8: Information Theoretic Lower Bounds

[A] Lower bounds for compression and Shannon's Source Coding Theorem

We saw that quantity $L_\varepsilon(X) = \min\{\lceil \log |S| \rceil : S \subseteq \mathcal{X}, P(S) \geq 1-\varepsilon\}$ is fundamental to the compression problem. We saw that

$$L_\varepsilon(X) \leq \min\{\lambda : P(\{x : -\log p(x) \leq \lambda\}) \geq 1-\varepsilon\}.$$

Using Chebyshev's inequality, we get

$$L_\varepsilon(X) \leq H(X) + \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\varepsilon}}$$

We now use Fano's inequality to show that this upper bound is almost tight. Consider a mapping $f : \mathcal{X} \to \{0,1\}^\ell$ and $\phi : \{0,1\}^\ell \to \mathcal{X}$ s.t.

$$P(X = \phi(f(X))) \geq 1-\varepsilon.$$

Choosing $S = \{x : x = \phi(f(x))\}$, $P(S) \geq 1-\varepsilon$. Further, since $f(x)$ can take at most 2^ℓ values, $|S| \leq 2^\ell$. Thus, $L_\varepsilon(X) \leq \ell$.

Also, given a set $S \subseteq \mathcal{X}$ s.t. $P(S) \geq 1-\varepsilon$, let

$$f(x) = \begin{cases} 1, & \text{if } x \notin S \\ i, & \text{if } x = x_i \in S, \quad 1 \leq i \leq |S|. \end{cases}$$

Thus, $f : \mathcal{X} \to \{0,1\}^{\lceil \log |S| \rceil}$, which gives

$\tilde{L}_\varepsilon(X) := \min\{\ell :$ there exists $f : \mathcal{X} \to \{0,1\}^\ell$ and $\phi : \{0,1\}^\ell \to \mathcal{X}$, $P(X = \phi(f(X))) \geq 1-\varepsilon\}$

$= L_\varepsilon(X)$.

We derive a lower bound for $\tilde{L}_\varepsilon(X)$.

For $f: \mathcal{X} \to \{0,1\}^\ell$, $\phi: \{0,1\}^\ell \to \mathcal{X}$, let $M = f(X)$ and $\hat{X} = \phi(M)$. Then, $P(X = \hat{X}) \geq 1-\varepsilon$.

$$\begin{aligned} \ell \geq H(M) \geq H(\hat{X}) &= H(X,\hat{X}) - H(X|\hat{X}) \\ &\geq H(X) - H(X|\hat{X}) \\ &\geq H(X) - \varepsilon \log(|\mathcal{X}|-1) - h(\varepsilon), \end{aligned}$$

where we used Fano's inequality in the last step.

Thus, $L_\varepsilon(X) \geq H(X) - \varepsilon \log(|\mathcal{X}|-1) - h(\varepsilon)$, which shows $L_\varepsilon(X) \approx H(X)$.

Shannon considered an asymptotic version of this problem where the goal is to characterize the number of bits per symbol needed to store $X_1, \ldots, X_n \sim$ iid P. Specifically, he looked at the optimal rate R^* given by

$$R^*(P) = \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} L_\varepsilon(X^n).$$

→ Using our upper bound, $L_\varepsilon(X^n) \leq nH(X) + \sqrt{\frac{n}{\varepsilon} \mathrm{Var}(-\log P(X_i))}$, giving $\lim_{n\to\infty} \frac{1}{n} L_\varepsilon(X^n) \leq H(X)$.

→ Using our lower bound, $L_\varepsilon(X^n) \geq nH(X) - n\varepsilon \log|\mathcal{X}| - 1$, giving $\lim_{n\to\infty} \frac{1}{n} L_\varepsilon(X^n) \geq H(X) - \varepsilon$.

Taking the limit $\varepsilon \to 0$, we get:

Theorem (Shannon's Source Coding Theorem)

$R^*(P) = H(P)$.

[B] Lower bound for hypothesis testing and Stein's Lemma

Recall the quantity $\beta_\varepsilon(P,Q) := \min\{Q(S) : P(S) \geq 1-\varepsilon\}$ which is fundamental for binary hypothesis testing (Neyman-Pearson formulation). We saw that

$$-\log \beta_\varepsilon(P,Q) \geq \max\{\lambda : P(\{x : \log \tfrac{P(x)}{Q(x)} \geq \lambda\}) \geq 1-\varepsilon\}$$

which by Chebyshev's inequality gives,

$$-\log \beta_\varepsilon(P,Q) \geq D(P\|Q) - \sqrt{\frac{1}{\varepsilon} \operatorname{Var}_P\left(-\log \frac{P(x)}{Q(x)}\right)}.$$

We now show (using the data processing inequality) that this bound is almost tight. Consider a test $T: \mathcal{X} \to \{0,1\}$. We can even allow a randomized test which can be viewed as a channel. We can interpret the output 0 as P and 1 as Q. Then, $\sum_x P(x) T(0|x) \geq 1-\varepsilon$. Then, by DPI, $D(TP\|TQ) \leq D(P\|Q)$. Further,

$$D(TP\|TQ) \geq (TP)(0) \log \frac{1}{(TQ)(0)} - h((TP)(0))$$

$$\geq (1-\varepsilon) \log \frac{1}{(TQ)(0)} - 1.$$

Thus, $-\log (T_0 Q)(0) \le \frac{D(P\|Q)+1}{1-\varepsilon}$.

Note that $(T_0 Q)(0) = \sum_x Q(x)\, T(0|x)$ is the prob of missed detection and $(T_0 P)(1) = \sum_x P(x)\, T(1|x)$ is the prob. of false alarm.

Thus, $-\log \beta_\varepsilon(P, Q) \le \frac{D(P\|Q)+1}{1-\varepsilon}$.

That is, the best prob. of missed detection $\approx 2^{-D(P\|Q)}$.

Stein's Lemma

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} -\log \beta_\varepsilon(P^n \| Q^n) = D(P\|Q).$$

(i.e., $\beta_\varepsilon(P^n \| Q^n) \approx 2^{-n D(P\|Q)}$ for iid observations.

To make this small, we must have $n \gtrsim \frac{1}{D(P\|Q)}$).

C Lower bound for randomness generation

Consider the uniformity generation problem. Suppose a mapping $f: \mathcal{X} \to \{0,1\}^\ell$ satisfies

$$d\left(P_{f(X)}, \text{unif}(\{0,1\}^\ell)\right) \le \varepsilon,$$

where $X \sim P$, then by the continuity bound for entropy,

$$H(f(X)) \ge H(\text{unif}(\{0,1\}^\ell)) - \varepsilon \log |\mathcal{X}| - h(\varepsilon)$$

$$= \ell - \varepsilon \log |\mathcal{X}| - 1$$

$$\Rightarrow \quad \ell \le H(X) + \varepsilon \log |\mathcal{X}| + 1.$$

(no more $\approx H(X)$ random bits can be generated) ∎

D Strong Converse for Shannon's Source Coding Theorem

We saw earlier that

$$\lim_{n\to\infty} \frac{1}{n} L_\varepsilon(X^n) \le H(X)$$

and

$$\lim_{n\to\infty} \frac{1}{n} L_\varepsilon(X^n) \ge H(X) - \varepsilon \log|\mathcal{X}|.$$

In fact, it can be shown that

$$\lim_{n\to\infty} \frac{1}{n} L_\varepsilon(X^n) = H(X), \quad \text{for every } 0<\varepsilon<1.$$

Thus $\lim_{n\to\infty} \frac{1}{n} L_\varepsilon(X^n) = R^* \quad \forall \ 0<\varepsilon<1$. Such results are called strong converse results. We show this now.

Consider a λ s.t. $P(\{x : -\log p(x) \ge \lambda\}) \ge 1-\delta$

(we used $-\log p(x) \le \lambda$ for the upper bound)

Denoting $B_\lambda = \{x : -\log p(x) \ge \lambda\}$, for any $S \subseteq \mathcal{X}$ s.t. $P(S) \ge 1-\varepsilon$, we have $P(B \cap S) \ge P(B) + P(S) - 1 \ge 1-\varepsilon-\delta$.

Thus, $1-\varepsilon-\delta \le \sum_{x \in B\cap S} P(x) \le \sum_{x \in B\cap S} 2^{-\lambda} \le 2^{-\lambda}|B\cap S|$,

since $p(x) \le 2^{-\lambda}$ for every $x \in B$. We have,

$$\log|S| \ge \log|B\cap S| \ge \lambda - \log\frac{1}{1-\varepsilon-\delta}.$$

We have shown:

$$L_\varepsilon(X) \geq \max\left\{\lambda : P(\{x : -\log p(x) \geq \lambda\}) \geq \frac{1-\varepsilon}{2}\right\} - \log\frac{2}{1-\varepsilon}.$$

By using this bound with Chebyshev's inequality,

$$\frac{1}{n} L_\varepsilon(X^n) \geq H(X) - \sqrt{\frac{1}{n\varepsilon}\mathrm{Var}(-\log P(X))} - \frac{1}{n}\log\frac{2}{1-\varepsilon}.$$

$$\Rightarrow \lim_{n\to\infty} \frac{1}{n} L_\varepsilon(X^n) \geq H(X).$$

[E] Minmax lower bound for estimation

The final lower bound we present is very useful in statistics. We present it using an example. Consider the problem of estimating the mean of a Gaussian random variable. Let $X_1, \ldots, X_n \sim$ iid $N(\mu, I)$. The empirical mean $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$ satisfies

$$\mathbb{E}\left[\|\mu - \hat{\mu}\|_2^2\right] = \sum_{i=1}^{d} \mathbb{E}\left[(\mu_i - \hat{\mu}_i)^2\right] = \sum_{i=1}^{d} \frac{1}{n^2}\sum_{t=1}^{n} \mathrm{Var}(X_t) = \frac{d}{n}.$$

Thus, we found an estimator with MSE $\boxed{d/n}$.

Can we do better? Not in the worst-case.

For any estimator $\hat{\mu} : \mathcal{X}^n \to \mathbb{R}^d$, we have

$$\max_{\mu} \mathbb{E}_{P_\mu}\left[\|\mu - \hat{\mu}\|_2^2\right] \geq \max_{\mu \in \{\mu_1, \ldots, \mu_M\}} \mathbb{E}_{P_\mu}\left[\|\mu - \hat{\mu}\|_2^2\right]$$

$$\geq \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{P_{\mu_i}}\left[\|\mu_i - \hat{\mu}\|_2^2\right]$$

$$\geq \delta^2 \cdot \left(\frac{1}{M} \sum_{i=1}^{M} P_{\mu_i}\left(\|\mu_i - \hat{\mu}\|_2^2 > \delta^2 \right) \right)$$

(This is a different application of Markov's inequality)

This bound holds for any selection of $\mu_1, \ldots, \mu_M$. The expression on the right-side is probability of error for an M-ary hypothesis testing problem. We choose $\mu_1, \ldots, \mu_M$ to get a very difficult hypothesis testing problem.

Specifically, we choose $\mu_1, \ldots, \mu_M$ s.t.

(i) $\|\mu_i\|_2 \leq 4\delta$, $1 \leq i \leq M$;

(ii) $\|\mu_i - \mu_j\|_2 \geq 2\delta$, $i \neq j$.

We can use our estimator $\hat{\mu}$ to form a test for our M-ary hypothesis testing problem. Specifically, we find the unique i s.t. $\|\hat{\mu} - \mu_i\|_2 < \delta$. Note that by (ii) we can only find one such i. Let $J \sim \text{unif}\{1, \ldots, M\}$. We generate $X_1, \ldots, X_n \sim \text{iid } N(\mu_J, I)$ and find $\hat{J} = i$ s.t. $\|\hat{\mu} - \mu_i\|_2 < \delta$, we get

$$\frac{1}{M}\sum_{i=1}^{M} P_{\mu_i}\left(\|\mu_i - \hat{\mu}\|_2 > \delta\right) \geq P(J \neq \hat{J})$$

$$\geq 1 - \frac{I(J \wedge X^n) + 1}{\log M},$$

where the final bound is by Fano's inequality.

Note that $I(J \wedge X^n) \leq \max_{P_J} I(J \wedge X^n)$

(using the information-radius interpretation of capacity) $\leq \max_{1 \leq i \neq j \leq M} D(P_{\mu_i}^n \| P_{\mu_j}^n)$

$$= n \max_{1 \leq i \neq j \leq M} \frac{\|\mu_i - \mu_j\|_2^2}{2}$$

$$\leq 8 n \delta^2.$$

Thus, the overall bound gives

$$\max_{\mu} \mathbb{E}_{P_\mu}\left[\|\mu - \hat{\mu}\|_2^2\right] \geq \delta^2 \cdot \left(1 - \frac{8n\delta^2 + 1}{\log M}\right)$$

We want M to be as large as possible. We can simply choose these points in a greedy manner by selecting a point and removing the ball of radius 2δ around it. Note that $\text{vol}(B_d(r)) = c_d r^d$. Thus, we can find at least $M \geq \frac{c_d (4\delta)^d}{c_d (2\delta)^d}$ points, which gives $\log M \geq d \log 2 = d$. With this choice, we have

$$\max_{\mu} \mathbb{E}_{P_\mu}\left[\|\mu - \hat{\mu}\|_2^2\right] \geq \delta^2\left(1 - \frac{8n\delta^2}{d} - \frac{1}{d}\right) \qquad (9)$$

Choosing $\delta^2 = \frac{d}{16n}$, the right-side is bounded by

$$\geq \frac{d}{16n}\left(\frac{1}{2} - \frac{1}{d}\right) \geq \frac{d}{(6\times 16)n} \quad \text{if } d \geq 3.$$

$$\Rightarrow \boxed{\text{MSE} \gtrsim \frac{d}{n}}$$