A Distributional Framework For Data Valuation Supplement*

Amirata Ghorbani Michael P. Kim James Zou

ICML 2020

A Review of Shapley Axioms

Here, we provide a high-level review of the axioms that Shapley used to describe an equitable valuation function [Sha53]. We consider the data Shapley setting, letting $v(z; U, D)$ denote the value of $z \in D$ for a finite subset $D \subseteq \mathcal{Z}$ with respect to potential $U : \mathcal{Z}^* \to [0, 1]$.

- *Symmetry* – Consider $z_i, z_j \in D$; suppose for all $S \subseteq D \setminus \{z_i, z_j\}$, $U(S \cup \{z_i\}) = U(S \cup \{z_j\})$. Then,
$$v(z_i; U, D) = v(z_j; U, D).$$
That is, if two data points are equivalent, then they should receive the same value.

- *Null player* – Consider $z \in D$; suppose for all $S \subseteq D \setminus \{z\}$, $U(S \cup \{z\}) = U(S)$. Then,
$$v(z; U, D) = 0.$$
That is, if a data point contributes no marginal gain in potential to any nontrivial subset, then it receives no value.

- *Additivity* – Consider two potentials U_1, U_2. For all $z \in D$,
$$U(z; U_1 + U_2, D) = U(z; U_1, D) + U(z; U_2, D).$$
That is, the value of a data point with respect to the combination of two tasks (addition of two potentials) is the sum of the values with respect to each task (potential) separately.

Theorem A.1 ([Sha53])**.** *The Shapley value is the unique valuation function that satisfies the symmetry, null player, and additivity axioms.*

Additionally, the Shapley value satisfies the desirable property that it allocates all of the value to the contributors.

*Code is available on Github at https://github.com/amiratag/DistributionalShapley

- *Efficiency* – The sum of the individuals' Shapley values equals the value of the coalition. That is,
$$\sum_{z \in D} v(z; U, D) = U(D).$$

It is straightforward to verify that the distributional Shapley value immediately inherits the properties of symmetry, null player, and additivity (by linearity of expectation). Further, it satisfies an on-average variant of efficiency.

Proposition A.2. *Given a potential U and a data distribution $\mathcal{D}$, for $m \in \mathbb{N}$,*

$$\mathop{\mathbf{E}}_{z \sim \mathcal{D}} [\nu(z; U, \mathcal{D}, m)] = \frac{\mathbf{E}_{D \sim \mathcal{D}^m} [U(D)] - U(\emptyset)}{m}.$$

Proof. We expand the expected distributional Shapley value with its definition and then apply linearity of expectation.

$$\begin{aligned}
\mathop{\mathbf{E}}_{z \sim \mathcal{D}} [\nu(z; U, \mathcal{D}, m)] &= \mathop{\mathbf{E}}_{z \sim \mathcal{D}} \left[\mathop{\mathbf{E}}_{\substack{k \sim [m] \\ S \sim \mathcal{D}^{k-1}}} [U(S \cup \{z\}) - U(S)] \right] \\
&= \frac{1}{m} \cdot \sum_{k=1}^{m} \left(\mathop{\mathbf{E}}_{\substack{z \sim \mathcal{D} \\ S \sim \mathcal{D}^{k-1}}} [U(S \cup \{z\})] - \mathop{\mathbf{E}}_{S \sim \mathcal{D}^{k-1}} [U(S)] \right) \\
&= \frac{1}{m} \cdot \sum_{k=1}^{m} \left(\mathop{\mathbf{E}}_{S_k \sim \mathcal{D}^k} [U(S_k)] - \mathop{\mathbf{E}}_{S_{k-1} \sim \mathcal{D}^{k-1}} [U(S_{k-1})] \right) \\
&= \frac{1}{m} \cdot \left(\mathop{\mathbf{E}}_{D \sim \mathcal{D}^m} [U(D)] - U(\emptyset) \right)
\end{aligned}$$

□

B Distributional Shapley Value for Mean Estimation

Proposition (Restatement of Proposition 2.4)**.** *Suppose $\mathcal{D}$ has bounded second moments. Then for $z \in \mathcal{Z}$ and $m \in \mathbb{N}$, $\nu(z; U_\mu, \mathcal{D}, m)$ for mean estimation over $\mathcal{D}$ is given by*

$$\frac{\mathbf{E}_{S \sim \mathcal{D}^m} [U(S)]}{m} + \frac{C_m}{m} \cdot \left(\mathop{\mathbf{E}}_{s \sim \mathcal{D}} \left[\|s - \mu\|^2 \right] - \|z - \mu\|^2 \right)$$

for an explicit constant $C_m = \Theta(1)$ determined by m.

Proof. Consider the unsupervised learning task of mean estimation using the empirical estimator. Specifically, suppose we receive samples from some distribution $\mathcal{D}$ supported on $\mathbb{R}^d$ with mean $\mu = \mathbf{E}_{s \sim \mathcal{D}}[s]$ and bounded second moments. Given a subset $S \subseteq \mathbb{R}^d$, we consider the empirical

estimator $\hat{\mu}_S = \frac{1}{|S|} \cdot \sum_{s \in S} s$. We define a potential $U(S)$ by the performance of the empirical estimator.

$$\begin{aligned} U(S) &= \mathop{\mathbf{E}}_{s \sim \mathcal{D}} \left[\|s - \mu\|^2 \right] - \|\hat{\mu}_S - \mu\|^2 \\ &= B^2 - \|\hat{\mu}_S - \mu\|^2 \end{aligned}$$

By convention, we will assume that $U(\emptyset) = 0$. For notational convenience, let $\mathbf{E}_{s \sim \mathcal{D}} \left[\|s - \mu\|^2 \right] = B^2$ for some $B = \Theta(1)$. As such, we can evaluate the difference in potentials as follows.

$$\begin{aligned} &= \left(B^2 - \left\|\mu - \hat{\mu}_{S \cup \{z\}}\right\|^2 \right) - \left(B^2 - \|\mu - \hat{\mu}_S\|^2 \right) \\ &= \|\mu - \hat{\mu}_S\|^2 - \left\|\mu - \hat{\mu}_{S \cup \{z\}}\right\|^2 \end{aligned}$$

Importantly, note that we can relate $\hat{\mu}_{S \cup \{z\}}$ to $\hat{\mu}_S$.

$$\hat{\mu}_{S \cup \{z\}} = \hat{\mu}_S + \frac{1}{k} \cdot (z - \hat{\mu}_S)$$

Using these expressions, we can expand the distributional Shapley value into a form that will be convenient to work with.

$$\begin{aligned} \nu(z; U, \mathcal{D}, m) &= \mathop{\mathbf{E}}_{\substack{k \sim [m] \\ S \sim \mathcal{D}^{k-1}}} \left[U(S \cup \{z\}) - U(S) \right] \\ &= \frac{1}{m} \cdot \sum_{k=1}^{m} \mathop{\mathbf{E}}_{S \sim \mathcal{D}^{k-1}} \left[U(S \cup \{z\}) - U(S) \right] \\ &= \frac{1}{m} \cdot \left(U(\{z\}) - U(\emptyset) + \sum_{k=2}^{m} \mathop{\mathbf{E}}_{S \sim \mathcal{D}^{k-1}} \left[U(S \cup \{z\}) - U(S) \right] \right) \\ &= \frac{1}{m} \cdot \left(B^2 - \|z - \mu\|^2 + \sum_{k=2}^{m} \mathop{\mathbf{E}}_{S \sim \mathcal{D}^{k-1}} \left[\|\mu - \hat{\mu}_S\|^2 - \left\|\mu - \hat{\mu}_{S \cup \{z\}}\right\|^2 \right] \right) \end{aligned}$$

We, thus, focus our efforts on bounding the summation from $k = 2$ to m. As such, we can evaluate the difference in potentials within the expectation as follows.

$$\begin{aligned} &\|\mu - \hat{\mu}_S\|^2 - \left\|\mu - \hat{\mu}_{S \cup \{z\}}\right\|^2 \\ &= \|\mu - \hat{\mu}_S\|^2 - \left\| \mu - \hat{\mu}_S - \frac{1}{k} \cdot (z - \hat{\mu}_S) \right\|^2 \\ &= \|\mu - \hat{\mu}_S\|^2 - \left(\|\mu - \hat{\mu}_S\|^2 + \frac{1}{k^2} \cdot \|z - \hat{\mu}_S\|^2 - \frac{2}{k} \cdot \langle \mu - \hat{\mu}_S, z - \hat{\mu}_S \rangle \right) \\ &= \frac{2}{k} \cdot \langle \mu - \hat{\mu}_S, z - \hat{\mu}_S \rangle - \frac{1}{k^2} \cdot \|z - \hat{\mu}_S\|^2 \end{aligned}$$

Taking an expectation over $S \sim \mathcal{D}^{k-1}$, we can simplify each term in the summation separately; first, some identities that will be useful and hold for all $n \in \mathbb{N}$:

$$\mathop{\mathbf{E}}_{S\sim\mathcal{D}^n}[\hat{\mu}_S] = \mu \tag{1}$$

$$\mathop{\mathbf{E}}_{S\sim\mathcal{D}^n}[\langle \mu - \hat{\mu}_S, q\rangle] = 0 \tag{2}$$

$$\mathop{\mathbf{E}}_{S\sim\mathcal{D}^n}\left[\|\hat{\mu}_S\|^2 - \|\mu\|^2\right] = \mathop{\mathbf{E}}_{S\sim\mathcal{D}^n}\left[\|\hat{\mu}_S - \mu\|^2\right] = \frac{1}{n}\cdot\mathop{\mathbf{E}}_{s\sim\mathcal{D}}\left[\|s-\mu\|^2\right] \tag{3}$$

where (1) follows because $\hat{\mu}_S$ an unbiased estimator of μ; (2) holds for all $q \in \mathbb{R}^d$; and (3) is a well-known fact that can be derived using (1) and (2).

Beginning with the first inner product.

$$\begin{aligned}\mathop{\mathbf{E}}_{S\sim\mathcal{D}^{k-1}}[\langle \mu - \hat{\mu}_S, z - \hat{\mu}_S\rangle] &= \mathop{\mathbf{E}}_{S\sim\mathcal{D}^{k-1}}[\langle \mu - \hat{\mu}_S, \mu - \hat{\mu}_S\rangle] &(4)\\ &= \mathop{\mathbf{E}}_{S\sim\mathcal{D}^{k-1}}\left[\|\mu - \hat{\mu}_S\|^2\right]\\ &= \frac{1}{k-1}\cdot\mathop{\mathbf{E}}_{s\sim\mathcal{D}}\left[\|s-\mu\|^2\right] &(5)\end{aligned}$$

where (4) applies (2) with $q = \mu - z$ and (5) applies (3).

Expanding the next term.

$$\begin{aligned}\mathop{\mathbf{E}}_{S\sim\mathcal{D}^{k-1}}\left[\|z - \hat{\mu}_S\|^2\right] &= \mathop{\mathbf{E}}_{S\sim\mathcal{D}^{k-1}}\left[\|z\|^2 + \|\hat{\mu}_S\|^2 - 2\langle z, \hat{\mu}_S\rangle + \|\mu\|^2 - \|\mu\|^2\right]\\ &= \mathop{\mathbf{E}}_{S\sim\mathcal{D}^{k-1}}\left[\|z\|^2 - 2\langle z, \hat{\mu}_S\rangle + \|\mu\|^2\right] + \mathop{\mathbf{E}}_{S\sim\mathcal{D}^{k-1}}\left[\|\hat{\mu}_S\|^2 - \|\mu\|^2\right]\\ &= \|z-\mu\|^2 + \frac{1}{k-1}\cdot\mathop{\mathbf{E}}_{s\sim\mathcal{D}}\left[\|s-\mu\|^2\right] &(6)\end{aligned}$$

where (6) follows by applying linearity of expectation and (1) to the first term and (3) to the second term.

Thus, in all, the value can be expressed as follows.

$$\begin{aligned}&\sum_{k=2}^{m}\mathop{\mathbf{E}}_{S\sim\mathcal{D}^{k-1}}\left[\frac{2}{k}\cdot\langle \mu - \hat{\mu}_S, z - \hat{\mu}_S\rangle - \frac{1}{k^2}\cdot\|z - \hat{\mu}_S\|^2\right]\\ &= \sum_{k=2}^{m}\left(\frac{2}{k\cdot(k-1)}\cdot\mathop{\mathbf{E}}_{s\sim\mathcal{D}}\left[\|s-\mu\|^2\right] - \frac{1}{k^2\cdot(k-1)}\cdot\mathop{\mathbf{E}}_{s\sim\mathcal{D}}\left[\|s-\mu\|^2\right] - \frac{1}{k^2}\cdot\|z-\mu\|^2\right)\\ &= \sum_{k=2}^{m}\left(\frac{2B^2 - \|z-\mu\|^2}{k\cdot(k-1)} - \frac{B^2 - \|z-\mu\|^2}{k^2\cdot(k-1)}\right)\\ &= \frac{m-1}{m}\cdot\left(2B^2 - \|z-\mu\|^2\right) + c(m)\cdot\left(B^2 - \|z-\mu\|^2\right)\\ &= \frac{m-1}{m}\cdot B^2 + (1 - 1/m + c(m))\cdot\left(B^2 - \|z-\mu\|^2\right)\end{aligned}$$

where $\sum_{k=2}^{m}\frac{1}{k\cdot(k-1)} = \frac{m-1}{m}$ and we take $c(m) = \sum_{k=2}^{m}\frac{1}{k^2\cdot(k-1)}$.

Thus, plugging this expression back into our original expansion.

$$\begin{aligned}
&\nu(z; U, \mathcal{D}, m) \\
&= \frac{1}{m} \cdot \left(\frac{m-1}{m} \cdot \left(2B^2 - \|z-\mu\|^2 \right) + (1 + c(m)) \cdot \left(B^2 - \|z-\mu\|^2 \right) \right) \\
&= \frac{m-1}{m^2} \cdot B^2 + \frac{C(m)}{m} \cdot \left(B^2 - \|z-\mu\|^2 \right) \\
&= \frac{1}{m} \cdot \left(C(m) \cdot \left(\mathop{\mathbf{E}}_{s \sim \mathcal{D}} \left[\|s-\mu\|^2 \right] - \|z-\mu\|^2 \right) + \left(\mathop{\mathbf{E}}_{s \sim \mathcal{D}} \left[\|s-\mu\|^2 \right] - \mathop{\mathbf{E}}_{S \sim \mathcal{D}^m} \left[\|\hat{\mu}_S - \mu\|^2 \right] \right) \right)
\end{aligned}$$

where $C(m) = 2 - 1/m + c(m) = \Theta(1)$ is an explicit function of m, and we use the fact that $\mathbf{E}_{S \sim \mathcal{D}^m} \left[\|\hat{\mu}_S - \mu\|^2 \right] = \frac{1}{m} \cdot B^2$. □

C Stability of Distributional Shapley Values – Omitted Proofs

C.1 RKHS empirical risk minimization is Lipschitz stable.

Reproducing Kernel Hilbert Spaces. A number of our theoretical results focus on supervised learning over Reproducing Kernel Hilbert Spaces (RKHS). Suppose $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$; let $\mathcal{F}$ to denote a RKHS, with associated feature map $\varphi : \mathcal{X} \to \mathcal{F}$, inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$, and norm $\|\cdot\|_{\mathcal{F}}$, such that for all $f \in \mathcal{F}$,

$$f(x) = \langle f, \varphi(x) \rangle_{\mathcal{F}}$$

Given $\mathcal{F}$, we define a natural metric over $\mathcal{Z}$, where for labeled pairs $z_i = (x_i, y_i)$ and $z_j = (x_j, y_j)$,

$$d_{\mathcal{F}}(z_i, z_j) = \begin{cases} \|\varphi(x_i) - \varphi(x_j)\|_{\mathcal{F}} & \text{if } y_i = y_j \\ +\infty & \text{o.w.} \end{cases}$$

That is, the distance is given by the RKHS norm if x_i and x_j have the same label, and are arbitrarily dissimilar otherwise.

We define the potential of a subset $S \subseteq \mathcal{Z}$ for an RKHS learning problem as the performance achieved when training using S. Specifically, suppose $\ell : [0,1] \times [0,1] \to \mathbb{R}^+$ is an L-Lipschitz, convex loss function and $\mathcal{D}$ is a distribution supported on $\mathcal{Z}$. We define the potential function $U_{\mathcal{F}} : \mathcal{Z}^* \to [0,1]$ in terms of the population loss over $\mathcal{D}$ of the following regularized ERM.

$$\begin{aligned}
U_{\mathcal{F}}(S) &= 1 - \mathop{\mathbf{E}}_{(x,y) \sim \mathcal{D}} \left[\ell(f_S(x), y) \right] \\
\text{where } \quad f_S &= \mathop{\mathrm{argmin}}_{f \in \mathcal{F}} \left\{ \mathrm{er}_S(f) + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2 \right\}
\end{aligned}$$

where $\mathrm{er}_S(f) = \frac{1}{|S|} \sum_{(x,y) \in S} \ell(f(x), y)$ and $\lambda > 0$.

Lemma C.1 (Learning RKHS is Lipschitz stable)**.** *Suppose $\mathcal{F}$ is a RKHS with feature map $\phi : \mathcal{X} \to \mathcal{F}$. Let $\mathcal{D}$ be a distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ such that $\mathbf{E}_{(x,y) \sim \mathcal{D}} [\|\phi(x)\|_{\mathcal{F}}] = B$. Then, $U_{\mathcal{F}} : \mathcal{Z}^* \to [0,1]$ as defined above is $(2L^2 B / \lambda k)$-Lipschitz stable.*

In other words, stability depends on the Lipschitz constant of the loss, the expected norm over $\mathcal{D}$, and the degree of regularization.

Proof. We follow the proof of replacement stability of RKHS from [Aga11] closely. For notational convenience, we drop reference to $\mathcal{F}$ in $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$.

Suppose $S \in \mathcal{Z}^{k-1}$ and suppose $z, z' \in \mathcal{Z}$ are two points such that $z = (x, y)$ and $z' = (x', y)$; i.e. they share the same label. By the definition of $d_{\mathcal{F}}$, this is the only case we need to consider. Let $f, f' \in \mathcal{F}$ denote the empirical risk minimizers over $S \cup \{z\}$ and $S \cup \{z'\}$, respectively.

$$f = \operatorname*{argmin}_{g \in \mathcal{F}} \left\{ \mathrm{er}_{S\cup\{z\}}(g) + \frac{\lambda}{2}\|g\|_{\mathcal{F}}^2 \right\}$$
$$f' = \operatorname*{argmin}_{g \in \mathcal{F}} \left\{ \mathrm{er}_{S\cup\{z'\}}(g) + \frac{\lambda}{2}\|g\|_{\mathcal{F}}^2 \right\}$$

For $\alpha \in [0, 1]$, let

$$f_\alpha = \alpha \cdot f + (1-\alpha)\cdot f'$$
$$f'_\alpha = (1-\alpha)\cdot f + \alpha \cdot f'.$$

By the assumption that ℓ is convex, we can derive the following inequalities for any $(x, y) \in \mathcal{Z}$ and any $S \subseteq \mathcal{Z}$.

$$\begin{aligned} &\ell(f_\alpha(x), y) \le \alpha\cdot\ell(f(x), y) + (1-\alpha)\cdot\ell(f'(x), y) \\ &\implies \mathrm{er}_S(f_\alpha) \le \alpha\cdot\mathrm{er}_S(f) + (1-\alpha)\cdot\mathrm{er}_S(f') \end{aligned} \tag{7}$$

Note that f_α and f'_α are also feasible hypotheses in the Hilbert space. Thus, by the fact that f, f' are ERMs,

$$\mathrm{er}_{S\cup\{z\}}(f) + \frac{\lambda}{2}\|f\|^2 \le \mathrm{er}_{S\cup\{z\}}(f_\alpha) + \frac{\lambda}{2}\|f_\alpha\|^2$$
$$\mathrm{er}_{S\cup\{z'\}}(f') + \frac{\lambda}{2}\|f'\|^2 \le \mathrm{er}_{S\cup\{z'\}}(f'_\alpha) + \frac{\lambda}{2}\|f'_\alpha\|^2$$

Rearranging, and applying convexity, we derive the following inequality.

$$\frac{\lambda}{2}\cdot\left(\|f\|^2 + \|f'\|^2 - \|f_\alpha\|^2 - \|f'_\alpha\|^2\right) \le \mathrm{er}_{S\cup\{z\}}(f_\alpha) - \mathrm{er}_{S\cup\{z\}}(f) + \mathrm{er}_{S\cup\{z'\}}(f'_\alpha) - \mathrm{er}_{S\cup\{z'\}}(f') \tag{8}$$

We can simplify the inner term of the LHS of (8) as follows.

$$\begin{aligned} \|f\|^2 + \|f'\|^2 - \|f_\alpha\|^2 - \|f'_\alpha\|^2 &= \|f\|^2 + \|f'\|^2 - \|f' - \alpha(f'-f)\|^2 - \|f + \alpha(f'-f)\|^2 \\ &= 2\alpha\langle f'-f, f'-f\rangle - 2\alpha^2\|f'-f\|^2 \\ &= 2\alpha(1-\alpha)\|f-f'\|^2 \end{aligned}$$

Then, we can bound the RHS of (8) as follows.

$$\begin{aligned}
&\mathrm{er}_{S\cup\{z\}}(f_\alpha) - \mathrm{er}_{S\cup\{z\}}(f) + \mathrm{er}_{S\cup\{z'\}}(f'_\alpha) - \mathrm{er}_{S\cup\{z'\}}(f') \\
&\qquad \le (1-\alpha)\cdot\big(\mathrm{er}_{S\cup\{z\}}(f') - \mathrm{er}_{S\cup\{z\}}(f) + \mathrm{er}_{S\cup\{z'\}}(f) - \mathrm{er}_{S\cup\{z'\}}(f')\big) &(9)\\
&\qquad = \frac{1-\alpha}{k}\cdot\big(\ell(f'(x),y) - \ell(f(x),y) + \ell(f(x'),y) - \ell(f'(x'),y) + (k-1)\cdot 0\big) &(10)\\
&\qquad = \frac{(1-\alpha)}{k}\cdot\big(\ell(f'(x),y) - \ell(f'(x'),y) - \ell(f(x),y) + \ell(f(x'),y)\big) \\
&\qquad \le \frac{(1-\alpha)L}{k}\cdot\langle f'-f, \varphi(x)-\varphi(x')\rangle &(11)\\
&\qquad \le \frac{(1-\alpha)L}{k}\cdot\|f-f'\|\cdot\|\varphi(x)-\varphi(x')\| &(12)
\end{aligned}$$

where (9) follows by expanding according to (7), and then simplifying; (10) follows by the fact that $\mathrm{er}_{S\cup\{z\}}(f) = \frac{1}{k}\cdot\ell(f(x),y) + \frac{k-1}{k}\mathrm{er}_S(f)$, but the errors on S cancel to 0, $\mathrm{er}_S(f') - \mathrm{er}_S(f) + \mathrm{er}_S(f) - \mathrm{er}_S(f')$; (11) follows by the Lipschitzness of ℓ; and (12) follows by Cauchy-Schwarz.

The analysis above applied for any $\alpha \in [0,1]$; taking $\alpha = 1/2$ implies

$$\begin{aligned}
\frac{\lambda}{4}\cdot\|f-f'\|^2 &\le \frac{L}{2k}\cdot\|f-f'\|\cdot\|\varphi(x)-\varphi(x')\| \\
\implies \|f-f'\| &\le \frac{2L}{\lambda k}\cdot d_{\mathcal{F}}(z,z').
\end{aligned}$$

Because ℓ is L-Lipschitz, we obtain Lipschitz-stability of $U_{\mathcal{F}}$.

$$\begin{aligned}
U_{\mathcal{F}}(S\cup\{z\}) - U_{\mathcal{F}}(S\cup\{z'\}) &= \mathop{\mathbf{E}}_{(x_0,y_0)\sim\mathcal{D}}\big[\ell(f'(x_0),y_0) - \ell(f(x_0,y_0)\big] \\
&\le \mathbf{E}\left[\|\varphi(x_0)\|\right]\cdot\frac{2L^2}{\lambda k}\cdot d_{\mathcal{F}}(z,z') \\
&= \frac{2L^2B}{\lambda k}\cdot d_{\mathcal{F}}(z,z')
\end{aligned}$$

□

C.2 Similar distributions yield similar valuation functions.

For two distributions $\mathcal{D}_s, \mathcal{D}_t$ over $\mathcal{Z}$, let Γ_{st} be the collection of joint distributions over $\mathcal{Z}\times\mathcal{Z}$, whose marginals are $\mathcal{D}_s$ and $\mathcal{D}_t$.[1] Fixing a metric d over $\mathcal{Z}$, the Wasserstein distance is the infimum over all such couplings $\gamma\in\Gamma_{st}$ of the expected distance between $(s,t)\sim\gamma$.

$$W_1(\mathcal{D}_s,\mathcal{D}_t) \triangleq \inf_{\gamma\in\Gamma_{st}} \mathop{\mathbf{E}}_{(s,t)\sim\gamma}[d(s,t)] \tag{13}$$

With this metric, we show the following theorem that formalizes the idea that distributional Shapley values are stable under small perturbations to the underlying data distribution.

[1] That is, for all $\gamma\in\Gamma_{st}$, if $(s,t)\sim\gamma$, then $s\sim\mathcal{D}_s$ and $t\sim\mathcal{D}_t$.

Theorem (Restatement of Theorem 2.7)**.** *Fix a metric space $(\mathcal{Z}, d)$ and let $U : \mathcal{Z}^* \to [0,1]$ be $\beta(k)$-Lipschitz stable with respect to d. Suppose $\mathcal{D}_s$ and $\mathcal{D}_t$ are two distributions over $\mathcal{Z}$. Then, for all $m \in \mathbb{N}$ and all $z \in \mathcal{Z}$,*

$$|\nu(z; U, \mathcal{D}_s, m) - \nu(z; U, \mathcal{D}_t, m)| \leq \frac{2}{m} \sum_{k=1}^{m-1} k\beta(k) \cdot W_1(\mathcal{D}_s, \mathcal{D}_t).$$

Proof. For notational convenience, for any $z \in \mathcal{Z}$ and subset $S \subseteq \mathcal{Z}$, we denote $\Delta_z U(S) = U(S \cup \{z\}) - U(S)$. Thus, fixing $z \in \mathcal{Z}$, we can write $\nu(z; U, \mathcal{D}, m)$ as $\mathbf{E}_{k\sim[m]} \mathbf{E}_{S\sim\mathcal{D}^{k-1}} [\Delta_z U(S)]$. We analyze $\mathbf{E}_{S\sim\mathcal{D}^{k-1}} [\Delta_z U(S)]$ for each fixed $k \in \{2, \ldots, m\}$ separately.[2]

Let $\gamma \in \Gamma_{st}$ be some coupling of $\mathcal{D}_s$ and $\mathcal{D}_t$. Then, we can expand the expectation as follows.

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim\mathcal{D}_s^{k-1}} [\Delta_z U(S)] &= \mathop{\mathbf{E}}_{S\times T\sim\gamma^{k-1}} [\Delta_z U(S)] && (14)\\
&= \mathop{\mathbf{E}}_{S\times T} [\Delta_z U(S) - \Delta_z U(T)] + \mathop{\mathbf{E}}_{S\times T} [\Delta_z U(T)] && (15)\\
&= \mathop{\mathbf{E}}_{S\times T} [\Delta_z U(S) - \Delta_z U(T)] + \mathop{\mathbf{E}}_{T\sim\mathcal{D}_t^{k-1}} [\Delta_z U(T)] && (16)
\end{aligned}$$

where (14) and (16) follow by the assumption that the marginals of γ are $\mathcal{D}_s$ and $\mathcal{D}_t$; and (15) follows by linearity of expectation.

To bound the first term of (16), we expand the difference between $\Delta_z U(S)$ and $\Delta_z U(T)$ into a telescoping sum of k pairs of terms, where we bound each pair to depend on a single draw $(s_i, t_i) \sim \gamma$. For $S, T \in \mathcal{Z}^k$ and $i \in \{0, \ldots, k\}$, denote by $Z_i = \left(\bigcup_{j=i+1}^{k} s_j\right) \cup \left(\bigcup_{j=1}^{i} t_j\right)$; note that $Z_0 = S$ and $Z_k = T$. Then, we can rewrite $\Delta_z U(S) - \Delta_z U(T)$ as follows.

$$\Delta_z U(S) - \Delta_z U(T) = \sum_{i=1}^{k} \Delta_z U(Z_{i-1}) - \Delta_z U(Z_i)$$

Now suppose U is $\beta(k)$-Lipschitz stable with respect to d; note that this implies $\Delta_z U$ is $2\beta(k)$-Lipschitz stable (because β is non-increasing). Then, we obtain the following bound.

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\times T\sim\gamma^{k-1}} [\Delta_z U(S) - \Delta_z U(T)] &= \mathop{\mathbf{E}}_{S\times T\sim\gamma^{k-1}} \left[\sum_{i=1}^{k-1} \Delta_z U(Z_{i-1}) - \Delta_z U(Z_i)\right]\\
&= \sum_{i=1}^{k-1} \mathop{\mathbf{E}}_{S,T\sim\gamma^{k-1}} [\Delta_z U(Z_{i-1}) - \Delta_z U(Z_i)]\\
&= \sum_{i=1}^{k-1} \mathop{\mathbf{E}}_{\substack{s_i,t_i\sim\gamma\\ R\in\mathcal{Z}^{k-2}}} [\Delta_z U(R \cup \{s_i\}) - \Delta_z U(R \cup \{t_i\})] && (17)\\
&\leq 2\beta(k-1) \cdot \sum_{i=1}^{k-1} \mathop{\mathbf{E}}_{(s_i,t_i)\sim\gamma} [d(s_i, t_i)] && (18)\\
&\leq 2(k-1)\beta(k-1) \cdot \mathop{\mathbf{E}}_{(s,t)\sim\gamma} [d(s, t)] && (19)
\end{aligned}$$

[2]Note that for a fixed potential U, $m = 1$ is uninteresting because both sides of the inequality are 0; in particular, $|S|$ is always 0, so the LHS is given by the difference $U(z) - U(z)$.

where (17) notes Z_{i-1} and Z_i differ on only the ith data point; (18) follows from the assumption that $\Delta_z U$ is $2\beta(k)$-Lischitz stable and linearity of expectation; and finally (19) follows by the fact that each draw from γ is i.i.d.

Finally, we note that the argument above worked for an arbitrary coupling in Γ_{st}; thus, we can express the difference in values in terms of the infimum over Γ_{st}.

$$\begin{aligned}
&\nu(z; U, \mathcal{D}_s, m) - \nu(z; U, \mathcal{D}_t, m) \\
&\leq \inf_{\gamma \in \Gamma_{st}} \mathop{\mathbf{E}}_{k \sim [m]} \left[\mathop{\mathbf{E}}_{S \times T \sim \gamma^{k-1}} \left[\Delta_U(S) - \Delta_z U(T) \right] \right] \\
&\leq \frac{2}{m} \sum_{k=2}^{m} (k-1)\beta(k-1) \inf_{\gamma \in \Gamma_{st}} \mathop{\mathbf{E}}_{(s,t) \sim \gamma} [d(s,t)] \\
&= \frac{2}{m} \sum_{k=1}^{m-1} k\beta(k) \cdot W_1(\mathcal{D}_s, \mathcal{D}_t)
\end{aligned}$$

where the first summation is taken over $k \in \{2, \ldots, m\}$ as the term associated with $k = 1$ is 0. □

Recall that often in applications, we take the potential function to depend on the underlying data distribution. For instance, we may take $U_{\mathcal{D}}(S) = \mathbf{E}_{z \sim \mathcal{D}} [\ell_S(z)]$, where $\ell_S(z)$ is the loss on a point $z \in \mathcal{Z}$ achieved by a model trained on the data set $S \subseteq \mathcal{Z}$. In the case where we only have access to samples from $\mathcal{D}_s$, we still may want to guarantee that $\nu(z; U_{\mathcal{D}_s}, \mathcal{D}_s, m)$ and $\nu(z; U_{\mathcal{D}_t}, \mathcal{D}_t, m)$ are close. Thankfully, such a result follows by showing that $U_{\mathcal{D}_s}$ is close to $U_{\mathcal{D}_t}$, and another application of the triangle inequality. As an example, we can show the following guarantee for RKHS.

Corollary C.2. *For two distributions $\mathcal{D}_s, \mathcal{D}_t$ over $\mathcal{X} \times \mathcal{Y}$, let $U_s^{\mathcal{F}}$ and $U_t^{\mathcal{F}}$ be defined as in Appendix D.1 over $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively. Then, for all $z \in \mathcal{X} \times \mathcal{Y}$,*

$$\left| \nu(z; U_s^{\mathcal{F}}, \mathcal{D}_s, m) - \nu(z; U_t^{\mathcal{F}}, \mathcal{D}_t, m) \right| \leq 2(L^2 B + L) \cdot W_1(\mathcal{D}_s, \mathcal{D}_t).$$

Proof. First, we apply Theorem 2.7 with $U = U_s^{\mathcal{F}}$ to bound $\left| \nu(z; U_s^{\mathcal{F}}, \mathcal{D}_s, m) - \nu(z; U_s^{\mathcal{F}}, \mathcal{D}_t, m) \right|$, and then bounding $\left| \nu(z; U_s^{\mathcal{F}}, \mathcal{D}_t, m) - \nu(z; U_t^{\mathcal{F}}, \mathcal{D}_t, m) \right|$. For all $Z \subseteq \mathcal{Z}$,

$$\begin{aligned}
U_s^{\mathcal{F}}(Z) - U_t^{\mathcal{F}}(Z) &= \mathop{\mathbf{E}}_{s \sim \mathcal{D}_s} [\ell_Z(s)] - \mathop{\mathbf{E}}_{t \sim \mathcal{D}_t} [\ell_Z(t)] \\
&= \inf_{\gamma \in \Gamma_{st}} \mathop{\mathbf{E}}_{(s,t) \sim \gamma} [\ell_Z(s) - \ell_Z(t)] \\
&\leq \inf_{\gamma \in \Gamma_{st}} \mathop{\mathbf{E}}_{s,t} [L \cdot d(s,t)] \\
&\leq L \cdot W_1(\mathcal{D}_s, \mathcal{D}_t).
\end{aligned}$$

As $\nu(z; U_s^{\mathcal{F}}, \mathcal{D}_t, m) - \nu(z; U_s^{\mathcal{F}}, \mathcal{D}_t, m)$ can be written as the expectation of two differences between in $U_s^{\mathcal{F}}(Z)$ and $U_t^{\mathcal{F}}(Z)$, applying the triangle inequality implies the claimed inequality. □

D Estimating Distributional Shapley Values – Omitted Proofs

We require the following standard concentration inequality.

Theorem (Hoeffding's Inequality)**.** *Suppose $X_1, \ldots, X_T$ are independent random variables, where X_t is bounded in the range $[-b_t, b_t]$. Let $\overline{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$. Then,*

$$\mathbf{Pr}\left[\left|\overline{X} - \mathbf{E}[\overline{X}]\right| > \varepsilon\right] \leq 2 \cdot \exp\left(\frac{-T^2\varepsilon^2}{2 \cdot \sum_{t=1}^{T} b_t^2}\right)$$

D.1 Iteration complexity of Algorithm 1.

Theorem (Restatement of Theorem 3.1)**.** *Fixing a potential U and distribution $\mathcal{D}$, and $Z \subseteq \mathcal{Z}$, suppose $T \geq \Omega\left(\frac{\log(|Z|/\delta)}{\varepsilon^2}\right)$. Algorithm 1 produces unbiased estimates and with probability at least $1-\delta$, $|\nu(z; U, \mathcal{D}, m) - \nu_T(z)| \leq \varepsilon$. for all $z \in Z$.*

Proof. The theorem follows from the analysis of Theorem D.1, by taking the stability to be trivial, $\beta(k) = 1$, and using uniform sampling $w_k = 1/m$ for all $k \in [m]$. □

D.2 Running time analysis under stability.

Theorem D.1 (Generalizes Theorem 3.2)**.** *Suppose U is $\beta(k)$-deletion stable and for all sets $S \subseteq \mathcal{Z}$ of cardinality $|S| = k$, $U(S)$ can be evaluated in time $R(k)$; consider a set of positive weights $\{w_k : k \in [m]\}$ and $p \in [0,1]$. Algorithm 2 produces unbiased estimates of $\nu(z; U, \mathcal{D}, m)$ that with probability at least $1-\delta$ are ε-accurate for all $z \in Z_p$ and runs in expected time*

$$RT_w(m) \leq O\left(p \cdot |Z| \cdot \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot \left(\sum_{k=1}^{m} \frac{\beta(k)^2}{w_k}\right) \cdot \left(\sum_{k=1}^{m} w_k \cdot R(k)\right)\right)$$

Proof. First, we bound the iteration complexity and time complexity to evaluate models at each iteration within Algorithm 2. The running time bound then follows by the fact that we need to evaluate a model per $z \in Z_p$, per iteration where the expected cardinality of $|Z_p| = p \cdot |Z|$.

Abusing notation, denote by

$$\Delta_z U(k) = \mathop{\mathbf{E}}_{S \sim \mathcal{D}^{k-1}}\left[U(S \cup \{z\}) - U(S)\right].$$

Suppose we sample k according to a possibly non-uniform discrete distribution where $\mathbf{Pr}[k \in m] = w_k$; we denote a random drawn from this distribution by $k \sim [m]_w$. Then, by sampling $k \sim [m]_w$, computing $\Delta_z U(S)$ for $S \sim \mathcal{D}^k$, and reweighting, we obtain an unbiased estimate of $\nu(z; U, \mathcal{D}, m)$.

$$\begin{aligned}
\nu(z; U, \mathcal{D}, m) &= \mathop{\mathbf{E}}_{k \sim [m]}\left[\Delta_z U(k)\right] \\
&= \frac{1}{m}\sum_{k=1}^{m} \Delta_z U(k) \\
&= \sum_{k=1}^{m} w_k \frac{\Delta_z U(k)}{w_k m} \\
&= \mathop{\mathbf{E}}_{k \sim [m]_w}\left[\frac{\Delta_z U(k)}{w_k m}\right]
\end{aligned}$$

For simplicity, we analyze a sampling scheme where we sample T_k sets with cardinality k for $T_k \geq w_k \cdot T$. (By the multiplicative Chernoff bound, this event will occur with high probability.) That is, for all $z \in Z$ and for each $k \in [m]$, we sample T_k subsets $S \sim \mathcal{D}^k$ and compute $\Delta_z U(S)$. For each $z \in Z$, each such sample is an independent unbiased estimate of $\Delta_z U(k)$, so reweighting according to w_k and averaging over $k \in [m]$ gives an unbiased estimate of $\nu(z; U, \mathcal{D}, m)$.

$$\nu_T(z) = \frac{1}{T} \sum_{k=1}^{m} \sum_{t=1}^{T_k} \frac{\Delta_z U(S_t)}{w_k m}$$

Note that by $\beta(k)$-deletion stability, for each k, the terms in the summation associated with $\Delta_z U(k)$ are bounded in magnitude by $\frac{\beta(k)}{w_k m}$. Thus, we can apply Hoeffding's inequality to derive the following bound to obtain ε-error with probability at least $1 - \delta_0$.

$$\begin{aligned}
\delta_0 &\geq 2 \cdot \exp\left(\frac{-\varepsilon^2 T^2}{2 \cdot \sum_{k=1}^{m} \sum_{t=1}^{T_k} \left(\frac{\beta(k)}{w_k m} \right)^2} \right) \\
&\geq 2 \cdot \exp\left(\frac{-\varepsilon^2 T^2}{\frac{2}{m^2} \cdot \sum_{k=1}^{m} w_k \cdot T \cdot \left(\frac{\beta(k)}{w_k} \right)^2} \right) \\
&= 2 \cdot \exp\left(\frac{-\varepsilon^2 m^2 T}{2 \cdot \sum_{k=1}^{m} \frac{\beta(k)^2}{w_k}} \right)
\end{aligned}$$

Thus, taking the failure probability $\delta_0 = \delta / |Z|$ small enough to union bound over all $z \in Z$, we derive the following bound on T.

$$T \geq \Omega\left(\frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot \sum_{k=1}^{m} \frac{\beta(k)^2}{w_k} \right)$$

Using this bound on T, we can compute the necessary running time for Algorithm 2 in terms of $R(k)$ per $z \in Z_p$.

$$\begin{aligned}
T_w(m) &= T \cdot \sum_{k=1}^{m} w_k \cdot R(k) \\
&= \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot \left(\sum_{k=1}^{m} \frac{\beta(k)^2}{w_k} \right) \cdot \left(\sum_{k=1}^{m} w_k \cdot R(k) \right)
\end{aligned}$$

Thus, we can compare various sampling schemes for different stability factors. Note that in the case of the uniform sampling scheme, where $w_k = 1/m$ for all $k \in [m]$, the sampling probabilities cancel.

$$\begin{aligned}
T_u(m) &= \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot \left(\sum_{k=1}^{m} \frac{\beta(k)^2}{1/m} \right) \cdot \left(\sum_{k=1}^{m} 1/m \cdot R(k) \right) \\
&= \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot \left(\sum_{k=1}^{m} \beta(k)^2 \right) \cdot \left(\sum_{k=1}^{m} R(k) \right)
\end{aligned}$$

Thus, the overall running time is given as $RT_w(m) = p \cdot |Z| \cdot \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot \left(\sum_{k=1}^{m} \frac{\beta(k)^2}{w_k} \right) \cdot (\sum_{k=1}^{m} w_k \cdot R(k))$. □

Concretely, to see the special case of the theorem stated as Theorem 3.2, suppose that $\beta(k) = k^{-b}$ for $b \geq 1/2$ and $R(k) = k^c$ for $c \geq 1$. With these settings of the parameters, uniform sampling takes time

$$\begin{aligned} T_u(m) &= \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot \left(\sum_{k=1}^{m} k^{-2b} \right) \cdot \left(\sum_{k=1}^{m} k^c \right) \\ &= \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot O(\log(m)) \cdot O(m^{c+1}) \\ &\leq \frac{\log(|Z|/\delta)}{\varepsilon^2} \cdot \tilde{O}\left(m^{c-1}\right). \end{aligned}$$

Suppose, instead, we take $w_k \propto k^{1-2b}$; that is, we choose w_k such that the first summation will still be bounded by $H_m = \Theta(\log(m))$. Under such a sampling scheme, the running time is bounded as

$$\begin{aligned} T_w(m) &= \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot \left(\sum_{k=1}^{m} k^{-1} \right) \cdot \left(\sum_{k=1}^{m} k^{c+1-2b} \right) \\ &= \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot O(\log(m)) \cdot O(m^{c+2-2b}) \\ &\leq \frac{\log(|Z|/\delta)}{\varepsilon^2} \cdot \tilde{O}\left(m^{c-2b}\right). \end{aligned}$$

In other words, the biased sampling scheme allows us to save roughly a factor-m^{2b-1} in computation time. Thus, if U is $O(1/k)$-deletion stable, then the biased sampling scheme saves roughly a factor m in computation time.

D.3 Finite Sample Approximation to $\mathcal{D}$

While the analysis of Theorem D.1 focuses on the running time of Algorithm 2, we can equally interpret it as a sample complexity bound. In particular, taking $R(k) = k$ corresponds to the sample complexity of taking a fresh sample $S \sim \mathcal{D}^k$ per iteration. Thus, using Algorithm 2, the naive sample complexity given by resampling for each iteration gives the bound of

$$M \approx \frac{\log(|Z|/\delta)}{\varepsilon^2 m^2} \cdot \left(\sum_{k=1}^{m} \frac{\beta(k)^2}{w_k} \right) \cdot \left(\sum_{k=1}^{m} w_k \cdot k \right)$$

Taking $\beta(k) = w_k = 1/k$ yields

$$\begin{aligned} M &\approx \frac{\log(|Z|/\delta)}{\varepsilon^2 m} \cdot \left(\sum_{k=1}^{m} \frac{1}{k} \right) \\ &\approx \frac{\log(m) \cdot \log(|Z|/\delta)}{\varepsilon^2 m} \end{aligned}$$

E Empirical Performance

E.1 Empirical performance of speed-up methods

Following the point removal experiment in [GZ19], for Adult Income prediction task and the UK Biobank breast cancer prediction task, we examine each speed-up trick's (interpolation and biased sampling) for different levels of computational savings. To have an extensive view of the performance of these methods, we examine each task using four different predictive models. As it is shown in Fig. 1, there is no significant drop in performance by making the computation of D-Shapley even 10 times faster using either of the tricks.

E.2 Speeding-up Distributional Shapley for Cifar10

In this experiment, we apply both speed-up methods to compute the D-Shapley values for Cifar10 dataset. We use biased sampling with a speed-up factor of 10 and interpolation with a speed-up factor of 50 (we compute values for 1000 data points). The model is an Inception-v3 [SVI$^+$16] model pretrained on ILSVRC2012 (Imagenet) [RDS$^+$15] dataset with all layers frozen except the last layer. Fig. 2 shows the point-removal results for the complete dataset (e.g. removing 50% of the points with the highest D-Shapley value drops the prediction accuracy from 77% to 68%.)

F Case Study

We report the complete results for Section 4 experimental setting. As mentioned, we use four large scale datasets from the UCI repository [DG17]: 1- Covertype dataset with 581012 samples where each sample contains 54 visual features of forest images and the task is to detect the type of forest cover (from a set of 7 different covers). We use a Random Forest model. 2- Diabetes130 dataset [SDG$^+$14] that with 100000 samples. Each sample containts 54 patient and hospital features. The task is to predict whether the patient will be readmitted to the hospital. We use an AdaBoost model. 3- Wearable Computing: Classification of Body Postures and Movements (PUC-Rio) Data Set which has 165632 points where each point has 18 attributes and has one of the 5 postures. We use a multinomial logistic regression model. 4- Dataset for Sensorless Drive Diagnosis Data Set that contains 58509 data points. Each data point has 48 features. The dataset has 11 classes. We use a Gradient Boosting model.

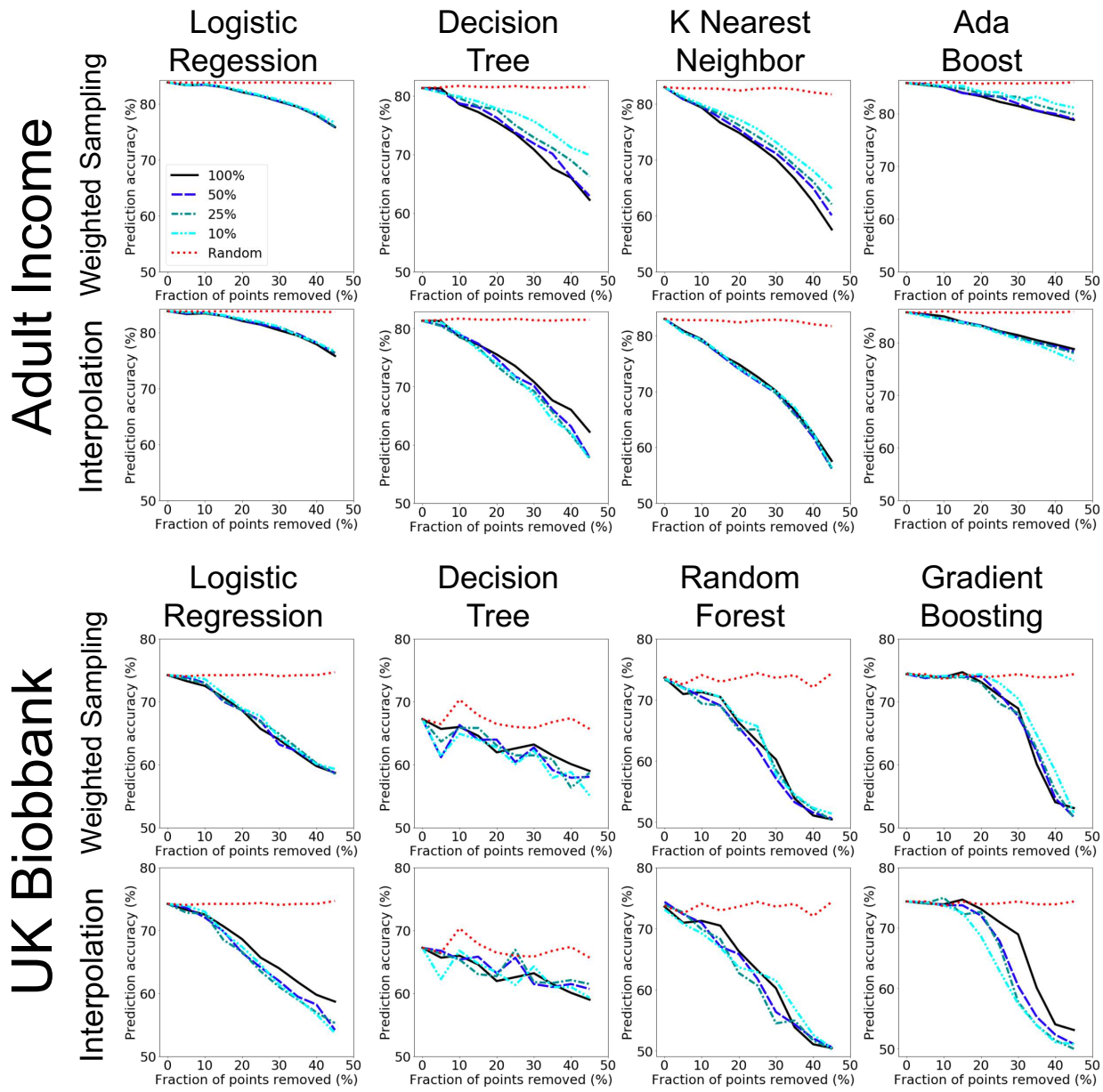

Figure 1: **Speedup performance** For two ML tasks, we apply the interpolation and biased sampling speed-up tricks separately. Here we depict the performance (in rank correlation and R^2 coefficient metrics) of this two methods for the two tasks. For each task, results for four different learning algorithms are shown. We also show the point removal experiment introduced in [GZ19] for several levels of computational cost.

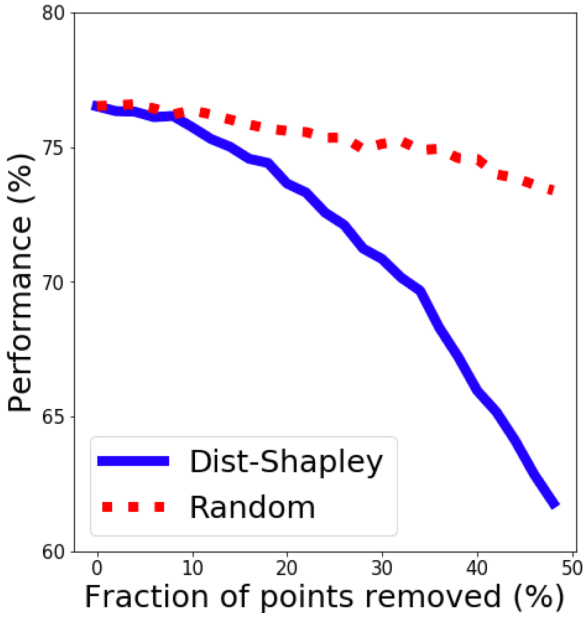

Figure 2: **Fast-$\mathcal{D}$-Shpaley for CIFAR10**

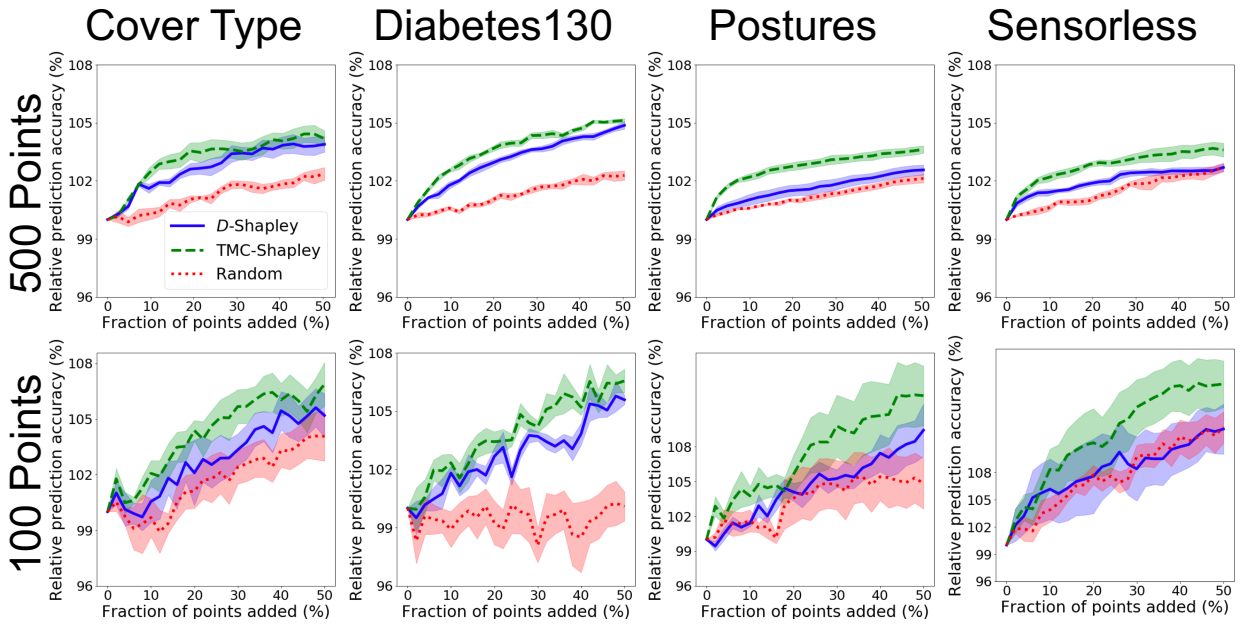

Figure 3: **Consistent data pricing** Points from an acquired set are added to the buyer's initial dataset in three different orders: according to ν ($\mathcal{D}$-Shapley), according to ϕ (TMC), and randomly. The plot shows the change in the accuracy of the model, relative to its performance using the buyer's initial dataset, as the points are added; shading indicates standard error of the mean.

References

[Aga11] Shivani Agarwal. Algorithmic stability. Lecture notes on Statistical Learning Theory, 2011.

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[GZ19] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251, 2019.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[SDG+14] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[Sha53] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[SVI+16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.