Eigen-Value: Efficient Domain-Robust Data Valuation via Eigenvalue-Based Approach

Supplementary Material

A. Theoretical Analysis

This section provides the theoretical proof of EV, as detailed in Section 4 Efficient Domain Robust Data Valuation.

A.1. Estimating Domain Discrepancy using Eigenvalue Shifts Induced by NCE

Normalized Cross-Entropy (NCE) is defined as

$$\text{NCE}(\theta) = \frac{-\sum_{k=1}^{K} q(y=k|x)\log p_\theta(k|x)}{-\sum_{j=1}^{K}\sum_{k=1}^{K} q(y=j|x)\log p_\theta(k|x)}, \quad (7)$$

with $0 \le \text{NCE}(\theta) \le 1$.
Since model parameter θ is trained on in-distribution (ID) data, it is assumed that NCE on OOD data (NCE_{OOD}) is larger than NCE on ID data (NCE_{ID})

$$0 < \text{NCE}_{\text{ID}}(\theta) \le \text{NCE}_{\text{OOD}}(\theta) \le 1.$$

Using the above relation, the domain discrepancy between OOD and ID can be defined as

$$\begin{aligned}\Gamma(\mathcal{D}_{\text{OOD}}, \mathcal{D}_{\text{ID}}) &= \sup_\theta \Big(\text{NCE}_{\text{OOD}}(\theta) - \text{NCE}_{\text{ID}}(\theta)\Big) \\ &\le \sup_\theta \frac{\text{NCE}_{\text{OOD}}(\theta)}{\text{NCE}_{\text{ID}}(\theta)}.\end{aligned} \quad (8)$$

Assuming an optimal model θ_0 for both domains, a Taylor expansion around θ_0 yields

$$\begin{aligned}\text{NCE}(\theta) \approx \text{NCE}(\theta_0) &+ (\theta-\theta_0)^\top \nabla_\theta \text{NCE}(\theta_0) \\ &+ \frac{1}{2}(\theta-\theta_0)^\top \nabla_\theta^2 \text{NCE}(\theta_0)(\theta-\theta_0).\end{aligned} \quad (9)$$

Since $\text{NCE}(\theta_0) \approx 0$ and $\nabla_\theta \text{NCE}(\theta_0) \approx 0$, it follows that

$$\text{NCE}(\theta) \approx \frac{1}{2}(\theta-\theta_0)^\top H (\theta-\theta_0),$$

where $H := \nabla_\theta^2 \text{NCE}(\theta_0)$.

Thus, the ratio can be approximated by the Hessian of each distribution ($H_{\text{OOD}}, H_{\text{ID}}$)

$$\sup_\theta \frac{\text{NCE}_{\text{OOD}}(\theta)}{\text{NCE}_{\text{ID}}(\theta)} \approx \sup_\theta \frac{\frac{1}{2}(\theta-\theta_0)^\top H_{\text{OOD}}(\theta-\theta_0)}{\frac{1}{2}(\theta-\theta_0)^\top H_{\text{ID}}(\theta-\theta_0)}. \quad (10)$$

Using the Rayleigh quotient property, for any nonzero vector $v \in \mathbb{R}^d$

$$\lambda_{\min}(H) \le \frac{v^\top H\, v}{v^\top v} \le \lambda_{\max}(H).$$

Then, under the assumption that the Hessian is positive semi-definite, we approximate the ratio of NCE between distributions using the ratio of their maximum ($\lambda_{\max}$) and minimum ($\lambda_{\min}$) eigenvalues.

$$\frac{\text{NCE}_{\text{OOD}}(\theta)}{\text{NCE}_{\text{ID}}(\theta)} \le \frac{\lambda_{\max}(H_{\text{OOD}})}{\lambda_{\min}(H_{\text{ID}})}. \quad (11)$$

where $\lambda_{\min}(H_{\text{ID}}) > 0$.

A.2. Using Logistic Regression Hessian as a Covariance Approximation

In logistic regression, the negative log-likelihood is given by

$$-\ell(\theta) = -\sum_{i=1}^{n}\Big[y_i \log\sigma(\theta^\top x_i) + (1-y_i)\log(1-\sigma(\theta^\top x_i))\Big].$$

Thus, the Hessian of the NCE (a variant of logistic regression) is upper bounded by the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$

$$\begin{aligned}H &= \sum_{i=1}^{n}\sigma(\theta_0^\top x_i)\big(1-\sigma(\theta_0^\top x_i)\big)x_i x_i^\top \\ &\le \frac{1}{4}\sum_{i=1}^{n} x_i x_i^\top = \frac{n}{4}\Sigma,\end{aligned} \quad (12)$$

since $\sigma(\theta_0^\top x_i)(1-\sigma(\theta_0^\top x_i)) \le \frac{1}{4}$.

Thus, it follows that

$$\frac{\text{NCE}_{\text{OOD}}}{\text{NCE}_{\text{ID}}} \le \frac{\lambda_{\max}(H_{\text{OOD}})}{\lambda_{\min}(H_{\text{ID}})} \le \frac{\lambda_{\max}(\Sigma_{\text{OOD}})}{\lambda_{\min}(\Sigma_{\text{ID}})}, \quad (13)$$

where Σ_{ID} and Σ_{OOD} are covariance matrices of data from ID and OOD, respectively.

With the eigendecomposition $\Sigma = Q\Lambda Q^\top$, the Frobenius norm is given by

$$\|\Sigma\|_F^2 = \text{tr}(\Sigma\Sigma^\top) = \text{tr}(\Lambda\Lambda^\top) = \sum(\text{eigenvalues})^2. \quad (14)$$

In addition, we have the inequality

$$\sqrt{\lambda_{\max}^2(\Sigma)} \le \|\Sigma\|_F \le \sqrt{\text{rank}(\Sigma)\cdot\lambda_{\max}^2(\Sigma)}. \quad (15)$$

We assume that the ID and OOD covariance matrices satisfy the matching marginal condition, which means that

the two distributions have identical marginal variances. In other words, the diagonal elements of their covariance matrices are the same, although the off-diagonal entries may differ. This condition preserves the variances of individual features across domains while allowing feature correlations to vary. In our study, this assumption is reasonable because we use normalized embeddings, which naturally align marginal variances. We empirically validate this condition on real datasets in Appendix C.1. This condition is formalized as:

$$\Sigma_{\text{OOD}} = \Sigma_{\text{ID}} + E,$$

where $E \in \mathbb{R}^{d\times d}$ is a matrix that captures domain-specific differences. By assumption, E has zero diagonal entries and non-zero off-diagonal entries, meaning it only affects feature correlations while preserving individual feature variances. By the triangle inequality,

$$\|\Sigma_{\text{OOD}}\|_F \leq \|\Sigma_{\text{ID}}\|_F + \|E\|_F,$$

and if $\|E\|_F \leq \sqrt{d^2-d}$ (with $|E_{ij}| \leq 1$ for $i \neq j$), one can bound the maximum singular value of Σ_{OOD}. This leads to the bound with $\mathcal{L}_{\text{OOD}}$ and $\mathcal{L}_{\text{ID}}$, which are losses of θ on OOD data and ID data, respectively.

$$\mathcal{L}_{\text{OOD}} \leq \mathcal{L}_{\text{ID}} + \frac{\lambda_{\max}(\Sigma_{\text{ID}}) \times \sqrt{d} + \sqrt{d^2-d}}{\lambda_{\min}(\Sigma_{\text{ID}})}. \tag{16}$$

A.3. Approximating Marginal Contributions of the Eigenvalue Term

Problem Statement: How can we use perturbation to compute the marginal value of a data point?

Given a normalized embedding dataset $\{x_1, x_2, \ldots, x_n\}$ of ID, the covariance matrix is defined as

$$\Sigma_{\text{ID}} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top. \tag{17}$$

When one data point x_k is removed, the new covariance matrix becomes

$$\Sigma_{-k} = \frac{1}{n-1}\sum_{i\neq k} x_i x_i^\top = \frac{n}{n-1}(\Sigma_{\text{ID}} + \Delta_k) \approx \Sigma_{\text{ID}} + \Delta_k, \tag{18}$$

where $\Delta_k = -\frac{1}{n}\, x_k x_k^\top$, $\frac{n}{n-1} \approx 1$.

Let $\lambda_{\max}(\Sigma_{\text{ID}})$ and $\lambda_{\min}(\Sigma_{\text{ID}})$ be the maximum and minimum eigenvalues of Σ_{ID}, with corresponding normalized eigenvectors $u_{\max}$ and $u_{\min}$. A first-order perturbation yields

$$\lambda_{\max}(\Sigma_{-k}) \approx \lambda_{\max}(\Sigma_{\text{ID}}+\Delta_k) \approx \lambda_{\max}(\Sigma_{\text{ID}}) + u_{\max}^\top \Delta_k\, u_{\max},$$

$$\lambda_{\min}(\Sigma_{-k}) \approx \lambda_{\min}(\Sigma_{\text{ID}}+\Delta_k) \approx \lambda_{\min}(\Sigma_{\text{ID}}) + u_{\min}^\top \Delta_k\, u_{\min}.$$

Define

$$\delta_{\max}^{(k)} := u_{\max}^\top \Delta_k\, u_{\max}, \quad \delta_{\min}^{(k)} := u_{\min}^\top \Delta_k\, u_{\min}.$$

Let $f(\Sigma_{\text{ID}})$ denote the approximated domain discrepancy function from Eq. 16:

$$f(\Sigma_{\text{ID}}) = \frac{\lambda_{\max}(\Sigma_{\text{ID}}) \times \sqrt{d} + \sqrt{d^2-d}}{\lambda_{\min}(\Sigma_{\text{ID}})}.$$

After removing x_k, we have

$$f(\Sigma_{-k}) \approx \frac{[\lambda_{\max}(\Sigma_{\text{ID}}) + \delta_{\max}^{(k)}] \times \sqrt{d} + \sqrt{d^2-d}}{\lambda_{\min}(\Sigma_{\text{ID}}) + \delta_{\min}^{(k)}}. \tag{19}$$

Define

$$A = \lambda_{\max}(\Sigma_{\text{ID}}) \times \sqrt{d} + \sqrt{d^2-d}, \quad B = \lambda_{\min}(\Sigma_{\text{ID}}).$$

A first-order expansion of the denominator gives:

$$\frac{1}{B+\delta_{\min}^{(k)}} \approx \frac{1}{B}\left(1 - \frac{\delta_{\min}^{(k)}}{B}\right).$$

Thus,

$$\begin{aligned} f(\Sigma_{-k}) &\approx \frac{A + \sqrt{d}\times\delta_{\max}^{(k)}}{B}\left(1 - \frac{\delta_{\min}^{(k)}}{B}\right) \\ &\approx \frac{A}{B} + \frac{\sqrt{d}\times\delta_{\max}^{(k)}}{B} - \frac{A\times\delta_{\min}^{(k)}}{B^2}. \end{aligned} \tag{20}$$

Therefore, the change in the function, which approximates the marginal OOD-robust data value of x_k, is

$$\begin{aligned} f(\Sigma_{-k}) - f(\Sigma_{\text{ID}}) &\approx \frac{\sqrt{d}\times\delta_{\max}^{(k)}}{B} - \frac{A\times\delta_{\min}^{(k)}}{B^2} \\ &= \frac{\sqrt{d}\times\delta_{\max}^{(k)}}{\lambda_{\min}(\Sigma_{\text{ID}})} - \frac{(\lambda_{\max}(\Sigma_{\text{ID}})\times\sqrt{d} + \sqrt{d^2-d}\,)\times\delta_{\min}^{(k)}}{\lambda_{\min}(\Sigma_{\text{ID}})^2}. \end{aligned} \tag{21}$$

The proposed term quantifies the marginal data value with respect to domain discrepancy, rather than the ID loss. Accordingly, it can be integrated into the marginal values derived from existing ID-based data valuation methods. Under the assumption that the OOD loss can be approximated by the sum of the ID loss and domain discrepancy, this enables principled data valuation in OOD settings.

The derivations above demonstrate how domain discrepancy can be bounded by the eigenvalue ratio of the Hessians, how the OOD covariance matrix can be related to the ID covariance matrix, and how perturbation analysis

yields an approximation of the marginal data value. In addition, we examine the potential bias arising from dominant eigenvalues, as previously discussed in PCA-related analyses [43]. Our formulation indicates that robustness improves when the spectral distribution is balanced, with sufficiently large minimum eigenvalues and no excessively dominant maximum eigenvalues. This observation aligns with prior findings, suggesting that over-reliance on dominant directions can lead to biased representations, while a more uniform eigenvalue spectrum contributes to improved stability and generalization.

B. Additional Experiment Setting

B.1. Dataset

CIFAR-10. A widely used image classification dataset consisting of natural images from ten classes. We use it as the source domain for training.

CIFAR-10 C. A corrupted version of CIFAR-10 that introduces common distribution shifts through 15 corruption types, each with multiple severity levels. We use the 5 severity level. CIFAR-10 serves as the target domain for evaluating robustness under distribution shift.

VLCS. A domain generalization benchmark composed of four visual domains: VOC2007, LabelMe, Caltech101, and SUN09. In each evaluation setting, one domain is held out as the target while the model is trained on the remaining three. The target domain is rotated across all four domains.

Amazon Reviews. A sentiment classification dataset organized by product category, with each category treated as a separate domain. We convert the 5-point rating into three sentiment classes (negative: 1–2, neutral: 3, positive: 4–5) and perform 3-class classification. Models are trained on one or more source categories and evaluated on a disjoint target category to assess cross-domain generalization.

ImageNet. A large-scale image classification dataset with 1,000 classes. For scalability experiments, we use a subset of the training split. Robustness is measured under domain shifts. For this benchmark, we performed the data valuation experiment using a subset of 30,000 samples.

DomainNet. A large-scale benchmark for multi-domain learning, containing six stylistically distinct domains: clipart (C), infograph (I), painting (P), quickdraw (Q), real (R), and sketch (S). We evaluate generalization by holding out one domain as the target and training on the remaining five. For this benchmark, we performed the data valuation experiment using a subset of 2,000 samples for each domain.

Unlike prior OOD-aware valuation methods that require access to OOD data during estimation [2, 58], EV operates solely on ID data and therefore remains valid even when no OOD samples are available. This makes EV inherently robust to unseen OOD scenarios, a setting that is common in real-world deployments. All experiments use the train split of each dataset. Performance is also evaluated on randomly sampled data from the train split of the target domain (i.e., the domain not used for training), using a fixed random seed of 42 except for the instability ranking experiment.

B.2. Experiment setting

Evaluation protocols. We use three procedures.

- **Point addition.** We sample 2,000 in distribution examples. We compute values with each method. We form an initial training set of 1,000 and retrain while adding the highest value samples from the remaining pool. We evaluate on a different target domain.
- **Data removal.** We sample 1,000 of 2,000 in distribution points. We score them, remove the top 50 percent by value, and train on the rest. We evaluate on the target domain. A larger drop in accuracy indicates a better ability to identify low utility samples.
- **Instability.** We assess sensitivity to small changes in the training set. We fix 180 of 200 indices, resample the remaining 20, repeat valuation five times, and compute the standard deviation of value rankings on the fixed indices. These five runs use different random seeds (8, 18, 29, 39, 58).

Baselines and parameters. In this work, we limit the EV-integrated methods to LAVA, KNN Shapley, and Data-OOB, selected based on a balance of performance and computational efficiency. For KNN Shapley, we use a validation set of 1,000 examples and set the neighborhood size to 1,000. For Data-OOB, we follow the original paper with num models = 800. We train a logistic regression classifier for 10 epochs with a learning rate of 0.001.

Hardware setting. We set seed 42 on a single RTX 4090 GPU and an Intel Xeon Gold 6426Y CPU with 32 cores.

B.3. Weight parameter

EV is combined with a baseline valuation score (V_{EV}). Since the EV term may have a different scale from other methods(V_{base}), we center and scale it using the baseline statistics to make the two terms comparable. Specifically, $\tilde{V}_{\text{EV}} = \frac{V_{\text{EV}} - \mu_{\text{base}}}{\sigma_{\text{base}}}$, where $\mu_{\text{base}} = \text{mean}(V_{\text{base}})$ and $\sigma_{\text{base}} = \text{std}(V_{\text{base}})$. The final score is $V_{\text{final}} = V_{\text{base}} + w\,\tilde{V}_{\text{EV}}$. In our experiments, we set w$\leq$1.

C. Supplementary Experiments

C.1. Experiment on ImageNet and DomainNet

We extend the Data Removal experiment from the main paper to more challenging benchmarks. For ImageNet, data valuation was conducted on 30,000 training samples from the train split of ImageNet. For DomainNet, 2,000

Acc (%) (↓)	ImageNet				DomainNet					
Method	V2	S	R	A	C	I	P	Q	R	S
Random	65.5	28.2	29.9	9.0	26.4	13.7	31.7	2.2	43.8	18.6
InfluenceFunction	65.6	28.3	29.5	8.7	25.6	12.6	30.9	2.2	41.2	19.7
KNN Shapley	40.3	18.0	17.0	7.8	21.5	9.5	24.7	2.7	34.0	14.2
Data-OOB	59.2	23.8	25.1	6.5	15.2	6.0	16.5	2.1	20.4	10.4
EV + KNN Shapley	**40.3**	**17.9**	**16.9**	7.8	20.2	9.2	23.7	2.6	33.0	13.5
EV + Data-OOB	54.7	21.8	22.8	**5.4**	**14.5**	**5.8**	**15.1**	**2.0**	**18.6**	**10.2**

Table 3. Data removal experiment. Train the model with 50% of the data, which is the lowest data value in the ID set, and evaluate performance on different domain data. **Lower is better.** Across both large and real benchmarks, EV augmented variants consistently achieve the lowest error, which means EV achieves stronger OOD robustness than other methods. Because of their prohibitive time complexity on large, high-cardinality datasets, LAVA and Deviation are omitted.

samples were drawn from each domain, and data valuation was performed using the remaining 10,000 samples, excluding the target domain. The experimental setup follows that of Table 2 in the main paper. As shown in Table 3, EV continues to outperform other methods in OOD domains, and the performance gain from integrating EV is consistently observed compared to the base methods without EV. While Table 2 reports the averaged performance over DomainNet due to space constraints, per-domain results in the appendix also confirm that EV consistently improves performance across individual domains. Notably, on V2, EV + KNN Shapley slightly outperforms KNN Shapley alone, even at the second decimal place. Due to computational constraints, LAVA was excluded due to its sensitivity to the number of labels, and Deviation was excluded because it scales poorly with dataset size.

C.2. Experiment under Indefinite Hessian

When the Hessian is indefinite, Eq. 11 does not directly apply because the minimum eigenvalue can be negative. By following the same derivation as in the positive semi-definite (PSD) case and restricting the analysis to the subspace spanned by negative eigenvectors, a Rayleigh–based upper bound can still be obtained, where the bound depends on the largest negative eigenvalue. For simplicity, we use the same formulation as in the PSD case. Empirically, incorporating EV consistently improved performance even when the Hessian was indefinite. Table 4 reports the results under the same setup as Table 1 using a 2-layer MLP.

C.3. Experiment under Different Embedder

Our experimental pipeline relies on embedding representations, which raises a natural concern that the performance of EV might depend on the quality or choice of the embedder. To examine this, we repeated the CIFAR-10 data-removal experiments (Table 1) using a variety of embedding models. As reported in Table 5, EV consistently maintains strong robustness to OOD data across different embedders. This result indicates that the OOD robustness of EV is not tied to a specific embedding model and that EV remains effective even when alternative feature extractors are used.

Method	**CIFAR-10 C ↓**
Random	51.9
InfluenceFunction	50.7
Deviation	51.0
LAVA	44.5
KNN Shapley	41.3
Data-OOB	51.8
EV + LAVA	**40.1**
EV + KNN Shapley	40.2
EV + Data-OOB	44.0

Table 4. Data removal experiment. Comparison of data valuation methods on CIFAR-10 C using a 2-layer MLP model under domain shift. Lower accuracy indicates more effective identification of low-utility samples. The experimental setup, except the model, follows the same configuration as described in Table 1.

C.4. Different Removal Rate Results on CIFAR-10 C

While Table 1 reports the results of removing the top 50% highest-value samples, we additionally evaluate a less extreme setting to show that EV is not effective only under large removal rates. Table 6 presents the results when the removal rate is reduced to 10%. EV continues to deliver strong performance at this lower rate, exhibiting consistent improvements across methods. These results indicate that EV is broadly applicable and remains effective under a range of removal rates.

Method / Embedder	ResNet-18	ResNet-50	ViT-B/16	ViT-L/16	**Avg**
Random	48.7	47.0	67.0	70.9	58.7
InfluenceFunction	46.3	47.8	67.5	70.2	57.9
Deviation	43.6	46.5	66.3	71.6	57.0
LAVA	42.7	46.3	66.6	69.9	56.4
KNN Shapley	33.8	38.8	60.4	63.3	49.1
Data-OOB	36.3	44.6	67.8	70.4	54.8
EV + LAVA	42.4	43.9	**58.1**	**58.6**	50.8
EV + KNN Shapley	33.9	**38.6**	60.3	62.5	**48.8**
EV + Data-OOB	**33.6**	43.6	63.9	66.6	51.9

Table 5. Data removal experiment. Comparison of data valuation methods on CIFAR-10 C using different embedders. Except for the choice of embedder, the experimental setup follows the same configuration as in Table 1.

Method	**CIFAR-10 C** ↓
Random	50.1
InfluenceFunction	49.3
Deviation	50.7
LAVA	50.03
KNN Shapley	48.19
Data-OOB	48.5
EV + LAVA	48.6
EV + KNN Shapley	**45.8**
EV + Data-OOB	46.9

Table 6. Data removal experiment. Comparison of data valuation methods on CIFAR-10 C using different removal rates (10%). Lower accuracy indicates more effective identification of low-utility samples. The experimental setup, except removal rate, follows the same configuration as described in Table 1.

C.5. Additional Baseline Results on CIFAR-10 C

We extended the data removal experiment of Table 1 by applying the same evaluation procedure to additional baselines on CIFAR-10 C. As summarized in Table 7, we included both DAVINZ [56] and Banzhaf [54] under the identical experimental setup. Across these baselines, incorporating EV led to consistent improvements, indicating that the proposed approach effectively enhances the ability of diverse valuation methods to identify low-utility samples. These results confirm that the benefit of EV is not limited to a specific algorithm but generalizes across methods with different underlying principles.

C.6. Additional metrics for the Instability Ranking experiment

In Figure 4, we examine how sensitive each valuation method is to small perturbations in the training subset by computing the standard deviation of the resulting data value rankings across multiple resamples. This analysis illustrates how much the ranking produced by each method

Method	**CIFAR-10 C** ↓
Random	47.0
InfluenceFunction	47.8
Deviation	46.5
LAVA	46.3
KNN Shapley	38.8
Data-OOB	44.6
Banzhaf	46.7
DAVINZ	47.3
EV + LAVA	43.9
EV + KNN Shapley	38.6
EV + Data-OOB	43.6
EV + Banzhaf	45.4
EV + DAVINZ	**35.6**

Table 7. Data removal experiment. Comparison of data valuation methods on CIFAR-10-C under domain shift, including additional results for Banzhaf and DAVINZ methods. The experimental setup follows the same configuration as described in Table 1.

Method / Metric	Std. (↓)	Variance (↓)	Spearman (↑)	Kendall (↑)
Random	48.58	2582.65	0.05	0.03
Deviation	34.25	1489.68	0.45	0.32
EV + LAVA	**7.60**	**75.45**	**0.97**	**0.86**
EV + KNN Shapley	7.84	81.77	**0.97**	**0.86**
EV + Data-OOB	19.18	592.09	0.77	0.62

Table 8. Additional metrics for the Instability Ranking experiment

fluctuates when the underlying subset used for valuation is slightly altered. Table 8 provides a more comprehensive evaluation based on the rankings obtained in Figure 4. In addition to reporting the standard deviation, we also compute the variance, Spearman correlation, and Kendall correlation across repeated runs. These complementary metrics reveal that EV achieves substantially greater stability than Deviation, demonstrating that EV retains robust OOD performance even when the training subset undergoes minor changes.