Lecture-25: Mutual information and channel capacity

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. Consider state space $\mathcal{X} \triangleq \{0,1\}$, Bernoulli random variable $X : \Omega \to \mathcal{X}$ with distribution $P_X = Q_X \in \mathcal{M}(\mathcal{X})$ having mean $\lambda \in [0,1]$. Let $P_0, P_1, Q_0, Q_1 \in \mathcal{M}(\mathcal{Y})$ and define Markov kernels

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

The divergence of two joint distributions $P_{X,Y}$ and $Q_{X,Y}$ in terms of conditional divergence, is given by

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

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

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

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 $P_{X|Y} = \prod_{i=1}^m P_{X_i|Y}$-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

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

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). We define *conditional mutual information* between random variables X and Y given Z as

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

where the product $P_{X|Z}P_{Y|Z}$ is a conditional distribution 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

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

□

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 $X \to Z \to Y$ is also a Markov chain.*

Proof. Since $X \to Y \to Z$ is a Markov chain, random variables X and Z are conditionally independent given Y, and hence $I(X;Z \mid Y) = 0$. 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 $I(X;Y \mid Z) \geqslant 0$ with equality iff $X \to Z \to Y$ is also a Markov chain. □

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

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

$$I(Z;X) = D(P_{Z|X}P_X\|P_ZP_X) \leqslant \bar{\lambda}D(P_{Y_0|X}P_X\|P_{Y_0}P_X) + \lambda D(P_{Y_1|X}P_X\|P_{Y_1}P_X) = \bar{\lambda}I(Y_0;X) + \lambda I(Y_1;X).$$

(b) Consider random variables X_0, X_1 such that $X = \bar{W}X_0 + WX_1$, and hence $P_X = \bar{\lambda}P_{X_0} + \lambda P_{X_1}$. Then $W \to X \to Y$ is a Markov chain and $I(W;Y \mid X) = 0$. Therefore, by chain rule of mutual information and the fact that mutual information is non negative, we get

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

Since $I(X;Y \mid W) = \bar{\lambda}I(X_0;Y) + \lambda I(X_1;Y)$, we get the result.

□

A Channel capacity

Definition A.1. Define a bivariate function $\mathrm{Log}\frac{a}{b}: \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R} \cup \{\pm\infty\}$ by

$$\mathrm{Log}\frac{a}{b} = -\infty \mathbb{1}_{\{a=0,b>0\}} + \infty \mathbb{1}_{\{a>0,b=0\}} + 0\mathbb{1}_{\{a=0,b=0\}} + \ln\frac{a}{b}\mathbb{1}_{\{a>0,b>0\}}.$$

Remark 3. Let $P, Q, R \ll \mu$ and f_P, f_Q, f_R denote their densities relative to μ.

(a) $\mathbb{E}_P \mathrm{Log}\frac{f_R}{f_Q}$ exists and $D(P\|Q) - D(P\|R) = \mathbb{E}_P \mathrm{Log}\frac{f_R}{f_Q}$ if at least one of the divergences is finite.

(b) $\mathbb{E}_P \mathrm{Log}\frac{f_P}{f_Q}$ is well-defined but possibly infinite, and $D(P\|Q) = \mathbb{E}_P \mathrm{Log}\frac{f_P}{f_Q}$. In particular, when $P \ll Q$ we have $D(P\|Q) = \mathbb{E}_P \ln\frac{dP}{dQ}$.

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".

Exercise A.2. Show that conditioning increases divergence. That is, consider an input X and output Y under two different channels $P_{Y|X}, Q_{Y|X}$ that lead to output distributions P_Y, Q_Y respectively. Then, show that $D(P_Y\|Q_Y) \leqslant D(P_{Y|X}\|Q_{Y|X} \mid P_X)$ with equality iff $D(P_{X|Y}\|Q_{X|Y} \mid P_Y) = 0$.

Theorem A.3 (Golden formula). *For any Q_Y we have*

$$D(P_{Y|X}\|Q_Y \mid P_X) = I(X;Y) + D(P_Y\|Q_Y). \tag{1}$$

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. We observe that

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

The argument below is a rigorous implementation of this idea.

From the fact that conditioning increases divergence, 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 (1) are infinite. Thus, we assume $D(P_Y\|Q_Y) < \infty$ and in particular $P_Y \ll Q_Y$. Hence, we can define $\lambda(y) \triangleq \frac{dP_Y}{dQ_Y}(y)$ for each $y \in \mathcal{Y}$. Rewriting LHS of (1) via the chain rule of divergence, we see that Theorem amounts to proving

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

The case of $D(P_{X,Y}\|P_X Q_Y) = D(P_{X,Y}\|P_X P_Y) = \infty$ is clear. Thus, we can assume at least one of these divergences is finite. Since $P_Y \ll Q_Y$, we have $P_X P_Y \ll P_X Q_Y$, and hence we can assume $P_{X,Y} \ll P_X Q_Y$ without loss of any generality. Since $\lambda(Y) > 0, P_Y$-a.s., applying the definition of Log in Definition A.1, we write

$$\mathbb{E}_{P_Y} \ln\lambda(Y) = \mathbb{E}_{P_{X,Y}} \mathrm{Log}\frac{\lambda(Y)}{1}. \tag{2}$$

Notice that the same $\lambda(y)$ is also the density $\frac{dP_X P_Y}{dP_X Q_Y}(x,y)$ of the product measure $P_X P_Y$ with respect $P_X Q_Y$. Therefore, the RHS of (2) by Remark 3(a) applied with $\mu = P_X Q_Y$ coincides with $D(P_{X,Y}\|P_X Q_Y) - D(P_{X,Y}\|P_X P_Y)$, while the LHS of (2) by Remark 3(b) equals $D(P_Y\|Q_Y)$. Thus, we have shown the required $D(P_Y\|Q_Y) = D(P_{X,Y}\|P_X Q_Y) - D(P_{X,Y}\|P_X P_Y)$. □

Corollary A.4 (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.5. *Let $\mathcal{A} \triangleq \left\{Q_{X|Y} : Q_{X|Y=y} \ll P_X \text{ for } P_Y - \text{ a.e. } y\right\}$.*

(a) For any Markov kernel $Q_{X|Y} \in \mathcal{A}$, we have $I(X;Y) \geqslant \mathbb{E}_{P_{X,Y}} \ln \frac{dQ_{X|Y}}{dP_X}$.

(b) If $I(X;Y) < \infty$, then $I(X;Y) = \sup_{Q_{X|Y}\in\mathcal{A}} \mathbb{E}_{P_{X,Y}} \ln \frac{dQ_{X|Y}}{dP_X}$.

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.

(a) If $I(X;Y) = \infty$ then there is nothing to prove. So we assume $I(X;Y) < \infty$, which implies $P_{X,Y} \ll P_X P_Y$. Recall that $P_{X,Y} \ll P_X P_Y$ iff $P_{X|Y=y} \ll P_X$ for all P_Y-a.e. $y \in \mathcal{X}$. For any such y, apply Remark 3(a) with $\mu = P_X$, and observe that $\operatorname{Log} \frac{dQ_{X|Y=y}/dP_X}{1} = \ln \frac{dQ_{X|Y=y}}{dP_X}$, to get

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

which is applicable since the first term is finite for a.e. y by the definition of mutual information. Taking expectation of the previous identity over y, we obtain

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

(b) 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.6. Let $\mathcal{P} \subseteq \mathcal{M}(\mathcal{X})$ be a convex set. Suppose there exists $P_X^* \in \mathcal{P}$, called a *capacity-achieving input distribution*, such that

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

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

Theorem A.7 (Saddle point). *Let $\mathcal{P} \subseteq \mathcal{M}(\mathcal{X})$ be a convex set. Then for all $P_X \in \mathcal{P}$ and for all $Q_Y \in \mathcal{M}(\mathcal{Y})$,*

$$D(P_{Y|X}\|P_Y^* \mid P_X) \leqslant D(P_{Y|X}\|P_Y^* \mid P_X^*) \leqslant D(P_{Y|X}\|Q_Y \mid P_X^*). \tag{3}$$

Proof. Right inequality in (3) follows from $C = I(P_X^*, P_{Y|X}) = \min_{Q_Y} D(P_{Y|X}\|Q_Y \mid P_X^*)$ from Corollary A.4. The left inequality in (3) is trivial when $C = \infty$. Hence, we assume that $C < \infty$ without any loss of generality. Therefore, we assume $I(P_X, P_{Y|X}) < \infty$ for all $P_X \in \mathcal{P}$. Let $\lambda \in (0,1)$ and define $P_{X_\lambda} \triangleq \lambda P_X + \bar{\lambda} P_X^* \in \mathcal{P}$ and $P_{Y_\lambda} = \mathbb{E}_{X\sim P_{X_\lambda}} P_{Y|X}$. Clearly, $P_{Y_\lambda} = \lambda P_Y + \bar{\lambda} P_Y^*$, where $P_Y = \mathbb{E}_{X\sim P_X} P_{Y|X}$. Consequently, we have the following chain

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

where inequality follows from the second inequality of (3) which is already shown. Thus, subtracting $\bar{\lambda} C$ and dividing by λ we get $D(P_{X,Y}\|P_X P_{Y_\lambda}) \leqslant C$ and the proof is completed by taking $\liminf_{\lambda\to 0}$ and applying the lower semincontinuity of divergence (Theorem 4.9). □

Corollary A.8. *In addition to the assumptions of Theorem A.7, suppose $C < \infty$.*
(a) The capacity-achieving output distribution P_Y^* is unique.
(b) Let $P_X \in \mathcal{P}$ and $P_Y = \mathbb{E}_{X\sim P_X} P_{Y|X}$, then $D(P_Y\|P_Y^*) \leqslant C < \infty$ and in particular $P_Y \ll P_Y^*$.

Proof. Let $C = D(P_{Y|X}\|P_Y \mid P_X) < \infty$.

(a) Indeed, from the left inequality in (3) of Theorem A.7, we get

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

(b) The statement $D(P_Y\|P_Y^*) \leqslant C < \infty$ follows from the left inequality in (3) and "conditioning increases divergence" property.

□

A.3 Gaussian channel capacity

Theorem A.9 (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 $Y_g \triangleq X_g + N_g$. Denoting the relative density of Y_g as p_{Y_g} and the relative conditional density of Y_g given $X_g = x$ as $p_{Y_g|X_g=x}$, both with respect to Lebesgue measure, we recall that

$$\ln p_{Y_g}(y) = -\frac{1}{2}\ln 2\pi(\sigma_X^2 + \sigma_N^2) - \frac{y^2}{2(\sigma_X^2 + \sigma_N^2)}, \qquad \ln p_{Y_g|X_g=x}(y) = -\frac{1}{2}\ln 2\pi\sigma_N^2 - \frac{1}{2}\frac{(y-x)^2}{\sigma_N^2}.$$

We 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 $I(X_g; X_g + N_g) = \mathbb{E}_{X_g \sim P_{X_g}} f(X_g) = C$.

(b) Recall the inf-representation from Corollary A.4 that implies $I(X;Y) = \min_Q D(P_{Y|X} \| Q_Y \mid P_X)$, i.e.

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

Furthermore, if $I(X; X + N_g) = C$, then from the uniqueness of the capacity-achieving output distribution in Corollary A.8, we get $P_Y = P_{Y_g}$. Since $Y = X + N_g$ where N_g is independent of X, we can write the characteristic function of Y as

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

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.5

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

From Theorem A.5, 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 4. 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.