Faster Hoeffding racing: Bernstein races via jackknife estimates

Po-Ling Loh[1] and Sebastian Nowozin[2]

[1] Department of Statistics, University of California, Berkeley, CA 97420, USA,
ploh@berkeley.edu
[2] Microsoft Research Cambridge, 21 Station Road, Cambridge, UK,
Sebastian.Nowozin@microsoft.com

Abstract. Hoeffding racing algorithms are used to achieve computational speedups in settings where the goal is to select a "best" option among a set of alternatives, but the amount of data is so massive that scoring all candidates using every data point is too costly. The key is to construct confidence intervals for scores of candidates that are used to eliminate options sequentially as more samples are processed. We propose a tighter version of Hoeffding racing based on empirical Bernstein inequalities, where a jackknife estimate is used in place of the unknown variance. We provide rigorous proofs of the accuracy of our confidence intervals in the case of U-statistics and entropy estimators, and demonstrate the efficacy of our racing algorithms with synthetic experiments.

Keywords: Bernstein inequalities, racing algorithms, bandits, jackknife, decision tree induction

1 Introduction

Many present-day machine learning algorithms suffer from significant computational challenges due to massive amounts of data. Whereas traditional statistical problems were limited by the cost of acquiring more samples, modern applications involve processing millions of cheaply-acquired samples, generating a large computational burden. Standard statistical methods must therefore be adjusted to focus on making inferences as efficiently as possible based on very large amounts of data.

Although having access to an essentially infinite pool of data should make statistical inference easier, the challenge is to determine how much data to process in what order before making a correct conclusion with sufficiently high probability. For instance, suppose the goal is to determine the best model amongst a set of M candidates based on N observations. The classical approach is to evaluate M likelihood functions upon all available samples, leading to an $\mathcal{O}(MN)$ computation. When N is extremely large, one may instead choose to evaluate certain samples only on certain models, leading to significant computational speedups.

In order to address these statistical problems in a mathematically rigorous manner, we adopt the terminology of the classical multi-armed bandit [22], [5].

In this setting, a gambler needs to select an optimal arm among a finite set in order to maximize the reward after sequential arm pulls. However, the gambler's decisions are made based on random data from an unknown distribution, resulting in a tradeoff between exploration (drawing more data to estimate arm values more accurately) and exploitation (choosing to pull an arm based on currently available samples, incurring a penalty from choosing the wrong arm).

Traditionally, the gambler aims to minimize expected *regret*, the difference between the reward accrued from picking the optimal arm on each pull and the reward accrued by the algorithm. In this paper, we consider a slightly different setting where the gambler wants to identify an almost optimal arm with high confidence using as few resources as possible, and then pulls the arm. More precisely, the gambler needs to determine how many samples to acquire before concluding that with probability at least $1-\delta$, the selected arm is optimal. In the probably approximately correct (PAC) model, the gambler may choose any arm with value within ϵ of the optimum, and all such arms are equally acceptable.

A promising approach for constructing PAC algorithms in bandit problems utilizes a technique known as the racing algorithm. Racing algorithms were first introduced by Maron and Moore [16] in the context of minimizing expected risk among a collection of models based on cross-validation scores. The key is to construct confidence intervals for the population-level quantities that shrink as the number of samples increases. Since the only objective is to find an optimal arm, the user may eliminate arms with low values after drawing only a few samples, then "race" the best candidates against each other. The Hoeffding race [16] derives its name from Hoeffding's inequality, which is used to construct confidence intervals; Bernstein races improve upon Hoeffding races by constructing tighter confidence intervals via Bernstein's inequality [18], [20]. Racing algorithms lead to computational speedups in settings where optimal arms have fairly close values, since most of the computation focuses on distinguishing amongst top arms, rather than being wasted on accurate estimation of low-performing arms.

Domingos and Hulten [8] introduced the idea of using Hoeffding racing to speed up decision tree learning, where successive splits are selected from steaming data. Their paper has sparked great interest in the online learning community [14], [21], [13], [24]. However, the form of the information gain estimator used to choose decision tree splits lies beyond the domain of Hoeffding's inequality. Hence, although Hoeffding racing has been applied with much empirical success, the theoretical justification in these settings has not been rigorous.

In this paper, we show how to extend the theory of racing algorithms to broader classes of statistics including discrete entropy—extending easily to estimation of the information gain statistic used in decision tree induction. In fact, we propose tighter variants of Hoeffding races based on empirical Bernstein inequalities, known to provide significant speedups in many applicaitons of interest. Whereas previous work on empirical Bernstein inequalities [3], [20] has relied heavily on finding an appropriate variance surrogate based on the specific type of estimator, our main contribution is to present a novel method for constructing confidence intervals based on a jackknife estimate, applicable to

an extremely broad class of statistics. We prove rigorously that our jackknifed Bernstein bounds are accurate in the case of U-statistics and discrete entropy estimators, and our proof techniques have natural generalizations to other types of statistics. We validate our theoretical results on synthetic data sets, and demonstrate that our methods yield vast computational savings.

2 Preliminaries

Recall the setting of a multi-armed bandit. Let $X_1, \ldots, X_N$ be i.i.d. data from an unknown distribution with density q, and consider a family of functions $\{f_m\}_{m=1}^M$ defined on q, forming the arms of the bandit. The goal is to determine the optimal arm $m^* \in \arg\max_m f_m(q)$ based on the X_i's. For instance, we may estimate m^* using $\widehat{m} \in \arg\max_m f_m(\widehat{q})$, where $\widehat{q}$ is the empirical distribution of the X_i's. In settings of where N is very large, it is computationally expensive to evaluate all N samples on all M functions before estimating m^*. Our goal is to decrease the number of function evaluations, while guaranteeing that the probability of picking an optimal arm is at least $1-\delta$ for some fixed $\delta \in (0,1)$.

The original Hoeffding racing paper [16] considers $f_m(q) = \mathbb{E}_{X\sim q}[g_m(X)]$; i.e., the f_m's are means of known functions $\{g_m\}_{m=1}^M$. In its most general form, a racing algorithm operates by maintaining a confidence interval $[a_m, b_m]$ for each f_m, which is updated according to the samples evaluated on arm m, as well as an active set $\mathcal{S}$, which is initialized with $\mathcal{S}_1 = \{1, \ldots, M\}$. At step i of the algorithm, a data sample X_i is drawn and evaluated on all arms in the current active set $\mathcal{S}_i$. Then the confidence intervals for each arm are updated to $[a_m^i, b_m^i]$. Letting $a_0^i := \max_m a_m^i$, we set $\mathcal{S}_{i+1} = \mathcal{S}_i \backslash \{m : b_m^i < a_0^i\}$. The algorithm terminates if either all N samples have been used, or only one arm remains.

To maximize efficiency, we wish to construct intervals $[a_m^i, b_m^i]$ of minimal width, while maintaining the correctness of our overall algorithm with probability at least $1-\delta$. When the X_i's are independent and $f_m(q) = \mathbb{E}_{X\sim q}[g_m(X)]$, with $|g_m| \le B$ for all m, Hoeffding's inequality [12] gives the $1-\delta$ confidence intervals $[a_m, b_m]$ defined by

$$\frac{1}{n_m}\sum_{i=1}^{n_m} g_m(X_i) \pm \sqrt{\frac{2B^2}{n_m}\log\left(\frac{2}{\delta}\right)}, \tag{1}$$

where n_m is the number of samples evaluated on arm m. Using the confidence intervals (1) with δ replaced by $\frac{\delta}{NM}$ then yields an algorithm that succeeds with probability at least $1-\delta$. This is the traditional Hoeffding race.

When we know in addition that $\mathrm{Var}[g_m(X_i)] \le \sigma_m^2$, however, we may use Bernstein's inequality [4] to obtain the tighter $1-\delta$ confidence interval

$$\frac{1}{n_m}\sum_{i=1}^{n_m} g_m(X_i) \pm \left(\sqrt{\frac{2\sigma_m^2}{n_m}\log\left(\frac{2}{\delta}\right)} + \frac{4B}{3n_m}\log\left(\frac{2}{\delta}\right)\right). \tag{2}$$

However, in general, σ_m^2 is not known a priori and must be estimated based on observed samples. Audibert et al. [3] developed an *empirical* Bernstein bound that replaces the unknown variance σ^2 in equation (2) by an estimate $\widehat{\sigma}(X)$.

When the f_m's are not simple empirical averages of independent observations, however, Hoeffding's and Bernstein's inequalities *do not* apply, undermining the validity of the confidence intervals (1) and (2). Examples of such functions include the following, where we suppress the dependence on m to simplify notation.

Example 1 (U-statistics). Recall that $f(X_1, \ldots, X_n)$ is a U-statistic of order k with kernel g if

$$f(X_1, \ldots, X_n) = \frac{1}{n \cdots (n-k+1)} \sum_{i_1, \ldots, i_k} g(X_{i_1}, \ldots, X_{i_k}), \tag{3}$$

where the sum is taken over all ordered k-tuples of distinct integers in $\{1, \ldots, n\}$. For instance, the sample variance is a U-statistic of order 2 with kernel $g(x_i, x_j) = \frac{(x_i - x_j)^2}{2}$. Note that individual terms of U-statistics are *not* independent.

Example 2 (Discrete entropy). Suppose the X_i's take values in $\{1, \ldots, K\}$, and let $\{\widehat{p}_k\}_{k=1}^K$ denote empirical proportions. The plugin entropy estimator is

$$f(X_1, \ldots, X_n) = -\sum_{k=1}^{K} \widehat{p}_k \log \widehat{p}_k, \tag{4}$$

which *cannot* be written as a simple empirical average.

Example 3 (Cross-validation). In the context of model selection, suppose the score of a model is given by

$$f(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \gamma(X_i; g_{(X_{\setminus i})}),$$

where g is the estimator based on leave-one-out data and γ is the cross-validation error function. Although $f(X_1, \ldots, X_n)$ is an empirical average of cross-validation errors over n data samples, the quantities in the summand are *not* independent. Maron and Moore [16] show that Hoeffding racing appears to behave well empirically and select the optimal model with high probability.

One major obstacle in developing Bernstein-type inequalities for the estimators described in the examples above is that unlike in the case of empirical averages, there is no natural analog of the sample variance as an estimator for $\mathrm{Var}[f(X_1, \ldots, X_n)]$. As argued by previous authors [18], [20] and demonstrated by our simulations (see Table 1), Bernstein inequalities often yield much greater gains in efficiency than Hoeffding inequalities. Our goal is to establish empirical Bernstein inequalities for statistics such as the ones mentioned above.

3 Results

Our main result concerns an empirical Bernstein bound involving a jackknife estimate for the variance of the statistic. Following the statement of the main theorem, we demonstrate the applicability of our result via corollaries for U-statistics and the discrete entropy estimator. Proofs are provided in Section 4.

Analogous to the idea behind Hoeffding and Bernstein races, we may use our bounds to construct $1-\frac{\delta}{MN}$ confidence intervals at each step of the race, yielding a racing algorithm that successfully selects the optimal arm with probability at least $1-\delta$. As before, we suppress dependence on the arm index m and consider i.i.d. samples $\{X_1, \ldots, X_n\}$. We have the following definition, which will be useful in prescribing sufficient conditions for our confidence intervals to hold:

Definition 1. *A statistic $f(X_1, \ldots, X_n)$ satisfies the **bounded difference condition** with parameter b if for each j and all X_i's,*

$$|f(X_1, \ldots, X_j, \ldots, X_n) - f(X_1, \ldots, X_j', \ldots, X_n)| \leq b,$$

where the statistics are evaluated on data sets differing in only one position.

Let $Z := f(X_1, \ldots, X_n)$ denote the statistic evaluated from the data. Recall that the jackknife estimate of variance [10], [7] is given by

$$V_n^J := \frac{n-1}{n} \sum_{i=1}^{n} (Z_{(i)} - \overline{Z}_{(\cdot)})^2 = \frac{n-1}{n^2} \sum_{i<j} (Z_{(i)} - Z_{(j)})^2, \tag{5}$$

where $Z_{(i)} := f(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and $\overline{Z}_{(\cdot)} := \frac{1}{n}\sum_{i=1}^{n} Z_{(i)}$. Furthermore, we have the Efron-Stein inequality [10], [25]:

$$\mathrm{Var}[Z] \leq \left(\frac{n}{n-1}\right)^2 \cdot \mathbb{E}[V_n^J]. \tag{6}$$

We make the following assumptions on the rate of concentration of Z and V_n^J:

Assumption 1 (Bernstein bound on Z) *For all $\delta > 0$, the following inequality holds for any constant $\overline{V} \geq \left(\frac{n}{n-1}\right)^2 \cdot \mathbb{E}[V_n^J] + f_0(n)$:*

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq c_1 \sqrt{\left(\overline{V} + \frac{f_1(n)}{\overline{V}} \log\left(\frac{c_2}{\delta}\right)\right) \log\left(\frac{c_2}{\delta}\right)} + f_2(n) \log\left(\frac{c_2}{\delta}\right)\right) \leq \delta, \tag{7}$$

for some constants $c_i > 0$ and functions $f_i(n)$.

Assumption 2 (Concentration of V_n^J) *For all $\delta > 0$, we have the inequality*

$$\mathbb{P}\left(|V_n^J - \mathbb{E}[V_n^J]| \geq f_3(n) \sqrt{\log\left(\frac{c_3}{\delta}\right)}\right) \leq \delta, \tag{8}$$

for some constant $c_3 > 0$ and function $f_3(n)$.

By a simple union bound, we have the following result:

Theorem 1. *Suppose Assumptions 1 and 2 hold. Then*

$$Z \pm \left\{ c_1 \sqrt{\left(\widetilde{V} + \frac{f_1(n)}{\widetilde{V}} \log\left(\frac{c_2}{\delta}\right)\right) \log\left(\frac{c_2}{\delta}\right)} + f_2(n) \log\left(\frac{c_2}{\delta}\right) \right\}$$

is a $1 - 2\delta$ *confidence interval for* $\mathbb{E}[Z]$*, where*

$$\widetilde{V} := \left(\frac{n}{n-1}\right)^2 \left(V_n^J + f_3(n) \sqrt{\log\left(\frac{c_3}{\delta}\right)} \right) + f_0(n).$$

In particular, if $\mathbb{E}[V_n^J] = \Omega\left(\frac{1}{n}\right)$ *and* $f_1(n), f_2(n) = o\left(\frac{1}{n^2}\right)$ *and* $f_0(n), f_3(n) = o\left(\frac{1}{n}\right)$*, then*

$$Z \pm c_1 \sqrt{V_n^J \log\left(\frac{c_2}{\delta}\right)} \tag{9}$$

is an asymptotic $1 - 2\delta$ *confidence interval for* $\mathbb{E}[Z]$*.*

The main work comes in verifying that Assumptions 1 and 2 hold in settings of interest. In the proofs of the corollaries below, we illustrate two different techniques for establishing the required assumptions.

We begin with a corollary about U-statistics:

Corollary 1. *Suppose* Z *is a* U*-statistic of order* k *and bounded kernel* $|g| \le B$*. Then*

$$Z \pm \left(\sqrt{2\left(\frac{n}{n-1}\right)^2 \left(V_n^J + \sqrt{\frac{B^2 c_k}{n^3} \left(\frac{n}{n-1}\right)^{2k-1} \log\left(\frac{2}{\delta}\right)} \right) \log\left(\frac{4}{\delta}\right)} + \frac{b_k}{n} \log\left(\frac{4}{\delta}\right) \right)$$

is a $1 - 2\delta$ *confidence interval for* $\mathbb{E}[Z]$*. Here,*

$$b_k := 2^{k+3} k^k + \frac{2}{3k}, \qquad \text{and} \qquad c_k := \frac{k(k+1)(k!)^2}{(2k-2)!}.$$

Remark 1. The value of c_k based on a very rough bound and could be sharpened. However, in this paper we are more concerned with establishing asymptotically accurate Bernstein bounds, so we we will not worry about optimizing constants. In particular, it is clear from the above expression that

$$Z \pm \sqrt{2 V_n^J \log\left(\frac{4}{\delta}\right)}$$

is an asymptotic $1 - 2\delta$ confidence interval for $\mathbb{E}[Z]$. When $Z = \frac{1}{n}\sum_{i=1}^n X_i$ is a simple empirical average, we have $V_n^J = \frac{\widehat{\sigma}^2}{n}$, where $\widehat{\sigma}$ is the sample variance, so the bound in Corollary 1 agrees with the familiar empirical Bernstein bound [3] up to constant factors. For higher-order U-statistics, the confidence intervals are

of the same order as the empirical Bernstein bounds proposed by Peel et al. [20], and V_n^J is not unrelated to the variance surrogate used there. However, the main point of Corollary 1 is to demonstrate the broad applicability of our jackknife method for constructing empirical Bernstein confidence intervals that may be rigorously proven to provide accurate coverage.

For the discrete entropy, given by equation (4), we have the following result:

Corollary 2. *Suppose Z is the discrete entropy estimator over K classes. Then*

$$Z \pm \frac{5}{2}\sqrt{\left(\widetilde{V} + \frac{f_1(n)}{\widetilde{V}}\log\left(\frac{2}{\delta}\right)\right)\log\left(\frac{2}{\delta}\right)}$$

is a $1-2\delta$ confidence interval for $\mathbb{E}[Z]$, where

$$\widetilde{V} := \left(\frac{n}{n-1}\right)^2\left(V_n^J + f_3(n)\sqrt{\log\left(\frac{2}{\delta}\right)}\right) + \frac{4K\log^2 n}{n^{3/2}}$$

and

$$f_1(n) := \frac{2(16\log^2 n + 32\log n + 64 + \frac{8}{n}\log^2 n)^2 + 64K^2\log^4 n}{n^3},$$
$$f_3(n) := \frac{16\log^2(n-1) + 32\log(n-1) + 64 + \frac{8}{n-1}\log(n-1)}{2(n-1)\sqrt{2n}}.$$

Remark 2. Since $f_1(n) = o\left(\frac{1}{n^2}\right)$ and $f_3(n) = o\left(\frac{1}{n}\right)$, the second part of Theorem 1 regarding asymptotic intervals (9) holds. Figure 1 provides simulations confirming the accuracy of asymptotic intervals even for moderate n.

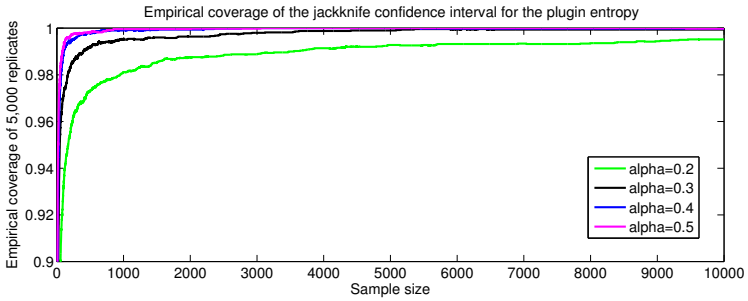

Fig. 1. Coverage of asymptotic jackknife intervals for the discrete entropy with $K = 4$ and probability vector generated from a Dirichlet distribution parametrized by α. Smaller α corresponds to more peaky distributions and less accurate intervals. In all cases, the coverage probability quickly rises above 0.95.

4 Proofs

In this section, we provide the proofs of the corollaries to Theorem 1 in Section 3. We comment that although our proofs are specific to the form of estimators used in the corollaries, the proof ideas provide two separate methods that may be used to verify Assumptions 1 and 2. In particular, both proofs establish Assumption 2 by showing that V_n^J satisfies the bounded difference condition. While the proof of Corollary 1 uses a known Bernstein bound to establish Assumption 1, the proof of Corollary 2 uses an entropy technique due to Boucheron et al. [6] involving bounding the mgf of variance surrogates.

4.1 Proof of Corollary 1

We begin by establishing the Bernstein bound of Assumption 1 from known results. From Theorem 2 of Arcones [2], we have the Bernstein bound

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq \sqrt{\frac{2k^2\zeta_1}{n}\log\left(\frac{4}{\delta}\right)} + \frac{kb_k}{n}\log\left(\frac{4}{\delta}\right)\right) \leq \delta, \tag{10}$$

where $\zeta_1 := \mathrm{Var}[\mathbb{E}[g(X_1, \ldots, X_k) \mid X_1]]$ and $b_k := 2^{k+3}k^{k-1} + \frac{2}{3k^2}$. By Lemma A (p.183) of Serfling [23], we have the bound $\frac{k^2}{n}\zeta_1 \leq \mathrm{Var}[Z]$. Hence, inequality (10), together with the Efron-Stein inequality (6), implies

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq \sqrt{2\left(\frac{n}{n-1}\right)^2\mathbb{E}[V_n^J]\log\left(\frac{4}{\delta}\right)} + \frac{kb_k}{n}\log\left(\frac{4}{\delta}\right)\right) \leq \delta. \tag{11}$$

We now establish Assumption 2 by showing that V_n^J satisfies a bounded difference condition. We have the following formula for V_n^J from Lee [15]:

$$V_n^J = \frac{n-1}{n^2}\binom{n-1}{k}^{-2}\sum_{\ell=0}^{k}(\ell n - k^2)Z_\ell, \tag{12}$$

where $Z_\ell := \sum_{|S\cap T|=\ell} g(X_S)g(X_T)$ is the sum over all pairs of subsets with exactly ℓ indices in common. We may use the formula (12) to establish the bounded difference condition. Indeed, a crude upper bound shows that altering one variable X_j changes the value of each Z_ℓ by at most $2B^2\binom{n}{2k-2}(k!)^2$. Hence, the overall change in V_n^J is upper-bounded by

$$\frac{n-1}{n^2}\binom{n-1}{k}^{-2} \cdot kn \cdot (k+1) \cdot 2B^2\binom{n}{2k-2}(k!)^2.$$

Using the bounds

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \frac{n^k}{k!},$$

we have the rough upper bound

$$\frac{n-1}{n}\left(\frac{k}{n-1}\right)^{2k} 2k(k+1)B^2\frac{n^{2k-2}}{(2k-2)!}(k!)^2 = \frac{2B^2k(k+1)(k!)^2}{n^2(2k-2)!}\left(\frac{n}{n-1}\right)^{2k-1},$$

so Lemma 1.2 of McDiarmid [17] gives Assumption 2. We now apply Theorem 1 to obtain the desired result.

4.2 Proof of Corollary 2

In the case of the discrete entropy, we employ more advanced machinery to establish the assumptions. Following Boucheron et al. [6], define

$$V_+ := \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})^2\mathbb{I}\{Z > Z^{(i)}\} \mid X_1, \ldots, X_n\right], \tag{13a}$$

$$V_- := \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})^2\mathbb{I}\{Z < Z^{(i)}\} \mid X_1, \ldots, X_n\right], \tag{13b}$$

where $Z^{(i)}$ denotes the random variable obtained by replacing X_i with an independent copy X_i'. We may verify via a Hoeffding decomposition [10] that

$$\mathbb{E}[V_+] = \mathbb{E}[V_-] = \left(\frac{n}{n-1}\right)^2 \mathbb{E}[V_n^J].$$

We then use the following lemma:

Lemma 1. *Suppose V_+ and V_- satisfy the mgf bounds*

$$\log\mathbb{E}[\exp(\lambda'(V_+ - \mathbb{E}[V_+] - f_0(n)))],\ \log\mathbb{E}[\exp(\lambda'(V_- - \mathbb{E}[V_-] - f_0(n)))] \le \lambda'^2 \cdot \frac{b_1^2}{2n^3}.$$

Then Assumption 1 holds with $c_1 = \frac{5}{2}$, $c_2 = 2$, $f_1(n) = \frac{b_1^2}{2n^3}$, and $f_2(n) = 0$.

Proof. Consider $\theta > 0$ and $\lambda \in \left(0, \frac{1}{\theta}\right)$. Setting $\lambda' = \frac{\lambda}{\theta}$ and using Theorem 2 of Boucheron et al. [6], we then have

$$\log\mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \le \frac{\lambda\theta}{1-\lambda\theta}\left\{\frac{\lambda}{\theta}\cdot\mathbb{E}[V_+] + \frac{\lambda^2}{\theta^2}\cdot\frac{b_1^2}{2n^3}\right\}.$$

Then by a Chernoff bound and the fact that $\mathbb{E}[V_+] + f_0(n) \le \bar{V}$ by assumption,

$$\mathbb{P}(Z - \mathbb{E}[Z] \ge t) \le \exp\left\{-\lambda t + \frac{\lambda\theta}{1-\lambda\theta}\left(\frac{\lambda}{\theta}\cdot\bar{V} + \frac{\lambda^2}{\theta^2}\cdot\frac{b_1^2}{2n^3}\right)\right\}.$$

Setting $V' = \bar{V} + \frac{b_1^2\log(2/\delta)}{2n^3\bar{V}}$, $\theta = \sqrt{V'/\log(2/\delta)}$, and $\lambda = \frac{1}{\theta}\left(1 - \left(1 + \frac{t\theta}{V'}\right)^{-1/2}\right)$, we have

$$\frac{\lambda}{\theta} \le \frac{1}{\theta^2} = \frac{\log(2/\delta)}{V'} \le \frac{\log(2/\delta)}{\bar{V}},$$

so

$$\mathbb{P}(Z - \mathbb{E}(Z) \geq t) \leq \exp\left\{-\lambda t + \frac{\lambda\theta}{1-\lambda\theta} \cdot \frac{\lambda}{\theta}\left(\overline{V} + \frac{b_1^2 \log(2/\delta)}{2n^3\overline{V}}\right)\right\}$$
$$= \exp\left\{-\lambda t + \frac{\lambda^2}{1-\lambda\theta} V'\right\}.$$

Finally, by a bit of algebra (cf. Lemma 11 of [6]), the last quantity is bounded by $\exp\left(\frac{-t^2}{2(2V'+t\theta/3)}\right)$, and the choice $t = \frac{5}{2}\sqrt{V' \log\left(\frac{2}{\delta}\right)}$ yields the probability bound $\frac{\delta}{2}$, since $1 - \lambda\theta = \left(1 + \frac{5}{2}\right)^{-1/2}$ and

$$t^2 \Big/ \left(4V' + \frac{2t\theta}{3}\right) = \frac{25}{4} \log\left(\frac{2}{\delta}\right) \Big/ \left(4 + \frac{5}{3}\right) \geq \log\left(\frac{2}{\delta}\right).$$

Repeating the argument for V_- and combining tail bounds yields the inequality.

We now establish an mgf bound for V_+; the argument for V_- is nearly identical. For $1 \leq k \leq K$, let $Y_k := n\widehat{p}_k$. Let $h_n(x) = -\left(\frac{x}{n}\right)\log\left(\frac{x}{n}\right)$, and define

$$\Delta_{k,k'}(Z) := \Big(h_n(Y_k) - h_n(Y_k - 1)\Big) + \Big(h_n(Y_{k'}) - h_n(Y_{k'} + 1)\Big),$$

the difference incurred on the statistic Z by moving a single observation from bin k to bin k'. Using this notation, we have

$$V_+ = \sum_k Y_k \sum_{k'} p_{k'} \Delta_{k,k'}^2(Z) \mathbb{I}\{\Delta_{k,k'}(Z) > 0\}. \tag{14}$$

We consequently define the plugin estimator

$$V_n^{PI} = \frac{1}{n} \sum_{k=1}^{K} Y_k \sum_{k'=1}^{K} Y_{k'} \Delta_{k,k'}^2(Z) \mathbb{I}\{\Delta_{k,k'}(Z) > 0\}, \tag{15}$$

which does not depend on the unknown proportions $\{p_k\}_{k=1}^K$.

We first show that V_n^{PI} satisfies a bounded difference condition with parameter $\frac{16\log^2 n + 32\log n + 64 + \frac{8}{n}\log^2 n}{n^2}$. For $k_1 \neq k_2$, note that

$$|\Delta_{k_1,k_2}(V_n^{PI})| \leq \frac{1}{n} \sum_{k,k'} |\Delta_{k_1,k_2}(Y_k Y_{k'} \Delta_{k,k'}^2(Z) \mathbb{I}\{\Delta_{k,k'}(Z) > 0\})|.$$

We now use the fact that

$$|xy - (x - \delta_1)(y - \delta_2)| \leq |\delta_1(y - \delta_2)| + |\delta_2 x|,$$

with

$$x = Y_k Y_{k'}, \qquad y = \Delta_{k,k'}^2(Z) \mathbb{I}\{\Delta_{k,k'}(Z) > 0\},$$

and

$$\delta_1 = \Delta_{k_1,k_2}(Y_k Y_{k'}), \qquad \delta_2 = \Delta_{k_1,k_2}\Big(\Delta^2_{k,k'}(Z)\mathbb{I}\left\{\Delta_{k,k'}(Z) > 0\right\}\Big).$$

It is easy to check that $|\Delta_{k,k'}(Z)| \leq \frac{2\log n}{n}$ using the mean value theorem on h_n, so $|y - \delta_2| \leq \frac{4\log^2 n}{n^2}$. Furthermore, a small calculation shows that

$$|\delta_1| \leq Y_k \mathbb{I}\left\{k' \in \{k_1, k_2\}\right\} + Y_{k'}\mathbb{I}\left\{k \in \{k_1, k_2\}\right\} + \mathbb{I}\left\{k \in \{k_1, k_2\}\right\}\mathbb{I}\left\{k' \in \{k_1, k_2\}\right\}.$$

To bound $|\delta_2|$, note that if the indicator is unaltered by the transition $k_1 \to k_2$,

$$|\Delta_{k_1,k_2}\left(\Delta^2_{k,k'}(Z)\mathbb{I}\left\{\Delta_{k,k'}(Z) > 0\right\}\right)| \leq |\Delta_{k_1,k_2}\left(\Delta^2_{k,k'}(Z)\right)|, \tag{16}$$

and if the value of the indicator does change, we have

$$|\Delta_{k,k'}(Z)| \leq |\Delta_{k_1,k_2}(\Delta_{k,k'}(Z))|,$$

so combined with the straightforward bound

$$|\Delta_{k_1,k_2}\left(\Delta^2_{k,k'}(Z)\mathbb{I}\left\{\Delta_{k,k'}(Z) > 0\right\}\right)| \leq |\Delta^2_{k,k'}(Z)| + |\Delta_{k_1,k_2}(\Delta^2_{k,k'}(Z))|,$$

we obtain

$$|\Delta_{k_1,k_2}\left(\Delta^2_{k,k'}(Z)\mathbb{I}\left\{\Delta_{k,k'}(Z) > 0\right\}\right)| \leq |\Delta^2_{k_1,k_2}(\Delta_{k,k'}(Z))| + |\Delta_{k_1,k_2}(\Delta^2_{k,k'}(Z))|,$$

which subsumes inequality (16).

From the simple bound $|(x+\epsilon)^2 - x^2| \leq 2|x\epsilon| + \epsilon^2$, we have

$$\begin{aligned}|\Delta_{k_1,k_2}(\Delta^2_{k,k'}(Z))| &\leq 2|\Delta_{k,k'}(Z)| \cdot |\Delta_{k_1,k_2}(\Delta_{k,k'}(Z))| + |\Delta^2_{k_1,k_2}(\Delta_{k,k'}(Z))| \\ &\leq \frac{4\log n}{n} \cdot |\Delta_{k_1,k_2}(\Delta_{k,k'}(Z))| + |\Delta^2_{k_1,k_2}(\Delta_{k,k'}(Z))|.\end{aligned}$$

Furthermore,

$$|\Delta_{k_1,k_2}(\Delta_{k,k'}(Z))| \leq \left|\Delta_{k_1,k_2}\Big(h_n(Y_k) - h_n(Y_k - 1)\Big)\right| + \left|\Delta_{k_1,k_2}\Big(h_n(Y_{k'}) - h_n(Y_{k'} + 1)\Big)\right|.$$

Note that if $k \notin \{k_1, k_2\}$, the first term is 0; and if $k' \notin \{k_1, k_2\}$, the second term is 0. Suppose $k = k_1$. Then the first term becomes

$$\left|\Big(h_n(Y_{k_1}) - h_n(Y_{k_1} - 1)\Big) + \Big(h_n(Y_{k_1} - 2) - h_n(Y_{k_1} - 1)\Big)\right|,$$

which may be bounded as

$$\left|\frac{1}{n}\log\left(\frac{n}{Y'_{k_1}}\right) - \frac{1}{n}\log\left(\frac{n}{Y''_{k_1}}\right)\right| = \frac{1}{n}\left|\log\left(\frac{Y''_{k_1}}{Y'_{k_1}}\right)\right| \leq \frac{1}{n}\left(\frac{Y''_{k_1}}{Y'_{k_1}} - 1\right) \leq \frac{2}{nY_{k_1}},$$

for $Y'_{k_1} \in [Y_{k_1} - 2, Y_{k_1} - 1]$ and $Y''_{k_1} \in [Y_{k_1} - 1, Y_{k_1}]$, using the mean value theorem. The other cases may be considered similarly, implying that

$$|\Delta_{k_1,k_2}(\Delta_{k,k'}(Z))| \leq \frac{2}{nY_k} \cdot \mathbb{I}\left\{k \in \{k_1, k_2\}\right\} + \frac{2}{nY_{k'}} \cdot \mathbb{I}\left\{k' \in \{k_1, k_2\}\right\} \leq \frac{4}{n}.$$

Altogether, we conclude that

$$|\delta_2| \leq \left(\frac{8}{n} + \frac{4\log n}{n}\right) \cdot \left(\frac{2}{nY_k} \cdot \mathbb{I}\left\{k \in \{k_1, k_2\}\right\} + \frac{2}{nY_{k'}} \cdot \mathbb{I}\left\{k' \in \{k_1, k_2\}\right\}\right),$$

and summing up, we obtain

$$\begin{aligned}
|\Delta_{k_1,k_2}(V_n^{PI})| &\leq \frac{4\log^2 n}{n^2}\left(4 + \frac{2}{n}\right) + \left(\frac{16}{n^3} + \frac{8\log n}{n^3}\right) \sum_{k,k'} \left(Y_{k'}\mathbb{I}\left\{k \in \{k_1, k_2\}\right\} + Y_k\mathbb{I}\left\{k' \in \{k_1, k_2\}\right\}\right) \\
&\leq \frac{16\log^2 n}{n^2} + \frac{8\log^2 n}{n^3} + \frac{64}{n^2} + \frac{32\log n}{n^2} \\
&\leq \frac{16\log^2 n + 32\log n + \frac{8}{n}\log^2 n}{n^2}.
\end{aligned}$$

Hence, we have the mgf bound

$$\mathbb{E}[\exp(2\lambda V_n^{PI})] \leq \exp\left(2\lambda\mathbb{E}(V_n^{PI}) + \lambda^2 \cdot \frac{c_n^2}{n^3}\right),$$

where we define $c_n := 16\log^2 n + 32\log n + 64 + \frac{8}{n}\log^2 n$.

However, we actually want control on the mgf of V_+. Note that

$$V_+ - V_n^{PI} = \sum_{k'=1}^{K}\left(p_{k'} - \frac{Y_{k'}}{n}\right) \underbrace{\sum_{k=1}^{K} Y_k \Delta_{k,k'}^2(Z)\mathbb{I}\left\{\Delta_{k,k'}(Z) > 0\right\}}_{W_{k,k'}}, \tag{17}$$

with $0 \leq W_{k,k'} \leq \frac{4\log^2 n}{n}$. By convexity of the exponential, we then have

$$\begin{aligned}
\exp(\lambda(V_+ - V_n^{PI})) &\leq \frac{1}{K}\sum_{k'=1}^{K} \exp\left(K\lambda\left(p_{k'} - \frac{Y_{k'}}{n}\right)W_{k,k'}\right) \\
&\leq \frac{1}{K}\sum_{k'=1}^{K} \exp\left(\lambda \cdot \frac{4K\log^2 n}{n} \cdot \left|p_{k'} - \frac{Y_{k'}}{n}\right|\right).
\end{aligned}$$

By the easily verified mgf bound

$$\mathbb{E}[\exp(\lambda|X|)] \leq \exp(\lambda\mathbb{E}|X| + 2n\lambda^2),$$

for $X \sim Bin(n, p)$, we have

$$\mathbb{E}\left[\exp\left(\frac{4K\lambda\log^2 n}{n} \cdot \left|p_{k'} - \frac{Y_{k'}}{n}\right|\right)\right] \leq \exp\left(\frac{4K\lambda\log^2 n}{n} \cdot \mathbb{E}\left|p_{k'} - \frac{Y_{k'}}{n}\right| + 2\lambda^2 \cdot \frac{16K^2\log^4 n}{n^3}\right)$$

for all k'. Hence,

$$\mathbb{E}[\exp(\lambda(V_+ - V_n^{PI}))] \leq \exp\left(\lambda \cdot \frac{4K\log^2 n}{n} \cdot \mu_{k'} + \lambda^2 \cdot \frac{32K^2\log^4 n}{n^3}\right),$$

where $\mu_{k'} := \mathbb{E}\left|p_{k'} - \frac{Y_{k'}}{n}\right|$. By Cauchy-Schwarz, we then have

$$\begin{aligned}\mathbb{E}[\exp(\lambda V_+)] &\leq \mathbb{E}[\exp(2\lambda V_n^{PI})]^{1/2}\mathbb{E}[\exp(2\lambda(V_+ - V_n^{PI}))]^{1/2} \\ &\leq \exp\left(\lambda\mathbb{E}(V_n^{PI}) + \lambda^2 \cdot \frac{c_n^2}{2n^3}\right)\exp\left(\lambda \cdot \frac{4K\log^2 n}{n} \cdot \mu_{k'} + \lambda^2 \cdot \frac{64K^2\log^4 n}{n^3}\right) \\ &= \exp\left(\lambda\left(\mathbb{E}(V_n^{PI}) + \frac{4K\log^2 n}{n} \cdot \mu_{k'}\right) + \lambda^2\left(\frac{c_n^2}{2n^3} + \frac{64K^2\log^4 n}{n^3}\right)\right).\end{aligned}$$

Finally, note that by equation (17), we have

$$|\mathbb{E}(V_+) - \mathbb{E}(V_n^{PI})| \leq \frac{4\log^2 n}{n}\sum_{k'}\mathbb{E}\left|p_{k'} - \frac{Y'_k}{n}\right| \leq \frac{4K\log^2 n}{n} \cdot \mu_{k'},$$

and

$$\mu_{k'} \leq \frac{1}{n}\operatorname{Var}(Bin(n, p_{k'}))^{1/2} \leq \frac{1}{n} \cdot \sqrt{\frac{n}{4}} = \frac{1}{2\sqrt{n}}.$$

It follows that

$$\mathbb{E}[\exp(\lambda V_+)] \leq \exp\left(\lambda\left(\mathbb{E}(V_+) + \frac{4K\log^2 n}{n^{3/2}}\right) + \lambda^2\left(\frac{2c_n^2}{n^3} + \frac{64K^2\log^4 n}{n^3}\right)\right).$$

An identical argument establishes the analogous mgf bound for V_-. We then apply Lemma 1 with $f_0(n) = \frac{4K\log^2 n}{n^{3/2}}$ to obtain Assumption 1.

Finally, note that the same argument used to establish a bounded difference condition for V_n^{PI} also shows that V_n^J satisfies a bounded difference inequality with parameter $\frac{n-1}{2n} \cdot \frac{c_{n-1}}{(n-1)^2}$, since we may write

$$\begin{aligned}V_n^J &= \frac{n-1}{n^2}\sum_{i\neq j}\left(Z_{(i)} - Z_{(j)}\right)^2 \\ &= \frac{n-1}{2n^2}\sum_{k,k'} Y_k Y_{k'}\Big([h_{n-1}(Y_{k'}) - h_{n-1}(Y_{k'} - 1)] + [h_{n-1}(Y_k - 1) - h_{n-1}(Y_k)]\Big)^2,\end{aligned}$$

analogous to equation (15). Assumption 2 follows from bounded differences [17].

5 Experiments

Here, we describe the results of experiments we performed to check the validity of our theoretical results. We report the results of simulation races for the discrete entropy estimator on M categorical probability vectors from a Dirichlet distribution over $K \in \{4, 40\}$ categories and concentration parameters chosen as a uniform random vector of K elements, multiplied by a constant α. A straightforward application of the bounded difference inequality provides a Hoeffding inequality of width $\sqrt{\frac{2\log^2 n}{n}\log\left(\frac{2}{\delta}\right)}$ [1]. We ran racing algorithms with both Hoeffding and asymptotic jackknife intervals (9) for comparison, using $\delta = 0.05$. The speedup achieved by each algorithm is the ratio between the number of samples processed over all arms during the race and the maximum number MN. Table 1 summarizes results numerically and Figure 2 provides a visualization.

α	M	K	N	Hoeffding	Jackknife
0.1	10	4	10000	1.00±0.00	20.06±22.67
0.5	10	4	10000	1.01±0.02	6.38±4.50
1.0	10	4	10000	1.00±0.00	4.38±2.72
5.0	10	4	10000	1.00±0.00	2.23±1.42
0.1	10	40	10000	1.03±0.04	3.59±3.94
0.5	10	40	10000	1.00±0.00	2.24±1.03
1.0	10	40	10000	1.00±0.00	1.42±0.25
5.0	10	40	10000	1.00±0.00	1.04±0.03

Table 1. Simulation results for racing with discrete entropy. We report the mean and standard deviation over 10 runs. The speedup is defined as the ratio between the number of sample evaluations used and the maximum MN required without racing.

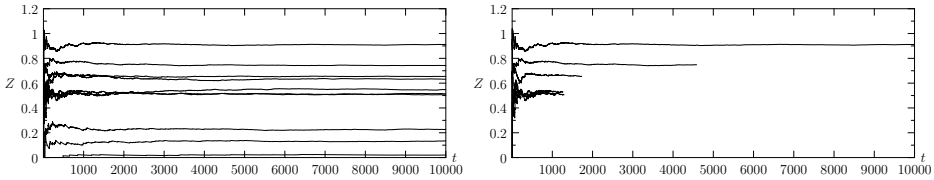

Fig. 2. Entropy race with Hoeffding and asymptotic jackknife confidence intervals ($\alpha = 0.5$, $K = 4$, $M = 10$). As shown in panel (a), the Hoeffding intervals are too conservative to eliminate any of the 10 arms, even after 10,000 evaluations. In contrast, the Bernstein race (panel (b)) terminates after 4,586 observations.

6 Conclusion

We proposed a generalization of racing algorithms to much broader classes of statistics than those previously considered in the literature. Our novel method of constructing empirical Bernstein bounds based on a jackknife estimate of variance has been shown to be theoretically rigorous in a variety of settings, and we have also shown through empirical simulations that our asymptotic Bernstein bounds lead to massive speedups in practice. We expect that similar types of arguments used to establish fast concentration of the jackknife estimate of variance could be used to prove the validity of asymptotic jackknife intervals for many relevant statistics. Even in settings where theoretical results are hard to derive, however, our method provides a general technique for producing empirical Bernstein intervals, which may be extremely useful for the practitioner.

Acknowledgments. We would like to thank Jamie Shotton for general advice regarding decision tree learning. We also thank Yevgeny Seldin, Aaditya Ramdas, and Richard Samworth for helpful discussions.

References

1. Antos, A., Kontoyiannis, I.: Convergence properties of functional estimates for discrete distributions. Random Structures and Algorithms 19:3-4, 163–193. John Wiley & Sons, New York, NY (2001)
2. Arcones, M.: A Bernstein-type inequality for U-statistics and U-processes. Statistics and Probability Letters 22:3, 239–247 (1995)
3. Audibert, J.-Y., Bubeck, S., Munos, R.: Optimization for Machine Learning: Bandit view on noisy optimization (2010)
4. Bernstein, S.N.: The Theory of Probabilities. Gastehizdat Publishing House, Moscow (1946)
5. Berry, D.A., Fristedt, B.: Bandit Problems. Chapman and Hall (1985)
6. Boucheron, S., Lugosi, G., Massart, P.: Concentration inequalities using the entropy method. Annals of Probability 31:3, 1583–1614 (2003)
7. DasGupta, A.: Asymptotic Theory of Statistics and Probability. Springer (2008)
8. Domingos, P., Hulten, G.: Mining high-speed data streams. KDD, pp. 71–80 (2000)
9. Dubhashi, D.P., Panconesi, A.: Concentration of Measure for the Analysis of Randomized Algorithms. Cambridge University Press (2009)
10. Efron, B., Stein, C.: The jackknife estimator of variance. Annals of Statistics 9, pp. 586–596 (1981)
11. Even-Dar, E., Mannor, S., Mansour, Y.: Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. JMLR 7, 1079–1105 (2006)
12. Hoeffding, W.: Probability inequalities for sums of bounded random variables. JASA 58:301, 13–30 (1963)
13. Ikonomovska, E., Gama., J., Zenko., B., Dzeroski, S.: Speeding up Hoeffding-based regression trees with options. Proceedings of ICML, pp. 537–544 (2011)
14. Jin, R., Agrawal, G.: Efficient decision tree construction on streaming data. Proceedings of the 9th ACM SIGKDD, pp. 571–576. ACM, New York, NY (2003)
15. Lee, A.J.: U-statistics: Theory and Practice. CRC Press (1990)
16. Maron, O., Moore, A.W.: Hoeffding races: Accelerating model selection search for classification and function approximation. NIPS, pp. 59–66 (1993)
17. McDiarmid, C.: On the method of bounded differences. Surveys in Combinatorics 141, 148–188 (1989)
18. Mnih, V., Szepesvári, C., Audibert, J.-Y.: Empirical Bernstein stopping. Proceedings of ICML 307, 672-679. ACM (2008)
19. Paninski, L.: Estimation of entropy and mutual information. Neural Computation 15:6, 1191–1253. MIT Press, Cambridge, MA (2003)
20. Peel, T., Anthoine, S., Ralaivola, L.: Empirical Bernstein inequalities for U-statistics. Advances in NIPS 23, 1903–1911 (2010)
21. Pfahringer, B., Holmes, G., Kirkby, R.: New options for Hoeffding trees. Australian Conference on Artificial Intelligence, pp. 90–99 (2007)
22. Robbins, H.: Some aspects of the sequential design of experiments. Bulletin of the AMS 58, 527–535 (1952)
23. Serfling, R.: Approximation theorems of mathematical statistics. Wiley Series in Probability and Mathematical Statistics. New York, NY (1980)
24. Stahl, F., Gaber, M.M., Bramer, M., Yu, P.S.: Distributed Hoeffding trees for pocket data mining. High-Performance Computing and Simulation, pp. 686-692. IEEE Press (2011)
25. Steele, J.M.: An Efron-Stein inequality for nonsymmetric statistics. The Annals of Statistics 14:2, 753–758 (1986)