An information-theoretic proof of a finite de Finetti theorem[*]

Lampros Gavalakis[†]   Ioannis Kontoyiannis[‡]

Abstract

A finite form of de Finetti's representation theorem is established using elementary information-theoretic tools: The distribution of the first k random variables in an exchangeable binary vector of length $n \geq k$ is close to a mixture of product distributions. Closeness is measured in terms of the relative entropy and an explicit bound is provided.

Keywords: Exchangeability; de Finetti theorem; entropy; mixture; relative entropy; mutual information.
MSC2020 subject classifications: 60G09; 62B10; 94A15; 94A17.
Submitted to ECP on April 8, 2021, final version accepted on September 12, 2021.

1 Introduction

A finite sequence of random variables $(X_1, X_2, \ldots, X_n)$ is *exchangeable* if it has the same distribution as $(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)})$ for every permutation π of $\{1, 2, \ldots, n\}$. An infinite sequence $\{X_k \, ; k \geq 1\}$ is exchangeable if $(X_1, X_2, \ldots, X_n)$ is exchangeable for all n. The celebrated representation theorem of de Finetti [8, 9] states that the distribution of any infinite exchangeable sequence of binary random variables can be expressed as a mixture of the distributions corresponding to independent and identically distributed (i.i.d.) Bernoulli trials. For discussions of the role of de Finetti's theorem in connection with the foundations of Bayesian statistics and subjective probability see, e.g, [10, 5] and the references therein.

Although it is easy to see via simple examples that de Finetti's theorem may fail for finite binary exchangeable sequences, for large but finite n the distribution of the first k random variables of an exchangeable vector of length n admits an approximate de Finetti-style representation. Quantitative versions of this statement have been established by Diaconis [10] and Diaconis and Freedman [13]. The approach of Diaconis' proof in [10] is based on a geometric interpretation of the set of exchangeable measures as a convex subset of the probability simplex.

[†]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K. E-mail: lg560@cam.ac.uk.

[‡]Statistical Laboratory, DPMMS, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, U.K. E-mail: yiannis@maths.cam.ac.uk. Web: http://www.dpmms.cam.ac.uk/person/ik355.

The purpose of this note is to provide a new information-theoretic proof of a related finite version of de Finetti's theorem. For each $p \in [0,1]$, let P_p denote the Bernoulli probability mass function with parameter p, $P_p(1) = 1 - P_p(0) = p$, and write $D(P\|Q) = \sum_{x \in B} P(x) \log[P(x)/Q(x)]$ for the relative entropy (or Kullback-Leibler divergence) between two probability mass functions P, Q on the same discrete set B; throughout, 'log' denotes the natural logarithm to base e.

Theorem. *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 Q_k of $(X_1, X_2, \ldots, X_k)$ and the mixture $M_{k,\mu} := \int P_p^k d\mu(p)$ satisfies:*

$$D(Q_k\|M_{k,\mu}) \leq \frac{5k^2 \log n}{n - k}. \tag{1.1}$$

By Pinsker's inequality [7, 19], $\|P - Q\|^2 \leq 2D(P\|Q)$, the theorem also implies that,

$$\|Q_k - M_{k,\mu}\| \leq k\Big(\frac{10 \log n}{n - k}\Big)^{\frac{1}{2}}, \tag{1.2}$$

where $\|P - Q\| := 2\sup_B |P(B) - Q(B)|$ denotes the total variation distance between P and Q. This bound is suboptimal in that, as shown by Diaconis and Freedman [13], the correct rate with respect to the total variation distance in (1.2) is $O(k/n)$. On the other hand, (1.1) gives an explicit bound for the stronger notion of relative entropy 'distance.'

Rather than to obtain optimal rates, our primary motivation is to illustrate how elementary information-theoretic ideas can be used to provide an alternative proof strategy for de Finetti's theorem, following a long series of works developing this point of view, including information-theoretic proofs of Markov chain convergence [21, 16], the central limit theorem [4, 2], Poisson and compound Poisson approximation [18, 3], and the Hewitt-Savage 0-1 law [20].

Before turning to the proof, we mention that there are numerous generalisations and extensions of de Finetti's classical theorem and its finite version along different directions; see, e.g., [11] and the references therein. The classical de Finetti representation theorem has been shown to hold for exchangeable processes with values in much more general spaces than $\{0, 1\}$ [15], and for mixtures of Markov chains [12]. Recently, an elementary proof of de Finetti's theorem for the binary case was given in [17], a more analytic proof appeared in [1], and connections with category theory were drawn in [14].

2   Proof of the finite de Finetti theorem

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 the lemma 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. Recall that the entropy $H(X)$ of X, often viewed as a measure of the inherent "randomness" of X [6], is defined as, $H(X) = H(P_X) = -\sum_x P_X(x) \log P_X(x)$, where the sum is over all possible values of X with nonzero probability. Similarly, the conditional entropy of Y given X is,

$$H(Y|X) = -\sum_x P_X(x) \sum_y P_{Y|X}(y|x) \log P_{Y|X}(y|x),$$

where $P_{Y|X}(y|x) = P_{XY}(x, y)/P_X(x)$.

The mutual information between X and Y is $I(X; Y) = H(Y) - H(Y|X)$, and it can also be expressed as:

$$I(X; Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y) = D(P_{XY} \| P_X P_Y).$$

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.

From the definition, an obvious interpretation of $I(X; Y)$ is as a measure of the amount of "common randomness" in X and Y. Additionally, since $I(X; Y)$ is always nonnegative and equal to zero iff X and Y are independent, the mutual information can be viewed as a universal, nonlinear measure of dependence between X and Y. See [6] for standard properties of the entropy, relative entropy and mutual information.

Finally, we record an elementary bound that will be used in the proof of the lemma. Write $h(p) = -p \log p - (1-p) \log(1-p)$, $p \in [0, 1]$, for the binary entropy function. Then a simple Taylor expansion gives:

$$|h(p) - h(q)| \leq |p - q| \times \max\left\{ \left|\log\left(\frac{1-p}{p}\right)\right|, \left|\log\left(\frac{1-q}{q}\right)\right| \right\}, \qquad p, q \in (0, 1). \qquad (2.1)$$

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

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

Proof. We assume without loss of generality that $k \leq n/2$, for otherwise the result is trivially true since the mutual information in the statement is always no greater than 1. Also, if $\ell = 0$ or n the conditional mutual information is zero and the result is again trivially true. Let Q_n denote the p.m.f. of X_1^n. By exchangeability, conditional on A_ℓ, all sequences in $\{0, 1\}^n$ with exactly ℓ 1s have the same probability under Q_n, so X_1^n conditional on A_ℓ is uniformly distributed among all such sequences. This implies that for all $1 \leq k \leq n/2$, $1 \leq i \leq k-1$, and $1 \leq \ell \leq n-1$,

$$\mathbb{P}(X_i = 1 | N_{1,n} = \ell, N_{i+1,k}) = \frac{\binom{n-(k-i)-1}{\ell - N_{i+1,k} - 1}}{\binom{n-(k-i)}{\ell - N_{i+1,k}}} = \frac{\ell - N_{i+1,k}}{n - (k-i)}.$$

For the mutual information we have:

$$
\begin{aligned}
I(X_i; X_{i+1}^k | A_\ell) &= \mathbb{E}\left(h\left(\frac{\ell}{n}\right) - h\left(\frac{\ell - N_{i+1,k}}{n - (k-i)}\right) \middle| N_{1,n} = \ell \right) \\
&\leq \mathbb{E}\left(\left| h\left(\frac{\ell}{n}\right) - h\left(\frac{\ell - N_{i+1,k}}{n - (k-i)}\right) \right| \mathbb{I}_{\{\ell+k-i-n+1 \leq N_{i+1,k} < \ell\}} \middle| N_{1,n} = \ell \right) \\
&\quad + h\left(\frac{\ell}{n}\right) \mathbb{P}(N_{i+1,k} = \ell | A_\ell) + h\left(\frac{\ell}{n}\right) \mathbb{P}(N_{i+1,k} \leq \ell + k - i - n | A_\ell). \quad (2.2)
\end{aligned}
$$

If the probability in the third term above is nonzero, then necessarily $\ell \geq n - k + 1$ and thus, using $n \geq 2k$, $h(\ell/n) \leq 2\frac{k}{n-k} \log n$. On the other hand, if $\ell + k - i - n + 1 \leq N_{i+1,k} < \ell$, then both ℓ/n and $(\ell - N_{i+1,k})/(n - (k-i))$ are between $1/n$ and $(n-1)/n$, so from (2.1) the first term in (2.2) is bounded above by,

$$\frac{\ell(k-i) + n\mathbb{E}(N_{i+1,k} | N_{1,n} = \ell)}{n(n - (k-i))} \log n = \frac{2\ell(k-i)}{n(n - (k-i))} \log n.$$

Finally, by Markov's inequality, the probability in the second term is no more than k/n, while the binary entropy is always bounded above by 1. Combining these three estimates yields,

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

The result follows. □

We are now ready to prove the theorem. By the bound in the lemma,

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

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,

$$\sum_{i=1}^{k-1} I(X_i; X_{i+1}^k | A_\ell) = \sum_{i=1}^{k-1} \left[H(X_i|A_\ell) + H(X_{i+1}^k|A_\ell) - H(X_i^k|A_\ell) \right]$$

$$= \left[\sum_{i=1}^{k} H(X_i|A_\ell) \right] - H(X_1^k|A_\ell)$$

$$= D\big(Q_{X_1^k|A_\ell} \big\| Q_{X_1|A_\ell} \times \cdots \times Q_{X_k|A_\ell}\big),$$

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

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

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

Remarks. 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. Moreover, 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 P_p^k is a bounded and continuous function of $p \in [0,1]$, we have for any $x_1^k \in \{0,1\}^k$,

$$M_{k,\mu_n}(x_1^k) = \int P_p^k(x_1^k)d\mu_n(p) \to M_{k,\mu}(x_1^k) = \int P_p^k(x_1^k)d\mu(p),$$

and by our theorem, $\|Q_k - M_{n,k}\| = O(\sqrt{(\log n)/n})$. Therefore, we can conclude that,

$$Q_k = \int P_p^k d\mu(p),$$

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

Finally we note that the argument used in the proof of the lemma as well as the proof of our theorem can easily be extended to provide corresponding results for exchangeable vectors taking values in any finite set. But as the the constants involved become quite cumbersome and our main motivation is to illustrate the connection with information-theoretic ideas (rather to obtain the most general possible results), we have chosen to restrict attention to the binary case.

References

[1] I. Alam. A nonstandard proof of de Finetti's theorem for Bernoulli random variables. *Journal of Stochastic Analysis*, 1(4):15, 2020. MR4188596

[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. MR2083473

[3] A.D. Barbour, O. Johnson, I. Kontoyiannis, and M. Madiman. Compound Poisson approximation via information functionals. *Electron. J. Probab*, 15:1344–1369, 2010. MR2721049

[4] A.R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, 14(1):336–342, January 1986. MR0815975

[5] M. Bayarri and J.O. Berger. The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004. MR2082147

[6] T.M. Cover and J.A. Thomas. *Elements of information theory*. J. Wiley & Sons, New York, second edition, 2012. MR2239987

[7] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967. MR0219345

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

[9] B. De Finetti. La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré*, 7(1):1–68, 1937. MR1508036

[10] P. Diaconis. Finite forms of de Finetti's theorem on exchangeability. *Synthese*, 36(2):271–281, 1977. MR0517222

[11] P. Diaconis. Recent progress on de Finetti's notions of exchangeability. *Bayesian Statistics*, 3:111–125, 1988. MR1008047

[12] P. Diaconis and D.A. Freedman. De Finetti's theorem for Markov chains. *Ann. Probab.*, 8(1):115–130, 1980. MR0556418

[13] P. Diaconis and D.A. Freedman. Finite exchangeable sequences. *Ann. Probab.*, 8(4):745–764, 1980. MR0577313

[14] T. Fritz, T. Gonda, and P. Perrone. De Finetti's Theorem in Categorical Probability. *preprint*, arXiv:2105.02639, 2021.

[15] E. Hewitt and L.J. Savage. Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.*, 80(2):470–501, 1955. MR0076206

[16] 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. MR0171317

[17] W. Kirsch. An elementary proof of de Finetti's theorem. *Statist. Probab. Lett.*, 151:84–88, 2019. MR3938896

[18] I. Kontoyiannis, P. Harremoës, and O. Johnson. Entropy and the law of small numbers. *IEEE Trans. Inform. Theory*, 51(2):466–472, February 2005. MR2236061

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

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

[21] 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. MR0132570

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards

- Free for authors, free for readers

- Quick publication (no backlog)

- Secure publication (LOCKSS[1])

- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4] , ProjectEuclid[5]

- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)

- Submit your best articles to EJP-ECP

- Choose EJP-ECP over for-profit journals

---

[1]LOCKSS: Lots of Copies Keep Stuff Safe http://www.lockss.org/
[2]EJMS: Electronic Journal Management System http://www.vtex.lt/en/ejms.html
[3]IMS: Institute of Mathematical Statistics http://www.imstat.org/
[4]BS: Bernoulli Society http://www.bernoulli-society.org/
[5]Project Euclid: https://projecteuclid.org/
[6]IMS Open Access Fund: http://www.imstat.org/publications/open.htm