Lecture-29: Assouad's lemma

1 Assouad's Lemma

We saw that Le Cam's two-point method effectively only perturbs one out of d coordinates, leaving the remaining $d-1$ coordinates unexplored; this is the source of its suboptimality. In order to obtain a lower bound that scales with the dimension, it is necessary to randomize all d coordinates. Our next topic Assouad's Lemma is an extension in this direction.

Definition 1.1 (Hamming distance). We define the hypercube as the space of d length binary strings $H_d \triangleq \{0,1\}^d$. Hamming distance $\ell_H : H_d \times H_d \to \{0,\dots,d\}$ is defined as $\ell_H(b,b') \triangleq \sum_{i=1}^d \mathbb{1}_{\{b_i \neq b'_i\}}$ for any binary strings $b, b' \in H_d$.

Theorem 1.2 (Assouad's lemma). *Consider a simple statistical decision theory setting with $\Theta = \hat{\Theta}$ where*
(a) the loss function $\ell : \Theta \times \Theta \to \mathbb{R}_+$ that is an α-metric on parameter space Θ, and
(b) the parameter space Θ contains a subset $\Theta' \triangleq \{\theta_b \in \Theta : b \in H_d\}$ indexed by the hypercube H_d such that $\ell(\theta_b, \theta_{b'}) \geqslant \beta \ell_H(b,b')$ for all $b \neq b' \in H_d$ and some $\beta > 0$.
Then, the minimax risk $R^*(\Theta) \triangleq \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \ell(\theta,\hat{\theta})$ satisfies

$$R^*(\Theta) \geqslant \frac{\beta d}{4\alpha}\Big(1 - \max_{\ell_H(b,b')=1} \mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}})\Big). \tag{1}$$

Proof. We lower bound the minimax risk with the Bayes risk and that with the minimum risk for the uniform prior over Θ'. Given any estimator $\hat{\theta}(X)$, define $\hat{b}(X) \in \arg\min_{b\in H_d} \ell(\hat{\theta}(X), \theta_b)$. Then for any $b \in H_d$,

$$\beta \ell_H(\hat{b}(X), b) \leqslant \ell(\theta_{\hat{b}(X)}, \theta_b) \leqslant \alpha(\ell(\theta_{\hat{b}(X)}, \hat{\theta}(X)) + \ell(\hat{\theta}(X), \theta_b)) \leqslant 2\alpha \ell(\hat{\theta}(X), \theta_b).$$

Let $B : \Omega \to H_d$ be a discrete uniform random variable, and we have a Markov chain $B \to \theta_B \to X \to \hat{B}$. Then lower bounding the minimum average probability of error $P\left\{\hat{B}_i(X) \neq B_i\right\} \geqslant \frac{1}{2}(1 - \mathrm{TV}(P_{X|B_i=0}, P_{X|B_i=1}))$ in binary hypothesis testing for each $i \in [d]$, we obtain

$$\mathbb{E}\ell(\hat{\theta}(X), \theta_B) \geqslant \frac{\beta}{2\alpha}\mathbb{E}\ell_H(\hat{B}(X), B) = \frac{\beta}{2\alpha}\sum_{i=1}^d P\left\{\hat{B}_i(X) \neq B_i\right\} \geqslant \frac{\beta}{4\alpha}\sum_{i=1}^d (1 - \mathrm{TV}(P_{X|B_i=0}, P_{X|B_i=1})).$$

From the Bayes' theorem, we have $P_{X|B_i=0} = \frac{\sum_{b\in H_d: b_i=0} P_{X|\theta_b} P_B(b)}{P_{B_i=0}} = \frac{1}{2^{d-1}}\sum_{b\in H_d:b_i=0} P_{\theta_b}$. Similarly, we have $P_{X|B_i=1} = \frac{1}{2^{d-1}}\sum_{b\in H_d:b_i=1} P_{\theta_b}$. Recall that f-divergence is convex in both arguments, and hence the total variation distance is convex in both arguments. Therefore, the total variation term for each $i \in [d]$ can be upper bounded as

$$\mathrm{TV}(P_{X|B_i=0}, P_{X|B_i=1}) = \mathrm{TV}\Big(\frac{1}{2^{d-1}}\sum_{b\in H_d:b_i=0} P_{\theta_b}, \frac{1}{2^{d-1}}\sum_{b\in H_d:b_i=1} P_{\theta_b}\Big) \leqslant \frac{1}{2^{d-1}}\sum_{b,b'\in H_d:b'-b=e_i} \mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}})$$

Since $\cup_{i=1}^d \left\{(b,b') \in H_d^2 : b'-b = e_i\right\} = \left\{(b,b') \in H_d^2 : \ell_H(b,b') = 1\right\}$, we obtain that for each $i \in [d]$

$$\mathrm{TV}(P_{X|B_i=0}, P_{X|B_i=1}) \leqslant \max_{b,b'\in H_d:b'-b=e_i} \mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}}) \leqslant \max_{b,b'\in H_d:\ell_H(b,b')=1} \mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}}).$$

Hence, the result follows. □

Example 1.3 (d-dimensional GLM). Consider *i.i.d.* observation sample $X : \Omega \to \mathcal{X}^m$ with common distribution $\mathcal{N}(\theta, I_d)$ for $\theta \in \Theta \triangleq \mathbb{R}^d$. Considering the sufficient statistic $\bar{X} \triangleq \frac{1}{m}\sum_{i=1}^m X_i$, the model is simply $\left\{\mathcal{N}(\theta, \frac{1}{m}I_d) : \theta \in \mathbb{R}\right\}$. We observe that $\sqrt{m}(\bar{X} - \theta_0) \sim \mathcal{N}(\sqrt{m}(\theta - \theta_0), I_d)$. Let $\theta \triangleq \sqrt{m}(\theta_1 - \theta_0)$. From the shift and scale invariance of the total variation distance and rotational invariance of isotropic Gaussians, we have

$$\mathrm{TV}(\mathcal{N}(\theta_0, \frac{1}{m}I_d), \mathcal{N}(\theta_1, \frac{1}{m}I_d) = \mathrm{TV}(\mathcal{N}(0, I_d), \mathcal{N}(\theta, I_d)) = \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(\|\theta\|_2, 1)).$$

Consider the discrete parameter $\theta_b \triangleq \epsilon b \in \Theta$, for each binary string b in hypercube H_d and $\epsilon > 0$. For the quadratic loss $\ell : \Theta \times \Theta \to \mathbb{R}_+$ and $b, b' \in H_d$, we have

$$\ell(\theta_b, \theta_{b'}) \triangleq \|\theta_b - \theta_{b'}\|_2^2 = \epsilon^2 \left\|b - b'\right\|_2^2 = \epsilon^2 \sum_{i=1}^d (b_i - b_i')^2 = \epsilon^2 \sum_{i=1}^d \mathbb{1}_{\{b_i \neq b_i'\}} = \epsilon^2 \ell_H(b, b').$$

Applying Theorem 1.2 with $\beta = \epsilon^2$, using the fact that loss function ℓ is a 2-metric on Θ, observing $\left\|\sqrt{m}(\theta_b - \theta_{b'})\right\|_2 = \epsilon\sqrt{m}\sqrt{\ell_H(b, b')}$, defining $s \triangleq \epsilon\sqrt{m}$, from the invariance of total variation distance under scaling and shifting, and rotational invariance of total variation for Gaussian distribution, we get

$$R^* \geqslant \frac{\epsilon^2 d}{8}\left(1 - \max_{b,b' \in H_d : \ell_H(b,b') = 1} \mathrm{TV}\Big(\mathcal{N}(\epsilon b, \frac{1}{m}I_d), \mathcal{N}(\epsilon b', \frac{1}{m}I_d)\Big)\right) = \frac{s^2 d}{8m}\Big(1 - \mathrm{TV}\Big(\mathcal{N}(0,1), \mathcal{N}(s,1)\Big)\Big).$$

Recall that $1 - \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(s,1)) = Q(\frac{s}{2})$ and $\sup_{s>0} \frac{1}{2}s^2 Q(\frac{s}{2}) = c$ for some absolute constant $c \approx 0.083$. Therefore, we have $R^* \geqslant \frac{cd}{4m}$.

Next, let's consider the loss function $\ell(\theta, \theta') \triangleq \|\theta - \theta'\|_\infty$ for all $\theta, \theta' \in \Theta$. In the same setup as before where $\theta_b = \epsilon b$ for each $b \in H_d$ and some $\epsilon > 0$. Then, we have

$$\|\theta_b - \theta_{b'}\|_\infty = \epsilon \left\|b - b'\right\| = \epsilon \sup_{i \in [d]} \left|b_i - b_i'\right| = \epsilon \sup_{i \in [d]} \mathbb{1}_{\{b_i \neq b_i'\}} \geqslant \frac{\epsilon}{d} \sum_{i=1}^d \mathbb{1}_{\{b_i \neq b_i'\}} = \frac{\epsilon}{d} \ell_H(b, b').$$

Applying Theorem 1.2 with $\beta = \frac{\epsilon}{d}$, using the fact that loss function ℓ is a 1-metric on Θ, observing $\left\|\sqrt{m}(\theta_b - \theta_{b'})\right\|_2 = \epsilon\sqrt{m}\sqrt{\ell_H(b, b')}$, and defining $s \triangleq \epsilon\sqrt{m}$, we get

$$R^* \geqslant \sup_{s>0} \frac{s}{4\sqrt{m}}(1 - \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(s,1)) = \frac{1}{2\sqrt{m}} \sup_{s>0} \frac{s}{2} Q\Big(\frac{s}{2}\Big) = \frac{c'}{2\sqrt{m}},$$

where $c' \triangleq \sup_{s>0} \frac{s}{2} Q(\frac{s}{2})$ is a universal constant. Then Assouad's lemma yields $R^* \geqslant \frac{c'}{2\sqrt{m}}$, which does not depend on dimension d. In fact, $R^* \asymp \sqrt{\frac{\ln d}{m}}$ as shown before. In the next section, we will discuss Fano's method which can resolve this deficiency.

2 Assouad's Lemma from the mutual information method

One can integrate the Assouad's idea into the mutual information method.

Definition 2.1 (Binary entropy). Consider a binary random variable $X : \Omega \to \mathcal{X} \triangleq \{0,1\}$ with probability mass function $(p, 1-p) \in \mathcal{M}(\mathcal{X})$ for any $p \in [0,1]$. Then binary entropy $h : [0,1] \to [0,1]$ is defined as $h(p) \triangleq H(X) = -p\ln p - (1-p)\ln(1-p)$ for all $p \in [0,1]$.

Remark 1. Recall that the binary entropy function h is concave with unique maximum of $\ln 2$ achieved at $p = .5$, and increasing in $p \in [0, .5]$. It follows that we can define the inverse map $h^{-1} : [0, \ln 2] \to [0, .5]$ an increasing function.

Definition 2.2. We define $f : [0,1] \to [0, .5]$ for each $t \in [0, 2\ln 2]$ as $f(t) \triangleq h^{-1}((1-t)\ln 2)$ where h^{-1} is the inverse of the restricted binary entropy function.

Theorem 2.3 (Assouad). *Consider a simple statistical decision theory setting with $\Theta = \hat{\Theta}$ where*
(a) the loss function $\ell : \Theta \times \Theta \to \mathbb{R}_+$ is an α-metric on parameter space Θ, and
(b) the parameter space Θ contains a subset $\Theta' \triangleq \{\theta_b \in \Theta : b \in H_d\}$ indexed by the hypercube $H_d \triangleq \{0,1\}^d$ such that $\ell(\theta_b, \theta_{b'}) \geqslant \beta \ell_H(b,b')$ for all $b, b' \in H_d$ and some $\beta > 0$.
Then, the minimax risk $R^*(\Theta) \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \ell(\theta, \hat{\theta})$ satisfies the following inequality in terms of f from Definition 2.2,

$$R^*(\Theta) \geqslant \frac{\beta d}{2\alpha} f\left(\max_{\ell_H(b,b')=1} \mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}}) \right). \tag{2}$$

Proof. Let $B : \Omega \to \{0,1\}^d$ be an *i.i.d.* Bernoulli random vector with common mean $\frac{1}{2}$. Using the same "hypercube embedding $B \to \theta_B$", we have the Markov chain $B \to \theta_B \to X \to \hat{B}$. From the independence of random vector B we have[1] for all $i \in [d]$,

$$I(B_i; X \mid B^{i-1}) = I(B_i; X, B^{i-1}) \leqslant I(B_i; X, B_{\setminus\{i\}}) = I(B_i; B_{\setminus\{i\}}) + (B_i; X \mid B_{\setminus\{i\}}) = I(B_i; X \mid B_{\setminus\{i\}}).$$

We note that the mutual information is expressed as the Jensen-Shannon divergence as $2I(B_i; X \mid B_{\setminus\{i\}}) = \mathrm{JS}(P_{X|B_i=0}, P_{X|B_i=1})$. From the upper bound on Jensen-Shannon divergence in (5), we obtain $I(B_i; X \mid B_{\setminus\{i\}}) \leqslant \mathrm{TV}(P_{X|B_i=0}, P_{X|B_i=1}) \ln 2$. This results, together with the application of the chain rule to mutual information $I(B;X)$, and convexity of f-divergences in both arguments, leads to the following upper bound

$$I(B;X) = \sum_{i=1}^d I(B_i; X \mid B^{i-1}) \leqslant \sum_{i=1}^d I(B_i; X \mid B_{\setminus\{i\}}) \leqslant d \ln 2 \max_{\ell_H(B,B')=1} \mathrm{TV}(P_{X|B}, P_{X|B'}). \tag{3}$$

From Corollary A.3, it follows that for any estimate $\hat{B}(X)$ and $\tau \in [0,1]$ such that $I(B;X) \leqslant d(1-\tau)\ln 2$, we have $\mathbb{E}\ell_H(\hat{B}, B) \geqslant d h^{-1}(\tau \ln 2)$. Substituting this fact in (3), we obtain from the mutual information method

$$\mathbb{E}\ell_H(B, \hat{B}(X)) \geqslant d h^{-1}((1 - \max_{\ell_H(B,B')=1} \mathrm{TV}(P_{X|B}, P_{X|B'}))\ln 2) = d f\Big(\max_{\ell_H(B,B')=1} \mathrm{TV}(P_{X|B}, P_{X|B'}) \Big).$$

Following the same steps as in the proof of Theorem 1.2, we obtain the result

$$\mathbb{E}\ell(\hat{\theta}(X), \theta_B) \geqslant \frac{\beta}{2\alpha} \mathbb{E}\ell_H(\hat{B}(X), B) \geqslant \frac{\beta d}{2\alpha} f\left(\max_{\ell_H(b,b')=1} \mathrm{TV}(P_{\theta_b}, P_{\theta_{b'}}) \right).$$

□

Remark 2. Note that (2) is slightly weaker than (1). Nevertheless, as seen in Example 31.4, Assouad's lemma is typically applied when the pairwise total variation is bounded away from one by a constant, in which case (2) and (1) differ by only a constant factor. In all, we may summarize Assouad's lemma as a convenient method for bounding $I(B;X)$ away from the full entropy (d bits) on the basis of distances between $P_{X|B}$ corresponding to adjacent b's.

A Evaluation of rate-distortion function

Definition A.1 (Rate distortion function). Recall that rate-distortion function $R : \mathbb{R}_+ \to \mathbb{R}$ for a loss function $\ell : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$ is defined as $R(D) \triangleq \inf_{P_{\hat{X}|X} : \mathbb{E}\ell(X,\hat{X}) \leqslant D} I(X;\hat{X})$.

A.1 Bernoulli source

Consider an *i.i.d.* observation $X : \Omega \to \mathcal{X}^m$ with common mean $\mathbb{E}X_1 = p$ and its estimate $\hat{X} : \Omega \to (\hat{\mathcal{X}}^m)^{\mathcal{X}^m}$ for alphabets $\mathcal{X} = \hat{\mathcal{X}} \triangleq \{0,1\}$, with Hamming loss $\ell_H(X, \hat{X}) \triangleq \sum_{i=1}^m \mathbb{1}_{\{X \neq \hat{X}\}}$. We define the bit-error rate or fraction of erroneously decoded bits as $\ell(X, \hat{X}) \triangleq \frac{1}{m} \ell_H(X, \hat{X})$. By symmetry, we assume that $p \leqslant \frac{1}{2}$.

Theorem A.2. *Let $h : [0,1] \to [0, \ln 2]$ be the binary entropy function defined in Definition 2.1, then the rate-distortion function defined in Definition A.1 for a random variable $X : \Omega \to \{0,1\}$ with mean $\mathbb{E}X = p$ is*

$$R(D) \triangleq (h(p) - h(D))_+.$$

[1]Equivalently, this also follows from the convexity of the mutual information in the channel (cf. Theorem 5.3).

Proof. Consider an estimate $\hat{X} \triangleq 0$ independent of X. Then, distortion $\mathbb{E}\ell(X, \hat{X}) = P\{X = 1\} = p$, and it follows that $D_{\max} = p$. Hence, we can assume $D \leqslant p$ for otherwise there is nothing to show.

(a) For the converse, consider any $P_{\hat{X}|X}$ such that $\mathbb{E}\ell(X, \hat{X}) = P\{X \neq \hat{X}\} \leqslant D \leqslant p \leqslant \frac{1}{2}$. It follows that $H(X) = h(p)$ and since h is increasing for $p \leqslant 1/2$, we have $h(P\{X \neq \hat{X}\}) \leqslant h(D)$. Then from the fact that conditioning reduces entropy, we get

$$I(X; \hat{X}) = H(X) - H(X \mid \hat{X}) = H(X) - H(X \oplus \hat{X} \mid \hat{X}) \geqslant H(X) - H(X \oplus \hat{X}) \geqslant h(p) - h(D).$$

(b) In order to achieve this bound, we need to saturate the above chain of inequalities, in particular, choose $P_{\hat{X}|X}$ so that the difference $X \oplus \hat{X}$ is independent of $\hat{X}$. Let $X = \hat{X} \oplus Z$, where $\hat{X} \sim \mathrm{Ber}(p')$ and is independent of $Z \sim \mathrm{Ber}(D)$, and p' is such that the convolution gives exactly $\mathrm{Ber}(p)$, namely, $p' * D \triangleq p'(1-D) + (1-p')D = p$, i.e., $p' = \frac{p-D}{1-2D}$. In other words, the backward channel $P_{X|\hat{X}}$ is exactly $\mathrm{BSC}(D)$ and the resulting $P_{\hat{X}|X}$ is our choice of the forward channel $P_{\hat{X}|X}$. For this forward channel, we have $\mathbb{E}\ell(X, \hat{X}) = P\{X \neq \hat{X}\} = P\{Z = 1\} = D$. Then,

$$R(D) \leqslant I(X; \hat{X}) = H(X) - H(X \mid \hat{X}) = H(X) - H(Z) = h(p) - h(D).$$

□

Corollary A.3. *Consider an* i.i.d. *Bernoulli random vector $B : \Omega \to H_d$ with common mean $\mathbb{E}B_1 = \frac{1}{2}$, a finite set of parameters $\{\theta_b : b \in H_d\} \subset \Theta \triangleq \mathbb{R}^d$, observation $X : \Omega \to \mathcal{X}$ under statistical model $\mathcal{P}(\Theta)$, and loss function $\ell : H_d \times H_d \to [0,1]$ defined as $\ell(B, \hat{B}(X)) \triangleq \frac{1}{d}\ell_H(B, \hat{B})$ for any estimate $\hat{B} : \Omega \to \mathcal{X}^{\mathcal{X}}$. Let $h : [0, \frac{1}{2}] \to [0, \ln 2]$ be the binary entropy function defined in Definition 2.1 for all $p \in [0, \frac{1}{2}]$. If $I(B; X) \leqslant d(1-\tau)\ln 2$ for some $\tau \in [0,1]$, then for any estimator $\hat{B}(X)$, we have*

$$\mathbb{E}\ell(B, \hat{B}) \triangleq \frac{1}{d}\mathbb{E}\ell_H(\hat{B}, B) \geqslant \tau' \triangleq h^{-1}(\tau \ln 2). \tag{4}$$

Proof. We observe that $B \to \theta_B \to X \to \hat{B}$ is a Markov chain. From the rate-distortion function of the Bernoulli source in Section A.1, we know that $R(D) = d(\ln 2 - h(D))$ for $p = \frac{1}{2}$. Recall that $D_{\max} \leqslant p = \frac{1}{2}$ and h is increasing in $[0, \frac{1}{2}]$. It follows that $R^{-1}(y) = h^{-1}(\ln 2 - \frac{y}{d})$ for $y \in [0, d\ln 2]$, and hence $R^{-1}(d(1-\tau)\ln 2) = h^{-1}(\tau \ln 2) = \tau'$ for $\tau \in [0,1]$. From the definition of rate distortion function, data processing inequality for mutual information, and the monotonic decrease of rate distortion function, we obtain

$$\mathbb{E}\ell(B, \hat{B}) = \frac{1}{d}\mathbb{E}\ell_H(B, \hat{B}) \geqslant R^{-1}(I(B; \hat{B})) \geqslant R^{-1}(I(B; X)) \geqslant R^{-1}(d(1-\tau)\ln 2) = \tau'.$$

□

Remark 3. Here is a more general strategy also implemented in the Gaussian case. Denote the optimal forward channel from the achievability proof by $P^*_{\hat{X}|X}$ and the associated backward channel by $P^*_{X|\hat{X}}$ which is BSC(D). We need to show that there is no better $P_{\hat{X}|X}$ with $P\{X \neq \hat{X}\} \leqslant D$ and a smaller mutual information. From the fact that $P\{X \neq \hat{X}\} \leqslant D \leqslant \frac{1}{2}$ and monotonicity of h in $[0, \frac{1}{2}]$, we obtain

$$\begin{aligned} I(P_X, P_{\hat{X}|X}) = D(P_{X|\hat{X}} \| P_X \mid P_{\hat{X}}) &= D(P_{X|\hat{X}} \| P^*_{X|\hat{X}} \mid P_{\hat{X}}) + \mathbb{E}_P \ln \frac{P^*_{X|\hat{X}}}{P_X} \\ &\geqslant H(X) + \mathbb{E}_P[\ln D \mathbb{1}_{\{X \neq \hat{X}\}} + \ln(1-D)\mathbb{1}_{\{X = \hat{X}\}}] \geqslant h(p) - h(D). \end{aligned}$$

Example A.4. For example, when $p = \frac{1}{2}, D = .11$, we have $R(D) \approx \frac{1}{2}$ bits. In the Hamming game described in Section 24.2 where we aim to compress 100 bits down to 50, we indeed can do this while achieving 11% average distortion, compared to the naive scheme of storing half the string and guessing on the other half, which achieves 25% average distortion. Note that we can also get very tight non-asymptotic bounds, cf. Exercise V.3.

Remark 4. By WLLN, the distribution $P_X \triangleq \mathrm{Ber}(p)^{\otimes m}$ concentrates near the Hamming sphere of radius mp as m grows large. Recall that in proving Shannon's rate distortion theorem, the optimal codebook are drawn independently from $P_{\hat{X}} \triangleq \mathrm{Ber}(p')^{\otimes m}$ with $p' = \frac{p-D}{1-2D}$. Note that $p' = \frac{1}{2}$ if $p = \frac{1}{2}$ but $p' < p$ if $p < \frac{1}{2}$. In the latter case, the reconstruction points concentrate on a smaller sphere of radius mp' and *none* of them are typical source realizations, as illustrated in Figure 26.1.

B Jensen-Shannon Divergence

Definition B.1 (Jensen-Shannon divergence). Jensen-Shannon divergence $\mathrm{JS} : \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}_+$ is an f divergence $D_f(P\|Q) \triangleq \mathbb{E}_{X\sim Q} f(\frac{dP}{dQ}(X))$ for $P, Q \in \mathcal{M}(\mathcal{X})$ and a convex function $f : \mathcal{X} \to \mathbb{R}_+$, defined for $x \in \mathbb{R}_+$ as

$$f(x) \triangleq x \ln \frac{2x}{x+1} + \ln \frac{2}{x+1}.$$

Remark 5. Let $P, Q \in \mathcal{M}(\mathcal{X})$ and consider a uniform random variable $M : \Omega \to \{0,1\}$ and channel $P_{X|M} = \bar{M}P + MQ \in \mathcal{M}(\mathcal{X})$ for each random M. We observe that $P_X = \frac{1}{2}(P+Q) \in \mathcal{M}(\mathcal{X})$ and

$$I(M;X) = \mathbb{E}[\mathbb{E}[\ln \frac{dP_{X|M}}{dP_X} \mid M]] = \mathbb{E}\Big[\bar{M}\mathbb{E}_{X\sim P} \ln \frac{2\frac{dP}{dQ}(X)}{\frac{dP}{dQ}(X)+1} + M\mathbb{E}_{X\sim Q} \ln \frac{2}{\frac{dP}{dQ}(X)+1}\Big] = \frac{1}{2}\mathrm{JS}(P,Q).$$

Exercise B.2. For the Jensen-Shannon divergence $\mathrm{JS} : \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}_+$, show the following.
(a) For all $P, Q \in \mathcal{M}(\mathcal{X})$, we have

$$\mathrm{JS}(P,Q) \triangleq D(P\|\frac{1}{2}(P+Q)) + D(Q\|\frac{1}{2}(P+Q)).$$

(b) Show that $\sqrt{\mathrm{JS}}$ is a metric on the space of probability distributions $\mathcal{M}(\mathcal{X})$.

Exercise B.3. If $D_f(P\|Q)$ is an f-divergence, then show that $D_f(\lambda P + \bar{\lambda}Q\|Q)$ and $D_f(P\|\lambda P + \bar{\lambda}Q)$ are f-divergences for all $\lambda \in [0,1]$. In particular, $D_f(Q\|P) = D_{\tilde{f}}(P\|Q)$ with $\tilde{f}(x) \triangleq x f(\frac{1}{x})$.

Lemma B.4 (JS vs TV divergence). *The full joint region is given by*

$$2d\Big(\frac{1}{2}(1-\mathrm{TV}(P,Q))\|\frac{1}{2}\Big) \leqslant \mathrm{JS}(P,Q) \leqslant \mathrm{TV}(P,Q)2\ln 2. \quad (5)$$

Proof. Consider a uniform random variable $M : \Omega \to \{0,1\}$ and a channel $P_{X|M} = \bar{M}P + MQ \in \mathcal{M}(\mathcal{X})$.

(a) The lower bound is a consequence of Fano's inequality. Consider a random estimator $\hat{M}(X)$ such that $M \to X \to \hat{M}$ is a Markov chain. Consider two joint distributions $P_{M,X,\hat{M}} = P_M P_{X|M} P_{\hat{M}|X}$ and $R_{M,X,\hat{M}} = P_M P_X P_{\hat{M}|X}$. Under the joint distribution R, the random variables $M, \hat{M}$ are independent and uniform, and hence $R\{M = \hat{M}(X)\} = R\{M = \hat{M} = 0\} + R\{M = \hat{M} = 1\} = \frac{1}{2}$. Further, we recall that $P_e = P\{\hat{M}(X) \neq M\} = \frac{1}{2}(1-\mathrm{TV}(P,Q)) < \frac{1}{2}$. Therefore, we can write

$$I(M;\hat{M}) = D(P_{M,X,\hat{M}}\|R_{M,X,\hat{M}}) \geqslant d(P\{M=\hat{M}\}\|R\{M=\hat{M}\}) = d\Big(P_e\|\frac{1}{2}\Big).$$

The result follows from the fact that $\mathrm{JS}(P,Q) = 2I(M;\hat{M})$ and the monotonicity of binary relative entropy d in the first argument for $[0,\frac{1}{2}]$

(b) For the upper bound, we notice that $\mathrm{JS}(P,Q) = 2\ln 2 - \mathbb{E}_P \ln(1+\frac{dQ}{dP}(X)) - \mathbb{E}_Q \ln(1+\frac{dP}{dQ}(X))$. We will show this for the case when $\mathcal{X} \triangleq \{0,1\}$ and $P \triangleq (1-p,p)$ and $Q \triangleq (1-q,q)$ for some $p,q \in [0,1]$. Let $\tau \triangleq |p-q| \in [0,1]$, then we have $\mathrm{TV}(P,Q) = \tau$ and $\mathrm{JS}(P,Q) = d(p\|\frac{p+q}{2}) + d(q\|\frac{p+q}{2})$. From symmetry of $\mathrm{JS}(P,Q)$, we can take $q = p+\tau$ without any loss of generality, and hence $\mathrm{JS}(P,Q) = f(p,\tau) \triangleq d(p\|p+\frac{\tau}{2}) + d(p+\tau\|p+\frac{\tau}{2})$. We define $f(\tau) \triangleq \sup_{p\in[0,1-\tau]} f(p,\tau)$, and observe that $f(p,0) = 0$ for all $p \in [0,1]$ and $f(0,1) = d(0\|\frac{1}{2}) + d(1\|\frac{1}{2}) = 2\ln 2$. Therefore, we have $f(0) = 0$ and $f(1) = 2\ln 2$, and it follows from the convexity of d that $f(\tau) \leqslant 2\tau\ln 2$.

□

Lemma B.5. *Consider an* i.i.d. *random vector $B : \Omega \to H_d$ with common mean $\mathbb{E}B_1 = \frac{1}{2}$, embedding $b \mapsto \theta_b \in \Theta \triangleq \mathbb{R}^d$, and Markov chain $B \to \theta_B \to X$. Then, for each $i \in [d]$*

$$I(B_i; X \mid B_{\setminus\{i\}}) \leqslant \mathrm{TV}(P_{X|B_i=0}, P_{X|B_i=1}) \ln 2.$$

Proof. The result follows from (5) by noting that the mutual information is expressed as the Jensen-Shannon divergence as $2I(B_i; X \mid B_{\setminus\{i\}}) = \mathrm{JS}(P_{X|B_i=0}, P_{X|B_i=1})$. □