ISSN 1066-5307, Mathematical Methods of Statistics, 2018, Vol. 27, No. 2, pp. 145–161. © Allerton Press, Inc., 2018.

The Deficiency Introduced by Resampling

T. Wiklund[1*]

[1] Dept. Math., Uppsala Univ., Uppsala, Sweden

Received October 24, 2017; in final form, February 6, 2018

Abstract—When the classical nonparametric bootstrap is implemented by a Monte-Carlo procedure one resamples values from a sequence of, typically, independent and identically distributed ones. But what happens when a decision has to be taken based on such resampled values? One way to quantify the loss of information due to this resampling step is to consider the deficiency distance, in the sense of Le Cam, between a statistical experiment of n independent and identically distributed observations and the one consisting of m observations taken from the original n by resampling with replacement. By comparing with an experiment where only subsampling with a random subsampling size has been performed one can bound the deficiency in terms of the amount of information contained in additional observations. It follows for certain experiments that the deficiency distance is proportional to the expected fraction of observations missed when resampling.

Keywords: deficiency, resampling, bootstrap, comparison of experiments.

2010 Mathematics Subject Classification: 62B15.

DOI: 10.3103/S1066530718020047

1. INTRODUCTION

The use of the bootstrap and similar procedures appears to have become almost synonymous with their approximation by Monte-Carlo methods. While this may be what make them tractable in practice they entail some form of resampling which is a perturbation of the data on top of whatever noise and general uncertainty is already inherent in the statistical experiment. Our goal here is to quantify the amount of information relevant to inference that is lost due to this randomization when resampling with replacement.

Formally we consider a statistical experiment $\mathcal{E} = (X, P_\theta; \theta \in \Theta)$ given by a family of probability measures $(P_\theta)_{\theta\in\Theta}$ on a sample space X indexed by a parameter space Θ. This description captures all relevant information of the scenario where one is observing a quantity from X with probability law governed by P_θ for some unknown $\theta \in \Theta$. Based on this experiment we may form two different experiments. First $\mathcal{E}^{\otimes n}$ will correspond to observing n independent and identically distributed observations from $\mathcal{E}$. Secondly $\mathsf{R}_m\mathcal{E}^{\otimes n}$ will be obtained by observing a vector of m values given by resampling with replacement from a collection of n independent and identically distributed values as in $\mathcal{E}^{\otimes n}$. Typically one would have $m = an$ for some positive integer a.

The deficiency $\delta(\mathsf{R}_m\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n})$ as introduced by Le Cam [11] can then be given the following operational interpretation. It is the smallest $\varepsilon \geq 0$ such that any decision, for any suitable decision space and normalized loss function, based on n independent and identically distributed observations is matched by a decision with a risk at most ε greater based on m resampled observations. From this point of view our goal is to bound this deficiency in terms of m, n, and $\mathcal{E}$. Another way to pose our question is given by asking how well, in the worst case scenario, a procedure based on resampling can perform relative to an optimal one that does not. Such procedures include ideas such as bootstrap aggregation (bagging) and at least roughly describe many more traditional applications of the bootstrap. Bootstrap procedures are often studied, in an appropriate sense, conditionally on the underlying vector of observations resampled from. We are here interested in the unconditional behavior. Indeed, conditioning

[*] E-mail: tilo.wiklund@math.uu.se

on the underlying vector would eliminate the dependence on Θ and thus tell us little regarding the effect on inference about Θ.

Our primary tool will be an approximation given by considering only the information lost due to samples not being included when resampling. This means ignoring any additional confusion introduced by certain values appearing in a larger or smaller proportion than in the original sample. When the original is known to contain no duplicates, for example when working with continuous measures on the real numbers, this approximation is actually exact, since any duplicates among the resampled values must be artefacts of resampling. When duplicates are allowed, for example in the discrete case, it is still the case that any observations not resampled cannot be recovered. In the general case one therefore gets a bound from below. Under very mild regularity conditions an experiment is equivalent to a continuous one, in an appropriate sense. In other words, the experiment is equivalent to an experiment where the inequality is an equality.

The above argument can be formalized by somewhat cumbersome but essentially elementary means. Doing so yields non-asymptotic inequalities relating the deficiency of the resampled experiment to the amount of information gained, again in the sense of deficiency, from additional observations. This quantity is somewhat better understood thanks to the work of Torgersen [25, 27], Le Cam [12], Helgeland [6], and Mammen [13].

If every observation is essential, so that removing even one will give a maximal deficiency, the question is closely related to the coupon collectors problem. A different situation arises in the parametric setting studied in the above papers by Le Cam, Helgeland, and Mammen. Here additional observations become less and less useful so that missing a few has a less dramatic effect. Indeed, we show that for such experiments the rate at which the number of resamples must increase with the original sample size for the deficiency to tend to zero is different from the rate in the coupon collectors problem.

Recall that the empirical measure of an independent and identically distributed sample will always be a sufficient statistic assuming the original sample size is known. In particular, any decision based on the original is matched by the one at least as good based only on the empirical measure and the deficiency from observing it instead of the original sample is zero. What we are considering here is thus *only* the effect of resampling, not the behavior of decision based on substituting the empirical measure for the unknown underlying measure. Results in the latter direction can be found, for example, in Putter and Van Zwet [19] or Mammen [14] and references therein.

While the author is unaware of any previous results in exactly this direction, it is conceptually similar to a number of other publications that consider approximate sufficiency of an experiment after some perturbation. First of all there are the already mentioned results of Torgersen, Le Cam, Helgeland, and Mammen on the amount of information lost due to dropping observations. Beyond these, without making any claim as to completeness, it is worth mentioning results by Marohn [17] and Falk and Marohn [3] on censoring, a series of papers by Reiss, Falk and Weller [21], Reiss [20], Janssen and Reiss [7], Marohn [16] and Janssen and Marohn [8] on information in various order statistics, papers by Milstein and Nussbaum [18], Konakov, Mammen and Woerner [10], Mariucci [15] and Genon-Catalot and Larédo [4] on the deficiency introduced by discretely observing diffusion processes and Strasser [24] on quantization.

The structure of the paper is as follows. Section 2 defines our notation and very briefly surveys some definitions. Section 3 presents the main results, given as Theorems 9 and 11, and examples collected in Corollary 12. Sections 4 and 5 contain the bulk of theory and proofs, the former with the goal of proving Theorem 9 and the latter towards Theorem 11 and examples collected in Corollary 12. After some concluding remarks in Section 6 one finds Appendix A containing a number of auxiliary results used in the paper, but which do not quite belong to the main narrative, and Appendix B to which the more technical and routine proofs have been condemned.

2. NOTATION AND CONVENTIONS

Throughout let $X = (\mathcal{X}, \mathscr{X})$, $Y = (\mathcal{Y}, \mathscr{Y})$, and $Z = (\mathcal{Z}, \mathscr{Z})$ be measurable spaces and let $\mathsf{S}\colon X \to_k Y$ and $\mathsf{T}\colon Y \to_k Z$ be *Markov kernels*.

If $\mathcal{B}([0,1])$ is the σ-algebra of Borel sets on $[0,1]$, we have that S is, formally, a function $\mathcal{X} \times \mathscr{Y} \to [0,1]$, that is $(\mathscr{X}, \mathcal{B}([0,1]))$-measurable when the second argument is fixed and defines a probability measure whenever the first argument is fixed. We will often write $\mathsf{S}_x(B) := \mathsf{S}(x, B)$ and treat S as a measurable map from X into the space of probability measures on Y, equipped with the smallest σ-algebra such that for all $B \in \mathscr{Y}$ the map $\mu \mapsto \mu(B)$ is measurable. Markov kernels act on probability measures with $\mathsf{S}P$ being the measure on Y defined by $\mathsf{S}P(A) = \int \mathsf{S}_x(A)\, P(dx)$ for $A \in \mathscr{Y}$.

Given two measures P on X and Q on Y their product measure on $X \otimes Y$ is denoted by $P \otimes Q$ with powers denoted by $P^{\otimes n} = P \otimes \cdots \otimes P$ (where the right-hand side has n factors P). Their direct sum $P \oplus Q$ on the disjoint(!) union of X and Y is determined by $(P \oplus Q)(A) = P(A \cap X) + Q(A \cap Y)$. If P and Q are probability measures and $\alpha \in [0, 1]$, then $\alpha P \oplus (1 - \alpha)Q$ is also a probability measure. The notion extends in the obvious way to finite indexed families $(P_i)_{i \in I}$ and probability measures μ on I with the corresponding measure denoted by $\bigoplus_{i \in I} \mu(\{i\})P_i$. In other words if I is a finite set with some probability measure P_i for each $i \in I$ on disjoint $(\mathcal{X}_i, \mathscr{X}_i)$, then for any $A \in \sigma(\bigcup_{i \in I} \mathscr{X}_i)$ we have $(\bigoplus_{i \in I} \mu(\{i\})P_i)(A) := \sum_{i \in I} \mu(\{i\})P_i(A \cap X_i)$. The use of $\bigoplus$ rather than $\sum$ is due to the latter being reserved for taking (not necessarily direct) sums as signed measures.

An experiment with parameter space Θ is specified as $(X, P_\theta; \theta \in \Theta)$ or $(P_\theta; \theta \in \Theta)$ depending on whether the underlying measurable space is required. A generic experiment is denoted by $\mathcal{E}$. For experiments $\mathcal{E}$ and $\mathcal{F}$ with a common parameter space Θ and Markov kernel S, we let $\mathcal{E} \otimes \mathcal{F}$, $\mathsf{S}\mathcal{E}$ and so on be defined point-wise for each parameter $\theta \in \Theta$.

The deficiency of $\mathcal{E} = (P_\theta; \theta \in \Theta)$ with respect to $\mathcal{F} = (Q_\theta; \theta \in \Theta)$ will be denoted by $\delta(\mathcal{E}, \mathcal{F})$. It is characterized either as a supremum over the difference in achievable risk over a good class of decision problems or as the infimum over an appropriate class of linear maps, often given by Markov kernels, sending P_θ to a measure close in total variation to Q_θ for every $\theta \in \Theta$ (see a standard reference such as the textbook of Torgersen [28]). If $\delta(\mathcal{E}, \mathcal{F}) = 0$ the latter experiment $\mathcal{F}$ is said to be *less informative* than $\mathcal{E}$ and will be denoted $\mathcal{F} \prec \mathcal{E}$. If also $\mathcal{E} \prec \mathcal{F}$, the experiments are said to be equivalent, denoted $\mathcal{E} \sim \mathcal{F}$.

The total variation distance between measures, and consequently the deficiency between experiments, is taken to be normalized between 0 and 1. Note that it is also common to normalize the quantities to lie between 0 and 2, which corresponds to allowing negative test functions and losses.

Finally $\mathrm{Urn}(n, m)$ will denote the law of the number of non-empty urns after independently and uniformly throwing m balls into n initially empty urns. For a standard reference see Johnson, Kemp and Kotz [9]. At times we will (explicitly) introduce the abbreviation $v = v(n, m) = \mathrm{Urn}(n, m)$, depending on whether dependence on n, m needs to be made explicit.

3. MAIN RESULTS AND EXAMPLES

For positive integers m and n let $F_{m,n}$ denote the set of functions from $\{1, \ldots, m\}$ to $\{1, \ldots, n\}$. One may think of a probability measure α on $F_{m,n}$ as a sampling strategy: if f has law α, it corresponds to picking $f(1), \ldots, f(m)$ from the population $\{1, \ldots, n\}$. Replacing $F_{m,n}$ by $F_n := \bigcup_m F_{m,n}$ one can allow for random sample sizes.

Proposition 6. *Let $m, n \in \mathbb{N}$ be positive integers, α a probability measure on $F_{m,n}$ and X a measurable space with underlying set $\mathcal{X}$. One can then define a Markov kernel $\mathsf{R}_{\alpha,X} \colon X^{\otimes n} \to_k X^{\otimes m}$ for $x \in \mathcal{X}^n$ by $(\mathsf{R}_{\alpha,X})_x = \sum_{f \in F_{m,n}} \alpha(\{f\})\delta_{x_f}$, where $x_f = (x_{f(1)}, \ldots, x_{f(m)})$ and δ_{x_f} denotes the point measure at x_f.*

More generally, if α is a probability measure on $F_n = \bigcup_m F_{m,n}$, the above definition gives a Markov kernel $\mathsf{R}_{\alpha,X} \colon X^{\otimes n} \to_k \bigoplus_m X^{\otimes m}$.

Note that the construction is contravariant in the sense that picking functions going from $\{1, \ldots, m\}$ to $\{1, \ldots, n\}$ gives a Markov kernel from $X^{\otimes n}$ to $X^{\otimes m}$.

We will be concerned almost exclusively with two special cases $\mathsf{R}_{m,n,X}$ and $\mathsf{S}_{m,n,X}$. The former corresponds to having α be the uniform measure on $F_{m,n}$ and the latter is defined when $m \le n$ and then corresponds to taking α as the uniform measure of the subset of injective functions. These special cases correspond to uniformly resampling m elements from a vector of n with and without replacement, respectively.

Proposition 7. *Let $m, n, X, \mathcal{X}$ be as in Proposition 6 and let $\mathsf{R}_{m,n,X} = \mathsf{R}_{\alpha,X}$, where α is the uniform measure on $F_{m,n}$. For any $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, it then holds that $(\mathsf{R}_{m,n,X})_x$ is the law of a vector of size m taken uniformly and* with *replacement from x.*

In case $m \leq n$ let $\mathsf{S}_{m,n,X} = \mathsf{R}_{\beta,X}$ with β the uniform measure on the set of injective functions in $F_{m,n}$. Analogously $(\mathsf{S}_{m,n,X})_x$ is the law of a vector of size m taken uniformly without *replacement from x.*

For some experiment $\mathcal{E} = (X, P_\theta; \theta \in \Theta)$ our objective is now to study $\mathsf{R}_m \mathcal{E}^{\otimes n}$, an experiment given by first taking n independent and identically distributed observations according to $\mathcal{E}$ and observing a vector of m values taken uniformly, with replacement, from this vector. When resampling with replacement one will, in general, miss some values from the original vector even if $m \geq n$. Thus if one takes m values from $(x_1, \ldots, x_n)$ one is implicitly choosing randomly some subvector $(x_{i_1}, \ldots, x_{i_\kappa})$ of random size $\kappa \leq n$ from which to sample exhaustively. Thinking of sampling as throwing m balls into n urns, the random size κ corresponds to the number of non-empty urns in the classical uniform urn occupancy problem (see Section 2), the law of which we denote $\mathrm{Urn}(n, m)$.

Proposition 8. *Fix any positive integers $n, m \in \mathbb{N}$ and introduce the shorthand $\upsilon = \mathrm{Urn}(n, m)$. For any measurable space X one then has*

$$\mathsf{R}_{m,n,X} = \mathsf{D}\mathsf{S}_{\upsilon,n,X}, \tag{1}$$

where $\mathsf{S}_{\upsilon,n,X} = \sum_{m=1}^{n} \upsilon(\{m\}) \mathsf{S}_{m,n,X}$, $\mathsf{R}_{m,n,X}$ and $\mathsf{S}_{m,n,X}$ are as in Proposition 6, *and D is a Markov kernel.*

The Markov kernel D above simply corresponds to randomly padding a vector of length k to a vector of length m by copying values.

In general the effect of resampling will depend in a non-trivial way on the structure of the underlying set of measures. The most manifest case is an experiment $\mathcal{E} = (X, P_\theta; \theta \in \Theta)$, where X is *countably separating* and $(P_\theta)_{\theta \in \Theta}$ is *continuous* in the sense that $P_\theta(\{x\}) = 0$ for each x and θ. Recall that being countably separating means that there exists a countable family of measurable sets that separates the point of X. This is equivalent to the diagonal set in $X^{\otimes 2}$ being measurable and implies that all singleton sets are measurable as well as continuous measures being the same as *non-atomic* measures [1]. In this situation any duplicates found in the resampled vector are almost surely artefacts of the resampling procedure and one can recover the exact random subset of values not missed during resampling. In this case the resampled experiment $\mathsf{R}_{m,n,X}\mathcal{E}^{\otimes n}$ is equivalent to the experiment $\mathsf{S}_{\upsilon,n,X}\mathcal{E}^{\otimes n}$, where only subsampling, of size governed by $\mathrm{Urn}(n, m)$, has been performed.

Theorem 9. *Fix some $n, m \in \mathbb{N}$, a measurable space X, and let $\upsilon = \mathrm{Urn}(n, m)$ and $\mathsf{S}_{\upsilon,n,X}$ be as in Proposition* 8. *For any experiment $\mathcal{E}$ on X one has*

$$\delta(\mathsf{R}_{m,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) \geq \delta(\mathsf{S}_{\upsilon,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}). \tag{2}$$

Moreover if X is countably separating, then the equality holds at least whenever $\mathcal{E}$ is continuous or non-atomic and the lower bound is the best that depends only on the equivalence class of $\mathcal{E}$.

Note that countably separating is certainly weaker than being countably generated (separable). It therefore holds at least for Polish spaces and thus in many, if not most, practical examples.

Applying a projection onto a set of k coordinates chosen independently of the observation gives an experiment equivalent to $\mathcal{E}^{\otimes k}$, so that $\mathsf{S}_{\upsilon,n,X}\mathcal{E}^{\otimes n}$ is equivalent to a mixture experiment, where one first picks κ according to $\mathrm{Urn}(n, m)$ and then observes $\mathcal{E}^{\otimes \kappa}$. Since deficiency is convex with respect to taking mixtures this allows one to get an upper bound on the right-hand side in terms of the expected value of $\delta(\mathcal{E}^{\otimes \kappa}, \mathcal{E}^{\otimes n})$. To get an approximately matching lower bound it suffices for those situations we shall consider to use a very weak type of concavity of the deficiency.

Proposition 10. *Fix some $n, m \in \mathbb{N}$, a measurable space X, and let $v = \operatorname{Urn}(n,m)$ and $\mathsf{S}_{v,n,X}$ be as in Proposition 8. For any experiment $\mathcal{E}$ on X one then has*

$$\max_k\bigl(\delta(\mathcal{E}^{\otimes k}, \mathcal{E}^{\otimes n})\mathbb{P}(\kappa \le k)\bigr) \le \delta(\mathsf{S}_{v,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) \le \mathbb{E}(\delta(\mathcal{E}^{\otimes \kappa}, \mathcal{E}^{\otimes n})), \tag{3}$$

where the probability on the left-hand side and the expectation on the right-hand side are taken with respect to $\kappa \sim \operatorname{Urn}(n,m)$.

As a consequence of Theorem 9 and Proposition 10 we have that $\delta(\mathsf{R}_{m,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) = 0$ only if $\delta(\mathcal{E}^{\otimes k}, \mathcal{E}^{\otimes n}) = 0$ for all k. Such experiments are degenerate in the sense that any pair of measures are either equal or singular [28, p. 285].

Consider a more general situation, where one has access to some appropriate bounds on $\delta(\mathcal{E}^{\otimes k}, \mathcal{E}^{\otimes n})$. Plugging such bounds into Proposition 10 one may attempt to balance k on the left-hand side of Proposition 10 so as to match the right-hand side up to some constant. This balancing act works at least in the following situation.

Theorem 11. *Consider a measurable space X, abbreviate $v(m,n) = \operatorname{Urn}(n,m)$, $m, n \in \mathbb{N}$, and let $\mathsf{S}_{v(m,n),n,X}$ be as in Proposition 8.*

If $\mathcal{E}$ is an experiment on X satisfying $c_0 l/n \le \delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) \le C_0 l/n$ for some constants $C_0 \ge c_0 > 0$, then

$$c_1 E_{m,n} \le \delta(\mathsf{S}_{v(m,n),n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) \le C_0 E_{n,m} \tag{4}$$

for some constant $c_1 > 0$ depending only on c_0, where $E_{n,m} = (1 - 1/n)^m$ is the expected proportion of observations missed when resampling m elements from n with replacement.

The rate $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) \asymp l/n$ is not chosen simply to give a nice result but appears naturally. Explicit examples are given in Corollary 12 below and the rate has been established by Mammen [13] and Helgeland [6] to be, at least asymptotically, the correct rate for all finite-dimensional exponential families. Based on the ideas by Le Cam [12] Mammen generalized this upper bound to experiments satisfying certain metric dimensionality conditions with respect to Hellinger distance.

Generalizing the lower bound is a less well understood problem, but we recall the following, abbreviated, partial justification due to Helgeland [6]. Assume for simplicity that the experiment $\mathcal{E}$ is identifiable. By restricting to finite subsets of the parameter space and using results concerning the convergence of experiments with finite parameter spaces one sees that the only possible limit experiment of $\mathcal{E}^{\otimes n}$ is a totally informative experiment $\mathcal{E}_\top$, where the measures are pairwise singular [27]. It is also known that $\delta(\mathcal{E}^{\otimes n}, \mathcal{E}_\top) = 1$ for all n as soon as, for example, $\mathcal{E}$ has an accumulation point for set-wise convergence or contains an uncountable dominated subexperiment [27, Proposition 5.6]. Thus one should not, in general, expect $\mathcal{E}^{\otimes n}$ to have a limit unless the parameter space is finite. Therefore, since the metric space of experiment types is complete, the sequence $\mathcal{E}^{\otimes n}$, in general, will not be a Cauchy sequence. But for $n, k > 0$ we have $\delta(\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n+k}) \le \sum_{i=1}^k \delta(\mathcal{E}^{\otimes n+i-1}, \mathcal{E}^{\otimes n+i})$, so that $\delta(\mathcal{E}^{\otimes n+i-1}, \mathcal{E}^{\otimes n+i})$ must at least decrease more slowly than $n^{-(1+\alpha)}$ for any $\alpha > 0$.

Torgersen [26, pp.1393–1394] derived a number of relatively explicit formulas for the deficiency for a number of examples. Using these we can apply Theorem 11 to two concrete examples. These examples turn out to be particularly nice in that $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n})$ can be written as a convex function of l/n.

Corollary 12. *Let $\mathcal{E}$ be one of the following two experiments:*

1. *The normal location experiment $(\mathrm{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R})$, where $\mathrm{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2;*
2. *The experiment $(\operatorname{Unif}(0,\theta); \theta \in \mathbb{R}_{>0})$, where $\operatorname{Unif}(0,\theta)$ is the uniform distribution on $[0,\theta]$.*

Letting $\mathsf{R}_{m,n,\mathbb{R}}$ be as in Proposition 6, one has $\delta(\mathsf{R}_{m,n,\mathbb{R}}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) \asymp E_{n,m}$, where $E_{n,m} = (1 - 1/n)^m$. In particular, $\delta(\mathsf{R}_{(a_i n_i), n_i, \mathbb{R}}\mathcal{E}^{\otimes n_i}, \mathcal{E}^{\otimes n_i}) \to 0$ if and only if $a_i \to \infty$.

As mentioned in the Introduction, the rate in Corollary 12 is different from the rate in the classical coupon collectors problem. In our setting the coupon collectors problem corresponds to considering a continuous/non-atomic experiment $\mathcal{E}$ satisfying $\delta(\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n+1}) = 1$. In this case one would need to have $a_i / \log(n) \to \infty$, as can be seen from the results of Erdős and Rényi [2].

4. RESAMPLING AND SUBSAMPLING

The goal of this section is to give the proof of Theorem 9, leaving some details for the appendices. The proofs of Propositions 6 and 7 are routine and will therefore be postponed until Appendix B.

While the exact form of the Markov kernel D in Proposition 8 is, for present purposes, not conceptually important for our main results, we will need it for some of the proofs leading up to Theorem 9. In this section we will therefore work with the following slightly more specific statement.

Proposition 13. *Let α_k be the uniform measure on the set of surjective functions from $\{1,\ldots,m\}$ to $\{1,\ldots,k\}$ and let $\mathsf{R}_{\alpha_k,X}\colon X^{\otimes k} \to_k X^{\otimes m}$ be as in Proposition* 6. *Define the Markov kernel $\mathsf{D}\colon \bigoplus_{k=1}^{n} X^{\otimes k} \to_k X^{\otimes m}$ such that for each $x \in \mathcal{X}^k$ one has $\mathsf{D}_x = (\mathsf{R}_{\alpha_k,X})_x$. The statement of Proposition* 8 *holds with this* D.

Given a vector of length $k \le m$, the Markov kernel D will pad it to the one of length m by copying values, generally changing the order while doing so.

The proof boils down to the basic that a function f from $\{1,\ldots,m\}$ to $\{1,\ldots,n\}$ factors as a surjective map onto $\{1,\ldots,k\}$ followed by an injective map from $\{1,\ldots,k\}$ into $\{1,\ldots,n\}$, where k is the cardinality of the image of f. The two factors D and $\mathsf{S}_{\upsilon,n,X}$ in the right-hand side of (1) correspond then to the surjective and injective parts of the function picked in the left-hand side kernel $\mathsf{R}_{m,n,X}$. Beyond this observation the proof of Proposition 13 is only bookkeeping and is therefore relegated to Appendix B.

Inequality (2) in Theorem 9 follows immediately from Proposition 13 by the general fact that applying a Markov kernel, here D, gives rise to a less informative experiment.

In general the diagonal set in $X^{\otimes 2}$ need not be measurable. Its non-measurability is, for example, inevitable in spaces of large cardinality [22, p. 550]. In practical terms this means that given two X-valued random variables the set of outcomes such that the two variables are equal is not necessarily an event. Recall from Section 3 that we get around this issue by assuming the space to be countably separating. Recall also that being countably separating implies that singleton sets are measurable and that continuous and non-atomic measures coincide.

When taking independent and identically distributed samples from a non-atomic measure on a countably separating space one will almost surely not observe any duplicate values. In particular, therefore, any duplicates when resampling must be artefacts of the resampling. If one is resampling exhaustively, so that all values are guaranteed to be picked at least once, the original collection of values can be recovered by simply throwing out any duplicates, only the original order is lost.

Lemma 14. *Let $m,n \in \mathbb{N}$, X be a countably separating measurable space and* D *be as in Proposition* 13. *There exists a Markov kernel D' such that $\mathsf{D}'\mathsf{D}P = P$ for any mixture $P = \sum_{i=1}^{r} \alpha_i P_i^{\otimes k_i}$, where P_i are continuous/non-atomic probability measures on X and $k_1,\ldots,k_r \in \{1,\ldots,n\}$.*

Intuitively the proof is simply what was said above, one may throw out any duplicates in order to recover the original vector. The details of the proof are left for Appendix B.

The usefulness of Lemma 14 stems from the fact that for experiments $\mathcal{E}$ on countably separating spaces, where all measures are non-atomic, the Markov kernel D in Proposition 13 can be cancelled. This is exactly what is needed for the second half of Theorem 9.

Before proceeding we need to note that subsampling acts on independent and identically distributed sequences simply by giving a shorter sequence of still independent and identically distributed values.

Proposition 15. *Let X and $\mathsf{S}_{\upsilon,n,X}$ be as in Proposition* 8 *and let P be a probability measure on X. Then $\mathsf{S}_{\upsilon,n,X}P^{\otimes n} = \sum_{k=1}^{n} \upsilon(\{k\})P^{\otimes k}$.*

Proof. Recall that $\mathsf{S}_{\upsilon,n,X} = \sum_{k=1}^{n} \upsilon(\{k\})\mathsf{S}_{k,n,X}$ by definition. By linearity we need therefore only show that $\mathsf{S}_{k,n,X}P^{\otimes n} = P^{\otimes k}$. But this follows from the fact that a subvector of independent and identically distributed random variables is also independent and identically distributed. □

We can now fulfil the stated goal of this section.

Proof of Theorem 9. Write $\mathsf{R}_{m,n,X} = \mathsf{DS}_{\upsilon,n,X}$ as in Proposition 13. This directly implies the inequality $\delta(\mathsf{R}_{m,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) = \delta(\mathsf{DS}_{\upsilon,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) \geq \delta(\mathsf{S}_{\upsilon,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n})$ in (2) since applying a Markov kernel gives rise to a less informative experiment (see any standard text such as Torgersen [28, Corollary 6.2.26]).

For the remaining part, let $\mathcal{E} = (X, P_\theta; \theta \in \Theta)$ with X countably separating, so that we may take D' to be as in Lemma 14.

Assume now that all P_θ are, moreover, continuous/non-atomic. Using again that applying a Markov kernel cannot result in a more informative experiment we have $\delta(\mathsf{DS}_{\upsilon,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) \leq \delta(\mathsf{D}'\mathsf{DS}_{\upsilon,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n})$. By Proposition 15 and Lemma 14 the experiment $\mathsf{D}'\mathsf{DS}_{\upsilon,n,X}\mathcal{E}^{\otimes n}$ is equal to $\mathsf{S}_{\upsilon,n,X}\mathcal{E}^{\otimes n}$, so that

$$\begin{aligned}\delta(\mathsf{S}_{\upsilon,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) &\leq \delta(\mathsf{R}_{m,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) \\ &\leq \delta(\mathsf{D}'\mathsf{DS}_{\upsilon,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) = \delta(\mathsf{S}_{\upsilon,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}),\end{aligned}$$

which means that (2) is an equality.

By Proposition 18 in Appendix A there exists an experiment $\mathcal{E}' = (Y, Q_\theta; \theta \in \Theta)$, where Y is countably separating, $\mathcal{E} \sim \mathcal{E}'$ and Q_θ is continuous/non-atomic. This concludes the proof, since $\mathcal{E}$ and $\mathcal{E}'$ are equivalent and (2) is an equality for $\mathcal{E}'$. □

Proposition 10 now follows from convexity properties of the deficiency.

Proof of Proposition 10. Since Markov kernels act point-wise, we have by Proposition 15 that $\mathsf{S}_{\upsilon,n,X}\mathcal{E}^{\otimes n} = \sum_{k=1}^{n} \upsilon(\{k\})\mathcal{E}^{\otimes k}$.

The upper bound of (3) is now a direct consequence of convexity of the deficiency with respect to mixtures of experiments [28, Corollary 6.3.22]. For the lower bound apply Proposition 20 and then maximize over K. □

5. BOUNDS IN THE PARAMETRIC CASE

The goal for this section is to prove Theorem 11 and its consequences in Corollary 12.

Getting something useful out of Proposition 10 boils down to choosing a k in the left-hand side of (3) yielding a value sufficiently close to the maximum. In this section we will perform this balancing act for experiments $\mathcal{E}$, where $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) \asymp l/n$.

In order to make clear what we are trying to achieve in the lower bound we begin by computing the upper bound.

Lemma 16. *Let $\mathcal{E}$ be an experiment such that there exists some constant C satisfying*

$$\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) \leq Cl/n$$

for all $n \in \mathbb{N}$ and $k = 0, \ldots, n$. Then $\mathbb{E}(\delta(\mathcal{E}^{\otimes \kappa}, \mathcal{E}^{\otimes n})) \leq CE_{n,m}$, where the expectation is taken with respect to $\kappa \sim \mathrm{Urn}(n, m)$ and $E_{n,m} = (1 - 1/n)^m$.

Proof. The expected value of an $\mathrm{Urn}(n, m)$-distributed random variable is $n - nE_{n,m}$, as can be seen by a short computation or using the general (factorial) moment formula [9, p. 416]. Plugging this into our expectation gives

$$\mathbb{E}(\delta(\mathcal{E}^{\otimes \kappa}, \mathcal{E}^{\otimes n})) \leq C\,\mathbb{E}(n - \kappa)/n = CnE_{n,m}/n = CE_{n,m}.$$

□

Note that the constant being independent of n and m is crucial as the conclusion holds trivially for $C = E_{n,m}^{-1}$, it always being true that $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) \leq 1 \leq nl/n$.

Given an experiment $\mathcal{E}$ such that $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) > cl/n$ for some $c > 0$ we wish now to match this bound by finding a c' such that $\max_k \delta(\mathcal{E}^{\otimes k}, \mathcal{E}^{\otimes n})\,\mathbb{P}(\kappa \leq k) \geq c' E_{n,m}$.

Depending on the size of $E_{m,n}$ we will use different strategies. When $E_{m,n}$ is very small the result is fairly direct, since one needs only consider the probability of any value at all being lost. Otherwise one either bounds the deficiency factor while keeping the probability $\mathbb{P}(\kappa \leq k)$ fixed (utilizing Lemma 21) or bounds the probability while the deficiency factor is fixed (utilizing Lemma 23).

Proof of Theorem 11. The upper bound is Lemma 16, so that only the lower bound remains. Applying Proposition 10 gives $\delta(\mathsf{S}_{v,n,X}\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) \geq \max_k \delta(\mathcal{E}^{\otimes k}, \mathcal{E}^{\otimes n})\mathbb{P}(\kappa \leq k)$, where $\kappa \sim \mathrm{Urn}(n, m)$. We need therefore bound $\max_k(1 - k/n)\,\mathbb{P}(\kappa \leq k)$ from below. The proof proceeds for three different cases corresponding to the marginal probability of missing $E_{n,m} = (1 - 1/n)^m$ being "very small", "small" and "large".

For the "very small" case, when m is much larger than n, one can directly apply the Bonferroni inequality after setting $k = n - 1$. Write $\kappa = \sum_{i=1}^n \kappa_i$, where κ_i are Bernoulli $1 - E_{m,n}$ indicator variables corresponding to the individual urns and compute

$$\begin{aligned}
\max_k \frac{n-k}{n}\mathbb{P}(\kappa \leq k) &\geq \frac{1}{n}\mathbb{P}(\kappa \neq n) = \frac{1}{n}\mathbb{P}\Big(\bigcup\nolimits_i \kappa_i = 0\Big) \\
&\geq \frac{1}{n}\Big(\sum\nolimits_i \mathbb{P}(\kappa_i = 0) - \sum\nolimits_{i<j} \mathbb{P}(\kappa_i = 0, \kappa_j = 0)\Big) \\
&= \mathbb{P}(\kappa_1 = 0)\big(1 - (n-1)\mathbb{P}(\kappa_2 = 0 \mid \kappa_1 = 0)\big) \\
&\geq \mathbb{P}(\kappa_1 = 0)(1 - E_{n,m}n) = E_{n,m}(1 - E_{n,m}n),
\end{aligned} \tag{5}$$

where the last inequality is due to the conditional distribution of $\kappa \sim \mathrm{Urn}(n, m)$ given $\kappa_1 = 0$ being $\mathrm{Urn}(n - 1, m)$.

We proceed with the "small" case where the idea is to control the first factor $1 - k/n$. Applying Lemma 23 with $\xi = E_{m,n}$ and then Lemma 24 gives

$$\max_k (1 - k/n)\mathbb{P}(\kappa \leq k) \geq E_{m,n}\mathbb{P}(\kappa \leq \mathbb{E}(\kappa)) \geq E_{m,n}\left(\frac{1}{4} + \frac{\log(E_{n,m})E_{n,m}}{1 - E_{n,m}}\right), \tag{6}$$

whenever $\log(4/3)/n \leq E_{m,n}$ and where we recall that $\mathbb{E}(\kappa) = \sum_i \mathbb{E}(\kappa_i) = n(1 - E_{n,m})$.

For the "large" case let κ' be an independent copy of κ, so that by Lemma 21 we have

$$\max_k (1 - k/n)\mathbb{P}(\kappa \leq K) \geq (1 - \mathbb{E}(\max(\kappa, \kappa'))/n)/2.$$

Recall that $E_{n,m} = n - \mathbb{E}(\kappa_1) = (1 - \mathbb{E}(\kappa)/n)$, so that proving $1 - \mathbb{E}(\max(\kappa, \kappa'))/n \geq c' E_{n,m}$ for some $c' > 0$ is equivalent to proving $\mathbb{E}(\max(\kappa, \kappa')) \leq (1 - c')n - c'\,\mathbb{E}(\kappa)$. Applying Lemma 22 yields

$$\begin{aligned}
\mathbb{E}(\max(\kappa, \kappa')) &\leq (1 - E_{n,2m})n = (1 - E_{n,m}^2)n = (1 + E_{n,m})(1 - E_{n,m})n \\
&= (1 - E_{n,m})n + E_{n,m}\,\mathbb{E}(\kappa),
\end{aligned}$$

so that

$$\max_k (1 - k/n)\mathbb{P}(\kappa \leq k) \geq E_{n,m}\frac{E_{n,m}}{2}. \tag{7}$$

It remains to combine (5), (6) and (7) to get a lower bound on $A_{m,n} := \max_k(1 - k/n)\mathbb{P}(\kappa \leq k)$ independent of m and n.

From (5) we have that $A_{m,n} \geq c'' E_{m,n}$ if $E_{n,m} \leq (1 - c'')/n$. Next (7) gives $A_{m,n} \geq c'' E_{m,n}$ if $E_{m,n} \geq 2c''$. We need therefore choose a $c'' > 0$ such that (6) gives $A_{m,n} \geq c'' E_{m,n}$ when $(1 - c'')/n < E_{n,m} < 2c''$.

First of all (6) holds only if $\log(4/3)/n \leq E_{m,n}$. But choosing $c'' \leq 1 - \log(4/3) \approx 0.7$ means the case $\log(4/3)/n \geq E_{m,n}$ is covered by Eq. (5). It remains to pick any $c'' \leq 1 - \log(4/3)$ such that

$$g(E_{m,n}) := \frac{1}{4} + \frac{\log(E_{n,m})E_{n,m}}{1 - E_{n,m}} \geq c''$$

when $(1 - c'')/n < E_{m,n} < 2c''$. Since g is monotone decreasing, we need only verify

$$g(2c'') = \frac{1}{4} + \frac{\log(2c'')c''}{1 - c''} \geq c''$$

for some $c'' > 0$. But this can be verified, for example, by $c'' = 1/20$. The result then follows by taking $c_1 = c''c_0$. □

The factor $c'' = 1/20$ is probably overly conservative. It arises due to the unsharp inequality in Lemma 24, where numerical investigation suggests that the extra term $\log(E_{n,m})E_{n,m}/(1 - E_{n,m})$ should not be necessary.

Both examples in Corollary 12 satisfy the requirements of the above theorem by way of the following simple observation.

Lemma 17. *Let $\mathcal{E}$ be an experiment such that for all $n \in \mathbb{N}$ and $0 \leq l \leq n$*

$$\delta(\mathcal{E}^{\otimes n-l}, E \otimes n) = f(l/n)$$

for some convex function $f\colon [0,1] \to [0,1]$. Then

$$cl/n \leq \delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) \leq l/n,$$

where c is any subdifferential of f at 0.

Proof. First note that $f(1) = \delta(\mathcal{E}^{\otimes 0}, \mathcal{E}^{\otimes n}) \leq 1$ and $f(0) = \delta(\mathcal{E}^{\otimes n}, \mathcal{E}^{\otimes n}) = 0$, so that by convexity $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) = f(l/n) \leq (1 - l/n)f(0) + (l/n)f(1) \leq l/n$.

Similarly for the lower bound we use convexity by recalling that a convex function always lies above its support lines $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) = f(l/n) \geq f(0) + c(l/n) = cl/n$. □

We will now make use of this observation together with Torgersen's computations to prove that both examples in Corollary 12 fall within the purview of Theorem 11.

Proof of Corollary 12. By Lemma 17 it suffices to find a convex f with appropriate subdifferentials such that $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) = f(l/n)$. Both formulas are due to Torgersen [26, pp.1393–1394]. The extra factor 1/2 in the second example stems from the different normalising constant for the total variation distance.

In Example 1 we have $\mathcal{E} = (\mathrm{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R})$ and

$$\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) = \|\,\mathrm{N}(0, 1/n) - \mathrm{N}(0, 1/(n-l)) = \|\,\mathrm{N}(0,1) - \mathrm{N}(0, 1 - l/n)\| = f(l/n).$$

For Example 2 we have $\mathcal{E} = (\mathrm{Unif}(0,\theta); \theta \in \mathbb{R}_{>0})$ and

$$\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) = \frac{1}{2}\int_0^1 |(n-l)x^{n-l-1} - nx^{n-1}|\,dx = (1 - l/n)^{n/l-1}\,l/n = g(l/n).$$

We wish to show that f and g are convex with appropriate subdifferentials at 0.

For f we may as well prove convexity of $x \mapsto f(1-x)$. Call $\mu_x = \mathrm{N}(0, 1/x)$, then

$$f(1-x) = \|\,\mathrm{N}(0,1) - \mathrm{N}(0,x)\| = \|\,\mathrm{N}(0,1/x) - \mathrm{N}(0,1)\| = \|\mu_1 - \mu_x\|.$$

By definition $\|\mu_1 - \mu_x\| = \sup_A |\mu_1(A) - \mu_x(A)|$, where the supremum ranges over all measurable sets. Computing the intersection points of the normal distribution densities one finds that the supremum is achieved on the symmetric interval $(-s(x), s(x))$, where $s(x) = \sqrt{-\log(x)/2(1-x)}$, so that

$$f(1-x) = \sup_{a>0} |\mu_x((-a,a)) - \mu_1((-a,a))| = |\mu_1((-s(x), s(x))) - \mu_x((-s(x), s(x)))|.$$

For any $a > 0$ and $x \in (0,1)$

$$|\mu_1((-a,a)) - \mu_x((-a,a))| = \mu_1((-a,a)) - \mu_x((-a,a)) = \mathrm{erf}(a) - \mathrm{erf}(a\sqrt{x}).$$

Setting $a = s(x)$ and computing the right derivative of f at 0 gives $1/\sqrt{2\pi e} \approx 0.24 > 0$. Since $f(1-x) = \sup_{a>0}(\mathrm{erf}(a) - \mathrm{erf}(a\sqrt{x}))$ convexity of f would follow from convexity of $x \mapsto \mathrm{erf}(a) - \mathrm{erf}(a\sqrt{x})$ for each fixed $a > 0$. But $\sqrt{x}$ is concave on $[0,1]$ and $-\,\mathrm{erf}$ is convex and decreasing on $[0,\infty)$, so the required convexity does hold.

Next consider $g(1-x)/2 = x^{x/(1-x)} - x^{1/(1-x)}$. Computing the right derivative gives $e^{-1} \approx 0.37 > 0$. On the analogy of the previous example we may rewrite $x^{x/(1-x)} - x^{1/(1-x)}$ as a difference of exponential terms $\exp(x\log(x)/(1-x)) - \exp(\log(x)/(1-x))$. Similarly to the normal location experiment we have

$$\exp(x\log(x)/(1-x)) - \exp(\log(x)/(1-x)) = \sup_{a\in\mathbb{R}}(\exp(xa) - \exp(a))$$

so that convexity follows by convexity of $x \mapsto \exp(xa)$ for each $a \in \mathbb{R}$. □

One cannot in general find a single function, whether convex or not, that describes the deficiency of $\mathcal{E}^{\otimes n-l}$ with respect to $\mathcal{E}^{\otimes n}$ as a function of l/n. To find a counterexample, one can use another example of Torgersen's. Take $\mathcal{E}$ to be given by the family of exponential distributions. In this case $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n})$ is not equal, for example, to $\delta(\mathcal{E}^{\otimes 2n-2l}, \mathcal{E}^{\otimes 2n})$. Numerical computations and heuristics suggest, though, that one can find a decreasing sequence of convex functions $f_1, f_2, \ldots$ such that $f_n(l/n) = \delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n})$. If this is indeed true then the same idea as in the proof of Corollary 12 can be applied.

6. CONCLUDING REMARKS

The results presented here suggest at least two questions. What upper bounds can be given for the deficiency in Theorem 9 when the experiment is not sufficiently regular to have an equality in (2) and what happens to the deficiency in Theorem 11 when $\mathcal{E}$ is not of the "parametric" type $\delta(\mathcal{E}^{\otimes n-l}, \mathcal{E}^{\otimes n}) \asymp l/n$.

For the latter there is no immediate reason for why the same idea of the proof as was used here could not be applied to some other rates. That being said, fairly little appears to be known about what rates can appear. Rates significantly faster than l/n, as with a finite parameter space, imply convergence to a totally informative limit. This case is therefore somewhat uninteresting, the deficiency tends to zero as soon as m tends to infinity at an arbitrary rate relative to n. At the other extreme one may construct experiments that are very large in the sense that $\delta(\mathcal{E}^{\otimes n-1}, \mathcal{E}^{\otimes n}) = 1$. In this case the deficiency term in (3) vanishes and the optimal K is $n-1$. One is then left with the coupon collector problem.

As for the former question recall why (2) was an equality for continuous/non-atomic experiments. The vector of values actually resampled from was reconstructed simply by throwing away any duplicates. In the general case such a reconstruction can still be attempted, though it will no longer be exact. Indeed, one cannot even be certain of the size of this unknown vector. A naive application of this idea can be realized using a fairly simple concentration argument. The resulting upper bound is of the type $\exp(-Cm/n^2)$ for some $C > 0$. While this bound holds for arbitrary experiments it does not seem sharp for any.

Finally it should be noted that while the results are general in that they cover any type of procedure based purely on resampling one should take care when interpreting their meaning. The deficiencies being small does not directly entail that any a priori chosen bootstrap-type method works, only that there always exists some procedure based on resampled values. Conversely, the deficiency being large does not mean the situation will always be hopeless. Rather it means that there exist *some* problem for which the situation is hopeless. Compare the performance, for example, on any trivial decision problem. Moreover, in general, the problem for which the difference is large may well change with the number of observations.

It may therefore be interesting to compare more systematically the performance of specific bootstrap-type procedures relative to optimal ones and see if the rates agree.

APPENDIX A: AUXILIARY RESULTS

Proposition 18. *Any experiment $\mathcal{E}$ is equivalent to a continuous experiment. Moreover if $\mathcal{E}$ is countably separating then it is equivalent to a continuous/non-atomic countably separating experiment.*

Proof. Let $\mathcal{E} = (\mathcal{X}, \mathscr{X}, P_\theta; \theta \in \Theta)$ and $\mathcal{U} = (\mathbb{R}, Q_\theta; \theta \in \Theta)$, where all Q_θ are the same $\mathrm{Unif}(0,1)$ distribution. Since $\mathcal{U}$ is non-informative, $\mathcal{E}$ is equivalent to $\mathcal{E} \otimes \mathcal{U}$. For any $(x,r) \in \mathcal{X} \times \mathbb{R}$ the event $\mathcal{X} \times \{r\}$ certainly contains (x,r) and $(P_\theta \otimes Q_\theta)(\mathcal{X} \times \{r\}) = 0$.

To see that the above construction preserves being countable separating, it suffices to show that the product of two countably separating spaces is countably separating. Both X and Y being countably separating is equivalent to the diagonals A and B in $X^{\otimes 2}$ and $Y^{\otimes 2}$ being countably separating [1]. But $(X \otimes Y)^{\otimes 2}$ is Borel-isomorphic to $X^{\otimes 2} \otimes Y^{\otimes 2}$ in a way that preserves the diagonal, where the diagonal in the latter is simply $A \times B$ and thus measurable. □

Lemma 19. *Let $\mathcal{E}$, $\mathcal{F}$ and $\mathcal{F}'$ be experiments on a common parameter space such that $\mathcal{F} \preceq \mathcal{F}'$. For any $\alpha \in [0,1]$ one has*

$$\delta(\alpha\mathcal{E} + (1-\alpha)\mathcal{F}', \mathcal{F}) = \alpha\delta(\mathcal{E}, \mathcal{F}).$$

Proof. Note that the upper bound $\delta(\beta\mathcal{E} + (1-\beta)\mathcal{F}', \mathcal{F}) \le \beta\delta(\mathcal{E}, \mathcal{F})$ follows directly by convexity.

Call the parameters Θ and write $\mathcal{E} = (P_\theta; \theta \in \Theta)$, $\mathcal{F} = (Q_\theta; \theta \in \Theta)$ and $\mathcal{F}' = (Q'_\theta; \theta \in \Theta)$. Using the randomization characterization of deficiency we compute

$$\begin{aligned}
\alpha \inf_{\mathsf{T}} \max_\theta \|\mathsf{T}P_\theta - Q_\theta\| &= \inf_{\mathsf{T}} \max_\theta \|\alpha\mathsf{T}P_\theta + (1-\alpha)Q_\theta - Q_\theta\| \\
&= \inf_{\mathsf{T},\mathsf{T}'} \max_\theta \|\alpha\mathsf{T}P_\theta + (1-\alpha)\mathsf{T}'Q'_\theta - Q_\theta\| \\
&= \inf_{\mathsf{T},\mathsf{T}'} \max_\theta \|(\mathsf{T} \oplus \mathsf{T}')(\alpha P_\theta + (1-\alpha)Q'_\theta) - Q_\theta\|, \\
&= \inf_{\mathsf{T}''} \max_\theta \|\mathsf{T}''(\alpha P_\theta + (1-\alpha)Q'_\theta) - Q_\theta\|.
\end{aligned}$$

The second equality uses the fact that $\mathcal{F} \preceq \mathcal{F}'$. □

Proposition 20. *Let $\mathcal{E}$ be an arbitrary experiment and $\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(n)}$ a finite collection of experiments, all with a common parameter space Θ. For β any probability measure on $\{1, \ldots, n\}$ one has*

$$\delta\Big(\sum_{k=1}^n \beta(\{k\})\mathcal{E}^{(k)}, \mathcal{E}\Big) \ge \beta(\{k\})\delta(\mathcal{E}^{(k)}, \mathcal{E}) \qquad \text{for } k = 1, \ldots, n.$$

Moreover, if $\mathcal{E}^{(1)} \preceq \cdots \preceq \mathcal{E}^{(n)}$ then

$$\delta\Big(\sum_{k=1}^n \beta(\{k\})\mathcal{E}^{(k)}, \mathcal{E}\Big) \ge \beta([1,K])\delta(\mathcal{E}^{(K)}, \mathcal{E}) \qquad \text{for } K = 1, \ldots, n.$$

Proof. Let $\mathcal{E}_\top$ be a fully informative experiment on Θ. The main trick for the first part of the theorem is to inflate the information of all experiments except $\mathcal{E}_k$ by substituting the fully informative experiment. Formally we have $\sum_{k=1}^n \beta(\{k\})\mathcal{E}^{(k)} \preceq \beta(\{k\})\mathcal{E}^{(k)} + (1 - \beta(\{k\}))\mathcal{E}_\top$. It remains then only to realize that $\delta(\beta(\{k\})\mathcal{E}^{(k)} + (1-\beta(\{k\}))\mathcal{E}_\top, \mathcal{E}) = \beta(\{k\})\delta(\mathcal{E}^{(k)}, \mathcal{E})$ by Lemma 19.

For the second statement consider first any K such that $0 < \beta([1,K]) < 1$ and collect the smaller terms

$$\sum_{k=1}^n \beta(\{k\})\mathcal{E}^{(k)} \sim \beta([1,K])\Big(\sum_{k=1}^K \frac{\beta(\{k\})}{\beta([1,K])}\mathcal{E}^{(k)}\Big) + \big(1 - \beta((K,n])\big)\Big(\sum_{k=K+1}^n \frac{\beta(\{k\})}{1 - \beta((K,n])}\mathcal{E}^{(k)}\Big).$$

Using the first part of this statement one finds

$$\delta\bigg(\sum_{k=1}^{n}\beta(\{k\})\mathcal{E}^{(k)},\mathcal{E}\bigg)\geq\beta([1,K])\delta\bigg(\sum_{k=1}^{K}\frac{\beta(\{k\})}{\beta([1,K])}\mathcal{E}^{(k)},\mathcal{E}\bigg).$$

Finally one applies the assumption that $\mathcal{E}^{(k)}$ is increasing, so that

$$\sum_{k=1}^{K}\frac{\beta(\{k\})}{\beta([1,K])}\mathcal{E}^{(k)}\preceq\sum_{k=1}^{K}\frac{\beta(\{k\})}{\beta([1,K])}\mathcal{E}^{(K)}\sim\mathcal{E}^{(K)}.$$

This concludes the proof since the inequality trivially holds when $\beta([1,K])\in\{0,1\}$. □

Lemma 21. *For any real-valued random variable X and bounded function $f\colon\mathbb{R}\to\mathbb{R}$ one has*

$$\sup_x f(x)\mathbb{P}(X\leq x)\geq\frac{1}{2}\,\mathbb{E}\left(f(\max(X,X'))\right),$$

where X' is an independent copy of X.

Proof. Let X' be an independent copy of X and bound the supremum by an average value, $\sup_x f(x)\mathbb{P}(X\leq x)\geq\mathbb{E}\left(f(X')\mathbb{P}(X\leq X')\right)$. By Fubini's theorem this can be rewritten as

$$\mathbb{E}\left(f(X')\mathbb{P}(X\leq X')\right)=\mathbb{E}\left(f(X')\mathbf{1}_{X\leq X'}\right),$$

where $\mathbf{1}_{X\leq X'}$ is the indicator function of the event $X\leq X'$. Now use that $X\leq X'$ if and only if $X'=\max(X,X')$ and that $\mathbf{1}_{X\leq X'}$ is independent of $\max(X,X')$ to compute

$$\mathbb{E}\left(f(X')\mathbf{1}_{X\leq X'}\right)=\mathbb{E}\left(f(\max(X,X'))\mathbf{1}_{X\leq X'}\right)=\mathbb{E}(f(\max(X,X')))\mathbb{P}(X\leq X').$$

Since X and X' are independent, $\mathbb{P}(X\leq X')=(1+\mathbb{P}(X=X'))/2\geq 1/2$ and the desired result follows. □

Lemma 22. *Let $\kappa,\kappa'\sim\mathrm{Urn}(n,m)$ be independent. Then $\mathbb{P}(\max(\kappa,\kappa')\geq k)\leq\mathbb{P}(\kappa''\geq k)$ for any $k\in\mathbb{N}$, where $\kappa''\sim\mathrm{Urn}(n,2m)$.*

Proof. Write $\kappa=|\{Z_1,\ldots,Z_m\}|$ and $\kappa'=|\{Z_1',\ldots,Z_m'\}|$, where (Z_i) and (Z_i') are pairwise independent uniform on $\{1,\ldots,n\}$. Consider the explicit coupling given by taking $\kappa''=|\{Z_1,\ldots,Z_m\}\cup\{Z_1',\ldots,Z_m'\}|$. By definition κ'' is $\mathrm{Urn}(n,2m)$-distributed such that the stochastic domination follows from

$$\{Z_1,\ldots,Z_m\},\{Z_1',\ldots,Z_m'\}\subseteq\{Z_1,\ldots,Z_m\}\cup\{Z_1',\ldots,Z_m'\}.$$

□

Lemma 23. *Let $f\colon\mathbb{R}\to\mathbb{R}$ be a left-continuous nonnegative decreasing function and X a (real-valued) random variable. For any $\xi\geq\min_x f(x)$ one has*

$$\sup_x f(x)\mathbb{P}(X\leq x)\geq\xi\mathbb{P}(f(X)\geq\xi).$$

Proof. Let $x'=\sup\{x\mid f(x)\geq\xi\}$, then

$$\sup_x f(x)\mathbb{P}(X\leq x)\geq f(x')\mathbb{P}(X\leq x')=\xi\mathbb{P}(X\leq x')=\xi\mathbb{P}(f(X)\geq\xi).$$

□

Lemma 24. *Let $\kappa\sim\mathrm{Urn}(n,m)$ with $n>1$ and m are such that $\log(4/3)/n\leq E_{n,m}$, where $E_{n,m}=(1-1/n)^m$ is the marginal probability of missing an urn. Then*

$$\mathbb{P}(\kappa\leq\mathbb{E}(\kappa))\geq\frac{1}{4}+\frac{\log(E_{n,m})E_{n,m}}{1-E_{n,m}}.$$

Proof. For brevity, let $p = 1 - E_{n,m}$ and take $X \sim \operatorname{Bin}(n, p)$. Then

$$\mathbb{P}(\kappa \leq \mathbb{E}(\kappa)) \geq \mathbb{P}(X \leq \mathbb{E}(X)) - \| \operatorname{Urn}(n, m) - \operatorname{Bin}(n, 1 - E_{n,m})\|.$$

Using [5], $\mathbb{P}(X \leq \mathbb{E}(X)) \geq \frac{1}{4}$ as long as $p \leq 1 - \frac{1}{n}$. The result is easily extended to $p \leq 1 - \frac{\log(4/3)}{n}$ as follows. Let c' be such that $p = 1 - \frac{c'}{n} > 1 - \frac{1}{n}$. Then in particular $\mathbb{E}(X) > n - 1$ and

$$\mathbb{P}(X \leq \mathbb{E}(X)) = 1 - \mathbb{P}(X = n) = 1 - (1 - c'/n)^n \geq 1 - e^{-c'}.$$

One needs now only note that $1 - e^{-c'} \geq \frac{1}{4}$ as long as $c' \geq \log(4/3)$.

For the total variation term $\| \operatorname{Urn}(n, m) - \operatorname{Bin}(n, 1 - E_{n,m})\|$ we use the Chen–Stein bound established in [23, Example 3],

$$\| \operatorname{Urn}(n, m) - \operatorname{Bin}(n, 1 - E_{n,m})\| \leq m E_{n,m} \frac{1 - (1 - E_{n,m})^{n+1}}{(n + 1)(1 - E_{n,m})}.$$

Rewriting $m = \frac{\log(E_{n,m})}{\log(n-1)-\log(n)}$ one has

$$m E_{n,m} \frac{1 - (1 - E_{n,m})^{n+1}}{(n + 1)(1 - E_{n,m})} = \frac{-\log(E_{n,m}) E_{n,m}}{1 - E_{n,m}} \frac{1 - (1 - E_{n,m})^{n+1}}{(n + 1)(\log(n) - \log(n - 1))},$$

so that it remains only to realize that for any $c \in (0, 1)$

$$\frac{1 - c^{n+1}}{(n + 1)(\log(n) - \log(n - 1))}$$

is increasing in n, tending to 1. □

APPENDIX B: ADDITIONAL PROOFS

Proof of Proposition 6. Assume first that α is finitely supported, which especially covers the simple case with fixed sample size, since $|F_{m,n}| = n^m < \infty$.

Let $f \colon \{1, \ldots, m\} \to \{1, \ldots, n\}$. For any measurable space X the projection map $(x_1, \ldots, x_n) \mapsto (x_{f(i)})$ from $X^{\otimes n}$ to X is measurable meaning also that $(x_1, \ldots, x_n) \mapsto (x_{f(1)}, \ldots, x_{f(m)})$ is measurable. Denoting $x_f = (x_{f(1)}, \ldots, x_{f(m)})$ for $x = (x_1, \ldots, x_n)$, this means that $x \mapsto \delta_{x_f}$ defines a Markov kernel. Since $\mathsf{R}_{\alpha,X}$ in the statement is simply a convex combination of a finite number of these it must also be a Markov kernel.

If α is not finitely supported, then $(\mathsf{R}_{\alpha,X})_x$ is still a probability measure for every $x \in \mathcal{X}$, so that we need only argue for measurability. But $\bigcup_m F_{m,n}$ is countably infinite, so that by restricting α to, say, $\bigcup_{m=1}^N F_{m,n}$ and letting N tend to infinity, one finds $x \mapsto \mathsf{R}_{\alpha,X}(A)$ for any measurable set A as a monotone increasing sequence of measurable functions. □

Proof of Proposition 7. Let $(X_1, \ldots, X_m) = (x_{f(1)}, \ldots, x_{f(m)})$, where f is uniformly chosen in $F_{m,n}$. By definition $(X_1, \ldots, X_m) \sim (\mathsf{R}_{m,n,X})_x$. We wish to prove that $\{X_1, \ldots, X_m\}$ are independent with marginal law $P = \sum_{i=1}^n n^{-1} \delta_{x_i}$. Let S_n and S_m be the groups of permutations on $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$, respectively. For any $\sigma \in S_n$ we have $\sigma \circ F_{m,n} := \{\sigma \circ g \mid g \in F_{m,n}\} = F_{m,n}$ and similarly for any $\pi \in S_m$ it holds that $F_{m,n} \circ \pi = F_{m,n}$. Since f is uniform on $F_{m,n}$, this means that $\sigma \circ f$ and $f \circ \pi$ are both equal in law to f. The latter says exactly that $(X_1, \ldots, X_m)$ is exchangeable. For any $1 \leq i \leq m$ and $1 \leq j, j' \leq n$ let τ be the transposition such that $\tau(j) = j'$. The former invariance then implies $\mathbb{P}(X_i = x_j) = \mathbb{P}(x_{\tau \circ f(i)} = x_{\tau(j)}) = \mathbb{P}(x_{f(i)} = x_{j'})$. This means that $\mathbb{P}(X_i = x_j) = |\{j' \mid x_{j'} = x_j\}|/n$, which is exactly the uniform distribution in the statement.

For $(\mathsf{R}_{m,n,X})_x$ it remains therefore only to show independence. We do this by proving that $(X_1, \ldots, X_{m-1}, X_m)$ is equal in law to $(X_1, \ldots, X_{m-1}, X'_m)$, where X'_m is an independent copy of X_m. Using exchangeability one can then inductively replace each X_i by an independent copy. The above can be realized by taking $(X_1', X_2, \ldots, X_m) = (x_{f'(1)}, x_{f'(2)}, \ldots, x_{f'(m)})$, where $f'|_{\{1,\ldots,m-1\}} = f|_{\{1,\ldots,m-1\}}$ and $f'(m)$ is independent of f and uniform on $\{1, \ldots, n\}$. For any $f_0 \in F_{m-1,n}$, we compute

$$\begin{aligned}\mathbb{P}(f|_{\{1,\ldots,m-1\}} = f_0) &= \mathbb{P}(f(1) = f_0(1), \ldots, f(m-1) = f_0(m-1)) \\ &= \sum_{i=1}^{n} \mathbb{P}(f(1) = f_0(1), \ldots, f(m-1) = f_0(m-1), f(m) = i) \\ &= nn^{-m} = n^{-(m-1)},\end{aligned}$$

so that $f|_{\{1,\ldots,m-1\}}$ is uniform on $F_{m-1,n}$. For any $f_1 \in F_{m,n}$ it then follows that

$$\begin{aligned}\mathbb{P}(f' = f_1) &= \mathbb{P}\big(f'|_{\{1,\ldots,m-1\}} = f_1|_{\{1,\ldots,m-1\}}, f'(m) = f_1(m)\big) \\ &= \mathbb{P}\big(f|_{\{1,\ldots,m-1\}} = f_1|_{\{1,\ldots,m-1\}}, f'(m) = f_1(m)\big) \\ &= \mathbb{P}(f|_{\{1,\ldots,m-1\}} = f_1|_{\{1,\ldots,m-1\}})\mathbb{P}(f'(m) = f_1(m)) \\ &= n^{-(m-1)}n^{-1} = n^{-m}.\end{aligned}$$

Consider now the case $m < n$ and let $(Y_1, \ldots, Y_m) = (x_{f(1)}, \ldots, x_{f(m)})$, where f is chosen uniformly among injective functions in $F_{m,n}$. By definition $(Y_1, \ldots, Y_m)$ has law $(\mathsf{S}_{m,n,X})_x$. Since the set of injective functions is preserved under pre- and post-composition by permutations, we have exactly as in the previous case that $(Y_1, \ldots, Y_m)$ is exchangeable with the correct marginal laws. We need therefore only show that the conditional law of $f|_{\{1,\ldots,m-1\}}$ on the event $f(m) = n$ is uniform on injective functions in $F_{m-1,n-1}$. But this follows directly from the fact that the restriction of a uniform measure to a subset is still uniform. □

Proof of Proposition 13. Our first step is to reduce the problem to the one not involving the arbitrary measurable space X. Rather we will directly deal with random functions from $F_{m,n}$, where we recall that $F_{m,n}$ denotes the set of functions from $\{1, \ldots, m\}$ to $\{1, \ldots, n\}$.

Fix some $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and let $\kappa \sim \mathrm{Urn}(n, m)$. Conditionally on $\kappa = k$ let f be uniform on injective functions in $F_{k,n}$. Then $x_f = (x_{f(1)}, \ldots, x_{f(\kappa)})$ by definition has law $(S_{\upsilon,n,X})_x$, where we recall the shorthand $\upsilon = \mathrm{Urn}(n, m)$. Independently of f but still conditionally on $\kappa = k$ let g be uniform among surjective functions in $F_{m,k}$. Then $(x_f)_g = (x_{f(g(1))}, \ldots, x_{f(g(m))}) = x_{f\circ g}$ has law $(\mathsf{DS}_{\upsilon,n,X})_x$.

Let h be uniform on $F_{m,n}$. Then $x_h = (x_{h(1)}, \ldots, x_{h(m)})$ has law $(R_{m,n,X})_x$. To prove the statement it suffices therefore to show that h is equal in law to $f \circ g$. To see this we will construct κ', f' and g' such that $h = f' \circ g'$ and (κ', f', g') equal in law to (κ, f, g).

Let $A = \{f(g(1)), \ldots, f(g(m))\}$ be the image of $f \circ g$ and $A' = \{h(1), \ldots, h(m)\}$ be the image of h. Since both the law of h and the law of $f \circ g$ are invariant under post-composition by permutations of $\{1, \ldots, n\}$ the probabilities $\mathbb{P}(A = A_0)$ and $\mathbb{P}(A' = A_0)$ for $A_0 \subset \{1, \ldots, n\}$ depend only on the cardinality of A_0. The law of $\kappa = |A|$ is $\mathrm{Urn}(n, m)$ by construction, while the law of $\kappa' := |A'|$ may be taken as the definition of $\mathrm{Urn}(n, m)$.

We proceed now conditionally on $A' = A_0$ for some non-empty $A_0 \subset \{1, \ldots, n\}$ such that, in particular, $\kappa' = |A_0|$. Since h is uniform on $F_{m,n}$, the conditional law is the uniform measure on the set of surjective functions from $\{1, \ldots, m\}$ onto A_0. Let f' be conditionally independent of everything else with conditional law the uniform measure on the set of injective functions from $\{1, \ldots, \kappa'\}$ to $\{1, \ldots, n\}$ with image A_0.

Let π be a random permutation on $\{1, \ldots, n\}$ such that $\pi(x) = y$ when $f'(y) = x$ and defined arbitrarily when $x \notin A_0$. Define $g' = \pi \circ h$ and note that it does not depend on the arbitrary extension used in the definition of π. For each fixed realization of π the map $h' \mapsto \pi \circ h'$ defines a bijection between surjective functions from $\{1, \ldots, m\}$ onto A_0 and surjective functions from $\{1, \ldots, m\}$ onto $\pi(A_0) = \{1, \ldots, \kappa'\}$. Since π and h are conditionally independent, it then follows that the conditional law of g' is the uniform distribution on surjective functions in $F_{m,\kappa'}$.

It remains now to establish that f' and g' have the correct marginal laws and are independent also conditionally on $\kappa' = k_0$, rather than on $A' = A_0$.

Consider any $1 \leq k_0 \leq n$ and $g_0, g_1 \in F_{m,k_0}$. Using the above we have

$$\begin{aligned}\mathbb{P}(g' = g_0 \mid \kappa' = k_0) &= \sum_{|A_0|=k_0} \mathbb{P}(g' = g_1 \mid A' = A_0)\mathbb{P}(A' = A_0 \mid \kappa' = k_0)\\ &= \sum_{|A_0|=k_0} \mathbb{P}(g' = g_0 \mid A' = A_0)\mathbb{P}(A' = A_0 \mid \kappa' = k_0)\\ &= \mathbb{P}(g' = g_1 \mid \kappa' = k_0),\end{aligned}$$

so that g' is uniform on F_{m,k_0} also conditionally on $\kappa' = k_0$.

Similarly consider any pair of injective functions $f_0, f_1 \in F_{\kappa',n}$ with images $A_0 = \{f_0(1), \dots, f_0(\kappa')\}$ and $A_1 = \{f_1(1), \dots, f_1(\kappa')\}$. Then

$$\begin{aligned}\mathbb{P}(f' = f_0 \mid \kappa' = k_0) &= \mathbb{P}(f' = f_0, A' = A_0 \mid \kappa' = k_0)\\ &= \mathbb{P}(f' = f_0 \mid A' = A_0)\mathbb{P}(A' = A_0 \mid \kappa' = k_0)\\ &= \mathbb{P}(f' = f_1 \mid A' = A_1)\mathbb{P}(A' = A_1 \mid \kappa' = k_0)\\ &= \mathbb{P}(f' = f_1 \mid \kappa' = k_0),\end{aligned}$$

so that f' must be uniform on the set of injective functions in $F_{k_0,n}$ conditionally on $\kappa' = k_0$.

If we can prove that f' and g' are independent conditionally on $\kappa' = k_0$, we are done. We have

$$\begin{aligned}&\mathbb{P}(f' = f_0, g' = g_0 \mid \kappa' = k_0)\\ &\quad= \mathbb{P}(f' = f_0, g' = g_0 \mid A' = A_0)\mathbb{P}(A' = A_0 \mid \kappa' = k_0)\\ &\quad= \mathbb{P}(f' = f_0, h = f' \circ g_0 \mid A' = A_0)\mathbb{P}(A' = A_0 \mid \kappa' = k_0)\\ &\quad= \mathbb{P}(h = f' \circ g_0 \mid f' = f_0)\mathbb{P}(f' = f_0 \mid A' = A_0)\mathbb{P}(A' = A_0 \mid \kappa' = k_0)\\ &\quad= \mathbb{P}(h = f' \circ g_0 \mid f' = f_0)\mathbb{P}(f' = f_0 \mid \kappa' = k_0).\end{aligned}$$

Finally,

$$\begin{aligned}\mathbb{P}(h = f' \circ g_0 \mid f' = f_0) &= \mathbb{P}(h = f_0 \circ g_0 \mid f' = f_0)\\ &= \mathbb{P}(h = f_0 \circ g_0 \mid A' = A_0) = \mathbb{P}(h = f' \circ g_0 \mid A' = A_0)\\ &= \sum_{|A_1|=k_0} \mathbb{P}(h = f' \circ g_0 \mid A' = A_0)\mathbb{P}(A' = A_1 \mid \kappa' = k_0)\\ &= \sum_{|A_1|=k_0} \mathbb{P}(h = f' \circ g_0 \mid A' = A_1)\mathbb{P}(A' = A_1 \mid \kappa' = k_0)\\ &= \mathbb{P}(h = f' \circ g_0 \mid \kappa' = k_0),\end{aligned}$$

where the third equality follows from the uniform law of h and conditional independence of h and f'. □

Proof of Lemma 14. We wish to pick D' by sending any vector of values to a subvector consisting of all distinct values.

Specify $X = (\mathcal{X}, \mathscr{X})$ and define a function $f \colon \mathcal{X}^m \to \bigcup_k \mathcal{X}^k$ as follows. For any $x = (x_1, \dots, x_m) \in \mathcal{X}^m$ let $k(x) = |\{x_1, \dots, x_m\}|$ be the number of distinct components and define $i_1(x) < \cdots < i_{k(x)}(x)$ as follows. Let $i_1(x) = 1$ and then recursively let $i_{r+1}(x)$ be the smallest value such that $x_{i_{r+1}} \notin \{x_1, \dots, x_{i_r}\}$. For later convenience let $i_{k(x)+r+1}(x) = i_{k(x)+r} + 2$. By assumption

$$c_{i,j}(x) = \begin{cases}1, & x_i \neq x_j,\\ 0, & \text{otherwise},\end{cases}$$

is measurable for $1 \leq i, j \leq m$. Consequently also

$$n_i(x) = \prod_{j=1}^{i-1} c_{i,j}(x) = \begin{cases}1, & x_i \notin \{x_1, \dots, x_{i-1}\},\\ 0, & \text{otherwise},\end{cases}$$

is measurable for $1 \leq i \leq m$. Since $k(x) = \sum_{i=1}^{n} n_i(x)$ and $i_r(x) = r + \sum_{i=1}^{r}(1 - n_i(x))$, they are also measurable. Any projection $\pi_r(x) = x_r$ is measurable. For any measurable $\iota\colon X^{\otimes m} \to \{1, \ldots, n\}$ the map $\pi_\iota(x) = x_{\iota(x)}$ is measurable, since for any $A \in \mathscr{X}$ one has $\pi_\iota^{-1}(A) = \bigcup_{i=1}^{n} \iota^{-1}(i) \cap \pi_i^{-1}(A)$. This means that the function $f'(x) = (x_{i_1(x)}, \ldots, x_{i_n(x)})$ is measurable. Using the same argument, mutatis mutandis, the map $f(x) = (x_{i_1(x)}, \ldots, x_{i_{k(x)}(x)})$ is measurable. By construction $f(x)$ sends x to a subvector consisting of all distinct values in order of appearance.

Let $\mathsf{D}'_x = \delta_{f(x)}$ be the deterministic Markov kernel induced by f. For any x consisting of distinct values, $(\mathsf{D}'\mathsf{D})_x = \sum_{\sigma \in S_m} \beta(\{\sigma\})\delta_{x_\sigma}$ is the law of a random permutation $x_\sigma = (x_{\sigma(1)}, \ldots, x_{\sigma(m)})$ of x, where β is some probability measure on the symmetric group S_m on $\{1, \ldots, m\}$.

If a probability measure P on X is non-atomic, we claim that $P^{\otimes k}(c_{i,j}) = 0$ for all $1 \leq i < j \leq k$ from which it follows that the probability of all values being distinct is 1. It suffices to show this for the case $k = 2$, $i = 1$ and $j = 2$.

Take any $\varepsilon > 0$. By assumption for each $x \in \mathcal{X}$ there exists a set $A_x \in \mathscr{X}$ such that $x \in A_x$ and $P(A_x) < \varepsilon$. By Fubini's theorem $P^{\otimes 2}(c_{1,2}) = \iint c_{1,2}(x_1, x_2)\, P(dx_1)P(dx_2) \leq \int P(A_{x_2})\, P(dx_2) \leq \varepsilon$. Since ε is arbitrary, we have $P^{\otimes 2}(c_{1,2}) = 0$.

By the above two facts, $\mathsf{D}'\mathsf{D}P^{\otimes k} = \sum_{\sigma \in S_m} \beta(\{\sigma\})P^{\otimes k} = P^{\otimes k}$. The result now follows by linearity. □

ACKNOWLEDGMENTS

I would like to thank my supervisor Silvelyn Zwanzig for suggesting and getting me interested in this question. She as well as Rolf Larsson are also to be thanked for helpful comments on an earlier version of the manuscript.

REFERENCES

1. W. Adamski, "On the Relations between Continuous and Nonatomic Measures", Math. Nachr. **99**, 55–60 (1980).
2. P. Erdős and A. Rényi, "On a Classical Problem of Probability Theory", Magyar Tud. Akad. Mat. Kutató Int. Közl. **6**, 215–220 (1961).
3. M. Falk and F. Marohn, "On the Loss of Information due to Nonrandom Truncation", J. Multivar. Anal. **72**, 1–21 (2000).
4. V. Genon-Catalot and C. Larédo, "Asymptotic Equivalence of Nonparametric Diffusion and Euler Scheme Experiments", Ann. Statist. **42**, 1145–1165 (2014).
5. S. Greenberg and M. Mohri, "Tight Lower Bound on the Probability of a Binomial Exceeding its Expectation", Statist. Probab. Lett. **86**, 91–98 (2014).
6. J. Helgeland, "Additional Observations and Statistical Information in the Case of 1-Parameter Exponential Distributions", Z. Wahrsch. und Verw. Gebiete **59**, 77–100 (1982).
7. A. Janssen and R.-D. Reiss, "Comparison of Location Models of Weibull Type Samples and Extreme Value Processes", Probab. Theory Rel. Fields **78**, 273–292 (1988).
8. A. Janssen and F. Marohn, "On Statistical Information of Extreme Order Statistics, Local Extreme Value Alternatives, and Poisson Point Processes", J. Multivar. Anal. **48**, 1–30 (1994).
9. N. L. Johnson, A. W. Kemp, and S. Kotz, *Univariate Discrete Distributions*, 2nd. ed. (Wiley, New York, 2005).
10. V. Konakov, E. Mammen, and J. Woerner, "Statistical Convergence of Markov Experiments to Diffusion Limits", Bernoulli **20**, 623–644 (2014).
11. L. Le Cam, "Sufficiency and Approximate Sufficiency", Ann. Math. Statist. **35**, 1419–1455 (1964).
12. L. Le Cam, "On the Information Contained in Additional Observations", Ann. Statist. **2**, 630–649 (1974).
13. E. Mammen, "The Statistical Information Contained in Additional Observations", Ann. Statist. **14**, 665–678 (1986).
14. E. Mammen, *When Does Bootstrap Work? Asymptotic Results and Simulations* (Springer, New York, 1992).
15. E. Mariucci, "Asymptotic Equivalence of Discretely Observed Diffusion Processes and Their Euler Scheme: Small Variance Case", Statist. Inference Stoch. Process. **19**, 71–91 (2016).
16. F. Marohn, "Global Sufficiency of Extreme Order Statistics in Location Models of Weibull Type", Probab. Theory Rel. Fields **88**, 261–268 (1991).

17. F. Marohn, "Neglecting Observations in Gaussian Sequences of Statistical Experiments", Statist. Decisions **13**, 83–92 (1995).
18. G. Milstein and M. Nussbaum, "Diffusion Approximation for Nonparametric Autoregression", Probab. Theory Rel. Fields **112**, 535–543 (1998).
19. H. Putter and W. R. van Zwet, "Resampling: Consistency of Substitution Estimators", Ann. Statist. **24**, 2297–2318 (1996).
20. R.-D. Reiss, "A New Proof of the Approximate Sufficiency of Sparse Order Statistics", Statist. Probab. Lett. **4**, 233–235 (1986).
21. R.-D. Reiss, M. Falk, and M. Weller, "Inequalities for the Relative Sufficiency between Sets of Order Statistics", in *Statistical Extremes and Applications* (Springer, Dordrecht, 1984), pp. 597–610.
22. E. Schechter, *Handbook of Analysis and Its Foundations* (Academic Press, San Diego, 1997).
23. S. Y. T. Soon, "Binomial Approximation for Dependent Indicators", Statist. Sinica **6**, 703–714 (1996).
24. H. Strasser, "Towards a Statistical Theory of Optimal Quantization", in *Data Analysis: Scientific Modeling and Practical Application* (Springer, Berlin–Heidelberg, 2000), pp. 369–383.
25. E. N. Torgersen, "Comparison of Experiments when the Parameter Space is Finite", Z. Wahrsch. und Verw. Gebiete **16**, 219–249 (1970).
26. E. N. Torgersen, "Comparison of Translation Experiments", Ann. Math. Statist. **43**, 1383–1399 (1972).
27. E. N. Torgersen, "Measures of Information Based on Comparison with Total Information and with Total Ignorance", Ann. Statist. **9**, 638–657 (1981).
28. E. N. Torgersen, *Comparison of Statistical Experiments*, in *Encyclopedia of Mathematics and its Applications* (Cambridge Univ. Press, Cambridge, 1991).