40110437 پارسا ترابیان

1)

1.1) $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma) \Rightarrow M_{\vec{X}}(\vec{t}) = \mathbb{E}_X(e^{t^T X}) = \exp(t^T\mu + \frac{1}{2} t^T \Sigma t)$

$\vec{Y} = A\vec{X} + \vec{b} \sim A \Rightarrow M_{\vec{Y}}(\vec{t}) = \mathbb{E}_X(\exp(t^T A x + t^T b)) \quad t \in \mathbb{R}^m$

$$= e^{t^T b}\, \mathbb{E}_X(e^{t^T A x})$$

$$\mathbb{E}_X(\exp(t^T A x)) = \int_{\mathbb{R}^d} \exp(t^T A x) \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))\, dx$$

$$= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \int_{\mathbb{R}^d} \exp(-\frac{1}{2} x^T \Sigma^{-1} x - \frac{1}{2} \mu^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} x + t^T A x)\, dx$$

$$= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \int_{\mathbb{R}^d} \exp(-\frac{1}{2} x^T \Sigma^{-1} x - \frac{1}{2} \mu^T \Sigma^{-1} \mu + \underbrace{(\mu + \Sigma A^T t)^T}_{\eta} \Sigma^{-1} x)\, dx$$

$$= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \int_{\mathbb{R}^d} \exp(-\frac{1}{2} x^T \Sigma^{-1} x - \frac{1}{2} \eta^T \Sigma^{-1} \eta + \eta^T \Sigma^{-1} x + \frac{1}{2} t^T A \Sigma^T \Sigma^{-1} \Sigma A^T t + \mu^T \Sigma^{-1} \Sigma A^T t)\, dx$$

$$= \exp(\frac{1}{2} t^T A \Sigma A^T t + \mu^T A^T t) \int_{\mathbb{R}^d} \underbrace{\mathcal{N}(\eta, \Sigma)\, dx}_{1}$$

$$\Rightarrow M_Y(t) = \exp(t^T(A\mu + b) + \frac{1}{2} t^T A \Sigma A^T t)$$

$$\Rightarrow Y \sim \mathcal{N}(A\mu + b, A\Sigma A^T) \quad \square$$

1.2)

$$\vec{X} \sim \mathcal{N}(\mu_X, \Sigma_X) \qquad \vec{Z} = a\vec{X} + b\vec{Y}$$

$$\vec{Y} \sim \mathcal{N}(\mu_Y, \Sigma_Y)$$

$$M_Z(t) = \mathbb{E}_Z(\exp(t^T Z)) = \mathbb{E}_{X,Y}(\exp(at^T X)\exp(bt^T Y))$$

$t_1 = at, \; t_2 = bt$

$$= \mathbb{E}_X(\exp(t_1^T X))\,\mathbb{E}_Y(\exp(t_2 Y))$$

$$= M_X(t_1)\,M_Y(t_2)$$

$$= \exp(t_1^T \mu_X + \tfrac{1}{2} t_1^T \Sigma_X t_1 + t_2^T \mu_Y + \tfrac{1}{2} t_2^T \Sigma_Y t_2)$$

$$\Rightarrow M_Z(t) = \exp(t^T(a\mu_X + b\mu_Y) + \tfrac{1}{2} t^T(a^2\Sigma_X + b^2\Sigma_Y)t)$$

$$\Rightarrow Z \sim \mathcal{N}(a\mu_X + b\mu_Y, \; a^2\Sigma_X + b^2\Sigma_Y) \qquad \square$$

1.3)

$$M_Y(t) = \mathbb{E}_Y(\exp(t^T Y)) = \mathbb{E}_X(\mathbb{E}_{Y|X}(\exp(t^T Y)\,|\,X))$$

$$= \mathbb{E}_X(\exp(t^T(AX+b) + \tfrac{1}{2} t^T \Sigma_{Y|X} t))$$

$$= \mathbb{E}_X(\exp(t^T AX + t^T b + \tfrac{1}{2} t^T \Sigma_{Y|X} t))$$

$$= \exp(t^T b + \tfrac{1}{2} t^T \Sigma_{Y|X} t)\, M_X(A^T t)$$

$$= \exp(t^T b + \tfrac{1}{2} t^T \Sigma_{Y|X} t)\exp(t^T A \mu_X + \tfrac{1}{2} t^T A \Sigma_{XX} A^T t)$$

$$\Rightarrow M_Y(t) = \exp(t^T(A\mu_X + b) + \tfrac{1}{2} t^T(\Sigma_{Y|X} + A\Sigma_{XX}A^T)t)$$

$$\Rightarrow Y \sim \mathcal{N}(A\mu_X + b, \; A\Sigma_{XX}A^T)$$

$\square$

1.4)

$$M_X(t) = \mathbb{E}_X(\exp(t^T X)) = \overset{\text{marginalize out}}{\mathbb{E}_Y}\left[\mathbb{E}_X(\exp(t^T X))\right] = \mathbb{E}_{X,Y}(\exp(t^T X))$$

$$M_Z\left(\begin{pmatrix} t_X \\ t_Y \end{pmatrix}\right) = \mathbb{E}_{X,Y}(\exp(t_X^T X + t_Y^T Y))$$

$$\Rightarrow M_X(t) = M_Z\left(\begin{pmatrix} t \\ 0 \end{pmatrix}\right) = \exp\left(\begin{pmatrix} t \\ 0 \end{pmatrix}^T \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} + \frac{1}{2}\begin{pmatrix} t \\ 0 \end{pmatrix}^T \Sigma \begin{pmatrix} t \\ 0 \end{pmatrix}\right)$$

$$= \exp\left(t^T \mu_X + \frac{1}{2} t^T \Sigma_{XX} t\right)$$

$$\Rightarrow X \sim \mathcal{N}(\mu_X, \Sigma_{XX})$$

$$\Rightarrow f_{Y|X}(Y|X) = \frac{f_{XY}(X,Y)}{f_X(X)} = \frac{\sqrt{(2\pi)^{d_X}|\Sigma_{XX}|}}{\sqrt{(2\pi)^{d_X+d_Y}|\Sigma|}} \cdot \frac{\exp\left(-\frac{1}{2}\begin{pmatrix} X-\mu_X \\ Y-\mu_Y \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} X-\mu_X \\ Y-\mu_Y \end{pmatrix}\right)}{\exp\left(-\frac{1}{2}(X-\mu_X)^T \Sigma_{XX}^{-1}(X-\mu_X)\right)}$$

$$\Sigma^{-1} = \begin{pmatrix} \Lambda_{XX} & \Lambda_{XY} \\ \Lambda_{YX} & \Lambda_{YY} \end{pmatrix}$$

$$\Lambda_{XX} = \Sigma_{XX}^{-1} + \Sigma_{XX}^{-1}\Sigma_{XY} M^{-1} \Sigma_{YX} \Sigma_{XX}^{-1}$$

$$\Lambda_{XY} = -\Sigma_{XX}^{-1}\Sigma_{XY} M^{-1}$$

$$\Lambda_{YX} = -M^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}$$

$$\Lambda_{YY} = M^{-1} \qquad M = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

$$\begin{pmatrix} X-\mu_X \\ Y-\mu_Y \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} X-\mu_X \\ Y-\mu_Y \end{pmatrix} = (X-\mu_X)^T \Lambda_{XX}(X-\mu_X) + (Y-\mu_Y)^T \Lambda_{YY}(Y-\mu_Y)$$

$$+ (X-\mu_X)^T \Lambda_{XY}(Y-\mu_Y) + (Y-\mu_Y)^T \Lambda_{YX}(X-\mu_X)$$

$$= (X-\mu_X)^T \Sigma_{XX}^{-1}(X-\mu_X) + (X-\mu_X)^T \Sigma_{XX}^{-1}\Sigma_{XY}\underset{(**)}{M^{-1}}\Sigma_{YX}\Sigma_{XX}^{-1}(X-\mu_X)$$

$$+ (Y-\mu_Y)^T \underset{(*)}{M^{-1}}(Y-\mu_Y) - (X-\mu_X)^T \underset{(**)}{\Sigma_{XX}^{-1}\Sigma_{XY}} M^{-1}(Y-\mu_Y) - (Y-\mu_Y)^T M^{-1} \underset{(*)}{\Sigma_{YX}}\Sigma_{XX}^{-1}(X-\mu_X)$$

$$= (X-\mu_X)^T \Sigma_{XX}^{-1}(X-\mu_X) + (Y-\mu_Y)^T M^{-1}(\underbrace{Y-\mu_Y - \Sigma_{YX}\Sigma_{XX}^{-1}(X-\mu_X)}_{a}) +$$

$$(x-\mu_X)^T\Sigma_{XX}^{-1}\Sigma_{XY}\, M^{-1}\Big(\underbrace{\Sigma_{YX}\Sigma_{XX}^{-1}(x-\mu_X)-(Y-\mu_Y)}_{-\alpha}\Big)$$

$$=(x-\mu_X)^T\Sigma_{XX}^{-1}(x-\mu_X)+$$

$$(Y-\mu_Y-\Sigma_{XY}^T\Sigma_{XX}^{-1}(x-\mu_X))^T M^{-1}(Y-\mu_Y-\Sigma_{YX}\Sigma_{XX}^{-1}(x-\mu_X))=g(x,Y)$$

Σ is symmetric: $\begin{pmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{pmatrix}=\begin{pmatrix}\Sigma_{XX}^T & \Sigma_{YX}^T\\ \Sigma_{XY}^T & \Sigma_{YY}^T\end{pmatrix}$

$\Rightarrow \Sigma_{YX}=\Sigma_{XY}^T$

$$\Rightarrow f_{Y|X}(Y|X)=\frac{\sqrt{(2\pi)^{d_X}|\Sigma_{XX}|}}{\sqrt{(2\pi)^{d_X+d_Y}|\Sigma|}}\exp\left(-\frac{1}{2}\left(g(x,Y)-(x-\mu_X)^T\Sigma_{XX}^{-1}(x-\mu_X)\right)\right)$$

$g(x,Y)-(x-\mu_X)^T\Sigma_{XX}^{-1}(x-\mu_X)=$

$$(Y-\mu_Y-\Sigma_{YX}\Sigma_{XX}^{-1}(x-\mu_X))^T M^{-1}(Y-\mu_Y-\Sigma_{YX}\Sigma_{XX}^{-1}(x-\mu_X))$$

$\mu_{Y|X}=\mu_Y+\Sigma_{YX}\Sigma_{XX}^{-1}(x-\mu_X)$ $\qquad\qquad$ $\Sigma_{Y|X}=M=\Sigma_{YY}-\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$

$|\Sigma|=|\Sigma_{XX}||M|$

$$\Rightarrow f_{Y|X}(Y|X)=\frac{1}{\sqrt{(2\pi)^{d_Y}|M|}}\exp\left[-\frac{1}{2}(Y-\mu_{Y|X})^T\Sigma_{Y|X}^{-1}(Y-\mu_{Y|X})\right]$$

$\Rightarrow Y|X\sim\mathcal{N}(\mu_{Y|X},\Sigma_{Y|X})$ $\qquad\qquad$ $\square$

1.5)
$$D_{KL}(\overbrace{\mathcal{N}(\mu_0,\Sigma_0)}^{P_1}\,||\,\overbrace{\mathcal{N}(\mu_1,\Sigma_1)}^{P_2}) = \mathbb{E}_{X\sim\mathcal{N}(\mu_0,\Sigma_0)}\left[\log\left(\frac{P(x)}{Q(x)}\right)\right]$$

$$= \mathbb{E}_{P_1}\left[\log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \frac{1}{2}\left((x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)\Sigma_0^{-1}(x-\mu_0)\right)\right]$$
$$= \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \mathbb{E}_{P_1}\left(\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right) - \mathbb{E}_{P_1}\left(\frac{1}{2}(x-\mu_0)\Sigma_0^{-1}(x-\mu_0)\right)$$

$$x\sim\mathcal{N}(m,\Sigma):\ \mathbb{E}_x\left((x-m')^TA(x-m')\right) = (m-m')^TA(m-m') + tr(A\Sigma)$$

proof: $\mathbb{E}_x\left((x-m')^TA(x-m')\right) = m'^TAm' - 2m^TAm' + \mathbb{E}_x(\underbrace{x^TAx}_{tr(x^TAx)=tr(Axx^T)})$

$$= m'^TAm' - 2m^TAm' + tr(A\,\mathbb{E}_x(xx^T)) =$$

$$= m'^TAm' - 2m^TAm' + tr(A\,\mathbb{E}_x((x-m)(x-m)^T - mm^T + xm^T + mx^T))$$

$$= m'^TAm' - 2m^TAm' + tr(A\Sigma - Amm^T + 2Amm^T)$$

$$= m'^TAm' - 2m^TAm' + tr(A\Sigma + Amm^T)$$

$$= m'^TAm' - 2m^TAm' + tr(A\Sigma) + m^TAm = (m-m')$$

$$\Rightarrow D_{KL}(P_1||P_2) = \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) - \frac{1}{2}tr(\Sigma_1^{-1}\Sigma_1) + \frac{1}{2}(\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1)$$
$$+ \frac{1}{2}tr(\Sigma_1^{-1}\Sigma_0)$$

$$\Rightarrow D_{KL}(\mathcal{N}(\mu_0,\Sigma_0)||\mathcal{N}(\mu_1,\Sigma_1)) = \frac{1}{2}\left[\log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + (\mu_0-\mu_1)^T\Sigma_1^{-1}(\mu_0-\mu_1)\right.$$
$$\left. + tr(\Sigma_1^{-1}\Sigma_0) - d\right]$$

$\square$

2)

2.1) $\mathcal{L}(y_1, \dots, y_n; \phi, \sigma^2) = P(\varepsilon_1 = y_1) P(\varepsilon_2 = y_2 - \phi_1 y_1) \cdots P(\varepsilon_n = y_n - \phi_1 y_{n-1} + \cdots + \phi_p y_{n-p})$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\frac{y_1^2}{\sigma^2}\right) \exp\left(-\frac{1}{2}\frac{(y_2-\phi_1 y_1)^2}{\sigma^2}\right) \cdots$$

$$\exp\left(-\frac{1}{2}\left(\frac{y_n - (\phi_1 y_{n-1} + \cdots + \phi_p y_{n-p})}{\sigma}\right)^2\right)$$

$$\Rightarrow \log \mathcal{L}(Y; \phi, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{y_1^2}{\sigma^2} - \frac{1}{2}\frac{(y_2 - \phi_1 y_1)^2}{\sigma^2} - \cdots$$

$$- \frac{1}{2}\frac{(y_n - (\phi_1 y_{n-1} + \cdots + \phi_p y_{n-p}))^2}{\sigma^2}$$

2.2)

maximizing $\mathcal{L}$ w.r.t ϕ:

$$\min_{\phi} (Y - X\phi)^T (Y - X\phi) : \qquad X = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ y_1 & 0 & \cdots & 0 \\ y_2 & y_1 & \cdots & \\ \vdots & & & \\ y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{pmatrix}$$

$$\min_{\phi} Y^T Y + \phi^T X^T X \phi - 2 Y^T X \phi$$

$$\min_{\phi} \phi^T X^T X \phi - 2 Y^T X \phi \Rightarrow \frac{\partial}{\partial \phi}\left[\phi^T X^T X \phi - 2 Y^T X \phi\right] = 0$$

$$\Rightarrow 2 X^T X \phi = 2 X^T Y \Rightarrow \boxed{\phi^* = (X^T X)^{-1} X^T Y}$$

maximizing $\mathcal{L}$ w.r.t σ:

$$\max_{\sigma} n \log(\sigma^2) + \frac{(Y - X\phi)^T (Y - X\phi)}{\sigma^2}$$

$$\frac{\partial}{\partial \sigma^2}\left(n \log(\sigma^2) + \frac{(Y - X\phi)^T (Y - X\phi)}{\sigma^2}\right) = 0 \Rightarrow \frac{n}{\sigma^2} - \frac{(Y - X\phi)^T (Y - X\phi)}{\sigma^4} = 0$$

$$\Rightarrow \sigma^2 = \frac{1}{n}(Y - X\phi)^T (Y - X\phi)$$

$$\Rightarrow \sigma^{2*} = \frac{1}{n}\left(Y - (X^TX)^{-1}X^TY\right)^T\left(Y - (X^TX)^{-1}X^TY\right)$$

2.3)

(a) $P(Y_t \mid Y_{t-1}) = P_\varepsilon(Y_t - \phi Y_{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(Y_t - \phi Y_{t-1})^2}{\sigma^2}\right)$

(b) $\sum_{y'} \pi(y')\, P(y \mid y') = \pi(y)$

$$\pi^{(n)}(y) = P(Y_0) \prod_{t=1}^{n-1} P(Y_t \mid Y_{t-1}) \qquad P(Y \mid Y_{n-1}) = P(Y_0)\, C_1 \exp\left(-\frac{(Y - \phi Y_{n-1})^2}{2\sigma^2}\right)$$

$$M_{Y_T}(t) = \mathbb{E}_{Y_T}[\exp(tY_T)] = \mathbb{E}_{Y_{T-1},\varepsilon_T}\left[\exp(t\phi Y_{T-1} + t\varepsilon_T)\right]$$

$$= \mathbb{E}_{Y_{T-1}}\left[\exp(t\phi Y_{T-1})\right] \mathbb{E}_{\varepsilon_T}\left[(\exp(t\varepsilon_T)\right]$$

$$= M_{Y_{T-1}}(t\phi) \quad \exp\left(\tfrac{1}{2}t^2\sigma^2\right)$$

$$\Rightarrow M_{Y_T}(t) = \exp\left(\tfrac{1}{2}t^2\sigma^2\right)\exp\left(\tfrac{1}{2}(t\phi)^2\sigma^2\right)\exp\left(\tfrac{1}{2}(t\phi^2)^2\sigma^2\right)\cdots$$

$$\exp\left(\tfrac{1}{2}(t\phi^{T-1})^2\sigma^2\right)\ \exp\left(t\phi^T Y_0\right)$$

$$= \exp\left(\tfrac{1}{2}\sigma^2 t^2 \sum_{i=1}^{T}(\phi^2)^{i-1}\right)\exp(t\phi^T Y_0) \underset{|\phi|<1}{=} \exp\left(\frac{1}{2}\sigma^2 t^2\, \frac{1-\phi^{2T}}{1-\phi^2}\right)\exp(t\phi^T Y_0)$$

$$\Rightarrow M_Y(t) = \lim_{t\to\infty} M_{Y_T}(t) = \exp\left(\frac{1}{2}\,\frac{\sigma^2 t^2}{1-\phi^2}\right) \Rightarrow Y \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right) \quad \square$$

3)

3.1) $$\mathcal{L}_\theta(x_{1:T}) = -\log\left(p_\theta(x_{1:T} : \theta, \sigma^2)\right)$$

$$= -\log\left(\prod_{i=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - f_\theta(x_{j-1},\dots,x_{j-p}))^2}{\sigma^2}\right)\right)$$

$$= \frac{T}{2}\log(2\pi\sigma^2) + \frac{1}{2}\sum_{i=1}^{T} \frac{(x_i - f_\theta(x_{j-1},\dots,x_{j-p}))^2}{\sigma^2}$$

$$\frac{\partial \mathcal{L}_\theta}{\partial \theta} = \frac{-1}{\sigma^2}\sum_{i=1}^{T} (x_i - f_\theta(x_{i-1},\dots,x_{i-p})) \frac{\partial f_\theta(x_{i-1},\dots,x_{i-p})}{\partial \theta}$$

در هر زمان $(x_i - f_\theta(x_{i-1},\dots,x_{i-p})) \frac{\partial f_\theta(x_{i-1},\dots,x_{i-p})}{\partial \theta}$ را محاسبه می‌کنیم و مقدارش را جمع می‌کنیم.

اگر p بزرگ باشد مدل θ به پارامترهای زیادی نیاز پیدا می‌کند که یادگیری‌اش را سخت‌تر می‌کند.

اگر T زیاد باشد مقدار گرادیان تجمیع یافته زیاد می‌شود و مشکل exploding gradient ایجاد می‌شود.

3.2)

$p \gg 1$: از Transformer ها برای یادگیری ارتباط اجزا با فاصله‌های زیاد استفاده می‌کنیم.
اما به تعداد پارامترهای بسیار زیادی برای یادگیری نیاز خواهیم داشت.

همچنین از GRU یا LSTM نیز می‌توان برای یادگیری سری‌های زمانی استفاده کرد اما مشکل vanishing gradient ممکن است رخ دهد.

small p: می‌توان از MLP و CNN برای یادگیری ارتباط بین سری‌های زمانی با p کوچک استفاده کرد. اما ممکن است روابط و اطلاعات پیچیده‌تر را نتوانند به خوبی دریافت کنند.

3.3) $x_1, \dots, x_t$

$$\hat{x}_{t+1} \sim \mathcal{N}(f_\theta(x_t, \dots, x_{t-p+1}), \sigma^2)$$

$$\vdots$$

$$\hat{x}_{t+k} \sim \mathcal{N}(f_\theta(\hat{x}_{t+k-1}, \dots, x_{t+k-p}), \sigma^2)$$

$$\varepsilon_{t+1} = x_{t+1} - \hat{x}_{t+1} \Rightarrow \bar{\varepsilon}_{t+1} = \mathbb{E}[\varepsilon_{t+1}] = x_{t+1} - f_\theta(x_t, \dots, x_{t-p+1})$$

بطور مشابه $\Rightarrow \mathbb{E}[\varepsilon_{t+2}] = x_{t+2} - f_\theta(\underbrace{\hat{x}_{t+1}}_{x_{t+1}-\bar{\varepsilon}_t}, x_t, \dots, x_{t-p+2})$

$$\approx x_{t+2} - f_\theta(x_{t+1}, x_t, \dots, x_{t-p+2}) + \frac{\partial f_\theta(x_{t+1}, \dots, x_{t-p+2})}{\partial x_{t+1}} \bar{\varepsilon}_{t+1}$$

بطور مشابه: $\mathbb{E}[\varepsilon_{t+k}] = x_{t+k} - f_\theta(x_{t+k-1}, \dots, x_{t+k-p}) + \nabla_x f_\theta(x_{t+k-1}, \dots, x_{t+k-p}) \cdot \bar{\varepsilon}$ $\quad \bar{\varepsilon} = \begin{bmatrix} \bar{\varepsilon}_{t+k-1} \\ \vdots \\ \bar{\varepsilon}_{t+k-p} \end{bmatrix}$

اگر $\frac{\partial f_\theta(x_{t+1}, \dots, x_{t-p+2})}{\partial x_{t+1}}$ زیاد باشد، خطای $\hat{x}_{t+k}$ با افزایش k بصورت تجمعی افزایش می‌یابد.

راه حل: ۱- بصورت مستقیم و با یک مرحله $\hat{x}_{t+1}, \dots, \hat{x}_{t+k}$ را تولید کنیم.

2- از چند مدل بصورت ensemble استفاده کنیم تا واریانس جواب و خطا را کاهش دهیم.

3.4)

$$L_\theta(x_{1:T}) = \frac{T}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{T}(x_i - f_\theta(x_{i-1}, \dots, x_{i-p}; c_{1:i}))^2$$

در هر نقطه باید $(x_i - f_\theta(x_{i-1}, \dots, x_{i-p}; c_{1:i}))^2$ را کمینه کنیم.

از این مدل می‌توان برای تولید یک speech مرتبط و نزدیک با یک متن استفاده کرد.

همچنین می‌توان با دادن متن به عنوان ورودی و به کمک یک کنترلر، حرکات مرتبط با متن را تولید کرد.

Factorization: Unconditional: $p(x_{1:T}) = \prod_{t=1}^{T} p(x_t | x_{<t})$

Conditional : $p(x_{1:T} | c_{1:T}) = \prod_{t=1}^{T} p(x_t | x_{<t}, c_{1:t})$

مدل Unconditional فقط به x های قبلی وابسته است اما مدل Conditional به context ها گذشته نیز وابسته است.

مدل Unconditional برای مدل کردن سری‌ها زمانی واحد که به سری زمانی دیگری وابسته نیستند مناسب است.

مدل Conditional برای مدل کردن سری‌ها وابسته به سری‌ها دیگر مانند text to speech مناسب است.

3.5)

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

$$= \mathbb{E}_{x \sim P_\theta}\left[\log\left(\prod_{t=1}^{T} \frac{P_\theta(x_t | x_{<t}, c_{1:t})}{q_\phi(x_t | x_{<t}, c_{1:t})}\right)\right]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{x_t \sim P_\theta}\left[\log\left(\frac{P_\theta(x_t | x_{<t}, c_{1:t})}{q_\phi(x_t | x_{<t}, c_{1:t})}\right)\right]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{x_t \sim P_\theta}\left[\frac{-1}{2\sigma^2}\left(x_t - f_\theta(x_{t-1}, \dots, x_{t-p}; c_{1:t})\right)^2 + \frac{1}{2\sigma^2}\left(x_t - g_\phi(x_{t-1}, \dots, x_{t-p}; c_{1:t})\right)^2\right]$$

$$\Rightarrow D_{KL}(P_\theta || Q_\phi) = \frac{1}{2\sigma^2} \sum_{t=1}^{T} \mathbb{E}_{x_t \sim P_\theta}\left[\left(x_t - g_\phi(x_{t-1}, \dots, x_{t-p}; c_{1:t})\right)^2 - \left(x_t - f_\theta(x_{t-1}, \dots, x_{t-p}; c_{1:t})\right)^2\right]$$

محاسبه $\mathbb{E}_{x_t \sim P_\theta}$ نیاز به انتگرال گیری از توزیع P_θ دارد که هم در $T \gg 1$ و هم به دلیل پیچیدگی موجود در f_θ , g_ϕ از نظر محاسباتی intractable می‌باشد.

راه حل: از روش Monte-Carlo برای محاسبه امید ریاضی استفاده می‌کنیم.

$$x_{1:T}^{(i)} \sim P_\theta \qquad i \in \{1, \dots, N\}$$

$$\Rightarrow D_{KL}(P_\theta \| Q_\phi) \approx \frac{1}{2\sigma^2} \frac{1}{N} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(x_t^{(i)} - g_\phi(x_{t-1}^{(i)}, \dots, x_{t-p}^{(i)}; c_{1:t})\right)^2 - \left(x_t^{(i)} - f_\theta(x_{t-1}^{(i)}, \dots, x_{t-p}^{(i)}; c_{1:t})\right)^2$$

4)

برای parametrize کردن مدل جدید باید بتوانیم c تا توزیع Guassian به همراه یک توزیع گسسته روی این توزیع‌های نرمال بسازیم.

Guassian (K):

$$\left.\begin{aligned} \mu_i^K &= v_{K_i}^T h_i + b_{K_i} \\ (\sigma_i^K)^2 &= \exp(u_{K_i}^T h_i + d_{K_i}) \end{aligned}\right\}$$

همان الگوی قبلی

Prior:

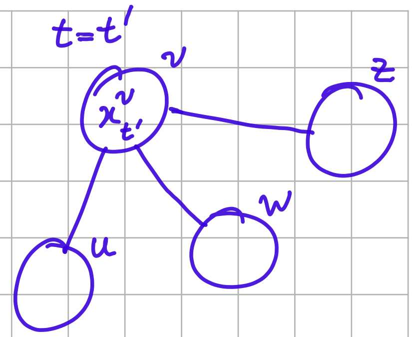
$$\pi_i = (\pi_i^1, \dots, \pi_i^C)^T \Rightarrow \pi_i^K = \mathrm{Softmax}(w_K^T h_i + c_K)$$

$$= \frac{\exp(w_K^T h_i + c_K)}{\sum_{c'=1}^{C} \exp(w_{c'}^T h_i + c_{c'})}$$

for a single $p(x_i | x_{<i})$:

$$\underbrace{C \times d + C}_{\mu_i} + \underbrace{C \times d + C}_{\sigma_i^2} + \underbrace{C \times d + C}_{\pi_i} = 3C(d+1)$$

5)



1. $$P(x_{1:T}) = \prod_{t=1}^{T} P(x_t \mid x_{<t})$$

$$P(x_t \mid x_{<t}) = P(x_t^1, \dots, x_t^d \mid x_{<t})$$
$$= \prod_{i=1}^{d} P(x_t^i \mid x_{<t}^i, x_{<t}^{N(i)})$$
$$= \prod_{v \in V} p(x_t^v \mid x_{<t}^v, x_{<t}^{N(v)})$$

$$\Rightarrow P(x_{1:T}) = \prod_{t=1}^{T} \prod_{v \in V} p(x_t^v \mid x_{<t}^v, x_{<t}^{N(v)})$$

2.

$$L_\theta(x_{1:T}) = \sum_{t=1}^{T} \sum_{v \in O_t} \log p_\theta(x_t^v \mid x_{<t}^v, x_{<t}^{N(v)})$$
$$= \sum_{t=1}^{T} \sum_{v \in V} m_t^v \log p_\theta(x_t^v \mid x_{<t}^v, x_{<t}^{N(v)})$$

اگر node های مختلف را در iteration های مختلف با یک احتمال p ماسک کنیم، درست مانند dropout مدل را مجبور می‌کنیم تا برای ساخت liklihood بالا تنها روی یک v خاص متمرکز نشود. درنتیجه مدل در برابر ورودی های ناقص مقاوم می‌شود و می‌تواند همچنان عملکرد مناسبی داشته باشد:

$$p(x_t^{M_t} \mid x_t^{O_t}) = \frac{p(x_t^{M_t}, x_t^{O_t})}{p(x_t^{O_t})}$$

$$\Rightarrow \hat{x}_t^{M_t} = \underset{x_t^{M_t}}{\operatorname{argmax}}\ p(x_t^{M_t}, x_t^{O_t}) \Rightarrow$$ مدل را با یک x^{M_t} اولیه به صورت iterative بروزرسانی می‌کنیم.

$$\forall v \in M_t : \hat{x}_t^v = GNN(x_{t-1}^v, \{x_{t-1}^u : u \in N(v)\})$$