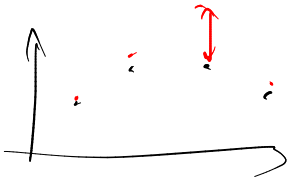
Choosing physically-meaningful metrics

- regression : $\mathcal{D} = \{(x_i, y_i)\}$

L^2 loss : $\sum_{i \in \mathcal{D}} \| \hat{y}_i - y_i \|^2$

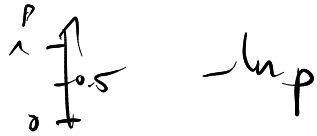
$\hat{y}_i$ ← prediction ; y_i ← target label



⚠ sensitive to outliers

(vs L^2 loss : $\sum_{i \in \mathcal{D}} |\hat{y}_i - y_i|$ → inducing sparsity → on weights rather than on outputs)

- classif : cross-entropy



↓

Kullback-Leibler divergence

$$KL(p||q) = \int_{\mathcal{X}} p \log \frac{p}{q} \rightarrow \sum_c p_c \log \frac{p_c}{q_c}$$

true label → $\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ ← your prediction

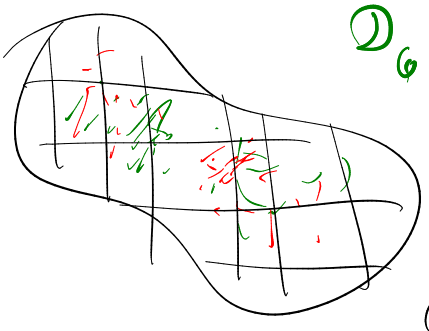
↳ $-\log q_{\text{true class}}$

⚠ not a distance

not symmetric

⚠ on a continuous space

$q = \begin{pmatrix} 0.1 \\ 0.2 \\ 0.01 \\ 0.5 \\ \vdots \end{pmatrix}$



$\mathcal{D}_\theta$ generated distrib°

$\mathcal{D}$ target distrib° (training set)

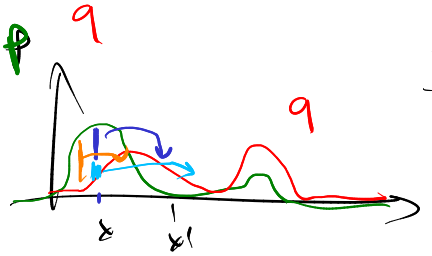
- no notion of geometry
- if $q=0$ for some point where p is not → too cost

in practice : estimations & estimate density

⚠ high-dim setting

Optimal transport

- Earth Mover distance
- Wasserstein
- Monge-Kantorovic

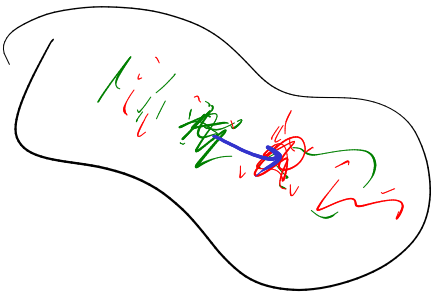


proba distrib°
- mass = 1
- ≥ 0

$$\inf_{M \to} C(m) = \int_{x'} \int_{x} \underbrace{m_{x \to x'}}_{\text{mass effort}} \|x - x'\| \, dx \, dx'$$

amount of work to do

constraints : $\int_x m_{x \to x'} dx' = p(x)$ & $\int_{x'} m_{x \to x'} dx = q(x')$

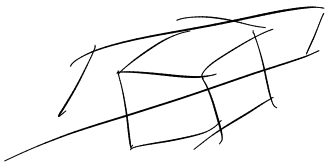


- much more robust to displacements than KL
- maths: distance

X

1D : closed-form solution

higher-dims : ?



approx° ←

regularized by entropy

$$O.T._{\varepsilon}(p,q) + \varepsilon H(\pi)$$

↑ small value

↳ efficient algo. to compute the global optimum

Sinkhorn algo

→ consists in a sequence of matrix-vector products

⚠ Sense in high-dims?

Minimum Mean Discrepancy (MMD)

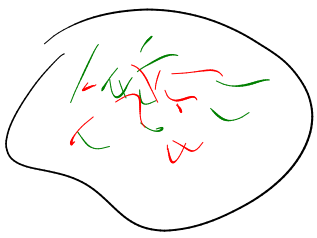
- compare statistics over distrib°

mean : $\mu_p = \mathbb{E}_{x \sim p}[x]$

$\mu_q = \mathbb{E}_{x \sim q}[x]$

$\|\mu_p - \mu_q\|$

$\|\mu_p^2 - \mu_q^2\|$



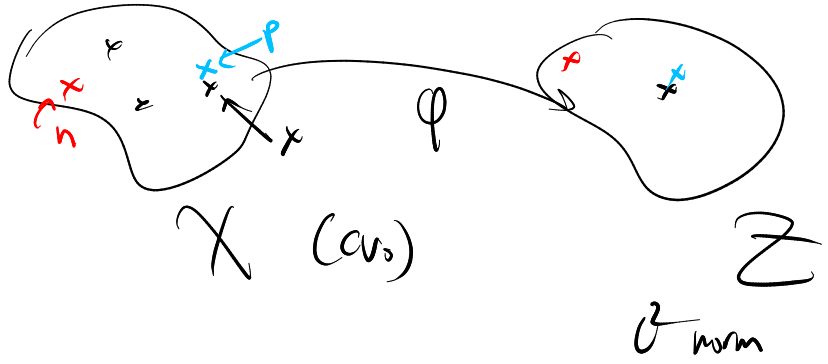
∃ high-dims

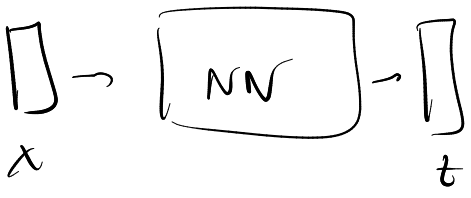
choose represent° space (features) that are very rich

↳ kernelize $k(x_1, x_2) = \phi(x_1) \cdot \phi(x_2)$ ∃ ϕ

↳ full discriminative power

Metrics learning [Siamese networks, ...]



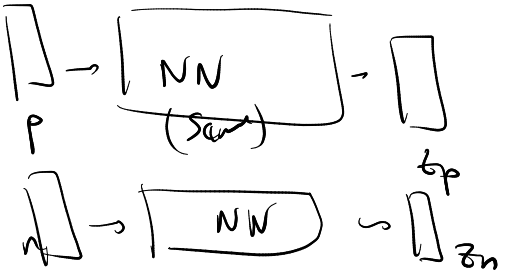


$$d(x_1, x_2) := \| \varphi(x_1) - \varphi(x_2) \|_2$$

Desired property:

$$d(x, p) \leq d(x, n) - m$$

(margin)



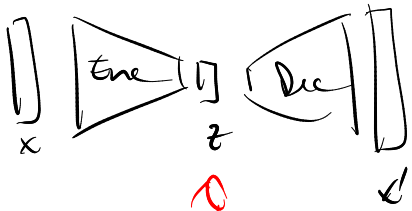
contrastive learning

loss:

$$\sum_{(x, p, n) \sim \mathcal{D}} \max\left(d(z_p) - d(x, n) + m, 0\right)$$

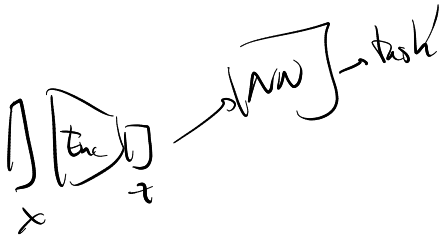
Generative models

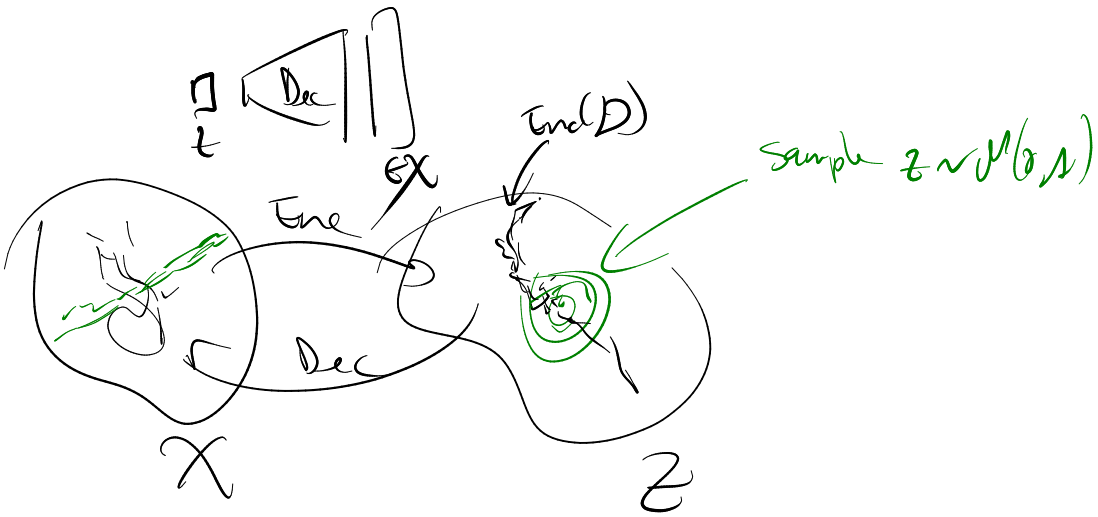
Auto-encoder:



$\|x - x'\|$

low-dim representation space





sample $z \sim \mathcal{N}(0, 1)$

Solve : VAE

Loss : $KL(D \| Dec(\mathcal{N}(0,1)))$ or $KL(Enc(D) \| \mathcal{N}(0,1))$

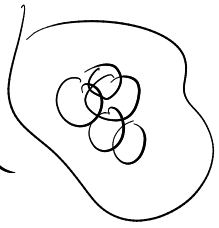
{

estimate through sampling? $\int_p \sim \sum_{i \in D}$ Monte Carlo

ELBO lower bound

$z \to$ NN $\to$ ~~x~~ distribution on $\mathcal{X}$

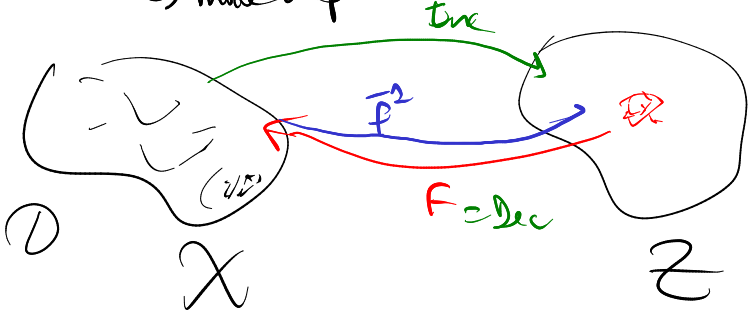
$\hookrightarrow \mathcal{N}(\mu_z, \tau_z) \sim$ sample



$\mathcal{X}$

~~GANs~~ : generated samples to be realistic

↳ mode drop



"normalizing" flows

invertible :

- learn f^{-1} : send D to a $\mathcal{N}$

↑ straightforward to optimize

↓

pros : compute exactly $p_\theta(x)$

↓

$\mathcal{N}$ closed form

↓

KL

Sample $z \sim \mathcal{N}(0,1)$

Goal : $f(\mathcal{N}(0,1)) \sim D$

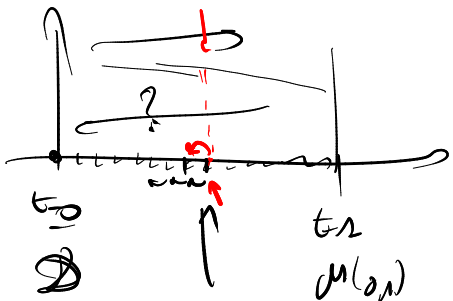
importance sampling ←

Diffusion models

- normalize : explicitly by adding Gaussian noise

diffusion → renormalized $\mathcal{N}(0,1)$

D



$\frac{\partial I_t}{\partial t} = NN_t(I_t)$

PDE

↓

$NN(t)$

↑ attention

NeuralODE

↓

10000 steps

↓

~ 10

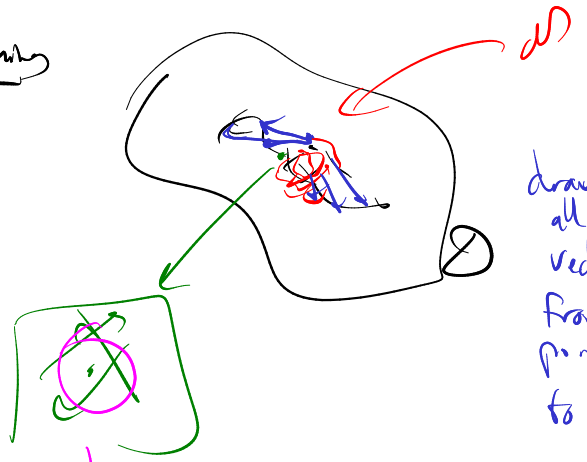
$x_t \to x_{t-1}$

$\hat{x}_0 - x_t$

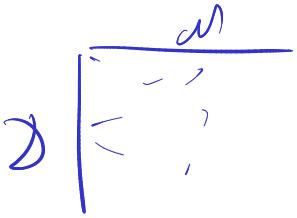
$\|\hat{x} - x_0\|$

↳ simplifies :

Flow matching



- N samples of $\mathcal{M}$ (as many as $\mathcal{D}$) all pairs



draw all vectors from points of $\mathcal{M}$ to points of $\mathcal{D}$

$\Rightarrow$ vector field

average of these vectors (locally)

& follow the flow : $\mathcal{M} \to \mathcal{D}$ th

instead: learn a NN to estimate

$x \longmapsto$ average flow