Approximating the Shapley Value Using Stratified Empirical Bernstein Sampling

Mark A. Burgess[1] and **Archie C. Chapman**[2]
[1]Australian National University, Canberra, Australia
[2]The University of Queensland, Brisbane, Australia
mark.burgess@anu.edu.au, archie.chapman@uq.edu.au

Abstract

The Shapley value is a well recognised method for dividing the value of joint effort in cooperative games. However, computing the Shapley value is known to be computationally hard, so stratified sample-based estimation is sometimes used. For this task, we provide two contributions to the state of the art. First, we derive a novel concentration inequality that is tailored to stratified Shapley value estimation using sample variance information. Second, by sequentially choosing samples to minimize our inequality, we develop a new and more efficient method of sampling to estimate the Shapley value. We evaluate our sampling method on a suite of test cooperative games, and our results demonstrate that it outperforms or is competitive with existing stratified sample-based estimation approaches to computing the Shapley value.

1 Introduction

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. It has many applications, including analyzing the power of voting blocks in weighted voting games [Bachrach *et al.*, 2009], in cost and surplus division problems [Soufiani *et al.*, 2014; O'Brien *et al.*, 2015; Aziz *et al.*, 2016; Chapman *et al.*, 2017], as a measure of network centrality [Michalak *et al.*, 2013]; and as a method of explaining the predictions of machine learning models [Lundberg and Lee, 2017]. Specifically, under the Shapley value, each player is allocated their average marginal contribution across every possible sequence of player join orderings. Although the Shapley value is conceptually simple, its use is hampered by the fact that its exact computation requires exponentially many evaluations of the marginal contributions of the players in the coalition.

Given this difficulty, one can exploit the fact that the Shapley value is an average by using estimation techniques to approximate it. In particular, the coalitions evaluated in the Shapley value computation can be naturally stratified by coalition size, allowing it to be reformulated as an average over strata averages. It is then possible to separately and efficiently estimate these strata averages via sample allocation techniques. Such techniques in literature include simple random sampling [Castro *et al.*, 2009], simple stratified random sampling, and a Neyman-type allocation [Castro *et al.*, 2017], and allocating stratified samples to minimize a Hoeffding type inequality [Maleki *et al.*, 2013].

In this paper, we improve on these approaches by developing a method for stratified sampling to maximally reduce an expression of the uncertainty in the Shapley value estimate. To do this, we develop a general expression associated with that uncertainty, which takes the form of a *concentration inequality*; specifically, a *stratified empirical Bernstein bound* (SEBB). 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 our SEBB, we propose an online method for sequentially sampling in order to maximally reduce the bound at each iteration, called the *stratified empirical Bernstein method* (SEBM). We numerically demonstrate the value of the SEBM by using it to compute the Shapley value in a suite of benchmark cooperative games. Our comparisons to existing sample-based approaches to computing the Shapley value show that our method is almost uniformly superior.

Next, Section 2 frames some context of the paper. Section 3 provides several component lemmas. From this Section 4 provides the derivation of our concentration inequality. In Section 5, we evaluate the performance of our bound in approximating the Shapley value. Section 6 discusses a multidimensional extension to the concentration inequality, and Section 7 concludes.

2 Background

Stratified sampling is a well known sampling technique, which estimates the mean of a population by partitioning it into mutually exclusive subgroups, or *strata*. It proceeds by applying a sampling estimator to each stratum, before weighting and combining these estimates to form an estimate of the population mean. If strata and their sizes are naturally given or determined, there exists a further question of how to allocate the sampling between the strata.

One way of deriving a sampling method is to use a concentration inequality as a confidence bound on the error of the population mean, and then selecting additional samples to minimise it. For instance, minimising Chebyshev's inequality on the variance of the estimation of the population mean results in the well-known *Neyman allocation* [1938], which has been inspiration as a method of estimating the Shapley value [Castro *et al.*, 2017]. Hoeffding's inequality is another commonly used concentration inequality whose minimisation has also been adapted in the context of Shaplay value sampling [Maleki *et al.*, 2013].

Recently, there has been interest in concentration inequalities called *empirical Bernstein bounds* (EBBs) [Maurer and Pontil, 2009], which are probabilistic bounds for the sample mean and sample variance. EBBs have been subject to rapid development [Audibert *et al.*, 2009; Audibert *et al.*, 2007; Bardenet and Maillard, 2015] and have replaced Hoeffding's inequality in a number of computational applications [Mnih *et al.*, 2008; Thomas *et al.*, 2015; Carpentier *et al.*, 2011].

Sampling *without replacement* offers the opportunity to further tighten the concentration bounds over the sampling-with-replacement case. The refinement was first demonstrated with a martingale argument by Serfling [1974] which was recently improved to created an EBB suitable for sampling *without* replacement by Bardenet and Maillard [2015].

Our observation, is that the components of these analyses can be used to create a variance-sensitive concentration inequality tailored for stratified random sampling, which then can be minimised in the context of Shapley value estimation.

3 Preliminaries

We now state lemmas which we use to derive our stratified empirical Bernstein bound (SEBB); the proofs of lemmas and theorems are as supplemental documentation (section A). 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:*

$$\mathbb{P}(a > c) \;\leq\; \mathbb{P}(a > b) + \mathbb{P}(b > c)$$

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

Lemma 3.2 (Variance Relation)**.** *For random variable X with mean μ, and n samples $\{x_k\}_{k=1,\ldots,n}$. The sample mean $\hat{\mu} = \frac{1}{n}\sum_k x_k$, sample variance, $\hat{\sigma}^2 = \frac{1}{n}\sum_k (x_k - \hat{\mu})^2$, centered average sample squares $\hat{\sigma}_0^2 = \frac{1}{n}\sum_k (x_k - \mu)^2$, obey:*

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

We use this result to create bounds for the sample variance from bounds on the sample squares. We also use the next lemma, which extends directly from Markov's inequality:

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*.

3.1 Bounds on the Moment Generating Function

The next three lemmas give three upper bounds for moment generating functions. The first is the famous *Hoeffding's lemma* [1963] which is essentially constructed by fitting a line over the exponential function of the moment generating function:

Lemma 3.4 (Hoeffding's Lemma)**.** *For random variable X bounded $a \leq X \leq b$ with $D = b - a$, for any $s > 0$:*

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

The next is a similar bound on the moment generating function that involves information about the variance.

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

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

The proof of this lemma is provided in supplementary material (section A), and essentially involves fitting a parabola (instead of a line) over the exponent in the moment generating function. This lemma is a simplified half-way result used in a derivation of Bennett's inequality as presented by Hoeffding [1963], and derived by Bennett [1962].

The next lemma stems from the creation of an upper bound on the random variable $-X^2$ instead of X. In this context the moment generating function becomes the expectation value of a Gaussian function which can be bound above by a parabola, yielding:

Lemma 3.6. *For random variable X bounded $a \leq X \leq b$, with $D = b - a$ and variance σ^2,and for any $q > 0$:*

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

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.

3.2 Moment Generating Function of Means

In the previous subsection we considered bounds on the moment generating function of random variables, but we must also relate these to bounds on the moment generating function of sample means from that of the random variables. The first is the most straightforward way to do this in the case of sampling with replacement, and directly assumes the independence of the samples:

Lemma 3.7 (Replacement Bound)**.** *For random variable X 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. 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:

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

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

However, for the context of sampling without replacement, there is an alternative result which can be substituted and may be (or may not be) tighter. In this context 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 have been produced with this judicious choice conducted. This result directly extends a reverse martingale argument from Bardenet and Maillard [2015]:

Lemma 3.8 (Martingale Bound)**.** *For finite data set $x_1, x_2, \dots x_n$ that is bounded $a \leq x_i \leq b$, and has a mean of zero and variance $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, denote $X_1, X_2, \dots, X_n$ the random variables corresponding to the data sequentially drawn randomly without replacement, and χ_m the average of the first m of them.*

$$\mathbb{E}[\exp(s\chi_m)] \leq \exp((\alpha D^2 \bar{\Omega}_m^n + \beta\sigma^2 \bar{\Psi}_m^n)s^2)$$

where

$$\bar{\Omega}_m^n = \sum_{k=m}^{n-1} \frac{1}{k^2} \leq \frac{(m+1)(1-m/n)}{m^2}$$

and

$$\bar{\Psi}_m^n = \sum_{k=m}^{n-1} \frac{n}{k^2(k+1)} \leq \frac{n+1-m}{m^2}.$$

Under the assumption that 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 $\mathbb{E}[\exp(sZ)] \leq \exp((\alpha D^2 + \beta\sigma_Z^2)s^2)$.

Proofs of both of these lemmas are found in supplementary material (section A).

4 The Stratified Finite Empirical Bernstein Bound and Sampling Method

We derive a novel probability bound for the error of the stratified sampling estimate, 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. Denote random variables for values sequentially drawn (with or without) replacement as $X_{i,1}, X_{i,2}, \dots, X_{i,n_i}$. Then, for the first m_i of these samples:

- $\chi_{i,m_i} = \frac{1}{m_i}\sum_{j=1}^{m_i} X_{i,j}$ is their average;
- 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;
- 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 $\{\tau_i\}_{i\in\{1\dots n\}}$. In the derivation we use arbitrary positive variables $\{\theta_i\}_{i\in\{1\dots n\}}$.

Given this context, the following two sections contain the derivation of the stratified empirical Bernstein 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 bounds 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 bounds the variance information in the context of sample variance information and the squared stratum mean errors.
3. Theorem 4.4 bounds the squared stratum mean errors.
4. Theorem 4.5 combines the three previous theorems together using union bounds (to eliminate the dependence on variance information and squared stratum mean errors), to create a concentration inequality for the error in the stratified population mean estimate given the sample variance information.

We begin with an expression for a probability bound on the absolute error of the weighted stratified sample means about the weighted strata means, and is developed from lemma 3.5:

Theorem 4.2. *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:*

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

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 from use of Lemma 3.6.

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

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

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, from lemma 3.4.

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

$$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\mu_i - \chi_{i,m_i})^2 \geq \frac{\log(2n/r)}{2}\sum_{i=1}^{n}\theta_i D_i^2\Omega_{m_i}^{n_i}\right) \leq r \tag{3}$$

This theorem bounds the weighted square error of the sample means. In the next, and final, step we combine the inequalities of Equations (1), (2) and (3) 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 $\Omega^{n_i}_{m_i}, \Psi^{n_i}_{m_i}$ per Lemma 3.7:*

$$\mathbb{P}\left(\frac{|\sum_{i=1}^{n}\tau_i(\chi_{i,m_i}-\mu_i)|}{\sqrt{\log(6/p)}} \geq \sqrt{\alpha+\left(\sqrt{\beta}+\sqrt{\gamma}\right)^2}\right) \leq p \tag{4}$$

where:

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

This completes the derivation. In Equation (4) 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 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 to minimize it.

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 various bounds on the moment generating functions that we developed in Section 3 use some loosening approximations in their derivations, and hence stronger and/or more representative bounds could be developed at the cost of greater mathematical complexity. Alternatively, integrating other kinds of inequalities such as entropic [Boucheron *et al.*, 2003] or Efron-Stein [Efron and Stein, 1981] could result in different and potentially tighter bounds. Nonetheless, when we use this bound in a sequential sampling algorithm, as described next, we see clear-cut estimation efficiency improvements in the Shapley value estimation task, as demonstrated in Section 5.

4.2 Sequential Sampling using the Stratified Empirical Bernstein 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.

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 $\hat{\bar{\sigma}}_i^2$, weights τ_i, widths D_i, sample budget B

1: **while** $\sum_i m_i < B$ **do**
2: $beststrata \leftarrow -1$
3: $lowestbound \leftarrow \infty$
4: **for** $k = 0$ to N **do**
5: $m_k \leftarrow m_k + 1$
6: $a \leftarrow [0,0]$, $b \leftarrow [0,0]$, $c \leftarrow [0,0]$, $d \leftarrow [0,0]$
7: **for** $i = 0$ to N **do**
8: $\Omega_{min} \leftarrow \min(\bar{\Omega}^{n_i}_{m_i}, \Omega^{n_i}_{m_i})$
9: $\Psi_{min} \leftarrow \min(\bar{\Psi}^{n_i}_{m_i}, \Psi^{n_i}_{m_i})$
10: $a_0 \leftarrow a_0 + \log(6N/p)D_i^2\bar{\Psi}^{n_i}_{m_i}\Omega_{min}\tau^2$
11: $a_1 \leftarrow a_1 + \log(6N/p)D_i^2\Psi^{n_i}_{m_i}\Omega_{min}\tau^2$
12: $b_0 \leftarrow \max(b_0, \log(3/p)D_i^2\bar{\Psi}^{n_i}_{m_i}\Psi_{min}\tau^2)$
13: $b_1 \leftarrow \max(b_1, \log(3/p)D_i^2\Psi^{n_i}_{m_i}\Psi_{min}\tau^2)$
14: $c_0 \leftarrow c_0 + 2\bar{\Psi}^{n_i}_{m_i}((m_i-1)\hat{\bar{\sigma}}_i^2/m_i)\tau^2$
15: $c_1 \leftarrow c_1 + 2\Psi^{n_i}_{m_i}((m_i-1)\hat{\bar{\sigma}}_i^2/m_i)\tau^2$
16: $d_0 \leftarrow d_0 + \frac{4}{17}D_i^2\bar{\Omega}^{n_i}_{m_i}\tau^2$
17: $d_1 \leftarrow d_1 + \frac{4}{17}D_i^2\Omega^{n_i}_{m_i}\tau^2$
18: **end for**
19: $w \leftarrow \sqrt{\min_j(d_j + (\sqrt{c_j + a_j + b_j} + \sqrt{b_j})^2)}$
20: **if** $w < lowestbound$ **then**
21: $beststrata \leftarrow k$
22: $lowestbound \leftarrow w$
23: **end if**
24: $m_k \leftarrow m_k - 1$
25: **end for**
26: take an extra sample from strata: $beststrata$
27: $m_{beststrata} \leftarrow m_{beststrata} + 1$
28: recalculate $\hat{\bar{\sigma}}^2_{beststrata}$
29: **end while**

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.

Algorithm 1 describes a process specific to sampling without replacement and involves the calculation of the SEBB with the tightest possible uses of Lemmas 3.8 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 3.8 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, using partial terms $\{a, b, c, d\}$. 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.

5 Numerical Evaluation: Shapley Value Approximation

We assess the benefits of using our sampling method by comparing its performance to other approaches in a set of example cooperative games.

For each game, we compute the exact Shapley value, and then the average absolute errors in the approximated Shapley value for a given budget of marginal-contribution samples across multiple computational runs. The results are shown in Table 1, where e^{SEBM} is the error associated with our method, SEBM, which is compared to the average absolute error in the Shapley value by sampling with:

- A Hoeffding-bound method [Maleki *et al.*, 2013], denoted e^{Ma},
- Simple random sampling of the without stratification, the described 'ApproShapley' method [Castro *et al.*, 2009], denoted e^{app}
- The stratified simple sampling method 'St-ApproShapley' [Castro *et al.*, 2017], denoted e^{sim}
- Castro's Neyman-type sampling method 'St-ApproShapley-opt' [Castro *et al.*, 2017], denoted e^{Ca}

Next, we describe the example cooperative games, and then discuss our results.

5.1 Example Cooperative Games

In general, 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 many attractive axioms. If $v_{i,k}$ is the average marginal contribution which player i can make across coalitions of size k:

$$v_{i,k} = \frac{1}{\binom{n-1}{k}} \sum_{S \subset N \setminus \{i\}, |S|=k} (v(S \cup \{i\}) - v(S)) \qquad (5)$$

Then the Shapley value can be expressed as an average:

$$\varphi_i(\langle N, v\rangle) = \frac{1}{n} \sum_{k=0}^{n-1} v_{i,k} \qquad (6)$$

The example games are described next, where w is a vector of weights in all the games. The first two are inspired by other benchmark games [Castro *et al.*, 2017].

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

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

where

$$w = [1, 1, 2, 2, 2, 3, 4, 5, 5, 5, 7, 8, 8, 8, 10]$$

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

$$w = [1, 3, 3, 6, 12, 16, 17, 19, 19, 19, 21, 22, 23, 24, 29]$$

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

$$w = [45, 41, 27, 26, 25, 21, 13, 13, 12, 12, 11, 11, 10, 10, 10]$$

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

$$w = [45, 41, 27, 26, 25, 21, 13, 13, 12, 12, 11, 11, 10, 10, 10]$$

In this game, we assign $D_i = 2$ for all i.

5.2 Results and Discussion

Overall, the results in Table 1 show that our method has excellent performance across the benchmark example games. Specifically, 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}) is sampled without replacement, while the other methods (per their design) are sampled with replacement.

Despite this performance, we make note of the computational overhead of iteratively minimizing (one sample at a

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^{app}	883.6	394.8	266.5	117.0	79.15
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^{app}	154.7	71.65	47.88	21.57	15.27
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^{app}	101.4	47.55	34.03	14.52	9.949
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^{app}	276.0	124.8	82.78	38.01	28.11
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). Lowest error results are boldened.

time) our inequality in the context of our simple example games, where this overhead can be a significant drawback. However, on more complicated games, such as where the characteristic function is slower to calculate (e.g. as in [Aziz *et al.*, 2016] or [O'Brien *et al.*, 2015]), any overhead associated with the sampling choice is expected to be much less relevant. We also note that our method's performance may 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). In practice, may also be advisable to run our method with an underestimate of the data widths D_i, as the sampling process is sensitive to the shape of the inequality and not necessarily its magnitude or accuracy as a bound. Finally, although there may be ways to further strengthen our concentration inequality at the cost of greater mathematical complexity, our computational results[1] show that using our bound greatly improves stratified sampling methods for Shapley value estimation.

6 Multidimensional Extension

Looking beyond Shapley value estimation, our concentration inequality and sampling method can also be extended directly to the context of multidimensional data. 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). And let $\hat{\sigma}^2_{i,j} = \frac{i}{m_i - 1}\sum_{k=1}^{m_i}(X_{i,k,j} - \chi_{i,m_i,j})^2$ be the unbiased sample variance of the m_i variables in the jth component. As before, we assume weights τ_i for each stratum.

In this context we have the following theorem (proof provided in supplimentary material):

Theorem 6.1 (Vector SEBM bound)**.** *In the context above, then with $\Omega^{n_i}_{m_i}, \Psi^{n_i}_{m_i}$ per Lemma 3.7:*

$$\mathbb{P}\left(\begin{matrix}\sum_{j=1}^{M}\left(\sum_{i=1}^{n}\tau_i(\chi_{i,m_i,j} - \mu_{i,j})\right)^2 \geq \\ \log(6/p)\sum_{j=1}^{M}\left(\alpha_j + \left(\sqrt{\beta_j} + \sqrt{\gamma_j}\right)^2\right)\end{matrix}\right) \leq Mp \quad (7)$$

where:

$$\alpha_j = \sum_{i=1}^{n}\frac{4}{17}\Omega^{n_i}_{m_i} D^2_{i,j}\tau_i^2,$$

$$\beta_j = \log(3/p)\left(\max_i \tau_i^2 {\Psi^{n_i}_{m_i}}^2 D^2_{i,j}\right)$$

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

[1] see: https://github.com/markopolo141/Stratified_Empirical_Bernstein_Sampling

The left hand side of (7) 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 $\sum_{i=1}^{n} \tau_i \mu_i$.

7 Conclusion

This paper develops an improved stratified sampling method for estimating the Shapley value of cooperative games. The sampling method is built on a novel empirical Bernstein bound, a concentration inequality for sampling from strata without replacement. This bound is used in a sampling strategy tailored to Shapley value estimation. Numerical results clearly demonstrate the benefit of our stratified sampling method for Shapley value estimation, by consistently outperforming the state of the art.

Acknowledgements

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

A Supplementary Material: Proofs

Proof of Lemma 3.1. For any events A and B,
$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ hence:
$\mathbb{P}((a > b) \cup (b > c)) \leq \mathbb{P}(a > b) + \mathbb{P}(b > c)$.
If $a > c$, then $(a > b) \cup (b > c)$ is true irrespective of b, so:
$\mathbb{P}(a > c) \leq \mathbb{P}((a > b) \cup (b > c))$ □

Proof of Lemma 3.2. By expanding terms:
$\hat{\sigma}^2 = \frac{1}{n}\sum_i \left(x_i - \frac{1}{n}\sum_j x_j\right)^2 = \frac{1}{n}\sum_i x_i^2 - \frac{1}{n^2}\sum_{i,j} x_i x_j$
$\hat{\sigma}_0^2 = \frac{1}{n}\sum_i (x_i - \mu)^2 = \frac{1}{n}\sum_i x_i^2 - \frac{2\mu}{n}\sum_i x_i + \mu^2$ therefore:
$\hat{\sigma}_0^2 - \hat{\sigma}^2 = \frac{1}{n^2}\sum_{i,j} x_i x_j - \frac{2\mu}{n}\sum_i x_i + \mu^2 = \left(\frac{1}{n}\sum_j x_j - \mu\right)^2$
□

Proof of Lemma 3.3. $\mathbb{P}(X \geq t) = \mathbb{P}(\exp(sX) \geq \exp(st))$
$\leq \mathbb{E}[\exp(sX)]\exp(-st)$ by Markov's inequality. □

Theorem A.1 (Parabola Fitting)**.** *For $a < b$ and $b, z > 0$, there exists α, β, γ that: $\alpha x^2 + \beta x + \gamma \geq \exp(x)$ for all $a \leq x \leq b$, and $z\alpha + \gamma = (z\exp(b) + b^2\exp(-z/b))(z + b^2)^{-1}$.*

Proof. Parabola $\alpha x^2 + \beta x + \gamma$ that satisfies these requirements touches the exponential curve at one point (at $x = f < b$) and intersects it at another (at $x = b$), per Figure 1, thus:

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

This gives 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 f occurs at $f = \frac{-z}{b}$ and gives the result. □

Proof of Lemma 3.5. Assume WLOG that X has a mean of zero. We construct an upper bound for $\mathbb{E}[\exp(sX)]$ by parabola over $\exp(sX)$. There exists a parabola defined by α, β, γ (Theorem A.1) and thus we expand: $E[\exp(sX)]$
$\leq \mathbb{E}[\alpha s^2 X^2 + \beta s X + \gamma] = \alpha s^2 \mathbb{E}[X^2] + \gamma = \alpha s^2 \sigma^2 + \gamma$
$= \left(\frac{\sigma^2}{b^2}\exp\left(s\left(b + \frac{\sigma^2}{b}\right)\right) + 1\right)\exp\left(-\frac{s\sigma^2}{b}\right)\left(\frac{\sigma^2}{b^2} + 1\right)^{-1}$.
This is monotonically increasing with b, and $D > b$.
Therefore: $\log(E[\exp(sX)]) \leq$
$\log\left(\frac{\sigma^2}{D^2}\exp\left(s\left(D + \frac{\sigma^2}{D}\right)\right) + 1\right) - \frac{s\sigma^2}{D} - \log\left(\frac{\sigma^2}{D^2} + 1\right)$
Using the fact that for any $\kappa, x \geq 0$:
$\log(\kappa\exp(x) + 1) \leq \log(\kappa + 1) + \frac{x\kappa}{\kappa + 1} + x^2\frac{\frac{1}{17} + \frac{\kappa}{2}}{(\kappa + 1)^2}$
The result follows using $\kappa = \frac{\sigma^2}{D^2}$ and $x = s(D + \sigma^2/D)$. □

Proof of Lemma 3.6. Assume WLOG X has a mean of zero. We construct an upper bound for $\mathbb{E}\left[\exp(-qX^2)\right]$ by parabola over $\exp(-qX^2)$. For α, γ such that $\alpha X^2 + \gamma \geq \exp(-qX^2)$ for all $a < X < b$. If we define $d = \max(b, -a)$ we can choose $\gamma = 1$ and $\alpha = (\exp(-qd^2) - 1)d^{-2}$
(see figure 2) which results in:
$\log \mathbb{E}[\exp(-qX^2)] \leq \log\left(\frac{\sigma^2}{D^2}\exp(-qD^2) - \frac{\sigma^2}{D^2} + 1\right)$
Given that for any $0 \leq \kappa \leq 0.5$ and $\gamma \leq 0$ that: $\log(\kappa\exp(\gamma) - \kappa + 1) \leq \kappa\gamma + \frac{1}{2}\kappa(1 - \kappa)\gamma^2$
Letting $\kappa = \frac{\sigma^2}{D^2}$ and $\gamma = -qD^2$
(which is valid by Popoviciu's inequality $\sigma^2 \leq D^2/4$)
$\mathbb{E}[\exp(-qX^2)] \leq \exp\left(\frac{1}{2}\sigma^2 q^2(D^2 - \sigma^2) - \sigma^2 q\right)$
$\leq \exp\left(\frac{1}{2}\sigma^2 q^2 D^2 - \sigma^2 q\right)$
and the result follows by multiplying by $\exp(q\sigma^2)$. □

Proof of Lemma 3.7. By the independence of samples:
$\mathbb{E}[\exp(s\chi_m)] = \mathbb{E}\left[\exp\left(\frac{s}{m}\sum_{i=1}^{m} X_i\right)\right] =$
$\prod_1^m \mathbb{E}\left[\exp\left(\frac{s}{m}X\right)\right] \leq \exp\left(\frac{s^2}{m^2}\sum_1^m\left(\alpha D^2 + \beta\sigma^2\right)\right)$ □

Proof of Lemma 3.8. $\chi_m = \frac{1}{m}\sum_{i=1}^{m} X_i$
$= \chi_{m+1} + \frac{1}{m}(\chi_{m+1} - X_{m+1})$
$= (\chi_m - \chi_{m+1}) + (\chi_{m+1} - \chi_{m+2}) + \cdots + (\chi_{n-1} - \chi_n)$
$= \frac{1}{m}(\chi_{m+1} - X_{m+1}) + \frac{1}{m+1}(\chi_{m+2} - X_{m+2}) + \cdots + \frac{1}{n-1}(\chi_n - X_n)$
Then because: $\exp(s\chi_m) = \prod_{k=m}^{n-1}\exp\left(\frac{s}{k}(\chi_{k+1} - X_{k+1})\right)$
$\mathbb{E}[\exp(s\chi_m)] =$
$\mathbb{E}\left[\prod_{k=m}^{n-1}\mathbb{E}\left[\exp\left(\frac{s}{k}(\chi_{k+1} - X_{k+1})\right)|\chi_{k+1}\cdots\chi_n\right]\right]$
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:
$\sigma_{k+1}^2 = \frac{n\sigma^2 - \sum_{j=k+1}^{n} X_j^2}{n - (n-k-1)} - \chi_k^2 \leq \frac{n\sigma^2}{k+1}$
by application of Lemma 3.2. Therefore:
$\mathbb{E}[\exp(s\chi_m)] \leq \exp\left(\sum_{k=m}^{n-1}\left(\alpha D^2 + \beta\frac{n\sigma^2}{k+1}\right)\frac{s^2}{k^2}\right)$ □

Proof of Theorem 4.2. 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 is sufficient for Lemmas 3.7, and 3.5 to apply giving:

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

Proof of Theorem 4.3. To bound the sum of variances (weighted by arbitrary positive θ_i), consider the average square of samples about the strata means. Applying Lemma 3.3 gives:
$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\sigma_i^2 - \frac{1}{m_i}\sum_{j=1}^{m_i}(X_{i,j}-\mu_i)^2) \geq y\right) \leq$
$\mathbb{E}\left[\exp\left(\sum_{i=1}^{n} s\theta_i\left(\sigma^2 - \frac{1}{m_i}\sum_{j=1}^{m_i}(X_{i,j}-\mu_i)^2\right)\right)\right]\exp(-sy)$
$\leq \exp(-sy)\prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{s\theta_i}{m_i}\sum_{j=1}^{m_i}(\sigma^2 - (X_{i,j}-\mu_i)^2)\right)\right]$
by independence of the sampling between the strata. This is sufficient for Lemma 3.7 with Lemma 3.6 to apply:
$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\sigma_i^2 - \frac{1}{m_i}\sum_{j=1}^{m_i}(X_{i,j}-\mu_i)^2) \geq y\right) \leq$
$\exp\left(\frac{1}{2}\sum_{i=1}^{n}\sigma_i^2\theta_i^2 s^2 D_i^2\Psi_{m_i}^{n_i} - sy\right)$
Minimizing with respect to s, rearranging, and applying Lemma 3.2 gives result. □

Proof of Theorem 4.4. 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.
Applying Lemma 3.3 together with Lemmas 3.7, 3.4, gives:
$\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\mu_i - \chi_{i,m_i})^2 \geq r\right)$
$\leq 1 - \prod_{i=1}^{n}\left(1 - 2\exp\left(-\frac{2r_i}{\theta_i D_i^2\Omega_{m_i}^{n_i}}\right)\right)$
Choosing r_i to minimize this expression gives:
$r_i = \left(r\theta_i D_i^2\Omega_{m_i}^{n_i}\right)\left(\sum_j \theta_j D_j^2\Omega_{m_j}^{n_j}\right)^{-1}$
Thus: $\mathbb{P}\left(\sum_{i=1}^{n}\theta_i(\mu_i - \chi_{i,m_i})^2 \geq r\right) \leq$
$1 - \prod_{i=1}^{n}\left(1 - 2\exp\left(\frac{-2r}{\sum_j \theta_j D_j^2\Omega_{m_j}^{n_j}}\right)\right)$
Using $\log(1-(1-\exp(x))^n) \leq x + \log(n)$ for negative x, and rearranging, gives result. □

Proof of Theorem 4.5. By widening the bound of Equation (2) we get:

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

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

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

Combining with Equation (3) with a union bound (Lemma 3.1) gives:

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

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 (1) 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. □

Proof of Theorem 6.1. Squaring (4) and applying it specifically to the jth component of all the vectors gives:

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

Taking a series of union bounds (Lemma 3.1) over j gives us our result. □

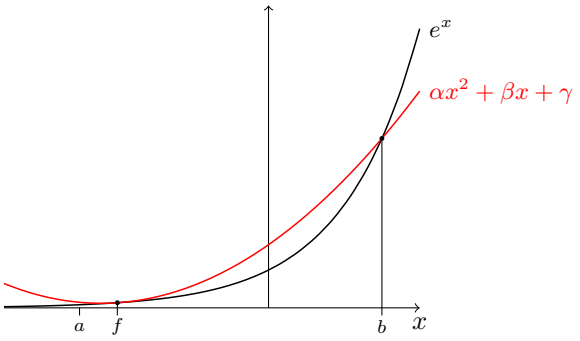

Figure 1: A parabola parametarized by touching and intercepting points f, b above an exponential curve for all $a \leq x \leq b$

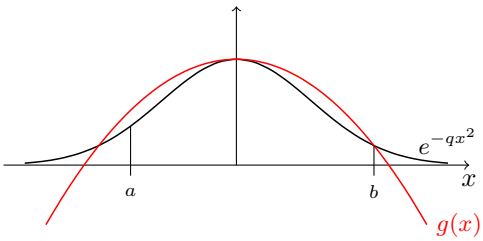

Figure 2: Parabola $g(x) = (\exp(-qd^2) - 1)d^{-2}x^2 + 1$ over function $\exp(-qx^2)$ for all $a \leq x \leq b$ where $d = \max(b, -a)$

References

[Audibert *et al.*, 2007] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT'07)*, pages 150–165, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[Audibert *et al.*, 2009] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[Aziz *et al.*, 2016] Haris Aziz, Casey Cahan, Charles Gretton, Philip Kilby, Nicholas Mattei, and Toby Walsh. A study of proxies for Shapley allocations of transport costs. *Journal of Artificial Intelligence Research*, 56:573–611, 2016.

[Bachrach *et al.*, 2009] Yoram Bachrach, Evangelos Markakis, Ezra Resnick, Ariel D. Procaccia, Jeffrey S. Rosenschein, and Amin Saberi. Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems (AAMAS)*, 20:105–122, 2009.

[Bardenet and Maillard, 2015] Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 08 2015.

[Bennett, 1962] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[Boucheron *et al.*, 2003] Stephane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

[Carpentier *et al.*, 2011] Alexandra Carpentier, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, and Peter Auer. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT'11)*, pages 189–203, Berlin, Heidelberg, 2011. Springer-Verlag.

[Castro *et al.*, 2009] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & OR*, 36(5):1726–1730, 2009.

[Castro *et al.*, 2017] Javier Castro, Daniel Gómez, Elisenda Molina, and Juan Tejada. Improving polynomial estimation of the Shapley value by stratified random sampling with optimum allocation. *Computers & Operations Research*, 82:180 – 188, 2017.

[Chapman *et al.*, 2017] Archie C. Chapman, Sleiman Mhanna, and Gregor Verbič. 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.

[Efron and Stein, 1981] Bradley Efron and Charles Stein. The jackknife estimate of variance. *Annals of Statistics*, 9(3):586–596, 05 1981.

[Hoeffding, 1963] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, Mar 1963.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[Maleki *et al.*, 2013] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the Estimation Error of Sampling-based Shapley Value Approximation. *arXiv e-prints*, page arXiv:1306.4265, June 2013.

[Maurer and Pontil, 2009] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009)*, June 2009.

[Michalak *et al.*, 2013] Tomasz P. Michalak, Karthik V. Aadithya, Piotr L. Szczepanski, Balaraman Ravindran, and Nicholas R. Jennings. Efficient computation of the Shapley value for game-theoretic network centrality. *Journal of Artificial Intelligence Research*, 46(1):607–650, January 2013.

[Mnih *et al.*, 2008] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, ICML '08, pages 672–679, New York, NY, USA, 2008. ACM.

[Neyman, 1938] J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116, 1938.

[O'Brien *et al.*, 2015] Geaorid O'Brien, Abbas El Gamal, and Ram Rajagopal. Shapley value estimation for compensation of participants in demand response programs. *IEEE Transactions on Smart Grid*, 6(6):2837–2844, 2015.

[Serfling, 1974] Robert J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 01 1974.

[Soufiani *et al.*, 2014] Hossein Azari Soufiani, Denis X Charles, David M Chickering, and David C Parkes. Approximating the Shapley value via multi-issue decomposition. In *Proc. Autonomous Agents and Multi-Agent Systems '14*. ACM, 2014.

[Thomas *et al.*, 2015] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 3000–3006, 2015.