Improvements to Sanov and PAC Sublevel-set Bounds for Discrete Random Variables

Michael A. Tope and Joel M. Morris
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County, Catonsville, MD 21250
Email: mtope1@umbc.edu, morris@umbc.edu

***Abstract*—We derive an improvement for probably approximately correct (PAC) sublevel-set bounds for the multinomial distributed discrete random variables. Previous bounds (including Sanov's Theorem) show that the Kullback Leibler (KL) divergence between the empirical probability mass function (*pmf*) and the true PMF converges with rate $O(\log(N)/N)$, where N is the number of independent and identically distributed (i.i.d.) samples used to compute the empirical *pmf*. We interpret the KL divergence as bounding the probability that a multinomial distributed random variable (RV) deviates into a halfspace and construct improved uniform PAC sublevel-set bounds that converge with rates $O(\log(\log(N))/N)$. These results bound the worst case performance for a number of machine learning algorithms. Finally, the 'halfspace bound' methodology suggests further improvements are possible for non-uniform bounds.**

In this paper, we derive an improvement (on the convergence rate) for various Probably Approximately Correct (PAC) bounds (including Sanov's Theorem) for multinomially distributed discrete random variables.

I. INTRODUCTION

Suppose Y is a discrete random variable, whose values (outcomes) are from a finite set $\mathcal{Y} \triangleq \{b_0, b_1, \ldots, b_{|\mathcal{Y}|-1}\}$. Further, suppose Y is distributed according to the probability mass function (*pmf*) $\mathbf{w} \triangleq [w_{b_0}, w_{b_1}, \ldots, w_{b_{|\mathcal{Y}|-1}}]$, where $w_y \triangleq \mathbb{P}\{Y = y\}\ \forall y \in \mathcal{Y}$ or $Y \sim \mathbf{w}$.

Suppose, we have a set $\mathcal{S}$ of samples (outcomes) from N i.i.d. random variables $Y_0, Y_1, \ldots, Y_N$ such that $Y_n \sim \mathbf{w}\ \forall n \in \{0, 1, \ldots, N-1\}$ and $\mathcal{S}_N = \{y_0, y_1, \ldots, y_{N-1}\}$ is the set of outcome values. We call $\mathbf{w}$ the 'generator' *pmf*.

From this set of samples $\mathcal{S}$, we compute the sample (empirical) *pmf* $\hat{\mathbf{w}} = [\hat{w}_{b_0}, \hat{w}_{b_1}, \ldots, \hat{w}_{b_{|\mathcal{Y}|-1}}]$, where

$$\hat{w}_y \triangleq \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{1}_{\{y_n = y\}},\ \forall y \in \mathcal{Y}. \tag{1}$$

Both $\mathbf{w}$ and $\hat{\mathbf{w}}$ lie in the probability space $\mathcal{P}_\mathcal{Y}$, and we think of $\hat{\mathbf{w}}$ as an estimate of $\mathbf{w}$. We define the 'empirical' or 'sample' probability space $\mathcal{P}_N$ as the set of all possible empirical *pmf*s (i.e. $\mathcal{P}_N = \{\hat{\mathbf{w}} \in \mathcal{P}_\mathcal{Y} : \mathbb{P}\{\hat{\mathbf{w}} |\ \mathcal{S}_N\} > 0\}$.

Define a region $\Gamma \subset \mathcal{P}_\mathcal{Y}$. We want to compute or bound the probability that the sample estimate $\hat{\mathbf{w}}$ falls within this region Γ i.e. $\mathbb{P}\{\hat{\mathbf{w}} \in \Gamma\}$. In this paper, we restrict the region Γ to be the interior of a convex level-set that is based on the Kullback-Leilbler (KL) divergence $D(\acute{\mathbf{w}} \,\|\mathbf{w}) \triangleq \sum_{y\in\mathcal{Y}} \acute{w}_y \ln\left(\frac{\acute{w}_y}{w_y}\right)$. That is we want to establish a tight bound of the form

$$\mathbb{P}\{\hat{\mathbf{w}} \in \mathit{\Gamma}_\xi(\mathbf{w})\} \geq 1 - \delta, \tag{2}$$

where the sublevel-set $\mathit{\Gamma}$, which is 'centered' on $\mathbf{w}$ with a 'size' ξ, is defined as

$$\mathit{\Gamma}_\xi(\mathbf{w}) \triangleq \{\acute{\mathbf{w}} \colon D(\acute{\mathbf{w}} \,\|\mathbf{w}) \leq \xi,\ \ \forall \mathbf{w} \in \mathcal{P}_\mathcal{Y}\}. \tag{3}$$

The remainder of this paper is as follows: After briefly describing some previous work, we develop a novel non-asymptotic bound on a halfspace (region) within the probability space. The halfspace bound yields an interpretation of Sanov's Theorem as a sublevel-set bound. In Section V, we develop a transparent methodology that invokes a plurality of halfspace bounds to 'wrap' a convex level-set $\mathit{\Gamma}_\circ \subset \mathit{\Gamma}_\xi(\mathbf{w})$. We compute the convergence rate performance of the 'new' sublevel-set bound. We finish with some conclusions and recommendations.

II. PREVIOUS WORK

Valiant [2] developed the probability approximately correct (PAC) concept, where a PAC bound refers to a bound that holds with a prescribed arbitrarily high probability. Langford [3] developed and outlined the application of PAC-bounds to machine-learning. Seldin and Tishby [4] derived PAC-bounds and PAC-Bayesian bounds for discrete RVs, and this paper significantly leverages that work. Our sublevel-set (PAC) bound is closely related to PAC-Bayesian bounds (see Guedj [5] for a survey and review).

III. HALFSPACE BOUND

We begin with the development of a multinomial halfspace bound (MHB).

Theorem III.1 *Multinomial Halfspace Bound*
Given the set $\mathcal{S}_N = \{y_0, y_1, \ldots, y_{N-1}\}$ of outcomes from N i.i.d. discrete random variables $Y_n \in \mathcal{Y}$ and $Y_n \sim \mathbf{w}$ for $n = 0, 1, \ldots, N-1$. Let $\hat{\mathbf{w}}$ be the sample (empirical) pmf. When given the halfspace $\mathit{\Lambda}$ (oriented to include the pmf $\mathbf{w}^* \in \mathcal{P}_\mathcal{Y}$) defined as

$$\mathit{\Lambda}(\mathbf{w}^*, \mathbf{w}) \triangleq \left\{ \hat{\mathbf{w}} \in \mathcal{P}_\mathcal{Y} : \sum_{y\in\mathcal{Y}} \hat{w}_y \ln\left(\frac{w^*_y}{w_y}\right) \leq \xi \right\} \tag{4}$$

where $\xi \triangleq D(\mathbf{w}^* \,\|\mathbf{w})$, then we have

$$\mathbb{P}\{\hat{\mathbf{w}} \notin \mathit{\Lambda}(\mathbf{w}^*, \mathbf{w})\} \leq \exp(-N D(\mathbf{w}^* \,\|\mathbf{w}))\,. \tag{5}$$

Proof. Theorem III.1

Start with the probability that the empirical *pmf* $\hat{\mathbf{w}}$ (generated from the *pmf* $\mathbf{w}$) does not 'fall' within the halfspace $\Lambda(\mathbf{w}^*, \mathbf{w})$, that is

$$\mathbb{P}\{\hat{\mathbf{w}} \notin \Lambda\} = \sum_{\hat{\mathbf{w}} \notin \Lambda} \binom{N}{N\hat{\mathbf{w}}} \prod_{x \in \mathcal{X}} {w_y}^{N\hat{w}_y}. \tag{6}$$

We multiple and divide by $\left(\frac{w_y^*}{w_y}\right)^{Nw_y^*}$ to get

$$= \sum_{\hat{\mathbf{w}} \in \Lambda} \binom{N}{N\hat{\mathbf{w}}} \prod_{y \in \mathcal{Y}} {w_y}^{N\hat{w}_y} \left(\frac{w_y^*}{w_y}\right)^{Nw_y^*} \left(\frac{w_y^*}{w_y}\right)^{-Nw_y^*}. \tag{7}$$

From the definition of the halfspace (eq. 4), we know that $\prod_{y \in \mathcal{Y}} \left(\frac{w_y^*}{w_y}\right)^{Nw_y^*} \leq \prod_{y \in \mathcal{Y}} \left(\frac{w_y^*}{w_y}\right)^{N\hat{w}_y}$ for all $\hat{\mathbf{w}}$ in the halfspace, i.e.

$$\leq \sum_{\hat{\mathbf{w}} \in \Lambda} \binom{N}{N\hat{\mathbf{w}}} \prod_{y \in \mathcal{Y}} {w_y}^{N\hat{w}_y} \left(\frac{w_y^*}{w_y}\right)^{N\hat{w}_x} \left(\frac{w_y^*}{w_y}\right)^{-Nw_y^*}. \tag{8}$$

Canceling the ${w_y}^{N\hat{w}_y}$ term in the numerator and denominator, we get

$$= \sum_{\hat{\mathbf{w}} \in \Lambda} \binom{N}{N\hat{\mathbf{w}}} \prod_{y \in \mathcal{Y}} {w_y^*}^{N\hat{w}_y} \left(\frac{w_y^*}{w_y}\right)^{-Nw_y^*}. \tag{9}$$

We expand the summation to include entire discrete probability space, and get

$$\leq \sum_{\hat{\mathbf{w}} \in \mathcal{P}_{\mathcal{Y}^N}} \binom{N}{N\hat{\mathbf{w}}} \prod_{y \in \mathcal{Y}} {w_y^*}^{N\hat{w}_y} \left(\frac{w_y^*}{w_y}\right)^{-Nw_y^*}. \tag{10}$$

Rearranging into two products yields

$$= \sum_{\hat{\mathbf{w}} \in \mathcal{P}_N} \binom{N}{N\hat{\mathbf{w}}} \prod_{y \in \mathcal{Y}} {w_y^*}^{N\hat{w}_y} \left(\prod_{y' \in \mathcal{Y}} \left(\frac{w_{y'}^*}{w_{y'}}\right)^{-Nw_{y'}^*} \right). \tag{11}$$

Then, recognizing the multinomial *pmf* and that the summation over the discrete probability space is equal to one, i.e. $\sum_{\hat{\mathbf{w}} \in \mathcal{P}_N} \binom{N}{N\hat{\mathbf{w}}} \prod_{y \in \mathcal{Y}} {w_y^*}^{N\hat{w}_y} = 1$, we get

$$= \prod_{y' \in \mathcal{Y}} \left(\frac{w_{y'}^*}{w_{y'}}\right)^{-Nw_{y'}^*}. \tag{12}$$

We insert the exponentiation of the log function to render the product as a summation, i.e.

$$= \exp\left(\sum_{y' \in \mathcal{Y}} -Nw_{y'}^* \ln\left(\frac{w_{y'}^*}{w_{y'}}\right) \right). \tag{13}$$

Finally, we recognize the definition of KL divergence (to complete the proof),

$$= \exp(-ND(\mathbf{w}^* \| \mathbf{w})). \tag{14}$$

□

This halfspace bound is non-asymptotic (holds for any given number of samples $N \in [1, \infty)$). The MHB has the same form as a well-known KL divergence bound for binary (binomial) RVs [3].

Consider the halfspace $\Lambda(\mathbf{w}^*, \mathbf{w})$ and $\hat{\mathbf{w}} \notin \Lambda(\mathbf{w}^*, \mathbf{w})$, then

$$D(\hat{\mathbf{w}} \| \mathbf{w}) = \sum_{y \in \mathcal{Y}} \hat{w}_y \ln\left(\frac{\hat{w}_y}{w_y}\right) \tag{15}$$

$$= \sum_{y \in \mathcal{Y}} \hat{w}_y \ln\left(\frac{\hat{w}_y}{w_y^*}\frac{w_y^*}{w_y}\right) \tag{16}$$

$$= D(\hat{\mathbf{w}} \| \mathbf{w}^*) + \sum_{y \in \mathcal{Y}} \hat{w}_y \ln\left(\frac{w_y^*}{w_y}\right) \tag{17}$$

$$= D(\hat{\mathbf{w}} \| \mathbf{w}^*) + \sum_{y \in \mathcal{Y}} ({w^*}_y + \vec{e}_y) \ln\left(\frac{w_y^*}{w_y}\right) \tag{18}$$

$$= D(\hat{\mathbf{w}} \| \mathbf{w}^*) + D(\mathbf{w}^* \| \mathbf{w}) + \sum_{y \in \mathcal{Y}} \vec{e}_y \ln\left(\frac{w_y^*}{w_y}\right), \tag{19}$$

where $\overrightarrow{\mathbf{e}}$ is an offset ($\sum_{y \in \mathcal{Y}} \vec{e}_y = 0$). If $\hat{\mathbf{w}}$ is in the halfspace $\Lambda(\mathbf{w}^*, \mathbf{w})$, then $\sum_{y \in \mathcal{Y}} \vec{e}_y \log\left(\frac{w_y^*}{w_y}\right) \leq 0$ (by definition, see eq. 4); therefore, we have alternative simple proof of the well-known Pythagorean-like inequality [6]

$$D(\hat{\mathbf{w}} \| \mathbf{w}) \leq D(\hat{\mathbf{w}} \| \mathbf{w}^*) + D(\mathbf{w}^* \| \mathbf{w}). \tag{20}$$

IV. Sanov Bound Revisited

Using the MHB, we can easily prove Sanov's Theorem [1].

Theorem IV.1 *Sanov's Theorem (see [1] section 11.4) Given the set $\mathcal{S}_N = \{y_0, y_1, \ldots, y_{N-1}\}$ of outcomes from N i.i.d. discrete random variables $Y_n \in \mathcal{Y}$ and $Y_n \sim \mathbf{w}$ for $n = 0, 1, \ldots, N-1$. Let $\hat{\mathbf{w}}$ be the empirical pmf of $\mathcal{S}_N$. When given **any** region $\Gamma \subset \mathcal{P}_{\mathcal{Y}}$ and $\mathbf{w}^*$ is the 'closest' pmf among all $\acute{\mathbf{w}} \in \Gamma$ to $\mathbf{w}$ in terms of the KL divergence*

$$\mathbf{w}^* = \arg \min_{\acute{\mathbf{w}} \in \Gamma} D(\acute{\mathbf{w}} \| \mathbf{w}), \tag{21}$$

then we have

$$\mathbb{P}\{\hat{\mathbf{w}} \notin \Gamma\} \leq (N+1)^{|\mathcal{Y}|} \exp(-ND(\mathbf{w}^* \| \mathbf{w})). \tag{22}$$

One difference between Theorem IV.1 and the Sanov theorem (as stated in [1] section 11.4.1) is that here Sanov's Theorem is claimed valid for *any* region (convex or not).

Proof. Theorem IV.1

Find the 'closest' *pmf* $\mathbf{w}^* = \arg \min_{\acute{\mathbf{w}} \in \Gamma} D(\acute{\mathbf{w}} \| \mathbf{w})$ and define $\xi_\perp \triangleq D(\mathbf{w}^* \| \mathbf{w})$. Now, for every possible 'empirical' *pmf* $\hat{\mathbf{w}}$ that is not in the region Γ (suppose $\hat{\mathbf{w}}'$ is one such *pmf*), we construct a ray from $\mathbf{w}$ towards $\hat{\mathbf{w}}'$, that is $ray_1 \triangleq \{\acute{\mathbf{w}} : \acute{\mathbf{w}} = t\hat{\mathbf{w}}' + (1-t)\mathbf{w}\ \forall t \in [0,1]\}$. Then we find the *pmf* ${\mathbf{w}^*}'$ along this ray_1 such that $D\left({\mathbf{w}^*}' \middle\| \mathbf{w}\right) = \xi_\perp$ (which we know exists), and at that position, we place the halfspace $\Lambda\left({\mathbf{w}^*}', \mathbf{w}\right)$. Let $\{\Lambda\}$ be the set of all such halfspaces (one for every $\hat{\mathbf{w}} \notin \Gamma$).

Given the MHB, we know that for each and every $\hat{\mathbf{w}}' \notin \Gamma$,

$$\mathbb{P}\{\hat{\mathbf{w}}' \notin \Lambda\} \leq \mathbb{P}\left\{{\mathbf{w}^*}' \notin \Lambda\right\} \leq \exp(-N\xi_\perp).$$

We invoke the Union Bound to bound the probability of $\hat{\mathbf{w}}$ not being in the region Γ, as

$$\delta_\Gamma \triangleq \mathbb{P}\{\hat{\mathbf{w}} \notin \Gamma\} \leq |\{\Lambda\}| \exp(-N\xi_\perp) \tag{23}$$

and solving for ξ, we get

$$\xi_\perp = D(\mathbf{w}^* \| \mathbf{w}) \leq \frac{\ln(|\{\Lambda\}|) - \ln(\delta_\Gamma)}{N}. \tag{24}$$

We know that there are less than $(N+1)^{|\mathcal{Y}|}$ *pmf*s $\hat{\mathbf{w}}$ in $\mathcal{P}_N$; therefore, we arrive at

$$\xi_\perp = D(\mathbf{w}^* \| \mathbf{w}) \leq \frac{|\mathcal{Y}| \ln(N+1) - \ln(\delta_\Gamma)}{N}, \tag{25}$$

and solving for δ_Γ completes the proof of Theorem IV.1. □

This proof of Sanov's Theorem via the MHB reveals a potential improvement because each halfspace could 'cover' many more than one single 'empirical' *pmf* (all the $\hat{\mathbf{w}}$ *pmf*s not in the halfspace). The main result of this paper is constructing a method to 'trim' down the number of halfspaces to potentially 'tighten' the Sanov bound.

V. Improved Sublevel-set Bound

In this section, we develop a bound for the sublevel-set $\Gamma_\xi(\mathbf{w})$ (see eq. 3). Note: this sublevel-set is the 'worst-case' region regarding Sanov's Theorem (i.e. it requires the greatest number of halfspaces).

Theorem V.1 *Sublevel-set Bound*
Given the set $\mathcal{S}_N = \{y_0, y_1, \ldots, y_{N-1}\}$ of outcomes from N i.i.d. discrete random variables $Y_n \in \mathcal{Y}$ and $Y_n \sim \mathbf{w}$ for $n = 0, 1, \ldots, N-1$. Let $\hat{\mathbf{w}}$ be the empirical pmf of $\mathcal{S}_N$, and select any $\delta_\Gamma \in (0, 1]$, then $\mathbb{P}\{\hat{\mathbf{w}} \notin \Gamma_\xi(\mathbf{w})\} \leq \delta_\Gamma$ for the sublevel-set $\Gamma_\xi(\mathbf{w})$ (see eq. 3) with 'size'

$$\xi \geq \frac{1}{N}\left(\frac{1}{2}\ln(2|\mathcal{Y}|) - \frac{3}{2}\ln\left(\frac{\delta_\Gamma}{2}\right) + |\mathcal{Y}| \ln\left(\log_2(\log_2(N)) + \kappa_1\sqrt{|\mathcal{Y}|} + \log_2(\kappa_2|\mathcal{Y}|) + 2\right)\right) \tag{26}$$

where $\kappa_1 = 2\sqrt{24}\left(1+\sqrt{2}\right)$ and $\kappa_2 = 24$.

As a corollary, solving eq. 26 for δ_Γ and setting $\xi = D(\mathbf{w}^* \| \mathbf{w})$ results in an improved Sanov's Theorem (for large N).

Proof. Theorem V.1

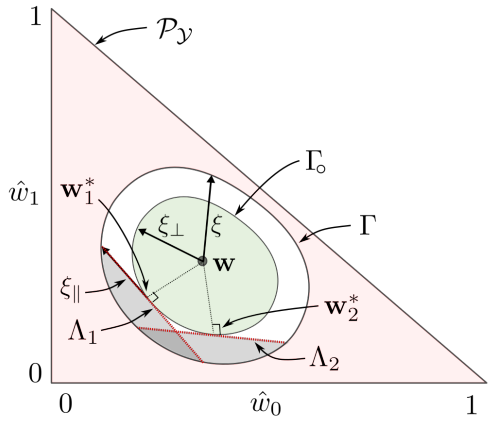

Fig. 1. Level Set Bound Schematic

Fig. 1 depicts the strategy. We want to bound $\mathbb{P}\{\hat{\mathbf{w}} \notin \Gamma_\xi(\mathbf{w})\}$ using a plurality of MHBs such that each halfspace is positioned against (tangent to) an 'inner' level-set $\Gamma_\circ \triangleq \{\acute{\mathbf{w}} : D(\acute{\mathbf{w}} \| \mathbf{w}) = \xi_\perp\}$. We define a 'shingle' as

$$\Phi_k \triangleq \left\{\acute{\mathbf{w}} \notin \Lambda_k(\mathbf{w}_k^*, \mathbf{w}) : D(\acute{\mathbf{w}} \| \mathbf{w}) \leq \xi_\perp + \xi_\parallel\right\} \tag{27}$$

To construct a suitable set of shingles $\{\Phi\}$ to cover all *pmf*s $\hat{\mathbf{w}} \notin \Gamma_\xi(\mathbf{w})$, we need to determine which *pmf*s $\hat{\mathbf{w}}$ each shingle with extent $\xi_\parallel$ can cover.

We will divide the probability space into hyper-boxes Ξ called 'cells.' Each cell's position and dimension is designed such that the 'extent' between any two *pmf*s within the cell is less than $\xi_\parallel$. The 'extent' is based on the KL Divergence; therefore, the cell dimensions (in an Euclidean sense) vary with the position of the cell within the probability space.

Suppose we have constructed a set of suitable cells.

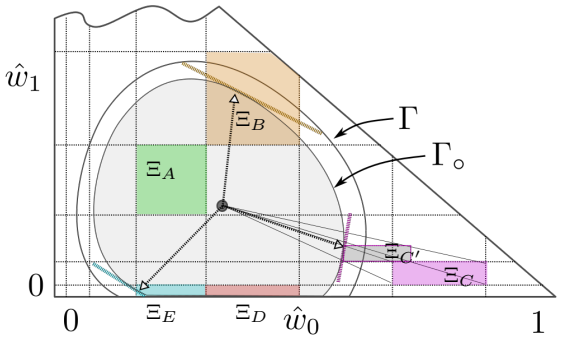

Fig. 2. Cell Cover Illustration

Fig. 2 depicts several situations regarding a cell Ξ and the level-set $\Gamma_\circ$: case I = all *pmf*s $\hat{\mathbf{w}}$ of the cell are inside $\Gamma_\circ$, case II = some *pmf*s $\hat{\mathbf{w}}$ of the cell are on the surface of $\Gamma_\circ$, and case III = all *pmf*s $\hat{\mathbf{w}}$ of the cell are outside $\Gamma_\circ$

The 'green' cell (Ξ_A) is case I, all the *pmf*s $\hat{\mathbf{w}}$ lie within the level-set $\Gamma_\circ$, and we do not require a shingle to cover any of these *pmf*s. The 'red' cell Ξ_D is also case I.

The 'orange' cell Ξ_B is case 2, we can choose any *pmf* $\hat{\mathbf{w}}$ 'point' within Ξ_B that is on the level-set $\Gamma_\circ$ and place the shingle's halfspace on that point. The shingle will 'shade' all the *pmf*s in the cell outside the shingle. The 'blue' cell Ξ_E is also case 2.

The 'purple' cell Ξ_C is case III, where all the *pmf*s in this cell are outside of the $\Gamma_\circ$ level-set. We create a 'scaled' cell $\Xi_{C'}$ by linearly scaling the endpoints of the cell Ξ_C towards the *pmf* $\mathbf{w}$ such that one single point of the scaled cell $\Xi_{C'}$ touches $\Gamma_\circ$. The shingle will have an extent to cover all the *pmf*s in the 'gray' scaled or reduced cell $\Xi_{C'}$. And all *pmf*s in the original Ξ_C lie in the 'shadow' of the $\Xi_{C'}$ cell's shingle.

We don't have to actually construct these shingles as knowing that they exist is sufficient.

The next step is to construct a set of cells to cover the entire probability space.

Given N and ϖ, Algorithm 1 outputs a set of $\{\nu_i\}_{i=0}^{V-1}$ increasing cell boundary values (to apply to each coordinate dimension).

Algorithm 1: Cell Positions/Boundaries

Input: $N \in \mathbb{N}, \varpi \in (0,1)$
Output: set of points $\{\nu_i\}_{i=0}^{V}$

1 $n_c \leftarrow \lceil \log_2(\log_2(N)) \rceil$;
2 $n_o \leftarrow \lceil -\log_2(\varpi) \rceil$;
3 $\nu_0 \leftarrow 0$;
4 $i \leftarrow 1$;
5 **for** $j \in \{0, 1, \ldots, n_c\}$ **do**
6 $\quad \nu_i \leftarrow 2^{-n_o - 2^{n_c - j}}$;
7 $\quad i \leftarrow i + 1$;
8 **end**
9 **for** $k \in \{0, 1, \ldots, n_o\}$ **do**
10 $\quad \nu_i \leftarrow 2^{-n_o + k}$;
11 $\quad i \leftarrow i + 1$;
12 $\quad$ **while** $\nu_{i-1} + 2^{-n_o + \frac{k}{2}} < 2^{-n_o + k + 1}$ **do**
13 $\quad\quad \nu_i \leftarrow 2^{-n_o + \frac{k}{2}}$;
14 $\quad\quad i \leftarrow i + 1$;
15 $\quad$ **end**
16 **end**
17 $\nu_i \leftarrow 1$;

Fig. 3 illustrates the output from Algorithm 1. After selecting n_o such that $2^{-n_o} < \varpi$, the first 'For' loop (line: 5) creates $n_c = \lceil \log_2(\log_2(N)) \rceil$ cell boundary values. Define the y-component chi-square extent as $\chi^2(i) \triangleq \frac{(\nu_{i+1} - \nu_i)^2}{\nu_i}$.

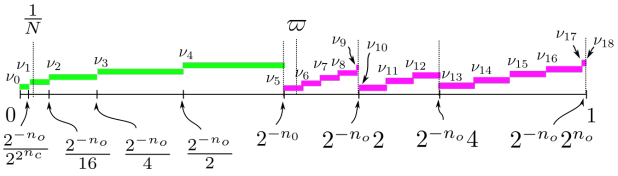

Fig. 3. Output from Algorithm 1

The n_c 'green' cell 'widths' are doubly exponentially increasing, and we have

$$\chi^2(i) = \frac{\left(2^{-n_o - 2^{n_c - i - 1}} - 2^{-n_o - 2^{n_c - i}}\right)^2}{2^{-n_o - 2^{n_c - i}}} \tag{28}$$

$$= 2^{-n_o}\left(1 - 2^{-2^{-n_c + i - 1}}\right)^2 \leq \varpi \;\; \forall i = 1, 2, \ldots, n_c \tag{29}$$

For the 'purple' cell boundary values, the algorithm's second 'For' loop (line: 9) creates $n_o = \lceil -\log_2(\varpi) \rceil$ segments, and each segment is further divided into cell boundary values (via the While loop (line: 12)).

Segment k has $width_k = 2^{-k+1} - 2^{-k}$ for $k = 1, 2, \ldots, n_o$, and 'purple' cell boundaries within a segment are created with constant increment values of $inc_k = \sqrt{2^{-k} 2^{-n_o}}$ ($\nu_j = 2^{-k}$ is the smallest cell boundary value within the segment). And so for all $\nu_j \in \left[2^{-k}, 2^{-k+1}\right]$ (i.e segment k), we have

$$\chi^2{}_y(j) = \frac{\left(\left(\nu_j + 2^{-n_o/2} 2^{-k/2}\right) - \nu_j\right)^2}{\nu_j} \tag{30}$$

$$\leq \frac{\left(2^{-n_o/2} 2^{-k/2}\right)^2}{2^{-k}} = 2^{-n_o} \leq \varpi. \tag{31}$$

So all y-component chi-square extents are bounded by ϖ.

To count the number of 'purple' cell boundary values, start with a bound on the number of cell boundary values per segment

$$n_{values/segment} = \lceil \frac{width_j}{inc_j} \rceil \leq \lceil \frac{2^{-k+1} - 2^{-k}}{2^{-n_o/2} 2^{-j/2}} \rceil \tag{32}$$

$$= \lceil 2^{n_o/2} \frac{1}{2^{-j/2}} \rceil \leq 2^{n_o/2} 2^{-j/2} + 1. \tag{33}$$

Summing over the n_o segments, we get

$$n_{purple} = \sum_{k=1}^{n_o} \left(2^{n_o/2} 2^{-j/2} + 1\right) \tag{34}$$

$$\leq n_o + \sum_{j=1}^{\infty} 2^{n_o/2} 2^{-j/2} = \left(1 + \sqrt{2}\right) 2^{n_o/2} + n_o. \tag{35}$$

So the total number of cell boundary points (component-wise) is

$$n_y = V - 1 \leq n_c + \left(1 + \sqrt{2}\right) 2^{n_o/2} + n_o + 1 \tag{36}$$

$$= \lceil \log_2(\log_2(N)) \rceil + \left(1 + \sqrt{2}\right) 2^{\lceil -\log_2\left(\sqrt{\varpi}\right) \rceil} + \lceil -\log_2(\varpi) \rceil \tag{37}$$

We define a vector $\underline{\mathbf{i}} \triangleq \left[i_0, i_1, \ldots, i_{|\mathcal{Y}|-1}\right]$ to index each cell and construct a set of cell indices $I \triangleq \{\underline{\mathbf{i}} : i_y \in \{0, 1, \ldots, V-2\}\}$. Then we define each cell as

$$\Xi_{\underline{\mathbf{i}}} \triangleq \left\{\acute{\mathbf{w}} : \acute{w}_y \in \left(\nu_{i_y}, \nu_{i_y + 1}\right]\right\} \; \forall y \in \{0, 1, \ldots, |\mathcal{Y}| - 1\} \tag{38}$$

for all $\underline{\mathbf{i}} \in I$, and $\left\{\Xi_{\underline{\mathbf{i}}}\right\}_{\underline{\mathbf{i}} \in I}$ is the set of all cells. The total number of cells ($|\mathcal{Y}|$ dimensions) is

$$\left|\left\{\Xi_{\underline{\mathbf{i}}}\right\}_{\underline{\mathbf{i}} \in I}\right| = (n_y)^{|\mathcal{Y}|}. \tag{39}$$

We assign one halfspace (i.e. shingle) per cell, so the *log* number of halfspaces is bounded (after using $\lceil x \rceil \leq x + 1$) as

$$\ln(|\{\Lambda\}|) \leq |\mathcal{Y}| \ln\Bigg(\log_2(\log_2(N)) + 2\left(1 + \sqrt{2}\right)\frac{1}{\sqrt{\varpi}} - \log_2(\varpi) + 2\Bigg). \tag{40}$$

As 'chi-square divergence' upper bounds 'KL divergence' [6], the worst case extent (KL divergence) across any cell $\Xi_{\underline{\mathbf{i}}}$, is

$$\xi_{\|}\left(\Xi_{\underline{\mathbf{i}}}\right) \leq \chi^2\left(\Xi_{\underline{\mathbf{i}}}\right) \triangleq \sum_{y \in \mathcal{Y}} \frac{\left(\nu_{i_y + 1} - \nu_{i_y}\right)^2}{\nu_{i_y}} \mathbb{1}_{\{i_y \neq 0\}} \tag{41}$$

$$= \sum_{y \in \mathcal{Y}} \chi^2{}_y(i_y) \, \mathbb{1}_{\{i_y \neq 0\}} \leq \varpi \, |\mathcal{Y}| \;\; \forall \underline{\mathbf{i}} \in I \tag{42}$$

For cells with one or more components of their indices with

$i_{y'} = 0$, we know that $\nu_0 = 0$ and $\nu_1 \leq 1/N$ along that y'-dimension, and so all possible $\hat{\mathbf{w}}$ *pmf*s within such a cell have $\hat{w}_{y'} = 0$. We construct a suitable halfspace and shingle (and avoid dividing by zero in calculating $\chi^2\left(\Xi_{\underline{\mathbf{i}}}\right)$) by setting $w_y = 0$ whenever $i_y = 0$.

For a 'scaled' cell (see $\Xi_{C'}$ in Fig. 2), we shift the cell boundary values as $\nu'_{i_y} = \lambda\nu_{i_y} + (1-\lambda)\,\mathbf{w}_y$ for any $\lambda \in [0,1]$, and then

$$\chi^2\left(\Xi_{\underline{\mathbf{i}}}, \mathbf{w}, \lambda\right) = \sum_{y\in\mathcal{Y}} \frac{\left(\nu_{i_y+1}\lambda - \nu_{i_y}\lambda\right)^2}{\nu_{i_y}\lambda + w_y\,(1-\lambda)} \tag{43}$$

$$\leq \sum_{y\in\mathcal{Y}} \frac{\left(\nu_{i_y+1} - \nu_{i_y}\right)^2\lambda^2}{\nu_{i_y}\lambda} = \chi^2\left(\Xi_{\underline{\mathbf{i}}}\right)\lambda, \tag{44}$$

which shows that the extent of the scaled cell is not increased.

The cells are constructed without regard to the *pmf* $\mathbf{w}$; therefore, this sublevel-set bound is uniform over $\mathbf{w}$. One improvement (perhaps) could be to 'tune' Algorithm 1 (customized the cell boundaries) for a specific *pmf* $\mathbf{w}$.

Recall that the level-set $\Gamma_\xi(\mathbf{w})$ must encase the shingles over all the cells, and the KL divergence from the *pmf* $\mathbf{w}$ to the level-set $\Gamma_\xi(\mathbf{w})$ equals the KL divergence from $\mathbf{w}$ to the 'inner' level-set $\Gamma_\circ(\xi_\perp)$ plus the extent of the shingle $\xi_\parallel$.

As N increases $\xi_\perp$ decreases, and the shingles (always touching against $\Gamma_\circ$) move inward towards *pmf* $\mathbf{w}$; however, we also need the cell extents $\xi_\parallel$ to decrease and shrink $\Gamma_\xi(\mathbf{w})$.

To accomplish this, we shall construct a 'bounding box' over the entire set of cells to squeeze (or shrink) every cell towards the *pmf* $\mathbf{w}$. The bounding box 'rules-out' 'distant' areas of the probability space according to the following Chernoff-Hoeffding relative-deviation-about-the-mean concentration inequality (see [7] Appendix A).

If $X_0, X_1, \ldots, X_N$ are i.i.d. RVs with values in $[0,1]$, and $X \triangleq \sum_{x\in\mathcal{X}} X_x$, then for any $\epsilon > 0$

$$\mathbb{P}\left\{X \gtrless (1\pm\epsilon)\,\mathbb{E}\{X\}\right\} < \exp\left(-\epsilon^2\,\mathbb{E}\{X\}/3\right). \tag{45}$$

Define $\delta_1 \triangleq \exp(-\epsilon\,\mathbb{E}\{X\}/3)$ and $\gamma \triangleq \sqrt{\frac{-\log(\delta_1)}{2N}}$. Select one component of the *pmf* $\hat{\mathbf{w}}$ (i.e. $\hat{w}_y$) to be the RV X and $w_y = \mathbb{E}\{X\}$, then after substituting into the 'relative-deviation' concentration inequality, we have

$$\mathbb{P}\left\{\hat{w}_y \leq \max\left(w_y - \gamma\sqrt{6w_y}, 0\right)\right\} \leq \delta_1 \tag{46}$$

$$\mathbb{P}\left\{\hat{w}_y \geq \min\left(w_y + \gamma\sqrt{6w_y}, 1\right)\right\} \leq \delta_1. \tag{47}$$

We define a 'trajectory' with parameter $\gamma \in [0,1]$ that sets the position of each cell boundary value within the bounding box as

$$\nu'(\gamma, w_y, i_y) \triangleq \nu_{i_y}\min\left(w_y + \gamma\sqrt{6w_y}, 1\right) + \left(1-\nu_{i_y}\right)\max\left(w_y - \gamma\sqrt{6w_y}, 0\right). \tag{48}$$

To ensure that $\hat{\mathbf{w}}$ does not lie outside of any of the $2\,|\mathcal{Y}|$ 'sides' of the bounding box with probability greater than δ_{bb}, we need for each concentration inequality to hold (be valid) with probability $\delta_1 = \frac{\delta_{bb}}{2|\mathcal{Y}|}$. Solving the concentration inequality (eq. 45), we get

$$\gamma = \sqrt{\frac{\ln(2\,|\mathcal{Y}|) - \ln(\delta_{bb})}{2N}}\,. \tag{49}$$

Our plan is to set the parameter ϖ of Algorithm 1 such that the extents are within the rate γ^2 for all $\gamma \in (0,1]$.

The y-component-wise chi-square extent of cell $\Xi_{\underline{\mathbf{i}}}$ (along the trajectory) is

$$\chi^2{}_y(\gamma, w_y, i_y) \triangleq \frac{\left(\nu'(\gamma, w_y, i_y) - \nu'(\gamma, w_y, i_y)\right)^2}{\nu'(\gamma, w_y, i_y)}\,. \tag{50}$$

Define the function

$$g(w,\gamma) \triangleq \max_{i_y\in\{1,2,\ldots,V-1\}} \frac{\chi^2{}_y(\gamma, w_y, i_y)}{\gamma^2\frac{\left(\nu_{i_y}-\nu_{i_y}\right)^2}{\nu_{i_y}}}\,. \tag{51}$$

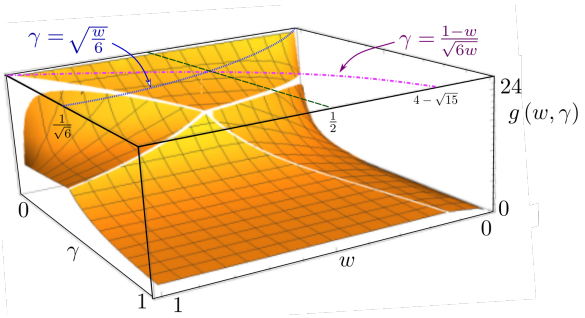

Fig. 4. Trajectory Convergence Bound

Fig. 4 shows a 3D plot of $g(w,\gamma)$. Over all trajectories and cells, we find (skipping the algebra) that

$$G_0 = \max_{\substack{\gamma\in(0,1]\\ w\in(0,1]}} g(w,\gamma) = 24.$$

Recall that the cells were constructed such that $\xi_\parallel\left(\Xi_{\underline{\mathbf{i}}}\right) \leq \chi^2\left(\Xi_{\underline{\mathbf{i}}}\right) \leq \varpi\,|\mathcal{Y}|$ (see eq. 41); therefore, if we set $\varpi = \frac{1}{G_0|\mathcal{Y}|}$, then we have

$$\chi^2{}_y(i_y, w_y, \gamma) \leq \frac{1}{|\mathcal{Y}|}\gamma^2 \quad \forall\gamma \in (0,1]\,. \tag{52}$$

So a bound on the extents of each shingle over all specified trajectories (after inserting eq. 49) is

$$\xi_\parallel = \chi^2\left(\Xi_{\underline{\mathbf{i}}}, \gamma\right) \leq \gamma^2 \tag{53}$$

$$= \frac{1}{2N}\left(\ln(2\,|\mathcal{Y}|) - \ln(\delta_{bb})\right) = O\left(\frac{1}{N}\left(\log(|\mathcal{Y}|)\right)\right), \tag{54}$$

and inserting the selected ϖ value into eq. 40, we get the *log* total number of halfspaces,

$$\ln(|\{\Lambda\}|) \leq |\mathcal{Y}|\ln\left(\log_2(\log_2(N)) + \kappa_1\sqrt{|\mathcal{Y}|} + \log_2(\kappa_2\,|\mathcal{Y}|) + 2\right) \tag{55}$$

where $\kappa_1 = 2\sqrt{24}\left(1+\sqrt{2}\right)$ and $\kappa_2 = 24$.

The probability δ_Γ must be distributed to ensure that each and every halfspace (shingle) is accounted for along with the concentration inequalities that 'size' then bounding box. Define parameter $\varrho \in [0,1]$ and set $\delta_{bb} = (1-\varrho)\,\delta_\Gamma$ and

$\delta_{hs} = \varrho\delta_{\Gamma}$. The δ_{hs} probability is divided evenly over each and every halfspace, and we compute the required 'size' $\xi_{\perp}$ of the level-set $\Gamma_{\circ}$ as

$$\xi_{\perp} = \frac{1}{N}\left(\ln(|\{\Lambda\}|) - \ln(\delta_{hs})\right) \tag{56}$$

$$= \frac{1}{N}\Bigg(|\mathcal{Y}| \ln\Bigg(\log_2(\log_2(N)) + \kappa_1\sqrt{|\mathcal{Y}|} + \ln(2)\kappa_2 |\mathcal{Y}| + 2\Bigg) - \ln(\delta_{hs})\Bigg) \tag{57}$$

where $\kappa_1 = 2\sqrt{24}\left(1+\sqrt{2}\right)$, $\kappa_2 = 24$. Using $\log(x+y) = \log((x/y+1)\,y) \leq x/y + \log(y)$, we get

$$\xi_{\perp} \leq \frac{1}{N}\Bigg(\frac{\sqrt{|\mathcal{Y}|}}{\kappa_1}\log_2(\log_2(N)) + |\mathcal{Y}|\ln\Big(\kappa_1\sqrt{|\mathcal{Y}|} + \log_2(\kappa_2|\mathcal{Y}|) + 2\Big) - \ln(\delta_{hs})\Bigg)$$

$$= O\left(\frac{1}{N}\left(\sqrt{|\mathcal{Y}|}\log(\log(N)) + |\mathcal{Y}|\log(|\mathcal{Y}|)\right)\right), \tag{58}$$

which proves that $\xi_{\perp}$ is within the $O(\log(\log(N))/N)$ convergence rate.

Finally, we set $\vartheta = 1/2$ and sum $\xi_{\perp}$ and $\xi_{\parallel}$ to determine the overall sublevel-set $\Gamma_{\xi}(\mathbf{w})$ such that $\mathbb{P}\{\hat{\mathbf{w}} \notin \Gamma_{\xi}(\mathbf{w}) \leq \delta_{\Gamma}\}$, with $\xi = \xi_{\parallel} + \xi_{\perp}$, i.e.

$$\xi = \frac{1}{N}\Bigg(\frac{1}{2}\ln(2|\mathcal{Y}|) - \frac{3}{2}\ln\left(\frac{\delta_{\Gamma}}{2}\right) + |\mathcal{Y}|\log\Big(\log_2(\log_2(N)) + \kappa_1\sqrt{|\mathcal{Y}|} + \log_2(\kappa_2|\mathcal{Y}|) + 2\Big)\Bigg) \tag{59}$$

□

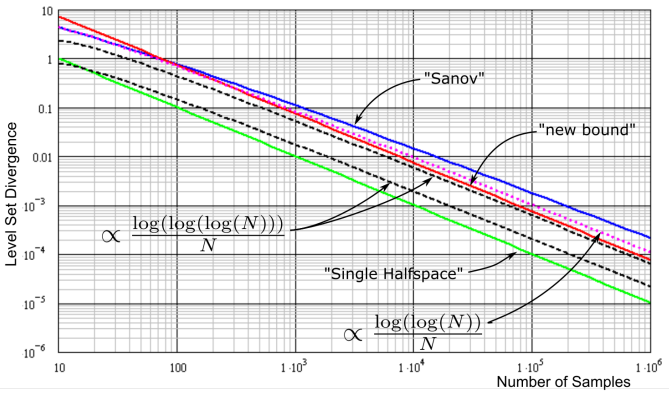

Fig. 5. Convergence Results

Fig. 5 shows an example of Thm. V.1 sublevel-set bound's convergence (for $|\mathcal{Y}| = 10$ and $\delta = 0.01$). The 'Sanov' curve $O(\log(N+1)/N)$ follows eq. 25. The 'single halfspace' curve is $O(1/N)$ for reference. The 'new bound' curve (see eq. 59) lies between a curve with $O(\log(\log(N+1))/N)$ and $O(\log(\log(\log(N+1)))/N)$, and this example demonstrates the new bound's convergence rate as better than $O(\log(\log(N+1))/N)$.

VI. CONCLUSIONS

In summary, we improved the PAC sublevel-set bound's convergence rate (for discrete RVs) from $O(\log(N)/N)$ to $O(\log(\log(N))/N)$ (for large N). Further, we developed a transparent methodology for leveraging a plurality of halfspaces to construct bounds (including the sublevel-set bound). For some practical scenarios or applications, we expect that the number and positions of the halfspaces can be further tuned to tighten the overall bound.

Several improvements are likely possible. For example: (1) in eq. 45 (in favor of simplification), we did not take advantage of an asymmetric concentration that is tighter, (2) the algorithm for computing cell boundary values is sub-optimal because we focused on providing a 'simple' analytical function to count the number of cells, (3) one may be able to drop $|\mathcal{Y}|$ to $|\mathcal{Y}| - 1$ using arguments in [4], and (4) a more efficient algorithm is likely feasible to further trim the number of required halfspace bounds (rather than place a bound at every cell). While these improvements may tighter the bounds, we believe that the the convergence rate (as N increases) will match the same 'big-Oh' order as the bounds provided here.

Finally, examining eq. 58, we see that many cells are required to cover the entire $|\mathcal{Y}|$-dimensional probability space (to cover all 'directions' that $\hat{\mathbf{w}}$ could falls relative to the generator *pmf* $\mathbf{w}$); however, if for a specific scenario the number of cells (and halfspaces and/or shingles) could be greatly reduced if one is only interested in whether $\hat{\mathbf{w}}$ falls within a convex subspace of the probability space.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, ISBN 0-471-06259-6, John Wiley and Sons Inc., New York, 1991

[2] L. G. Valiant, "A theory of the learnable," Communications of the ACM, Vol 27, No. 11, pp. 1134–1142, 1984.

[3] J. Langford, "Tutorial on practical prediction theory for classification," Journal of Machine Learning Research, 6, pp. 273–306, 2005

[4] Y. Seldin and N. Tishby, "PAC-bayesian analysis of co-Clustering and beyond," Journal of Machine Learning Research Vol 11, pp 3595-3636, Dec 2010

[5] B. Guedj, "A Primer on PAC-Bayesian Learning," arXiv preprint arXiv:1901.05353, 2019

[6] T. van Erven and P. Harremoës, "Rényi Divergence and Kullback-Leibler Divergence," IEEE Transactions on Information Theory, vol. 60, no. 7, pp. 3797-3820, 2014

[7] Devdatt P. Dubhashi and Alessandro Panconesi, *Concentration of Measure for the Analysis of Randomized, ISBN 978-0-521-88427-3 Algorithms* Cambridge University Press, New York, NY 2009