 doi:10.20944/preprints201901.0202.v2

Stratified Finite Empirical Bernstein Sampling

Mark Alexander Burgess[a] and Archie C. Chapman[b]

[a]Australian National University, College of Engineering and Computer Science, Canberra;
[b]The University of Sydney, School of Electrical and Information Engineering, Sydney

ARTICLE HISTORY
Compiled May 31, 2019

ABSTRACT
We derive a concentration inequality for the uncertainty in the mean computed by stratified random sampling, and provide an online sampling method based on this inequality. Our concentration inequality is versatile and considers a range of factors including: the data ranges, weights, sizes of the strata, the number of samples taken, the estimated sample variances, and whether strata are sampled with or without replacement. Sequentially choosing samples to minimize this inequality leads to a online method for choosing samples from a stratified population. We evaluate and compare the effectiveness of our method against others for synthetic data sets, and also in approximating the Shapley value of cooperative games. Results show that our method is competitive with the performance of Neyman sampling with perfect variance information, even without having prior information on strata variances. We also provide a multidimensional extension of our inequality and discuss future applications.

KEYWORDS
Concentration Inequality; Empirical Bernstein Bound; Stratified Random Sampling; Shapley Value Approximation;

1. Introduction

Stratified sampling is a statistical method for estimating the mean of a population by partitioning it into mutually exclusive subgroups, or strata, and applying a sampling estimator to each stratum, before weighting and combining these estimates to form an estimate of the population mean. For example, to poll the population of a country's support for a particular government policy, we can selectively poll the different demographic regions within the country. For instance, if we know that regions A,B and C contain 10%, 40% and 50% of the population, and sampling shows support levels for a policy of 2%, 70% and 30%, respectively, then we can estimate that 43.2% of the total population supports the policy.

Stratified sampling can lead to improved reliability in estimation under certain conditions, such as: when the population is easily divided into strata, in which there is less variance in each stratum than across them all; when the size of the strata are known, and; when sampling selectively from each strata is possible [1,2]. If it is possible to sample selectively between the strata, then there is a further question of how to

ORCID: 0000-0003-1627-9125 and 0000-0002-5055-3004 respectively
Corresponding author: Mark Alexander Burgess. Email: markburgess1989@gmail.com

 doi:10.20944/preprints201901.0202.v2

conduct that sampling most efficiently.

In this paper we propose a process of sampling in order to maximally reduce the uncertainty in the population mean estimate, and to do this we develop a expression associated with that uncertainty. The expression takes the form of a *concentration inequality*, specifically, a *stratified empirical Berstein bound* (SEBB), 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 range of data of each strata; any additional importance weightings on the strata, and; whether any (or all) of the strata are sampled with or without replacement.

Using this inequality, we then propose an online method for sequentially sampling in order to maximally reduce this bound at each iteration, called the *stratified empirical Berstein method* (SEBM). By numerically evaluating the SEBM, we demonstrate its value in sampling and estimation applications. Moreover, we show that it can assist in computational tasks — particularly those that involve the calculation of expectation values, as sampling is a straightforward way of approximating them. Specifically, in this work, we consider the calculation of the *Shapley value*, a solution concept from cooperative game theory, as a computational task to which we can apply our sampling method and demonstration of its utility. Taken together, these results show that SEBM is competitive with the widely-used Neyman sampling method, that assumes perfect variance information, even when SEBM does not have prior information on strata variances.

The remainder of the paper is as follows. The next section reviews background material and sets the context of the paper. Section 3 provides several lemmas that are components of our derivation. Building on this, in Section 4 we provide the full derivation of our concentration inequality, SEBB, and online sampling method, SEBM, which are the main technical contributions of the paper. For evaluation, Section 5 examines the effectiveness of SEBM in the context of synthetic data sets, while Section 6 evaluates its performance on the task of approximating the Shapley value via sampling estimation. Section 7 discusses the results of these two sets of numerical evaluations, and analyze the reasons for the effectiveness of our sampling method. Finally, Section 8 gives an extension of SEBB and SEBM to multidimensional data. Section 9 concludes.

2. Background

Stratified sampling is a well known sampling technique in statistics and research, with many applications, including polling [3], auditing [4,5] and medical trials [6–8]. Stratified sampling is often conducted as a two-stage process, particularly when it is unclear how to stratify the population, or how large the resulting strata would be. Under this process, the first stage consists of sampling the population uniformly at random, and collecting the values of readily observable auxiliary variables, in order to estimate the sizes of potential strata by those variables. In the second stage, the strata are chosen and sampled with respect to the information gathered in the first stage, and the total population estimate is computed [9].

One well-known method of estimating efficient strata sizes is via *inverse probability weighting* methods such as the *Horvitz-Thompson estimator* [10]. However, these estimators are sometimes seen to perform quite badly in practice [11,12]. Moreover, these estimator do not directly address the prior problem of how to optimally break the population into strata based on the values of the auxiliary variables [3,13,14].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

In other situations, the strata and their sizes are naturally given, or the first stage of the two-stage process above may be assumed to have been conducted ideally. In these situation, there exists a further problem of how to allocate the finite number of 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) [1,2]. This approach assumes knowledge of the variances of the strata; however, in practice it is often the case that strata variance can only be estimated, either as part of the first stage or as the sampling proceeds [15,16].

Finally, even once the samples are taken from the various strata, there is still the question of how to compute appropriate confidence bounds on the final estimate. In the voting verification context, there exist some specialized bounds [5,17], but in the more general case there is some degree of discussion of what bounds should be used [4]. The confidence bounds that are derived critically depend on what assumptions are made about the underlying populations.

In particular, *Hoeffding's inequality* [18] is used as a bound under the assumption that the underlying population has bounded data support, or are drawn from a finite interval [4,17]. 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 [19] 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 (some similar bounds being published about that time, [20,21]). EBBs are is sensitive to the data width and sample variance. They have replaced Hoeffding's inequality in a number of computational applications [22–25]. The derivation of Maurer and Pontil's EBB has been extended to entropic [26] and Chernoff concentration inequalities, which are combined using union bounds.

Beyond this, sampling *without replacement* offers the opportunity to further tighten the bounds over the sampling-with-replacement case. For example, the refinement that is possible was first demonstrated by Serfling [27] with a martingale argument. More recently, Bardenet and Maillard [28] 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 a primary subject of this paper.

3. Preliminaries

We now state nine lemmas across the next three sections, which we use later to derive our stratified empirical Bernstein bound (SEBB) and method (SEBM). Specifically, in Section 3.1, we provide three lemmas, which show how probability bounds are related to the moment generating function, how probability bounds can be combined, and a useful algebraic relationship regarding the sample variance. Then, in Section 3.2, we provide three bounds on the moment generating functions of random variables. Last, in Section 3.3, we derive three lemmas that relate the moment generating function to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

the sample means of random variables with and without replacement.

3.1. Fundamental results

The first lemma is an often-used and rather weak result used to fuse simple statements of probability:

Lemma 3.1 (Probability Union)**.** *For any random variables a, b, c:*
$\mathbb{P}(a > c) \leq \mathbb{P}(a > b) + \mathbb{P}(b > c)$

This relationship is a well known and useful tool for settings where the probability relationship between a and c is unknown but the relationship between a and some b, and also between that b and c is known.

The next lemma is a straightforward result of algebra that relates the sample squares about the mean to the sample variance.

Lemma 3.2 (Variance Relation)**.** *Let X be a random variable with mean μ. For n samples of X, $\{x_k\}_{k=1,\dots,n}$, the sample mean, $\hat{\mu} = \frac{1}{n}\sum_k x_k$, biased sample variance, $\hat{\sigma}^2 = \frac{1}{n}\sum_k (x_k - \hat{\mu})^2$, and average of sample squares about the mean, $\hat{\sigma}_0^2 = \frac{1}{n}\sum_k (x_k - \mu)^2$, are related such that:*

$$\hat{\sigma}_0^2 - \hat{\sigma}^2 = (\hat{\mu} - \mu)^2 .$$

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, which extends directly from Markov's inequality and encompasses a range of inequalities called *Chernoff bounds*:

Lemma 3.3 (Chernoff Bound)**.** *For a random variable X, and for any $s > 0$ and t:*

$$\mathbb{P}(X \geq t) \leq \mathbb{E}\left[\exp(sX)\right] \exp(-st)$$

Many well-known inequalities follow from upper bounds for $\mathbb{E}\left[\exp(sX)\right]$, also known as the *moment generating function*.

3.2. Bounds on 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* [18] and is stated here without proof:

Lemma 3.4 (Hoeffding's Lemma)**.** *For a random variable X that is of finite support on the interval $a \leq X \leq b$, with width $D = b - a$, and for any $s > 0$:*

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

The second lemma is very much like Hoeffding's Lemma, except it involves information about the variance of the random variable. The proof of this result is included because it is useful in explaining our own approach.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

Lemma 3.5. *For a random variable X that is bounded on an interval $a \leq X \leq b$ with $D = b - a$ and variance σ^2, and any $s > 0$:*

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

Proof. We assume without loss of generality 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.

There exists an α, β, γ (see AppendixA) such that $\alpha s^2 X^2 + \beta s X + \gamma \geq \exp(sX)$, and for all $a \leq X \leq b$:

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

Where it follows that:

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

The expression in (1) is monotonically increasing with b, and $D > b$, therefore:

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

Given that for any $\kappa, x \geq 0$, that:

$$\log(\kappa\exp(x) + 1) \leq \log(\kappa + 1) + \frac{x\kappa}{\kappa + 1} + x^2\frac{\frac{1}{17} + \frac{\kappa}{2}}{(\kappa + 1)^2} \quad (2)$$

Thus letting $\kappa = \frac{\sigma^2}{D^2}$ and $x = s(D + \sigma^2/D)$ if follows that:

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

□

We note that this process of fitting a parabola over the exponential function bears a strong conceptual relationship with a famous bound developed by Hoeffding [18] and Bennett [29].

The third bound that we present, on the moment generating function, is similar again, however this time we consider the random variable X^2 instead of X.

Lemma 3.6. *Let X be a random variable of finite support 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 for any $q > 0$:*

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

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

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

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

If $d = \max(b, -a)$ we can choose $\gamma = 1$ and $\alpha = (\exp(-qd^2) - 1)d^{-2}$ (see figure A2), Thus:

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

Given that for any $0 \leq \kappa \leq 0.5$ and $x \leq 0$ that:

$$\log\left(\kappa\exp(x) - \kappa + 1\right) \leq \kappa x + \frac{1}{2}\kappa(1-\kappa)x^2$$

Letting $\kappa = \frac{\sigma^2}{D^2}$ and $x = -qD^2$, which is valid by Popoviciu's inequality [30] $\sigma^2 \leq D^2/4$, then:

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

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

The three inequalities above, Lemmas 3.4,3.5 and 3.6, are used in the derivation of our stratified sampling concentration inequality in Section 4.

3.3. The Moment Generating Function of Sample Means

In order to use the previous bonds on the moment generating function we need 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 state three further inequalities, where the first one (Lemma 3.7) is most appropriate for sampling with replacement, and the second and third (Lemma 3.8 and Lemma 3.9) can optionally be used in the context of sampling without replacement.

Lemma 3.7 (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 $\chi_m = \frac{1}{m}\sum_{i=1}^{m} X_i$ 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:*

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

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

Proof. By the independence of samples, we have:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

Thus:

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

These inequalities are sufficient for all the further derivations that we conduct. However, for the case of sampling without replacement, there is an alternative result that can be directly substituted, given in Lemma 3.9, below, which can be tighter in certain cases. We give its form and derivation, 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 3.7 can be used in the context of either sampling with or without replacement. In contrast, Lemma 3.9 can only be used when sampling without replacement. This distinction was shown to be true by Hoeffding [18], and is stated here without proof:

Lemma 3.8 (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:*

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

we now state an inequality regarding the moment generating function of the average of samples taken specifically *without replacement.*

When the sampling takes place without replacement the inequality of Lemma 3.7 can potentially be improved to take advantage of the finite size of the population. This inequality extends an important martingale inequality from [28]:

Lemma 3.9 (Martingale Bound)**.** *For finite data $x_1, x_2, \ldots x_n$ 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 $X_1, X_2, \ldots, X_n$ the random variables corresponding to the data sequentially drawn randomly without replacement, and χ_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:*

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

Proof. Observe that:

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

Then because:

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

we also have that:

$$\mathbb{E}[\exp(s\chi_m)] = \mathbb{E}\left[\prod_{k=m}^{n-1} \mathbb{E}\left[\exp\left(\frac{s}{k}(\chi_{k+1} - X_{k+1})\right) | \chi_{k+1} \cdots \chi_n\right]\right]$$

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

$$\mathbb{E}[X_{k+1} | \chi_{k+1} \cdots \chi_n] = \chi_{k+1},$$

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

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

by application of Lemma 3.2. Therefore:

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

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. We note that this result could be made much stronger by working around the use of Equation (4), but this comes at a cost of increased mathematical complexity.

Since Lemmas 3.9 and 3.7 share a common form, and because of Hoeffding's reduction (Lemma 3.8), all the derivations that follow that invoke Lemma 3.7 have direct analogues using Lemma 3.9 for the context of sampling without replacement. Note, however, that the bound without replacement (Lemma 3.9) may or may not be tighter than the bound with replacement (Lemma 3.7). However, the process of substituting one for the other can be done judiciously on a case-by-case basis to create the tightest possible bound. All the numerical results in this paper (relevant to sampling without replacement) have been produced with this judicious choice conducted.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

4. The Stratified Finite Empirical Bernstein Bound and Sampling Method

This section contains our main results: we derive a novel probability bound for the error of the stratified random sampling estimate, and use it to define a sequential stratified sampling algorithm. Before this, we begin by precisely defining the context of our derivations, to which our bound applies.

Definition 4.1 (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, denote values randomly and sequentially drawn (with or without) replacement by $X_{i,1}, X_{i,2}, \ldots, X_{i,n_i}$. Then, for the first m_i of these sample: (i) $\chi_{i,m_i} = \frac{1}{m_i}\sum_{j=1}^{m_i} X_{i,j}$ is their average; (ii) their biased sample variance is $\hat{\sigma}_i^2 = \frac{1}{m_i}\sum_{j}^{m_i}(X_{i,j}-\chi_{i,m_i})^2$, and; (iii) their unbiased sample variance is $\hat{\hat{\sigma}}_i^2 = m_i\hat{\sigma}_i^2/(m_i-1)$. We are interested in the average of the means of the strata as weighted by constant positive factors $\{\tau_i\}_{i\in\{1,\ldots,n\}}$. Throughout our derivation, we temporarily use arbitrary positive variables $\{\theta_i\}_{i\in\{1,\ldots,n\}}$.

Given this context, the following two sections contain the derivation of the stratified empirical Berstein bound (SEBB) and the sequential sampling method (SEBM), respectively.

4.1. Bound derivation

The bound is now developed in four theorems, which build on each other in sequence:

(1) Theorem 4.2 develops a concentration inequality for the error in the stratified population mean estimate $\sum_{i=1}^{n}\tau_i\chi_{i,m_i}$ in the context of variance information.
(2) Theorem 4.3 is a concentration inequality of the difference between the stratum variances and sample variances in the context the sum of squared stratum mean errors.
(3) Theorem 4.4 is an inequality directly about the sum of sample squared stratum mean errors.
(4) Theorem 4.5 combines the three previous theorems together using union bounds to create a concentration inequality for the error in the stratified population mean estimate given the sample variances.

We begin with an expression for a probability bound on the absolute error of the weighted stratified sample means about the weighted strata means, which we call a variance-assisted SEBB (stratified empirical Bernstein bound).

Theorem 4.2 (Variance-assisted SEBB)**.** *Assuming the context given in Definition 4.1, and let $\Omega_{m_i}^{n_i}$ and $\Psi_{m_i}^{n_i}$ be given as in Lemma 3.7, then:*

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

Proof. Applying Lemma 3.3:

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

by independence of the sampling between the strata. This form is sufficient for Lemma 3.7 with Lemma 3.5 to apply, resulting in a double-sided tail bound:

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

Minimizing 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 weighted sum of variances must be estimated from the data collected, and to include this factor we develop and incorporate a probability bound for the estimate of the sum of variances (as weighted by arbitrary θ_i), as follows.

Theorem 4.3. *Assuming the context given in Definition 4.1. Then with $\Psi_{m_i}^{n_i}$ per Lemma 3.7:*

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

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 3.3 gives:

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

by independence of the sampling between the strata. This is sufficient for Lemma 3.7

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

with Lemma 3.6 to apply, giving:

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

Minimizing with respect to s, rearranging, and applying Lemma‘3.2 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 the weighted square error of the sample means may go to zero quickly as additional samples are taken, we nonetheless develop another probability bound to incorporate specific consideration of it.

Theorem 4.4. *Assuming the context given in Definition 4.1. Then with $\Omega_{m_i}^{n_i}$ as in Lemma 3.7:*

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

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

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

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

such that $\sum r_i = r$, by independence of the sampling and probability complementarities. This is sufficient for us to apply Lemma 3.3 together with Lemmas 3.7 and 3.4, giving:

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

Next, choosing r_i to minimize this expression gives:

$$r_i=\frac{r\theta_iD_i^2\Omega_{m_i}^{n_i}}{\sum_j\theta_jD_j^2\Omega_{m_j}^{n_j}}$$

Thus:

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

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 of the SEBB.

Theorem 4.5 (Stratified Empirical Bernstein Bound (SEBB))**.** *Assuming the context given in Definition 4.1. Then with* $\Omega_{m_i}^{n_i}, \Psi_{m_i}^{n_i}$ *per Lemma 3.7:*

$$\mathbb{P}\left(\frac{\left|\sum_{i=1}^{n}\tau_i(\chi_{i,m_i}-\mu_i)\right|}{\sqrt{\log(6/p)}} \geq \sqrt{\alpha+\left(\sqrt{\beta}+\sqrt{\gamma}\right)^2}\right) \leq p \tag{8}$$

where:

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

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

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

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

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

Combining with Equation (7) with a union bound (Lemma 3.1) gives:

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

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 3.1), and then assigning $r = t = y = p/3$ and rewriting in terms of unbiased sample variance, gives the result. □

This completes the derivation. In Equation (8) of Theorem 4.5, we have a concentration inequality for the sum of weighted strata sample mean errors relative to the sample variances. In this context, the weights τ_i are flexible but would naturally be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

probability weights proportional to strata size, $\tau_i = n_i/(\sum_{j=1}^{n} n_j)$, in which case the inequality provides a concentration of measure in stratified random sampling. Based on this bound, we proceed to propose an online process of sequentially choosing samples from the strata in order maximally minimize it.

4.2. Sequential Sampling Using the Stratified Empirical Berstein Method

We introduce a method of sampling, the *stratified empirical Bernstein method* (SEBM) which sequentially minimizes the bound in Theorem 4.5 (SEBB). Pseudocode for the calculation of the bound and the process of sampling to minimize it, is given in Algorithm 1.

Specifically, Algorithm 1 is a repetitive process involving a scan through the possible strata and then the selection of one stratum to sample from to minimize the SEBB under mild assumptions. The process of scanning involves calculating the confidence bound width (SEBB) that would result if an additional sample were to be taken from that stratum without changing its sample variance (line numbers 5-17 in Algorithm 1). The stratum that yields the smallest confidence bound width in the context of an additional sample is then selected (line 18-21) and sampled (line 24), the sample variance of that stratum is updated (line 26); this process repeats until the maximum sample budget is reached (per the outer loop, line 1). In this way the process attempts to iteratively minimize the SEBB in expectation with each additional sample taken; and hence lead to potentially greater accuracy in stratified sampling as a result.

We note that computing the SEBB requires the sample variances of all the strata having been calculated. Accordingly, Algorithm 1 must be initialized with at least two samples from each stratum so that sample variance can be calculated. This is a standard requirement of the many reinforcement learning algorithms that use variance in their sampling policies.

Algorithm 1 describes a process specific to sampling without replacement and involves the calculation of the SEBB with the tightest possible uses of Lemmas 3.9 and 3.7. In particular, for any stratum i that is sampled without replacement, any specific bound with an 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, and this corresponds to choice of Lemma 3.9 or Lemma 3.7 in the bound's derivation. Since the SEBB is a composition of such bounds with such choices throughout, there is a structure of valid pairs of substitutions Ω, Ψ for $\bar{\Omega}, \bar{\Psi}$ in the optimal calculation of the SEBB, which is shown in the steps 8-15 of Algorithm 1. The equivalent algorithm for sampling with replacement simply is the same algorithm altered by replacing all use of $\bar{\Omega}, \bar{\Psi}$ with Ω, Ψ.

In the next Section 5 we evaluate the performance of SEBM with other methods in the context of synthetic data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

Algorithm 1 Stratified Empirical Bernstein Method (SEBM) with replacement

Require: probability p, strata number N, stratum sizes n_i, initial sample numbers m_i, initial stratum sample variances $\hat{\hat{\sigma}}_i^2$, weights τ_i, widths D_i, maximum sample budget B

while $\sum_i m_i < B$ **do**
  $beststrata \leftarrow -1$
  $lowestbound \leftarrow \infty$
  for $k = 0$ to N **do**
    $m_k \leftarrow m_k + 1$
    $a \leftarrow [0,0]$, $b \leftarrow [0,0]$, $c \leftarrow [0,0]$, $d \leftarrow [0,0]$
    for $i = 0$ to N **do**
      $a_0 \leftarrow a_0 + \log(6N/p) D_i^2 \bar{\Psi}_{m_i}^{n_i} \min(\bar{\Omega}_{m_i}^{n_i}, \Omega_{m_i}^{n_i}) \tau^2$
      $a_1 \leftarrow a_1 + \log(6N/p) D_i^2 \Psi_{m_i}^{n_i} \min(\bar{\Omega}_{m_i}^{n_i}, \Omega_{m_i}^{n_i}) \tau^2$
      $b_0 \leftarrow \max(b_0, \log(3/p) D_i^2 \bar{\Psi}_{m_i}^{n_i} \min(\bar{\Psi}_{m_i}^{n_i}, \Psi_{m_i}^{n_i}) \tau^2)$
      $b_1 \leftarrow \max(b_1, \log(3/p) D_i^2 \Psi_{m_i}^{n_i} \min(\bar{\Psi}_{m_i}^{n_i}, \Psi_{m_i}^{n_i}) \tau^2)$
      $c_0 \leftarrow c_0 + 2\bar{\Psi}_{m_i}^{n_i}((m_i - 1)\hat{\hat{\sigma}}_i^2/m_i)\tau^2$
      $c_1 \leftarrow c_1 + 2\Psi_{m_i}^{n_i}((m_i - 1)\hat{\hat{\sigma}}_i^2/m_i)\tau^2$
      $d_0 \leftarrow d_0 + \frac{4}{17} D_i^2 \bar{\Omega}_{m_i}^{n_i} \tau^2$
      $d_1 \leftarrow d_1 + \frac{4}{17} D_i^2 \Omega_{m_i}^{n_i} \tau^2$
    end for
    $boundwidth \leftarrow \sqrt{\log(6/p) \min_j (d_j + (\sqrt{c_j + a_j + b_j} + \sqrt{b_j})^2)}$
    if $boundwidth < lowestbound$ **then**
      $beststrata \leftarrow k$
      $lowestbound \leftarrow boundwidth$
    end if
    $m_k \leftarrow m_k - 1$
  end for
  take an extra sample from strata: $beststrata$
  $m_{beststrata} \leftarrow m_{beststrata} + 1$
  recalculate $\hat{\hat{\sigma}}_{beststrata}^2$
end while

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

5. Numerical Evaluation

In this section assess the value of SEBM as an online method of sampling from stratified data. First we outline the benchmark algorithms used to evaluate our method's performance. Then in Section 5.2 we describe two synthetic data sets and report the distribution of errors under our method and the benchmarks. Following this, in Section 6, we evaluate our method in an example application — that of calculating the Shapley value of a cooperative game. Discussion and analysis of all the numerical results is left to Section 7.

5.1. Benchmarks algorithms

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

- SEBM (Stratified empirical Bernstein method, without replacement): our SEBM method (per Algorithm 1) of iteratively choosing samples from strata to minimize the SEBB, given in Equation (8). An initial sample of two data points from each strata is used to initialize the sample variances of each, with additional samples made to maximally minimize the inequality at each step. All samples are drawn *without* replacement.
- 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 3.9.
- Sim (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 (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 assistance): the method of iteratively choosing samples *without* replacement from strata to minimize Equation (5), utilizing martingale Lemma 3.9.

Note that the last three methods (Ney,Ney-W and SEBM*) assume and utilize prior perfect knowledge of the variance of each of the strata, and that for methods SEBM, SEBM-W and SEBM* (which use Equations (8) and (5)) we selected for minimising a 50% confidence interval (i.e. constant $p = 0.5$ and $t = 0.5$).

Also note that these methods provide comparisons of different algorithm 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 for a case of uniform-and-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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

methods of choosing finite sequences of samples. In this section, we described the two types of synthetic data sets used in this evaluation, namely: (i) beta distributed stratum data, which are intended reflect real-world data, and (ii) a particular form of uniform and Bernoulli distributed stratum data, in which our sampling method (SEBM) performs poorly.

5.2.1. Beta-Distributed Data

The first pool of synthetic data 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, with classic probability density function:

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

with α and β parameters drawn uniformly between 0 and 4 for each stratum, and Γ is gamma function.

Figure 1 compares the distribution of absolute error achieved by each of the sampling methods over 5000 rounds of these data sets. Each panel 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 plots in Figure 1, we can see that our sampling technique (SEBM and SEBM-W) performs comparably to Neyman sampling (Ney and Ney-W) despite not having access to knowledge of stratum variances. Also, there is a notable similarity between SEBM* and SEBM. 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.

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) performs poorly, particularly compared to Neyman sampling (Ney). 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 Bernoulli distributed, with all zeros except for a specified small number, a, of successes with value 1. For this problem, we conduct stratified random sampling with a budget of 300 samples, comparing the SEBM*, SEBM and Ney methods. The average error returned by the methods across 20,000 realizations of this problem, plotted against the number of successes a, are shown as a graph in Figure 2.

This figure demonstrates that SEBM and SEBM* perform poorly when the strata contain only very small numbers of successes. This under-performance is not simply a result of the SEBM 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

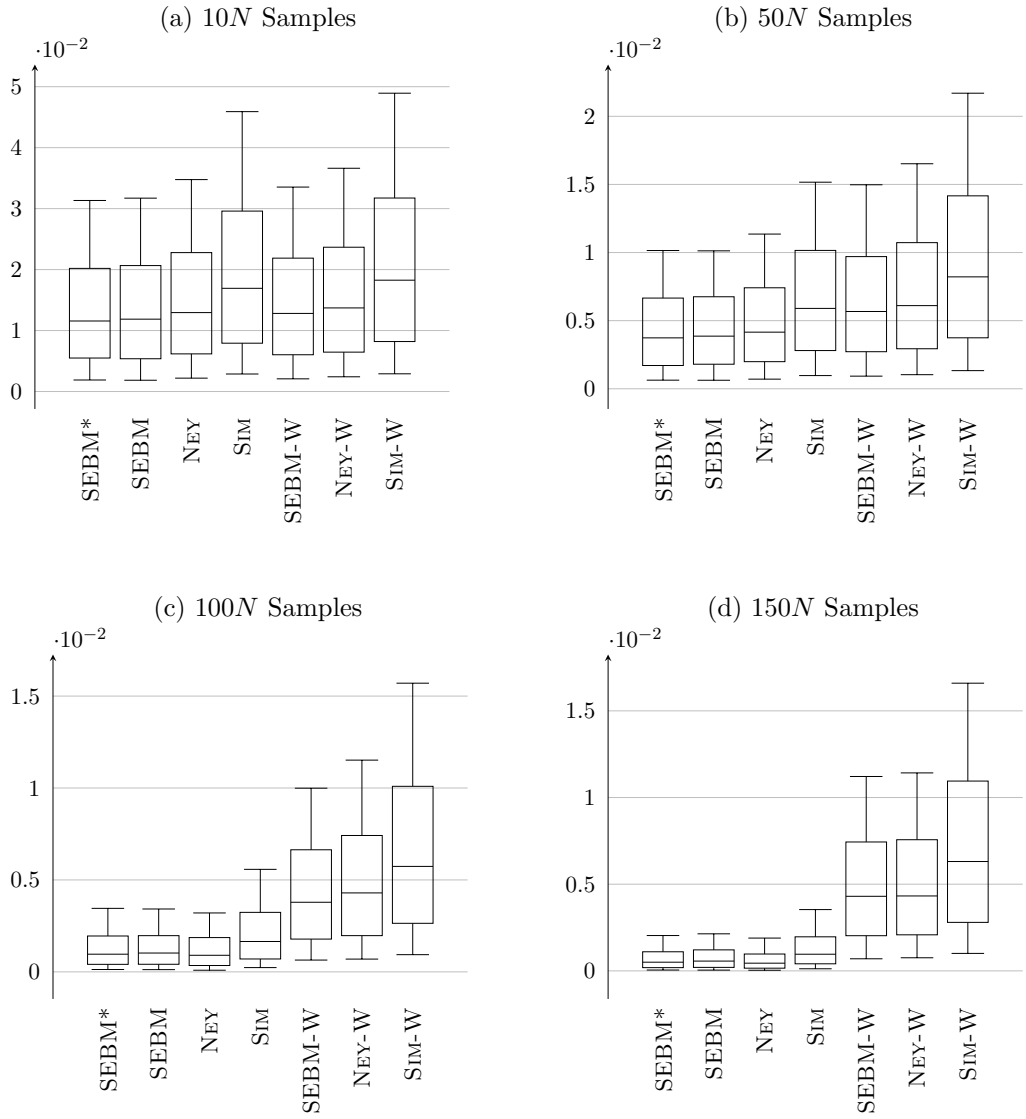

Figure 1.: Distribution of numerical absolute errors across 5000 rounds of beta-distributed data, for different methods of stratified sampling. Each plot shows absolute errors for different numbers of samples multiplied by the number of strata, N, e.g. $10N$ samples means that the test problem has a 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

Section 7. Before this, we considered the calculation of the Shapley value as an example computational application of our stratified sampling method.

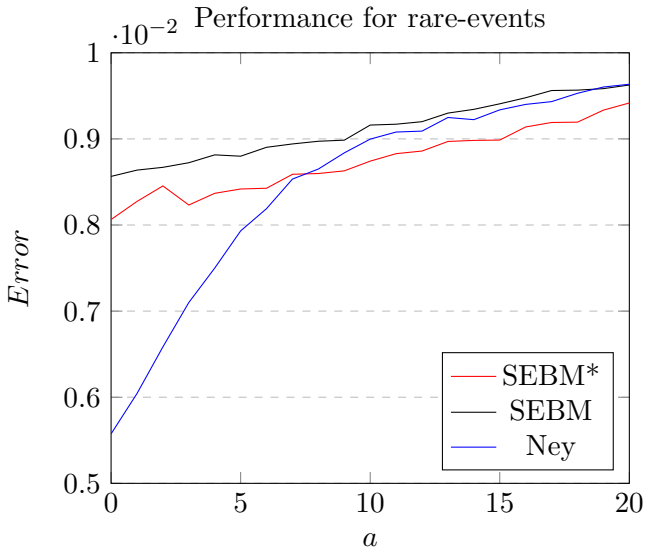

Figure 2.: Average error of three stratified random sampling methods for the uniform-Bernoulli data sets of Section 5.2.2, plotted against success parameter a, across 20,000 rounds.

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 surplus or profit a group or "coalition" of participants could achieve by themselves [31]. It has many applications, including analyzing the power of voting blocks in weighted voting games [32], in cost and surplus division problems [33,34], and as a measure of network centrality [35].

Formally, a *cooperative game*, $\langle N, v\rangle \in \mathbb{G}_N$, 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 φ is a unique mapping from cooperative games to the player rewards $\mathbb{G}_N \to \mathbb{R}^n$ which satisfies axioms:

- **Efficiency**: That the total reward is divided up: $\sum_i \varphi_i(\langle N, v\rangle) = v(N)$
- **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 $\varphi_i(\langle N, v\rangle) = \varphi_i(\langle N, v\rangle)$
- **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 $\varphi_i(\langle N, v\rangle) = 0$
- **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: $\varphi(\langle N, v_1 + v_2\rangle) = \varphi(\langle N, v_1\rangle) + \varphi(\langle N, v_2\rangle)$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

Specifically, the Shapley value is expressed as:

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

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

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

then the Shapley value can be expressed as an average:

$$\varphi_i(\langle N, v\rangle) = \frac{1}{n} \sum_{k=0}^{n-1} v_{i,k} \tag{11}$$

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 these averages are naturally stratified by coalition size. In previously published literature, other techniques have been used to allocate samples in this context, particularly simple sampling [36], Neyman allocation [16,37], and allocation to minimize Hoeffding's inequality [38].

We assess the benefits of using our bound by comparing its performance to the approaches above in the context of some example cooperative games, with results analyzed in Section 7. The example games are 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:

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

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 charac-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

teristic 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 absolute errors in the approximated Shapley value for a given budget m of marginal-contribution samples across multiple computational runs. The results are shown in Table 1, where the average absolute error in the Shapley value is computed by sampling with Maleki's method [38] is denoted e^{Ma}, e^{sim} is Castro's stratified simple sampling method [36], e^{Ca} is Castro's Neyman sampling method [37], and e^{SEBM} is the error associated with our method, SEBM. 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.

7. Discussion

In this section we give considerations to the numerical results of the paper. In general, from the results across Figures 1 and 2 and Table 1, the main observation is that our sampling technique, SEBM or SEBM-W, performs competitively to Neyman sampling (Ney 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* to Ney, which both have access to strata variances, then both methods use the same information and the results are even stronger for our method. The reasons for this performance are interesting, and we now discuss them in more detail.

From Figure 1 we 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 even 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

(a) Airport Game Average Errors

m/n^2	10	50	100	500	1000
e^{Ma}	298.4	133.1	99.64	41.96	29.26
e^{sim}	357.8	146.1	106.2	44.55	36.33
e^{Ca}	325.7	115.8	75.85	31.01	22.12
e^{SEBM}	259.2	73.8	54.76	7.71	1.30

(b) Voting Game Average Errors

m/n^2	10	50	100	500	1000
e^{Ma}	131.0	57.78	41.52	18.66	13.18
e^{sim}	145.7	59.72	40.31	17.56	12.84
e^{Ca}	142.1	47.35	31.05	14.08	9.800
e^{SEBM}	122.8	47.44	33.18	8.55	1.995

(c) Simple Reward Division Game average errors

m/n^2	10	50	100	500	1000
e^{Ma}	25.68	11.62	7.792	3.481	2.290
e^{sim}	22.10	9.045	6.218	2.642	1.938
e^{Ca}	22.37	8.925	6.692	2.727	1.940
e^{SEBM}	19.25	7.044	5.158	1.183	0.2817

(d) Complex Reward Division Game average errors

m/n^2	10	50	100	500	1000
e^{Ma}	276.1	118.9	87.00	40.15	27.44
e^{sim}	251.4	108.0	78.63	34.64	26.82
e^{Ca}	290.5	116.5	81.82	35.70	26.50
e^{SEBM}	214.2	78.47	54.10	12.45	2.711

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).

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 in the case of high number of samples and sampling without replacement, as illustrated in Figure 1. This indicates that the use of sub-optimal equation 4 in the derivation of Lemma 3.9 might have some negative effect, by distorting the shape of the functions that the sampling processes then minimizes.

Furthermore, if the data features very rare events, then SEBM 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). This shortcoming can be partly explained by noting that SEBM unnecessarily wastes samples on the Bernoulli stratum of rare events in the process of learning that the variance is almost zero, whereas Ney can avoid this because it has prior knowledge of the variances to begin with. As such, this factor explains the difference between the performance of SEBM and SEBM* in the context of Figure 1 and also in Figure 2. What is surprising is how small

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

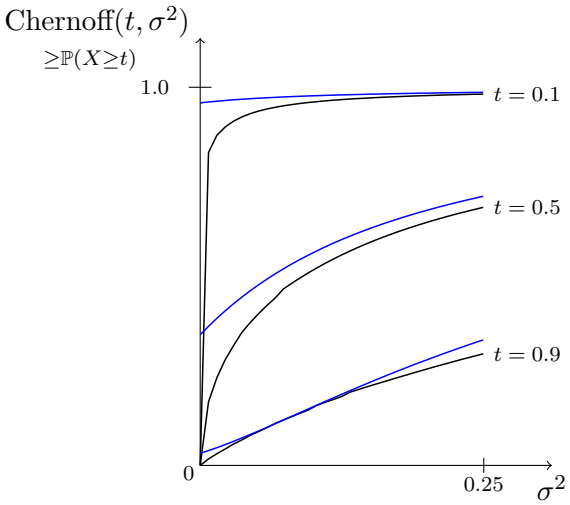

Figure 3.: Chernoff upper bounds derived directly from the moment generating functions of equations Equation (1) and (3) (in black and blue, resp.) with $D = 1$; Plotted for various t against the variance σ^2. Note that Equation (3) generally captures the shape and magnitude of the more accurate equation, except in the region of small σ^2 where the bound is overly weakened.

the difference in performance between SEBM and SEBM* is. This indicates that as additional samples are taken, the original 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 in Figure 2 is not explained by this argument, as they use the same information. Instead, the reason for this difference in performance is found by considering the simplifying approximation of Equation (2) in the derivation of Lemma 3.5. Specifically, (2) introduces a particular distortion into the shape of Equation (8) which our sampling seeks to minimize. 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, 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 minimizing (one sample at a time) our inequality in the context of our simple example games. This overhead can be a significant drawback, however on more complicated games such as where the characteristic function is slower to calculate, any overhead associated with the sampling choice is expected to 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

In practice, it might also 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.

8. Multidimensional Extension

Our 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, let $X_{i,1}, X_{i,2}, \ldots, X_{i,n_i}$ (where $X_{i,k,j}$ is the jth component, of the kth vector from stratum i) be vector random variables corresponding to those data values randomly and sequentially drawn (with or without) replacement. Denote the average of the first m_i of these random variables from the ith stratum by $\chi_{i,m_i} = \frac{1}{m_i}\sum_{k=1}^{m_i} X_{i,k}$ (with $\chi_{i,m_i,j}$ being the jth component of that vector average). Let $\hat{\hat{\sigma}}^2_{i,j} = \frac{i}{m_i - 1}\sum_{k=1}^{m_i}(X_{i,k,j} - \chi_{i,m_i,j})^2$ be the unbiased sample variance of the first m_i of these random variables in the jth component. As before, we assume weights τ_i for each stratum.
In this context we have the following theorem:

Theorem 8.1 (Vector SEBM bound)**.** *In the context above, then with* $\Omega^{n_i}_{m_i}, \Psi^{n_i}_{m_i}$ *per Lemma 3.7:*

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

where:

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

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

$$\mathbb{P}\left(\frac{\left(\sum_{i=1}^{n}\tau_i(\chi_{i,m_i}-\mu_i)\right)^2}{\log(6/p)} \geq \alpha_j + \left(\sqrt{\beta_j} + \sqrt{\gamma_j}\right)^2\right) \leq p$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

Taking a series of union bounds (Lemma 3.1) 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 $\sum_{i=1}^{n} \tau_i \chi_{i,m_i}$ and the true mean stratified vector $\sum_{i=1}^{n} \tau_i \mu_i$. In this context, an example sampling process might consist of sampling to maximally minimize the right hand side of the inequality (similar to our SEBM process, described in Section 4.2). This formulation can be applied to more involved computational tasks that involve sampling data with multiple features or auxiliary variables.

9. Conclusions and Future Work

The derivation of our inequality extends from consideration of Chernoff bounds and probability unions in a similar vein to other EBB derivations [19,28]. However, the bounds on the moment generating functions that we developed in Section 3 use loosening approximations, and hence stronger and/or more representative bounds could be developed at the cost of greater mathematical complexity. Alternatively, an approach utilizing entropic [39] or Efron-Stein inequalities [40] could result in different and potentially tighter results.

We now consider prospective applications of our 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). Two examples of this are the problems of pricing (i) logistics, involving solutions to the travelling salesman problem [33], and (ii) electricity networks, which requires solving optimzation problems that incorporate the power flow equations [16,34,41,42]. Focusing on electricity networks in particular: these are complicated technical system used to transport electrical power from generators to loads, subject to the non-linear 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 [16], demand or load [34], supply or generation [41], and potentially all simultaneously [42]. 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 [43]. 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 has been called SGD, where one of the hyperparameters is the size of the batches, see [44,45]. 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 [46–48]. 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 potentially compatible with other performance-enhancing techniques in the literature on neural networks [49,50].

Sourcecode for all the experiments in this paper are available at:
https://github.com/Markopolo141/Stratified_Empirical_Bernstein_Sampling

Acknowledgements

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

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Neyman J. Contribution to the theory of sampling human populations. Journal of the American Statistical Association. 1938;33(201):101–116.
[2] Wright T. The equivalence of neyman optimum allocation for sampling and equal proportions for apportioning the U.S. house of representatives. The American Statistician. 2012;66(4):217–224.
[3] Hillson R, Alejandre JD, Jacobsen KH, et al. Stratified sampling of neighborhood sections for population estimation: A case study of Bo city, Sierra Leone. PLoS ONE. 2015 Jul; 10(7):e0132850.
[4] Stark PB. Risk-limiting postelection audits: Conservative p-values from common probability inequalities. IEEE Transactions on Information Forensics and Security. 2009 Dec; 4(4):1005–1014.
[5] Miratrix LW, Stark PB. Election audits using a trinomial bound. IEEE Transactions on Information Forensics and Security. 2009 Dec;4(4):974–981.
[6] Hu W, Cai J, Zeng D. Sample size/power calculation for stratified casecohort design. Statistics in Medicine. 2014 1;33(23):3973–3985.
[7] Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986;73(1):1–11.
[8] Borgan O, Langholz B, Samuelsen SO, et al. Exposure stratified case-cohort designs. Lifetime Data Analysis. 2000 Mar;6(1):39–58.

[9] Legg JC, Fuller WA. Two-phase sampling. In: Rao C, editor. Handbook of statistics. (Handbook of Statistics; Vol. 29); Chapter 3. Elsevier; 2009. p. 55 – 70.
[10] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1952;47(260):663–685.
[11] Saegusa T, Wellner JA. Weighted likelihood estimation under two-phase sampling. The Annals of Statistics. 2013 02;41(1):269–295.
[12] Breslow NE, Hu J, Wellner JA. Z-estimation and stratified samples: application to survival models. Lifetime Data Analysis. 2015 Oct;21(4):493–516.
[13] Khan MGM, Ahmad N, Khan S. Determining the optimum stratum boundaries using mathematical programming. Journal of Mathematical Modelling and Algorithms. 2009; 8(4):409–423.
[14] Kozak M. Optimal stratification using random search method in agricultural surveys. Statistics in Transition. 2004;6(5):797–806.
[15] Étoré P, Jourdain B. Adaptive optimal allocation in stratified sampling methods. Methodology and Computing in Applied Probability. 2010 Sep;12(3):335–360.
[16] O'Brien G, Gamal AE, Rajagopal R. Shapley value estimation for compensation of participants in demand response programs. IEEE Transactions on Smart Grid. 2015;6(6):2837–2844.
[17] Bentkus V, van Zuijlen M. On conservative confidence intervals. Lithuanian Mathematical Journal. 2003 Apr;43(2):141–160.
[18] Hoeffding W. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association. 1963 Mar;58(301):13–30.
[19] Maurer A, Pontil M. Empirical bernstein bounds and sample variance penalization. In: Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009); June; 2009.
[20] Audibert JY, Munos R, Szepesvári C. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science. 2009;410(19):1876–1902.
[21] Audibert JY, Munos R, Szepesvári C. Tuning bandit algorithms in stochastic environments. In: Hutter M, Servedio RA, Takimoto E, editors. Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT'07); Berlin, Heidelberg. Springer Berlin Heidelberg; 2007. p. 150–165.
[22] Rehman MZ, Li T, Li T. Exploiting empirical variance for data stream classification. Journal of Shanghai Jiaotong University (Science). 2012 Apr;17(2):245–250.
[23] Mnih V, Szepesvári C, Audibert JY. Empirical bernstein stopping. In: Proceedings of the 25th International Conference on Machine Learning (ICML); New York, NY, USA. ACM; 2008. p. 672–679; ICML '08.
[24] Thomas PS, Theocharous G, Ghavamzadeh M. High-confidence off-policy evaluation. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.; 2015. p. 3000–3006.
[25] Carpentier A, Lazaric A, Ghavamzadeh M, et al. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In: Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT'11); Berlin, Heidelberg. Springer-Verlag; 2011. p. 189–203.
[26] Maurer A. Concentration inequalities for functions of independent variables. Random Structures and Algorithms. 2006;29(2):121–138.
[27] Serfling RJ. Probability inequalities for the sum in sampling without replacement. The Annals of Statistics. 1974 01;2(1):39–48.
[28] Bardenet R, Maillard OA. Concentration inequalities for sampling without replacement. Bernoulli. 2015 08;21(3):1361–1385.
[29] Bennett G. Probability inequalities for the sum of independent random variables. Journal of the American Statistical Association. 1962;57(297):33–45.
[30] Sharma R, Gupta M, Kapoor G. Some better bounds on the variance with applications. Journal of Mathematical Inequalities. 2010;4(3):355–363.
[31] Chalkiadakis G, Elkind E, Wooldridge M. Computational aspects of cooperative game

theory. Morgan and Claypool Publishers; 2012. Synthesis Lectures on Artificial Intelligence and Machine Learning.

[32] Bachrach Y, Markakis E, Resnick E, et al. Approximating power indices: theoretical and empirical analysis. Autonomous Agents and Multi-Agent Systems (AAMAS). 2009; 20:105–122.

[33] Aziz H, Cahan C, Gretton C, et al. A study of proxies for shapley allocations of transport costs. Journal of Artificial Intelligence Research. 2016;56:573–611.

[34] Chapman AC, Mhanna S, Verbič G. Cooperative game theory for non-linear pricing of load-side distribution network support. In: Proceedings of the 10th Bulk Power Systems Dynamics and Control Symposium (IREP'17); 2017.

[35] Michalak TP, Aadithya KV, Szczepanski PL, et al. Efficient computation of the shapley value for game-theoretic network centrality. Journal of Artificial Intelligence Research. 2013 Jan;46(1):607–650.

[36] Castro J, Gómez D, Tejada J. Polynomial calculation of the shapley value based on sampling. Computers & OR. 2009;36(5):1726–1730.

[37] Castro J, Gómez D, Molina E, et al. Improving polynomial estimation of the shapley value by stratified random sampling with optimum allocation. Computers & Operations Research. 2017;82:180 – 188.

[38] Maleki S, Tran-Thanh L, Hines G, et al. Bounding the Estimation Error of Sampling-based Shapley Value Approximation. arXiv e-prints. 2013 Jun;:arXiv:1306.4265.

[39] Boucheron S, Lugosi G, Massart P. Concentration inequalities using the entropy method. The Annals of Probability. 2003;31(3):1583–1614.

[40] Efron B, Stein C. The jackknife estimate of variance. Annals of Statistics. 1981 05; 9(3):586–596.

[41] Acuña LG, Ríos DR, Arboleda CP, et al. Cooperation model in the electricity energy market using bi-level optimization and shapley value. Operations Research Perspectives. 2018;5:161–168.

[42] Burgess MA, Chapman AC, Scott P. The generalized N&K value: An axiomatic mechanism for electricity trading. Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'18). 2018 Jul;17:1883–1885.

[43] Ruder S. An overview of gradient descent optimization algorithms. arXiv e-prints. 2016 Sep;:arXiv:1609.04747.

[44] Keskar NS, Mudigere D, Nocedal J, et al. On large-batch training for deep learning: Generalization gap and sharp minima. In: Proceedings of the 5th International Conference on Learning Representations (ICLR); April; 2017.

[45] Smith SL, Kindermans PJ, Le QV. Don't decay the learning rate, increase the batch size. In: Proceedings of the 6th International Conference on Learning Representations (ICLR); April; 2018.

[46] Zhang C, Kjellström H, Mandt S. Stochastic learning on imbalanced data: Determinantal point processes for mini-batch diversification. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI'17); 2017.

[47] Zhang C, Öztireli C, Mandt S, et al. Active mini-batch sampling using repulsive point processes. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19, accepted); 2019.

[48] Zhao P, Zhang T. Accelerating Minibatch Stochastic Gradient Descent using Stratified Sampling. arXiv e-prints. 2014 May;:arXiv:1405.3080.

[49] Papa G, Bianchi P, Clémençon S. Adaptive sampling for incremental optimization using stochastic gradient descent. In: Proceedings of the 25th International Conference on Algorithmic Learning Theory (ALT'15). Springer International Publishing; 2015. p. 317–331.

[50] Clémençon S, Bellet A, Jelassi O, et al. Scalability of stochastic gradient descent based on smart sampling techniques. Procedia Computer Science. 2015 12;53:308–315.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

Appendix A. Parabola Fitting

Theorem A.1. *For $b > 0$ and $a < b$ and $z > 0$, there exists an α, β, γ such that: $\alpha x^2 + \beta x + \gamma \geq \exp(x)$ for all $a \leq x \leq b$, and where:*

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

.

Proof. An example parabola $\alpha x^2 + \beta x + \gamma$ which that satisfies these requirements tangentially touches the exponential curve at one point (at $x = f < b$) and intersects it at another (at $x = b$), as illustrated in Figure A1. Thus the parabola's intersection at $x = b$ and its tangential intersection at $x = f$ can be written in matrix algebra:

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

This gives our parabola parameters α, β, γ, in terms of f and b, hence:
$z\alpha + \gamma = (((z + fb - b)(f - b - 1) - b)e^f + (f^2 + z)e^b)(b - f)^{-2}$
Minimizing with respect to f occurs at $f = \frac{-z}{b}$ and gives the result. □

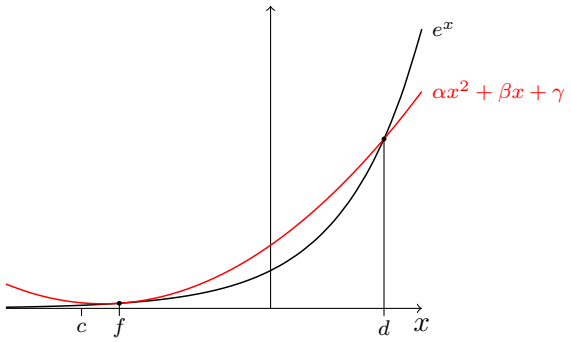

Figure A1.: A parabola parametarized by touching and intercepting points f, b above an exponential curve for all $a \leq x \leq b$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2019 doi:10.20944/preprints201901.0202.v2

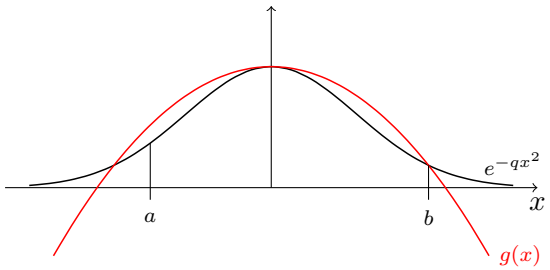

Figure A2.: Parabola $g(x) = (\exp(-qd^2)-1)d^{-2}x^2+1$ over function $f(x) = \exp(-qx^2)$ for all $a \leq x \leq b$ where $d = \max(b, -a)$; in the case $a = -1, b = 1.3, q = 1$