Bernoulli **21**(3), 2015, 1361–1385
DOI: 10.3150/14-BEJ605

Concentration inequalities for sampling without replacement

RÉMI BARDENET[1] and ODALRIC-AMBRYM MAILLARD[2]

[1]*Department of Statistics, University of Oxford, 1 South Parks Road, OX1 3TG Oxford, UK.*
E-mail: remi.bardenet@gmail.com
[2]*Faculty of Electrical Engineering, The Technion, Fishbach Building, 32000 Haifa, Israel.*
E-mail: odalric.maillard@ee.technion.ac.il

Concentration inequalities quantify the deviation of a random variable from a fixed value. In spite of numerous applications, such as opinion surveys or ecological counting procedures, few concentration results are known for the setting of sampling without replacement from a finite population. Until now, the best general concentration inequality has been a Hoeffding inequality due to Serfling [*Ann. Statist.* **2** (1974) 39–48]. In this paper, we first improve on the fundamental result of Serfling [*Ann. Statist.* **2** (1974) 39–48], and further extend it to obtain a Bernstein concentration bound for sampling without replacement. We then derive an empirical version of our bound that does not require the variance to be known to the user.

Keywords: Bernstein; concentration bounds; sampling without replacement; Serfling

1. Introduction

Few results exist on the concentration properties of sampling without replacement from a finite population $\mathcal{X}$. However, potential applications are numerous, from historical applications such as opinion surveys (Kish [9]) and ecological counting procedures (Bailey [2]), to more recent approximate Monte Carlo Markov chain algorithms that use subsampled likelihoods (Bardenet, Doucet and Holmes [3]). In a fundamental paper on sampling without replacement, Serfling [14] introduced an efficient Hoeffding bound, that is, one which is a function of the range of the population. Bernstein bounds are typically tighter when the variance of the random variable under consideration is small, as their leading term is linear in the standard deviation of $\mathcal{X}$, while the range only influences higher-order terms. This paper is devoted to Hoeffding and Bernstein bounds for sampling without replacement.

Setting and notations

Let $\mathcal{X} = (x_1, \dots, x_N)$ be a finite population of N real points. We use capital letters to denote random variables on $\mathcal{X}$, and lower-case letters for their possible values. Sampling without replacement a list $(X_1, \dots, X_n)$ of size n from $\mathcal{X}$ can be described sequentially as follows: let first $\mathcal{I}_1 = \{1, \dots, n\}$, sample an integer I_1 uniformly on $\mathcal{I}_1$, and set X_1 to be x_{I_1}. Then, for each $i = 2, \dots, n$, sample I_i uniformly on the remaining indices $\mathcal{I}_i = \mathcal{I}_{i-1} \setminus \{I_{i-1}\}$. Hereafter, we assume that $N \geq 2$.

Previous work

There have been a few papers on concentration properties of sampling without replacement; see, for instance, Hoeffding [7], Serfling [14], Horvitz and Thompson [8], McDiarmid [13]. One notable contribution is the following reduction result in Hoeffding's seminal paper (Hoeffding [7], Theorem 4):

Lemma 1.1. *Let $\mathcal{X} = (x_1, \dots, x_N)$ be a finite population of N real points, $X_1, \dots, X_n$ denote a random sample without replacement from $\mathcal{X}$ and $Y_1, \dots, Y_n$ denote a random sample with replacement from $\mathcal{X}$. If $f : \mathbb{R} \to \mathbb{R}$ is continuous and convex, then*

$$\mathbb{E} f\left(\sum_{i=1}^{n} X_i\right) \leq \mathbb{E} f\left(\sum_{i=1}^{n} Y_i\right).$$

Lemma 1.1 implies that the concentration results known for sampling with replacement as Chernoff bounds (Boucheron, Lugosi and Massart [4]) can be transferred to the case of sampling without replacement. In particular, Proposition 1.2, due to Hoeffding [7], holds for the setting without replacement.

Proposition 1.2 (Hoeffding's inequality). *Let $\mathcal{X} = (x_1, \dots, x_N)$ be a finite population of N points and $X_1, \dots, X_n$ be a random sample drawn without replacement from $\mathcal{X}$. Let*

$$a = \min_{1 \leq i \leq N} x_i \quad \text{and} \quad b = \max_{1 \leq i \leq N} x_i.$$

Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon\right) \leq \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right), \tag{1}$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the mean of $\mathcal{X}$.

The proof of Proposition 1.2 (see, e.g., Boucheron, Lugosi and Massart [4]) relies on a classical bound on the moment-generating function of a random variable, which we restate here as Lemma 1.3 for further reference.

Lemma 1.3. *Let X be a real random variable such that $\mathbb{E}X = 0$ and $a \leq X \leq b$ for some $a, b \in \mathbb{R}$. Then, for all $s \in \mathbb{R}$,*

$$\log \mathbb{E} e^{sX} \leq \frac{s^2(b-a)^2}{8}.$$

When the variance of $\mathcal{X}$ is small compared to the range $b - a$, another Chernoff bound, known as Bernstein's bound (Boucheron, Lugosi and Massart [4]), is usually tighter than Proposition 1.2.

Proposition 1.4 (Bernstein's inequality). *With the notations of Proposition 1.2, let*

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

be the variance of $\mathcal{X}$. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + (2/3)(b-a)\varepsilon}\right).$$

Although these are interesting results, it appears that the bounds in Propositions 1.2 and 1.4 are actually very conservative, especially when n is large, say, $n \geq N/2$. Indeed, Serfling [14] proved that the term n in the RHS of (1) can be replaced by $\frac{n}{1-(n-1)/N}$; see Theorem 2.4 below, where the result of Serfling is restated in our notation and slightly improved. As n approaches N, the bound of Serfling [14] improves dramatically, which corresponds to the intuition that when sampling without replacement, the sample mean becomes a very accurate estimate of μ as n approaches N.

Contributions and outline

In Section 2, we slightly modify Serfling's result, yielding a Hoeffding–Serfling bound in Theorem 2.4 that dramatically improves on Hoeffding's in Proposition 1.2. In Section 3, we contribute in Theorem 3.5 a similar improvement on Proposition 1.4, which we call a Bernstein–Serfling bound. To allow practical applications of our Bernstein–Serfling bound, we finally provide an *empirical* Bernstein–Serfling bound in Section 4, in the spirit of Maurer and Pontil [12], which does not require the variance of $\mathcal{X}$ to be known beforehand. In Section 5, we discuss direct applications and potential further improvements of our results.

Illustration

To give the reader a visual intuition of how the above mentioned bounds compare in practice and motivate their derivation, in Figure 1, we plot the bounds given by Proposition 1.2 and Theorem 2.4 for Hoeffding bounds, and Proposition 1.4 and Theorem 3.5 for Bernstein bounds for $\varepsilon = 10^{-2}$, in some common situations. We set $\mathcal{X}$ to be an independent sample of size $N = 10^4$ from each of the following four distributions: unit centered Gaussian, log-normal with parameters $(1, 1)$, and Bernoulli with parameter $1/10$ and $1/2$. An estimate of the probability $\mathbb{P}(n^{-1}\sum_{i=1}^{n} X_i - \mu \geq 10^{-2})$ is obtained by averaging over 1000 repeated samples of size n taken without replacement. In Figures 1(a), 1(b) and 1(c), Hoeffding's bound and the Hoeffding–Serfling bound of Theorem 2.4 are close for $n \leq N/2$, after which the Hoeffding–Serfling bound decreases to zero, outperforming Hoeffding's bound. Bernstein's and our Bernstein–Serfling bound behave similarly, both outperforming their counterparts that do not make use of the variance of $\mathcal{X}$. However, Figure 1(d) shows that one should not always prefer Bernstein bounds.

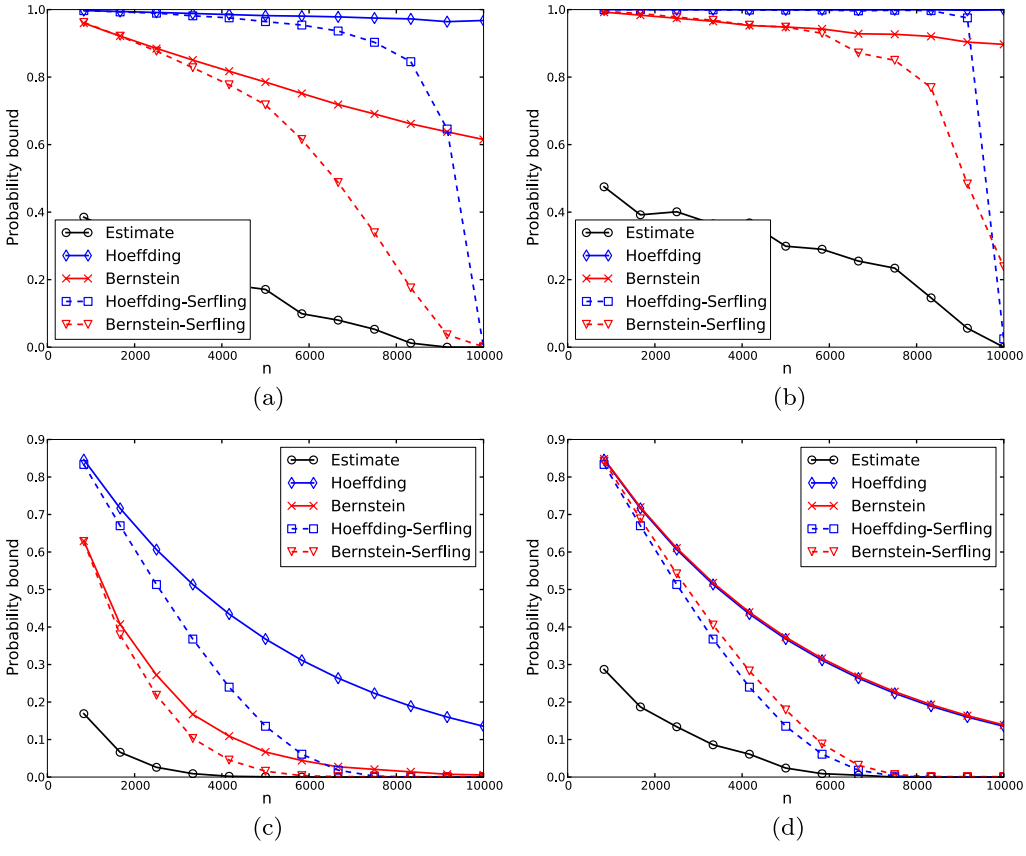

Figure 1. Comparing known bounds on $p = \mathbb{P}(n^{-1}\sum_{i=1}^{n} X_i - \mu \geq 0.01)$ with our Hoeffding–Serfling and Bernstein–Serfling bounds. $\mathcal{X}$ is here a sample of size $N = 10^4$ from each of the four distributions written below each plot. An estimate (black plain line) of p is obtained by averaging over 1000 repeated subsamples of size n, taken from $\mathcal{X}$ uniformly without replacement. (a) Gaussian $\mathcal{N}(0, 1)$. (b) Log-normal $\ln\mathcal{N}(1, 1)$. (c) Bernoulli $\mathcal{B}(0.1)$. (d) Bernoulli $\mathcal{B}(0.5)$.

In this case, the standard deviation is as large as roughly half the range, making Hoeffding's and Bernstein's bounds identical, and Hoeffding–Serfling actually slightly better than Bernstein–Serfling. We emphasize here that Bernstein bounds are typically useful when the variance is small compared to the range.

2. A reminder of Serfling's fundamental result

In this section, we recall an initial result and proof by Serfling [14], and slightly improve on his final bound.

We start by identifying the following martingales structures. Let us introduce, for $1 \leq k \leq N$,

$$Z_k = \frac{1}{k}\sum_{t=1}^{k}(X_t - \mu) \quad \text{and} \quad Z_k^\star = \frac{1}{N-k}\sum_{t=1}^{k}(X_t - \mu), \qquad \text{where } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2}$$

Note that by definition $Z_N = 0$, so that the σ-algebra $\sigma(Z_{k+1}, \ldots, Z_N)$ is equal to $\sigma(Z_{k+1}, \ldots, Z_{N-1})$.

Lemma 2.1. *The following forward martingale structure holds for* $\{Z_k^\star\}_{k\leq N}$:

$$\mathbb{E}\big[Z_k^\star | Z_{k-1}^\star, \ldots, Z_1^\star\big] = Z_{k-1}^\star. \tag{3}$$

Similarly, the following reverse martingale structure holds for $\{Z_k\}_{k\leq N}$:

$$\mathbb{E}[Z_k | Z_{k+1}, \ldots, Z_{N-1}] = Z_{k+1}. \tag{4}$$

Proof. We first prove (3). Let $1 \leq k \leq N$. We start by noting that

$$\begin{aligned} Z_k^\star &= \frac{1}{N-k}\sum_{t=1}^{k-1}(X_t - \mu) + \frac{X_k - \mu}{N-k} \\ &= \frac{N-k+1}{N-k} Z_{k-1}^\star + \frac{X_k - \mu}{N-k}. \end{aligned} \tag{5}$$

Since X_k is uniformly distributed on the remaining elements of $\mathcal{X}$ after $X_1, \ldots, X_{k-1}$ have been drawn, its conditional expectation given $X_1, \ldots, X_{k-1}$ is the average of the $N-k+1$ remaining points in $\mathcal{X}$. Since points in $\mathcal{X}$ add up to $N\mu$, we obtain

$$\begin{aligned} \mathbb{E}\big[X_k | Z_{k-1}^\star, \ldots, Z_1^\star\big] &= \mathbb{E}[X_k | X_{k-1}, \ldots, X_1] \\ &= \frac{N\mu - \sum_{i=1}^{k-1} X_i}{N-k+1} \\ &= \mu - Z_{k-1}^\star. \end{aligned} \tag{6}$$

Combined with (5), this yields (3).

We now turn to proving (4). First, let $1 \leq k \leq N$. Since

$$(k+1)Z_{k+1} = (N-k-1)\mu - X_{k+2} - \cdots - X_N,$$

$\sigma(Z_{k+1}, \ldots, Z_{N-1})$ is equal to $\sigma(X_{k+2}, \ldots, X_N)$. Now, let us remark that the indices of $(X_1, \ldots, X_N)$ are uniformly distributed on the permutations of $\{1, \ldots, N\}$, so that $(X_1, \ldots, X_{N-k})$ and $(X_{k+1}, \ldots, X_N)$ have the same marginal distribution. Consequently,

$$\mathbb{E}[X_{k+1} | Z_{k+1}, \ldots, Z_{N-1}] = \mathbb{E}[X_{k+1} | X_{k+2} \ldots, X_N] = \frac{S_{k+1}}{k+1},$$

where we introduced the sum $S_{k+1} = \sum_{t=1}^{k+1} X_t$. Finally, we prove (4) along the same lines as (3):

$$\begin{aligned}\mathbb{E}[Z_k | Z_{k+1}, \dots, Z_{N-1}] &= \mathbb{E}\left[\frac{S_k - k\mu}{k} \middle| Z_{k+1}, \dots, Z_{N-1} \right] \\ &= \mathbb{E}\left[\frac{S_{k+1} - X_{k+1}}{k} \middle| Z_{k+1}, \dots, Z_N \right] - \mu \\ &= \frac{S_{k+1}}{k} - \frac{S_{k+1}}{k(k+1)} - \mu \\ &= Z_{k+1}. \end{aligned}$$

□

A Hoeffding–Serfling inequality

Let us now state the main result of Serfling [14]. This is a key result to derive a concentration inequality, a maximal concentration inequality and a self-normalized concentration inequality, as explained in Serfling [14].

Proposition 2.2 (Serfling [14]). *Let us denote* $a = \min_{1 \le i \le N} x_i$, *and* $b = \max_{1 \le i \le N} x_i$. *Then, for any* $\lambda > 0$, *it holds that*

$$\log \mathbb{E} \exp(\lambda n Z_n) \le \frac{(b-a)^2}{8} \lambda^2 n \left(1 - \frac{n-1}{N} \right).$$

Moreover, for any $\lambda > 0$, *it also holds that*

$$\log \mathbb{E} \exp\left(\lambda \max_{1 \le k \le n} Z_k^\star \right) \le \frac{(b-a)^2}{8} \frac{\lambda^2}{(N-n)^2} n \left(1 - \frac{n-1}{N} \right).$$

Proof. First, (5) yields that for all $\lambda' > 0$,

$$\lambda' Z_k^\star = \lambda' Z_{k-1}^\star + \lambda' \frac{X_k - \mu + Z_{k-1}^\star}{N-k}. \tag{7}$$

Furthermore, we know from (6) that $-Z_{k-1}^\star$ is the conditional expectation of $X_k - \mu$ given $X_1, \dots, X_{k-1}$. Thus, since $X_k - \mu \in [a - \mu, b - \mu]$, Lemma 1.3 applies and we get that, for all $2 \le k \le n$,

$$\log \mathbb{E}\left[\exp\left(\lambda' \frac{X_k - \mu + Z_{k-1}^\star}{N-k} \right) \middle| Z_1^\star, \dots, Z_{k-1}^\star \right] \le \frac{(b-a)^2}{8} \frac{\lambda'^2}{(N-k)^2}. \tag{8}$$

Similarly, we can apply Lemma 1.3 to $Z_1^\star = (X_1 - \mu)/(N-1)$ to obtain

$$\log \mathbb{E} \exp\left(\lambda' Z_1^\star \right) \le \frac{(b-a)^2}{8} \frac{\lambda'^2}{(N-1)^2}. \tag{9}$$

Upon noting that $Z_n = \frac{N-n}{n} Z_n^\star$, and combining (8) and (9) together with the decomposition (7), we eventually obtain the bound

$$\log \mathbb{E} \exp\left(\lambda' \frac{n}{N-n} Z_n\right) \leq \frac{(b-a)^2}{8} \sum_{k=1}^{n} \frac{\lambda'^2}{(N-k)^2}.$$

In particular, for λ such that $\lambda' = (N-n)\lambda$, the RHS of this equation contains the quantity

$$\begin{aligned} \sum_{k=1}^{n} \frac{(N-n)^2}{(N-k)^2} &= 1 + (N-n)^2 \sum_{k=N-n+1}^{N-1} \frac{1}{k^2} \\ &\leq 1 + (N-n)^2 \frac{((N-1)-(N-n))}{(N-n)N} = 1 + \frac{(N-n)(n-1)}{N} \\ &= 1 + n - 1 - n\frac{n-1}{N} = n\left(1 - \frac{n-1}{N}\right), \end{aligned} \tag{10}$$

where we used in the second line the following approximation from (Serfling [14], Lemma 2.1): for $1 \leq j \leq m$, it holds

$$\sum_{k=j+1}^{l} \frac{1}{k^2} \leq \frac{l-j}{j(l+1)}.$$

This concludes the proof of the first result of Proposition 2.2. The second result follows from applying Doob's maximal inequality for martingales combined with the previous derivation. □

The result of Proposition 2.2 reveals a powerful feature of the no replacement setting: the factor $n(1 - \frac{n-1}{N})$ in the exponent, as opposed to n in the case of sampling with replacement. This leads to a dramatic improvement of the bound when n is large, as can be seen on Figure 1. Serfling [14] mentioned that a factor $1 - \frac{n}{N}$ would be intuitively more natural, as indeed when $n = N$ the mean μ is known exactly, so that Z_N is deterministically zero.

Serfling did not publish any result with $1 - \frac{n}{N}$. However, it appears that a careful examination of the previous proof and of the use of equation (4), in lieu of (3), allows us to get such an improvement. We detail this in the following proposition. More than a simple cosmetic modification, it is actually a slight improvement on Serfling's original result when $n > N/2$.

Proposition 2.3. *Let (Z_k) be defined by (2). For any $\lambda > 0$, it holds that*

$$\log \mathbb{E} \exp(\lambda n Z_n) \leq \frac{(b-a)^2}{8} \lambda^2 (n+1)\left(1 - \frac{n}{N}\right).$$

Moreover, for any $\lambda > 0$, it also holds that

$$\log \mathbb{E} \exp\left(\lambda \max_{n \leq k \leq N-1} Z_k\right) \leq \frac{(b-a)^2}{8} \frac{\lambda^2}{n^2} (n+1)\left(1 - \frac{n}{N}\right).$$

Proof. Let us introduce the notation $Y_k = Z_{N-k}$ for $1 \leq k \leq N-1$. From (4), it comes

$$\mathbb{E}[Y_{N-k}|Y_1, \dots, Y_{N-k-1}] = Y_{N-k-1}.$$

By a change of variables, this can be rewritten as

$$\mathbb{E}[Y_k|Y_1, \dots, Y_{k-1}] = Y_{k-1}.$$

Now we remark that the following decomposition holds:

$$\begin{aligned} \lambda Y_k &= \lambda \frac{\sum_{i=1}^{N-k}(X_i - \mu)}{N-k} \\ &= \lambda Y_{k-1} - \lambda \frac{X_{N-k+1} - \mu - Y_{k-1}}{N-k}. \end{aligned} \tag{11}$$

Since Y_{k-1} is the conditional mean of $X_{N-k+1} - \mu \in [a-\mu, b-\mu]$, Lemma 1.3 yields that, for all $2 \leq k \leq n$,

$$\log \mathbb{E}\left[\exp\left(\lambda' \frac{X_{N-k+1} - \mu - Y_{k-1}}{N-k}\right) \middle| Y_1, \dots, Y_{k-1}\right] \leq \frac{(b-a)^2}{8} \frac{\lambda'^2}{(N-k)^2}. \tag{12}$$

On the other hand it holds by definition of Y_1 that

$$Y_1 = Z_{N-1} = \frac{\sum_{i=1}^{N-1}(X_i - \mu)}{N-1} \in [a-\mu, b-\mu].$$

Along the lines of the proof of Proposition 2.2, we obtain

$$\log \mathbb{E} \exp\left(\lambda' Y_1\right) \leq \frac{(b-a)^2}{8} \frac{\lambda'^2}{(N-1)^2}. \tag{13}$$

Combining equations (12) and (13) with the decomposition (11), it comes

$$\begin{aligned} \log \mathbb{E} \exp\left(\lambda' Y_n\right) &\leq \frac{(b-a)^2}{8} \sum_{k=1}^{n} \frac{\lambda'^2}{(N-k)^2} \\ &\leq \frac{(b-a)^2}{8} \frac{\lambda'^2}{(N-n)^2} n \left(1 - \frac{n-1}{N}\right), \end{aligned}$$

where in the last line we made use of (10). Rewriting this inequality in terms of Z, we obtain that, for all $1 \leq n \leq N-1$,

$$\log \mathbb{E} \exp\left(\lambda(N-n) Z_{N-n}\right) \leq \frac{(b-a)^2}{8} \lambda^2 n \left(1 - \frac{n-1}{N}\right),$$

that is, by resorting to a new change of variable,

$$\begin{aligned}\log \mathbb{E}\exp(\lambda n Z_n) &\leq \frac{(b-a)^2}{8}\lambda^2(N-n)\left(1-\frac{N-n-1}{N}\right)\\ &\leq \frac{(b-a)^2}{8}\lambda^2(N-n)\frac{n+1}{N}\\ &\leq \frac{(b-a)^2}{8}\lambda^2(n+1)\left(1-\frac{n}{N}\right).\end{aligned}$$

The second part of the proposition follows from applying Doob's maximal inequality for martingales to Y_n, similarly to Proposition 2.2. □

Theorem 2.4 (Hoeffding–Serfling inequality). *Let $\mathcal{X} = (x_1, \ldots, x_N)$ be a finite population of $N > 1$ real points, and $(X_1, \ldots, X_n)$ be a list of size $n < N$ sampled without replacement from $\mathcal{X}$. Then for all $\varepsilon > 0$, the following concentration bounds hold*

$$\mathbb{P}\left(\max_{n\leq k\leq N-1}\frac{\sum_{t=1}^k (X_t-\mu)}{k}\geq\varepsilon\right)\leq\exp\left(-\frac{2n\varepsilon^2}{(1-n/N)(1+1/n)(b-a)^2}\right),$$

$$\mathbb{P}\left(\max_{1\leq k\leq n}\frac{\sum_{t=1}^k (X_t-\mu)}{N-k}\geq\frac{n\varepsilon}{N-n}\right)\leq\exp\left(-\frac{2n\varepsilon^2}{(1-(n-1)/N)(b-a)^2}\right),$$

where $a = \min_{1\leq i\leq N} x_i$ and $b = \max_{1\leq i\leq N} x_i$.

Proof. Applying Proposition 2.3 together with Markov's inequality, we obtain that, for all $\lambda > 0$,

$$\mathbb{P}\left(\max_{n\leq k\leq N-1}\frac{\sum_{t=1}^k (X_t-\mu)}{k}\geq\varepsilon\right)\leq\exp\left(-\lambda\varepsilon+\frac{(b-a)^2}{8}\frac{\lambda^2}{n^2}(n+1)(1-n/N)\right).$$

We now optimize the previous bound in λ. The optimal value is given by

$$\lambda^\star = \varepsilon\frac{4}{(b-a)^2}\frac{n^2}{(n+1)(1-n/N)}.$$

This gives the first inequality of Theorem 2.4. The proof of the second inequality follows the very same lines. □

Inverting the result of Theorem 2.4 for $n < N$ and remarking that the resulting bound still holds for $n = N$, we straightforwardly obtain the following result.

Corollary 2.5. *For all $n \leq N$, for all $\delta \in [0, 1]$, with probability higher than $1-\delta$, it holds*

$$\frac{\sum_{t=1}^n (X_t-\mu)}{n}\leq(b-a)\sqrt{\frac{\rho_n\log(1/\delta)}{2n}},$$

where we define

$$\rho_n = \begin{cases} \left(1 - \frac{n-1}{N}\right) & \text{if } n \leq N/2, \\ \left(1 - \frac{n}{N}\right)(1 + 1/n) & \text{if } n > N/2\,. \end{cases} \tag{14}$$

3. A Bernstein–Serfling inequality

In this section, we consider $\sigma^2 = N^{-1}\sum_{i=1}^N (x_i - \mu)^2$ is known, and extend Theorem 2.4 to that situation.

Similarly to Lemma 2.1, the following structural lemma will be useful:

Lemma 3.1. *It holds*

$$\mathbb{E}\big[(X_k - \mu)^2 | Z_1, \ldots, Z_{k-1}\big] = \sigma^2 - Q^\star_{k-1}, \qquad \text{where } Q^\star_{k-1} = \frac{\sum_{i=1}^{k-1}((X_i - \mu)^2 - \sigma^2)}{N - k + 1},$$

where the Z_i's are defined in (2). Similarly, it holds

$$\mathbb{E}\big[(X_{k+1} - \mu)^2 | Z_{k+1}, \ldots, Z_{N-1}\big] = \sigma^2 + Q_{k+1}, \qquad \text{where } Q_{k+1} = \frac{\sum_{i=1}^{k+1}((X_i - \mu)^2 - \sigma^2)}{k+1}.$$

Proof. We simply remark again that, conditionally on $X_1, \ldots, X_{k-1}$, the variable X_k is distributed uniformly over the remaining points in $\mathcal{X}$, so that

$$\begin{aligned} \mathbb{E}\big[(X_k - \mu)^2 | Z_1, \ldots, Z_{k-1}\big] &= \mathbb{E}\big[(X_k - \mu)^2 | X_1, \ldots, X_{k-1}\big] \\ &= \frac{1}{N - k + 1}\left[N\sigma^2 - \sum_{i=1}^{k-1}(X_i - \mu)^2\right] \\ &= \sigma^2 - Q^\star_{k-1}. \end{aligned}$$

The second equality of Lemma 3.1 follows from the same argument, as in the proof of Lemma 2.1. □

Let us now introduce the following notations:

$$\begin{aligned} \mu_{<,k+1} &= \mathbb{E}[X_{k+1} - \mu | Z_{k+1}, \ldots, Z_{N-1}], \\ \mu_{>,k} &= \mathbb{E}[X_k - \mu | Z_1, \ldots, Z_{k-1}], \\ \sigma^2_{<,k+1} &= \mathbb{E}\big[(X_{k+1} - \mu)^2 | Z_{k+1}, \ldots, Z_{N-1}\big] - \mu^2_{<,k+1}, \\ \sigma^2_{>,k} &= \mathbb{E}\big[(X_k - \mu)^2 | Z_1, \ldots, Z_{k-1}\big] - \mu^2_{>,k}. \end{aligned}$$

We are now ready to state Proposition 3.2, which is a Bernstein version of Proposition 2.2.

Proposition 3.2. *For any $\lambda > 0$, it holds that*

$$\log \mathbb{E} \exp\left(\lambda n Z_n - \lambda^2 \sum_{k=1}^{N-n} \varphi\left(\frac{2(b-a)\lambda}{N-k} \right) \frac{\sigma_{<,N-k+1}^2 n^2}{(N-k)^2} \right) \leq 0,$$

$$\log \mathbb{E} \exp\left(\lambda n Z_n - \lambda^2 \sum_{k=1}^{n} \varphi\left(2(b-a)\lambda \frac{N-n}{N-k} \right) \frac{\sigma_{>,k}^2 (N-n)^2}{(N-k)^2} \right) \leq 0,$$

where we introduced the function $\varphi(c) = \frac{\mathrm{e}^c - 1 - c}{c^2}$. Moreover, for any $\lambda > 0$, it also holds that

$$\log \mathbb{E} \exp\left(\lambda \Big(\max_{1 \leq k \leq n} Z_k^\star \Big) - \sum_{k=1}^{n} \varphi\left(\frac{2(b-a)\lambda}{N-k} \right) \frac{\sigma_{>,k}^2 \lambda^2}{(N-k)^2} \right) \leq 0,$$

$$\log \mathbb{E} \exp\left(\lambda \Big(\max_{n \leq k \leq N-1} Z_k \Big) - \sum_{k=1}^{N-n} \varphi\left(\frac{2(b-a)\lambda}{N-k} \right) \frac{\sigma_{<,N-k+1}^2 \lambda^2}{(N-k)^2} \right) \leq 0.$$

Proof. The key point is to replace equations (8) and (9) in the proof of Proposition 2.2, which make use of the range of $\mathcal{X}$, by equivalent ones that involve the variance. We only detail the proof of the first inequality, the proof of the three others follows the same steps.

A standard result from the proof of Bennett's inequality (see Lugosi [10], page 11, or Boucheron, Lugosi and Massart [4], proof of Theorem 2.9) applied to the random variable $X_{N-k+1} - \mu$, with conditional mean $\mu_{<,N-k+1}$ and conditional variance $\sigma_{<,N-k+1}^2$, yields

$$\begin{aligned} \mathbb{E}\Bigg[\exp\Bigg(\lambda' \frac{X_{N-k+1} - \mu + Y_{k-1}}{N-k} \\ - \sigma_{<,N-k+1}^2 \varphi\left(\frac{2(b-a)\lambda'}{N-k} \right) \frac{\lambda'^2}{(N-k)^2} \Bigg) \Big| Y_1, \ldots, Y_{k-1} \Bigg] \leq 1, \end{aligned} \tag{15}$$

where we used the notation $Y_k = Z_{N-k}$ of Proposition 2.3, and the function φ defined in the statement of the proposition

$$\varphi(c) = \frac{\mathrm{e}^c - 1 - c}{c^2}.$$

Similarly, Y_1 satisfies

$$\log \mathbb{E} \exp\big(\lambda' Y_1\big) = \log \mathbb{E} \exp\left(\lambda' \frac{\mu - X_N}{N-1} \right) \leq \sigma_{<,N}^2 \varphi\left(\frac{2(b-a)\lambda'}{N-1} \right) \frac{\lambda'^2}{(N-1)^2}, \tag{16}$$

where $\sigma_{<,N}^2 = \sigma^2$ is deterministic.

Thus, combining (15) and (16) together with the decomposition (11), we eventually get the bound

$$\log \mathbb{E} \exp\left(\lambda' Y_n - \sum_{k=1}^{n} \varphi\left(\frac{2(b-a)\lambda'}{N-k} \right) \frac{\sigma^2_{<,N-k+1} \lambda'^2}{(N-k)^2} \right) \leq 0. \qquad \square$$

Using the result of Proposition 3.2, we could immediately derive a simple Bernstein inequality for sampling without replacement via an application of Theorem 2.4 to the random variables $Z_i = (X_i - \mu)^2$. However, Maurer and Pontil [12] and Audibert, Munos and Szepesvári [1] showed that, in the case of sampling with replacement, a careful use of self-bounded properties of the variance yields better bounds. We now explain how to get a similar improvement on the naive Bernstein inequality in the case of sampling without replacement. We start with a technical lemma.

Lemma 3.3. *For all $\delta \in [0,1]$, with probability larger than $1-\delta$, it holds*

$$\max_{1 \leq k \leq n} \sigma^2_{>,k} \leq \sigma^2 + \frac{\sigma(b-a)(n-1)}{N-n+1} \sqrt{\frac{2\log(1/\delta)}{n-1}}. \tag{17}$$

Similarly, with probability larger than $1-\delta$, it holds

$$\max_{n \leq k \leq N-1} \sigma^2_{<,k+1} \leq \sigma^2 + \frac{\sigma(b-a)(N-n-1)}{n+1} \sqrt{\frac{2\log(1/\delta)}{N-n-1}}. \tag{18}$$

Remark 3.4. When $N \to \infty$, the upper bound on $\max_{1 \leq k \leq n} \sigma^2_{>,k}$ reduces to σ^2. Indeed, this limit case intuitively corresponds to sampling with replacement, for which the conditional variance equals σ^2.

Proof of Lemma 3.3. We first prove (17). By definition and Lemma 3.1, it holds that

$$\begin{aligned} \sigma^2_{>,k} &= \sigma^2 - Q^\star_{k-1} - {Z^\star_{k-1}}^2 \\ &\leq \sigma^2 - \frac{1}{N-k+1} \sum_{i=1}^{k-1} \big[(X_i - \mu)^2 - \sigma^2 \big]. \end{aligned} \tag{19}$$

Let $V_{k-1} = \frac{1}{k-1} \sum_{i=1}^{k-1} (X_i - \mu)^2$. Equation (19) yields

$$\max_{1 \leq k \leq n} \sigma^2_{>,k} \leq \sigma^2 + \max_{1 \leq k \leq n} \frac{k-1}{N-k+1} \big(\sigma^2 - V_{k-1} \big).$$

The rest of the proof proceeds by establishing a suitable maximal concentration bound for the quantity V_{k-1}, the mean of which is σ^2.

We remark that $-Q_{k-1}^\star = \frac{k-1}{N-k+1}(\sigma^2 - V_{k-1})$ is a martingale. Indeed, it satisfies

$$\begin{aligned}
&\mathbb{E}\big[-Q_{k-1}^\star | Q_{k-2}^\star, \dots, Q_1^\star\big] \\
&= \frac{1}{N-k+1}\mathbb{E}\left[\sum_{i=1}^{k-1}\big(\sigma^2 - (X_i-\mu)^2\big) | Q_{k-2}^\star, \dots, Q_1^\star\right] \\
&= \frac{1}{N-k+1}\sum_{i=1}^{k-2}\big(\sigma^2 - (X_i-\mu)^2\big) + \frac{1}{N-k+1}\mathbb{E}\big[\big(\sigma^2 - (X_{k-1}-\mu)^2\big) | Q_{k-2}^\star, \dots, Q_1^\star\big] \\
&= -\frac{N-k+2}{N-k+1}Q_{k-2}^\star + \frac{1}{N-k+1}Q_{k-2}^\star \\
&= -Q_{k-2}^\star,
\end{aligned}$$

where we applied Lemma 3.1 in the third line. Doob's maximal inequality thus yields that, for all $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}\Big(\max_{1\le k\le n} -Q_{k-1}^\star \ge \varepsilon\Big) &= \mathbb{P}\Big(\max_{1\le k\le n} \exp\big(-\lambda Q_{k-1}^\star\big) \ge \exp(\lambda\varepsilon)\Big) \\
&\le \mathbb{E}\big[\exp\big(-\lambda Q_{n-1}^\star - \lambda\varepsilon\big)\big] \\
&= \mathbb{E}\left[\exp\left(\lambda\frac{n-1}{N-n+1}\left(\sigma^2 - V_{n-1} - \frac{N-n+1}{n-1}\varepsilon\right)\right)\right].
\end{aligned}$$

At this point, we fix $\lambda > 0$ and apply Lemma 1.1 to the random variables $X_i' = (X_i-\mu)^2$ and function $f : x \to \exp(-\lambda(n-1)x)$. We deduce that, for all $\varepsilon' > 0$ and $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}\left(\max_{1\le k\le n} \sigma_{>,k}^2 - \sigma^2 \ge \frac{n-1}{N-n+1}\varepsilon'\right) &\le \mathbb{E}\big[\exp\big(-\lambda\big(V_{n-1} - \sigma^2 + \varepsilon'\big)\big)\big] \\
&\le \mathbb{E}\big[\exp\big(-\lambda\big(\tilde{V}_{n-1} - \sigma^2 + \varepsilon'\big)\big)\big],
\end{aligned} \tag{20}$$

where we introduced in the last line the notation $\tilde{V}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1}(Y_i-\mu)^2$, with the $\{Y_i\}_{1\le i\le n-1}$ being sampled from $\mathcal{X}$ *with replacement*. Note that $\tilde{V}_{n-1}$ has mean σ^2 too.

Now, we check that the assumptions of Theorem 13 of Maurer [11] hold. We first introduce the modification

$$\mathbf{Y}_{1:n-1}^{j,y} = \{Y_1, \dots, Y_{j-1}, y, Y_{j+1}, \dots, Y_{n-1}\}$$

of $\mathbf{Y}_{1:n-1}$, where Y_j is replaced by $y \in \mathcal{X}$. Writing $\tilde{V}_{n-1} = \tilde{V}_{n-1}(\mathbf{Y}_{1:n-1})$ to underline the dependency on the sample set $\mathbf{Y}_{1:n-1}$, it straightforwardly comes, on the one hand, that for all $y \in \mathcal{X}$

$$\begin{aligned}
\tilde{V}_{n-1}(\mathbf{Y}_{1:n-1}) - \tilde{V}_{n-1}\big(\mathbf{Y}_{1:n-1}^{j,y}\big) &= \frac{1}{n-1}\big((Y_j-\mu)^2 - (y-\mu)^2\big) \\
&\le \frac{1}{n-1}(Y_j-\mu)^2 \le \frac{1}{n-1}(b-a)^2,
\end{aligned}$$

and, on the other hand, that the following self-bounded property holds:

$$\sum_{j=1}^{n-1}\Big(\tilde{V}_{n-1}(\mathbf{Y}_{1:n-1}) - \inf_{y\in\mathcal{X}} \tilde{V}_{n-1}\big(\mathbf{Y}^{j,y}_{1:n-1}\big)\Big)^2 \leq \frac{1}{(n-1)^2}\sum_{j=1}^{n-1}(Y_j-\mu)^4 \leq \frac{(b-a)^2}{n-1}\tilde{V}_{n-1}(\mathbf{Y}_{1:n-1}).$$

We now apply of the proof of Theorem 13 of Maurer [11][1] to $Z = \frac{n-1}{(b-a)^2}\tilde{V}_{n-1}$, together with (20), which yields

$$\mathbb{P}\bigg(\max_{1\leq k\leq n}\sigma^2_{>,k} - \sigma^2 \geq \frac{(b-a)^2}{N-n+1}\varepsilon\bigg) \leq \exp\bigg(-\lambda\varepsilon + \frac{\lambda^2}{2}\mathbb{E}[Z]\bigg) = \exp\bigg(-\frac{(b-a)^2\varepsilon^2}{2(n-1)\sigma^2}\bigg),$$

where we used the same value $\lambda = \frac{\varepsilon}{\mathbb{E}[Z]} = \frac{(b-a)^2\varepsilon}{(n-1)\sigma^2}$ as in Maurer [11], Theorem 13.

Finally, we have proven that for all $\delta\in[0,1]$, with probability higher than $1-\delta$,

$$\max_{1\leq k\leq n}\sigma^2_{>,k} \leq \sigma^2 + 2\sqrt{\sigma^2}\frac{(b-a)(n-1)}{N-n+1}\sqrt{\frac{\log(1/\delta)}{2(n-1)}},$$

which concludes the proof of (17).

We now turn to proving (18). First, we remark that

$$\begin{aligned}\sigma^2_{<,k+1} &\leq \mathbb{E}\big[(X_{k+1}-\mu)^2|Z_{k+1},\dots,Z_{N-1}\big]\\ &= \mathbb{E}\big[(X_{k+1}-\mu)^2|X_{k+2},\dots,X_N\big]\\ &= \mathbb{E}\big[(Y_{N-k}-\mu)^2|Y_1,\dots,Y_{N-k-1}\big],\end{aligned}$$

where in the second line we used that $Z_{k+1} = \mu - X_N - \cdots - X_{k+2}$, and in the third line we used the change of variables $Y_u = X_{N-u+1}$. It follows that

$$\begin{aligned}\max_{n\leq k\leq N-1}\sigma^2_{<,k+1} &\leq \max_{n\leq k\leq N-1}\mathbb{E}\big[(Y_{N-k}-\mu)^2|Y_1,\dots,Y_{N-k-1}\big]\\ &= \max_{1\leq k\leq N-n}\mathbb{E}\big[(Y_k-\mu)^2|Y_1,\dots,Y_{k-1}\big].\end{aligned}$$

Now $(Y_1,\dots,Y_{N-n})$ has the same marginal distribution as $(X_1,\dots,X_{N-n})$, so that the proof of (17) applies and yields the result. □

[1]The theorem is stated for $\mathbb{P}[\mathbb{E}[Z]-Z\geq\varepsilon]$ but, actually, $\mathbb{E}[\exp(-\lambda(Z-\mathbb{E}[Z]+\varepsilon))]$ is bounded in the proof.

We emphasize that we used Hoeffding's reduction Lemma 1.1 in the proof of Lemma 3.3. This allowed us to apply the key result from Maurer [11]. We will discuss alternatives to this proof in Section 5. We can now state our Bernstein–Serfling bound.

Theorem 3.5 (Bernstein–Serfling inequality). *Let $\mathcal{X} = (x_1, \dots, x_N)$ be a finite population of $N > 1$ real points, and $(X_1, \dots, X_n)$ be a list of size $n < N$ sampled without replacement from $\mathcal{X}$. Then, for all $\varepsilon > 0$ and $\delta \in [0, 1]$, the following concentration inequality holds*

$$\mathbb{P}\left(\max_{1\le k\le n} \frac{\sum_{t=1}^{k}(X_t - \mu)}{N-k} \ge \frac{n\varepsilon}{N-n}\right) \le \exp\left[\frac{-n\varepsilon^2/2}{\gamma^2 + (2/3)(b-a)\varepsilon}\right] + \delta, \tag{21}$$

where

$$\gamma^2 = (1 - f_{n-1})\sigma^2 + f_{n-1}c_{n-1}(\delta),$$

$c_n(\delta) = \sigma(b-a)\sqrt{\frac{2\log(1/\delta)}{n}}$, *and* $f_{n-1} = \frac{n-1}{N}$. *Similarly, it holds*

$$\mathbb{P}\left(\max_{n\le k\le N-1} \frac{\sum_{t=1}^{k}(X_t - \mu)}{k} \ge \varepsilon\right) \le \exp\left[\frac{-n\varepsilon^2/2}{\tilde{\gamma}^2 + (2/3)(b-a)\varepsilon}\right] + \delta, \tag{22}$$

where

$$\tilde{\gamma}^2 = (1 - f_n)\left(\frac{n+1}{n}\sigma^2 + \frac{N-n-1}{n}c_{N-n-1}(\delta)\right).$$

Proof. We first prove (22). Applying Proposition 3.2 together with Markov's inequality, we obtain that for all $\lambda, \delta > 0$,

$$\mathbb{P}\left(\max_{n\le k\le N-1} \frac{\sum_{t=1}^{k}(X_t - \mu)}{k} \ge \frac{\log(1/\delta)}{\lambda} + \lambda \sum_{k=1}^{N-n} \varphi\left(\frac{2(b-a)\lambda}{N-k}\right)\frac{\sigma^2_{<,N-k+1}}{(N-k)^2}\right) \le \delta. \tag{23}$$

Thus, combining equations (23) and (18) with a union bound, we get that for all $\lambda > 0$ for all δ, δ', with probability higher than $1 - \delta - \delta'$, it holds that

$$\begin{aligned}
&\max_{n\le k\le N-1} \frac{\sum_{t=1}^{k}(X_t - \mu)}{k} \\
&\le \frac{\log(1/\delta)}{\lambda} + \lambda \sum_{k=1}^{N-n} \varphi\left(\frac{2(b-a)\lambda}{N-k}\right)\frac{1}{(N-k)^2}\left[\sigma^2 + \frac{N-n-1}{n+1}c_{N-n-1}\big(\delta'\big)\right] \\
&\le \frac{\log(1/\delta)}{\lambda} + \frac{\lambda}{n^2}\varphi\left(\frac{2(b-a)\lambda}{n}\right)\left[\sigma^2 + \frac{N-n-1}{n+1}c_{N-n-1}\big(\delta'\big)\right]\sum_{k=1}^{N-n}\frac{n^2}{(N-k)^2} \\
&\le \frac{\log(1/\delta)}{\lambda} + \frac{\lambda}{n^2}\varphi\left(\frac{2(b-a)\lambda}{n}\right)\left[\sigma^2 + \frac{N-n-1}{n+1}c_{N-n-1}\big(\delta'\big)\right](n+1)\left(1 - \frac{n}{N}\right),
\end{aligned}$$

where we introduced

$$c_{N-n-1}(\delta') = \sigma(b-a)\sqrt{\frac{2\log(1/\delta')}{N-n-1}},$$

where we used in the second line the fact that φ is nondecreasing and where we applied (10) in the last line. For convenience, let us now introduce the quantities $f_n = \frac{n}{N}$ and

$$\tilde{\gamma}^2 = (1-f_n)\left[\sigma^2 + \frac{N-n-1}{n+1} c_{N-n-1}(\delta')\right].$$

The previous bound can be rewritten in terms of $\varepsilon > 0$ and δ' only, in the form

$$\mathbb{P}\left(\max_{n\le k\le N-1} \frac{\sum_{t=1}^{k}(X_t-\mu)}{k} \ge \varepsilon\right) \le \exp\left(-\lambda\varepsilon + \frac{\lambda^2(n+1)}{n^2}\varphi\left(\frac{2(b-a)\lambda}{n}\right)\tilde{\gamma}^2\right) + \delta'. \quad (24)$$

We now optimize the bound (24) in λ. Let us introduce the function

$$f(\lambda) = -\lambda\varepsilon + \frac{\lambda^2(n+1)}{n^2}\varphi\left(\frac{2(b-a)\lambda}{n}\right)\tilde{\gamma}^2,$$

corresponding to the term in brackets in (24). By definition of φ, it comes

$$\begin{aligned} f(\lambda) &= -\lambda\varepsilon + \frac{\lambda^2}{n^2}\varphi\left(\frac{2(b-a)\lambda}{n}\right)\tilde{\gamma}^2(n+1) \\ &= -\lambda\varepsilon + \left(\exp\left(\frac{2(b-a)\lambda}{n}\right) - 1 - \frac{2(b-a)\lambda}{n}\right)\frac{\tilde{\gamma}^2}{4(b-a)^2}(n+1). \end{aligned}$$

Thus, the derivative of f is given by

$$f'(\lambda) = -\varepsilon + \left(\exp\left(\frac{2(b-a)\lambda}{n}\right) - 1\right)\frac{\tilde{\gamma}^2(n+1)}{2(b-a)n},$$

and the value $\lambda^\star$ that optimizes f is given by

$$\lambda^\star = \frac{n}{2(b-a)}\log\left(1 + \frac{2(b-a)\varepsilon n}{\tilde{\gamma}^2(n+1)}\right).$$

Let us now introduce for convenience the quantity $u = \frac{2(b-a)n}{\tilde{\gamma}^2(n+1)}$. The corresponding optimal value $f(\lambda^\star)$ is given by

$$\begin{aligned} f(\lambda^\star) &= -\varepsilon\frac{n}{2(b-a)}\log(1+u\varepsilon) + \frac{\tilde{\gamma}^2}{4(b-a)^2}(n+1)\big(u\varepsilon - \log(1+u\varepsilon)\big) \\ &= \frac{\tilde{\gamma}^2(n+1)}{4(b-a)^2}\big[-u\varepsilon\log(1+u\varepsilon) + u\varepsilon - \log(1+u\varepsilon)\big] \\ &= -\frac{n}{2(b-a)u}\zeta(u\varepsilon), \end{aligned}$$

where we introduced in the last line the function $\zeta(u) = (1+u)\log(1+u) - u$. Now, using the identify $\zeta(u) \geq u^2/(2+2u/3)$ for $u \geq 0$, we obtain

$$\mathbb{P}\left(\max_{n\leq k\leq N-1} \frac{\sum_{t=1}^{k}(X_t-\mu)}{k} \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon}{2(b-a)}\frac{u\varepsilon}{2+2u\varepsilon/3}\right) + \delta'$$
$$\leq \exp\left(-\frac{n\varepsilon^2}{2\tilde{\gamma}^2(n+1)/n + \frac{4}{3}(b-a)\varepsilon}\right) + \delta',$$

which concludes the proof of (22). The proof of (21) follows the very same lines, simply using (17) instead of (18). □

Inverting the bounds of Theorem 3.5, we obtain Corollary 3.6.

Corollary 3.6. *Let $n \leq N$ and $\delta \in [0,1]$. With probability larger than $1-2\delta$, it holds that*

$$\frac{\sum_{t=1}^{n}(X_t-\mu)}{n} \leq \sigma\sqrt{\frac{2\rho_n \log(1/\delta)}{n}} + \frac{\kappa_n (b-a)\log(1/\delta)}{n},$$

where we remind the definition of ρ_n (14)

$$\rho_n = \begin{cases} (1-f_{n-1}) & \text{if } n \leq N/2, \\ (1-f_n)(1+1/n) & \text{if } n > N/2, \end{cases}$$

and where we introduced the quantity

$$\kappa_n = \begin{cases} \frac{4}{3} + \sqrt{\frac{f_n}{g_{n-1}}} & \text{if } n \leq N/2, \\ \frac{4}{3} + \sqrt{g_{n+1}(1-f_n)} & \text{if } n > N/2, \end{cases} \tag{25}$$

with $f_n = n/N$ and $g_n = N/n - 1$.

Proof. Let $\delta, \delta' \in [0,1]$. From (21) in Theorem 3.5, it comes that, with probability higher than $1-\delta-\delta'$,

$$\frac{\sum_{t=1}^{n}(X_t-\mu)}{N-n} \leq \varepsilon_\delta, \qquad \text{where } \gamma^2 + B\frac{N-n}{n}\varepsilon_\delta = \frac{(N-n)^2}{2n\log(1/\delta)}\varepsilon_\delta^2,$$

where we introduced for convenience $B = \frac{2}{3}(b-a)$ and

$$\gamma^2 = (1-f_{n-1})\sigma^2 + f_{n-1}\sigma(b-a)\sqrt{\frac{2\log(1/\delta')}{n-1}}.$$

Solving this equation in ε leads to

$$\begin{aligned}\varepsilon_\delta &= n\log(1/\delta)\frac{B((N-n)/n)+\sqrt{B^2((N-n)/n)^2+4((N-n)^2/(2n\log(1/\delta)))\gamma^2}}{(N-n)^2}\\ &= \frac{1}{N-n}\Big(\sqrt{B^2\log(1/\delta)^2+2\gamma^2\log(1/\delta)n}+B\log(1/\delta)\Big)\\ &\leq \frac{n}{N-n}\left(\sqrt{\frac{2\gamma^2\log(1/\delta)}{n}}+\frac{2B\log(1/\delta)}{n}\right).\end{aligned}$$

On the other hand, following the same lines but starting from (22) in Theorem 3.5, it holds that, with probability higher than $1-\delta-\delta'$,

$$\frac{\sum_{t=1}^n(X_t-\mu)}{n}\leq\sqrt{\frac{2\tilde{\gamma}^2\log(1/\delta)}{n}}+\frac{2B\log(1/\delta)}{n},$$

where we introduced this time

$$\tilde{\gamma}^2=(1-f_n)\left((1+1/n)\sigma^2+\frac{N-n-1}{n}\sigma(b-a)\sqrt{\frac{2\log(1/\delta')}{N-n-1}}\right).$$

Finally, we note that

$$\sqrt{\tilde{\gamma}^2}\leq\sqrt{(1-f_n)(1+1/n)}\left(\sigma+\frac{N-n-1}{n+1}(b-a)\sqrt{\frac{\log(1/\delta')}{2(N-n-1)}}\right).$$

Thus, when $n\leq N/2$, we deduce that for all $1\leq n\leq N-1$, with probability higher than $1-2\delta$, it holds

$$\begin{aligned}\frac{\sum_{t=1}^n(X_t-\mu)}{n} &\leq \sqrt{1-f_{n-1}}\left(\sigma\sqrt{\frac{2\log(1/\delta)}{n}}+\frac{n-1}{N-n+1}\frac{(b-a)\log(1/\delta)}{\sqrt{n(n-1)}}\right)\\ &\quad+\frac{2B\log(1/\delta)}{n}\\ &\leq \sigma\sqrt{\frac{2(1-f_{n-1})\log(1/\delta)}{n}}+\frac{(b-a)\log(1/\delta)}{n}\left(\frac{4}{3}+\sqrt{\frac{n(n-1)}{N(N-n+1)}}\right);\end{aligned}$$

whereas when $N>n>N/2$, it holds, with probability higher than $1-2\delta$, that

$$\begin{aligned}\frac{\sum_{t=1}^n(X_t-\mu)}{n} &\leq \sqrt{(1-f_n)(1+1/n)}\left(\sigma\sqrt{\frac{2\log(1/\delta)}{n}}+\frac{N-n-1}{n+1}\frac{(b-a)\log(1/\delta)}{\sqrt{n(N-n-1)}}\right)\\ &\quad+\frac{2B\log(1/\delta)}{n}\end{aligned}$$

$$\leq \sigma\sqrt{\frac{2(1-f_n)(1+1/n)\log(1/\delta)}{n}} + \frac{(b-a)\log(1/\delta)}{n}\left(\frac{4}{3}+\sqrt{\frac{(N-n-1)(N-n)}{(n+1)N}}\right).$$

Finally we note that when $n = N$, $g_{n+1}(1-f_n) = 0$ and $\rho_n = 0$. So the bound is still satisfied. □

4. An empirical Bernstein–Serfling inequality

In this section, we derive a practical version of Theorem 3.5 where the variance σ^2 is replaced by an estimate. A natural (biased) estimator is given by

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2 = \frac{1}{n^2}\sum_{i,j=1}^n \frac{(X_i - X_j)^2}{2}, \qquad \text{where } \widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i. \tag{26}$$

We also define, for notational convenience, the quantity $\widehat{\sigma}_n = \sqrt{\widehat{\sigma}_n^2}$.

Before proving our empirical Bernstein–Serfling inequality, we first need to control the error between $\widehat{\sigma}_n$ and σ. For instance, in the standard case of sampling with replacement, it can be shown (Maurer and Pontil [12]) that, for all $\delta \in [0,1]$,

$$\mathbb{P}\left(\sigma \geq \frac{n}{n-1}\widehat{\sigma}_n + (b-a)\sqrt{\frac{2\ln(1/\delta)}{n-1}}\right) \leq \delta.$$

We now show an equivalent result in the case of sampling without replacement.

Lemma 4.1. *When sampling without replacement from a finite population $\mathcal{X} = (x_1, \ldots, x_N)$ of size N, with range $[a,b]$ and variance σ^2, the empirical variance $\widehat{\sigma}_n^2$ defined in (26) using $n < N$ samples satisfies the following concentration inequality (using the notation of Corollary 2.5)*

$$\mathbb{P}\left(\sigma \geq \widehat{\sigma}_n + (b-a)(1+\sqrt{1+\rho_n})\sqrt{\frac{\log(3/\delta)}{2n}}\right) \leq \delta.$$

Remark 4.2. We conjecture that it is possible, at the price of a more complicated analysis, to reduce the term $(1+\sqrt{1+\rho_n})$ to $\sqrt{4\rho_n}$, which would then be consistent with the analogous result for sampling with replacement in Maurer and Pontil [12]. We further discuss this technically involved improvement in Section 5.

Proof of Lemma 4.1. In order to prove Lemma 4.1, we again use Lemma 1.1, which allows us to relate the concentration of the quantity $V_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ to that of its equivalent

$$\tilde{V}_n = \tilde{V}_n(\mathbf{Y}_{1:n}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \mu)^2,$$

where the Y_is are drawn from $\mathcal{X}$ *with replacement*. Let us introduce the notation $Z = \frac{n}{(b-a)^2}\tilde{V}_n(\mathbf{Y}_{1:n})$. We know from the proof of Lemma 3.3 that Z satisfies the conditions of application of Maurer [11], Theorem 13. Let us also introduce for convenience the constant $\lambda = -\frac{\varepsilon}{\mathbb{E}[Z]} = -\frac{(b-a)^2\varepsilon}{n\sigma^2}$. Using these notations, it comes

$$\begin{aligned}
\mathbb{P}\left(\sigma^2 - V_n \geq \frac{(b-a)^2}{n}\varepsilon\right) &\leq \mathbb{E}\left[\exp\left(-\lambda\left(\frac{n}{(b-a)^2}\sigma^2 - \frac{n}{(b-a)^2}V_n - \varepsilon\right)\right)\right] \\
&\leq \mathbb{E}\big[\exp\big(-\lambda\big(\mathbb{E}[Z] - Z - \varepsilon\big)\big)\big] \\
&\leq \exp\left(\lambda\varepsilon + \frac{\lambda^2}{2}\mathbb{E}[Z]\right) \\
&= \exp\left(-\frac{(b-a)^2\varepsilon^2}{2n\sigma^2}\right).
\end{aligned}$$

The first line results of the application of Markov's inequality. The second line follows from the application of Lemma 1.1 to $X_i' = (X_i - \mu)^2$ and $f(x) = \exp(-\lambda\frac{n}{(b-a)^2}x)$. The last steps are the same as in the proof of Lemma 3.3.

So far, we have shown that, with probability at least $1-\delta$,

$$\sigma^2 - 2\sqrt{\sigma^2}(b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \leq V_n. \tag{27}$$

Let us remark that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2 = (\widehat{\mu}_n - \mu)^2,$$

that is, $V_n = (\widehat{\mu}_n - \mu)^2 + \widehat{\sigma}_n^2$. In order to complete the proof, we thus resort twice to Theorem 2.4 to obtain that, with probability higher than $1-\delta$, it holds

$$(\widehat{\mu}_n - \mu)^2 \leq (b-a)^2\frac{\rho_n\log(2/\delta)}{2n}. \tag{28}$$

Combining equations (27) and (28) with a union bound argument yields that, with probability at least $1-\delta$,

$$\begin{aligned}
\widehat{\sigma}_n^2 &\geq \sigma^2 - 2\sqrt{\sigma^2}\sqrt{(b-a)^2\frac{\log(3/\delta)}{2n}} - (b-a)^2\frac{\rho_n\log(3/\delta)}{2n} \\
&= \left(\sigma - \sqrt{(b-a)^2\frac{\log(3/\delta)}{2n}}\right)^2 - (b-a)^2(1+\rho_n)\frac{\log(3/\delta)}{2n}.
\end{aligned}$$

Finally, we obtain

$$\mathbb{P}\left(\sigma \geq \widehat{\sigma}_n + (1+\sqrt{1+\rho_n})\sqrt{(b-a)^2\frac{\log(3/\delta)}{2n}}\right) \leq \delta. \qquad \square$$

Eventually, combining Theorem 3.5 and Lemma 4.1 with a union bound argument, we finally deduce the following result.

Theorem 4.3 (An empirical Bernstein–Serfling inequality). *Let $\mathcal{X} = (x_1, \dots, x_N)$ be a finite population of $N > 1$ real points, and $(X_1, \dots, X_n)$ be a list of size $n \leq N$ sampled without replacement from $\mathcal{X}$. Then for all $\delta \in [0, 1]$, with probability larger than $1 - 5\delta$, it holds*

$$\frac{\sum_{t=1}^{n}(X_t - \mu)}{n} \leq \widehat{\sigma}_n \sqrt{\frac{2\rho_n \log(1/\delta)}{n}} + \frac{\kappa (b-a) \log(1/\delta)}{n},$$

where we remind the definition of ρ_n (14)

$$\rho_n = \begin{cases} \left(1 - \dfrac{n-1}{N}\right) & \text{if } n \leq N/2, \\ \left(1 - \dfrac{n}{N}\right)(1 + 1/n) & \text{if } n > N/2\,, \end{cases}$$

and $\kappa = \frac{7}{3} + \frac{3}{\sqrt{2}}$.

Remark 4.4. First, Theorem 4.3 has the familiar form of Bernstein bounds. The alternative definition of ρ_n guarantees that we get the best reduction out of the no replacement setting. In particular, when n is large, the factor $(1 - f_n)$ replaces $(1 - f_{n-1})$ and the corresponding factor eventually equals 0 when $n = N$, a feature that was missing in Proposition 2.2. Second, the constant κ is to relate to the constant $7/3$ in Maurer and Pontil [12], Theorem 11, for sampling with replacement.

Proof of Theorem 4.3. First, by application of Corollary 3.6, it holds for all $\delta \in [0, 1]$ that, with probability higher than $1 - 2\delta$,

$$\frac{\sum_{t=1}^{n}(X_t - \mu)}{n} \leq \sigma \sqrt{\frac{2\rho_n \log(1/\delta)}{n}} + \frac{\kappa_n (b-a) \log(1/\delta)}{n},$$

where we remind the definition of ρ_n (14)

$$\rho_n = \begin{cases} (1 - f_{n-1}) & \text{if } n \leq N/2, \\ (1 - f_n)(1 + 1/n) & \text{if } n > N/2, \end{cases}$$

and the definition of κ_n (25)

$$\kappa_n = \begin{cases} \dfrac{4}{3} + \sqrt{\dfrac{f_n}{g_{n-1}}} & \text{if } n \leq N/2, \\ \dfrac{4}{3} + \sqrt{g_{n+1}(1 - f_n)} & \text{if } n > N/2. \end{cases}$$

We then apply Lemma 4.1 to get that, with probability higher than $1-5\delta$, if $n \leq N/2$, then

$$\begin{aligned}\frac{\sum_{t=1}^n (X_t-\mu)}{n} &\leq \sqrt{\widehat{\sigma}_n^2}\sqrt{\frac{2\log(1/\delta)}{n}}\sqrt{1-f_{n-1}} \\ &\quad + \frac{(b-a)\log(1/\delta)}{n}\left(\frac{4}{3} + \sqrt{\frac{f_n}{g_{n-1}}}\right. \\ &\quad \left. + (1+\sqrt{2-f_{n-1}})\sqrt{1-f_{n-1}}\right),\end{aligned} \tag{29}$$

and if $n > N/2$, then

$$\begin{aligned}&\frac{\sum_{t=1}^n (X_t-\mu)}{n} \\ &\leq \sqrt{\widehat{\sigma}_n^2}\sqrt{\frac{2\log(1/\delta)}{n}}\sqrt{(1-f_n)(1+1/n)} \\ &\quad + \frac{(b-a)\log(1/\delta)}{n}\left(\frac{4}{3} + \sqrt{g_{n+1}(1-f_n)}\right. \\ &\quad \left. + \sqrt{(1-f_n)(1+1/n)}\left(1+\sqrt{1+(1-f_n)(1+1/n)}\right)\right).\end{aligned} \tag{30}$$

We now simplify this result. Assume first that $n \leq N/2$. We thus get

$$\frac{f_n}{g_{n-1}} \leq \frac{1}{2g_{n-1}} = \frac{n-1}{2(N-n+1)} \leq \frac{1}{2},$$

so that we deduce

$$\frac{4}{3} + (1+\sqrt{2-f_{n-1}})\sqrt{1-f_{n-1}} + \sqrt{\frac{f_n}{g_{n-1}}} \leq 2 + \frac{1}{3} + \sqrt{2} + \frac{1}{\sqrt{2}}. \tag{31}$$

Assume now that $n > N/2$. In this case, it holds

$$\begin{aligned}g_{n+1}(1-f_n) &= \frac{N-n-1}{n+1}\frac{N-n}{N} \leq \frac{N-n}{N} \leq \frac{1}{2}, \\ (1-f_n)(1+1/n) &= \left(1-\frac{n}{N}\right)(1+1/n) \leq \frac{1}{2}\left(1+\frac{2}{N}\right),\end{aligned}$$

so that we deduce, since $N \geq 2$,

$$\frac{4}{3} + \sqrt{g_{n+1}(1-f_n)} + \sqrt{(1-f_n)(1+1/n)}(1+\sqrt{2-f_{n-1}}) \leq 2 + \frac{1}{3} + \frac{1}{\sqrt{2}} + \sqrt{2}. \tag{32}$$

Respectively combining (31) and (32) with equations (29) and (30) concludes the proof. □

5. Discussion

In this section, we discuss the bounds of Theorem 3.5 and Theorem 4.3 from the perspective of both theory and application.

First, both bounds involve either the factor $1 - f_{n-1}$ or $1 - f_n$, thus leading to a dramatic improvement on the usual Bernstein or empirical Bernstein bounds, which do not make use of the no replacement setting. This is crucial, for instance, when the user needs to rapidly compute an empirical mean from a large number of samples up to some precision level. To better understand the improvement of Serfling bounds, we plot in Figure 2 the bounds of Corollaries 2.5 and 3.6, and Theorem 4.3 for an example where $\mathcal{X}$ is a sample of size $N = 10^6$ from each of the following four distributions: unit centered Gaussian, log-normal with parameters (1, 1), and Bernoulli with parameter $1/10$ and $1/2$. As n increases, we keep sampling without replacement

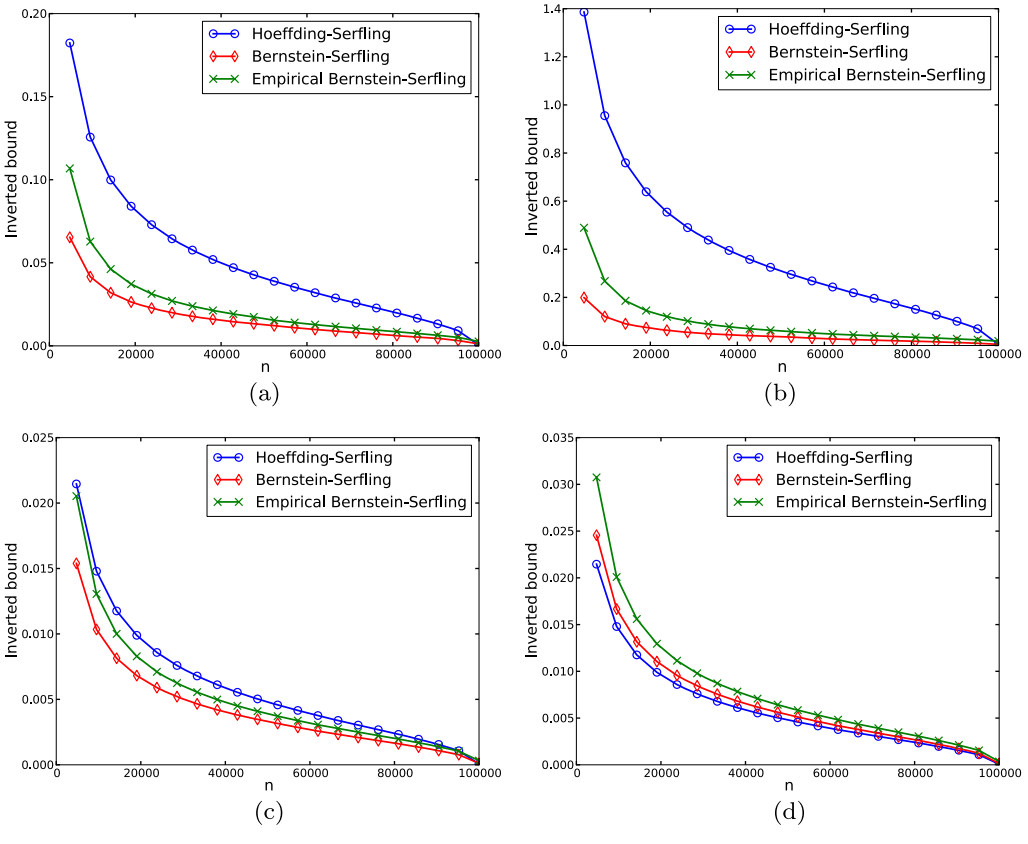

Figure 2. Comparing the bounds of Corollaries 2.5 and 3.6, and Theorem 4.3. $\mathcal{X}$ is here a sample from each of the four distributions written below each plot, of size $N = 10^6$. Unlike Figure 1, as n increases, we keep sampling here without replacement until exhaustion. (a) Gaussian $\mathcal{N}(0, 1)$. (b) Log-normal $\ln \mathcal{N}(1, 1)$. (c) Bernoulli $\mathcal{B}(0.1)$. (d) Bernoulli $\mathcal{B}(0.5)$.

from $\mathcal{X}$ until exhaustion, and report the corresponding bounds. Note that all our bounds have their leading term exactly equal to zero when $n = N$, though our Hoeffding–Serfling bound only is exactly zero. In all experiments, the loss of tightness as a result of using the empirical variance is small. Our empirical Bernstein–Serfling demonstrates here a dramatic improvement on the Hoeffding–Serfling bound of Corollary 2.5 in Figures 2(a) and 2(b). A slight improvement is demonstrated in Figure 2(c) where the standard deviation of $\mathcal{X}$ is roughly a third of the range. Finally, Bernstein–Serfling itself does not improve on Hoeffding–Serfling in Figure 2(d), where the standard deviation is roughly half of the range, again indicating that Bernstein bounds are not uniformly better than Hoeffding bounds.

There is a number of nontrivial applications of our bounds. *Scratch games*, for instance, were introduced in Féraud and Urvoy [6] as a variant of the multi-armed bandit problem, to model two real world problems: selecting ads to display on web pages and optimizing e-mailing campaigns. In particular, Féraud and Urvoy [6] discuss practical situations where an upper confidence bound algorithm based on a Hoeffding–Serfling inequality outperforms a standard algorithm based on Hoeffding's inequality. Similar improvements should appear in practice when using our empirical Bernstein–Serfling inequality. As another application, our results could be useful in optimization. The stochastic dual-coordinate ascent algorithm (SDCA; Shalev-Shwartz and Zhang [15]) is a state-of-the-art optimization algorithm used in machine learning. Shalev-Shwartz and Zhang [15] introduce a variant of SDCA called SDCA-Perm, which – unlike SDCA – relies on sampling without replacement, and achieves better empirical performance than SDCA. However, the analysis in Shalev-Shwartz and Zhang [15] does not cover SDCA-Perm. We believe that the use of Serfling bounds is an appropriate tool for that purpose.

To conclude, we discuss potential improvements of our bounds. A careful look at Lemmas 3.3 and 4.1 indicates that our bounds may be further improved, though at the price of a more intricate analysis. Indeed, these two lemmas both resort to Hoeffding's reduction Lemma 1.1, in order to be able to apply concentration results known for self-bounded random variables to the setting of sampling without replacement. As a result, we lose here a potential factor ρ_n for the confidence bound around the variance, and we conjecture that the term $1 + \sqrt{1 + \rho_n}$ in Lemma 4.1 could ultimately be replaced with $2\sqrt{\rho_n}$. A natural tool for this would be a dedicated *tensorization inequality* for the entropy in the case of sampling without replacement (Boucheron, Lugosi and Massart [4], Maurer [11], Bousquet [5]). Indeed, it is not difficult to show that $\widehat{\sigma}_n^2$ satisfies a self-bounded property similar to that of Maurer and Pontil [12], Theorem 11, involving the factor ρ_n. Thus, in order to be able to get a version of Maurer and Pontil [12], Theorem 11, in our setting, a specific so-called tensorization inequality would be enough. Unfortunately, we are unaware of the existence of such an inequality for sampling without replacement, where the samples are strongly dependent. We are also unaware of any tensorization inequality designed for U-statistics, which could be another possible way to get the desired result. Although we believe this is possible, developing such tools goes beyond the scope of this paper, and the current results of Theorem 3.5 and Theorem 4.3 are already appealing without resorting to further technicalities, which would only affect second-order terms in the end.

Acknowledgements

This work was supported by both the *2020 Science* program, funded by EPSRC grant number EP/I017909/1, and the Technion.

References

[1] Audibert, J.-Y., Munos, R. and Szepesvári, Cs. (2009). Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoret. Comput. Sci.* **410** 1876–1902. MR2514714
[2] Bailey, N.T.J. (1951). On estimating the size of mobile populations from recapture data. *Biometrika* **38** 293–306.
[3] Bardenet, R., Doucet, A. and Holmes, C. (2014). Towards scaling up MCMC: An adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning (ICML). JMLR W&CP* **32** 405–413. Brookline, MA: Microtome Publishing.
[4] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. With a foreword by Michel Ledoux. MR3185193
[5] Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic Inequalities and Applications*. *Progress in Probability* **56** 213–247. Basel: Birkhäuser. MR2073435
[6] Féraud, R. and Urvoy, T. (2013). Exploration and exploitation of scratch games. *Machine Learning* 1–25.
[7] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363
[8] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460
[9] Kish, L. (1965). *Survey Sampling*. New York: Wiley.
[10] Lugosi, G. (2009). Concentration-of-measure inequalities. Lecture notes. Available at www.econ.upf.edu/~lugosi/anu.pdf.
[11] Maurer, A. (2006). Concentration inequalities for functions of independent variables. *Random Structures Algorithms* **29** 121–138. MR2245497
[12] Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample-variance penalization. In *COLT* 2009 – *The* 22*nd Conference on Learning Theory*, *Montreal*, *Quebec*, *Canada*, *June* 18–21, 2009 1–9. Available at http://www.cs.mcgill.ca/~colt2009/papers/012.pdf.
[13] McDiarmid, C. (1997). Centering sequences with bounded differences. *Combin. Probab. Comput.* **6** 79–86. MR1436721
[14] Serfling, R.J. (1974). Probability inequalities for the sum in sampling without replacement. *Ann. Statist.* **2** 39–48. MR0420967
[15] Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *J. Mach. Learn. Res.* **14** 567–599. MR3033340

Received September 2013 and revised January 2014