Information-Theoretic de Finetti-style Theorems

Lampros Gavalakis   Ioannis Kontoyiannis

University of Cambridge

ITW, 1- 2 November 2022
Mumbai, India

Outline

1. De Finetti's Representation Theorem
2. Information-Theoretic Approach
   - First Proof
   - Second Proof

De Finetti's Representation Theorem

$X_1, X_2, \ldots, X_n$ random variables in A

De Finetti's Representation Theorem

$X_1, X_2, \ldots, X_n$ random variables in A

Exchangeability: $X_1, X_2, \ldots, X_n$ and $X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}$ have the same distribution for every permutation π on $\{1, 2, \ldots, n\}$.

De Finetti's Representation Theorem

$X_1, X_2, \ldots, X_n$ random variables in A

Exchangeability: $X_1, X_2, \ldots, X_n$ and $X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}$ have the same distribution for every permutation π on $\{1, 2, \ldots, n\}$.

$\{X_i\}$ is *exchangeable*, if $X_1, X_2, \ldots, X_n$ are exchangeable for every $n \geq 1$

De Finetti's Representation Theorem

Theorem (de Finetti, 1930s)

$\{X_i\}$ *is exchangeable if and only if it is a mixture of i.i.ds:*

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

for some measure μ on the simplex $\mathcal{P}$

De Finetti's Representation Theorem

Theorem (de Finetti, 1930s)

$\{X_i\}$ *is exchangeable if and only if it is a mixture of i.i.ds:*

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

for some measure μ *on the simplex* $\mathcal{P}$

Fails if a *finite* collection of random variables $(X_1, X_2, \ldots, X_n)$ is exchangeable for some *fixed* n [Diaconis & Freedman, '80]:

De Finetti's Representation Theorem

Theorem (de Finetti, 1930s)

$\{X_i\}$ is exchangeable if and only if it is a mixture of i.i.ds:

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

for some measure μ on the simplex $\mathcal{P}$

Fails if a *finite* collection of random variables $(X_1, X_2, \ldots, X_n)$ is exchangeable for some *fixed* n [Diaconis & Freedman, '80]:

$$\mathbb{P}(X_1 = 1, X_2 = 0) = \mathbb{P}(X_1 = 0, X_2 = 1) = 1/2$$
$$\mathbb{P}(X_1 = 0, X_2 = 0) = \mathbb{P}(X_1 = 1, X_2 = 1) = 0$$

De Finetti's Representation Theorem

Theorem (de Finetti, 1930s)

$\{X_i\}$ is exchangeable if and only if it is a mixture of i.i.ds:

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

for some measure μ on the simplex $\mathcal{P}$

Fails if a *finite* collection of random variables $(X_1, X_2, \ldots, X_n)$ is exchangeable for some *fixed n* [Diaconis & Freedman, '80]:

$$\mathbb{P}(X_1 = 1, X_2 = 0) = \mathbb{P}(X_1 = 0, X_2 = 1) = 1/2$$
$$\mathbb{P}(X_1 = 0, X_2 = 0) = \mathbb{P}(X_1 = 1, X_2 = 1) = 0$$

However, $P_{X_1^k}$ is *close* to $M_{k,\mu}$

Information-Theoretic Approach: First Proof

Closeness will be measured in terms of relative entropy

$$D(P\|Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}$$

Information-Theoretic Approach: First Proof

Closeness will be measured in terms of relative entropy

$$D(P\|Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}$$

Theorem

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 $1 \leq k \leq n$,

$$D(P_{X_1^k} \| M_{k,\mu}) \leq \frac{5k^2 \log n}{n-k}$$

Information-Theoretic Approach: First Proof

Closeness will be measured in terms of relative entropy

$$D(P\|Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}$$

Theorem

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 $1 \le k \le n$,

$$D(P_{X_1^k} \| M_{k,\mu}) \le \frac{5k^2 \log n}{n-k}$$

- By Pinsker's inequality

$$\|P_{X_1^k} - M_{k,\mu}\| \le \sqrt{10}k \left(\frac{\log n}{n-k}\right)^{1/2},$$

[Diaconis & Freedman, '80]: $O(k/n)$ for the total variation distance

Information-Theoretic Approach: Proof ideas

- Condition on the *type*

Information-Theoretic Approach: Proof ideas

- Condition on the *type*

Lemma

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

$$I\left(X_i; X_{i+1}^k \middle| \sum_{j=1}^n X_j = \ell\right) \leq \frac{5k \log n}{n-k}.$$

Information-Theoretic Approach: Proof ideas

- Condition on the *type*

Lemma

For all $1 \le k \le n$, all $1 \le i \le k-1$, and any $0 \le \ell \le n$:

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

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

Information-Theoretic Approach: Proof ideas

- Condition on the *type*

Lemma

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

$$I\left(X_i; X_{i+1}^k \,\middle|\, \sum_{j=1}^n X_j = \ell\right) \leq \frac{5k \log n}{n-k}.$$

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

- Joint convexity of relative entropy

Information-Theoretic Approach: Second Proof

Notation: $\hat{P}_{X_1^n}(a) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{X_i=a\}}$ (empirical distribution of X_1^n)

Information-Theoretic Approach: Second Proof

Notation: $\hat{P}_{X_1^n}(a) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{X_i=a\}}$ (empirical distribution of X_1^n)

Theorem (Gibbs' conditioning principle)

Let X_1^n IID with distribution Q, having full support on A. Suppose E is a closed and convex set of probability distributions on A, which has nonempty interior, and does not contain Q. If $Q^n(\hat{P}_n \in E) \neq 0$ eventually, then

$$\mathbb{E}(\hat{P}_n(a)|\hat{P}_n \in E) = \mathbb{P}(X_1 = a|\hat{P}_n \in E) \to P^*(a)$$

where $P^* \in E$ uniquely achieves $D(P^*||Q) = inf_{P\in E} D(P||Q)$

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 1:*

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 1:*
  - $\{X_n\}$ exchangeable $\Rightarrow$ stationary

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 1:*
  - $\{X_n\}$ exchangeable $\Rightarrow$ stationary
    $\Rightarrow \hat{P}_{X_1^n} \to P$ (random on A) a.s. by the ergodic theorem

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 1:*
  - $\{X_n\}$ exchangeable $\Rightarrow$ stationary
    $\Rightarrow \hat{P}_{X_1^n} \to P$ (random on A) a.s. by the ergodic theorem
  - Let $\{Y_n\}$ IID uniformly distributed on A. By exchangeability

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

for any n-type Q_n

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 1:*
  - $\{X_n\}$ exchangeable $\Rightarrow$ stationary
    $\Rightarrow \hat{P}_{X_1^n} \to P$ (random on A) a.s. by the ergodic theorem
  - Let $\{Y_n\}$ IID uniformly distributed on A. By exchangeability

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

    for any n-type Q_n
- *Step 2*:

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 1:*
  - $\{X_n\}$ exchangeable $\Rightarrow$ stationary
    $\Rightarrow \hat{P}_{X_1^n} \to P$ (random on A) a.s. by the ergodic theorem
  - Let $\{Y_n\}$ IID uniformly distributed on A. By exchangeability

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

    for any n-type Q_n
- *Step 2*: Fix $\{Q_n\}$ along which $\hat{P}_{X_1^n} \to$ some Q .

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 1:*
  - $\{X_n\}$ exchangeable $\Rightarrow$ stationary
    $\Rightarrow$ $\hat{P}_{X_1^n} \to P$ (random on A) a.s. by the ergodic theorem
  - Let $\{Y_n\}$ IID uniformly distributed on A. By exchangeability

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

    for any n-type Q_n

- *Step 2*: Fix $\{Q_n\}$ along which $\hat{P}_{X_1^n} \to$ some Q . By symmetry

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

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 1:*
  - $\{X_n\}$ exchangeable $\Rightarrow$ stationary
    $\Rightarrow \hat{P}_{X_1^n} \to P$ (random on A) a.s. by the ergodic theorem
  - Let $\{Y_n\}$ IID uniformly distributed on A. By exchangeability

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

    for any n-type Q_n

- *Step 2*: Fix $\{Q_n\}$ along which $\hat{P}_{X_1^n} \to$ some Q . By symmetry

$$\mathbb{P}(X_1 = a | \hat{P}_{X_1^n} = Q_n) = \mathbb{E}\left(\mathbb{I}_{\{Y_1 = a\}} \middle| \hat{P}_{Y_1^n} = Q_n\right)$$
$$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Y_i = a\}} \middle| \hat{P}_{Y_1^n} = Q_n\right) = Q_n(a)$$

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

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 3*: Generalise to blocks (of size 2 for simplicity)

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 3*: Generalise to blocks (of size 2 for simplicity)
  Fix one of the almost all realisations $Q_n \to$ some Q

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 3*: Generalise to blocks (of size 2 for simplicity)
  Fix one of the almost all realisations $Q_n \to$ some Q
  Define $Z_n = (Y_{2n-1}, Y_{2n})$ IID uniform on $A \times A$.

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 3*: Generalise to blocks (of size 2 for simplicity)
  Fix one of the almost all realisations $Q_n \to$ some Q
  Define $Z_n = (Y_{2n-1}, Y_{2n})$ IID uniform on $A \times A$.

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

where

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

Information-Theoretic Approach: Second Proof (Heuristic)

- *Step 3*: Generalise to blocks (of size 2 for simplicity)
  Fix one of the almost all realisations $Q_n \to$ some Q
  Define $Z_n = (Y_{2n-1}, Y_{2n})$ IID uniform on $A \times A$.

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

  where

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

  $W^*_\ell = Q \times Q$ uniquely achieves $\min_{W \in E(Q)} D(W \| U)$,
  where $U =$ uniform on $A \times A$

Information-Theoretic Approach: Second Proof (Heuristic)

Apply the *Gibbs' conditioning principle***

Information-Theoretic Approach: Second Proof (Heuristic)

Apply the *Gibbs' conditioning principle***

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

Information-Theoretic Approach: Second Proof (Heuristic)

Apply the *Gibbs' conditioning principle***

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

and thus

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

Information-Theoretic Approach: Second Proof (Heuristic)

Apply the *Gibbs' conditioning principle***

$$\Rightarrow \lim_{\ell\to\infty} \mathbb{P}\big(Z_1 = (a_1, a_2)\big|\hat{P}_{Z_1^\ell} \in E(Q_{2\ell})\big)$$

$$= \lim_{\ell\to\infty} W_\ell^*(a_1, a_2)$$

$$= \lim_{\ell\to\infty} Q_{2\ell}(a_1)Q_{2\ell}(a_2)$$

$$= Q(a_1)Q(a_2),$$

and thus

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

Similarly, for any $k \geq 1$,

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

- *Step 4*:

- *Step 4*:
  Bounded converge theorem

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

- *Step 4*:
  Bounded converge theorem

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

  which is de Finetti's theorem!

- *Step 4*:
  Bounded converge theorem

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

  which is de Finetti's theorem!

  ** But $E(Q)$ has empty interior!

Nevertheless,

Theorem (Second finite de Finetti theorem)

If X_1^n are exchangeable, for any $1 \leq k \leq (n/100)^{1/3}$,

$$D(P_{X_1^k} \| M_{k,\mu}) \leq 2\delta_{n,k} + ke^{-\frac{n}{k}\delta_{n,k}} \left(\frac{n}{k} + 1\right)^{2|A|^k} \log n$$

with $\alpha_{n,k} = \left[\frac{2k}{\sqrt{n}}\left(\frac{1+2k}{\sqrt{n}} + 1\right)\right]^{1/2}$ and $\delta_{n,k} = \alpha_{n,k} \log(|A|^k / \alpha_{n,k})$

Further Work

- Applications (Machine Learning, Quantum Information Theory, etc.)

Further Work

- Applications (Machine Learning, Quantum Information Theory, etc.)
- Extensions to more general spaces

Further Work

- Applications (Machine Learning, Quantum Information Theory, etc.)
- Extensions to more general spaces
- Third proof ?

Thank you!