Electron. Commun. Probab. **18** (2013), no. 35, 1–6.
DOI: 10.1214/ECP.v18-2045
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

The BK inequality for pivotal sampling a.k.a. the Srinivasan sampling process

Johan Jonasson[*]

Abstract

The pivotal sampling algorithm, a.k.a. the Srinivasan sampling process, is a simply described recursive algorithm for sampling from a finite population a fixed number of items such that each item is included in the sample with a prescribed desired inclusion probability. The algorithm has attracted quite some interest in recent years due to the fact that despite its simplicity, it has been shown to satisfy strong properties of negative dependence, e.g. conditional negative association. In this paper it is shown that (tree-ordered) pivotal/Srinivasan sampling also satisfies the BK inequality. This is done via a mapping from increasing sets of samples to sets of match sequences and an application of the van den Berg-Kesten-Reimer inequality. The result is one of only very few non-trivial situations where the BK inequality is known to hold.

Keywords: Srinivasan sampling ; negative association ; Reimer's inequality ; LaTeX.
AMS MSC 2010: 60C05; 60K35.
Submitted to ECP on May 23, 2012, final version accepted on March 18, 2013.

1 Introduction

Let n be a positive integer and let $S = \{0,1\}^n$ with the usual coordinatewise partial order. For $\omega \in S$ and $K \subseteq [n]$, let $\omega_K = (\omega_k)_{k \in K}$. Define the subset $[\omega]_K$ as

$$[\omega]_K = \{\alpha \in S : \alpha_K = \omega_K\}.$$

The operation $\Box$ on pairs of subsets of S is given by

$$A \Box B = \{\omega \in S : \exists K, L \subset [n] : K \cap L = \emptyset,\ [\omega]_K \subseteq A, [\omega]_L \subseteq B\}.$$

Loosely speaking, $A \Box B$ is the set of ω's for which A and B occur disjointly. When $[\omega]_K \subseteq A$ we will sometimes say that ω_K *guarantees* A. Note that if A and B depend on disjoint sets of indices, then $A \Box B = A \cap B$. A subset A of S is said to be *increasing* if for all $\alpha, \omega \in S$, we have $\alpha \in A$, $\alpha \le \omega \Rightarrow \omega \in A$.

Let $X = (X_1, \dots, X_n)$ be a family of binary random variables and let $\mu(\cdot) = \mathbb{P}(X \in \cdot)$ be its law. We say that X (or μ) is BK, or that X (or μ) satisfies the BK inequality, if for every pair of increasing events, A and B,

$$\mathbb{P}(X \in A \Box B) \le \mathbb{P}(X \in A)\mathbb{P}(X \in B). \tag{1.1}$$

Recall also that X is said to be *negatively associated* (NA) whenever (1.1) holds for all A and B which depend on disjoint sets of indices (i.e. whenever $\mathbb{P}(X \in A \cap B) \le \mathbb{P}(X \in$

[*] Chalmers University of Technology and University of Gothenburg, Sweden.
E-mail: jonasson@chalmers.se

$A)\mathbb{P}(X \in B)$ for all such A and B). Hence BK is trivially a property which is at least as strong as NA.

The BK inequality is known to hold when the X_k's are independent; this is the classical BK inequality of van den Berg and Kesten [3], a result which has turned out to be of fundamental importance in e.g. percolation theory and reliability theory, see e.g. [10]. In fact, when the X_k's are independent, (1.1) holds for *all* sets A and B. This was a long standing open problem until finally solved by David Reimer [13] (2000). Consequently, this fact is now known as the *van den Berg-Kesten-Reimer inequality*, a result that will be of fundamental importance in this paper.

Clearly, if a family has the BK property, this means that it is negatively dependent in some sense. For example, as noted above, any BK family is NA. In recent years, it has become a challenge to understand how the BK property fits into the theory of negative dependence. The chase for a theory of negative dependence started out a decade or so ago with the pioneering papers [12] and [9]. A major step forward was taken by Borcea, Brändén and Liggett in [4]. Their work was based on an algebraic/analytic approach involving the zeros of the generating polynomials. This approach in turn was based on a series of papers of Borcea and Brändén, see the bibliography of [4]. The generating polynomial approach is powerful, see e.g. [5], where a number of important sampling techniques were easily shown to satisfy a strong form of negative dependence, the *strong Rayleigh property*, This property implies in particular CNA, i.e. that the conditional distribution, given any subset of the variables, of the remaining variables is NA. However, the BK property has so far resisted the analytic approach and it is unclear how it would fit into this framework. Markström [11] gave examples that showed that the BK property is neither closed under conditioning, nor under external fields. He also showed that there are examples of NA families which are not BK. If CNA is sufficient for BK, remains an open question.

A moments thought reveals that the van den Berg-Kesten-Reimer inequality can only be satisfied for product measures. However it is intuitively clear that (1.1) should hold for all increasing events for many classes of negatively dependent binary random variables. Until quite recently however, the BK inequality was not known to hold for any substantial classes of measures apart from product measures, despite the efforts of several researchers (oral communication). The first substantial new contribution came in [2], where uniform samples of, say, k items from a population of size n, were shown to be BK. This was also shown to be true for weighted versions of uniform k-out-of-n samples and products of such measures. The still more recent work [1] proves that the anti-ferromagnetic Ising Curie-Weiss model satisfies the BK inequality. It is also shown that if (1.1) is modified in a natural way, then it holds also for the ferromagnetic Ising model.

In this paper, we show that the important pivotal sampling procedure, also known as Srinivasan sampling, satisfies the BK inequality. As in [2], this is done via an application of Reimer's results. A difference however, is that here we apply the van den Berg-Kesten-Reimer inequality directly, whereas [2] refers to the key ingredient of Reimer's proof, namely the set-theoretic fact known as Reimer's Butterfly Theorem.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the sampling process and state the main result. Section 3 is then devoted to the proof.

2 Notation and statements

Pivotal sampling is an important algorithm in sampling theory. It was introduced by Deville and Tillé [7] in 1998. In the computer science community, which was generally not aware of the Deville-Tillé paper, the method was independently rediscovered and introduced by Srinivasan [14] and is consequently known there as Srinivasan sampling. The pivotal/Srinivasan algorithm is an efficient method for picking fixed size samples with the exact right pre-specified inclusion probabilities, that despite its simplicity enjoys all the virtues of negative association. One drawback is that the entropy of the resulting sample is fairly low. For example, in a population of n items, there will typically be $n-1$ pairs of items such that the two items in a given pair either cannot both be included in the sample or cannot both be outside the sample.

The algorithm is recursive and works as follows. Suppose that we have a population of n items from which we want to draw a sample of exactly k items, in such way that for each item i, the probability that this item is included in the sample is exactly a pre-specified number π_i. (To make this possible, we of course need that $\sum_i \pi_i = k$.) Order the items linearly as item 1, item 2, ..., item n in some arbitrary way. Suppose that $\pi_1 + \pi_2 \leq 1$. Play a "match" between items 1 and 2, with 1 as the winner with probability $\pi_1/(\pi_1+\pi_2)$ and 2 as the winner with the remaining probability $\pi_2/(\pi_1+\pi_2)$. The loser is now ruled out from being included in the sample (i.e. one sets $X_1 = 0$ if item 1 lost the match and $X_2 = 0$ otherwise), whereas the winner gets the new inclusion probability $\pi'_2 = \pi_1 + \pi_2$ and is relabelled as item $2'$ in a new population consisting of $2'$ together with $3, \ldots, n$. Now apply the algorithm recursively and independently from the result of the first match, on the new population with inclusion probabilities $\pi'_2, \pi_3 \ldots, \pi_n$. In case $\pi_1 + \pi_2 > 1$, declare instead 1 the winner with probability $(1-\pi_2)/(2-\pi_1-\pi_2)$ and declare 2 the winner with the remaining probability $(1-\pi_1)/(1-\pi_1-/\pi_2)$. The winner is now given a secure place in the sample (i.e. one sets $X_1 = 1$ if item 1 won the match and $X_2 = 1$ otherwise), whereas the loser plays on, as above under the new identity $2'$, in the new population $2', 3, \ldots, n$, with inclusion probabilities $\pi'_2 = \pi_1 + \pi_2 - 1$ for $2'$ and π_i for $i = 3, \ldots, n$.

Note that the i'th match of the process is always played between the item initially labelled $i+1$ and one item with a lower initial label, whose identity is determined by the results of the previous $i-1$ matches and that the probabilities with which these two items compete is non-random. In particular the final sample is determined by the $n-1$ *independent* matches. Note also that the process can equivalently be described with the matches in reverse order. Indeed, with the process as described above, we can equally well pick the sample on $2', 3, \ldots, n$ first and then use the first match (which now becomes the last match) to decide the true identity of $2'$.

That this indeed produces a sample of exactly k items and with the desired inclusion probabilities follows from a short induction argument. Indeed, by induction it follows that it suffices to note that the inclusion probability of item 1 is π_1 under the induction hypothesis that the algorithm works as claimed for populations of size $n-1$. This however, is obvious.

One variant of the pivotal sampling algorithm, which may raise the entropy, is to replace the linear order of the items with a tree order. I.e. place the items at the leafs of a binary tree (i.e. a tree where all vertices have degree 1 or 3) with n leafs, in some deterministic way. Then play the first match between two predetermined vertices at two leafs with a common neighbor. Place the winner or the loser, depending on the total probability of the given match, at the common neighbor and erase the two leafs. Then repeat recursively as above. (Another variant, which stands out naturally, is to order the items linearly in a uniform random way. Then, of course, our results apply given the order, but unfortunately they do not apply to the whole procedure including the randomness in order.)

Pivotal sampling/Srinivasan's process was shown to be CNA under linear order in [8]. This was extended to the tree-ordered case in [6]. These results were further strengthened in [5], where it was shown that pivotal sampling is in fact strongly

Rayleigh. Here we prove that it is also BK:

Theorem 2.1. *Let $X = (X_1, \dots, X_n)$ be the indicator random variables of a pivotal sample on n items, either linearly ordered or tree-ordered. I.e. let $\pi_1, \dots, \pi_n$ be the given inclusion probabilities (satisfying $\sum_i = k$, $k \in [n]$) and let $X_i = 1$ if item i gets included in the sample and $X_i = 0$ otherwise. Then X is BK.*

3 Proof of Theorem 2.1

The sample X is determined by $n-1$ matches. Let the i'th match be denoted by m_i and let $M = \{m_1, \dots, m_{n-1}\}$ be the set of matches. Let $Y = (Y_1, \dots, Y_{n-1})$ be the binary random variables given by $Y_i = 1$ if match number i is won by the item with the smallest label and $Y_i = 0$ otherwise. Let $f : \{0,1\}^M \to S_k = \{\omega \in S : \sum_i \omega_i = k\}$ be the function given by letting $f(y)$ be the pivotal sample that results from $Y = y$. In particular $X = f(Y)$. (Note that f is neither injective nor surjective. E.g. if $\pi_1 + \pi_2 > 1$, then $f(y)_1 + f(y)_2 \geq 1$ for all y and it is easily seen that $f(y)_1 = f(y)_2 = 1$ for at least two different y's). For an event $A \subseteq S$, let $\widehat{A} := f^{-1}(A) = \{y \in \{0,1\}^M : f(y) \in A\}$. The key result is the following lemma.

Lemma 3.1. *Let A and B be two increasing subsets of $\{0,1\}^n$. Then*

$$\widehat{A \Box B} \subseteq \widehat{A} \Box \widehat{B}.$$

Proof. We will do this by induction over n. It is trivial to check this for $n = 1$ (which makes perfect sense), so we can focus on the induction step. Fix an integer $r \geq 2$ and assume that the lemma holds for $n = 1, \dots, r-1$ and consider the case $n = r$. Pick an arbitrary $y \in \widehat{A \Box B}$. Let $x := f(y)$. By definition we have $x \in A \Box B$. Hence there are two disjoint index sets $I, J \subseteq [n]$ such that $[x]_I \subseteq A$, $[x]_J \subseteq B$ and $x_I \equiv x_J \equiv 1$. We want to show that $y \in \widehat{A} \Box \widehat{B}$. The crucial step is to use that if we write $y = (y_1, u)$, $u \in \{0,1\}^{n-2}$, then the match sequence u gives, by the very definition of pivotal sampling given above, rise to a pivotal sample x' on the items $2', 3, \dots, n$ which agrees with x on $3, \dots, n$, but may possibly differ on items 1 and 2. Since this latter sample is from a population of $n-1$ items, the induction hypothesis applies.

Consider first the case $\pi_1 + \pi_2 < 1$. Then, since at least one of x_1 and x_2 is 0, we cannot have both items 1 and 2 in $I \cup J$. Let $I' = I$ if $1, 2 \notin I$, and $I' = I \setminus \{i\} \cup \{2'\}$ if $i \in I$, $i \in \{1, 2\}$. Define J' identically. Then I' and J' are disjoint. Hence by the induction hypothesis, there are disjoint sets of matches, K' and L' (subsets of $\{m_2, \dots, m_{n-1}\}$) such that $f(w)_{I'} \equiv 1$ for $w \in [u]_{K'}$ and $f(w)_{J'} \equiv 1$ for $w \in [u]_{L'}$.

Now if neither 1 nor 2 is in $I \cup J$, then $I' = I$ and $J' = J$, so with $K = K'$ and $L = L'$, $f(w)_I \equiv 1$ whenever $w \in [u]_K$ and $f(w)_J \equiv 1$ whenever $w \in [u]_L$. Hence if $z = (z_1, w) \in [y]_K$, then $w \in [u]_K$, so $f(z)_I \equiv 1$ and analogously $f(z)_J \equiv 1$ when $z \in [y]_L$. Assume now that $2 \in I$. Then y_1 must equal 0, since otherwise $f(y)_2 = 0$, a contradiction. Let $K = K' \cup \{m_1\}$ and $L = L'$. It is obvious that $f(z)_J \equiv 1$ for $z \in [y]_L$. If $z = (z_1, w) \in [y]_K$, then $z_1 = 0$ and $w \in [u]_{K'}$. Therefore $f(z)_2 = 1$ and $f(z)_{I' \setminus \{2'\}} \equiv 1$, i.e. $f(z)_I \equiv 1$. Of course, the case $2 \in J$ is analogous, with $K = K'$ and $L = L' \cup \{m_1\}$. If instead $1 \in I$ (or $1 \in J$), then just replace 0 with 1 for y_1 and z_1 and repeat the argument. In all cases, we have found disjoint sets $K, L \subseteq \{m_1, \dots, m_{n-1}\}$ such that $f(z)_I \equiv 1$ whenever $z \in [y]_K$ and $f(z)_J \equiv 1$ whenever $z \in [y]_L$. Hence $y \in \widehat{A} \Box \widehat{B}$.

Now consider the case $\pi_1 + \pi_2 > 1$. Assume first that $y_1 = 1$, so that 1 wins match m_1. Hence match m_1 guarantees that $f(y)_1 = 1$ and we have $f(y)_2 = 1$ iff $f(u)_{2'} = 1$. Let $I' = I \setminus \{1\}$ if $2 \notin I$ and $I' = I \setminus \{1, 2\} \cup \{2'\}$ if $2 \in I$ and define J' identically. Then I' and J' are disjoint subsets of $\{m_2, \dots, m_n\}$, so the induction hypothesis allows us to assume that there are disjoint sets $K', L' \subseteq \{m_2, \dots, m_{n-1}\}$ such that $f(w)_{I'} \equiv 1$ for

$w \in [u]_{K'}$ and $f(w)_{J'} \equiv 1$ for $w \in [u]_{L'}$. If neither 1 nor 2 is in $I \cup J$, then we can take $K = K'$ and $L = L'$ exactly as for $\pi_1 + \pi_2 \leq 1$. If $1 \in I$ (or $1 \in J$) and $2 \notin I \cup J$, then take $K = K' \cup \{m_1\}$ and $L = L'$ (or $K = K'$, $L = L' \cup \{m_1\}$). If $2 \in I$ (or $2 \in J$) and $1 \notin J$, then take $K = K'$ and $L = L'$. If 1 and 2 are both in $I \cup J$, then take $K = K \cup \{m_1\}$, $L = L'$ if $1 \in I$ and $K = K'$ and $L = L' \cup \{m_1\}$ if $1 \in J$. In all cases K and L are disjoint, $f(z)_I \equiv 1$ for $z \in [y]_K$ and $f(z)_J \equiv 1$ for $z \in [y]_L$. This proves that $y \in \widehat{A} \square \widehat{B}$.

Finally if $y_1 = 1$, then repeat the same analysis with the rôles of 1 and 2 changed.

□

Now the proof of the main result is very short. Since the Y_i's are independent,

$$\begin{aligned} \mathbb{P}(X \in A \square B) &= \mathbb{P}(Y \in \widehat{A \square B}) \leq \mathbb{P}(Y \in \widehat{A} \square \widehat{B}) \\ &\leq \mathbb{P}(Y \in \widehat{A})\mathbb{P}(Y \in \widehat{B}) = \mathbb{P}(X \in A)\mathbb{P}(X \in B), \end{aligned}$$

where the second inequality is the van den Berg-Kesten-Reimer's inequality. This completes the proof. □

Remark. It may be tempting to believe that $\widehat{A \square B} = \widehat{A} \square \widehat{B}$. However the inclusion $\widehat{A} \square \widehat{B} \subseteq \widehat{A \square B}$ fails. The following example is due to an anonymous referee. Let $n = 4$, $\pi_1 = \pi_2 = 1/3$ and $\pi_3 = \pi_4 = 2/3$. Let A be the event that $X_2 + X_3 + X_4 \geq 2$ and let B be the event that $X_1 + X_2 + X_3 \geq 2$. Then $A \square B$ is the event that $X_1 = X_2 = X_2 = X_4 = 1$ and since all match sequences result samples of size 2, $\widehat{A \square B} = \emptyset$. However any match sequence $z = (z_1, z_2.z_3)$ with $z_1 = 0$ entails $X_1 = 0$ and hence A occurs. Analogously $z_3 = 1$ implies B. Hence $\widehat{A} \square \widehat{B} \supseteq \{(0,0,1),(0,1,1)\}$.

Acknowledgments. I would like to thank the two anonymous referees whose comments helped me to substantially improve this paper. I am particularly grateful to the referee who discovered an important mistake in the proof of the main lemma in an earlier version.

References

[1] J. van den Berg and A. Gandolfi (2012), BK-type inequalities and generalized random-cluster representations. Available at arXiv:1203.3665v1

[2] J. van den Berg and J. Jonasson (2011), A BK inequality for randomly drawn subsets of fixed size, *Probab. Th. Rel. Fields*, to appear. Available at arXiv:1105.3862 MR-3000563

[3] J. van den Berg and H. Kesten (1985), Inequalities with applications to percolation and reliability, *J. Appl. Probab.* **22**, 556-569. MR-0799280

[4] J. Borcea, P. Brändén and T. M. Liggett (2009), Negative dependence and the geometry of polynomials, *J. Amer. Math. Soc.* **22**, 521-567. MR-2476782

[5] P. Brändeén and J. Jonasson (2011), Negative dependence in sampling, to appear in *Scand. J. Statist.*, to appear. Available at http://www.math.chalmers.se/~jonasson/recent.html MR-3000852

[6] J. Brown Cramer, J. Cutler and A. J. Radcliffe (2011), Negative dependence and Srinivasan's sampling process, *Combin. Probab. Comput.* **20**, 347-361. MR-2784632

[7] J-C. Deville and Y. Tillé (1998), Unequal probability sampling without replacement through a splitting method, *Biometrika* **85**, 89-101. MR-1627234

[8] D. Dubhasi, J. Jonasson, D. Ranjan (2007), Positive influence and negative dependence, *Combin. Probab. Comput.* **16**, 29-41. MR-2286510

[9] D. Dubhashi and D. Ranjan (1998), Balls and bins: A study in negative dependence, *Random Structures Algorithms* **13**, 99-124. MR-1642566

[10] G. Grimmett, "Percolation," Springer Verlag, 1999. MR-1707339

[11] K. Markström (2009), Closure Properties and Negatively Associated Measures violating the van den Berg-Kesten Inequality, *Electronic Commun. Probab.* **15**, 449-456. MR-2726091

[12] R. Pemantle (2000), Towards a theory of negative dependence, J. Math. Phys. **41**, 1371-1390. MR-1757964

[13] D. Reimer (2000), A proof of the van den Berg-Kesten conjecture, *Combin. Probab. Computing* **9**, 27-32. MR-1751301

[14] A. Srinivasan (2001), Distributions on level sets with applications to approximation algorithms, in *Proc. 42nd IEEE Symposium on the Foundations of Computer Science (FOCS).* MR-1948748