Information measures in estimation

COM-406 | Foundations of Data Science

Millen Kanabar

October 1, 2025

1 Problem setup

Say we get random variables X^n sampled i.i.d. from Bernoulli(θ), where $\theta \in \Theta := [0, 1]$. We want to estimate θ from this sequence. One can come up with a large variety of procedures to do this. In this note, we are interested in whether we can find bounds on the 'goodness' of our estimate $\hat{\theta}$ *no matter what procedure we use.* As you can probably tell, this is determined by the way we get our data.

Note that we will return to this problem in more detail later in this class, in Chapters 6 and 7.

1.1 What is a good estimator?

We will use the expected squared error $E[(\hat{\theta} - \theta)^2]$ to quantify the goodness of our estimate. This is a loss/cost function used extensively in estimation problems, and many of you will already have seen it in probability classes. Clearly, a smaller squared error is better, since this implies that the estimate is, on average, closer to θ. Now, it is important to note that we can construct estimators $\hat{\theta}$ that do not use the information from the samples at all and yet perform very well for some subset of θ's.

Consider an estimator that maps every sequence of samples to $\hat{\theta} = 1/2$. This will outperform every estimator that actually does use the information from X^n whenever $\theta = 1/2$, however, we do not want to classify such estimators as 'good' estimators. To ensure that such 'cheats' do not give us a false sense of hope, we will consider the worst-case squared error over all values θ can take,

$$R = \sup_{\theta} E[(\hat{\theta} - \theta)^2]. \tag{1}$$

2 Divergences between Bernoulli distributions

We will first quantify how divergences between two different Bernoulli distributions relate to their parameters. Fix $\delta \in (0, 1/2)$. Consider two Bernoulli distributions P_+ and P_- with parameters $1/2 + \delta$ and $1/2 - \delta$. The KL divergence between them is given as

$$D(P_+\|P_-) = \left(\frac{1}{2} + \delta\right) \log \frac{1/2 + \delta}{1/2 - \delta} + \left(\frac{1}{2} - \delta\right) \log \frac{1/2 - \delta}{1/2 + \delta}$$

Using the fact that $\log x \leq x - 1$ for every $x \in \mathbb{R}^+$,

$$\begin{aligned} D(P_+\|P_-) \leq& \left(\frac{1}{2} + \delta\right) \left(\frac{1/2 + \delta}{1/2 - \delta} - 1\right) + \left(\frac{1}{2} - \delta\right) \left(\frac{1/2 - \delta}{1/2 + \delta} - 1\right) \\ =& \left(\frac{1}{2} + \delta\right) \frac{2\delta}{1/2 - \delta} + \left(\frac{1}{2} - \delta\right) \frac{(-2\delta)}{1/2 + \delta} \\ =& 2\delta \left(\frac{1/2 + \delta}{1/2 - \delta} - \frac{1/2 - \delta}{1/2 + \delta}\right) = 2\delta \frac{(1/2 + \delta)^2 - (1/2 - \delta)^2}{1/4 - \delta^2} \\ =& \frac{4\delta^2}{1/4 - \delta^2}. \end{aligned} \tag{2}$$

Consequently, we can easily find an upper bound on $D(P_+^n\|P_-^n)$, where the distributions in the arguments are n-fold product distributions: it is simply n times the right hand side of (2) due to the chain rule of KL divergences. Note that it is crucial to use the chain rule *before* using the approximation of the log function; the upper bound will be exponential in n instead of linear otherwise.

Note: we found a slightly different version of (2) in class, but the heart of the bounding techniques is the same and leads to roughly the same constants.

3 Le Cam's method

We will now derive a lower bound on the worst-case error. This technique, due to Lucien Le Cam, relies on the fact that if one has a 'good' estimator, one can use it to distinguish between distributions P and Q simply by estimating the distribution from the samples, and guessing the distribution closer to the estimate.

Let $\tilde{\theta}$ be a random variable taking values $1/2 + \delta$ and $1/2 - \delta$ with equal probability. Using the fact that $E[f(\tilde{\theta})] \leq \sup_{\theta \in \Theta} f(\theta)$,

$$R = \sup_\theta E_{\hat{\theta}} \left[\left(\hat{\theta} - \theta\right)^2\right] \geq E_{\tilde{\theta}} \left[E_{\hat{\theta}} \left[\left(\hat{\theta} - \tilde{\theta}\right)^2\right]\right].$$

Now, consider the events $A := \{\hat{\theta} \in [0, 1/2]\}$ and $A^{\complement}$. We can see that for both values that $\tilde{\theta}$ takes, the minimum expected error when the 'wrong' event occurs is at least δ^2.

Let us write this formally. Let the distribution of $\hat{\theta}$ conditioned on $\tilde{\theta} = 1/2 + \delta$ be denoted by $\bar{P}_+$ and conditioned on $\tilde{\theta} = 1/2 - \delta$ be denoted by $\bar{P}_-$. Then,

$$\begin{aligned}
R \geq & E_{\tilde{\theta}}\left[E_{\hat{\theta}}\left[\left(\hat{\theta}-\tilde{\theta}\right)^2\right]\right] \\
= & \frac{1}{2}\left(E_{\hat{\theta}\sim\bar{P}_+}\left[\left(\hat{\theta}-\left(\frac{1}{2}+\delta\right)\right)^2\right] + E_{\hat{\theta}\sim\bar{P}_-}\left[\left(\hat{\theta}-\left(\frac{1}{2}-\delta\right)\right)^2\right]\right) \\
\overset{(a)}{\geq} & \frac{\delta^2}{2}\left(\bar{P}_+(A) + \bar{P}_-\left(A^{\complement}\right)\right) = \frac{\delta^2}{2}\left(1 - \left(\bar{P}_-(A) - \bar{P}_+(A)\right)\right) \\
\geq & \frac{\delta^2}{2}\left(1 - \|\bar{P}_+ - \bar{P}_-\|_{TV}\right), \qquad (3)
\end{aligned}$$

where (a) holds because if $\tilde{\theta}$ and $\hat{\theta}$ are on different sides of $1/2$, the error is *at least* δ^2, and it is at least 0 otherwise. This is the crux of Le Cam's method. More general results can be found in the sources mentioned in the Further Reading section; we will only consider the above proofs in this note.

4 Putting it all together

First, we find a good upper bound for the total variation distance in (3).

$$\begin{aligned}
\|\bar{P}_+ - \bar{P}_-\|_{TV} & \overset{(a)}{\leq} \sqrt{\frac{1}{2}D(\bar{P}_+\|\bar{P}_-)} \\
& \overset{(b)}{\leq} \sqrt{\frac{1}{2}D(P_+^n\|P_-^n)} \\
& = \sqrt{\frac{n}{2}D(P_+\|P_-)} \\
& \overset{(c)}{\leq} \sqrt{\frac{n}{2}\cdot\frac{4\delta^2}{1/4-\delta^2}}, \qquad (4)
\end{aligned}$$

where (a) follows from Pinsker's inequality, (b) follows from the Data Processing Inequality applied to the channel $P_{\hat{\theta}|X^n}$, and (c) follows from (2). Substituting (4) in (3),

$$R \geq \frac{\delta^2}{2}\left(1 - \sqrt{\frac{n}{2}\cdot\frac{16\delta^2}{1-4\delta^2}}\right).$$

For any $n > 0$, choosing δ such that $\delta^2 = 1/64n$ will make the expression inside the square root smaller than $1/4$, and therefore,

$$R \geq \frac{1}{128n}\left(1 - \frac{1}{2}\right) \geq \frac{1}{256n}.$$

This general framework allows us to systematically prove nontrivial lower bounds on the worst-case losses without worrying about the specifics of the estimation scheme being used. It is remarkable that it captures the correct scaling: This bound shows that no matter how well one designs the estimator, the worst-case error will not decay faster than $\Omega(1/n)$, and one can find estimators that match that [1]. Can you think of a simple estimator that has similar performance?

5 Further reading

This example simply demonstrates how we can use the ideas learned in class to say interesting and useful things about problems that, at first glance, might not seem very relevant. **We will not be asking questions specifically from this session/note or from any of the references, and only the content of Chapters 6 and 7 is in the syllabus for the exams.**

That said, those of you who are interested in reading more about how one can find lower bounds on estimation loss are encouraged to read the excellent write-up by Bin Yu [4] covering three different methods (including Le Cam's method discussed here). These can also be found with more background and wider coverage in books by Tsybakov [3, Chapter 2] and Polyanskiy and Wu [2, Chapter 31].

[1] for a full treatment of the algorithmic side of this problem, see [1]

References

[1] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1066–1100, Paris, France, 03–06 Jul 2015. PMLR.

[2] Yury Polyanskiy and Yihong Wu. Information Theory: From Coding to Learning, January 2025. ISBN: 9781108966351 Publisher: Cambridge University Press.

[3] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009.

[4] Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and Grace L. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435. Springer, New York, NY, 1997.