Generative Models: GANs and NFs

$\{x_1, x_2, \ldots, x_N\}$ – dataset

Goal: $x \sim p(x)$

$z \in \mathbb{R}^d$, $z \sim p(z) = \mathcal{N}(z \mid 0, I)$

$X = g(z, \theta)$

$p(x \mid \theta)$

ML: $p(x_1, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta) \to \max_\theta$

$\int f(x)\delta_a(x)\,dx = f(a)$

$p(x \mid \theta) =$

$= \int \delta_x\big(g(z,\theta)\big)\,p(z)\,dz$

$$\delta_a(x) = \begin{cases} +\infty, & \text{if } x = a \\ 0, & \text{if } x \neq a \end{cases}$$

Discriminator $d(x) \begin{cases} \to \text{"real"} \\ \to \text{"fake"} \end{cases}$ $d(x) \in (0, 1)$

Fixed g : $\mathbb{E}_{p(x)} \log d(x) + \mathbb{E}_{q(x)} \log(1 - d(x)) \to \max_d$

data distr

generator distr

$$V(g, d) = \mathbb{E}_{p(x)} \log d(x) + \mathbb{E}_{p(z)} \log(1 - d(g(z, \theta))) \longrightarrow \max_d \min_g$$

GAN scheme

$g(z, \theta)$
$d(x, \lambda)$
Init. λ, θ
Iterations until convergence:

maybe several iters.

$x_1, x_2, \ldots, x_M \sim p_{data}(x)$
$z_1, z_2, \ldots, z_M \sim p(z) = \mathcal{N}(z|0, I)$
$$\mathcal{L}_\lambda = \frac{1}{M}\sum_{j=1}^{M}\left[\log d(x_j, \lambda) + \log(1 - d(g(z_j, \theta), \lambda))\right]$$
$\lambda = \text{opt\_step}(\nabla_\lambda \mathcal{L}_\lambda)$

maybe several iters.

$z_1, z_2, \ldots, z_M \sim p(z)$
$$\mathcal{L}_\theta = \frac{1}{M}\sum_{j=1}^{M}\log(1 - d(g(z_j, \theta), \lambda))$$
$\theta = \text{opt\_step}(\nabla_\theta \mathcal{L}_\theta)$

Fix g: $\int p(x) \log d(x)\, dx + \int q(x) \log(1 - d(x))\, dx \to \max_d$

Fix x: $p \cdot \log d + q \log(1-d) \to \max_{d \in (0,1)}$

$$\frac{\partial}{\partial d} = \frac{p}{d} - \frac{q}{1-d} = 0 \Rightarrow p(1-d) = dq$$

$$\Rightarrow d = \frac{p}{p+q}$$

$$d_{opt}(x) = \frac{p(x)}{p(x)+q(x)}$$

Fix $d_{opt}(x) = \frac{p(x)}{p(x)+q(x)}$

Hypothesis: $q_{opt}(x) = p(x)$

$$V(q_{opt}, d_{opt}) = \int p(x) \log \frac{p(x)}{p(x)+q(x)} dx + \int q(x) \log \frac{q(x)}{p(x)+q(x)} dx =$$

$$= \{ q(x) = p(x) \} = \int p(x) \log \frac{1}{2} dx + \int q(x) \log \frac{1}{2} dx =$$

$$= 2 \log \frac{1}{2}$$

$$V(q, d_{opt}) - V(q_{opt}, d_{opt}) =$$

$$= \int p(x) \log \frac{p(x)}{p(x)+q(x)} dx + \int q(x) \log \frac{q(x)}{p(x)+q(x)} dx - 2 \log \frac{1}{2} =$$

$$= \int p(x) \log \frac{p(x)}{\frac{1}{2}(p(x)+q(x))} dx + \int q(x) \log \frac{q(x)}{\frac{1}{2}(p(x)+q(x))} dx =$$

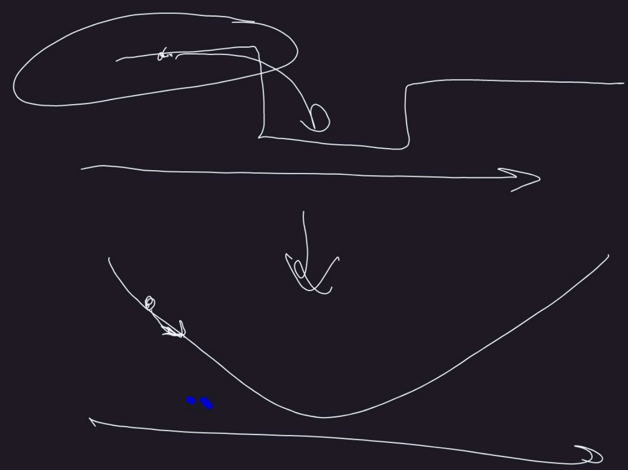
$$= KL\left(p(x) \,\|\, \frac{1}{2}(p(x)+q(x))\right) + KL\left(q(x) \,\|\, \frac{1}{2}(p(x)+q(x))\right) = JS(p(x) \| q(x)) \geq 0$$

$$\Rightarrow q_{opt}(x) = p(x)$$

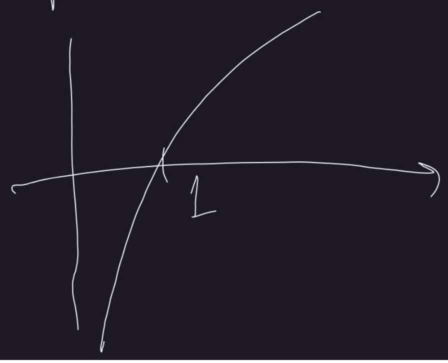
Trick 1: use diff. losses for disc. and generator

$$\log(1 - d(g(z))) \approx 0$$

$$d(g(z)) \approx 0$$



$$\mathbb{E}_{p(z)} \log d(g(z)) \to \max_g$$



$p(x|y)$ $g(z, y)$

$d(x, y)$

Trick 2: make discr. not so certain

- label smoothing $1 \to 0.7$, $1 \to 0.3$

$$\log d(x) \longrightarrow 0.7 \log d(x) + 0.3 \log(1 - d(x))$$

- label swapping
- add supervised loss $\log p(y = y_i | x_i) + \log p_{real}(x_i) + \log(1 - p_{real}(x_{fake}))$

$\{x_i, y_i\}$ $d(x) = [p(y=1|x), \ldots, p(y=k|x), p_{real}(x)]$

- use LeakyReLu instead ReLu, AvgPool instead Max-Pool

ELU conv. with strides

Trick 3: add random noise to feature maps in generator

mode-collapse

$$x \sim p(x)$$

$$\cancel{\text{comp } p(x)}$$

$$z \sim p_z(z) = \mathcal{N}(z \mid 0, I)$$

$$x = f(z)$$

$$z = f^{-1}(x)$$

flow: $z_0 \sim p_0(z_0) = \mathcal{N}(z_0 \mid 0, I)$

$$z_1 = f_1(z_0)$$

$$z_2 = f_2(z_1)$$

$$\cdots$$

$$x = z_K = f_K(z_{K-1})$$

$$p_x(x) = p_z\Big(\overbrace{f^{-1}(x)}^{z}\Big) \cdot \left| \det \frac{\partial f^{-1}(x)}{\partial x} \right| =$$

$$= p_z(z) \left| \det \frac{\partial f(z)}{\partial z} \right|^{-1}$$

dim x × dim x
jacobian

$$q_0(z_0) = p_0(z_0)$$

$$q_1(z_1) = p_0(z_0)\left|\det \frac{\partial f_1}{\partial z_0}\right|^{-1}$$

$$q_2(z_2) = q_1(z_1)\left|\det \frac{\partial f_2}{\partial z_1}\right|^{-1} = p_0(z_0)\left|\det \frac{\partial f_1}{\partial z_0}\right|^{-1} \cdot \left|\det \frac{\partial f_2}{\partial z_1}\right|^{-1}$$

$$\ldots$$

$$\log q_K(z_K) = \log p_0(z_0) - \sum_{k=1}^{K} \log\left|\det \frac{\partial f_k}{\partial z_{k-1}}\right|$$

NF scheme

$f_K(z, \theta)\ \forall K$

Init. θ

Iterations until convergence:

$x_1, x_2, \dots, x_M \sim p(x)$

for each x_i:

inversed flow: $z_K^i = x_i,\ z_{K-1}^i = f_K^{-1}(z_K^i), \dots, z_0^i = f_1^{-1}(z_1^i)$

$$L_\theta = \frac{1}{M} \sum_{i=1}^{M} \left(\log p_0(z_0^i) - \sum_{K=1}^{K} \log \left| \det \frac{\partial f_K}{\partial z_{K-1}} \right|_{z_{K-1} = z_{K-1}^i} \right)$$

$\theta = \text{opt\_step}(\nabla_\theta L_\theta)$

Classical model: $f(z) = z + u\,h(w^T z + b)$

$u, w \in \mathbb{R}^{\dim z}$

$b \in \mathbb{R}$, h – scalar func.

$$\det \frac{\partial f}{\partial z} = 1 + \underset{\mathbb{R}}{(w^T u)} \cdot h'(w^T z + b)$$

$$df = \frac{\partial f}{\partial z} \cdot dz = dz + u\, h'(w^T z + b)\, w^T dz$$

$$\Rightarrow \frac{\partial f}{\partial z} = I + u \cdot \underbrace{h'(w^T z + b)}_{\mathbb{R}} \cdot w^T$$

$$\det(\underset{A}{I} + \underset{U}{u} \cdot \underset{C}{h'} \cdot \underset{V}{w^T}) = 1 \cdot h' \cdot \det\left((h')^{-1} + w^T \cdot u\right) =$$

$$= h'\left(\frac{1}{h'} + w^T u\right) = 1 + w^T u \cdot h'$$

$$A + UCV$$

$n \times n \quad n \times m \quad m \times m \quad m \times n$

$m << n$

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(\underbrace{C^{-1}+VA^{-1}U}_{m \times m})^{-1}VA^{-1}$$

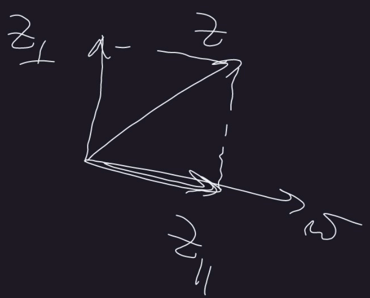
Woodbury identity

$$\det(A+UCV) = \det A \cdot \det C \det(C^{-1}+VA^{-1}U)$$

determinant lemma

$$y = f(z) = z + u\,h(w^T z + b)$$

$$z = f^{-1}(y) - ?$$



$$z = z_\perp + z_\parallel \quad , \quad z_\perp^T w = 0$$

$$z_\parallel = \text{const} \cdot w$$

$$y = z_\perp + z_\parallel + u\,h\left(w^T(z_\perp + z_\parallel) + b\right)$$

$$\Rightarrow z_\perp = y - z_\parallel - u\,h(w^T z_\parallel + b)$$

$$z_\parallel = \frac{\alpha}{\|w\|^2} \cdot w$$

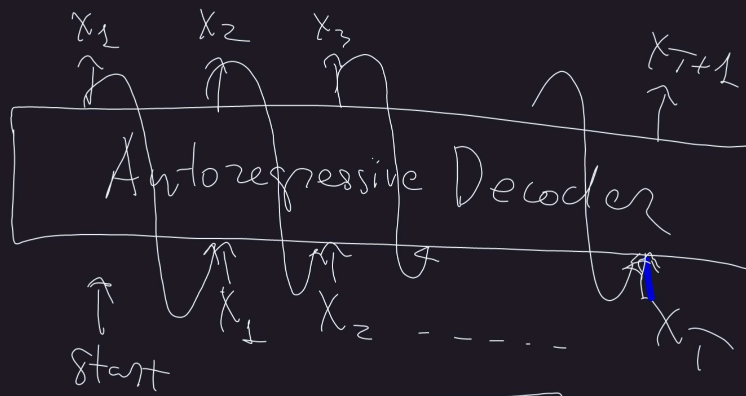
$$y^T w = \cancel{w^T} \cdot \frac{\alpha}{\cancel{\|w\|^2}} \cdot \cancel{w} + (w^T u)\,h(\underbrace{z_\parallel^T w}_{\alpha} + b)$$

$$y^T w = \alpha + (w^T u) \cdot h(\alpha + b)$$

$$\alpha = \text{solution}(y^T w, w^T u, b)$$

Inversed Autoregressive Flow for distillation of autoregr. generative model

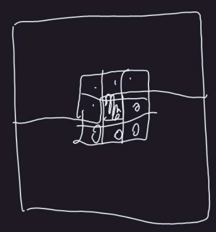
GPT



$$p(x_{1:T+1} \mid x_{0:T}) \text{ — fast}$$

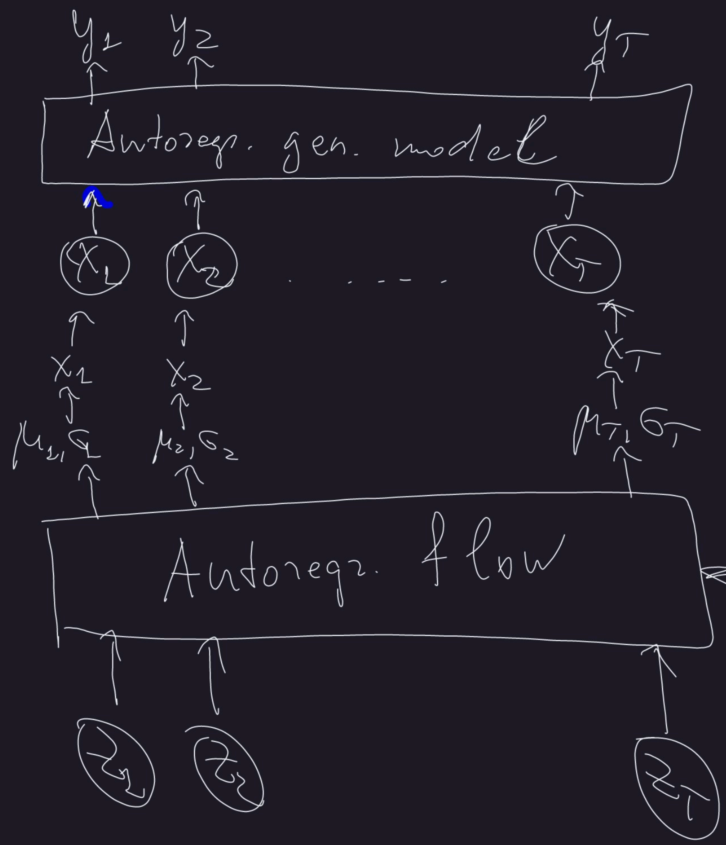
$$x \sim p(x) \text{ — slow}$$

PixelCNN



$$p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}) =$$

$$= \prod_{t=1}^{T} \mathcal{N}\left(x_t \mid \mu_t(x_{1:t-1}, \theta), \sigma_t(x_{1:t-1}, \theta)\right)$$



$$p(x_2 \dots x_{T+1} \mid x_1 \dots x_T)$$

$p_T(x)$ – teacher model

$$x_t \sim \mathcal{N}\big(x_t \mid \mu_t(z_1, \dots z_{t-1}, \theta), \sigma_t(z_1, \dots z_{t-1}, \theta)\big)$$

$$x \sim \mathcal{N}(x \mid \mu, \sigma^2) \iff$$

$$z \sim \mathcal{N}(z \mid 0, 1), \quad x = \sigma \cdot z + \mu$$

$$\boxed{x_t = \sigma_t(z_{1:t-1}, \theta) \cdot z_t + \mu_t(z_{1:t-1}, \theta)}$$

$$z_t \sim \mathcal{N}(z_t \mid 0, 1)$$

$$x = f(z):$$

$$x_t = \sigma_t(z_{1:t-1}) \cdot z_t + \mu_t(z_{1:t-1}) \qquad \forall t$$

$$x_1 = \sigma_1 \cdot z_1 + \mu_1 \quad \Rightarrow \quad z_1 = \frac{x_1 - \mu_1}{\sigma_1}$$

$$x_2 = \sigma_2(z_1) \cdot z_2 + \mu_2(z_1) \quad \Rightarrow \quad z_2 = \frac{x_2 - \mu_2(z_1)}{\sigma_2(z_1)}$$

$$\cdots$$

$$\det\left(\frac{\partial x}{\partial z}\right) = \prod_{t=1}^{T} \frac{\partial x_t}{\partial z_t} = \prod_{t=1}^{T} \sigma_t(z_{1:t-1}, \theta)$$

$$\parallel$$

$$\begin{pmatrix} & 0 \\ & \end{pmatrix}$$

$$\log\left|\det \frac{\partial x}{\partial z}\right| = \sum_{t=1}^{T} \log \sigma_t(z_{1:t-1}, \theta)$$

$p_T(x)$ – teacher

$p_S(x|\theta)$ – student

$$KL\left(p_S(x|\theta) \,\|\, p_T(x)\right) \to \min_\theta$$

$$\mathbb{E}_{p_S(x|\theta)} \log \frac{p_S(x|\theta)}{p_T(x)} = \mathbb{E}_{p_S(x|\theta)} \underbrace{\log p_S(x|\theta)} - \mathbb{E}_{p_S(x|\theta)} \log p_T(x)$$

$$\mathbb{E}_{p_S(x|\theta)} \log p_S(x|\theta) = \mathbb{E}_{p_S(x|\theta)} \sum_{t=1}^{T} \log \mathcal{N}\left(x_t \,|\, \mu_t(z_{<t},\theta), \sigma_t(z_{<t},\theta)\right) =$$

$$= \sum_{t=1}^{T} \mathbb{E}_{p_S(x|\theta)} \left(-\frac{1}{2}\log 2\pi - \log \sigma_t(z_{<t},\theta) - \frac{1}{2\sigma_t^2}(x_t-\mu_t)^2\right) =$$

$$= \sum_{t=1}^{T} \left(-\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma_t(z_{<t}, \theta) - \frac{1}{2\sigma_t^2} \underbrace{\mathbb{E}_{p_s(x|\theta)} \left(x_t^2 - 2 x_t \cdot \mu_t + \mu_t^2 \right)}_{=} \right) =$$

$$\underbrace{\mathbb{E} x_t^2}_{= \sigma_t^2 + \mu_t^2} - 2 \underbrace{\mathbb{E} x_t}_{= \mu_t} \cdot \mu_t + \mu_t^2 = \sigma_t^2$$

$$= \sum_{t=1}^{T} \left(-\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma_t(z_{<t}, \theta) - \frac{1}{2} \right)$$

$$\mathbb{E}_{p_S(x|\theta)} \log p_T(x) = \mathbb{E}_{p_S(x|\theta)} \sum_{t=1}^{T} \log p_T(x_t | x_{<t}) =$$

$$= \sum_{t=1}^{T} \mathbb{E}_{p_S(x|\theta)} \log p_T(x_t | x_{<t}) = \sum_{t=1}^{T} \mathbb{E}_{p_S(z)} \log p_T\left(\sigma_t(z_{<t},\theta) \cdot z_t + \mu_t(z_{<t},\theta) \;\middle|\; \sigma_{<t} \circ z_{<t} + \mu_{<t}\right)$$

$$z_1, \ldots, z_M \sim \mathcal{N}(z | 0, I)$$

$$L_\theta = \frac{1}{M} \sum_{i=1}^{M} \log p_T\left(\sigma_t(z_i) \cdot z_i + \mu_t(z_i) \;\middle|\; \ldots\ldots \right)$$