Lecture-25: Assouad's lemma

1 Assouad's Lemma

From Example 31.3 we see 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.

Theorem 1.1 (Assouad's lemma). *Assume that the loss function L satisfies the α-triangle inequality. Suppose Θ contains a subset $\Theta' \triangleq \left\{\theta_b : b \in \{0,1\}^d\right\}$ indexed by the hypercube, such that $L(\theta_b, \theta_{b'}) \geqslant \beta d_H(b,b')$ for all b, b' and some $\beta > 0$. Then*

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

Proof. We lower bound the Bayes risk with respect to the uniform prior over Θ'. Given any estimator $\hat{\theta}(X)$, define $\hat{b} \in \arg\min L(\hat{\theta}, \theta_b)$. Then for any $b \in \{0,1\}^d$,

$$\beta d_H(\hat{b}, b) \leqslant L(\theta_{\hat{b}}, \theta_b) \leqslant \alpha(L(\theta_{\hat{b}}, \hat{\theta}) + L(\hat{\theta}, \theta_b)) \leqslant 2\alpha L(\hat{\theta}, \theta_b).$$

Let $b : \Omega \to \{0,1\}^d$ be a discrete uniform random variable, and we have $b \to \theta_b \to X$. Then

$$\mathbb{E}L(\hat{\theta}, \theta_b) \geqslant \frac{\beta}{2\alpha}\mathbb{E}d_H(\hat{b}, b) = \frac{\beta}{2\alpha}\sum_{i=1}^d P\left\{\hat{b}_i \neq b_i\right\} \geqslant \frac{\beta}{4\alpha}\sum_{i=1}^d (1 - \mathrm{TV}(P_{X|b_i=0}, P_{X|b_i=1})),$$

where the last step is again by Theorem 7.7, just like in the proof of Theorem 31.1. Each total variation can be upper bounded as

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

where the equality follows from the Bayes rule, and the inequality follows from the convexity of total variation (Theorem 7.5). This completes the proof. □

Example 1.2 (d-dimensional GLM). Consider the quadratic loss first. To apply Theorem 1.1, consider the hypercube $\theta_b = \epsilon b$, where $b \in \{0,1\}^d$. Then $\|\theta_b - \theta_{b'}\|_2^2 = \epsilon^2 d_H H(b,b')$. Applying Theorem 1.1 yields

$$R^* \geqslant \frac{\epsilon^2 d}{4}\left(1 - \max_{b,b' \in \{0,1\}^d : d_H(b,b')=1} \mathrm{TV}\Big(\mathcal{N}(\epsilon b, \frac{1}{m} I_d), \mathcal{N}(\epsilon b', \frac{1}{m} I_d)\Big)\right) = \frac{\epsilon^2 d}{4}\left(1 - \mathrm{TV}\Big(\mathcal{N}(0, \frac{1}{m}), \mathcal{N}(\epsilon, \frac{1}{m})\Big)\right),$$

where the last step applies (7.11) for f-divergence between product distributions that only differ in one coordinate. Setting $\epsilon = \frac{1}{\sqrt{m}}$ and by the scale-invariance of TV, we get the desired $R^* \geqslant\approx \frac{d}{m}$. Next, let's consider the loss function $\|\theta_b - \theta_{b'}\|_\infty$. In the same setup, we only have $\|\theta_b - \theta_{b'}\|_\infty \geqslant \epsilon d_H(b,b')$. Then Assouad's lemma yields $R^* \geqslant\approx \frac{1}{\sqrt{m}}$, which does not depend on dimension d. In fact, $R^* \asymp \sqrt{\frac{\ln d}{m}}$ as shown in Corollary 28.8. 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. Consider the Bayesian setting of Theorem 1.1, where $B : \Omega \to \{0,1\}^d$ is *i.i.d.* Bernoulli random vector with mean $\frac{1}{2}$. From the rate-distortion function of the Bernoulli source in Section 26.1.1, we know that for any $\hat{B}$ and $\tau > 0$ there is some $\tau' > 0$ such that $I(B;X) \leqslant d(1-\tau)\ln 2$ which implies that

$$\mathbb{E}d_H(\hat{B}, B) \geqslant d\tau'. \quad (31.6)$$

Here τ' is related to τ by $\tau \ln 2 = h(\tau')$. Thus, using the same "hypercube embedding $B \to \theta_B$", the bound similar to (1) will follow once we can bound $I(B;X)$ away from $d\ln 2$. Can we use the pairwise total variation bound in (1) to do that? Yes! Notice that thanks to the independence of b_i's we have[1]

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

Applying the chain rule leads to the upper bound

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

where in the last step we used the fact that whenever $B_i \sim \mathrm{Ber}(1/2)$,

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

which follows from (7.39) by noting that the mutual information is expressed as the Jensen- Shannon divergence as $2I(B;X) = \mathrm{JS}(P_{X|B_i=0}, P_{X|B_i=1})$. Combining (31.6) and (31.7), the mutual information method implies the following version of the Assouad's lemma. Under the assumption of Theorem 31.2 and defining $f(t) \triangleq h^{-1}\left(\frac{(1-t)}{2}\ln 2\right)$ for $h^{-1} : [0, \ln 2] \to [0, \frac{1}{2}]$ being the inverse of the binary entropy function, we get

$$R^*(\Theta) \geqslant \frac{\beta}{4\alpha} f\left(\max_{d_H(\theta,\theta')=1} \mathrm{TV}(P_\theta, P_{\theta'})\right).$$

Note that (31.9) is slightly weaker than (31.5). 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 (31.9) and (31.5) 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

Recall that rate-distortion function $R : \mathbb{R}_+ \to \mathbb{R}$ is defined as

$$R(D) \triangleq \inf_{P_{\hat{X}|X} : \mathbb{E}d(\hat{X},X) \leqslant D} I(X;\hat{X}).$$

A.1 Bernoulli Source

Consider an *i.i.d.* random vector $X : \Omega \to \mathcal{X}^m$ with common mean $\mathbb{E}X_1 = p$ and its estimate $\hat{X} : \mathcal{X}^m \to \hat{\mathcal{X}}^m$ for alphabets $\mathcal{X} = \hat{\mathcal{X}} \triangleq \{0,1\}$, with Hamming distortion $d_H(X,\hat{X}) = \sum_{i=1}^m \mathbb{1}_{\{X \neq \hat{X}\}}$. Then $d(X,\hat{X}) = \frac{1}{m} d_H(X,\hat{X})$ is the bit-error rate (fraction of erroneously decoded bits). By symmetry, we may assume that $p \leqslant \frac{1}{2}$.

Theorem A.1. *Let $h : [0,1] \to \mathbb{R}_+$ be binary entropy function defined for each $p \in [0,1$ as $h(p) \triangleq -p\ln p - \bar{p}\ln\bar{p}$, then rate-distortion function 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. Since $D_{\max} = p$, in the sequel we can assume $D < p$ for otherwise there is nothing to show. For the converse, consider any $P_{\hat{X}|X}$ such that $d(X, \hat{X}) = P\{X \neq \hat{X}\} \leqslant D \leqslant p \leqslant \frac{1}{2}$. Then

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X \mid \hat{X}) = H(X) - H(X + \hat{X} \mid \hat{X}) \\ &\geqslant H(X) - H(X + \hat{X}) = h(p) - h(P\{X \neq \hat{X}\}) \geqslant h(p) - h(D). \end{aligned}$$

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 $X + \hat{X}$ is independent of $\hat{X}$. Let $X = \hat{X} + Z$, 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,

$$p' * D \triangleq p'(1-D) + (1-p')D = p, \text{ i.e., } p' = \frac{p-D}{1-2D}.$$

In other words, the backward channel $P_{X|\hat{X}}$ is exactly BSC(D) and the resulting $P_{\hat{X}|X}$ is our choice of the forward channel $\mathrm{P}_{\hat{X}|X}$. Then,

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

yielding the upper bound $R(D) \leqslant h(p) - h(D)$. □

Remark 1. Here is a more general strategy (which we will later implement 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. Then

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

where the last inequality uses $P\{X \neq \hat{X}\} \leqslant D \leqslant \frac{1}{2}$.

Example A.2. 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 2. 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.