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 real-world use cases such as collaborative machine learning, federated learning, trusted data sharing, data marketplaces. The value of data is often associated with the learning performance (e.g., validation accuracy) of a model trained on the data, which introduces a close coupling between data valuation and validation. However, 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. Another practical issue is that of data replication: Given the value of some data points, a dishonest data provider may replicate these data points to exploit the valuation for a larger reward/payment. We observe that the diversity of the data points is an inherent property of a dataset that is independent of validation. We formalize diversity via the *volume* of the data matrix (i.e., determinant of its left Gram), which allows us to establish a formal connection between the diversity of data and learning performance without requiring validation. Furthermore, we propose a *robust volume* measure with a theoretical guarantee on the replication robustness by following the intuition that copying the same data points does not increase the diversity of data. We perform extensive experiments to demonstrate its consistency in valuation and practical advantages over existing baselines and show that our method is model- and task-agnostic and can be flexibly adapted to handle various neural networks.

1 Introduction

Data is increasingly recognized as a valuable resource [19], so we need a principled measure of its worth. A suitable data valuation has wide-ranging applications such as fairly compensating clinical trial researchers for their collected data [12, 16, 25], fostering collaborative machine learning and federated learning among industrial organizations [35, 36, 39], encouraging trusted data sharing and building data marketplaces [7, 30, 32, 37], among others.

A popular viewpoint is that the value of data should correlate with the learning performance of a model trained on the data [14, 18], which enforces a close coupling between data valuation and validation. However, a validation set may not always be available in practice [35]. Also, as different choices of the validation set can lead to different data valuations, it is challenging for the data providers to agree on the choice of such a validation set [35]. Since valuation is coupled with validation, if the

[*]Equal contribution.

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

validation set is not sufficiently representative of the distribution of test queries in a learning task, the resulting valuation may not be as accurate/useful [40]. We adopt a different perspective: The value of data should be related to its intrinsic properties and valuation can be decoupled from validation by considering the inherent diversity of the data. Intuitively, a more diverse collection of data points corresponds to a higher-quality dataset and thus yields a larger value. This perspective circumvents the above practical limitations and allows our valuation method to be model- and task-agnostic. We formalize diversity via the *volume* of the data matrix (i.e., determinant of its left Gram).

Data replication is another practical issue in data valuation due to the digital nature and anonymous setting of data marketplaces [15]. Supposing a dataset has some value and a data provider instead offers one containing two copies of every data point in this dataset, 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 by constructing a compressed version of the original data to assign little value to replicated data and still preserve its inherent diversity, hence guaranteeing replication robustness.

We provide theoretical justifications for formalizing diversity via volume: Firstly, diversity should be non-negative and monotonic [14, 18, 35, 38] and volume satisfies both properties. Secondly, a greater diversity should lead to a better learning performance [23]: We formally show that a larger volume generally leads to a better performance using the *ordinary least squares* (OLS) framework and our method can be flexibly adapted to handle more complex machine learning models (i.e., various neural networks) in our experiments. Specifically, data with a larger volume can lead to a more accurate pseudo-inverse (i.e., a key component of the least squares solution) and a smaller mean squared error.

To ensure 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 the notion of replication robustness via the asymptotic value attainable through replication. Unfortunately, the conventional definition of volume does not have this property. So, we propose a *robust volume* (RV) measure by constructing a compressed version of the original data that groups similar data via discretized cubes of the input feature space and represents those in each cube via a statistic. The RV measure offers practitioners the flexibility to trade off between diversity representation and replication robustness via the cube's width. 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.

The specific contributions of our work here include:

- Formalizing a measure of data diversity via the volume of data (Sec. 2) and justifying the suitability of volume for data valuation both theoretically (Sec. 3) and empirically (Sec. 5);
- Formalizing the notion of replication robustness and designing a data valuation method based on the *robust volume* (RV) measure with a theoretical guarantee on replication robustness (Sec. 4);
- Performing extensive empirical comparisons with baselines to demonstrate that our method is consistent in valuation without validation, replication robust, and can be flexibly adapted to handle complex machine learning models such as various neural networks (Sec. 5).

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 input feature vectors, respectively. Let $\mathbf{P}_S \coloneqq [\mathbf{X}_S^\top \ \mathbf{0}]^\top \in \mathbb{R}^{n\times d}$ be the zero-padded version of $\mathbf{X}_S \in \mathbb{R}^{s\times d}$. We concatenate the data submatrices 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 $n = s + s'$. We denote the corresponding observed labels/responses 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 $\mathbf{w}_S \coloneqq \mathbf{X}_S^+\mathbf{y}_S$. To ease notations, let $V \coloneqq \mathrm{Vol}(\mathbf{X})$ and $V_S \coloneqq \mathrm{Vol}(\mathbf{X}_S)$ where $\mathrm{Vol}()$ is defined below. Let $|\mathbf{A}|$ denote the determinant of a square matrix $\mathbf{A}$. The left Gram matrix of $\mathbf{X}$ is $\mathbf{G} \coloneqq \mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{d\times d}$, so for data submatrix $\mathbf{X}_S$, $\mathbf{G}_S \coloneqq \mathbf{X}_S^\top\mathbf{X}_S \in \mathbb{R}^{d\times d}$.

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 the above definition of volume for several reasons: (a) Often, the input feature space of the data is pre-determined and fixed due to the data collection process. But, new data can stream in and

so, n can grow indefinitely while d remains fixed [9, 10]. (b) By leveraging the formal connection between volume and learning performance (Sec. 3), we can design a validation free volume-based data valuation to assign a larger value to data leading to a better learning performance. (c) This affords 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 [9, 10]. In practice, we perform pre-processing such as principal component analysis to reduce the dimension of the input feature space 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 using other features. For instance, if a dataset already contains monthly salaries, then an annual salary would be redundant.

3 Larger Volume Entails Better Learning Performance

The value of a data (sub)matrix depends on the learning performance trained on it [14, 18] which, we will show, depends on its volume. Simply put, the larger the volume, the better the learning performance. In this section, we will formalize this claim through the *ordinary least squares* (OLS) framework. In particular, we will investigate two metrics for learning performance: (a) the quality of the pseudo-inverse represented by $\text{bias}_S \coloneqq \left\|\mathbf{P}_S^+ - \mathbf{X}^+\right\|$ because estimating $\mathbf{X}^+$ accurately is important to achieving small *mean squared error* (MSE) [9] and where $\mathbf{P}_S^+ \coloneqq (\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{P}_S^\top$, and (b) the MSE denoted as $L(\mathbf{w}_S) \coloneqq \left\|\mathbf{y} - \mathbf{X}\mathbf{w}_S\right\|^2$.

3.1 Larger Volume Entails Smaller Bias

In regression problems, the closed-form optimal solution is constructed via $\mathbf{X}^+$ computed using $\mathbf{X}$. So, the bias of $\mathbf{P}_S^+$ from $\mathbf{X}^+$ indirectly determines the value of $\mathbf{X}_S$ [9], i.e., a smaller bias means a larger value. We show in Proposition 1 below that 'a larger volume means a smaller bias' always holds for $d = 1$. 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 $\mathbf{X} \in \mathbb{R}^{n\times 1}$, $V_S \geq V_{S'} \iff \text{bias}_S - \text{bias}_{S'} \leq 0$.*

The above result can be generalized to $M > 2$ non-zero data submatrices: Let $\mathbf{X} \coloneqq [\mathbf{X}_{S_1}^\top \ \mathbf{X}_{S_2}^\top \ \cdots \ \mathbf{X}_{S_M}^\top]^\top$ and w.l.o.g., suppose that $V_{S_1} \geq V_{S_2} \geq \ldots \geq V_{S_M}$. Then, $\text{bias}_{S_1} \leq \text{bias}_{S_2} \leq \ldots \leq \text{bias}_{S_M}$. For $d > 1$, counterexamples exist (see Fig. 1), so we instead compare bias_S and $\text{bias}_{S'}$ in the next result:

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

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

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. $\mathbf{G}_S, \mathbf{G}_{S'}$, $\mathbf{P}_S$ and $\mathbf{P}_{S'}$ are, respectively, zero-padded versions of $\mathbf{X}_S$ and $\mathbf{X}_{S'}$, 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 (Appendix A.1) relies on a key observation that for $d = 1$, the left Gram matrix is a number and the rest of the proof follows. However, it cannot be generalized to that for $d > 1$, so we resort to a different proof technique. The proof of Proposition 2 requires Lemma 2 in Appendix A.1 which establishes a formal connection between volume and $\mathbf{G}^{-1}$ using the 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) $\left\|\mathbf{Q}_S\mathbf{X}_S^\top\right\| \approx \left\|\mathbf{Q}_{S'}\mathbf{X}_{S'}^\top\right\|$ and $V \gg \max(V_S, V_{S'})$. Case A is intuitive: $V_S \gg V_{S'}$ means $\mathbf{X}_S$ is much "larger" in volume 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 data distribution). The intuition is that the first difference term will be relatively large in magnitude (so, its sign will dominate the overall expression), while the second inner product term will be relatively small in magnitude. This is because the first difference term involves $1/V_S^4$ and $1/V_{S'}^4$ but the second inner

product term involves $1/(V^2 \times V_S^2)$ and $1/(V^2 \times V_{S'}^2)$, and we show $V \gg \max(V_S, V_{S'})$ (Lemma 3 in Appendix A.1). Subsequently, $\left\|\mathbf{Q}_S \mathbf{X}_S^\top\right\| \approx \left\|\mathbf{Q}_{S'} \mathbf{X}_{S'}^\top\right\|$ and $V_S \geq V_{S'}$ suggest that the first difference term (and thus the overall expression) is likely negative. We empirically verify in Fig. 1 that $V_S \geq V_{S'} \implies \text{bias}_S \leq \text{bias}_{S'}$ holds for more than 80% of times.

3.2 Larger Volume Entails Smaller MSE

In Proposition 3 (see proof in Appendix A.2) below, we will show a similar result (to Proposition 1) theoretically analyzing the connection between volume and MSE when $d = 1$, which may be surprising since $\text{Vol}()$ (Definition 1) does not consider $\mathbf{y}$ at all and can yet determine which data submatrix offers better predictions on the rest of the (unobserved) data. Unfortunately, such a result does not directly generalize to $d > 1$ or beyond two submatrices. Nevertheless, we will analyze the effect of volume on the learning performance (i.e., MSE) in general.

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, the above result 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 have derived in Appendix A.2 that

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

and also shown in Appendix A.2 that since $L(\mathbf{w}_S) - L(\mathbf{w}_{S'})$ explicitly depends on $\mathbf{y}$ (1) and $\text{Vol}()$ does not include $\mathbf{y}$ at all, it is possible to adversarially construct $\mathbf{y}$ s.t. $L(\mathbf{w}_S) - L(\mathbf{w}_{S'}) > 0$ or $L(\mathbf{w}_S) - L(\mathbf{w}_{S'}) < 0$ for some fixed $\mathbf{X}_S, \mathbf{X}_{S'}$.

The adversarial cases notwithstanding, volume is regarded as a good surrogate measure of the quality of data applied to active learning and matrix subsampling with theoretical performance guarantees [11, 28]. Similarly, we can adopt the perspective that $\text{Vol}()$ is a measure of the diversity in the input features [23], which provides an intuitive interpretation for Proposition 3: A more diverse dataset with a larger volume gives a better learning performance (i.e., smaller MSE). We will show in Sec. 5.2 that not requiring labels/responses can be an advantage in practice if the labels/responses are noisy/corrupted or there is a distributional difference between the validation and test sets.

We conclude Sec. 3 by empirically verifying whether the additional assumptions described in the last paragraph of Sec. 3.1 are satisfied by checking the percentage of times that $V_S \geq V_{S'} \implies \text{bias}_S - \text{bias}_{S'} \leq 0$ holds. To elaborate, we randomly and identically sample equal-sized $\mathbf{X}_S, \mathbf{X}_{S'}$ over 500 independent trials and compute the percentage of times that a larger volume leads to better learning performance (vertical axis) against the size of $\mathbf{X}_S, \mathbf{X}_{S'}$ (horizontal axis). We consider sampling $\mathbf{X}_S, \mathbf{X}_{S'}$ from either a uniform or normal distribution of varying dimensions: In Fig. 1, for example, '$\mathcal{N}\ d = 1$' denotes $\mathbf{X}_S, \mathbf{X}_{S'}$ being sampled from 1-dimensional standard normal distribution. For MSE, the response y of 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 $U(0, 2)^d$. Fig. 1 (left) shows that a larger volume leads to a smaller bias for more than 80% of times, thus verifying that the additional assumptions in Sec. 3.1 are satisfied. Fig. 1 (right) shows that a larger volume leads to a smaller MSE for more than 50% of times for $d \leq 10$, which is consistent with the above implications from (1).

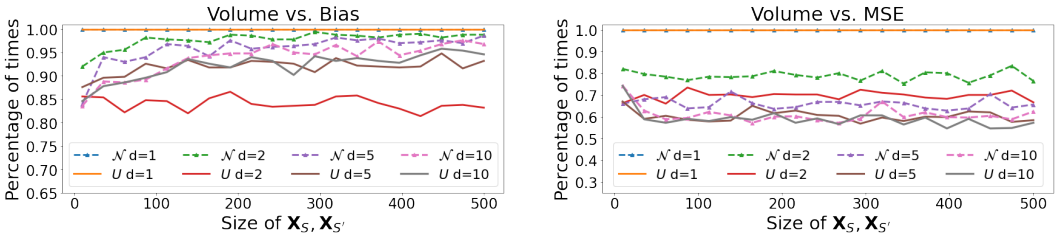

Figure 1: Volume vs. bias (left) and volume vs. MSE (right) for both identically sampled, equal-sized datasets $\mathbf{X}_S, \mathbf{X}_{S'}$ from either a uniform $U(0, 1)^d$ or normal $\mathcal{N}(0, 1)^d$ distribution. The vertical axis shows the percentage of times over 500 independent trials that the dataset with a larger volume leads to a better learning performance (i.e., smaller bias or MSE).

4 Robustifying Volume-based Data Valuation

As a larger volume can entail a better learning performance (Sec. 3), we consider a volume-based data valuation method. Unfortunately, volume (Definition 1) is *not* robust to replication via direct data copying. Hence, we will introduce a modified volume measure that can trade off a more refined representation of diversity for greater robustness to replication.

4.1 First Attempt of Volume-based Data Valuation

Directly using $\mathrm{Vol}(\mathbf{X})$ as a valuation of $\mathbf{X}$ satisfies both non-negativity and monotonicity which follow directly from Definition 1 and the matrix determinant lemma, respectively:

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* $\mathbf{x} \in \mathbb{R}^{1\times d}$ *is a new data point.*

The properties of $\mathrm{Vol}()$ in Proposition 4 imply that a bigger-sized $\mathbf{X}$ (i.e., more data) should yield a larger value [14, 18, 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 data by replicating the data infinitely, as shown in the following result (see proof in Appendix A.3):

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 replicated for* $m \geq 1$ *times and* $\mathbf{X}_{\text{rep}} \coloneqq [\mathbf{X}^\top\ \mathbf{x}_q^\top\ \ldots\ \mathbf{x}_q^\top]^\top \in \mathbb{R}^{(n+m)\times d}$. *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. We define a measure of *inflation* as the ratio $\nu(\text{replicate}(\mathbf{X}, c))/\nu(\mathbf{X})$ where $\nu()$ is a data valuation function (e.g., $\mathrm{Vol}()$) mapping a data matrix to a real value, the function $\text{replicate}(\mathbf{X}; c)$ directly copies the data in $\mathbf{X}$ and appends them back to $\mathbf{X}$ to output $\mathbf{X}_{\text{rep}} \in \mathbb{R}^{(nc)\times d}$, and the *replication factor* c denotes the amount of replication. One way of replication is to copy the entire $\mathbf{X}$ for c times. Another way is to copy some data submatrix for a certain number of times s.t. $\mathbf{X}_{\text{rep}} \in \mathbb{R}^{(nc)\times d}$. We consider the second way because replicating different data increases the value differently (Lemma 1). We define below a measure of replication robustness to formalize the intuition that greater robustness should guarantee smaller inflation:

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

The theoretically optimal robustness is $\gamma_\nu = 1$, which implies no additional gain from replicating data, hence discouraging replication completely. In contrast, the worst-case robustness is $\gamma_\nu = 0$, which is the case for any $\nu()$ that strictly monotonically increases with replication and, in particular, $\gamma_{\mathrm{Vol}} = 0$ by applying Lemma 1. As a result, a replication robust data valuation function must have a diminishing marginal value from replication: The additional gain from having more copies of the same data converges asymptotically to 0 w.r.t. c. This aligns with what we observe in practice: Repeatedly adding the same data to a training set does not improve the learning performance indefinitely.

4.2 Replication Robust Volume (RV)-Based Data Valuation

We will propose an RV measure by constructing a compressed version of original data matrix $\mathbf{X}$ that groups similar data points via discretized cubes of the input feature space and represents those in each cube via a statistic. The RV measure offers practitioners the flexibility to trade off a more refined diversity representation for greater replication robustness by increasing the cube's width.

Definition 3 (**Replication Robust Volume (RV)).** *Let the* d*-dimensional input feature space/domain for* $\mathbf{X}$ *be discretized into a set* Ψ *of* d*-cubes of width/discretization coefficient* ω, ϕ_i *denote the number of data points in* d*-cube* $i \in \Psi$, $\boldsymbol{\mu}_i \in \mathbb{R}^{1\times d}$ *be a statistic (e.g., mean vector) of the data points in* d*-cube* i, *and* $\widetilde{\mathbf{X}} \coloneqq [\boldsymbol{\mu}_i^\top]_{i\in\Psi:\phi_i\neq 0}^\top$ *be a compressed version of* $\mathbf{X}$ *s.t. each row of* $\widetilde{\mathbf{X}}$ *is a statistic* $\boldsymbol{\mu}_i$ *of the data points in non-empty* d*-cube* i. *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 $\rho_i \coloneqq \sum_{p=0}^{\phi_i} \alpha^p$ *with hyperparameter* $\alpha \in [0, 1]$ *controlling the degree of robustness.*

In contrast to 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 the 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 any robustness guarantee, we will first show in Proposition 5 (see proof in Appendix A.3) below that RV (Definition 3) preserves the original volume in a relative sense, i.e., the ratio $V_S/V_{S'}$ is preserved. The implication is a similar effect of RV on the learning performance (Sec. 3), as empirically demonstrated in Sec. 5.1.

Proposition 5 (Bounded Distortion of $\mathrm{RV}(\mathbf{X}_S;\omega)/\,\mathrm{RV}(\mathbf{X}_{S'};\omega)$). *Define distortion $\delta(\omega) := [\mathrm{RV}(\mathbf{X}_S;\omega)/\,\mathrm{RV}(\mathbf{X}_{S'};\omega)]/[\mathrm{Vol}(\mathbf{X}_S)/\,\mathrm{Vol}(\mathbf{X}_{S'})]$. Then, $(\exp(\beta^{-1}))^{-1} \le \delta(\omega) \le \exp(\beta^{-1})$ for any $\omega > 0$ where $\beta = 1/(\alpha n)$. For example, $\beta = 10$ bounds $\delta(\omega) \in [0.905, 1.105]$ approximately.*

Near-optimal robustness by upper-bounding inflation. We have previously defined 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 when the data in d-cube i is replicated 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}). *For $\alpha \in [0,1)$, $\gamma_{\mathrm{RV}} \ge (1-\alpha)^{|\Psi|}$ where, with a slight abuse of notation, Ψ denotes the set of non-empty d-cubes. For $\alpha = 1, \gamma_{\mathrm{RV}} = 0$.*

Its proof is in Appendix A.3. Recall from Definition 2 that $\gamma_{\mathrm{RV}} = 1$ is optimal robustness. From Proposition 6, reducing α achieves a smaller upper bound on inflation and greater 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 case, 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 should be set to achieve a certain upper bound on inflation but should not be 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 $\exp(\beta^{-1})$ on the inflation, as proven in Lemma 5 in Appendix A.3. For instance, setting $\beta = 10$ and $\alpha = 1/(\beta n)$ guarantees $\mathrm{RV}(\mathrm{replicate}(\mathbf{X}, c);\omega) \le 110\% \times \mathrm{RV}(\mathbf{X};\omega)$. However, it requires us to know the true n *without* any replication. In practice, as we can only observe the data with replication (if any) [15], we estimate n with the number $|\Psi|$ of rows in $\widetilde{\mathbf{X}}$.

Trading off diversity representation for replication robustness 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 representing them via a statistic. On the other hand, a larger ω means a less refined diversity representation but greater replication robustness. In the extreme case, a sufficiently large ω results in grouping all data points together and representing them all using a single statistic, hence foregoing the diversity in data. So, a practitioner should determine the trade-off between diversity representation vs. replication robustness based on the requirements of the real-world use case. The following result (see proof in Appendix A.3) formalizes both extremes of the trade-off:

Proposition 7 (Reduction to $\mathrm{Vol}()$ vs. Optimal Robustness). *Set ω to be s.t. each d-cube only contains completely identical data points, and*

1. *set ρ_i to some constant $K_{\widetilde{\mathbf{X}},i}$ for $i \in \Psi$ based on a recursive application of Lemma 1. Then, $\mathrm{RV}(\cdot;\omega) = \mathrm{Vol}()$;*
2. *set $\alpha = 0$. So, $\rho_i = \mathbb{1}(\phi_i \ne 0)$ 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 [15]). Intuitively, $\mathrm{RV}_{\mathbb{1}}(\cdot;\omega)$ is not robust to noisy replication as the replicated data are perturbed. Our preliminary empirical study in Appendices B.2 and B.3 shows that RV is robust to noisy replication if the noise magnitude is small relative to ω. So, a future work is to devise a way to optimize the trade-off between diversity representation and replication robustness via ω. In our work here, we empirically find $\omega = 0.1$ suitable for the case of standardized input features.

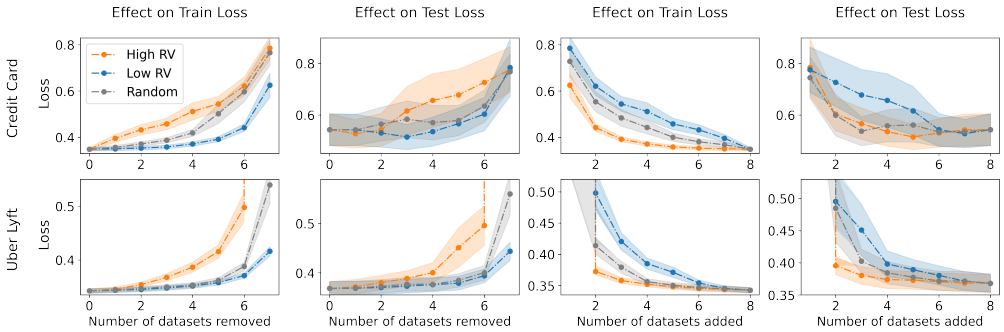

Figure 2: Effect of removing/adding dataset with highest/lowest RV on train/test loss for real-world credit card and Uber Lyft datasets. Plots show the average and standard errors over 50 random trials.

In using standardized input features, we implicitly assume that the input features follow a normal distribution. This makes the data further away from the mean (i.e., statistically rarer) more valuable in learning [11]. We also observe this in Sec. 5.2 where data closer to the mean are valued to be smaller across all baselines and our method. Our work here excludes considerations of outliers as they are not truly representative of the true data distribution.

5 Experiments and Discussion

In this section, we will first verify our claim in Sec. 3 that a larger volume leads to a better learning performance and reveal some interesting practical perspectives in Sec. 5.1. Then, in Sec. 5.2, we will show that RV produces results consistent with existing baseline methods and also demonstrate the limitations of these baselines. In particular, RV is model- and task-agnostic while another baseline with an explicit dependence on the validation set is shown to have some deviation in data valuation as the validation set changes. Lastly, in Sec. 5.3, we will verify our robustness guarantee by analyzing its asymptotic behavior under replication. Importantly, our empirical study has gone beyond the OLS framework used for the theoretical analysis in Sec. 3 as our method can be flexibly adapted to handle various neural network architectures on different machine learning tasks including both image classification and natural language processing. All experiments have been run on a server with Intel(R) Xeon(R)@ 2.70GHz processor and 256GB RAM. Our code is publicly available at: `https://github.com/ZhaoxuanWu/VolumeBased-DataValuation`.

5.1 Effect of Robust Volume (RV) on Learning Performance

In this subsection, we use RV and volume interchangeably as replication is not considered here and Proposition 5 guarantees that RV preserves the original volume. We consider the setting of sequentially adding/removing the dataset with highest/lowest RV to analyze the effect of RV on the learning performance [14]. We include random selection as a baseline. We simulate 8 data providers to make the results more generalizable. In this experiment, we use two real-world datasets: credit card fraud detection [2] (i.e., transaction amount prediction) and Uber & Lyft [5] (i.e., carpool ride price prediction) which are pre-processed to contain 8 and 12 standardized input features, respectively. Fig. 2 shows the results. Additional results on two other real-world datasets are in Appendix B.4.

It can be observed that adding (resp., removing) a dataset with a larger RV leads to a smaller (resp., larger) train loss, thus verifying Proposition 2 that a larger volume leads to a more accurate pseudo-inverse and smaller train loss in terms of mean squared error. This observation is also consistent with the results on the test loss, albeit with larger standard errors. This confirms (1) that in a higher dimensional input feature space, a larger volume does not immediately guarantee a smaller test loss.

Interesting practical perspectives. The results on adding datasets provide justification for a data buyer with a limited budget to spend on datasets with larger RVs first to achieve the best learning performance, thereby resonating with the active learning paradigm [29]. On the other hand, the results on removing datasets sheds light on the following question: If training on all collected datasets is too costly due to memory or time constraints, then which dataset should be removed first without compromising the learning performance much (i.e., the dataset with smallest RV)?

5.2 Empirical Comparison of Robust Volume (RV) Shapley Value with Baselines

We will demonstrate that RV without validation gives results consistent with existing baseline methods which may require validation. Then, we will empirically show the limitations of these baselines.

To design principled, fair payments to the data providers, we use (robust) volume as the characteristic function in the commonly used Shapley value to measure the expected marginal contributions of their datasets [14, 18, 35]. Our *robust volume Shapley value* (RVSV) is defined as follows [33]:

$$\text{RVSV}_m \coloneqq (1/M!) \sum_{\mathcal{C} \subseteq \mathcal{M} \setminus \{S_m\}} [|\mathcal{C}|! \times (M - |\mathcal{C}| - 1)!] \times [\text{RV}(\mathbf{X}_{\mathcal{C} \cup \{S_m\}}; \omega) - \text{RV}(\mathbf{X}_{\mathcal{C}}; \omega)] \quad (3)$$

where $\mathcal{M} \coloneqq \{S_1, \ldots, S_M\}$ denotes a set of M data providers/datasets and $\mathbf{X}_{\mathcal{C}}$ denotes a data matrix constructed from concatenating the data matrix $\mathbf{X}_{S_{m'}}$ of every data provider $S_{m'} \in \mathcal{C} \subseteq \mathcal{M}$. Our *volume Shapley value* (VSV) is computed by replacing $\text{RV}(\cdot; \omega)$ in (3) with $\text{Vol}()$. We compare VSV and RVSV with the following baselines: validation loss *leave-one-out* (LOO) value [21, 27], *validation loss Shapley value* (VLSV) [14, 18], and *information gain Shapley value* (IGSV) [35]. We consider the contributions of $M = 3$ data providers/matrices/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 from baseline distributions. We first consider simpler experimental settings on synthetic data drawn from the 6D Hartmann function [24] defined over $[0, 1]^6$ with four baseline data distributions for $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$: (A) *independent and identical distribution* (i.i.d.) where $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ contain 200 i.i.d. samples each; (B) ascending dataset 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 input 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 the respective 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 in Fig. 3 show that both VSV and RVSV are generally consistent with IGSV. For (B) ascending dataset size, VSV, RVSV, and IGSV increase from $\mathbf{X}_{S_1}$ to $\mathbf{X}_{S_3}$, while VLSV surprisingly values the contributions of $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ to be nearly equal; the latter may be due to VLSV's sensitivity to the definition of the value $\nu(\emptyset)$ of an empty dataset/matrix $\emptyset$ when calculating the Shapley value. Fig. 4 illustrates that for i.i.d., VLSV is sensitive to the definition of $\nu(\emptyset)$: For example, setting $\nu(\emptyset)$ to 0 [18], 1.06 (by initializing parameters to zeros), and 8.75 (by initializing parameters randomly using $\mathcal{N}(0, 1)$ [14]) yield different VLSVs of 0.346, 0.183, and 0.330 for $\mathbf{X}_{S_1}$, respectively. These conflicting choices of $\nu(\emptyset)$ add to the difficulties of applying VLSV in practice.

Interestingly, under (C) disjoint input domains, all methods unanimously value the contribution of $\mathbf{X}_{S_2}$ to be the lowest despite their input domains to be of the same size, which is due to the standardization of the input features and so offers the following interpretation: The data in the "center" is the most common if we assume the true data distribution follows a normal one. 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 with this distribution are reported in Appendix B.5. It is counter-intuitive to see that for i.i.d., LOO values the contribution of $\mathbf{X}_{S_1}$ to be 0, which may be due to instability from the calculation of their contributions [8].

Real-world datasets with different preferences of validation sets. We use two real-world datasets: UK used car dataset [1] (i.e., car price prediction) and credit card fraud detection dataset [2] (i.e., transaction amount prediction) where there are different preferences of validation sets [35]. For instance, car dealers for different manufacturers such as Audi, Ford, and Toyota may have different preferences over data. So, we construct two different validation sets comprising cars from different manufacturers. Similarly, different financial institutions may differ in their interests of the transaction amounts. For example, smaller banks typically manage and focus on smaller transaction amounts, so we construct two different validation sets comprising large (i.e., $> \$1000$) vs. small transaction amounts. The results in Fig. 5 show that the effect of different preferences of validation sets on LOO is pronounced, as expected. The effect on VLSV is less due to the averaging of marginal contributions. On the other hand, there is no effect on IGSV, VSV, and RVSV as they do not require a validation set.

5.3 Replication Robustness

We first perform a simpler experiment to demonstrate the effect of replication and then perform more extensive experiments under more complex settings to show the asymptotic behavior of RVSV and existing baseline methods under replication.

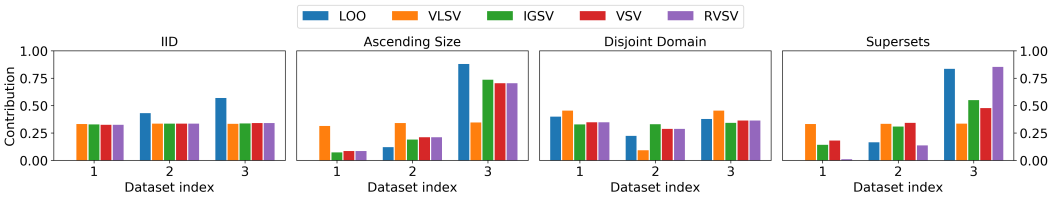

Figure 3: Contributions of $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ from Hartmann function with baseline data distributions: (A) i.i.d., (B) ascending dataset size, (C) disjoint input domains, and (D) supersets.

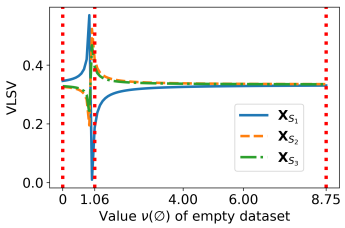

Figure 4: Sensitivity of VLSV to varying $\nu(\emptyset)$ (e.g., 3 red dotted lines).

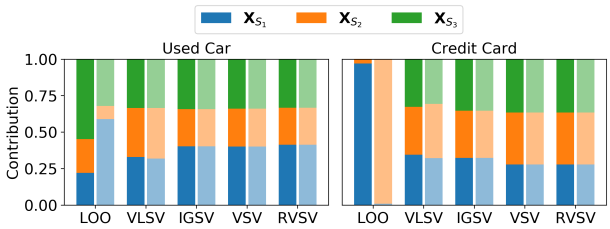

Figure 5: Contributions of $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ for 2 validation sets distinguished by darker vs. lighter shades.

Contributions of $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ under i.i.d. setting. We perform this experiment on the Trip Advisor hotel reviews dataset [4] (i.e., numerical rating prediction) which contains text reviews data. We utilize the GloVe [31] word embeddings and a bidirectional long short-term memory model with a fully-connected layer of 8 hidden units. Regression is performed over the 8-dimensional latent features from this model. Data matrices $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, 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 results in Fig. 6 show noticeable increases in the contribution of $\mathbf{X}_{S_3}$ for IGSV and VSV, which implies that they are not replication robust. On the other hand, both VLSV and RVSV appear robust.

Contributions of $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ under non-i.i.d. settings. As our replication robustness includes $\sup_c$ (Definition 2), we investigate large replication factors c of up to 100. Since the previous experiment shows that VLSV is robust, we use it as the baseline for comparison. We additionally consider two non-i.i.d. data distributions extended from the previous setting: *supersets* and *disjoint input domains* for 4 real-world datasets: California housing price prediction (CaliH) [20], Kings county housing sales prediction (KingH) [3], US census income prediction (USCensus) [6], 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 $\mathbf{X}_{S_2} = \mathbf{X}_{S_1}$ s.t. $\mathbf{X}_{S_2}$ simulates an honest data provider and we examine the effect of replicating $\mathbf{X}_{S_1}$. For supersets, we vary the proportion of data from $\mathbf{X}_{S_1}$ that is 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 input domains, we vary how disjoint they are for $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ via a ratio: 0 (resp., 1) means that $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, and $\mathbf{X}_{S_3}$ have completely disjoint (resp., overlapped) input domains. In other words, with ratio 0, they do not contain any similar data, while with ratio 1, they may contain some similar data. Fig. 7 shows results for two datasets with i.i.d. data distribution. For CaliH, we use the latent features from the last layer of a neural network with 2 fully connected layers of 64 and 10 hidden units and

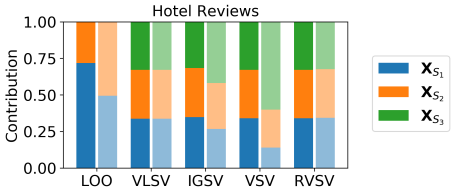

Figure 6: Effect of replication on contributions of $\mathbf{X}_{S_1}$, $\mathbf{X}_{S_2}$, $\mathbf{X}_{S_3}$. Darker (lighter) shade denotes before (after) replication.

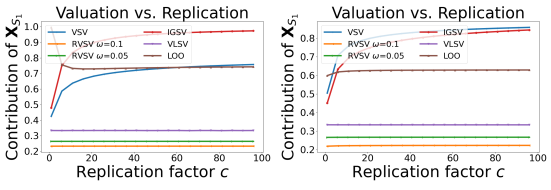

Figure 7: Contribution of the replicated $\mathbf{X}_{S_1}$ with varying replication factors c for CaliH (left) and FaceA (right) datasets.

the *rectified linear unit* (ReLU) as the activation function. Additional details on data distributions, datasets, and models are in Appendix B.6.

Next, we compare the similarity of RVSV and baseline methods to VLSV using similarity measures such as the Pearson correlation coefficient (r_{p}) [18], cosine similarity (cos), and the reciprocal of the l_2 norm of the difference [36]. For RVSV, we set $\omega = 0.05$ and 0.1, which are respectively denoted by RVSV-005 and RVSV-01. Table 1 shows results averaged over varying replication factors c for CaliH; the other results are reported in Appendix B.6. VSV and IGSV are not robust and may be exploited as both increase relatively quickly with replication for $c < 20$ (Fig. 7). Furthermore, our additional experiments on varying hyperparameter choices (Appendix B.7) show that IGSV is sensitive to the choice of hyperparameter whereas RVSV is consistent, even with varying ω. From Fig. 7, RVSV is replication robust. RVSV can also achieve a high degree of similarity to VLSV without requiring validation, as seen in Table 1.

Table 1: Effect of replication on similarity of RVSV and existing baseline methods to VLSV for CaliH dataset. 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 Work

Data valuation methods assign a larger value to data that leads to a better learning performance [14, 18, 35, 40]. Existing methods such as leave-one-out approaches [18], the Shapley value-based methods [14, 17], and a reinforcement learning framework [40] require validation. Due to the tight coupling between valuation and validation, these methods may face practical limitations arising from using a validation set (Sec. 1). The work of [35] has proposed an information-theoretic approach to valuing data based on the *information gain* (IG) on the model parameters to avoid the need for validation. However, it has not proven that a larger IG (value) leads to a better learning performance. Our method has this desirable theoretical property without needing validation (Sec. 3). While existing methods demonstrate some effectiveness against replication using carefully selected validation sets [14, 18], our method achieves such a guarantee without needing validation. The work of [15] has considered replication from a different perspective and is thus not directly comparable to our method.

7 Conclusion and Future Work

This paper describes a model- and task-agnostic replication robust data valuation method that requires no validation. In particular, we value data based on its inherent diversity formalized as the volume of the data matrix because we have shown in Sec. 3 that a larger volume entails a better learning performance. We have identified that volume is not robust to replication, so we design a data valuation method based on the novel *robust volume* (RV) measure with a theoretical guarantee on replication robustness (Sec. 4). In our experiments (Sec. 5), we have used RV as a characteristic function in the Shapley value and empirical comparison with existing baseline methods verifies its effectiveness in data valuation and its robustness guarantee. Importantly, we have tested on various real-world datasets and our robust volume data valuation method can be flexibly adapted to handle machine learning models more complex than OLS (i.e., various neural networks) to demonstrate its practical applicability. Current works on data pricing may build on our perspective to ease the dependence on the validation set. For future work, we plan to consider more sophisticated replication techniques and investigate how to optimize the trade-off between diversity representation vs. replication 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 recom-

mendations 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] 100,000 UK used car dataset. URL `https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes`.

[2] Credit card fraud detection. URL `https://www.kaggle.com/mlg-ulb/creditcardfraud`.

[3] House sales in King County, USA. URL `https://www.kaggle.com/harlfoxem/housesalesprediction`.

[4] Trip Advisor hotel reviews. URL `https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews`.

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

[6] US census demographic data. URL `https://www.kaggle.com/muonneutrino/us-census-demographic-data`.

[7] A. Agarwal, M. Dahleh, and T. Sarkar. A marketplace for data: An algorithmic solution. In *Proc. ACM EC*, pages 701–726, 2019.

[8] S. Basu, P. Pope, and S. Feizi. Influence functions in deep learning are fragile. In *Proc. ICLR*, 2021.

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

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

[11] M. Dereziński, M. K. Warmuth, and D. Hsu. Leveraged volume sampling for linear regression. In *Proc. NeurIPS*, pages 2510–2519, 2018.

[12] P. Devereaux, G. Guyatt, H. Gerstein, S. Connolly, and S. Yusuf. Toward fairness in data sharing. *New England Journal of Medicine*, 375(5):405–407, 2016.

[13] J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, 1991.

[14] A. Ghorbani and J. Zou. Data Shapley: Equitable valuation of data for machine learning. In *Proc. ICML*, pages 2242–2251, 2019.

[15] D. Han, M. Wooldridge, A. Rogers, S. Tople, O. Ohrimenko, and S. Tschiatschek. Replication-robust payoff-allocation for machine learning data markets. arXiv:2006.14583, 2021.

[16] C. A. Jackevicius, J. An, D. T. Ko, J. S. Ross, S. Angraal, J. D. Wallach, M. Koh, J. Song, and H. M. Krumholz. Submissions from the sprint data analysis challenge on clinical risk prediction: A cross-sectional evaluation. *BMJ Open*, 9(3), 2019.

[17] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gurel, B. Li, C. Zhang, C. Spanos, and D. Song. Efficient task-specific data valuation for nearest neighbor algorithms. In *Proc. VLDB Endowment*, pages 1610–1623, 2019.

[18] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gurel, B. Li, C. Zhang, D. Song, and C. Spanos. Towards efficient data valuation based on the Shapley value. In *Proc. AISTATS*, pages 1167–1176, 2019.

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

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

[21] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *Proc. ICML*, pages 1885–1894, 2017.

[22] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008.

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

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

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

[26] L. Lovász. Submodular functions and convexity. In A. Bachem, B. Korte, and M. Grötschel, editors, *Mathematical Programming: The State of the Art – Bonn 1982*, pages 235–257. Springer, Berlin, Heidelberg, 1983.

[27] L. Lyu, X. Xu, Q. Wang, and H. 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.

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

[29] C. Orhan and Ö. Taştan. ALEVS: Active learning by statistical leverage sampling. In *Proc. ICML Workshop on Active Learning*, 2015.

[30] J. Pei. A survey on data pricing: From economics to data science. *IEEE TKDE*, 2020.

[31] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proc. EMNLP*, pages 1532–1543, 2014.

[32] R. Raskar, P. Vepakomma, T. Swedish, and A. Sharan. Data markets to support AI for all: Pricing, valuation and governance. arXiv:1905.06462, 2019.

[33] L. S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games*, volume 2, pages 307–317. Princeton Univ. Press, 1953.

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

[35] R. H. L. Sim, Y. Zhang, M. C. Chan, and B. K. H. Low. Collaborative machine learning with incentive-aware model rewards. In *Proc. ICML*, pages 8927–8936, 2020.

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

[37] S. Tay, X. Xu, C. S. Foo, and B. K. H. Low. Incentivizing collaboration in machine learning via synthetic data rewards. In *Proc. AAAI*, 2022.

[38] T. Wang, J. Rausch, C. Zhang, R. Jia, and D. 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] X. Xu, L. Lyu, X. Ma, C. Miao, C. S. Foo, and B. K. H. Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Proc. NeurIPS*, 2021.

[40] J. Yoon, S. O. Arik, and T. Pfister. Data valuation using reinforcement learning. In *Proc. ICML*, pages 10842–10851, 2020.

[41] Z. Zhang, Y. Song, and H. 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 a robustness guarantee 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 input 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.

A Proofs and Derivations

A.1 Larger Volume Entails Smaller Bias

Proof of Proposition 1. Let $\mathbf{P}_S$ be the zero-padded version of $\mathbf{X}_S$ such that $\mathbf{P}_S \in \mathbb{R}^{n\times d}$. In this proof, we only consider $d = 1$. Note $\|\mathbf{P}_S\|^2 = \|\mathbf{X}_S\|^2 = |\mathbf{X}_S^\top \mathbf{X}_S| = \mathrm{Vol}(\mathbf{X}_S)^2$ for $d = 1$. Recall the pseudo-inverse $\mathbf{X}^+ \coloneqq (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

$$
\begin{aligned}
\left\|\mathbf{X}^+ - \mathbf{P}_S^+\right\|^2 &= \left\| \frac{1}{\|\mathbf{X}\|^2}\mathbf{X}^\top - \frac{1}{\|\mathbf{X}_S\|^2}\mathbf{P}_S^\top \right\|^2 \\
&= \frac{1}{\|\mathbf{X}\|^4}\left\|\mathbf{X}^\top\right\|^2 + \frac{1}{\|\mathbf{X}_S\|^4}\left\|\mathbf{P}_S^\top\right\|^2 - \frac{2}{\|\mathbf{X}\|^2\|\mathbf{X}_S\|^2}\langle \mathbf{X}^\top, \mathbf{P}_S^\top\rangle \\
&= \frac{1}{\|\mathbf{X}\|^2} + \frac{1}{\|\mathbf{X}_S\|^2} - \frac{2}{\|\mathbf{X}\|^2\|\mathbf{X}_S\|^2}\|\mathbf{X}_S\|^2 \\
&= \frac{1}{\|\mathbf{X}_S\|^2} - \frac{1}{\|\mathbf{X}\|^2} \\
&= \frac{1}{\mathrm{Vol}(\mathbf{X}_S)^2} - \frac{1}{\mathrm{Vol}(\mathbf{X})^2}
\end{aligned}
$$

Since $\mathrm{Vol}(\mathbf{X})$ is constant, larger the $\mathrm{Vol}(\mathbf{X}_S)$, smaller the square of the bias $\left\|\mathbf{X}^+ - \mathbf{P}_S^+\right\|^2$. The proof of the proposition is complete. □

The proof above establishes a direct connection between bias_S and $1/V_S^2 - 1/V^2$. Therefore, we can extend to $M > 2$ non-zero submatrices such that, for any $\mathbf{X}_S, \mathbf{X}_{S'} \in \{\mathbf{X}_{S_1}, \mathbf{X}_{S_2}, \ldots, \mathbf{X}_{S_M}\}$, Proposition 1 still holds.

Proof of Proposition 2. The proof is relatively straightforward and follows the expansion of the l.h.s and substituting $\mathbf{X}^+$ by $\mathbf{G}^{-1}\mathbf{X}^\top$ and $\mathbf{X}_S^+$ by $\mathbf{G}_S^{-1}\mathbf{P}_S^\top$. Here, $\mathbf{G}_S \coloneqq \mathbf{X}_S^\top\mathbf{X}_S$ and $\mathbf{P}_S$ is the zero-padded version of $\mathbf{X}_S$. We use the expression and definitions of $\mathbf{Q}, \mathbf{Q}_S, \mathbf{Q}_{S'}$ in Lemma 2:

$$
\begin{aligned}
\mathrm{bias}_S^2 - \mathrm{bias}_{S'}^2 &\coloneqq \left\|\mathbf{X}^+ - \mathbf{P}_S^+\right\|^2 - \left\|\mathbf{X}^+ - \mathbf{P}_{S'}^+\right\|^2 \\
&= \left\|\mathbf{X}^+\right\|^2 - 2\langle \mathbf{X}^+, \mathbf{P}_S^+\rangle + \left\|\mathbf{P}_S^+\right\|^2 - \left\|\mathbf{X}^+\right\|^2 + 2\langle \mathbf{X}^+, \mathbf{P}_{S'}^+\rangle - \left\|\mathbf{P}_{S'}^+\right\|^2 \\
&= \left\|\mathbf{P}_S^+\right\|^2 - \left\|\mathbf{P}_{S'}^+\right\|^2 + 2\langle \mathbf{X}^+, \mathbf{P}_{S'}^+ - \mathbf{P}_S^+\rangle \\
&= \frac{1}{V_S^4}\left\|\mathbf{Q}_S\mathbf{P}_S^\top\right\|^2 - \frac{1}{V_{S'}^4}\left\|\mathbf{Q}_{S'}\mathbf{P}_{S'}^\top\right\|^2 + 2\langle \frac{1}{V^2}\mathbf{Q}\mathbf{X}^\top, \frac{1}{V_{S'}^2}\mathbf{Q}_{S'}\mathbf{P}_{S'}^\top - \frac{1}{V_S^2}\mathbf{Q}_S\mathbf{P}_S^\top\rangle \\
&= \frac{1}{V_S^4}\left\|\mathbf{Q}_S\mathbf{X}_S^\top\right\|^2 - \frac{1}{V_{S'}^4}\left\|\mathbf{Q}_{S'}\mathbf{X}_{S'}^\top\right\|^2 + 2\langle \frac{1}{V^2}\mathbf{Q}\mathbf{X}^\top, \frac{1}{V_{S'}^2}\mathbf{Q}_{S'}\mathbf{P}_{S'}^\top - \frac{1}{V_S^2}\mathbf{Q}_S\mathbf{P}_S^\top\rangle.
\end{aligned}
$$

□

Theorem 1 (Sylvester's Matrix Theorem). *Given a diagonalizable square matrix* $\mathbf{A}$ *and an analytic function* $f(\cdot)$*, we have,*

$$
f(\mathbf{A}) = \sum_{l=1}^{k} f(\lambda_l)\mathbf{A}_l \tag{4}
$$

where λ_l *is the* l*-th distinct eigenvalue of* $\mathbf{A}$ *and* $\mathbf{A}_l$ *is the Frobenius covariant defined as follows,*

$$
\mathbf{A}_l \coloneqq \prod_{j=1, j\neq l}^{k} \frac{1}{\lambda_l - \lambda_j}(\mathbf{A} - \lambda_j\mathbf{I}).
$$

Corollary 1. *Suppose* $f(\mathbf{A}) = \mathbf{A}^{-1}$*, then*

$$
\mathbf{A}^{-1} = f(\mathbf{A}) \coloneqq \sum_{l=1}^{k} \frac{1}{\lambda_l}\mathbf{A}_l.
$$

Lemma 2 (Expressing $\mathbf{G}^{-1}$ in $\mathrm{Vol}(\mathbf{X})$). *With $\mathbf{G} \coloneqq \mathbf{X}^\top\mathbf{X}$, then its inverse has*

$$\mathbf{G}^{-1} = V^{-2}\sum_{l=1}^{k}(\lambda_l\sigma_l)^{-1}\prod_{j=1,j\neq l}^{k}(\mathbf{G}-\lambda_j\mathbf{I}) \tag{5}$$

where λ_l is the l-th distinct eigenvalue of $\mathbf{G}$, and constant σ_l is defined as follows,

$$\sigma_l \coloneqq \sum_{g=1}^{k}(-1)^{g+1}\lambda_l^{k-g}\left[\sum_{\mathcal{H}\subseteq\{1,\ldots,k\}\setminus\{l\},|\mathcal{H}|=g-1}\left(\prod_{h\in\{1,\ldots,k\}\setminus\mathcal{H}}\lambda_h^{-1}\right)\right].$$

We define $\mathbf{Q} \coloneqq \sum_{l=1}^{k}(\lambda_l\sigma_l)^{-1}\prod_{j=1,j\neq l}^{k}(\mathbf{G}-\lambda_j\mathbf{I})$ for convenience.

Proof of Lemma 2. The proof uses a key result, Sylvester's matrix theorem, specifically the corollary for the inverse of a matrix (reproduced above in Corollary 1) and properties of the left Gram matrix $\mathbf{G} \coloneqq \mathbf{X}^\top\mathbf{X}$ such as invertibility and positive definiteness when $\mathbf{X}$ is full-rank.

First, observe since $\mathbf{G}$ is a real symmetric matrix, it is diagonalizable, hence a direct application of the corollary above gives

$$\mathbf{G}^{-1} = \sum_{l=1}^{k}\frac{1}{\lambda_l}\mathbf{G}_l \tag{6}$$

where $\mathbf{G}_l$ is the Frobenius covariant defined in the above theorem and we consider $\mathbf{G}_l$ on its own,

$$\begin{aligned}\mathbf{G}_l &\coloneqq \prod_{j=1,j\neq l}^{k}\frac{1}{\lambda_l-\lambda_j}(\mathbf{G}-\lambda_j\mathbf{I}) \\ &= \underbrace{\prod_{j=1,j\neq l}^{k}\frac{1}{\lambda_l-\lambda_j}}_{p_l}\times\underbrace{\prod_{j=1,j\neq l}^{k}(\mathbf{G}-\lambda_j\mathbf{I})}_{\mathbf{M}_l}\end{aligned} \tag{7}$$

Observe the denominator of the expanded p_l is a summation of terms that are products of multiple λ_l's and all with coefficient either 1 or -1. To see from a specific and self-contained example: suppose $k = 4 = l$, so

$$\begin{aligned}p_l &= \frac{1}{\lambda_4-\lambda_1}\times\frac{1}{\lambda_4-\lambda_2}\times\frac{1}{\lambda_4-\lambda_3} \\ &= \frac{1}{\lambda_4^3-\lambda_4^2\lambda_1-\lambda_4^2\lambda_2-\lambda_4^2\lambda_3+\lambda_4\lambda_1\lambda_2+\lambda_4\lambda_1\lambda_3+\lambda_4\lambda_2\lambda_3-\lambda_1\lambda_2\lambda_3}.\end{aligned}$$

Since it is easier to work with $\frac{1}{p_l}$, we derive the following formula for it by extracting a common factor of $\Lambda \coloneqq \prod_{i=l}^{k}\lambda_l$ to give

$$\frac{1}{p_l} = \underbrace{\Lambda\sum_{g=1}^{k}(-1)^{g+1}\lambda_l^{k-g}\left[\sum_{\mathcal{H}\subseteq\{1,\ldots,k\}\setminus\{l\},|\mathcal{H}|=g-1}\left(\prod_{h\in\{1,\ldots,k\}\setminus\mathcal{H}}\frac{1}{\lambda_h}\right)\right]}_{\sigma_l}. \tag{8}$$

Using the result that determinant of the left Gram matrix is the product of its eigenvalues, we have $|\mathbf{G}| = \Lambda$, and substituting the definition of σ_l, we rewrite (7) as follows,

$$\mathbf{G}_l = \frac{1}{|\mathbf{G}|}\frac{1}{\sigma_l}\mathbf{M}_l. \tag{9}$$

Recalling $\mathrm{Vol}(\mathbf{X})^2 = |\mathbf{G}|$ and plugging (9) back into (6) gives

$$\mathbf{G}^{-1} = \frac{1}{\mathrm{Vol}(\mathbf{X})^2}\underbrace{\sum_{l=1}^{k}\frac{1}{\lambda_l}\frac{1}{\sigma_l}\mathbf{M}_l}_{\mathbf{Q}}.$$

□

Examining the additional scenario of case 2) in Proposition 2. The scenario is where $\mathbf{X}_S, \mathbf{X}_{S'}$ are similar in the sense that they may contain a similar number of rows, and are drawn from the same distribution. We focus on showing $V \gg \max(V_S, V_{S'})$ and assume $\left\|\mathbf{Q}_S \mathbf{X}_S^\top\right\| \approx \left\|\mathbf{Q}_{S'} \mathbf{X}_{S'}^\top\right\|$ (which we verify empirically later by showing Proposition 2 is true most of the time in Fig.1).

Lemma 3 states that V^2 is larger than $V_S^2, V_{S'}^2$ by a multiplicative factor which is exponential in the number of rows in $V_{S'}, V_S$. See Fig. 8 for an illustration.

Lemma 3 (V **vs.** $V_S, V_{S'}$)**.** *Let $V, V_S, V_{S'}$ be the respective volumes of $\mathbf{X}, \mathbf{X}_S, \mathbf{X}_{S'}$ and let s, s' denote the respective number of rows in $\mathbf{X}_S, \mathbf{X}_{S'}$. Assume $\mathbf{X}, \mathbf{X}_S, \mathbf{X}_{S'}$ are all full-rank, we have*

$$V^2 > \max((1+\xi_{S'})^{s'} V_S^2, (1+\xi_S)^s V_{S'}^2)$$

where $\xi_{S'} \coloneqq \min_{\mathbf{x}_q \in \mathbf{X}_{S'}} \mathbf{x}_q (\mathbf{X}_S^\top \mathbf{X}_S)^{-1} \mathbf{x}_q^\top > 0$ and $\xi_S \coloneqq \min_{\mathbf{x}_q \in \mathbf{X}_S} \mathbf{x}_q (\mathbf{X}_{S'}^\top \mathbf{X}_{S'})^{-1} \mathbf{x}_q^\top > 0$.

Proof of Lemma 3. This is a constructive proof. We will add rows one by one from $\mathbf{X}_{S'}$ to $\mathbf{X}_S$ to finally construct $\mathbf{X}$. For an arbitrary row $\mathbf{x}_q$ from $\mathbf{X}_{S'}$, we have

$$\begin{aligned}
|\mathbf{X}_{S\cup\{q\}}^\top \mathbf{X}_{S\cup\{q\}}| &= |\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{x}_q^\top \mathbf{x}_q| \\
&= \underbrace{(1 + \mathbf{x}_q (\mathbf{X}_S^\top \mathbf{X}_S)^{-1} \mathbf{x}_q^\top)}_{\text{coeff}_\emptyset} |\mathbf{X}_S^\top \mathbf{X}_S| \\
&\geq (1+\xi_{S'}) V_S^2
\end{aligned}$$

The second equality uses the matrix determinant lemma. We can repeat this addition for every row in $\mathbf{X}_{S'}$. Note after adding $\mathbf{x}_q$, and we want to add a different row $\mathbf{x}_{q'}$, in the second line the new coefficient (having added $\mathbf{x}_q$) is $\text{coeff}_{\{q\}} = (1 + \mathbf{x}_{q'} (\mathbf{X}_{S\cup\{q\}}^\top \mathbf{X}_{S\cup\{q\}})^{-1} \mathbf{x}_{q'}^\top)$ and $\text{coeff}_{\{q\}} \geq \text{coeff}_\emptyset$ because $(\mathbf{X}_{S\cup\{q\}}^\top \mathbf{X}_{S\cup\{q\}})^{-1}$ is positive definite and the previously added row $\mathbf{x}_q$ now makes a non-negative contribution to the sum, therefore $(1+\xi_{S'})$ is still a valid lower-bound for $\text{coeff}_{\{q\}}$. In addition, obviously $|\mathbf{X}_S^\top \mathbf{X}_S| \leq |\mathbf{X}_{S\cup\{q\}}^\top \mathbf{X}_{S\cup\{q\}}|$ so V_S^2 is a valid lower bound. Recursively adding this for s' times gives the desired lower bound of $(1+\xi_{S'})^{s'} V_S^2$. The result then follows. □

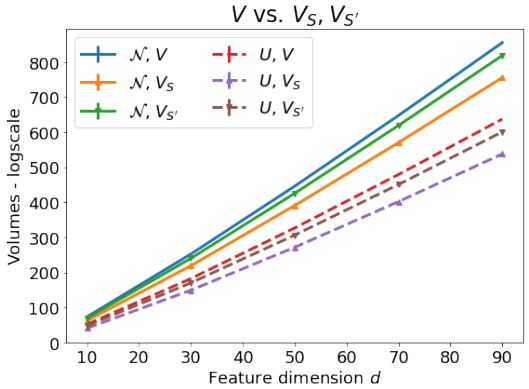

Figure 8: The volume of full matrix $\mathbf{X}$ against the volumes of submatrices $\mathbf{X}_S, \mathbf{X}_{S'}$. $\mathbf{X}_S, \mathbf{X}_{S'}$ are randomly sampled from normal ($\mathcal{N}$, solid lines) and uniform (U, dashed lines) distributions and concatenated to form $\mathbf{X}$, with $\mathbf{X}_{S'}$ containing twice the number of rows as in $\mathbf{X}_S$. The results are in log-scale. The volume of the full matrix $\mathbf{X}$ is noticeably larger than $V_S, V_{S'}$ even in log-scale, indicating the actual volume is significantly larger, validating our claim that $V \gg \max(V_S, V_{S'})$.

A.2 Larger Volume Entails Smaller MSE

Proof of Proposition 3. Recall the least squares solution from OLS on training set $(\mathbf{X}, \mathbf{y})$ is $\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. In the case of $d = 1$, the least squared solution $\mathbf{w}$ is a scalar. We simplify the

notation by letting w and w' be the least squared solutions on training set $(\mathbf{X}_S, \mathbf{y}_S)$ and $(\mathbf{X}_{S'}, \mathbf{y}_{S'})$. Now,

$$w = \frac{1}{\|\mathbf{X}_S\|^2}\mathbf{X}_S^\top \mathbf{y}_S, \quad w' = \frac{1}{\|\mathbf{X}_{S'}\|^2}\mathbf{X}_{S'}^\top \mathbf{y}_{S'}$$

Then,

$$\begin{aligned}
L(\mathbf{w}_S) &= \|\mathbf{y} - \mathbf{X}w\|^2 \\
&= \|\mathbf{y}_S - \mathbf{X}_S w\|^2 + \|\mathbf{y}_{S'} - \mathbf{X}_{S'} w\|^2 \\
&= \|\mathbf{y}_S\|^2 + w^2\|\mathbf{X}_S\|^2 - 2w\langle \mathbf{X}_S, \mathbf{y}_S\rangle + \|\mathbf{y}_{S'}\|^2 + w^2\|\mathbf{X}_{S'}\|^2 - 2w\langle \mathbf{X}_{S'}, \mathbf{y}_{S'}\rangle \\
&= \|\mathbf{y}_S\|^2 + \|\mathbf{y}_{S'}\|^2 + \frac{1}{\|\mathbf{X}_S\|^4}(\mathbf{X}_S^\top \mathbf{y}_S)^2\left(\|\mathbf{X}_S\|^2 + \|\mathbf{X}_{S'}\|^2\right) - \frac{2}{\|\mathbf{X}_S\|^2}\mathbf{X}_S^\top \mathbf{y}_S\left(\mathbf{X}_S^\top \mathbf{y}_S + \mathbf{X}_{S'}^\top \mathbf{y}_{S'}\right) \\
&= \|\mathbf{y}_S\|^2 + \|\mathbf{y}_{S'}\|^2 - \frac{(\mathbf{X}_S^\top \mathbf{y}_S)^2}{\|\mathbf{X}_S\|^2} + \frac{\|\mathbf{X}_{S'}\|^2(\mathbf{X}_S^\top \mathbf{y}_S)^2}{\|\mathbf{X}_S\|^4} - \frac{2\left(\mathbf{X}_S^\top \mathbf{y}_S\right)\left(\mathbf{X}_{S'}^\top \mathbf{y}_{S'}\right)}{\|\mathbf{X}_S\|^2} \\
&= \|\mathbf{y}_S\|^2 + \|\mathbf{y}_{S'}\|^2 + \frac{(\mathbf{X}_S^\top \mathbf{y}_S)^2}{\|\mathbf{X}_S\|^4}\left(\|\mathbf{X}_{S'}\|^2 - \|\mathbf{X}_S\|^2\right) - \frac{2\left(\mathbf{X}_S^\top \mathbf{y}_S\right)\left(\mathbf{X}_{S'}^\top \mathbf{y}_{S'}\right)}{\|\mathbf{X}_S\|^2}
\end{aligned}$$

Similarly,

$$L(\mathbf{w}_{S'}) = \|\mathbf{y}_S\|^2 + \|\mathbf{y}_{S'}\|^2 + \frac{(\mathbf{X}_{S'}^\top \mathbf{y}_{S'})^2}{\|\mathbf{X}_{S'}\|^4}\left(\|\mathbf{X}_S\|^2 - \|\mathbf{X}_{S'}\|^2\right) - \frac{2\left(\mathbf{X}_{S'}^\top \mathbf{y}_{S'}\right)\left(\mathbf{X}_S^\top \mathbf{y}_S\right)}{\|\mathbf{X}_{S'}\|^2}$$

Subtracting,

$$\begin{aligned}
L(\mathbf{w}_S) - L(\mathbf{w}_{S'}) &= \left[\frac{(\mathbf{X}_S^\top \mathbf{y}_S)^2}{\|\mathbf{X}_S\|^4} + \frac{(\mathbf{X}_{S'}^\top \mathbf{y}_{S'})^2}{\|\mathbf{X}_{S'}\|^4} - \frac{2(\mathbf{X}_S^\top \mathbf{y}_S)(\mathbf{X}_{S'}^\top \mathbf{y}_{S'})}{\|\mathbf{X}_S\|^2\|\mathbf{X}_{S'}\|^2}\right]\left(\|\mathbf{X}_{S'}\|^2 - \|\mathbf{X}_S\|^2\right) \\
&= \left[\frac{\mathbf{X}_S^\top \mathbf{y}_S}{\|\mathbf{X}_S\|^2} - \frac{\mathbf{X}_{S'}^\top \mathbf{y}_{S'}}{\|\mathbf{X}_{S'}\|^2}\right]^2 \left(\mathrm{Vol}(\mathbf{X}_{S'})^2 - \mathrm{Vol}(\mathbf{X}_S)^2\right)
\end{aligned}$$

The last step follows from the fact that $\mathrm{Vol}(\mathbf{X})^2 = |\mathbf{X}^\top \mathbf{X}| = \|\mathbf{X}\|^2$ when $d = 1$. Therefore, we have $L(\mathbf{w}_S) \leq L(\mathbf{w}_{S'})$ if and only if $\mathrm{Vol}(\mathbf{X}_S) \geq \mathrm{Vol}(\mathbf{X}_{S'})$. □

Alternate Proof of Proposition 3. Following Lemma 4, we denote $\Gamma \coloneqq \mathbf{X}_S^\top \mathbf{y}_S,\ \Gamma' \coloneqq \mathbf{X}_{S'}^\top \mathbf{y}_{S'}$ to consider $L(w) - L(w')$ as follows,

$$\begin{aligned}
L(w) - L(w') &= [\|\mathbf{y}\|^2 - w^2(V_S^2 - V_{S'}^2) - 2w\Gamma'] - [\|\mathbf{y}\|^2 - w'^2(V_{S'}^2 - V_S^2) - 2w'\Gamma] \\
&= -w^2(V_S^2 - V_{S'}^2) + w'^2(V_{S'}^2 - V_S^2) - 2w\Gamma' + 2w'\Gamma \\
&= (w^2 + w'^2)(V_{S'}^2 - V_S^2) - 2(w\Gamma' - w'\Gamma) \\
&= (w^2 + w'^2)(V_{S'}^2 - V_S^2) - 2(w\frac{\Gamma'}{V_{S'}^2}V_{S'}^2 - w'\frac{\Gamma}{V_S^2}V_S^2) \\
&= (w^2 + w'^2)(V_{S'}^2 - V_S^2) - 2(ww'V_{S'}^2 - w'wV_S^2) \quad \text{Noting } w = \frac{\Gamma}{V_S^2}, w' = \frac{\Gamma'}{V_{S'}^2} \\
&= (w^2 + w'^2)(V_{S'}^2 - V_S^2) - 2ww'(V_{S'}^2 - V_S^2) \\
&= (w - w')^2(V_{S'}^2 - V_S^2)
\end{aligned}$$

Since $(w - w')^2 \geq 0$, we have $L(w) - L(w') \geq 0 \iff V_{S'}^2 - V_S^2 \geq 0$ or equivalently, $L(w) \geq L(w') \iff V_{S'}^2 \geq V_S^2$. □

Lemma 4 (MSE of the least-squares solution on $\mathbf{X}_S$ for $d = 1$). *Let S, S' be a partition of the rows of the matrix $\mathbf{X}$, so that $\mathbf{X}_S, \mathbf{X}_{S'}$ are submatrices of $\mathbf{X}$, i.e. $\mathbf{X} = [\mathbf{X}_S^\top\ \mathbf{X}_{S'}^\top]^\top$. Let $\mathbf{y}_S, \mathbf{y}_{S'}$ be defined similarly. Further, let w denote the least squares solution (note it is a scalar for $d = 1$) on the submatrix $\mathbf{X}_S$ with labels $\mathbf{y}_S$, i.e. $w = (\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{X}_S^\top \mathbf{y}_S = \Gamma/V_S^2$. Then the mean squared loss on the full matrix $L(w) \coloneqq \|\mathbf{y} - \mathbf{X}w\|^2$ is*

$$L(w) = ||\mathbf{y}||^2 - w^2(V_S^2 - V_{S'}^2) - 2w\Gamma' \tag{10}$$

where $V_S \coloneqq \mathrm{Vol}(\mathbf{X}_S), V_{S'} \coloneqq \mathrm{Vol}(\mathbf{X}_{S'})$ and $\Gamma' \coloneqq \mathbf{X}^\top \mathbf{y} - \mathbf{X}_S^\top \mathbf{y}_S$.

Proof of Lemma 4. For $d = 1$, observe the least squares solution $\mathbf{w} \coloneqq (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$ is a scalar and $(\mathbf{X}^\top \mathbf{X})^{-1} = 1/(||\mathbf{X}||^2)$. Further, note that $\mathrm{Vol}(\mathbf{X})^2 \coloneqq \det(\mathbf{X}^\top \mathbf{X}) = ||\mathbf{X}||^2$. Subsequently, the least squares solution w for $(\mathbf{X}_S, \mathbf{y}_S)$ is $w = (\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{X}_S^\top \mathbf{y}_S = \Gamma / V_S^2$, where $\Gamma \coloneqq \mathbf{X}_S^\top \mathbf{y}_S$, and let w', Γ' be defined similarly for $(\mathbf{X}_{S'}, \mathbf{y}_{S'})$. Then we have

$$\begin{aligned}
L(w) &= \|\mathbf{y} - \mathbf{X}w\|^2 \\
&= \|\mathbf{y}\|^2 - 2w\langle \mathbf{X}, \mathbf{y}\rangle + w^2 \|\mathbf{X}\|^2 \\
&= \|\mathbf{y}\|^2 - 2w\langle \mathbf{X}, \mathbf{y}\rangle + w^2 \|\mathbf{X}_S\|^2 + w^2 \|\mathbf{X}_{S'}\|^2 \\
&= \|\mathbf{y}\|^2 - 2w(\Gamma + \Gamma') + w^2 \|\mathbf{X}_S\|^2 + w^2 \|\mathbf{X}_{S'}\|^2 \\
&= \|\mathbf{y}\|^2 - 2w^2 V_S^2 - 2w\Gamma' + w^2 V_S^2 + w^2 V_{S'}^2 \\
&= \|\mathbf{y}\|^2 - w^2(V_S^2 - V_{S'}^2) - 2w\Gamma'.
\end{aligned}$$

□

Derivation of (1). We will expand the $L(\mathbf{w}_S)$ into several terms and show there are common terms to both $L(\mathbf{w}_S)$ and $L(\mathbf{w}_{S'})$ which get canceled in $L(\mathbf{w}_S) - L(\mathbf{w}_{S'})$ and we analyze the difference in the remainder terms.

$$L(\mathbf{w}_S) \coloneqq ||\mathbf{y} - \mathbf{X}\mathbf{w}_S||^2 = \|\mathbf{y}_S - \mathbf{X}_S\mathbf{w}_S\|^2 + \underbrace{\|\mathbf{y}_{S'} - \mathbf{X}_{S'}\mathbf{w}_S\|^2}_{A}$$

Next, we want to write $A = \|\mathbf{y}_{S'} - \mathbf{X}_{S'}\mathbf{w}_{S'}\|^2 + R_S$, letting R_S denote the remainder term. Note with this expression we have $L(\mathbf{w}_S) - L(\mathbf{w}_{S'}) = R_S - R_{S'}$. We first expand A as follows,

$$A = \underbrace{\|\mathbf{y}_{S'}\|^2}_{A_1} \underbrace{-2\langle \mathbf{y}_{S'}, \mathbf{X}_{S'}\mathbf{w}_S\rangle}_{B} + \underbrace{\|\mathbf{X}_{S'}\mathbf{w}_S\|^2}_{C}$$

Next to rewrite B, C as follows,

$$\begin{aligned}
B &= -2\langle \mathbf{y}_{S'}, \mathbf{X}_{S'}(\mathbf{w}_{S'} - \mathbf{w}_{S'} + \mathbf{w}_S)\rangle = \underbrace{-2\langle \mathbf{y}_{S'}, \mathbf{X}_{S'}\mathbf{w}_{S'}\rangle}_{B_1} + \underbrace{-2\langle \mathbf{y}_{S'}, \mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\rangle}_{B_2}, \\
C &= \|\mathbf{X}_{S'}\mathbf{w}_S\|^2 = \|\mathbf{X}_{S'}(\mathbf{w}_{S'} - \mathbf{w}_{S'} + \mathbf{w}_S)\|^2 \\
&= \underbrace{\|\mathbf{X}_{S'}\mathbf{w}_{S'}\|^2}_{C_1} + \underbrace{2\langle \mathbf{X}_{S'}\mathbf{w}_{S'}, \mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\rangle + \|\mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\|^2}_{C_2}.
\end{aligned}$$

We can collect A_1, B_1, C_1 to complete the square of $\|\mathbf{y}_{S'} - \mathbf{X}_{S'}\mathbf{w}_{S'}\|^2$ and naturally collect B_2, C_2 to form the remainder R_S as follows,

$$\begin{aligned}
R_S &= -2\langle \mathbf{y}_{S'}, \mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\rangle + 2\langle \mathbf{X}_{S'}\mathbf{w}_{S'}, \mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\rangle + \|\mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\|^2 \\
&= 2\langle \mathbf{X}_{S'}\mathbf{w}_{S'} - \mathbf{y}_{S'}, \mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\rangle + \|\mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\|^2 \\
&= \langle 2(\mathbf{X}_{S'}\mathbf{w}_{S'} - \mathbf{y}_{S'}) + \mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'}), \mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\rangle \\
&= \langle \mathbf{X}_{S'}(\mathbf{w}_S + \mathbf{w}_{S'}) - 2\mathbf{y}_{S'}, \mathbf{X}_{S'}(\mathbf{w}_S - \mathbf{w}_{S'})\rangle \\
&= \langle \mathbf{X}_{S'}^\top[\mathbf{X}_{S'}(\mathbf{w}_S + \mathbf{w}_{S'}) - 2\mathbf{y}_{S'}], \mathbf{w}_S - \mathbf{w}_{S'}\rangle.
\end{aligned}$$

Now we can derive $R_S - R_{S'}$ as follows,

$$\begin{aligned}
R_S - R_{S'} &= \langle \mathbf{X}_{S'}^\top[\mathbf{X}_{S'}(\mathbf{w}_S + \mathbf{w}_{S'}) - 2\mathbf{y}_{S'}], \mathbf{w}_S - \mathbf{w}_{S'}\rangle - \langle \mathbf{X}_S^\top[\mathbf{X}_S(\mathbf{w}_{S'} + \mathbf{w}_S) - 2\mathbf{y}_S], \mathbf{w}_{S'} - \mathbf{w}_S\rangle \\
&= \langle (\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{X}_{S'}^\top \mathbf{X}_{S'})(\mathbf{w}_{S'} + \mathbf{w}_S) - 2\mathbf{X}_S^\top \mathbf{y}_S - 2\mathbf{X}_{S'}^\top \mathbf{y}_{S'}, \mathbf{w}_S - \mathbf{w}_{S'}\rangle \\
&= \langle (\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{X}_{S'}^\top \mathbf{X}_{S'})(\mathbf{w}_{S'} + \mathbf{w}_S) - 2\mathbf{X}^\top \mathbf{y}, \mathbf{w}_S - \mathbf{w}_{S'}\rangle
\end{aligned}$$

□

Adversarially constructed counter-examples to achieve arbitrary signs of $L(\mathbf{w}_S) - L(\mathbf{w}_{S'})$. Given $\mathbf{X}_S, \mathbf{X}_{S'}$ as follows, we construct two sets of labels $\mathbf{y}_1, \mathbf{y}_2$ where $\mathbf{y}_1 = [\mathbf{y}_{1,S}^\top \ \mathbf{y}_{1,S'}^\top]^\top$ and $\mathbf{y}_2 = [\mathbf{y}_{2,S}^\top \ \mathbf{y}_{2,S'}^\top]^\top$ such that $L(\mathbf{w}_S) < L(\mathbf{w}_{S'})$ on $\mathbf{y}_1$ while $L(\mathbf{w}_S) > L(\mathbf{w}_{S'})$ on $\mathbf{y}_2$.

Note this example is adversarially constructed to show that a larger volume does not necessarily lead to a smaller MSE because $\mathrm{Vol}()$ does *not* take the labels into consideration.

Let $d = 2,\ n = 6$ and fix $\mathbf{X}_S = \begin{bmatrix} 0.4197849, 0.82836752 \\ 0.8393158, 0.24545882 \\ 0.8544813, 0.72294841 \end{bmatrix}, \mathbf{X}_{S'} = \begin{bmatrix} 0.40205988, 0.44985846 \\ 0.36588236, 0.33433118 \\ 0.79521338, 0.34753677 \end{bmatrix}$.

Set two sets of labels $\mathbf{y}_1, \mathbf{y}_2$ as follows, $\mathbf{y}_{1,S} = \mathbf{0}$ and $\mathbf{y}_{1,S'} = \{\exp(10 \times x[1]) | x \in S'\}$; and $\mathbf{y}_{2,S} = \{\exp(10 \times x[1]) | x \in S\}$ and $\mathbf{y}_{2,S'} = \mathbf{0}$. The $\exp(10 \times x[1])$ refers to taking the exponential of the product between 10 and the second value of a datum $x \in \mathbb{R}^2$. An observation is that since the true function/labels only depend on the second feature of each data point, the first feature is redundant/unnecessary to achieve a small MSE and yet included in the volume calculation. This is how we can construct the adversarial labels.

With this setting, we have $\mathrm{Vol}(\mathbf{X}_S) > \mathrm{Vol}(\mathbf{X}_{S'})$ and $L(\mathbf{w}_S) < L(\mathbf{w}_{S'})$ on $\mathbf{y}_1$ while $L(\mathbf{w}_S) > L(\mathbf{w}_{S'})$ on $\mathbf{y}_2$.

A.3 Replication Robustness

Proof of Lemma 1. Using the same notation for $\mathbf{X}_{\text{rep}} = [\mathbf{X}^\top\ \mathbf{x}_q^\top\ \dots\ \mathbf{x}_q^\top]^\top \in \mathbb{R}^{(n+m)\times d}$, we can write $\mathrm{Vol}(\mathbf{X}_{\text{rep}})^2 \coloneqq |\mathbf{X}_{\text{rep}}^\top \mathbf{X}_{\text{rep}}|$. Consider the simple case where $m = 1$, we have

$$\begin{aligned} |\mathbf{X}_{\text{rep}}^\top \mathbf{X}_{\text{rep}}| &= \left| [\mathbf{X}^\top\ \mathbf{x}_q^\top] \begin{bmatrix} \mathbf{X} \\ \mathbf{x}_q \end{bmatrix} \right| = |\mathbf{X}^\top \mathbf{X} + \mathbf{x}_q^\top \mathbf{x}_q| \\ &= (1 + \mathbf{x}_q (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_q^\top) |\mathbf{X}^\top \mathbf{X}| \end{aligned}$$

where the last equality uses the matrix determinant lemma. Note with a full-rank $\mathbf{X}$, $\mathbf{X}^\top \mathbf{X}$ is invertible, symmetric and positive semi-definite, thus diagonalizable, as required by the matrix determinant lemma.

In general,

$$\begin{aligned} |\mathbf{X}_{\text{rep}}^\top \mathbf{X}_{\text{rep}}| &= \left| [\mathbf{X}^\top\ \underbrace{\mathbf{x}_q^\top\ \dots\ \mathbf{x}_q^\top}_{m \text{ terms}}] \begin{bmatrix} \mathbf{X} \\ \mathbf{x}_q \\ \vdots \\ \mathbf{x}_q \end{bmatrix} \right| = |\mathbf{X}^\top \mathbf{X} + m \times \mathbf{x}_q^\top \mathbf{x}_q| \\ &= (1 + m \times \mathbf{x}_q (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_q^\top) |\mathbf{X}^\top \mathbf{X}| \end{aligned}$$

In other words, $\mathrm{Vol}(\mathbf{X}_{\text{rep}}) = \mathrm{Vol}(\mathbf{X}) \times (1 + m \times \mathbf{x}_q (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_q^\top)^{1/2}$. □

Lemma 5 (**Inflation vs. α**). *When $\alpha = 1/(\beta n)$, the inflation of replicating $\mathbf{X} \in \mathbb{R}^{n\times d}$ with a replication factor c is upper bounded as follows,*

$$\lim_{n\to\infty} \left(\sum_{p=0}^{nc} (\frac{1}{\beta n})^p \right)^n = \exp(\beta^{-1}).$$

Proof of Lemma 5. Suppose some rows of $\mathbf{X}$ are copied and appended back to get $\mathbf{X}_{\text{rep}} \in \mathbb{R}^{(nc)\times d}$ such that the replication factor is c. For simplicity, we set ω such that each d-cube contains identical data points. As there are n rows in the original $\mathbf{X}$, there can be at most n non-empty d-cubes for any ω by the pigeon-hole principle. Note since the replication is by direct copying, the number of non-empty d-cubes for $\mathbf{X}_{\text{rep}}$ is upper bounded by n.

We next consider the inflation, which is the ratio $\mathrm{RV}(\mathbf{X}_{\text{rep}};\omega)/\,\mathrm{RV}(\mathbf{X};\omega)$ as follows,

$$\begin{aligned}
\frac{\mathrm{RV}(\mathbf{X}_{\text{rep}};\omega)}{\mathrm{RV}(\mathbf{X};\omega)} &= \frac{\mathrm{Vol}(\widetilde{\mathbf{X}}_{\text{rep}}) \times \prod_{i\in\Omega_{\text{rep}}} \rho_{i,\text{rep}}}{\mathrm{Vol}(\widetilde{\mathbf{X}}) \times \prod_{i\in\Omega} \rho_i} \\
&= \frac{\prod_{i\in\Omega_{\text{rep}}} \rho_{i,\text{rep}}}{\prod_{i\in\Omega} \rho_i} \\
&\leq \frac{\prod_{i\in\Omega_{\text{rep}}} \rho_{i,\text{rep}}}{1} \triangleq \prod_{i\in\Omega_{\text{rep}}} \sum_{p=0}^{\phi_{i,\text{rep}}} \alpha^p \\
&\leq \prod_{i\in\Omega_{\text{rep}}} \sum_{p=0}^{nc} \alpha^p \triangleq \prod_{i\in\Omega_{\text{rep}}} \sum_{p=0}^{nc} (\frac{1}{\beta n})^p \\
&\leq (\sum_{p=0}^{nc} (\frac{1}{\beta n})^p)^n
\end{aligned}$$

The first line is by definition; the second equality is by observing that $\widetilde{\mathbf{X}}_{\text{rep}} = \widetilde{\mathbf{X}}$ due to direct copying and that each d-cube contains identical data points; the next inequality is by observing $\prod_{i\in\Omega} \rho_i \geq 1$; the next equality is by definition of ρ_i; the next inequality uses nc to upper bound the number of data points in any d-cube; the next equality substitutes $\alpha = 1/(\beta n)$ and the last inequality is by bounding the number of non-empty d-cubes by n.

We upper-bound $\sum_{p=0}^{nc} (1/\beta n)^p$ with respect to $c \to \infty$ as follows,

$$\sum_{p=0}^{nc} (\frac{1}{\beta n})^p \leq \frac{1}{1-\frac{1}{\beta n}} = \frac{\beta n}{\beta n - 1} = 1 + \frac{1}{\beta n - 1}.$$

Next, apply the limit of $n \to \infty$ to give the following [2]:

$$\lim_{n\to\infty} (1 + \frac{1}{\beta n - 1})^n = \lim_{n\to\infty} (1 + \frac{1}{n} \times \frac{1}{\beta - \frac{1}{n}})^n = \exp(\beta^{-1}).$$

The last equality is by first considering $(\beta - \frac{1}{n})^{-1} \to \beta^{-1}$ as $n \to \infty$ and then using a known result of $\lim_{n\to\infty} (1 + \frac{x}{n})^n = \exp(x)\ \forall x$. □

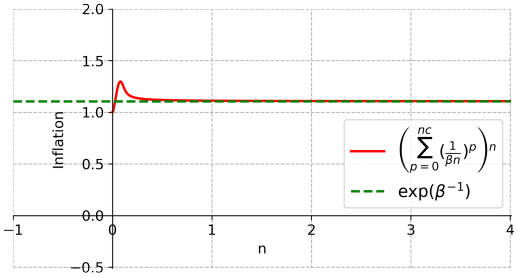

Figure 9: Inflation vs. n. With $\beta = 10, c = 100$, the inflation (red) quickly decays and converges to below the green line where $\exp(\beta^{-1}) \approx 1.105$. Note x-axis denotes n, the number of data points.

Note Lemma 5 upper bounds the inflation is with respect to the replication factor c, while Lemma 6 upper bounds the RV of a $\mathbf{X}$ which may not contain replicated data.

Lemma 6 ($\mathrm{RV}(\cdot;\omega)$ **vs.** ω)**.** *The growth of* $\mathrm{RV}(\mathbf{X};\omega)$ *with respect to* ω *is slow and upper bounded as follows,*

$$\sup_{\omega} \lim_{n\to\infty} \frac{\mathrm{RV}(\mathbf{X};\omega)}{\mathrm{Vol}(\mathbf{X})} \leq \exp(\beta^{-1})$$

[2] The asymptotic condition on n is included for theoretical rigor and may be easily removed in practice for any reasonable n (larger than 10). See Fig. 9.

where n is the number of rows in $\mathbf{X}$ and $\beta = (\alpha n)^{-1}$ where α is a user-determined robustness coefficient.

Proof of Lemma 6. Recall $\mathrm{RV}(\mathbf{X};\omega) \coloneqq \mathrm{Vol}(\widetilde{\mathbf{X}}) \times \prod_{i\in\Omega} \rho_i$. We assume $\mathrm{Vol}(\widetilde{\mathbf{X}}) \leq \mathrm{Vol}(\mathbf{X})$ because $\widetilde{\mathbf{X}}$ possibly has a smaller number of rows. Due to Lemma 1, we know that reducing a row decreases the total volume. We subsequently verify the result of Lemma 6 to confirm this assumption is satisfied.

To derive an upper bound on $\prod_{i\in\Omega} \rho_i$, we first consider the fact that given an ω, the number of non-empty d-cubes is upper bounded by n, the number of rows in $\mathbf{X}$. Then, for each of these d-cubes, the up-weight constant ρ_i is upper bounded by $1/(1-\alpha)$ due to the geometric series sum. Therefore,

$$\prod_{i\in\Omega} \rho_i \leq (\frac{1}{1-\frac{1}{\beta n}})^n.$$

Using the similar technique in proof of Lemma 5, we have

$$\lim_{n\to\infty} (\frac{1}{1-\frac{1}{\beta n}})^n = \exp(\beta^{-1}),$$

and the result follows.

Fig. 10 shows a numerical experiment where 50 matrices $\mathbf{X}$ are independently and randomly drawn, and for each we compute the ratio $\mathrm{RV}(\mathbf{X};\omega)/\,\mathrm{Vol}(\mathbf{X})$ (in y-axis) over a range of ω (in x-axis). In particular, $\beta = 10$ and we see the upper bound is in fact followed. □

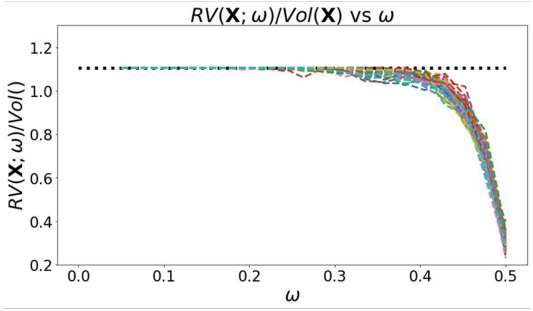

Figure 10: Growth of $\mathrm{RV}(\mathbf{X};\omega)$ is upper bounded by $\exp(\beta^{-1})$ (black dotted line). $\beta = 10$ and the discretization coefficient ω takes range $(0, 0.5)$.

Proof of Proposition 5 (Bounded Distortion). We re-arrange the distortion $\delta(\omega) \coloneqq [\mathrm{RV}(\mathbf{X}_S;\omega)/\,\mathrm{RV}(\mathbf{X}_{S'};\omega)]/[\mathrm{Vol}(\mathbf{X}_S)/\,\mathrm{Vol}(\mathbf{X}_{S'})]$ as follows,

$$\delta(\omega) = \underbrace{\frac{\mathrm{RV}(\mathbf{X}_S;\omega)}{\mathrm{Vol}(\mathbf{X}_S)}}_{\leq \exp(\beta^{-1}) \text{ by Lemma 6}} \times \underbrace{\frac{\mathrm{Vol}(\mathbf{X}_{S'})}{\mathrm{RV}(\mathbf{X}_{S'};\omega)}}_{\leq 1 \text{ by Proposition 8}} \leq \exp(\beta^{-1}). \tag{11}$$

The ≤ 1 is from the fact that replication-robustness valuation on an original matrix (without replication) is no smaller than the original volume (with two mild conditions that $\mathbf{X}$ contains sufficient diversity to *not* resemble a replicated dataset, and α is not too small, see Proposition 8, we empirically verify these assumptions by verifying the upper and lower bounds of $\delta(\omega)$ in Fig. 11), implying the RV will not reduce the information contained in the original matrix.

The lower bound of $(\exp(\beta^{-1}))^{-1}$ is by an argument by symmetry. Specifically, taking reciprocal on both sides of

$$\frac{\mathrm{RV}(\mathbf{X}_S;\omega)}{\mathrm{Vol}(\mathbf{X}_S)} \times \frac{\mathrm{Vol}(\mathbf{X}_{S'})}{\mathrm{RV}(\mathbf{X}_{S'};\omega)} \leq \exp(\beta^{-1})$$

gives

$$\frac{\mathrm{RV}(\mathbf{X}_{S'};\omega)}{\mathrm{Vol}(\mathbf{X}_{S'})} \times \frac{\mathrm{Vol}(\mathbf{X}_S)}{\mathrm{RV}(\mathbf{X}_S;\omega)} \geq (\exp(\beta^{-1}))^{-1}.$$

Since the reference on S, S' is arbitrary, we arrive at the lower bound by switching the indexing.

Furthermore, we verify the two conditions required to apply Proposition 8 by empirically verifying the distortion is bounded. We construct two equal-sized, independent and identically sampled d-dimensional matrices, $\mathbf{X}_S, \mathbf{X}_{S'}$ and plot the distortion over a range of ω. Note we have considered varied $d \in \{1, 2, 5, 10\}$ and two such distributions: d dimensional $\mathcal{N}(0, 1)$ or uniform distribution $U(0, 1)$. The result is in Fig. 11. The black dotted lines are the theoretical upper and lower bounds. Fig. 11 suggests the bound on distortion may be tighter, implying in practice the consistency in relative valuation is preserved. □

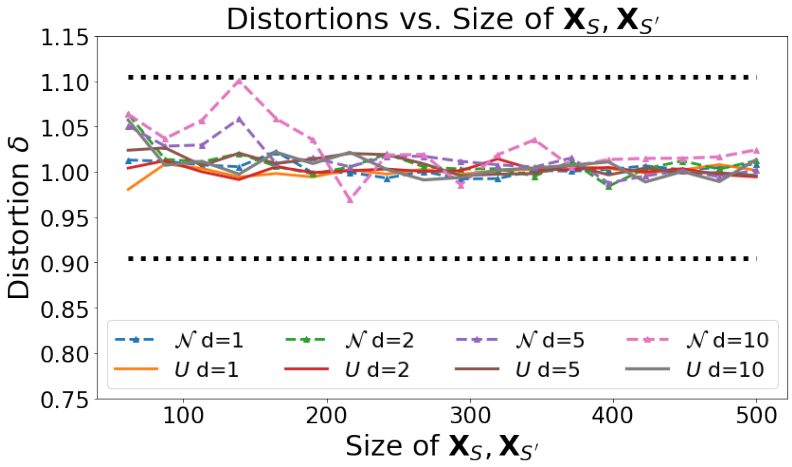

Figure 11: Distortion $\delta(\omega)$ vs. size of $\mathbf{X}_S, \mathbf{X}_{S'}$. $\mathbf{X}_S, \mathbf{X}_{S'}$ are equal-sized, independent and identically sampled from d-dimensional normal distribution $\mathcal{N}(0, 1)$ or uniform distribution $U(0, 1)$. Black dotted lines are $\exp(\beta^{-1})^{-1}, \exp(\beta^{-1})$ where $\beta = 10$. The discretization coefficient $\omega = 0.1$.

Proof of Proposition 6. By definition and simple rearranging, we can write

$$\gamma_{\text{RV}} \triangleq \frac{\text{RV}(\mathbf{X}; \omega)}{\sup_c \text{RV}(\text{replicate}(\mathbf{X}, c); \omega)} = \frac{\text{Vol}(\widetilde{\mathbf{X}})}{\text{Vol}(\widetilde{\mathbf{X}}_{\text{rep}})} \times \prod_{i \in \Psi} \frac{\rho_i}{\rho_i'}$$

where ρ_i' denotes the coefficient for the d-cube after replication. We calculate these two terms separately.

Consider $\boldsymbol{\mu}_i$ for some d-cube, and let $\boldsymbol{\mu}_i'$ denote the statistic after replication. We have $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i'$ because each data point in the d-cube is replicated for equal number of times (i.e., $\to \infty$) due to $\sup_c$. This implies $\widetilde{\mathbf{X}} = \widetilde{\mathbf{X}}_{\text{rep}}$ and $\text{Vol}(\widetilde{\mathbf{X}}) / \text{Vol}(\widetilde{\mathbf{X}}_{\text{rep}}) = 1$.

For a non-empty d-cube, $\rho_i \geq 1$ and $\rho_i' \leq 1/(1-\alpha)$ due to the geometric series. So we have $\frac{\rho_i}{\rho_i'} \geq (1-\alpha)$. Multiplying all $|\Psi|$ such ratios we have $\prod_{i \in \Psi} \frac{\rho_i}{\rho_i'} \geq (1-\alpha)^{|\Psi|}$. Combining this with the previous result completes the proof. □

Proof of Proposition 7. We outline the proof ideas and omit the tedious details.

1. Reduction of $\text{RV}(\cdot; \omega)$ to $\text{Vol}()$ requires careful tracing of the constants $K_{\widetilde{\mathbf{X}}, i}$ using Lemma 1. There may be different ways to set the values of $K_{\widetilde{\mathbf{X}}, i}$ for this equality to hold from different order of tracing. We provide one such construction.

   First, construct $\mathbf{A} = \widetilde{\mathbf{X}}$. We will use $\mathbf{A}$ to represent the intermediate construction of $\mathbf{X}$ for the calculation of $\text{Vol}()$. Also, set $K_{\widetilde{\mathbf{X}}, i} = 1, i \in \Psi$. This step fills up each non-empty d-cube with one data point.

   The second step exhausts the remaining data points $\mathbf{B}$, one distinct data point at a time. The remaining data points from the first step are all the data points excluding one data

point for each d-cube. We abuse notation a little to write $\mathbf{B} \coloneqq \mathbf{X} \setminus \mathbf{A} = \mathbf{X} \setminus \widetilde{\mathbf{X}}$. For each distinct $\mathbf{x} \in \mathbf{B}$, suppose it goes into the i-th d-cube and it has m copies in $\mathbf{B}$: update $K_{\widetilde{\mathbf{X}},i} \leftarrow (1+m\times\mathbf{x}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{x}^\top)^{1/2}$ and then update $\mathbf{A} \leftarrow [\mathbf{A}^\top \underbrace{\mathbf{x}^\top \dots \mathbf{x}^\top}_{m \text{ times}}]^\top$ by Lemma 1 and remove all copies of $\mathbf{x}$ from $\mathbf{B}$. Iterate until $\mathbf{B}$ is empty.

2. $\gamma_{\mathrm{RV}_1} = 1$ is from a direct application of Proposition 6.

□

Proposition 8 (Conditions for $\mathrm{RV}(\mathbf{X};\omega) \geq \mathrm{Vol}(\mathbf{X})$). *For a given ω and $\mathbf{X}$, $\mathrm{RV}(\mathbf{X};\omega) \geq \mathrm{Vol}(\mathbf{X})$ requires the following two conditions:*

1. *The original $\mathbf{X}$ contains sufficient diversity that it does not resemble replicated copies of data.*
2. *The α is not too small (so $\prod_{\omega_i} \rho_i$ will not be too small).*

Proof of Proposition 8. For a given ω, let the rows which occupy d-cubes alone be in $\mathbf{X}_A$ and the rest in $\mathbf{X}_B$. Each row in $\mathbf{X}_A$ occupies some d-cube by itself, while each row in $\mathbf{X}_B$ has to "share" a d-cube with another row in $\mathbf{X}_B$. With this, we rearrange and rewrite $\mathbf{X} = \begin{bmatrix}\mathbf{X}_A \\ \mathbf{X}_B\end{bmatrix}$ and we have

$$\mathrm{Vol}(\mathbf{X})^2 = |\mathbf{G}_A| \times |\mathbf{I} + \mathbf{G}_A^{-1}\mathbf{G}_B| \tag{12}$$

from Lemma 7 where $\mathbf{G}_A \coloneqq \mathbf{X}_A^\top\mathbf{X}_A, \mathbf{G}_B \coloneqq \mathbf{X}_B^\top\mathbf{X}_B$.

Let $\widetilde{\mathbf{X}}_B$ be the estimated $\mathbf{X}_B$ according to the d-cubes: for each d-cube containing more than one row, take the average of the rows in a d-cube, and put this as one row in $\widetilde{\mathbf{X}}_B$. With this, we write $\widetilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X}_A \\ \widetilde{\mathbf{X}}_B\end{bmatrix}$. Note $\widetilde{\mathbf{X}}$ is part of the calculation for $\mathrm{RV}(\mathbf{X};\omega)$.

We have

$$\mathrm{Vol}(\widetilde{\mathbf{X}})^2 = |\mathbf{G}_A| \times |\mathbf{I} + \mathbf{G}_A^{-1}\widetilde{\mathbf{G}}_B|$$

similarly as above where $\widetilde{\mathbf{G}}_B \coloneqq \widetilde{\mathbf{X}}_B^\top\widetilde{\mathbf{X}}_B$. Substituting this into the definition of RV gives

$$\mathrm{RV}(\mathbf{X}) = \mathrm{Vol}(\widetilde{\mathbf{X}}) \times (\prod_{i\in\Omega} \rho_i) = |\mathbf{G}_A| \times |\mathbf{I} + \mathbf{G}_A^{-1}\widetilde{\mathbf{G}}_B| \times (1+\epsilon) \tag{13}$$

Here, we write $\prod_{i\in\Omega} \rho_i = (1+\epsilon)$ where $\epsilon \geq 0$ is a constant which depends on α. A smaller α leads to a smaller ϵ.

We compare (12) and (13) as follows,

$$|\mathbf{G}_A| \times |\mathbf{I} + \mathbf{G}_A^{-1}\mathbf{G}_B| \leq |\mathbf{G}_A| \times |\mathbf{I} + \mathbf{G}_A^{-1}\widetilde{\mathbf{G}}_B| \times (1+\epsilon)$$
$$|\mathbf{I} + \mathbf{G}_A^{-1}\mathbf{G}_B| \leq |\mathbf{I} + \mathbf{G}_A^{-1}\widetilde{\mathbf{G}}_B| \times (1+\epsilon).$$

While we want to find the conditions under which above inequality is satisfied, it may be more intuitive to consider the conditions which can dissatisfy it. So we consider

$$|\mathbf{I} + \mathbf{G}_A^{-1}\mathbf{G}_B| > |\mathbf{I} + \mathbf{G}_A^{-1}\mathbf{G}_{\hat{B}}| \times (1+\epsilon).$$

Due to the previous discussion in Lemma 7, we know $|\mathbf{I} + \mathbf{G}_A^{-1}\mathbf{G}_B|$ is large if $\mathbf{X}_A$ covers little relative to $\mathbf{X}_B$, similarly for $|\mathbf{I} + \mathbf{G}_A^{-1}\widetilde{\mathbf{G}}_B|$. Therefore, we are left to find out the conditions where $\mathbf{X}_A$ covers little relative to $\mathbf{X}_B$ but $\mathbf{X}_A$ covers a lot relative to $\widetilde{\mathbf{X}}_B$ as this will lead to the l.h.s to be larger than r.h.s. The extreme case of $\mathbf{X}_A$ is empty while $\mathbf{X}_B$ contains multiple copies of the same datum, fits the described scenario very well.

In summary, under two conditions: 1) the original $\mathbf{X}$ contains sufficient diversity that it does not resemble replicated copies of data; 2) the α is not too small (so $\prod_{\omega_i} \rho_i$ will not be too small); then we have $\mathrm{RV}(\mathbf{X};\omega) \geq \mathrm{Vol}(\mathbf{X})$. The first condition is intuitive. The second one on α can be understood

as a trade-off between enforcing robustness (smaller α) vs. representing data (larger α). In practice, we should set α to achieve the desired robustness guarantee but not extremely small as it would have an over-correcting effect of mistakenly reducing the value of an honest dataset.

□

B Additional Experimental Results

B.1 Dataset License

Credit Card [2]: Database Contents License (DbCL); Uber & Lyft [5]: CC0 1.0 Universal (CC0 1.0); Used Car [1]: CC0 1.0 Universal (CC0 1.0); Hotel Reviews [4]: Attribution-NonCommercial 4.0 International (CC BY-NC 4.0); CaliH [20]: CC0 1.0 Universal (CC0 1.0); KingH [3]: CC0 1.0 Universal (CC0 1.0); USCensus [6]: CC0 1.0 Universal (CC0 1.0); The FaceA dataset [41]: non-commercial research purposes only.

B.2 Simulation for Replication Experiments on Volume and RV

We illustrate that the inflation of robust volumes (RV) can be controlled. We generate a dataset with 200 i.i.d. samples uniformly drawn from the synthetic 6D Friedman [13] function. As shown in Fig. 12, the volume of the dataset explodes exponentially with the number of full dataset replications. On the contrary, robust volume controls the volume explosion through α and the resultant RV stays almost constant with replications. Similar behavior is observed when we randomly select data rows to replicate instead.

In a more realistic adversarial replication setting, random noises could be injected into the replicated data rows. Interestingly, RV remains robust when the magnitude of the injected noise is small relative to ω, further demonstrating RV's practical utility. Specifically in this experimental setting when ω is set to 0.1, the RV is kept almost constant when the random Gaussian noise has $\sigma = 0.01$ (see pink dashed line in Fig. 12). On the other hand, RV does not inflate as much as Vol when σ is increased to 0.03 (compare gray and red lines). However, we observe if σ becomes too large, the robustness degrades. This is because when the magnitude of injected noise is larger than the actual data, it effectively becomes "new" but noisy data. Consequently, it is still practically challenging when the Gaussian noise is large. As large noise could dilute the data row's original information, it is difficult to distinguish whether the row is replicated (with noise).

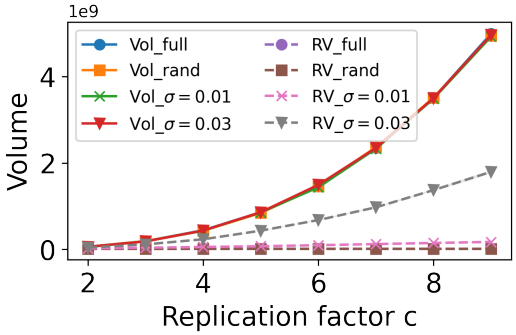

Figure 12: Vol & RV vs. replication. `full`, `rand` and σ denote three types of replications, where `full` (*resp.* `rand`) replicates all data (*resp.* random rows) and σ denotes noisy replication with noise $\sim \mathcal{N}(0, \sigma^2)$. Here, $\omega = 0.1$.

B.3 Selection of the Discretization Coefficient

Throughout the work, we set $\omega = 0.1$ for standardized features, including for real-world datasets which may contain unknown noise in labels. Through synthetic experiments, we show empirically that $\omega \in [0, 0.5]$ is a suitable range for standardized features when noise is small. Intuitively, feature standardization "squashes" most of the data to a relatively small range. For instance, the $[-2, 2]$

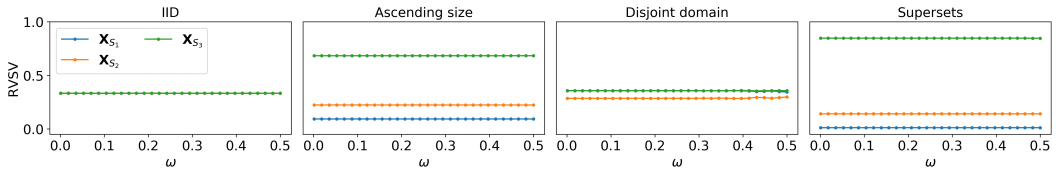

Figure 13: The choice of ω does not change values much under (A) i.i.d., (B) ascending dataset size, (C) disjoint input domain and (D) supersets. Setting $\omega > 0.5$ is not recommended as it results in an overly "compressed" $\widetilde{\mathbf{X}}$.

range is the 95% confidence interval for standardized features. As illustrated in Fig. 13, the relative RVSV does not vary much for any $\omega < 0.5$ under all 4 settings: (A) i.i.d., (B) ascending dataset sizes, (C) disjoint input domain and (D) supersets.

Note that in the disjoint input domain setting, $\mathbf{X}_{S_1}$ and $\mathbf{X}_{S_3}$ are valued more because of feature standardization. The standardized features have the following interpretation: values that are very positive or negative are statistically rare, and thus are more valuable. In contrast, values which are close to the mean (a value of 0) are statistically common, and thus less valuable. The implication is less common data are given higher values. We find $\omega = 0.1$ suitable under standard scaling and in practice it may be adjusted based on the prior on the amount of noise expected in the feature/input, i.e., larger noise requires a larger ω.

B.4 Robust Volume and Learning Performance

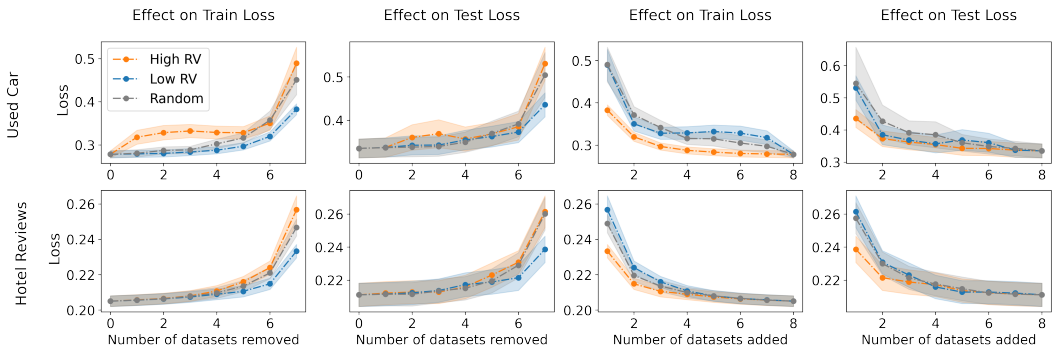

Figure 14: The effect of removing/adding the dataset with the highest/lowest RV on the train/test loss for two additional real-world datasets, the UK used car and the hotel reviews. The plots show the average and standard errors over 50 random trials.

To further verify that a larger RV leads to a better learning performance, we present results on two additional real-world datasets, the UK used car dataset [1] (i.e., car price prediction) and the Trip Advisor hotel reviews dataset [4] (i.e., numerical rating prediction). The two datasets are pre-processed through feature selection or neural networks to contain 5 and 8 standardized features respectively. Other experimental setups follow that of Sec. 5.1. The results are in Fig. 14. We observe a consistent general trend, adding (*resp.* removing) a dataset with high RV leads to lower (*resp.* higher) train and test loss. These results (along with the previous results) suggest that RV is a good indicator for the learning performance on the data without requiring validation, and is thus a good data valuation method.

B.5 Overlap of Input Domains

To extend the "disjoint input domain" case discussed in Sec. 5.2, we set the input range of $\mathbf{X}_{S_1}, \mathbf{X}_{S_2}$ to be $[0, 0.5 + z]^6$ and $\mathbf{X}_{S_3}$ to be $[0.5, 1]^6$, where z is the amount of domain overlap. To interpret, a larger z implies a less "unique" dataset $\mathbf{X}_{S_3}$ and thus less value for $\mathbf{X}_{S_3}$. We show in Fig. 15 that all methods including RVSV observe the correct trend, except that VLSV is still dominated by $\nu(\emptyset)$

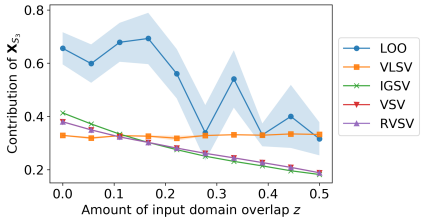

Figure 15: The effect of an increasing amount of input domain overlap on the valuation of $\mathbf{X}_{S_3}$, across different valuation methods. The values are averaged over 10 trails.

discussed in Sec. 5.2. Interestingly, we also observe a similar rate of decrease of the relative value of $\mathbf{X}_{S_3}$ for IGSV, VSV and RVSV.

B.6 Replication Experiments

Experimental Settings. For CaliH and KingH, we train a neural network (NN) with 2 hidden layers consisting of 64 and 10 hidden units, respectively. For USCensus, we train an NN with 2 hidden layers consisting of 128 and 16 hidden units, respectively. For FaceA, we train a convolutional NN with 3 convolutional layers (the first two of which each followed by 2-dimensional batch normalization and max pooling), followed by 3 fully connected layers consisting of 1024, 64, and 10 hidden units respectively. The activation function used is the rectified linear unit (ReLU). Consequently, the features after the last hidden layer may contain completely zeros, so we remove features that contain only zero. Alternatively, leaky ReLU can be used instead to prevent this issue.

Additional Results. The results on four datasets (four rows) CaliH, KingH, USCensus and FaceA for the two non-i.i.d. distributions: *supersets* and *disjoint* are in Fig. 16. We can make these observations from the results: LOO's behavior is unstable, in particular for supersets ratio of 0.1 (leftmost column) where 10% of $\mathbf{X}_{S_1}$ is contained in $\mathbf{X}_{S_3}$. VSV and IGSV both increase quite quickly for $c \leq 20$. RVSV and VLSV are consistent regardless of c. The tabulated SV similarity results for KingH, USCensus and FaceA are in Tables 2, 3 and 4. We find that RVSV generally performs the best, in terms of similarity in relative valuations to the validation-based VLSV.

Table 2: Similarity with VLSV under replication for **KingH**. Bold values indicate 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	0.693	0.215	0.381	-0.008	0.483	0.964	**0.704**	0.752	1.971	-0.114	0.051	0.091	**0.883**	0.524	1.068
IGSV	-0.898	0.639	1.569	-0.937	0.637	1.547	-0.998	0.629	1.520	-0.962	0.641	1.597	0.592	0.660	1.723
VSV	-0.902	0.741	2.055	-0.932	0.743	2.084	-0.999	0.727	1.954	-0.963	0.750	2.140	0.547	0.782	2.415
RVSV-005	**0.819**	**0.985**	**9.890**	**0.283**	**0.998**	**28.557**	-0.908	**0.998**	**26.040**	**0.668**	**0.980**	**8.425**	-0.978	**0.940**	**4.757**
RVSV-01	0.817	0.977	7.965	-0.000	0.996	19.873	-0.874	0.995	16.779	0.662	0.962	6.109	-0.977	0.892	3.416

Table 3: Similarity with VLSV under replication for **USCensus**. Bold values indicate best results. The disjoint ratio 0 for USCensus dataset leads to NaN values for VSV so we show from disjoint ratio 0.2.

Method	i.i.d.			disjoint 0.2			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	**0.891**	0.816	2.493	-0.127	0.517	1.632	-0.950	0.164	0.290	**0.845**	0.211	0.396	**0.682**	0.602	1.344
IGSV	0.302	0.640	1.582	0.589	0.640	1.579	-0.999	0.641	1.599	0.475	0.640	1.596	-0.022	0.658	1.713
VSV	0.292	0.739	2.064	0.606	0.739	2.052	-0.998	0.739	2.073	0.482	0.740	2.061	-0.099	0.779	2.368
RVSV-005	-0.975	**0.963**	**6.223**	**0.758**	**0.998**	**30.976**	**0.687**	**0.999**	**46.907**	-0.999	**0.974**	**7.456**	-0.846	**0.915**	**3.920**
RVSV-01	-0.976	0.958	5.782	0.756	0.997	24.403	0.602	0.998	28.317	-0.999	0.973	7.257	-0.847	0.908	3.753

Table 4: Similarity with VLSV under replication for **FaceA**. Bold values indicate 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	-0.080	0.558	1.165	**0.823**	0.709	1.854	0.903	0.690	1.662	-1.000	0.948	5.157	**0.693**	**0.946**	**5.084**
IGSV	0.019	0.732	2.088	0.229	0.733	2.100	-0.898	0.731	2.096	-0.838	0.735	2.137	0.632	0.760	2.347
VSV	0.013	0.701	1.807	0.226	0.706	1.847	-0.894	0.702	1.816	-0.836	0.709	1.869	0.603	0.745	2.095
RVSV-005	-0.867	**0.941**	**4.828**	0.340	**1.000**	**69.502**	**0.923**	**1.000**	**100.947**	**0.869**	**0.943**	**4.909**	-0.957	0.880	3.209
RVSV-01	-0.877	0.884	3.275	0.140	0.997	20.856	0.921	0.998	30.487	0.859	0.915	3.926	-0.953	0.830	2.586

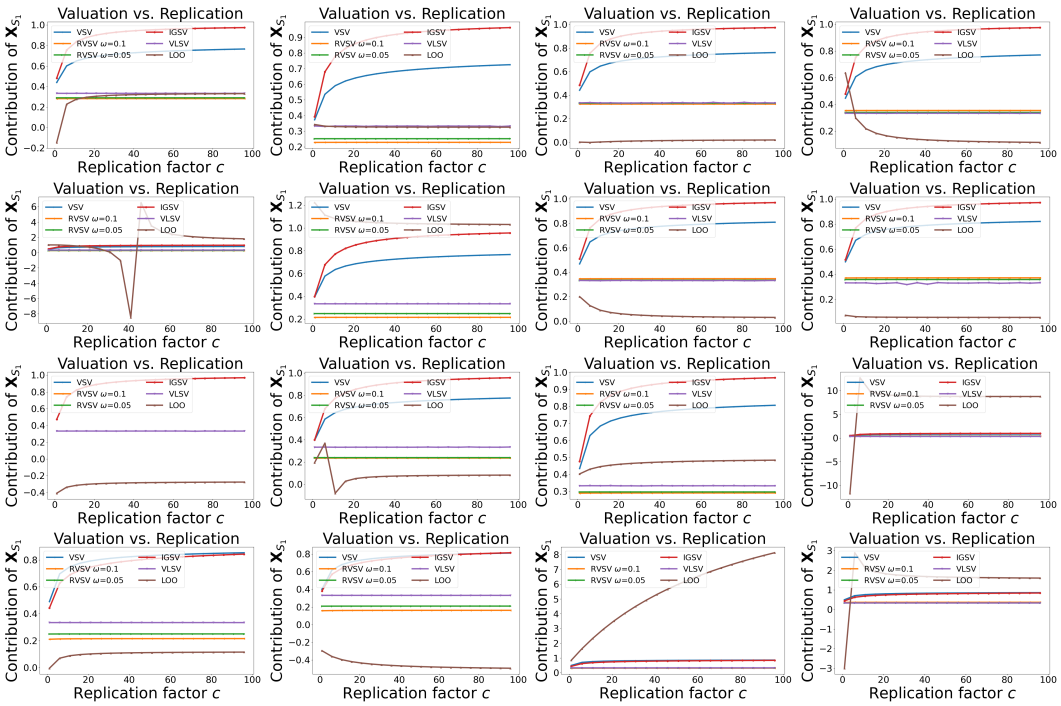

Figure 16: Valuations for the replicated dataset $\mathbf{X}_{S_1}$ under two data distributions: *supersets* of ratios 0.1 & 1 (left two) and *disjoint* of ratios 0 & 1 (right two). The vertical axis shows $\mathbf{X}_{S_1}$'s value. The horizontal axis shows the replication factor c. From first row to last row: CaliH, KingH, USCensue, FaceA. Note The disjoint ratio 0 for USCensus dataset leads to NaN values for VSV so we show from disjoint ratio 0.2.

B.7 IGSV vs. RVSV

Effect of hyperparameters on the SV. The setting of the experiments is consistent as described previously, including data distributions and models used. In IGSV, a crucial hyperparameter is the user-specified prior σ in the covariance, namely assuming the random variables of interest follow $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ [35]. In this case, the random variables are the parameters of the linear regressor. Here we vary the $\sigma \in \{0.0001, 0.001, 0.01, 0.1, 1\}$. For RVSV, we vary the $\omega \in \{0.01, 0.05, 0.1, 0.2, 0.25\}$ and calculate the $\log()$ of RV for numerical stability and due to Lemma 7 below which sheds some light on the similarity between RV and IG. The results on four datasets CaliH, KingH, USCensus and FaceA for the two non-i.i.d. distributions: *supersets* and *disjoint* are in Fig. 17.

We make two observations. 1): For IGSV, different priors lead to different SV with the same $\mathbf{X}_{S_1}, \mathbf{X}_{S_2}, \mathbf{X}_{S_3}$. The data providers will thus want to use the prior which leads to their SV being the highest, and may result in disagreement. For RVSV, all the experimented ω values coincide with the same SV, avoiding this potential selection over ω. 2): $\ln()$ function is plotted as a growth rate reference. While in some cases IGSV grows slower than $\ln()$, it does not converge, implying that a data provider can always have non-zero additional gain with more replication, especially for $c \leq 20$. Confirming the result of our previous replication experiments that IGSV may be less robust. In contrast, RVSG stays consistent regardless of c.

An interesting connection from volume to information gain. IGSV leverages the information gain criterion, or alternatively the conditional entropy criterion. The intuitive interpretation is given $\mathbf{X}_{S_1}, \mathbf{X}_{S_2}$ is valuable if $\mathbf{X}_{S_2}$ provides additional and new information that is *not* captured by $\mathbf{X}_{S_1}$. Interestingly, $\log \mathrm{Vol}()$ offers an echoing interpretation via Lemma 7. The intuition is also similar, given $\mathbf{X}_{S_1}, \mathbf{X}_{S_2}$ is valuable if $\mathbf{X}_{S_2}$ "occupies" additional space that is *not* "occupied" by $\mathbf{X}_{S_1}$.

Furthermore, this similarity inspired us to relate to the duality of *maximum entropy sampling* [34], which gives rise to the following practical implication: a greedy iterative approach to maximize the

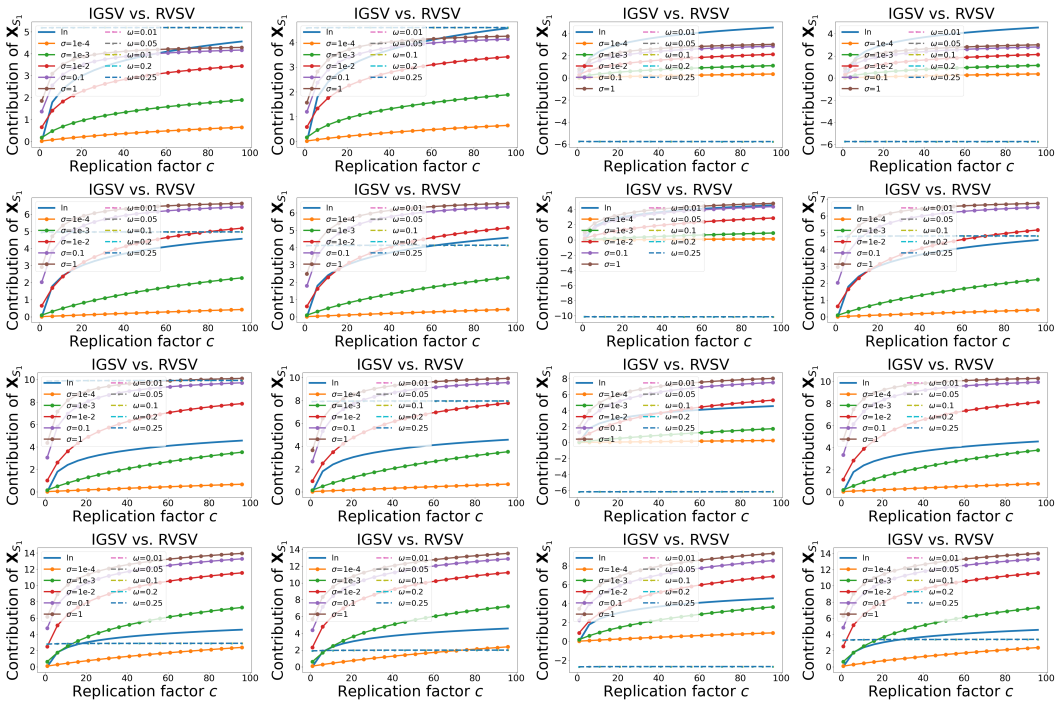

Figure 17: IGSV (solid lines with dot) and RVSV (dashed lines) vs. replication factor c. σ denotes the prior on the standard deviation for IGSV. ω is the discretization for RVSV. Valuations for the replicated dataset $\mathbf{X}_{S_1}$ under two data distributions:*supersets* of ratios 0.1 & 1 (left two) and *disjoint* of ratios 0 & 1 (right two). The y-axis shows $\mathbf{X}_{S_1}$'s value. The x-axis shows the replication factor c. The first to last row: CaliH, KingH, USCensus, FaceA.

$\log \mathrm{Vol}()$ in data collection/purchase gives a near-optimal solution ($(1-1/e)$-approximation) in terms maximizing the volume. See detailed discussion below.

Lemma 7 (**Duality of Volume Decomposition**). *For full-rank $\mathbf{X}_S, \mathbf{X}_{S'}$ of $\mathbf{X}$, we have*

$$\log V_{S\cup S'} = \log V_S + 0.5 \times \log V_{S|S'}$$

where $\log V_{S|S'} \coloneqq \log|\mathbf{I} + \mathbf{G}_S^{-1}\mathbf{G}_{S'}|$.

Proof of Lemma 7.

$$\begin{aligned} V_{S\cup S'}^2 &\coloneqq |\mathbf{X}^\top\mathbf{X}| = |\mathbf{G}_S + \mathbf{G}_{S'}| = |\mathbf{G}_S \times (\mathbf{I} + \mathbf{G}_S^{-1}\mathbf{G}_{S'})| \\ &= |\mathbf{G}_S| \times |\mathbf{I} + \mathbf{G}_S^{-1}\mathbf{G}_{S'}| = V_S^2|\mathbf{I} + \mathbf{G}_S^{-1}\mathbf{G}_{S'}| \quad \text{Taking log on both sides} \\ 2\log(V_{S\cup S'}) &= 2\log(V_S) + \log(|\mathbf{I} + \mathbf{G}_S^{-1}\mathbf{G}_{S'}|) \\ \log(V_{S\cup S'}) &= \log(V_S) + 0.5 \times \log(|\mathbf{I} + \mathbf{G}_S^{-1}\mathbf{G}_{S'}|) \end{aligned}$$

□

We explicitly write $V = V_{S\cup S'}$ and define a notation $V_{S|S'}$ to better illustrate the similarity to the duality in *maximum entropy sampling* where maximizing the entropy of the selected set ($\mathbb{H}[\mathbf{f}_\mathcal{O}]$) minimizes the conditional entropy ($\mathbb{H}[\mathbf{f}_{\mathcal{X}\setminus\mathcal{O}}|\mathbf{f}_\mathcal{O}]$) on the remaining set as follows,

$$\mathbb{H}[\mathbf{f}_\mathcal{X}] = \mathbb{H}[\mathbf{f}_\mathcal{O}] + \mathbb{H}[\mathbf{f}_{\mathcal{X}\setminus\mathcal{O}}|\mathbf{f}_\mathcal{O}]$$

where $\mathcal{X}$ denotes the input space for the input locations to be observed, and $\mathcal{O} \subseteq \mathcal{X}$ is a selected subset of the input locations and $\mathbf{f}()$ is the unknown function of interest. $\mathbb{H}(\cdot)$ is the standard differential entropy.

V_S depends on the relation between $\mathbf{X}_S$ and $\mathbf{X}_{S'}$ in how well $\mathbf{X}_S$ covers the space relative to $\mathbf{X}_{S'}$ as reflected in $|\mathbf{I} + \mathbf{G}_{S'}\mathbf{G}_S^{-1}|$. V_S is large if $\mathbf{X}_S$ covers the remaining subset $\mathbf{X}_{S'}$ (implying $V_{S'}$ will be

small). This is because $V_{S\cup S'}$ is a constant. Similarly, in MES, if $\mathbb{H}[\mathbf{f}_{\mathcal{O}}]$ is large, then $\mathbb{H}[\mathbf{f}_{\mathcal{X}\setminus\mathcal{O}}|\mathbf{f}_{\mathcal{O}}]$ will be small, because $\mathbb{H}[\mathbf{f}_{\mathcal{X}}]$ is a constant.

A practical implication is thus: suppose a data provider already knows what data can be collected (denoted by $S \cup S'$ to be consistent with previous discussion), and wants to maximize the value of collected data (denoted by S) under a constrained budget (e.g. time, memory, and other processing costs), an iterative greedy approach to update S, S' as follows:

$$\begin{aligned}\underset{\mathbf{x}_i}{\operatorname{argmax}}\ \log V_S - \log V_{S|S'}\\ S \leftarrow S \cup \{\mathbf{x}_i\}\end{aligned}$$

yields an $(1-1/e)$-approximation [22] if $\log\mathrm{Vol}()$ as in our definition is submodular. Note our proof is different from the examples [26] which define the squared volume to be the determinant of the (right) Gram matrix. That definition admits a simpler proof technique by a geometric argument that is not applicable to our definition of volume.

First, we recall the definition of submodularity. Let $[n]$ denote the set $\{1,\ldots,n\}$ as the indices of the rows/data points in $\mathbf{X}$, there are thus 2^n possible subsets of $[n]$. A set function $g: 2^n \mapsto \mathbb{R}$ is called *submodular* if for any $S \subseteq S^+ \subseteq [n]$ and $\forall \mathbf{x} \in \mathbf{X}$ ($\mathbf{x}$ is a data point in $\mathbf{X}$),

$$\underbrace{g(S\cup\{\mathbf{x}\}) - g(S)}_{\Delta_S} \geq \underbrace{g(S^+\cup\{\mathbf{x}\}) - g(S^+)}_{\Delta_{S^+}}.$$

Proposition 9 (Submodularity of $\log\mathrm{Vol}()$)**.** *If for any $S \subseteq S^+ \subseteq [n]$, $\mathbf{G}^+ - \mathbf{G}$ is positive semi-definite where $\mathbf{G} \coloneqq \mathbf{X}_S^\top\mathbf{X}_S$ and $\mathbf{G}^+ \coloneqq \mathbf{X}_{S^+}^\top\mathbf{X}_{S^+}$, then $\log\mathrm{Vol}()$ is submodular.*

Proof. By letting $g() = \log\mathrm{Vol}()$ and a direct application of Lemma 1 , we have: $\Delta_S = 1/2 \times \log(1+\mathbf{x}(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}^\top)$ and $\Delta_{S^+} = 1/2\times\log(1+\mathbf{x}(\mathbf{X}_{S^+}^\top\mathbf{X}_{S^+})^{-1}\mathbf{x}^\top)$. We want to show:

$$\begin{aligned}\Delta_S &\geq \Delta_{S^+}\\ 1/2\times\log(1+\mathbf{x}(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}^\top) &\geq 1/2\times\log(1+\mathbf{x}(\mathbf{X}_{S^+}^\top\mathbf{X}_{S^+})^{-1}\mathbf{x}^\top)\\ \mathbf{x}(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}^\top &\geq \mathbf{x}(\mathbf{X}_{S^+}^\top\mathbf{X}_{S^+})^{-1}\mathbf{x}^\top\\ \mathbf{x}\left[(\mathbf{X}_S^\top\mathbf{X}_S)^{-1} - (\mathbf{X}_{S^+}^\top\mathbf{X}_{S^+})^{-1}\right]\mathbf{x}^\top &\geq 0.\end{aligned}$$

Next, by a direct application of Lemma 8 below: substituting $\mathbf{A} = \mathbf{X}_S^\top\mathbf{X}_S, \mathbf{B} = \mathbf{X}_{S^+}^\top\mathbf{X}_{S^+}$, we can show $(\mathbf{X}_S^\top\mathbf{X}_S)^{-1} - (\mathbf{X}_{S^+}^\top\mathbf{X}_{S^+})^{-1}$ is positive semi-definite and the proof is complete. □

The assumption $\mathbf{G}^+ - \mathbf{G}$ is positive semi-definite, has the interpretation that the left Gram of a larger dataset ($\mathbf{X}_{S^+}$) has "more" information (the entire result is verified empirically after the proof). Fig. 18 shows the proportion of $\log\mathrm{Vol}()$ is submodular over 500 independent trials of randomly sampled matrices $\mathbf{X}_S$ (and subsequently constructed $\mathbf{X}_{S^+}$). We observe $\log\mathrm{Vol}()$ is almost always submodular. The exceptions may be attributed to that randomly drawn matrices may sometimes violate the condition of $\mathbf{G}, \mathbf{G}^+$ being positive definite required by Lemma 1, or equivalently $\mathbf{X}_S$ may not be full-rank if n is small relative to d.

While the lemma below is with respect to matrices, if we consider the scalar version, it is much more intuitive: for two positive scalars $a, b > 0$, $a \geq b \implies 1/b \geq 1/a$. The result essentially generalizes this idea to symmetric and positive definite matrices. We will use $\mathbf{A} \succeq \mathbf{0}$ to denote $\mathbf{A}$ is positive semi-definite.

Lemma 8. *Given $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ are both symmetric and positive definite, then*

$$\mathbf{B} - \mathbf{A} \succeq \mathbf{0} \implies \mathbf{A}^{-1} - \mathbf{B}^{-1} \succeq \mathbf{0}$$

Proof.

$$\begin{aligned}&\mathbf{B} - \mathbf{A} \succeq \mathbf{0}\\ \implies &\mathbf{A}^{-1/2}(\mathbf{B}-\mathbf{A})\mathbf{A}^{-1/2} \succeq \mathbf{0}\\ \implies &\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} \succeq \mathbf{I}\\ \implies &\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2} \preceq \mathbf{I},\end{aligned}$$

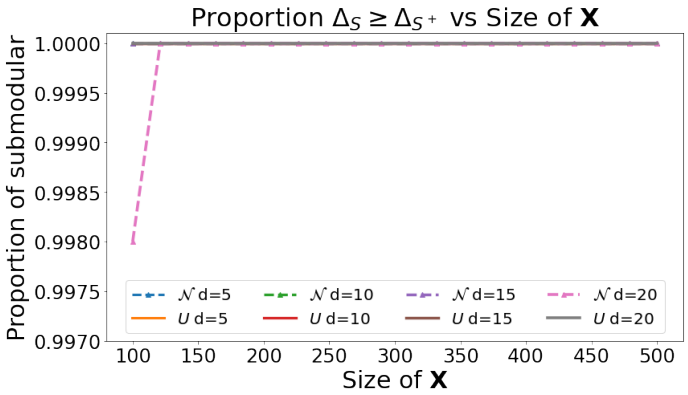

Figure 18: The proportion of $\log \mathrm{Vol}()$ is submodular with respect to randomly drawn $\mathbf{X}_S, \mathbf{X}_{S^+}, S \subset S^+$. d denotes feature dimension. $\mathcal{N}$ denotes $\mathbf{X}$ is sampled from the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and U denotes $\mathbf{X}$ is sampled from the uniform distribution $U(0, 1)$. $\mathbf{X}_S$ is then a randomly sampled submatrix of $\mathbf{X}$, containing $0.2 \times n$ number of data points of $\mathbf{X}$. $\mathbf{X}_S^+$ is constructed by appending a randomly select data point from $\mathbf{X}$ to $\mathbf{X}_S$. The proportion is calculated over 500 independent random trials. We observe that $\log \mathrm{Vol}()$ is almost always submodular, except in some degenerate cases where the randomly drawn $\mathbf{X}_S$ is not full-rank.

therefore,

$$\begin{aligned}\mathbf{B}^{-1} &= \mathbf{A}^{-1/2}\mathbf{A}^{1/2}\mathbf{B}^{-1}\mathbf{A}^{1/2}\mathbf{A}^{-1/2} \\ &\preceq \mathbf{A}^{-1/2}\mathbf{I}\mathbf{A}^{-1/2} \\ &\preceq \mathbf{A}^{-1}.\end{aligned}$$

The first implication is by using the fact that $\mathbf{B} - \mathbf{A} \succeq \mathbf{0} \implies \mathbf{C}^\top\mathbf{A}\mathbf{C} \preceq \mathbf{C}^\top\mathbf{B}\mathbf{C}$ for any conformable matrix $\mathbf{C}$ and viewing $\mathbf{A}^{-1/2}$ as a conformable matrix. The second implication is by considering the relationship between a symmetric and positive definite matrix $\mathbf{B}$ and $\mathbf{I}$, where $\mathbf{I} \preceq \mathbf{B} \implies \mathbf{B}^{-1} \preceq \mathbf{I}$. The following steps are substitutions and applications of the definition of $\succeq$. □