Electrical Engineering 229A Lecture 27 Notes

Daniel Raban

December 2, 2021

1 I-Projection in Sanov's Theorem and Hypothesis Testing

1.1 Properties of I-projection in Sanov's theorem

Last time, we proved Sanov's theorem:

Theorem 1.1 (Sanov)**.** *Let $X_1, X_2, \dots \overset{\text{iid}}{\sim} Q$ be $\mathscr{X}$-valued random variables, and let P_{x^n} be the type of x^n: $P_{x^n}(x) = \frac{N(x|x^N)}{n}$. Let $\mathcal{P}$ be the set of probability distributions on $\mathcal{X}$, and assume that $E \subseteq \mathcal{P}$ is the closure of its interior. Then*

$$\lim_{n \to \infty} \frac{1}{n} \log Q^n(P_{X^n} \in E) = -D(P^* \,||\, Q),$$

where

$$P^* = \operatorname*{arg\,min}_{P \in E} D(P \,||\, Q).$$

P^* is called the ***I*-projection** of Q onto E.

Definition 1.1. Let $\mathscr{X}$ be finite. Given $Q \in \mathcal{P}$ and $h : \mathscr{X} \to \mathbb{R}$, the probability distribution of the form

$$\frac{Q(x)e^{\lambda h(x)}}{\sum_{a \in \mathscr{X}} Q(a)e^{\lambda h(a)}}$$

is called an **exponential transform** of Q.

Proposition 1.1. *Suppose E is defined as*

$$E = \left\{ P : \sum_x g_j(x)P(x) \geq \alpha_j, j = 1, \dots, k \right\}.$$

Then P^* will be an exponential transform of Q.

Proof. Assume $Q(x) > 0$ for all x. We want

$$\max \sum_x P(x) \log \frac{P(x)}{Q(x)},$$

subject to

$$\begin{cases} \sum_a P(a) g_j(a) \geq \alpha_j, & j = 1, \ldots, k \\ P(x) \geq 0 & x \in \mathscr{X} \\ \sum_x P(x) = 1. \end{cases}$$

where the variables are $(P(x), x \in \mathscr{X})$ and $Q \in \mathcal{P}$ is fixed. The correct Lagrangian is

$$\sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_{j=1}^k \lambda_j \left(\sum_x P(x) g_j(x) - \alpha_j \right) - \sum_x \mu_x P(x) + \nu \left(\sum_x P(x) - 1 \right).$$

Write the KKT conditions for this:

$$\lambda *_j \geq 0,$$

$$\mu_x^* \geq 0,$$

$$\lambda_j^* \left(\alpha_j - \sum_x P^*(x) g_j(x) \right) = 0 \qquad \forall j,$$

$$\mu_x^* P^*(x) = 0 \qquad \forall x.$$

Differentiate this to get

$$\log \frac{P^*(x)}{Q(x)} + 1 + \sum_j \lambda_j g_j(x) - \mu_x^* + \nu^* = 0 \qquad \forall x.$$

Since $P^*(x)$ cannot be 0 for any x, we must have $\mu_x^* = 0$. □

We also can show the following.

Theorem 1.2.

$$\lim_{n \to \infty} Q^n(X_1 = a \mid P_{X^n} \in E) = P^*(a) \qquad \forall a \in \mathscr{X}.$$

Proof. Given $\delta > 0$, let $A = \{P \in E : D(P \,||\, Q) \leq D(P^* \,||\, Q) + 2\delta\}$. The Sanov theorem calculation tells us that

$$Q^n(E \setminus A) \leq (n+1)^{|\mathscr{X}|} 2^{-n(D^*(P||Q) + 2\delta)}.$$

For large enough n,

$$Q^n(A) \geq \frac{1}{(n+1)^{|\mathscr{X}|}} 2^{-n(D^*(P||Q) + \delta)}.$$

This proves that

$$Q^n(P_{X^n} \in A \mid P_{X^n} \in E) \xrightarrow{n\to\infty} 1.$$

If E is convex, we can use $D(P \,||\, P^*) + D(P^* \,||\, Q) \le D(P \,||\, Q)$ for all $P \in E$ to show that

$$Q^n(D(P_{X^n} \,||\, P^*) \le 2\delta \mid P_{X^n} \in E) \xrightarrow{n\to\infty} 1.$$

Then use Pinsker's inequality:

$$D(P_1 \,||\, P_2) \ge \frac{1}{2\ln 2}\|P_1 - P_2\|_1^2 \qquad \forall P_1, P_2. \qquad \square$$

1.2 The Neyman-Pearson framework of hypothesis testing

Here is the Neyman-Pearson formulation of hypothesis testing with two hypotheses H_1 and H_2. Under H_1, assume that $X_1, X_2, \dots,$ are iid $\mathscr{X}$-valued with $X_i \sim P_1$. Under H_2, assume that $X_1, X_2, \dots,$ are iid $\mathscr{X}$-valued with $X_i \sim P_2$. Given a "threshold" T, define

$$A_n(T) = \left\{ x^n : \frac{P_1^n(x^n)}{P_2^n(x^n)} > T \right\}.$$

Definition 1.2. A **hypothesis test** is a function $\mathscr{X}^n \to \{1, 2\}$.

Equivalently, it means we choose a set $B \subseteq \mathscr{X}^n$ on which to decide H_1, and on B^c we decide H_2.

Let $\mathbb{1}_B$ denote the indicator function of B. Observe that

$$(\mathbb{1}_{A_n(T)}(x^n) - \mathbb{1}_B(x^n))(P_1^n(x^n) - TP_2^n(x^n)) \ge 0 \qquad \forall x^n.$$

Summing this up over x^n,

$$\underbrace{\underbrace{\sum_{x^n \in A_n(T)} P_1(x^n)}_{1-\mathbb{P}_1^n(X^n \notin A_n(T))}}_{\alpha^*} - T\underbrace{\sum_{x^n \in A_n(T)} P_2^n(x^n)}_{\beta^*} - \underbrace{\sum_{x^n \in B} P_1^n(x^n)}_{1-\alpha} + T\underbrace{\sum_{x^n \in B} P_2^n(x^n)}_{\beta} \ge 0.$$

We get

$$T(\beta - \beta^*) - (\alpha^* - \alpha) \ge 0,$$

so if $\alpha \le \alpha^*$, then $\beta \ge \beta^*$. Hence, if one tries to minimize $\mathbb{P}(\text{error} \mid H_2)$ given a bound on $\mathbb{P}(\text{error} \mid H_1)$, then we use a threshold test.

Theorem 1.3 (Stein's lemma)**.** *For any $\varepsilon > 0$, let*

$$\beta_n^\varepsilon := \min_{B \subseteq \mathscr{X}^n} \{\beta_n : \alpha_n \le \varepsilon\}.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n^\varepsilon = -D(P_1 \,||\, P_2).$$

The intuition is that for all $\delta > 0$, the ball $C_n = \{P \in \mathcal{P} : D(P \,||\, P_1) \leq \delta\}$ has $P_1^n(C_n) \to 1$ as $n \to \infty$ and

$$\liminf_{n\to\infty} -\frac{1}{n} \log P_2^n(C_n) \geq D(P_1 \,||\, P_2) - \eta,$$

where $\eta \to 0$ as $\delta \to 0$.

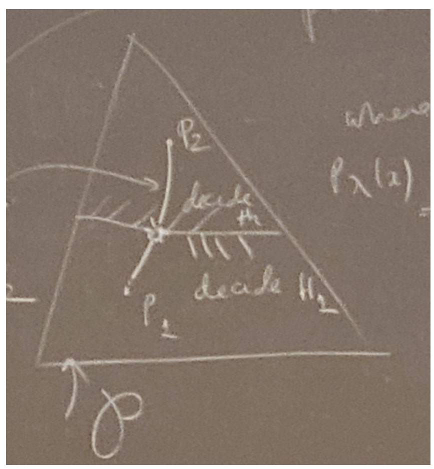
1.3 The Bayesian framework of hypothesis testing

In the Bayesian view of hypothesis testing, π_1 is the prior probability of H_1, and π_2 is the prior of H_2. The optimal test is to decide H_1 if

$$\frac{\pi_1 P_1^n(x^n)}{\pi_2 P_2^n(x^n)} \geq 1.$$

This is related to information geometry, which is about the space of probability distributions with separation defined by relative entropy. If $P_1, P_2 \in \mathcal{P}$, then there is a statistically natural path connecting them, parameterized by $\lambda \in [0,1]$, where

$$P_\lambda(x) = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_a P_1^\lambda(a) P_2^{1-\lambda}(a)}.$$



P_λ arises by studying the minimum of $D(P \,||\, P_2)$ subject to $D(P \,||\, P_2) - D(P \,||\, P_1) = K$. Why this constraint? This is because

$$\left\{x^n : \frac{P_1(x^n)}{P_2(x^n)} \geq T\right\} = \left\{x^n : D(P_{x^n} \,||\, P_2) - D(P_{x^n} \,||\, P_1) \geq \frac{1}{n} \log T\right\}.$$

Theorem 1.4. *Assume that* $\pi_1 > 0$ *and* $\pi_2 > 0$*. Let* $\alpha_n^* = P_1^n(A_n(\frac{\pi_2}{\pi_1})^c)$*, and let* $\beta_n^* = P_2^n(A_n(\frac{\pi_2}{\pi_1}))$*. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log(\pi_1\alpha_n^* + \pi_2\beta_n^*) \to -D(P_{\lambda^*} \mid\mid P_2),$$

where $D(P_{\lambda^*} \mid\mid P_2) = D(P_{\lambda^*} \mid\mid P_1)$*.*