H(X|Y) + H(Y)$$

- $$H(X,Y) = -\sum_{x,y} \mathbb{P}(x,y) \log \left(\mathbb{P}(y|x)\, \mathbb{P}(x) \right) = -\sum_{x,y} \mathbb{P}(x,y) \log \mathbb{P}(y|x) -$$

$$\sum_{x} \sum_{y} \mathbb{P}(x,y) \log \mathbb{P}(x) = H(Y|X) - \sum_{x} \log \mathbb{P}(x) \underbrace{\sum_{y} \mathbb{P}(x,y)}_{\mathbb{P}(x)} =$$

$$H(Y|X) - \sum_{x} \mathbb{P}(x) \log \mathbb{P}(x) = H(Y|X) + H(X)$$

4 $$D_{KL}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Jensen inequality : $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$ if $g(\cdot)$ is convex .

$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$ if $g(\cdot)$ is concave .

$$-D_{KL}(P\|Q) = -\sum_x P(x) \log \frac{P(x)}{Q(x)} = \sum_x P(x) \log \frac{Q(x)}{P(x)} \leq$$

($\log(\cdot)$ is concave)

$$\log \sum_x P(x) \frac{Q(x)}{P(x)} = \log \sum_x Q(x) = -\log 1 = 0$$

$$\Rightarrow -D_{KL}(P\|Q) \leq 0 \Longrightarrow \boxed{D_{KL}(P\|Q) \geq 0}$$

Jensen inequality is equality iff $g(\cdot)$ is affine or X is constant .

$$\Longrightarrow 0 = D_{KL}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \geq 0 \xrightarrow[\text{affine}]{\log \text{ is not}} \frac{P(x)}{Q(x)} = \text{Const.} = 1$$

(since if $P(x) = c\,Q(x)$, due to the $\sum_x P(x) = 1 = c\sum_x Q(x) = c \Longrightarrow c = 1$)

$$\Rightarrow \boxed{P(x) = Q(x)}$$

5 $$H(X) = -\sum_x \mathbb{P}(x) \log_2 \mathbb{P}(x) = -\mathop{\mathbb{E}}_X [\log \mathbb{P}(X)] = \mathop{\mathbb{E}}_X \left[\log \frac{1}{\mathbb{P}(X)}\right]$$

log is a concave function ; Jensen inequality :

$$\mathop{\mathbb{E}}_X \left[\log \frac{1}{\mathbb{P}(X)}\right] \leqslant \log \mathop{\mathbb{E}}_X \left[\frac{1}{\mathbb{P}(X)}\right] = \log |\mathcal{X}|$$ where $|\mathcal{X}|$ is the number of states of X.

$$\Rightarrow \boxed{H(X) \leqslant \log |\mathcal{X}|}$$

Another way :

$$U(x) := \frac{1}{|\mathcal{X}|}, \quad \forall x \in X$$

$$H(\mathbb{P}) = -\sum_x \mathbb{P}(x) \log \mathbb{P}(x) =$$

$$-\sum_x \mathbb{P}(x) \log \mathbb{P}(x) + \sum_x \mathbb{P}(x) \log u(x) - \sum_x \mathbb{P}(x) \log u(x)$$

$$= -D_{KL}(\mathbb{P} \| U) - \sum_x \mathbb{P}(x) \log \frac{1}{|\mathcal{X}|}$$

$$= -D_{KL}(\mathbb{P} \| U) + \log |\mathcal{X}| \sum_x \mathbb{P}(x) = \log |\mathcal{X}| - D_{KL}(\mathbb{P} \| U)$$

$$\Rightarrow H(\mathbb{P}) = \log |\mathcal{X}| - \underbrace{D_{KL}(\mathbb{P} \| U)}_{\geqslant 0} \leqslant \log |\mathcal{X}|$$

equality : $D_{KL}(\mathbb{P} \| U) = 0 \Longrightarrow$ $\boxed{\mathbb{P}(x) = u(x) = \frac{1}{|\mathcal{X}|} : \mathbb{P} \text{ is uniform}}$

6 We want to show that $H(X|Y) \leqslant H(X)$

$$I(X;Y) = D_{KL}\left(\mathbb{P}(X,Y) \,\|\, \mathbb{P}(X)\mathbb{P}(Y)\right) = \sum_{x,y} \mathbb{P}(x,y) \log \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)}$$

$$= \sum_{x,y} \mathbb{P}(x,y) \log \frac{\mathbb{P}(x|y)\mathbb{P}(y)}{\mathbb{P}(x)\,\mathbb{P}(y)} = \sum_{x,y} \mathbb{P}(x,y) \log \frac{\mathbb{P}(x|y)}{\mathbb{P}(x)} =$$

$$\sum_{x,y} \mathbb{P}(x,y) \log \mathbb{P}(x|y) - \sum_{x,y} \mathbb{P}(x,y) \log \mathbb{P}(x) = -H(X|Y) - \sum_{x} \log \mathbb{P}(x) \sum_{y} \mathbb{P}(x,y)$$

$$= -H(X|Y) - \sum_{x} \mathbb{P}(x) \log \mathbb{P}(x) = H(X) - H(X|Y)$$

Also, since $I(X;Y)$ is a KL divergence, it is $\geqslant 0$ (as proved in 4):

$$I(X;Y) = H(X) - H(X|Y) \geqslant 0 \implies$$

$$\boxed{H(X) \geqslant H(X|Y)}$$

7 ▪ $I(X,Y;Z) \geq I(X;Z)$ $= H(X|Z) + H(Y|X,Z)$

$$I(X,Y;Z) = \underbrace{H(X,Y)}_{= H(X) + H(Y|X)} - H(X,Y|Z) =$$

$$\underbrace{\left[H(X) - H(X|Z)\right]}_{I(X;Z)} + \underbrace{\left[H(Y|X) - H(Y|X,Z)\right]}_{I(Z;Y|X) \geq 0} \Longrightarrow$$

$$I(X,Y;Z) - I(X;Z) \geq 0 \Longrightarrow \boxed{I(X,Y;Z) \geq I(X;Z)}$$

equality : $I(Z;Y|X) = 0 \Rightarrow D_{KL}\left(\mathbb{P}(Z,Y|X) \,\|\, \mathbb{P}(Z|X)\mathbb{P}(Y|X)\right) = 0$

$$\Rightarrow \boxed{\mathbb{P}(Z,Y|X) = \mathbb{P}(Z|X)\,\mathbb{P}(Y|X)}$$

▪ $H(X,Y|Z) \geq H(X|Z)$

$$H(X,Y|Z) \underset{\text{chain rule}}{=} H(X|Z) + \underbrace{H(Y|X,Z)}_{\geq 0} \geq H(X|Z)$$

equality : $H(Y|X,Z) = 0 \Longrightarrow \boxed{Y = f(X,Z)}$

Bonus

8

- $I(X;Y) = 0, \quad I(X;Y|Z) = 1$

$I(X;Y) = 0 \Longrightarrow H(X) = H(X|Y) \Longrightarrow X$ & Y are independent.

$I(X;Y,Z) = \cancel{I(X;Y)}^{0} + I(X;Z|Y) = \underbrace{I(X;Z)}_{*=0} + \cancel{I(X;Y|Z)}^{1}$

* For simplicity, we assume $I(X;Z) = 0 \Longrightarrow X$ & Z are independent.

$\Rightarrow X \sim \text{Bernulli}(1/2), \; Z \sim \text{Bernulli}(1/2), \; Y = \begin{cases} X & \text{if } Z=0 \\ 1-X & \text{if } Z=1 \end{cases}$

Z \ xy	00	01	10	11
0	1/4	0	0	1/4
1	0	1/4	1/4	0

« example taken from wikipedia »

- $I(X;Y|Z) = 0 \Longrightarrow (X,Y)$ is independent of Z.

$I(X;Y) = 1 \Longrightarrow \sum_{a,y} p(a,y) \log \frac{p(a,y)}{p(a)p(y)} = 1$

Z \ xy	00	01	10	11
0	0	1/4	1/4	0
1	1/4	0	0	1/4

Probabilistic Models & Latent Variables

1 $p \sim$ true (unknown) distribution , $q_\theta \sim$ our estimation of p

$$D_{KL}(p \| q) = \mathop{\mathbb{E}}_{x \sim p}\left[\log \frac{p(x)}{q_\theta(x)}\right] = \mathop{\mathbb{E}}_{x \sim p}[\log p(x)] - \mathop{\mathbb{E}}_{x \sim p}[\log q_\theta(x)]$$

$$\arg\min_\theta D_{KL}(p \| q) = \arg\min_\theta \Big(\underbrace{\mathop{\mathbb{E}}_{x \sim p}[\log p(x)]}_{\text{does not depend on } \theta} - \mathop{\mathbb{E}}_{x \sim p}[\log q_\theta(x)] \Big)$$

$$= \arg\min_\theta - \mathop{\mathbb{E}}_{x \sim p}[\log q_\theta(x)] = \arg\max_\theta \mathop{\mathbb{E}}_{x \sim p}[\log q_\theta(x)]$$

Also, by using the Law of large numbers, we can say:

$$\arg\max_\theta \mathop{\mathbb{E}}_{x \sim p}[\log q_\theta(x)] = \arg\max_\theta \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \log q_\theta(x_i) =$$

$$\arg\max_\theta \log \prod_{i=1}^{n} q_\theta(x_i) = \arg\max_\theta \log q(x|\theta) = \arg\max_\theta q(x|\theta) = \hat{\theta}_{ML}$$

$$\Longrightarrow \boxed{\arg\min_\theta D_{KL}(p \| q_\theta) = \arg\max_\theta q(x|\theta)}$$

2 $$l(\theta; X) = \log \mathbb{P}_\theta(X) = \log \sum_z \mathbb{P}_\theta(X, z) = \log \sum_z \frac{\mathbb{P}_\theta(X, z)}{Q(z)} Q(z) =$$

$$\log \mathop{\mathbb{E}}_{z \sim Q} \left[\frac{\mathbb{P}_\theta(X, z)}{Q(z)} \right] \overset{*}{\geqslant} \mathop{\mathbb{E}}_{z \sim Q} \left[\log \frac{\mathbb{P}_\theta(X, z)}{Q(z)} \right]$$

* inequality holds since log is concave (Jensen inequality)

$$\mathop{\mathbb{E}}_{z \sim Q} \left[\log \frac{\mathbb{P}_\theta(X, z)}{Q(z)} \right] = \mathop{\mathbb{E}}_{z \sim Q} \left[\log \mathbb{P}_\theta(X, z) \right] - \mathop{\mathbb{E}}_{z \sim Q} \left[\log Q(z) \right] =$$

$$\mathop{\mathbb{E}}_{z \sim Q} \left[\log \mathbb{P}_\theta(X, z) \right] + H(Q) = \mathcal{L}(\theta, Q)$$

$$\Longrightarrow \boxed{l(\theta, X) = \log \mathbb{P}_\theta(X) \geqslant \mathcal{L}(\theta, Q) = \mathop{\mathbb{E}}_{z \sim Q} \left[\log \mathbb{P}_\theta(X, z) \right] + H(Q)}$$

So, we have found a lower bound for log-likelihood parameterized by a distribution $Q(z)$ over z (latent variables).

3 For a given θ, to minimize the distance between log-likelihood and lower bound, we must maximize $\mathcal{L}(\theta, Q)$:

(Constraint for a valid $Q(z)$)

$$\max \; \mathcal{L}(Q;\theta)$$
$$\text{s.t.} \; \sum_z Q(z) = 1$$

$$\Longrightarrow \text{Lagrange multipliers: } h(Q,\lambda) = \mathcal{L}(Q;\theta) + \lambda\left(\sum_z Q(z) - 1\right)$$

$$= \mathop{\mathbb{E}}_{z\sim Q}[\log P_\theta(x,z)] - \mathop{\mathbb{E}}_{z\sim Q}[\log Q(z)] + \lambda\left(\sum_z Q(z) - 1\right) =$$

$$\sum_z Q(z)\log P_\theta(x,z) - \sum_z Q(z)\log Q(z) + \lambda\left(\sum_z Q(z) - 1\right)$$

$$\frac{\partial h(Q,\lambda)}{\partial Q(z)} = 0 \Longrightarrow \sum_z \log P_\theta(x,z) - \left(\sum_z \log Q(z) + \sum_z \frac{Q(z)}{Q(z)}\right)$$

$$+ \lambda \sum_z 1 = 0 \xrightarrow[\text{out}]{\text{take } \Sigma} \log P_\theta(x,z) - \log Q(z) - 1 + \lambda = 0 \Longrightarrow$$

$$\log Q(z) = \log \mathbb{P}_\theta(x,z) + \lambda - 1 \implies$$

$$Q(z) = e^{\lambda-1}\, \mathbb{P}_\theta(x,z) \implies Q(z) \propto \mathbb{P}_\theta(x,z) = \mathbb{P}_\theta(z|x)\, \mathbb{P}_\theta(x)$$

$$Q(z) = e^{\lambda-1} \underbrace{\mathbb{P}_\theta(x)}_{\text{const.}}\, \mathbb{P}_\theta(z|x) \implies Q^*(z) = \mathbb{P}_\theta(z|x)$$

$$\implies \mathcal{L}(\theta, Q) = \mathop{\mathbb{E}}_{z\sim Q^*}\left[\log \frac{\mathbb{P}_\theta(x,z)}{Q^*(z)}\right] = \sum_z Q^*(z) \log \frac{\mathbb{P}_\theta(x,z)}{Q^*(z)}$$

$$= \sum_z \mathbb{P}_\theta(z|x) \log \frac{\mathbb{P}_\theta(x,z)}{\mathbb{P}_\theta(z|x)} = \sum_z \mathbb{P}_\theta(z|x) \log \mathbb{P}_\theta(x) =$$

$$\log \mathbb{P}_\theta(x) \sum_z \mathbb{P}_\theta(z|x) = \log \mathbb{P}_\theta(x) = \mathcal{L}(\theta, x)$$

$\implies$ The distance is 0 and the bound is tight.

4 As described, EM can be seen as [E] optimizing the lower bound of log-likelihood with respect to $Q(z)$ for a fixed θ which would be the posterior distribution $\mathbb{P}(z|x)$, and [M] optimizing with respect to θ for a fixed Q of the previous step.

The problem with the above procedure would be the interactability of the posterior $\mathbb{P}_\theta(z|x)$; since:

$$\mathbb{P}_\theta(z|x) = \frac{\mathbb{P}_\theta(x|z)\,P_\theta(z)}{\mathbb{P}_\theta(x)} = \frac{\mathbb{P}_\theta(x|z)\;P_\theta(z)}{\sum_z \mathbb{P}_\theta(x|z)\,P_\theta(z)}$$

The denominator of the above posterior is interactable; since it must be calculated for any possible z in the latent space!, which is almost impossible.

5 $$\mathcal{L}(\theta,\phi) = \mathop{\mathbb{E}}_{z\sim Q_\phi}\left[\log \mathbb{P}_\theta(x,z)\right] + H(Q_\phi) = \mathop{\mathbb{E}}_{z\sim Q_\phi}\left[\log \frac{\mathbb{P}_\theta(x,z)}{Q_\phi(z)}\right]$$

- $$\log \mathbb{P}_\theta(x) = \log \sum_z \mathbb{P}_\theta(x|z)\,\mathbb{P}_\theta(z) = \log \mathop{\mathbb{E}}_{z\sim \mathbb{P}_\theta}\left[\mathbb{P}_\theta(x|z)\right]$$

$$= \log \mathop{\mathbb{E}}_{z\sim \mathbb{P}_\theta}\left[\frac{\mathbb{P}_\theta(x|z)}{Q_\phi(z|x)}\, Q_\phi[z|x]\right] = \log \mathop{\mathbb{E}}_{Q_\phi(z|x)}\left[\frac{\mathbb{P}_\theta(x|z)\,\mathbb{P}_\theta(z)}{Q_\phi(z|x)}\right]$$

$$\overset{*}{\geq} \mathbb{E}\left[\log\left(\frac{\mathbb{P}_\theta(x|z)\,\mathbb{P}_\theta(z)}{Q_\phi(z|x)}\right)\right] = \mathbb{E}\left[\log\left[\mathbb{P}_\theta(x|z)\right]\right] -$$

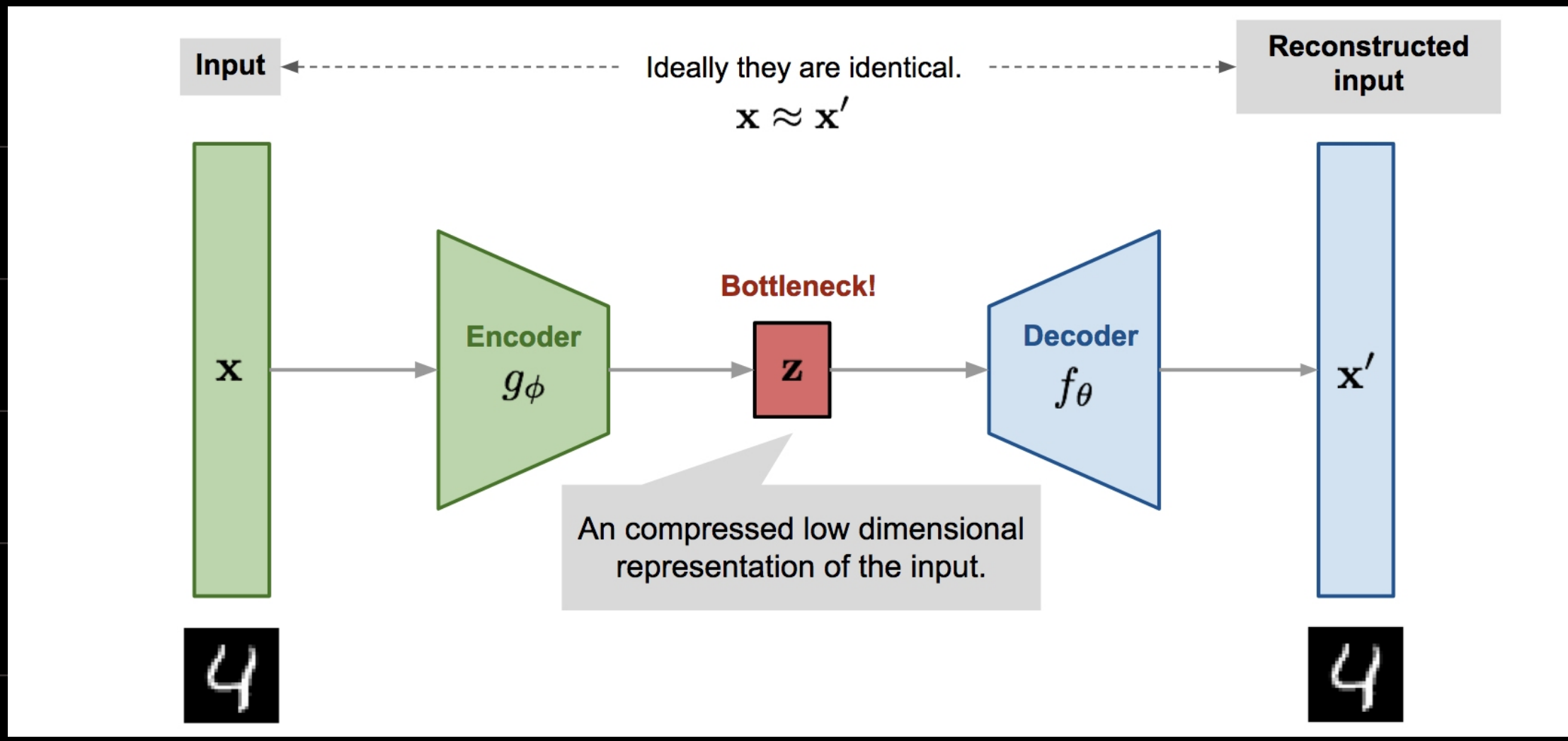
Jensen

$$\mathbb{E}\left[\log \frac{Q_\phi(z|x)}{\mathbb{P}_\theta(z)}\right] = \mathbb{E}\left[\log\left(\mathbb{P}_\theta(x|z)\right)\right] - D_{KL}\left(Q_\phi(z|x)\,\|\,\mathbb{P}_\theta(z)\right)$$

$$= \mathcal{L}(\theta,\phi)$$

$$\Longrightarrow \boxed{\mathcal{L}(\theta,\phi) = \mathop{\mathbb{E}}_{Q_\phi(z|x)}\left[\log\left(\mathbb{P}_\theta(x|z)\right)\right] - D_{KL}\left(Q_\phi(z|x)\,\|\,\mathbb{P}_\theta(z)\right)}$$

6

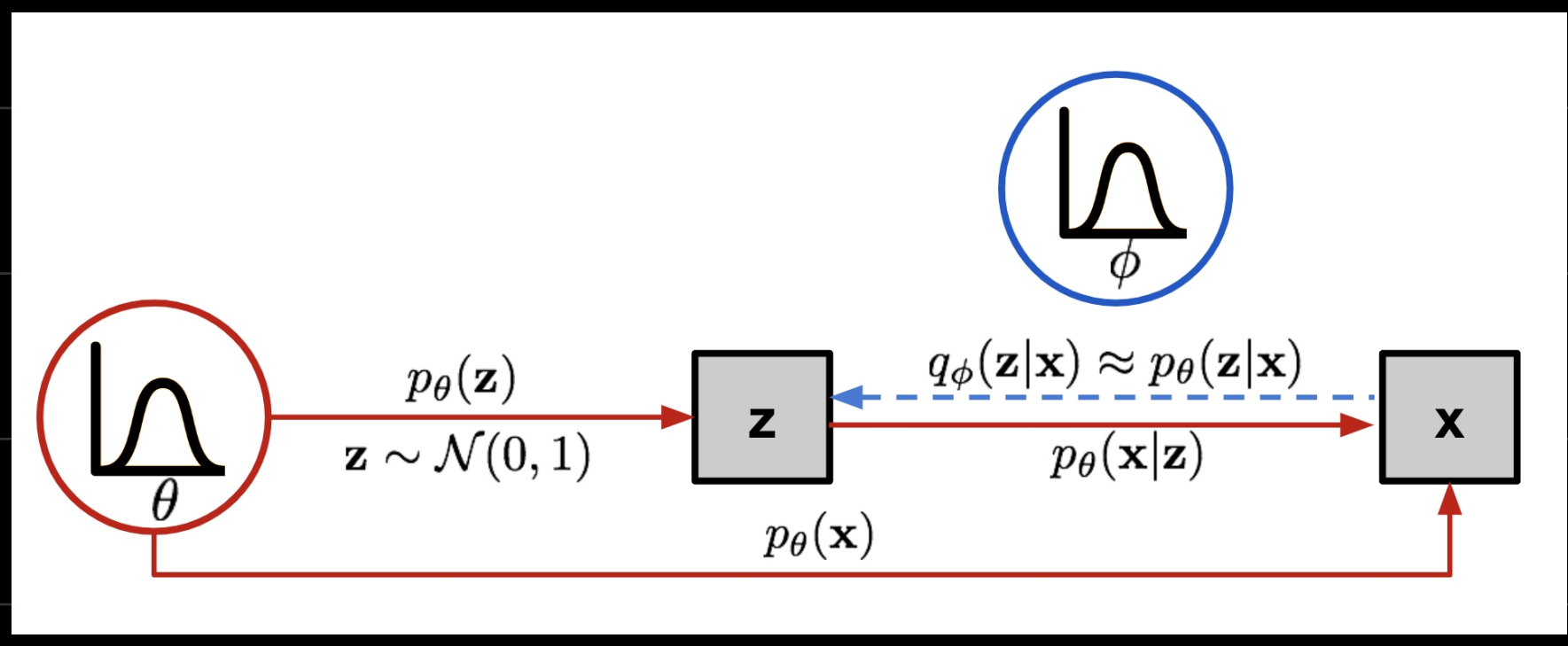


The training of the AE consists of minimizing a loss function through a deep neural network with encoder and decoder. A common loss could be MSE or ZINB (zero-inflated negative binomial, used for datasets with excessive dropout events, such as scRNA-seq data.)

$$\theta, \phi = \arg\min_{\theta,\phi} \mathcal{L}\left(f_\theta(g_\phi(\overbrace{x}^{x})), x\right)$$

However, in the VAE, instead of mapping the input into a fixed vector in the latent space, we want to map it into a distribution.

As stated in the previous parts, maximizing the likelihood is equivalent to maximizing the ELBO (which is computable in contrast with interactable posterior); So, we estimate the underlying distribution with $q_\phi(z|x)$. The graphical model would be:



$$ELBO = \underset{Q_\phi(z|x)}{\mathbb{E}}\left[\log P_\theta(x|z)\right] - D_{KL}\left(Q_\phi(z|x) \,||\, P_\theta(z)\right)$$

likelihood — regularization : KLD

Minimizing the first term implies lower reconstruction error; while the 2nd term is a regularizer, forcing the latent distribution to be as close as possible to the $P_\theta(z)$, which is usually a standard normal distribution.

* The problem of not being able to backpropagate after sampling is addressed by reparam. trick)

In other words, the encoder takes the input X and outputs a mean μ and a variance σ^2. Then, sampling occurs with $N(\mu, \sigma^2)$ through reparametrization trick: $z' \sim N(0,1) \Longrightarrow z = \sigma \cdot z' + \mu \sim N(\mu, \sigma^2)$. This random vector of latent space is passed through the decoder, which tries to reconstruct the original X. The KL divergence ensures the latent space distribution is close to the standard Normal distribution.

7 $$L(\theta, \phi) = \mathop{\mathbb{E}}_{z \sim Q_\phi(z|x)} \left[\log P_\theta(x|z) \right] - D_{KL}\left(Q_\phi(z|x) \,||\, P_\theta(z) \right)$$

$$\mathop{\mathbb{E}}_{x \sim Q_\phi} \left[D_{KL}\left(Q_\phi(z|x) \,||\, P_\theta(z) \right) \right] = \sum_x Q_\phi(x) \sum_z Q_\phi(z|x) \log \frac{Q_\phi(z|x)}{P_\theta(z)} =$$

$$\sum_x Q_\phi(x) \sum_z Q_\phi(z|x) \log \frac{Q_\phi(z|x)}{P_\theta(z)} \times \frac{Q_\phi(z)}{Q_\phi(z)} = \sum_{x,z} Q_\phi(x,z) \log \frac{Q_\phi(x,z)}{Q_\phi(x) Q_\phi(z)} +$$

$$\sum_{x,z} Q_\phi(x,z) \log \frac{Q_\phi(x)}{P_\theta(z)} = \mathop{I}_{Q_\phi}(X;Z) + D_{KL}\left(Q_\phi(z) \,||\, P_\theta(z) \right)$$

$$\Longrightarrow \boxed{L(\theta, \phi) = \mathop{\mathbb{E}}_{Q_\phi(z|x)} \left[\log P_\theta(x|z) \right] - D_{KL}\left(Q_\phi(z) \,||\, P_\theta(z) \right) - \mathop{I}_{Q_\phi}(X;Z)}$$

8

- By minimizing the KL divergence, we are implicitly minimizing the mutual information as well! This results in a poor reconstruction of the input, since the mutual information is an indicator of information of z in x and by reducing this quantity, we are encouraging our model to generate latent vectors independent of the input x. Hence, the reconstructed output would be more gaussian (more smooth) but has a higher error.

  Solution:

$$ELBO := \mathop{\mathbb{E}}_{x\sim Q_\phi}\left[\mathop{\mathbb{E}}_{z\sim Q_\phi(z|x)}[\log P_\theta(x|z)] - \beta D_{KL}\left(Q_\phi(z|x) \,\|\, P_\theta(z)\right)\right]$$

, where β controls the mentioned tradeoff; the lower the β, the higher the quality of reconstructed image & vice versa. (β-VAE)

- The noise variance estimation ϵ through maximizing the ELBO is biassed.