Lecture-26: Mutual Information Method

1 Rate-distortion theory

Definition 1.1 (Rate distortion). Consider parameter space Θ, prediction space Θ', and loss function $\ell : \Theta \times \Theta' \to \mathbb{R}$. We define the rate distortion function $\phi_\theta : \mathbb{R} \to \mathbb{R}$ for each $D \in \mathbb{R}$ as

$$\phi_\theta(D) \triangleq \inf_{P_{\hat{\theta}|\theta} : \mathbb{E}\ell(\theta,\hat{\theta}) \leqslant D} I(\theta;\hat{\theta}). \tag{1}$$

Theorem 1.2 (General converse). *Suppose $X \to W \to \hat{X}$, where $W \in [M]$ and $\mathbb{E}\ell(X,\hat{X}) \leqslant D$. Then $\ln M \geqslant \phi_X(D)$.*

Proof. For a feasible solution $P_{\hat{X}|X}$, we get $\ln M \geqslant H(W) \geqslant I(X;W) \geqslant I(X;\hat{X}) \geqslant \phi_X(D)$. □

Definition 1.3. We define maximum distortion as $D_{\max} \triangleq \inf_{\hat{\theta}} \mathbb{E}_{\theta\sim\pi}\ell(\theta,\hat{\theta})$ over all pre-determined estimators $\hat{\theta}$ without any observation X.

Remark 1. By definition, $D_{\max}$ is the distortion attainable without any information. Indeed, if $D_{\max} = \mathbb{E}\ell(\theta,\hat{\theta})$ for some fixed $\hat{\theta}$, then this $\hat{\theta}$ is the "default" reconstruction of θ, i.e., the best estimate when we have no information about θ. Therefore $D \geqslant D_{\max}$ can be achieved for free. This is the reason for the notation $D_{\max}$ despite that it is defined as an infimum.

Theorem 1.4 (Properties). *The following properties are true for rate distortion function $\phi_\theta : \mathbb{R} \to \mathbb{R}$.*
(a) The map ϕ_θ is convex and non-increasing.
(b) $\phi_\theta(D) = 0$ for all $D > D_{\max}$.

Proof. Let ϕ_θ be rate distortion function as defined in (1) for Markov chain $\theta \to X \to \hat{\theta}$. For a prior $\pi \in \mathcal{M}(\Theta)$, we have $I(\theta;\hat{\theta}) = D(P_{\hat{\theta}|\theta}\|P_{\hat{\theta}} \mid \pi) = \mathbb{E}_{\theta\sim\pi}\mathbb{E}_{P_{\hat{\theta}|\theta}} \ln \frac{dP_{\hat{\theta}|\theta}}{dP_{\hat{\theta}}}$.

(a) Since infimum is a non-increasing function of the set size, we obtain that ϕ_θ is non-increasing in D. Next, we define $\mathcal{A}(D) \triangleq \{\hat{\theta} : \mathbb{E}\ell(\theta;\hat{\theta}) \leqslant D\}$. Let $D_1, D_2 < D_{\max}$, estimators $(\hat{\theta}_1,\hat{\theta}_2) \in \mathcal{A}(D_1) \times \mathcal{A}(D_2)$, and an independent uniform random variable $W : \Omega \to [0,1]$ with mean $\lambda \in [0,1]$. We define another estimator $\hat{\theta} \triangleq \bar{W}\hat{\theta}_1 + W\hat{\theta}_2$, and observe that $\hat{\theta} \in \mathcal{A}(\bar{\lambda}D_1 + \lambda D_2)$, i.e.

$$\mathbb{E}\ell(\theta;\hat{\theta}) = \bar{\lambda}\mathbb{E}\ell(\theta;\hat{\theta}_1) + \lambda\mathbb{E}\ell(\theta,\hat{\theta}_2) \leqslant \bar{\lambda}D_1 + \lambda D_2.$$

Further, it follows from the convexity of mutual information in channel $P_{\hat{\theta}|\theta}$ for a fixed prior $\pi \in \mathcal{M}(\Theta)$, that $I(\theta;\hat{\theta}) \leqslant \bar{\lambda}I(\theta;\hat{\theta}_1) + \lambda I(\theta;\hat{\theta}_2)$. Summarizing both results, we observe that

$$\bar{\lambda}\phi_\theta(D_1) + \lambda\phi_\theta(D_2) = \inf_{(\hat{\theta}_1,\hat{\theta}_2)\in\mathcal{A}(D_1)\times\mathcal{A}(D_2)} \bar{\lambda}I(\theta;\hat{\theta}_1) + \lambda I(\theta;\hat{\theta}_2) \geqslant \inf_{\hat{\theta}\in\mathcal{A}(\bar{\lambda}D_1+\lambda D_2)} I(\theta;\hat{\theta}) = \phi_\theta(\bar{\lambda}D_1 + \lambda D_2).$$

(b) For any $D > D_{\max}$ we can set $\hat{\theta}$ without any information about θ. Thus $I(\theta;\hat{\theta}) = 0$.

□

Theorem 1.5 (Single-letterization). *For stationary memoryless source $S : \Omega \to \mathcal{S}^m$ with common distribution $P_{S_1} \in \mathcal{M}(\mathcal{S})$ and separable loss ℓ such that $\ell(S,\hat{S}) = \frac{1}{m}\sum_{i=1}^m \ell_1(S_i,\hat{S}_i)$, then $\phi_S(D) = m\phi_{S_1}(D)$ for every m. Thus,*

$$R^{(I)}(D) \triangleq \limsup_{m\to\infty} \frac{1}{m}\phi_S(D) = \phi_{S_1}(D).$$

Proof. We will show this in two steps. Let $\mathcal{A}_1(D) \triangleq \{\hat{S}_i : \mathbb{E}\ell_1(S_i,\hat{S}_i) \leqslant D\}$ and $\mathcal{A}(D) \triangleq \{\hat{S} : \mathbb{E}\ell(S,\hat{S}) \leqslant D\}$.

(a) Consider an estimate $\hat{S}$ such that $P_{\hat{S}|S} \triangleq P_{\hat{S}_1|S_1}^{\otimes m}$ where $\hat{S}_i \in \mathcal{A}_1(D)$ for all $i \in [m]$. Then $\hat{S}$ is a feasible estimate with $S \in \mathcal{A}(D)$. Since S is memoryless and stationary and $P_{\hat{S}|S}$ has the product form, the estimate $\hat{S}$ is memoryless and stationary. It follows that $I(S;\hat{S}) = \sum_{i=1}^m I(S_i;\hat{S}_i)$. Recall that the rate distortion for m-sized source S is defined as

$$\phi_S(D) \triangleq \inf_{\hat{S}\in\mathcal{A}(D)} I(S;\hat{S}) \leqslant \inf_{P_{\hat{S}|S}=P_{\hat{S}_1|S_1}^{\otimes m}:\hat{S}_i\in\mathcal{A}_1(D),i\in[m]} \sum_{i=1}^m I(S_i;\hat{S}_i) \leqslant \sum_{i=1}^m \inf_{\hat{S}_i\in\mathcal{A}_1(D)} I(S_i;\hat{S}_i) = m\phi_{S_1}(D).$$

Diving by m on both sides and taking limit $m \to \infty$, we obtain $R^{(I)}(D) \leqslant \phi_{S_1}(D)$.

(b) For the converse, we focus on any estimator $\hat{S} \in \mathcal{A}(D)$, i.e. Markov kernel $P_{\hat{S}|S}$ satisfies the constraint $\mathbb{E}\ell(S,\hat{S}) \leqslant D$. From the super-additivity property of mutual information for memoryless source, we obtain $I(S;\hat{S}) \geqslant \sum_{i=1}^m I(S_i;\hat{S}_i)$. From the definition of rate distortion function, we obtain $\phi_{S_1}(\mathbb{E}\ell_1(S_i;\hat{S}_i)) \leqslant I(S_i;\hat{S}_i)$ for each $i \in [m]$. From convexity and non-increasing property of rate distortion function in Theorem 1.4, we obtain

$$I(S;\hat{S}) \geqslant \sum_{i=1}^m I(S_i;\hat{S}_i) \geqslant \sum_{i=1}^m \phi_{S_1}(\mathbb{E}\ell_1(S_i;\hat{S}_i)) \geqslant m\phi_{S_1}\Big(\frac{1}{m}\sum_{i=1}^m \mathbb{E}\ell_1(S_i;\hat{S}_i)\Big) \geqslant m\phi_{S_1}(D).$$

The result follows from taking infimum over all such Markov kernels $P_{\hat{S}|S}$ and the definition of rate distortion function.

□

Theorem 1.6 (Rate distortion for Gaussian sources). *Let $S \sim \mathcal{N}(0,\sigma^2 I_d)$ and $\ell(s,\hat{s}) \triangleq \|s-\hat{s}\|_2^2$ for $s,\hat{s} \in \mathbb{R}^d$, then rate distortion function $R(D) \triangleq \inf_{P_{\hat{S}|S}:\mathbb{E}\ell(S,\hat{S})\leqslant D} I(S;\hat{S}) = \frac{d}{2}\ln^+\frac{d\sigma^2}{D}$.*

Proof. We first show the result for $d = 1$. Since $D_{\max} = \sigma^2$, we can assume $D < \sigma^2$ for otherwise there is nothing to show.

(a) **Achievability.** Choose $S = \hat{S} + Z$, where $\hat{S} \sim \mathcal{N}(0,\sigma^2 - D)$ and independent of $Z \sim \mathcal{N}(0,D)$. In other words, the backward channel $P_{S|\hat{S}}$ is AWGN with noise power D. Since S is Gaussian with mean 0 and variance σ^2, we can write the conditional density for forward channel as

$$\begin{aligned} f_{\hat{S}|S}(\hat{s}\mid s) = \frac{f_{S,\hat{S}}(s,\hat{s})}{f_S(s)} &= \frac{1}{\sqrt{2\pi\frac{(\sigma^2-D)}{\sigma^2}D}}\exp\Big(\frac{s^2}{2\sigma^2} - \frac{\hat{s}^2}{2(\sigma^2-D)} - \frac{(s-\hat{s})^2}{2D}\Big) \\ &= \frac{1}{\sqrt{2\pi\frac{(\sigma^2-D)}{\sigma^2}D}}\exp\Big(-\frac{s^2(\sigma^2-D)}{2\sigma^2 D} - \frac{\hat{s}^2\sigma^2}{2(\sigma^2-D)D} + \frac{s\hat{s}}{D}\Big) \\ &= \frac{1}{\sqrt{2\pi\frac{(\sigma^2-D)}{\sigma^2}D}}\exp\left(-\frac{1}{2\frac{\sigma^2-D}{\sigma^2}D}\Big(\hat{s}^2 + \frac{s^2(\sigma^2-D)^2}{\sigma^4} - 2\frac{s\hat{s}(\sigma^2-D)}{\sigma^2}\Big)\right). \end{aligned}$$

It follows that the forward channel is $P_{\hat{S}|S} = \mathcal{N}(\frac{\sigma^2-D}{\sigma^2}S, \frac{\sigma^2-D}{\sigma^2}D)$, and hence $R(D) \leqslant I(S;\hat{S}) = \frac{1}{2}\ln\frac{\sigma^2}{D}$.

(b) **Converse.** Let $S \sim \mathcal{N}(0,\sigma^2)$ and $P_{\hat{S}|S}$ be Markov kernel associated with an estimator $\hat{S}$. We denote the joint distribution of source S and estimate $\hat{S}$ by $P_{S,\hat{S}}$ or simply by P, such that $\mathbb{E}_{(S,\hat{S})\sim P}\ell(S,\hat{S}) \leqslant D$. Denote the forward channel in the above achievability by $P^*_{\hat{S}|S}$. Then, we have

$$I(S;\hat{S}) = \mathbb{E}_P \ln\frac{dP_{S|\hat{S}}}{dP^*_{S|\hat{S}}} + \mathbb{E}_P\ln\frac{dP^*_{S|\hat{S}}}{dP_S} = D(P_{S|\hat{S}}\|P^*_{S|\hat{S}}\mid P_{\hat{S}}) + \mathbb{E}_P\ln\frac{dP^*_{S|\hat{S}}}{dP_S}.$$

From the non-negativity of KL divergence and definition of $P^*_{\hat{S}|S}$ such that $\mathbb{E}_P\ell(S,\hat{S}) \leqslant D$, we write

$$I(S;\hat{S}) \geqslant \mathbb{E}_P\ln\frac{dP^*_{S|\hat{S}}}{dP_S} = \frac{1}{2}\ln\frac{\sigma^2}{D} + \frac{1}{2}\mathbb{E}_P\Big[\frac{S^2}{\sigma^2} - \frac{(S-\hat{S})^2}{D}\Big] \geqslant \frac{1}{2}\ln\frac{\sigma^2}{D} \geqslant 0.$$

Finally, for the vector case follows from the scalar case and the same single-letterization argument in Theorem 1.5 using the convexity of the rate-distortion function. □

2 Mutual information method

Definition 2.1. The quantity $I(\theta;X)$ is the amount of information provided by the data X about the latent parameter θ. We define the capacity of the channel $P_{X|\theta}$ by maximizing over all priors, i.e.

$$I(\theta;X) \leqslant \sup_{\pi\in P(\Theta)} I(\theta;X) \triangleq C. \tag{2}$$

Theorem 2.2 (Mutual information method (MIM)). *Consider a simple statistical decision theory setting with parameter space Θ, prediction space $\hat{\Theta}$, estimator $\hat{\theta}$ represented by a Markov kernel $P_{\hat{\theta}|X} : \mathcal{X} \to \mathcal{M}(\hat{\Theta})$, and loss function $\ell : \Theta \times \hat{\Theta} \to \mathbb{R}$. If $\pi \in \mathcal{M}(\Theta)$ is a prior on the parameter space, then minimax and Bayes risk are lower bounded as*

$$R^* \geqslant R^*_\pi = \inf_{P_{\hat{\theta}|\theta}} \mathbb{E}\ell(\theta,\hat{\theta}) \geqslant \phi^{-1}(I(\theta;X)) \geqslant \phi^{-1}(C). \tag{3}$$

Proof. Fix some prior $\pi \in \mathcal{M}(\Theta)$ and we will lower bound the Bayes risk R^*_π of estimating $\theta \sim \pi$ on the basis of observation X with respect to loss function ℓ. Consider the Markov chain $\theta \to X \to \hat{\theta}$, where $\hat{\theta}(X)$ is a random estimator with Markov kernel $P_{\hat{\theta}|X}$ such that $\mathbb{E}[\ell(\theta,\hat{\theta})] \leqslant D$. Applying the data processing inequality for mutual information, we have

$$\phi_\theta(D) \triangleq \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}\ell(\theta,\hat{\theta})\leqslant D} I(\theta;\hat{\theta}) \leqslant I(\theta;\hat{\theta}) \leqslant I(\theta;X) \leqslant \sup_{\pi\in\mathcal{M}(\Theta)} I(\theta;X) = C. \tag{4}$$

Taking $D = \mathbb{E}\ell(\theta;\hat{\theta})$ for any estimator $\hat{\theta}$, we obtain $\phi_\theta(\mathbb{E}\ell(\theta,\hat{\theta})) \leqslant I(\theta;\hat{\theta}) \leqslant I(\theta;X) \leqslant C$. Since the rate-distortion function ϕ_θ is non-increasing, we obtain that

$$\mathbb{E}\ell(\theta,\hat{\theta}) \geqslant \phi^{-1}(I(\theta;\hat{\theta})) \geqslant \phi^{-1}(I(\theta;X)) \geqslant \phi^{-1}(C).$$

Minimizing the loss $\mathbb{E}\ell(\theta,\hat{\theta})$ over all estimation kernels $P_{\hat{\theta}|\theta}$, we obtain the lower bound on the Bayes and hence the minimax risk. □

Remark 2. We observe the following for the above inequality.

(a) The quantity $\inf_{P_{\hat{\theta}|\theta}:\mathbb{E}\ell(\theta,\hat{\theta})\leqslant D} I(\theta;\hat{\theta})$ is the minimum amount of information required to achieve a given estimation accuracy, which is precisely the rate-distortion $\phi(D) \equiv \phi_\theta(D)$.

(b) The reasoning of the mutual information method is reminiscent of the converse proof for joint-source channel coding. As such, the argument here retains the flavor of "source-channel separation", in that the lower bound in (4) depends only on the prior (source) and the loss function, while the capacity upper bound (2) depends only on the statistical model (channel).

We will discuss three popular approaches for, namely, *Le Cam's method*, *Assouad's lemma*, and *Fano's method*. All three follow from the mutual information method, corresponding to different choice of prior $\pi \in \mathcal{M}(\theta)$, namely, the uniform distribution over a two-point set $\{\theta_0,\theta_1\}$, the hypercube $\{0,1\}^d$, and a packing. While these methods are highly useful in determining the minimax rate for many problems, they are often loose with constant factors compared to the MIM.

2.1 GLM revisited and the Shannon lower bound

Example 2.3 (GLM). Consider the d-dimensional GLM for statistical decision theory simple setting $\Theta = \hat{\Theta} = \mathcal{X} \triangleq \mathbb{R}^d$, where we observe an *i.i.d.* sample $X : \Omega \to \mathcal{X}^m$ with common distribution $\mathcal{N}(\theta, I_d)$ and parameter $\theta \in \Theta$. Denote by $R^*(\Theta)$ the minimax risk with respect to the quadratic loss $\ell : (\theta,\hat{\theta}) \mapsto \left\|\hat{\theta} - \theta\right\|_2^2$. First, let us consider the unconstrained model where $\Theta \triangleq \mathbb{R}^d$.

(a) **Upper bound.** Estimating using the sample mean, i,e. $\hat{\theta} \triangleq \bar{X} \triangleq \frac{1}{m}\sum_{i=1}^m X_i \sim \mathcal{N}(\theta, \frac{1}{m} I_d)$, we achieve the upper bound $R^*(\mathbb{R}^d) \leqslant \frac{d}{m}$. This turns out to be the exact minimax risk, as seen by computing the Bayes risk for Gaussian priors. Next we apply the mutual information method to obtain the same matching lower bound without evaluating the Bayes risk.

(b) **Lower bound.** Again, let us consider $\theta \sim \mathcal{N}(0, sI_d)$ for some $s > 0$. We know from the Gaussian rate-distortion function that

$$\phi(D) \triangleq \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}\left\|\theta-\hat{\theta}\right\|_2^2\leqslant D} I(\theta;\hat{\theta}) = \frac{d}{2}\ln\frac{sd}{D}\mathbb{1}_{\{D<sd\}}.$$

It follows that $\phi^{-1}(x) = sde^{-\frac{2x}{d}}$ for all $x \in \mathbb{R}_+$. Using the sufficiency of sample mean $\bar{X}$ and the formula of Gaussian channel capacity, the mutual information between the parameter and the data can be computed as

$$I(\theta; X) = I(\theta; \bar{X}) = \frac{d}{2}\ln(1+sm).$$

It then follows from (3) that $R^*_\pi \geqslant \phi^{-1}(I(\theta;X)) = \frac{sd}{1+sm}$, which in fact matches the exact Bayes risk. Sending $s \to \infty$ we recover the result $R^*(\mathbb{R}^d) = \frac{d}{m}$.

Remark 3. In the above unconstrained GLM, we are able to compute everything in closed form when applying the mutual information method. Such exact expressions are rarely available in more complicated models in which case various bounds on the mutual information will prove useful.

Definition 2.4. Let $B(\rho) \triangleq \left\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leqslant \rho\right\}$ is the ℓ_2-ball of radius ρ centered at zero.

Theorem 2.5 (Bounded GLM). $R^*(B(\rho)) \asymp \frac{d}{m} \wedge \rho^2$.

Proof. We will show that upper and lower bound for the minimax risk when parameter space $\Theta \triangleq B(\rho)$.

(a) The upper bound $R^*(B(\rho)) \leqslant \frac{d}{m} \wedge \rho^2$ follows from considering the estimator $\hat{\theta} = \bar{X}$ and $\hat{\theta} = 0$.

(b) To prove the lower bound, we apply the mutual information method with a uniform prior $\theta \sim U$ where $U : \Omega \to B(r)$ is a uniform random variable and $r \in [0, \rho]$ is to be optimized. The mutual information can be upper bounded using the AWGN capacity as

$$I(\theta; X) = I(\theta; \bar{X}) \leqslant \sup_{\pi \in \mathcal{M}(\Theta): \mathbb{E}\|\theta\|_2^2 \leqslant r} I\Big(\theta; \theta + \frac{1}{\sqrt{m}}Z\Big) = \frac{d}{2}\ln\Big(1 + \frac{mr^2}{d}\Big) \leqslant \frac{mr^2}{2},$$

where $Z \sim \mathcal{N}(0, I_d)$. Alternatively, we can use Corollary A.1 to bound the capacity (as information radius) by the KL diameter, which yields the same bound within constant factors,

$$I(\theta; X) \leqslant \sup_{\pi \in \mathcal{M}(\Theta): \mathbb{E}\|\theta\|_2^2 \leqslant r} I(\theta; \theta + \frac{1}{\sqrt{m}}Z) \leqslant \max_{\theta, \theta' \in B(r)} D(\mathcal{N}(\theta, \frac{1}{m}I_d) \| \mathcal{N}(\theta', \frac{1}{m}I_d)) = 2mr^2.$$

For the lower bound, due to the lack of closed-form formula for the rate-distortion function for uniform distribution over Euclidean balls, we apply the Shannon lower bound (SLB) from Section 26.1. Since θ has an isotropic distribution, applying Theorem 26.3 yields

$$\inf_{P_{\tilde{\theta}|\theta}: \mathbb{E}\left\|\theta - \tilde{\theta}\right\|^2 \leqslant D} I(\theta; \tilde{\theta}) \geqslant h(\theta) + \frac{d}{2}\ln\frac{2\pi e d}{D} \geqslant \frac{d}{2}\ln\frac{cr^2}{D},$$

for some universal constant c, where the last inequality is because for $\theta \sim U$ uniformly distributed over $B(r)$, $h(\theta) = \ln \mathrm{vol}(B(r)) = d\ln r + \ln \mathrm{vol}(B(1))$ and the volume of a unit Euclidean ball in d dimensions satisfies (recall (27.14)) $\mathrm{vol}(B(1))^{\frac{1}{d}} \asymp \frac{1}{\sqrt{d}}$. Finally, applying (3) yields $\frac{1}{2}\ln\frac{cr^2}{R^*} \leqslant \frac{mr^2}{2}$, i.e., $R^* \geqslant cr^2 e^{-\frac{mr^2}{d}}$. Optimizing over r and using the fact that $\sup_{x \in (0,1)} xe^{-ax} = \frac{1}{ea}\mathbb{1}_{\{a \geqslant 1\}} + e^{-a}\mathbb{1}_{\{a<1\}}$, we have

$$R^* \geqslant \sup_{r \in [0, \rho]} cr^2 e^{-\frac{mr^2}{d}} \asymp \frac{d}{m} \wedge \rho^2.$$

□

Remark 4. Comparing the bounded GLM with unconstrained GLM case, we see that if $\rho^2 > \frac{d}{m}$, it is rate-optimal to ignore the bounded-norm constraint. If $\rho^2 < \frac{d}{m}$, we can discard all observations and estimate by zero, because data do not provide a better resolution than the prior information.

A Capacity as information radius

Definition A.1. For state spaces $\mathcal{X}$ and $\mathcal{Y}$ and conditional distribution $P_{Y|X=x} \in \mathcal{M}(\mathcal{Y})$, and a KL divergence ball with center $P_{Y|X=x}$, we define the radius r and the diameter d as

$$r \triangleq \inf_Q \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| Q), \qquad d \triangleq \sup_{x, x' \in \mathcal{X}} D(P_{Y|X=x} \| P_{Y|X=x'})$$

Corollary A.2 (Radius less than diameter). *Let* $\left\{P_{Y|X=x} \in \mathcal{M}(\mathcal{Y}) : x \in \mathcal{X}\right\}$ *be a set of distributions. Then*

$$C = \sup_{P_X} I(X;Y) \leqslant \inf_{Q} \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| Q) \leqslant \sup_{x,x' \in \mathcal{X}} D(P_{Y|X=x} \| P_{Y|X=x'}).$$

Proof. By the golden formula Corollary 4.2, we have

$$I(X;Y) = \inf_{Q} D(P_{Y|X} \| Q \mid P_X) \leqslant \inf_{Q} \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| Q) \leqslant \inf_{x' \in \mathcal{X}} \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| P_{Y|X=x'}).$$

□