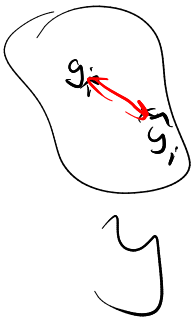
Modeling Tasks & Losses

Choosing physically-meaningful metrics

eg: regression: $\mathcal{D} = \{(x_i, y_i)\} \longrightarrow \sum_i \|\hat{y}_i - y_i\|^2$ MSE



classific°: $\hat{y}_i = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_C \end{pmatrix}$ ← probs distrib° over classes; ↑ sample

target: $\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ ← true class

$p - p'$ $\quad p/p'$

0.0001 ! 0.0100 !

$\simeq 0.01$ ×100 ~ frequency

0.5 ↑ 0.5 ↓

$\Rightarrow -\log p$

$-\log \hat{p} \iff -\log p$

$\log \hat{p}/p$

From information theory

$q = \hat{p}$ $\quad p = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}$

Kullback-Leibler divergence:

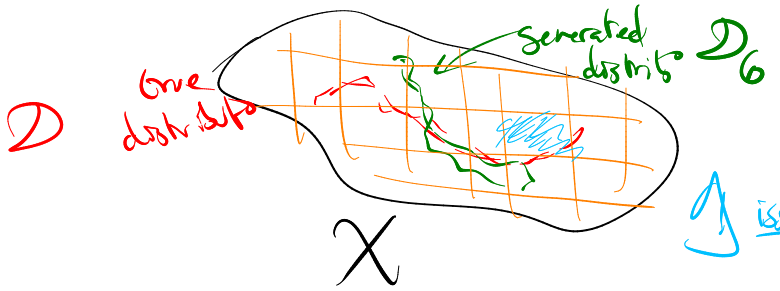
$$KL(p\|q) = \int p \log p/q = \sum_{\text{classes } c} p_c \log p_c/q_c = -\log \hat{p}_{\text{true class}}$$

cross-entropy loss

⚠ not a distance: not symmetric, no triangular inequality

Compare distributions

eg: generative models

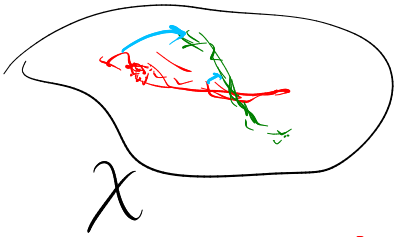


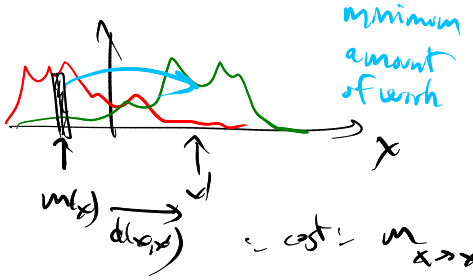
$KL(\mathcal{D}\|\mathcal{D}_\theta)$

$KL(\mathcal{D}_\theta\|\mathcal{D})$

⚠ issue: difficult to estimate / empty bins

Optimal transport



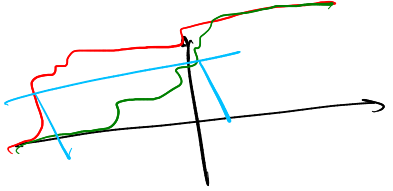


$\Rightarrow$ cost: $m_{x\to x'}\, d(x,x')$

$$\text{optimal transport}(\mathcal{D}_p, \mathcal{D}_q) = \inf_m \iint m_{x\to x'}\, d(x,x')\, dx\, dx' \quad \left(+ \varepsilon H(q)\right)$$

under constraints $\int_{x'} m_{x\to x'}\, dx' = p(x)$ $\quad \int_x m_{x\to x'}\, dx = q(x')$

entropy regularizer ↑



1D: very easy

higher dim: very difficult

↳ [Marco Cuturi, 2013] : Sinkhorn algorithm to solve an approx° (very fast) ← regularized

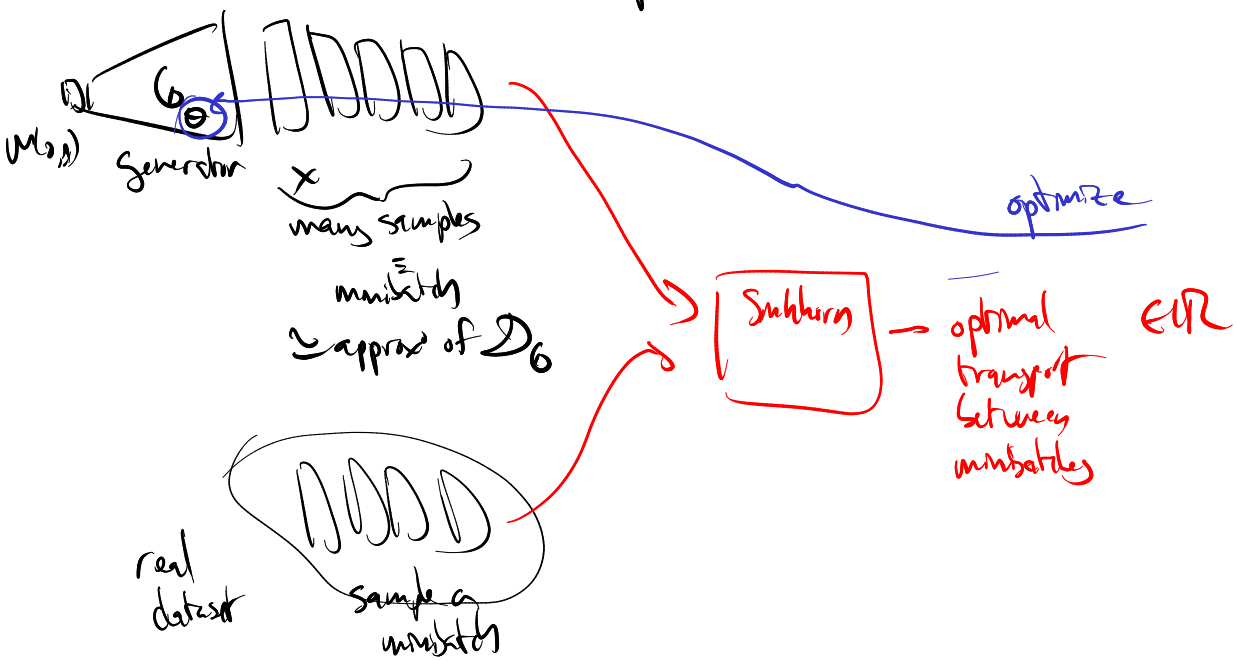
$$K = e^{-d(\cdot,\cdot)/\varepsilon}$$

init°: $u, v = (1, \dots, 1)$

loop ↻ { $u \leftarrow Kv$; $v \leftarrow Ku$

in Pytorch
for a fixed # of iterations
or until convergence

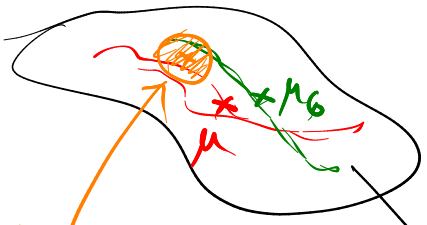
⇒ optimizable by ∇ descent



⚠ Beware: # of samples ∝ exp(intrinsic dimension of $\mathcal{D}$)

Maximum Mean Discrepancy (MMD)

→ to compare statistics of p & q



$\|\mu - \mu_\theta\|^2$ distance between statistics → minimize

$\mu = \mathbb{E}_{x \sim \mathcal{D}(p)}[x]$ $\quad \mu_\theta = \mathbb{E}_{x \sim \mathcal{D}(p_\theta)}[x]$

density around point x_0 → sufficient statistics

φ

$\mathcal{F}$ (higher-dim)

$\|\mu'_\theta - \mu'\|^2$ ↑ very high dim

if φ well chosen ↙ sufficient

Kernel trick: $k(x, x') = \varphi(x) \cdot \varphi(x')$

don't design φ but choose k

similarity

typically: $k(x,x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$

$$\left(\mathbb{E}_{\mathcal{D}}[\varphi] - \mathbb{E}_{\mathcal{D}_0}[\varphi] \right)^2$$

μ μ_0

rewrite using only k