How do we know if our model $p(x)$ is good?

measure "distance" between $p(x)$ and $p^*(x)$ (even if we do not know $p^*(x)$)

1. use as objective function (loss) for training
2. use for validation after training

forward KL divergence (Kullback-Leibler)

$$KL[p^* \| p] := \int p^*(x) \log \frac{p^*(x)}{p(x)} dx = \mathbb{E}_{x \sim p^*(x)} \left[\log \frac{p^*(x)}{p(x)} \right]$$

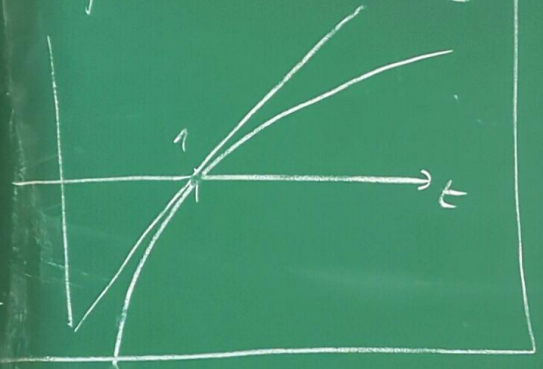
["forward"

$$KL[p^* \| p] = 0 \iff p^*(x) = p(x)$$

$$KL[p^* \| p] \geq 0$$

$$KL = \int p^*(x) \log \frac{p^*(x)}{p(x)} dx \Rightarrow -KL = \int p^*(x) \log \frac{p(x)}{p^*(x)} dx \leq \int p^*(x) \left[\frac{p(x)}{p^*(x)} - 1 \right] dx$$

$\log t \leq t - 1$



$$= \underbrace{\int p(x)\,dx}_{=1} - \underbrace{\int p^*(x)\,dx}_{=1 \text{ (normal.)}}$$

$$= 1 - 1 = 0$$

$$-KL \leq 0 \qquad \boxed{KL \geq 0}$$

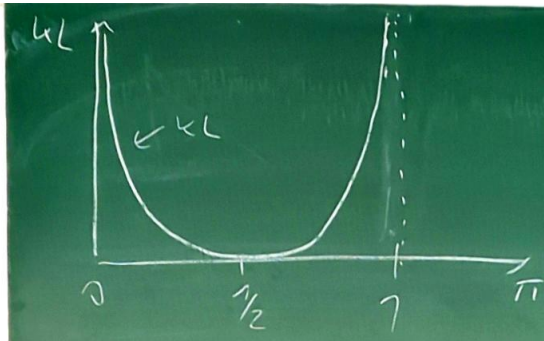
KL is not a distance ($\hat{=}$ metric): not symmetric, no triangular inequality.

$\Rightarrow$ "divergence" $\hat{=}$ distance-like

- <u>example</u> $p^*(x) = \text{discrete}\left(\frac{1}{2}, \frac{1}{2}\right)$ $\quad p(x) = \text{discrete}(\pi, 1-\pi)$

$$KL[p^* \| p] = p^*(-1) \cdot \log \frac{p^*(-1)}{p(-1)} + p^*(+1) \log \frac{p^*(1)}{p(1)}$$

$$= \frac{1}{2} \log \frac{1}{2\pi} + \frac{1}{2} \log \frac{1}{2(1-\pi)} = -\frac{1}{2} \log 4\pi(1-\pi)$$



$$\frac{\partial KL}{\partial \pi} = -\frac{1}{2}\frac{1}{4\pi(1-\pi)}\cdot(4-8\pi) \doteq 0 \qquad \frac{\partial}{\partial \pi}(4\pi - 4\pi^2) = 4 - 8\pi$$

$$\Rightarrow 4 - 8\pi = 0 \qquad \pi = \frac{1}{2}$$

<u>caveat:</u> if there exist data points x such that $p^*(x) > 0$, but $p(x) = 0$

$$\Rightarrow \log \frac{p^*(x)}{p(x)} = \log \infty = \infty \qquad \Rightarrow KL = \infty$$

$\Rightarrow$ cannot use $\frac{\partial KL}{\partial \theta}$ as training gradient

$\Rightarrow$ use model families such that $\mathrm{dom}(p^*(x)) \subseteq \mathrm{dom}(p(x))$

- relationship between forward KL and maximum likelihood training

$$KL[p^* \| p] = \underbrace{\int p^*(x) \log p^*(x)\, dx}_{-H[p^*] \text{ (entropy)}} \underbrace{- \int p^*(x) \log p(x)\, dx}_{\mathbb{E}_{x \sim p^*(x)}[-\log p(x)]}$$

$$KL[p^* \| p] = \mathbb{E}_{x \sim p^*(x)}\left[-\log p(x)\right] + \text{const}$$

const: independent of $p(x)$ $\Rightarrow$ drop

$\Rightarrow$ optimization problem

$$\hat{p}(x) = \arg\min_{p(x) \in \mathcal{F}} \mathbb{E}_{x \sim p^*(x)}\left[-\log p(x)\right]$$

minimize KL $\Leftrightarrow$ minimize NLL $\Leftrightarrow$ maximize $p(x)$ for TS

given $TS = \{x_i \sim p^*(x)\}_{i=1}^{N}$

$$\hat{p}(x) \approx \arg\min_{p(x)} \frac{1}{N} \sum_{i=1}^{N} -\log p(x_i)$$

- does not contain $p^*(x)$ $\Rightarrow$ we can optimize <u>without</u> knowing $p^*(x)$
- sufficient to have sample from $p^*(x)$ $\hat{=}$ TS
- <u>but</u>: we need to calculate $p(x_i)$ during training
  $\Rightarrow$ $p(x)$ must be capable to do "inference"
  (a model that only generates data $x \sim p(x)$ is not enough)

reverse KL divergence

$$KL[p \,\|\, p^*] = \int p(x) \log \frac{p(x)}{p^*(x)} \, dx = \mathbb{E}_{x \sim p(x)} \left[\log \frac{p(x)}{p^*(x)} \right]$$

empirical approximation: (train $p^{(T)}(x)$ to minimize $KL(p^{(T)} \,\|\, p^*)$)

iterate $t = 1 \dots T$

- current guess $p^{(t-1)}(x)$: draw batch $\{x_i \sim p^{(t-1)}(x)\}_{i=1}^{N}$
- $p^{(t)}(x) = \underset{p(x)}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i)}{p^*(x_i)}$

$\Rightarrow$ need to know $p^*(x)$ to calculate $\frac{p(x_i)}{p^*(x_i)}$

$\Rightarrow$ reverse KL cannot be used in many applications

important exception: Gibbs distribution in physics & chemistry

$$p_T^*(x) = \frac{1}{Z(T)} \exp\left(-\frac{E(x)}{k T}\right)$$

$$Z(T) = \int \exp\left(-\frac{E(x)}{kT}\right) dx$$

$E(x)$: energy of state x

T: temperature, k: Boltzmann's constant

$Z(T)$ partition function $\hat{=}$ normalization (usually intractable and thus unknown)

$p(x) \simeq p_T^*(x)$ is then a "surrogate" model, often faster, interpretable,

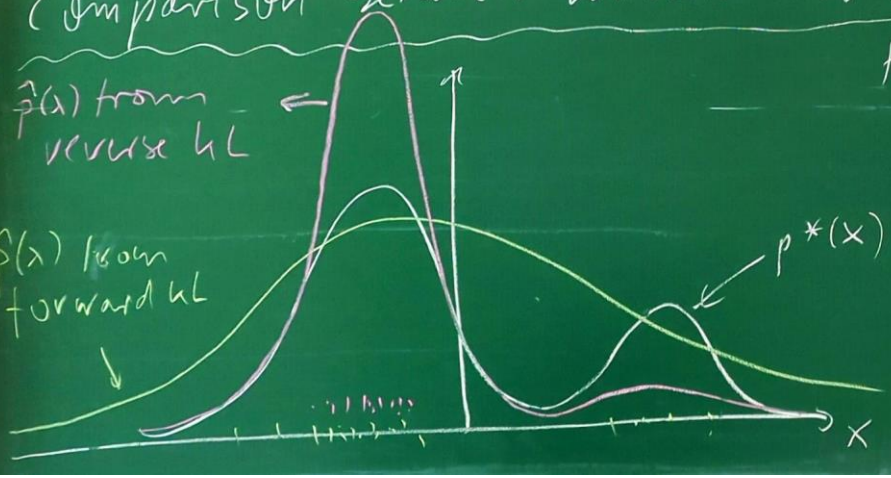
insert Gibbs into optimization objective:

$$p^{(t)}(x) = \arg\min_{p(x)} \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i)}{\frac{1}{Z(T)} \exp\left(-\frac{E(x_i)}{k_B T}\right)} = \arg\min_{p(x)} \frac{1}{N} \sum_{i=1}^{N} \log p(x_i) + \frac{E(x_i)}{k_B T}$$

$$\log p(x_i) + \frac{E(x_i)}{k_B T} + \log Z(T)$$

independent of our model $\Rightarrow$ drop

comparison between forward and reverse KL:



forward KL – tends to smooth out modes of $p^*(x)$
$\hat{=}$ puts probability mass where should be none
"mode covering" behavior

reverse KL – tends to focus on a single (or few highest modes) and forgets the others
$\hat{=}$ has no mass where we should be
"mode seeking", "mode collapse" behavior

Maximum Mean Discrepancy (MMD) [Gretton et al. 2012]

idea: use kernel trick

- without kernel
  - define $\varphi : \mathrm{dom}(x) \to \mathbb{R}$ $\quad \varphi(x) = \text{real}$
  - two data sets $\{x_i^* \sim p^*(x)\}_{i=1}^{N}$, $\{x_i \sim p(x)\}_{i=1}^{M}$
  - calculate $\tilde{x}_i = \varphi(x_i)$, $\tilde{x}_i^* = \varphi(x_i^*)$
  - $$MMD = \max_{\varphi \in F} \left[\frac{1}{M} \sum_{i=1}^{M} \varphi(x_i) - \frac{1}{N} \sum_{i=1}^{N} \varphi(x_i^*) \right]$$

φ: "witness function"

↓

certifies that p and p^* are different, when MMD is big

$MMD \geq 0$ (if $-\varphi \in F$ when $\varphi \in F$)

$MMD = 0 \iff p(x) = p^*(x)$

MMD is a metric

replace explicit mapping $\varphi(x)$ with a kernel function $k(x, x')$

$$MMD^2 = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{i' \neq i} k(x_i, x_{i'}) + \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{i' \neq i} k(x_i^*, x_{i'}^*) - \frac{2}{M \cdot N} \sum_{i=1}^{M} \sum_{i'=1}^{N} k(x_i, x_{i'}^*)$$

repulsive between synthetic x_i | independent of $p(x)$ $\Rightarrow$ drop | attractive force between synthetic & real x

typical kernels:

- squared exponential $k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2h}\right)$ ← bandwidth hyper param.
- inverse multi-quadratic $k(x, x') = \frac{1}{\frac{\|x - x'\|^2}{h} + 1}$
- multi-scale squared expon. $k(x, x') = \sum_{\ell=1}^{L} \exp\left(-\frac{\|x - x'\|^2}{2h_\ell}\right)$
- generally: $k(x, x') \approx 1$ if $x \approx x'$
  ≈ 0 if x far away from x'