5: Information and Statistical Inference-2

[A] How much "information" is revealed by one coin toss?

Let us revisit our coin-bias estimation problem.

Suppose that we observe independent tosses of a coin with bias $\frac{1}{2}$ or $\frac{1}{2}+\varepsilon$. We assume a Bayesian setup with a uniform prior on two coins. In our language, $X \sim \text{unif}\{0,1\}$ and $Y|X \sim \text{iid Ber}(p_X)$ with $p_0 = \frac{1}{2}$ and $p_1 = \frac{1}{2}+\varepsilon$.

Initial uncertainty of X is $H(X) = 1$. How much is this uncertainty reduced when we toss the coin, namely

how much information is revealed per coin toss?

One answer to this question can be seen from our formula: $P_e^*(\text{unif}) = \frac{1}{2}\left(1 - d\left(W_0^n, W_1^n\right)\right)$

We can look at how fast $d(W_0^n, W_1^n)$ increases per sample. Specifically, how large is $d(W_0^n, W_1^n)/n$?

Earlier we saw that $\frac{d(W_0^n, W_1^n)}{n} \le d(W_0, W_1) = \varepsilon$.

But this bound is weak. Indeed, by Pinsker's ineq

$$\frac{d(W_0^n, W_1^n)}{n} \le \frac{1}{n}\sqrt{\frac{\ln 2}{2} D(W_0^n \| W_1^n)}$$

$$= \frac{1}{n}\sqrt{\frac{n \ln 2}{2} D(W_0 \| W_1)} \quad \text{(additivity of divergence)}$$

$$= \sqrt{\frac{\ln 2}{2n} \log \frac{1}{1-\varepsilon^2}} \approx \frac{\varepsilon}{\sqrt{n}} \quad \text{when } \varepsilon \text{ is small}$$

Thus, the "information" revealed per toss is much smaller than what was suggested by our earlier bound.

In particular, we need at least (roughly) $\frac{1}{\varepsilon^2}$ tosses to get $P_e^*(\text{unif}) \le 1/3$.

→ This notion of information is not very formal, but gives us an idea how $D(P\|Q)$ better captures information than $d(P,Q)$.

[B] M-ary hypothesis testing: Maximum Likelihood (ML) and Maximum Aposteriori Probability (MAP) tests

We now move to M-ary hypothesis testing. We consider a Bayesian setting where it is assumed that the unknown X is generated from a fixed distribution P_X on $\mathcal{X} = \{1, \dots, M\}$. The observation Y is generated from W_x when $X = x$. A test now is a mapping $g: \mathcal{Y} \to \mathcal{X}$ and its probability of error is given by

$$\mathbb{P}(g(Y) \neq X) = \sum_x P_X(x) \sum_{y: g(y) \neq x} W(y|x).$$

We denote this error by $P_e(g|P_X)$, and denote by $P_e^*(P_X)$ the minimum of $P_e(g|P_X)$ over g.

* P_X is uniform

For $|\mathcal{X}| = 2$, we saw the LLRT given by

$$\log \frac{W_0(y)}{W_1(y)} \underset{1}{\overset{0}{\gtrless}} 0.$$

We can reinterpret this test as

$$g(y) = \underset{x \in \{0,1\}}{\operatorname{argmax}} \; W_x(y) = \underset{x \in \{0,1\}}{\operatorname{argmax}} \; \log W_x(y).$$

This test is much more general and is called the Maximum Likelihood (ML) test:

$$g_{ML}(y) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \; \log W_x(y).$$

* General P_X

We can consider a natural generalization of the ML test where discount $\log W_x(y)$ by $\lambda_x \geq 0$ and use

$$g(y) = \max_{x \in \mathcal{X}} \; \log W_x(y) - \lambda_x.$$

This λ is often called a regularizer and its role is to make the test favor certain hypotheses $\mathcal{X}_0 \subseteq \mathcal{X}$. Indeed, regularized ML is a very popular algorithm in statistics and machine learning.

Note that $\log W_x(y) - \lambda_x = \log 2^{-\lambda_x} W_x(y)$, where $2^{-\lambda_x} \in [0, 1]$. Thus,

$$g(y) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \; 2^{-\lambda_x} W_x(y) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \; 2^{-\lambda_x} W_x(y) \Big/ \sum_{x'} 2^{-\lambda_{x'}}.$$

We can interpret $2^{-\lambda x} / \sum_{x'} 2^{-\lambda x'}$ as a distribution on $\mathcal{X}$.

A natural choice is $P_X(x)$, for which

$$P_X(x)\, W(y|x) = P_{X,Y}(x,y) = P_{X|Y}(x|y)\, P_Y(y).$$

Thus,

$$g(y) = \operatorname*{argmax}_{x \in \mathcal{X}} P_{X|Y}(x|y)\, P_Y(y) = \operatorname*{argmax}_{x \in \mathcal{X}} P_{X|Y}(x|y).$$

Since $P_{X|Y}(x|y)$ denote the a posteriori prob. of X given Y, the test above is called the MAP test, denoted

$$g_{MAP}(y) = \operatorname*{arg\,max}_{x \in \mathcal{X}} P_{X|Y}(x|y).$$

In fact, this 1 test attains $P_e^*(P_X)$.

Theorem (Optimality MAP)

$$P_e(g_{MAP}|P_X) = P_e^*$$

Proof. For any $g: \mathcal{Y} \to \mathcal{X}$,

$$P_e(g|P_X) = \mathbb{P}(g(Y) \neq X) = P_Y(y)\, \mathbb{P}(X \neq g(y)|Y=y)$$

$$\geq \sum_y P_Y(y)\, \mathbb{P}(X \neq g(y)|Y=y)$$

$$\geq \sum_y P_Y(y) \min_g \mathbb{P}(X \neq g(y)|Y=y)$$

(in fact, this is an equality. why?)

$$= \sum_y P_Y(y) \min_{\tilde{x}} \mathbb{P}(X \neq \tilde{x}|Y=y)$$

Note that $\min_{\tilde{x}} \mathbb{P}(X \neq \tilde{x} \mid Y=y) = 1 - \max_{\tilde{x}} \mathbb{P}(X=\tilde{x} \mid Y=y)$

$$= 1 - \max_{\tilde{x}} P_{X|Y}(\tilde{x}|y).$$

Thus, the right-side of the previous inequality coincides with $P_e(g_{MAP}|P_X)$, giving

$$P_e(g|P_X) \geq P_e(g_{MAP}|P_X).$$ ∎

C Which distance governs the minimum probability of error for M-ary HT?

We saw two variants of binary hypothesis testing: for the first, the min. prob of error was determined by $d(W_0, W_1)$, and for the second, the min prob. of error (of type II) was determined by $D(W_0 \| W_1)$.

Is there a similar geometric "notion" of distance that governs $P_e^*(P_X)$?

We build using an heuristic argument.

$$P_e(g|P_X) = \sum_x P_X(x) \sum_{x' \neq x} \mathbb{P}(g(Y)=x' \mid X=x)$$

$$\leq (M-1) \max_{x,x':\, x \neq x'} \mathbb{P}(g(Y)=x' \mid X=x)$$

Consider

$$\max_{x \neq x'} \mathbb{P}(g(Y)=x' \mid X=x) = 2 \max_{x \neq x'} \frac{1}{2} \mathbb{P}(g(Y)=x' \mid X=x) + \frac{1}{2} \mathbb{P}(g(Y)=x \mid X=x')$$

We can roughly approximate this expression as

$$1 - \min_{x \neq x'} d(W_x, W_{x'}).$$

Alternatively, we can use heuristics from Stein's lemma and approximate $\mathbb{P}(g(Y) = x' \mid X = x) \approx 2^{-D(W_{x'} \| W_x)}$, saying that

$$P_e^*(P_X) \approx M \, 2^{-\min_{x \neq x'} D(W_x \| W_{x'})}.$$

Thus:

→ Some notion of how spread out W_x are in a statistical distance determines the error

→ We can't distinguish more that $\min_{x \neq x'} D(W_x \| W_{x'})$ bits of uncertainty $\left(\log M \leq \min_{x \neq x'} D(W_x \| W_{x'}) \right)$.

Of course, this answer is only heuristic. But it gives us an idea of how to characterize the "information revealed by Y about X."

D Information revealed by Y about X

We now return to the question we have been pursuing. First, we define a notion of uncertainty remaining in X given Y.

* Conditional entropy of X given Y

$H(X|Y) \equiv H(P_{XY}|P_Y)$ is defined as

$$:= \sum_y P_Y(y)\, H(P_{X|Y=y}),$$

namely the average entropy of the conditional distribution.

At this point, this definition appears out of the blue. But soon we will see that this is a good definition.

Note

$$H(X|Y) = \sum_y P_Y(y) \sum_x P_{X|Y}(x|y) \log \frac{1}{P_{X|Y}(x|y)}$$

$$= \sum_{y,x} P_{XY}(x,y) \log \frac{1}{P_{X|Y}(x|y)} = \mathbb{E}\left[\log \frac{1}{P_{X|Y}(X|Y)}\right]$$

(here $P_{X|Y}(\cdot|\cdot)$ is a fixed function of (X,Y))

$$= \sum_{x,y} P_{XY}(x,y) \log \frac{P_Y(y)}{P_{XY}(x,y)}$$

$$= \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{XY}(x,y)} - \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_Y(y)} = H(X,Y) - H(Y)$$

$\sum_y P_Y(y) \log \frac{1}{P_Y(y)}$

uncertainty in (X,Y) − uncertainty in $H(Y)$

In the previous calculation, we have used the Bayes rule several times: $P_{X,Y}(x,y) = P_{X|Y}(x|y)\,P_Y(y)$

$= P_{Y|X}(y|x)\,P_X(x)$

* Information

Now that we have defined the notion of residual uncertainty, the conditional entropy of X given Y, we can define information revealed by Y about X as

(uncertainty of X) − (uncertainty remaining in X after Y is revealed)

$= H(X) - H(X|Y) =: \underline{I(X \wedge Y)}$ ⟶ mutual information b/w X and Y

$= H(X) + H(Y) - H(X,Y)$

$= H(Y) - H(Y|X)$ ⟶ Thus, information revealed by Y about X equals information revealed by X about Y

$= H(X,Y) - H(X|Y) - H(Y|X)$

Note $I(X \wedge Y) = H(X) - H(X|Y) = \mathbb{E}\left[\log \frac{1}{P_X(X)} - \log \frac{1}{P_{X|Y}(X|Y)}\right]$

$= \mathbb{E}\left[\log \frac{P_{X|Y}(X|Y)}{P_X(X)}\right]$

$= \mathbb{E}\left[\log \frac{P_{X,Y}(X,Y)}{P_X(X)\,P_Y(Y)}\right]$

$= D(P_{X,Y} \| P_X\,P_Y)$ ⟶ divergence b/w joint and independent dist.

$= \mathbb{E}_Y\left[D(P_{X|Y} \| P_X)\right]$

→ Given a channel $W: \mathcal{X} \to \mathcal{Y}$ and a distribution P_X on $\mathcal{X}$, let Y be the output of the channel when the input $X \sim P_X$.

Then,

$$I(X \wedge Y) = \mathbb{E}_X \left[D(W_X \| W \circ P_X) \right]$$

$$= \sum_x P_X(x)\, D\left(W_x \Big\| \sum_{x'} P_X(x')\, W_{x'}\right).$$

In particular, for $P_X = \text{unif}(\{1, \ldots, M\})$,

$$I(X \wedge Y) = \frac{1}{M} \sum_x D\left(W_x \Big\| \frac{1}{M} \sum_{x'} W_{x'}\right)$$

avg. distance from the centroid

It is this avg. distance that will be seen to govern the difficulty of M-ary HT

[E] Information-error bound: Fano's inequality

Fano's inequality (without proof)

For $P_X \equiv \text{unif}(\{1, \ldots, M\})$,

→ $h(q) = q \log \frac{1}{q} - (1-q) \log \frac{1}{1-q}$

$$P_e^*(\text{unif}) \geq 1 - \frac{I(X \wedge Y) + h(P_e^*(\text{unif}))}{\log M}$$

$$\geq 1 - \frac{I(X \wedge Y) + 1}{\log M}$$

Thus, if $P_e^*(\text{unif}) \leq \frac{1}{3}$,

$$\frac{3}{2}\left(I(X \wedge Y) + 1\right) \geq \log M$$

→ mutual information limits the number of values of X that can be resolved by Y