Lecture-21: Mutual information and rate-distortion

1 Mutual information

Lemma 1.1. *Let $P, Q \in \mathcal{M}(\mathcal{Y})$ be two measures on space $\mathcal{Y}$, then the map $(P, Q) \mapsto D(P\|Q)$ is convex.*

Proof. Let $\mathcal{X} \triangleq \{0,1\}$ and let $P_X = Q_X \in \mathcal{M}(\mathcal{X})$ be a Bernoulli distribution with mean $\lambda \in [0,1]$. Let $P_0, P_1, Q_0, Q_1 \in \mathcal{M}(\mathcal{Y})$ and define Markov kernels

$$P_{Y|X=0} \triangleq P_0, \qquad P_{Y|X=1} \triangleq P_1, \qquad Q_{Y|X=0} \triangleq Q_0, \qquad Q_{Y|X=1} \triangleq Q_1.$$

We can write the divergence of two joint distributions $P_{X,Y}$ and $Q_{X,Y}$ in terms of conditional divergence, and as

$$D(P_{X,Y}\|Q_{X,Y}) = D(P_{Y|X}\|Q_{Y|X} \mid P_X) = \bar{\lambda} D(P_0\|Q_0) + \lambda D(P_1\|Q_1).$$

We get the result from the data processing inequality $D(P_{X,Y}\|Q_{X,Y}) \geqslant D(P_Y\|Q_Y)$ for KL divergence and recalling that $P_Y = \mathbb{E}_{X\sim P_X} P_{Y|X}$. □

Remark 1. The proof shows that for an arbitrary measure of similarity $D(P\|Q)$, the convexity of $(P, Q) \mapsto D(P\|Q)$ is equivalent to *conditioning increases divergence* property of D. Convexity can also be understood as *mixing decreases divergence*.

Definition 1.2. For a random vector $(X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ with joint distribution $P_{X,Y} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$, the mutual information is defined as

$$I(X;Y) \triangleq D(P_{X,Y}\|P_X \otimes P_Y).$$

Lemma 1.3. *For a random vector $(X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ with joint distribution $P_{X,Y} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$, the mutual information $I(X;Y) = D(P_{Y|X}\|P_Y \mid P_X)$.*

Proof. From the definition of mutual information and tower property of conditional expectation, we write $I(X;Y) = \mathbb{E}_{P_X P_{Y|X}} \ln \frac{dP_{Y|X}}{dP_Y} = \mathbb{E}_{P_X} D(P_{Y|X}\|P_Y) = D(P_{Y|X}\|P_Y \mid P_X)$. □

Theorem 1.4 (Joint vs marginal mutual information). *Consider a random vector $(X, Y) : \Omega \to (\mathcal{X} \times \mathcal{Y})^m$.*

(a) If the channel is memoryless, i.e., $P_{Y|X} = \prod_{i=1}^m P_{Y_i|X_i}$, then $I(X;Y) \leqslant \sum_{i=1}^m I(X_i;Y_i)$, with equality iff $P_Y = \prod_{i=1}^m P_{Y_i}$. Consequently, the (unconstrained) capacity is additive for memoryless channels, i.e.

$$\max_{P_X} I(X;Y) = \sum_{i=1}^m \max_{P_{X_i}} I(X_i;Y_i).$$

(b) If the source is memoryless, i.e., $P_X = \prod_{i=1}^m P_{X_i}$, then $I(X;Y) \geqslant \sum_{i=1}^m I(X_i;Y)$ with equality iff $P_{X|Y} = P_Y \prod_{i=1}^m P_{X_i|Y}$-almost surely. Consequently,

$$\min_{P_{Y|X}} I(X;Y) = \sum_{i=1}^m \min_{P_{Y|X_i}} I(X_i;Y).$$

Proof. We utilize the definition of mutual information.

(a) From the definition of mutual information, we write

$$I(X;Y) - \sum_{i=1}^m I(X_i,Y_i) = \mathbb{E}_{P_X}\mathbb{E}_{P_{Y|X}} \ln \frac{dP_{Y|X}}{dP_Y} - \sum_{i=1}^m \mathbb{E}_{P_{X_i}}\mathbb{E}_{P_{Y_i|X_i}} \ln \frac{dP_{Y_i|X_i}}{dP_{Y_i}} = \mathbb{E}_{P_X}\mathbb{E}_{P_{Y|X}}\Big[\ln \frac{dP_{Y|X}}{dP_Y} - \ln \frac{\prod_{i=1}^m dP_{Y_i|X_i}}{\prod_{i=1}^m dP_{Y_i}}\Big].$$

We can rearrange the terms and observe that $\ln \frac{P_Y}{\prod_{i=1}^m P_{Y_i}}$ only depends on P_Y, to get

$$I(X;Y) - \sum_{i=1}^m I(X_i,Y_i) = D(P_{Y|X}\|\prod_{i=1}^m P_{Y_i|X_i} \mid P_X) - D(P_Y\|\prod_{i=1}^m P_{Y_i}).$$

When channel is memoryless, $D(P_{Y|X}\|\prod_{i=1}^m P_{Y_i|X_i} \mid P_X) = 0$, and we get the result.

(b) Similarly, switching the role of X and Y, we can write

$$I(X;Y) - \sum_{i=1}^{m} I(X_i,Y) = \mathbb{E}_{P_Y}\mathbb{E}_{P_{X|Y}}\Big[\ln\frac{dP_{X|Y}}{dP_X} - \ln\frac{\prod_{i=1}^m dP_{X_i|Y}}{\prod_{i=1}^m dP_{X_i}}\Big] = D(P_{X|Y}\|\prod_{i=1}^m P_{X_i|Y} \mid P_Y) - D(P_X\|\prod_{i=1}^m P_{X_i}).$$

When source is memoryless, $D(P_X\|\prod_{i=1}^m P_{X_i}) = 0$, and we get the result.

□

Remark 2. We observe the following.
(a) For a product channel, the input maximizing the mutual information is a product distribution.
(b) For a product source, the channel minimizing the mutual information is a product channel.

Definition 1.5 (Conditional mutual information). If $X,Y,Z : \Omega \to \mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$, then we define

$$I(X;Y \mid Z) \triangleq D(P_{X,Y|Z}\|P_{X|Z}P_{Y|Z} \mid P_Z) = \mathbb{E}_{z\sim P_Z} I(X;Y \mid Z=z),$$

where the product $P_{X|Z}P_{Y|Z}$ is a conditional distribution such that $(P_{X|Z}P_{Y|Z})(A\times B \mid z) = P_{X|Z}(A \mid z)P_{Y|Z}(B \mid z)$, under which X and Y are independent conditioned on Z.

Lemma 1.6 (Chain rule). *For random variables X,Y,Z, we have $I(Y,Z;X) = I(X;Y) + I(X;Z \mid Y)$.*

Proof. By the definition of conditional mutual information and mutual information, we get

$$I(X;Z \mid Y) = \mathbb{E}_{P_{X,Y,Z}} \ln\frac{dP_{Y,Z|X}}{dP_{Y|X}dP_{Z|X}} = \mathbb{E}_{P_{X,Y,Z}} \ln\frac{dP_{X,Y,Z}}{dP_{Y|X}dP_{Z,X}} = \mathbb{E}_{P_{X,Y,Z}} \ln\frac{dP_{X,Z|Y}dP_Y}{dP_{X,Z}dP_{Y|X}} = I(X;Z \mid Y) - I(X;Y).$$

□

Theorem 1.7 (Data processing inequality). *If $X \to Y \to Z$ is a Markov chain, then $I(X;Z) \leqslant I(X;Y)$ with equality iff $X \to Z \to Y$.*

Proof. Since $X \to Y \to Z$ is a Markov chain. Hence, X and Z are conditionally independent given Y, and $I(X;Z \mid Y) = 0$. Applying Kolmogorov identity to $I(Y,Z;X)$, we get

$$I(Y,Z;X) = I(X;Y) + I(X;Z \mid Y) = I(X;Z) + I(X;Y \mid Z).$$

The result follows from the observation that $I(X;Z \mid Y) = 0$ and $I(X;Y \mid Z) \geqslant 0$. □

Lemma 1.8. *For a random vector $(X,Y) : \Omega \to \mathcal{X}\times\mathcal{Y}$ with joint distribution $P_{X,Y} \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})$, the mutual information $I(X;Y)$ is convex in $P_{Y|X}$.*

Proof. Consider three random variables X, Y_0, Y_1 and two Markov kernels $P_{Y_0|X}, P_{Y_1|X} : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ and $\lambda \in [0,1]$. Let $W : \Omega \to \{0,1\}$ be an independent Bernoulli random variable with mean $\mathbb{E}W = \lambda$, to define $Z \triangleq \bar{W}Y_0 + WY_1$. Then, we observe that $P_{Z|X} = \bar{\lambda}P_{Y_0|X} + \lambda P_{Y_1|X}$. Since $\mathbb{E}_{X\sim P_X}\mathbb{E}P_{Y|X} = P_Y$, we get $P_Z = \mathbb{E}_{X\sim P_X}\mathbb{E}P_{Z|X} = \bar{\lambda}P_{Y_0} + \lambda P_{Y_1}$. Recall that the map $(P,Q) \mapsto D(P\|Q)$ is convex, we have

$$D(P_{Z|X}\otimes P_X\|P_Z\otimes P_X) \leqslant \bar{\lambda}D(P_{Y_0|X}\otimes P_X\|P_{Y_0}\otimes P_X) + \lambda D(P_{Y_1|X}\otimes P_X\|P_{Y_1}\otimes P_X).$$

The result follows from recognizing that $I(X;Y) = D(P_{X,Y}\|P_X\otimes P_Y)$. □

2 Rate-distortion theory

Definition 2.1 (Rate distortion). Consider parameter space Θ, prediction space Θ', and loss function $L : \Theta\times\Theta' \to \mathbb{R}$. We define the rate distortion function $\phi_\theta : \mathbb{R} \to \mathbb{R}$ for each $D \in \mathbb{R}$ as

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

Theorem 2.2 (General converse). *Suppose $X \to W \to \hat{X}$, where $W \in [M]$ and $\mathbb{E}L(X,\hat{X}) \leqslant D$. Then*

$$\ln M \geqslant \phi_X(D) \triangleq \inf_{P_{\hat{X}|X}:\mathbb{E}L(X,\hat{X})\leqslant D} I(X;\hat{X}).$$

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

Definition 2.3. We define maximum distortion as $D_{\max} \triangleq \inf_{\hat{\theta}} \mathbb{E}_{\theta\sim\pi} L(\theta,\hat{\theta})$ for a deterministic $\hat{\theta}$.

Remark 3. By definition, $D_{\max}$ is the distortion attainable without any information. Indeed, if $D_{\max} = \mathbb{E}L(X,\hat{\theta})$ 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 2.4 (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. Recall that $I(\theta;\hat{\theta}) = \mathbb{E}_{\theta\sim\pi} D(P_{\theta,\hat{\theta}} \| P_\theta \otimes P_{\hat{\theta}}) = D(P_{\hat{\theta}|\theta} \| P_{\hat{\theta}} \mid \pi) = \mathbb{E}_{\theta\sim\pi}\mathbb{E}_{P_{\hat{\theta}|\theta}} \ln \frac{dP_{\hat{\theta}|\theta}}{dP_{\hat{\theta}}}$

(a) Since $I(\theta;\hat{\theta}) = D(P_{\hat{\theta}|\theta} \| P_{\hat{\theta}} \mid \pi)$ and $P_{\hat{\theta}}$ is linear in $P_{\hat{\theta}|\theta}$, it follows that the map $P_{\hat{\theta}|\theta} \mapsto D(P_{\hat{\theta}|\theta} \| P_{\hat{\theta}})$ is convex in $P_{\hat{\theta}|\theta}$ for fixed prior $\pi \in \mathcal{M}(\Theta)$, as shown in Lemma 1.8. Since ϕ_θ is infimum of $I(\theta;\hat{\theta})$ over a set of Markov kernels $P_{\hat{\theta}|\theta}$, and the infimum of convex functions is convex, the result follows.

(b) For any $D > D_{\max}$ we can set $\hat{\theta}$ deterministically. Thus $I(\theta;\hat{\theta}) = 0$.

□

Theorem 2.5 (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 $L(S,\hat{S}) = \frac{1}{m}\sum_{i=1}^m L_1(S_i,\hat{S}_i)$, 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. Consider an estimate $\hat{S}$ such that $P_{\hat{S}|S} \triangleq P_{\hat{S}_1|S_1}^{\otimes m}$ where $\mathbb{E}L_1(S_i,\hat{S}_i) \leqslant D$ for all $i \in [m]$. Then $\hat{S}$ is a feasible estimate with $\mathbb{E}L(S,\hat{S}) \leqslant D$. 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 S is defined as

$$\phi_S(D) \triangleq \inf_{P_{\hat{S}|S}:\mathbb{E}L(S,\hat{S})\leqslant D} I(S;\hat{S}) \leqslant \inf_{P_{\hat{S}|S}=P_{\hat{S}_1|S_1}^{\otimes m}:\mathbb{E}L_1(S_i,\hat{S}_i)\leqslant D, i\in[m]} \sum_{i=1}^m I(S_i;\hat{S}_i) \leqslant \sum_{i=1}^m \inf_{P_{\hat{S}_i|S_i}:\mathbb{E}L_1(S_i,\hat{S}_i)\leqslant D} I(S_i;\hat{S}_i) = m\phi_{S_1}(D).$$

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

For the converse, we focus on any Markov kernel $P_{\hat{S}|S}$ that satisfies the constraint $\mathbb{E}L(S,\hat{S}) \leqslant D$. From the super-additivity property of mutual information for memoryless source in Theorem 1.4 (b), 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 $\phi_{S_1}(\mathbb{E}L_1(S_i;\hat{S}_i)) \leqslant I(S_i;\hat{S}_i)$. From convexity and non-increasing property of rate distortion function in Theorem 2.4, we obtain

$$I(S;\hat{S}) \geqslant \sum_{i=1}^m I(S_i;\hat{S}_i) \geqslant \sum_{i=1}^m \phi_{S_1}(\mathbb{E}L_1(S_i;\hat{S}_i)) \geqslant m\phi_{S_1}\Big(\frac{1}{m}\sum_{i=1}^m \mathbb{E}L_1(S_i;\hat{S}_i)\Big) \geqslant m\phi_{S_1}(D).$$

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

Theorem 2.6 (Rate distortion for Gaussian sources). *Let $S \sim \mathcal{N}(0,\sigma^2 I_d)$ and $L(s,\hat{s}) \triangleq \|s-\hat{s}\|_2^2$ for $s,\hat{s} \in \mathbb{R}^d$, then rate distortion function $R(D) \triangleq \inf_{P_{\hat{S}|S}:\mathbb{E}L(S,\hat{S})\leqslant D} I(S;\hat{S}) = \frac{d}{2}\ln^+ \frac{d\sigma^2}{D}$.*

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, and the forward channel is $P_{\hat{S}|S} = \mathcal{N}(\frac{\sigma^2-D}{\sigma^2}S, \frac{\sigma^2-D}{\sigma^2}D)$. This is due to the fact that S is Gaussian with mean 0 and variance σ^2,

and the conditional density is

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

Then $R(D) \leqslant I(S;\hat{S}) = \frac{1}{2}\ln\frac{\sigma^2}{D}$.

(b) **Converse.** Let $S \sim \mathcal{N}(0,\sigma^2)$ and $P_{\hat{S}|S}$ be any conditional distribution such that $\mathbb{E}_P L(S,\hat{S}) \leqslant D$. 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 $\mathbb{E}_P L(S,\hat{S}) \leqslant D$, 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 2.5 using the convexity of the rate-distortion function. □