Uniform, nonparametric, non-asymptotic confidence sequences

Hsin-Po Wang

May 10, 2019

Abstract

This is a summarize of mainly arXiv:1810.08240. It focuses on how to estimate the mean (and sometimes variance) of a given sequence of variables in an online learning setup. By online learning setup we meant that at every time t we are given an instance of X_t and are to output an interval $[a_t, b_t] \subset \mathbb{R}$ depending on $X_1, \dots, X_t$. The interval should have contained a certain estimand; we fail if at any time t our output does not contain so. The main issue is to minimize $b_t - a_t$ as well as the failure probability.

Problem Setup

Consider the following almost trivial setup: Let X be a single point $\{*\}$ (the observable is trivial). Let Y be $\mathbb{R}$. Let $\mathcal{F}$ be $\mathbb{R}$, functions from the singleton to reals. We use the quadratic loss, i.e., $\ell(f) := (y - f(*))^2$. Then the risk minimizer is $f^* := \mathbb{E}Y$. So far this setup recovers the classical setup in statistics where a series of i.i.d. random variables are given and we want to estimates the hidden parameter. For instance, we may assume that X_i follows Bernoulli(μ) for some unknown μ; or we assume X_i follows Normal$(\mu, 1)$ for some unknown μ. In the machine learning setup, our algorithm takes $X_1, X_2, \dots$ and outputs an $\hat{f}$, and how good $\hat{f}$ is is measured by, say,

$$\sup_\mu \mathbb{E}(Y - \hat{f})^2 - \mathbb{E}(Y - f^*)^2 = \sup_\mu \mathbb{E}(\hat{f} - \mathbb{E}Y)^2.$$

Here Y is a fresh sample and the supremum is taken over all possible hidden parameters. Other ways to measure how good $\hat{f}$ is include finding δ, ε such that

$$\sup_\mu \mathbb{P}\{|\hat{f} - \mathbb{E}Y| > \varepsilon\} < \delta.$$

In classical statistics setting, the interval $[\hat{f}-\varepsilon, \hat{f}+\varepsilon]$ is called the *confident interval* and $1-\delta$ is called the *confident level*. It can also be seen, for instance, as the 0-1 loss of $\hat{f}$. There are countless results regarding how to give confident interval with a certain confident interval under different assumption on X_i: Bernoulli, normal with or without fixed variance, sub-gaussian, sub-gamma, etc.

Classical Approach

The first step of the classical approach is to find the sufficient statistic. Formally speaking, a *sufficient statistic* for the hidden parameter μ is an algorithm T outputting a number $T_n = T(X_1, \dots, X_n)$ such that the conditional distribution of $X_1, \dots, X_n$ given T_n does not depends on μ anymore. In other words, if a wise algorithm is presented T_n, knowing $X_1, \dots, X_n$ does not help improving the estimate of μ. For example, if we want to estimate the mean of Bernoulli or Normal, knowing the empirical average is sufficient. If we want to estimate the upper-bound of a uniform distribution, then knowing the maximum among X_i's is sufficient. The concept of sufficient statistic provides some insight into the converse results: if T_n as a random variable does not behave well, then there is very little we can say about μ.

Let us continue the examples regarding estimating mean where the empirical average is a sufficient statistic. We have the following result regarding averages:
Law of large numbers: $\sum X_i/n \to \mu$.
Central limit theorem: $\sum X_i/\sqrt{n} \to \text{Normal}(\mu, \sigma^2)$.
Large deviations principle: $\mathbb{P}\{\sum X_i/n > \mu + v\} \to \exp(-nI(v))$ for some rate function $I(v)$.
Finally law of iterated log: $\limsup \sum(X_i - \mu)/\sigma n\sqrt{2n \log\log n} = 1$. Except that LLN is an optimistic result, CLT, LDP, and LIL all contain pessimistic results. They are saying that the empirical average *must* deviate from the true mean by a certain amount so the estimate of μ is never accurate. Any result regarding the confident interval and level must follow.

New Approach

The generalization by Howard et. al. is threefold: One, they are using an online learning setup. That is, at every time t they want to produce a confident interval $[a_t, b_t]$ based on the history $X_1, \dots, X_t$. And the confident level is defined as the probability that all (but finitely many) intervals contain the estimand. Two, while the width $b_t - a_t$ heavily depends on the variance, the algorithm does not need to know the variance in the first place; it can estimate it on the fly. This is inspired

by the fact that the empirical deviation is a sufficient statistic of the variance in, say, the normal case. Three, they allow the mean and variance to change from time to time without notifying the algorithm. This is similar to Zinkevich's adversarial framework of online learning where X_n is chosen evilly to maximize the regret of $\hat{f}_n$.

We now define explicitly what we want to estimate. Let our estimand, μ_t, be $\sum_{i=1}^{t} \mathbb{E}[X_i|X_1 \cdots X_{i-1}]/t$. Notice that this quantity is hisotry-dependent: the underlying story is that everyday there is a fresh, fine, mean of X_i conditioning on the known history $X_1, \dots, X_{i-1}$; and we are not interested in the coarse mean $\mathbb{E}X_i$. In the sense of Doob decomposition theorem, μ_t is the predictable/drift part of X_i. Let S_t be $\sum_{i=1}^{t} X_i - \mathbb{E}[X_i|X_1 \cdots X_{i-1}]$; this is the martingale part of the decomposition of X_i. Similarly, let V_t be the sum of conditional (fine) variances. Now an LDP-flavor bound should look like $\mathbb{P}\{S_t > 0 + u(V_t)\} < \delta$ or like

$$\mathbb{P}\left\{\frac{1}{t}\sum_{i=0}^{t} X_i > \mu_t + \frac{u(V_t)}{t}\right\} < \delta.$$

Here u is a function that scales the variance properly such that the final probability is δ (c.f. Chebyshev inequality).

Let $\psi : [0, \lambda_{\text{max}}) \to \mathbb{R}$ be a function. This function lives in the MGF world and is used to bound other MGFs from above. We now state Howard's Assumption 1: for every $\lambda \in [0, \lambda_{\text{max}})$, there exists a supermartingale $L_t(\lambda)$ such that $\mathbb{E}L_0(\lambda)$ is a constant not depending on λ, and such that

$$\exp(\lambda S_t - \psi(\lambda)V_t) \leq L_t(\lambda).$$

This inequality handles two things at once: On one hand, if X_t are i.i.d. random variables, the LHS falls back to $\exp(\lambda S_t - \log \text{MGF}(\lambda)n)$ (where $V_n = n$), which appears in the derivation of the Cramér (rate) function in LDP. On the other hand, if X_i is very general (mean and variance change on the fly, historically dependent), then the fact that $L_t(\lambda)$ is a supermartingale provides first-moment bounds (c.f. Hoeffding, Azuma, McDiarmid, etc).

Main Theorems

Theorem 1: Let $h(k)$ be such that $\sum 1/h(k) < 1$. Let $S_\alpha(v)$ be a function of the form $k_1\sqrt{v\ell(v)} + ck_2\ell(v)$ where $\ell(v)$ is of the form $\log h(\log_\eta v) + \log(\mathbb{E}L_0/\alpha)$ and c, k_1, k_2, η are properly chosen constants. Then

$$\mathbb{P}\left\{\frac{1}{t}\sum_{i=0}^{t} X_i > \mu_t + \frac{S_\alpha(1 \vee V_t)}{t} \text{ infinitely many times}\right\} = 0.$$

Notice how this bound avoid LIL: If we let $h(k) = k$, then $\ell(v) \approx \log\log v$ so $S_\alpha(v) \approx \sqrt{v \log\log v}$ and $S_\alpha(1 \vee V_t) \approx O(\sqrt{t \log\log t})$ since $V_t \approx t$. Then the bound says that the average of X_t deviates from the mean by $\sqrt{t \log\log t}$ finitely many times, while LIL says infinitely many times. The bug here is that $\sum h(k)$ does not converge. Should $\sum h(k)$ converge, then h and ℓ and S_α increase fast enough to avoid LIL.

In particular, if we let $h(k) = k^s$ for some $s > 1$, then S_t has estimate

$$\mathbb{P}\left\{S_t \sqrt{t \log\log(2t) + 0.72t \log(5.2/\delta)} \text{ for some } t\right\} < \delta$$

Theorem 2: Let w_k be some properly chosen weights related to the pdf of X_i. Then

$$\text{DM}_\alpha(v) := \sup\left\{ s \in \mathbb{R} : \sum_{k=0}^{\infty} w_k \exp(\lambda_k s - \psi(\lambda_k) v) < \frac{\mathbb{E}L_0}{\alpha} \right\}$$

is such that

$$\mathbb{P}\left\{ \frac{1}{t}\sum_{i=0}^{t} X_i > \mu_t + \frac{\text{DM}_\alpha(V_t)}{t} \text{ once} \right\} = 0.$$

This bound is closely related to LPD for that if X_i are i.i.d. random variables then $\text{DM}_\alpha(v)$ is just the "inverse function" of the rate function $I(v)$, which used to be defined as $\sup_\lambda \lambda x - \log \text{MGF}(\lambda)$.