

Weighted sampling without replacement

Anna Ben-Hamou^a, Yuval Peres^b and Justin Salez^a

^a*Université Paris Diderot*

^b*Microsoft Research*

Abstract. Comparing concentration properties of uniform sampling with and without replacement has a long history which can be traced back to the pioneer work of Hoeffding [8]. The goal of this note is to extend this comparison to the case of non-uniform weights, using a coupling between samples drawn with and without replacement. When the items' weights are arranged in the same order as their values, we show that the induced coupling for the cumulative values is a submartingale coupling. As a consequence, the powerful Chernoff-type upper-tail estimates known for sampling with replacement automatically transfer to the case of sampling without replacement. For general weights, we use the same coupling to establish a sub-Gaussian concentration inequality. As the sample size approaches the total number of items, the variance factor in this inequality displays the same kind of sharpening as Serfling [17] identified in the case of uniform weights. We also construct an other martingale coupling which allows us to answer a question raised by Luh and Pippenger [12] on sampling in Polya urns with different replacement numbers.

1 Introduction

In a celebrated paper [8], Hoeffding first singled out a fruitful comparison between sampling with and without replacement: any linear statistics induced by uniform sampling without replacement in a finite population is less, in the convex order, than the one induced by sampling with replacement. In particular, all the Chernoff-type tail estimates that apply to sampling with replacement (the sample then being i.i.d.) automatically apply to sampling without replacement. As the sample size increases, it is natural to expect that sampling without replacement should concentrate even more, in the sense that, when the sample size approaches the total number of items, the variance should not be of the order of the number of sampled items, but of the number of *unsampled* items. This was verified by Serfling in [17].

One natural question is to determine whether a similar comparison also holds when the sampling procedure is no longer uniform and when different items have different weights.

More precisely, consider a collection of N items $1 \leq i \leq N$, each equipped

with a *weight* $\omega(i) > 0$ and a *value* of interest $\nu(i) \in \mathbb{R}$. We assume that

$$\sum_{i=1}^N \omega(i) = 1.$$

Let X be the cumulative value of a sample of length $n \leq N$ drawn without replacement and with probability proportional to weights, i.e.

$$X := \nu(\mathbf{I}_1) + \cdots + \nu(\mathbf{I}_n),$$

where for each n -tuple (i_1, \dots, i_n) of distinct indices in $\{1, \dots, N\}$,

$$\mathbb{P}((\mathbf{I}_1, \dots, \mathbf{I}_n) = (i_1, \dots, i_n)) = \prod_{k=1}^n \frac{\omega(i_k)}{1 - \omega(i_1) - \cdots - \omega(i_{k-1})}.$$

A much simpler statistic is the one that arises when the sample is drawn with replacement, namely

$$Y := \nu(\mathbf{J}_1) + \cdots + \nu(\mathbf{J}_n),$$

where now for each n -tuple $(j_1, \dots, j_n) \in \{1, \dots, N\}^n$,

$$\mathbb{P}((\mathbf{J}_1, \dots, \mathbf{J}_n) = (j_1, \dots, j_n)) = \prod_{k=1}^n \omega(j_k).$$

One particular case, referred to as the monotone case, is when weights and values are arranged in the same order, *i.e.*

$$\omega(i) > \omega(j) \implies \nu(i) \geq \nu(j). \quad (1)$$

Theorem 1. *Assume that condition (1) holds. Then X is less than Y in the increasing convex order, i.e. for every non-decreasing, convex function $f: \mathbb{R} \rightarrow \mathbb{R}$,*

$$\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]. \quad (2)$$

Our second result is a sub-Gaussian concentration inequality for X in the case of arbitrary weights $(\omega(i))_{i=1}^N$. Define

$$\Delta := \max_{1 \leq i \leq N} \nu(i) - \min_{1 \leq i \leq N} \nu(i) \quad \text{and} \quad \alpha = \frac{\min_{1 \leq i \leq N} \omega(i)}{\max_{1 \leq i \leq N} \omega(i)}.$$

The case $\alpha = 1$ (uniform sampling) was analysed by Serfling [17].

Theorem 2. *Assume $\alpha < 1$. For all $t > 0$,*

$$\max \{ \mathbb{P}(X - \mathbb{E}X > t), \mathbb{P}(X - \mathbb{E}X < -t) \} \leq \exp \left(-\frac{t^2}{2v} \right),$$

with

$$v = \min \left(4\Delta^2 n, \frac{1 + 4\alpha}{\alpha(1 - \alpha)} \Delta^2 N \left(\frac{N - n}{N} \right)^\alpha \right) \quad (3)$$

We also answer a question raised by Luh and Pippenger [12]. The problem is to compare linear statistics induced by sampling in Polya urns with replacement number d versus D , for positive integers d, D with $D > d \geq 1$.

Let \mathbf{C} be a population of N items, labelled from 1 to N , each item i being equipped with some value $\nu(i)$. Let $d < D$ be two positive integers. For $n \geq 1$, let (K_1, \dots, K_n) and (L_1, \dots, L_n) be samples generated by sampling in Polya urns with initial composition \mathbf{C} and replacement numbers d and D respectively, *i.e.* each time an item is picked, it is replaced along with $d - 1$ (resp. $D - 1$) copies. We say that (K_1, \dots, K_n) (resp. (L_1, \dots, L_n)) is a d -Polya (resp. D -Polya) sample. Let

$$\begin{aligned} W &= \nu(K_1) + \dots + \nu(K_n), \\ Z &= \nu(L_1) + \dots + \nu(L_n). \end{aligned}$$

Theorem 3. *The variable W is less than Z in the convex order, *i.e.* for every convex function $f: \mathbb{R} \rightarrow \mathbb{R}$,*

$$\mathbb{E}[f(W)] \leq \mathbb{E}[f(Z)].$$

Remark 1. [12] proved a similar result in the case where the first sample is drawn without replacement in \mathbf{C} and the second is a D -Polya sample, for $D \geq 1$.

The rest of the paper is organized as follows. In Section 2, we give a brief review of the literature on sampling without replacement and stochastic ordering. Both Theorems 1 and 2 rely on a coupling between samples drawn with and without replacement, which is constructed in Section 3. In Section 4, we consider the monotone case and prove Theorem 1. Section 5 investigates the concentration properties of weighted sampling without replacement in the general case and is devoted to the proof of Theorem 2. Theorem 3 is proved in Section 6. Section 7 contains concluding remarks.

2 Related work

Weighted sampling without replacement, also known as *successive sampling*, appears in a variety of contexts. In statistical ecology for instance, the Horvitz-Thompson estimator [10] is a well-known estimator for the population total $\sum_{i=1}^N \nu(i)$ and takes the form of a cumulative value of samples drawn without replacement:

$$HT_n = \sum_{k=1}^n \frac{\nu(\mathbf{I}_k)}{P_n(\mathbf{I}_k)},$$

where, $P_n(i) = \mathbb{P}(i \in \{\mathbf{I}_1, \dots, \mathbf{I}_n\})$ is the inclusion probability of item $i \in \{1, \dots, N\}$.

Allowing heterogeneous weights can make the analysis of sampling without replacement challenging. Even simple quantities such as the expectation $\mathbb{E}X$ or the inclusion probabilities $\mathbb{P}(i \in \{\mathbf{I}_1, \dots, \mathbf{I}_n\})$ can be hard to track, even asymptotically [16, 21]. The limiting distribution of X has been deeply investigated and central limit theorems were obtained under various regimes [7, 9, 6]. However, in situations where the sample size is relatively small, one may also be interested in non-asymptotic results, in the form of deviation or concentration inequalities. When $n \ll N$, it is natural to expect Y to be a good approximation of X . For instance, the total-variation distance between $\mathbb{P}(\mathbf{I}_{n+1} \in \cdot \mid (\mathbf{I}_k)_{k=1}^n)$ and $\mathbb{P}(\mathbf{J}_1 \in \cdot)$ is given by $\sum_{k=1}^n \omega(\mathbf{I}_k)$, which is $O(n/N)$ provided all the weights are $O(1/N)$.

Under the monotonicity assumption (1), Theorem 1 establishes an exact strong stochastic ordering between X and Y . Since $\mathbf{J}_1, \dots, \mathbf{J}_n$ are independent copies of \mathbf{I}_1 , the innumerable results on sums of independent and identically distributed random variables apply to Y . In particular, Chernoff's bound

$$\mathbb{P}(Y \geq a) \leq \exp(n\Lambda(\theta) - \theta a), \quad (4)$$

yields a variety of sharp concentration results based on efficient controls on the log-Laplace transform $\Lambda(\theta) = \ln \mathbb{E}[e^{\theta\nu(\mathbf{I}_1)}]$. This includes the celebrated Hoeffding and Bernstein inequalities, see the book [5]. Theorem 1 implies in particular that all upper-tail estimates derived from Chernoff's bound (4) apply to X without modification.

The condition (1) describes a sampling procedure which is sometimes referred to as size-biased sampling without replacement. It arises in many situations, including ecology, oil discovery models, in the construction of

the Poisson-Dirichlet distribution [15, 14], or in the configuration model of random graphs [3, 4].

Stochastic orders provide powerful tools to compare distributions of random variables and processes, and they have been used in various applications [20, 13, 18]. As other stochastic relations, the increasing convex order is only concerned with marginal distributions. One way of establishing (2) is thus to carefully construct two random variables X and Y with the correct marginals on a common probability space, in such a way that

$$X \leq \mathbb{E}[Y|X] \quad (5)$$

holds almost-surely. The existence of such a *submartingale coupling* clearly implies (2), thanks to Jensen's inequality. Quite remarkably, the converse is also true, as proved by Strassen [19]. Similarly, Theorem 3 is equivalent to the existence of a *martingale coupling* (W, Z) .

Remark 2 (The uniform case). *When ω is constant, i.e. $\alpha = 1$, the sequence $(\mathbf{I}_1, \dots, \mathbf{I}_n)$ is exchangeable. In particular, $\mathbb{E}[X] = \mathbb{E}[Y]$, forcing equality in (5). Thus, (2) automatically extends to arbitrary convex functions. This important special case was established five decades ago by Hoeffding in his seminal paper [8]. Since then, improvements have been found as n/N approaches 1 [17, 2]. Another remarkable feature of uniform sampling without replacement is the negative association of the sequence $(\nu(\mathbf{I}_1), \dots, \nu(\mathbf{I}_n))$ [11]. However, this result seems to make crucial use of the exchangeability of $(\mathbf{I}_1, \dots, \mathbf{I}_n)$, and it is not clear whether it can be extended to more general weights, e.g. to monotone weights satisfying (1). Non-uniform sampling without replacement can be more delicate and induce counter-intuitive correlations, as highlighted by Alexander [1], who showed that for two fixed items, the indicators that each is in the sample can be positively correlated.*

In the non-uniform case, $\mathbb{E}X$ and $\mathbb{E}Y$ need not be equal. Hence, Theorem 1 may not entail concentration inequalities for X around its mean, but rather bounds on $\mathbb{P}(X \geq \mathbb{E}Y + t)$, for $t > 0$. Theorem 2 then establishes a sub-Gaussian concentration inequality for X . It holds under the only assumption that $\alpha < 1$, but the domain of application that we have in mind is when α is bounded away from 0 and 1. In this domain, when $n \leq qN$, for some fixed $0 < q < 1$, equation (3) gives $v = O(\Delta^2 n)$, which corresponds to the order of the variance factor in the classical Hoeffding inequality. When $n/N \rightarrow 1$ as $N \rightarrow +\infty$, it can be improved up to $v = O\left(\Delta^2 n \left(\frac{N-n}{N}\right)^\alpha\right)$. In the uniform case $\alpha = 1$, Serfling [17] showed that X satisfies a sub-Gaussian inequality with $v = \Delta^2 n \frac{N-n+1}{4N}$, implying that the variance factor has the order of the minimum between the number of sampled and unsampled items.

3 A coupling between samples drawn with and without replacement

The proofs of Theorem 1 and 2 rely on a particular coupling of samples drawn with and without replacement. This coupling is inspired by the one described in [12] for the uniform case.

First generate an infinite sequence $(\mathbf{J}_k)_{k \geq 1}$ by sampling with replacement and with probability proportional to $(\omega(i))_{i=1}^N$. Now, “screen” this sequence, starting at \mathbf{J}_1 as follows: for $1 \leq k \leq N$, set

$$\mathbf{I}_k = \mathbf{J}_{T_k},$$

where T_k is the random time when the k^{th} distinct item appears in $(\mathbf{J}_i)_{i \geq 1}$.

The sequence $(\mathbf{I}_1, \dots, \mathbf{I}_n)$ is then distributed as a sample without replacement. As above, we define $X = \sum_{k=1}^n \nu(\mathbf{I}_k)$ and $Y = \sum_{k=1}^n \nu(\mathbf{J}_k)$.

4 The monotone case

Consider the coupling of X and Y described above (Section 3). Under the monotonicity assumption (1), we show that (X, Y) is a submartingale coupling in the sense of (5). As the sequence $(\mathbf{J}_1, \dots, \mathbf{J}_n)$ is exchangeable and as permuting \mathbf{J}_i and \mathbf{J}_j in this sequence does not affect X , it is sufficient to show that $\mathbb{E}[\nu(\mathbf{J}_1)|X] \geq X/n$.

Proof of Theorem 1. Let $\{i_1, \dots, i_n\} \subset \{1, \dots, N\}$ be a set of cardinality n , and let \mathbf{A} be the event $\{\mathbf{I}_1, \dots, \mathbf{I}_n\} = \{i_1, \dots, i_n\}$.

$$\mathbb{E}[\nu(\mathbf{J}_1)|\mathbf{A}] = \sum_{j=1}^n \mathbb{P}(\mathbf{J}_1 = i_j | \mathbf{A}) \nu(i_j).$$

Let us now show that, for all $1 \leq k \neq \ell \leq n$, if $\nu(i_k) \geq \nu(i_\ell)$, then $\mathbb{P}(\mathbf{J}_1 = i_k | \mathbf{A})$ is not smaller than $\mathbb{P}(\mathbf{J}_1 = i_\ell | \mathbf{A})$. First, by (1), one has $\omega(i_k) \geq \omega(i_\ell)$. Letting \mathfrak{S}_n be the set of permutations of n elements, one has

$$\mathbb{P}(\{\mathbf{J}_1 = i_k\} \cap \mathbf{A}) = \sum_{\pi \in \mathfrak{S}_n, \pi(1)=k} p(\pi),$$

where

$$p(\pi) := \omega(i_{\pi(1)}) \frac{\omega(i_{\pi(2)})}{1 - \omega(i_{\pi(1)})} \cdots \frac{\omega(i_{\pi(n)})}{1 - \omega(i_{\pi(1)}) - \omega(i_{\pi(2)}) - \cdots - \omega(i_{\pi(n-1)})}$$

Now, each permutation π with $\pi(1) = k$ can be uniquely associated with a permutation π^* such that $\pi^*(1) = \ell$, by performing the switch: $\pi^*(\pi^{-1}(\ell)) = k$, and letting $\pi(j) = \pi^*(j)$, for all $j \notin \{1, \pi^{-1}(\ell)\}$. Observe that $p(\pi) \geq p(\pi^*)$. Thus

$$\mathbb{P}(\mathbf{J}_1 = i_k | \mathbf{A}) - \mathbb{P}(\mathbf{J}_1 = i_\ell | \mathbf{A}) = \frac{1}{\mathbb{P}(\mathbf{A})} \sum_{\pi \in \mathfrak{S}_n, \pi(1)=k} (p(\pi) - p(\pi^*)) \geq 0.$$

Consequently, by Chebyshev's sum inequality,

$$\begin{aligned} \mathbb{E}[\nu(\mathbf{J}_1) | \mathbf{A}] &= n \frac{1}{n} \sum_{j=1}^n \mathbb{P}(\mathbf{J}_1 = i_j | \mathbf{A}) \nu(i_j) \\ &\geq n \left(\frac{1}{n} \sum_{j=1}^n \mathbb{P}(\mathbf{J}_1 = i_j | \mathbf{A}) \right) \left(\frac{1}{n} \sum_{j=1}^n \nu(i_j) \right) \\ &= \frac{\sum_{j=1}^n \nu(i_j)}{n}, \end{aligned}$$

and $\mathbb{E}[Y | X] \geq X$, which concludes the proof of Theorem 1. \square

5 A sub-Gaussian concentration inequality

We now consider general weights, and proceed with the proof of Theorem 2. We only need to show that the bound in Theorem 2 holds for $\mathbb{P}[X - \mathbb{E}X > t]$. Indeed, replacing X by $-X$ (i.e. changing all the values to their opposite) does not affect the proof. Hence, the bound on $\mathbb{P}[X - \mathbb{E}X < -t]$ will follow directly.

Theorem 2 is proved using the same coupling between sampling with and without replacement as described in Section 3.

Note that, in this coupling, X is a function of the i.i.d. variables $(\mathbf{J}_i)_{i \geq 1}$:

$$X = \sum_{i=1}^{+\infty} \nu(\mathbf{J}_i) \mathbb{1}_{\{\mathbf{J}_i \notin \{\mathbf{J}_1, \dots, \mathbf{J}_{i-1}\}\}} \mathbb{1}_{\{T_n \geq i\}}. \quad (6)$$

As such, one may obtain concentration results for X by resorting to the various methods designed for functions of independent variables.

Proof of Theorem 2. The proof relies on the *entropy method* as described in Chapter 6 of [5]. We will show that X is such that, for all $\lambda > 0$,

$$\lambda \mathbb{E}[X e^{\lambda X}] - \mathbb{E}[e^{\lambda X}] \log \mathbb{E}[e^{\lambda X}] \leq \frac{\lambda^2 v}{2} \mathbb{E}[e^{\lambda X}], \quad (7)$$

for v as in (3). Then, a classical argument due to Herbst (see [5], Proposition 6.1) ensures that, for all $\lambda > 0$,

$$\log \mathbb{E} \left[e^{\lambda(X - \mathbb{E}X)} \right] \leq \frac{\lambda^2 v}{2},$$

and thus, for all $t > 0$,

$$\mathbb{P}(X - \mathbb{E}X > t) \leq \exp\left(-\frac{t^2}{2v}\right),$$

that is, the upper-tail of X is sub-Gaussian with variance factor v . Let us establish inequality (7). For $t \geq 1$, consider the truncated variable X_t defined by summing only from 1 to t in (6), i.e.

$$\begin{aligned} X_t &= \sum_{i=1}^t \nu(\mathbf{J}_i) \mathbb{1}_{\{\mathbf{J}_i \notin \{\mathbf{J}_1, \dots, \mathbf{J}_{i-1}\}\}} \mathbb{1}_{\{T_n \geq i\}} \\ &:= f(\mathbf{J}_1, \dots, \mathbf{J}_t). \end{aligned}$$

Note that X_t converges to X almost surely as $t \rightarrow +\infty$. Then, for all $1 \leq i \leq t$, consider the perturbed variable X_t^i which is obtained by replacing \mathbf{J}_i by an independent copy \mathbf{J}'_i , i.e.

$$X_t^i = f(\mathbf{J}_1, \dots, \mathbf{J}_{i-1}, \mathbf{J}'_i, \mathbf{J}_{i+1}, \dots, \mathbf{J}_t),$$

and let X^i be the almost sure limit of X_t^i , as $t \rightarrow +\infty$. Theorem 6.15 of [5] implies that, for all $\lambda > 0$,

$$\lambda \mathbb{E} \left[X_t e^{\lambda X_t} \right] - \mathbb{E} \left[e^{\lambda X_t} \right] \log \mathbb{E} \left[e^{\lambda X_t} \right] \leq \sum_{i=1}^t \mathbb{E} \left[\lambda^2 e^{\lambda X_t} (X_t - X_t^i)_+^2 \right] \quad (8)$$

We now show that this inequality still holds when we let t tend to $+\infty$. Let $\nu_{\max} = \max_{1 \leq j \leq N} \nu(j)$. For all $t \geq 1$, the variable X_t is almost surely bounded by $n\nu_{\max}$. Hence, the left-hand side of (8) tends to the left-hand side of (7). As for the right-hand side, we have that, for all $1 \leq i \leq t$,

$$\mathbb{E} \left[\lambda^2 e^{\lambda X_t} (X_t - X_t^i)_+^2 \right] \leq \lambda^2 e^{\lambda n\nu_{\max}} \Delta^2 \mathbb{P}(i \leq T_n),$$

and $\sum_{i=1}^{+\infty} \mathbb{P}[i \leq T_n] = \mathbb{E}[T_n] < +\infty$. Hence, by dominated convergence, the right-hand side also converges, and we obtain

$$\lambda \mathbb{E} \left[X e^{\lambda X} \right] - \mathbb{E} \left[e^{\lambda X} \right] \log \mathbb{E} \left[e^{\lambda X} \right] \leq \sum_{i=1}^{+\infty} \mathbb{E} \left[\lambda^2 e^{\lambda X} (X - X^i)_+^2 \right].$$

Recall that $(\mathbf{I}_1, \dots, \mathbf{I}_n)$ is the sequence of the first n distinct items in $(\mathbf{J}_i)_{i \geq 1}$ and that X is measurable with respect to $\sigma(\mathbf{I}_1, \dots, \mathbf{I}_n)$, so that

$$\sum_{i=1}^{+\infty} \mathbb{E} \left[\lambda^2 e^{\lambda X} (X - X^i)_+^2 \right] = \mathbb{E} \left[\lambda^2 e^{\lambda X} \mathbb{E} \left[\sum_{i=1}^{+\infty} (X - X^i)_+^2 \middle| \mathbf{I}_1, \dots, \mathbf{I}_n \right] \right].$$

Thus, letting

$$V := \mathbb{E} \left[\sum_{i=1}^{+\infty} (X - X^i)_+^2 \middle| \mathbf{I}_1, \dots, \mathbf{I}_n \right],$$

our task comes down to showing that

$$V \leq \frac{v}{2} \quad \text{a.s..}$$

Observe that for all $i \geq 1$, we have $(X - X^i)_+^2 \leq \Delta^2$ and that $X = X^i$ unless $i \leq T_n$ and one of the following two events occurs:

- $\mathbf{J}'_i \notin \{\mathbf{I}_1, \dots, \mathbf{I}_n\}$;
- the item \mathbf{J}_i occurs only once before T_{n+1} .

Let us define

$$A = \sum_{i=1}^{+\infty} \mathbb{E} \left[\mathbb{1}_{\{\mathbf{J}'_i \notin \{\mathbf{I}_1, \dots, \mathbf{I}_n\}\}} \mathbb{1}_{i \leq T_n} \middle| \mathbf{I}_1, \dots, \mathbf{I}_n \right],$$

and

$$B = \sum_{k=1}^n \mathbb{E} \left[\mathbb{1}_{\{\exists! i < T_{n+1}, \mathbf{J}_i = \mathbf{I}_k\}} \middle| \mathbf{I}_1, \dots, \mathbf{I}_n \right],$$

so that $V \leq \Delta^2 (A + B)$. Since \mathbf{J}'_i is independent of everything else and since $\sigma_n := \omega(\mathbf{I}_1) + \dots + \omega(\mathbf{I}_n)$ is a measurable function of $(\mathbf{I}_1, \dots, \mathbf{I}_n)$, we have

$$A = (1 - \sigma_n) \mathbb{E} \left[T_n \middle| \mathbf{I}_1, \dots, \mathbf{I}_n \right].$$

We use the following fact.

Lemma 3. *For $1 \leq k \leq n$, let $\tau_k = T_k - T_{k-1}$. Conditionally on $(\mathbf{I}_1, \dots, \mathbf{I}_n)$, the variables $(\tau_k)_{k=1}^n$ are independent and for all $1 \leq k \leq n$, τ_k is distributed as a Geometric random variables with parameters $1 - \sigma_{k-1}$.*

Proof. Let (i_1, \dots, i_n) be an n -tuple of distinct elements of $\{1, \dots, N\}$ and let $t_1, \dots, t_n \geq 1$. Let also $(G_k)_{k=1}^n$ be independent Geometric random variables with parameter $(1 - \omega(i_1) - \dots - \omega(i_{k-1}))$. We have

$$\begin{aligned} \mathbb{P}((\tau_1, \dots, \tau_n) = (t_1, \dots, t_n), (\mathbf{I}_1, \dots, \mathbf{I}_n) = (i_1, \dots, i_n)) \\ &= \mathbb{1}_{\{t_1=1\}} \omega(i_1) \prod_{k=2}^n (\omega(i_1) + \dots + \omega(i_{k-1}))^{t_k-1} \omega(i_k) \\ &= \prod_{k=1}^n \frac{\omega(i_k)}{1 - \omega(i_1) - \dots - \omega(i_{k-1})} \prod_{k=1}^n \mathbb{P}(G_k = t_k) \\ &= \mathbb{P}((\mathbf{I}_1, \dots, \mathbf{I}_n) = (i_1, \dots, i_n)) \prod_{k=1}^n \mathbb{P}(G_k = t_k), \end{aligned}$$

and we obtain the desired result. \square

Lemma 3 implies that

$$\mathbb{E}[T_n | \mathbf{I}_1, \dots, \mathbf{I}_n] = \sum_{k=1}^n \frac{1}{1 - \sigma_{k-1}}.$$

In particular, $A \leq n$. We also have

$$A \leq \frac{1}{\alpha} \sum_{k=1}^n \frac{N-n}{N-k+1} \leq \frac{1}{\alpha} (N-n) \log\left(\frac{N}{N-n}\right). \quad (9)$$

It remains to control B . Clearly $B \leq n$, which shows that $V \leq 2\Delta^2 n$. Moreover, for $1 \leq k \leq n$, we have

$$\mathbb{P}\left(\exists! i < T_{n+1}, \mathbf{J}_i = \mathbf{I}_k | \mathbf{I}_1, \dots, \mathbf{I}_n\right) = \mathbb{E}\left[\prod_{j=k}^n \left(1 - \frac{\omega(\mathbf{I}_k)}{\sigma_j}\right)^{\tau_{j+1}-1} | \mathbf{I}_1, \dots, \mathbf{I}_n\right].$$

Using Lemma 3 and the fact that the generating function of a geometric variable G with parameter p is given by $\mathbb{E}[x^G] = \frac{px}{1-(1-p)x}$, we obtain

$$B = \sum_{k=1}^n \prod_{j=k}^n \frac{1}{1 + \frac{\omega(\mathbf{I}_k)}{1-\sigma_j}}.$$

Thanks to the inequality the inequality $\log(1+x) \geq x - x^2/2$ for $x \geq 0$,

$$B \leq \sum_{k=1}^n \prod_{j=k}^n \frac{1}{1 + \frac{\alpha}{N-j}} \leq \sum_{k=1}^n \exp\left(-\alpha \sum_{j=k}^n \frac{1}{N-j} + \frac{1}{2} \sum_{j=k}^n \frac{1}{(N-j)^2}\right).$$

The second term in the exponent is always smaller than $1/2$. Using Riemann sums, we get

$$\begin{aligned} B &\leq 2 \sum_{k=1}^n \exp\left(-\alpha \log\left(\frac{N-k+1}{N-n}\right)\right) = 2 \sum_{k=1}^n \left(\frac{N-n}{N-k+1}\right)^\alpha \\ &\leq \frac{2}{1-\alpha} N \left(\frac{N-n}{N}\right)^\alpha, \end{aligned}$$

Combined with (9), this yields

$$\begin{aligned} V &\leq \left(\frac{1}{\alpha} \left(\frac{N-n}{N}\right)^{1-\alpha} \log\left(\frac{N}{N-n}\right) + \frac{2}{1-\alpha}\right) \Delta^2 N \left(\frac{N-n}{N}\right)^\alpha \\ &\leq \left(\frac{e^{-1}}{\alpha(1-\alpha)} + \frac{2}{1-\alpha}\right) \Delta^2 N \left(\frac{N-n}{N}\right)^\alpha \\ &\leq \frac{1/2 + 2\alpha}{\alpha(1-\alpha)} \Delta^2 N \left(\frac{N-n}{N}\right)^\alpha, \end{aligned}$$

where the second inequality is due to the fact that $\log(x)/x^{1-\alpha} \leq e^{-1}/(1-\alpha)$ for all $x > 0$. \square

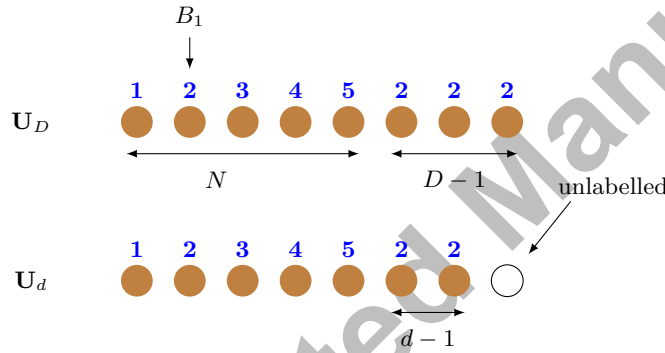
6 A martingale coupling for Polya urns

The proof of Theorem 3 relies on the construction of a martingale coupling (W, Z) , *i.e.* of a coupling of W and Z such that $\mathbb{E}[Z|W] = W$.

Proof of Theorem 3. Consider two urns, \mathbf{U}_d and \mathbf{U}_D , each of them initially containing N balls, labelled from 1 to N . In each urn, arrange the balls from left to right by increasing order of their label. Then arrange \mathbf{U}_D and \mathbf{U}_d on top of one another. Each time we will pick a ball in \mathbf{U}_D , we will pick the ball just below it in \mathbf{U}_d . More precisely, we perform an infinite sequence of steps as follows: at step 1, we pick a ball B_1 uniformly at random in \mathbf{U}_D and pick the ball just below it in \mathbf{U}_d . They necessarily have the same label, say j . We let $K_1 = L_1 = j$, and add, on the right part of \mathbf{U}_D , $D-1$ balls with label j , and, on the right part of \mathbf{U}_d , $d-1$ balls with label j and $D-d$ unlabelled balls. Note that, at the end of this step, the two urns still have the same number of balls, $N+D-1$. The first step is depicted in Figure 1. Then, at each step t , we pick a ball B_t at random among the $N+(t-1)(D-1)$ balls of \mathbf{U}_D and choose the ball just below it in \mathbf{U}_d . There are two different possibilities:

- if the ball drawn in \mathbf{U}_d is unlabelled and the one drawn in \mathbf{U}_D has label j , we let $L_t = j$ and add $D - 1$ balls with label j on the right part of \mathbf{U}_D , and $D - 1$ unlabelled balls on the right part of \mathbf{U}_d ,
- if both balls have label j , and if t corresponds to the i^{th} time a labelled ball is drawn in \mathbf{U}_d , we let $L_t = K_i = j$ and add $D - 1$ balls with label j on the right part of \mathbf{U}_D , and $d - 1$ balls with label j and $D - d$ unlabelled balls on the right part of \mathbf{U}_d ;

Figure 1 The ball B_1 has label 2 ($N = 5$, $d = 3$, $D = 4$).



The sequence (K_1, \dots, K_n) records the labels of the first n labelled balls picked in \mathbf{U}_d , and (L_1, \dots, L_n) the labels of the first n balls picked in \mathbf{U}_D . Observe that (K_1, \dots, K_n) (resp. (L_1, \dots, L_n)) is distributed as a d -Polya (resp. D -Polya) sample. Define

$$\begin{aligned} W &= \nu(K_1) + \dots + \nu(K_n), \\ Z &= \nu(L_1) + \dots + \nu(L_n). \end{aligned}$$

Let us show that $1 \leq i \leq n - 1$, $\mathbb{E}[\nu(L_{i+1})|W] = \mathbb{E}[\nu(L_i)|W]$. Let $\{k_1, \dots, k_n\}$ be a multiset of cardinality n of elements of $\{1, \dots, N\}$, and let \mathbf{A} be the event $\{K_1, \dots, K_n\} = \{k_1, \dots, k_n\}$ (accounting for the multiplicity of each label). Denote by \mathcal{C}_i the set of $D - 1$ balls added at step i . Observe that, if $B_{i+1} \in \mathcal{C}_i$, then $L_{i+1} = L_i$. Hence

$$\mathbb{E}[\nu(L_{i+1})|\mathbf{A}] = \mathbb{E}[\nu(L_i)\mathbb{1}_{\{B_{i+1} \in \mathcal{C}_i\}}|\mathbf{A}] + \mathbb{E}[\nu(L_{i+1})\mathbb{1}_{\{B_{i+1} \notin \mathcal{C}_i\}}|\mathbf{A}].$$

We have

$$\mathbb{E}[\nu(L_{i+1})\mathbb{1}_{\{B_{i+1} \notin \mathcal{C}_i\}}|\mathbf{A}] = \frac{1}{\mathbb{P}(\mathbf{A})} \sum_{k=1}^N \nu(k) \sum_{\ell=1}^N \mathbb{P}(L_i = \ell, L_{i+1} = k, B_{i+1} \notin \mathcal{C}_i, \mathbf{A}).$$

Notice that, on the event $B_{i+1} \notin \mathcal{C}_i$, the balls B_i and B_{i+1} are exchangeable. Hence $\mathbb{P}(L_i = \ell, L_{i+1} = k, B_{i+1} \notin \mathcal{C}_i) = \mathbb{P}(L_i = k, L_{i+1} = \ell, B_{i+1} \notin \mathcal{C}_i)$. Moreover, permuting B_i and B_{i+1} can not affect the multiset $\{K_1, \dots, K_n\}$. Hence

$$\mathbb{E} \left[\nu(L_{i+1}) \mathbb{1}_{\{B_{i+1} \notin \mathcal{C}_i\}} \middle| \mathbf{A} \right] = \mathbb{E} \left[\nu(L_i) \mathbb{1}_{\{B_{i+1} \notin \mathcal{C}_i\}} \middle| \mathbf{A} \right],$$

and $\mathbb{E} \left[\nu(L_{i+1}) \middle| W \right] = \mathbb{E} \left[\nu(L_i) \middle| W \right]$. We get that, for all $1 \leq i \leq n$,

$$\mathbb{E} \left[\nu(L_i) \middle| W \right] = \mathbb{E} \left[\nu(L_1) \middle| W \right] = \mathbb{E} \left[\nu(K_1) \middle| W \right] = W/n,$$

where the last equality comes from the exchangeability of (K_1, \dots, K_n) . \square

7 Concluding remarks

To summarize the three contributions of the paper,

- we proved that, when weights and values are arranged in the same order, then the sum of values arising from sampling without replacement is less, in the increasing order, than the sum arising from sampling with replacement;
- in the general case, we established a sub-Gaussian concentration inequality for sampling without replacement, in which the variance factor sharpens as the sample size approaches the total number of items;
- we answered a question posed by Luh and Pippenger [12], by constructing a martingale coupling between Polya urns with different replacement numbers.

Various questions are left open. The variance factor in Theorem 2, for instance, is probably not order-optimal, and it would be interesting to see whether such a concentration inequality holds with $v = O\left(n\left(1 - \frac{n}{N}\right)\right)$, as in Serfling's inequality. Also, the dependence properties of the sequence $(\nu(\mathbf{I}_1), \dots, \nu(\mathbf{I}_n))$ are far from being well understood, even in the monotone case.

References

- [1] K. S. Alexander et al. A counterexample to a correlation inequality in finite sampling. *The Annals of Statistics*, 17(1):436–439, 1989.
- [2] R. Bardenet and O.-A. Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.
- [3] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

- [4] B. Bollobás. *Random graphs*. Springer, 1998.
- [5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013.
- [6] L. Gordon. Successive sampling in large finite populations. *The Annals of Statistics*, 11(2):702–706, 1983.
- [7] J. Hájek. Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5(361):74, 1960.
- [8] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
- [9] L. Holst. Some limit theorems with applications in sampling theory. *The Annals of Statistics*, 1:644–658, 1973.
- [10] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.
- [11] K. Joag-Dev and F. Proschan. Negative association of random variables with applications. *The Annals of Statistics*, pages 286–295, 1983.
- [12] K. Luh and N. Pippenger. Large-deviation bounds for sampling without replacement. *American Mathematical Monthly*, 121(5):449–454, 2014.
- [13] A. Müller and D. Stoyan. *Comparison methods for stochastic models and risks*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2002.
- [14] J. Pitman, N. M. Tran, et al. Size-biased permutation of a finite sequence with independent and identically distributed terms. *Bernoulli*, 21(4):2484–2512, 2015.
- [15] J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.
- [16] B. Rosén. Asymptotic theory for successive sampling with varying probabilities without replacement. I, II. *Annals of Mathematical Statistics*, 43:373–397; *ibid.* 43 (1972), 748–776, 1972.
- [17] R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2:39–48, 1974.
- [18] M. Shaked and J. G. Shanthikumar. *Stochastic orders*. Springer Series in Statistics. Springer, New York, 2007.
- [19] V. Strassen. The existence of probability measures with given marginals. *Annals of Mathematical Statistics*, 36:423–439, 1965.
- [20] R. Szekli. *Stochastic ordering and dependence in applied probability*, volume 97 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1995.
- [21] Y. Yu. On the inclusion probabilities in some unequal probability sampling plans without replacement. *Bernoulli*, 18(1):279–289, 2012.

Université Paris Diderot, LPMA
 Bâtiment Sophie Germain
 Paris 75013, France.
 E-mail: benhamou@math.univ-paris-diderot.fr
 E-mail: salez@math.univ-paris-diderot.fr

Microsoft Research
 One Microsoft Way
 Redmond, WA 98052, USA.
 E-mail: peres@microsoft.com