Optimal Confidence Sets for the Multinomial Parameter

Matthew L. Malloy, Ardhendu Tripathy and Robert D. Nowak

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517964

***Abstract*—Construction of tight confidence sets and intervals is central to statistical inference and decision making. This paper develops new theory showing minimum average volume confidence sets for categorical data. More precisely, consider an empirical distribution $\widehat{p}$ generated from n iid realizations of a random variable that takes one of k possible values according to an unknown distribution p. This is analogous to a single draw from a multinomial distribution. A confidence set is a subset of the probability simplex that depends on $\widehat{p}$ and contains the unknown p with a specified confidence. This paper shows how one can construct minimum average volume confidence sets. The optimality of the sets translates to improved sample complexity for adaptive machine learning algorithms that rely on confidence sets, regions and intervals.**

I. Introduction

This paper shows an optimal confidence set construction for the parameter of a multinomial distribution. The confidence sets, a generalization of the famous Clopper-Pearson confidence interval for the binomial [2], are optimal in the sense of having minimal average volume in the probability simplex for a prescribed confidence level.

Consider an empirical distribution $\widehat{\boldsymbol{p}}$ generated from n i.i.d. samples of a discrete random variable X that takes one of k values according to an unknown distribution $\boldsymbol{p}$. A confidence set for $\boldsymbol{p}$ is a subset of the k-simplex that depends on $\widehat{\boldsymbol{p}}$, and includes the unknown true distribution $\boldsymbol{p}$ with a specified confidence. More precisely, $\mathcal{C}_\delta(\widehat{\boldsymbol{p}}) \subset \Delta_k$ is a confidence set at confidence level $1-\delta$ if

$$\sup_{\boldsymbol{p}\in\Delta_k} \mathbb{P}_{\boldsymbol{p}}\left(\boldsymbol{p} \notin \mathcal{C}_\delta(\widehat{\boldsymbol{p}})\right) \leq \delta\,, \tag{1}$$

where Δ_k denotes the k-simplex, and $\mathbb{P}_{\boldsymbol{p}}(\cdot)$ is the probability measure under the multinomial parameter $\boldsymbol{p}$.

Construction of tight confidence sets for categorical distributions is a long standing problem dating back nearly a hundred years [2]. The goal is to construct sets that are as small as possible, but still satisfy (1). Broadly speaking, approaches for constructing confidence sets can be classified into: *(i)* approximate methods that fail to guarantee coverage (i.e, (1) fails to hold for all $\boldsymbol{p}$), and *(ii)* methods that succeed in guaranteeing coverage, but have excessive volume – for example, approaches based on Sanov or Hoeffding-Bernstein type inequalities. Recent approaches based on combinations of methods [3] have shown improvement through numerical experiment, but do not provide theoretical guarantees on the volume of the confidence sets. To the best of our knowledge, construction of confidence sets for the multinomial parameter that have minimal volume and guarantee coverage is an open problem.

One construction that has shown promise empirically is the *level-set* approach of [4]. The level-set confidence regions (or confidence sets[1]) are similar to 'exact' and Clopper-Pearson[2] regions [2] as they involve inverting tail probabilities, but are applicable beyond the binomial case, *i.e.*, they are defined for $k > 2$. Clopper-Pearson, exact, and level-set confidence sets are closely related to statistical significance testing; the confidence set defined by these approaches is synonymous with the range of parameters over which the outcome *is not statistically significant at an exact p-value of* $1-\delta$. For a discussion of these relationships in the binomial case, see [6], [5] and references therein.

This paper proves that the *level-set* confidence sets of [4], which are extensions of Clopper-Pearson regions, are optimal in that they have minimal average volume among any confidence set construction. More precisely, when averaged across either *i)* the possible empirical outcomes, or *ii)* a uniform prior on the unknown parameter $\boldsymbol{p}$, the level-set confidence sets have minimal volume among any confidence set construction that satisfies the coverage guarantee. The proof first involves showing that arbitrary confidence sets can be expressed as the inversion of a set mapping. The level-set confidence sets are minimal in this setting by design, and the minimal average volume property follows. As the authors of [4] observe through numerical experiment, the level-set confidence sets have small volume when compared with a variety of other approaches. Indeed this observation is correct; the sets minimize average volume among *any* construction of confidence sets.

Confidence intervals for functionals such as the mean, variance, and median can be derived from confidence sets for the multinomial parameter by finding the range of values assumed by the functional in the confidence set. When compared against other confidence intervals based on e.g. Hoeffding's inequality or the empirical Bernstein bound [7], [8], [9], the method can obtain tighter intervals as it accounts for the shape of the distribution in the simplex. In an extended version of this paper [1] we show that swapping our confidence intervals in place of

M. Malloy and R. Nowak are with the Electrical & Computer Engineering Department at University of Wisconsin-Madison. emails: {mmalloy, rdnowak}@wisc.edu. A. Tripathy is with the Computer Science Department at Missouri University of Science & Technology. email: astripathy@mst.edu

An extended version of this manuscript is available [1].

[1]The phrase *confidence region* and *confidence set* are used interchangeably in literature, although *region* can imply a connected set. As the sets discussed herein may not be connected, we prefer *confidence sets*.

[2]Note that 'exact' and Clopper-Pearson are often used synonymously [5].

the those used in standard best-arm identification algorithms [10], [11] in multi-armed bandits can lead to faster termination.

Direct computation of the minimum volume confidence sets involves enumerating empirical outcomes and computing partial sums. In the small sample regime (e.g., $n = 50, k = 5$) computation of the minimal volume sets is straightforward. As computation scales as $O(n^k)$, this becomes prohibitive for modest k. To aid in computation, we show an outer bound based on the Kullback Leibeler divergence that can be used to accelerate computation. We also note that the large sample regime, where computation is prohibitive, is well-served by traditional confidence sets based on asymptotic statistics.

II. Preliminaries

Let $X = X_1, \ldots, X_n \in \mathcal{X}^n$ be a i.i.d. sample of a categorical random variable where X_i takes one of k possible values from a set of categories $\mathcal{X}$. The empirical distribution of X is the relative proportion of occurrences of each element of $\mathcal{X}$ in X. More precisely, let $\mathcal{X} =: \{x_1, \ldots, x_k\}$ and define $n_i = \sum_{j=1}^{n} \mathbf{1}_{\{X_j = x_i\}}$ for $i = 1, \ldots k$. Then $\widehat{\boldsymbol{p}}(X) = [n_1/n, \ldots, n_k/n] \in \Delta_{k,n}$, where $\Delta_{k,n}$ is the discrete simplex from n samples over k categories:

$$\Delta_{k,n} = \left\{ \widehat{\boldsymbol{p}} \in \{0, 1/n, \ldots, 1\}^k : \sum_i \widehat{p}_i = 1 \right\}.$$

To simplify notation in what follows, we write $\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}})$ as shorthand for $\mathbb{P}_{\boldsymbol{p}}\left(\{X \in \mathcal{X}^n : \widehat{\boldsymbol{p}}(X) = \widehat{\boldsymbol{p}}\}\right)$ where $\mathbb{P}_{\boldsymbol{p}}(\cdot)$ denotes the probability measure under $\boldsymbol{p} \in \Delta_k$ and Δ_k is the k-dimensional probability simplex:

$$\Delta_k = \left\{ \boldsymbol{p} \in [0,1]^k : \sum_i p_i = 1 \right\}.$$

We refer to the powerset of Δ_k as $\mathcal{P}(\Delta_k)$, and likewise, $\mathcal{P}(\Delta_{k,n})$ as the power set of $\Delta_{k,n}$. We also write $\mathbb{P}_{\boldsymbol{p}}(\mathcal{S})$ for $\mathcal{S} \subset \Delta_{k,n}$ as shorthand for $\mathbb{P}_{\boldsymbol{p}}\left(\{X \in \mathcal{X}^n : \widehat{\boldsymbol{p}}(X) \in \mathcal{S}\}\right)$. $\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}})$ is fully characterized by the multinomial distribution with parameter $\boldsymbol{p} \in \Delta_k$:

$$\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}}) = \frac{n!}{(n\widehat{p}_1)! \ldots (n\widehat{p}_k)!} p_1^{n\widehat{p}_1} \cdots p_k^{n\widehat{p}_k}.$$

The parameter $\boldsymbol{p}$ specifies the unknown distribution over $\mathcal{X}$.

The focus of this paper is construction of confidence sets for $\boldsymbol{p}$ from a sample $X_1, \ldots, X_n$. Since $\widehat{\boldsymbol{p}}$ is a sufficient statistic for $X_1, \ldots, X_n$, we focus on construction of confidence sets that are functions of $\widehat{\boldsymbol{p}}$ with no loss of generality.

Definition 1. *Confidence set. Let $\mathcal{C}_\delta(\widehat{\boldsymbol{p}}) : \Delta_{k,n} \to \mathcal{P}(\Delta_k)$ be a set valued function that maps an observed empirical distribution $\widehat{\boldsymbol{p}}$ to a subset of the k-simplex. $\mathcal{C}_\delta(\widehat{\boldsymbol{p}})$ is a* confidence set *at confidence level $1 - \delta$ if (1) holds.*

Observation 1. *Equivalent Characterization via Covering Collections. Let $\mathcal{S}(\boldsymbol{p}) : \Delta_k \to \mathcal{P}(\Delta_{k,n})$ be given as:*

$$\mathcal{S}(\boldsymbol{p}) = \{\widehat{\boldsymbol{p}} \in \Delta_{k,n} : \boldsymbol{p} \in \mathcal{C}_\delta(\widehat{\boldsymbol{p}})\}. \tag{2}$$

Then

$$\boldsymbol{p} \in \mathcal{C}_\delta(\widehat{\boldsymbol{p}}) \Leftrightarrow \widehat{\boldsymbol{p}} \in \mathcal{S}(\boldsymbol{p}) \tag{3}$$

and

$$\mathcal{C}_\delta(\widehat{\boldsymbol{p}}) = \{\boldsymbol{p} \in \Delta_k : \widehat{\boldsymbol{p}} \in \mathcal{S}(\boldsymbol{p})\}. \tag{4}$$

We refer to $\mathcal{S}(\boldsymbol{p})$ as a covering collection [4], and observe that any confidence set construction can be equivalently expressed in terms of its covering collection according to (4). Note that for any valid confidence set, $\mathbb{P}_{\boldsymbol{p}}(\mathcal{S}(\boldsymbol{p})) \geq 1 - \delta$ holds for all $\boldsymbol{p}$, since $\mathbb{P}_{\boldsymbol{p}}\left(\boldsymbol{p} \in \mathcal{C}_\delta(\widehat{\boldsymbol{p}})\right) = \mathbb{P}_{\boldsymbol{p}}\left(\mathcal{S}(\boldsymbol{p})\right)$ by (3).

Next we define the *minimal volume confidence set* constructions, which are termed the *level-set* region in [4]. The sets are defined in terms of their covering collection. We note that construction is different from the definition in [4] to facilitate the main theorem of this paper. We discuss this difference in Section IV.

Definition 2. *Minimal volume confidence set. Let $\mathcal{S}^\star(\boldsymbol{p}) : \Delta_k \to \mathcal{P}(\Delta_{k,n})$ be any set valued function that satisfies*

$$\mathcal{S}^\star(\boldsymbol{p}) = \arg \min_{\{\mathcal{S} \in \mathcal{P}(\Delta_{k,n}) : \mathbb{P}_{\boldsymbol{p}}(\mathcal{S}) \geq 1-\delta\}} |\mathcal{S}| \tag{5}$$

for all $\boldsymbol{p}$. Then the minimal volume confidence set is given as

$$\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}}) := \{\boldsymbol{p} \in \Delta_k : \widehat{\boldsymbol{p}} \in \mathcal{S}^\star(\boldsymbol{p})\}. \tag{6}$$

$\mathcal{S}^\star(\boldsymbol{p})$ is a set valued function, mapping $\boldsymbol{p}$ to a subset of empirical distributions with minimal number of elements among subsets whose probability under $\boldsymbol{p}$ equals or exceeds $1 - \delta$. $\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})$ is the subset of the simplex for which the set valued function $\mathcal{S}^\star(\boldsymbol{p})$ includes the observation $\widehat{\boldsymbol{p}}$.

Note that $\mathcal{S}^\star(\boldsymbol{p})$ is in general not unique, and many subsets of $\Delta_{k,n}$ can have minimal cardinality and sufficient probability. As we develop in what follows, any subset of $\Delta_{k,n}$ that satisfies (5) must have minimal average volume, and thus, *equal* average volume. We discuss this in section IV. Before proceeding, we note that the construction creates confidence sets with sufficient coverage, by definition.

Observation 2. *$\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})$ is a confidence set at level $1 - \delta$ since $\mathbb{P}_{\boldsymbol{p}}\left(\boldsymbol{p} \in \mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})\right) = \mathbb{P}_{\boldsymbol{p}}\left(\mathcal{S}^\star(\boldsymbol{p})\right) \geq 1 - \delta$.*

III. Results

The main result of the paper shows that the confidence set $C^\star_\delta(\widehat{\boldsymbol{p}})$ of Definition 2 have on average minimal volume among all confidence sets at level $1 - \delta$.

Theorem 1. *Let $\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})$ be a confidence set given by Definition 2 and define $\mu(\cdot)$ as the Lebesgue measure on the simplex Δ_k. Then*

$$\sum_{\widehat{\boldsymbol{p}} \in \Delta_{k,n}} \mu(\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})) \leq \sum_{\widehat{\boldsymbol{p}} \in \Delta_{k,n}} \mu(\mathcal{C}_\delta(\widehat{\boldsymbol{p}}))$$

for any confidence set $\mathcal{C}_\delta(\widehat{\boldsymbol{p}})$.

Proof. Note that for any confidence set

$$\sum_{\widehat{\boldsymbol{p}} \in \Delta_{k,n}} \mu(\mathcal{C}_\delta(\widehat{\boldsymbol{p}})) = \int_{\Delta_k} |\mathcal{S}(\boldsymbol{p})| d\boldsymbol{p} \tag{7}$$

since

$$\begin{aligned}\sum_{\widehat{\boldsymbol{p}}\in\Delta_{k,n}}\mu(\mathcal{C}_\delta(\widehat{\boldsymbol{p}})) &= \sum_{\widehat{\boldsymbol{p}}\in\Delta_{k,n}}\int_{\mathcal{C}_\delta(\widehat{\boldsymbol{p}})} d\boldsymbol{p}\\ &= \sum_{\widehat{\boldsymbol{p}}\in\Delta_{k,n}}\int_{\boldsymbol{\Delta}_k}\mathbb{1}_{\{\boldsymbol{p}\in\mathcal{C}_\delta(\widehat{\boldsymbol{p}})\}}d\boldsymbol{p}\\ &= \int_{\boldsymbol{\Delta}_k}\sum_{\widehat{\boldsymbol{p}}\in\Delta_{k,n}}\mathbb{1}_{\{\boldsymbol{p}\in\mathcal{C}_\delta(\widehat{\boldsymbol{p}})\}}d\boldsymbol{p}\\ &= \int_{\boldsymbol{\Delta}_k}|\{\widehat{\boldsymbol{p}}:\boldsymbol{p}\in\mathcal{C}_\delta(\widehat{\boldsymbol{p}})\}|\,d\boldsymbol{p}\\ &= \int_{\boldsymbol{\Delta}_k}|\mathcal{S}(\boldsymbol{p})|d\boldsymbol{p}\end{aligned}$$

where last equality follows from (4). By definition, $|\mathcal{S}(\boldsymbol{p})| \geq |\mathcal{S}^\star(\boldsymbol{p})|$ for all $\boldsymbol{p}$. This implies

$$\int_{\boldsymbol{\Delta}_k}|\mathcal{S}(\boldsymbol{p})|d\boldsymbol{p} \;\geq \int_{\boldsymbol{\Delta}_k}|\mathcal{S}^\star(\boldsymbol{p})|d\boldsymbol{p}. \tag{8}$$

Given that any confidence set construction can be defined in terms of its covering collection according to Observation 1, together with (7), this implies the result. □

Theorem 1 shows that, averaged over empirical distributions, the confidence sets defined in (2) have minimal volume. The main idea of the proof is to count the sum of the Lebesgue measure of the confidence sets in two different ways. The LHS in (7) obtains the sum by adding up the areas of the confidence sets corresponding to each point in $\boldsymbol{\Delta}_{k,n}$. The RHS in (7) obtains the same sum by integrating, over all $\boldsymbol{p}\in\boldsymbol{\Delta}_k$, the count of elements in $\boldsymbol{\Delta}_{k,n}$ that include $\boldsymbol{p}$ in their confidence set (i.e, integrating the size of the covering collection (2) over $\boldsymbol{p}$). Fig. 1 can be used to visualize the steps of the proof. We next show that if the multinomial parameter is chosen with uniform probability over the simplex, then the optimal properties of the set still apply.

Proposition 1. *Let $\boldsymbol{p}$ be drawn uniformly at random from $\boldsymbol{\Delta}_k$ and denote $\mathbb{E}_{\boldsymbol{p}}$ expectation with respect to the multinomial parameter $\boldsymbol{p}$. For $\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})$ given by Def. (2) we have that*

$$\mathbb{E}_{\boldsymbol{p}}\left[\mu(\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}}))\right] \leq \mathbb{E}_{\boldsymbol{p}}\left[\mu(\mathcal{C}_\delta(\widehat{\boldsymbol{p}}))\right],$$

where $\mathcal{C}_\delta(\widehat{\boldsymbol{p}})$ is any confidence set at level $1-\delta$.

Proof. Suppose $\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}}) = 1/|\boldsymbol{\Delta}_{k,n}|$. Then

$$\begin{aligned}\mathbb{E}_{\boldsymbol{p}}[\mu(\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}}))] &= \frac{1}{|\boldsymbol{\Delta}_{k,n}|}\sum_{\widehat{\boldsymbol{p}}\in\Delta_{k,n}}\mu(\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}}))\\ &\leq \frac{1}{|\boldsymbol{\Delta}_{k,n}|}\sum_{\widehat{\boldsymbol{p}}\in\Delta_{k,n}}\mu(\mathcal{C}_\delta(\widehat{\boldsymbol{p}}))\\ &= \mathbb{E}_{\boldsymbol{p}}\left[\mu(\mathcal{C}_\delta(\widehat{\boldsymbol{p}}))\right],\end{aligned}$$

where the inequality is due to Theorem 1. Now we describe why $\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}}) = 1/|\boldsymbol{\Delta}_{k,n}|$. A multinomial parameter drawn uniformly at random in $\boldsymbol{\Delta}_k$ induces a uniform distribution over the set of empirical distributions. This is because the resulting distribution on $\widehat{\boldsymbol{p}}$ is the Dirichlet-Multinomial distribution, or a compound Dirichlet distribution [12] with a uniform Dirichlet. □

As noted in Sec. II, the minimal volume confidence construction is under-specified. In general there are many covering collections $\mathcal{S}^\star(\boldsymbol{p})$, each of which results in equal and minimal volume confidence sets.

A simple way to fully specify the confidence sets is to order the empirical distributions based on their probability under $\boldsymbol{p}$ (with ties broken randomly), and construct $\mathcal{S}^\star(\boldsymbol{p})$ by including the most probable empirical distributions until a mass of $1-\delta$ is obtained. This results in covering collections that satisfy (2) and also have an additional guarantee on their coverage probability. We capture this in the following corollary.

Proposition 2. *For any $\boldsymbol{p}$, let $\widehat{\boldsymbol{p}}_1, \widehat{\boldsymbol{p}}_2 \ldots$ be an ordering of the elements of $\Delta_{k,n}$ such that $\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}}_1) \geq \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}}_2) \geq \ldots$, and let ℓ be the smallest integer that satisfies*

$$\sum_{i=1}^{\ell}\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}}_i) \geq 1-\delta. \tag{9}$$

Define $\mathcal{S}^{\star\star}(\boldsymbol{p}) = \{\widehat{\boldsymbol{p}}_1,\ldots,\widehat{\boldsymbol{p}}_\ell\}$ and $\mathcal{C}^{\star\star}_\delta(\widehat{\boldsymbol{p}}) := \{\boldsymbol{p}\in\Delta_k : \widehat{\boldsymbol{p}}\in\mathcal{S}^{\star\star}(\boldsymbol{p})\}$. Then

$$\mathbb{P}_{\boldsymbol{p}}(\boldsymbol{p}\in\mathcal{C}^{\star\star}_\delta(\widehat{\boldsymbol{p}})) \geq \mathbb{P}_{\boldsymbol{p}}(\boldsymbol{p}\in\mathcal{C}^{\star}_\delta(\widehat{\boldsymbol{p}})) \geq 1-\delta$$

holds for all $\boldsymbol{p}$.

Proof. Since $\mathbb{P}_{\boldsymbol{p}}(\boldsymbol{p}\in\mathcal{C}^{\star\star}_\delta(\widehat{\boldsymbol{p}})) = \mathbb{P}_{\boldsymbol{p}}(\mathcal{S}^{\star\star}(\boldsymbol{p}))$ by the relationship in (3), and since $\mathbb{P}_{\boldsymbol{p}}(\mathcal{S}^{\star\star}(\boldsymbol{p})) \geq \mathbb{P}_{\boldsymbol{p}}(\mathcal{S}^{\star}(\boldsymbol{p}))$ by the ordering above, the proof follows immediately. □

Proposition 2 shows that a particular choice for construction of the covering collection $\mathcal{S}^{\star\star}(\boldsymbol{p})$ also satisfies a secondary optimality property – among all confidence sets that have minimal (and equal) average volume, $\mathcal{C}^{\star\star}_\delta(\widehat{\boldsymbol{p}})$ has maximal coverage probability for all $\boldsymbol{p}$. Several confidence set constructions can have equal average minimal volume. This occurs because the average is taken over the set of possible empirical distributions. Provided the minimal cardinality requirement is employed in the construction, the average volume is constant, but the coverage probability may vary.

Proposition 2 also highlights the difference between the definition of the minimal volume confidence sets defined here, and the level-set construction in [4]. In the level-set construction, equiprobable outcomes are either all included or excluded in the covering collections, which precludes the construction from having minimal average volume in this case.

IV. Discussion and Extensions

A. Relationship to Significance Testing

The confidence sets in this paper and in [4] are closely related to p-values in statistical significance testing. Often, the phrase p-value is used to describe an approximate p-value based on a normal approximation. A more precise interpretation of a p-value can be related to the construction of $\mathcal{C}_\delta(\widehat{\boldsymbol{p}})$.

Definition 3. *p-value. The p-value of an outcome $\widehat{\boldsymbol{p}}$ (under the hypothesis $\boldsymbol{p}$) is:*

$$p(\widehat{\boldsymbol{p}};\boldsymbol{p}) = \sum_{\widehat{\boldsymbol{q}}\in\Delta_{k,n}:\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{q}})\leq\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}})}\mathbb{P}_{\boldsymbol{p}}\left(\widehat{\boldsymbol{q}}\right).$$

A p-value has the following interpretation in statistical significance testing: p is the *probability that the observed outcome or something less probable* occurred under the hypothesis $\boldsymbol{p}$. A small p-value corresponds to a strange outcome under the null, and thus corresponds to rejection of the null hypothesis. The level-set confidence sets described in this paper and in [4] can be stated in terms of covering collection based on p-values: $\mathcal{C}_\delta(\widehat{\boldsymbol{p}}) = \{\boldsymbol{p} : p(\widehat{\boldsymbol{p}}; \boldsymbol{p}) > \delta\}$.

We note that the level-set confidence sets and their expressions herein are closely related to 'exact' confidence sets defined in [13] for the specific case when $k = 2$. The confidence set defined by an exact test is the *range of parameters over which the outcome is not statistically significant at a p-value of* $1 - \delta$. Extending this to the multinomial setting is the essence of the level-set confidence sets.

B. Relationship to Sanov Confidence Sets

Sanov's theorem (Theorem 11.4.1 in [14]) allows us to bound the probability of observing a set of empirical distributions using its Kullback Leibler distance to the data-generating distribution. Since the statement of the theorem involves an infimum over Kullback Leibler distances, we can use it to obtain the following inequality:

$$\mathbb{P}_{\boldsymbol{p}}(\mathrm{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p}) > z) \leq (n+1)^k e^{-nz}$$

which implies

$$\mathbb{P}_{\boldsymbol{p}}\left(\mathrm{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p}) \leq \frac{\log((n+1)^k/\delta)}{n}\right) \geq 1 - \delta$$

where

$$\mathrm{KL}(\boldsymbol{p}, \boldsymbol{p}') := \sum_{i=1}^{k} p_i \log\left(\frac{p_i}{p_i'}\right)$$

is the Kullback Leibler divergence. One can view the previous inequality as a concentration result for the Kullback Leibler divergence between the observed empirical distribution and the true distribution. The work done in [15] has sharpened these types of results in several parameter ranges. For example, when $k \leq e\sqrt[3]{n/8\pi}$, [15] shows that

$$\mathbb{P}_{\boldsymbol{p}}(\mathrm{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p}) > z) \leq 2(k-1)e^{-nz/(k-1)}$$

which implies

$$\mathbb{P}_{\boldsymbol{p}}\left(\mathrm{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p}) \leq (k-1)\frac{\log(2(k-1)/\delta)}{n}\right) \geq 1 - \delta. \quad (10)$$

Thus using Sanov's theorem gives us a choice for a confidence set of level $1 - \delta$. Another approach used by [3] to obtain a confidence set is to obtain bounds on the marginal probabilities $\{p_i : i \in \{1, 2, \ldots, k\}\}$. This can be done as $n\hat{p}_i$ corresponds to n i.i.d. realizations of a Bernoulli random variable having mean as p_i. By allocating δ/k error probability in bounding each of the marginal parameters, we get using the Bernoulli-KL inequality [16] that for each $i \in \{1, 2, \ldots, k\}$

$$\mathbb{P}_{p_i}(\mathrm{KL}([\hat{p}_i, 1-\hat{p}_i], [p_i, 1-p_i]) > z) \leq 2e^{-nz} \quad (11)$$

which implies

$$\mathbb{P}_{\boldsymbol{p}}\left(\bigcap_i \mathrm{KL}([\hat{p}_i, 1-\hat{p}_i], [p_i, 1-p_i]) \leq \frac{\log(2k/\delta)}{n}\right) \geq 1 - \delta.$$

Both (10) and (11) give us valid confidence sets for the multinomial parameter. We plot these sets along with the proposed set in Figure 2 in Sec. IV-D.

C. Computation

Computation of $\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})$ requires enumerating all empirical outcomes and computing partial sums. In our experiments, enumerating and ordering the empirical distributions for $k = 5$ and $n = 50$ and checking membership in $\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})$ completes in around two seconds on a laptop. Regardless, as computation scales as n^k, computation of membership in $\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})$ becomes prohibitive for a modest number of categories. We note that the large sample regime, which is not the focus of the work here, is served well by traditional confidence regions based on asymptotic statistics.

There are a number of ways in which computation of the proposed confidence sets can be accelerated. First, in the numerical experiments, we use the approximate p-values returned by Pearson's χ^2 test to obtain a course estimate of the confidence sets, and refine it using exhaustive computation only when needed.

To further aid in computation, we show an outer bound based on the Kullback Leibler divergence that can be used to accelerate computation of the sets. The bound provides a way to confirm if a particular $\boldsymbol{p}$ is outside $\mathcal{C}^\star_\delta(\widehat{\boldsymbol{p}})$.

Theorem 2. Outer bound. *The following inequality holds:*

$$p(\widehat{\boldsymbol{p}}; \boldsymbol{p}) \leq (n+1)^{2k} \exp\left(-n \ \mathrm{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p})\right)$$

Proof. From [14] (Theorem 11.1.4), we can bound the probability of any empirical distribution $\widehat{\boldsymbol{q}}$ under $\boldsymbol{p}$:

$$\frac{1}{(n+1)^k} \exp\left(-n\mathrm{KL}(\widehat{\boldsymbol{q}}, \boldsymbol{p})\right) \leq \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{q}}) \leq \exp\left(-n\mathrm{KL}(\widehat{\boldsymbol{q}}, \boldsymbol{p})\right). \quad (12)$$

Thus, for any $\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{q}}) \leq \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}})$,

$$\frac{1}{(n+1)^k} \exp\left(-n\mathrm{KL}(\widehat{\boldsymbol{q}}, \boldsymbol{p})\right) \leq \exp\left(-n\mathrm{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p})\right)$$

which implies the following. Let $\mathcal{S} \subset \Delta_{k,n}$ be a set of empirical distributions that satisfies $\mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{q}}) \leq \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}})$ for all $\widehat{\boldsymbol{q}} \in \mathcal{S}$. Then,

$$\min_{\widehat{\boldsymbol{q}} \in \mathcal{S}} \mathrm{KL}(\widehat{\boldsymbol{q}}, \boldsymbol{p}) \geq \mathrm{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p}) - \frac{k}{n}\log(n+1). \quad (13)$$

Next, we require Sanov's Theorem, [14] (Theorem 11.4.1), which states the following. Let $\mathcal{S} \subset \Delta_{k,n}$ be a set of empirical distributions. Then

$$\mathbb{P}_{\boldsymbol{p}}(\mathcal{S}) \leq (n+1)^k \exp\left(-n \min_{\widehat{\boldsymbol{q}} \in \mathcal{S}} \mathrm{KL}(\widehat{\boldsymbol{q}}, \boldsymbol{p})\right). \quad (14)$$

Choosing $\mathcal{S} = \{\widehat{\boldsymbol{q}} \in \Delta_{k,n} : \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{q}}) \leq \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}})\}$ and combining (13) and (14), we conclude

$$p(\widehat{\boldsymbol{p}}; \boldsymbol{p}) = \sum_{\widehat{\boldsymbol{q}} \in \Delta_{k,n} : \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{q}}) \leq \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}})} \mathbb{P}_{\boldsymbol{q}}(\widehat{\boldsymbol{q}}) \leq (n+1)^{2k} e^{(-n\mathrm{KL}(\widehat{\boldsymbol{p}}, \boldsymbol{p}))}.$$

□

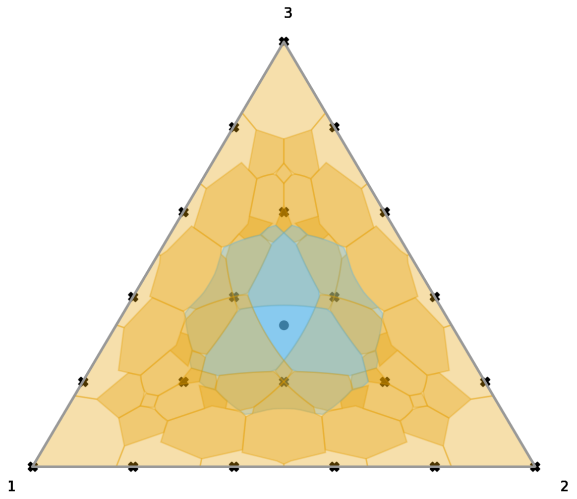

Fig. 1. All confidence sets $\{\mathcal{C}^{\star\star}_{0.7}(\widehat{\boldsymbol{p}}) : \widehat{\boldsymbol{p}} \in \Delta_{3,5}\}$ are shaded over a picture of the three dimensional simplex. The figure depicts the 3-simplex with black crosses indicating the empirical proportions that could be observed in 5 trials (their total number is $\binom{5+3-1}{3-1} = 21$). The number of confidence sets that cover a parameter in Δ_3 vary based on where the parameter lies within the simplex. As an example, the uniform parameter $\boldsymbol{p}_u = [1/3, 1/3, 1/3]$ is shown by a blue dot in the center of the simplex. $\boldsymbol{p}_u$ is covered by the confidence sets of three empirical distributions: $\mathcal{S}^{\star\star}(\boldsymbol{p}_u) = \{[1/5, 2/5, 2/5], [2/5, 1/5, 2/5], [2/5, 2/5, 1/5]\}$. The confidence sets associated with these three empirical distributions are indicated in blue. The main idea in the proof of Thm. 1 is to count the sum of Lebesgue measure of the confidence sets in two ways. The LHS in (7) obtains the sum by adding up the shaded areas corresponding to each point in $\boldsymbol{\Delta}_{3,5}$. The RHS in (7) obtains the same sum by integrating, over all $\boldsymbol{p} \in \boldsymbol{\Delta}_3$, the count of elements in $\boldsymbol{\Delta}_{3,5}$ that include $\boldsymbol{p}$ in their confidence set (i.e, integrating the size of the covering collection over $\boldsymbol{p}$).

Note that the above bound has an additional factor of two in the second term, beyond what arises from directly inverting Sanov's Theorem [14]. This arises from the fact that $\widehat{\boldsymbol{p}}$ is not necessarily the minimal empirical distribution in KL divergence, i.e, it is not necessary true that $\widehat{\boldsymbol{p}}$ equals

$$\arg \min_{\{\widehat{\boldsymbol{q}} \in \Delta_{k,n} : \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{q}}) \leq \mathbb{P}_{\boldsymbol{p}}(\widehat{\boldsymbol{p}})\}} \mathrm{KL}(\widehat{\boldsymbol{q}}, \boldsymbol{p}). \tag{15}$$

Further discussion of computation of exact p-values can be found in [17], [18].

D. Numerical Experiments

We begin with a visualization of the proposed confidence sets $\mathcal{C}^{\star\star}_{\delta}(\widehat{\boldsymbol{p}})$ for a small scale experiment with $n = 5$ samples of a $k = 3$ categorical random variable. Figure 1 shows the confidence sets at level $1 - \delta = 0.3$ for all possible empirical distributions in the discrete simplex $\Delta_{3,5}$ overlaid on top of each other. We also show the uniform parameter $[1/3, 1/3, 1/3] \in \Delta_3$ and indicate the sets that include it at the chosen confidence level, i.e., its covering collection. In this example, from the figure, we can see that $|\mathcal{S}^{\star\star}([1/3, 1/3, 1/3])| = 3$.

Next, in Fig. 2, we show an illustration of the proposed set contrasted with the Sanov and polytope confidence sets of (10) and (11) for different problem parameters. The illustration highlights the significant difference in volume of the proposed set when compared against the Sanov and polytope sets.

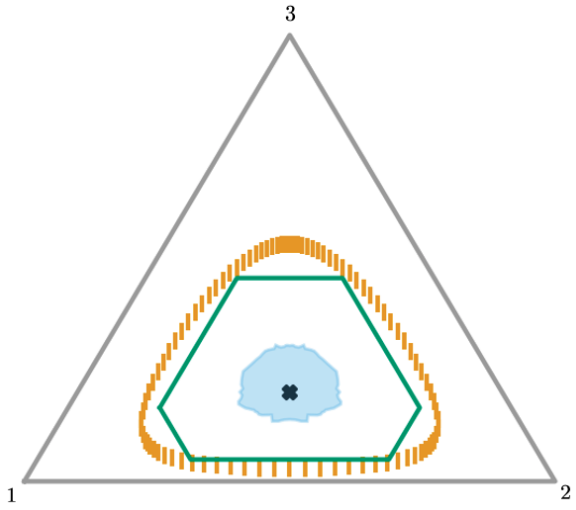

Fig. 2. Proposed confidence set (Proposition 2) shown in blue with the Sanov confidence set (10) in orange and the polytope confidence set (11) in green. The black cross is the observed empirical distribution $\widehat{\boldsymbol{p}} = [6/15, 6/15, 3/15]$ of 15 realizations of a categorical random variable. All confidence sets are shown at 30% confidence level.

V. Summary

Construction of tight confidence sets is a challenging problem with a long history. The problem has seen increased interest, as confidence bounds are central to the analysis and operation of many learning algorithms, especially sequential methods such as active learning, bandit problems, and more generally, reinforcement learning.

This paper shows an optimal construction for confidence sets for the parameter of a multinomial distribution. The sets, termed *minimal volume confidence sets* are optimal in the sense of having minimal volume in the probability simplex, on average, for a prescribed coverage (i.e., confidence). More precisely, when averaged across the possible empirical outcomes or a uniform prior on the unknown parameter $\boldsymbol{p}$, the sets have minimal volume among any confidence set construction that satisfies the coverage guarantee. The *minimal volume confidence sets* (or level-set sets, [4]) are a generalization of the famous Clopper-Pearson confidence interval for the binomial [2]. Clopper-Pearson, exact, and minimum volume confidence sets are closely related to statistical significance testing.

While computation of the sets is straightforward for modest n and k through direct enumeration of the sample space, it can become prohibitive for problems with a large number of categories and samples. To aid in computation, we relate the sets to p-values, and derive a bound based on Kullback Leibler divergence that can be used to accelerate computation, which complements the work in [18]. In this paper we focused our attention on the multinomial parameter due to its wide applicability and importance across reinforcement and adaptive machine learning. We note that the techniques can be extended to more general measure spaces equipped with a conditional probability measure, which we leave for future work.

REFERENCES

[1] M. L. Malloy, A. Tripathy, and R. D. Nowak, "Optimal confidence regions for the multinomial parameter," 2021.

[2] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[3] R. Nowak and E. Tànczos, "Tighter confidence intervals for rating systems," 2019.

[4] D. Chafai and D. Concordet, "Confidence regions for the multinomial parameter with small sample size," *Journal of the American Statistical Association*, vol. 104, no. 487, pp. 1071–1079, 2009.

[5] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, vol. 52, no. 2, pp. 119–126, 1998.

[6] A. Agresti, "Dealing with discreteness: making exact confidence intervals for proportions, differences of proportions, and odds ratios more exact," *Statistical Methods in Medical Research*, vol. 12, no. 1, pp. 3–21, 2003.

[7] V. Mnih, C. Szepesvári, and J.-Y. Audibert, "Empirical Bernstein stopping," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 672–679.

[8] A. Maurer and M. Pontil, "Empirical Bernstein bounds and sample variance penalization," in *COLT 2009-The 22nd Conference on Learning Theory*, 2009.

[9] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876–1902, 2009.

[10] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil'UCB: An optimal exploration algorithm for multi-armed bandits," in *Conference on Learning Theory*, 2014, pp. 423–439.

[11] ——, "On finding the largest mean among many," *arXiv preprint arXiv:1306.3917*, 2013.

[12] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the dirichlet distribution and related processes," *Department of Electrical Engineering, University of Washington, UWEETR-2010-0006*, no. 0006, pp. 1–27, 2010.

[13] C. R. Blyth and H. A. Still, "Binomial confidence intervals," *Journal of the American Statistical Association*, vol. 78, no. 381, pp. 108–116, 1983.

[14] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[15] J. Mardia, J. Jiao, E. Tánczos, R. D. Nowak, and T. Weissman, "Concentration inequalities for the empirical distribution," *arXiv preprint arXiv:1809.06522*, 2018.

[16] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th annual Conference On Learning Theory*, 2011, pp. 359–376.

[17] M. L. Malloy, S. Alfeld, and P. Barford, "Contamination estimation via convex relaxations," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1189–1193.

[18] J. Resin, "A simple algorithm for exact multinomial tests," *arXiv preprint arXiv:2008.12682*, 2020.