Validation Free and Replication Robust Volume-based Data Valuation

Xinyi Xu[†§*], **Zhaoxuan Wu**[‡¶*], **Chuan Sheng Foo**[§], **Bryan Kian Hsiang Low**[†]
Dept. of Computer Science, National University of Singapore, Republic of Singapore[†]
Institute of Data Science, National University of Singapore, Republic of Singapore[‡]
Integrative Sciences and Engineering Programme, NUSGS, Republic of Singapore[¶]
Institute for Infocomm Research, A*STAR, Republic of Singapore[§]
{xuxinyi,lowkh}@comp.nus.edu.sg[†]
wu.zhaoxuan@u.nus.edu[‡¶]
foo_chuan_sheng@i2r.a-star.edu.sg[§]

Abstract

Data valuation arises as a non-trivial challenge in use cases such as collaborative data sharing, data markets, among others. The value of data is often associated with the learning performance (e.g., validation accuracy) of the model trained on the data. This intuitive methodology introduces a high coupling between data valuation and validation. This may be undesirable because a validation set may not be available in practice, and it can be challenging for the data providers to reach an agreement on the choice of the validation set. A separate but practical issue is data replication. Given the value of some data points, a dishonest data provider may replicate these data points to exploit the valuation for a higher reward/payment. We observe that the diversity of the data points is an inherent property of the dataset that is independent of validation. We formalize diversity via the *volume* of the data matrix (determinant of its left Gram). This allows us to formally connect the diversity of data to the learning performance without requiring validation. Furthermore, we propose a *robust volume* with theoretical replication robustness guarantees by following the intuition that copying the same data points does not increase the diversity in data. We perform extensive experiments to demonstrate its consistency and practical advantages over existing baselines and show that our method is model- and task-agnostic and flexibly adaptable to various neural networks.

1 Introduction

Data is increasingly recognized as a valuable resource [17], so we need a principled way to measure its worth. A suitable data valuation has wide-ranging applications such as fairly compensating clinical trial researchers for their collected data [9, 14, 24], fostering collaborative machine learning among industrial organizations [35, 36, 40], and formulating data markets and a data economy [2, 4, 30, 32].

A popular viewpoint is that the value of data should correlate with the learning performance of the model trained on it [11, 16]. These intuitive methods enforce a high coupling between data valuation and validation, so they face certain practical limitations. In practice, a validation set may not always be available [35]. Furthermore, as different choices of the validation set can lead to different data valuations, it is difficult for the data providers to agree on the choice of such a validation set [35]. Since the valuation is coupled with the validation, if the validation set is *not* a good representation

[*]Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

of the actual application scenario, then the obtained valuation may be less useful [39]. We adopt a different perspective that data value should be related to the intrinsic properties of data, and decouple valuation from validation by considering the inherent diversity in the data. Intuitively, a more diverse collection of data points corresponds to a higher-quality dataset and thus a higher value. This approach circumvents the above practical limitations and allows our valuation method to be model- and task-agnostic. The diversity is formalized by the *volume* of the data matrix.

Data replication is another practical issue due to the digital nature and anonymous settings of data markets [12]. Supposing a dataset has some value and a data provider instead offers one containing two copies of every data point, is this "new" dataset twice as valuable as the original one? Intuitively, the answer should be no as replication adds no new data and so does not increase diversity. We formalize this intuition to guarantee replication robustness. Specifically, we construct a compressed version of the original data to preserve its inherent diversity and assign little value to replicated data.

We provide theoretical justification for formalizing diversity with volume. Firstly, the diversity should be non-negative and monotonic [11, 16, 35, 38], and the volume satisfies both of these properties. Subsequently, a higher diversity should indicate better learning performance [21]. We formally show that larger volumes generally correspond to better performance using the *ordinary least squares* (OLS) framework and extend our method to more complex models (i.e., various neural networks) in our empirical investigation. Specifically, data with larger volumes can lead to more accurate pseudo-inverses (a key part of the least squares solution) and lower mean squared errors.

To guarantee replication robustness, we find that the marginal increase in value from replication must diminish to zero. Otherwise, a data provider can exploit this valuation by making infinite copies of the data to achieve infinite value. We thus formalize a notion of replication robustness via the asymptotic value attainable through replication. Unfortunately, the conventional volume definition does not have this property. We propose a *robust volume* (RV) which groups similar data together to construct a compressed version of the original data using the statistics of these data groups. Moreover, we show that RV leads to similar valuations as the conventional volume when there is no replication and is robust if replication exists. We perform extensive experiments on synthetic and real-world datasets to demonstrate that our method produces consistent valuations with existing methods while making fewer assumptions.

Our specific contributions include:

- Formalizing a measure of data diversity via the volume of data and justifying (both theoretically and empirically) the suitability of volume for data valuation;
- Formalizing a notion of replication robustness and designing a *robust volume* (RV) valuation with robustness guarantees;
- Performing extensive empirical comparisons with baselines to demonstrate our method's consistency in valuation, replication robustness without validation, and flexible adaptability to more complex machine learning models, i.e., various *neural networks*.

2 Problem Setting and Notations

Consider two data submatrices $\mathbf{X}_S$ and $\mathbf{X}_{S'}$ to be valued that contain s and s' rows of d-dimensional feature vectors, respectively. We concatenate them along the rows to form the full data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, i.e., $\mathbf{X} \coloneqq [\mathbf{X}_S^\top \ \mathbf{X}_{S'}^\top]^\top$ and $s + s' = n$. Similarly, we denote the corresponding observations as $\mathbf{y} \coloneqq [\mathbf{y}_S^\top \ \mathbf{y}_{S'}^\top]^\top \in \mathbb{R}^{n \times 1}$. The least squares solution from OLS is $\mathbf{w} \coloneqq \mathbf{X}^+\mathbf{y} = \operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ where $\mathbf{X}^+ \coloneqq (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the pseudo-inverse of $\mathbf{X}$. Similarly, we denote $\mathbf{X}_S^+$ as the pseudo-inverse of $\mathbf{X}_S$ and $\mathbf{w}_S \coloneqq \mathbf{X}_S^+\mathbf{y}_S$. For notational brevity, let $V \coloneqq \mathrm{Vol}(\mathbf{X})$ and $V_S \coloneqq \mathrm{Vol}(\mathbf{X}_S)$ where Vol() is defined below. Let $|\mathbf{A}| \coloneqq \det(\mathbf{A})$. The left Gram matrix of $\mathbf{X}$ is $\mathbf{G} \coloneqq \mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{d \times d}$, so for submatrix $\mathbf{X}_S$, $\mathbf{G}_S \coloneqq \mathbf{X}_S^\top\mathbf{X}_S$.

Definition 1 (Volume). *For a full-rank* $\mathbf{X} \in \mathbb{R}^{n \times d}$ *with* $n \geq d$, $\mathrm{Vol}(\mathbf{X}) \coloneqq \sqrt{|(\mathbf{X}^\top\mathbf{X})|} = \sqrt{|\mathbf{G}|}$.

We adopt this volume definition for several reasons: (a) Often, the feature space of data is predetermined and fixed due to the data collection process. But, new data can become available incrementally, thus implying that n can grow indefinitely with d fixed [6, 7]. (b) By leveraging the theoretical connections between volume and learning performance, we can design a volume-based data valuation to assign higher value for data that lead to better learning performance. (c) This allows

an intuitive interpretation between volume and diversity: Adding a data point to a dataset can increase the diversity/volume depending on the data points already in the dataset (Lemma 1).

We restrict our discussion to full-rank matrices $\mathbf{X}$, $\mathbf{X}_S$, and $\mathbf{X}_{S'}$ since otherwise we can adopt the Gram-Schmidt process to remove the linearly dependent columns [6, 7]. In practice, we perform pre-processing such as principal component analysis to reduce the feature space dimension to ensure that this assumption is satisfied. This assumption is to ensure that there are no redundant features, namely, features that can be exactly reconstructed with other features. For instance, if a dataset already contains monthly salaries, then an annual salary would be redundant.

3 Larger Volumes Yield Better Learning Performance

The value of a data (sub)matrix depends on the learning performance trained on it [11, 16], which we will show depends on its volume. Simply put, the larger the volume, the better the learning performance. In this section, we formalize this claim through the OLS framework. In particular, we investigate two metrics for learning performance: (a) the quality of the pseudo-inverse formalized as $\text{bias}_S \coloneqq \left\|\mathbf{X}_S^+ - \mathbf{X}^+\right\|$ because estimating $\mathbf{X}^+$ accurately is important in achieving low *mean squared error* (MSE) [6], and (b) the MSE as $L(\mathbf{w}_S) \coloneqq \|\mathbf{y} - \mathbf{X}\mathbf{w}_S\|^2$.

3.1 A Larger Volume Corresponds to a Smaller Bias

In regression problems, the closed-form optimal solution is constructed via $\mathbf{X}^+$ computed on $\mathbf{X}$, so the bias between $\mathbf{X}_S^+$ and $\mathbf{X}^+$ indirectly determines the value of $\mathbf{X}_S$ [6], i.e., smaller bias means higher value. We show that 'a larger volume means a smaller bias' is always true for $d = 1$. But, for $d > 1$, it requires additional assumptions which are mostly satisfied via empirical verification (Fig. 1).

Proposition 1 (Volume vs. Bias for $d = 1$). *For non-zero $\mathbf{X}_S, \mathbf{X}_{S'}$ of $\mathbf{X} \in \mathbb{R}^{n\times 1}$, $V_S \geq V_{S'} \iff \text{bias}_S - \text{bias}_{S'} \leq 0$.*

We can generalize to $M > 2$ non-zero submatrices: Let $\mathbf{X} = [\mathbf{X}_{S_1}^\top \ \mathbf{X}_{S_2}^\top \ \cdots \ \mathbf{X}_{S_M}^\top]^\top$ and w.l.o.g., assume $V_{S_1} \geq V_{S_2} \geq \ldots \geq V_{S_M}$. Then, $\text{bias}_{S_1} \leq \text{bias}_{S_2} \leq \ldots \leq \text{bias}_{S_M}$. For $d > 1$, there exist counterexamples (see Fig. 1), so we compare the bias, as follows:

Proposition 2 (Volume vs. Bias in General). *For full-rank $\mathbf{X}_S, \mathbf{X}_{S'}$ of $\mathbf{X} \in \mathbb{R}^{n\times d}$,*

$$\text{bias}_S^2 - \text{bias}_{S'}^2 = \frac{1}{V_S^4}\left\|\mathbf{X}_S\mathbf{Q}_S\right\|^2 - \frac{1}{V_{S'}^4}\left\|\mathbf{X}_{S'}\mathbf{Q}_{S'}\right\|^2 + 2\left\langle \frac{1}{V^2}\mathbf{Q}\mathbf{X}^\top, \frac{1}{V_{S'}^2}\mathbf{Q}_{S'}\mathbf{X}_{S'}^\top - \frac{1}{V_S^2}\mathbf{Q}_S\mathbf{X}_S^\top \right\rangle$$

where $\mathbf{Q} \coloneqq \sum_{l=1}^k (\lambda_l\sigma_l)^{-1}\prod_{j=1,j\neq l}^k(\mathbf{G} - \lambda_j\mathbf{I})$, $\{\lambda_l\}_{l=1}^k$ denotes the k unique eigenvalues of the left Gram matrix $\mathbf{G}$ of $\mathbf{X}$, $\mathbf{Q}_S, \mathbf{Q}_{S'}$ are similarly defined w.r.t. $\mathbf{G}_S, \mathbf{G}_{S'}$, and $\sigma_l \coloneqq \sum_{g=1}^k (-1)^{g+1}\lambda_l^{k-g}[\sum_{\mathcal{H}\subseteq\{1,\ldots,k\}\setminus\{l\},|\mathcal{H}|=g-1}(\prod_{h\in\{1,\ldots,k\}\setminus\mathcal{H}}\lambda_h^{-1})]$.

The proof of Proposition 1 relies on a key observation that for $d = 1$, the left Gram matrix is a scalar and the rest of the proof follows. However, it cannot be generalized to that for $d > 1$, so we resort to a different approach. The proof of Proposition 2 requires Lemma 2 in Appendix A.1 which establishes the connection between volume and $\mathbf{G}^{-1}$ using Sylvester's formula. To obtain $V_S \geq V_{S'} \implies \text{bias}_S \leq \text{bias}_{S'}$, there are two cases requiring different additional assumptions: (A) $V_S \gg V_{S'}$, and (B) $\|\mathbf{X}_S\mathbf{Q}_S\| \approx \|\mathbf{X}_{S'}\mathbf{Q}_{S'}\|$ and $V \gg \max(V_S, V_{S'})$. Case A is intuitive: $V_S \gg V_{S'}$ means $\mathbf{X}_S$ is much 'larger' than $\mathbf{X}_{S'}$, so bias_S is smaller. Case B is when $\mathbf{X}_S$ and $\mathbf{X}_{S'}$ are similar, e.g., when they are sampled from the same distribution. In this case, we show $V \gg \max(V_S, V_{S'})$ (Lemma 3 in Appendix A.1) and verify that the implication is true most of the time (over 80% of random trials) in Fig. 1.

3.2 A Larger Volume Corresponds to a Lower MSE

We show a similar result for MSE when $d = 1$, which may be surprising since $\text{Vol}()$ does not consider $\mathbf{y}$ at all and yet, it can determine which submatrix predicts better on the rest of the unseen data. Unfortunately, this result does not directly generalize to $d > 1$ or beyond two submatrices. Nevertheless, we analyze how the difference in MSEs depends on volumes to shed light on how volume affects learning performance in more complicated scenarios.

Proposition 3 (Volume vs. MSE for $d = 1$). *For non-zero* $\mathbf{X}_S, \mathbf{X}_{S'}$ *of* $\mathbf{X} \in \mathbb{R}^{n\times 1}$, $V_S \geq V_{S'} \iff L(\mathbf{w}_S) - L(\mathbf{w}_{S'}) \leq 0$.

Unfortunately, it does not generalize to $d > 1$. For full-rank $\mathbf{X}_S, \mathbf{X}_{S'}$ of $\mathbf{X} \in \mathbb{R}^{n\times d}$, we rewrite $L(\mathbf{w}_S) - L(\mathbf{w}_{S'})$, as follows (derivations in Appendix A.2):

$$L(\mathbf{w}_S) - L(\mathbf{w}_{S'}) = \langle \mathbf{w}_S - \mathbf{w}_{S'}, (\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{X}_{S'}^\top \mathbf{X}_{S'})(\mathbf{w}_S + \mathbf{w}_{S'}) - 2\mathbf{X}^\top \mathbf{y} \rangle \,. \tag{1}$$

As $L(\mathbf{w}_S) - L(\mathbf{w}_{S'})$ explicitly depends on $\mathbf{y}$ and $\mathrm{Vol}()$ does not include $\mathbf{y}$ at all, it is possible to adversarially construct $\mathbf{y}$ to have $L(\mathbf{w}_S) - L(\mathbf{w}_{S'}) > 0$ or < 0 for fixed $\mathbf{X}_S, \mathbf{X}_{S'}$ (Appendix A.2).

The adversarial cases notwithstanding, volume is still considered a good indirect measure of the quality of data applied in active learning and matrix subsampling with theoretical performance guarantees [8, 27]. We can also adopt the perspective that $\mathrm{Vol}()$ is a measure of the diversity in the features [21], which provides an intuitive interpretation for the theoretical guarantee: A more diverse dataset (i.e., larger volume) gives better learning performance. We will demonstrate in Sec. 5.2 that not requiring labels can be an advantage in practice if the obtained labels are noisy/corrupted or there is a distributional difference between the validation sets and the test set.

We conclude this section by empirically verifying the additional required assumptions. We randomly and identically sample equal-sized $\mathbf{X}_S, \mathbf{X}_{S'}$ for 500 independent trials and compute the percentage of times that a larger volume leads to better performance (vertical axis) against the size of $\mathbf{X}_S, \mathbf{X}_{S'}$ (horizontal axis). We consider uniform and normal distributions of various dimensions: '$\mathcal{N}$ $d = 1$' denotes $\mathbf{X}_S, \mathbf{X}_{S'}$ drawn from 1-dimensional standard normal distribution. For MSE, the response y for a data point $\mathbf{x}$ is calculated from $y = \sin(\langle \mathbf{w}^*, \mathbf{x} \rangle)$ where the true parameters $\mathbf{w}^*$ are randomly sampled from $U(0, 2)^d$. The left figure shows that a larger volume leads to a smaller bias for more than 80% of times, thus verifying that our assumptions are satisfied. The right figure shows that a larger volume leads to a lower MSE for more than 50% of times for $d \leq 10$, which is consistent with what we expect from (1).

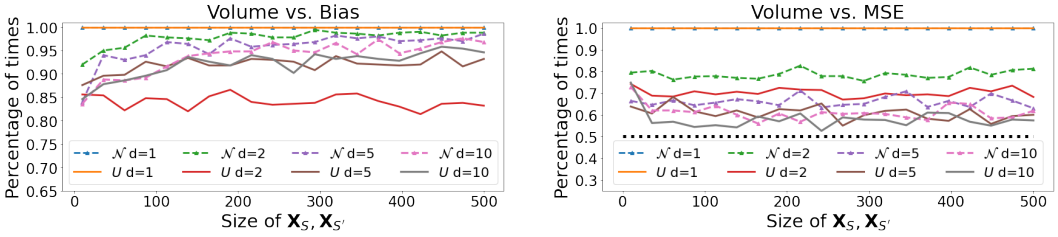

Figure 1: Volume vs. bias (left) and vs. MSE (right) for two identically sampled, equal-sized datasets $\mathbf{X}_S, \mathbf{X}_{S'}$ under two distributions (Uniform: $U(0, 1)^d$, Normal: $\mathcal{N}(0, 1)^d$). The vertical axis shows the percentage of times where the dataset of a larger volume leads to better performance: lower bias or MSE (from 500 independent trials). Here, d denotes the dimension of the dataset.

4 Robustifying the Volume-based Valuation

As larger volumes can indicate better learning performance, we consider a volume-based data valuation method [11, 16, 35]. Unfortunately, the volume (Definition 1) is *not* robust to replication via direct data copying. Hence, we propose a modification to ensure robustness to replication. We conclude by raising an interesting relation involving replication robustness and diversity representation.

4.1 First Attempt of Volume-based Data Valuation

Directly using $\mathrm{Vol}(\mathbf{X})$ as a valuation for $\mathbf{X}$ satisfies both non-negativity and monotonicity:

Proposition 4 (Non-negativity and Monotonicity of $\mathrm{Vol}()$). *For full-rank* $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathrm{Vol}(\mathbf{X}) \geq 0$ *and* $\mathrm{Vol}([\mathbf{X}^\top \ \mathbf{x}^\top]^\top) \geq \mathrm{Vol}(\mathbf{X})$ *where* $\mathbf{x}$ *is a new data point.*

These properties imply that a larger $\mathbf{X}$ (i.e., more data) should correspond to a higher value [11, 16, 35]. However, $\mathrm{Vol}()$ is unbounded and has a multiplicative scaling factor w.r.t. replication. The implication

is that a data provider can arbitrarily "inflate" the volume or value of the data by replicating data infinitely:

Lemma 1 (Unbounded Multiplicative Scaling of $\mathrm{Vol}(\mathbf{X})$ from Replication). *For full-rank $\mathbf{X} \in \mathbb{R}^{n\times d}$, let $\mathbf{x}_q \in \mathbb{R}^{1\times d}$ be a data point that is replicated for $m \geq 1$ times and so, $\mathbf{X}_{\text{rep}} \coloneqq [\mathbf{X}^\top\ \mathbf{x}_q^\top\ \ldots\ \mathbf{x}_q^\top]^\top \in \mathbb{R}^{(n+m)\times d}$. Then, $\mathrm{Vol}(\mathbf{X}_{\text{rep}}) = \mathrm{Vol}(\mathbf{X}) \times (1 + m \times \mathbf{x}_q(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_q^\top)^{1/2}$.*

Replication robustness defined via inflation. Following the previous discussion, we use the term *inflation* to denote the ratio $\nu(\text{replicate}(\mathbf{X}, c))/\nu(\mathbf{X})$ where $\nu()$ is a data valuation function that maps a data matrix to a real value (e.g., $\mathrm{Vol}()$). The function $\text{replicate}(\mathbf{X}; c)$ means directly copying the data in $\mathbf{X}$ and appending them back to $\mathbf{X}$ so that $\mathbf{X}_{\text{rep}} \in \mathbb{R}^{(nc)\times d}$, and the *replication factor* c denotes the amount of replication. One choice of replication is to copy the entire $\mathbf{X}$ for c times. Another way is to copy some selected subset for a certain number of repetitions so that $\mathbf{X}_{\text{rep}} \in \mathbb{R}^{(nc)\times d}$. The second way is considered because replicating different data increases the value differently (Lemma 1). We propose the replication robustness definition to formalize the intuition that a high robustness should guarantee low inflation:

Definition 2 (Replication Robustness of Valuation $\nu()$). *Define the replication robustness as $\gamma_\nu \coloneqq \nu(\mathbf{X})/(\sup_{c\geq 1} \nu(\mathbf{X}_{\text{rep}}))$ where the replicated matrix $\mathbf{X}_{\text{rep}} \coloneqq \text{replicate}(\mathbf{X}, c) \in \mathbb{R}^{(nc)\times d}$.*

The theoretical optimal robustness is $\gamma_\nu = 1$, which implies that there is no additional gain from replicating data, so it eliminates any motivation for replication. In contrast, the worst case is $\gamma_\nu = 0$. It is the case for any valuation that strictly monotonically increases with replication and, in particular, $\gamma_{\mathrm{Vol}} = 0$ by applying Lemma 1. Consequently, a replication robust valuation must have diminishing marginal values from replication. In other words, the additional gain from having more copies of the same data converges asymptotically to 0 with respect to c. This is congruent with what we observe in practice: Adding the same data to the training set repeatedly does not improve performance infinitely.

4.2 Replication Robust Volume-Based Valuation

We propose a robust definition to balance the value of diversity and repetition in data. Specifically, we construct a compressed version of the original data matrix $\mathbf{X}$ by grouping and representing the data points via discretized cubes of the input space:

Definition 3 (Replication Robust Volume (RV)). *Given a discretization coefficient ω, the input domain for $\mathbf{X}$ is discretized into a set of d-cubes with sides of length ω and Ψ denotes the set of indices of these d-cubes. Let ϕ_i denote the number of data points in the i-th d-cube. The replication robust volume is*

$$\mathrm{RV}(\mathbf{X}; \omega) \coloneqq \mathrm{Vol}(\widetilde{\mathbf{X}}) \times \textstyle\prod_{i\in\Psi} \rho_i \tag{2}$$

where $\rho_i \coloneqq \sum_{p=0}^{\phi_i} \alpha^p$, $\widetilde{\mathbf{X}} \coloneqq \{\text{mean}_i | \phi_i \neq 0, i \in \Psi\}$ is a compressed version of $\mathbf{X}$, $\alpha \in (0, 1)$ controls the degree of robustness, and mean_i is a statistic of the data points in the i-th d-cube.

$\widetilde{\mathbf{X}}$ "compresses" $\mathbf{X}$ by grouping similar data together and each row in $\widetilde{\mathbf{X}}$ is constructed via a statistic (e.g., average, trimmed mean) of the data points in a non-empty d-cube. Therefore, the number of rows in $\widetilde{\mathbf{X}}$ is equal to the number of non-empty d-cubes. In contrast, with the unbounded $\mathrm{Vol}()$, we ensure that $\mathrm{RV}(;\omega)$ is bounded by setting $\prod_{i\in\Psi} \rho_i$ to be bounded and convergent w.r.t. the size of replicated data. Note that $\phi_i = 0 \implies \rho_i = 1$ (i.e., an empty d-cube) and $\phi_i > 0 \implies \rho_i > 1$.

Before considering robustness guarantees, we first demonstrate that Definition 3 preserves the original volume in a relative sense, i.e., the ratio between V_S and $V_{S'}$ is preserved. This ensures that Definition 3 has similar guarantees on learning performance from Sec. 3 (empirically demonstrated in Sec. 5.1).

Proposition 5 (Bounded Distortion of $\mathrm{RV}(\mathbf{X}_S;\omega)/\mathrm{RV}(\mathbf{X}_{S'};\omega)\ \forall\omega$). *Set $\alpha = 1/(\beta n)$. Define distortion $\delta(\omega) \coloneqq [\mathrm{RV}(\mathbf{X}_S;\omega)/\mathrm{RV}(\mathbf{X}_{S'};\omega)]/[\mathrm{Vol}(\mathbf{X}_S)/\mathrm{Vol}(\mathbf{X}_{S'})]$. Then, for all ω, $(\exp(\beta^{-1}))^{-1} \leq \delta(\omega) \leq \exp(\beta^{-1})$. For example, $\beta = 10$ bounds $\delta(\omega) \in [0.905, 1.105]$ approximately.*

Achieving near-optimal robustness by upper-bounding the inflation. We define robustness (Definition 2) as the maximum attainable inflation via replication. Since ρ_i and inflation are monotonic in ϕ_i, we consider the asymptotic inflation ($\phi_i \to \infty$). In Definition 3, even if the data in i-th d-cube is

replicated for infinitely many times, the inflation from this d-cube is still upper-bounded by a constant. This can be generalized to all the d-cubes as each can be considered independently and there is a constant number of d-cubes for a fixed $\mathbf{X}$ and ω.

Proposition 6 (Robustness γ_{RV}). *Let $\rho_i \coloneqq \sum_{p=0}^{\phi_i} \alpha^p$. For $\alpha \in (0,1)$, $\gamma_{\mathrm{RV}} \geq (1-\alpha)^{|\Psi|}$ where $|\Psi|$ is the number of non-empty d-cubes; for $\alpha \geq 1, \gamma_{\mathrm{RV}} = 0$. Note that $\gamma = 1$ is optimal.*

Reducing α achieves a lower upper bound on inflation and a higher/better robustness. However, if α is too small, then it may have an undesirable effect: $\mathrm{RV}(\mathbf{X};\omega) < \mathrm{Vol}(\mathbf{X})$ for some $\mathbf{X}$ (with similar data points) from an honest provider without replication. In this situation, RV has an over-correcting effect: RV is designed to avoid exploitation of $\mathrm{Vol}()$ due to replication but mistakenly leads to a decrease in the value of an honest dataset. Therefore, α is set to achieve a certain upper bound on inflation but is not unnecessarily small; more details are given in Proposition 8 in Appendix A.3. In particular, setting $\alpha = 1/(\beta n)$ guarantees a constant upper bound on the inflation of $\exp(\beta^{-1})$ (see Lemma 5 in Appendix A.3). For instance, setting $\beta = 10$ and $\alpha = 1/(\beta n)$ guarantees $\mathrm{RV}(\mathrm{replicate}(\mathbf{X}, c);\omega) \leq 110\% \times \mathrm{RV}(\mathbf{X};\omega)$. However, it requires us to know the true n *without* any replication. In practice, as we only observe the data with replication (if any) [12], we estimate n with the number of rows in $\widetilde{\mathbf{X}}$ (i.e., $|\Psi|$).

Diversity representation vs. replication robustness balance via ω. A smaller ω means that the d-cubes are more refined and RV can better represent the original data instead of crudely grouping many data points together and estimating them via a statistic. On the other hand, a larger ω means lower diversity representation but higher replication robustness. In the extreme case, a sufficiently large ω results in grouping all data points together and representing them all using one single statistic, hence foregoing the diversity in data. Therefore, a balance between them should be achieved depending on the practical requirements of the problem. We formalize this discussion with the following proposition:

Proposition 7 (Reduction to $\mathrm{Vol}()$ vs. Achieving $\gamma = 1$ Robustness). *Set ω to be such that each d-cube only contains completely identical data points, and*

1. *set $\rho_i = K_{\widetilde{\mathbf{X}},i}$ which is a constant from recursive application of Lemma 1. Then, $\mathrm{RV}(\cdot;\omega) = \mathrm{Vol}()$;*
2. *set $\alpha = 0$. So, $\rho_i = \mathbb{1}(\phi_i \neq 0)$ and name this formulation $\mathrm{RV}_{\mathbb{1}}(\cdot;\omega)$. Then, $\gamma_{\mathrm{RV}_{\mathbb{1}}} = 1$.*

$\mathrm{RV}_{\mathbb{1}}(\cdot;\omega)$ can be seen as reducing all potential replications to one data point. It achieves robustness but loses the density information of each d-cube due to the indicator function. Specifically, the true distribution may have different densities at different d-cubes, which is reflected via ϕ_i's. But, this information is completely lost in $\mathrm{RV}_{\mathbb{1}}(\cdot;\omega)$. In contrast, $\mathrm{Vol}()$ represents all the data indiscriminately, hence sacrificing robustness. Furthermore, while we restrict our consideration of replication to direct copying, it is natural to additionally consider a noisy replication (i.e., adding small random perturbations to copies [12]). Intuitively, $\mathrm{RV}_{\mathbb{1}}(\cdot;\omega)$ is not robust to noisy replication as the replicated data are perturbed. Preliminary empirical exploration in Appendices B.2 and B.3 shows that RV is robust to noisy replication if the noise magnitude is small relative to ω. Consequently, an interesting future exploration is to formalize the strategies for striking a balance between diversity representation and replication robustness by modifying ω. For this work, we empirically find $\omega = 0.1$ suitable for standardized features.

In using standardized features, we implicitly assume that the features follow a normal distribution. This makes data further away from the mean (i.e., statistically rarer) more valuable in learning [8]. We also observe this in Sec. 5.2 where data closer to the mean are valued to be lower across all baselines and our method. This work excludes considerations of outliers as they are not truly representative of the actual distribution. All proofs and derivations are in Appendix A.

5 Experiments and Discussion

In this section, we first verify our claim in Sec. 3 that a larger volume leads to better learning performance, and derive some interesting practical perspectives in Sec. 5.1. Subsequently, in Sec. 5.2, we show that RV produces valuations consistent with baseline methods, and additionally demonstrate the limitations of existing methods. In particular, RV is model- and task-agnostic while another baseline with an explicit dependence on the validation set is shown to have some deviations in data

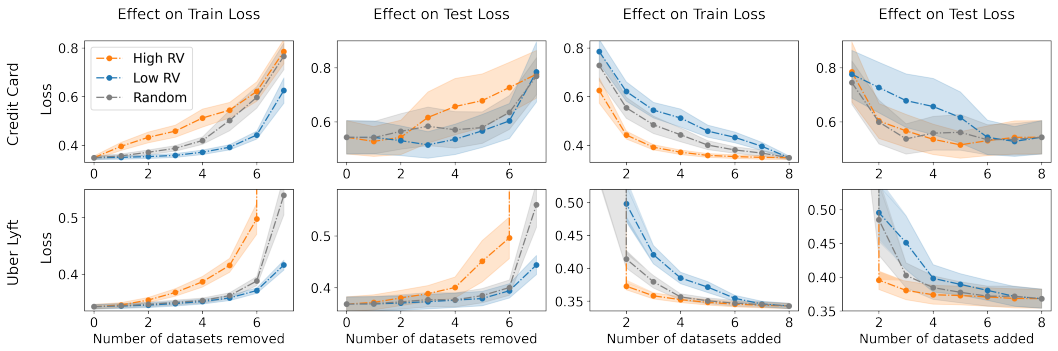

Figure 2: Effect of removing/adding the dataset with the highest/lowest RV on the train/test loss for two real-world datasets: credit card and Uber Lyft. The plots show the average and standard errors over 50 random trials.

valuation as the validation set changes. Lastly, in Sec. 5.3, we verify our robustness guarantees via approximated asymptotic performance. Importantly, our empirical investigation has gone beyond the OLS framework for the theoretical analysis in Sec. 3 as we have adapted our method to various neural network architectures on different machine learning tasks including both image and natural language processing. All experiments were run on a server with Intel(R) Xeon(R)@ 2.70GHz processor and 256GB RAM.

5.1 Robust Volume and Learning Performance

In this subsection, *robust volume* (RV) and volume are interchangeable as we do not consider replication and Proposition 5 guarantees their similarity. We consider the paradigm of sequentially adding/removing the dataset with the highest/lowest RV to observe the trend in model performance [11]. We also include random selection as a baseline. We simulate 8 data providers, so the results are more generalizable. For this experiment, we use two real-world datasets: credit card fraud detection [37] (i.e., transaction amount prediction) and Uber & Lyft [5] (i.e., carpool rides price prediction) which are pre-processed to contain 8 and 12 standardized features, respectively. The results are in Fig. 2. Additional results on two more real-world datasets are in Appendix B.4.

We observe adding (*resp*., removing) a dataset with a high RV leads to a lower (*resp*., higher) train loss, thus verifying Proposition 2 that a larger volume leads to a more accurate pseudo-inverse and lower train loss in terms of mean squared error. Furthermore, a consistent trend is observed for test loss, albeit with higher standard errors. This confirms (1) that in a higher dimensional feature space, a higher volume does not immediately guarantee a lower test loss.

Interesting practical perspectives. The additional experiment provides a justification for data buyers under a constrained budget to spend their budget on datasets with high RVs first in order to get the best performance, thereby resonating with the active learning framework [29]. On the other hand, the removal experiment sheds light on the following question: If the collected datasets are too computationally expensive to learn altogether due to memory or time constraints, then which dataset should be removed first without hurting learning performance (i.e., the dataset with the lowest RV)?

5.2 Robust Volume Shapley Value vs. Baselines

We demonstrate that RV without validation gives results consistent with other methods which may require validation. Subsequently, we show the limitations of the baselines, as revealed in the experiments.

We combine (robust) volume with the commonly used Shapley value to design principled and fair payments to the providers (i.e., relative valuations for the datasets) [11, 16, 35]. We compare with the following baselines: the validation loss *leave-one-out* (LOO) value [19, 26], the *validation loss Shapley value* (VLSV) [11, 16], and the *information gain Shapley value (IGSV)* [35]. Our *volume Shapley value* (VSV) and *robust volume Shapley value* (RVSV) are as follows [33]:

$$\text{RVSV}_m := 1/(M!) \sum_{\mathcal{C} \subseteq \mathcal{M} \setminus \{S_m\}} [|\mathcal{C}|! \times (M - |\mathcal{C}| - 1)!] \times [\text{RV}(\mathbf{X}_{\mathcal{C} \cup \{S_m\}}; \omega) - \text{RV}(\mathbf{X}_{\mathcal{C}}; \omega)]$$

where M is the total number of datasets/data providers, $\mathcal{C} \subseteq \mathcal{M} := \{S_1, \ldots, S_M\}$, $\mathbf{X}_{S_m}$ denotes provider m's data matrix, and we abuse the notation slightly by using $\mathbf{X}_{\mathcal{C}}$ to denote the matrix constructed by concatenating all the data matrices from $\mathcal{C}$. VSV is computed by using $\mathrm{Vol}()$ instead of $\mathrm{RV}(\cdot;\omega)$. We set $M = 3$ and investigate the relative valuations for 3 datasets $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ [35]. The input features are standardized and we set $\omega = 0.1$. LOO and VLSV use MSE on a validation set.

Synthetic data on baseline distributions. We first investigate simpler scenarios on synthetic data and baseline distributions for $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$. We consider the 6D Hartmann function [23] defined over $[0, 1]^6$ and four baseline data distributions: (A) *independent and identical distribution* (i.i.d.) where each of $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ contains 200 samples; (B) ascending size where $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ contain 20, 50, and 200 i.i.d. samples, respectively; (C) disjoint domains where $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ are sampled from the input domains of $[0, 1/3]^6$, $[1/3, 2/3]^6$, and $[2/3, 1]^6$, respectively; and (D) supersets $\mathbf{X}_{S_1} \subset \mathbf{X}_{S_2} \subset \mathbf{X}_{S_3}$ with sizes 200, 400, and 600 where $\mathbf{X}_{S_2}$ (*resp.*, $\mathbf{X}_{S_3}$) has 200 i.i.d. data samples in addition to $\mathbf{X}_{S_1}$ (*resp.*, $\mathbf{X}_{S_2}$). The results are in Fig. 3.

Results in Fig. 3 show that both VSV and RVSV are generally consistent with VLSV and IGSV. For (B) ascending size distribution, both RVSV and IGSV show an increasing trend, while VLSV surprisingly shows approximately equal valuations for all three sizes. It may be attributed to VLSV's sensitivity to the definition of the value on an empty set of data (i.e., $\nu(\emptyset)$ in the Shapley value calculation). Fig. 4 illustrates that VLSV is sensitive to $\nu(\emptyset)$ definitions for i.i.d. $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$. Setting $\nu(\emptyset)$ to 0 [16], to 1.06 (i.e., by initializing parameters to zeros), and to 8.75 (i.e., by random parameter initialization from $\mathcal{N}(0, 1)$ [11]) gives very different VLSV for $\mathbf{X}_{S_1}$ of 0.346, 0.183, and 0.330, respectively. These conflicting choices of $\nu(\emptyset)$ add to the difficulties of applying VLSV in practice.

Interestingly, under (C) disjoint domain distribution, all methods unanimously value $\mathbf{X}_{S_2}$ to be the lowest despite identically sized input domain ranges. This is due to the standardization of the features which offers the following interpretation: The data in the "center" is the most common if we assume the true population follows a normal distribution. Therefore, the most common data are valued less while the statistically "rarer" data at the two tails of the distribution are valued more. Additional experimental results under this distribution are reported in Appendix B.5. The counter-intuitive LOO valuation of $\mathbf{X}_{S_1}$ with 0 under i.i.d. may be attributed to instability due to the calculation of relative values [3].

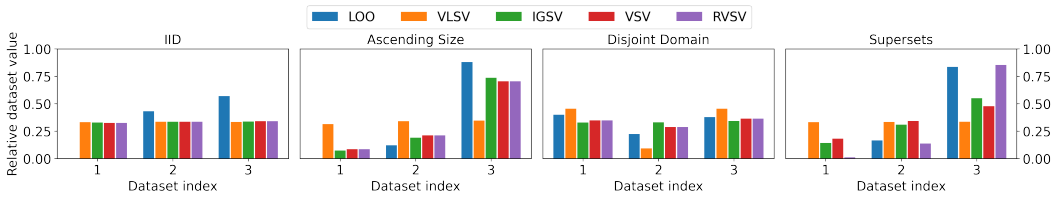

Figure 3: Relative values of $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ on Hartmann function under baseline distributions: (A) i.i.d., (B) ascending dataset size, (C) disjoint input domain, and (D) supersets.

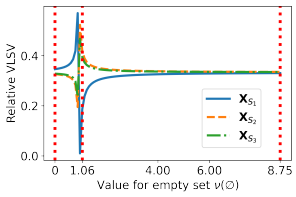

Figure 4: VLSV is sensitive to $\nu(\emptyset)$ definitions. Red dotted lines denote 3 proposed $\nu(\emptyset)$ definitions.

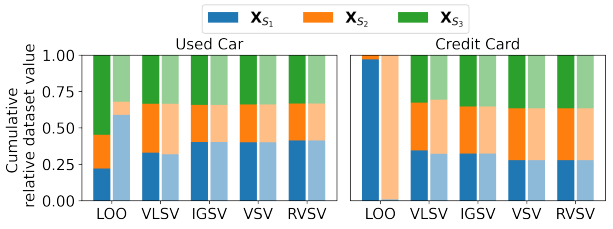

Figure 5: Relative valuations for two different validation sets denoted by darker/lighter shades.

Real-world datasets with different preferences on validation sets. We investigate two real-world datasets: the UK used car dataset [1] (i.e., car price prediction) and the credit card fraud detection dataset [37] (i.e., transaction amount prediction) where there are differences in the choice of the validation set [35]. For instance, car dealers for different manufacturers such as Audi, Ford, and Toyota may have different preferences over data. Thus, we construct two different validation sets

composing cars from different manufacturers. Similarly, different financial institutions may differ in their interests in the size of transactions. For example, smaller banks typically manage and focus on smaller transactions, so we construct two different validation sets composing high ($> \$1000$) or low transactions. The results are in Fig. 5 where the three colors represent $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ and the two shades denote the two different validation sets. The effect of validation choice on LOO is very pronounced, as expected. The effect on VLSV is less due to the averaging of marginal contributions. On the other hand, IGSV, VSV, and RVSV stay consistent as they do not require validation.

5.3 Replication Robustness

We first conduct a simpler experiment to demonstrate the effect due to replication and then perform more extensive experiments under more complicated settings to show asymptotic performance.

Relative valuations in i.i.d. setting. We conduct this experiment on the Trip Advisor hotel reviews dataset [22] (i.e., numerical rating prediction) which contains text reviews data. We utilize the GloVe [31] word embeddings and a *bidirectional long short-term memory* (LSTM) model with a fully-connected layer of 8 hidden units. Regression is performed over the 8-dimensional features from this model. $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ follow an i.i.d. partition of the processed data and subsequently, $\mathbf{X}_{S_2}$ and $\mathbf{X}_{S_3}$ are replicated for 2 and 10 times, respectively. The relative valuations are in Fig. 6 where the darker shaded plots denote *without* replication. The noticeable increases in $\mathbf{X}_{S_3}$'s value from IGSV and VSV imply that they are not replication robust. Both VLSV and RVSV appear to be robust.

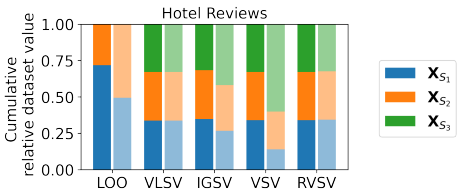

Figure 6: Effect of replication on valuation; darker/lighter shades denotes before/after replication.

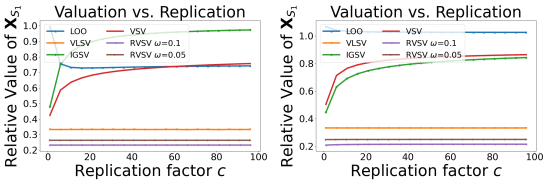

Figure 7: Relative value for the replicated $\mathbf{X}_{S_1}$ on CaliH (left) and FaceA (right). The vertical axis shows its value and horizontal axis shows the replication factor c.

Asymptotic valuations in non-i.i.d. settings. As our replication robustness includes $\sup_c$, we investigate large replication factors c (i.e, up to 100). As the previous experiment demonstrates that VLSV is robust, we use it as the comparison baseline. We additionally consider two non-i.i.d. distributions extended from the previous setting: *supersets* and *disjoint* on four separate real-world datasets: California housing price prediction (CaliH) [18], Kings county housing sales prediction (KingH) [13], US census income prediction (USCensus) [28], and age estimation from facial images (FaceA) [41]. We use 60% of data to construct $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ and the remaining 40% as the validation set for LOO and VLSV. For i.i.d. and supersets, we set $\mathbf{X}_{S_2} = \mathbf{X}_{S_1}$ such that $\mathbf{X}_{S_2}$ simulates an honest provider and we examine the relative effect of replicating $\mathbf{X}_{S_1}$. For supersets, we vary the proportion of data from $\mathbf{X}_{S_1}$ contained in $\mathbf{X}_{S_3}$. If the ratio is 0.1, then $\mathbf{X}_{S_3}$ contains 10% data from $\mathbf{X}_{S_1}$. If the ratio is 1, then $\mathbf{X}_{S_1} \subset \mathbf{X}_{S_3}$. For disjoint, we have three different datasets $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ and vary the degree of disjoint via a ratio: 0 (*resp.*, 1) means that the sampling spaces for $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ are completely disjoint (*resp.*, overlapped). In other words, with ratio 0, they do *not* contain any same data and are completely disjoint; with ratio 1, they may contain some same data. The results on i.i.d. for two datasets are in Fig. 7. For CaliH, we use the features from the last layer of a neural network with two fully connected layers of 64 and 10 hidden units and rectified linear units as the activation function. Additional details on distributions, datasets, and models are in Appendix B.6.

Next, we compare the relative valuations, as follows: compute the similarity (between the valuations by VLSV and those of another method) averaged over different replication factors c. We consider several such similarity measures including the Pearson correlation coefficient (r_{p}) [16], cosine similarity (cos), and the reciprocal of the l_2 norm of the difference [36]. For RVSV, we set $\omega = 0.05, 0.1$, denoted by RVSV-005 and RVSV-01. The results for CaliH are in Table 1 and other results are in Appendix B.6. VSV and IGSV are not robust and may be exploited as Fig. 7 shows both increase with replication relatively quickly for $c < 20$. Furthermore, our additional

experiments (Appendix B.7) on hyperparameter selection comparison show that IGSV is sensitive to the hyperparameter whereas RVSV is consistent, even with varying ω. RVSV is replication robust from Fig. 7 and the high similarity with VLSV in Table 1 shows that it achieves similar performance to VLSV *without* requiring validation.

Table 1: Similarity with VLSV under replication for CaliH. Values in bold indicate the best results.

Method	i.i.d.			disjoint 0			disjoint 1			supersets 0.1			supersets 1		
	r_{p}	cos	$1/l_2$	r_{p}	cos	$1/l_2$	r_{p}	cos	$1/l_2$	r_{p}	cos	$1/l_2$	r_{p}	cos	$1/l_2$
LOO	-0.991	0.730	1.894	-0.459	0.816	2.457	**-0.488**	0.406	0.770	-0.339	0.801	2.362	-0.590	0.771	2.100
IGSV	-0.903	0.637	1.591	0.640	0.639	1.583	-0.763	0.636	1.589	-0.893	0.636	1.580	-0.716	0.653	1.687
VSV	-0.886	0.787	2.493	0.644	0.784	2.415	-0.780	0.775	2.335	-0.892	0.779	2.389	-0.660	0.813	2.696
RVSV-005	**0.767**	**0.959**	**5.857**	**0.700**	**1.000**	**77.714**	-0.784	**0.998**	**28.479**	**0.810**	**0.983**	**9.314**	**0.918**	**0.946**	**5.051**
RVSV-01	**0.767**	0.920	4.055	0.351	0.999	47.066	-0.939	0.997	20.845	0.808	0.976	7.839	0.917	0.914	3.901

6 Related Works

Data valuation methods assign high values to data which lead to high performance [11, 16, 35, 39]. Existing methods such as leave-one-out approaches [16], the Shapley value-based methods [11, 15] and a reinforcement learning framework [39] require validation. Due to the tight coupling between valuation and validation, these approaches may face practical limitations of acquiring a good validation set [35]. The work of [35] has proposed an information-theoretic approach by valuing data based on the *information gain* (IG) on a Bayesian prior to avoid the need for validation. However, they have not theoretically shown a higher IG (value) leads to a better predictive performance. Our method has this theoretical property without needing validation. While existing methods demonstrate some effectiveness against replication using carefully selected validation sets [11, 16], our method achieves such guarantees without needing validation. The work of [12] has considered replication from a different perspective and is thus not directly comparable to our method.

7 Conclusion and Future Work

This paper proposes a replication robust data valuation method that requires no validation (i.e., model- and task-agnostic). In particular, we value data based on the inherent diversity formalized as the volume of the data matrix because we show that larger volumes correspond to better learning performance. We identify that volume is not robust to replication, so we propose a novel *robust volume* (RV) to provide replication robustness guarantees. In our experiments, we demonstrate that RV can be combined with the Shapley value and empirical comparison with baselines verifies its effectiveness in data valuation and its robustness guarantees. Importantly, we investigated various real-world datasets and adapted our proposed robust volume to machine learning models more complex than OLS (i.e., various neural networks) to demonstrate the practical applicability of robust volume. Current works on data pricing may build upon our perspective to reduce the dependence on validation and auxiliary datasets. For future work, we plan to consider more sophisticated replication techniques and investigate the balance between diversity representation vs. robustness.

Acknowledgments and Disclosure of Funding

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (Award No: AISG2-RP-2020-018). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Xinyi Xu is supported by the Institute for Infocomm Research of Agency for Science, Technology and Research (A*STAR). The authors thank Fusheng Liu for many interesting discussions.

References

[1] Aditya. 100,000 UK Used Car Dataset. URL `https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes`.

[2] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proc. ACM EC*, pages 701–726, 2019.

[3] Samyadeep Basu, Phil Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *Proc. ICLR*, 2021.

[4] Eric Bax. Computing a Data Dividend. arxiv:1905.01805v1, 2019.

[5] BM. Uber and Lyft Dataset Boston, MA. URL `https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma`.

[6] Michał Dereziński and Manfred K. Warmuth. Unbiased estimates for linear regression via volume sampling. In *Proc. NeurIPS*, pages 3087–3096, 2017.

[7] Michał Dereziński and Manfred K. Warmuth. Reverse iterative volume sampling for linear regression. *Journal of Machine Learning Research*, 19(23):1–39, 2018.

[8] Michal Dereziński, Manfred K. Warmuth, and Daniel Hsu. Leveraged volume sampling for linear regression. In *Proc. NeurIPS*, pages 2510–2519, 2018.

[9] P.J. Devereaux, Gordon Guyatt, Hertzel Gerstein, Stuart Connolly, and Salim Yusuf. Toward fairness in data sharing. *New England Journal of Medicine*, 375(5):405–407, 2016.

[10] Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67, 1991.

[11] Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proc. ICML*, pages 2242–2251, 2019.

[12] Dongge Han, Michael Wooldridge, Alex Rogers, Shruti Tople, Olga Ohrimenko, and Sebastian Tschiatschek. Replication-robust payoff-allocation for machine learning data markets. arXiv:2006.14583, 2021.

[13] Harlfoxem. House Sales in King County, USA. URL `https://www.kaggle.com/harlfoxem/housesalesprediction`.

[14] Cynthia A. Jackevicius, Jaejin An, Dennis T. Ko, Joseph S. Ross, Suveen Angraal, Joshua D. Wallach, Maria Koh, Jeeeun Song, and Harlan M. Krumholz. Submissions from the sprint data analysis challenge on clinical risk prediction: A cross-sectional evaluation. *BMJ Open*, 9(3), 2019.

[15] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. In *Proc. VLDB Endowment*, pages 1610–1623, 2019.

[16] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gurel, Bo Li, Ce Zhang, Dawn Song, and Costas Spanos. Towards Efficient Data Valuation Based on the Shapley Value. In *Proc. AISTATS*, pages 1167–1176, 2019.

[17] Sam Jossen. The world's most valuable resource is no longer oil, but data. *The Economist*, 2017.

[18] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

[19] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proc. ICML*, pages 1885–1894, 2017.

[20] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008.

[21] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, 2012.

[22] Larxel. Trip Advisor Hotel Reviews. URL https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews.

[23] D. J. Lizotte. *Practical Bayesian optimization*. PhD thesis, University of Alberta, Canada, 2008.

[24] Bernard Lo and David L. DeMets. Incentives for clinical trialists to share data. *New England Journal of Medicine*, 375(12):1112–1115, 2016.

[25] L. Lovász. *Submodular functions and convexity*, pages 235–257. Springer Berlin Heidelberg, 1983.

[26] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. In Q. Yang, L. Fan, and H. Yu, editors, *Federated Learning: Privacy and Incentive*, Lecture Notes in Computer Science, pages 189–204. Springer International Publishing, Cham, 2020.

[27] A. Mikhalev and I. V. Oseledets. Rectangular maximum-volume submatrices and their applications. *Linear Algebra and its Applications*, 538:187–211, 2018.

[28] MuonNeutrino. US Census Demographic Data. URL https://www.kaggle.com/muonneutrino/us-census-demographic-data.

[29] Cem Orhan and Öznur Taştan. Alevs: Active learning by statistical leverage sampling. In *Proc. ICML Active Learning Workshop*, 2015.

[30] Jian Pei. Data pricing – from economics to data science. In *Proc. KDD*, 2020.

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proc. EMNLP*, pages 1532–1543, 2014.

[32] Ramesh Raskar, Praneeth Vepakomma, Tristan Swedish, and Aalekh Sharan. Data markets to support AI for all: Pricing, valuation and governance. arXiv:1905.06462, 2019.

[33] L. S. Shapley. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, 2016.

[34] M. C. Shewry and H. P. Wynn. Maximum entropy sampling. *Journal of Applied Statistics*, 14(2):165–170, 1987.

[35] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In *Proc. ICML*, pages 8927–8936, 2020.

[36] Tianshu Song, Yongxin Tong, and Shuyue Wei. Profit allocation for federated learning. In *Proc. IEEE Big Data*, pages 2577–2586, 2019.

[37] Machine Learning Group ULB. Credit Card Fraud Detection. URL https://www.kaggle.com/mlg-ulb/creditcardfraud.

[38] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. In Q. Yang, L. Fan, and H. Yu, editors, *Federated Learning: Privacy and Incentive*, Lecture Notes in Computer Science, pages 153–167. Springer International Publishing, Cham, 2020.

[39] Jinsung Yoon, Sercan O. Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *Proc. ICML*, pages 10842–10851, 2020.

[40] Han Yu, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. A fairness-aware incentive scheme for federated learning. In *Proc. AIES*, pages 393–399, 2020.

[41] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proc. CVPR*, pages 4352–4360, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We clearly describe the problem of data valuation and give an overview of our proposed method and what it achieves - a diversity-based data valuation method without validation and with robustness guarantees to replication. We summarize our contributions in point forms in the introduction section.
   (b) Did you describe the limitations of your work? [Yes] See Sec. 3, we show the theoretical guarantees require complicated assumptions to generalize to high-dimensional feature spaces, but demonstrate in Sec. 5 that our method works well empirically. See the last part of Sec. 4.2 that we restrict our consideration to data that follow a normal distribution and do not contain outliers.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All assumptions are clearly stated.
   (b) Did you include complete proofs of all theoretical results? [Yes] All theoretical results are proven. Complete proofs are given in the Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We submit our code as supplementary materials. Instructions on getting the datasets, processing the datasets and running the code are given.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Experiment settings including datasets and models are described in Sec. 5 with additional details in Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Fig. 2 and additional figures in Appendix B.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Sec. 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Our work uses existing datasets. We cite creators for all datasets clearly. URLs are also provided.
   (b) Did you mention the license of the assets? [Yes] See Appendix B.1.
   (c) Did you include any new assets either in the supplemental material or as a URL?[N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] No crowdsourcing or human subjects were involved.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No crowdsourcing or human subjects were involved.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] No crowdsourcing or human subjects were involved.