Finite sample versus asymptotic theory

Art Owen
Stanford University
September 2021

After the 2021 ICML workshop on reinforcement learning, we had a panel discussion. One of the topics that came up was the difference between finite sample theory and asymptotic theory. Theoretical computer science puts relatively more emphasis on finite sample theory, while statistics has relatively more emphasis on asymptotics. So, why is that?

Habituation

Maybe it's like people's choice of computer environment, e.g., python versus R versus Matlab versus Mathematica et cetera. We learn it one way and get used to it that way, and get skilled at using those tools. Then for the other kind of theory we are slightly ill at ease. The formulations don't quite look right, we have to translate from one to the other, it's not quite our skill set and it doesn't look like the papers we have read. I think that habituation plays a role but that it is not the only reason.

Consequences

Many of the problems in statistics are simply about estimating a parameter θ by some $\hat\theta$. Reinforcement learning includes settings where you or a computer take some action. In some of the statistical settings the action might essentially reduce to just believing that $\theta \approx \hat\theta$. You might then want to accurately estimate $|\hat\theta - \theta|$ and pick a method that does not overestimate it. When as is common, the CLT or other asymptotics set in pretty fast then you can use them without harm. If instead, the consequence of your algorithm is to move a robot arm or decide who goes into the ICU then you want a wider berth around $\hat\theta$ than a 95% or 99% confidence interval that probably undercovers slightly will give.

Problem complexity

If we take an average of IID values, then the variance is exactly σ^2/n so the asymptotic and finite sample values match. A $100(1-\alpha)\%$ confidence interval based on the CLT using an estimate $\hat{\sigma}$ attains coverage $1-\alpha+O(1/n)$ and is in that sense even more accurate than the center point which has root mean square error $O(1/\sqrt{n})$. Very old text books would have rules of thumb like using the CLT when there are $\geqslant 30$ observations. That hides the problem of extremely skewed data or extremely sharp confidence level demands, but they'd work well enough and often enough.

Many of the classical statistical problems with modest dimensional parameters have estimates that are not so different from an average.

Data set size and null hypotheses

Much of the early emphasis in statistics was about testing whether a null hypothesis that $\theta = 0$ holds, based on small samples of noisy data. That task remains prominent in introductory courses. Of course we usually know for sure that $\theta \neq 0$ before seeing any data at all. The real point about testing $\theta = 0$ is that if you cannot reject $\theta = 0$ then you cannot know the sign of θ, at least not based on your data. There can be genuine prior uncertainty about $\mathrm{sign}(\theta)$. The hardest positive/negative value to reject is $\theta = \pm\epsilon$ as $\epsilon \to 0$. So, you're back to testing whether 0 is plausible. If it is, then you don't want to start interpreting it strongly because there might be nothing there or you might even have the wrong sign. In problems like that you don't want to greatly overestimate $\mathrm{var}(\hat{\theta})$, so plugging in an asymptotically justified estimate is better than using a conservative bound. That is, false non-discoveries have a cost too, not just false discoveries.

In a large data set we might find that a low dimensional parameter is estimated with extreme accuracy. Then testing the null is irrelevant compared to handling lack of fit. Ironically, large data sets could reduce the value of asymptotic methods. Or we might find that the individual components of a vector parameter are not interesting in and of themselves but only through some derived properties like an AUC. For something like predictive accuracy we might be dealing with binary variables that make finite sample theory applicable.

Sequencing

Maybe it's like compile time versus run time. When we are desiging an algorithm to use in the future we know very little about the setting it will face. Something like PAC bounds handle many eventualities. Later when we have a specific data set at hand, our function class might narrow towards a singleton about just that day's problem and we want to estimate its error. In a setting like that we might use finite sample theory to decide on what is a good method to use. Later on the user, who might be us or somebody else, could use an asymptotic method to estimate their attained accuracies.

Infinite regress

This is commonly presented as 'turtles all the way down'. An estimate of error might require its own error estimate and so on ad infinitum. One usually has to truncate this to some level. An alternative is to get something stronger theoretically under some assumptions, such as random variables with a known bound, and independence. At that point the uncertainty transfers to the assumptions. If the assumptions are rock solid, that's good but the result might be something conservative enough to include quite implausible settings. If there is doubt about the assumptions, then we can test them but that triggers an infinite regress.

Unbounded variables

Much of statistics is derived for unbounded random variables, using Gaussian or gamma distributions and similar things. Binary and categorical and other bounded random variables also appear in statistics but they are much more dominant in computer science. There are many more finite sample results available for bounded random variables than for unbounded ones. Nonparametric confidence intervals for the mean of unbounded IID random variables do not exist, so asymptotics are the only choice.

Many of the phenomena in statistics are not quite unbounded even though models allow them to be unbounded. It is better to say that they are 'indefinite'. For instance, the height of any future human is a bounded random variable. It's not clear how to choose a value for that bound if one has to. We might be safe using 4 meters as the bound but 3.5 or 3 meters might give a sharper answer while not being quite as safe. For long-tailed quantities especially those measured in currency (income, wealth, market capitalization) picking a bound would be very hard. In the correct units, an indefinite variable belongs to $[0, 1]$ making the data amenable to theory for bounded variables, but the user in an applied context has to choose a scale. Likewise we can devise theory for sub-Gaussian random variables. The user might have to be more precise than that and supply the constants. That could be pretty hard.

Empirical Bernstein

I learned about the empirical Bernstein inequality while preparing for my talk. I like how it manages to be less conservative while still satisfying finite sample properties. So it helps in having things both ways. It does require a known bound but it can then be much less conservative through the use of a sample variance.

Acknowledgments

I thank Csaba Szepesvai and Lester Mackey for some email discussions on this topic. This is not meant to imply that they concur. This note was supported by the U.S. NSF under grant IIS-1837931.