

$$\text{EBM} \quad p(x) = \frac{e^{-f(x)}}{Z \text{ unknown}}$$

Training: various are contrastive.

$$f = \arg \min_{\substack{x \sim x^+ \\ x \sim x^-}} \left| \mathbb{E}_{x \sim x^+} f(x) - \mathbb{E}_{x \sim x^-} f(x) \right|$$

normalized

Sampling MCMC

$$x_0 \sim p_0 \quad x_{t+1} \sim h(\cdot | x_t)$$

s.t. p is stationary for M

Detailed Balance:

$$\forall x, x' \quad p(x) p_r[x \rightarrow x'] = p(x') p_r[x' \rightarrow x]$$

Lemma: $DB \Rightarrow p$ is stationary

$p \in$ if $x_t \sim p$

$$P_r[x_{t+1} = x'] = \sum_x p(x) P[x \rightarrow x']$$

$$= \sum_{x'} p(x') P[x' \rightarrow x]$$

$$= p(x') \underbrace{\sum_{\substack{x \\ \text{---}}} P[x' \rightarrow x]}_1$$



Example: Metropolis Hastings

Symmetric dist $D_{x,x'}$ ($\sum_{x'} D_{xx'} = 1$)

Given x_t : (1) pick $x' \sim D_{x,x'}$ $\ell^{f(x') - f(x)}$

(2) accept w.p. $\min \left\{ 1, \frac{p(x')}{p(x)} \right\}$

Lemma: MH satisfies DB

PF: $p(x) P[x \rightarrow x'] = p(x) D_{x,x'} \min \left\{ 1, \frac{p(x')}{p(x)} \right\} = \dots$

Example: Langevin Dynamics

Well: $\nabla \log p = -\nabla f$ remember: $e^{-\|x - \mu\|^2/2}$
 $p(x) = e^{-\|x - \mu\|^2/2}$

$$X_{t+1} = X_t + \eta \nabla \log p + \sqrt{2\eta} N(0, I)$$

Lemma: LD satisfies DB (uniqueness)

PF:

$$\Pr[X \rightarrow x + \delta] = \Pr[N(\eta \nabla \log p, \sqrt{2\eta} I) = f]$$

$$= e^{-\|\delta + \eta \nabla \log p\|^2 / 2 \cdot 2\eta}$$

$$= e^{\frac{-\langle \delta, \eta \nabla \log p \rangle}{2\eta} + \frac{\|\delta\|^2}{4\eta} + \frac{\eta \|\nabla \log p\|^2}{4}}$$

$$\Pr[X \rightarrow x + \delta] = \frac{-\langle \delta, \nabla \log p \rangle}{2} + \frac{\|\delta\|^2}{4\eta} + \frac{\eta \|\nabla \log p\|^2}{4}$$

$$p(x + \delta) \approx e^{\log p(x) + \langle \delta, \nabla \log p \rangle}$$

$$p(x) \Pr[x \rightarrow x+\delta] \approx e^{\log p(x) + \frac{\langle \delta, \nabla \log p \rangle}{\varepsilon}}$$

$$p(x+\delta) \Pr[x+\delta \rightarrow x] \approx e^{\log p(x) + \langle \delta, \nabla \log p \rangle - \frac{\langle \delta, \nabla \log p \rangle}{\varepsilon}}$$

Stochastic Preference Equation (SPE)

$$\frac{dx}{dt} = -\nabla f + \frac{dB}{dt}$$

Why: energy, temperature

Statistical Physics

prob dist over states

System state $x \in \mathcal{R}$

Forces on system determine energy $f: \mathcal{R} \rightarrow \mathbb{R}$

f_{ext} is better

Temperature T is tendency of system

to move around increasing entropy $H(p)$

$L(p)$

$$P_* = \arg \min_{x \sim p} [E_f(x) - T \cdot H(p)]$$

internal energy

canonical entropy

Thm (Variational Principle):

$$-\frac{1}{T} \mathbb{E} f(x)$$

p_* is unique dist s.t. $p_*(x) \propto e^{-\frac{1}{T} f(x)}$

$$\text{i.e. } p_*(x) = \frac{e^{-\frac{1}{T} f(x)}}{\sum_I} = e^{-\frac{1}{T} f(x) - A_T}$$

\downarrow
partition function

\log partition
function

free energy

$$\text{Lemma: } T \cdot A_T = -\mathcal{L}(p_*) = T \cdot H(p_*) - \mathbb{E}_{x \sim p_*} f(x)$$

$$\text{Pf: } H(p^*) = -\mathbb{E}_{x \sim p^*} \log p^*(x) = \frac{1}{T} \mathbb{E}_{x \sim p^*} f(x) + A_T$$

free energy + internal energy \rightleftharpoons canonical entropy

$$\text{NOTE: } \beta = \frac{1}{T} \quad \frac{dA_\beta}{d\beta} = -\mathbb{E} f(x)$$

$\beta \rightarrow \infty$ system cools

$$(\text{cool } \nabla f = \nabla (\log A \cdot f))$$

Thm from Lemma⁰

For every dist q

$$0 \leq \mathbb{E}_{x \sim q} \text{KL}(q \| p^*) = \mathbb{E}_{x \sim q} \log \frac{q(x)}{p^*(x)} = \mathbb{E}_{x \sim q} \log q(x) - \underbrace{\mathbb{E}_{x \sim q} \log p(x)}_{= f(x) - A_T} =$$

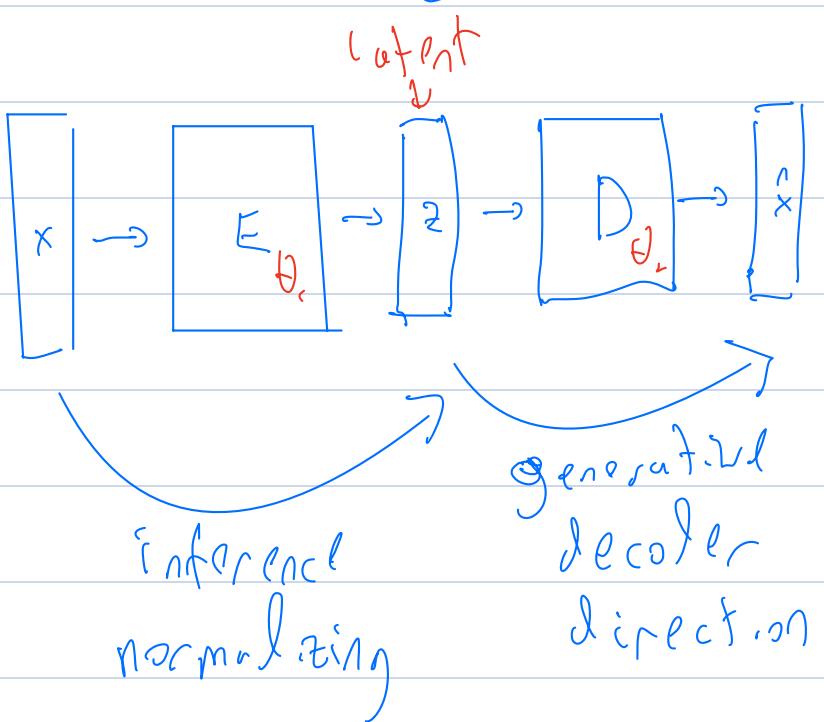
$$-\mathbb{E}_{x \sim q} H(q) + \mathbb{E}_{x \sim q} f(x) + T \cdot A_T \geq 0$$

$$\lambda(q) - \lambda(p^*) \geq 0$$

equal 0 iff $\text{KL}(q \| p^*) = 0$



Auto Encoders



WANT TO TRAIN E_{θ}, D_{θ} s.t.

$$(1) \quad D(E(x)) \approx x \quad (\text{reconstruction})$$

$$(2) \quad z = E(x) \quad \text{"interesting" (inference)}$$

$$(3) \quad D(\mu(\theta, I)) \approx p \quad (\text{generalization})$$

or another "nice" dist

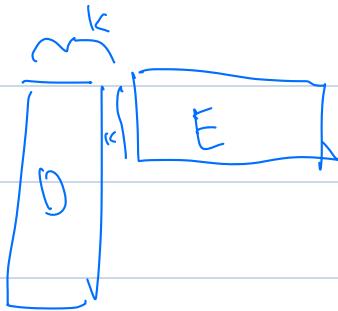
P: empirical dist

minimum: rule out $E, D = \text{identity}$

Example: E, D linear

restriction is $\dim(z) \leq k$

$$\arg \min_{x \in E} \|Dx - z\|^2$$



$$\arg \min_{x \in L} \|Lx - z\|^2 =$$

L

rank k

$$E \|Lx - z\|_V^2 + E \|x\|_{V^+}^2$$

optimized by $L = I_{\{v_1, \dots, v_n\}}$
top e-space

Generative model

Given samples $x_1, \dots, x_n \sim p$

build q_θ to minimize

compute

$$[LL(p || q_\theta)] = \underbrace{\mathbb{E}_{x \sim p} [\log p(x)] - \mathbb{E}_{x \sim p} [\log q_\theta(x)]}$$

~~Unknown p but independent of θ~~

get samples can optimize
if q_θ computable

log likelihood
cross entropy

VAE Two equivalent views

(1) combine (a) $\min \mathbb{E} \|x - \hat{x}\|^2$

$$(b) \text{KL}(D(x) \parallel N(0, I))$$

(2) use estimate for $\mathbb{E}_{\hat{x} \sim p} \log q_{\theta}(\hat{x})$

specifically

$$\text{arg} \max \text{ELBO}(\theta), \text{ELBO}(\theta) \leq \mathbb{E}_{\hat{x} \sim p} \log q_{\theta}(\hat{x})$$

ELBO Thm: If $q_{\theta} = D(N(0, I))$ then

$$\log q_{\theta}(x) \geq - \text{KL}(E(x) \parallel N(0, I)) + \mathbb{E}_{z \sim E(x)} \left[\log P(D_2 | z) \right]$$

Divergence Reconstruction

$$\lambda(z) := \frac{P_r[D(z)=x]}{D} \quad q(z) = \frac{N(z)\lambda(z)}{\frac{P_r[D(z)=x]}{D}} \quad P_r[q_G(U)=x]$$

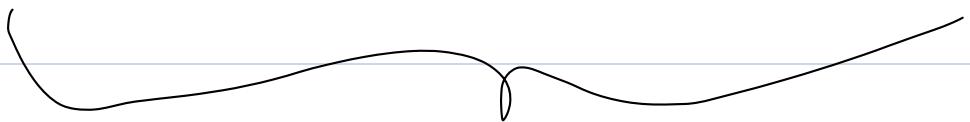
$$PF; \quad 0 \leq KL\left(E(x) \parallel \underset{z \sim D(z)}{\underset{q}{\mathbb{E}}} \right) = \underset{z \sim E(x)}{\mathbb{E}} \log \frac{P(E(x) > z)}{q(z)}$$

$$= \underset{z \sim E(x)}{\mathbb{E}} \log P(E(x) > z) - \log N(z) \sim \log \lambda(z)$$

$$+ \log q_\theta(x)$$

$$0 \leq KL\left(E(x) \parallel N(0, I)\right) - \underset{z \sim E(x)}{\mathbb{E}} \log P_r[D(z)=x] + \log q_\theta(x)$$

$$\log q_\theta(x) \geq \underset{z \sim D(E(x))}{\mathbb{E}} \log P_r[D(E(x))=x] - KL\left(E(x) \parallel N(0, I)\right)$$



ELBO



Issues with VAE

- (1) mode collapse
- (2) global objective
- (3) force $N(0, I)$ $\longrightarrow p$
(not necessarily blurry,
see GANs)

Normalizing Flows

Find T_θ s.t. can compute exactly

$$q_\theta(x) := P_{T_\theta(N(0, I))}(x)$$

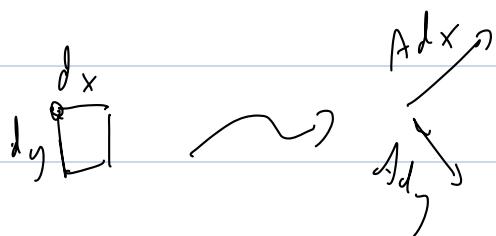
$$\arg \max_{\theta} \mathbb{E}_{\mathcal{D}} \log q_\theta(x)$$

Exercise: Suppose J is invertible linear
transformation

$$q = J(N(0, I))$$

Then for every x

$$q(x) = \frac{N(x)}{\det J}$$



If f is differentiable then locally,

f is linear

$$f(x+\delta) \approx f(x) + J\delta$$

;

J is Jacobian
i.e. $\begin{pmatrix} \frac{\partial F_i}{\partial x_j} \end{pmatrix}$

Thm: If p is dist $q = F(p)$

then for every y

$$p(y) = \frac{q(x)}{\det JF(x)}$$
 where $x = F^{-1}(y)$

Joint Sequence of transformations

$$f_0 \ f_1 \ f_2 \ , \ \sim \sim \sim$$

$$P_0 = N(0, I)$$

$$P_{t+1} = f_t(P_{t+k})$$

$$\log P_{t+1}(x_{t+1}) = \log P_t(x_t) - \log \det \left(\frac{\partial f_t}{\partial x} \right)$$

VDE

$$\frac{d \log P_t}{dt} = - \log \det \left(\frac{\partial f}{\partial x} \right)$$

$$= - \text{Tr} \left(\log \frac{\partial f}{\partial x} \right)$$

Making transformations invertible

(See also quantum, crypto)

CLAIM: For every $f: \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$

The map $g(x_1, x_2) = (x_1, x_2 + f(x_1))$
is invertible.

$$J_g = \begin{pmatrix} x_1 & x_2 \\ I & 0 \\ x_2 Jf & I \end{pmatrix}$$

More generally $f: \mathbb{R}^{d_1} \rightarrow$ invertible
linear
transforms
on \mathbb{R}^{d_2}

$$g(x_1, x_2) = (x_1, f(x_1) x_2)$$

$$J_g = \begin{pmatrix} I & 0 \\ * & f(x_1) \end{pmatrix}$$

Auto regressive flow

$$d_1 = 1, 2, 3,$$

$$d_2 = 1$$

CONS ~ Force "unnatural" structure

- training painful

- can't "skip steps"
in inference

Teaser: Diffusion

Flow + EBM

$P_1 P_2 P_3 -$ —

$P_{+1} \approx$ Langevin