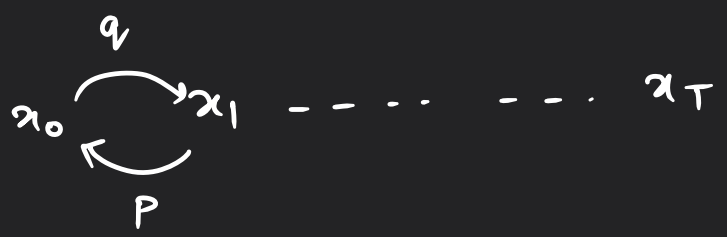
Diffusion Models and their use in Language Modelling

cumulating loss problem
fast inference
control over generation
↳ Main observation points

#DDPM:



$$q(x_t|x_{t-1}) = N(x_t|\sqrt{1-\beta_t}\,x_{t-1}, \beta_t I)$$

$$x_t = \sqrt{1-\beta_t}\,x_{t-1} + \sqrt{\beta_t}\,\varepsilon \qquad \varepsilon \sim N(0,I)$$

$$q(x_t|x_0) = N(x_t|\sqrt{\bar{\alpha}_t}\,x_0, (1-\bar{\alpha}_t)I)$$

$$x_t = \sqrt{\bar{\alpha}_t}\,x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon \qquad \varepsilon \sim N(0,I)$$

objective: $\max_\theta E_{x_0 \sim p_r}[\log p_\theta(x_0)]$

$$p_\theta(x_0) = p_\theta(x_0, z) \qquad z = x_{1:T}$$

$$\log p_\theta(x_0) = \log \int p_\theta(x, z)\,dz$$

$$= \log \int \frac{q(z|x)}{q(z|x)} p_\theta(x, z)\,dz$$

$$= \log E_q\left(\frac{p_\theta(x,z)}{q(z|x)}\right)$$

$$\geq E_q\left(\log \frac{p_\theta(x,z)}{q(z|x)}\right) \rightarrow \text{ELBO}$$

$$p_\theta(x_0, z) = p_\theta(x_0, x_1 \cdots x_T) = p(x_T) \prod_{t=1}^{T} p(x_{t-1}|x_t)$$

$$q(z|x) = q(x_1 \cdots x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

$$E_q\left(\log \frac{P(x_T)\prod_{t=1}^{T} P(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}\right)$$

$$= E_q\left(\sum_{t=1}^{T} \log \frac{P(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right)$$

$$= E_q\left(\log \frac{P(x_0|x_1)}{q(x_1|x_0)} + \sum_{t=2}^{\infty} \log \frac{P(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right)$$

$$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0)$$

$$= \frac{q(x_{t-1}|x_t, x_0)\, q(x_t|x_0)}{q(x_{t-1}|x_0)}$$

$$= E_q\left(\sum_{t=2}^{\infty} \log \frac{P(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \log \prod_{t=2}^{T} \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} + \log \frac{P(x_0|x_1)}{q(x_1|x_0)}\right)$$

$\frac{q(x_1|x_0)}{q(x_T|x_0)}$

$$= E_q\left(\sum_{t=2}^{T} \log \frac{P(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} + \log \frac{P(x_0|x_1)}{q(x_T|x_0)}\right)$$

$$= \sum_{t=2}^{T} E_{x_t \sim q_t,\, x_{t-1} \sim q_{t-1}}\left(\log \frac{P(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\right)$$

$$+ \; E_{x_1 \sim q_1} \log P(x_0|x_1)$$

$$= \sum_{t=2}^{T} E_{x_t \sim q(x_t|x_0)} \Big(-D_{KL}\big(q(x_{t-1}|x_t, x_0) \,\|\, P(x_{t-1}|x_t) \big) \Big)$$

$$+ E_{x_1 \sim q_1} \log P(x_0|x_1)$$

$$\text{obj} \quad \min_{\theta} L_\theta$$

$$L(\theta) = \sum_{t=2}^{T} E_{q(x_t|x_0)} \Big(D_{KL}\big(q(x_{t-1}|x_t, x_0) \,\|\, P(x_{t-1}|x_t) \big) \Big)$$

$$- E_{q(x_1|x_0)} \log P(x_0|x_1)$$

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)\, q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$$= \frac{N\left(\sqrt{\alpha_t}\, x_{t-1}, (1-\alpha_t) I\right) N\left(\sqrt{\bar{\alpha}_{t-1}}\, x_0, (1-\bar{\alpha}_{t-1}) I\right)}{N\left(\sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t) I\right)}$$

$$\propto \exp\left(-\frac{(x_t - \sqrt{\alpha_t}\, x_{t-1})^2}{2(1-\alpha_t)} - \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\, x_0)^2}{2(1-\bar{\alpha}_{t-1})} \right.$$

$$\left. + \frac{(x_t - \sqrt{\bar{\alpha}_t}\, x_0)^2}{2(1-\bar{\alpha}_t)} \right)$$

$$\propto \exp\left(\frac{1}{2(1-\alpha_t)} \left(-\alpha_t x_{t-1}^2 + 2\sqrt{\alpha_t}\, x_t\, x_{t-1} \right) \right.$$

$$\left. - \frac{1}{2(1-\bar{\alpha}_{t-1})} \left(x_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}\, x_0\, x_{t-1} \right) \right)$$

$$\propto \exp\left(\frac{-1}{2}\left[\left(\frac{\alpha_t}{(1-\alpha_t)} + \frac{1}{(1-\bar{\alpha}_{t-1})}\right) x_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}}{1-\alpha_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}} x_0\right) x_{t-1}\right]\right)$$

$$\propto \exp\left(\frac{-1}{2}\left[\frac{\alpha_t - \bar{\alpha}_t + 1 - \alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})} x_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t) x_0}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right) x_{t-1}\right]\right)$$

$$\propto \exp\left(\frac{-1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[x_{t-1}^2 - \frac{2\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) \cdots}{(1-\bar{\alpha}_t)} x_{t-1}\right]\right)$$

$$\tilde{\Sigma} = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} I =$$

$$\tilde{\mu} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t) x_0}{1-\bar{\alpha}_t}$$

$$\tilde{\mu} = \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) x_t}{1-\bar{\alpha}_t}$$

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon \qquad x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon}{\sqrt{\bar{\alpha}_t}}$$

$$\tilde{\mu} = \frac{(x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon)}{\sqrt{\alpha_t}} \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t$$

$$= \frac{1}{\sqrt{\alpha_t}}\left(-\frac{(1-\alpha_t)\,\varepsilon}{\sqrt{1-\bar{\alpha}_t}}\right) + \frac{1}{\sqrt{\alpha_t}}\left(\frac{x_t(1-\alpha_t)}{1-\bar{\alpha}_t} + \frac{\alpha_t(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)} x_t\right)$$

$$\frac{x_t - \cancel{\alpha_t x_t} + \cancel{\alpha_t x_t} - \bar{\alpha}_t x_t}{(1-\bar{\alpha}_t)} = x_t$$

$$\tilde{\mu} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\varepsilon\right) \qquad \tilde{\Sigma} = \frac{\beta_t(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)} I$$

$$\text{obj} = E_{x_0 \sim p^2,\, t \sim U[0,T],\, x_t \sim q_t}\left[D_{KL}(q_t \| p_t)\right]$$

$$q_t \sim N(\tilde{\mu}, \tilde{\Sigma}) \qquad p_t \sim N(\mu_\theta, \Sigma_\theta)$$

decoder model

#MDLM:

obj:

$$E_{x_0 \sim P_d}\Big[\sum_{t=2}^{\infty} E_{q(x_t|x_0)}\big[D_{KL}\big(q(x_{t-1}|x_t, x_0) \,\|\, P(x_{t-1}|x_t)\big)\big] + E_{q(x_1|x_0)}[\log P(x_0|x_1)] \Big]$$

$$= E_{t \in \{\frac{1}{T}, \frac{2}{T} \ldots 1\}} x_0 \sim P_d \; x_t \sim q(x_t|x_0)\,[D_{KL}(q \| P)]$$

instead of Gaussian, Categorical

$$q(x_t|x_{t-1}) = \text{cat}\left(Q_t^T x_{t-1}\right)$$

$$q(x_t|x_0) = \text{cat}\left((Q_1 \cdots Q_t)^T x_0\right)$$

$$= \text{cat}\left(\bar{Q}_t^T x_0\right)$$

$$q(x_t|x_0) = \text{cat}\left(\alpha_t x_0 + (1-\alpha_t) m\right) \quad m \to \text{mask}$$

$$q(x_t|x_{t-1}) = \text{cat}\left(\alpha_{t|t-1}\, x_{t-1} + (1-\alpha_{t|t-1}) m\right)$$

$s<t$ $\quad q(x_t|x_s) = \text{cat}\left(\alpha_{t|s}\, x_s + (1-\alpha_{t|s}) m\right)$

$$Q_{t|s} = \alpha_{t|s} I + (1-\alpha_{t|s})\, 1 m^T$$

$$q(x_t|x_s) = \text{cat}\left(Q_{t|s}^T x_s\right)$$

$$q(x_s|x_t, x_0) = \frac{q(x_t|x_s)\, q(x_s|x_0)}{q(x_t|x_0)}$$

$$= \frac{cat(Q_{t|s}\, x_t)\, cat(Q_s^T x_0)}{cat(Q_t^T x_0)}$$

$$= cat\left(\frac{Q_{t|s}\, x_t \odot Q_s^T x_0}{x_t^T Q_t^T x_0}\right)$$

Model: given x_t it predicts x_0 which is put in forward process.

$$x_0 \longrightarrow f_\theta(x_t, t)$$

obj: $D_{KL}\left(q(x_s|x_t, x_0) \,\|\, P(x_s|x_t)\right) \quad s<t$

$$P(x_s|x_t) = q(x_s|x_t, f_\theta(x_t, t))$$

$$P(x_s|x_t \neq m) = cat(x_t)$$

$$P(x_s|x_t = m)$$

$$= cat\left(\frac{\alpha_t m \odot x_\theta + (\alpha_s - \alpha_t) x_\theta + (1-\alpha_s) m}{\alpha_t \langle x_\theta, m\rangle + 1 - \alpha_t}\right)$$

$$L(\theta) = \sum_{x_s} q(x_s|x_t, x_0) \log \frac{q(x_s|x_t, x_0)}{P(x_s|x_t)}$$

if $x_t = x \quad L_1(\theta) = 0 \quad \because P(x_s|x_t) = q(x_s|x_t, x_0)$

$$= cat(x_t)$$

if $x_t = m$

$$L_2(\theta) = \sum_{x_s} q(x_s | x_t, x_0) \log \frac{q(x_s | x_t, x_0)}{p(x_s | x_t)}$$

$$= \frac{\alpha_s - \alpha_t}{1 - \alpha_t} \log \frac{\alpha_t \langle x_\theta, m \rangle + (1 - \alpha_t)}{(1 - \alpha_t) \langle x_\theta, x \rangle}$$

$$+ \frac{1 - \alpha_s}{1 - \alpha_t} \log \frac{(1 - \alpha_s)(\alpha_t \langle x_\theta, m \rangle) + (1 - \alpha_t)}{(1 - \alpha_t)(\alpha_s \langle x_\theta, m \rangle + (1 - \alpha_s)}$$

$$L(\theta) = E_t \, E_{q(x_t|x)} \left(\cancel{L_1(\theta) \langle x_t, x \rangle}^{0} + L_2(\theta) \langle x_t, m \rangle \right)$$

$$= E_{t \sim \{\frac{1}{T} \cdots 1\}} E_{q(x_t|x)} \left[T \, L_2(\theta) \langle x_t, m \rangle \right]$$

SUBS parameterization

① zero-masking probs:

observing $\langle x, m \rangle = 0$ we put $\langle x_\theta, m \rangle = 0$

② carry-over unmasking:

$x_\theta(x_t, t) = x_t$ if $x_t \neq m$.

(i.e predict only for the masked tokens

subs. in $L(\theta)$ to get

$$L(\theta) = E_{t \sim \{\frac{1}{T} \cdots 1\}} \, x_t \sim q_t \left[T \left[\frac{\alpha_s - \alpha_t}{1 - \alpha_t} \log \langle x_\theta, x \rangle \right] \right]$$

↓ Rao-Blackwellized bound

↑ C.E over masked x_ts

Flex-MDLM :

unmasking posterior $\rightarrow P(x_s | x_t = m)$

Insertion expectation $\rightarrow$ exp. number of tokens between x_{i-1} & x_i $\forall i$

$$L_\theta = L_{MDLM} + L_{insertion}$$

↓

unmasking loss

use Bregman divergence for ins. loss

$$L_{MDLM} = E_t \, E_{q(x_t|x)} \left[\frac{\dot{\beta}_t}{1-\beta_t} \sum_{i \in \{x_t^i = m\}} \log \left(f_\theta^i(x_t, t), x_0^i \right) \right]$$

$$L_{ins} = E_t \, E_{q(x_t|x)} \left[\frac{\dot{\alpha}_t}{1-\alpha_t} \sum_{i=0}^{len(x_t)} \phi\left(s_t(i) - s_t(i-1) - 1, \; g_\theta^i(x_t, t) \right) \right]$$