REPRESENTER POINT SELECTION VIA TRACKING REGULARIZED GRADIENT TRAJECTORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the individual contribution of training samples to model predictions is essential for enhancing model transparency and facilitating data-driven debugging. While recent advancements, such as Representer Point Selection and Neural Tangent Kernels, offer efficient frameworks for sample-based attribution, they face significant practical hurdles. Specifically, existing methods often necessitate performance-altering regularization that deviates from the original model, rely on computationally prohibitive Hessian-Vector Products, or fail to faithfully reconstruct the original parameters. Consequently, they explain a surrogate approximation rather than the actual model's predictions. To address these challenges, we propose an alternative derivation for Representer Point Selection via Tracking Regularized Gradient Trajectory (RPS-TrReg), which identifies sample importance by tracing a regularized optimization path toward the original model's parameter state. Unlike previous methods that optimize a surrogate model in isolation, RPS-TrReg incorporates a regularization term to ensure the tracking phase remains strictly anchored to the target model's weights. This enables a high-fidelity decomposition of the target model into a linear combination of training samples without the need for expensive second-order derivatives. Empirically, our results demonstrate that RPS-TrReg consistently outperforms state-of-the-art explanation methods in data debugging tasks across vision and language classification benchmarks.

1 INTRODUCTION

As machine learning systems become deeply embedded in high-stakes decision-making, the demand for transparency and accountability has reached an unprecedented level. Within the field of Explainable Artificial Intelligence (XAI), sample-based explanations have emerged as a powerful tool for understanding model behavior by identifying the most influential training data points behind a specific prediction. Unlike feature-based methods that highlight salient input regions Ribeiro et al. (2016); Selvaraju et al. (2017); Lundberg & Lee (2017); Smilkov et al. (2017), sample-based attribution methods allow practitioners to perform data curation and model debugging by pinpointing specific examples in the training set that drive incorrect or biased outputs Koh & Liang (2017); Yeh et al. (2018); Pruthi et al. (2020); Sui et al. (2021); Tsai et al. (2023a;b).

A particularly efficient and theoretically grounded framework for these explanations is Representer Point Selection (RPS), which leverages the dual form of optimization problems to express model parameters as a linear combination of training samples Yeh et al. (2018); Sui et al. (2021); Tsai et al. (2023b). However, existing methods face critical challenges that limit their practical utility. Previous implementations Yeh et al. (2018); Tsai et al. (2023b) typically require fine-tuning the model with an performance-altering regularization term. This induces a substantial parameter shift between the original network and the modified explanatory model. Consequently, the resulting attributions risk misrepresenting the original decision process, ultimately explaining the behavior of a drifted surrogate rather than the true prediction itself. To mitigate these issues, recent work improves the fidelity of explanations by using a Taylor expansion near the saddle point, but heavily relies on the calculation of the Hessian inverse matrix Sui et al. (2021). This requirement is often computationally prohibitive for large-scale neural networks and can be numerically unstable in non-convex landscapes. Alternative approaches, such as those based on the Neural Tangent Kernel (NTK) Tsai et al. (2023a); Engel et al. (2023), offer another path but often incur computational costs tied to the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

complexity of the selected kernel. More importantly, when training a surrogate model to mimic the target, these methods often fail to remain strictly anchored to the original model's weights, resulting in explanations that lack high-fidelity parameter alignment.

To address these significant limitations, we propose an alternative derivation for **R**epresenter **P**oint **S**election via **Tr**acking **Reg**ularized Gradient Trajectory (**RPS-TrReg**), which identifies sample influence by tracing a regularized optimization path during the fine-tuning stage. Unlike previous surrogate-based methods that may optimize in isolation, RPS-TrReg ensures the parameters of the explanatory model remain strictly anchored to the weights of the original network via a proximity-preserving regularization term. This constraint enables high-fidelity decomposition of the target model while minimizing the reliance on computationally expensive second-order derivatives and intrusive weight modifications that alter the original model's behavior.

In summary, our main contributions are as follows: 1) we introduce the RPS-TrReg framework, which utilizes trajectory regularization to ensure high-fidelity sample-based attribution; 2) we provide a computationally efficient derivation that bypasses the need for Hessian-Vector Products; and 3) empirical evaluations across vision (CIFAR-10) and language (IMDB) benchmarks demonstrate that RPS-TrReg effectively captures sample influence and achieves superior alignment with the original model compared to state-of-the-art sample-based attribution methods.

2 Related Works and Preliminaries

Sample-based attribution methods aim to identify specific training instances that exert the most significant influence on a model's prediction for a given test point Hammoudeh & Lowd (2024). Early model-agnostic approaches focused on prototype selection or using prototypes and criticisms to summarize dataset coverage Bien & Tibshirani (2011); Kim et al. (2014; 2016). In contrast, model-aware methods characterize influence by measuring how the inclusion or exclusion of a specific sample alters the objective loss or prediction behavior Koh & Liang (2017); Anirudh et al. (2017). While theoretically grounded, leave-one-out approaches often incur prohibitive computational overhead because they require retraining the model for each training sample. Recently, frameworks such as Representer Point Selection (RPS) Yeh et al. (2018); Tsai et al. (2023b) and Neural Tangent Kernels (NTK) Tsai et al. (2023a); Engel et al. (2023) have emerged as more efficient alternatives, leveraging optimization dual forms and gradient tracing to attribute predictions to training data.

Beyond attributing individual test predictions, another line of research seeks to explain global model behavior by tracing training dynamics across time. For instance, ExPLAIND Eichin et al. (2025) unifies model and training attribution, MAGIC Ilyas & Engstrom (2025) develops near-optimal attribution frameworks for deep learning, and exact kernel equivalence formulations for finite classification models Bell et al. (2023) bridge empirical models with functional kernel representations. These efforts pursue an orthogonal objective to our work. While they analyze global model properties and optimization dynamics by tracking intermediate checkpoints at high computational cost, our work focuses strictly on efficient, instance-level prediction attribution.

2.1 Problem Formulation

The objective of sample-based explanation is to reformulate the model's output as an interpretable sum of training contributions Schölkopf et al. (2001). Specifically, for a classifier $\mathcal{M}_{\mathbf{\Theta}^\dagger}$, we seek to decompose the pre-activation prediction $\hat{y}_t^\dagger = \Theta^\dagger \phi(\mathbf{x}_t)$ into a linearly separable format:

$$\hat{y}_t^\dagger = \sum_{i=1}^{n} \alpha_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}_t), \tag{1}$$

where $\Theta^\dagger \subset \mathbf{\Theta}^\dagger$ represents the weights of the final classification layer, and $\phi(\cdot)$ denotes the feature embedding function. For clarity in the subsequent sections, the superscript $\dagger$ denotes the original target model, while $*$ denotes the model optimized within each respective attribution method. Here, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_t)$ is a kernel function (typically the inner product $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_t)$) that measures the similarity between the training point $\mathbf{x}_i$ and the test point $\mathbf{x}_t$ (Local Importance), while α_i represents the importance weight of training sample i (Global Importance) Yeh et al. (2018). This decomposition allows practitioners to perform data-driven debugging and curation by pinpointing the specific historical samples that drive the model's current behavior.

2.2 REPRESENTER POINT SELECTION METHODS

The original RPS-l_2 framework Yeh et al. (2018) achieves this decomposition by introducing an l_2 regularization term to ensure the existence of a representer dual form. The model is fine-tuned using a modified objective function:

$$\Theta^* = \arg\min_{\Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\hat{y}_i, \hat{y}_i^\dagger) + \lambda \|\Theta\|_2 \right\}, \ \hat{y}_i = \Theta\phi(\mathbf{x}_i), \tag{2}$$

where Θ represents the parameters of the classification layer being optimized. By assuming Θ^* is at a saddle point where the gradient is close to zero, the global importance is defined as $\alpha_i = -\frac{1}{2\lambda n} \frac{\partial \mathcal{L}(\sigma(\hat{y}_i^*), y_i)}{\partial \hat{y}_i^*}$, where $\hat{y}_i^* = \Theta^*\phi(\mathbf{x}_i)$, while σ and y_i denote the activation function and the ground-truth label, respectively. While efficient, a key drawback is that the explanations correspond to the modified l_2-regularized classification layer Θ^* rather than the target layer parameters $\Theta^\dagger$.

To improve fidelity, RPS-LJE Sui et al. (2021) further develops this approach by utilizing a Local Jacobian Taylor expansion around a nearby parameter Θ^*. Assuming the target model $\Theta^\dagger$ is at a saddle point where the total gradient is close to zero, and can be reconstructed from a nearby parameter state Θ^* (obtained via one-step gradient descent from $\Theta^\dagger$) through a Taylor expansion:

$$\Theta^\dagger = \Theta^* - \sum_{i=1}^{n} \frac{1}{n} \mathcal{H}_{\Theta^*}^{-1} \nabla_{\Theta^*} \mathcal{L}(\sigma(\hat{y}_i^*), y_i) + \xi, \ \hat{y}_i^* = \Theta^*\phi(\mathbf{x}_i). \tag{3}$$

Here, $\mathcal{H}^{-1}$ denotes the inverse Hessian matrix, and the global importance is approximated as $\alpha_i = \Theta^* \frac{1}{\phi(\mathbf{x}_i)n} - \frac{1}{n} \mathcal{H}_{\Theta^*}^{-1} \frac{\partial \mathcal{L}(\sigma(\hat{y}_i^*), y_i)}{\partial \hat{y}_i^*}$. This approach can be viewed as an evolution of the Influence Function method Koh & Liang (2017), sharing a similar mathematical structure that estimates parameter entry correlations. The primary distinction lies in RPS-LJE's use of an additive factor $\frac{1}{n}\Theta^*\phi(\mathbf{x}_t)$ and its application of derivatives to the nearby parameters Θ^* rather than the original target parameters. Despite these advancements, RPS-LJE still requires the calculation of the Hessian matrix, which is often computationally prohibitive and numerically unstable for modern deep architectures with high-dimensional parameter spaces.

2.3 NEURAL TANGENT KERNEL METHODS

Neural Tangent Kernel methods approximate the training process as a kernelized regression in a Reproducing Kernel Hilbert Space (RKHS) Schölkopf et al. (2001). Under this paradigm, the global importance α is derived by solving:

$$\alpha^* = \underset{\alpha}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L} \left(\sum_{j=1}^{n} \alpha_j \mathcal{K}(\mathbf{x}_j, \mathbf{x}_i), \hat{y}_i^\dagger \right) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha \right\}, \tag{4}$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the kernel gram matrix with $K_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$. While optimal for the RKHS surrogate, this objective optimizes an idealized kernel loss rather than faithfully reconstructing the target model's actual predictions.

3 METHODOLOGY

To address the fidelity limitations inherent in RPS-l_2 Yeh et al. (2018) and the computational bottlenecks of Hessian-based approaches like RPS-LJE Sui et al. (2021), we propose **R**epresenter **P**oint **S**election through **Tr**acking **Reg**ularized Gradient Descent Trajectories (**RPS-TrReg**). Our framework satisfies two critical requirements for effective sample-based attribution: maintaining high alignment with the target model state $\Theta^\dagger$ and ensuring computational scalability by avoiding second-order derivatives. Unlike existing methods, RPS-TrReg introduces a regularization term that explicitly anchors the gradient trajectory to the target model.

3.1 OBJECTIVE FUNCTION

We consider a model $\mathcal{M}_{\mathbf{\Theta}^\dagger}$ where the pre-activation output for an input $\mathbf{x}$ is given by $\hat{y}^\dagger = \Theta^\dagger\phi(\mathbf{x})$, where $\Theta^\dagger \in \mathbb{R}^{c \times d}$ represents the last linear layer parameters from $\mathbf{\Theta}^\dagger$, c denotes the number of

classes and d represents the feature dimension. To maintain proximity to the $\Theta^\dagger$, we define the objective function as:

$$\Theta^* = \arg\min_{\Theta} \{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\sigma(\hat{y}_i), y_i) + \frac{1}{2} \lambda \|\Theta - \Theta^\dagger\|_2 \}, \tag{5}$$

where $\mathcal{L}(\sigma(\hat{y}_i), y_i)$ denotes the loss function associated with the training task (e.g., Binary Cross-Entropy for classification) for the i-th sample, and λ represents the regularization strength governing the proximity to the original weights. It should be noted that the parameters are initialized with $\mathbf{\Theta}^\dagger$ and remain fixed throughout the optimization, with the exception of the last-layer parameters Θ, which are updated to minimize the objective function.

3.2 Regularized Gradient Descent Trajectory

To find the optimal parameters Θ^* defined in the objective function, we analyze the optimization trajectory under gradient-based updates. The gradient of the loss function $\mathcal{L}$ with respect to the last-layer parameters at training step t, denoted as $\nabla\mathcal{L}(\Theta_t)$, is composed of the empirical data gradient and the regularization term:

$$\nabla\mathcal{L}(\Theta_t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \mathcal{L}(\sigma(\hat{y}_{i,t}), y_i)}{\partial \hat{y}_{i,t}} \frac{\partial \hat{y}_{i,t}}{\partial \Theta_t} + \lambda(\Theta_t - \Theta^\dagger) \tag{6}$$

$$= \frac{1}{n} \sum_{i=1}^{n} g_{i,t} \phi(\mathbf{x}_i)^\top + \lambda(\Theta_t - \Theta^\dagger), \tag{7}$$

where $g_{i,t} = \frac{\partial \mathcal{L}(\sigma(\hat{y}_{i,t}), y_i)}{\partial \hat{y}_{i,t}} \in \mathbb{R}^{c \times 1}$ denotes the gradient of the loss $\mathcal{L}(\sigma(\hat{y}_{i,t}), y_i))$ with respect to the pre-activation output $\hat{y}_{i,t} = \Theta_t \phi(\mathbf{x}_i)$ for sample i at step t. For the sake of analytical clarity and without loss of generality, we assume the use of Stochastic Gradient Descent (SGD) with a learning rate η to find the optimal parameters Θ^*. Under this optimization protocol, the update rule for the regularized trajectory is expressed as:

$$\Theta_{t+1} = \Theta_t - \eta \left[\frac{1}{n} \sum_{i=1}^{n} g_{i,t} \phi(\mathbf{x}_i)^\top + \lambda(\Theta_t - \Theta^\dagger) \right]. \tag{8}$$

The recursive relation can be further simplified by introducing the contraction coefficient $k = 1 - \eta\lambda$:

$$\Theta_{t+1} = k\Theta_t + (1-k)\Theta^\dagger - \frac{\eta}{n} \sum_{i=1}^{n} g_{i,t} \phi(\mathbf{x}_i)^\top. \tag{9}$$

By expanding this recurrence over T iterations, we reveal how the final state Θ^* is constructed as a cumulative result of the target parameters $\Theta^\dagger$ and the weighted training data. We assume an initial condition of $\Theta_0 = \mathbf{0}$, which is a natural choice as the last layer can be explicitly zero-initialized, or alternatively, a standard zero-mean initialization can be effectively viewed as this state for the purpose of the tracking derivation. The resulting closed-form expression is:

$$\Theta^* = \left[\sum_{j=0}^{T-1} k^j (1-k) \right] \Theta^\dagger - \frac{\eta}{n} \sum_{t=0}^{T-1} k^{T-1-t} \left(\sum_{i=1}^{n} g_{i,t} \phi(\mathbf{x}_i)^\top \right). \tag{10}$$

3.3 Reconstruction of $\Theta^\dagger$ via Stationary Assumption

Assuming the model reaches a near-saddle point at the final step, we presume that the loss derivative at the end of training is negligible ($0 \approx \nabla\mathcal{L}(\Theta^*)$). This yields a stationary balance:

$$0 \approx \frac{1}{n} \sum_{i=1}^{n} g_{i,T} \phi(\mathbf{x}_i)^\top + \lambda(\Theta^* - \Theta^\dagger), \tag{11}$$

which provides the steady-state representation of the parameters:

$$\Theta^* \approx \Theta^\dagger - \frac{1}{n\lambda} \sum_{i=1}^{n} g_{i,T} \phi(\mathbf{x}_i)^\top. \tag{12}$$

Substituting the steady-state representation of Θ^* in Eq. 12 into the path integral formulation in Eq. 10 and leveraging the geometric series sum $\sum_{j=0}^{T-1} k^j(1-k) = 1 - k^T$, we derive the final reconstruction of the original parameters $\Theta^\dagger$:

$$\Theta^\dagger \approx \frac{1}{k^T n\lambda} \sum_{i=1}^{n} g_{i,T} \phi(\mathbf{x}_i)^\top - \frac{\eta}{k^T n} \sum_{t=0}^{T-1} k^{T-1-t} \left(\sum_{i=1}^{n} g_{i,t} \phi(\mathbf{x}_i)^\top \right). \tag{13}$$

By regrouping the terms, we can represent the influence of each training point i through its cumulative and terminal representer weights:

$$\Theta^\dagger \phi(\mathbf{x}_t) \approx \sum_{i=1}^{n} \underbrace{\left[\frac{g_{i,T}}{k^T n\lambda} - \frac{\eta}{k^T n} \sum_{t=0}^{T-1} k^{T-1-t} g_{i,t} \right]}_{\text{Global Importance } \alpha_i} \underbrace{\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_t)}_{\text{Local Importance } \beta_i} \tag{14}$$

This formulation fulfills our primary objective of decomposing the original model's state into a weighted linear combination of training samples. By tracing the regularized gradient trajectory rather than relying on expensive second-order derivatives, RPS-TrReg provides a high-fidelity interpretation of how specific training data points contribute to the target model parameters.

4 Experiments

In this section, we evaluate the performance of RPS-TrReg against state-of-the-art baselines across several critical dimensions. First, we assess the ability of each method to quantify training sample influence in a data debugging scenario (Sec. 4.2), simulating an application where human experts inspect the annotation quality of a training set to mitigate detrimental impacts on test performance. In Sec. 4.3, we demonstrate the superior fidelity of our approach by measuring the l_1 alignment between the reconstructed weights and the original target model parameters. Subsequently, we demonstrate the computational efficiency of our approach, particularly its scalability relative to methods requiring Hessian-Vector Products (HVP), by evaluating the runtime across two datasets (Sec. 4.4). Finally, we qualitatively analyze the numerical stability and explanation fidelity of the candidate models in Sec. 4.5 and Sec. 4.6, respectively.

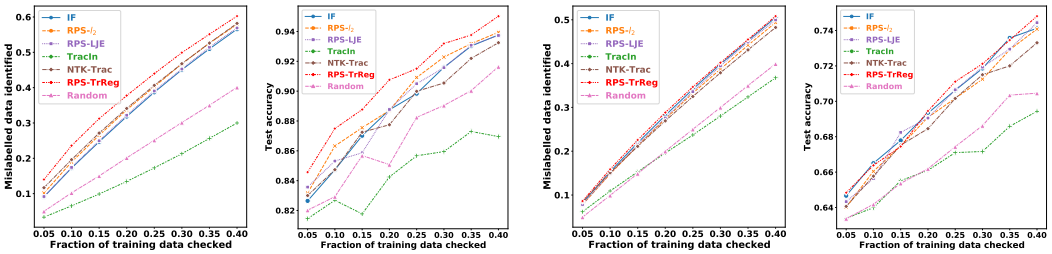
4.1 Experimental Setups

Following the experimental protocols established in RPS-LJE Sui et al. (2021), we evaluate our framework on both vision and language classification tasks: 1) binary image classification using ResNet-20 He et al. (2016) on the CIFAR-10 dataset Krizhevsky et al. (2009) (specifically the "horses vs. cars" subset), and 2) sentiment analysis using a Bi-LSTM architecture Graves (2012) on the IMDB dataset Maas et al. (2011). To provide a comparative analysis, we include several state-of-the-art data attribution methods as baselines: Influence Function (IF) Koh & Liang (2017), RPS-l_2 Yeh et al. (2018), RPS-LJE Sui et al. (2021), TracIn Pruthi et al. (2020), and NTK-Trac Tsai et al. (2023a). For RPS-l_2 and RPS-TrReg, the regularization parameter λ is set to 3×10^{-3}. All experiments were conducted using a single NVIDIA GeForce RTX 4090 (24GB).

Notably, while evaluating on massive architectures and large-scale datasets is valuable, our study primarily focuses on the fundamental fidelity of prediction reconstruction. Consistent with established attribution literature, these controlled benchmarks allow us to isolate mathematical fidelity from confounding optimization factors, providing a clear testbed to study prediction alignment where existing surrogate formulations face noticeable challenges.

4.2 Main Results

To evaluate the efficacy of our method in data debugging, we intentionally corrupt the training datasets by randomly flipping 30% of the training labels, which significantly degrades model per-



(a) CIFAR-10 Image Classification

(b) IMDB Sentiment Analysis

Figure 1: Each group displays the fraction of mislabeled data identified versus training data inspected (left) and the resulting test accuracy after correction (right).

formance. The primary objective is to identify and rectify these harmful samples as efficiently as possible. In our experimental protocol, we inspect training samples in increments: at each iteration $i = 0, 1, \ldots, 8$, we select a batch of $k_i = 0.05i \times n$ samples (where n represents the total number of training points) that exhibit the highest global importance value α. For Influence Function and TracIn, which do not permit the decomposition of influence into separate local and global components, we utilize their total influence values for this selection process. Following each 5% increment of data correction, the model is retrained, and the resulting test accuracy is recorded. We report the mean accuracy improvement across 10 random trials for both the ResNet-20 and Bi-LSTM architectures. As expected, model accuracy increases as corrupted labels are systematically corrected. A random baseline, where samples are selected for inspection without prioritization, is included to provide a lower bound for comparison. The effectiveness of RPS-TrReg is evaluated through two complementary lenses: the efficiency of error discovery (left side of each sub-figure in Fig. 1) and the subsequent recovery of model performance (right side of each sub-figure in Fig. 1).

Efficiency of Error Discovery. As illustrated in the left panels of Fig. 1, RPS-TrReg maintains a clear margin above all baselines on CIFAR-10 and a slight advantage on IMDB, demonstrating its efficiency in identifying mislabelled samples across different architectures. Notably, TracIn consistently exhibits poor performance across two benchmarks. A possible hypothesis is that TracIn's reliance on the unconstrained accumulation of gradient dot products makes it vulnerable to the noisy early stages of training. In the presence of high label corruption, early gradients are often erratic as the model attempts to fit contradictory labels Arpit et al. (2017). While NTK-Trac and RPS-TrReg also involve gradient accumulation, they incorporate mathematical constraints that anchor the trajectory to either the final prediction or the target model state. In contrast, TracIn's unconstrained summation is more susceptible to being polluted by unstable historical gradients, leading to a loss of fidelity when identifying corrupted samples.

Impact on Model Performance. The practical utility of data debugging is best reflected in the accuracy gains achieved after retraining the model on the rectified dataset. As shown in the right panels of Fig. 1, the higher discovery rates of RPS-TrReg translate directly into superior performance recovery, where our method achieves the highest overall accuracy improvements on both the CIFAR-10 and IMDB datasets. While Influence Function and RPS-LJE marginally outperform our method at the 15% inspection mark on the IMDB dataset, their reliance on HVP makes them computationally expensive and less scalable for larger tasks. Consistent with the discovery trends discussed above, TracIn yields the lowest performance gains across both datasets.

4.3 Weight Alignment with the Target Model

To evaluate the fidelity of the candidate methods, we measure the alignment between the reconstructed parameters $\sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i)^\top$ and the target model state $\Theta^\dagger$. A high-fidelity attribution method should yield reconstructed weights that closely approximate the actual parameters of the model being explained. We quantify this alignment by calculating the average L_1 difference across the CIFAR-10 and IMDB datasets. Note that Influence Function and TracIn are excluded from this comparison, as they are not designed to explicitly reconstruct target model parameters. As illustrated in Fig. 2, RPS-TrReg achieves the lowest L_1 difference among all evaluated methods. Notably, on the CIFAR-10 dataset, our method maintains a significant margin of superior alignment compared

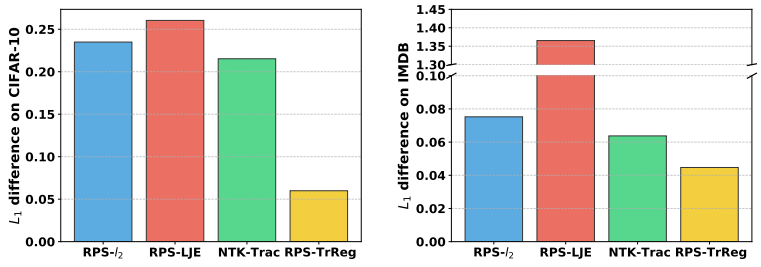


Figure 2: Comparison of the average L_1 difference between reconstructed representer weights and the actual model state.

Table 1: The time required for computing importance values across all training and test points is reported in milliseconds (ms).

		Influence Function	TracIn	RPS-l_2	RPS-LJE	NTK-Trac	RPS-TrReg
CIFAR-10	Fine-tuning	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$11.54_{\pm 0.66}$	$3.09_{\pm 1.66}$	$9.89_{\pm 0.57}$	$16.84_{\pm 0.37}$
	Computation	$239.48_{\pm 5.71}$	$56.09_{\pm 3.86}$	$24.13_{\pm 0.17}$	$322.95_{\pm 9.52}$	$23.69_{\pm 0.31}$	$23.64_{\pm 0.21}$
	Overall	$239.48_{\pm 5.71}$	$56.09_{\pm 3.86}$	$35.67_{\pm 0.62}$	$326.05_{\pm 10.29}$	$33.58_{\pm 0.33}$	$40.49_{\pm 0.38}$
IMDB	Fine-tuning	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$12.66_{\pm 0.79}$	$43.12_{\pm 0.52}$	$10.04_{\pm 0.66}$	$22.95_{\pm 0.51}$
	Computation	$1512.19_{\pm 143.43}$	$750.58_{\pm 7.94}$	$361.18_{\pm 0.94}$	$896.37_{\pm 27.97}$	$361.65_{\pm 1.29}$	$363.37_{\pm 11.16}$
	Overall	$1512.19_{\pm 143.43}$	$750.58_{\pm 7.94}$	$373.84_{\pm 0.50}$	$939.49_{\pm 28.24}$	$371.69_{\pm 0.84}$	$386.33_{\pm 11.18}$

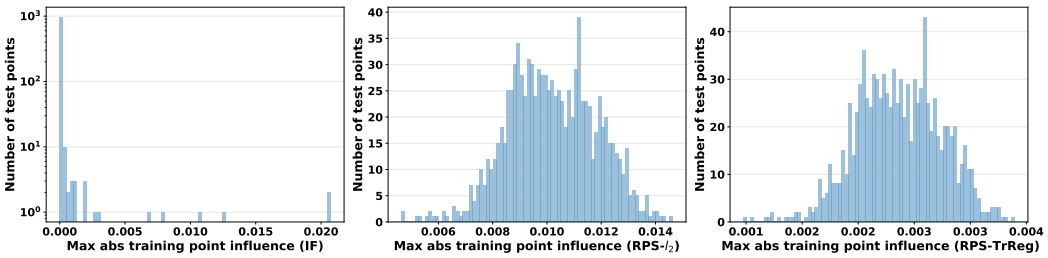
to the l_2-regularized RPS-l_2 and the surrogate-based NTK-Trac. Furthermore, while RPS-LJE was specifically designed to improve upon the fidelity of RPS-l_2 through Taylor expansion, it unexpectedly results in the highest alignment gap in our experiments. These results demonstrate that by tracing a regularized gradient trajectory anchored to the final weights, RPS-TrReg provides a more accurate decomposition of the target model's knowledge.

4.4 Computational Cost

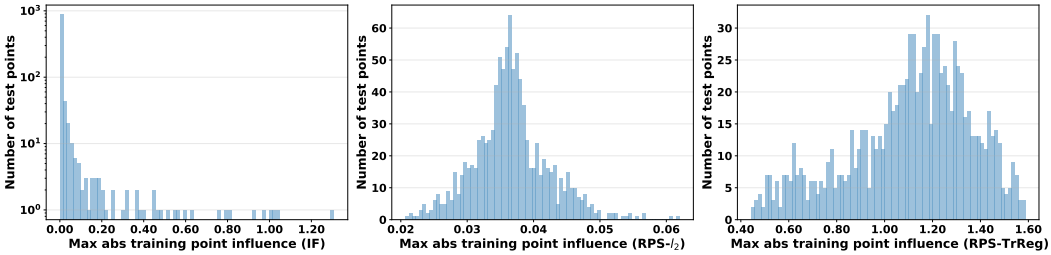
Computational efficiency is a critical factor when calculating importance values for large-scale datasets, as the resource requirements for processing each data point can quickly become prohibitive. We report a comparison of computation times in Tab. 1, measured across the CIFAR-10 and IMDB benchmarks. To ensure reliability, we conducted ten independent runs to compute importance values for all training and test data, reporting the average execution time and standard deviation. Notably, Influence Function and TracIn do not require a fine-tuning step when provided with a pre-trained model and its corresponding checkpoints. Consequently, their fine-tuning costs are recorded as zero. In contrast, the other evaluated methods require varying forms of preprocessing or optimization before importance values can be computed. It is important to emphasize that while fine-tuning represents a one-time overhead, the computation time is incurred for every sample analyzed. Our method, RPS-TrReg, demonstrates superior scalability compared to Influence Function and RPS-LJE, as it entirely avoids the use of HVP. This efficiency advantage becomes even more pronounced as the dataset size increases. While RPS-l_2 and NTK-Trac require slightly less overall running time than our approach, RPS-TrReg provides superior data debugging performance (Sec. 4.2) and significantly better alignment (Sec. 4.3), justifying the marginal difference in fine-tuning overhead.

4.5 The Distribution of Importance Value

When ranking training samples by their overall importance values ($\alpha \times \beta$), we observed significant variations in numerical stability across methods. The distribution of maximum absolute importance values for 1,000 randomly sampled test points from CIFAR-10 and IMDB is illustrated in Fig. 3. Compared to Influence Function (IF) and RPS-l_2, RPS-TrReg produces more stable and consistent values across the test sets. For both datasets, the vast majority of test points for IF fall into the first bin, indicating that most test instances have influential training points with low sparsity. This



(a) The distribution of importance values for a set of randomly selected 1,000 test points on CIFAR-10



(b) The distribution of importance values for a set of randomly selected 1,000 test points on IMDB

Figure 3: RPS-TrReg exhibits a more balanced distribution of importance values across test points, especially when compared to the importance values of Influence Function (IF) which can be either really small or become zero for most points, as seen in the left-most bin.



Figure 4: For each test sample, the upper row shows positive influential points, whereas the bottom row shows negative influential points. Compared to RPS-l_2, RPS-TrReg identifies a more diverse set of influential training points.

suggests that IF often fails to identify meaningful influential examples, a limitation further validated by our case study in Sec. 4.6. Conversely, RPS-TrReg maintains a more balanced distribution, ensuring that influential training samples are effectively captured for a broader range of test cases.

4.6 Case Study

In this section, we qualitatively analyze the prediction explanation ability of candidate sample-based attribution models on the CIFAR-10 and IMDB datasets. We compare the results of our proposed

Table 2: The "Sentiment" column denotes the target label, and "Raw Reviews" represents the Bi-LSTM input. RPS-TrReg provides influential training points that closely mirror the nuanced linguistic style of the test sample.

		ID	Sentiment	Raw Reviews
Sample 1	Test point	289	positive	There was a lot about Little Vera that was strange to me. All in all I did enjoy this movie, but a lot of the way the characters behaved was not ...
Influence Function	Explanation 1	15600	positive	Though this movie has a first rate roster of fine actors, special effects that are excellent, and a story line that is full of surprises, it was ...
	Explanation 2	7999	negative	this movie was fantastic great movie all through scary as hell. and i mean it freaked me out as much as pulse and IT and the omen etc. but ...
	Explanation 3	6123	negative	What happened? What we have here is basically a solid and plausible premise and with a decent and talented cast, but somewhere the movie ...
RPS-l_2	Explanation 1	17368	negative	The story goes something like this: A small town girl, Katie (Jessica Simpson), decides to visit her boyfriend in the big city. When she ...
	Explanation 2	13466	negative	Highly implausible, unbelievable, and incoherent Spanish production about ... well, let me see just how close I can get to it. The film opens with ...
	Explanation 3	9920	positive	Why a stupid, boring, crappy overrated film series like "Star Wars" gets all the hype, and a truly amazing film like this one goes completely ...
RPS-TrReg	Explanation 1	2833	positive	This movie is a ripoff of James Cain 's novel, THE POSTMAN ALWAYS RINGS TWICE. Apparently, the director and producer never bothered ...
	Explanation 2	5741	positive	OK, this movie is stupid. I mean that in a good way though. It was stupid on purpose, and was one of the better "stupid" movies I 've seen ...
	Explanation 3	10797	positive	When it comes to movies I can be pretty picky, and I 'll complain about anything and everything that is done wrong. While every movie has its ...
Sample 2	Test point	450	positive	This is an excellent film. The aerial scenes were well-done. It was also the right balance of war and love. The film gives meaning to the phrase...
Influence Function	Explanation 1	12091	negative	I think this cartoon is one of the worst cartoons I have ever watched. I would recommend this cartoon to people who are under 5. I did used to ...
	Explanation 2	9920	positive	Why a stupid, boring, crappy overrated film series like "Star Wars" gets all the hype, and a truly amazing film like this one goes completely ...
	Explanation 3	1635	negative	I feel much less generous with this film than others of its ilk. The portrayal of madmen in this century is always done with them being so totally ...
RPS-l_2	Explanation 1	17368	negative	The story goes something like this: A small-town girl, Katie (Jessica Simpson), decides to visit her boyfriend in the big city. When she ...
	Explanation 2	5851	negative	I saw this film opening weekend in Australia anticipating with an excellent cast of Ledger, Edgerton, Bloom, Watts and Rush that the definitive ...
	Explanation 3	16053	negative	The highlight of this movie for me was without doubt Tom Hanks. As Mike Sullivan, he was definitely cast against type and showed that he can ...
RPS-TrReg	Explanation 1	13150	positive	Robert De Niro, Cuba Gooding Jr, Hal Holbrook, and all the rest of the actors and actresses in "Men of Honor" have combined to make this a ...
	Explanation 2	10797	positive	When it comes to movies I can be pretty picky, and I 'll complain about anything and everything that is done wrong. While every movie has its ...
	Explanation 3	5111	negative	Vodka Lemon is a charming, yet extremely uneven Arminian film that will delight some and bore others. Though reasonably well - crafted, the ...
Sample 3	Test point	12532	negative	I would label this show as horrendous if it weren't for the fact that it's on the same network as Arrested Development. Because it is on FOX and ...
Influence Function	Explanation 1	7794	negative	This is a wonderfully goofy example of a self produced, written and directed vanity project ... while I was working as a crew member John ...
	Explanation 2	7316	negative	Bridges of madison county is a better made version of this story. I felt the ending of this movie was not handled sensitively as they did in the ...
	Explanation 3	8217	negative	Alfred Hitchcok is not my favorite director by any means but imagine what he could have done with this! The plot holds much potential for ...
RPS-l_2	Explanation 1	12932	positive	So, Todd Sheets once stated that he considers his 1993, shot-on-video-Z-epic, Zombie Bloodbath to be his first feature film. Anyone who's ...
	Explanation 2	22	positive	THE DEVIL'S PLAYTHING is my second attempt at a Joseph Sarno production and although I will say it is far more enjoyable than the painfully ...
	Explanation 3	2817	positive	This time around, Blackadder is no longer royal (or even particularly close to being any such thing) ... instead, rather a butler to the ...
RPS-TrReg	Explanation 1	11431	negative	This film features two of my favorite guilty pleasures. Sure , the effects are laughable, the story confused, but just watching Hasselhoff in his ...
	Explanation 2	8128	positive	I am from Sweden and i have just seen this movie and the thing is that i thought it was okay. I have seen many bad comments about it but you ...
	Explanation 3	7848	negative	This should have been a movie about Sam and his wife, the glorious Peter Falk and equally glorious Olympia Dukakis. That would have been a ...

RPS-TrReg with the Influence Function (IF) and RPS-l_2. As RPS-LJE serves as an alternative to IF Sui et al. (2021), it is omitted here to focus on distinct methodological frameworks.

Influential Points on CIFAR-10 Image Classification with ResNet. Fig. 4 visualizes the top-4 training points (both positive and negative) that exert the strongest overall importance value on the prediction of a test point. We observe several key phenomena: 1) IF provides incorrect positive or negative influential points; for instance, identifying "horse" as positive influential points for a "car" test sample (ID-912). This likely stems from the numerical instabilities discussed in Sec. 4.5, where many training points yield zero importance values toward a specific test point. 2) RPS-l_2 results exhibit low diversity, with training points appearing repeatedly across the explanation lists for different test points, while RPS-TrReg produces more diverse and relevant explanation results.

Influential Points on IMDB Sentiment Analysis with LSTM. Tab. 2 displays the top-3 explanations for the IMDB sentiment analysis task. Our analysis highlights two primary findings: 1) Similar to the vision task, IF often provides positive influential points with a sentiment opposite to the test point. While RPS-based methods may also show opposite labels, this is mathematically consistent rather than an error. The global importance value in RPS is calculated as $\alpha_i \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_t)$; therefore, a negative global importance α_i paired with a negative local importance $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_t)$ can result in a positive overall importance value. 2) When the sentiment aligns, RPS-TrReg better matches the expression style of the test point. For example, in Sample 1, our explanations consistently mirror the test point's nuanced style, employing praise through disparagement, whereas baselines lack this level of stylistic fidelity.

5 Conclusion

In this paper, we introduced RPS-TrReg, a sample-based attribution framework that mitigates the trade-off between fidelity and computational efficiency in representer point selection. By reformulating attribution as a regularized gradient descent trajectory, our approach anchors the optimization path to the target model state without requiring second-order derivatives. Experimental results across vision and language tasks demonstrate that RPS-TrReg achieves superior parameter alignment and error discovery compared to both l_2-regularized and Hessian-based methods in practical data debugging. By bypassing Hessian matrix computations, RPS-TrReg provides an efficient bridge between theoretical rigor and practical interpretability.

AI USE STATEMENT

During the preparation of this manuscript, generative AI tools were used solely for language editing and polishing to improve readability. Generative AI was not used for "generate synthetic data sets", "help develop theoretical models or conceptual frameworks", "formulate mathematical claims", "provide critical ingredients for proving mathematical claims", "assist in the writing of proofs", "propose or refine hypotheses", "design or provide feedback on research methodology or experiments", "implement methods", "assist with translation", "clean and reformat dataset", "support qualitative and thematic data analysis", or "interpret results". We have reviewed and edited all generated content and take full responsibility for the final manuscript, including its integrity, accuracy, and claims.

REFERENCES

Rushil Anirudh, P Bremer, Rahul Sridhar, and JJ Thiagarajan. Influential sample selection: A graph signal processing approach. Technical report, Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 2017.

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.

Brian Wesley Bell, Michael Geyer, David Glickenstein, Amanda S Fernandez, and Juston Moore. An exact kernel equivalence for finite classification models. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 206–217. PMLR, 2023.

Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. 2011.

Florian Eichin, Yupei Du, Philipp Mondorf, Maria Matveev, Barbara Plank, and Michael A Hedderich. Explaind: Unifying model, data, and training attribution to study model behavior. *arXiv preprint arXiv:2505.20076*, 2025.

Andrew Engel, Zhichao Wang, Natalie S Frank, Ioana Dumitriu, Sutanay Choudhury, Anand Sarwate, and Tony Chiang. Faithful and efficient explanations for neural networks via neural tangent kernel surrogate models. *arXiv preprint arXiv:2305.14585*, 2023.

Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Andrew Ilyas and Logan Engstrom. Magic: Near-optimal data attribution for deep learning. *arXiv preprint arXiv:2504.16430*, 2025.

Been Kim, Cynthia Rudin, and Julie Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems*, 27, 2014.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pp. 416–426. Springer, 2001.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Yi Sui, Ga Wu, and Scott Sanner. Representer point selection via local jacobian expansion for post-hoc classifier explanation of deep neural networks and ensemble models. *Advances in neural information processing systems*, 34:23347–23358, 2021.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Sample based explanations via generalized representers. *Advances in Neural Information Processing Systems*, 36:23485–23498, 2023a.

Che-Ping Tsai, Jiong Zhang, Hsiang-Fu Yu, Eli Chien, Cho-Jui Hsieh, and Pradeep Kumar Ravikumar. Representer point selection for explaining regularized high-dimensional models. In *International Conference on Machine Learning*, pp. 34469–34490. PMLR, 2023b.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems*, 31, 2018.