Supplemental Material for ISF Proposal: Large Alphabet Inference

I. A Proof for Theorem 3

We begin with the following proposition.

Proposition I.1. *Let* $\delta_2 > 0$. *Then, with probability* $1 - \delta_2$,

$$\sum_i \sum_{k=1}^{m/2} k^{m-k}(np_i(1-p_i))^k \leq \frac{n}{n-1}\left(\sum_i \sum_{k=1}^{m/2} k^{m-k}(n\hat{p}_i(1-\hat{p}_i))^k + \epsilon\right) \tag{1}$$

for every even m, *where*

$$\epsilon = \sqrt{\frac{n}{2}\log(1/\delta_2)} \sum_{k=1}^{d} k^{m-k} n^k \left(\frac{k}{n4^{k-1}} + \frac{3k(k-1)(k-2)}{n^3 \cdot 2^{2k-5}}\right) \tag{2}$$

.

Proof. Define $\psi(n,d,\hat{p}) = \sum_i \sum_{k=1}^{d} k^{m-k}(n\hat{p}_i(1-\hat{p}_i))^k$. McDiarmind's inequality suggests that

$$\mathrm{P}\left(\psi(n,d,\hat{p}) - \mathbb{E}\left(\psi(n,d,\hat{p})\right) \leq -\epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{j=1}^{n} c_j^2}\right)$$

where

$$\sup_{x'_j \in \mathcal{X}} \left|\psi(n,d,\hat{p}) - \psi(n,d,\hat{p}')\right| \leq c_j. \tag{3}$$

where $\hat{p}'$ is the MLE over the same sample x^n, but with a different j^{th} observation, x'_j.

First, let us find c_j. We have

$$\sup_{x'_j \in \mathcal{X}} \left|\psi(n,d,\hat{p}) - \psi(n,d,\hat{p}')\right| \overset{\text{(i)}}{\leq} \tag{4}$$

$$\sup_{p\in[0,1-1/n]} 2\Big|\sum_{k=1}^{d} k^{m-k}\left(np(1-p)\right)^k - \sum_{k=1}^{d} k^{m-k}\left(n(p+1/n)(1-(p+1/n)))^k\right)\Big| =$$

$$\sup_{p\in[0,1-1/n]} 2\Big|\sum_{k=1}^{d} k^{m-k}n^k\left(p(1-p)\right)^k - \left((p+1/n)(1-(p+1/n)))^k\right)\Big| \overset{\text{(ii)}}{\leq}$$

$$2\sum_{k=1}^{d} k^{m-k}n^k\left(\frac{k}{n\cdot 4^{k-1}} + \frac{3k(k-1)(k-2)}{n^3\cdot 2^{2k-5}}\right)$$

where

(i) Changing a single observation effects only two symbols (for example, $\hat{p}_l$ and $\hat{p}_t$), where the change is $\pm 1/n$.

(ii) Please refer to Appendix A below.

Next, we have

$$\mathbb{E}(\psi(n,d,\hat{p})) \geq \sum_i \sum_{k=1}^{d} k^{m-k}n^k\left(\mathbb{E}(\hat{p}_i(1-\hat{p}_i))\right)^k =$$

$$\sum_i \sum_{k=1}^{d} k^{m-k}n^k\left(\left(1-\frac{1}{n}\right)p_i(1-p_i)\right)^k \geq$$

$$\left(1-\frac{1}{n}\right)\sum_i \sum_{k=1}^{d} k^{m-k}(np_i(1-p_i))^k \tag{5}$$

where the first inequality follows from Jensen Inequality and the equality that follows is due to $\mathbb{E}(\hat{p}_i(1-\hat{p}_i)) = p(1-p)(1-1/n)$. Going back to McDiarmind's inequality, we have

$$\mathbb{P}\left(\mathbb{E}\psi(n,d,\hat{p}) \geq \psi(n,d,\hat{p}) + \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{nc_j^2}\right) \tag{6}$$

In word, the probability that the random variable $Z = \psi(n,d,\hat{p})$ is smaller than a constant $C = \mathbb{E}(\psi(n,d,\hat{p})) - \epsilon$ is not greater that $\nu = \exp\left(-2\epsilon^2/\sum_{j=1}^{n} c_j^2\right)$. Therefore, it necessarily means that the probability that Z is smaller than a constant smaller than

C, is also not greater than ν. Hence, plugging (5) we obtain

$$\mathbb{P}\left(\left(1-\frac{1}{n}\right)\psi(n,d,p) \geq \psi(n,d,\hat{p})+\epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_j c_j^2}\right)$$

Setting the right hand side to equal δ_2 we get

$$\epsilon = \sqrt{\frac{n}{2}\log(1/\delta_2)}\sum_{k=1}^{d} k^{m-k}n^k\left(\frac{k}{n\cdot 4^{k-1}}+\frac{3k(k-1)(k-2)}{n^3\cdot 2^{2k-5}}\right) \tag{7}$$

and with probability $1-\delta_2$,

$$\sum_i\sum_{k=1}^{d} k^{m-k}(np_i(1-p_i))^k \leq \frac{n}{n-1}\left(\sum_i\sum_{k=1}^{d} k^{m-k}(n\hat{p}_i(1-\hat{p}_i))^k+\epsilon\right) \tag{8}$$

□

Finally, we apply the union bound with $\delta=\delta_1$ and Proposition I.1 to obtain the stated result.

II. A Proof for Corollary 4

We prove the Corollary with two propositions.

Proposition II.1. *Let* $\delta_1 > 0$. *Then, with probability* $1-\delta_1$,

$$\sup_{i\in\mathcal{X}}|p_i-\hat{p}_i(X^n)| \leq \frac{m}{2n}\left(\frac{1}{\delta_1}\right)^{1/m}\left(\sum_i\sum_{k=1}^{m/2}(np_i(1-p_i))^k\right)^{1/m} \tag{9}$$

for every even $m>0$.

Proof. First, we have

$$\mathbb{E}\big(\sup_i|p_i-\hat{p}_i(X^n)|\big)^m \overset{\text{(i)}}{\leq} \frac{1}{n^m}\sum_i\sum_{k=1}^{d}k^{m-k}(np_i(1-p_i))^k \overset{\text{(ii)}}{\leq}$$
$$\left(\frac{d}{m}\right)^m\sum_i\sum_{k=1}^{d}(np_i(1-p_i))^k$$

where $d=n/2$ and

(i) follows from (15) in the main text .

(ii) follows from $k^{m-k} \leq d^m$ for every $k \in \{1, ..., d\}$.

Applying Markov's inequality we obtain

$$\mathbb{P}\left(\sup_i |p_i - \hat{p}_i(X^n)| \geq a\right) \leq \frac{1}{a^m}\mathbb{E}\left(\sup_i |p_i - \hat{p}_i(X^n)|\right)^m \leq \qquad (10)$$
$$\frac{1}{a^m}\left(\frac{d}{n}\right)^m \sum_i \sum_{k=1}^{d} (np_i(1-p_i))^k.$$

Setting the right hand side to equal δ_1 yields

$$a = \left(\frac{1}{\delta_1}\left(\frac{d}{n}\right)^m \sum_i \sum_{k=1}^{d} (np_i(1-p_i))^k\right)^{1/m} = \frac{m}{2n}\left(\frac{1}{\delta_1}\sum_i \sum_{k=1}^{m/2} (np_i(1-p_i))^k\right)^{1/m}.$$

□

Proposition II.2. *Let $\delta_2 > 0$. Then, with probability $1 - \delta_2$,*

$$\sum_i \sum_{k=1}^{d} (np_i(1-p_i))^k \leq \qquad (11)$$
$$\frac{n}{n-1}\left(\sum_i \sum_{k=1}^{d} (n\hat{p}_i(1-\hat{p}_i))^k + d\sqrt{\frac{1}{2}\log(1/\delta_2)}\left(2n^{d-1/2} + 48n^{d-5/2}\right)\right)$$

for every even m.

Proof. McDiarmind's inequality suggests that

$$\mathbb{P}\left(\sum_i \sum_{k=1}^{d} (n\hat{p}_i(1-\hat{p}_i))^k - \mathbb{E}\left(\sum_i \sum_{k=1}^{d} (n\hat{p}_i(1-\hat{p}_i))^k\right) \leq -\epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{j=1}^{n} c_j^2}\right)$$

where

$$\sup_{x'_j \in \mathcal{X}} \Big| \sum_i \sum_{k=1}^{d} (n\hat{p}_i(1-\hat{p}_i))^k - \sum_i \sum_{k=1}^{d} (n\hat{p}'_i(1-\hat{p}'_i))^k) \Big| \leq c_j. \qquad (12)$$

First, let us find c_j. We have

$$\sup_{x_j'\in\mathcal{X}}\Big|\sum_i\sum_{k=1}^{d}\left(n\hat{p}_i(1-\hat{p}_i)\right)^k-\sum_i\sum_{k=1}^{d}\left(n\hat{p}_i'(1-\hat{p}_i')\right)^k\Big|\overset{\text{(i)}}{\leq} \tag{13}$$
$$2\sup_{p\in[0,1-1/n]}\Big|\sum_{k=1}^{d}\left(np(1-p)\right)^k-\sum_{k=1}^{d}\left(n(p+1/n)(1-(p+1/n))\right)^k\Big|=$$
$$2\sup_{p\in[0,1-1/n]}\Big|\sum_{k=1}^{d}n^k\left(p(1-p)\right)^k-\left((p+1/n)(1-(p+1/n))\right)^k\Big|\leq$$
$$2\sum_{k=1}^{d}n^k\sup_{p\in[0,1-1/n]}\Big|\left(p(1-p)\right)^k-\left((p+1/n)(1-(p+1/n))\right)^k\Big|\overset{\text{(ii)}}{=}$$
$$2\sum_{k=1}^{d}n^k\left(\frac{k}{n\cdot 4^{k-1}}+\frac{3k(k-1)(k-2)}{n^3\cdot 2^{2k-5}}\right)\leq$$
$$n^{d-1}\sum_{k=1}^{d}\left(\frac{2k}{4^{k-1}}+\frac{3k(k-1)(k-2)}{n^2\cdot 2^{2k-4}}\right)\overset{\text{(iii)}}{\leq}2dn^{d-1}+48dn^{d-3}$$

where

(i) Changing a single observation effects only two symbols (for example, $\hat{p}_l$ and $\hat{p}_t$), where the change is $\pm 1/n$.

(ii) Please refer to Appendix A.

(iii) Follows from $\sum_{k=1}^{d}\frac{k}{4^{k-1}}=4\sum_{k=1}^{d}\frac{k}{4^k}\leq d$ and

$$\sum_{k=1}^{d}\frac{k(k-1)(k-2)}{4^{k-2}}\leq\sum_{k=1}^{d}\frac{k^3}{4^{k-2}}\leq d\max_{k\in[1,d]}\frac{k^3}{4^{k-2}}\leq 16\frac{2\exp(-3)}{\log(4)}\leq 16 \tag{14}$$

where the maximum is obtain for $k^*=3/\log(4)$.

next, we have

$$\mathbb{E}\left(\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k\right)\geq\sum_i\sum_{k=1}^{d}\left(\mathbb{E}(n\hat{p}_i(1-\hat{p}_i))\right)^k= \tag{15}$$
$$\sum_i\sum_{k=1}^{d}n^k\left(\left(1-\frac{1}{n}\right)p_i(1-p_i)\right)^k\geq\left(1-\frac{1}{n}\right)\sum_i\sum_{k=1}^{d}(np_i(1-p_i))^k$$

Going back to McDiarmind's inequality, we have

$$\mathbb{P}\left(\mathbb{E}\left(\sum_i \sum_{k=1}^{d} (n\hat{p}_i(1-\hat{p}_i))^k\right) \geq \sum_i \sum_{k=1}^{d} (n\hat{p}_i(1-\hat{p}_i))^k + \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{j=1}^{n} c_j^2}\right) \tag{16}$$

Plugging (15) we obtain

$$\mathbb{P}\left(\left(1-\frac{1}{n}\right)\sum_i \sum_{k=1}^{d} (np_i(1-p_i))^k \geq \sum_i \sum_{k=1}^{d} (n\hat{p}_i(1-\hat{p}_i))^k + \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_j c_j^2}\right)$$

Setting the right hand side to equal δ_2 we get

$$\epsilon = \sqrt{\frac{n}{2}\log(1/\delta_2)}\left(2dn^{d-1} + 48dn^{d-3}\right)$$

and with probability $1-\delta_2$,

$$\begin{aligned}
&\sum_i \sum_{k=1}^{d} (np_i(1-p_i))^k \leq \qquad (17)\\
&\frac{n}{n-1}\left(\sum_i \sum_{k=1}^{d} (n\hat{p}_i(1-\hat{p}_i))^k + d\sqrt{\frac{1}{2}\log(1/\delta_2)}\left(2n^{d-1/2} + 48n^{d-5/2}\right)\right)
\end{aligned}$$

□

Finally, we apply the union bound to Propositions II.1 and II.2 to obtain

$$\begin{aligned}
&\sup_{i\in\mathcal{X}} |p_i - \hat{p}_i(X^n)| \leq \\
&\frac{m}{2n}\left(\frac{1}{\delta_1}\frac{n}{n-1}\left(\sum_i \sum_{k=1}^{d} (n\hat{p}_i(1-\hat{p}_i))^k + d\sqrt{\frac{1}{2}\log(1/\delta_2)}\left(2n^{d-1/2} + 48n^{d-5/2}\right)\right)\right)^{1/m} \leq \\
&\frac{m}{2\delta_1^{1/m}}\frac{1}{n}\left(\frac{n}{n-1}\right)^{1/m}\left(\sum_i \sum_{k=1}^{m/2} (n\hat{p}_i(1-\hat{p}_i))^k\right)^{1/m} + \\
&\frac{m}{2\delta_1^{1/m}}\frac{1}{n}\left(\frac{n}{n-1}\right)^{1/m}(m/2)^{1/m}\left(\frac{1}{2}\log\left(\frac{1}{\delta_2}\right)\right)^{1/m}\left(2n^{\frac{1}{2}-\frac{1}{2m}} + 48n^{\frac{1}{2}-\frac{5}{2m}}\right)
\end{aligned}$$

with probability $1-\delta_1-\delta_2$. Define $g(m,\delta_1) = m/\delta_1^{1/m}$. Further, it is immediate to show that $(m/2)^{1/m} \leq \sqrt{\exp(1/\exp(1))}$. Hence, with probability $1-\delta_1-\delta_2$,

$$\sup_{i\in\mathcal{X}} |p_i - \hat{p}_i(X^n)| \leq \frac{g(m,\delta_1)}{n}\left(\sum_i \sum_{k=1}^{m/2} (n\hat{p}_i(1-\hat{p}_i))^k\right)^{1/m} + \\ bg(m,\delta_1)(\log(1/\delta_2))^{1/2m}\left(n^{-\frac{1}{2}\left(1+\frac{1}{m}\right)} + 24n^{-\frac{1}{2}\left(1+\frac{5}{m}\right)}\right)$$

for every even m, where $b = \sqrt{2\exp(1/\exp(1))}$. Finally, we would like to choose m which minimizes $g(m,\delta_1)$. We show in Appendix B that $\inf_m g(m,\delta_1) = \exp(1)\log(1/\delta_1)$, where and the infimum is obtained for a choice of $m^* = \log(1/\delta_1)$.

III. A Proof of Theorem 5

Let us first introduce some auxiliary results and background

A. Auxiliary Results

Lemma III.1 (contained in the proof of Lemma 10, [1]). *Let $Y_{i\in I\subseteq\mathbb{N}}$ be random variables such that, for each $i \in I$, there are $v_i > 0$ and $a_i \geq 0$ satisfying*

$$\mathbb{P}\left(Y_i \geq \varepsilon\right) \quad \leq \quad \exp\left(-\frac{\varepsilon^2}{2(v_i + a_i\varepsilon)}\right), \qquad \varepsilon \geq 0. \tag{18}$$

Put

$$v^* := \sup_{i\in I} v_i, \quad V^* := \sup_{i\in I} v_i \log(i+1), \quad a^* := \sup_{i\in I} a_i, \quad A^* := \sup_{i\in I} a_i \log(i+1). \tag{19}$$

Then

$$\mathbb{P}\left(\sup_{i\in I} Y_i \geq 2\sqrt{V^* + v^*\log\frac{1}{\delta}} + 4A^* + 4a^*\log\frac{1}{\delta}\right) \quad \leq \quad \delta.$$

Remark III.1. *When considering the random variable $Z = \sup_{i\in\mathbb{N}} |\hat{p}_i - p_i|$, there is no loss of generality in assuming that $p_i \leq 1/2$, $i \in \mathbb{N}$. Indeed, $|Y_i| = |\hat{p}_i - p_i|$ is distributed as $|n^{-1}\operatorname{Bin}(n,p_i) - p_i|$, and the latter distribution is invariant under the transformation $p_i \mapsto 1 - p_i$.*

Lemma III.2. *For any distribution $p_{i\in\mathbb{N}}$,*

$$V(p) \quad \leq \quad \phi(v^*(p)).$$

Proof. (This elegant proof idea is due to Václav Voráček.) There is no loss of generality in assuming $p = p^{\downarrow}$. The claim then amounts to

$$\sup_{i \in \mathbb{N}} v_i \log(i+1) \leq v^* \log \frac{1}{v^*}.$$

The monotonicity of the p_i implies $p_i \leq (p_1 + \ldots + p_i)/i \leq 1/i$. Now $x \leq 1/i \implies x(1-x) \leq 1/(i+1)$ for $i \in \mathbb{N}$, and hence $v_i \leq 1/(i+1)$. Thus, $v_i \log(i+1) \leq v_i \log \frac{1}{v_i}$. Finally, since $x \log(1/x)$ is increasing on $[0, 1/4]$, which is the range of the v_i, we have $\sup_{i \in \mathbb{N}} v_i \log \frac{1}{v_i} \leq v^* \log \frac{1}{v^*}$. □

Remark III.2. *There is no reverse inequality of the form $\phi(v^*(p)) \leq F(V^*(p))$, for any fixed $F : \mathbb{R}_+ \to \mathbb{R}_+$. This can be seen by considering p supported on $[k]$, with $p_1 = \log(k)/k$ and the remaining masses uniform. Then $V^*(p) \approx \log(k)/k$ while $\phi(v^*(p)) \approx \log(k) \log(k/\log k)/k$.*

Proposition III.1. *Let $n \geq 10$ and $\beta = \log(n)$. Then,*

$$f(n) = \frac{\beta^{-\beta} n^2 \left(\frac{n-\beta}{n}\right)^{\beta - n}}{2^\beta - 2} \leq \frac{81}{2}.$$

Proof. To prove the above, we show that $f(n)$ is decreasing for $n > 200$. This means that the maximum of $f(n)$ may be numerically evaluated in the range $n \in \{10, ..., 200\}$. Finally, we verify that the maximum of $f(n)$ is attained for $n = 33$, and is bounded from above by $81/2$ as desired. It remains to verify that $f(n)$ is decreasing for $n > 200$. Since $f(n)$ is non-negative, it is enough to show that $g(n) = \log f(n)$ is decreasing. Denote

$$g(n) = -\beta \log \beta + 2 \log n + (n - \beta) \log(n - \beta) + (n - \beta) \log n - \log(2^\beta - 2). \quad (20)$$

Taking the derivative of $g(n)$ we have,

$$
\begin{aligned}
&g'(n) = \qquad (21)\\
&-\frac{1}{n}(\log\beta+1)+\frac{2}{n}+\left(1-\frac{1}{n}\right)(-\log(n-\beta)-1+\log n)+\frac{n-\beta}{n}-\frac{1}{n}\frac{2^\beta\log 2}{2^\beta-2}=\\
&\frac{1}{n}\left((n-1)\log\frac{n}{n-\beta}-\log\beta-\beta+2-\frac{2^\beta\log 2}{2^\beta-2}\right)\leq\\
&\frac{1}{n}\left(n\log\frac{n}{n-\beta}-\log\beta-\beta+2-\log 2\right)\leq\frac{1}{n}\left(\frac{n\beta}{n-\beta}-\log\beta-\beta+2-\log 2\right)=\\
&\frac{1}{n}\left(\frac{\beta^2}{n-\beta}-\log\beta+2-\log 2\right),
\end{aligned}
$$

where the first inequality follows from $\log(n/(n-\beta))\geq 1$ and $2^\beta/(2^\beta-2)\geq 1$, while the second inequality is due to Bernoulli's inequality, $(n/(n-\beta))^n\leq\exp(n\beta/(n-\beta))$. Finally, it is easy to show that $\beta^2/(n-\beta)$ is decreasing for $n\geq 10$. This means that $\beta^2/(n-\beta)\leq(\log 10)^2/(10-\log(10))$ and $g'(n)<0$ for $n>200$. $\square$

Lemma III.3 (generalized Fano method [2], Lemma 3)**.** *For $r\geq 2$, let $\mathcal{M}_r$ be a collection of r probability measures $\nu_1,\nu_2,...,\nu_r$ with some parameter of interest $\theta(\nu)$ taking values in pseudo-metric space (Θ,ρ) such that for all $j\neq k$, we have*

$$\rho(\theta(\nu_j),\theta(\nu_k))\geq\alpha$$

and

$$D(\nu_j\parallel\nu_k)\leq\beta.$$

Then

$$\inf_{\hat\theta}\max_{j\in[d]}\mathbb{E}_{Z\sim\mu_j}\rho(\hat\theta(Z),\theta(\nu_j))\geq\frac{\alpha}{2}\left(1-\left(\frac{\beta+\log 2}{\log r}\right)\right),$$

where the infimum is over all estimators $\hat\theta: Z\mapsto\Theta$.

Proposition III.2. *Let p and q be two distributions with support size n. Define p by*

$$p_1=\frac{\log n}{2n\log\log n},\quad p_i=\frac{1-p_1}{n-1},\qquad i>1,$$

and q by $q_2=p_1$, and $q_i=p_2$ for $i\neq 2$. Then,

(i) $\|p-q\|_\infty\geq c\frac{\log n}{n\log\log n}$ for some $c>0$ and all n sufficiently large.

(ii) $\lim_{n\to\infty} \frac{n}{\log n} D(p||q) = \frac{1}{2}$

Proof. For the first part, it is enough to show that

$$|p_1 - p_2| \geq c\log(n)/n\log\log n$$

for some $c > 0$ and sufficiently large n. First, we show that $p_1 \geq p_2$ for $n \geq (\log n)^2$. That is,

$$p_1 - \frac{1-p_1}{n-1} = \frac{np_1 - 1}{n-1} > 0 \tag{22}$$

for $np_1 > 1$. Next, fix $0 < c \leq 1/2$. We have,

$$\begin{aligned} |p_1 - p_2| - \frac{c\log(n)}{n\log\log n} =& \frac{ap_1 - 1}{n-1} - \frac{c\log n}{n\log\log n} = \\ & \frac{1}{n-1}\left(\frac{\log n}{2\log\log n} - 1 - \frac{n-1}{n}\frac{c\log n}{\log\log n}\right) = \\ & \frac{1}{(n-1)2\log\log n}\left(\log n\left(1 - \frac{n-1}{n}2c\right) - 2\log\log n\right) > 0 \end{aligned} \tag{23}$$

where the last inequality holds for $c(n-1)/n < 1/2$ and sufficiently large n, as desired. We now proceed to the second part of the proof.

$$\frac{n}{\log n}D(p||q) = \frac{n}{\log n}\left(p_1\log\frac{p_1}{q_1} + p_2\log\frac{p_2}{q_2}\right) = \frac{n}{\log n}(p_1 - p_2)\log\frac{p_1}{p_2}. \tag{24}$$

First, we have

$$\begin{aligned} \frac{n}{\log n}(p_1 - p_2) =& \frac{n}{\log n}\left(p_1 - \frac{1-p_1}{n-1}\right) = \frac{n}{\log n}\left(\frac{np_1 - 1}{n-1}\right) = \\ & \frac{n}{\log n}\frac{\log n/2n\log\log n - 1}{n-1} = \frac{n}{n-1}\left(\frac{1}{2\log\log n} - \frac{1}{\log n}\right). \end{aligned} \tag{25}$$

Next,

$$\begin{aligned} \log\frac{p_1}{p_2} =& \log(n-1) + \log\frac{p_1}{1-p_1} = \log(n-1) + \log\frac{\log n}{2n\log\log n - \log n} = \\ & \log(n-1) + \log\log n - 2\log(2n\log\log n - \log n). \end{aligned} \tag{26}$$

Putting it all together we obtain

$$\frac{n}{\log n} D(p||q) = \tag{27}$$

$$\frac{n}{n-1}\left(\frac{1}{2\log\log n} - \frac{1}{\log n}\right)(\log(n-1) + \log\log n - 2\log(2n\log\log n - \log n)) =$$

$$\frac{n}{n-1}\Bigg(\frac{\log(n-1)}{2\log\log n} - \frac{\log(n-1)}{\log n} + \frac{1}{2} - \frac{\log\log n}{\log n} -$$

$$\frac{\log(2n\log\log n - \log n)}{2\log\log n} + \frac{\log(2n\log\log n - \log n)}{\log n}\Bigg) =$$

$$\frac{n}{n-1}\Bigg(\frac{1}{2} + \frac{\log(n-1) - \log(2n\log\log n - \log n)}{2\log\log n} +$$

$$\frac{\log(2n\log\log n - \log n) - \log(n-1)}{\log n} - \frac{\log\log n}{\log n}\Bigg).$$

It is straightforward to show that the last three terms in the parenthesis above converge to zero for sufficiently large n, which leads to the stated result. □

Lemma III.4 ([3]). *When estimating a single Bernoulli parameter in the range* $[0, p_0]$*,* $\Theta(p_0\varepsilon^{-2}\log(1/\delta))$ *draws are both necessary and sufficient to achieve additive accuracy* ε *with probability at least* $1-\delta$*.*

B. Bernstein inequalities

Background: Let $Y \sim \text{Bin}(n, \theta)$ be a Binomial random variable and let $\hat{\theta} = Y/n$ be the its MLE.

- Classic Bernstein [4]:

$$\mathbb{P}\left(\hat{\theta} - \theta \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(\theta(1-\theta) + \varepsilon/3)}\right) \tag{28}$$

  with an analogous bound for the left tail. This implies:

$$|\theta - \hat{\theta}| \leq \sqrt{\frac{2\theta(1-\theta)}{n}\log\frac{2}{\delta}} + \frac{2}{3n}\log\frac{2}{\delta}. \tag{29}$$

- Empirical Bernstein [5, Lemma 5]:

$$|\theta - \hat{\theta}| \leq \sqrt{\frac{5\hat{\theta}(1-\hat{\theta})}{n}\log\frac{2}{\delta}} + \frac{5}{n}\log\frac{2}{\delta}. \tag{30}$$

We are now ready to present the proof of Theorem 5.

C. *Proof of Theorem* 5

Theorem 5. *Let $p = p_{i\in\mathbb{N}}$ be a distribution over $\mathbb{N}$ and put $v^* = v^*(p)$, $V^* = V(p)$. For $n \geq 81$ and $\delta \in (0,1)$, we have that*

$$\|p-\hat{p}\|_\infty \leq 2\sqrt{\frac{V^*}{n}+\frac{v^*}{n}\log\frac{2}{\delta}}+\frac{4}{3n}\log\frac{2(n+1)}{\delta}+\frac{\log n}{n} \leq \tag{31}$$

$$2\sqrt{\frac{\phi(v^*)}{n}+\frac{v^*}{n}\log\frac{2}{\delta}}+\frac{4}{3n}\log\frac{2(n+1)}{\delta}+\frac{\log n}{n}; \tag{32}$$

$$\|p-\hat{p}\|_\infty \leq 2\sqrt{\frac{v^*\log(n+1)}{n}+\frac{v^*}{n}\log\frac{2}{\delta}}+\frac{4}{3n}\log\frac{2(n+1)}{\delta}+\frac{\log n}{n} \tag{33}$$

holds with probability at least $1-\delta-81/n$.

Proof. We assume without loss of generality that p is sorted in descending order: $p_1 \geq p_2 \geq \ldots$ and further, as per Remark III.1, that $p_1 \leq 1/2$. The estimate $\hat{p}_i$ is just the MLE based on n iid draws.

Our strategy for analyzing $\sup_{i\in\mathbb{N}}|\hat{p}_i - p_i|$ will be to break up p into the "heavy" masses, where we apply a maximal Bernstein-type inequality, and the "light" masses, where we apply a multiplicative Chernoff-type bound.

We define the "heavy" masses as those with $p_i \geq 1/n$. Denote by $I \subset \mathbb{N}$ the set of corresponding indices and note that $|I| \leq n$. For $i \in I$, put $Y_i = \hat{p}_i - p_i$. Then (28) implies that each Y_i satisfies (18) with $v_i = p_i(1-p_i)/n$ and $a_i = 1/(3n)$; trivially, $\max_{i\in I} a_i \log(i+1) = \log(n+1)/(3n)$. Invoking Lemma III.1 twice (once for Y_i and again for $-Y_i$) together with the union bound,

we have, with probability $\geq 1-\delta$,

$$\max_{i\in I}|\hat{p}_i - p_i| \leq 2\sqrt{\frac{V^*}{n}+\frac{v^*}{n}\log\frac{2}{\delta}}+\frac{4\log(n+1)}{3n}+\frac{4}{3n}\log\frac{2}{\delta}. \tag{34}$$

Next, we analyze the light masses. Our first "segment" consisted of the $p_i \in [n^{-1}, 1]$; these were the heavy masses. We take the next segment to consist of $p_i \in [(2n)^{-1}, n^{-1}]$, of which there are at most $2n$ atoms. The segment after that will be in the range $[(4n)^{-1}, (2n)^{-1}]$, and, in general, the kth segment is in the range $[(2^k n)^{-1}, (2^{k-1}n)^{-1}]$, and will contain at most $2^k n$ atoms. To the kth segment, we apply the Chernoff bound $\mathbb{P}(\hat{p} \geq p+\varepsilon) \leq \exp(-nD(p+\varepsilon||p))$, where $p = (2^k n)^{-1}$ and $\varepsilon = \varepsilon_k = 2^k p\beta - p$, for some β to be specified below. [Note that $D(\alpha p||p)$ is monotonically increasing in p for fixed α, so we are justified in taking the left endpoint.] For this choice, in the kth segment we have

$$\begin{aligned} D(p+\varepsilon||p) &= D(2^k p\beta||p) = D\left(\frac{\beta}{n}\middle\|\frac{1}{2^k n}\right) \\ &= \frac{(n-\beta)\log\left(\frac{2^k(n-\beta)}{2^k n-1}\right) + \beta\log\left(2^k\beta\right)}{n} \\ &\geq \frac{(n-\beta)\log\left(\frac{n-\beta}{n}\right) + \beta\log\left(2^k\beta\right)}{n}, \end{aligned}$$

since neglecting the $-1/2^k$ additive term in the denominator decreases the expression. Let E be the event that *any* of the p_is in any of the segments $k = 1, 2, \ldots$ has a corresponding $\hat{p}_i$ that exceeds β/n. Then

$$\mathbb{P}(E) \leq \sum_{k=1}^{\infty} 2^k n \exp\left(-(n-\beta)\log\left(\frac{n-\beta}{n}\right) - \beta\log\left(2^k\beta\right)\right) = \frac{2\beta^{-\beta} n \left(\frac{n-\beta}{n}\right)^{\beta-n}}{2^\beta - 2}.$$

For the choice $\beta = \log n$, we have

$$\mathbb{P}(E) \leq \frac{2\beta^{-\beta} n \left(\frac{n-\beta}{n}\right)^{\beta-n}}{2^\beta - 2} \leq \frac{81}{n}, \qquad n \geq 10, \tag{35}$$

which is proved in Proposition III.1. Now E is the event that $\sup_{i: p_i < 1/n}(\hat{p}_i - p_i) \geq \log(n)/n$. Since $p_i < 1/n$, there is no need to consider the left-tail deviation at this scale, as all of the probabilities will be zero. Combining (34) with (35) yields (31). Since Lemma III.2 implies that $V^* \leq \phi(v^*)$, (32) follows from (31). Finally, (33) follows from (31) via the obvious relation $V^* \leq \log(n+1)v^*$. □

IV. A Proof for Theorem 6

We begin with an elementary observation: for $N \in \mathbb{N}$ and $a, b \in [0,1]^N$, we have

$$\left|\max_{i\in[N]} a_i(1-a_i) - \max_{i\in[N]} b_i(1-b_i)\right| \leq \max_{i\in[N]} |a_i - b_i|,$$

and this also carries over to $a, b \in [0,1]^{\mathbb{N}}$. Let us denote $v^* := \sup_{i\in\mathbb{N}} p_i(1-p_i)$ and $\hat{v}^* := \sup_{i\in\mathbb{N}} \hat{p}_i(1-\hat{p}_i)$.

Together with (33), this implies

$$|v^* - \hat{v}^*| \leq \|p - \hat{p}\|_\infty \leq a + b\sqrt{v^*}$$

where

$$\begin{aligned} a &= \frac{4}{3n}\log\frac{2(n+1)}{\delta} + \frac{\log n}{n}, \\ b &= 2\sqrt{\frac{\log(n+1)}{n} + \frac{1}{n}\log\frac{2}{\delta}}. \end{aligned}$$

Following the proof of Lemma 5 in [5],

$$\begin{aligned} |v^* - \hat{v}^*| &\leq a + b\sqrt{v^*} \\ &\leq a + b\sqrt{\hat{v}^* + |v^* - \hat{v}^*|} \\ &\leq a + b\sqrt{\hat{v}^*} + b\sqrt{|v^* - \hat{v}^*|}, \end{aligned}$$

where we used $v^* \leq \hat{v}^* + |v^* - \hat{v}^*|$ and $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$.

Now we have an expression of the form

$$A \leq B\sqrt{A} + C,$$

where $A = |v^* - \hat{v}^*|$, $B = b$, $C = a + b\sqrt{\hat{v}^*}$, which implies $A \leq B^2 + B\sqrt{C} + C$, or

$$|v^* - \hat{v}^*| \leq b^2 + a + b\sqrt{\hat{v}^*} + b\sqrt{a + b\sqrt{\hat{v}^*}}.$$

Using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and $\sqrt{xy} \leq (x+y)/2$,

$$\begin{aligned} |v^* - \hat{v}^*| &\leq b^2 + a + b\sqrt{\hat{v}^*} + b\sqrt{a} + b\sqrt{b\sqrt{\hat{v}^*}} \\ &\leq b^2 + a + b\sqrt{\hat{v}^*} + b\sqrt{a} + b(b + \sqrt{\hat{v}^*})/2 \\ &= a + 3b^2/2 + b\sqrt{a} + 3b\sqrt{\hat{v}^*}/2. \end{aligned}$$

We still have

$$a + b\sqrt{v^*} \leq a + 3b^2/2 + b\sqrt{a} + 3b\sqrt{\hat{v}^*}/2,$$

whence, with probability $1-\delta$,

$$\|p-\hat{p}\|_\infty \le a + 3b^2/2 + b\sqrt{a} + 3b\sqrt{\hat{v}^*}/2. \tag{36}$$

V. A Proof for Theorem 7

We begin with the following proposition.

Proposition V.1. *Assume there exists $V_\delta(X^n)$ such that*

$$\mathbb{P}\left(|p_j - \hat{p}_j| \ge V_\delta(X^n) | p_j = p_{[1]}\right) \le \delta. \tag{37}$$

Then,

$$\mathbb{E}(V_\delta(X^n)) \ge z_{\delta/2}\sqrt{\frac{p_{[1]}(1-p_{[1]})}{n}} + O\left(\frac{1}{n}\right).$$

Proof. Assume there exists $V_\delta(X^n)$ that satisfies (37) and

$$\mathbb{E}(V_\delta(X^n)) < z_{\delta/2}\sqrt{\frac{p_{[1]}(1-p_{[1]})}{n}} + O\left(\frac{1}{n}\right).$$

From (37), we have that

$$\mathbb{P}\left(|p_j - \hat{p}_j| \ge U_\delta(X^n) | p_j = p_{[1]}\right) = \mathbb{P}\left(|p_{[1]} - \hat{p}_j| \ge U_\delta(X^n) | p_j = p_{[1]}\right) \le \delta. \tag{38}$$

Now, consider $Y \sim \text{Bin}(n, p_{[1]})$. Let Y^n be a sample of n independent observations. Notice we can always extend the Binomial case to a multinomial setup with parameters p, over any alphabet size $||p||_0$. That is, given a sample Y^n, we may replace every $Y=0$ (or $Y=1$) with a sample from a multinomial distribution over an alphabet size $||p||_0 - 1$. Further, we may focus on samples for which $p_{[1]}$ is the most likely event in the alphabet, and construct a CI for $p_{[1]}$ following (38). This means that we found a CI for $p_{[1]}$ with an expected length that is shorter than the CP CI, which contradicts its optimality. □

Now, assume there exists $U_\delta(X^n)$ that satisfies

$$\mathbb{P}\left(|p_j - \hat{p}_j| \ge U_\delta(X^n)\right) \le \delta. \tag{39}$$

and

$$\mathbb{E}(U_\delta(X^n)) < z_{\delta/2}\sqrt{\frac{p_{[1]}(1-p_{[1]})}{n}} + O\left(\frac{1}{n}\right). \tag{40}$$

For simplicity of notation, denote $v = \arg\max_i p_i$ as the symbol with the greatest probability in the alphabet. That is, $p_v = p_{[1]}$. We implicitly assume that v is unique, although the proof holds in case of several maxima as well. We have that

$$\begin{aligned} &\mathbb{P}\left(|p_j - \hat{p}_j| \geq U_\delta(X^n)\right) = \\ &\qquad \sum_{u\in\mathcal{X}} \mathbb{P}\left(|p_j - \hat{p}_j| \geq U_\delta(X^n)|j=u\right)\mathbb{P}(j=u) = \\ &\qquad \mathbb{P}\left(|p_{[1]} - \hat{p}_j| \geq U_\delta(X^n)|j=v\right)\mathbb{P}(j=v) + \\ &\qquad \sum_{u\neq v} \mathbb{P}\left(|p_j - \hat{p}_j| \geq U_\delta(X^n)|j=u\right)\mathbb{P}(j=u). \end{aligned} \tag{41}$$

Proposition V.1 together with assumption (40) suggest that

$$\mathbb{P}\left(|p_{[1]} - \hat{p}_j| \geq U_\delta(X^n)|j=v\right) > \delta.$$

On the other hand, it is well-known that $\hat{p}_{[1]} \to p_{[1]}$ for sufficiently large n [6], [7], [8]. This means that $\mathbb{P}(j=u) \to 1$ and (41) is bounded from below by δ, for sufficiently large n. This contradicts (38) as desired.

Appendix A

We show that

$$\sup_{p\in[0,1-1/n]} \left|\, (p(1-p))^k - ((p+1/n)(1-(p+1/n)))^k)\right| \leq \frac{k}{n\cdot 4^{k-1}} + \frac{3k(k-1)(k-2)}{n^3\cdot 2^{2k-5}}$$

Let $0 \leq p \leq 1/2 - 1/n$. Denote $f_k(p) = ((p(1-p))^k$. Applying Taylor series to $f_k(p+1/n)$ around $f_k(p)$ yields

$$f_k\left(p+\frac{1}{n}\right) = f_k(p) + \frac{1}{n}f_k'(p) + r(p)$$

where $r(p) = \frac{1}{3!}\frac{1}{n^3}f'''(c)$ is the residual and $c \in [p, p+1/n]$ [9]. We have

$$f'_k(p) = k\left(p(1-p)\right)^{k-1}(1-2p) \leq k\left(p(1-p)\right)^{k-1} \tag{42}$$
$$f'''_k(p) = k(k-1)(k-2)p^{k-3}(1-p)^{k-3}(1-2p)^3 - 6k(k-1)p^{k-2}(1-p)^{k-2}(1-2p) \leq$$
$$k(k-1)p^{k-3}(1-p)^{k-3}\left((k-2)+6p(1-p)\right).$$

Hence,

$$\sup_{p\in[0,1/2-1/n]} \left| \left(p(1-p)\right)^k - \left((p+1/n)(1-(p+1/n))\right)^k) \right| = \tag{43}$$
$$\sup_{p\in[0,1/2-1/n]} \left| -\frac{1}{n}f'_k(p) - \frac{1}{3!}\frac{1}{n^3}f'''(c) \right| \leq \sup_{p\in[0,1/2-1/n]} \frac{1}{n}\left|f'_k(p)\right| + \frac{1}{3!}\frac{1}{n^3}\left|f'''(c)\right| \overset{\text{(i)}}{\leq}$$
$$\sup_{p\in[0,1/2-1/n]} \frac{k}{n}\left(p(1-p)\right)^{k-1} + k(k-1)p^{k-3}(1-p)^{k-3}\left((k-2)+6p(1-p)\right) \overset{\text{(ii)}}{\leq}$$
$$\frac{k}{n\cdot 4^{k-1}} + \frac{3k(k-1)(k-2)}{n^3\cdot 2^{2k-5}}$$

where

(i) follows from (42).

(ii) follows from the concavity of $(p(1-p))^k$ for $k \geq 1$.

Appendix B

We study $\min_m m/a^{1/m}$ for some positive a. This problem is equivalent to

$$\min_m \log(m) - \frac{1}{m}\log(a).$$

Taking its derivative with respect to m and setting it to zero yields

$$\frac{d}{dm}\log(m) - \frac{1}{m}\log(a) = \frac{1}{m} + \frac{1}{m^2}\log(a) = 0.$$

Hence, $m^* = \log(1/a)$. Therefore,

$$\min_m m/a^{1/m} = \exp(\log(m^*) - (1/m^*)\log(a)) = \exp(1)\log(1/a). \tag{44}$$

APPENDIX C

We study

$$\min_{m\in\mathbb{R}^+}\left(\frac{\sqrt{m/2}}{\delta^{1/m}}\right)\exp\left(-\frac{1}{2}+\frac{1}{m}\right) \tag{45}$$

This problem is equivalent to

$$\min_{d\in\mathbb{R}^+}\frac{1}{2}\log(d)+\frac{1}{2d}\log\left(\frac{1}{\delta}\right)-\frac{1}{2}+\frac{1}{2d} \tag{46}$$

where $d=m/2$. Taking its derivative with respect to d and setting it to zero yields

$$\frac{1}{2d}-\frac{1}{2d^2}\left(\log\left(\frac{1}{\delta}\right)+1\right)=0.$$

Hence, $d^*=\log(1/\delta)+1$. Therefore,

$$\min_{d\in\mathbb{R}^+}\frac{1}{2}\log(d)+\frac{1}{2d}\log\left(\frac{1}{\delta}\right)-\frac{1}{2}+\frac{1}{2d}=\frac{1}{2}\log(\log(1/\delta)+1) \tag{47}$$

and

$$\min_{m\in\mathbb{R}^+}\left(\frac{\sqrt{m/2}}{\delta^{1/m}}\right)\exp\left(-\frac{1}{2}+\frac{1}{m}\right)=\sqrt{\log\left(\frac{1}{\delta}\right)+1}. \tag{48}$$

APPENDIX D

Proposition V.2. *Let $p_{i\in\mathbb{N}}$ be a probability distribution over $\mathbb{N}$. Then,*

$$p_{[1]}=\max_{i\in\mathbb{N}}p_i(1-p_i) \tag{49}$$

where $p_{[1]}=\max_{i\in\mathbb{N}}p_i$ is the largest element in p.

Proof. Let us first consider the case where $p_i\le 1/2$ for all $i\in\mathbb{N}$. Then (49) follows directly from the montonicity of $p_i(1-p_i)$ for $p_i\in[0,1/2]$. Next, assume there exists a single $p_j>1/2$. Specifically, $p_j=1/2+a$ for some positive a. Then, the remaining p_i's are necessarily smaller than $1/2$. Further, the maximum of $p_i(1-p_i)$ over $i\neq j$ is obtained for $p_i=1/2-a$, from the same monotonicity reason. This means that $\max_{i\neq j}p_i(1-p_i)=(1/2-a)(1-(1/2-a))=(1/2+a)(1-(1/2+a))$ where the second equality follows from the symmetry of $p_i(1-p_i)$ around $p_i = 1/2$, which concludes the proof. $\square$

References

[1] D. Cohen and A. Kontorovich, "Local glivenko-cantelli," in *The Thirty Sixth Annual Conference on Learning Theory, COLT*, vol. 195, 2023, p. 715.

[2] B. Yu, "Assouad, Fano, and Le Cam," *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435, 1997.

[3] Y. Peres, "Learning a coin's bias (localized)," Theoretical Computer Science Stack Exchange, 2017, uRL:https://cstheory.stackexchange.com/q/38931 (version: 2017-08-28).

[4] S. Boucheron, G. Lugosi, and O. Bousquet, "Concentration inequalities," in *Summer school on machine learning*. Springer, 2003, pp. 208–240.

[5] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, 2008, pp. 208–215.

[6] A. Gelfand, J. Glaz, L. Kuo, and T.-M. Lee, "Inference for the maximum cell probability under multinomial sampling," *Naval Research Logistics (NRL)*, vol. 39, no. 1, pp. 97–114, 1992.

[7] X. Shifeng and L. Guoying, "Testing for the maximum cell probabilities in multinomial distributions," *Science in China Series A: Mathematics*, vol. 48, pp. 972–985, 2005.

[8] S. Xiong and G. Li, "Inference for ordered parameters in multinomial distributions," *Science in China Series A: Mathematics*, vol. 52, no. 3, pp. 526–538, 2009.

[9] K. R. Stromberg, *An introduction to classical real analysis*. American Mathematical Soc., 2015, vol. 376.