Check for updates

J. Appl. Probab. **62,** 1578–1593 (2025)
*doi:*10.1017/jpr.2025.27

RELATIVE ENTROPY BOUNDS FOR SAMPLING WITH AND WITHOUT REPLACEMENT

OLIVER JOHNSON,* *University of Bristol*
LAMPROS GAVALAKIS,** *Université Gustave Eiffel*
IOANNIS KONTOYIANNIS,*** *University of Cambridge*

Abstract

Sharp, nonasymptotic bounds are obtained for the relative entropy between the distributions of sampling with and without replacement from an urn with balls of $c \geq 2$ colors. Our bounds are asymptotically tight in certain regimes and, unlike previous results, they depend on the number of balls of each color in the urn. The connection of these results with finite de Finetti-style theorems is explored, and it is observed that a sampling bound due to Stam (1978) combined with the convexity of relative entropy yield a new finite de Finetti bound in relative entropy, which achieves the optimal asymptotic convergence rate.

Keywords: Relative entropy; sampling; replacement; hypergeometric; binomial; de Finetti

2010 Mathematics Subject Classification: Primary 60E05
Secondary 60G09; 94A17

1. Introduction

1.1. The problem

Consider an urn containing n balls, each ball having one of $c \geq 2$ different colors, and suppose we draw out $k \leq n$ of them. We will compare the difference in sampling with and without replacement. Write $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_c)$ for the vector representing the number of balls of each color in the urn, so that there are ℓ_i balls of color i, $1 \leq i \leq c$, and $\ell_1 + \ell_2 + \cdots + \ell_c = n$. Write $\boldsymbol{s} = (s_1, s_2, \ldots, s_c)$ for the number of balls of each color drawn out, so that $s_1 + s_2 + \cdots + s_c = k$.

When sampling without replacement, the probability that the colors of the k balls drawn are given by $\boldsymbol{s}$ is given by the multivariate hypergeometric probability mass function (p.m.f.)

$$H(n, k, \boldsymbol{\ell}; \boldsymbol{s}) := \frac{\prod_{i=1}^{c} \binom{\ell_i}{s_i}}{\binom{n}{k}} = \frac{\binom{k}{\boldsymbol{s}} \binom{n-k}{\boldsymbol{\ell}-\boldsymbol{s}}}{\binom{n}{\boldsymbol{\ell}}}, \tag{1}$$

Received 9 April 2024; accepted 28 February 2025.
* Postal address: School of Mathematics, Fry Building, Woodland Road, Bristol, BS8 1UG, UK. Email: O.Johnson@bristol.ac.uk
** Postal address: Université Gustave Eiffel, Université Paris Est Creteil, CNRS, LAMA UMR8050 F-77447 Marne-la-Vallée, France. Email: lg560@cam.ac.uk
*** Postal address: Statistical Laboratory, DPMMS, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, UK. Email: yiannis@maths.cam.ac.uk

© The Author(s), 2025. Published by Cambridge University Press on behalf of Applied Probability Trust. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

for all $\boldsymbol{s}$ with $0 \le s_i \le \ell_i$ for all i, and $s_1 + \cdots + s_c = k$. In (1) and throughout, we write $\binom{v}{\boldsymbol{u}} = v!/\prod_{i=1}^{c} u_i!$ for the multinomial coefficient for any vector $\boldsymbol{u} = (u_1, \ldots, u_c)$ with $u_1 + \cdots + u_c = v$.

Our goal is to compare $H(n, k, \boldsymbol{\ell}; \boldsymbol{s})$ with the corresponding p.m.f. $B(n, k, \boldsymbol{\ell}; \boldsymbol{s})$ of sampling with replacement, which is given by the multinomial distribution p.m.f.

$$B(n, k, \boldsymbol{\ell}; \boldsymbol{s}) := \binom{k}{\boldsymbol{s}} \prod_{i=1}^{c} \left(\frac{\ell_i}{n}\right)^{s_i} \tag{2}$$

for all $\boldsymbol{s}$ with $s_i \ge 0$ and $s_1 + \cdots + s_c = k$.

The study of the relationship between sampling with and without replacement has a long history and numerous applications in both statistics and probability; see, e.g., the classic review in [22], or [24]. Beyond the elementary observation that B and H have the same mean, namely that $\sum_{\boldsymbol{s}} B(n, k, \boldsymbol{\ell}; \boldsymbol{s}) s_i = \sum_{\boldsymbol{s}} H(n, k, \boldsymbol{\ell}; \boldsymbol{s}) s_i = k\ell_i/n$ for each i, it is well known that in certain limiting regimes the p.m.f.s themselves are close in a variety of senses. For example, [12, Theorem 4] showed that H and B are close in the total variation distance $\|P - Q\| := \sup_A |P(A) - Q(A)|$:

$$\|H(n, k, \boldsymbol{\ell}; \cdot) - B(n, k, \boldsymbol{\ell}; \cdot)\| \le \frac{ck}{n}. \tag{3}$$

In this paper we give bounds on the relative entropy (or Kullback–Leibler divergence) between H and B, which for brevity we denote as

$$D(n, k, \boldsymbol{\ell}) := D(H(n, k, \boldsymbol{\ell}; \cdot) \,\|\, B(n, k, \boldsymbol{\ell}; \cdot)) = \sum_{\boldsymbol{s}} H(n, k, \boldsymbol{\ell}; \boldsymbol{s}) \log\left(\frac{H(n, k, \boldsymbol{\ell}; \boldsymbol{s})}{B(n, k, \boldsymbol{\ell}; \boldsymbol{s})}\right). \tag{4}$$

[All logarithms are natural logarithms to base e.] Clearly, if $\ell_r = 0$ for some r, then bounding $D(n, k; \boldsymbol{\ell})$ becomes equivalent to the same problem with a smaller number of colors c, so we may assume that each $\ell_r \ge 1$ for simplicity.

Interest in controlling the relative entropy derives in part from the fact that it naturally arises in many core statistical problems, e.g. as the optimal hypothesis testing error exponent in Stein's lemma; see, e.g., [9, Theorem 12.8.1]. Further, bounding the relative entropy also provides control of the distance between H and B in other senses. For example, Pinsker's inequality, e.g. [26, Lemma 2.9.1], states that, for any two p.m.f.s P and Q,

$$2\|P - Q\|^2 \le D(P \,\|\, Q), \tag{5}$$

while the Bretagnolle–Huber bound [7, 8] gives $\|P - Q\|^2 \le 1 - \exp\{-D(P \,\|\, Q)\}$.

Our results also follow a long line of work that has been devoted to establishing probabilistic theorems in terms of relative entropy. Information-theoretic arguments often provide insights into the underlying reasons why the result at hand holds. Among many others, examples include the well-known work of [4] on the information-theoretic central limit theorem; information-theoretic proofs of Poisson convergence and Poisson approximation [20]; compound Poisson approximation bounds [3]; convergence to Haar measure on compact groups [16]; connections with extreme value theory [18]; and the discrete central limit theorem [15]. We refer to [14] for a detailed review of this line of work.

The relative entropy in (4) has been studied before. Stam [23] established the bound

$$D(n, k, \boldsymbol{\ell}) \le \frac{(c-1)k(k-1)}{2(n-1)(n-k+1)} \tag{6}$$

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

uniformly in $\boldsymbol{\ell}$, and also provided an asymptotically matching lower bound,

$$D(n, k, \boldsymbol{\ell}) \geq \frac{(c-1)k(k-1)}{4(n-1)^2}, \tag{7}$$

indicating that, in the regime $k = o(n)$ and fixed c, the relative entropy $D(n, k, \boldsymbol{\ell})$ is of order $(c-1)k^2/2n^2$.

More recently, related bounds were established by [17], showing that (see [17, Theorem 4.5])

$$D(n, k, \boldsymbol{\ell}) \leq (c-1)\left(\log\left(\frac{n-1}{n-k}\right) - \frac{k}{n} + \frac{1}{n-k+1}\right), \tag{8}$$

and (see [17, Theorem 4.4])

$$D(n, k, \boldsymbol{\ell}) \geq \frac{(c-1)}{2}(r - 1 - \ln r) \quad \text{for } r = \frac{n-k+1}{n-1}, \tag{9}$$

both results also holding uniformly in $\boldsymbol{\ell}$. Moreover, in the regime where $k/n \to s$ and $0 < \epsilon \leq \ell_r/n$ for all r, [17] showed that the lower bound in (9) is asymptotically sharp and the upper bound in (8) is sharp to within a factor of two by proving that

$$\lim_{n\to\infty} D(n, k, \boldsymbol{\ell}) = \frac{c-1}{2}\Big(-\log(1-s) - s\Big) \approx (c-1)\left(\frac{s^2}{4} + \frac{s^3}{6}\right). \tag{10}$$

1.2. Main results

Despite the fact that (6) and (8) are both asymptotically optimal in the sense described above, it turns out it is possible to obtain more accurate upper bounds on $D(n, k, \boldsymbol{\ell})$ if we allow them to depend on $\boldsymbol{\ell}$. Indeed, as remarked in [17], 'a good upper bound should depend on $\boldsymbol{\ell}$" (in our notation). In this vein, our main result is the following; it is proved in Section 2.2.

Theorem 1. *For any $1 \leq k \leq n/2$, $c \geq 2$, and $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_c)$ with $\ell_1 + \cdots + \ell_c = n$ and $\ell_r \geq 1$ for each r, the relative entropy $D(n, k, \boldsymbol{\ell})$ between H and B satisfies*

$$\begin{aligned} D(n, k, \boldsymbol{\ell}) \leq{}& \frac{c-1}{2}\left(\log\left(\frac{n}{n-k}\right) - \frac{k}{n-1}\right) \\ &+ \frac{k(2n+1)}{12n(n-1)(n-k)} \sum_{i=1}^{c} \frac{n}{\ell_i} + \frac{1}{360}\left(\frac{1}{(n-k)^3} - \frac{1}{n^3}\right) \sum_{i=1}^{c} \frac{n^3}{\ell_i^3}. \end{aligned} \tag{11}$$

While the bound of Theorem 1 is simple, straightforward, and works well for the balanced cases described below, it is least accurate in the case of small ℓ_i; see Figure 1. We therefore also provide an alternative expression in Proposition 1. For simplicity, we only give the result in the case $c = 2$, though similar arguments will work in general, see Remark 6. It is proved in Section 3.

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

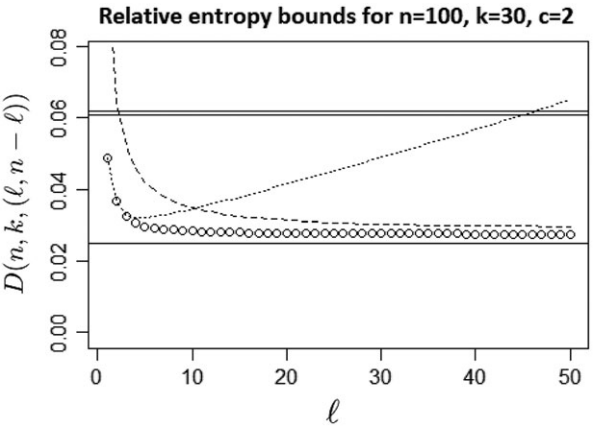


FIGURE 1. Comparison of the true values of $D(100, 30, (\ell, 100-\ell))$, plotted as circles, with the uniform upper bounds in (6), (8), and the uniform lower bound (9), all plotted as straight lines, and also with the new bounds in Theorem 1 (dashed line) and Proposition 1 (dotted line). The combination of the two new bounds outperforms those of [17, 23] for the whole range of the values of ℓ in this case.

Proposition 1. *Under the assumptions of Theorem 1, if* $c = 2$ *and* $\ell \le n/2$,

$$\begin{aligned} D(n, k, (\ell, n-\ell)) \le \ell \Bigg[\left(1 - \frac{k}{n}\right) \log \left(1 - \frac{k}{n}\right) + \frac{k}{n} - \frac{k}{2n(n-1)} \Bigg] \\ + \frac{k\ell}{(n-1)(n-\ell)(n-k)} \\ - \frac{k(k-1)}{2n(n-1)} \ell(\ell-1) \log \left(\frac{\ell}{\ell-1}\right) \\ - \frac{k(k-1)(k-2)}{6n(n-1)(n-2)} \ell(\ell-1)(\ell-2) \log \left(\frac{(\ell-1)^2}{\ell(\ell-2)}\right). \end{aligned}$$

Figure 1 shows a comparison of the earlier uniform bounds (6), (7), and (8) with the new bounds in Theorem 1 and Proposition 1 for the case $n = 100$, $k = 30$, and $\boldsymbol{\ell} = (\ell, 1-\ell)$, for $1 \le \ell \le 50$. It can be seen that the combination of the two new upper bounds outperforms the earlier ones over the entire range of values considered.

Remark 1. The dependence of the bound (11) in Theorem 1 on $\boldsymbol{\ell}$ is only via the quantities

$$\Sigma_1(n, c, \boldsymbol{\ell}) := \sum_{r=1}^{c} \frac{n}{\ell_r}, \qquad \Sigma_2(n, c, \boldsymbol{\ell}) := \sum_{r=1}^{c} \frac{n^3}{\ell_r^3}. \tag{12}$$

For Σ_1, in the 'balanced' case where each $\ell_r = n/c$, we have $\Sigma_1(n, c, \boldsymbol{\ell}) = c^2$, and using Jensen's inequality it is easy to see that we always have $\Sigma_1(n, c, \boldsymbol{\ell}) \ge c^2$. In fact, the Schur–Ostrowski criterion shows that (12) is Schur convex, and hence respects the majorization order, being largest in 'unbalanced' cases. For example, if $\ell_1 = \cdots = \ell_{c-1} = 1$ and $\ell_c = n - c + 1$, then $\Sigma_1(n, c, \boldsymbol{\ell}) = (c-1)n + 1/(n-c+1) \sim n(c-1) \to \infty$ as $n \to \infty$. On the other hand, in the regime considered in [17] where all $\ell_r \ge \epsilon n$ for some $0 < \epsilon \le 1/c$, we have $\Sigma_1(n, c, \boldsymbol{\ell}) \le c/\epsilon$, which is bounded in n.

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

Similar comments apply to $\Sigma_2(n, c, \boldsymbol{\ell})$: it is bounded in balanced cases, including the scenario of [17], but can grow like $n^3(c-1)$ in the very unbalanced case.

Remark 2. The first term in the bound (11) of Theorem 1,

$$\frac{c-1}{2}\left(\log\left(\frac{n}{n-k}\right)-\frac{k}{n-1}\right),$$

matches the asymptotic expression in (10), so Theorem 1 may be regarded as a nonasymptotic version version of (10). In particular, (10) is only valid in the asymptotic case where all $\ell_r \geq \epsilon n$, when, as discussed above, $\Sigma_1(n, c, \boldsymbol{\ell})$ is bounded. Therefore, in this balanced case Theorem 1 recovers the asymptotic result of (10).

Remark 3. Any bound that holds uniformly in $\boldsymbol{\ell}$ must hold in particular for the binary $c=2$ case with $\ell_1 = 1$, $\ell_2 = n-1$. Then H is Bernoulli with parameter k/n and B is binomial with parameters k and $1/n$, and direct calculation gives the exact expression

$$D(n, k, (1, n-1)) = \left(1-\frac{k}{n}\right)\log\left(1-\frac{k}{n}\right) - k\left(1-\frac{1}{n}\right)\log\left(1-\frac{1}{n}\right).$$

Taking $n\to\infty$ while $k/n \to s$ for some fixed s, in this unbalanced case where $\ell_1 = 1$ and $\ell_2 = n-1$, we obtain

$$\lim_{n\to\infty} D(n, ns, (1, n-1)) = s + (1-s)\log(1-s) \approx \frac{s^2}{2} + \frac{s^3}{6}. \tag{13}$$

Compared with the analogous limiting expression (10) in the balanced case where ℓ_1, ℓ_2 are both bounded below by ϵn, this suggests that the asymptotic behavior of $D(n, k, \boldsymbol{\ell})$ in the regime $k/n \to s$ may vary by a factor of two over different values of ℓ.

Remark 4. Comparing the earlier asymptotic expression (10) with (13) reveals some interesting behavior of $D(n, k, (\ell, n-\ell))$. There is a unique value $s^* \approx 0.8834$ such that, for $s < s^*$, corresponding to small k, the limiting expression in (13) is larger than that in (10), whereas, for $s > s^*$, the expression in (10) is larger.

1.3. Finite de Finetti theorems

An interesting and intimate connection was revealed in [11, 12] between the problem of comparing sampling with and without replacement and finite versions of de Finetti's theorem. Recall that de Finetti's theorem says that the distribution of an infinite collection of exchangeable random variables can be represented as a mixture over independent and identically distributed (i.i.d.) sequences; see, e.g., [19] for an elementary proof in the binary case. Although such a representation does not always exist for *finite* exchangeable sequences, approximate versions are possible: If $(X_1, \ldots, X_n)$ are exchangeable, then the distribution of $(X_1, \ldots, X_k)$ is close to a mixture of independent sequences as long as k is relatively small compared to n. Indeed, [12] proved such a finite de Finetti theorem using the sampling bound (3).

Let $\mathcal{A} = \{a_1, \ldots, a_c\}$ be an alphabet of size c, and suppose that the random variables $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, with values $X_i \in \mathcal{A}$, are exchangeable, that is, the distribution of $\boldsymbol{X}_n$ is the same as that of $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ for any permutation π on $\{1, \ldots, n\}$. Given a sequence $\boldsymbol{y} \in \mathcal{A}^m$ of length m, its *type* $\mathcal{T}_m(\boldsymbol{y})$ [10] is the vector of its empirical frequencies: the rth component of $\mathcal{T}_m(\boldsymbol{y})$ is $(1/m)\sum_{i=1}^m \mathbf{1}(y_i = a_r)$, $1 \leq r \leq c$. The sequence $\boldsymbol{X}_n$ induces a measure μ on p.m.f.s $\boldsymbol{p}$ on $\mathcal{A}$ via

$$\mu(\boldsymbol{p}) = \mathbb{P}(\mathcal{T}_n(\boldsymbol{X}_n) = \boldsymbol{p}). \tag{14}$$

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

This is the law of the empirical measure induced by $\boldsymbol{X}_n$ on $\mathcal{A}$.

For each $1 \le k \le n$, let P_k denote the joint p.m.f. of $\boldsymbol{X}_k = (X_1, \ldots, X_k)$, and write

$$M_{k,\mu}(z) = \int \boldsymbol{p}^k(z)\, \mathrm{d}\mu(\boldsymbol{p}), \quad z \in \mathcal{A}^k, \tag{15}$$

for the mixture of the i.i.d. distributions $\boldsymbol{p}^k$ with respect to the mixing measure μ. A key step in the connection between sampling bounds and finite de Finetti theorems is the simple observation that, given that its type $\mathcal{T}_k(\boldsymbol{X}_k) = \boldsymbol{q}$, the exchangeable sequence $\boldsymbol{X}_k$ is uniformly distributed on the set of sequences with type $\boldsymbol{q}$. In other words, for any $\boldsymbol{x} \in \mathcal{A}^k$,

$$P_k(\boldsymbol{x}) = \frac{1}{\binom{k}{\boldsymbol{s}}} \sum_{\boldsymbol{p}} H(n, k, n\boldsymbol{p}; \boldsymbol{s})\mu(\boldsymbol{p}), \tag{16}$$

where $\boldsymbol{s} = k\mathcal{T}_k(\boldsymbol{x})$. This elementary and rather obvious (by symmetry) observation already appears implicitly in the literature, e.g. in [11, 14].

An analogous representation can be easily seen to hold for $M_{k,\mu}$. Since the probability of an i.i.d. sequence only depends on its type, for any $\boldsymbol{x} \in \mathcal{A}^k$ we have, with $\boldsymbol{s} = k\mathcal{T}_k(\boldsymbol{x})$,

$$M_{k,\mu}(\boldsymbol{x}) = \frac{1}{\binom{k}{\boldsymbol{s}}} \sum_{\boldsymbol{p}} B(n, k, n\boldsymbol{p}; \boldsymbol{s})\mu(\boldsymbol{p}). \tag{17}$$

The following simple proposition clarifies the connection between finite de Finetti theorems and sampling bounds. Its proof is a simple consequence of (16) combined with (17) and the log-sum inequality [9].

Proposition 2. *Suppose $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ is a finite exchangeable sequence of random variables with values in $\mathcal{A}$, and let μ denote the law of the empirical measure defined by (14). Then, for any $k \le n$,*

$$D(P_k \,\|\, M_{k,\mu}) \le \sum_{\boldsymbol{p}} \mu(\boldsymbol{p}) D(n, k, n\boldsymbol{p}) \le \max_{\boldsymbol{\ell}/n \in \mathrm{supp}(\mu)} D(n, k, \boldsymbol{\ell}),$$

where $M_{k,\mu}$ is the mixture defined in (15) and the maximum is taken over all p.m.f.s of the form $\boldsymbol{\ell}/n$ in the support of μ.

Proof. Fix $\boldsymbol{s}$ with $\sum_{i=1}^{c} s_i = k$. The log-sum inequality gives, for any $\boldsymbol{x}$ of type $\mathcal{T}_k(\boldsymbol{x}) = \boldsymbol{s}/k$,

$$\begin{aligned} P_k(\boldsymbol{x}) \log\left(\frac{P_k(\boldsymbol{x})}{M_{k,\mu}(\boldsymbol{x})}\right) &= \binom{k}{\boldsymbol{s}}^{-1} \left(\sum_{\boldsymbol{p}} H(n, k, n\boldsymbol{p}; \boldsymbol{s})\mu(\boldsymbol{p})\right) \log\left(\frac{\sum_{\boldsymbol{p}} H(n, k, n\boldsymbol{p}; \boldsymbol{s})\mu(\boldsymbol{p})}{\sum_{\boldsymbol{p}} B(n, k, n\boldsymbol{p}; \boldsymbol{s})\mu(\boldsymbol{p})}\right) \\ &\le \binom{k}{\boldsymbol{s}}^{-1} \sum_{\boldsymbol{p}} \mu(\boldsymbol{p}) H(n, k, n\boldsymbol{p}; \boldsymbol{s}) \log\left(\frac{H(n, k, n\boldsymbol{p}; \boldsymbol{s})}{B(n, k, n\boldsymbol{p}; \boldsymbol{s})}\right). \end{aligned}$$

Hence, summing over the $\binom{k}{\boldsymbol{s}}$ vectors $\boldsymbol{x}$ of type $\mathcal{T}_k(\boldsymbol{x}) = \boldsymbol{s}/k$ we obtain

$$\sum_{\boldsymbol{x}: \mathcal{T}(\boldsymbol{x}) = \boldsymbol{s}/k} P_k(\boldsymbol{x}) \log\left(\frac{P_k(\boldsymbol{x})}{M_{k,\mu}(\boldsymbol{x})}\right) \le \sum_{\boldsymbol{p}} \mu(\boldsymbol{p}) H(n, k, n\boldsymbol{p}; \boldsymbol{s}) \log\left(\frac{H(n, k, n\boldsymbol{p}; \boldsymbol{s})}{B(n, k, n\boldsymbol{p}; \boldsymbol{s})}\right).$$

Finally, summing over $\boldsymbol{s}$ yields $D(P_k \,\|\, M_{k,\mu}) \le \sum_{\boldsymbol{p}} \mu(\boldsymbol{p}) D(n, k, n\boldsymbol{p})$, and each of the relative entropy terms can be bounded by the maximum value. □

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

Combining Stam's bound (6) and Proposition 2 immediately gives the following result.

Corollary 1. (Sharp finite de Finetti.) *Under the assumptions of Proposition 2, for* $1 \le k \le n$,

$$D(P_k \parallel M_{k,\mu}) \le \frac{(c-1)k(k-1)}{2(n-1)(n-k+1)}. \tag{18}$$

Pinsker's inequality (5) and (18) gives

$$\|P_k - M_{k,\mu}\| \le \sqrt{\frac{(c-1)k(k-1)}{2(n-1)(n-k+1)}}.$$

In the $k = o(n)$ regime this is in fact optimal, in that it is of the same order as the bound in [12],

$$\|P_k - M_{k,\mu}\| \le \frac{ck}{n}, \tag{19}$$

which was shown to be of optimal order in k and n when $k = o(n)$.

For certain applications, for example to approximation schemes for minimization of specific polynomials [5], the dependence on the alphabet size is also of interest; see also [6] and the references therein. The lower bound (7) shows that the linear dependence on the alphabet size c is optimal for the sampling problem, in the sense that any upper bound that holds for any c, k, and n must have at least linear dependence on c. However, we do not know whether the linear dependence on the alphabet size c is optimal for the de Finetti problem under optimal rates in k,n.

Information-theoretic proofs of finite de Finetti theorems have lately been developed in [6, 13, 14]. In [13], the bound

$$D(P_k \parallel M_{k,\mu}) \le \frac{5k^2 \log n}{n-k} \tag{20}$$

was obtained for binary sequences, and in [14] the weaker bound

$$D(P_k \parallel M_{k,\mu}) = O\left(\left(\frac{k}{\sqrt{n}}\right)^{1/2} \log \frac{n}{k}\right) \tag{21}$$

was established for finite-valued random variables. Finally, the sharper bound

$$D(P_k \parallel M_{k,\mu}) \le \frac{k(k-1)}{2(n-k-1)} \log c \tag{22}$$

was derived in [6], where random variables with values in abstract spaces were also considered. Although the derivations of (20)–(22) are interesting in that they employ purely information-theoretic ideas and techniques, the actual bounds are of strictly weaker rate than the sharp $O(k^2/n^2)$ rate we obtained in Corollary 1 via sampling bounds.

1.3.1. *A question on monotonicity.* A significant development in the area of information-theoretic proofs of probabilistic limit theorems was in 2004, when it was shown [2] that the convergence in relative entropy established in [4] was in fact monotone. In the present setting we observe that, while the total variation bound (19) of [12], $\|P_k - M_{k,\mu}\| \le 2ck/n$, is nonincreasing in n, we do not know whether the total variation distance is itself monotone.

Writing $D(P_r \parallel M_{r,\mu}) = D((X_1, \ldots, X_r) \parallel (\tilde{X}_1, \ldots, \tilde{X}_r))$, where the vector $(\tilde{X}_1, \ldots, \tilde{X}_r)$ is distributed according to $M_{r,\mu}$, a direct application of the data processing inequality for the

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

relative entropy [9] gives, for any integer $r, s \geq 0$, $D(P_r \,\|\, M_{r,\mu}) \leq D(P_{r+s} \,\|\, M_{r+s,\mu})$. In other words, $D(P_k \,\|\, M_{k,\mu})$ is nondecreasing in k. Since the total variation is an f-divergence and therefore satisfies the data processing inequality, the same argument shows that $\|P_k - M_{k,\mu}\|$ is also nondecreasing in k. In view of the monotonicity in the information-theoretic central limit theorem and the convexity properties of the relative entropy it is tempting to conjecture that $D(P_k \,\|\, M_{k,\mu})$ is *also nonincreasing in n*. However, this relative entropy depends on n through the choice of the mixing measure μ, which makes the problem significantly harder. We therefore pose the following questions.

- Let $\boldsymbol{X}_{n+1} = (X_1, \ldots, X_{n+1})$ be a finite exchangeable random sequence and let the mixing measures μ_n, μ_{n+1} be the laws of the empirical measures induced by $\boldsymbol{X}_n$ and $\boldsymbol{X}_{n+1}$, respectively. For a fixed $k \leq n$, is $D(P_k \,\|\, M_{k,\mu_n}) \leq D(P_k \,\|\, M_{k,\mu_{n+1}})$?
- Are there possibly different measures ν_n such that $D(P_k \,\|\, M_{k,\nu_n})$ is vanishing for $k = o(n)$ and nonincreasing in n?

2. Upper bounds on relative entropy

The proof of Theorem 1 in this section will be based on the decomposition of $D(n, k, \boldsymbol{\ell})$ as a sum of expectations of terms involving the quantity $U(a,b)$ introduced in Definition 1. We tightly approximate these terms U within a small additive error (see Proposition 3), and control the expectations of the resulting terms using Lemmas 3 and 4.

2.1. Hypergeometric properties

We first briefly mention some standard properties of hypergeometric distributions that will help our analysis.

Notice that (1) and (2) are both invariant under permutation of color labels. That is, $H(n, k, \boldsymbol{\ell}; \boldsymbol{s}) = H(n, k, \widetilde{\boldsymbol{\ell}}; \widetilde{\boldsymbol{s}})$ and $B(n, k, \boldsymbol{\ell}; \boldsymbol{s}) = B(n, k, \widetilde{\boldsymbol{\ell}}; \widetilde{\boldsymbol{s}})$ whenever $\widetilde{\ell}_i = \ell_{\pi(i)}$ and $\widetilde{s}_i = s_{\pi(i)}$ for the same permutation π. Hence, $D(n, k, \boldsymbol{\ell})$ from (4) is itself permutation invariant,

$$D(n, k, \boldsymbol{\ell}) = D(n, k, \widetilde{\boldsymbol{\ell}}), \tag{23}$$

so we may assume, without loss of generality, that $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_c$. Similarly, if we swap the roles of balls that are 'drawn' and 'not drawn', then direct calculation using the second representation in (1) shows that

$$H(n, n-k, \boldsymbol{\ell}; \boldsymbol{\ell} - \boldsymbol{s}) = \frac{\binom{n-k}{\boldsymbol{\ell}-\boldsymbol{s}}\binom{k}{\boldsymbol{s}}}{\binom{n}{\boldsymbol{\ell}}} = H(n, k, \boldsymbol{\ell}; \boldsymbol{s}). \tag{24}$$

Let $\boldsymbol{S} = (S_1, \ldots, S_c) \sim H(n, k, \boldsymbol{\ell}; \cdot)$. Then each S_r has a multinomial distribution with $c = 2$, with p.m.f. given by $\mathbb{P}(S_r = s_r) = H(n, k, \ell_r; s_r) := H(n, k, (\ell_r, n - \ell_r); (s_r, k - s_r))$.

We use standard notation and write $(x)_r = x(x-1)\cdots(x-r+1) = x!/(x-r)!$ for the falling factorial, and note that on marginalizing we can apply the following standard result.

Lemma 1. *For any i and r, the factorial moments of the hypergeometric $H(n, k, \boldsymbol{\ell}; \cdot)$ satisfy*

$$\mathbb{E}[(S_i)_r] = \sum_{\boldsymbol{s}} H(n, k, \boldsymbol{\ell}; \boldsymbol{s})(s_i)_r = \frac{(\ell_i)_r (k)_r}{(n)_r}.$$

Hence, as mentioned in the introduction, the mean $\mathbb{E}(S_i) = k\ell_i/n$. In general, we write $M_r(S_i) = \mathbb{E}[(S_i - k\ell_i/n)^r]$ for the rth centered moment of S_i and note that, by expressing

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

$(s_i - k\ell_i/n)^r$ as a linear combination of factorial moments of s_i, we obtain

$$M_2(S_i) = \mathrm{Var}(S_i) = \frac{k(n-k)\ell_i(n-\ell_i)}{n^2(n-1)}, \tag{25}$$

$$M_3(S_i) = \frac{k\ell_i(n-k)(n-2k)(n-\ell_i)(n-2\ell_i)}{n^3(n-1)(n-2)}. \tag{26}$$

2.2. Proof of Theorem 1

A key role in the proof will be played by the analysis of the following quantity, defined in terms of the gamma function Γ.

Definition 1. For $0 \le b \le a$, define the function

$$U(a,b) := \log\left(\frac{a^b\Gamma(a-b+1)}{\Gamma(a+1)}\right). \tag{27}$$

Using the standard fact that $\Gamma(n+1) = n!$, when a and b are both integers we can write

$$U(a,b) := \log\left(\frac{a^b(a-b)!}{a!}\right) = \log\left(\frac{a^b}{a(a-1)\cdots(a-b+1)}\right) \ge 0,$$

and note that $U(a,0) = U(a,1) \equiv 0$ for all a. Using Stirling's approximation for the factorials suggests that $U(a,b) \sim (a-b+1/2)\log((a-b)/a) + b$, but we can make this precise using results from [1], for example. We will need the following proposition.

Proposition 3. *Writing*

$$A(a,b) := \left(a-b+\frac{1}{2}\right)\log\left(\frac{a-b}{a}\right) + b + \frac{1}{12(a-b)} - \frac{1}{12a},$$
$$\varepsilon(a,b) := \frac{1}{360}\left(\frac{1}{(a-b)^3} - \frac{1}{a^3}\right) \ge 0$$

for any $0 \le b \le a-1$, we have the bounds $A(a,b) - \varepsilon(a,b) \le U(a,b) \le A(a,b)$.

The key to the proof of the proposition is to work with the logarithmic derivative of the gamma function $\psi(x) := (\mathrm{d}/\mathrm{d}x)\log\Gamma(x)$, $x > 0$. For completeness, we state and prove the following standard bound on ψ.

Lemma 2. *For any $y > 0$,*

$$0 \le \psi(y) - \log y + \frac{1}{2y} + \frac{1}{12y^2} \le \frac{1}{120y^4}.$$

Proof. Note first that, by [1, (2.2)], for example, $-1/y \le \psi(y) - \log(y) \le -1/2y$, meaning that $\psi(y) - \log(y) \to 0$ as $y \to \infty$, so we can write

$$\psi(y) - \log(y) = \int_y^\infty \left(-\psi'(x) + \frac{1}{x}\right)\mathrm{d}x. \tag{28}$$

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

Further, taking $k = 1$ in [1, Theorem 9], we can deduce that, for all $x > 0$,

$$-\frac{1}{2x^2} - \frac{1}{6x^3} \le -\psi'(x) + \frac{1}{x} \le -\frac{1}{2x^2} - \frac{1}{6x^3} + \frac{1}{30x^5}. \tag{29}$$

Substituting (29) in (28) and integrating yields the claimed result. □

Proof of Proposition 3. Combining (27) with the standard fact that $\Gamma(n+1) = n\Gamma(n)$, we can write

$$\begin{aligned} U(a, b) &= b \log a + \log \Gamma(a-b) + \log (a-b) - \log \Gamma(a) - \log a \\ &= (b-1) \log a + \log (a-b) + \int_{a-b}^{a} -\psi(x)\, \mathrm{d}x. \end{aligned}$$

We can now provide an upper bound on this using Lemma 2, to deduce that

$$\begin{aligned} U(a, b) &\le (b-1) \log a + \log (a-b) + \left[-\left(x - \frac{1}{2}\right) \log x + x - \frac{1}{12x} \right]_{a-b}^{a} \\ &= \left(a - b + \frac{1}{2}\right) \log \left(\frac{a-b}{a}\right) + b + \frac{1}{12(a-b)} - \frac{1}{12a}. \end{aligned}$$

Similarly, since $-\psi(x) \ge -\log x + 1/(2x) + 1/(12x^2) - 1/(120x^4)$ we can deduce that

$$U(a, b) \ge A(a, b) - \int_{a-b}^{a} \frac{1}{120x^4}\, \mathrm{d}x = A(a, b) - \frac{1}{360}\left(\frac{1}{(a-b)^3} - \frac{1}{a^3}\right),$$

and the result follows. □

Proof of Theorem 1. We can give an expression for the relative entropy $D(n, k, \boldsymbol{\ell})$ by using the second form of (1) and the fact that $\sum_{i=1}^{c} s_i = k$ to obtain

$$\begin{aligned} \log \left(\frac{H(n, k, \boldsymbol{\ell}; \boldsymbol{s})}{B(n, k, \boldsymbol{\ell}; \boldsymbol{s})}\right) &= \log \left(\frac{\binom{n-k}{\boldsymbol{\ell}-\boldsymbol{s}}}{\binom{n}{\boldsymbol{\ell}}} \prod_{i=1}^{c} \frac{1}{(\ell_i/n)^{s_i}} \right) \\ &= \log \left(\frac{n^k (n-k)!}{n!} \right) + \sum_{i=1}^{c} \log \left(\frac{\ell_i!}{\ell_i^{s_i} (\ell_i - s_i)!} \right) = U(n, k) - \sum_{i=1}^{c} U(\ell_i, s_i), \end{aligned}$$

so that we can write

$$D(n, k, \boldsymbol{\ell}) = U(n, k) - \sum_{i=1}^{c} \mathbb{E} U(\ell_i, S_i), \tag{30}$$

where S_i has a hypergeometric $H(n, k, \ell_i; \cdot)$ distribution. Note that since $S_i \le \min(k, \ell_i)$, in the case $k = 1$ all of the values involved in this calculation are of the form $U(a, 0)$ or $U(a, 1)$, so (30) is identically zero as we would expect – we only sample one item, so it does not matter whether that is with or without replacement. We can now approximate $\mathbb{E} U(\ell_i, S_i)$ by $U(\ell_i, \mathbb{E} S_i) = U(\ell_i, \ell_i k/n)$, to rewrite

$$D(n, k, \boldsymbol{\ell}) = \left(U(n, k) - \sum_{i=1}^{c} U(\ell_i, \ell_i k/n) \right) + \sum_{i=1}^{c} (-\mathbb{E} U(\ell_i, S_i) + U(\ell_i, \ell_i k/n)). \tag{31}$$

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

We bound the two parts of (31) in Lemmas 3 and 4, respectively. The result follows on combining these two expressions, noting that the result also includes the negative term

$$\frac{k}{12n(n-k)} - \frac{ck}{4(n-k)(n-1)} = -\frac{k(3cn-n+1)}{12n(n-1)(n-k)} \leq -\frac{k(3c-1)}{12(n-1)(n-k)} < 0,$$

which may be ignored. □

Lemma 3. *The first term of (31), $U(n,k) - \sum_{i=1}^{c} U(\ell_i, \ell_i k/n)$, is bounded above by*

$$\frac{c-1}{2}\log\left(\frac{n}{n-k}\right) + \frac{k}{12n(n-k)}\left(1 - \sum_{i=1}^{c}\frac{n}{\ell_i}\right) + \frac{1}{360}\left(\frac{n^3}{(n-k)^3} - 1\right)\sum_{i=1}^{c}\frac{1}{\ell_i^3}.$$

Proof. Using the fact that $\sum_{i=1}^{c} \ell_i = n$, we first simplify the corresponding approximation terms from Proposition 3 into the form

$$\begin{aligned} A(n,k) &- \sum_{i=1}^{c} A(\ell_i, \ell_i k/n) \\ &= \left(n - k + \frac{1}{2} - \left(\sum_{i=1}^{c}\ell_i(1-k/n)\right) - c/2\right)\log\left(\frac{n-k}{n}\right) + k - \frac{k}{n}\sum_{i=1}^{c}\ell_i \\ &\quad + \frac{k}{12n(n-k)}\left(1 - \sum_{i=1}^{c}\frac{n}{\ell_i}\right) \\ &= \frac{c-1}{2}\log\left(\frac{n}{n-k}\right) + \frac{k}{12n(n-k)}\left(1 - \sum_{i=1}^{c}\frac{n}{\ell_i}\right). \end{aligned} \tag{32}$$

Then we can use Proposition 3 to obtain

$$\begin{aligned} U(n,k) - \sum_{i=1}^{c} U(\ell_i, \ell_i k/n) &\leq A(n,k) - \sum_{i=1}^{c} A(\ell_i, \ell_i k/n) + \sum_{i=1}^{c}\varepsilon(\ell_i, \ell_i k/n) \\ &= A(n,k) - \sum_{i=1}^{c} A(\ell_i, \ell_i k/n) + \frac{1}{360}\left(\frac{n^3}{(n-k)^3} - 1\right)\sum_{i=1}^{c}\frac{1}{\ell_i^3}, \end{aligned}$$

and the upper bound follows using (32). □

Lemma 4. *The second term in (31) is bounded above as*

$$\sum_{i=1}^{c}(-\mathbb{E}U(\ell_i, S_i) + U(\ell_i, \ell_i k/n)) \leq -\frac{k(c-1)}{2(n-1)} + \frac{k}{4(n-k)(n-1)}\sum_{i=1}^{c}\left(\frac{n}{\ell_i} - 1\right).$$

In order to prove Lemma 4 we first provide a Taylor-series-based approximation for the summands in (31).

Lemma 5. *Each term in the second sum in (31) satisfies*

$$\begin{aligned} -\mathbb{E}U(\ell_i, S_i) &+ U(\ell_i, \ell_i k/n) \\ &\leq -\psi'(\ell_i(1-k/n)+1)\frac{M_2(S_i)}{2} + \psi''(\ell_i(1-k/n)+1)\frac{M_3(S_i)}{6}. \end{aligned} \tag{33}$$

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

Proof. We can decompose the summand in (31) in the form

$$\begin{aligned}-\mathbb{E}U(\ell_i, S_i) + U(\ell_i, \ell_i k/n) &= -\mathbb{E}(S_i \log \ell_i + \log \Gamma(\ell_i - S_i + 1) - \log \Gamma(\ell + 1)) \\ &\quad + \left(\frac{\ell_i k}{n} \log \ell_i + \log \Gamma(\ell_i(1-k/n)+1) - \log \Gamma(\ell+1)\right) \\ &= -\log \Gamma(\ell_i - S_i + 1) + \log \Gamma(\ell_i(1-k/n)+1). \end{aligned} \tag{34}$$

Recalling the definition $\psi(x) := (\mathrm{d}/\mathrm{d}x) \log \Gamma(x)$, we have the Taylor expansion

$$\begin{aligned}&-\log \Gamma(\ell - s + 1) + \log \Gamma(\ell - \mu + 1) \\ &= \psi(\ell-\mu+1)(s-\mu) - \psi'(\ell-\mu+1)\frac{(s-\mu)^2}{2} + \psi''(\ell-\mu+1)\frac{(s-\mu)^3}{6} - \psi'''(\xi)\frac{(s-\mu)^4}{24} \\ &\le \psi(\ell-\mu+1)(s-\mu) - \psi'(\ell-\mu+1)\frac{(s-\mu)^2}{2} + \psi''(\ell-\mu+1)\frac{(s-\mu)^3}{6},\end{aligned}$$

where we used the fact (see, e.g., [1, Theorem 9]) that $\psi'''(x) \ge 0$. The result follows by substituting this in (34) and taking expectations. □

Proof of Lemma 4. We will bound the two terms from Lemma 5. Using [1, Theorem 9] we can deduce that, for any $x > 0$,

$$-\psi'(x+1) \le -\frac{1}{x+1} - \frac{1}{2(x+1)^2} \le -\frac{1}{x} + \frac{1}{2x^2}.$$

Hence, recalling the value of $M_2(S_i)$ from (25), we know that the first term of (33) contributes an upper bound of

$$\begin{aligned}&-\sum_{i=1}^{c} \psi'(\ell_i(1-k/n)+1)\frac{M_2(S_i)}{2} \\ &\quad \le \frac{k}{2n(n-1)} \sum_{i=1}^{c} (n-\ell_i)\ell_i(1-k/n)\left(-\frac{1}{\ell_i(1-k/n)} + \frac{1}{2\ell_i^2(1-k/n)^2}\right) \\ &\quad = \sum_{i=1}^{c} -\frac{k(n-\ell_i)}{2n(n-1)} + \frac{k(n-\ell_i)}{4n(n-1)\ell_i(1-k/n)} \qquad (35) \\ &\quad \le -\frac{k(c-1)}{2(n-1)} + \frac{k}{4(n-k)(n-1)} \sum_{i=1}^{c} \left(\frac{n}{\ell_i} - 1\right), \qquad (36)\end{aligned}$$

where we used the fact that $\sum_{i=1}^{c} (n-\ell_i) = n(c-1)$.

Since $k \le n/2$ by assumption, we know from (26) that $M_3(S_r)$ has the same sign as $(n-2\ell_r)$. And since, as described in (23), we can assume that $\ell_1 \le \ell_2 \le \cdots \le \ell_c$, we know that $\ell_r \le n/2$ for all $r \le c-1$, and hence $M_3(S_r) \ge 0$. This means that, since $\psi''(\ell_i(1-k/n)+1)$ is increasing in ℓ_i, we can bound each $\psi''(\ell_i(1-k/n)+1) \le \psi''(\ell_c(1-k/n)+1)$, so that

$$\frac{1}{6}\sum_{i=1}^{c} \psi''(\ell_i(1-k/n)+1)M_3(S_i) \le \frac{1}{6}\sum_{i=1}^{c} \psi''(\ell_c(1-k/n)+1)M_3(S_i). \tag{37}$$

Finally, writing $f(\ell) = \ell(n-\ell)(n-2\ell)$, we have that $f(\ell) + f(m-\ell)$ is concave and minimized at $\ell = 1$ and $\ell = m-1$. Therefore, the sum $\sum_{r=1}^{c} f(\ell_r)$ is minimized overall when $\ell_1 = \cdots = \ell_{c-1} = 1$ and $\ell_c = n-(c-1)$, giving the value $(c-1)(c-2)(3n-2c) \ge 0$. This means that $\sum_{i=1}^{c} M_3(S_i) \ge 0$, so (37) is negative, and we can simply use (36) as an upper bound. □

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

3. Proofs for small ℓ

Recall the form of the Newton series expansion; see, for example [21, (8), p. 59].

Lemma 6. *Consider a function $f:(a,\infty)\to\mathbb{R}$ for some $a>0$. Suppose that, for some positive integers k, m, and y, the $(k+1)$th derivative $f^{(k+1)}(z)$ of f is negative for all $m\le z\le y$. Then, for any integer x satisfying $m\le x\le y$, there exists some $\xi=\xi(x)\in(m,y)$ such that*

$$f(x)=\sum_{r=0}^{k}\frac{\Delta^r f(m)}{r!}(x-m)_r+\frac{f^{(k+1)}(\xi)}{(k+1)!}(x-m)_{k+1}\le\sum_{r=0}^{k}\frac{\Delta^r f(m)}{r!}(x-m)_r,$$

where as before $(x)_k$ represents the falling factorial and Δ^r is the rth compounded finite difference Δ where, as usual, $\Delta f(x)=f(x+1)-f(x)$.

Lemma 7. *For $\ell_i\ge 3$ we have the upper bound*

$$-\mathbb{E}U(\ell_i,S_i)\le\frac{(k)_2}{2(n)_2}\left(-\ell_i(\ell_i-1)\log\left(\frac{\ell_i}{\ell_i-1}\right)\right)+\frac{(k)_3}{6(n)_3}\left(-\ell_i(\ell_i-1)(\ell_i-2)\log\left(\frac{(\ell_i-1)^2}{\ell_i(\ell_i-2)}\right)\right).\tag{38}$$

Proof. We write $-U(\ell,s)=-s\log\ell-\log\Gamma(\ell-s+1)+\log\Gamma(\ell)$, and use Lemma 6 with $f(s)=-\log\Gamma(\ell-s+1)$ for fixed ℓ.

Since $\Delta f(0)=f(1)-f(0)=\log\Gamma(\ell+1)-\log\Gamma(\ell)=\log\ell$, the Newton series expansion of Lemma 6 with $m=0$ and $k=2$ gives

$$\begin{aligned}-U(\ell,s)&=f(s)-f(0)-s\Delta f(0)\\&\le-\frac{s(s-1)}{2}\log\left(\frac{\ell}{\ell-1}\right)-\frac{s(s-1)(s-2)}{6}\log\left(\frac{(\ell-1)^2}{\ell(\ell-2)}\right),\end{aligned}$$

where we used the fact that $f^{(3)}(\xi)\le 0$. The result follows on taking expectations, using the of the factorial moments from Lemma 1. □

Remark 5. Note that (38) remains valid for $\ell_i=1,2$ as well, as long as we interpret the two bracketed terms in the bound appropriately. For $\ell_i=1$ we assume that both brackets are zero, which is consistent with the fact that $U(1,0)=U(1,1)=0$, so the expectation is zero. For $\ell_i=2$, the first bracket equals $-2\log 2$ and again we assume the second bracket is zero. This is consistent with the fact that $U(2,0)=U(2,1)=0$ and $U(2,2)=\log 2$, coupled with the fact that $H(n,k,2;2)=k(k-1)/(n(n-1))$.

Proof of Proposition 1. The key is to deal with the two $\mathbb{E}U(\ell_i,S_i)$ terms in (30) using two separate arguments, and to think of $D(n,k,\boldsymbol{\ell})$ as

$$[U(n,k)-U(n-\ell,(n-\ell)k/n)]+[-\mathbb{E}U(\ell,S_1)]+[U(n-\ell,(n-\ell)k/n)-\mathbb{E}U(n-\ell,S_2)].\tag{39}$$

We treat the three terms of (39) separately. First, using the bounds of Proposition 3, we obtain that the first term of (39) is bounded above as

$$\begin{aligned}&U(n,k)-U(n-\ell,(n-\ell)k/n)\\&\quad\le A(n,k)-A(n-\ell,(n-\ell)k/n)+\varepsilon(n-\ell,(n-\ell)k/n)\\&\quad=\ell((1-k/n)\log(1-k/n)+k/n)-\frac{k\ell}{12n(n-\ell)(n-k)}+\varepsilon(n-\ell,(n-\ell)k/n)\\&\quad\le\ell((1-k/n)\log(1-k/n)+k/n),\end{aligned}\tag{40}$$

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

where the last inequality follows using the facts that $n-\ell \geq n/2$ and $1-k/n \geq 1/2$ to get

$$\begin{aligned}\varepsilon(n-\ell,(n-\ell)k/n) &\leq \frac{1}{360(n-\ell)^3(1-k/n)^3}\\ &\leq \frac{2}{45n^2(n-\ell)(1-k/n)} \leq \frac{k\ell}{12n(n-\ell)(n-k)}.\end{aligned}$$

For the second term of (39), we apply Lemma 7 with the same conventions as in Remark 5 to obtain the upper bound

$$\frac{(k)_2}{2(n)_2}\left(-\ell(\ell-1)\log\left(\frac{\ell}{\ell-1}\right)\right)+\frac{(k)_3}{6(n)_3}\left(-\ell(\ell-1)(\ell-2)\log\left(\frac{(\ell-1)^2}{\ell(\ell-2)}\right)\right). \tag{41}$$

Using Lemma 5 and taking $\ell_i = n-\ell$ in each of the summands in (35) shows that the third term of (39) is bounded above by

$$\begin{aligned}&-\frac{k\ell}{2n(n-1)}+\frac{k\ell}{4(n-1)(n-\ell)(n-k)}+\frac{M_3(S_2)}{6}\psi''((n-\ell)(1-k/n)+1)\\ &\quad\leq -\frac{k\ell}{2n(n-1)}+\frac{k\ell}{4(n-1)(n-\ell)(n-k)}+\frac{k\ell}{2n(n-1)(n-2)} \qquad (42)\\ &\quad\leq -\frac{k\ell}{2n(n-1)}+\frac{k\ell}{(n-1)(n-\ell)(n-k)}. \qquad (43)\end{aligned}$$

Here, we simplified (42) by using the fact that, by [1, Theorem 9], we have

$$-\psi''(x+1) \leq \frac{1}{(1+x)^2}+\frac{2}{(1+x)^3} \leq \frac{3}{x^3},$$

so that

$$M_3(S_2)(\psi''((n-\ell)(1-k/n)+1) \leq \frac{k(n-2k)\ell(n-2\ell)}{2n(n-1)(n-2)(n-k)(n-\ell)} \leq \frac{k\ell}{2n(n-1)(n-2)}.$$

The result follows on adding (40), (41), and (43). □

Remark 6. While Proposition 1 is only stated for the case $c=2$ for brevity, it would be possible to use a similar argument in the case of general c. The key would be to categorize the colors according to whether they have 'small' or 'large' ℓ. We can write T for the set of indices i for which ℓ_i is small, and deal with the $\mathbb{E}U(\ell_i, S_i)$ terms in (30) using two separate arguments, according to whether $i \in T$ or not. That is, we think of $D(n,k,\boldsymbol{\ell})$ as

$$\left[U(n,k)-\sum_{i\notin T}U(\ell_i,\ell_i k/n)\right]+\left[-\sum_{i\in T}\mathbb{E}U(\ell_i,S_i)\right]+\left[\sum_{i\notin T}-\mathbb{E}U(\ell_i,S_i)+U(\ell_i,\ell_i k/n)\right].$$

We would treat these three terms separately. We can bound the first term using Proposition 3, the second term using Lemma 7, and the third term using Lemma 5. However the resulting expression cannot be stated in a compact fashion.

Remark 7. The last two terms in the bound of Proposition 1 can be further simplified as follows. Recall, e.g. from [25, (3)], that, for all $x \geq 0$,

$$\frac{2x}{2+x} \leq \log(1+x) \leq \frac{x(2+x)}{2(1+x)}. \tag{44}$$

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

Taking $x = 1/(\ell - 1)$ in (44) and rearranging gives

$$0 \leq -\ell(\ell-1)\log\left(\frac{\ell}{\ell-1}\right) + \ell - \frac{1}{2} \leq \frac{1}{2(2\ell-1)}.$$

Similarly, taking $x = 1/(\ell(\ell-2))$ in (44) and rearranging gives

$$0 \leq \frac{1}{2(\ell-1)} \leq -\ell(\ell-1)(\ell-2)\log\left(\frac{(\ell-1)^2}{\ell(\ell-2)}\right) + \ell - 1 \leq \frac{\ell-1}{2\ell^2-4\ell+1}.$$

Funding Information

This research was funded by the EPSRC-funded INFORMED-AI project EP/Y028732/1. L.G. has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 101034255 and from the Bézout Labex, funded by ANR, reference ANR-10-LABX-58.

Competing Interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] ALZER, H. (1997). On some inequalities for the gamma and psi functions. *Math. Comput.* **66**, 373–389.
[2] ARTSTEIN, S., BALL, K., BARTHE, F. AND NAOR, A. (2004). Solution of Shannon's problem on the monotonicity of entropy. *J. Amer. Math. Soc.* **17**, 975–982.
[3] BARBOUR, A., JOHNSON, O., KONTOYIANNIS, I. AND MADIMAN, M. (2010). Compound Poisson approximation via information functionals. *Electron. J. Prob.* **15**, 1344–1369.
[4] BARRON, A. (1986). Entropy and the central limit theorem. *Ann. Prob.* **14**, 336–342.
[5] BERTA, M., BORDERI, F., FAWZI, O. AND SCHOLZ, V. (2022). Semidefinite programming hierarchies for constrained bilinear optimization. *Math. Program.* **194**, 781–829.
[6] BERTA, M., GAVALAKIS, L. AND KONTOYIANNIS, I. (2023). A third information-theoretic approach to finite de Finetti theorems. Preprint, arXiv:2304.05360 [cs.IT].
[7] BRETAGNOLLE, J. AND HUBER, C. (1978). Estimation des densités: Risque minimax. *Séminaire de probabilités XII 1976/77*, 342–363.
[8] CANONNE, C. (2022). A short note on an inequality between KL and TV. Preprint, arXiv:2202.07198[math.PR].
[9] COVER, T. AND THOMAS, J. (2012). *Elements of Information Theory*, 2nd edn. John Wiley, New York.
[10] CSISZÁR, I. (1998). The method of types. *IEEE Trans. Inform. Theory* **44**, 2505–2523.
[11] DIACONIS, P. (1977). Finite forms of de Finetti's theorem on exchangeability. *Synthese* **36**, 271–281.
[12] DIACONIS, P. AND FREEDMAN, D. (1980). Finite exchangeable sequences. *Ann. Prob.* **8**, 745–764.
[13] GAVALAKIS, L. AND KONTOYIANNIS, I. (2021). An information-theoretic proof of a finite de Finetti theorem. *Electron. Commun. Prob.* **26**, 1–5.
[14] GAVALAKIS, L. AND KONTOYIANNIS, I. (2023). Information in probability: Another information-theoretic proof of a finite de Finetti theorem. In *Mathematics Going Forward: Collected Mathematical Brushstrokes*, eds. J.-M. Morel and B. Teissier (Lect. Notes Math. **2313**). Springer, New York.
[15] GAVALAKIS, L. AND KONTOYIANNIS, I. (2024). Entropy and the discrete central limit theorem. *Stoch. Process. Appl.* **170**, 104294.
[16] HARREMOËS, P. (2009). Maximum entropy on compact groups. *Entropy* **11**, 222–237.
[17] HARREMOËS, P. AND MATÚŠ, F. (2020). Bounds on the information divergence for hypergeometric distributions. *Kybernetika* **56**, 1111–1132.
[18] JOHNSON, O. (2024). Information-theoretic convergence of extreme values to the Gumbel distribution. *J. Appl. Prob.* **61**, 244–254.
[19] KIRSCH, W. (2019). An elementary proof of de Finetti's theorem. *Statist. Prob. Lett.* **151**, 84–88.
[20] KONTOYIANNIS, I., HARREMOËS, P. AND JOHNSON, O. (2005). Entropy and the law of small numbers. *IEEE Trans. Inform. Theory* **51**, 466–472.

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press

[21] MILNE-THOMSON, L. (2000). *The Calculus of Finite Differences*. American Mathematical Society, Providence, RI.
[22] RAO, J. (1966). On the comparison of sampling with and without replacement. *Rev. Int. Statist. Inst.* **34**, 125–138.
[23] STAM, A. (1978). Distance between sampling with and without replacement. *Statistica Neerlandica* **32**, 81–91.
[24] THOMPSON, S. (2012). *Sampling*. John Wiley, Hoboken, NJ.
[25] TOPSØE, F. (2007). Some bounds for the logarithmic function. In *Inequality Theory and Applications*, Vol. 4, eds. Y. CHO, J. Kim, and S. Dragomir. Nova Science Publishers, New York, pp. 137–151.
[26] TSYBAKOV, A. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

https://doi.org/10.1017/jpr.2025.27 Published online by Cambridge University Press