Concentration Inequalities for the Empirical Distribution of Discrete Distributions : Beyond the Method of Types

Jay Mardia, Jiantao Jiao, Ervin Tánczos, Robert D. Nowak, Tsachy Weissman*

October 22, 2019

Abstract

We study concentration inequalities for the Kullback–Leibler (KL) divergence between the empirical distribution and the true distribution. Applying a recursion technique, we improve over the method of types bound uniformly in all regimes of sample size n and alphabet size k, and the improvement becomes more significant when k is large. We discuss the applications of our results in obtaining tighter concentration inequalities for L_1 deviations of the empirical distribution from the true distribution, and the difference between concentration around the expectation or zero. We also obtain asymptotically tight bounds on the variance of the KL divergence between the empirical and true distribution, and demonstrate their quantitatively different behaviors between small and large sample sizes compared to the alphabet size.

Contents

*Jay Mardia and Tsachy Weissman are with the Department of Electrical Engineering, Stanford University. Email: {jmardia, tsachy}@stanford.edu. Jiantao Jiao is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. Email: jiantao@berkeley.edu. Ervin Tánczos and Robert D. Nowak are with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison. Email: tanczos@wisc.edu, nowak@engr.wisc.edu.

1 Introduction and main results

Concentration inequalities of empirical distributions play fundamental roles in probability theory, statistics, and machine learning. For example, the Kolmogorov–Smirnov goodness-of-fit test relies on the Dvoretzky–Kiefer–Wolfowitz–Massart inequality [Mas90] to control the significance level, the widely used Sanov's theorem [CT12, Theorem 11.4.1] is proved via the method of types concentration inequality, and the Vapnik–Chervonenkis [VC15] inequalites, chaining [Tal14] ideas, among others, provide foundational tools for statistical learning theory, and allow us to control the deviation of the empirical distribution from the true distribution under integral probability metrics [Mül97]. There have been works before that seek to improve well known concentration inequalities by incorporating additional distributional information. For example, [BK+13] and [RS+13, p 24] both provide distribution dependent improvements to the Hoeffding bounds for the case of discrete distributions with support size $= 2$. Our focus, however, will be on improving uniform concentration inequalities for discrete distributions with known support size that are distribution independent.

This paper focuses on obtaining concentration inequalities of the Kullback–Leibler (KL) divergence between the empirical distribution and the true distribution for discrete distributions. The KL divergence is not an integral probability metric, which makes it difficult to apply the VC inequality and chaining to obtain tight bounds. However, there is fundamental importance in understanding the behavior of the KL divergence. For example, what Sanov's theorem reveals is that as the sample size $n \to \infty$, under any distribution P, the probability of observing an empirical distribution Q is characterized as $e^{-nD(Q\|P)+o(n)}$. Also, due to the Pinsker inequality [CT12, Chapter 11], a concentration inequality for KL divergence also implies a concentration inequality for the total variation distance, which will also be used in this paper to provide an improved concentration inequality for the L_1 deviation. Concentration inequalities for the KL divergence between the estimated distribution and the real distribution in general exponential families were considered in [Lai88, Mai18].

Concretely, we obtain n i.i.d. samples $X^n = (X_1, X_2, \ldots, X_n)$ following distribution $P = (p_1, p_2, \ldots, p_k) \in \mathrm{int}(\mathcal{M}_k)$[1], where $\mathcal{M}_k$ denotes the space of probability measures with alphabet size k. We are interested in concentration inequalities for the random variable

$$D(\hat{P}_{n,k}\|P) \triangleq \sum_{i=1}^{k} \hat{p}_i \log \frac{\hat{p}_i}{p_i},$$

[1]int stands for interior and this assumption simply means that all p_i are positive. This assumption is without loss of generality because otherwise we could work with a smaller value of k.

where $\hat{P}_{n,k} \triangleq (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k)$ is the empirical distribution obtained from the n samples. Throughout this paper, log is the natural logarithm.

Probably the most well-known result on the concentration of $D(\hat{P}_{n,k}\|P)$ is due to the method of types [Csi98, Lemma II.1], which is used in the proof of [CT12, Sanov's Theorem 11.4.1]). It states that for any $\epsilon > 0$, we have

$$\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \leq \binom{n+k-1}{k-1} e^{-n\epsilon} \tag{1}$$

here $\binom{x}{y}$ is the binomial coefficient.

This bound is tight asymptotically when k, ϵ are fixed and $n \to \infty$ in the sense that

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right)\right) = -\epsilon \tag{2}$$

as shown by [CT12, Theorem 11.4.1], but it does not capture the correct dependence on k in the non-asymptotic regime. In the modern era of big data, it is usually no longer a valid assumption that the sample size is significantly larger than the parameter dimension, and a clear understanding of the concentration inequalities in the non-asymptotic and large alphabet regime is becoming increasingly important.

Example 1. *When $k = 2$, it is well known that the upper bound can be improved [DZ10, Remark (c) Theorem 2.2.3] to*

$$\mathbb{P}\left(D(\hat{P}_{n,2}\|P) \geq \epsilon\right) \leq 2e^{-n\epsilon}, \tag{3}$$

rather than the $(n+1)e^{-n\epsilon}$ bound from Equation 1. This fact is a consequence of a union bound and the Sanov property of convex sets [Csi84, Theorem 1] and is proved in Lemma 8 in Appendix C for the convenience of the reader. It is also clear that if we want a uniform bound that works for all P and ϵ, then this bound is tight. Consider $P = (\frac{1}{2}, \frac{1}{2})$ and $\epsilon = \log 2$. Then $\mathbb{P}\left(D(\hat{P}_{n,2}\|P) \geq \epsilon\right) = \mathbb{P}\left(\hat{P}_{n,2} = (1,0) \cup \hat{P}_{n,2} = (0,1)\right) = 2 \cdot \frac{1}{2}^n = 2e^{-n\epsilon}$ where we have used the fact that $D\left(\hat{P}_{n,2}\|(\frac{1}{2}, \frac{1}{2})\right)$ has maximum value $\log 2$ which is attained at the extremal points of $\mathcal{M}_2$.

Following the asymptotic tightness of Sanov's theorem, we aim at bounds of the type

$$\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \leq f(n,k)e^{-n\epsilon} = e^{n(\epsilon_{\text{thresh}} - \epsilon)}, \tag{4}$$

where $\epsilon_{\text{thresh}} = \frac{\log f(n,k)}{n}$. The threshold ϵ_{thresh} can be interpreted as the lower bound on ϵ such that (4) becomes non-vacuous. Note that ϵ_{thresh} is defined with respect to a bound, but we will not introduce subscripts to denote this and let the corresponding upper bound be clear by context. For the method of types bound, we have

$$\frac{k-1}{n} \log\left(\frac{n+k-1}{k-1}\right) \leq \epsilon_{\text{thresh}} = \frac{\log\binom{n+k-1}{k-1}}{n}. \tag{5}$$

Before we talk about improving ϵ_{thresh}, however, we should see just how much we can hope to improve it. To this end, we observe that there also exists a lower bound for ϵ_{thresh} that follows from Equations (6) and (7) below. Indeed, since for non-negative random variable X one has

$\mathbb{E}[X^p] = \int_0^\infty pt^{p-1}\mathbb{P}(X \geq t)dt = \int_0^{\epsilon_{\text{thresh}}} pt^{p-1}\mathbb{P}(X \geq t)dt + \int_{\epsilon_{\text{thresh}}}^\infty pt^{p-1}\mathbb{P}(X \geq t)dt$, equation (4) implies that

$$\mathbb{E}[D(\hat{P}_{n,k}\|P)^p] \leq (\epsilon_{\text{thresh}})^p + \frac{p!}{n^p}. \tag{6}$$

It also follows from [JVHW17, Lemma 21] that when $n \geq 15k$ and P is the uniform distribution, we have

$$\mathbb{E}[D(\hat{P}_{n,k}\|P)] \geq \frac{k-1}{2n} + \frac{k^2}{20n^2} - \frac{1}{12n^2}, \tag{7}$$

and for any $n \geq 1$, [Pan03, Proposition 1] shows

$$\mathbb{E}[D(\hat{P}_{n,k}\|P)] \leq \log\left(1 + \frac{k-1}{n}\right) \leq \frac{k-1}{n}, \tag{8}$$

which means that Equation (7) is tight up to constants. Together they imply that when $15k \leq n \leq Ck$ and P is uniform, ϵ_{thresh} is at least a constant because $\epsilon_{\text{thresh}} \geq \frac{k-3}{2n}$. So we can't, in general, hope to get an ϵ_{thresh} that is smaller than $\frac{k-3}{2n}$.

The contribution of this paper can be understood as obtaining uniformly smaller ϵ_{thresh} for all configurations of n and k compared with the method of types bound in (5). These upper bounds on ϵ_{thresh} follow from the concentration inequality presented in Theorem 3 stated in Section 1.1. This inequality is complicated and we thus present slightly looser but easier to use bounds alongside, and spend Section 1.2 understanding how Theorem 3 compares to the method fo types bound.

Naturally, one may ask whether our Theorem 3 provides the best non-asymptotic bound on the KL deviation. Although this question appears non-trivial, we hope to demonstrate through the following results on the variance of $D(\hat{P}_{n,k}\|P)$ that bounding $\mathbb{P}(D(\hat{P}_{n,k}\|P) \geq \epsilon)$ may not be the right question to ask in this context.

Theorem 1. *We have the following upper and lower bounds on the variance of $D(\hat{P}_{n,k}\|P)$.*

- *There exists a universal constant C such that for any $P \in \mathcal{M}_k$ and any n,*

$$\mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right] \leq \min\left(\frac{6\cdot(3+\log k)^2}{n}, C\frac{k}{n^2}\right). \tag{9}$$

- *For a fixed k, for any $P \in \mathcal{M}_k$, asymptotically as n goes to infinity,*

$$2nD(\hat{P}_{n,k}\|P) \xrightarrow{\mathcal{D}} \chi^2_{k-1}. \tag{10}$$
$$2(k-1) \leq \lim_{n\to\infty} 4n^2\mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right].$$

  Here χ^2_{k-1} is the chi-square distribution with $k-1$ degrees of freedom and hence has variance $2(k-1)$. Notationally, $\xrightarrow{\mathcal{D}}$ means 'converges in distribution'.

Observe that the asymptotic lower bound of Theorem 1 implies that in the variance upper bound, the $C\frac{k}{n^2}$ terms is tight for large n. The variance bound $O\left(\frac{\log^2 k}{n}\right)$, is in fact tight when $n = 1$. We prove Theorem 1 in Section A.

Combining Theorem 1 with Equations (7) and (8), we observe that it requires at least $n \gg k$ to achieve vanishing expectation for $D(\hat{P}_{n,k}\|P)$ if P is uniform in $\mathcal{M}_k$, but it only requires $n \gg (\log k)^2$ to achieve vanishing variance.

So, if $n \geq 15k$, and P uniform, then we have

$$\mathbb{E}[D(\hat{P}_{n,k}\|P)] = \Theta\left(\frac{k}{n}\right)$$

and

$$\text{standard deviation of } D(\hat{P}_{n,k}\|P) = O\left(\frac{\sqrt{k}}{n}\right) \ll \Theta\left(\frac{k}{n}\right) = \mathbb{E}[D(\hat{P}_{n,k}\|P)].$$

In other words, in this regime the random variable $D(\hat{P}_{n,k}\|P)$ is concentrating very tightly around its expectation, and proving a concentration inequality of the type $\mathbb{P}(D(\hat{P}_{n,k}\|P) \geq \epsilon)$ fails to capture the different behavior of the expectation and the variance of $D(\hat{P}_{n,k}\|P)$. In this context, it may be more insightful to provide bounds for the centered concentration, i.e.,

$$\mathbb{P}\left(|D(\hat{P}_{n,k}\|P) - \mathbb{E}[D(\hat{P}_{n,k}\|P)]| \geq \epsilon\right), \tag{11}$$

for which Theorem 1 provides a bound via Chebyshev's inequality, but we suspect stronger (exponential) bounds are within reach.

The main technique we use in the variance bound for Theorem 1 is to break the quantity into smooth ($p_i > \frac{1}{n}$) and non-smooth ($p_i \leq \frac{1}{n}$) regions. We then apply polynomial approximation techniques inspired from [Pan03] / [BS04] when $p_i > \frac{1}{n}$ and then utilize the negative association properties of multinomial random variables [JDP83]. Note that in [Pan03] such a polynomial approximation technique provides a near-optimal scaling for $\mathbb{E}[D(\hat{P}_{n,k}\|P)]$ in the worst case, which suggests that it might be useful in analyzing the variance too. The reason we go through the effort of carefully handling dependent random variables in the proof of Theorem 1 is that the Poissonized version of this problem gives an incorrect (and worse) scaling for the variance. We show the following lower bound on the variance in the Poissonized version of the problem.

Theorem 2. *Let $\hat{P}^{\mathsf{Poi}}_{n,k} = \left(\hat{p}^{\mathsf{Poi}}_1, \hat{p}^{\mathsf{Poi}}_2, ..., \hat{p}^{\mathsf{Poi}}_k\right)$ where each $\hat{p}^{\mathsf{Poi}}_i$ is independently distributed as $\frac{\mathsf{Poi}(np_i)}{n}$, where $\mathsf{Poi}(\lambda)$ is a Poisson random variable with parameter λ. Then for a fixed k, as $n \to \infty$*

$$\sqrt{n}D(\hat{P}^{\mathsf{Poi}}_{n,k}\|P) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{12}$$
$$1 \leq \lim_{n\to\infty} n\mathsf{Var}\left[D(\hat{P}^{\mathsf{Poi}}_{n,k}\|P)\right].$$

This shows that the Poissonized version of the problem cannot give us the right scaling for the variance ($\frac{k}{n^2}$ as from Theorem 1) because of the asymptotic lower bound of $\frac{1}{n}$ in the Poissonized version.

The concentration inequality presented in Theorem 3 on KL divergence deviation can also be translated into a concentration inequality for the L_1 deviation via Pinsker's inequality. In fact, we use a strengthened version of Pinsker's inequality [WOS+03, Theorem 2.2] to obtain a bound that, to our knowledge, beats the best known concentration inequality for the L_1 distance between the empirical distribution and the true distribution for large k. We formally state these results in Section 1.3 in Lemma 1 and Theorem 4. To compare known results with our results from Theorem 4, we plot Figures 6 and 7.

The rest of the paper is organized as follows. We present the details of Theorem 3 in Section 1.1, compare the performance of this Theorem to the method of types in Section 1.2, present the improved L_1 deviation inequality in Section 1.3, and discuss future directions in Section 2. The

proofs of main results are collected in the Appendices A and B. Appendix C states some auxiliary lemmas and integrals that are used throughout this paper.

Notation: We use the notation $a_\gamma \lesssim b_\gamma$ to denote that there exists a universal constant C such that $\sup_\gamma \frac{a_\gamma}{b_\gamma} \leq C$. Notation $a_\gamma \asymp b_\gamma$ is equivalent to $a_\gamma \lesssim b_\gamma$ and $b_\gamma \lesssim a_\gamma$. Notation $a_\gamma \gg b_\gamma$ means that $\liminf_\gamma \frac{a_\gamma}{b_\gamma} = \infty$, and $a_\gamma \ll b_\gamma$ is equivalent to $b_\gamma \gg a_\gamma$. The sequences a_γ, b_γ are non-negative.

1.1 The Kullback–Leibler concentration inequality

Define $c_0 = \pi, c_1 = 2, K_{-1} = 1, d_0 = \max\{\pi, \frac{e^3}{2}\} = \frac{e^3}{2}$.

$$c_m \triangleq \begin{cases} \frac{1\times3\times5\ldots\times m-1}{2\times4\times6\times\ldots\times m} \cdot \pi & m \text{ is even}, m \geq 2 \\ \frac{2\times4\times6\times\ldots\times m-1}{1\times3\times5\ldots\times m} \cdot 2 & m \text{ is odd}, m \geq 3 \end{cases} \tag{13}$$

$$K_m \triangleq \prod_{j=0}^{m} c_j = \begin{cases} \frac{\pi(2\pi)^{\frac{m}{2}}}{2\times4\times\ldots\times m} \leq \sqrt{\frac{\pi}{m}}(\sqrt{\frac{2\pi e}{m}})^m & m \text{ is even} \\ \frac{(2\pi)^{\frac{m+1}{2}}}{1\times3\times\ldots\times m} \leq \sqrt{\frac{\frac{e^3}{2}}{m}}(\sqrt{\frac{2\pi e}{m}})^m & m \text{ is odd} \end{cases} \tag{14}$$

Observe that c_m behaves as $\sqrt{\frac{2\pi}{m}}$ for large m and that for all positive integers m, $K_m \leq \sqrt{\frac{d_0}{m}}(\sqrt{\frac{2\pi e}{m}})^m$.

Theorem 3. *For all $n, k \geq 2$ and $P \in \mathcal{M}_k$, we have, for universal constants $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$ and $C_1 = \frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}} \approx 2.9290$, the following, where c_m and K_m are defined as in (13) and (14).*

$$\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \leq e^{-n\epsilon}\left[\frac{3c_1}{c_2}\sum_{i=0}^{k-2} K_{i-1}(\frac{e\sqrt{n}}{2\pi})^i\right] \leq e^{-n\epsilon}\left[\frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}}\left(\sum_{i=1}^{k-2}\left(\sqrt{\frac{e^3 n}{2\pi i}}\right)^i + 1\right)\right] \tag{15}$$

The table that follows contains slightly looser but much more easily used and interpreted versions of the upper bound in (15).

More interpretable upper bounds for $\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right)$, $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$	
Parameter Range	*Upper Bound*
$3 \leq k \leq \sqrt{nC_0} + 2$	$C_1 e \left(\sqrt{\frac{C_0 n}{k}}\right)^k e^{-n\epsilon}$
$3 \leq k \leq \frac{nC_0}{e} + 2$	$C_1 k \left(\sqrt{\frac{C_0 n}{k}}\right)^k e^{-n\epsilon}$
$\frac{nC_0}{e} + 2 \leq k \leq nC_0 + 2$	$C_1 k e^{\frac{C_0 n}{2e}} e^{-n\epsilon}$
$k \geq nC_0 + 2$	$C_1 \left(nC_0 e^{\frac{C_0 n}{2e}} + k\right) e^{-n\epsilon}$

The key technique we employ to prove Theorem 3 is a recursive approach to reduce the problem with alphabet size k to a problem with alphabet size $k-1$. Since we have good bounds for $k = 2$, we can induct from this case to obtain bounds for higher values of k. We formalize this and present a proof in Appendix B.

1.2 Comparison with the method of types bound

In this section we compare the results of Theorem 3 to the Method of Types bound to better understand what improvements it yields.

The contribution of this theorem can be understood as obtaining uniformly smaller ϵ_{thresh} for all configurations of n and k compared with the method of types bound in (5). Our Theorem 3 implies the following ϵ_{thresh}:

ϵ_{thresh} improvement using the upper bound from Theorem 3, $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$, $C_1 \approx 2.9290$	
Parameter Range	ϵ_{thresh}(Theorem 3)
$3 \leq k \leq \sqrt{nC_0} + 2$	$\dfrac{k \log\left(\sqrt{\frac{C_0 n}{k}}\right) + \log(C_1 e)}{n}$
$3 \leq k \leq \dfrac{nC_0}{e} + 2$	$\dfrac{k \log\left(\sqrt{\frac{C_0 n}{k}}\right) + \log(C_1 k)}{n}$
$\dfrac{nC_0}{e} + 2 \leq k \leq nC_0 + 2$	$\dfrac{\frac{nC_0}{2e} + \log(C_1 k)}{n}$
$k \geq nC_0 + 2$	$\dfrac{\log\left(nC_0 e^{\frac{C_0 n}{2e}} + k\right) + \log(C_1)}{n}$

To better compare the improvement of ϵ_{thresh} in Theorem 3 with the method of types bound, we upper bound the ratio $\frac{\epsilon_{\text{thresh}}(\text{Theorem 3})}{\epsilon_{\text{thresh}}(\text{method of types})}$ for several scalings of n, k after letting $n \to \infty$.

ϵ_{thresh} improvement using the upper bound from Theorem 3, $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$, $C_1 \approx 2.9290$		
Parameter Range	$\dfrac{\epsilon_{\text{thresh}}(\text{Theorem 3})}{\epsilon_{\text{thresh}}(\text{method of types})}$	Upper bound on ratio
$k = o(n)$	$\dfrac{k \log\left(\sqrt{\frac{C_0 n}{k}}\right) + \log(C_1 e)}{\log\binom{n+k-1}{k-1}}$	$\dfrac{1}{2}$
$k = \dfrac{nC_0}{e}$	$\dfrac{k \log\left(\sqrt{\frac{C_0 n}{k}}\right) + \log(C_1 k)}{\log\binom{n+k-1}{k-1}}$	$\dfrac{1}{2} \cdot \dfrac{1}{\log\left(1 + \frac{e}{C_0}\right)} \approx 0.8125$
$k = nC_0$	$\dfrac{\frac{k}{2e} + \log(C_1 k)}{\log\binom{n+k-1}{k-1}}$	$\dfrac{1}{2e} \cdot \dfrac{1}{\log\left(1 + \frac{1}{C_0}\right)} \approx 0.6758$

To compute the upper bound on the asymptotic ration in the table, we use the lower bound on $\epsilon_{\text{thresh}}(\text{method of types})$ from equation 5.

We notice that the constant improvement on the exponent ϵ_{thresh} provided by Theorem 3 can be viewed as a power function improvement on the tail probability. For $k = o(n)$, the asymptotic ratio of at least $\frac{1}{2}$ implies at least a square root improvement on the prefactor to the method of types bound. In fact, for regimes where $k \leq (\frac{nC_0}{4})^{\frac{1}{3}}$, we can obtain a better ϵ_{thresh} than the one obtained using the bound in Theorem 3 by letting $F = \{\hat{P}_{n,k} \in \mathcal{M}_k : D(\hat{P}_{n,k} \| P) \geq \epsilon)\}$ in Lemma 6 which gives a different bound on $\mathbb{P}\left(D(\hat{P}_{n,k} \| P) \geq \epsilon\right)$ and is proved in Appendix C. The ϵ_{thresh} one obtains from Lemma 6 is

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

. One can observe that this is smaller than the ϵ_{thresh} values in the table above when $2(k-1) \leq \sqrt{\frac{nC_0}{k}}$, which is satisfied when $k \leq (\frac{nC_0}{4})^{\frac{1}{3}}$. In particular, this shows that for any $k \leq (\frac{nC_0}{4})^{\frac{1}{3}}$ we get

$$\frac{\epsilon_{\text{thresh}}(\text{method of types})}{\epsilon_{\text{thresh}}(\text{Lemma 6})} = \Theta\left(\frac{\log\left(\frac{n+k}{k}\right)}{\log k}\right).$$

For any k that is polylogarithmic in n, this is a superconstant improvement.

To further illustrate the results, we plot the log of the upper bounds we obtain from Theorem 3 and from the method of types bound (along with the trivial upper bound one on probability) in Figures 1, 2, 3, and 4. In the numerical plots in these figures which accompany the cartoon plots, we have set P to be the uniform distribution and used Monte Carlo simulations to calculate and plot the true probabilities.

$3 \leq k \leq \sqrt{nC_0} + 2$, $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$, $C_1 \approx 2.9290$

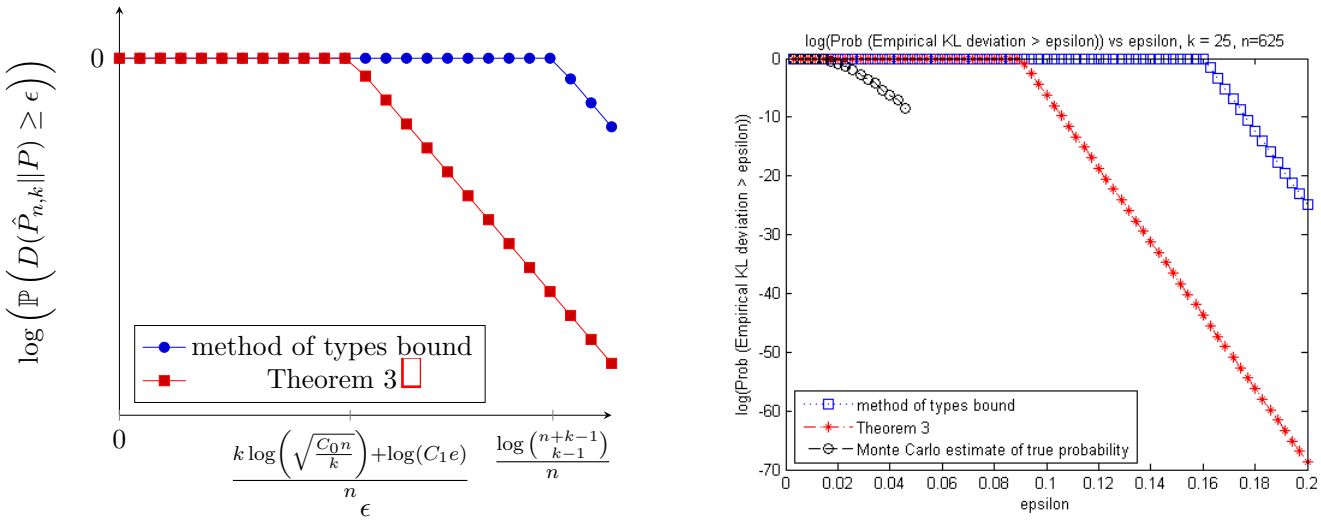

Figure 1: Cartoon plot and Numerical plot- $\log\left(\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right)\right)$ vs ϵ for large n
Observe that for large n, Theorem 3 is a significant improvement over the Method of types bound. Our cartoon plot on the left is corroborated with a numerical plot in which we fix P to be uniform and also plot a Monte Carlo estimate of the true probabilities.

$\sqrt{nC_0} + 2 \leq k \leq \frac{nC_0}{e} + 2$, $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$, $C_1 \approx 2.9290$

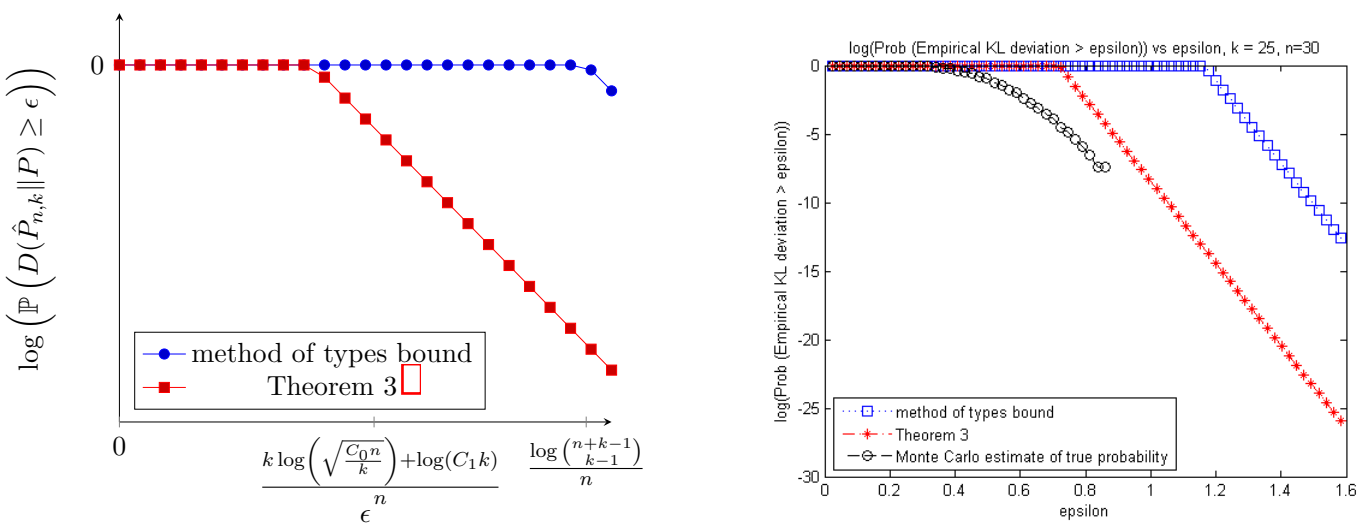

Figure 2: Cartoon plot and Numerical plot- $\log\left(\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right)\right)$ vs ϵ for medium n
When n is medium sized, Theorem 3 is a significant improvement over the method of types bound and we can see it is much closer to the true probabilities than in the regime of Figure 1. Our cartoon plot on the left is corroborated with a numerical plot in which we fix P to be uniform and also plot a Monte Carlo estimate of the true probabilities.

$\frac{nC_0}{e} + 2 \leq k \leq nC_0 + 2$, $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$, $C_1 \approx 2.9290$

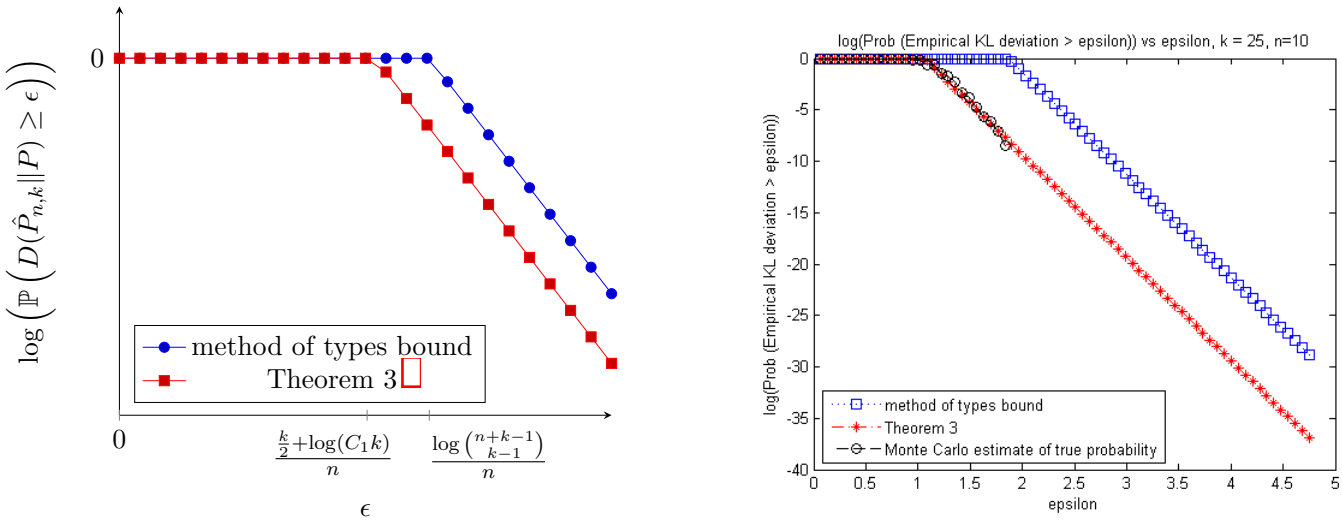

Figure 3: Cartoon plot and Numerical plot- $\log\left(\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right)\right)$ vs ϵ for small n
For small n, Theorem 3 is much better than the Method of types bound. As demonstrated by our numerical plot, Theorem 3 and the true probabilities (for P uniform and computed via a Monte Carlo simulation) are more or less the same. Hence in the regime of high dimensional distributions or very little data, Theorem 3 is essentially tight.

$nC_0 + 2 \leq k$, $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$, $C_1 \approx 2.9290$

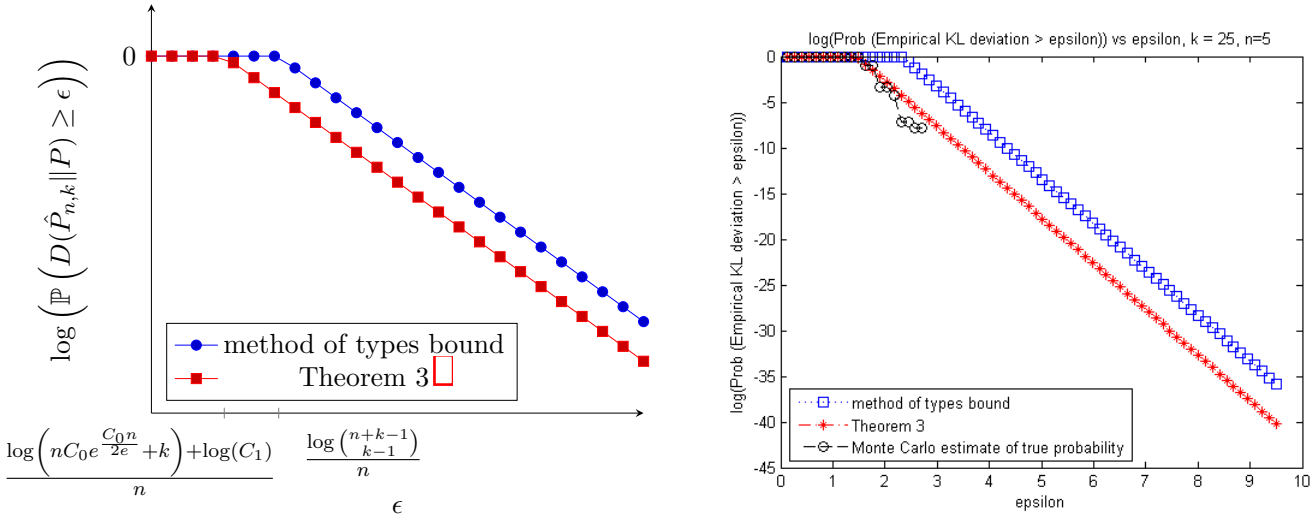

Figure 4: Cartoon plot and Numerical plot- $\log\left(\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right)\right)$ vs ϵ for very small n
When n is much smaller than k, qualitatively the picture is still the same as in Figure 3. Theorem 3's improvement over the method of types bound is enough that it numerically almost matches the Monte Carlo estimates of the true probabilities. (Plotted for P uniform.)

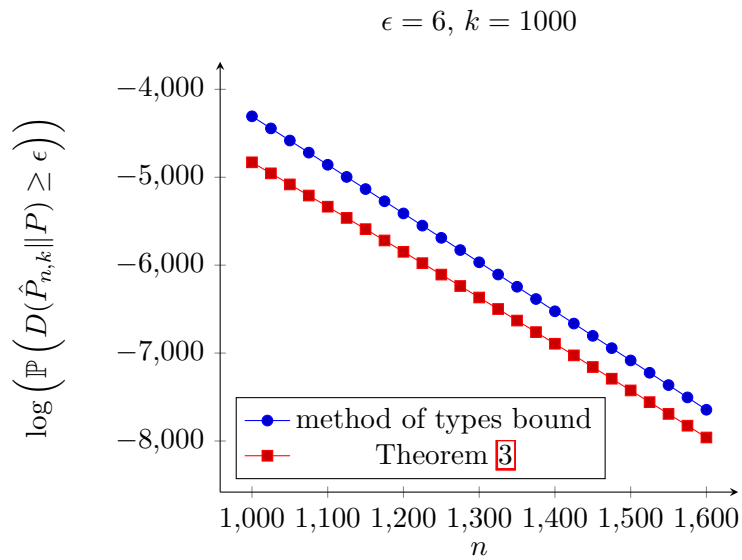

Figure 5: Numerical plot- $\log\left(\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right)\right)$ vs n. This plot illustrates the fact that as n decreases, Theorem 3 is a more and more significant improvement to the method of types bound.

1.3 Tightening L_1 deviation inequalities

Having obtained results bounding the probability of a large KL deviation of the empirical distribution from the true distribution, we can now make use of Pinsker's inequality relating the L_1 distance between two distributions to the KL divergence between them to obtain bounds on the probability of L_1 deviation between the empirical and true distributions. We see that our bounds improve on the state-of-the-art bound from [WOS+03, Theorem 2.1] in the regime when $k \asymp n$. We first state some definitions and known results.

$$\|\hat{P}_{n,k} - P\|_1 \triangleq \sum_{j=1}^{k} |\hat{p}_j - p_j|$$

Definition 1. *Suppose P is a discrete distribution with alphabet size k. Then,*

$$\pi_P \triangleq \max_{A \subseteq [k]} \min\left(\mathbb{P}(A), 1 - \mathbb{P}(A)\right).$$

Note that $\pi_P \leq \frac{1}{2}$ for any P.

Definition 2. *For $p \in [0, \frac{1}{2})$*

$$\varphi(p) \triangleq \frac{1}{1-2p} \log \frac{1-p}{p},$$

and by continuity set $\varphi(\frac{1}{2}) = 2$.

Observe that $\varphi(p) \geq 2$ for all $p \in [0, \frac{1}{2}]$.

Lemma 1. *From [WOS+03, Theorems 2.1 and 2.2], and with π_P, $\varphi(p)$ as defined in Definitions 1 and 2*

- *Let P be a probability distribution in $\mathcal{M}_k$. Then for all n, k, ϵ-*

$$\mathbb{P}\left(\|\hat{P}_{n,k} - P\|_1 \geq \epsilon\right) \leq \left(2^k - 2\right) e^{-\frac{n\varphi(\pi_P)\epsilon^2}{4}}.$$

- *Let P, Q be a probability distributions in $\mathcal{M}_k$. Then we have the following strengthened Pinsker inequality*

$$\|Q - P\|_1 \leq 2\sqrt{\frac{D(Q\|P)}{\varphi(\pi_P)}}.$$

Now, with the slightly strengthened version of the Pinkser inequality from Lemma 1 we can get the inequality

$$\mathbb{P}\left(\|\hat{P}_{n,k} - P\|_1 \geq \epsilon\right) \leq \mathbb{P}\left(2\sqrt{\frac{D(\hat{P}_{n,k}\|P)}{\varphi(\pi_P)}} \geq \epsilon\right),$$

which combined with Theorem 3 allows us to obtain the following bounds on the L_1 deviation probability.

Theorem 4. *For all $n, k \geq 2$ and $P \in \mathcal{M}_k$, we have, for universal constants $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$ and $C_1 = \frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}} \approx 2.9290$, the following, where c_m and K_m are defined as in Equations 13 and Equation 14 and with π_P, $\varphi(p)$ as defined in Definitions 1 and 2*

$$\mathbb{P}\left(\|\hat{P}_{n,k} - P\|_1 \geq \epsilon\right) \leq e^{-\frac{n\varphi(\pi_P)\epsilon^2}{4}} \left[\frac{3c_1}{c_2}\sum_{i=0}^{k-2} K_{i-1}(\frac{e\sqrt{n}}{2\pi})^i\right] \tag{16}$$

$$\leq e^{-\frac{n\varphi(\pi_P)\epsilon^2}{4}} \left[\frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}}\left(\sum_{i=1}^{k-2}\left(\sqrt{\frac{e^3 n}{2\pi i}}\right)^i + 1\right)\right]. \tag{17}$$

We can make these bounds more interpretable in the same way as in Theorem 3 in the table following it. Here we note the regimes of k and n for which Theorem 4 actually provides better bounds than Lemma 1 and compare the two in the following table.

Comparing upper bounds for $\mathbb{P}\left(\|\hat{P}_{n,k} - P\|_1 \geq \epsilon\right)$ from Lemma 1 and Theorem 4, $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$		
Parameter Range	Theorem 4	Lemma 1 [WOS+03, Theorem 2.1]
$\frac{nC_0}{4} + 2 \leq k \leq \frac{nC_0}{e} + 2$	$C_1 k\left(\sqrt{\frac{C_0 n}{k}}\right)^k e^{-\frac{n\varphi(\pi_P)\epsilon^2}{4}}$	$\left(2^k - 2\right) e^{-\frac{n\varphi(\pi_P)\epsilon^2}{4}}$
$\frac{nC_0}{e} + 2 \leq k$	$C_1 k e^{\frac{C_0 n}{2e}} e^{-\frac{n\varphi(\pi_P)\epsilon^2}{4}}$	$\left(2^k - 2\right) e^{-\frac{n\varphi(\pi_P)\epsilon^2}{4}}$

Hence for large enough k, Theorem 4 outperforms Lemma 1 in the regimes listed above. Of course, keeping in mind that L_1 distance between probability distributions takes a maximum value of 2, this improvement is only meaningful if either of these upper bounds does better than the trivial upper bound of 1 on any probability. However, since $\varphi(\pi_P) \geq 2$ for all $P \in \mathcal{M}_k$, this happens even for $n \asymp k$. We illustrate this and the difference between Lemma 1 and Theorem 4 in Figures 6 and 7. Observe that the smaller n becomes, the more significantly improved our results become compared to known bounds and approach the true probabilities (as computed using Monte Carlo simulations). For the Monte Carlo simulation of true probabilities, we have set P to be the uniform distribution.

$\frac{nC_0}{4} + 2 \leq k \leq \frac{nC_0}{e} + 2$, $C_0 = (\frac{e^3}{2\pi}) \approx 3.1967$, $C_1 \approx 2.9290$

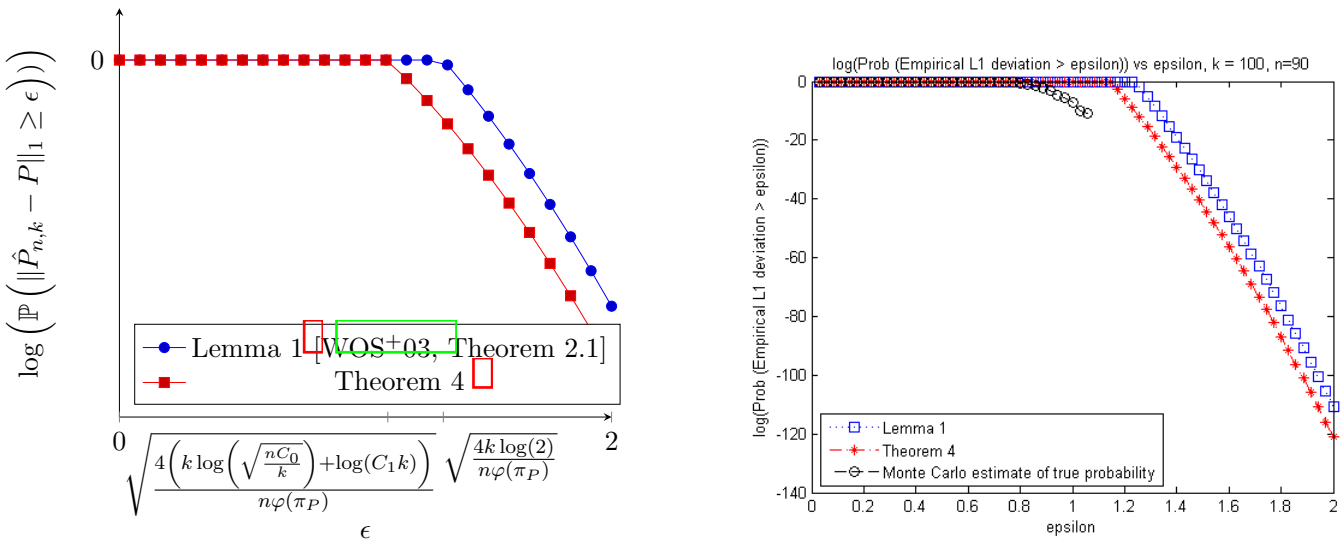

Figure 6: Cartoon and numerical plots- $\log\left(\mathbb{P}\left(\|\hat{P}_{n,k} - P\|_1 \geq \epsilon\right)\right)$ vs ϵ for medium n.
When n is comparable to k, we plot cartoon versions of the previously known bounds from Lemma 1 and our bounds from Theorem 4. The latter shows a non-trivial but slight improvement, and we corroborate this plot with a numerical plot for $k = 100$, $n = 90$ with P uniform (and hence $\varphi(\pi_P) = 2$). In the numerical figure we also plot a Monte Carlo estimate of the true probabilities, which provides a baseline to aim for.

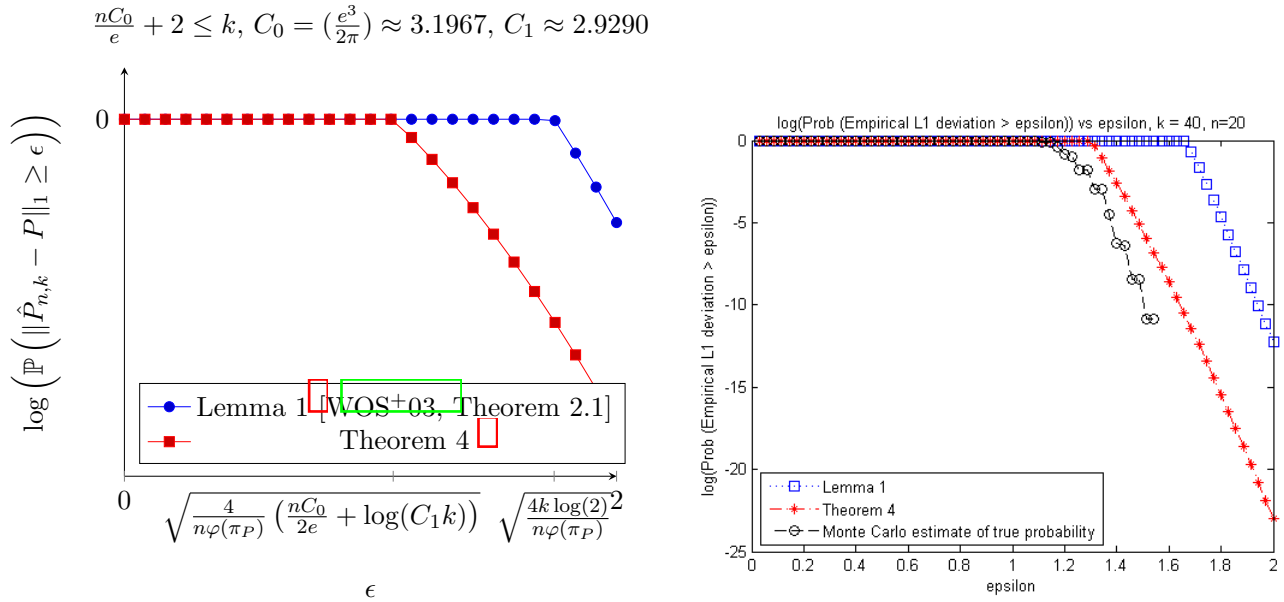

Figure 7: Cartoon and numerical plots- $\log\left(\mathbb{P}\left(\|\hat{P}_{n,k} - P\|_1 \geq \epsilon\right)\right)$ vs ϵ for small n.
When n is smaller to k, we plot cartoon versions of the previously known bounds from Lemma 1 and our bounds from Theorem 4. The latter shows a significant improvement, and we corroborate this plot with a numerical plot for $k = 40$, $n = 20$ with P uniform (and hence $\varphi(\pi_P) = 2$). In the numerical figure we also plot a Monte Carlo estimate of the true probabilities, which provides a baseline to aim for. We observe that the bounds from Theorem 4 are closer to the true probabilities compared with that from Lemma 1.

2 Future directions

We hope this work opens more doors than it closes. Motivated by equation (8), we conjecture the following bound, which appears non-trivial to prove or disprove for general n and k:

Conjecture 1.

$$\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \leq \left(1 + \frac{k-1}{n}\right)^n 2e^{-n\epsilon}. \tag{18}$$

Note that we can't hope to get ϵ_{thresh} values much better than those implied by Conjecture 1 because as n grows, for $k = o(n)$, the ϵ_{thresh} we get from Conjecture 1 goes to $\Theta\left(\frac{k}{n}\right)$ which we know is a lower bound on ϵ_{thresh}.

In fact, with the fact that $2n \cdot D(\hat{P}_{n,k}\|P)$ is asymptotically χ^2_{k-1} from Theorem 1, we can make a conjecture for the centralized concentration using known results for sub-exponential random variables.

Conjecture 2. *There exist two constants $g_1 > 0$, $g_2 > 0$ such that for any $t > 0$,*

$$\mathbb{P}\left(|D(\hat{P}_{n,k}\|P) - \mathbb{E}[D(\hat{P}_{n,k}\|P)]| \geq t\right) \leq g_1 e^{-g_2 \min\left\{\frac{n^2t^2}{k-1}, nt\right\}}. \tag{19}$$

Moreover, we note that the L_1 deviation inequalities depend on P, while the KL divergence deviation bounds are uniform. It might be interesting to look at how to get better KL divergence deviation bounds that depend on some parameter of the distribution P.

After initial dissemination of this work, in follow up work [Agr19, Thm 1.2] provided another KL divergence concentration bound. Their bound performs better than Theorem 3 when

$$\frac{k-1}{n}\left(\log 2+1\right) < \epsilon < \frac{k-1}{n}\left(\log 2+\sqrt{\frac{n}{k}}\right).$$

Moreover, although they do not state this, their result immediately implies sub-Gaussian concentration for the square root of the KL divergence $\left(\sqrt{D(\hat{P}_{n,k}\|P)}\right)$ for a restricted range of parameters. Their result states that for $\epsilon > \frac{k-1}{n}\left(\log 2+1\right)$,

$$\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \leq e^{-n\epsilon}\left(2e\left(\frac{\epsilon n}{k-1}-\log 2\right)\right)^{k-1}.$$

Using the facts that $\frac{x}{2}-\log(12x) \geq 0$ if $x \geq 10$ and that $\left(2e\left(\frac{\epsilon n}{k-1}-\log 2\right)\right) < 12\frac{\epsilon n}{k-1}$, we can see that their result implies that for $\epsilon > 10\frac{k-1}{n}$

$$\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \leq e^{-\frac{n\epsilon}{2}}.$$

Since both the total variation (L_1) distance and Hellinger distance are upper bounded by a constant times the square root of the KL divergence, this result implies sub-Gaussian concentration for the empirical distribution in both these distance metrics for a restricted range of parameters. It would be interesting to see if one can obtain such a scaling for the whole range of parameters.

3 Acknowledgments

The authors would like to thank Rohit Agrawal for pointing out a typo in a previous version of this paper; in Theorem 1 we had missed a factor of 2. JM would like to thank Tavor Baharav for catching a bug in an earlier version of Lemma 2.

References

[Agr19] Rohit Agrawal. Concentration of the multinomial in Kullback-Leibler divergence near the ratio of alphabet and sample sizes. *arXiv preprint arXiv:1904.02291*, 2019.

[BK+13] Daniel Berend, Aryeh Kontorovich, et al. On the concentration of the missing mass. *Electronic Communications in Probability*, 18, 2013.

[BS04] Dietrich Braess and Thomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

[Car00] N.L. Carothers. *Real Analysis*. Cambridge University Press, 2000.

[Csi84] Imre Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

[Csi98] Imre Csiszár. The method of types. *IEEE Transactions on Information Theory*, 44(6):2505–2523, 1998.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[DZ10] Amir Dembo and Ofer Zeitouni. *LDP for Finite Dimensional Spaces*, pages 11–70. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[JDP83] Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295, 1983.

[JK70] N.L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions*. Number v. 2 in Distributions in Statistics. Houghton Mifflin, 1970.

[JVHW17] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 63(10):6774–6798, 2017.

[Lai88] Tze Leung Lai. Boundary crossing problems for sample means. *The Annals of Probability*, pages 375–396, 1988.

[Mai18] O-A Maillard. Boundary crossing probabilities for general exponential families. *Mathematical Methods of Statistics*, 27(1):1–31, 2018.

[Mas90] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[Mül97] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[Pan03] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

[RS+13] Maxim Raginsky, Igal Sason, et al. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.

[Tal14] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[VC15] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[WOS+03] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L_1 deviation of the empirical distribution. 2003.

A Proof of Theorem 1

In this Section, we prove Theorem 1. In Subsection A.1 we prove Equation 9 which says that

$$\mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right] \leq \min\left(\frac{6\cdot(3+\log k)^2}{n}, C\frac{k}{n^2}\right),$$

and in Subsection A.2 we prove Equation 10 which says that asymptotically

$$2nD(\hat{P}_{n,k}\|P) \xrightarrow{\mathcal{D}} \chi^2_{k-1}$$

and

$$2(k-1) \leq \lim_{n\to\infty} 4n^2 D(\hat{P}_{n,k}\|P)].$$

Hence we have an upper bound and an asymptotic lower bound on $\mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right]$.

A.1 Showing $\mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right] \leq \min\left(\frac{6\cdot(3+\log k)^2}{n}, C\frac{k}{n^2}\right)$

To prove Equation 9 in Theorem 1, our goal is to upper bound

$$\mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right] = \mathsf{Var}\left(\sum_{i\in[k]} \hat{p}_i \log\frac{\hat{p}_i}{p_i}\right).$$

In what follows in Appendix A, c denotes a constant independent of k and n, though its value might change from line to line.

A.1.1 $\mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right] \leq \frac{6(\log k+3)^2}{n}$

We first show that

$$\mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right] \leq \frac{6(\log k+3)^2}{n}. \tag{20}$$

Indeed, decomposing

$$\mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right] = \mathsf{Var}\left(\sum_{i\in[k]} -\hat{p}_i\log\frac{1}{\hat{p}_i} + \sum_{i\in[k]}\hat{p}_i\log\frac{1}{p_i}\right) \leq 2\left(\mathsf{Var}\left(\sum_{i\in[k]} -\hat{p}_i\log\frac{1}{\hat{p}_i}\right) + \mathsf{Var}\left(\sum_{i\in[k]}\hat{p}_i\log\frac{1}{p_i}\right)\right).$$

The first term is the plug-in entropy estimator, whose variance was shown in [JVHW17, Lemma 15] to be upper bounded by

$$\mathsf{Var}\left(\sum_{i\in[k]} -\hat{p}_i\log\frac{1}{\hat{p}_i}\right) \leq \frac{2(\log k+3)^2}{n}.$$

Regarding the second term, since each term is increasing respect to $\hat{p}_i$, by the negative association property of multinomial distribution (There are many references for negative association properties. For example, one could consult [JDP83, Definition 2.1, Property P6, and Section 3.1(a)]) we obtain

$$\mathsf{Var}\left(\sum_{i\in[k]}\hat{p}_i\log\frac{1}{p_i}\right) \leq \sum_{i\in[k]}\frac{p_i(1-p_i)}{n}\log^2(1/p_i) \leq \sum_{i\in[k]}\frac{p_i}{n}\log^2(1/p_i) \leq \sum_{i\in[k]}\frac{p_i}{n}(1+\log(1/p_i))^2$$

$$= \frac{k}{n} \sum_{i \in [k]} \frac{1}{k} p_i (1 + \log(1/p_i))^2$$
$$\leq \frac{(1 + \log k)^2}{n},$$

where in the last step we used the fact that $x(1 + \ln(1/x))^2$ is a concave function on $(0, 1]$ along with Jensen's inequality.
These two variance upper bounds complete the proof of the fact $\mathsf{Var}\left[D(\hat{P}_{n,k} \| P)\right] \leq \frac{6(\log k+3)^2}{n}$.

A.1.2 $\mathsf{Var}\left[D(\hat{P}_{n,k} \| P)\right] \leq C \frac{k}{n^2}$

We now show the other part of Equation 9, that there exists some absolute constant C

$$\mathsf{Var}\left[D(\hat{P}_{n,k} \| P)\right] \leq C \frac{k}{n^2}. \tag{21}$$

We rewrite the KL-divergence as

$$\sum_{i \in [k]} \hat{p}_i \log \frac{\hat{p}_i}{p_i} = \sum_{p_i \geq 1/n} \left(\frac{(\hat{p}_i - p_i)^2}{p_i} + \hat{p}_i - p_i + \Delta_i \right) + \sum_{p_i < 1/n} \left(\hat{p}_i \log \frac{\hat{p}_i}{p_i} + \hat{p}_i - p_i + p_i - \hat{p}_i \right)$$
$$= \sum_{p_i \geq 1/n} \frac{(\hat{p}_i - p_i)^2}{p_i} + \sum_{p_i \geq 1/n} \Delta_i + \sum_{p_i < 1/n} \left(\hat{p}_i \left(\log \frac{\hat{p}_i}{p_i} \right) + p_i - \hat{p}_i \right) ,$$

where $\Delta_i = \hat{p}_i \log \frac{\hat{p}_i}{p_i} - \frac{(\hat{p}_i - p_i)^2}{p_i} + p_i - \hat{p}_i = \hat{p}_i \log(\hat{p}_i/p_i) - \hat{p}_i^2/p_i + \hat{p}_i$.

Hence

$$\mathsf{Var}\left[D(\hat{P}_{n,k} \| P)\right] = \mathsf{Var}\left(\sum_{p_i \geq 1/n} \frac{(\hat{p}_i - p_i)^2}{p_i} + \sum_{p_i \geq 1/n} \Delta_i + \sum_{p_i < 1/n} \left(\hat{p}_i \left(\log \frac{\hat{p}_i}{p_i} \right) + p_i - \hat{p}_i \right) \right)$$
$$\leq 9 \left(\underbrace{\mathsf{Var}\left(\sum_{p_i \geq 1/n} \frac{(\hat{p}_i - p_i)^2}{p_i} \right)}_{:= V_1} + \underbrace{\mathsf{Var}\left(\sum_{p_i \geq 1/n} \Delta_i \right)}_{:= V_2} + \underbrace{\mathsf{Var}\left(\sum_{p_i < 1/n} \left(\hat{p}_i \left(\log \frac{\hat{p}_i}{p_i} \right) + p_i - \hat{p}_i \right) \right)}_{:= V_3} \right) . \tag{22}$$

We bound the terms V_1, V_2 and V_3 separately.

Bounding V_1:

We first state the following result which we use here and prove in Lemma 7 in Appendix C.
Let $\hat{P}_{n,k} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k)$ be the empirical distribution. Denote

$$X_i = \frac{(\hat{p}_i - p_i)^2}{p_i} - \frac{1 - p_i}{n}.$$

Then, for any $i \neq j$,

$$\mathbb{E}[X_i^2] = \frac{(1 - p_i)(1 + 2(n - 3)p_i(1 - p_i))}{n^3 p_i}$$

$$\mathbb{E}[X_i X_j] = \frac{2(p_i + p_j) - 1 + 2(n-3)p_i p_j}{n^3}.$$

Now defining $I = \{i : np_i \geq 1\}$ we use Lemma 7 as follows. It is clear that $\sum_{i \in I} p_i \leq 1$ and $\sum_{i \in I} 1 \leq k$.
Now, using the fact that for the terms we are concerned with ($i \in I$), $np_i > 1$, that $(1 - p_i) < 1$ and $n - 3 < n$, we can write, using Lemma 7 that

$$\mathbb{E}[X_i^2] \leq \frac{3np_i}{n^3 p_i} = \frac{3}{n^2}. \tag{23}$$

Similarly, using Lemma 7 and the facts that $p_i p_j < p_i$, $n - 2 < n$, $-1 < 0$, we get

$$\mathbb{E}[X_i X_j] \leq 2\frac{p_j + np_i}{n^3}.$$

Using these two inequalities, we can now easily upper bound V_1

$$\begin{aligned} V_1 &= \sum_{i,j \in I} \mathbb{E}[X_i X_j] = \sum_{i \in I} \mathbb{E}[X_i^2] + \sum_{i \in I} \sum_{j \neq i} \mathbb{E}[X_i X_j] \\ &\leq \sum_{i \in I} \frac{3}{n^2} + \sum_{i \in I} \sum_{j \neq i} 2\frac{p_j + np_i}{n^3} \leq \frac{3k}{n^2} + \sum_{i \in I} 2\frac{1 + nkp_i}{n^3} \\ &\leq \frac{3k}{n^2} + 2\frac{k + nk}{n^3} \leq \frac{7k}{n^2}. \end{aligned}$$

$$V_1 \leq \frac{7k}{n^2}. \tag{24}$$

Bounding V_2:
We want to upper bound $V_2 = \mathsf{Var}\left(\sum_{p_i \geq 1/n} \Delta_i\right)$ where $\Delta_i = \hat{p}_i \log(\hat{p}_i/p_i) - \hat{p}_i^2/p_i + \hat{p}_i$. The first step we take towards this is to obtain the following upper bound -

$$V_2 = \mathsf{Var}\left(\sum_{p_i \geq 1/n} \Delta_i\right) \leq c \sum_{p_i \geq 1/n} \mathbb{E}\left[\Delta_i^2\right]. \tag{25}$$

Recall that throughout this Section, c is an absolute constant independent of n and k but may change from line to line. Equation 25 is saying is that we can decompose the variance of the sum in V_2 into a sum of squared expectations.
The standard way of obtaining such an inequality is via the Negative Association properties [JDP83] of multinomial random variables. In particular, we know that multinomial random variables are negatively associated, disjoint monotone functions of negatively associated random variables are negatively associated, and the variance of sums of negatively associated random variables is sub-additive. However, the hitch is that Δ_i's are not monotone in $\hat{p}_i$. Hence our strategy will be to decompose Δ_i as the sum of monotone functions (in $\hat{p}_i$), and use the properties of negatively associated random variables on those to decompose the variance of a sum into a sum of variances. We use the notation $f(x, y) = x \log(x/y) + x - x^2/y$. Then

$$\frac{\partial}{\partial x} f(x, y) = \log(x/y) + 2 - 2x/y = \log(1 + z) - 2z \ ,$$

with $z = (x/y) - 1$. The derivative is zero at $z = 0 \iff x = y$ and at some $z^* \in (-1, 0)$ satisfying $\log(1 + z^*) = 2z^*$, in which case $x = (1 + z^*)y$. Hence the function $f(x, y)$ is decreasing in x on the interval $[0, (1 + z^*)y]$, increasing on the interval $[(1 + z^*)y, y]$, and decreasing on $[y, 1]$. Also note that $f(0, y) = f(y, y) = 0$. Observe that $\Delta_i = f(\hat{p}_i, p_i)$.

Hence the decomposition $\Delta_i = \Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3)}$ with

$$\begin{aligned}
\Delta_i^{(1)} &= \mathbf{1}\{\hat{p}_i \le (1 + z^*)p_i\} \left(f(\hat{p}_i, p_i) - f((1 + z^*)p_i, p_i)\right), \\
\Delta_i^{(2)} &= \mathbf{1}\{\hat{p}_i \le (1 + z^*)p_i\} f((1 + z^*)p_i, p_i) + \mathbf{1}\{\hat{p}_i \in [(1 + z^*)p_i, p_i]\} f(\hat{p}_i, p_i), \\
\Delta_i^{(3)} &= \mathbf{1}\{\hat{p}_i > p_i\} f(\hat{p}_i, p_i)
\end{aligned}$$

holds. Furthermore, $\Delta_i^{(1)}, \Delta_i^{(2)}$ and $\Delta_i^{(3)}$ are all monotonic in $\hat{p}_i$. We make the following two observations that

$$\mathbb{E}[\Delta_i^{(1)} \Delta_i^{(3)}] = \mathbb{E}[\Delta_i^{(2)} \Delta_i^{(3)}] = 0,$$

because the support of $\Delta_i^{(3)}$ is disjoint from the supports of $\Delta_i^{(1)}$ and $\Delta_i^{(2)}$. Also,

$$\mathbb{E}[\Delta_i^{(1)} \Delta_i^{(2)}] = \mathbb{E}[\Delta_i^{(1)}] \mathbb{E}[\Delta_i^{(2)}],$$

because wherever $\Delta_i^{(1)}$ is non-zero, $\Delta_i^{(2)}$ is a constant and hence can be pulled out of the expectation. Using these two observations, and after some algebra, we can obtain the following inequality

$$\mathsf{Var}\left(\Delta_i^{(1)}\right) + \mathsf{Var}\left(\Delta_i^{(2)}\right) + \mathsf{Var}\left(\Delta_i^{(3)}\right) \le \mathbb{E}[\Delta_i^2]. \tag{26}$$

Thus using the negative association of $(\hat{p}_1, \ldots, \hat{p}_k)$, the fact that $\Delta_i^{(1)}, \Delta_i^{(2)}$ and $\Delta_i^{(3)}$ are all monotonic in $\hat{p}_i$, and Equation 26, we get

$$\begin{aligned}
V_2 &\le 3 \left(\mathsf{Var}\left(\sum_{p_i \ge 1/n} \Delta_i^{(1)} \right) + \mathsf{Var}\left(\sum_{p_i \ge 1/n} \Delta_i^{(2)} \right) + \mathsf{Var}\left(\sum_{p_i \ge 1/n} \Delta_i^{(3)} \right) \right) \\
&\le 3 \left(\sum_{p_i \ge 1/n} \mathsf{Var}\left(\Delta_i^{(1)}\right) + \sum_{p_i \ge 1/n} \mathsf{Var}\left(\Delta_i^{(2)}\right) + \sum_{p_i \ge 1/n} \mathsf{Var}\left(\Delta_i^{(3)}\right) \right) \\
&\le 3 \sum_{p_i \ge 1/n} \mathbb{E}\left[\Delta_i^2\right].
\end{aligned}$$

which proves Equation 25.

To further bound this quantity, we introduce the following lemma.

Lemma 2. *For every $x \ge -1$, these standard logarithmic inequalities*

$$x \le (1 + x) \log(1 + x) \le (1 + x)(x)$$

hold.

Proof. For $x = -1$, the inequality is true because $-1 \le 0 \le 0$. Hence now suppose $x \in (-1, 1]$. We then have that

$$1 + x > 0$$

so dividing on throughout by $1+x$, we only need to show that

$$\underbrace{\frac{x}{1+x} - \log(1+x)}_{f_1(x)} \leq 0 \leq \underbrace{x - \log(1+x)}_{f_2(x)}$$

for all $x \in (-1, 1]$. To do this, observe that $f_1(0) = 0$ ($f_2(0) = 0$). If we can show that

$$f_1'(x) \leq 0 \quad (0 \leq f_2'(x)) \text{ for } 0 \leq x \leq 1$$

and

$$0 \leq f_1'(x) \quad (f_2'(x) \leq 0) \text{ for } -1 < x \leq 0,$$

then we are done. Computing the derivative shows that these conditions hold because

$$f_1'(x) = \frac{-x}{(1+x)^2}$$

and

$$f_2'(x) = \frac{x}{1+x}.$$

This concludes the proof of the lemma.

□

Using the Lemma with $x = \frac{\hat{p}_i}{p_i} - 1$ yields

$$\left(\frac{\hat{p}_i}{p_i} - 1\right) \leq \frac{\hat{p}_i}{p_i} \log \frac{\hat{p}_i}{p_i} \leq \frac{\hat{p}_i}{p_i}\left(\frac{\hat{p}_i}{p_i} - 1\right) .$$

Rearranging and multiplying throughout by p_i yields

$$-\frac{(\hat{p}_i - p_i)^2}{p_i} \leq \Delta_i \leq 0.$$

Hence,

$$\mathbb{E}[\Delta_i^2] \leq \mathbb{E}\left[\left(\frac{(\hat{p}_i - p_i)^2}{p_i}\right)^2\right].$$

Now, using Lemma 7 and the relaxation in Equation 23, we obtain

$$\mathbb{E}[\Delta_i^2] \leq 2\left(\frac{3}{n^2} + \frac{(1-p_i)^2}{n^2}\right) \leq \frac{8}{n^2}.$$

Plugging this bound into the inequality $V_2 \leq 3\sum_{p_i \geq 1/n} \mathbb{E}\left[\Delta_i^2\right]$ we obtain

$$V_2 \leq \frac{ck}{n^2}. \tag{27}$$

Bounding V_3:
To upper bound V_3, we again split into a sum of monotone functions and use negative association of multinomial random variables. Note that

$$V_3 = \mathsf{Var}\left(\sum_{p_i < 1/n}\left(\hat{p}_i \log \frac{\hat{p}_i}{p_i}\right) + \sum_{p_i < 1/n}(p_i - \hat{p}_i)\right) \leq 2\left(\mathsf{Var}\left(\sum_{p_i < 1/n}\left(\hat{p}_i \log \frac{\hat{p}_i}{p_i}\right)\right) + \mathsf{Var}\left(\sum_{p_i < 1/n}(p_i - \hat{p}_i)\right)\right).$$

$p_i - \hat{p}_i$ is monotone decreasing in $\hat{p}_i$ and hence we can upper bound the variance of the sum by the sum of the variance using negative association. To do the same for the $\hat{p}_i \log \frac{\hat{p}_i}{p_i}$ term, note that

$$\frac{\partial}{\partial x}\left(x \log \frac{x}{p}\right) = \log \frac{x}{p} + 1 \ ,$$

which is positive when $x > p$. Since $p_i < 1/n$, whenever $\hat{p}_i \geq \frac{1}{n}$, we have $\hat{p}_i \log \frac{\hat{p}_i}{p_i}$ is increasing in $\hat{p}_i$. It is also clear that $0 = 0 \log \frac{0}{p} < \frac{1}{n} \log \frac{1}{np}$. Since the variance computation only cares about the values of $\hat{p}_i \log \frac{\hat{p}_i}{p_i}$ at $\hat{p}_i = \frac{l}{n}$ for $l \in \{0, 1, ..., n\}$ and we have shown that the function is monotone increasing when restricted to this set of inputs, we can again use negative association to upper bound the variance of the sum by the sum of the variance. This gives us

$$V_3 \leq 2\left(\sum_{p_i < 1/n} \mathsf{Var}\left(\hat{p}_i \log \frac{\hat{p}_i}{p_i}\right) + \sum_{p_i < 1/n} \mathsf{Var}\left(p_i - \hat{p}_i\right)\right). \tag{28}$$

Note that $\mathsf{Var}\left(p_i - \hat{p}_i\right) = \mathsf{Var}(\hat{p}_i) = \frac{p_i(1-p_i)}{n} < 1/n^2$ when $p_i < 1/n$. On the other hand

$$\begin{aligned}
\mathbb{E}\left[\hat{p}_i^2 \log^2 \frac{\hat{p}_i}{p_i}\right] &= \sum_{j=1}^{n} \binom{n}{j} p_i^j (1-p_i)^{n-j} \frac{j^2}{n^2} \log^2 \frac{j}{np_i} \\
&\leq \frac{1}{n^2} \sum_{j=1}^{n} \left(\frac{e}{j}\right)^j j^2 \log^2 j \\
&\leq \frac{c}{n^2} \ ,
\end{aligned}$$

using that $\binom{n}{j} \leq (en/j)^j$ and $np_i < 1$, and that $\max_{x \in (0,1)} \left(x \log^2 \frac{1}{x}\right) < 1$.

Using these bounds and the fact that $\mathsf{Var}\left(\hat{p}_i \log \frac{\hat{p}_i}{p_i}\right) \leq \mathbb{E}\left[\hat{p}_i^2 \log^2 \frac{\hat{p}_i}{p_i}\right]$, and that the summation in Equation 28 has at most k terms, we conclude that

$$V_3 \leq \frac{ck}{n^2}. \tag{29}$$

Having shown that for some constant c we have $V_i \leq c\frac{k}{n^2}$ for $i = 1, 2, 3$ (Equations 24, 27, 29), we now use Equation 22 to complete the proof of Equation 21 which finishes Subsubsection A.1.2.

Combining the results of Subsubsections A.1.1 and A.1.2 immediately completes the proof of Equation 9 in Theorem 1 which finishes Subsection A.1.

A.2 Showing $2(k-1) \leq \lim_{n\to\infty} 4n^2 \mathsf{Var}\left[D(\hat{P}_{n,k} \| P)\right]$

The idea behind Equation 10 in Theorem 1 is a use of the well known delta method in statistics. Here we need a second order multivariate version of the delta method. One can consult [VdV00, Chapter 3] for an introduction to the Delta method. The idea behind the delta method is that to understand the asymptotic distribution of a functional of a random variable whose asymptotic distribution we understand, one can use a Taylor approximation of the functional to the desired precision. We

want to show the following.
For a fixed k, for any $P \in \mathcal{M}_k$, asymptotically as n goes to infinity,

$$2nD(\hat{P}_{n,k}\|P) \xrightarrow{\mathcal{D}} \chi^2_{k-1},$$
$$2(k-1) \leq \lim_{n\to\infty} 4n^2 \mathsf{Var}\left[D(\hat{P}_{n,k}\|P)\right].$$

Here χ^2_{k-1} is the chi-square distribution with $k-1$ degrees of freedom and hence has variance $2(k-1)$. If we have

$$2nD(\hat{P}_{n,k}\|P) \xrightarrow{\mathcal{D}} \chi^2_{k-1},$$

then using Slutsky's Theorem [VdV00, 2.8] and the Continuous Mapping theorem [VdV00, 2.3], we obtain

$$\left(2nD(\hat{P}_{n,k}\|P) - \mathbb{E}[2nD(\hat{P}_{n,k}\|P)]\right)^2 \xrightarrow{\mathcal{D}} \left(\chi^2_{k-1} - \mathbb{E}[\chi^2_{k-1}]\right)^2.$$

Keeping in mind that

$$\mathsf{Var}\left(\chi^2_{k-1}\right) = 2(k-1),$$

the asymptotic variance lower bound in Theorem 1 follows from this by applying Fatou's Lemma [Car00, 18.13] on the sequence of positive random variables

$$\left(2nD(\hat{P}_{n,k}\|P) - \mathbb{E}[2nD(\hat{P}_{n,k}\|P)]\right)^2.$$

Hence we now only need to show that $2nD(\hat{P}_{n,k}\|P) \xrightarrow{\mathcal{D}} \chi^2_{k-1}$.

First, let us compute a Taylor series approximation of $D(Q\|P)$. To do that, we need to parametrize P with $k-1$ variables. Hence, let us consider

$$P = (p_1, p_2, ..., p_{k-1}, 1 - \sum_{j=1}^{k-1} p_j)$$

with parameters $p_1, ..., p_{k-1}$. Define

$$p_k = 1 - \sum_{j=1}^{k-1} p_j,$$

and given a $(k \times 1)$ vector V, let V_{-k} denote the $(k-1 \times 1)$ vector obtained from V by removing the kth component. We need the first and second order derivatives of

$$D(Q\|P) = \sum_{i=1}^{k-1} q_i \log \frac{q_i}{p_i} + q_k \log \frac{q_k}{1 - \sum_{j=1}^{k-1} p_j}$$

at P. Here $Q = (q_1, q_2, ..., q_k)$. Let

$$G \triangleq \nabla_{Q_{-k}} D(Q\|P)|_{Q_{-k}=P_{-k}}$$

be a $(k-1 \times 1)$ vector and denote its ith component by $\nabla_{Q_{-k}} D(Q\|P)|_{Q_{-k}=P_{-k}}(i)$ Let

$$H \triangleq \nabla^2_{Q_{-k}} D(Q\|P)|_{Q_{-k}=P_{-k}}$$

be a $(k-1\times k-1)$ matrix and denote its $(i,j)^{th}$ entry by $\nabla^2_{Q_{-k}} D(Q\|P)|_{Q_{-k}=P_{-k}}(i,j)$. For notation, let $\vec{0}$ be the $(k-1\times 1)$ all-zeros vector, $\vec{1}$ be the $(k-1\times 1)$ all-ones vector,

$$P_{-k}^{-1} \triangleq (\frac{1}{p_1}, \frac{1}{p_2}, ..., \frac{1}{p_{k-1}}),$$

and $\mathsf{diag}(P_{-k}^{-1})$ be the $(k-1\times k-1)$ matrix with P_{-k}^{-1} on its diagonal, and $\mathcal{I}_{k-1}$ be the $(k-1\times k-1)$ identity matrix.
We can calculate that

$$G(i) = \nabla_{Q_{-k}} D(Q\|P)|_{Q_{-k}=P_{-k}}(i) = \log\frac{q_i}{p_i} - \log\frac{1-\sum_{j=1}^{k-1} q_j}{1-\sum_{j=1}^{k-1} p_j} = \log\frac{p_i}{p_i} - \log\frac{1-\sum_{j=1}^{k-1} p_j}{1-\sum_{j=1}^{k-1} p_j} = 0.$$

So,

$$G = \nabla_{Q_{-k}} D(Q\|P)|_{Q_{-k}=P_{-k}} = \vec{0}.$$

$$H(i,j) = \nabla^2_{Q_{-k}} D(Q\|P)|_{Q_{-k}=P_{-k}}(i,j) = \begin{cases} \frac{1}{q_k} = \frac{1}{p_k} & i \neq j \\ \frac{1}{q_i} + \frac{1}{q_k} = \frac{1}{p_i} + \frac{1}{p_k} & i = j \end{cases}.$$

So,

$$H = \nabla^2_{Q_{-k}} D(Q\|P)|_{Q_{-k}=P_{-k}} = \frac{1}{p_k}\vec{1}\vec{1}^T + \mathsf{diag}(P_{-k}^{-1}).$$

For the rest of Subsection A.2, denote the empirical distribution with n draws $\hat{P}_{n,k}$ by $\hat{P}$, suppressing the subscripts which are obvious by context for this subsection.

Lemma 3. *[VdV00, 2.18] By the multivariate central limit theorem, we have that*

$$\sqrt{n}\left(\hat{P}_{-k} - P_{-k}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\vec{0}, \Sigma\right),$$

where Σ is a $(k-1\times k-1)$ matrix with

$$\Sigma(i,j) = \begin{cases} -p_i p_j & i \neq j \\ p_i(1-p_i) & i = j \end{cases}.$$

Observation 5.

$$\Sigma H = \mathcal{I}_{k-1}$$

Now we can write our Taylor Series expansion as

$$nD(\hat{P}_{n,k}\|P) = n\underbrace{D(P\|P)}_{0} + \underbrace{nG^T\left(\hat{P}_{-k} - P_{-k}\right)}_{0} + \underbrace{\frac{1}{2}\sqrt{n}\left(\hat{P}_{-k} - P_{-k}\right)^T H\sqrt{n}\left(\hat{P}_{-k} - P_{-k}\right)}_{:=\text{Quadratic term}} + \text{Higher order terms.}$$

Let $\mathcal{Z}$ be a random variable that is distributed as $\mathcal{N}\left(\vec{0}, \Sigma\right)$. Because of the fact that quadratic (and cubic and higher order) maps are continuous, we can use the Continuous Mapping theorem along with Lemma 3 to get

$$\text{Quadratic term} \xrightarrow{\mathcal{D}} \frac{1}{2}\mathcal{Z}^T H \mathcal{Z}.$$

Because the higher order terms all contain cubics or higher powers of $\left(\hat{P}_{-k} - P_{-k}\right)$, each of which are only pre-multiplied by an n, we again have (because of Lemma 3) , by the Continuous mapping theorem -

$$\text{Higher order terms} \xrightarrow{\mathcal{D}} 0.$$

Now, using Slutsky's Theorem, we get

$$nD(\hat{P}_{n,k}\|P) \xrightarrow{\mathcal{D}} \frac{1}{2}\mathcal{Z}^T H \mathcal{Z} + 0 = \frac{1}{2}\mathcal{Z}^T H \mathcal{Z}.$$

So all that remains to be done to complete the proof of Equation 10 of Theorem 1 is to show that $\mathcal{Z}^T H \mathcal{Z}$ is distributed as χ^2_{k-1}. Quadratic forms of multivariate Gaussian random variables are well studied and we have the following result from [JK70, Chapter 29 (Quadratic Forms in Normal Variables)] which we state here for the reader's convenience.

Lemma 4. *If $\mathcal{Z}$ is a $m \times 1$ multivariate Gaussian Random variable with mean $\vec{0}$ and nonsingular covariance matrix $\mathcal{V}$, then the quadratic form $\mathcal{Z}^T A \mathcal{Z}$ is distributed as*

$$\sum_{j=1}^{m} \lambda_j W_j^2,$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$ are the eigenvalues of $\mathcal{V}A$ and W_j's are i.i.d. $\mathcal{N}(0,1)$ random variables.

Applying Lemma 4 to our quantity of interest $\mathcal{Z}^T H \mathcal{Z}$, coupled with Observation 5 (which says that all the $k-1$ eigenvalues of ΣH are 1) says that $\mathcal{Z}^T H \mathcal{Z}$ is distributed as $\sum_{j=1}^{k-1} W_j^2$, which is exactly the definition of a χ^2_{k-1} random variable. This concludes our proof of the fact that $2nD(\hat{P}_{n,k}\|P) \xrightarrow{\mathcal{D}} \chi^2_{k-1}$, and concludes the proof of Theorem 1.

A.3 Variance lower bounds in the Poissonized model

In this section we prove Theorem 2 which demonstrates why it was necessary to handle the dependencies and cancellations in Theorem 1 as we did rather than simply use the Poissonization technique to work with independent random variables which are much easier to work with. We have $\hat{P}^{\mathsf{Poi}}_{n,k} = \left(\hat{p}^{\mathsf{Poi}}_1, \hat{p}^{\mathsf{Poi}}_2, ..., \hat{p}^{\mathsf{Poi}}_k\right)$ where each $\hat{p}^{\mathsf{Poi}}_i$ is independently distributed as $\frac{\mathsf{Poi}(np_i)}{n}$, where $\mathsf{Poi}(\lambda)$ is a Poisson random variable with parameter λ. We want to show that

$$\sqrt{n}D(\hat{P}^{\mathsf{Poi}}_{n,k}\|P) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$
$$1 \leq \lim_{n\to\infty} n\mathsf{Var}\left[D(\hat{P}^{\mathsf{Poi}}_{n,k}\|P)\right].$$

Just like in Subsection A.2, $1 \leq \lim_{n\to\infty} n\mathsf{Var}\left[D(\hat{P}^{\mathsf{Poi}}_{n,k}\|P)\right]$ follows from $\sqrt{n}D(\hat{P}^{\mathsf{Poi}}_{n,k}\|P) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$ using a combination of Slutsky's Theorem, the Continuous mapping theorem, and Fatou's Lemma. Hence we only need to show

$$\sqrt{n}D(\hat{P}^{\mathsf{Poi}}_{n,k}\|P) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

The key difference between the lower bound in this model and in the multinomial model is that for the Poissonized model, we do not have a constraint of the form $\sum_{i=1}^{k} \hat{p}^{\mathsf{Poi}}_i = 1$, which we did in the multinomial model in Subsection A.2. This means we use k parameters in our Taylor series expansion for the delta method, and most importantly, the first order derivative does NOT vanish. We first make the following two observations -

Observation 6. *As $n \to \infty$,*

$$\frac{\mathsf{Poi}(np_i) - np_i}{\sqrt{np_i}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

Proof. We can view $\mathsf{Poi}(np_i)$ as a sum of n i.i.d. $\mathsf{Poi}(p_i)$ random variables because of the properties of Poisson random variables. The statement then follows from a simple application of the Central Limit Theorem. □

Observation 7. $\nabla_Q D(Q\|P)(i) = 1 + \log \frac{q_i}{p_i} \Rightarrow \nabla_Q D(Q\|P)|_{Q=P} = \vec{1}$

Armed with these two observations, we write the Taylor series expansion -

$$\sqrt{n} D(\hat{P}^{\mathsf{Poi}}_{n,k}\|P) = \sqrt{n}\underbrace{D(P||P)}_{0} + \sqrt{n}\left(\nabla_Q D(Q\|P)|_{Q=P}\right)^T \left(\hat{P}^{\mathsf{Poi}}_{n,k} - P\right) + \text{Second and higher order terms.}$$

Using Observation 6 and the fact that 'Second and higher order terms' have quadratic or higher powers in $\left(\hat{P}^{\mathsf{Poi}}_{n,k} - P\right)$ only premultiplied by a $\sqrt{n}$, we obtain that Second and higher order terms $\xrightarrow{\mathcal{D}} 0$. Using Observations 6 and 7, we get

$$\sqrt{n}\left(\nabla_Q D(Q\|P)|_{Q=P}\right)^T \left(\hat{P}^{\mathsf{Poi}}_{n,k} - P\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

Using Slutsky's Theorem, this completes the proof of

$$\sqrt{n} D(\hat{P}^{\mathsf{Poi}}_{n,k}\|P) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

which completes the proof of Theorem 2 and ends Subsection A.3.

B Proof of Theorem 3

B.1 Using conditional probabilities and chain rule to reduce the size k problem to a size $k-1$ problem

Now we begin our proof of Theorem 3. In what follows, $\tilde{P}$ will be a distribution that may change from line to line and may even mean two different things in the same line. The only important thing is that it is a distribution and is of the right support size for the context it appears in. Suppose $\hat{p}_k \neq 1$, then by the chain rule of relative entropy we have, with a little algebra

$$\begin{aligned} D(\hat{P}_{n,k}\|P) &= \sum_{i=1}^{k} \hat{p}_i \log \frac{\hat{p}_i}{p_i} \\ &= \underbrace{D((\hat{p}_k, 1-\hat{p}_k)\|(p_k, 1-p_k))}_{:=A_{\hat{p}_k}} + (1-\hat{p}_k)\underbrace{D(\hat{P}_{n(1-\hat{p}_k),k-1}\|\tilde{P})}_{:=B_{\hat{p}_k}}, \end{aligned} \tag{30}$$

where $\hat{P}_{n(1-\hat{p}_k),k-1}$ and $\tilde{P}$ are both distributions in $\mathcal{M}_{k-1}$ and defined as

$$\hat{P}_{n(1-\hat{p}_k),k-1} \triangleq \left(\frac{\hat{p}_1}{1-\hat{p}_k}, \frac{\hat{p}_2}{1-\hat{p}_k}, \ldots, \frac{\hat{p}_{k-1}}{1-\hat{p}_k}\right),$$

$$\tilde{P} \triangleq \left(\frac{p_1}{1-p_k}, \frac{p_2}{1-p_k}, \ldots, \frac{p_{k-1}}{1-p_k}\right).$$

The idea now is to control the probability of $D(\hat{P}_{n,k}\|P)$ being large by conditioning on the value of $\hat{p}_k$. This fixes $A_{\hat{p}_k}$ and $1-\hat{p}_k$ and we can control $B_{\hat{p}_k}$ by using our control on the probability of $D(\hat{P}_{\cdot,k-1}\|\tilde{P})$ (by building up inductively).

The random variables $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$ are not independent of each other.

However, by the law of total probability, since $\hat{p}_k$ can take values in $\{\frac{0}{n}, \frac{1}{n}, ..., \frac{n-1}{n}, \frac{n}{n}\}$

$$\begin{aligned}
&\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \\
&= \sum_{l=0}^{n} \mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon | n\hat{p}_k = l\right) \mathbb{P}\left(n\hat{p}_k = l\right) \\
&= \sum_{l=0}^{n-1} \mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon | n\hat{p}_k = l\right) \mathbb{P}\left(n\hat{p}_k = l\right) + \mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon | n\hat{p}_k = n\right) \mathbb{P}\left(n\hat{p}_k = n\right) \\
&= \sum_{l=0}^{n-1} \mathbb{P}\left(B_{\hat{p}_k} \geq \frac{\epsilon - A_{\hat{p}_k}}{1-\hat{p}_k} | n\hat{p}_k = l\right) \mathbb{P}\left(n\hat{p}_k = l\right) + \mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon | n\hat{p}_k = n\right) \mathbb{P}\left(n\hat{p}_k = n\right) \\
&= \sum_{l=0}^{n-1} \mathbb{P}\left(D(\hat{P}_{n-l,k-1}\|\tilde{P}) \geq \frac{\epsilon - A_{\frac{l}{n}}}{1-\frac{l}{n}}\right) \mathbb{P}\left(n\hat{p}_k = l\right) + \underbrace{\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon | n\hat{p}_k = n\right) \mathbb{P}\left(n\hat{p}_k = n\right)}_{:=T}.
\end{aligned} \tag{31}$$

In the third equality, since for $l \in [n-1]$ we have $\hat{p}_k \neq 1$, we have used Equation 30.
We will plug in an upper bound for the $k-1$ sized problem into Equation 31 to do our computations.

Definition 3. *Let $\mathcal{E} \subseteq [n]$ be the set containing all $l \in [n]$ such that $A_{\frac{l}{n}} > \epsilon$.*

$$\begin{aligned}
&\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \\
&\leq \sum_{l \in [n]\setminus\mathcal{E}} \mathbb{P}\left(D(\hat{P}_{n-l,k-1}\|\tilde{P}) \geq \frac{\epsilon - A_{\frac{l}{n}}}{1-\frac{l}{n}}\right) \mathbb{P}\left(n\hat{p}_k = l\right) + \sum_{l \in \mathcal{E}} 1 \cdot \mathbb{P}\left(n\hat{p}_k = l\right) + T \\
&= \sum_{l \in [n]\setminus\mathcal{E}} \mathbb{P}\left(D(\hat{P}_{n-l,k-1}\|\tilde{P}) \geq \frac{\epsilon - A_{\frac{l}{n}}}{1-\frac{l}{n}}\right) \mathbb{P}\left(n\hat{p}_k = l\right) + \mathbb{P}\left(A_{\hat{p}_k} > \epsilon\right) + T \\
&= \underbrace{\sum_{l \in [n]\setminus\mathcal{E}} \mathbb{P}\left(D(\hat{P}_{n-l,k-1}\|\tilde{P}) \geq \frac{\epsilon - A_{\frac{l}{n}}}{1-\frac{l}{n}}\right) \mathbb{P}\left(n\hat{p}_k = l\right)}_{:=R_k} + \underbrace{\mathbb{P}\left(D((\hat{p}_k, 1-\hat{p}_k)\|(p_k, 1-p_k)) > \epsilon\right)}_{:=S} + T
\end{aligned} \tag{32}$$

In the first equality we have used the fact that KL divergence is always non-negative..
For S we can simply use our tight characterization of the distribution of types for $k=2$ and from Example 1 observe that

$$\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \geq \mathbb{P}\left(D((\hat{p}_k, 1-\hat{p}_k)\|(p_k, 1-p_k)) > \epsilon\right)$$

could be as large as $2e^{-n\epsilon}$ because all the terms in Equation 32 are non-negative.

Observation 8. *$S \leq 2e^{-n\epsilon}$, and it could be as large as $2e^{-n\epsilon}$ using Example 1.*

Observation 9. *Define*

$$p_{min} \triangleq \min_{i \in [k]} p_i.$$

Then

$$\max_{\text{empirical distributions } \hat{P}_{n,k}} D(\hat{P}_{n,k} \| P) \leq \max_{Q \in \mathcal{M}_k} D(Q \| P) = \log \frac{1}{p_{min}}.$$

We can see this because $D(Q\|P)$ is convex in Q over a compact convex set, $\mathcal{M}_k$ and hence must attain its maxima at an extreme point of $\mathcal{M}_k$. This means that the maxima must be attained at a distribution which puts all its mass in one spot. Of all these k distributions, the one which puts all its mass on p_{min} maximizes $D(Q\|P)$.

So far we have been agnostic to what the actual distribution P is. Since $\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right)$ is invariant under permutations of the support set, we might as well compute this quantity with any permutation of our choice. So assume that $p_{\min} = p_k$. This means we choose to condition on the value taken by the outcome which has least probability.
Using Observation 9, we get that

$$\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) = 0,$$

if $\epsilon \geq \log \frac{1}{p_k}$. Hence we only need to consider the following for $\epsilon \leq \log \frac{1}{p_k}$.

$$\begin{aligned} T &= \mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon | n\hat{p}_k = n\right) \mathbb{P}\left(n\hat{p}_k = n\right) \\ &= \mathbb{P}\left(\log \frac{1}{p_k} \geq \epsilon\right) \cdot p_k^n \leq 1 \cdot p_k^n = e^{-n \log \frac{1}{p_k}} \\ &\leq e^{-n\epsilon}. \end{aligned} \tag{33}$$

This holds for all ϵ because when $\epsilon \geq \log \frac{1}{p_k}$, $T = 0$.

We have bounded both S and T, and now turn to R_k from Equation 32.

B.2 Using $k = 3$ to understand the behaviour of R_k

Let $k = 3$. Using Equation 3 we have, using the definition of $A_{\frac{l}{n}}$ from Equation 30-

$$\begin{aligned} R_3 &= \sum_{l \in [n] \setminus \mathcal{E}} \mathbb{P}\left(D(\hat{P}_{n-l,k-1}\|\tilde{P}) \geq \frac{\epsilon - A_{\frac{l}{n}}}{1 - \frac{l}{n}}\right) \mathbb{P}\left(n\hat{p}_k = l\right) \\ &= \sum_{l \in [n] \setminus \mathcal{E}} \mathbb{P}\left(D(\hat{P}_{n-l,2}\|\tilde{P}) \geq \frac{\epsilon - A_{\frac{l}{n}}}{1 - \frac{l}{n}}\right) \mathbb{P}\left(n\hat{p}_k = l\right) \\ &\leq \sum_{l \in [n] \setminus \mathcal{E}} 2e^{-(n-l)\left(\frac{\epsilon - A_{\frac{l}{n}}}{1-\frac{l}{n}}\right)} \mathbb{P}\left(n\hat{p}_k = l\right) \\ &= \sum_{l \in [n] \setminus \mathcal{E}} 2e^{-n\epsilon} \{e^{nD((\frac{l}{n}, 1-\frac{l}{n})\|(p_k, 1-p_k))} \mathbb{P}\left(n\hat{p}_k = l\right)\} \end{aligned} \tag{34}$$

$$
\begin{aligned}
&\leq \sum_{l=0}^{n} 2e^{-n\epsilon}\{e^{nD((\frac{l}{n},1-\frac{l}{n})\|(p_k,1-p_k))}\mathbb{P}\left(n\hat{p}_k = l\right)\} \qquad (35)\\
&= 2e^{-n\epsilon}\underbrace{\mathbb{E}_{X\sim \mathsf{B}(n,p_k)}[e^{nD((\frac{X}{n},1-\frac{X}{n})\|(p_k,1-p_k))}]}_{:=E}\\
&\leq 2e^{-n\epsilon}(\frac{e}{2}\sqrt{n}) = 2e^{-n\epsilon}\frac{ec_0}{2\pi}\sqrt{n}.
\end{aligned}
$$

In the last inequality we need the expectation E which we upper bound in Lemma 5 in Section B.4 and use above.
Hence using Equation 32 and our results for R_3 (above), S (Observation 8) and T (Equation 33), we get

$$
\boxed{\mathbb{P}\left(D(\hat{P}_{n,3}\|P) \geq \epsilon\right) \leq 2e^{-n\epsilon}(\frac{e}{2}\sqrt{n}) + 3e^{-n\epsilon} = e^{-n\epsilon}\left[3\left(1 + K_0\frac{e\sqrt{n}}{2\pi}\right)\right].} \qquad (36)
$$

Here c_0 and K_0 are as defined in Equations 13 and 14, and this result gives Theorem 3 for the special case $k = 3$ (Note that the statement of Theorem 3 has an extra factor of $\frac{c_1}{c_2}$ for reasons that will become clearer later).

B.3 The case for general k

In what follows, c_m and K_m are defined as in Equations 13 and 14. Also, here we define the following

$$
h_m \triangleq \begin{cases} c_m & m \neq 2 \\ c_1 & m = 2 \end{cases}, \qquad (37)
$$

$$
H_m \triangleq \prod_{j=0}^{m} h_j. \qquad (38)
$$

We will also later use the fact that

$$
H_m \leq \frac{c_1}{c_2}K_m \text{ for } \ m \geq 0. \qquad (39)
$$

We have shown in Equation 36 that

$$
\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) \leq e^{-n\epsilon}\left[3\sum_{i=0}^{k-2} H_{i-1}(\frac{e\sqrt{n}}{2\pi})^i\right]
$$

holds for $k = 3$. So we will induct assuming it is true for $k-1$, and hope to show it for value k.

Using Equation 34, we have, using the definition of $A_{\frac{l}{n}}$ from Equation 30-

$$
\begin{aligned}
R_k &= \sum_{l\in[n]\setminus\mathcal{E}} \mathbb{P}\left(D(\hat{P}_{n-l,k-1}\|\tilde{P}) \geq \frac{\epsilon - A_{\frac{l}{n}}}{1-\frac{l}{n}}\right)\mathbb{P}\left(n\hat{p}_k = l\right)\\
&\leq \sum_{l=0}^{n} \mathbb{P}\left(D(\hat{P}_{n-l,k-1}\|\tilde{P}) \geq \frac{\epsilon - A_{\frac{l}{n}}}{1-\frac{l}{n}}\right)\mathbb{P}\left(n\hat{p}_k = l\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{l=0}^{n} e^{-(n-l)(\frac{\epsilon - A\frac{l}{n}}{1-\frac{l}{n}})} \left[3\sum_{i=0}^{k-3} H_{i-1}(\frac{e\sqrt{n-l}}{2\pi})^i \right] \mathbb{P}\left(n\hat{p}_k = l\right) && (40)\\
&= e^{-n\epsilon} \left[3\sum_{i=0}^{k-3} H_{i-1}(\frac{e\sqrt{n}}{2\pi})^i \left(\sum_{l=0}^{n} \left(\sqrt{1-\frac{l}{n}}\right)^i e^{nA\frac{l}{n}} \mathbb{P}\left(n\hat{p}_k = l\right) \right) \right] \\
&= e^{-n\epsilon} \left[3\sum_{i=0}^{k-3} H_{i-1}(\frac{e\sqrt{n}}{2\pi})^i \underbrace{\mathbb{E}_{X\sim \mathsf{B}(n,p_k)} \left[\left(\sqrt{1-\frac{X}{n}}\right)^i e^{nD((\frac{X}{n},1-\frac{X}{n})\|(p_k,1-p_k))} \right]}_{:=E_i} \right] && (41)\\
&\leq e^{-n\epsilon} \left[3\sum_{i=0}^{k-3} H_{i-1}(\frac{e\sqrt{n}}{2\pi})^i \left(h_i \frac{e\sqrt{n}}{2\pi} \right) \right] \\
&\leq e^{-n\epsilon} \left[3\sum_{i=1}^{k-2} H_{i-1}(\frac{e\sqrt{n}}{2\pi})^i \right].
\end{aligned}
$$

In Inequality 40, we have used our inductive assumption

$$
\mathbb{P}\left(D(\hat{P}_{n,k-1}\|P) \geq \epsilon \right) \leq e^{-n\epsilon} \left[3\sum_{i=0}^{k-3} H_{i-1}(\frac{c_0}{c_2}\frac{e\sqrt{n}}{2\pi})^i \right],
$$

and we have upper bounded the expectation E_i by $\left(h_i \frac{e\sqrt{n}}{2\pi}\right)$ in Lemma 5 in Section B.4.
Hence using Equation 32 and our upper bounds for R_k (above), S (Observation 8) and T (Equation 33), we get (with Equation 39)

$$
\begin{aligned}
\mathbb{P}\left(D(\hat{P}_{n,3}\|P) \geq \epsilon \right) &\leq e^{-n\epsilon} \left[3\sum_{i=1}^{k-2} H_{i-1}(\frac{e\sqrt{n}}{2\pi})^i \right] + 3e^{-n\epsilon} \\
&= e^{-n\epsilon} \left[3\sum_{i=0}^{k-2} H_{i-1}(\frac{e\sqrt{n}}{2\pi})^i \right] \\
&\leq e^{-n\epsilon} \left[\frac{3c_1}{c_2} \sum_{i=0}^{k-2} K_{i-1}(\frac{e\sqrt{n}}{2\pi})^i \right].
\end{aligned}
$$

This proves our inductive step, and using upper bounds on K_m from Equation 14, this completes the proof of Equation 15 in Theorem 3. To get more interpretable versions of this bound, we do the following.
Using the fact $K_m \leq \sqrt{\frac{d_0}{m}} \left(\sqrt{\frac{2\pi e}{m}}\right)^m$ in conjunction with this result, we obtain

$$
\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon \right) \leq e^{-n\epsilon} \left[\frac{3c_1}{c_2} \sqrt{\frac{d_0}{2\pi e}} \left(\sum_{i=1}^{k-2} \left(\sqrt{\frac{e^3 n}{2\pi i}} \right)^i + 1 \right) \right].
$$

Recalling that $C_1 = \frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}}$ and $C_0 = (\frac{e^3}{2\pi})$, the rest of the piecewise bounds in Theorem 3 follow straightforwardly after making the following observations-

- If $k \leq \sqrt{nC_0} + 2$, then $k \leq \frac{n}{k} \implies k\left(\sqrt{\frac{C_0 n}{k}}\right)^{k-2} \leq \left(\sqrt{\frac{C_0 n}{k}}\right)^{k}$.
- The function $\left(\sqrt{\frac{C_0 n}{i}}\right)^{i}$ is maximized at $i = \frac{C_0 n}{e}$.
- The function $\left(\sqrt{\frac{C_0 n}{i}}\right)^{i} \leq 1$ for $i \geq C_0 n$.
- $\left(\frac{k}{k-2}\right)^{k-2} \leq e^2$ for $k \geq 3$.

$$
\begin{aligned}
&\mathbb{P}\left(D(\hat{P}_{n,k} \| P) \geq \epsilon\right) \\
&\leq e^{-n\epsilon}\left[\frac{3c_1}{c_2}\sum_{i=0}^{k-2} K_{i-1}(\frac{e\sqrt{n}}{2\pi})^{i}\right] \leq e^{-n\epsilon}\left[\frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}}\left(\sum_{i=1}^{k-2}\left(\sqrt{\frac{e^3 n}{2\pi i}}\right)^{i} + 1\right)\right] \\
&\leq \begin{cases}
e^{-n\epsilon}\left[\frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}}(k-2)\left(\sqrt{\frac{e^3 n}{2\pi(k-2)}}\right)^{k-2}\right] \leq \boxed{C_1 e\left(\sqrt{\frac{C_0 n}{k}}\right)^{k} e^{-n\epsilon}} & 3 \leq k \leq \sqrt{nC_0} + 2 \\
e^{-n\epsilon}\left[\frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}}(k-2)\left(\sqrt{\frac{e^3 n}{2\pi(k-2)}}\right)^{k-2}\right] \leq \boxed{C_1 k\left(\sqrt{\frac{C_0 n}{k}}\right)^{k} e^{-n\epsilon}} & 3 \leq k \leq \frac{nC_0}{e} + 2 \\
e^{-n\epsilon}\left[\frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}}(k-2)e^{\frac{n(\frac{e^3}{2\pi})}{2e}}\right] \leq \boxed{C_1 k e^{\frac{C_0 n}{2e}} e^{-n\epsilon}} & \frac{nC_0}{e} + 2 \leq k \leq nC_0 + 2 \\
e^{-n\epsilon}\left[\frac{3c_1}{c_2}\sqrt{\frac{d_0}{2\pi e}}\left(nC_0 e^{\frac{n(\frac{e^3}{2\pi})}{2e}} + k - 2 - nC_0\right)\right] \leq \boxed{C_1\left(nC_0 e^{\frac{C_0 n}{2e}} + k\right)e^{-n\epsilon}} & k \geq nC_0 + 2
\end{cases}
\end{aligned}
$$

This completes the proof of Theorem 3.

B.4 Computing the expectation of the exponential of the KL Divergence in the binary alphabet

Note: Because we use $0 \log 0 = 0$, we also use $0^0 = 1$.

In this section, we prove the following upper bound on E_i (defined in Equation 41) which has been used in Sections B.2 and B.3.

Lemma 5.

$$
E_i = \mathbb{E}_{X \sim \mathsf{B}(n,p_k)}\left[\left(\sqrt{1 - \frac{X}{n}}\right)^{i} e^{nD((\frac{X}{n}, 1-\frac{X}{n}) \| (p_k, 1-p_k))}\right] \leq h_i \frac{e\sqrt{n}}{2\pi}
$$

Proof. Below, we use the following Stirling Approximation that is valid for all integers n on each of the 3 factorials involved in $\binom{n}{l}$.

$$
\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \leq n! \leq e n^{n+\frac{1}{2}} e^{-n}. \tag{42}
$$

$$
\begin{aligned}
E_i &= \mathbb{E}_{X\sim \mathsf{B}(n,p_k)}\left[\left(\sqrt{1-\frac{X}{n}}\right)^i e^{nD((\frac{X}{n},1-\frac{X}{n})\|(p_k,1-p_k))}\right] \\
&= \sum_{l=0}^{n}\binom{n}{l}p_k^l(1-p_k)^{n-l}e^{n(\frac{l}{n}\log\frac{l}{np_k}+\frac{n-l}{n}\log\frac{n-l}{n(1-p_k)})}\left(\sqrt{1-\frac{l}{n}}\right)^i \\
&= \sum_{l=0}^{n}\binom{n}{l}p_k^l(1-p_k)^{n-l}\frac{l^l}{n^l p_k^l}\frac{(n-l)^{n-l}}{n^{n-l}(1-p_k)^{n-l}}\left(\sqrt{1-\frac{l}{n}}\right)^i \\
&= \sum_{l=0}^{n}\binom{n}{l}\frac{l^l(n-l)^{n-l}}{n^n}\left(\sqrt{1-\frac{l}{n}}\right)^i \\
&\le \sum_{l=0}^{n}\frac{en^{n+\frac{1}{2}}e^{-n}}{\sqrt{2\pi}l^{l+\frac{1}{2}}e^{-l}\sqrt{2\pi}(n-l)^{n-l+\frac{1}{2}}e^{-(n-l)}}\frac{l^l(n-l)^{n-l}}{n^n}\left(\sqrt{1-\frac{l}{n}}\right)^i \\
&= \frac{e}{2\pi}\sum_{l=0}^{n}\frac{\sqrt{n}}{\sqrt{l}\sqrt{n-l}}\left(\sqrt{1-\frac{l}{n}}\right)^i \\
&= \frac{e\sqrt{n}}{2\pi}\sum_{l=0}^{n}\frac{\left(\sqrt{1-\frac{l}{n}}\right)^i}{\sqrt{\frac{l}{n}}\sqrt{1-\frac{l}{n}}}\frac{1}{n}
\end{aligned}
$$

For all non-negative integers $i \neq 2$, we can now use Lemmas 9 and 10 from Appendix C and certain definite integrals from Section C.1 to upper bound this sum.

$$
E_i \le \frac{e\sqrt{n}}{2\pi}\sum_{l=0}^{n}\frac{\left(\sqrt{1-\frac{l}{n}}\right)^i}{\sqrt{\frac{l}{n}}\sqrt{1-\frac{l}{n}}}\frac{1}{n} \le \frac{e\sqrt{n}}{2\pi}\int_0^1\frac{(1-x)^{\frac{i}{2}}}{\sqrt{x-x^2}}dx = c_i\frac{e\sqrt{n}}{2\pi} \le h_i\frac{e\sqrt{n}}{2\pi}
$$

For $i=2$, we observe that $E_2 \le E_1 \le c_1\frac{e\sqrt{n}}{2\pi} = h_2\frac{e\sqrt{n}}{2\pi}$. This completes the proof of Lemma 5. □

C Auxiliary lemmas

Lemma 6. *For any distribution $P \in \mathcal{M}_k$ and any subset $F \subseteq \mathcal{M}_k$, given n iid samples from P with $\hat{P}_{n,k} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k)$ denoting the empirical distribution, we have*

$$
\mathbb{P}\left(\hat{P}_{n,k} \in F\right) \le 2(k-1)e^{-n\frac{\inf_{P'\in F} D(P'\|P)}{k-1}} \tag{43}
$$

Proof. First, it is clear that

$$
\mathbb{P}\left(\hat{P}_{n,k} \in F\right) \le \mathbb{P}\left(D(\hat{P}_{n,k}\|P) \ge \inf_{P'\in F} D(P'\|P)\right).
$$

Hence we only need to focus on upper bounds for $\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \ge \epsilon\right)$ for some fixed $\epsilon = \inf_{P'\in F} D(P'\|P)$. The structure of this proof is the same as that of Theorem 3. We decompose $D(\hat{P}_{n,k}\|P)$ as in Equation 30, use the law of total probability and then use induction. The only difference is that now

we use a different inductive hypothesis and bound the terms differently. Our inductive hypothesis is that for all $\tilde{P} \in \mathcal{M}_{k-1}$ $(k \geq 3)$ and all positive integers n we have

$$\mathbb{P}\left(D(\hat{P}_{n,k-1}\|\tilde{P}) \geq \epsilon\right) \leq 2(k-2)e^{-\frac{n\epsilon}{k-2}}.$$

The base case for $k = 3$ is immediate from Lemma 8. Using Equation 30 we then have

$$\begin{aligned}
\mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon\right) &= \mathbb{P}\left(D(\hat{P}_{n,k}\|P) \geq \epsilon | n\hat{p}_k = n\right)\mathbb{P}\left(n\hat{p}_k = n\right) \\
&\quad + \mathbb{P}\left(D((\hat{p}_k, 1-\hat{p}_k)\|(p_k, 1-p_k)) + (1-\hat{p}_k)D(\hat{P}_{n(1-\hat{p}_k),k-1}\|\tilde{P}) \geq \epsilon | n\hat{p}_k \neq n\right)\mathbb{P}\left(n\hat{p}_k \neq n\right) \\
&\leq \mathbb{P}\left(D((\hat{p}_k, 1-\hat{p}_k)\|(p_k, 1-p_k)) \geq q_k\epsilon | n\hat{p}_k = n\right)\mathbb{P}\left(n\hat{p}_k = n\right) \\
&\quad + \mathbb{P}\left(D((\hat{p}_k, 1-\hat{p}_k)\|(p_k, 1-p_k)) \geq q_k\epsilon | n\hat{p}_k \neq n\right)\mathbb{P}\left(n\hat{p}_k \neq n\right) \\
&\quad + \mathbb{P}\left((1-\hat{p}_k)D(\hat{P}_{n(1-\hat{p}_k),k-1}\|\tilde{P}) \geq (1-q_k)\epsilon | n\hat{p}_k \neq n\right)\mathbb{P}\left(n\hat{p}_k \neq n\right) \\
&\leq \mathbb{P}\left(D((\hat{p}_k, 1-\hat{p}_k)\|(p_k, 1-p_k)) \geq q_k\epsilon\right) + \mathbb{P}\left((1-\hat{p}_k)D(\hat{P}_{n(1-\hat{p}_k),k-1}\|\tilde{P}) \geq (1-q_k)\epsilon\right) \\
&\leq 2e^{-nq_k\epsilon} + 2(k-2)e^{\frac{n(1-\hat{p}_k)(1-q_k)\epsilon}{(1-\hat{p}_k)k-2}} \\
&\leq 2(k-1)e^{-\frac{n\epsilon}{k-1}}.
\end{aligned}$$

Above we have used Lemma 8, the inductive hypothesis, and eventually set $q_k = \frac{1}{k-1}$. This completes the proof of Lemma 6. □

Lemma 7. *Let $\hat{P}_{n,k} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k)$ be the empirical distribution. Denote*

$$X_i = \frac{(\hat{p}_i - p_i)^2}{p_i} - \frac{1-p_i}{n}.$$

Then, for any $i \neq j$,

$$\begin{aligned}
\mathbb{E}[X_i^2] &= \frac{(1-p_i)(1+2(n-3)p_i(1-p_i))}{n^3 p_i}, \\
\mathbb{E}[X_i X_j] &= \frac{2(p_i+p_j) - 1 + 2(n-3)p_i p_j}{n^3}.
\end{aligned}$$

Proof. First, let us set up some notation. Here, $\hat{p}_i$ is distributed as $\frac{\mathsf{B}(n,p_i)}{n}$. Let $\mathsf{B}(n,p_i) = \sum_{l=1}^n Z_l^i$ where Z_l^i's are i.i.d. Bernoulli(p_i) random variables. Similarly, $\hat{p}_j$ is distributed as $\frac{\mathsf{B}(n,p_j)}{n}$. Let $\mathsf{B}(n,p_i) = \sum_{l=1}^n Z_m^j$ where Z_m^j's are i.i.d. Bernoulli(p_j) random variables. The correlation between Z_l^i's and Z_m^j's is the obvious one inherited from the fact that they are part of a multinomial distribution. In particular, this is their joint distribution-

$$\mathbb{P}\left((Z_l^i, Z_m^j) = (z_1, z_2)\right) = \begin{cases} (1-p_i-p_j)\delta_{lm} & (z_1,z_2) = (0,0) \\ p_i\delta_{lm} & (z_1,z_2) = (1,0) \\ p_j\delta_{lm} & (z_1,z_2) = (0,1) \\ 0 & \text{otherwise} \end{cases},$$

where δ_{lm} is the Kronecker delta.
We first note, using a rearrangement of terms and the fact about binomial distributions that

$\mathbb{E}[(\hat{p}_i - p_i)^2] = \frac{p_i(1-p_i)}{n}$, that the two statements in Lemma 7 are equivalent to the following two statements-

$$\mathbb{E}\left[(n\hat{p}_i - np_i)^4\right] = \mathbb{E}\left[\left(\sum_{l=1}^{n}\left(Z_l^i - p_i\right)\right)^4\right] = np_i(1-p_i) + 3n(n-2)p_i^2(1-p_i)^2 \tag{44}$$

and

$$\begin{aligned}\mathbb{E}\left[(n\hat{p}_i - np_i)^2(n\hat{p}_j - np_j)^2\right] &= \mathbb{E}\left[\left(\sum_{l=1}^{n}\left(Z_l^i - p_i\right)\right)^2\left(\sum_{m=1}^{n}\left(Z_m^j - p_j\right)\right)^2\right] \\ &= np_ip_j\left[(n-1) - (n-2)(p_i+p_j) + (3n-6)p_ip_j\right]. \end{aligned}\tag{45}$$

We will prove both these statements by induction on n.
First we prove Equation 44. The base case for $n = 1$ is the following

$$\mathbb{E}\left[\left(Z_1^i - p_i\right)^4\right] = p_i(1-p_i)^4 + (1-p_i)p_i^4 = p_i(1-p_i)(1-3p_i(1-p_i)).$$

We now make the inductive hypothesis that

$$\mathbb{E}\left[\left(\sum_{l=1}^{n-1}\left(Z_l^i - p_i\right)\right)^4\right] = (n-1)p_i(1-p_i) + 3(n-1)(n-3)p_i^2(1-p_i)^2.$$

Our inductive step then follows as

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{l=1}^{n}\left(Z_l^i - p_i\right)\right)^4\right] &= \mathbb{E}\left[\left(\left(\sum_{l=1}^{n-1}\left(Z_l^i - p_i\right) + \left(Z_n^i - p_i\right)\right)\right)^4\right] \\ &= \underbrace{\mathbb{E}\left[\left(\sum_{l=1}^{n-1}\left(Z_l^i - p_i\right)\right)^4\right]}_{(n-1)p_i(1-p_i)+3(n-1)(n-3)p_i^2(1-p_i)^2} + \underbrace{4\mathbb{E}\left[\left(\sum_{l=1}^{n}\left(Z_l^i - p_i\right)\right)^3\left(Z_n^i - p_i\right)\right]}_{0} \\ &\quad + \underbrace{6\mathbb{E}\left[\left(\sum_{l=1}^{n}\left(Z_l^i - p_i\right)\right)^2\left(Z_n^i - p_i\right)^2\right]}_{6(n-1)p_i^2(1-p_i)^2} + \underbrace{4\mathbb{E}\left[\left(\sum_{l=1}^{n}\left(Z_l^i - p_i\right)\right)\left(Z_n^i - p_i\right)^3\right]}_{0} \\ &\quad + \underbrace{\mathbb{E}\left[\left(Z_n^i - p_i\right)^4\right]}_{p_i(1-p_i)(1-3p_i(1-p_i))} \\ &= np_i(1-p_i) + 3n(n-2)p_i^2(1-p_i)^2.\end{aligned}$$

There are five terms of interest above in the expansion. In the 1st term, we use our inductive hypothesis. In the 2nd, 3rd, and 4th terms we use the fact that Z_n^i is uncorrelated with all the other Z_l^i's to split the expectation of the product into a product of expectations, each of which is simply a binomial random variable mean or variance computation. In the 5th term we use our base case. This completes the proof of Equation 44.
We now begin proving Equation 45. The base case for $n = 1$ follows as

$$\mathbb{E}\left[\left(Z_1^i - p_i\right)^2\left(Z_1^j - p_j\right)^2\right] = p_i(1-p_i)^2p_j^2 + p_j(1-p_j)^2p_i^2 + (1-p_i-p_j)p_i^2p_j^2 = p_ip_j\left[(p_i+p_j) - 3p_ip_j\right],$$

where we have used the joint distribution of Z_1^i and Z_1^j as specified above. We assume our inductive hypothesis that

$$\mathbb{E}\left[\left(\sum_{l=1}^{n-1}\left(Z_l^i-p_i\right)\right)^2\left(\sum_{m=1}^{n-1}\left(Z_m^j-p_j\right)\right)^2\right]=(n-1)p_ip_j\left[(n-2)-(n-3)(p_i+p_j)+(3n-9)p_ip_j\right].$$

Our inductive step then follows as

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{l=1}^{n}\left(Z_l^i-p_i\right)\right)^2\left(\sum_{m=1}^{n}\left(Z_m^j-p_j\right)\right)^2\right]\\
&=\mathbb{E}\left[\left(\sum_{l=1}^{n-1}\left(Z_l^i-p_i\right)+\left(Z_n^i-p_i\right)\right)^2\left(\sum_{m=1}^{n-1}\left(Z_m^j-p_j\right)+\left(Z_n^j-p_j\right)\right)^2\right]\\
&=\underbrace{\mathbb{E}\left[\left(\sum_{l=1}^{n-1}\left(Z_l^i-p_i\right)\right)^2\left(\sum_{m=1}^{n-1}\left(Z_m^j-p_j\right)\right)^2\right]}_{(n-1)p_ip_j[(n-2)-(n-3)(p_i+p_j)+(3n-9)p_ip_j]}+\underbrace{\mathbb{E}\left[\left(\sum_{l=1}^{n-1}\left(Z_l^i-p_i\right)\right)^2 2\left(\sum_{m=1}^{n-1}\left(Z_m^j-p_j\right)\right)\left(Z_n^j-p_j\right)\right]}_{0}\\
&\quad+\underbrace{\mathbb{E}\left[\left(\sum_{l=1}^{n-1}\left(Z_l^i-p_i\right)\right)^2\left(Z_n^j-p_j\right)^2\right]}_{(n-1)p_i(1-p_i)p_j(1-p_j)}+\underbrace{\mathbb{E}\left[\left(\sum_{m=1}^{n-1}\left(Z_m^j-p_j\right)\right)^2 2\left(\sum_{l=1}^{n-1}\left(Z_l^i-p_i\right)\right)\left(Z_n^i-p_i\right)\right]}_{0}\\
&\quad+\underbrace{\mathbb{E}\left[4\left(\sum_{m=1}^{n-1}\left(Z_m^j-p_j\right)\right)\left(\sum_{l=1}^{n-1}\left(Z_l^i-p_i\right)\right)\left(Z_n^i-p_i\right)\left(Z_n^j-p_j\right)\right]}_{4(n-1)(-p_ip_j)(-p_ip_j)}+\underbrace{\mathbb{E}\left[2\left(\sum_{l=1}^{n-1}\left(Z_l^i-p_i\right)\right)\left(Z_n^i-p_i\right)\left(Z_n^j-p_j\right)^2\right]}_{0}\\
&\quad+\underbrace{\mathbb{E}\left[\left(\sum_{m=1}^{n-1}\left(Z_m^j-p_j\right)\right)^2\left(Z_n^i-p_i\right)^2\right]}_{(n-1)p_j(1-p_j)p_i(1-p_i)}+\underbrace{\mathbb{E}\left[2\left(\sum_{m=1}^{n-1}\left(Z_m^j-p_j\right)\right)\left(Z_n^j-p_j\right)\left(Z_n^i-p_i\right)^2\right]}_{0}\\
&\quad+\underbrace{\mathbb{E}\left[\left(Z_1^i-p_i\right)^2\left(Z_1^j-p_j\right)^2\right]}_{p_ip_j[(p_i+p_j)-3p_ip_j]}\\
&=np_ip_j\left[(n-1)-(n-2)(p_i+p_j)+(3n-6)p_ip_j\right].
\end{aligned}$$

In the nine terms of the expansion above, we have repeatedly used the fact Z_n^i and Z_n^j are independent of the other Z_l^i's and Z_m^j's to reduce expectations of products to products of expectations. The first term follows from the inductive hypothesis and the last term follows from the base case. The rest of the terms require only the knowledge of the mean and covariance matrix of a multinomial random variable, which are well known.
This completes the proof of Equation 45 and hence of Lemma 7. □

Lemma 8. *Let $k=2$ and $P\in\mathcal{M}_k$ be any probability distribution. Then for all n,ϵ*

$$\mathbb{P}\left(D(\hat{P}_{n,2}\|P)\geq\epsilon\right)\leq 2e^{-n\epsilon}.$$

Proof. Let

$$\hat{P}_{n,2} = (\hat{p}_1, 1-\hat{p}_1), P = (p_1, 1-p_1).$$

$\mathcal{M}_2$ is a line segment and $\{\hat{P}_{n,2} : D(\hat{P}_{n,2}\|P) \geq \epsilon\} \in \mathcal{M}_2$ is a union of two line segments (which are, in particular, convex). In fact,

$$\{\hat{P}_{n,2} : D(\hat{P}_{n,2}\|P) \geq \epsilon\} = \underbrace{\{\hat{P}_{n,2} : D(\hat{P}_{n,2}\|P) \geq \epsilon, \hat{p}_1 > p_1\}}_{:=M_1} \cup \underbrace{\{\hat{P}_{n,2} : D(\hat{P}_{n,2}\|P) \geq \epsilon, \hat{p}_1 < p_1\}}_{:=M_2},$$

and so

$$\begin{aligned}\mathbb{P}\left(D(\hat{P}_{n,2}\|P) \geq \epsilon\right) &= \mathbb{P}\left(M_1 \cup M_2\right) \\ &\leq \mathbb{P}\left(M_1\right) + \mathbb{P}\left(M_2\right) \\ &\leq e^{-n\epsilon} + e^{-n\epsilon} = 2e^{-n\epsilon}.\end{aligned}$$

To bound $\mathbb{P}\left(M_1\right)$ and $\mathbb{P}\left(M_2\right)$ we use the fact that M_1 and M_2 are convex, $\inf_{Q\in M_1} D(Q\|P) = \inf_{Q\in M_2} D(Q\|P) = \epsilon$, and inequality (2.16) in [Csi84, Theorem 1]. □

Lemma 9. *Let $f(x)$ be a convex function on $(0,1)$. Let $l_1 < l_2 < n$ be some integers. Then we must have*

$$\sum_{l=l_1}^{l_2} f\left(\frac{l}{n}\right)\frac{1}{n} \leq \int_{\frac{l_1}{n}}^{\frac{l_2}{n}} f(x)dx.$$

Proof. The proof of this follows from using convexity and Jensen's Inequality to observe that

$$f(\frac{l}{n}) = f(\int_{\frac{l-\frac{1}{2}}{n}}^{\frac{l+\frac{1}{2}}{n}} x \cdot n \cdot dx) \leq n \cdot \int_{\frac{l-\frac{1}{2}}{n}}^{\frac{l+\frac{1}{2}}{n}} f(x)dx,$$

and then simply adding this inequality for all l from l_1 to l_2. □

Lemma 10. *$\frac{(1-x)^{\frac{m}{2}}}{\sqrt{x-x^2}}$ is convex for $x \in (0,1)$ for any non-negative integer m except $m = 2$.*

Proof. Let

$$f(x) = \frac{(1-x)^{\frac{m}{2}}}{\sqrt{x-x^2}}$$

and

$$g(x) = \log f(x).$$

Then

$$f'' = (g'' + g'^2)f.$$

Since $f > 0$ for all $x \in (0,1)$, f is convex iff $(g'' + g'^2) > 0$ for all $x \in (0,1)$.

$$g' = \frac{-\frac{m-1}{2}}{(1-x)} - \frac{1}{2x}, g'' = \frac{-\frac{m-1}{2}}{(1-x)^2} + \frac{1}{2x^2}$$

So,

$$g'' + g'^2 = \frac{3}{4x^2} + \frac{(\frac{m-1}{2})(\frac{m-1}{2}-1)}{(1-x)^2} + \frac{(\frac{m-1}{2})\frac{1}{2}}{x(1-x)}.$$

For $m \neq 0, 2$, observe that all the terms in $g'' + g'^2$ are positive. For $m = 0$, one can verify that though all the terms aren't positive, the resulting function is always positive. □

C.1 Some useful definite integrals

1.

$$c_m \triangleq \int_0^1 \frac{(1-x)^{\frac{m}{2}}}{\sqrt{x-x^2}} dx \tag{46}$$

$$= \int_0^{\frac{\pi}{2}} 2(\cos\theta)^m d\theta \tag{47}$$

$$= \begin{cases} \frac{1\times3\times5\ldots\times m-1}{2\times4\times6\times\ldots\times m} \cdot \pi & m \text{ is even} \\ \frac{2\times4\times6\times\ldots\times m-1}{1\times3\times5\ldots\times m} \cdot 2 & m \text{ is odd} \end{cases} \tag{48}$$

2.

$$\int_{-1}^{1} (\sqrt{1-x^2})^m dx = 2\int_0^{\frac{\pi}{2}} (\cos\theta)^{m+1} d\theta$$

$$= c_{m+1} \tag{49}$$