Electrical Engineering 229A Lecture 5 Notes

Daniel Raban

September 9, 2021

1 Sufficient Statistics, Fano's Inequality, and the Asymptotic Equipartition Property

1.1 Sufficient statistics

Last time, we discussed the data processing inequality. Given $Y - X - Z$ (i.e. Y and Z are conditionally independent given X), the data processing inequality says that

$$I(X; Z) \geq I(Y; Z).$$

The equality condition is when $I(X; Z \mid Y) = 0$, i.e. $X - Y - Z$.

We also discussed sufficient statistics. The idea is to think about learning about Θ by processing some observations X into $T(X)$, so $\Theta - X - T(X)$. Then $\Theta - T(X) - X$ if and only if $I(\Theta; X) = I(\Theta; T(X))$. Given $\Theta - X - T(X)$, we say that $T(X)$ is a **sufficient statistic** (for learning about Y from X).

If $|\mathscr{X}| = d$, let $u(x) = \frac{1}{d}$ for $x \in \mathscr{X}$. Given $(p(x), x \in \mathscr{X})$, then

$$D(p \mid\mid u) = \sum_x p(x) \log \frac{p(x)}{u(x)} = \log d - H((p(x), x \in \mathscr{X})).$$

So it is difficult to define entropy in non-discrete settings. Regardless, here is a non-discrete example of a sufficient statistic.

Example 1.1. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$ when $\Theta = \theta$, where $\Theta \in \{\theta_1, \dots, \theta_d\}$ is a random variable. Then $\frac{1}{n} \sum_{i=1}^n X_i$ is a sufficient statistic for Θ. To check this, we need to show that $\Theta - \overline{X} - (X_1, \dots, X_n)$, where $\overline{X} := \frac{1}{n}(X_1, \dots, X_n)$. The conditional joint density is

$$\begin{aligned} f(x_1, \dots, x_n \mid \theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i - \theta)^2/2} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i - \overline{x} + \overline{x} - \theta)^2/2} \end{aligned}$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^n (x_i-\overline{x})^2} e^{-\frac{n}{2}(\overline{x}-\theta)^2} \underbrace{e^{-\sum_{i=1}^n (x_i-\overline{x})(\overline{x}-\theta)}}_{=1}.$$

Now

$$\begin{aligned} f(x_1,\ldots,x_n \mid \overline{x},\theta) &= \frac{f(x_1,\ldots,x_n,\theta,\overline{x})}{f(\theta,x)} \\ &= \frac{f(x_1,\ldots,x_n,\overline{x} \mid \theta)}{f(\overline{x} \mid \theta)} \end{aligned}$$

And $f(\overline{x} \mid \theta) = e^{-\frac{n}{2}(\overline{x}-\theta)^2}$ by integrating over $x_1,\ldots,x_n$, so

$$= f(x_1,\ldots,x_n \mid \overline{x}).$$

1.2 Fano's inequality

In the data processing inequality, we had $Y - X - \widehat{Y}$, where $\widehat{Y}$ is viewed as derived from X to learn about Y. Suppose Y and $\hat{Y}$ take values in the same set and our goal is to try to get small $\mathbb{P}(\widehat{Y} \neq Y)$. Fano's inequality gives us an lower bound on this probability using the conditional entropy of Y given X.

Theorem 1.1 (Fano's inequality)**.** *Suppose $Y - X - \widehat{Y}$, and let $p_e = \mathbb{P}(Y \neq \widehat{Y})$. Then*

$$H(Y \mid X) \leq H(Y \mid \widehat{Y}) \leq h(p_e) + p_e \log(|\mathscr{Y}| - 1),$$

where $h(p_e) = -p_e \log p_e - (1-p_e)\log(1-p_e)$ is the binary entropy function.

Proof. Because $I(X;Y) = H(Y) - H(Y \mid X)$ and $I(\widehat{Y};Y) = H(Y) - H(Y \mid \widehat{Y})$, the data processing inequality gives $H(Y \mid X) \leq H(Y \mid \widehat{Y})$. Now consider $H(Y, E \mid \widehat{Y})$, where $E = \mathbb{1}_{\{Y \neq \widehat{Y}\}}$ is a $\{0,1\}$-valued random variable. Write this as

$$H(Y, E \mid \widehat{Y}) = H(Y \mid \widehat{Y}) + \underbrace{H(E \mid Y, \widehat{Y})}_{=0}.$$

We can also write this as

$$\begin{aligned} H(Y, E \mid \widehat{Y}) &= H(E \mid \widehat{Y}) + H(Y \mid E, \widehat{Y}) \\ &\leq H(E) + (1-p_e) H(Y \mid E = 1, \widehat{Y}) \\ &\leq h(p_e) + (1-p_e)\log(|\mathscr{Y}| - 1). \end{aligned}$$

□

1.3 The asymptotic equipartition property

Given $(p(x), x \in \mathscr{X})$ with $\mathscr{X}$ finite, let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} p$. Then

$$p^n(x_1^n) = \prod_{i=1}^n p(x_i)$$

$$= \prod_{x \in \mathscr{X}} p(x)^{N(x|x_1^n)},$$

where $N(x \mid x_1^n) = \sum_{i=1}^n \mathbb{1}_{\{x_i = x\}}$ is the number of times x shows up in $x_1, \ldots, x_n$.

$$= 2^{\sum_{x \in \mathscr{X}} N(x|x_1^n) \log p(x)}.$$

The Strong Law of Large Numbers tells us that $\frac{1}{n} N(x \mid X_1^n) \to p(x)$ almost surely as $n \to \infty$. This suggests that for large n, the realizations that "matter" are those x_1^n for which each $N(x \mid x_1^n)$ is roughly $np(x)$. The asymptotic equipartition property formalizes this statement in a weak way via the weak law of large numbers. The "method of types" formalizes this more carefully.

The asymptotic equipartition property comes from applying the weak law of large numbers to the iid equence of entropy densiies, i.e. to the sequence $\log \frac{1}{p(x_1)}, \log \frac{1}{p(x_2)}, \ldots$.

Lemma 1.1 (Weak law of large numbers)**.** *For any real-valued iid sequence $Z_1, Z_2, \ldots$ with $\mathbb{E}[|Z_1|] < \infty$,*

$$\frac{1}{n} \sum_{i=1}^n Z_i \xrightarrow{p} \mathbb{E}[Z_1].$$

That is, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[Z_1] \right| \leq \varepsilon \right) \to 1$$

as $n \to \infty$.

Theorem 1.2 (Asymptotic equipartition property)**.** *For all $\varepsilon > 0$,*

$$\mathbb{P}\left(\left| \frac{1}{n} \log p(X_1, \ldots, X_n) - H(X) \right| \leq \varepsilon \right) \to 1$$

as $n \to \infty$.

This leads us to define the following.

Definition 1.1. The set of **ε-weakly typical sequences** is

$$A_\varepsilon^{(n)} := \left\{ x_1^n : \left| \tfrac{1}{n} \log p^n(x_1^n) - H(x) \right| \leq \varepsilon \right\}.$$

We can see

$$x_1^n \in A_\varepsilon^{(n)} \iff 2^{-nH(x)} 2^{-n\varepsilon} \leq p^n(x_1^n) \leq 2^{-nH(X)} 2^{n\varepsilon}.$$

Proposition 1.1.

$$|A_\varepsilon^{(n)}| \leq 2^{nH} 2^{n\varepsilon}.$$

Proof. We must have $\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \leq 1$. □

The AEP says that

$$\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \to 1$$

as $n \to \infty$. The left hand side is equal to

$$\sum_{x_1^n \in A_\varepsilon^{(n)}} p^n(x_1^n).$$

Hence, for all $\varepsilon \to 0$, if n is large enough (how large depending on δ), $\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \geq 1-\delta$. Hence,

$$|A_\varepsilon^{(n)}| \geq (1-\delta)2^{nH}2^{-n\varepsilon}$$

for all large enough n.