Lecture-22: 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 mutual information data-processing inequality.

Definition 1.1. 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.

$$I(\theta;X) \leqslant \sup_{\pi\in P(\Theta)} I(\theta;X) \triangleq C. \tag{1}$$

Theorem 1.2 (Mutual information method (MIM)). *Consider a simple statistical decision theory setting with parameter space Θ, prediction space $\hat{\Theta}$, estimate $\hat{\theta} : \mathcal{X} \times [0,1] \to \hat{\Theta}$ for external randomness $U : \Omega \to [0,1]$ independent of everything, and loss function $L : \Theta \times \hat{\Theta} \to \mathbb{R}$. If $\pi \in \mathcal{M}(\Theta)$ is a prior on the parameter space, then minimax and Bayes risk are lower bounded as*

$$R^* \geqslant R^*_\pi = \inf_{P_{\hat{\theta}|\theta}} \mathbb{E}L(\theta,\hat{\theta}) \geqslant \phi^{-1}(I(\theta;X)) \geqslant \phi^{-1}(C). \tag{2}$$

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 $L : \Theta \times \hat{\Theta} \to \mathbb{R}$. Let $\hat{\theta}(X,U)$ 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

$$\phi_\theta(D) \triangleq \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}L(\theta,\hat{\theta})\leqslant D} I(\theta;\hat{\theta}) \leqslant I(\theta;\hat{\theta}) \leqslant I(\theta;X) \leqslant \sup_{\pi\in\mathcal{M}(\Theta)} I(\theta;X) = C. \tag{3}$$

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

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

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

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

(a) The quantity $\inf_{P_{\hat{\theta}|\theta}:\mathbb{E}L(\theta,\hat{\theta})\leqslant D} I(\theta;\hat{\theta})$ 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 (3) depends only on the prior (source) and the loss function, while the capacity upper bound (1) 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

Example 1.3 (GLM). 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 $\bar{X} \triangleq \frac{1}{m}\sum_{i=1}^m X_i \sim \mathcal{N}(\theta, \frac{1}{m}I_d)$, 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.

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 $\bar{X}$ and the formula of Gaussian channel capacity A.7, 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 (2) 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 $R^*(\mathbb{R}^d) = \frac{d}{m}$.

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. Next, let us consider the GLM with bounded means, where the parameter space $\Theta \triangleq B(\rho) \triangleq \left\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leqslant \rho\right\}$ is the ℓ_2-ball of radius ρ centered at zero. In this case there is no known closed- form formula for the minimax quadratic risk even in one dimension[1]. Nevertheless, the next result determines the sharp minimax rate, which characterizes the minimax risk up to universal constant factors.

Theorem 1.4 (Bounded GLM). *$R^*(B(\rho)) \asymp \frac{d}{m} \wedge \rho^2$.*

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

$$I(\theta; X) = I(\theta; \bar{X}) \leqslant \sup_{P_\theta : \mathbb{E}\|\theta\|_2^2 \leqslant r} I(\theta; \theta + \frac{1}{\sqrt{m}}Z) = \frac{d}{2}\ln\left(1 + \frac{mr^2}{d}\right) \leqslant \frac{mr^2}{2},$$

where $Z \sim \mathcal{N}(0, I_d)$. Alternatively, we can use Corollary 5.8 to bound the capacity (as information radius) by the KL diameter, which yields the same bound within constant factors,

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

[1] It is known that there exists some ρ_0 depending on d/m such that for all $\rho \leqslant \rho_0$, the uniform prior over the sphere of radius ρ is exactly least favorable (see [81] for $d = 1$ and [48] for $d > 1$.)

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

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

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 (2) yields $\frac{1}{2}\ln\frac{cr^2}{R^*}\leqslant\frac{mr^2}{2}$, i.e., $R^*\geqslant cr^2e^{-\frac{mr^2}{d}}$. 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

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

□

Remark 2. 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 Channel capacity

A.1 Geometric interpretation of channel capacity

Mutual information (MI) can be understood as a weighted “distance” from the conditional distributions to the marginal distribution. Indeed, for a discrete random variable $X:\Omega\to\mathcal{X}$, we have

$$I(X;Y)=D(P_{Y|X}\|P_Y\mid P_X)=\sum_{x\in\mathcal{X}}D(P_{Y|X=x}\|P_Y)P_X(x).$$

Furthermore, it turns out that P_Y, similar to the center of gravity, minimizes this weighted distance and thus can be thought as the best approximation for the “center” of the collection of distributions $\left\{P_{Y|X=x}:x\in\mathcal{X}\right\}$ with weights given by P_X. We formalize these results in this section and start with the proof of a “golden formula”.

Theorem A.1 (Golden formula). *For any Q_Y we have $D(P_{Y|X}\|Q_Y\mid P_X)=I(X;Y)+D(P_Y\|Q_Y)$. Thus, if $D(P_Y\|Q_Y)<\infty$, then $I(X;Y)=D(P_{Y|X}\|Q_Y\mid P_X)-D(P_Y\|Q_Y)$.*

Proof. In the discrete case and ignoring the possibility of dividing by zero, the argument is really simple. simple. We just need to write

$$I(X;Y)=\mathbb{E}_{P_{X,Y}}\ln\frac{P_{Y|X}}{P_Y}=\mathbb{E}_{P_{X,Y}}\ln\frac{P_{Y|X}Q_Y}{P_YQ_Y},$$

and then expand $\ln\frac{P_{Y|X}Q_Y}{P_YQ_Y}=\ln\frac{P_{Y|X}}{Q_Y}-\ln\frac{Q_Y}{P_Y}$. The argument below is a rigorous implementation of this idea.

First, notice that by Theorem 2.16(e) we have $D(P_{Y|X}\|Q_Y\mid P_X)\geqslant D(P_Y\|Q_Y)$ and thus if $D(P_Y\|Q_Y)=\infty$ then both sides of (4.2) are infinite. Thus, we assume $D(P_Y\|Q_Y)<\infty$ and in particular $P_Y\ll Q_Y$. Rewriting LHS of (4.2) via the chain rule (2.24) we see that Theorem amounts to proving

$$D(P_{X,Y}\|P_XQ_Y)=D(P_{X,Y}\|P_XP_Y)+D(P_Y\|Q_Y).$$

The case of $D(P_{X,Y}\|P_XQ_Y)=D(P_{X,Y}\|P_XP_Y)=\infty$ is clear. Thus, we can assume at least one of these divergences is finite, and, hence, also $P_{X,Y}\ll P_XQ_Y$. Let $\lambda(y)\triangleq\frac{dP_Y}{dQ_Y}(y)$. Since $\lambda(Y)>0$, P_Y-a.s., applying the definition of *Log* in (2.10), we can write

$$\mathbb{E}_{P_Y}\ln\lambda(Y)=\mathbb{E}_{P_{X,Y}}Log\frac{\lambda(Y)}{1}.$$

Notice that the same $\lambda(y)$ is also the density $\frac{dPXP_Y}{dP_XQ_Y}(x,y)$ of the product measure P_XP_Y with respect P_XQ_Y. Therefore, the RHS of (4.4) by (2.11) applied with $\mu = P_XQ_Y$ coincides with $D(P_{X,Y}\|P_XQ_Y) - D(P_{X,Y}\|P_XP_Y)$, while the LHS of (4.4) by (2.13) equals $D(P_Y\|Q_Y)$. Thus, we have shown the required $D(P_Y\|Q_Y) = D(P_{X,Y}\|P_XQ_Y) - D(P_{X,Y}\|P_XP_Y)$. □

Corollary A.2 (Mutual information as center of gravity). *For any Q_Y we have $I(X;Y) \leqslant D(P_{Y|X}\|Q_Y \mid P_X)$. Consequently $I(X;Y) = \min_{Q_Y} D(P_{Y|X}\|Q_Y \mid P_X)$. If $I(X;Y) < \infty$, the unique minimizer is $Q_Y = P_Y$.*

Theorem A.3. *For any Markov kernel $Q_{X|Y}$ such that $Q_{X|Y=y} \ll P_X$ for P_Y-a.e. y we have*

$$I(X;Y) \geqslant \mathbb{E}_{P_{X,Y}} \ln \frac{dQ_{X|Y}}{dP_X}.$$

If $I(X;Y) < \infty$, then $I(X;Y) = \sup_{Q_{X|Y} \ll P_X} \mathbb{E}_{P_{X,Y}} \ln \frac{dQ_{X|Y}}{dP_X}$, where the supremum is over Markov kernels $Q_{X|Y}$ as in the first sentence.

Proof. Since modifying $Q_{X|Y=y}$ on a negligible set of y's does not change the expectations, we will assume that $Q_{X|Y=y} \ll P_Y$ for every y. If $I(X;Y) = \infty$ then there is nothing to prove. So we assume $I(X;Y) < \infty$, which implies $P_{X,Y} \ll P_XP_Y$. Then by Lemma 3.3, we have that $P_{X|Y=y} \ll P_X$ for almost every y. Choose any such y and apply (2.11) with $\mu = P_X$ and noticing $\mathrm{Log}\frac{dQ_{X|Y=y}}{dP_X} = \ln dQ_{X|Y=y}dP_X$, we get

$$\frac{dQ_{X|Y=y}}{dP_X} = D(P_{X|Y=y}\|P_X) - D(P_{X|Y=y}\|Q_{X|Y=y}),$$

identity over y we obtain $\mathbb{E}P_{X|Y=y}$ which is applicable since the first term is finite for a.e. y by (3.1). Taking expectation of the previous identity over y, we obtain

$$\mathbb{E}_{P_{X,Y}}[\ln \frac{dQ_{X|Y}}{dP_X}] = I(X;Y) - D(P_{X|Y}\|Q_{X|Y} \mid P_Y) \leqslant I(X;Y),$$

implying the first part. The equality for $I(X;Y) < \infty$ follows by taking $Q_{X|Y} = P_{X|Y}$, which satisfies the conditions on Q when $I(X;Y) < \infty$. □

A.2 Saddle point of mutual information

Definition A.4. Let $\mathcal{P}$ be a convex set of distributions on $\mathcal{X}$. Suppose there exists $P_X^* \in \mathcal{P}$, called a *capacity-achieving input distribution*, such that

$$C \triangleq I(P_X^*, P_{Y|X}) = \sup_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}).$$

Then $P_Y^* \triangleq P_{Y|X} \circ P_X$ is called a *capacity-achieving output distribution*.

Theorem A.5 (Saddle point). *Let $\mathcal{P}$ be a convex set of distributions on $\mathcal{X}$. Then for all $P_X \in \mathcal{P}$ and for all Q_Y, we have*

$$D(P_{Y|X}\|P_Y^* \mid P_X) \leqslant D(P_{Y|X}\|P_Y^* \mid P_X^*) \leqslant D(P_{Y|X}\|Q_Y \mid P_X^*). \qquad (4)$$

Proof. Right inequality in (4) follows from $C = I(P_X^*, P_{Y|X}) = \min_{Q_Y} D(P_{Y|X}\|Q_Y \mid P_X^*)$ from Corollary A.2. The left inequality in (4) is trivial when $C = \infty$. Hence, we assume that $C < \infty$ without any loss of generality. Therefore, $I(P_X, P_{Y|X}) \leqslant C\infty$ for all $P_X \in \mathcal{P}$. Let $P_{X_\lambda} = \lambda P_X + \lambda P_X^* \in \mathcal{P}$ and $P_{Y_\lambda} = P_{Y|X} \circ P_{X_\lambda}$. Clearly, $P_{Y_\lambda} = \lambda P_Y + \lambda P_Y^*$, where $P_Y = P_{Y|X} \circ P_X$. Consequently, we have the following chain

$$\begin{aligned} C \geqslant I(X_\lambda; Y_\lambda) = D(P_{Y|X}\|P_{Y_\lambda} \mid P_{X_\lambda}) &= \lambda D(P_{Y|X}\|P_{Y_\lambda} \mid P_X) + \bar{\lambda} D(P_{Y|X}\|P_{Y_\lambda} \mid P_X^*) \\ &\geqslant \lambda D(P_{Y|X}\|P_{Y_\lambda} \mid P_X) + \bar{\lambda} C = \lambda D(P_{X,Y}\|P_XP_{Y_\lambda}) + \bar{\lambda} C, \end{aligned}$$

where inequality follows from the second inequality of (4) which is already shown. Thus, subtracting $\bar{\lambda} C$ and dividing by λ we get $D(P_{X,Y}\|P_XP_{Y_\lambda}) \leqslant C$ and the proof is completed by taking $\liminf_{\lambda \to 0}$ and applying the lower semincontinuity of divergence (Theorem 4.9). □

Corollary A.6. *In addition to the assumptions of Theorem A.5, suppose $C < \infty$. Then the capacity-achieving output distribution P_Y^* is unique. It satisfies the property that for any P_Y induced by some $P_X \in \mathcal{P}$, i.e. $P_Y = P_{Y|X} \circ P_X$, we have $D(P_Y\|P_Y^*) \leqslant C < \infty$ and in particular $P_Y \ll P_Y^*$.*

Proof. The statement is $I(P_X, P_{Y|X}) = C$ implies $P_Y = P_Y^*$. Indeed

$$C = D(P_{Y|X} \| P_Y \mid P_X) = D(P_{Y|X} \| P_Y^* \mid P_X) - D(P_Y \| P_Y^*) \leqslant D(P_{Y|X} \| P_Y^* \mid P_X^*) - D(P_Y \| P_Y^*) = C - D(P_Y \| P_Y^*)$$

implies $P_Y = P_Y^*$. The statement $D(P_Y \| P_Y^*) \leqslant C < \infty$ follows from the left inequality in (4) and "conditioning increases divergence" property in Theorem 2.16. □

A.3 Gaussian channel capacity

Theorem A.7 (Gaussian channel capacity). *Consider two independent zero mean Gaussian random variables $X_g \sim \mathcal{N}(0, \sigma_X^2)$ and $N_g \sim \mathcal{N}(0, \sigma_N^2)$. Then the following statement are true.*

(a) ***Gaussian capacity.*** $C = I(X_g; X_g + N_g) = \frac{1}{2} \ln\left(1 + \frac{\sigma_X^2}{\sigma_N^2}\right)$.

(b) ***Gaussian input is the best for Gaussian noise.*** *For all random variables X with variance $\mathrm{Var}(X) \leqslant \sigma_X^2$ independent of N_g, we have $I(X; X + N_g) \leqslant I(X_g; X_g + N_g)$ with equality iff $F_X = F_{X_g}$.*

(c) ***Gaussian noise is the worst for Gaussian input.*** *For all random variables N such that $\mathbb{E}X_g N = 0$ and $\mathbb{E}N^2 \leqslant \sigma_N^2$, we have $I(X_g; X_g + N) \geqslant I(X_g; X_g + N_g)$ with equality iff $F_N = F_{N_g}$ and N independent of X_g.*

Proof. WLOG, we assume that all random variables have zero mean. Let $Y_g \triangleq X_g + N_g$. Recall that $C = \frac{1}{2} \ln\left(1 + \frac{\sigma_X^2}{\sigma_N^2}\right)$, and define

$$f(x) \triangleq D(P_{Y_g|X_g = x} \| P_{Y_g}) = D(\mathcal{N}(x, \sigma_N^2) \| \mathcal{N}(0, \sigma_X^2 + \sigma_N^2)) = C + \frac{1}{2} \frac{(x^2 - \sigma_X^2)}{\sigma_X^2 + \sigma_N^2}.$$

(a) Compute $I(X_g; X_g + N_g) = \mathbb{E}f(X_g) = C$.

(b) Recall the inf-representation from Corollary A.2 that implies $I(X;Y) = \min_Q D(P_{Y|X} \| Q \mid P_X)$, i.e.

$$I(X; X + N_g) \leqslant D(P_{Y_g|X_g} \| P_{Y_g} \mid P_X) = \mathbb{E}f(X) \leqslant C < \infty.$$

Furthermore, if $I(X; X + N_g) = C$, then the uniqueness of the capacity-achieving output distribution from Corollary A.6, we get $P_Y = P_{Y_g}$. But $P_Y = P_X * \mathcal{N}(0, \sigma_N^2)$,where $*$ denotes convolution. Then it must be that $X \sim \mathcal{N}(0, \sigma_X^2)$ simply by considering characteristic functions,

$$\Psi_X(t) e^{-\frac{1}{2}\sigma_N^2 t^2} = e^{-\frac{1}{2}(\sigma_X^2 + \sigma_N^2) t^2}.$$

It follows that $\Psi_X(t) = e^{-\frac{1}{2}\sigma_X^2 t^2}$, and therefore $X \sim \mathcal{N}(0, \sigma_X^2)$.

(c) Let $Y = X_g + N$ and let $P_{Y|X_g}$ be the associated kernel such that $\mathbb{E}X_g N = 0$ and $\mathbb{E}N^2 \leqslant \sigma_N^2$. It follows that $\mathbb{E}Y^2 = \mathbb{E}N^2 + \mathbb{E}X_g^2 \leqslant \sigma_N^2 + \sigma_X^2$. Note that here we only assume that N is uncorrelated with X_g, and not necessarily independent. Since $P_{X_g|X_g + N_g} \ll P_{X_g}$, we get from Theorem A.3

$$\begin{aligned} I(X_g; Y) &\geqslant \mathbb{E}_{P_{X_g,Y}} \ln \frac{dP_{X_g|Y_g}(X_g \mid Y)}{dP_{X_g}(X_g)} = \mathbb{E}_{P_{X_g,Y}} \ln \frac{dP_{Y_g|X_g}(Y \mid X_g)}{dP_{Y_g}(Y)} = C + \frac{1}{2} \mathbb{E}\Big[\frac{Y^2}{\sigma_X^2 + \sigma_N^2} - \frac{N^2}{\sigma_N^2}\Big] \\ &= C + \frac{1}{2} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2} \Big(1 - \frac{\mathbb{E}N^2}{\sigma_N^2}\Big) \geqslant C. \end{aligned}$$

From Theorem A.3, the conditions for first equality in above equation requires

$$D(P_{X_g|Y} \| P_{X_g|Y_g} \mid P_Y) = 0.$$

Thus, $P_{X_g|Y} = P_{X_g|Y_g}$, i.e., X_g is conditionally Gaussian and $P_{X_g|Y=y} = \mathcal{N}(by, c^2)$ for some constants b and c. In other words, under $P_{X_g Y}$, we have $X_g = bY + cZ$ where Z is a Gaussian random variable independent of Y. This implies that Y must be Gaussian itself by Cramer's Theorem [106] or simply by considering characteristic functions, where $\Psi_Y(t) e^{ct^2} = e^{c' t^2}$ implies $\Psi_Y(t) = e^{c'' t^2}$, i.e. Y is Gaussian. Therefore, (X_g, Y) must be jointly Gaussian and hence $N = Y - X_g$ is Gaussian. Thus we conclude that it is only possible to attain $I(X_g; X_g + N) = C$ if N is Gaussian of variance σ_N^2 and independent of X_g.

□

Remark 3. This result encodes extremality properties of the normal distribution: for the AWGN channel, Gaussian input is the most favorable, i.e. attains the maximum mutual information or capacity, while for a general additive noise channel the least favorable noise is Gaussian. For a vector version of the former statement see Exercise I.9.