 doi:10.20944/preprints201901.0202.v1

Stratified Finite Empirical Bernstein Sampling

Mark Alexander Burgess* **and Archie C. Chapman**

Australian National University
College of Engineering and Computer Science
Canberra ACT 2600, Australia
e-mail: markburgess1989@gmail.com

The University of Sydney
School of Electrical and
Information Engineering
Sydney NSW 2006, Australia
e-mail: archie.chapman@sydney.edu.au

Abstract: We derive a concentration inequality for the uncertainty in stratified random sampling. Minimising this inequality leads to an iterated online method for choosing samples from the strata. The inequality is versatile and considers a range of factors including: the data ranges, weights, sizes of the strata, as well as the number of samples taken, the estimated sample variances and whether strata are sampled with or without replacement. We evaluate the improvement this method reliably offers against other methods over sets of synthetic data, and also in approximating the Shapley value of cooperative games. The method is seen to be competitive with the performance of perfect Neyman sampling, even without prior information on strata variances. We supply a multidimensional extension of our inequality and discuss some future applications.

MSC 2010 subject classifications: 94A20 91A12 60E15.
Keywords and phrases: Concentration Inequality, Empirical Bernstein Bound, Stratified Random Sampling, Shapley Value Approximation.

1. Introduction

Stratified sampling is a statistical method of estimating the mean of a population by dividing it into mutually exclusive subgroups (or 'strata') and applying a sampling estimator to each stratum, before weighting these estimates to form an estimate of the population mean. If the stratum sampling estimator is simple random sampling, then the resulting stratified sampling is called 'stratified random sampling'.

As an example: if we want to poll how much a county's population supports a particular government policy, it may make good sense to selectively poll the different voting blocks within the country. For instance, if we accurately estimate that blocks A, B and C, containing 10%, 40% and 50% of the population, show

*A great thanks to Sylvie Thiébaux and Paul Scott for academic advice, encouragement and support!

support of 2%, 70% and 30%, respectively, then we can reliably estimate that 43.2% of the total population supports the policy.

Stratifying the sampling in this way can lead to improved reliability in estimation especially under certain conditions, such as when: the population is easily divided into strata, in which there is less variance in each than across them; when the size of the strata are reasonably or accurately known or knowable, and; when it is readily possible to sample selectively between the strata, as considered by Neyman (1938); Wright (2012). If it is possible to sample selectively between the strata, then there is a further question of how to conduct that selection most effectively.

In this paper we propose a process of sampling in order to maximally reduce the uncertainty in the population estimate, and to do this we develop a expression associated with that uncertainty. The expression takes the form of a *concentration inequality*, developed under the assumption that the data values have bounded support. This inequality considers factors such as: the sizes of all the strata and the proportion of each that are sampled, the sample variances of the samples from each of the strata, the differences in the potential ranges of data values between the strata, any additional weightings between the strata, and whether any (or all) of the strata are sampled with or without replacement.

We then propose an online method of sampling in order to maximally-reduce this inequality in each iteration. Such a sampling method has applications in selectively sampling from real-world data sets, and moreover, it can also assist in computational tasks. Particularly computational tasks that involve the calculation of expectation values, as sampling is a straightforward way of approximating such values. We consider the calculation of the Shapley Value (a solution concept from cooperative game theory) as a task to which we can apply our method. And we use the calculation of the Shapley value as an example to demonstrate our technique.

The remainder of the paper is divided into the following sections:

- Section 2 reviews the background material and gives the context for the paper,
- Section 3 provides several lemmas that form the components of our derivation,
- In Section 4 we derive our concentration inequality, which is the main technical contribution of the paper,
- Section 5 evaluates the effectiveness of minimising our inequality as an online sampling method, in the context of synthetic data.
- Section 6 we introduce and evaluate the effectiveness of approximating the Shapley value via our method,
- Section 7 discusses the results and the reasons for the effectiveness of our method,
- Section 8 gives an easy extension of our method to multidimensional data, and
- in Section 9 we conclude by hinting at some future applications.

2. Background

Stratified sampling is a well known sampling technique in statistics and research, with many applications, including polling (Hillson et al., 2015), auditing (Stark, 2009; Miratrix and Stark, 2009) and medical trials (Hu, Cai and Zeng, 2014; Prentice, 1986; Borgan et al., 2000).

In practice, stratified sampling is often done as a two-stage process, particularly when it is unclear what variables the population should be stratified by, and how large the resultant strata would be. In the first stage, the population is sampled uniformly at random, and the values of readily observable auxiliary variables are collected in order to estimate the sizes of potential strata by those variables. In the second stage, the strata are sampled with respect to the information gathered in the first stage, and the total population estimate is computed; for example, see Legg and Fuller (2009).

One well-known, but basic estimator of strata size is the *Horvitz-Thompson estimator* (Horvitz and Thompson, 1952). This estimator is sometimes seen to perform quite badly in practice, as identified by Saegusa and Wellner (2013); Breslow, Hu and Wellner (2015). however, even despite such an estimator, there is the secondary problem of how to optimally break the population into strata based on the values of the auxiliary variables, identified and addressed by Hillson et al. (2015); Khan, Ahmad and Khan (2009); Kozak (2004).

However, in other situations, the strata and their sizes are naturally given, or the first stage may be assumed to have been conducted ideally. Nonetheless, even in that case there exists a further problem of how to allocate the second-stage samples between the strata; for instance, one could choose to sample equally between strata, proportional to the sizes of the strata, or proportional to the variance of the strata. The last option is often considered in theory and practice, and is called *Neyman allocation* (sometimes called 'optimum' allocation) (Neyman, 1938; Wright, 2012). This approach involves knowledge of the variances of the strata, which in practice can often only be estimated, either as a prior step or as the sampling proceeds (Étoré and Jourdain, 2010; O'Brien, Gamal and Rajagopal, 2015).

However, even once the samples are taken from the various strata, there is yet another question of how to compute appropriate confidence bounds on the final estimate. In the voting verification context, there exist some specialised bounds (for instance see Miratrix and Stark (2009); Bentkus and van Zuijlen (2003)), but in the more general case there is some degree of discussion of what bounds should be used, as considered by Stark (2009). The confidence bounds that are derived critically depend on what assumptions are made about the underlying populations. For instance, Hoeffding's inequality (Hoeffding, 1963) has been used variously as such a bound under the assumption that the underlying population has bounded data values (Bentkus and van Zuijlen, 2003; Stark, 2009). Hoeffding's inequality can be used to produce a very conservative confidence interval that is sensitive only to the width of the data value bounds and the number of samples taken, and it is also most directly suitable for sampling with replacement. In contrast, other concentration inequalities, such as

Chebychev's inequality, are sensitive to the sample variance but not the width of the data.

Recently, Maurer and Pontil (2009) developed a bound which they named as an *Empirical Bernstein Bound* (EBB) as a concentration of measure for the sample mean of a single (unstratified) population, which is sensitive to the data width and sample variance (some similar bounds being published around that time, Audibert, Munos and Szepesvári (2009); Audibert, Munos and Szepesvári (2007)). EBBs have replaced Hoeffding's inequality in a number of computational applications (Rehman, Li and Li, 2012; Mnih, Szepesvári and Audibert, 2008; Thomas, Theocharous and Ghavamzadeh, 2015; Carpentier et al., 2011). The derivation of the particular EBB in Maurer and Pontil (2009) extended entropic (Maurer, 2006) and Chernoff concentration inequalities, bound together using union bounds.

Beyond this, sampling *without replacement* offers the opportunity to further sharpen bounds over the sampling-with-replacement case. For example, the refinement that is possible was first demonstrated by Serfling (1974) with a martingale argument. More recently, Bardenet and Maillard (2015) improved on Serfling's result with a reverse martingale argument, and created an EBB suitable to the case of sampling without replacement.

Our key observation is that the components of these analyses can be combined together to create a closed form analytical concentration inequality tailored for stratified random sampling, which is the subject of this paper.

3. Components of the Bound

In this section, we provide several useful lemmas, which we combine in the derivation of our concentration inequality. The first is an often used and weak result between statements of probability:

Lemma 1 (Probability Union)**.** *For any random variables a, b, c, the following holds:* $\mathbb{P}(a > c) \leq \mathbb{P}(a > b) + \mathbb{P}(b > c)$

Proof. For any events A and B, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ hence:

$$\mathbb{P}\left((a > b) \cup (b > c)\right) \leq \mathbb{P}(a > b) + \mathbb{P}(b > c).$$

If $a > c$, then $(a > b) \cup (b > c)$ is true irrespective of b, so:

$$\mathbb{P}(a > c) \leq \mathbb{P}\left((a > b) \cup (b > c)\right). \qquad \square$$

This relationship gives us a useful tool for settings where $\mathbb{P}(a > c)$ is unknown but the relationship between a and some b, and also between b and c is known.

The next lemma is a quick result that relates the sample squares about the mean and the sample variance.

Lemma 2 (Variance Relation)**.** *For n samples x_i of a random variable X with mean μ, sample mean $\hat{\mu} = \frac{1}{n}\sum_i x_i$, biased sample variance $\hat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \hat{\mu})^2$, and average of sample squares about the mean $\hat{\sigma}_0^2 = \frac{1}{n}\sum_i (x_i - \mu)^2$, are related such that: $\hat{\sigma}_0^2 - \hat{\sigma}^2 = (\hat{\mu} - \mu)^2$*

Proof. By definition:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i \left(x_i - \frac{1}{n}\sum_j x_j\right)^2 = \frac{1}{n}\sum_i x_i^2 - \frac{1}{n^2}\sum_i \sum_j x_i x_j$$

and:

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_i (x_i - \mu)^2 = \frac{1}{n}\sum_i x_i^2 - \frac{2\mu}{n}\sum_i x_i + \mu^2$$

therefore:

$$\hat{\sigma}_0^2 - \hat{\sigma}^2 = \frac{1}{n^2}\sum_i \sum_j x_i x_j - \frac{2\mu}{n}\sum_i x_i + \mu^2 = \left(\frac{1}{n}\sum_j x_j - \mu\right)^2 \qquad \square$$

This result is used later to create bounds for the sample variance from bounds on the sample squares about the mean. In order to create such probability bounds, we make repeated use of the next lemma, which encapsulates a range of inequalities called Chernoff bounds:

Lemma 3 (Chernoff Bound)**.** *For a random variable X then for any $s > 0$ and t that:* $\mathbb{P}(X \geq t) \leq \mathbb{E}\left[\exp(sX)\right]\exp(-st)$

Proof. $\mathbb{P}(X \geq t) = \mathbb{P}\left(\exp(sX) \geq \exp(st)\right) \leq \mathbb{E}\left[\exp(sX)\right]\exp(-st)$ by Markov's inequality. $\square$

Many well-known inequalities follow from upper bounds for $\mathbb{E}\left[\exp(sX)\right]$, also known as the *moment generating function*. The next three lemmas give three of these upper bounds for moment generating functions, from which we create our probability inequalities of interest. The first is well known and sometimes called "Hoeffding's Lemma" (Hoeffding, 1963) and is stated here without proof:

Lemma 4 (Hoeffding's Lemma)**.** *For a random variable X that is bounded on an interval $a \leq X \leq b$ with $D = b - a$ that:*

$$E\left[\exp(s(X - \mathbb{E}[x]))\right] \leq \exp\left(\frac{1}{8}D^2 s^2\right)$$

The second is very much like Hoeffding's Lemma, except it involves information about the variance of the random variable:

Lemma 5. *For a random variable X that is bounded on an interval $a \leq X \leq b$ with $D = b - a$ and variance σ^2 that:*

$$E\left[\exp(s(X - \mathbb{E}[x]))\right] \leq \exp\left(\left(\frac{D^2}{17} + \frac{\sigma^2}{2}\right)s^2\right)$$

Proof. We assume without loss of generality (and for ease of presentation) that X is centered to have a mean of zero. Then we construct an upper bound for $\mathbb{E}\left[\exp(sX)\right]$ in terms of D by a parabola over $\exp(sX)$ for the permitted values of X.

For α, β, γ such that $\alpha s^2 X^2 + \beta s X + \gamma \geq \exp(sX)$ for all $a \leq X \leq b$:

$$E\left[\exp(sX)\right] \leq \mathbb{E}[\alpha s^2 X^2 + \beta s X + \gamma] = \alpha s^2 \mathbb{E}[X^2] + \gamma = \alpha s^2 \sigma^2 + \gamma$$

 doi:10.20944/preprints201901.0202.v1

Choosing α, β, γ to minimize the above expression (see appendix) leads to:

$$E\left[\exp(sX)\right] \leq \left(\frac{\sigma^2}{b^2}\exp\left(s\left(b+\frac{\sigma^2}{b}\right)\right)+1\right)\exp\left(-\frac{s\sigma^2}{b}\right)\left(\frac{\sigma^2}{b^2}+1\right)^{-1}.$$

This expression is monotonically increasing with b, therefore using the fact that $D > b$ and rearranging:

$$\log(E\left[\exp(sX)\right]) \leq \log\left(\frac{\sigma^2}{D^2}\exp\left(s\left(D+\frac{\sigma^2}{D}\right)\right)+1\right)-\frac{s\sigma^2}{D}-\log\left(\frac{\sigma^2}{D^2}+1\right) \tag{1}$$

Given that:

$$\log(a\exp(x)+1) \leq \log(a+1)+\frac{xa}{a+1}+x^2\frac{\frac{1}{17}+\frac{a}{2}}{(a+1)^2} \tag{2}$$

then:

$$\log(E\left[\exp(sX)\right]) \leq \left(\frac{D^2}{17}+\frac{\sigma^2}{2}\right)s^2 \tag{3}$$

□

We note that the derivation process of fitting a parabola over the exponential function was indirectly also conducted by Hoeffding (1963) and Bennett (1962). Our result here is a weakening of theirs, which is more tractable for manipulation in our subsequent algebra.

The third bound on the moment generating function is similar again, however this time we consider the random variable X^2 instead of X. These three bounds (lemmas 4,5 and 6) are folded into the derivation of our stratified sampling concentration inequality in the next Section 4.

Lemma 6. *Let X be a random variable that is bounded on an interval $a \leq X \leq b$ with $D = b - a$ and variance $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[x])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Then:*

$$\mathbb{E}[\exp(q(\sigma^2-(X-\mathbb{E}[X])^2))] \leq \exp\left(\frac{1}{2}\sigma^2q^2D^2\right)$$

Proof. We assume without loss of generality (and for ease of presentation) that X is centered to have a mean of zero. We construct an upper bound for $E\left[\exp(-qX^2)\right]$ in terms of D by a parabola over $\exp(-qX^2)$ for the permitted values of X.

For α, γ such that $\alpha qX^2 + \gamma \geq \exp(-qX^2)$ then:

$$\mathbb{E}[\exp(-qX^2)] \leq \alpha q\sigma^2 + \gamma.$$

Choosing an α, γ to minimize this expression irrespective of a, b consistent with a D, gives $\gamma = 1$ and $\alpha = (\exp(-qD^2) - 1)(qD^2)^{-1}$. Thus:

$$\begin{aligned}\mathbb{E}[\exp(-qX^2)] &\leq \frac{\sigma^2}{D^2}\exp(-qD^2)-\frac{\sigma^2}{D^2}+1\\ &\leq \exp\left(\log\left(\frac{\sigma^2}{D^2}\exp(-qD^2)-\frac{\sigma^2}{D^2}+1\right)\right)\end{aligned}$$

 doi:10.20944/preprints201901.0202.v1

Given that: $\log\left(a\exp(x) - a + 1\right) < ax + \frac{1}{2}a(1-a)x^2$ for negative x:

$$\mathbb{E}[\exp(-qX^2)] \leq \exp\left(\frac{1}{2}\sigma^2q^2(D^2-\sigma^2) - \sigma^2q\right) \leq \exp\left(\frac{1}{2}\sigma^2q^2D^2 - \sigma^2q\right),$$

and the result follows by multiplying by $\exp(q\sigma^2)$ □

These three inequalities on the moment generating function are used to create desirable probability inequalities in our derivation. However, in order to use them we needed an inequality relating the moment generating function of a random variable, with the moment generating function of the average of samples of that random variable. To do this we introduce two inequalities, the first one (lemma 7) is most appropriate for sampling with replacement, and the second (lemma 9) can optionally be used in the context of sampling without replacement.

Lemma 7 (Replacement Bound)**.** *Let X be a random variable that is bounded $a \leq X \leq b$ with a mean of zero, with $D = b - a$ and variance σ^2. Let $\Xi_m = \frac{1}{m}\sum_{i=1}^{m} X_i$ be the average of m independently drawn (with replacement) samples of this random variable. If there exists an $\alpha, \beta \geq 0$ such that for any $s > 0$ that $\mathbb{E}[\exp(sX)] \leq \exp((\alpha D^2 + \beta\sigma^2)s^2)$ then:*

$$\mathbb{E}[\exp(s\Xi_m)] \leq \exp(\alpha s^2 D^2 \tfrac{1}{m} + \beta s^2\sigma^2\tfrac{1}{m}) = \exp((\alpha D^2\Omega_m^n + \beta\sigma^2\Psi_m^n)s^2)$$

where $\Omega_m^n = \Psi_m^n = \frac{1}{m}$

Proof. By the independence of samples, we have:

$$\mathbb{E}[\exp(s\Xi_m)] = \mathbb{E}\left[\exp\left(\frac{s}{m}\sum_{i=1}^{m} X_i\right)\right] = \prod_{i=1}^{m}\mathbb{E}\left[\exp\left(\frac{s}{m}X\right)\right]$$

Thus:

$$\mathbb{E}[\exp(s\Xi_m)] \leq \exp\left(\sum_{i=1}^{m}\left(\alpha D^2 + \beta\sigma^2\right)\frac{s^2}{m^2}\right)$$ □

These inequalities are sufficient for all the further derivations that we conduct. However, for the case of sampling without replacement, there is an alternative and directly substitutable result, lemma 9, which can be somewhat sharper. We give its form and derivation in the next subsection, which is included for completeness but is not part of the main results presented in the paper.

Before this, particular note must be made that the inequality above, lemma 7 can be used in the context of either sampling with or without replacement. In contrast, the inequality in the next subsection can only be used when sampling without replacement. This distinction was shown to be true by Hoeffding (1963), and is stated here without proof:

Lemma 8 (Hoeffding's reduction)**.** *let $X = (x_1, \ldots, x_n)$ be a finite population of n real points, let $X_1, \ldots, X_n$ denote a random sample without replacement from X and $Y_1, \ldots, Y_n$ denote a random sample with replacement from X. If $f : \mathbb{R} \to \mathbb{R}$ is continuous and convex, then:*

$$\mathbb{E}\left[f\left(\textstyle\sum_{i=1}^{m} X_i\right)\right] \leq \mathbb{E}\left[f\left(\textstyle\sum_{i=1}^{m} Y_i\right)\right]$$

3.1. Preliminary results for sampling without replacement

In this subsection we state an inequality regarding the moment generating function of the average of samples taken *without replacement*.

When the sampling takes place without replacement the inequality of lemma 7 can potentially be improved to take advantage of the finiteness of the data set. This inequality extends an important martingale inequality from Bardenet and Maillard (2015):

Lemma 9 (Martingale Bound)**.** *For finite data $x_1, x_2, \dots x_n$ that is bounded $a \leq x_i \leq b$, and has a mean of zero and variance $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i$, denote $X_1, X_2, \dots, X_n$ the random variables corresponding to the data sequentially drawn randomly without replacement, and Ξ_m the average of the first m of them. If for any random variable Z with a mean of zero such that $a \leq Z \leq b$ and $D = b - a$, with variance σ_Z^2 that there exists an $\alpha, \beta \geq 0$ such that for any $s > 0$ that $\mathbb{E}[\exp(sZ)] \leq \exp((\alpha D^2 + \beta\sigma_Z^2)s^2)$ then:*

$$\begin{aligned}\mathbb{E}[\exp(s\Xi_m)] &\leq \exp\left(\alpha s^2 D^2 \sum_{k=m}^{n-1} \frac{1}{k^2} + \beta s^2\sigma^2 \sum_{k=m}^{n-1} \frac{n}{k^2(k+1)}\right) \\ &\leq \exp((\alpha D^2 \bar{\Omega}_m^n + \beta\sigma^2 \bar{\Psi}_m^n)s^2)\end{aligned}$$

where $\bar{\Omega}_m^n = \sum_{k=m}^{n-1} \frac{1}{k^2} \approx \frac{(m+1)(1-m/n)}{m^2}$ *and* $\bar{\Psi}_m^n = \sum_{k=m}^{n-1} \frac{n}{k^2(k+1)} \approx \frac{n+1-m}{m^2}$.

Proof. Observe that:

$$\begin{aligned}\Xi_m &= \frac{1}{m}\sum_{i=1}^{m} X_i = \Xi_{m+1} + \frac{1}{m}(\Xi_{m+1} - X_{m+1}) \\ &= (\Xi_m - \Xi_{m+1}) + (\Xi_{m+1} - \Xi_{m+2}) + \cdots + (\Xi_{n-1} - \Xi_n) \\ &= \frac{1}{m}(\Xi_{m+1} - X_{m+1}) + \frac{1}{m+1}(\Xi_{m+2} - X_{m+2}) + \cdots + \frac{1}{n-1}(\Xi_n - X_n).\end{aligned}$$

Then because:

$$\exp(s\Xi_m) = \prod_{k=m}^{n-1} \exp\left(\frac{s}{k}(\Xi_{k+1} - X_{k+1})\right),$$

we also have that:

$$\mathbb{E}[\exp(s\Xi_m)] = \mathbb{E}\left[\prod_{k=m}^{n-1} \mathbb{E}\left[\exp\left(\frac{s}{k}(\Xi_{k+1} - X_{k+1})\right) | \Xi_{k+1} \dots \Xi_n\right]\right]$$

by repeated application of the Law of total expectation. Since:

$$\mathbb{E}[X_{k+1} | \Xi_{k+1} \dots \Xi_n] = \Xi_{k+1},$$

then $\Xi_{k+1} - X_{k+1}$ is a random variable with a mean of zero bounded within width D, and it also has a variance given by:

$$\sigma_{k+1}^2 = \frac{n\sigma^2 - \sum_{j=k+1}^{n} X_j^2}{n - (n-k-1)} - \Xi_k^2 \leq \frac{n\sigma^2}{k+1} \tag{4}$$

by application of lemma 2. Therefore:

$$\mathbb{E}[\exp(s\Xi_m)] \leq \exp\left(\sum_{k=m}^{n-1}\left(\alpha D^2 + \beta\frac{n\sigma^2}{k+1}\right)\frac{s^2}{k^2}\right) \qquad \square$$

This martingale result relates the moment generating function bound of the average of finite variables relative to their mean, to the moment generating function bounds of the differences of the incremental averages relative to their mean. It is pertinent to note that this result could be made much stronger by working around Equation (4), but this comes at a cost of increased mathematical complexity.

Since lemmas 9 and 7 share a common form, and because of Hoeffding's reduction (lemma 8), all the derivations that follow that invoke lemma 7 have direct analogues using lemma 9 for the context of sampling without replacement. Note, however, that the bound without replacement (lemma 9) may or may-not be tighter than the bound with replacement (lemma 7), so the process of substituting one for the other should be done judiciously on a case-by-case basis to create the tightest possible bound. All the results in this paper (relevant to sampling without replacement) have been produced with this judicious choice been conducted.

4. The Stratified Finite Empirical Bernstein Bound

In this section we derive a novel probability bound for the error of the stratified random sampling estimate. We begin by precisely defining the context of our derivation and to which our bound applies.

Definition 1 (Problem context)**.** *Let a population consist of n number of strata of finite data points, where n_i is the number of data points in the ith stratum. All values in a stratum are bound within a finite support of width D_i. Denote the mean and variance of the ith stratum μ_i and σ_i^2, respectively. In this context, if $X_{i,1}, X_{i,2}, \ldots, X_{i,n_i}$ are random variables corresponding to those data values randomly and sequentially drawn (with or without) replacement, then $\Xi_{i,m_i} = \frac{1}{m_i}\sum_{j=1}^{m_i} X_{i,j}$ is the average of the first m_i of these random variables. And $\hat{\sigma}_i^2 = \frac{1}{m_i}\sum_{j}^{m_i}(X_{i,j} - \Xi_{i,m_i})^2$ is the biased sample variance of the first m_i of these samples. And $\hat{\hat{\sigma}}_i^2 = m_i\hat{\sigma}_i^2/(m_i - 1)$ is the unbiased sample variance of the first m_i of these samples.*

We are interested in the average of the means of the strata as weighted by constant positive factors $\{\tau_i\}_{i\in\{1,\ldots,n\}}$. In our derivation, we also consider intermediary weights $\{\theta_i\}_{i\in\{1,\ldots,n\}}$.

The bound is now developed in four theorems, which build on each other in sequence. The first is an expression for a probability bound on the absolute error of the weighted stratified sample means about the weighted strata means.

 doi:10.20944/preprints201901.0202.v1

Theorem 1 (SEBM* bound)**.** *Assuming the context given in Definition 1, and let* $\Omega_{m_i}^{n_i}$ *and* $\Psi_{m_i}^{n_i}$ *be given as in lemma 7, then:*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}\tau_i(\Xi_{i,m_i}-\mu_i)\right| \geq \sqrt{4\log(2/t)\sum_{i=1}^{n}\left(\frac{1}{17}D_i^2\Omega_{m_i}^{n_i}+\frac{1}{2}\sigma_i^2\Psi_{m_i}^{n_i}\right)\tau_i^2}\right) \leq t \tag{5}$$

Proof. Applying Lemma 3:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n}\tau_i\Xi_{i,m_i}-\sum_{i=1}^{n}\tau_i\mu_i \geq t\right) &\leq \mathbb{E}\left[\exp\left(\sum_{i=1}^{n}\tau_i s\left(\Xi_{i,m_i}-\mu_i\right)\right)\right]\exp(-st)\\ &= \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\tau_i s\left(\Xi_{i,m_i}-\mu_i\right)\right)\right]\exp(-st)\end{aligned}$$

by independence of the sampling between the strata. This form is sufficient for Lemma 7 with Lemma 5 to apply, resulting in a double-sided tail bound:

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}\tau_i(\Xi_{i,m_i}-\mu_i)\right| \geq t\right) \leq 2\exp\left(\sum_{i=1}^{n}\left(\frac{1}{17}D_i^2\Omega_{m_i}^{n_i}+\frac{1}{2}\sigma_i^2\Psi_{m_i}^{n_i}\right)\tau_i^2 s^2 - st\right)$$

Minimising with respect to s and rearranging gives result. □

In most cases, the weights τ_i can be considered as the probability weights $\tau_i = n_i/(\sum_{j=1}^{n} n_j)$, and in this context this probability bound can be used as-is for a measure of uncertainty in stratified random sampling if the true variances (or alternatively, upper bounds on the true variances) of the strata are known. However, in other contexts the sum of variances must be estimated from the data collected, and to rectify this, we develop and incorporate a probability bound for the estimate of the sum of variances, as follows.

Theorem 2. *Assuming the context given in Definition 1. Then with* $\Psi_{m_i}^{n_i}$ *per lemma 7:*

$$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\sigma_i^2-\hat{\sigma}_i^2-(\mu_i-\Xi_{i,m_i})^2)\geq\sqrt{2\log(1/y)\sum_{i=1}^{n}\sigma_i^2\theta_i^2D_i^2\Psi_{m_i}^{n_i}}\right)\leq y \quad (6)$$

Proof. To create a probability bound for the sum of variances (weighted by arbitrary positive θ_i), we consider the average square of samples about the strata means. Applying lemma 3 gives:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\sigma_i^2-\frac{1}{m_i}\sum_{j=1}^{m_i}(X_{i,j}-\mu_i)^2)\geq y\right)\\
&\leq\mathbb{E}\left[\exp\left(\sum_{i=1}^{n}s\theta_i\left(\sigma^2-\frac{1}{m_i}\sum_{j=1}^{m_i}(X_{i,j}-\mu_i)^2\right)\right)\right]\exp(-sy)\\
&=\exp(-sy)\prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{s\theta_i}{m_i}\sum_{j=1}^{m_i}(\sigma^2-(X_{i,j}-\mu_i)^2)\right)\right]
\end{aligned}$$

by independence of the sampling between the strata. This is sufficient for lemma 7 with lemma 6 to apply giving:

$$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\sigma_i^2-\frac{1}{m_i}\sum_{j=1}^{m_i}(X_{i,j}-\mu_i)^2)\geq y\right)\leq\exp\left(\frac{1}{2}\sum_{i=1}^{n}\sigma_i^2\theta_i^2s^2D_i^2\Psi_{m_i}^{n_i}-sy\right)$$

Minimising with respect to s, rearranging, applying lemma 2 gives result. □

This inequality gives the probability bound between the weighted variances of the strata, the weighted (biased) sample variances and the weighted square error of the sample means. Although we anticipate that the weighted square error of the sample means goes to zero rather fast as additional samples are taken, we nonetheless wish to develop and incorporate another probability bound to eliminate the specific consideration of it.

 doi:10.20944/preprints201901.0202.v1

Theorem 3. *Assuming the context given in Definition 1. Then with $\Omega_{m_i}^{n_i}$ per lemma 7:*

$$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\mu_i-\Xi_{i,m_i})^2 \geq \frac{\log(2n/r)}{2}\sum_{i=1}^{n}\theta_i D_i^2\Omega_{m_i}^{n_i}\right) \leq r \tag{7}$$

Proof. We consider the weighted square error of the sample means:

$$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\mu_i-\Xi_{i,m_i})^2 \geq r\right) \leq 1-\prod_{i=1}^{n}\mathbb{P}\left(\theta_i(\mu_i-\Xi_{i,m_i})^2 \leq r_i\right)$$

$$= 1-\prod_{i=1}^{n}\left(1-\mathbb{P}\left(\mu_i-\Xi_{i,m_i}\geq\sqrt{\frac{r_i}{\theta_i}}\right)-\mathbb{P}\left(\Xi_{i,m_i}-\mu_i\geq\sqrt{\frac{r_i}{\theta_i}}\right)\right),$$

such that $\sum r_i = r$, by independence of the sampling and probability complementaries. This is sufficient for us to apply lemma 3 together with lemma 7 and lemma 4, giving:

$$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\mu_i-\Xi_{i,m_i})^2 \geq r\right) \leq 1-\prod_{i=1}^{n}\left(1-2\exp\left(-\frac{2r_i}{\theta_i D_i^2\Omega_{m_i}^{n_i}}\right)\right)$$

Next, choosing r_i to minimise this expression gives:

$$r_i = \frac{r\theta_i D_i^2\Omega_{m_i}^{n_i}}{\sum_j \theta_j D_j^2\Omega_{m_j}^{n_j}}$$

Thus:

$$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\mu_i-\Xi_{i,m_i})^2 \geq r\right) \leq 1-\prod_{i=1}^{n}\left(1-2\exp\left(\frac{-2r}{\sum_j \theta_j D_j^2\Omega_{m_j}^{n_j}}\right)\right)$$

Using the knowledge that $\log(1-(1-\exp(x))^n) \leq x+\log(n)$ for negative x, and rearranging, gives result. □

This theorem bounds the weighted square error of the sample means. In the next, and final, step we combine the inequalities of Equations (5), (6) and (7) together, to complete our derivation.

 doi:10.20944/preprints201901.0202.v1

Theorem 4 (SEBM bound)**.** *Assuming the context given in Definition 1. Then with* $\Omega_{m_i}^{n_i}, \Psi_{m_i}^{n_i}$ *per lemma 7:*

$$\mathbb{P}\left(\frac{|\sum_{i=1}^{n}\tau_i(\Xi_{i,m_i}-\mu_i)|}{\sqrt{\log(6/p)}} \geq \sqrt{\alpha_{m_i}^{n_i}+\left(\sqrt{\beta_{m_i}^{n_i}}+\sqrt{\gamma_{m_i}^{n_i}}\right)^2}\right) \leq p \tag{8}$$

where:

$$\begin{aligned}
\alpha_{m_i}^{n_i} =& \sum_{i=1}^{n}\frac{4}{17}\Omega_{m_i}^{n_i}D_i^2\tau_i^2 \\
\beta_{m_i}^{n_i} =& \log(3/p)\left(\max_i \tau_i^2 {\Psi_{m_i}^{n_i}}^2 D_i^2\right) \\
\gamma_{m_i}^{n_i} =& 2\sum_{i=1}^{n}\tau_i^2\Psi_{m_i}^{n_i}(m_i-1)\hat{\sigma}_i^2/m_i + \log(6n/p)\sum_i \tau_i^2 D_i^2 \Omega_{m_i}^{n_i}\Psi_{m_i}^{n_i} \\
&+ \log(3/p)\left(\max_i \tau_i^2 {\Psi_{m_i}^{n_i}}^2 D_i^2\right)
\end{aligned}$$

Proof. By widening the bound of Equation (6) we get:

$$\mathbb{P}\left(\begin{array}{c}\sum_{i=1}^{n}\theta_i\sigma_i^2 - \sum_{i=1}^{n}\theta_i(\hat{\sigma}_i^2+(\mu_i-\Xi_{i,m_i})^2) \geq \\ \sqrt{2\log(1/y)(\max_i \theta_i D_i^2 \Psi_{m_i}^{n_i})\sum_{i=1}^{n}\theta_i\sigma_i^2}\end{array}\right) \leq y.$$

Completing the square gives for $\sqrt{\sum_{i=1}^{n}\theta_i\sigma_i^2}$ gives:

$$\mathbb{P}\left(\sqrt{\sum_{i=1}^{n}\theta_i\sigma_i^2} \geq \begin{array}{c}\sqrt{\sum_{i=1}^{n}\theta_i(\hat{\sigma}_i^2+(\mu_i-\Xi_{i,m_i})^2)} \\ +\frac{\log(1/y)}{2}\left(\max_i \theta_i D_i^2 \Psi_{m_i}^{n_i}\right) \\ +\sqrt{\frac{\log(1/y)}{2}\left(\max_i \theta_i D_i^2 \Psi_{m_i}^{n_i}\right)}\end{array}\right) \leq y.$$

Combining with Equation (7) with a union bound (lemma 1) gives:

$$\mathbb{P}\left(\sqrt{\sum_{i=1}^{n}\theta_i\sigma_i^2} \geq \begin{array}{c}\sqrt{\sum_{i=1}^{n}\theta_i\hat{\sigma}_i^2 + \frac{\log(2n/r)}{2}\sum_i \theta_i D_i^2\Omega_{m_i}^{n_i}} \\ +\frac{\log(1/y)}{2}\left(\max_i \theta_i D_i^2 \Psi_{m_i}^{n_i}\right) \\ +\sqrt{\frac{\log(1/y)}{2}\left(\max_i \theta_i D_i^2 \Psi_{m_i}^{n_i}\right)}\end{array}\right) \leq y + r$$

Which is a bound for the weighted sum variances in terms of the sample variances. Letting $\theta_i = \frac{1}{2}\tau_i^2\Psi_{m_i}^{n_i}$ and combining with (5) with a union bound (lemma 1) and assigning $r = t = y = p/3$ and rewriting in terms of unbiased sample variance gives the result. □

This completes the derivation. In Equation (8), we have a concentration inequality for the sum of weighted strata sample mean errors. In this context, the weights τ_i are flexible but would naturally be the probability weights $\tau_i = n_i/(\sum_{j=1}^{n} n_j)$, in which case the inequality gives us a measure of accuracy in stratified random sampling.

We propose an online process of choosing additional samples from the strata in order to to minimise this bound, which we henceforth refer to as the *stratified empirical Bernstein method*, or SEBM as shorthand.

Additionally, we note that for any strata i that is sampled without replacement, the associated $\Omega_{m_i}^{n_i}$ and $\Psi_{m_i}^{n_i}$ may be substituted for $\bar{\Omega}_{m_i}^{n_i}$ and $\bar{\Psi}_{m_i}^{n_i}$ to potentially tighten the bound. This corresponds to optional substitution of lemma 9 for lemma 7 at various points in the derivation.

5. Numerical Evaluation

In this section we consider the utility of minimising this concentration inequality as a method of choosing samples from the strata. First we outline the benchmarks used to evaluate our method's performance. Then we describe two synthetic data sets and report the distribution of errors under our method and the benchmarks. Following this, in Section 6, our method is evaluated in an example application — that of calculating the Shapley value of a cooperative game.

5.1. Benchmarks algorithms

In the numerical evaluations, we compare the following sampling methods:

- **SEBM-WO** (Stratified empirical Bernstein method without replacement): our method of iteratively choosing samples from strata to minimize our concentration inequality, Equation (8). An initial sample of two data points from each strata is used to initialise the sample variances of each, with additional samples made to maximally minimise the inequality at each step, before recomputing sample variances. All samples are drawn *without* replacement, and the inequality used involved the judicious use of martingale inequality lemma 9 (see the notice in Section 3.1).
- **SEBM-W** (Stratified empirical Bernstein method with replacement): as above, with the exception that all samples are drawn *with* replacement, and consequently the inequality does not utilize the martingale inequality given in lemma 9.
- **Sim-WO** (Simple random sampling without replacement): simple random sampling from the population irrespective of strata *without* replacement.
- **Sim-W** (Simple random sampling with replacement): simple random sampling from the population irrespective of strata *with* replacement.
- **Ney-WO** (Neyman sampling without replacement): the method of maximally choosing samples *without* replacement from strata proportional to the strata variance.
- **Ney-W** (Neyman sampling with replacement): the method of choosing samples *with* replacement proportional to the strata variance.
- **SEBM*** (Stratified empirical Bernstein method with variance information): the method of iteratively choosing samples *without* from strata to minimize (5), judiciously using martingale lemma 9

Note that the last three methods assume prior perfect knowledge of the variance of each of the strata, and that in Equations (8) and (5) we selected for minimising a 50% confidence interval (i.e. $p = 0.5$ and $t = 0.5$ respectively).

Between these methods there are compared different factors such as the dynamics of sampling: with and without replacement, with stratification and without, between our method and Neyman sampling, and with and without perfect knowledge of stratum variances. For these methods we consider the effectiveness against Beta distributed data and Uniform & Bernoulli data.

5.2. Synthetic Data

The first way we demonstrate the efficacy of our method is to generate sets of synthetic data, and then numerically examine the distribution of errors generated by different methods of choosing finite sequences of samples. In this subsection, we described the two types of synthetic data sets used in this evaluation, namely: (i) beta distributed stratum data, and (ii) a particular form of uniform and Bernoulli distributed stratum data.

5.2.1. Beta-Distributed Data

The first pool of synthetic data sets are intended to be representative of potential real-world data. These sets have between 5 and 21 strata, with the number of strata drawn with uniform probability, and each strata sub-population has sizes ranging from 10 to 201, also drawn uniformly. The data values in each strata are drawn from beta distributions:

$$\phi(x)_{\{\alpha,\beta\}} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

with α and β parameters drawn uniformly between 0 and 4 for each stratum.

Figure 1 compares the distribution of absolute error achieved by each of the sampling methods over 5000 rounds of these data sets. Each quadrant presents the results that the methods achieve for a given budget of samples, expressed as a multiple of the number of strata (noting that data sets where the sampling budget exceeded the volume of data were excluded). From the results of data in Figure 1, we can see that our sampling technique (SEBM-WO and SEBM-W) performs comparably to Neyman sampling (Ney-WO and Ney-W) despite not having access to knowledge of stratum variances. Also, there is a notable similarity between SEBM* and SEBM-WO. As expected, sampling without replacement always performs better than sampling with replacement for the same method, and this difference is magnified as the number of samples grows large in comparison to the population size. Finally, simple random sampling almost always performs worst, because it fails to take advantage of any variance information. These results and their interpretation are discussed and detailed in Section 7 along with results from the other test cases discussed below.

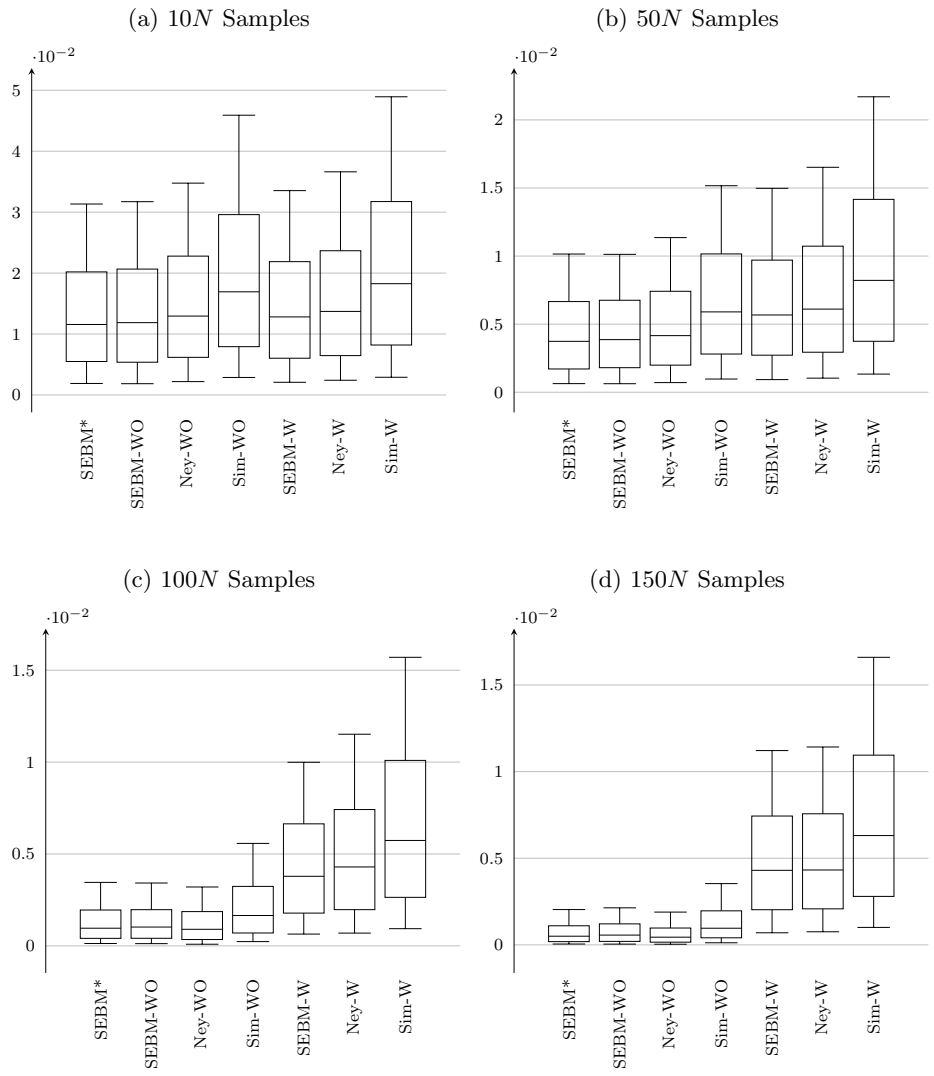

Fig 1: Distribution of numerical absolute errors across 5000 rounds of beta-Distributed data, for different methods of stratified sampling. Each plot shows absolute errors for different numbers of samples multiplied by the number of strata, N, e.g. $10N$ samples means that the test problem has a sample budget of ten times the number of strata. The whiskers show the 9th and 91st percentiles, data points outside this range are not shown.

5.2.2. A Uniform and Bernoulli Distribution

The aim of this section is to examine cases of distributions in which our sampling method (SEBM-WO) performs poorly, particularly compared to Neyman sampling (Ney-WO). For this purpose, a data set with two strata is generated, with each stratum containing 1000 points. The data in the first stratum is uniform continuous data in range $[0, 1]$, while the data in the second is all zeros except for a specified small number, a, of successes with value one. For this estimation problem, we conduct stratified random sampling with a budget of 300 samples, comparing the SEBM*, SEBM-WO and Ney-WO methods. The average error returned by the methods across 20000 realisations of this problem, plotted against the number of sucesses a, are shown as a graph in Figure 2.

This figure demonstrates that SEBM-WO and SEBM* perform poorly when the strata contain only very small numbers of sucesses. This under-performance is not simply a result of the SEBM-WO method oversampling in a process of learning the stratum variances, as the under-performance is present in SEBM* as well. The reasons for this under-performance are discussed in conjunction with other results in more detail in Section 7.

Next, we considered the calculation of the Shapley value as an example application of our stratified sampling method.

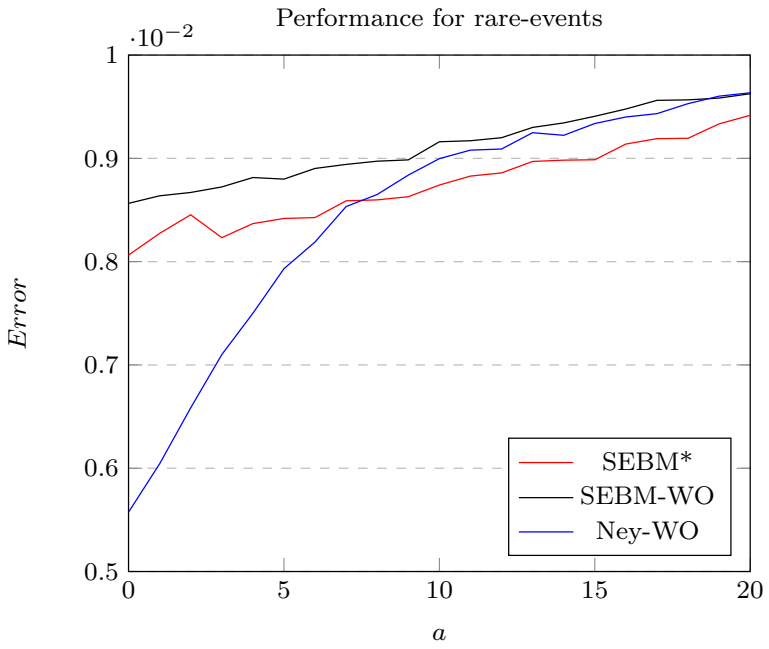

Fig 2: Average error of three stratified random sampling methods for the Uniform-Bernoulli data sets of Section 5.2.2, plotted against success parameter a, across 20000 rounds.

6. Shapley Value Approximation

The Shapley value is a cornerstone measure in cooperative game theory. It is an axiomatic approach to allocating a divisible reward or cost between participants, where there is a clearly defined notion of how much reward any group (or 'coalition') of participants could achieve by themselves (Chalkiadakis, Elkind and Wooldridge, 2012). The Shapley value has many applications, including analysing the power of voting blocks in weighted voting games (Bachrach et al., 2009), in cooperative cost and surplus division problems (Aziz et al., 2016; Chapman, Mhanna and Verbič, 2017), and as a measure of network centrality (Michalak et al., 2013).

Formally, a *cooperative game*, $(N, v) \in \mathbb{G}_N$, comprises a set of n players, $N = \{1, 2, \ldots, n\}$, and a *characteristic function*, $v : S \subset N \to \mathbb{R}$, which is a function specifying the reward which can be achieved if a subset of the players $S \subset N$ cooperate, where $v(\emptyset) = 0$. In this context the Shapley value is a unique allocation which satisfies the following set of natural axioms:

- **Efficiency**: That the total reward is divided up: $\sum_i \psi(\langle N, v\rangle)_i = v(N)$
- **Symmetry**: If two players i and j are totally equivalent 'substitutes' then the receive the same reward: ie. if $v(S \cup i) = v(S \cup j) \quad \forall S \subseteq N \setminus \{i, j\}$, then $\psi(\langle N, v\rangle)_i = \psi(\langle N, v\rangle)_j$
- **Null Player**: If the addition of a player i to any coalition brings nothing, and is a 'null player', then it receives reward of zero: i.e if $v(S \cup i) = v(S) \quad \forall S \subseteq N$ then $\psi(\langle N, v\rangle)_i = 0$
- **Additivity**: That for any two games the reward afforded each player is each is the sum of the games considered together: i.e. for any v_1 and v_2, that: $\psi(\langle N, v_1 + v_2\rangle) = \psi(\langle N, v_1\rangle) + \psi(\langle N, v_2\rangle)$

Specifically, the Shapley value is a mapping from cooperative games to the player rewards: $\mathrm{Sh}, \mathbb{G}_N \to \mathbb{R}^n$, given by:

$$\mathrm{Sh}_i((N, v)) = \sum_{S \subset N, i \notin S} \frac{(n - |S| - 1)!\,|S|!}{n!}(v(S \cup \{i\}) - v(S)) \tag{9}$$

That is, under the Shapley value each player is afforded their average marginal contribution across every possible sequence of player join orderings. Alternatively, if $v_{i,k}$ is the average marginal contribution of player i across coalitions of size k:

$$v_{i,k} = \sum_{\substack{S \subset N \\ |S|=k,\ i \notin S}} \frac{(n - |S| - 1)!\,|S|!}{n!}(v(S \cup \{i\}) - v(S)), \tag{10}$$

then the Shapley value can be expressed as:

$$\mathrm{Sh}_i((N, v)) = \frac{1}{n}\sum_{k=0}^{n-1} v_{i,k} \tag{11}$$

Though the Shapley value is conceptually simple, its use is hampered by the fact that its total expression involves exponentially many evaluations of the characteristic function (there are $n \times 2^{n-1}$ possible marginal contributions between n players).

However, since the Shapley value is expressible as an average over averages by Equation (11), it is possible to approximate these averages via sampling techniques, and particularly so as these averages are naturally stratified by size. In previously published literature, other techniques have been used to allocate samples in this context, particularly simple sampling (Castro, Gómez and Tejada, 2009), Neyman allocation (Castro et al., 2017; O'Brien, Gamal and Rajagopal, 2015), and allocation to minimize Hoeffding's inequality (Maleki et al., 2013). We assess the benefits of using our bound by comparing its performance to the approaches above in the context of some example cooperative games, as described below.

Example Game 1 (Airport Game)**.** *An $n = 15$ player game with characteristic function:*

$$v(S) = \max_{i \in S} w_i$$

where $w = \{w_1, \ldots, w_{15}\} = \{1, 1, 2, 2, 2, 3, 4, 5, 5, 5, 7, 8, 8, 8, 10\}$. The maximum marginal contribution is 10, so we assign $D_i = 10$ for all i.

Example Game 2 (Voting Game)**.** *An $n = 15$ player game with characteristic function:*

$$v(S) = \begin{cases} 1, & \text{if} \quad \sum_{i \in S} w_i > \sum_{j \in N} w_j / 2 \\ 0, & \text{otherwise} \end{cases}$$

where $w = \{w_1, \ldots, w_{15}\} = \{1, 3, 3, 6, 12, 16, 17, 19, 19, 19, 21, 22, 23, 24, 29\}$. The maximum marginal contribution is 1, so we assign $D_i = 1$ for all i.

Example Game 3 (Simple Reward Division)**.** *An $n = 15$ player game with characteristic function:*

$$v(S) = \frac{1}{2} \left(\sum_{i \in S} \frac{w_i}{100} \right)^2$$

where $w = \{w_1, \ldots, w_{15}\} = \{45, 41, 27, 26, 25, 21, 13, 13, 12, 12, 11, 11, 10, 10, 10\}$ The maximum marginal contribution is 1.19025, so we assign $D_i = 1.19025$ for all i.

Example Game 4 (Complex Reward Division)**.** *An $n = 15$ player game with characteristic function:*

$$v(S) = \left(\sum_{i \in S} \frac{w_i}{50} \right)^2 - \left\lfloor \sum_{i \in S} \frac{w_i}{50} \right\rfloor^2$$

where $w = \{w_1, \ldots, w_{15}\} = \{45, 41, 27, 26, 25, 21, 13, 13, 12, 12, 11, 11, 10, 10, 10\}$ In this game, we assign $D_i = 2$ for all i.

For each game, we compute the exact Shapley value, and then the average error in the approximated Shapley value for a given budget m of samples. The results are shown in Table 1, where the average absolute error in the Shapley value is computed by sampling with Maleki's method (Maleki et al., 2013) is denoted e^{Ma}, e^{sim} is Castro's simple sampling method (Castro, Gómez and Tejada, 2009), e^{Ca} is Castro's Neyman sampling method (Castro et al., 2017), and e^{SEBM} is the error associated with our method, SEBM-WO. The results in Table 1 show that our method performs well across the benchmarks. A discussion of all of the results is considered in the next section.

(a) Airport Game average errors

m/n^2	10	50	100	500	1000
e^{Ma}	298.36	133.07	99.639	41.963	29.257
e^{sim}	357.84	146.09	106.22	44.545	36.333
e^{Ca}	325.65	115.73	75.85	31.014	22.115
e^{SEBM}	259.24	73.754	54.756	7.7099	1.3038

(b) Voting Game average errors

m/n^2	10	50	100	500	1000
e^{Ma}	130.98	57.775	41.522	18.657	13.178
e^{sim}	145.72	59.716	40.306	17.563	12.835
e^{Ca}	142.1	47.35	31.048	14.08	9.7998
e^{SEBM}	122.79	47.435	33.179	8.5464	1.9953

(c) Simple Reward Division Game average errors

m/n^2	10	50	100	500	1000
e^{Ma}	25.678	11.615	7.7921	3.4805	2.2904
e^{sim}	22.102	9.0445	6.2178	2.6419	1.9379
e^{Ca}	22.367	8.925	6.6915	2.7267	1.9402
e^{SEBM}	19.254	7.0441	5.1578	1.1825	0.28173

(d) Complex Reward Division Game average errors

m/n^2	10	50	100	500	1000
e^{Ma}	276.13	118.88	86.993	40.148	27.44
e^{sim}	251.44	107.97	78.628	34.639	26.821
e^{Ca}	290.51	116.5	81.819	35.702	26.501
e^{SEBM}	214.21	78.467	54.101	12.447	2.7109

Table 1: Average absolute errors in the Shapley value calculation across all players in the four cooperative games (units in 10^{-4}), for the different sampling schemes with different sampling budgets m per number of strata (with $n^2 = 15^2$ for all).

7. Discussion

From the results across Figures 1 and 2 and Table 1, the main observation is that our sampling technique, SEBM-WO or SEBM-W, performs competitively to Neyman sampling (Ney-WO or Ney-W). This is despite SEBM not having access to knowledge of strata variances, which the Neyman methods do. If instead we compare SEBM*, which has access to strata variances, and Ney-WO then both methods use the same information, and the results are even more positive for our method. The reasons for this performance are interesting, and we now discuss them in more detail.

To begin, from Figure 1, observe that sampling without replacement always performs better than sampling with replacement for the same method. This improvement is magnified as the number of samples grows large relative to the size of the population. At the same time, simple random sampling almost always performs worst, because it fails to take advantage of any variance information. These results are as expected.

Next, note that the results of Figure 1 show that there is a mid-range of sample sizes where choosing a different method can have a greater impact on sampling efficiency and rate of average error reduction than the difference between sampling with or without replacement. This is an important insight, as sampling real-world data (e.g. surveys, polling, destructive testing, etc) can be an expensive and slow process. Accordingly an appropriate method of choosing numbers of samples can lead to a material difference in cost for the same accuracy. There is also a slight decrease in the performance of SEBM* in comparison with Ney-WO in the case of high number of samples and sampling without replacement, as illustrated in Figure 1. This potentially indicates that lemma 9 can be improved — as noted in Section 3.1.

Furthermore, if the data features very rare events, then SEBM-WO and SEBM* seem to perform in a manner less than ideal. These condition are illustrated in Figure 2, where the more rare the Bernoulli variable successes, the worse our methods perform in comparison with Neyman sampling (Ney-WO). This shortcoming can be partly explained by noting that SEBM-WO unnecessarily wastes samples on the Bernoulli stratum of rare events in the process of learning that the variance is almost zero, whereas Ney-WO can avoid this because it has prior knowledge of the variances to begin with. As such, this factor accounts for the difference between the performance of SEBM-WO and SEBM* in the context of Figure 1 and also in Figure 2. What is surprising is how small the difference in performance between SEBM-WO and SEBM* is. This indicates that as additional samples are taken that uncertainty about the strata variances have less and less effect upon the total numbers of samples that are eventually drawn from each of the strata.

However, the performance difference between SEBM* and Ney-WO in Figure 2 is not explained by this argument, as they use the same information. Instead, reason for this difference in performance is found by considering the simplifying approximation of Equation (2). Specifically, (2) introduces a particular distortion into the shape of Equation (8) which our sampling seeks to

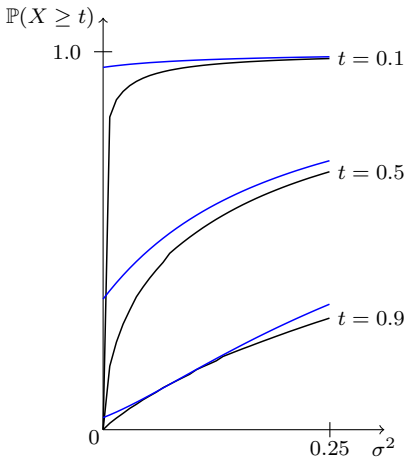

Fig 3: The plots against σ^2 of $\mathbb{P}(X \geq t) \leq \min_s \mathbb{E}[\exp(sX)] \exp(-st)$ with $\mathbb{E}[\exp(sX)]$ via Equations (1) (black) and (3) (blue) with $D = 1$. Note that Equation (3) generally captures the relevant shape and magnitude of the more accurate equation except in region of small σ^2 where the bound is overly weakened.

minimise. Figure 3 illustrates how the approximation (2) loosens the bound with respect to the variance. Observe that when the variances are very small that Equation (3) overly loosens the bounds, causing oversampling of strata with very small variances. It appears that this factor is at play in the under-performance shown in Figure 2 and also the slight under-performance of our method in the Voting Game in Table 1b. We note that there may be other corner-cases where our method also under-performs.

In comparison to existing approaches to approximating the Shapley value, our sampling method shows improved performance on almost all accounts, as shown in Table 1. This was particularly the case in the context of large sample budgets, as our method (SEBM-WO, with error e^{SEBM}) sampled without replacement, while the other methods sampled with replacement. However it would be remiss not to mention the computational overhead of iteratively minimising (one sample at a time) our inequality in the context of our simple example games. This overhead can, in practice, reduce the benefit of using a more efficient sampling method. However, on more complicated games where the characteristic function is slower to calculate, any overhead associated with the sampling choice will be much less relevant. We also note that our method's performance could potentially be further improved by selecting more refined D_i values for our example games.

One primary limitation of our method is that it rests on assumption of known data widths D_i (and in the case of sampling-without-replacement, also on strata sizes N_i), which may not be exactly known in practice. One way to overcome this may be to use our method with a reliable overestimate these parameters (by expert opinion or otherwise). This approximation or estimation may itself open consideration of other probability bounds and/or sampling methods, however we have not investigated this line of inquiry.

In practice, it may be advisable to run our method with an underestimate of the data widths, as the sampling process is fundamentally sensitive the the shape

of the inequality and not necessarily its magnitude or accuracy as a bound. Our concentration inequality, Equation (8), is derived by a combination of Chernoff bounds fused together with probability unions, so it is expected to give conservative confidence intervals on the error of the estimate in stratified random sampling, which may be useful outside of the context of sampling decisions.

8. Multidimensional Extension

The method of choosing samples can be extended to multidimensional data by a simple modification.

Specifically, instead of considering data that is single-valued, we consider data points that are vectors. Formally, for n strata of finite data points which are all vectors of size M, let n_i be the number of data points in the ith stratum. Let the data in the ith stratum have a mean vector values μ_i (with $\mu_{i,j}$ for the jth component of the vector), which are value bounded within a finite width $D_{i,j}$, and have vector value variances $\sigma^2_{i,j}$. Given this, if $X_{i,1}, X_{i,2}, \ldots, X_{i,n_i}$ (with $X_{i,1,j}$ being the jth component of such a vector) are vector random variables corresponding to those data values randomly and sequentially drawn (with or without) replacement, then denote the average of the first m_i of these random variables by $\Xi_{i,m_i} = \frac{1}{m_i}\sum_{j=1}^{m_i} X_{i,j}$ (with $\Xi_{i,m_i,j}$ being the jth component of that vector average). Let $\hat{\sigma}^2_{i,j} = \frac{i}{m_i - 1}\sum_{k=1}^{m_i}(X_{i,k,j} - \Xi_{i,m_i,j})^2$ be the unbiased sample variance of the first m_i of these random variables in the jth component. And again, assume we have weights τ_i for each stratum.
In this context we have the following theorem:

Theorem 5 (Vector SEBM bound)**.** *In the context above, then with* $\Omega^{n_i}_{m_i}, \Psi^{n_i}_{m_i}$ *per lemma 7:*

$$\mathbb{P}\left(\begin{matrix}\sum_{j=1}^{M}\left(\sum_{i=1}^{n}\tau_i(\Xi_{i,m_i,j} - \mu_{i,j})\right)^2 \geq \\ \log(6/p)\sum_{j=1}^{M}\left(\alpha^{n_i}_{m_i,j} + \left(\sqrt{\beta^{n_i}_{m_i,j}} + \sqrt{\gamma^{n_i}_{m_i,j}}\right)^2\right)\end{matrix}\right) \leq Mp \tag{12}$$

where:

$$\begin{aligned}
\alpha^{n_i}_{m_i,j} =& \sum_{i=1}^{n}\frac{4}{17}\Omega^{n_i}_{m_i}D^2_{i,j}\tau_i^2 \\
\beta^{n_i}_{m_i,j} =& \log(3/p)\left(\max_i \tau_i^2 {\Psi^{n_i}_{m_i}}^2 D^2_{i,j}\right) \\
\gamma^{n_i}_{m_i,j} =& 2\sum_{i=1}^{n}\tau_i^2\Psi^{n_i}_{m_i}(m_i - 1)\hat{\sigma}^2_{i,j}/m_i + \log(6n/p)\sum_i \tau_i^2 D^2_{i,j}\Omega^{n_i}_{m_i}\Psi^{n_i}_{m_i} \\
& + \log(3/p)\left(\max_i \tau_i^2 {\Psi^{n_i}_{m_i}}^2 D^2_{i,j}\right)
\end{aligned}$$

 doi:10.20944/preprints201901.0202.v1

Proof. Squaring (8) and applying it specifically to the jth component of all the vectors gives:

$$\mathbb{P}\left(\frac{\left(\sum_{i=1}^{n}\tau_i(\Xi_{i,m_i}-\mu_i)\right)^2}{\log(6/p)} \geq \alpha_{m_i,j}^{n_i} + \left(\sqrt{\beta_{m_i,j}^{n_i}}+\sqrt{\gamma_{m_i,j}^{n_i}}\right)^2\right) \leq p$$

Taking a series of union bounds (lemma 1) over j gives result. □

The left hand side of the inequality in (12) is the square euclidean distance between our weighted stratified sample vector estimate $\sum_{i=1}^{n}\tau_i\Xi_{i,m_i}$ and the true mean stratified vector $\sum_{i=1}^{n}\tau_i\mu_i$. In this context a sampling process consists (the same as before) of sampling to maximally minimise the right hand side of the inequality. This formulation can potentially be applied to more involved computational tasks and sampling data with multiple features.

9. Future Work and Applications

We begin this section with a discussion of the relationship of our bound to existing concentration inequalities, and some opportunities for future improvements. The derivation of our inequality extends from consideration of Chernoff bounds and probability unions in a similar vein to other EBB derivations (Maurer and Pontil, 2009; Bardenet and Maillard, 2015). However, the bounds on the moment generating functions that we developed in Section 3 are rife with loosening approximations, and stronger and/or more representative bounds could be developed at the cost of greater mathematical complexity. Alternatively, the approach used to derive the Entropic (Boucheron, Lugosi and Massart, 2003) or Effron-Stein methods (Efron and Stein, 1981) could result in different and possibly tighter results.

Additionally, although our method works generally, there may be better or more appropriate sampling methods in the event that there is more information known about the underlying distributions. It is sometimes possible to derive ideal concentration inequalities in restricted circumstances, and more broadly there exist some computational methods to numerically derive ideal bounds (Owhadi et al., 2013; Han et al., 2015). Using these techniques it may be possible to derive ideal numerical bounds, particularly for bounds considering very small numbers of samples.

We now consider two prospective applications of our existing method. First, the approach derived in this paper was motivated by the problem of approximating the Shapley value of cooperative games defined over complicated optimization problems (i.e. with characteristic functions given by the solution to non-trivial optimization problems). One example of this is the problem of pricing access and services in electricity networks. An electricity network is complicated systems used to transport electrical power from generators to loads, subject to the physical and operational constraints of the system's components. With the emergence of new technologies, electricity is now generated, monitored and used on neighborhood distribution networks by devices that are increasingly

responsive and interconnected to the network. Because of this, there are various emerging schemes of how a future distribution-network energy market platform might be designed. Within this context, the Shapley value has been proposed as a fair mechanism for the allocation of resources and costs on such networks. The Shapley value has been considered in different ways as a mechanism for pricing demand response (O'Brien, Gamal and Rajagopal, 2015), demand or load (Chapman, Mhanna and Verbič, 2017), supply or generation (Acuña et al., 2018), and potentially all simultaneously (Burgess, Chapman and Scott, 2018). Although computing the Shapley value exactly is impractical in these contexts, sample-based approximations are a promising avenue for implementing Shapely value pricing schemes in real-world electricity systems.

Second, a potential use of our stratified sampling method is in improving the performance of *stochastic gradient decent* (SGD) methods for training neural networks (Ruder, 2016). Neural networks are trained by iteratively refining their parameters — the weights and biases of the network — against a cost function of the network's performance against training data. One common method of training neural networks is gradient decent (GD). In each iteration of GD, the derivative of how much a change in any parameter would influence the average performance of the network across the training data is calculated as a gradient vector. Once this vector is calculated, each network parameter takes a small step in the direction of this gradient, to incrementally increase the performance of the network, and through many of these steps the network becomes trained.

However in many cases, the entire corpus of training data is not used in each iteration but only a fraction of the corpus is sampled (as a 'batch' or 'minibatch'), and the average gradient vector of improved performance across the samples of the batch is calculated as an approximation of the true gradient vector. This iterative process can be broadly called SGD, where one of the hyperparameters is the size of the batches, see Keskar et al. (2017); Smith, Kindermans and Le (2018). In the context of supervised learning, each element of the training data is labelled with the desired output of the neural network for it, and these labels can serve to naturally stratify the training data; or the data can be stratified by other means too (Zhang, Kjellström and Mandt, 2017; Zhang et al., 2019; Zhao and Zhang, 2014). In this setting, Equation 12 may be used to choose between samples of labelled training data, in order to sample batches that more-efficiently estimate of the performance gradient, and hence improve the efficiency of neural network training procedure. This idea of 'smart sampling' for neural network training is not particularly new, and our method is compatible with other performance-enhancing techniques in the literature on neural networks (Papa, Bianchi and Clémençon, 2015; Clmenon et al., 2015).

Full exploration of these potential applications are beyond the scope of this document and are left for future work. However, at present, we are pleased to present our analytic concentration inequality (Equation 8) as an immediately computable expression and practical method for choosing samples from strata, and all sourcecode is available at:
https://github.com/Markopolo141/Stratified_Empirical_Bernstein_Sampling

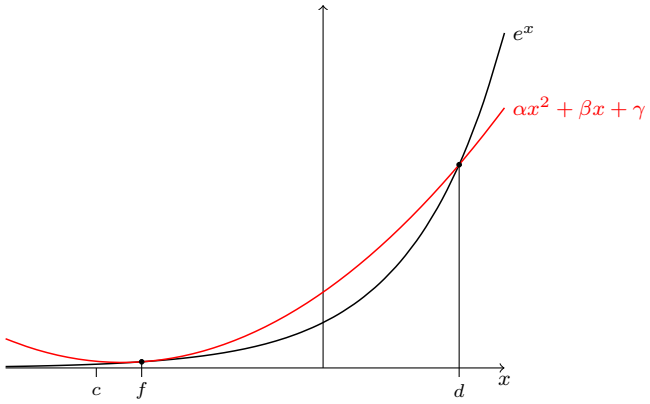

Fig 4: fitting a parabola above an exponential curve for all $c \leq x \leq d$

Appendix A: Parabola Fitting

In selecting an α, β, γ as the parameters[1] of a parabola $\alpha x^2 + \beta x + \gamma \geq \exp(x)$ for all $c \leq x \leq d$ which minimises $z\alpha + \gamma$ for constants c, d, z.
We witness that such a parabola may tangentially touch the exponential curve at one point (at $x = f < d$) and intersect it at another (at $x = d$), as illustrated in Figure 4.

The parabola's intersection at $x = d$ and its tangential intersection at $x = f$ can be written in matrix algebra:

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} d^2 & d & 1 \\ f^2 & f & 1 \\ 2f & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \exp(d) \\ \exp(f) \\ \exp(f) \end{bmatrix}$$

which gives our parabola parameters α, β, γ in terms of f and d, hence our objective function $z\alpha + \gamma$ can be written as:

$$z\alpha + \gamma = \frac{((z + fd - d)(f - d - 1) - d)e^f + (f^2 + z)e^d}{(d - f)^2}$$

since d is fixed, minimising with respect to f gives $f = \frac{-z}{d}$ where our objective function becomes:

$$z\alpha + \gamma = \frac{ze^d + d^2 e^{-z/d}}{z + d^2}.$$

[1] Here we carry the derivation with the explicit dependence on the s (seen in lemma 5) removed for simplicity.

References

Audibert, J.-Y., Munos, R. and Szepesvári, C. (2007). Tuning Bandit Algorithms in Stochastic Environments. In *Proc 18th Int. Conf. Algorithmic Learning Theory (ALT'07)* (M. Hutter, R. A. Servedio and E. Takimoto, eds.) 150–165. Springer Berlin Heidelberg, Berlin, Heidelberg.

Audibert, J.-Y., Munos, R. and Szepesvári, C. (2009). Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.* **410** 1876–1902.

Aziz, H., Cahan, C., Gretton, C., Kilby, P., Mattei, N. and Walsh, T. (2016). A study of proxies for shapley allocations of transport costs. *Journal of Artificial Intelligence Research* **56** 573-611.

Bachrach, Y., Markakis, E., Resnick, E., Procaccia, A. D., Rosenschein, J. S. and Saberi, A. (2009). Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems* **20** 105-122.

Bardenet, R. and Maillard, O.-A. (2015). Concentration inequalities for sampling without replacement. *Bernoulli* **21** 1361–1385.

Bennett, G. (1962). Probability Inequalities for the Sum of Independent Random Variables. *Journal of the American Statistical Association* **57** 33–45.

Bentkus, V. and van Zuijlen, M. (2003). On Conservative Confidence Intervals. *Lithuanian Mathematical Journal* **43** 141–160.

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pogoda, J. (2000). Exposure Stratified Case-Cohort Designs. *Lifetime Data Analysis* **6** 39–58.

Boucheron, S., Lugosi, G. and Massart, P. (2003). Concentration inequalities using the entropy method. *The Annals of Probability* **31** 1583-1614.

Breslow, N. E., Hu, J. and Wellner, J. A. (2015). Z-estimation and stratified samples: application to survival models. *Lifetime Data Analysis* **21** 493–516.

Burgess, M. A., Chapman, A. C. and Scott, P. (2018). The Generalized N&K Value:An Axiomatic Mechanism for Electricity Trading. *Proc. 17th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS'18)* **17** 1883–1885.

Carpentier, A., Lazaric, A., Ghavamzadeh, M., Munos, R. and Auer, P. (2011). Upper-confidence-bound Algorithms for Active Learning in Multi-armed Bandits. In *Proc. 22nd Int. Conf. Algorithmic Learning Theory (ALT'11)* 189–203. Springer-Verlag, Berlin, Heidelberg.

Castro, J., Gómez, D. and Tejada, J. (2009). Polynomial calculation of the Shapley value based on sampling. *Computers & OR* **36** 1726–1730.

Castro, J., Gmez, D., Molina, E. and Tejada, J. (2017). Improving polynomial estimation of the Shapley value by stratified random sampling with optimum allocation. *Computers & Operations Research* **82** 180 - 188.

Chalkiadakis, G., Elkind, E. and Wooldridge, M. (2012). *Computational Aspects of Cooperative Game Theory. Synthesis Lectures on Artificial Intelligence and Machine Learning.* Morgan and Claypool Publishers.

Chapman, A. C., Mhanna, S. and Verbič, G. (2017). Cooperative game theory for non-linear pricing of load-side distribution network support. In *Proc. 10th Bulk Power Systems Dynamics and Control Symposium (IREP'17)*.

Clmenon, S., Bellet, A., Jelassi, O. and Papa, G. (2015). Scalability of Stochastic Gradient Descent based on Smart Sampling Techniques. *Procedia Computer Science* **53** 308-315.

Efron, B. and Stein, C. (1981). The Jackknife Estimate of Variance. *Annals of Statistics* **9** 586–596.

Étoré, P. and Jourdain, B. (2010). Adaptive Optimal Allocation in Stratified Sampling Methods. *Methodology and Computing in Applied Probability* **12** 335–360.

Han, S., Tao, M., Topcu, U., Owhadi, H. and Murray, R. (2015). Convex Optimal Uncertainty Quantification. *SIAM Journal on Optimization* **25** 1368-1387.

Hillson, R., Alejandre, J. D., Jacobsen, K. H., Ansumana, R., Bockarie, A. S., Bangura, U., Lamin, J. M. and Stenger, D. A. (2015). Stratified Sampling of Neighborhood Sections for Population Estimation: A Case Study of Bo City, Sierra Leone. *PLoS ONE* **10** e0132850.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* **58** 13–30.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association* **47** 663–685.

Hu, W., Cai, J. and Zeng, D. (2014). Sample size/power calculation for stratified casecohort design. *Statistics in Medicine* **33** 3973–3985.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. and Tang, P. T. P. (2017). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *International Conference on Learning Representations*.

Khan, M. G. M., Ahmad, N. and Khan, S. (2009). Determining the Optimum Stratum Boundaries Using Mathematical Programming. *Journal of Mathematical Modelling and Algorithms* **8** 409–423.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition* **6** 797–806.

Legg, J. C. and Fuller, W. A. (2009). Chapter 3 - Two-Phase Sampling. In *Handbook of Statistics*, (C. R. Rao, ed.). *Handbook of Statistics* **29** 55 - 70. Elsevier.

Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T. and Rogers, A. (2013). Bounding the Estimation Error of Sampling-based Shapley Value Approximation. *arXiv e-prints* arXiv:1306.4265.

Maurer, A. (2006). Concentration inequalities for functions of independent variables. *Random Structures and Algorithms* **29** 121-138.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein Bounds and Sample Variance Penalization. stat. Conference On Learning Theory (COLT).

Michalak, T. P., Aadithya, K. V., Szczepanski, P. L., Ravindran, B.

and Jennings, N. R. (2013). Efficient Computation of the Shapley Value for Game-theoretic Network Centrality. *J. Artif. Int. Res.* **46** 607–650.

Miratrix, L. W. and Stark, P. B. (2009). Election Audits Using a Trinomial Bound. *IEEE Transactions on Information Forensics and Security* **4** 974–981.

Mnih, V., Szepesvári, C. and Audibert, J.-Y. (2008). Empirical Bernstein Stopping. In *Proc. 25th Int. Conf. Machine Learning. ICML '08* 672–679. ACM, New York, NY, USA.

Acuña, L. G., Ríos, D. R., Arboleda, C. P. and Ponzón, E. G. (2018). Cooperation model in the electricity energy market using bi-level optimization and Shapley value. *Operations Research Perspectives* **5** 161–168.

Neyman, J. (1938). Contribution to the Theory of Sampling Human Populations. *Journal of the American Statistical Association* **33** 101-116.

O'Brien, G., Gamal, A. E. and Rajagopal, R. (2015). Shapley Value Estimation for Compensation of Participants in Demand Response Programs. *IEEE Transactions on Smart Grid* **6** 2837–2844.

Owhadi, H., Scovel, C., Sullivan, T. J., McKerns, M. and Ortiz, M. (2013). Optimal Uncertainty Quantification. *SIAM Rev* **55** 271–345.

Papa, G., Bianchi, P. and Clémençon, S. (2015). Adaptive Sampling for Incremental Optimization Using Stochastic Gradient Descent. In *Proc. 25th Int. Conf. Algorithmic Learning Theory (ALT'15)* (K. Chaudhuri, C. GENTILE and S. Zilles, eds.) 317–331. Springer International Publishing, Cham.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73** 1-11.

Rehman, M. Z., Li, T. and Li, T. (2012). Exploiting empirical variance for data stream classification. *Journal of Shanghai Jiaotong University (Science)* **17** 245–250.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv e-prints* arXiv:1609.04747.

Saegusa, T. and Wellner, J. A. (2013). Weighted likelihood estimation under two-phase sampling. *The Annals of Statistics* **41** 269–295.

Serfling, R. J. (1974). Probability Inequalities for the Sum in Sampling without Replacement. *The Annals of Statistics* **2** 39–48.

Smith, S. L., Kindermans, P.-J. and Le, Q. V. (2018). Don't Decay the Learning Rate, Increase the Batch Size. In *International Conference on Learning Representations*.

Stark, P. B. (2009). Risk-limiting Postelection Audits: Conservative P-values from Common Probability Inequalities. *IEEE Transactions on Information Forensics and Security* **4** 1005–1014.

Thomas, P. S., Theocharous, G. and Ghavamzadeh, M. (2015). High-Confidence Off-Policy Evaluation. In *Proc. 29th AAAI Conf. Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. 3000–3006.

Wright, T. (2012). The Equivalence of Neyman Optimum Allocation for Sampling and Equal Proportions for Apportioning the U.S. House of Representatives. *The American Statistician* **66** 217–224.

Zhang, C., Kjellström, H. and Mandt, S. (2017). Stochastic Learning on

Imbalanced Data: Determinantal Point Processes for Mini-batch Diversification. In *Proc. Conf. Uncertainty in Artificial Intelligence (UAI'17)*.

Zhang, C., Öztireli, C., Mandt, S. and Salvi, G. (2019). Active Mini-Batch Sampling using Repulsive Point Processes. In *Proc. 33rd AAAI Conf. Artificial Intelligence (AAAI'19, accepted)*.

Zhao, P. and Zhang, T. (2014). Accelerating Minibatch Stochastic Gradient Descent using Stratified Sampling. *arXiv e-prints* arXiv:1405.3080.