Supplementary Material: Towards Efficient Data Valuation Based on the Shapley Value

Ruoxi Jia[1*], David Dao[2*], Boxin Wang[3], Frances Ann Hubis[2], Nick Hynes[1], Nezihe Merve Gurel[2], Bo Li[4], Ce Zhang[2], Dawn Song[1], Costas Spanos[1]
[1]University of California at Berkeley, [2]ETH, Zurich
[3]Zhejiang University, [4]University of Illinois at Urbana-Champaign

1 Proof of Lemma 1

Lemma 1. *For any $i, j \in I$ and $i \neq j$, the difference in Shapley values between i and j is*

$$s_i - s_j = \frac{1}{N-1} \sum_{S \subseteq I \setminus \{i,j\}} \frac{1}{\binom{N-2}{|S|}} \left[U(S \cup \{i\}) - U(S \cup \{j\}) \right]$$

Proof.

$$s_i - s_j = \sum_{S \subseteq I \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} \left[U(S \cup \{i\}) - U(S) \right] - \sum_{S \subseteq I \setminus \{j\}} \frac{|S|!(N-|S|-1)!}{N!} \left[U(S \cup \{j\}) - U(S) \right]$$
$$= \sum_{S \subseteq I \setminus \{i,j\}} \frac{|S|!(N-|S|-1)!}{N!} \left[U(S \cup \{i\}) - U(S \cup \{j\}) \right] + \sum_{S \in \{T | T \subseteq I, i \notin T, j \in T\}} \frac{|S|!(N-|S|-1)!}{N!} \left[U(S \cup \{i\}) - U(S) \right]$$
$$- \sum_{S \in \{T | T \subseteq I, i \in T, j \notin T\}} \frac{|S|!(N-|S|-1)!}{N!} \cdot \left[U(S \cup \{j\}) - U(S) \right]$$
$$= \sum_{S \subseteq I \setminus \{i,j\}} \frac{|S|!(N-|S|-1)!}{N!} \left[U(S \cup \{i\}) - U(S \cup \{j\}) \right]$$
$$+ \sum_{S' \subseteq I \setminus \{i,j\}} \frac{(|S'|+1)!(N-|S'|-2)!}{N!} \left[U(S' \cup \{i\}) - U(S' \cup \{j\}) \right]$$
$$= \sum_{S \subseteq I \setminus \{i,j\}} \left(\frac{|S|!(N-|S|-1)!}{N!} + \frac{(|S|+1)!(N-|S|-2)!}{N!} \right) \cdot \left[U(S \cup \{i\}) - U(S \cup \{j\}) \right]$$
$$= \frac{1}{N-1} \sum_{S \subseteq I \setminus \{i,j\}} \frac{1}{C_{N-2}^{|S|}} \left[U(S \cup \{i\}) - U(S \cup \{j\}) \right].$$

□

Loosely speaking, the proof distinguishes subsets S which include neither i nor j (such that the subset utility $U(S)$ of the marginal contribution directly cancels) and subsets including either i or j. In the latter case, S can be partitioned to a mock subset S' by excluding the respective point from S such that a common sum over S' again eliminates all terms other than $U(S' \cup \{i\}) - U(S' \cup \{j\})$.

2 Proof of Lemma 2

Lemma 2. *Suppose that C_{ij} is an $(\epsilon/(2\sqrt{N}), \delta/(N(N-1)))$-approximation to $s_i - s_j$. Then, the solution to the feasibility problem*

$$\sum_{i=1}^{N} \hat{s}_i = U_{tot} \tag{1}$$
$$|(\hat{s}_i - \hat{s}_j) - C_{i,j}| \leq \epsilon/(2\sqrt{N}) \quad \forall i, j \in \{1, \ldots, N\} \tag{2}$$

is an (ϵ, δ)-approximation to s with respect to l_2-norm.

Proof. Let $\epsilon' = \epsilon/(2\sqrt{N})$. Assume that $\hat{s}_i - s_i > \epsilon/\sqrt{N}$. Let $\hat{s}_i - s_i = c\epsilon'$ where $c > 2$.

Since $C_{i,j}$ is an $(\epsilon', \delta/(N(N-1)))$-approximation to $s_i - s_j$, we have that with probability at least $1 - \delta/(N(N-1))$,

$$|(s_i - s_j) - C_{i,j}| \leq \epsilon' \tag{3}$$

Moreover, the inequality (2) implies that

$$|(\hat{s}_i - \hat{s}_j) - C_{i,j}| \leq \epsilon'$$

Therefore,

$$\begin{aligned} |\hat{s}_i - s_i + s_j - \hat{s}_j| &= |\hat{s}_i - \hat{s}_j - C_{i,j} - (s_i - s_j - C_{i,j})| &(4)\\ &\leq |\hat{s}_i - \hat{s}_j - C_{i,j}| + |s_i - s_j - C_{i,j}| &(5)\\ &\leq 2\epsilon' &(6) \end{aligned}$$

with probability at least $1 - \delta/(N(N-1))$. By the assumption that $\hat{s}_i - s_i = c\epsilon'$ and $c > 2$, we have

$$(c-2)\epsilon' \leq \hat{s}_j - s_j \leq (c+2)\epsilon' \tag{7}$$

which further implies that $\hat{s}_j - s_j > 0$ for some $j \neq i$. Thus, with probability $1 - \delta/N$, we have $\hat{s}_j - s_j > 0$ for all $j \neq i$.

Then,

$$\sum_{j=1}^{N}(\hat{s}_j - s_j) = \sum_{j\neq i}(\hat{s}_j - s_j) + (\hat{s}_i - s_i) > 0 \tag{8}$$

Since $\sum_{j=1}^{N} s_j = U_{\text{tot}}$, it follows that $\sum_{j=1}^{N} \hat{s}_j > U_{\text{tot}}$, which contradicts with the fact that $\hat{s}_j$ $(j = 1, \ldots, N)$ is a solution to the feasibility problem (1) and (2).

The contradiction can be similarly established for $s_i - \hat{s}_i = c\epsilon'$. Therefore, we have that with probability at least $1-\delta/N$, $|s_i - \hat{s}_i| \leq 2\epsilon'$ for some i. This in turn implies that with probability at least $1-\delta$, $\|\hat{s} - s\|_\infty \leq 2\epsilon' = \epsilon/\sqrt{N}$. Moreover, since $\|\hat{s} - s\|_2 \leq \sqrt{N}\|\hat{s} - s\|_\infty = \epsilon$, we have that $\|\hat{s} - s\|_2 \leq \epsilon$ with probability at least $1 - \delta$.

□

3 Proof of Theorem 3

We prove Theorem 3, which specifies a lower bound on the number of tests needed for achieving a certain approximation error. Before delving into the proof, we first present a lemma that is useful for establishing the bound in Theorem 3.

Lemma 3 (Bennett's inequality [1])**.** *Given independent zero-mean random variables $X_1, \cdots, X_n$ satisfying the condition $|X_i| \leq a$, let $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$ be the total variance. Then for any $t \geq 0$,*

$$P[S_n > t] \leq \exp\left(-\frac{\sigma^2}{a^2} h\left(\frac{at}{\sigma^2}\right)\right)$$

where $h(u) = (1+u)\log(1+u) - u$.

We now restate Theorem 3 and proceed to the main proof.

Theorem 3. *Algorithm 1 returns an (ϵ, δ)-approximation to the Shapley value with respect to l_2-norm if the number of tests T satisfies $T \geq 8 \log \frac{N(N-1)}{2\delta} / \left((1-q_{tot}^2) h\left(\frac{\epsilon}{Zr\sqrt{N}(1-q_{tot}^2)}\right)\right)$, where $q_{tot} = \frac{N-2}{N} q(1) + \sum_{k=2}^{N-1} q(k)[1 + \frac{2k(k-N)}{N(N-1)}]$, $h(u) = (1+u)\log(1+u) - u$, $Z = 2\sum_{k=1}^{N-1} \frac{1}{k}$, and r is the range of the utility function.*

Proof. By Lemma 1, the difference in Shapley values between points i and j is given as

$$\begin{aligned} s_i - s_j &= \frac{1}{N-1} \sum_{S \subseteq I \setminus \{i,j\}} \frac{1}{C_{N-2}^{|S|}} \Big[U(S \cup \{i\}) - U(S \cup \{j\})\Big] \\ &= \frac{1}{N-1} \sum_{k=0}^{N-2} \frac{1}{C_{N-2}^{k}} \sum_{S \subseteq I \setminus \{i,j\}, |S| = k} \Big[U(S \cup \{i\}) - U(S \cup \{j\})\Big]. \end{aligned}$$

Let $\beta_1, \cdots, \beta_N$ denote N Boolean random variables drawn with the following sampler:

1. Sample the "length of the sequence" $\sum_{i=1}^{N} \beta_i = k \in \{1, 2, \cdots, N-1\}$, with probability $q(k)$.

2. Uniformly sample a length-k sequence from $\binom{N}{k}$ all possible length-k sequences

Then the probability of any given sequence $\beta_1, \cdots, \beta_N$ is

$$P[\beta_1, \cdots, \beta_N] = \frac{q(\sum_{i=1}^N \beta_i)}{C_N^{\sum_{i=1}^N \beta_i}}.$$

Now, we consider any two data points x_i and x_j where $i, j \in I = \{1, \cdots, N\}$ and their associated Boolean variables β_i and β_j, and analyze

$$\Delta = \beta_i U(\beta_1, \cdots, \beta_N) - \beta_j U(\beta_1, \cdots, \beta_N)$$

Consider the expectation of Δ. Obviously, only $\beta_i \neq \beta_j$ has non-zero contributions:

$$\begin{aligned}\mathbb{E}[\Delta] &= \sum_{k=0}^{N-2} \frac{q(k+1)}{C_N^{k+1}} \sum_{S \subseteq I \setminus \{i,j\}, |S|=k} \big[U(\beta_1, \cdots, \beta_{i-1}, 1, \beta_{i+1}, \cdots, \beta_{j-1}, 0, \beta_{j+1}, \cdots, \beta_N) \\ &\quad - U(\beta_1, \cdots, \beta_{i-1}, 0, \beta_{i+1}, \cdots, \beta_{j-1}, 1, \beta_{j+1}, \cdots, \beta_N)\big] \\ &= \sum_{k=0}^{N-2} \frac{q(k+1)}{C_N^{k+1}} \sum_{S \subseteq I \setminus \{i,j\}, |S|=k} \big[U(S \cup \{i\}) - U(S \cup \{j\})\big]\end{aligned}$$

We would like to have $Z\mathbb{E}[\Delta] = s_i - s_j$

$$Z\frac{q(k+1)}{C_N^{k+1}} = \frac{1}{(N-1)C_{N-2}^k}$$

which yields

$$q(k+1) = \frac{N}{Z(k+1)(N-k-1)} = \frac{1}{Z}\left(\frac{1}{k+1} + \frac{1}{N-k-1}\right)$$

for $k = 0, \cdots, N-2$. Equivalently,

$$q(k) = \frac{1}{Z}\left(\frac{1}{k} + \frac{1}{N-k}\right)$$

for $k = 1, \cdots, N-1$. The value of Z is given by

$$Z = \sum_{k=1}^{N-1}\left(\frac{1}{k} + \frac{1}{N-k}\right) = 2\sum_{k=1}^{N-1}\frac{1}{k} \leq 2(\log(N-1)+1)$$

Now, $\mathbb{E}[Z\Delta] = s_i - s_j$. Assume that the utility function ranges from $[0, r]$; then, we know from (**??**) that $Z\Delta$ is random variable ranges in $[-Zr, Zr]$.

Consider

$$\Delta := \beta_i U(\beta_1, \cdots, \beta_N) - \beta_j U(\beta_1, \cdots, \beta_N)$$

Note that $\Delta = 0$ when $\beta_i = \beta_j$. If $P[\beta_i = \beta_j]$ is large, then the variance of Δ will be much smaller than its range.

$$\begin{aligned}P[\beta_i = \beta_j] &= P[\beta_i = 1, \beta_j = 1] + P[\beta_i = 0, \beta_j = 0] \\ &= \left[\sum_{k=2}^{N-1} \frac{q(k)}{C_N^k} C_{N-2}^{k-2}\right] + \left[q(1) + \sum_{k=2}^{N-1} \frac{q(k)}{C_N^k} C_{N-2}^k\right]\end{aligned}$$

$$= \frac{N-2}{N} q(1) + \sum_{k=2}^{N-1} q(k) \Big[1 + \frac{2k(k-N)}{N(N-1)} \Big] \equiv q_{tot}$$

Let $W = \mathbb{1}[\Delta \neq 0]$ be an indicator of whether or not $\Delta = 0$. Then, $P[W = 0] = q_{tot}$ and $P[W = 1] = 1 - q_{tot}$.

Now, we analyze the variance of Δ. By the law of total variance,

$$\texttt{Var}[\Delta] = \mathbb{E}[\texttt{Var}[\Delta|W]] + \texttt{Var}[\mathbb{E}[\Delta|W]]$$

Recall $\Delta \in [-r, r]$. Then, the first term can be bounded by

$$\begin{aligned}
&\mathbb{E}[\texttt{Var}[\Delta|W]] = P[W=0]\texttt{Var}[\Delta|W=0] + P[W=1]\texttt{Var}[\Delta|W=1] \\
&= q_{tot}\texttt{Var}[\Delta|\Delta=0] + (1-q_{tot})\texttt{Var}[\Delta|\Delta\neq 0] \\
&= (1-q_{tot})\texttt{Var}[\Delta|\Delta\neq 0] \\
&\leq (1-q_{tot})r^2
\end{aligned}$$

where the last inequality follows from the fact that if a random variable is in the range $[m, M]$, then its variance is bounded by $\frac{(M-m)^2}{4}$.

The second term can be expressed as

$$\begin{aligned}
&\texttt{Var}[\mathbb{E}[\Delta|W]] = \mathbb{E}_W[(\mathbb{E}[\Delta|W] - \mathbb{E}[\Delta])^2] \\
&= P[W=0](\mathbb{E}[\Delta|W=0] - \mathbb{E}[\Delta])^2 + P[W=1](\mathbb{E}[\Delta|W=1] - \mathbb{E}[\Delta])^2 \\
&= q_{tot}(\mathbb{E}[\Delta|\Delta=0] - \mathbb{E}[\Delta])^2 + (1-q_{tot})(\mathbb{E}[\Delta|\Delta\neq 0] - \mathbb{E}[\Delta])^2 \\
&= q_{tot}(\mathbb{E}[\Delta])^2 + (1-q_{tot})(\mathbb{E}[\Delta|\Delta\neq 0] - \mathbb{E}[\Delta])^2
\end{aligned} \tag{9}$$

Note that

$$\begin{aligned}
\mathbb{E}[\Delta] &= P[W=0]\mathbb{E}[\Delta|\Delta=0] + P[W=1]\mathbb{E}[\Delta|\Delta\neq 0] \\
&= (1-q_{tot})\mathbb{E}[\Delta|\Delta\neq 0]
\end{aligned} \tag{10}$$

Plugging (10) into (9), we obtain

$$\texttt{Var}[\mathbb{E}[\Delta|W]] = (q_{tot}(1-q_{tot})^2 + q_{tot}^2(1-q_{tot}))(\mathbb{E}[\Delta|\Delta\neq 0])^2$$

Since $|\Delta| \leq r$, $(\mathbb{E}[\Delta|\Delta\neq 0])^2 \leq r^2$. Therefore,

$$\texttt{Var}[\mathbb{E}[\Delta|W]] \leq q_{tot}(1-q_{tot})r^2$$

It follows that

$$\texttt{Var}[\Delta] \leq (1-q_{tot}^2)r^2$$

Given T samples, the application of Bennett's inequality in Lemma 3 yields

$$P\Big[\sum_{t=1}^{T}(Z\Delta_t - \mathbb{E}[Z\Delta_t]) > \epsilon'\Big] \leq \exp\Big(-\frac{T(1-q_{tot}^2)}{4} h(\frac{2\epsilon'}{TZr(1-q_{tot}^2)})\Big)$$

By letting $\epsilon = \epsilon'/T$,

$$P\big[(Z\bar{\Delta} - \mathbb{E}[Z\Delta]) > \epsilon\big] \leq \exp\Big(-\frac{T(1-q_{tot}^2)}{4} h(\frac{2\epsilon}{Zr(1-q_{tot}^2)})\Big)$$

Therefore, the number of tests T we need in order to get an $(\epsilon/(2\sqrt{N}), \delta/(N(N-1)))$-approximation to the difference of two Shapley values for a single pair of data points is

$$T \geq \frac{4}{(1-q_{tot}^2)h(\frac{\epsilon}{Z\sqrt{N}r(1-q_{tot}^2)})} \log \frac{N(N-1)}{\delta}$$

By union bound, the number of tests T for achieving $(\epsilon/\sqrt{N}, \delta/N)$-approximation to the difference of the Shapley values for all $N(N-1)/2$ pairs of data points is

$$T \geq \frac{8}{(1-q_{tot}^2)h(\frac{\epsilon}{Z\sqrt{N}rC_\epsilon(1-q_{tot}^2)})} \log \frac{N(N-1)}{2\delta}$$

By Lemma 2, we approximate the Shapley value up to (ϵ, δ) with $(\epsilon/\sqrt{N}, \delta/(N(N-1)))$ approximations to all $N(N-1)/2$ pairs of data points. $\square$

4 Proof of Theorem 4

Theorem 4. *There exists some constant C' such that if $M \geq C'(K\log(N/(2K)) + \log(2/\delta))$ and $T \geq \frac{2r^2}{\epsilon^2}\log\frac{4M}{\delta}$, except for an event of probability no more than δ, the output of Algorithm* **??** *obeys*

$$\|\hat{s} - s\|_2 \leq C_{1,K}\epsilon + C_{2,K}\frac{\sigma_K(s)}{\sqrt{K}} \tag{11}$$

for some constants $C_{1,K}$ and $C_{2,K}$.

Proof. Due to the super-additivity of $U(\cdot)$, $\hat{y}_{m,t}$ can be lower bounded by $-\frac{1}{\sqrt{M}}\sum_{i=1}^N U(P_i^{\pi_t}\cup\{i\}) - U(P_i^{\pi_t}) = -\frac{1}{\sqrt{M}}U(\pi_t) \geq -\frac{r}{\sqrt{M}}$; the upper bound can be similarly analyzed. Thus, the range of $\hat{y}_{m,t}$ is $[-1/\sqrt{M}r, 1/\sqrt{M}r]$. Since $\mathbb{E}[\hat{y}_{m,t}] = \sum_{i=1}^N A_{m,i}\mathbb{E}[U(P_i^{\pi_t}\cup\{i\}) - U(P_i^{\pi_t})] = \sum_{i=1}^N A_{m,i}s_i$ for all $m = 1,\ldots,M$, an application of Hoeffding's bound gives

$$P[\|As - \bar{y}\|_2 \geq \epsilon] \leq P[\|As - \bar{y}\|_\infty \geq \frac{\epsilon}{\sqrt{M}}] \tag{12}$$

$$\leq \sum_{m=1}^M P[|A_m s - \bar{y}_m| \geq \frac{\epsilon}{\sqrt{M}}] \tag{13}$$

$$\leq 2M\exp(-\frac{\epsilon^2}{2r^2T}) \tag{14}$$

Let $s = \Delta s + \bar{s}$. Thus, $P[\|A(\bar{s}+\Delta s) - \bar{y}\|_2 \leq \epsilon]$ holds with probability at least $\delta/2$ provided

$$T \geq \frac{2r^2}{\epsilon^2}\log\frac{4M}{\delta}. \tag{15}$$

By the random matrix theory, the restricted isometry constant of A satisfies $\delta_{2K} \leq C_\delta = 0.465$ with probability at least $1-\delta/2$ if

$$M \geq CC_\delta^{-2}(2K\log(N/(2K)) + \log(2/\delta)) \tag{16}$$

where $C > 0$ is a universal constant.

Applying the Theorem 2.7 in [3], we obtain that the output of Algorithm 2 satisfies

$$\|\hat{s} - s\| = \|\Delta s^* - \Delta s\| \leq C_{1,K}\epsilon + C_{2,K}\frac{\sigma_K(s)}{\sqrt{K}} \tag{17}$$

with probability at least $1-\delta$ provided that (15) holds and $M \geq C'(K\log(N/(2K)) + \log(2/\delta))$ for some constant C'. $\square$

5 Proof of Theorem 5

For the proof of Theorem 5 we need the following definition of a *stable utility function*.

Definition 1. *A utility function $U(\cdot)$ is called λ-stable if*

$$\max_{i,j\in I, S\subseteq I\setminus\{i,j\}} |U(S\cup\{i\}) - U(S\cup\{j\})| \leq \frac{\lambda}{|S|+1}$$

Then, Shapley values calculated from λ-stable utility functions have the following property.

Proposition 1. *If $U(\cdot)$ is λ-stable, then for all $i,j\in I$ and $i\neq j$*

$$s_i - s_j \leq \frac{\lambda(1+\log(N-1))}{N-1}$$

Proof. By Lemma 1, we have

$$s_i - s_j \leq \frac{1}{N-1}\sum_{S\subseteq I\setminus\{i,j\}} \frac{1}{C_{N-2}^{|S|}}\frac{\lambda}{|S|+1} = \frac{1}{N-1}\sum_{|S|=0}^{N-2}\frac{\lambda}{|S|+1}$$

Recall the bound on the harmonic sequences

$$\sum_{k=1}^{N}\frac{1}{k} \leq 1+\log(N)$$

which gives us

$$s_i - s_j \leq \frac{\lambda(1+\log(N-1))}{N-1}$$

□

Then, we can prove Theorem 5.

Theorem 5. *For a learning algorithm $A(\cdot)$ with uniform stability $\beta = \frac{C_{stab}}{|S|}$, where $|S|$ is the size of the training set and C_{stab} is some constant. Let the utility of D be $U(D) = M - L_{test}(A(D), D_{test})$, where $L_{test}(A(D), D_{test}) = \frac{1}{N}\sum_{i=1}^{N} l(A(D), z_{test,i})$ and $0\leq l(\cdot,\cdot)\leq M$. Then, $s_i - s_j \leq 2C_{stab}\frac{1+\log(N-1)}{N-1}$ and the Shapley difference vanishes as $N\to\infty$.*

Proof. For any $i,j\in I$ and $i\neq j$,

$$\begin{aligned}
&|U(S\cup\{i\}) - U(S\cup\{j\})| \\
&= |\frac{1}{N}\sum_{i=1}^{N}[l(A(S\cup\{i\}), z_{\text{test},i}) - l(A(S\cup\{j\}), z_{\text{test},i})]| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}|l(A(S\cup\{i\}), z_{\text{test},i}) - l(A(S), z_{\text{test},i})| + |l(A(S), z_{\text{test},i}) - l(A(S\cup\{j\}), z_{\text{test},i})| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\frac{2C_{\text{stab}}}{|S|+1} = \frac{2C_{\text{stab}}}{|S|+1}
\end{aligned}$$

Combining the above inequality with Proposition 1 proves the theorem. □

6 Proof of Theorem 6

Theorem 6. *Consider the value attribution scheme that assign the value $\hat{s}(U,i) = C_U[U(S\cup\{i\}) - U(S)]$ to user i where $|S| = N-1$ and C_U is a constant such that $\sum_{i=1}^{N}\hat{s}(U,i) = U(I)$. Consider two utility functions $U(\cdot)$ and $V(\cdot)$. Then, $\hat{s}(U+V,i)\neq\hat{s}(U,i)+\hat{s}(V,i)$ unless $V(I)[\sum_{i=1}^{N}U(S\cup\{i\}) - U(S)] = U(I)[\sum_{i=1}^{N}V(S\cup\{i\}) - V(S)]$.*

Proof. Consider two utility functions $U(\cdot)$ and $V(\cdot)$. The values attributed to user i under these two utility functions are given by

$$\hat{s}(U,i) = C_U[U(S \cup \{i\}) - U(S)]$$

and

$$\hat{s}(V,i) = C_V[V(S \cup \{i\}) - V(S)]$$

where C_U and C_V are constants such that $\sum_{i=1}^N \hat{s}(U,i) = U(I)$ and $\sum_{i=1}^N \hat{s}(V,i) = V(I)$. Now, we consider the value under the utility function $W(S) = U(S) + V(S)$:

$$\hat{s}(U+V,i) = C_W[U(S \cup \{i\}) - U(S) + V(S \cup \{i\}) - V(S)]$$

where

$$C_W = \frac{U(I) + V(I)}{\sum_{i=1}^N [U(S \cup \{i\}) - U(S) + V(S \cup \{i\}) - V(S)]}$$

Then, $\hat{s}(U+V,i) = \hat{s}(U,i) + \hat{s}(V,i)$ if and only if $C_U = C_V = C_W$, which is equivalent to

$$V(I)[\sum_{i=1}^N U(S \cup \{i\}) - U(S)] = U(I)[\sum_{i=1}^N V(S \cup \{i\}) - V(S)]$$

$\square$

7 Theoretical Results on the Baseline Permutation Sampling

Let π_t be a random permutation of $D = \{z_i\}_{i=1}^N$ and each permutation has a probability of $\frac{1}{N!}$. Let $\phi_i^t = U(P_i^{\pi_t} \cup \{i\}) - U(P_i^{\pi_t})$, we consider the following estimator of s_i:

$$\hat{s}_i = \frac{1}{T} \sum_{t=1}^T \phi_i^t$$

Theorem 2. *Given the range of the utility function r, an error bound ϵ, and a confidence $1-\delta$, the sample size required such that*

$$P[\|\hat{s} - s\|_2 \geq \epsilon] \leq \delta$$

is

$$T \geq \frac{2r^2 N}{\epsilon^2} \log \frac{2N}{\delta}$$

Proof.

$$P[\max_{i=1,\cdots,N} |\hat{s}_i - s_i| \geq \epsilon] = P[\cup_{i=1,\cdots,N}\{|\hat{s}_i - s_i| \geq \epsilon\}] \leq \sum_{i=1}^N P[|\hat{s}_i - s_i| \geq \epsilon]$$
$$\leq 2N \exp\left(- \frac{2T\epsilon^2}{4r^2} \right)$$

The first inequality follows from the union bound and the second one is due to Hoeffding's inequality. Since $\|\hat{s} - s\|_2 \leq \sqrt{N}\|\hat{s} - s\|_\infty$, we have

$$P[\|\hat{s} - s\|_2 \geq \epsilon \leq P[\|\hat{s} - s\|_\infty \geq \epsilon/\sqrt{N}] \leq 2N \exp\left(- \frac{2T\epsilon^2}{4Nr^2} \right)$$

Setting $2N\exp(-\frac{T\epsilon^2}{2Nr^2}) \leq \delta$ yields

$$T \geq \frac{2r^2 N}{\epsilon^2} \log \frac{2N}{\delta}$$

$\square$

The permutation sampling-based method used as baseline in the experimental part of this work was adapted from Maleki et al. [2] and is presented in Algorithm 1.

Algorithm 1: Baseline: Permutation Sampling-Based Approach

input : Training set - $D = \{(x_i, y_i)\}_{i=1}^N$, utility function $U(\cdot)$, the number of measurements - M, the number of permutations - T

output : The Shapley value of each training point - $\hat{s} \in \mathbb{R}^N$

```
for t ← 1 to T do
    π_t ← GenerateUniformRandomPermutation(D);
    φ_i^t ← U(P_i^{π_t} ∪ {i}) − U(P_i^{π_t}) for i = 1, ..., N;
end
ŝ_i = (1/T) Σ_{t=1}^{T} φ_i^t for i = 1, ..., N;
```

References

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

[2] S. Maleki, L. Tran-Thanh, G. Hines, T. Rahwan, and A. Rogers. Bounding the estimation error of sampling-based shapley value approximation. *arXiv preprint arXiv:1306.4265*, 2013.

[3] H. Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.