Information-theoretic de Finetti-style theorems

Lampros Gavalakis
Department of Engineering
University of Cambridge
Cambridge, UK
lg560@cam.ac.uk

Ioannis Kontoyiannis
Statistical Laboratory
University of Cambridge
Cambridge, UK
yiannis@maths.cam.ac.uk

***Abstract*—We review information-theoretic approaches to obtaining simple probabilistic representations for sequences of exchangeable random variables. Specifically, we examine information-theoretic proofs of finite versions of de Finetti's celebrated representation theorem. Such results state, in a quantitative manner, that the joint distribution of the first k of $n > k$ exchangeable random variables is close to a mixture of product distributions. Closeness is measured in terms of the relative entropy and explicit bounds are typically provided. First we review a recent information-theoretic proof a finite de Finetti theorem for binary random variables, and then we give a different, new proof for the case of arbitrary finite alphabets. This second proof is nicely motivated by the Gibbs conditioning principle in connection with statistical mechanics, and it follows along an appealing sequence of steps. The technical estimates required for these steps are obtained via the method of types.**

A full version of this paper is available online as [23].

***Index Terms*—Exchangeability, de Finetti theorem, mixture, entropy, relative entropy, mutual information, Gibbs conditioning principle, conditional limit theorem, method of types**

I. Introduction

A. Exchangeability

Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables with values in the same finite alphabet A. The sequence is said to be *exchangeable*, if $X_1, X_2, \ldots, X_n$ and $X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}$ have the same distribution for every permutation π on $\{1, 2, \ldots, n\}$. An infinite sequence of random variables $\{X_i\}$ is said to be exchangeable, if $X_1, X_2, \ldots, X_n$ are exchangeable for every $n \geq 1$. A classical result derived by de Finetti [16], [17] in the 1930s states that an infinite sequence of random variables with values in A is exchangeable if and only if it is a mixture of independent and identically distributed (i.i.d.) sequences, that is, if and only if there is a probability measure $\bar{\mu}$ on the simplex $\mathcal{P}$ of probability distributions on A, such that, for any $k \geq 1$ and any $x_1^k = (x_1, x_2, \ldots, x_k) \in A^k$,

$$\mathbb{P}(X_1^k = x_1^k) = M_{k,\bar{\mu}}(x_1^k) := \int_{\mathcal{P}} Q^k(x_1^k) d\bar{\mu}(Q). \qquad (1)$$

L.G. was supported in part by EPSRC grant number RG94782.

I.K. was supported in part by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant, project number 1034."

Since its original appearance, de Finetti's representation theorem (1) has found many important extensions and generalisations. For example, it has been extended to random variables with values in compact Hausdorff spaces by Hewitt and Savage [24], a result that was later further generalised by Diaconis and Freedman [20]. And in a different direction, Diaconis and Freedman [19] also obtained a natural analogue of (1) to the case of Markov chain mixtures.

De Finetti's theorem plays an important role in the foundations of subjective probability and Bayesian statistics, see, e.g., the discussions in [7], [18]. But arguments about its practical relevance are limited by the fact that, as is well known [20], the representation (1) fails in general if it is only assumed that a *finite* collection of random variables $(X_1, X_2, \ldots, X_n)$ is exchangeable for some *fixed* n.

Nevertheless, it has been shown – first by Diaconis [18] and then by several others – that the distribution of the initial segment of a finite exchangeable sequence is *close* to a mixture of i.i.d. distributions. At the heart of Diaconis' [18] proof for binary random variables, lies the geometric interpretation of the set of exchangeable measures as a convex subset of the probability simplex and the representation of a point as a convex combination of extreme points. Here we describe two approaches to proving finite de Finetti-style representations based on information-theoretic ideas and techniques.

B. Related work and new results

Several alternative proofs of de Finetti's theorem have appeared in the literature. Recently, an elementary proof for the binary case was given in 2019 by Kirsch [28], another proof of a more analytic nature was developed in 2020 by Alam [1], and interesting connections with category theory were drawn in 2021 in [21].

Providing information-theoretic proofs of core results in probability theory has been a fruitful and active research area that has produced a wide and multifaceted collection or results. The long list of works that follow this development includes, to name only a few, information-theoretic proofs of Markov chain convergence [27], [32], of the central limit theorem [2], [4], and of Poisson and compound Poisson approximation [3], [26], [29]. And in connection with a different aspect of exchangeability, mutual information and its properties were employed by O'Connell [31] to give a very elegant, elementary proof the Hewitt-Savage zero-one law.

The interested reader may also consult Barron's reviews of information-theoretic proofs and connections with statistics and learning [5], [6], Csiszár's review of information-theoretic methods in probability [12], and Johnson's text [25].

In Section II we review a finite de Finetti-type theorem that was recently established using information-theoretic methods [22]. In the case of binary random variables, Theorem 2.1 states that, if $X_1, X_2, \ldots, X_n$ are exchangeable, then there is a probability measure μ on $[0,1]$ such that the distribution of the first k of the X_i is close to the mixture of the product Bern(p) distributions under μ. Closeness is measured in terms of relative entropy and an explicit bound (2) is provided in terms of n and k. The main idea in its proof is to measure the dependence of the random variables under an appropriate conditioning using mutual information.

Next, in Section III we give an analogous relative entropy bound in for random variables X_i taking values in any finite alphabet. The proof of this second result is completely different from that in [23]; it uses estimates provided by the method of types and by an argument based on a variation of the proof of the conditional limit theorem.

C. The method of types and large deviations

The celebrated method of types is a collection of combinatorial estimates for probabilities associated with discrete i.i.d. random variables and memoryless channels. Based in part on preliminary ideas of Wolfowitz [35], the method of types was fully developed in 1981 by Csiszár and Körner [15]. As described in Csiszár's review [13], the method of types has been employed very widely and with great success in numerous information-theoretic problems arising from different communication-theoretic scenarios.

Based in part on the method of types, Csiszár was able to establish a series of important results in large deviations. In 1975 [10] he identified the exponent in Sanov's theorem [33] as an extremum of relative entropies, and in 1984 [11] he proved a general, strong version of Sanov's theorem, by a combination of the method of types, discretisation arguments, and a general Pythagorean inequality for the relative entropy established by Topsøe [34]. He also gave a simpler proof along the same lines in his 2006 paper [14]. Moreover, in [11] Csiszár established a version of the *Gibbs conditioning principle*, also know as the *conditional limit theorem*, using the same tools.

The method of types and the Gibbs conditioning principle both play an important role in the proof of our finite de Finetti theorem in Section III.

II. A First Finite de Finetti Theorem

For each $p \in [0,1]$, let Q_p denote the Bernoulli probability mass function with parameter p, $Q_p(1) = 1 - Q_p(0) = p$, and write $D(P\|Q) = \sum_{x\in A} P(x)\log[P(x)/Q(x)]$ for the relative entropy (or Kullback-Leibler divergence) between two probability mass functions P, Q on the same discrete set A; throughout, 'log' denotes the natural logarithm to base e.

Theorem 2.1: Let $n \geq 2$. If the binary random variables $(X_1, X_2, \ldots, X_n)$ are exchangeable, then there is a probability measure μ on $[0,1]$, such that, for every $1 \leq k \leq n$, the relative entropy between the probability mass function P_k of $(X_1, X_2, \ldots, X_k)$ and the mixture $M_{k,\mu} := \int Q_p^k d\mu(p)$ satisfies:

$$D(P_k\|M_{k,\mu}) \leq \frac{5k^2 \log n}{n-k}. \tag{2}$$

By Pinsker's inequality [9], [30], $\|P - Q\|^2 \leq 2D(P\|Q)$, the theorem also implies that,

$$\|P_k - M_{k,\mu}\| \leq \sqrt{10}k\Big(\frac{\log n}{n-k}\Big)^{1/2}, \tag{3}$$

where $\|P - Q\| := \sum_x |P(x) - Q(x)|$ denotes the ℓ_1 distance between P and Q. This bound is suboptimal in that, as shown by Diaconis and Freedman [20], the correct rate with respect to the ℓ_1 distance in (3) is $O(k/n)$. On the other hand, (2) gives an explicit bound for the stronger notion of relative entropy 'distance.' Of course, rather than to obtain optimal rates, our primary motivation in this note is to illustrate how elementary information-theoretic ideas can be used to provide an alternative proof strategy for de Finetti's theorem.

Before sketching the main ideas of the proof of Theorem 2.1 from [22], we first need to introduce some notation. Let $n \geq 2$ be fixed. For any $1 \leq i \leq j \leq n$, write X_i^j for the block of random variables $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$. Denote by $N_{i,j}$ the number of 1s in X_i^j, so that $N_{i,j} = \sum_{k=i}^j X_k$, and for every $0 \leq \ell \leq n$ write A_ℓ for the event $\{N_{1,n} = \ell\}$.

The main step of the proof is the estimate in Lemma 2.2 below, which gives a bound on the degree of dependence between X_i and X_{i+1}^k, conditional on A_ℓ. This bound is expressed in terms of the mutual information. Let (X, Y) be two discrete random variables with joint probability mass function (p.m.f.) P_{XY} and marginal p.m.f.s P_X and P_Y, respectively. We write $H(X) = H(P_X)$ for the entropy of X, $H(Y|X)$ for the conditional entropy of Y given X, and $I(X;Y) = H(Y) - H(Y|X)$ for the mutual information. For any event A, we write $I(X;Y|A)$ for the mutual information between X and Y when all relevant p.m.f.s are conditioned on A.

Lemma 2.2: For all $1 \leq k \leq n$, all $1 \leq i \leq k-1$, and any $0 \leq \ell \leq n$:

$$I(X_i; X_{i+1}^k | A_\ell) \leq \frac{5k \log n}{n-k}. \tag{4}$$

The proof of Lemma 2.2 is quite technical, involving a combinatorial estimate of the probability that $X_i = 1$ under an appropriate conditioning and a Taylor expansion of the binary entropy function, in order to show that the entropy in the definition of the mutual information is close to the conditional one; see [22].

Theorem 2.1 follows rather quickly from Lemma 2.2. By the bound in (4),

$$\sum_{i=1}^{k-1} I(X_i; X_{i+1}^k | A_\ell) \leq \frac{5k^2 \log n}{n-k}.$$

Also, by definition of the mutual information, using the obvious notation $H(X|A)$ for the entropy of the conditional p.m.f. of X given A,

$$\begin{aligned}\sum_{i=1}^{k-1} I(X_i;X_{i+1}^k|A_\ell) &= \sum_{i=1}^{k-1}\Big[H(X_i|A_\ell)+H(X_{i+1}^k|A_\ell)\\ &\quad - H(X_i^k|A_\ell)\Big]\\ &= \Big[\sum_{i=1}^{k} H(X_i|A_\ell)\Big] - H(X_1^k|A_\ell)\\ &= D\big(P_{X_1^k|A_\ell}\big\|P_{X_1|A_\ell}\times\cdots\times P_{X_k|A_\ell}\big),\end{aligned}$$

where we write $P_{X_i^j|A_\ell}$ for the conditional p.m.f. of X_i^j given A_ℓ. Since $P_{X_i|A_\ell}=Q_{\ell/n}$, we have,

$$D\big(P_{X_1^k|A_\ell}\big\|Q_{\ell/n}^k\big) \le \frac{5k^2\log n}{n-k}.$$

Finally, writing μ for the distribution of $\ell/n=(1/n)\sum_{i=1}^n X_i$ on $\{0,1/n,2/n\dots,1\}$, averaging both sides with respect to ℓ and using the joint convexity of relative entropy yields the claimed result.

Remarks.

1) The mixing measure $\mu=\mu_n$ in the theorem is completely characterised in the proof as the distribution of $(1/n)\sum_{i=1}^n X_i$, and it is the same for all k.
2) If $\{X_n\ ;\ n\ge 1\}$ is an infinite exchangeable sequence, then it is also stationary, so, by the ergodic theorem $(1/n)\sum_{i=1}^n X_i$ converges a.s. to some X, and the μ_n converge weakly to the law, say μ, of X. For fixed k, since Q_p^k is a bounded and continuous function of $p\in[0,1]$, we have for any $x_1^k\in\{0,1\}^k$,

$$\begin{aligned}M_{k,\mu_n}(x_1^k) &= \int Q_p^k(x_1^k)d\mu_n(p)\to M_{k,\mu}(x_1^k)\\ &= \int Q_p^k(x_1^k)d\mu(p),\end{aligned}$$

and by Theorem 2.1, $\|P_k-M_{k,\mu_n}\|=O(\sqrt{(\log n)/n})$. Therefore, we can conclude that,

$$P_k=\int Q_p^k d\mu(p),$$

for each $k\ge 1$, and thus recover de Finetti's classical representation theorem.

III. A New Finite de Finetti Theorem

Suppose $X_1^n=(X_1,\dots,X_n)$, for some fixed n, are exchangeable, discrete random variables, with values in a finite alphabet A of $m=|A|$ elements. Let $\hat{P}_{X_1^n}$ denote the (random) type of X_1^n, and let the measure $\mu=\mu_n$ denote the law of $\hat{P}_{X_1^n}$ on the probability simplex $\mathcal{P}$. Using the same notation as in Theorem 2.1, we now establish the following new information-theoretic finite de Finetti theorem:

Theorem 3.1 (Finite de Finetti theorem): For any $1\le k\le (n/100)^{1/3}$, we have,

$$\begin{aligned}D(P_k\|M_{k,\mu}) &\le \epsilon(n,k)\\ &:= 2\delta + ke^{-\frac{n}{k}\delta}\Big(\frac{n}{k}+1\Big)^{2m^k}\log n,\end{aligned}\qquad (5)$$

with $\alpha=\alpha_{n,k}=\big[\frac{2k}{\sqrt{n}}\big(\frac{1+2k}{\sqrt{n}}+1\big)\big]^{1/2}$ and $\delta=\delta_{n,k}=\alpha\log(m^k/\alpha)$.

Remarks.

1) It can be seen from (5) that, if k stays bounded as $n\to\infty$, then:

$$\epsilon(n,k)=O(\delta_{n,k})=O\Big(\Big(\frac{k}{\sqrt{n}}\Big)^{1/2}\log\frac{n}{k}\Big)\to 0.$$

Moreover, in order for $\epsilon(n,k)$ to vanish, k can grow at most logarithmically with n. This is, at least asymptotically, weaker than the bound given in Theorem 2.1 for the binary case $m=2$. Furthermore, the proof of (5) given below is longer and more involved that the corresponding proof of (2). But the interest in Theorem 3.1 lies in the fact that the proof of (5) follows a completely different information-theoretic path, which consists of an appealing sequence of steps making interesting connections. So we first present a heuristic outline, and then discuss some of the key steps in the actual proof.
2) We have cheated slightly in the statement of the theorem, in that the proof below is only given for the case when n is a multiple of k. However, this is only a minor technical inconvenience; for example, we can replace n with an integer multiple of k which is no less than $n-k$, leading to the same bound with $\epsilon(n-k,k)$ in place of $\epsilon(n,k)$.
3) De Finetti's original theorem (1) easily follows from (5) as in Section II.

A. A heuristic proof of de Finetti's theorem (1).

Step 1: Since the sequence $\{X_n\}$ is exchangeable it is also stationary, therefore, by the ergodic theorem $\hat{P}_{X_1^n}$ converges as $n\to\infty$ a.s. to a (random) P on A. Let $\bar{\mu}$ denote the law of P, and let $\{Y_n\}$ be i.i.d. random variables uniformly distributed on A. Then, by exchangeability, we clearly have for any n, any $k\le n$, any n-type Q_n, and any $a_1^k\in A^k$,

$$\mathbb{P}(X_1^k=a_1^k|\hat{P}_{X_1^n}=Q_n)=\mathbb{P}(Y_1^k=a_1^k|\hat{P}_{Y_1^n}=Q_n). \qquad (6)$$

Step 2: Choose and fix any one of the almost all realisations $\{Q_n\}$ along which $\hat{P}_{X_1^n}$ converges to some Q as $n\to\infty$. By (6) and symmetry we have,

$$\begin{aligned}\mathbb{P}(X_1=a|\hat{P}_{X_1^n}=Q_n) &= \mathbb{E}\Big(\mathbb{I}_{\{Y_1=a\}}\Big|\hat{P}_{Y_1^n}=Q_n\Big)\\ &= \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{Y_i=a\}}\Big|\hat{P}_{Y_1^n}=Q_n\Big),\end{aligned}$$

so that,

$$\mathbb{P}(X_1=a|\hat{P}_{X_1^n}=Q_n)=\mathbb{E}\Big(\hat{P}_{Y_1^n}(a)\Big|\hat{P}_{Y_1^n}=Q_n\Big)=Q_n(a),$$

for any $a \in A$, and letting $n \to \infty$ yields,

$$\lim_{n\to\infty} \mathbb{P}(X_1 = a | \hat{P}_{Y_1^n} = Q_n) = Q(a). \tag{7}$$

Step 3: Next we generalise (7) to blocks of random variables. As before, choose and fix any one of the almost all realisations $\{Q_n\}$ of the random $\hat{P}_{X_1^n}$ such that $Q_n \to$ some Q as $n \to \infty$. Define a new sequence of i.i.d. random variables $Z_n = (Y_{2n-1}, Y_{2n})$, $n \geq 1$, so that each Z_n is uniformly distributed on $A \times A$. From (6), taking $k = 2$ and an arbitrary even $n = 2\ell$,

$$\begin{aligned}&\mathbb{P}\big((X_1, X_2) = (a_1, a_2)\big|\hat{P}_{X_1^{2\ell}} = Q_{2\ell}\big)\\ &\quad= \mathbb{P}\big(Z_1 = (a_1, a_2)\big|\hat{P}_{Z_1^\ell} \in E(Q_{2\ell})\big),\end{aligned} \tag{8}$$

where $E(Q)$ denotes the set of probability distributions W on $A \times A$ with the property that the average of the two marginals W_1 and W_2 of W equals Q,

$$E(Q) = \Big\{W \text{ on } A \times A \,:\, \frac{W_1 + W_2}{2} = Q\Big\}.$$

If we write U for the uniform distribution on $A \times A$, it is easy to check that the distribution W_ℓ^* that uniquely achieves the $\min_{W \in E(Q)} D(W\|U)$ is simply $Q \times Q$. At this point, we would wish to apply the conditional limit theorem [8] to the i.i.d. process $\{Z_n\}$, to obtain that,

$$\begin{aligned}\lim_{\ell\to\infty} &\mathbb{P}\big(Z_1 = (a_1, a_2)\big|\hat{P}_{Z_1^\ell} \in E(Q_{2\ell})\big)\\ &= \lim_{\ell\to\infty} W_\ell^*(a_1, a_2)\\ &= \lim_{\ell\to\infty} Q_{2\ell}(a_1) Q_{2\ell}(a_2)\\ &= Q(a_1)Q(a_2),\end{aligned}$$

and combining this with (8) would yield,

$$\mathbb{P}\big((X_1, X_2) = (a_1, a_2)\big|\hat{P}_{X_1^{2\ell}} = Q_{2\ell}\big) \to Q(a_1)Q(a_2),$$

as $\ell \to \infty$. The same argument can be used without difficulty to show that for any $k \geq 1$ and any $a_1^k \in A^k$,

$$\lim_{\ell\to\infty} \mathbb{P}(X_1^k = a_1^k | \hat{P}_{X_1^{k\ell}} = Q_{k\ell}) = Q^k(a_1^k). \tag{9}$$

Step 4: Since (9) holds for almost every sequence $\{Q_n\}$, letting $\ell \to \infty$, by the bounded converge theorem we have,

$$\mathbb{P}(X_1^k = a_1^k) = \mathbb{E}\Big(\mathbb{P}(X_1^k = a_1^k | \hat{P}_{X_1^{k\ell}})\Big) \to \int_{\mathcal{P}} Q^k(a_1^k) d\bar{\mu}(Q),$$

as required. $\square$

The only problem with the above argument is that the set $E(Q)$ has an empty interior so that the conditional limit theorem is not directly applicable. Nevertheless, in the next section where we take a finite-n approach, we are able to 'imitate' the proof of the conditional limit theorem and replace the step where the non-empty interior assumption is used with a different argument.

B. Proof outline of Theorem 3.1

Recall the notation and terminology for types described in the previous section. Let $\mu = \mu_n$ denote the law of $\hat{P}_{X_1^n}$ on $\mathcal{P}$, and let $\{Y_n\}$ be i.i.d. random variables uniformly distributed on A. For any $k \leq n$, any n-type Q_n, and any $a_1^k \in A^k$,

$$\mathbb{P}(X_1^k = a_1^k | \hat{P}_{X_1^n} = Q_n) = \mathbb{P}(Y_1^k = a_1^k | \hat{P}_{Y_1^n} = Q_n).$$

Foe $k = 1$ and any $a \in A$, by symmetry we have,

$$\mathbb{P}(X_1 = a | \hat{P}_{X_1^n}) \;=\; \hat{P}_{X_1^n}(a),$$

and taking the expectation of both sides with respect to μ shows that in fact $P_1 = M_{1,\mu}$.

For general $1 \leq k \leq n$ with $n = k\ell$, for any n-type Q we have,

$$\begin{aligned}\mathbb{P}(X_1^k = a_1^k | \hat{P}_{X_1^{k\ell}} = Q) &= \mathbb{P}(Z_1 = a_1^k | \hat{P}_{Z_1^\ell} \in E_k(Q))\\ &= \mathbb{E}\big(\hat{P}_{Z_1^\ell}(a_1^k)\big|\hat{P}_{Z_1^\ell} \in E_k(Q)\big),\end{aligned}$$

where now $\{Z_n\}$ is a sequence of i.i.d. random variables uniformly distributed on A^k, and $E_k(Q)$ consists of all probability distributions W on A^k with the property that the average of the k one-dimensional marginals of W equals Q. Taking expectations with respect to $\hat{P}_{X_1^{k\ell}} = \hat{P}_{X_1^n} \sim \mu = \mu_n$,

$$\mathbb{P}(X_1^k = a_1^k) = \int \mathbb{E}\big(\hat{P}_{Z_1^\ell}(a_1^k)\big|\hat{P}_{Z_1^\ell} \in E_k(Q)\big) d\mu(Q),$$

and by the joint convexity of relative entropy,

$$\begin{aligned}&D(P_k \| M_{k,\mu})\\ &= D\Big(\int \mathbb{E}\big(\hat{P}_{Z_1^\ell}\big|\hat{P}_{Z_1^\ell} \in E_k(Q)\big) d\mu(Q)\Big\| \int Q^k \, d\mu(Q)\Big)\\ &\leq \int D\Big(\mathbb{E}\big(\hat{P}_{Z_1^\ell}\big|\hat{P}_{Z_1^\ell} \in E_k(Q)\big)\Big\|Q^k\Big) d\mu(Q)\\ &\leq \int \mathbb{E}\Big(D(\hat{P}_{Z_1^\ell}\|Q^k)\,\Big|\hat{P}_{Z_1^\ell} \in E_k(Q)\Big) d\mu(Q).\end{aligned} \tag{10}$$

We will obtain an explicit bound for the relative entropy in (10). First, we construct a joint ℓ-type W with desirable properties. Let $\mathcal{P}_\ell$ denote the set of ℓ-types on A^k.

Lemma 3.2: For any $\ell > k \geq 1$ and any n-type Q, there is a $W \in E_k(Q) \cap \mathcal{P}_\ell$ with:

$$\max_{a_1^k} |W(a_1^k) - Q^k(a_1^k)| \leq M := \left[\frac{2}{\ell} + \frac{4k}{\ell} + 2\sqrt{\frac{k}{\ell}}\right]^{1/2}.$$

Moreover, for $2 \leq k \leq \sqrt{\ell}/10$,

$$|H(W) - H(Q^k)| \leq -M \log \frac{M}{m^k}.$$

The first part of the lemma can be proved using the second moment method and the second is a consequence of [15, Lemma 2.7]; see [22] for details.

Next we obtain an upper bound on the conditional expectation in (10).

Lemma 3.3: Suppose $n = \ell k$, with $2 \le k \le \sqrt{\ell}/10$. For any n-type Q we have:

$$\mathbb{E}\Big(D(\hat{P}_{Z_1^\ell}\|Q^k)\,\Big|\hat{P}_{Z_1^\ell} \in E_k(Q)\Big) \le \epsilon(n,k).$$

Proof: We follow the same steps as in the proof of the conditional limit theorem in [8]. Recall that if we write U_k for the uniform distribution on A^k, then the W_k^* that uniquely achieves $D^* = \min_{W\in E_k(Q)} D(W\|U_k)$ is $W_k^* = Q^k$. We partition $E_k(Q)$ into $B_{2\delta}$ and $C = E_k(Q) - B_{2\delta}$, where $B_{2\delta} = \{W \in E_k(Q) : D(W\|U_k) \le D^* + 2\delta\}$, with $\delta = \delta_{n,k}$. Then, writing ν_ℓ for the distribution of $\hat{P}_{Z_1^\ell}$,

$$\begin{aligned}
\nu_\ell(C|E_k(Q)) &= \frac{\nu_\ell(C\cap E_k(Q))}{\nu_\ell(E_k(Q))}\\
&\le \frac{\nu_\ell(C)}{\nu_\ell(B_{2\delta})}.
\end{aligned}$$

Next we bound the above numerator and denominator. For the numerator, writing again $\mathcal{P}_\ell$ for the set of ℓ-types on A^k,

$$\begin{aligned}
\nu_\ell(C) &\overset{(a)}{=} \sum_{W\in C\cap\mathcal{P}_\ell} U_k^\ell(T(W))\\
&\overset{(b)}{\le} \sum_{W\in C\cap\mathcal{P}_\ell} e^{-\ell D(W\|U_k)}\\
&\overset{(c)}{\le} |E_k(Q)\cap\mathcal{P}_\ell| e^{-\ell(D^*+2\delta)}\\
&\overset{(d)}{\le} (\ell+1)^{m^k} e^{-\ell(D^*+2\delta)}
\end{aligned}$$

where $T(W)$ in (a) denotes the type class of all strings of length ℓ in A^k with type W, (b) is a standard property [8], (c) follows from the definition of C and the fact that $E_k(Q)\cap\mathcal{P}_\ell \subset E_k(Q)$, and (d) follows from the standard observation that $|E_k(Q)\cap\mathcal{P}_\ell| \le |\mathcal{P}_\ell| \le (\ell+1)^{m^k}$. Similarly, letting W_0 denote the type from Lemma 3.2,

$$\begin{aligned}
\nu_\ell(B_{2\delta}) &\ge \nu_\ell(B_\delta)\\
&= \sum_{W\in B_\delta\cap\mathcal{P}_\ell} U_k^\ell(T(W))\\
&\ge U_k^\ell(T(W_0))\\
&\ge (\ell+1)^{-m^k} e^{-\ell D(W_0\|U)}\\
&\ge (\ell+1)^{-m^k} e^{-\ell(D^*+\delta)}.
\end{aligned}$$

Combining these bounds, we obtain,

$$\mathbb{P}\big(\hat{P}_{Z_1^\ell} \in C\big|\hat{P}_{Z_1^\ell} \in E_k(Q)\big) \le (\ell+1)^{2m^k} e^{-\ell\delta},$$

or,

$$\begin{aligned}
\mathbb{P}\Big(D(\hat{P}_{Z_1^\ell}\|U_k) > D^* + 2\delta\Big|\hat{P}_{Z_1^\ell} \in E_k(Q)\Big)&\\
\le (\ell+1)^{2m^k} e^{-\ell\delta}&.
\end{aligned}$$

Since the set $E_k(Q)$ is closed and convex, we may apply the Pythagorean identity for relative entropy [8] to conclude that:

$$\mathbb{P}\Big(D(\hat{P}_{Z_1^\ell}\|Q^k) > 2\delta|\hat{P}_{Z_1^\ell} \in E_k(Q)\Big) \le (\ell+1)^{2m^k} e^{-\ell\delta}.$$

Thus,

$$\begin{aligned}
&\mathbb{E}\Big(D(\hat{P}_{Z_1^\ell}\|Q^k)\big|\hat{P}_{Z_1^\ell} \in E_\ell(Q)\Big)\\
&\le (\ell+1)^{2m^k} e^{-\ell\delta} \max_{P\in E_k(Q)} D(P\|Q^k) + 2\delta.
\end{aligned}$$

The claimed bound now follows by Lemma 3.4 on taking $\ell = k/n$. ■

Lemma 3.4: For any n-type Q, we have:

$$\max_{W\in E_k(Q)} D(W\|Q^k) \le k\log n.$$

The proof of this technical lemma is omitted; see [22] for the details.

Theorem 3.1 follows from (10) combined with Lemma 3.3.

References

[1] I. Alam. A nonstandard proof of de Finetti's theorem for Bernoulli random variables. *Journal of Stochastic Analysis*, 1(4):15, 2020.

[2] S. Artstein, K. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *J. Amer. Math. Soc.*, 17(4):975–982, 2004.

[3] A.D. Barbour, O. Johnson, I. Kontoyiannis, and M. Madiman. Compound Poisson approximation via information functionals. *Electron. J. Probab*, 15:1344–1369, 2010.

[4] A.R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, 14(1):336–342, January 1986.

[5] A.R. Barron. Information theory in probability, statistics, learning, and neural nets. In Y. Freund and R.E. Schapire, editors, *Proceedings of the Tenth Annual Conference on Computational Learning Theory (COLT)*, Nashville, Tennessee, July 1997. Available at: http://www.stat.yale.edu/∼arb4/publications_files/COLT97.pdf.

[6] A.R. Barron. Limits of information, Markov chains, and projection. In *2000 IEEE International Symposium on Information Theory (ISIT)*, Sorrento, Italy, June 2000.

[7] M. Bayarri and J.O. Berger. The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

[8] T.M. Cover and J.A. Thomas. *Elements of information theory*. J. Wiley & Sons, New York, second edition, 2012.

[9] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[10] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, February 1975.

[11] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, 12(3):768–793, August 1984.

[12] I Csiszár. Information theoretic methods in probability and statistics. In *1997 IEEE International Symposium on Information Theory (ISIT)*, Ulm, Germany, June 1997.

[13] I. Csiszár. The method of types. *IEEE Trans. Inform. Theory*, 44(6):2505–2523, October 1998. Information theory: 1948–1998.

[14] I. Csiszár. A simple proof of Sanov's theorem. *Bulletin of the Brazilian Mathematical Society*, 37(4), 2006.

[15] I. Csiszár and J. Körner. *Information theory: Coding theorems for discrete memoryless systems*. Academic Press, New York, 1981.

[16] B. De Finetti. Sul significato soggettivo della probabilita. *Fundamenta Mathematicae*, 17(1):298–329, 1931.

[17] B. De Finetti. La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré*, 7(1):1–68, 1937.

[18] P. Diaconis. Finite forms of de Finetti's theorem on exchangeability. *Synthese*, 36(2):271–281, 1977.

[19] P. Diaconis and D.A. Freedman. De Finetti's theorem for Markov chains. *Ann. Probab.*, 8(1):115–130, 1980.

[20] P. Diaconis and D.A. Freedman. Finite exchangeable sequences. *Ann. Probab.*, 8(4):745–764, 1980.

[21] T. Fritz, T. Gonda, and P. Perrone. De Finetti's theorem in categorical probability. *arXiv e-prints*, `2105.02639 [math.PR]`, September 2021.

[22] L. Gavalakis and I. Kontoyiannis. An information-theoretic proof of a finite de Finetti theorem. *Electron. Comm. Probab.*, 26:1–5, 2021.

[23] L. Gavalakis and I. Kontoyiannis. Information in probability: Another information-theoretic proof of a finite de Finetti theorem. *arXiv e-prints*, `2204.05033 [math.PR]`, April 2022.

[24] E. Hewitt and L.J. Savage. Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.*, 80(2):470–501, 1955.

[25] O. Johnson. *Information theory and the central limit theorem*. Imperial College Press, London, U.K., 2004.

[26] O. Johnson, I. Kontoyiannis, and M. Madiman. Log-concavity, ultra-log-concavity, and a maximum entropy property of discrete compound Poisson measures. *Discrete Applied Mathematics*, 161(9):1232–1250, 2013.

[27] D.G. Kendall. Information theory and the limit-theorem for Markov chains and processes with a countable infinity of states. *Ann. Inst. Statist. Math.*, 15(1):137–143, May 1963.

[28] W. Kirsch. An elementary proof of de Finetti's theorem. *Statist. Probab. Lett.*, 151:84–88, 2019.

[29] I. Kontoyiannis, P. Harremoës, and O. Johnson. Entropy and the law of small numbers. *IEEE Trans. Inform. Theory*, 51(2):466–472, February 2005.

[30] S. Kullback. A lower bound for discrimination information in terms of variation. *IEEE Trans. Inform. Theory*, 13(1):126–127, January 1967.

[31] N. O'Connell. Information-theoretic proof of the Hewitt-Savage zero-one law. Technical report, Hewlett-Packard Laboratories, Bristol, U.K., June 2000.

[32] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 547–561. Univ. California Press, Berkeley, Calif., 1961.

[33] I.N. Sanov. On the probability of large deviations of random variables. *Mat. Sb.*, 42:11–44, 1957. English translation in *Sel. Transl. Math. Statist. Probab.* (1961) **1**, 213-244.

[34] F. Topsøe. Information-theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

[35] J. Wolfowitz. *Coding theorems of information theory*. Springer, Berlin, 1961.