# Lecture-21: Mutual Information Method

## 1 Introduction

In this chapter we describe a strategy for proving statistical lower bound we call the *mutual information method* (MIM), which entails comparing the amount of information data provides with the minimum amount of information needed to achieve a certain estimation accuracy. The main information-theoretical ingredient is the data-processing inequality, this time for mutual information as opposed to $f$-divergences.

Here is the main idea of the MIM. Fix some prior $\pi \in \mathcal{M}(\Theta)$ and we aim to lower bound the Bayes risk $R^*_\pi$ of estimating $\theta \sim \pi$ on the basis of $X$ with respect to some loss function $L : \Theta \times \hat{\Theta} \to \mathbb{R}$. Let $\hat{\theta}$ be an estimator such that $\mathbb{E}[L(\theta, \hat{\theta})] \leqslant D$. Then we have the Markov chain $\theta \to X \to \hat{\theta}$. Applying the data processing inequality for mutual information, we have

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

*Remark* 1. We observe the following for the above inequality.

(a) The leftmost quantity can be interpreted as the minimum amount of information required to achieve a given estimation accuracy. This is precisely the rate-distortion function $\phi(D) \equiv \phi_\theta(D)$.

(b) The rightmost quantity can be interpreted as the amount of information provided by the data about the latent parameter. Sometimes it suffices to further upper-bound it by the capacity of the channel $P_{X|\theta}$ by maximizing over all priors. That is,

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

Therefore, we arrive at the following lower bound on the Bayes and hence the minimax risks

$$R^*_\pi \geqslant \phi^{-1}(I(\theta;X)) \geqslant \phi^{-1}(C).$$

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 (1) depends only on the prior (source) and the loss function, while the capacity upper bound (b) depends only on the statistical model (channel).

We next discuss a sequence of examples to illustrate the MIM and its execution:

(a) Denoising a vector in Gaussian noise, where we will compute the exact minimax risk;

(b) Denoising a sparse vector, where we determine the sharp minimax rate;

(c) Community detection, where the goal is to recover a dense subgraph planted in a bigger Erdös-Rényi graph.

Subsequently, 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. We discuss the problem of how and when is non-trivial estimation achievable by applying the MIM. For this purpose, none of the three methods works.

### 1.1 GLM revisited and the Shannon lower bound

Consider the $d$-dimensional GLM, where we observe an *i.i.d.* sample $X : \Omega \to \mathbb{R}^m$ 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 $L : (\theta, \hat{\theta}) \mapsto \left\|\hat{\theta} - \theta\right\|_2^2$. First, let us consider the unconstrained model where $\Theta \triangleq \mathbb{R}^d$. Estimating using the sample mean $\bar{X} \triangleq \frac{1}{m}\sum_{i=1}^m X_i \sim N(\theta, \frac{1}{m} I_d)$, 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.

Again, let us consider $\theta \sim \mathcal{N}(0, sI_d)$ for some $s > 0$. We know from the Gaussian rate-distortion function (Theorem 26.2) that

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

# A Rate-distortion theory

**Lemma A.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 by applying the data processing inequality $D(P_{X,Y}\|Q_{X,Y}) \geqslant D(P_Y\|Q_Y)$ for divergences. □

*Remark* 2. 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 A.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

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

**Lemma A.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

$$I(X; Y) = \mathbb{E}_{P_X P_{Y|X}} \ln \frac{P_{Y|X}}{P_Y} = \mathbb{E}_{P_X} D(P_{Y|X}\|P_Y) = D(P_{Y|X}\|P_Y \mid P_X).$$

□

**Theorem A.4 (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 A.5.** *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 $P_{Y|X}$.*

**Definition A.6 (Rate distortion).** Consider parameter space $\Theta$, prediction space $\Theta'$, 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

$$\phi_\theta(D) \triangleq \inf_{P_{\hat{\theta}|\theta}: \mathbb{E} L(\theta, \hat{\theta}) \leqslant D} I(\theta; \hat{\theta}). \tag{2}$$

**Theorem A.7 (General converse).** *Suppose $X \to W \to \hat{X}$, where $W \in [M]$ and $\mathbb{E}L(X,\hat{X}) \leqslant D$. Then*

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

*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 A.8.** 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 $D_{\max} = \mathbb{E}d(X,\hat{\theta})$ for some fixed $\hat{\theta}$, then this $\hat{\theta}$ is the "default" reconstruction of $\theta$, i.e., the best estimate when we have no information about $\theta$. 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 A.9 (Properties).** *The following properties are true for rate distortion function $\phi_\theta : \mathbb{R} \to \mathbb{R}$.*
*(a) The map $\phi_\theta$ is convex and non-increasing.*
*(b) $\phi_\theta(D) = 0$ for all $D > D_{\max}$.*

*Proof.* Recall that $I(\theta;\hat{\theta}) = \mathbb{E}_{\theta\sim\pi} D(P_{\theta,\hat{\theta}} \| P_\theta \otimes P_{\hat{\theta}}) = D(P_{\hat{\theta}|\theta} \| P_{\hat{\theta}} \mid \pi) = \mathbb{E}_{\theta\sim\pi}\mathbb{E}_{P_{\hat{\theta}|\theta}} \ln \frac{P_{\hat{\theta}|\theta}}{P_{\hat{\theta}}}$

(a) Since $I(\theta;\hat{\theta}) = D(P_{\hat{\theta}|\theta} \| P_{\hat{\theta}} \mid \pi)$ and $P_{\hat{\theta}}$ is linear in $P_{\hat{\theta}|\theta}$, we get that $P_{\hat{\theta}|\theta} \mapsto D(P_{\hat{\theta}|\theta} \| P_{\hat{\theta}})$ is convex for each realization $\theta$. Infimum of convex functions is convex, and the result follows.
(b) For any $D > D_{\max}$ we can set $\hat{\theta}$ deterministically. Thus $I(\theta;\hat{\theta}) = 0$.

□

**Theorem A.10 (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,*

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

*Proof.* We utilize the definition of mutual information.
(a) From the definition of mutual information, we write

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

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

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

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

□

*Remark* 4. 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.

**Theorem A.11 (Single-letterization).** *For stationary memoryless source $S:\Omega\to\mathcal{S}^m$ with common distribution $P\in\mathcal{M}(\mathcal{S})$ and separable loss $L$ such that $L(S,\hat{S})=\frac{1}{m}\sum_{i=1}^m L_i(S_i,\hat{S}_i)$, 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 $P_{\hat{S}|S}\triangleq\prod_{i=1}^m P_{\hat{S}_i|S_i}$ where $\mathbb{E}L_i(S_i,\hat{S}_i)\leqslant D$ for all $i\in[m]$. Then $\hat{S}$ is a feasible estimate with $\mathbb{E}L(S,\hat{S})\leqslant D$. Further, for this definition of estimate $\hat{S}$ is memoryless and stationary, since $S$ 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

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

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, for any $P_{\hat{S}|S}$ satisfying the constraint $\mathbb{E}L(S,\hat{S})\leqslant D$, we have

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

In the first step we used the crucial super-additivity property of mutual information. □

**Theorem A.12 (Rate distortion for Gaussian sources).** *Let $S\sim\mathcal{N}(0,\sigma^2I_d)$ and $d(s,\hat{s})\triangleq\|s-\hat{s}\|_2^2$ for $s,\hat{s}\in\mathbb{R}^d$, then rate distortion function $R(D)\triangleq\inf_{P_{\hat{S}|S}:\mathbb{E}d(S,\hat{S})\leqslant D}I(S;\hat{S})=\frac{d}{2}\ln^+\frac{d\sigma^2}{D}$.*

*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 $P_{\hat{S}|S}=\mathcal{N}(\frac{\sigma^2-D}{\sigma^2}S,\frac{\sigma^2-D}{\sigma^2}D)$. 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 $\mathbb{E}_P(S-\hat{S})^2\leqslant D$. Denote the forward channel in the above achievability by $P^*_{\hat{S}|S}$. Then, we have

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

From the non-negativity of KL divergence and definition of $P^*_{\hat{S}|S}$, we write

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

Finally, for the vector case follows from the scalar case and the same single-letterization argument in Theorem 24.8 using the convexity of the rate-distortion function in Theorem 24.4(a). □