A Bayesian Information-Theoretic Approach to Data Attribution

Dharmesh Tailor[*]
University of Amsterdam

Nicolò Felicioni
Spotify

Kamil Ciosek
Spotify

Abstract

Training Data Attribution (TDA) seeks to trace model predictions back to influential training examples, enhancing interpretability and safety. We formulate TDA as a Bayesian information-theoretic problem: subsets are scored by the *information loss* they induce—the entropy increase at a query when removed. This criterion credits examples for resolving predictive uncertainty rather than label noise. To scale to modern networks, we approximate information loss using a Gaussian Process surrogate built from tangent features. We show this aligns with classical influence scores for single-example attribution while promoting diversity for subsets. For even larger-scale retrieval, we relax to an information-gain objective and add a variance correction for scalable attribution in vector databases. Experiments show competitive performance on counterfactual sensitivity, ground-truth retrieval and coreset selection, showing that our method scales to modern architectures while bridging principled measures with practice.

1 INTRODUCTION

The problem of training data attribution (TDA) is to trace a model's prediction back to its training data, thereby addressing a fundamental need for interpretability and transparency in modern machine learning systems. In contrast to approaches that seek to understand a model's internal workings in a *mechanistic* fashion, TDA is a form of explanation that is grounded in data. Attributed training examples can be used to provide explanations for recommendation systems, enable risk mitigation strategies that address safety, privacy, and fairness concerns, and facilitate appropriate compensation frameworks for creative work while detecting potential plagiarism or memorization.

The de-facto technique for TDA is influence function—a classical tool from robust statistics (Hampel, 1974) and regression diagnostics (Cook and Weisberg, 1980) revived for modern machine learning by Koh and Liang (2017). By linearizing the stationarity condition at the trained parameters, influence function provides a local, first-order estimate of the counterfactual change in a model's loss or prediction when a training point is infinitesimally upweighted or removed, thereby sidestepping costly retraining. This convenience comes with caveats: the derivation depends on fully-converged parameters which is not a realistic assumption in practice; and requires the computation and inversion of large curvature matrices which is impractical for modern architectures. A substantial body of work has since tackled some of these concerns by proposing alternate influence function estimators that relax such restrictions or exploiting numerical approximations for computational efficiency. This includes approaches that backtrack through training trajectories (Hara et al., 2019); gradient-accumulation methods like TraceIn (Pruthi et al., 2020); expressive curvature surrogates such as EK-FAC (Grosse et al., 2023); leveraging random projections or improved linear system solvers (Schioppa et al., 2022).

In this work, we replace the loss-based counterfactual notion of TDA with an explicitly information-theoretic one: for a test query we score a subset by the *information* lost about the latent prediction when that subset is withheld—*i.e.*, the increase in entropy at the query (measured in nats). This choice targets epistemic uncertainty attributable to missing training data, and places our formulation in a broader tradition of Bayesian and information-theoretic approaches to learning. Information measures have long been used to formalize principled learning objectives and guarantees—*e.g.*, information gain as an acqui-

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

[*] Part of this work was carried out when Dharmesh Tailor was on internship at Spotify.

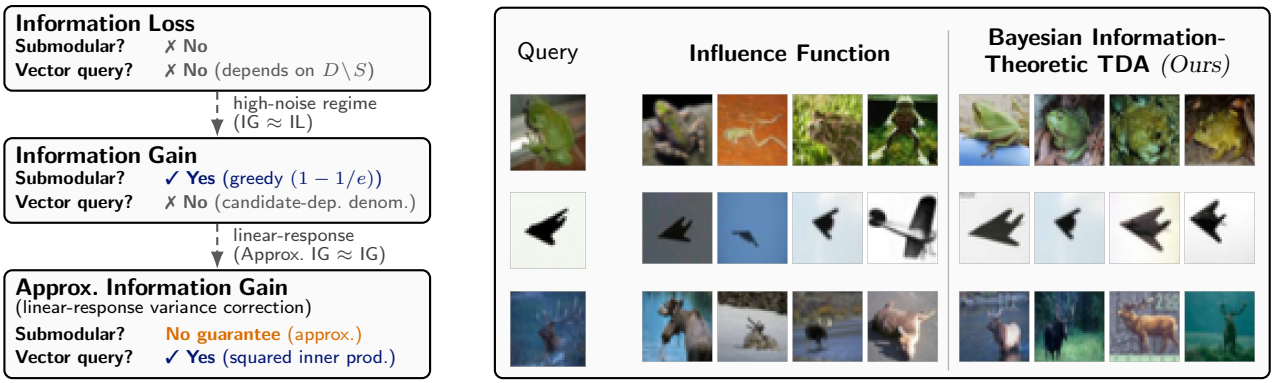


Figure 1: Bayesian information-theoretic training data attribution (TDA). For a query input $\mathbf{x}_*$, we quantify the contribution of a training subset S via its *information loss*: the increase in (posterior) predictive entropy at $\mathbf{x}_*$ when S is withheld from training, attributing credit to examples that resolve *epistemic* uncertainty rather than label noise. Left: we relate information loss to a submodular *information gain* relaxation whose leading-order term matches information loss in a high-noise regime, followed by a linear-response variance correction that implements greedy selection with a squared inner-product score, enabling efficient retrieval in a vector database. Right: CIFAR-10 (ResNet-9) examples showing the top-ranked training images under an influence-function estimator versus our information loss criterion.

sition rule in Bayesian experimental design and active learning (MacKay, 1992; Houlsby et al., 2011), information-theoretic analyses of generalization via mutual information (Xu and Raginsky, 2017; Steinke and Zakynthinou, 2020), and the information bottleneck view of representation learning and training dynamics (Shwartz-Ziv and Tishby, 2017). Closer in spirit, Bayesian viewpoints have studied TDA under training stochasticity, treating attribution scores as random variables rather than fixed point estimates (Nguyen et al., 2023), and data point selection through posterior inference over instance-wise weights (Xu et al., 2024). These perspectives are complementary to our query-specific information-theoretic formulation of subset attribution.

Contributions. *(i)* We pose TDA as *information loss* at a query, the entropy increase induced by withholding a subset. This can be implemented by instantiating a Gaussian Process surrogate from tangent features of the network (the empirical NTK), leading to a closed-form expression. *(ii)* We show that this criterion admits a principled *relaxation to information gain*: for fixed subset size and large observation noise, information loss and information gain share the same leading-order expansion. The score for a subset then depends only on the selected points themselves, not on the remaining training data. *(iii)* To make our algorithm amenable to fast lookup, we derive a *linear-response variance correction* that turns each step into a squared inner-product search between a residual query vector and (precomputed) sketched Jacobians. All inversions are performed in a low-dimensional sketch space, and we demonstrate how selection can be implemented via approximate nearest-neighbour search over a vector database (querying with both the residual and its negation, then merging top candidates), giving a computational footprint comparable to modern influence-based pipelines. *(iv)* Empirically, we evaluate three complementary settings: *leave-subset-out brittleness*—counterfactual sensitivity of predictions when top-ranked subsets are removed; *backdoor attribution*, where our method retrieves the ground-truth poisoned subset in controlled attacks; and *coreset selection*, where we test whether attributed subsets preserve downstream accuracy when used for retraining.

2 BACKGROUND

Setup. We consider supervised learning with training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, covering tasks from image classification to next-token prediction in language modeling. Let $f_{\boldsymbol{\theta}}$ denote a model with parameters $\boldsymbol{\theta} \in \mathbb{R}^P$ that is trained on $\mathcal{D}$ (*e.g.* by empirical risk minimization). Training data attribution (TDA) seeks to quantify the contribution of a subset $S \subset \mathcal{D}$ to the model's behaviour at a designated test query $\mathbf{x}_*$ (optionally with label y_*). We write this contribution as an *attribution score* $\mathcal{I}(S; \mathbf{x}_*)$, intentionally agnostic about the precise notion of "behaviour"—it may refer to loss, prediction, or another task-relevant functional of $f_{\boldsymbol{\theta}}$.

A counterfactual ideal and its relaxations. Let $\hat{\boldsymbol{\theta}}$ be the parameters obtained by training on $\mathcal{D}$ and $\hat{\boldsymbol{\theta}}^{\setminus S}$ the parameters obtained when the subset S is removed and the model is retrained under the same procedure. A natural "ground-truth" attribution for a loss-based notion of behaviour is the counterfactual change in test loss. Denoting $\mathbf{z}_* = (\mathbf{x}_*, y_*)$, we have,

$$\ell_*(\boldsymbol{\theta}) := \ell(\boldsymbol{\theta}; \mathbf{z}_*), \quad \mathcal{I}(S; \mathbf{x}_*) \triangleq \ell_*(\hat{\boldsymbol{\theta}}^{\setminus S}) - \ell_*(\hat{\boldsymbol{\theta}}), \quad (1)$$

which measures how the example-wise loss at $\mathbf{x}_*$ would differ if S were not available for training. This choice requires access to y_*; when labels are unavailable one can replace the loss with a prediction-based quantity, *e.g.* comparing model outputs $f_{\hat{\boldsymbol{\theta}}^{\setminus S}}(\mathbf{x}_*)$ and $f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_*)$ directly. Regardless of the choice of attribution score, computing (1) is expensive: identifying the most influential subset of size M,

$$\underset{S \subset \mathcal{D}:\, |S| = M}{\arg\max} \; \mathcal{I}(S; \mathbf{x}_*), \quad (2)$$

is combinatorial and intractable in general (except for the singleton case $M{=}1$). To avoid the cost, many practical TDA methods—most notably those based on first-order influence functions—adopt *additive* group scores. In this case, the subset attribution decomposes into pointwise terms and the size-M maximizer is obtained by simple top-M selection:

$$\mathcal{I}_{\mathrm{IF}}(S; \mathbf{x}_*) = \sum_{\mathbf{z}_i \in S} \mathcal{I}_{\mathrm{IF}}(\{\mathbf{z}_i\}; \mathbf{x}_*), \quad (3)$$

where $\mathbf{z}_i = (\mathbf{x}_i, y_i)$. This rank-and-pick rule is computationally attractive—pointwise scores can be precomputed once per test query and reused for any M—but it discards pairwise and higher-order interactions between training examples. The resulting redundancy (*e.g.* over-counting near-duplicates) has motivated extensions that account for interactions, such as higher-order influence functions (Basu et al., 2020). We return to this limitation when introducing our information-theoretic criteria in the next section.

3 INFORMATION-THEORETIC DATA ATTRIBUTION

Training Data Attribution (TDA) is fundamentally about quantifying and allocating information: which subsets of data best contribute to resolving uncertainty. The most principled toolkit for measuring information is information theory, which provides a mathematically rigorous framework for quantifying uncertainty and information in nats. A common critique of such an approach is its perceived lack of tractability, since information-theoretic objectives often involve intractable conditioning on large datasets. However, as we will show, the approximations developed later in this section reduce the implementation to efficient primitives–queries in a vector database and linear algebra over small matrices–bringing the framework into practical reach. With this motivation in place, we now introduce the information-theoretic framework underlying our approach, beginning with an ideal criterion and then developing its practical relaxation.

We propose to score a subset $S \subset \mathcal{D}$ at a test query $\mathbf{x}_*$ by the *information loss* (InfoLoss),

$$\mathcal{I}_{\mathrm{IL}}(S; \mathbf{x}_*) = -\mathcal{H}\big[f_* \,|\, \mathbf{x}_*, \mathcal{D}\big] + \mathcal{H}\big[f_* \,|\, \mathbf{x}_*, \mathcal{D} \setminus S\big], \quad (4)$$

that is, the amount of information (in nats) about the latent $f_* := f(\mathbf{x}_*)$ that would be lost if S were removed. Under a Gaussian model that we consider in this work, this directly corresponds to an increase in marginal uncertainty about the test query.

A Bayesian surrogate. In common practice with pretrained networks (or networks trained using standard learning algorithms such as SGD), parameters are available as a point estimate $\hat{\boldsymbol{\theta}}$. The mapping $\mathbf{x} \mapsto f_{\hat{\boldsymbol{\theta}}}(\mathbf{x})$ is therefore deterministic, rendering the entropy terms in Eq. (4) degenerate. To obtain meaningful information-theoretic quantities, we use a *Bayesian surrogate* that endows latent functions with uncertainty consistent with the local geometry of the trained model.

Neural networks as Gaussian processes. Our Bayesian surrogate is a Gaussian process induced by linearizing the network around $\hat{\boldsymbol{\theta}}$:

$$k(\mathbf{x}, \mathbf{x}') = \nabla_{\boldsymbol{\theta}} f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) \nabla_{\boldsymbol{\theta}} f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}')^\top \quad (5)$$

where $k(\cdot, \cdot)$ – often referred to as the (empirical) *neural tangent kernel* (NTK) (Jacot et al., 2018) – is a linear kernel with tangent features given by Jacobians of the neural network. In contrast to the infinite-width NTK at *initialization*—which converges to a deterministic kernel and remains constant during training under the NTK parameterization (Jacot et al., 2018; Yang, 2019)—evaluating the Jacobians at the *trained* $\hat{\boldsymbol{\theta}}$ ties the kernel to the representation the model actually uses. This is precisely what is justified by the linearized (Laplace) posterior around $\hat{\boldsymbol{\theta}}$ (Khan et al., 2019; Immer et al., 2021) and its function-space counterpart (Pan et al., 2020).

Since Eq. (4) depends only on the entropy of the latent f_* under the GP surrogate, it suffices to compute the marginal variance. For a Gaussian likelihood (squared-error regression) with noise variance σ^2, the standard Gaussian process regression (GPR) expressions apply (Rasmussen and Williams, 2006). Crucially, in this regression setting the marginal variance

is *independent of the targets*, so the entropy term—and therefore $\mathcal{I}_{\text{IL}}$—depends only on inputs, the kernel and σ^2.

To keep the objective and its updates simple and to retain the Gaussian setting, we recast classification as regression (*e.g.* with mean-centered one-hot targets) and use the GPR variance irrespective of the underlying likelihood. Under this surrogate, the entropy difference in Eq. (4) reduces to a log-variance ratio.

Relation to influence-based TDA. It turns out that in the singleton case ($|S| = 1$), Eq. (4) reduces to a form resembling a cross-influence term normalized by self-influence terms, *i.e.* the cosine-style normalization used in *RelatIF* (Barshan et al., 2020), especially when combined with a Gauss-Newton Hessian approximation. In App. D, we show RelatIF is label-independent (assuming a single-output case), which provides an explanation for its reported behaviour of selecting fewer "globally" influential examples that are often outliers or mislabelled examples. However, an important difference is that RelatIF is *signed*—it distinguishes positively vs. negatively influential examples—whereas our (singleton) score is *unsigned*.

Relaxation to information gain. While the information loss criterion in Eq. (4) is intuitively appealing, it has two critical issues. First, we have to condition on the near-complete dataset which is expensive. Second, information loss is not sub-modular and requires an intractable combinatorial optimization (we cannot expect a performance guarantee when optimising for it greedily). In principle, one could tackle the first problem in an analogous way to how recent linear system solvers handles Hessian-vector products (HVPs), that is replacing full-batch accumulations by sub-sampling as is done with the LiSSA scheme by Agarwal et al. (2017))[1]. However, the second problem is more challenging to address.

To sidestep these issues, we instead take a conceptually simpler approach and propose to use the *information gain* (InfoGain) criterion. This scores a subset $S \subset \mathcal{D}$ at a test query $\mathbf{x}_*$ by,

$$\mathcal{I}_{\text{IG}}(S; \mathbf{x}_*) = -\mathcal{H}\big[f_* \,|\, \mathbf{x}_*, S\big] + \mathcal{H}\big[f_* \,|\, \mathbf{x}_*\big], \qquad (6)$$

that is, the reduction (in nats) of the marginal uncertainty about $f_* := f(\mathbf{x}_*)$ if we trained only on S. Information gain is a standard acquisition criterion in Bayesian experimental design and active learning (MacKay, 1992); a key difference here is that we evaluate it directly at the test marginal rather than on parameters.

Crucially, for a fixed subset size and large observation noise, information loss and information gain share the same leading-order expansion. This makes InfoGain a principled high-noise relaxation of InfoLoss, formalized below.

Lemma 1. *Fix a query $\mathbf{x}_*$ and subset size M, and let $\mathcal{S}_M := \{S \subseteq \mathcal{D} : |S| = M\}$. Then, as $\sigma^2 \to \infty$,*

$$\mathcal{I}_{\text{IG}}(S; \mathbf{x}_*) = \frac{\|\mathbf{k}_{S*}\|^2}{2k_{**}\sigma^2} + \mathcal{O}(\sigma^{-4}),$$
$$\mathcal{I}_{\text{IL}}(S; \mathbf{x}_*) = \frac{\|\mathbf{k}_{S*}\|^2}{2k_{**}\sigma^2} + \mathcal{O}(\sigma^{-4}),$$

uniformly over $S \in \mathcal{S}_M$. In particular, if $\|\mathbf{k}_{S*}\|^2$ has a unique maximizer $S^\dagger$ over $\mathcal{S}_M$, then both criteria are maximized by $S^\dagger$ for sufficiently large σ^2.

For proof, see App. A. The lemma justifies using InfoGain as a high-noise surrogate for InfoLoss: we no longer need to condition on the whole dataset, and the new objective is sub-modular, allowing for the standard $1-\frac{1}{e}$ guarantee for greedy algorithms (Nemhauser et al., 1978).

Greedy algorithm for information gain. Justified by its sub-modularity, we now optimize Eq. (6) under the GP surrogate by greedy selection. Let $\mathcal{A}$ denote the set of acquired points (initially $\emptyset$). At each of the M rounds we pick the example whose inclusion most reduces the marginal variance at $\mathbf{x}_*$. Using the formula for inverse of a partitioned matrix, we can express this with $\mathcal{A}$ as conditioning set throughout:

$$\mathbf{x}^{(m)} \leftarrow \operatorname*{arg\,min}_{\mathbf{x} \in \mathcal{D} \setminus \mathcal{A}} v_*^{\mathcal{A} \cup \{\mathbf{x}\}} \qquad (7)$$

This is equivalent (see App. B for a full derivation) to

$$\mathbf{x}^{(m)} \leftarrow \operatorname*{arg\,max}_{\mathbf{x} \in \mathcal{D} \setminus \mathcal{A}} \frac{\left(v_{\mathbf{x},*}^{\mathcal{A}}\right)^2}{\sigma^2 + v_{\mathbf{x}}^{\mathcal{A}}}, \qquad (8)$$

where the covariance terms are given by:

$$v_{\mathbf{x},*}^{\mathcal{A}} = k_{\mathbf{x},*} - \mathbf{k}_{\mathcal{A},\mathbf{x}}^\top \left(\mathbf{K}_{\mathcal{A}} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{k}_{\mathcal{A},*}, \qquad (9)$$
$$v_{\mathbf{x}}^{\mathcal{A}} = k_{\mathbf{x},\mathbf{x}} - \mathbf{k}_{\mathcal{A},\mathbf{x}}^\top \left(\mathbf{K}_{\mathcal{A}} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{k}_{\mathcal{A},\mathbf{x}}. \qquad (10)$$

Here $k_{\cdot,\cdot} = k(\cdot,\cdot)$ are kernel evaluations from Eq. (5), $\mathbf{K}$ is the corresponding Gram matrix, and $\mathbf{I}$ is the identity. Intuitively, the numerator is the squared (conditional) covariance between $f_{\mathbf{x}}$ and f_*, and the denominator contains the marginal variance of the candidate point. After choosing $\mathbf{x}^{(m)}$, we update $\mathcal{A} \leftarrow \mathcal{A} \cup \{\mathbf{x}^{(m)}\}$ and iterate until $|\mathcal{A}| = M$.

[1] One could go even further and consider reducing the candidate space via heuristic prefiltering strategies (Grosse et al., 2023) or sparse kernel/GP techniques to obtain *landmark* points (*e.g.* leverage scores by Alaoui and Mahoney (2015)).

Scalability via Jacobian sketches. To apply the greedy rule, we must repeatedly evaluate the covariance terms, which involves per-example Jacobians in $\mathbb{R}^P$ for the kernel evaluations and inversions whose dimension scales with M. While computing per-example Jacobians is tractable on modern accelerators, storing $O(N)$ of them is prohibitive in memory at contemporary scales (*e.g.* pre-training corpora). We therefore adopt random-projection Jacobian sketches: draw a sketching matrix $\mathbf{A} \in \mathbb{R}^{K\times P}$ with $K \ll P$ from a sub-gaussian family (*e.g.*, Gaussian or Rademacher), and define

$$\boldsymbol{\phi}_{\mathbf{x}} = {}^{1}/\sqrt{K}\mathbf{A}\nabla_{\boldsymbol{\theta}} f_{\hat{\boldsymbol{\theta}}}(\mathbf{x})^{\top} \quad \text{so that} \quad k(\mathbf{x},\mathbf{x}') \approx \boldsymbol{\phi}_{\mathbf{x}}^{\top}\boldsymbol{\phi}_{\mathbf{x}'}, \tag{11}$$

which (by Johnson–Lindenstrauss) preserves pairwise inner products with high probability. Substituting these features into the "weight-space" form (via Woodbury formula) yields the scalable approximations to Eqs. (9) and (10),

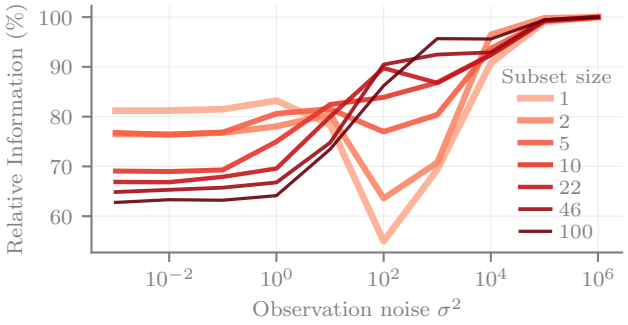
$$v^{\mathcal{A}}_{\mathbf{x},*} \approx \boldsymbol{\phi}_{\mathbf{x}}^{\top}\mathbf{S}_{\mathcal{A}}^{-1}\boldsymbol{\phi}_{*}, \quad v^{\mathcal{A}}_{\mathbf{x}} \approx \boldsymbol{\phi}_{\mathbf{x}}^{\top}\mathbf{S}_{\mathcal{A}}^{-1}\boldsymbol{\phi}_{\mathbf{x}}, \tag{12}$$

where

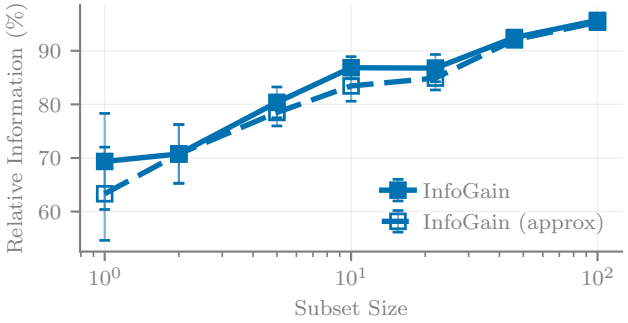
$$\mathbf{S}_{\mathcal{A}} := \tfrac{1}{\sigma^2}\sum_{\mathbf{x}'\in\mathcal{A}} \boldsymbol{\phi}_{\mathbf{x}'}\boldsymbol{\phi}_{\mathbf{x}'}^{\top} + \mathbf{I}. \tag{13}$$

so all dot-products and inversions are only in a K-dimensional space with $K \ll P$. Here $\mathbf{S}_{\mathcal{A}}$ is the posterior precision matrix of a Bayesian linear regressor built from the projected Jacobians $\boldsymbol{\phi}_{\mathbf{x}}$ on $\mathcal{A}$; consequently, removing a single example in subsequent greedy steps corresponds to a rank-one update, for which Sherman-Morrison formula yields efficient $O(K^2)$ maintenance of $\mathbf{S}_{\mathcal{A}}^{-1}$. This addresses the memory consideration (no need to retain full P-dimensional Jacobians) and maintains the reasonable compute (cheap solves in the sketch space). Practically, the projected Jacobians $\boldsymbol{\phi}_{\mathbf{x}}$ can be obtained without materializing full Jacobians, an approach advocated in recent scalable TDA pipelines (Choe et al., 2024; Schioppa, 2024).

Vector database. Even under the InfoGain relaxation, each greedy step in Eq. (8) still requires scanning all candidates in $\mathcal{D} \setminus \mathcal{A}$. Approximate nearest-neighbour (ANN) indices in high-performance vector databases (e.g., FAISS (Johnson et al., 2019)) can substantially reduce this cost by turning the search into vector similarity queries. This perspective has recently been drawn on to scale TDA to LLM regimes (Sun et al., 2025; Liu et al., 2025). However, such system optimizations only apply to primitives that reduce to vector similarity search. While a full database implementation is beyond our scope, we use this as a design constraint. In Eq. (8), the numerator $\left(v^{\mathcal{A}}_{\mathbf{x},*}\right)^2$ is amenable to this setting,[2] whereas the denominator $v^{\mathcal{A}}_{\mathbf{x}}$ is a candidate-specific quadratic form and is not directly expressible as a standard vector-search primitive.



(a) Information efficiency vs. observation noise



(b) Information efficiency vs. subset size

Figure 2: Information efficiency experiments on binary CIFAR-10. Top: selecting subsets via greedy InfoGain and varying the observation noise. Bottom: fixed observation noise level ($\sigma^2 = 10^3$), using InfoGain and approximate InfoGain.

Linear-response variance correction. In order to have an algorithm that can be implemented as queries to a vector database, we now derive a further approximation (see App. C for further details). Specifically, we derive a first-order (linear-response) perturbation of the test marginal variance in Eq. (7) by introducing a scalar weight on the candidate point and differentiating with respect to it, in the spirit of Giordano et al. (2018) (cf. Opper and Winther, 2003; Welling and Teh, 2004):

$$v_{*}^{\mathcal{A}\cup\{\mathbf{x}\}} \approx v_{*}^{\mathcal{A}} - \tfrac{1}{\sigma^2}\left(v^{\mathcal{A}}_{\mathbf{x},*}\right)^2. \tag{14}$$

Substituting this approximation into Eq. (7) removes the candidate-dependent denominator and yields the

[2] FAISS (Johnson et al., 2019) supports an absolute inner-product metric for some indexes, which preserves the ranking induced by a squared inner product. Otherwise, a squared inner product can be emulated via two *vanilla* inner-product queries.

following greedy rule for the *variance-corrected* information gain,

$$\mathbf{x}^{(*)} \leftarrow \arg\max_{\mathbf{x}\in\mathcal{D}\setminus\mathcal{A}} \left(\boldsymbol{\phi}_{\mathbf{x}}^{\top}\mathbf{r}_{*}^{\mathcal{A}}\right)^{2}, \tag{15}$$

where

$$\mathbf{r}_{*}^{\mathcal{A}} = \mathbf{S}_{\mathcal{A}}^{-1}\,\boldsymbol{\phi}_{*} = \boldsymbol{\phi}_{*} \;-\; \boldsymbol{\Phi}_{\mathcal{A}}^{\top}\left(\boldsymbol{\Phi}_{\mathcal{A}}\boldsymbol{\Phi}_{\mathcal{A}}^{\top} + \sigma^{2}\mathbf{I}\right)^{-1}\boldsymbol{\Phi}_{\mathcal{A}}\,\boldsymbol{\phi}_{*}.$$

Here $\boldsymbol{\Phi}_{\mathcal{A}}$ stacks the projected Jacobians for $\mathcal{A}$ as an $(M \times K)$ matrix. The second line shows that $\mathbf{r}_{*}^{\mathcal{A}}$ is a "residual" query obtained by removing from $\boldsymbol{\phi}_{*}$ the energy explained by the acquired set (a kernel-ridge projection). Operationally, each greedy step reduces to a squared inner-product search over the fixed database $\{\boldsymbol{\phi}_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{D}\setminus\mathcal{A}}$ with the single evolving query $\mathbf{r}_{*}^{\mathcal{A}}$—exactly the primitive supported by vector databases.

Approximation Quality. To evaluate how well InfoGain and approximate InfoGain with the linear-response correction track InfoLoss, we introduce a metric which we call relative information (Figure 2). Relative information tracks the fraction of information (measured in nats) recovered by approximations, relative to the information loss criterion (approximated by greedy selection for tractability purposes). We evaluate this on binary CIFAR-10 (class 0 vs. class 1) trained on a 3-layer MLP. Varying the observation noise (top row of Figure 2), InfoGain approaches InfoLoss across subset sizes, converging near 100% with $\sigma^2 \geq 10^4$. This is consistent with the high-noise asymptotic equivalence in Lemma 1. At a fixed noise level ($\sigma^2 = 10^3$, bottom row of Figure 2), the approximate InfoGain closely tracks exact InfoGain across subset budgets, indicating that the variance-correction introduces minimal degradation while enabling vector-database-friendly selection.

Computational footprint. This approximate InfoGain pipeline requires one extra pass over the training set to compute a K-dimensional projected Jacobian for each of the N training examples, which has the same order of cost as one training epoch and is comparable to influence-based pipelines that likewise rely on a preprocessing pass to compute projected features or approximate curvature factors. Storing these features costs $O(NK)$ memory. For a query and subset budget M, maintaining the sketch-space precision matrix and residual via rank-one updates costs $O(MK^2)$ and is independent of N; the only remaining N-dependent operation is candidate retrieval. In the most naive implementation this retrieval is $O(NK)$ per greedy step, whereas ANN indices can avoid full scans and make the search sublinear in N.

Summary of methods and practical guidance. We have developed three complementary instantiations of our information-theoretic criterion, which differ in how they balance faithfulness to the leave-subset-out counterfactual objective, greedy optimization guarantees and scalability. Greedy InfoLoss most directly matches the counterfactual notion of entropy increase when subsets are withheld, but it is not submodular and therefore lacks greedy optimization guarantees. Greedy InfoGain is a principled high-noise relaxation whose leading-order term matches InfoLoss, while enabling submodular optimization with the standard $(1-1/e)$ guarantee. Finally, approximate greedy InfoGain adds the linear-response variance correction, reducing each step to a squared inner-product query against a vector database. In practice, this makes approximate greedy InfoGain approx the default choice: Fig. 2 shows that it closely tracks exact greedy InfoGain, while being the only variant that maps directly to standard vector-database primitives and thus cleanly scales to very large candidate pools. Exact greedy InfoGain is preferable when the candidate set is moderate and one wants the exact submodular objective, whereas InfoLoss is most appropriate when staying as close as possible to the leave-subset-out counterfactual definition matters more than greedy optimization guarantees or retrieval efficiency.

4 EXPERIMENTS

We evaluate our Bayesian information-theoretic attribution methods—greedy[3] *InfoLoss*, greedy *InfoGain*, and approximate greedy *InfoGain* (implementable with a vector database)—on three complementary tasks: (i) *leave-subset-out brittleness*, which probes how well an attribution method identifies training examples whose removal most degrades performance; (ii) *retrieving ground-truth attribution via a synthetic backdoor*, which provides a controlled setting with known influential training instances; and (iii) *coreset selection*, which asks whether attributed subsets suffice for accurate retraining. All experiments operate in a *final-model-only* regime (Wei et al., 2024) where only the final checkpoint is available and methods must avoid retraining the baseline model for scoring (for instance as done in Choe et al. (2024)). We treat the observation noise σ^2 in our attribution methods as a hyperparameter and tune this on a held-out split using a log-scaled grid from 10^{-6} to 10^{6} (see App. E for full details).

Baselines. We compare against standard TDA baselines used in prior work: RANDOM, representational similarity (REPSIM / cosine similarity on penultimate-layer features) (Hanawa et al., 2021), gra-

[3]While information loss is not sub-modular (unlike information gain), we can still optimize it greedily.

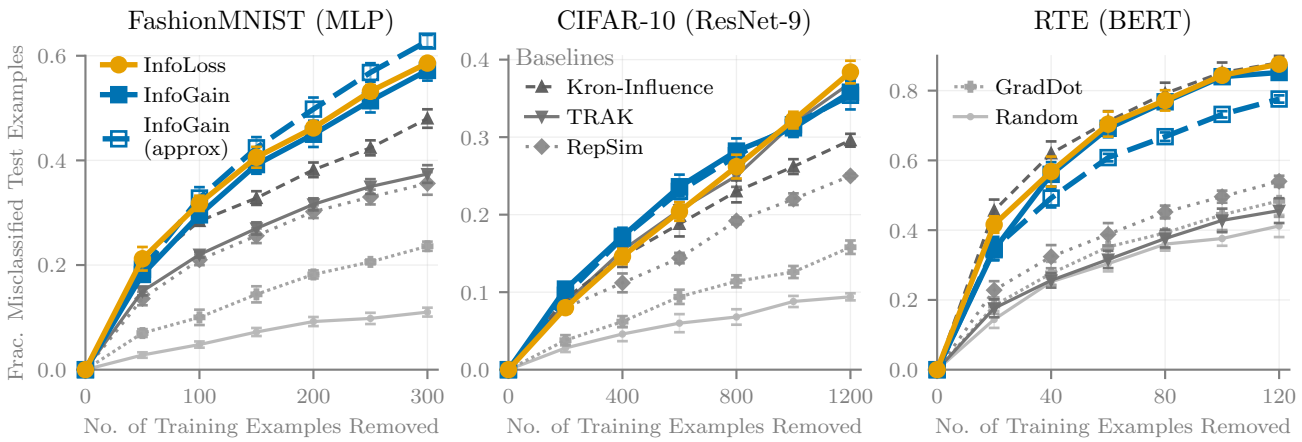


Figure 3: Our Bayesian information-theoretic methods—InfoLoss, InfoGain, and InfoGain (approx)—recover strong attribution signal in identified subsets, as reflected by higher brittleness. We plot the fraction of previously correct test queries that become misclassified after removing the *same-label* training examples attributed by each method and retraining. Left: Fashion-MNIST (MLP), where our methods dominate and KronInfluence is the closest baseline. Middle: CIFAR-10 (ResNet-9), where our methods lead across budgets with TRAK most competitive. Right: RTE (BERT), where KronInfluence is strongest at smaller budgets, while InfoLoss/InfoGain catch up and are competitive at larger budgets. Error bars represent standard error across repeated runs.

dient dot-product (GradDot / *i.e.* TracIn with final checkpoint only) (Pruthi et al., 2020), TRAK (Park et al., 2023), and KronInfluence (influence function with EK-FAC approximation) (Grosse et al., 2023)[4].

Implementation. We avoid explicit multi-output Jacobians by working with a scalar measurement function. For brittleness and backdoor retrieval, we compute training-example Jacobians with respect to the logit corresponding to the query's ground-truth class (rather than Jacobians of all logits). For the coreset experiment, where attribution is defined with respect to a set of queries, we instead use the same multiclass reduction as TRAK. All Jacobians are projected using a Rademacher sketch, shared across projection-based methods (ours and TRAK) within each dataset; we use projection dimension 4096 for CIFAR-10/Fashion-MNIST and 20480 for RTE. We reuse the efficient CUDA Jacobian/sketch implementation from Park et al. (2023). Given the sketched features, our methods solve a Bayesian linear regression in closed form; all variance computations use a numerically stable Cholesky factorization. For the vision models in brittleness/backdoor (CIFAR-10 and Fashion-MNIST), before computing Jacobians we convert the trained network to the NTK parameterization (Jacot et al., 2018), which preserves predictions but introduces width-dependent scaling of Jacobians; we apply this reparameterization *post hoc* solely for attribution (we do not train in NTK parameterization). For the CIFAR-10 coreset experiment, we keep the original parameterization. For BERT on RTE, we do not apply NTK parameterization and restrict Jacobians to linear layers.

[4]For the brittleness task, influence-based methods directly target loss increases upon removal, whereas our methods target information-theoretic criteria that are only indirectly related to accuracy. We nevertheless include this evaluation because it has become standard in the attribution literature.

4.1 Leave-subset-out brittleness

Setup. We follow the subset-removal counterfactual protocol first introduced in Singla et al. (2023) but the experimental setup is adapted from Bae et al. (2024). We evaluate on CIFAR-10 and Fashion-MNIST (the latter subsampled to 25% of the original training set) using a ResNet-9 and MLP respectively, and on GLUE-RTE using BERT (see App. E for further details). In contrast to Bae et al. (2024), the candidate pool for subset selection is restricted to *same-class* training examples as the query, following Singla et al. (2023), to mitigate the bias of certain methods towards attribution in the same-class (Hanawa et al., 2021).

Evaluation protocol. Here we describe the details of our evaluation protocol:

(1) Train a base model on the full training set.

(2) From the test split, identify correctly classified examples as queries (100 for CIFAR-10/Fashion-MNIST and 50 for RTE).

Figure 4: Examples of backdoored CIFAR-10 images with triggers. The trigger is a 3×3 black/white random pattern placed at the bottom-right corner. This causes the model to misclassify the image to the corrupted class.

(3) For each query:

   (a) Perform attribution according to each method (for a fixed subset size).

   (b) Remove the attributed subset and retrain from scratch (performed for the number of query points).

   (c) Predict on the query point and flag if misclassification occurs.

(4) Finally we report the proportion of all query points that are misclassified. This is repeated for a linearly-spaced grid of subset sizes (if a query point becomes misclassified for a given subset size then this is terminated early).

Results. In Fig. 3, we observe that across all three datasets our Bayesian information-theoretic methods perform strongly on the brittleness metric. On Fashion-MNIST (left), InfoLoss, InfoGain, and InfoGain (approx) all induce substantially more counterfactual errors than the baselines at every removal budget, with the gap widening as the subset size grows; InfoGain (approx) is typically the top curve and closely tracks (or slightly exceeds) the exact InfoGain. Among baselines, KronInfluence is the most competitive on Fashion-MNIST. On CIFAR-10 (middle), our methods again lead throughout; InfoGain is strongest at small–mid budgets while InfoLoss edges out others at the largest budget. Here, however, TRAK performs very close to our methods, with curves that mostly overlap. On RTE with BERT (right), the strongest baseline (KronInfluence) is highly competitive and tends to lead at smaller budgets, while InfoLoss and InfoGain close the gap and are competitive at larger budgets; InfoGain (approx) remains weaker than its exact counterpart but still outperforms the other baselines. Despite influence-derived approaches (*i.e.* TRAK and KronInfluence) being tailored to loss increases, our information-centric objectives remain competitive and are often superior, indicating that the information captured by our scores translates into substantial counterfactual sensitivity under subset removal (although the same-class candidate restriction may also amplify this effect).

Method	Recall@50 (↑)	MRR (↑)
Random	0.011 ± 0.003	0.045 ± 0.016
GradDot	0.010 ± 0.002	0.217 ± 0.033
RepSim	0.985 ± 0.004	**1.000** ± 0.000
TRAK	0.009 ± 0.002	0.189 ± 0.036
KronInfluence	0.058 ± 0.011	0.504 ± 0.034
InfoLoss	**1.000** ± 0.000	**0.999** ± 0.001
InfoGain	0.974 ± 0.007	**0.998** ± 0.001
InfoGain (approx)	0.974 ± 0.007	**0.998** ± 0.001

Table 1: Our Bayesian information-theoretic method InfoLoss retrieves the ground-truth backdoor examples almost perfectly on CIFAR-10—achieving the highest recall and matching the best baseline on MRR (no significant difference). Moreover, InfoGain and InfoGain (approx) show no significant difference from the best methods on MRR, and their recall is close to the best baseline. We compare standard TDA baselines to our proposed methods (shaded). We denote the best within each block (Baselines / Proposed) with underline and bold denotes the best overall. Reported metrics are accompanied by standard error from repeated runs.

4.2 Retrieving ground-truth attribution via backdoor

While brittleness measures whether the subsets identified by an attribution method matter counterfactually, it does not tell us whether the retrieved examples are truly the ones driving the prediction. To test that more directly, we next turn to a synthetic backdoor setting where the relevant training instances are known by construction.

Backdoor design. We use CIFAR-10 with a class-conditional (cyclic) BadNets (Gu et al., 2019) rule: for base class k, a backdoored image is relabelled to $(k+1) \bmod 10$. The backdoor trigger is a 3×3 black/white randomly-generated grid placed at the bottom-right corner, that is unique to each class (see Fig. 4). This is applied to 1% of the training set, distributed evenly across classes (50 per class). With this configuration, the attack success rate on the poisoned test set is near-perfect whilst the clean test accuracy is unchanged.

Retrieval task and metrics. For each backdoored test query from base class k, the ground-truth attribution set are those poisoned training examples whose

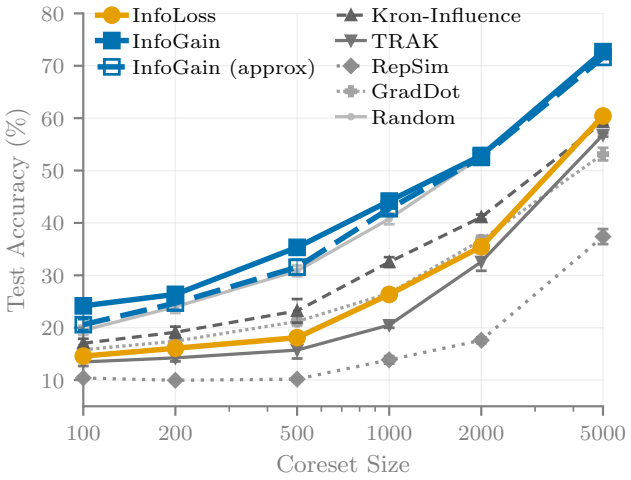


Figure 5: InfoGain and InfoGain (approx) construct substantially better CIFAR-10 coresets than existing TDA baselines which perform worse than random selection. We plot test accuracy after retraining on the selected subset for coreset sizes ranging from 0.2% to 10% of the training set. Error bars show standard error across repeated runs.

base class is k. We report the standard IR metrics, *recall@50* and mean reciprocal rank (*MRR*), averaged over 100 queries (see App. E for further details). All other aspects of the experimental setup are identical to the brittleness experiment on CIFAR-10.

Results. In Table 1, InfoLoss attains near-perfect retrieval and yields significantly higher recall than all baselines in this synthetic backdoor-retrieval task. For ranking quality (MRR), the top methods—RepSim, InfoLoss, and both InfoGain variants—are statistically indistinguishable. The InfoGain and InfoGain (approx) methods also deliver performance close to the strongest baseline while maintaining MRR on par with the best methods. In contrast, influence-based TDA baselines struggle to retrieve the true influential training points, underscoring the advantage of our Bayesian information-theoretic approach for this retrieval setting.

4.3 Coreset selection

We next consider coreset selection through the lens of TDA: identifying a small training subset that still supports accurate retraining.

Setup. We consider CIFAR-10 with ResNet-9 and coresets ranging from 0.2% to 10% of the training set. We extract 500 examples from the test set as the query set for attribution, and report test accuracy on the remaining held-out test examples after retraining on the selected coreset. For the TDA baselines (TRAK, KronInfluence, GradDot, and RepSim), which do not explicitly define attribution to multiple query points at once, we average the per-query attribution scores before selecting the top-M examples. This is consistent with the usual additivity assumption in influence-based TDA and with the linear datamodelling score used to evaluate such methods. App. E gives the full protocol and derives the tractable multi-query greedy selection criterion used by our methods.

Results. Fig. 5 shows that InfoGain is the strongest method across essentially the entire budget range, with the clearest gains at small coresets. At $M = 500$ (1% of the training set), InfoGain reaches 35.3% accuracy, compared with 30.8% for Random, 23.2% for KronInfluence, and 15.7% for TRAK; at $M = 1000$, the gap over Random remains substantial (44.2% vs. 40.8%). InfoGain (approx) closely tracks the exact greedy variant, especially at larger budgets. In contrast, InfoLoss is markedly weaker on this task, which is consistent with its leave-subset-out objective being tailored to identifying removals that maximally disrupt a prediction rather than subsets that are jointly informative for many queries.

The TDA baselines all underperform Random, indicating that influence-based rankings do not cleanly transfer to coreset construction. At the same time, a comparison with class-balanced variants (see App. E) shows that much of this degradation is due to imbalance in the selected subsets: enforcing per-class quotas substantially improves all TDA baselines and often lifts the strongest ones above greedy InfoLoss. Nevertheless, even after this correction they still remain below both InfoGain variants.

5 CONCLUSION

We proposed a Bayesian information-theoretic approach to training data attribution (TDA), framing attribution as the information loss incurred when subsets of training data are withheld. Using Gaussian process surrogates from neural tangent features, we derived closed-form estimators and scalable relaxations to information gain, enabling efficient retrieval with Jacobian sketches and vector database queries. Empirical results on brittleness, backdoor retrieval and coreset selection show that our methods reliably identify influential subsets, recover ground-truth poisoned data and construct small training sets that preserve downstream accuracy, while maintaining computational efficiency comparable to influence-based baselines. This work bridges principled information measures with scalable attribution, offering a foundation for future applications in auditing, safety, and interpretability.

Acknowledgements

DT would like to thank Alice Wang and the other members of the Simplex Lab at Spotify. DT also acknowledges Emtiyaz Khan and Eric Nalisnick for discussions around sensitivity-based variance estimates.

References

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-Order Stochastic Optimization for Machine Learning in Linear Time. *Journal of Machine Learning Research*, 2017.

Ahmed Alaoui and Michael W Mahoney. Fast Randomized Kernel Ridge Regression with Statistical Guarantees. *Advances in Neural Information Processing Systems*, 2015.

Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger Grosse. Training Data Attribution via Approximate Unrolled Differentiation. *ArXiv e-Prints*, 2024.

Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. RelatIF: Identifying Explanatory Training Examples via Relative Influence. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Samyadeep Basu, Xuchen You, and Soheil Feizi. On Second-Order Group Influence Functions for Black-Box Predictions. In *International Conference on Machine Learning*, 2020.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, et al. What is Your Data Worth to GPT? LLM-Scale Data Valuation with Influence Functions. *ArXiv e-Prints*, 2024.

R Dennis Cook and Sanford Weisberg. Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics*, 1980.

Ryan Giordano, Tamara Broderick, and Michael I Jordan. Covariances, Robustness, and Variational Bayes. *Journal of Machine Learning Research*, 2018.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying Large Language Model Generalization with Influence Functions. *ArXiv e-Prints*, 2023.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, 2019.

Frank R Hampel. The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, 1974.

Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, and Kentaro Inui. Evaluation of similarity-based explanations. In *International Conference on Learning Representations*, 2021.

Satoshi Hara, Atsushi Nitanda, and Takanori Maehara. Data Cleansing for Models Trained with SGD. In *Advances in Neural Information Processing Systems*, 2019.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *ArXiv e-Prints*, 2011.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of Bayesian neural nets via local linearization. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Advances in Neural Information Processing Systems*, 2018.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 2019.

Mohammad Emtiyaz Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate Inference Turns Deep Networks into Gaussian Processes. *Advances in Neural Information Processing Systems*, 2019.

Pang Wei Koh and Percy Liang. Understanding Black-Box Predictions via Influence Functions. In *International Conference on Machine Learning*, 2017.

Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite Versus Infinite Neural Networks: an Empirical Study. *Advances in Neural Information Processing Systems*, 2020.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, Yen-Sung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Yu-Yen Cheng, Karen Farley, Taira Anderson, David Albright, Carissa Schoenick, Luca Soldaini, Dirk Groeneveld, Rock Yuren Pang, Pang Wei Koh, Noah A. Smith, Sophie Lebrecht, Yejin Choi, Hannaneh Hajishirzi, Ali Farhadi, and Jesse Dodge. OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Association for Computational Linguistics, 2025.

David JC MacKay. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 1992.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 1978.

Elisa Nguyen, Minjoon Seo, and Seong Joon Oh. A Bayesian Approach to Analysing Training Data Attribution in Deep Learning. *Advances in Neural Information Processing Systems*, 2023.

Manfred Opper and Ole Winther. Variational Linear Response. *Advances in Neural Information Processing Systems*, 16, 2003.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner, and Mohammad Emtiyaz Khan. Continual Deep Learning by Functional Regularisation of Memorable Past. *Advances in Neural Information Processing Systems*, 2020.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing Model Behavior at Scale. *ArXiv e-Prints*, 2023.

Daryl Pregibon. Logistic Regression Diagnostics. *The Annals of Statistics*, 1981.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems*, 2020.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Andrea Schioppa. Efficient Sketches for Training Data Attribution and Studying the Loss Landscape. *Advances in Neural Information Processing Systems*, 2024.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling Up Influence Functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of Deep Neural Networks via Information. *ArXiv e-Prints*, 2017.

Vasu Singla, Pedro Sandoval-Segura, Micah Goldblum, Jonas Geiping, and Tom Goldstein. A Simple and Efficient Baseline for Data Attribution on Images. *ArXiv e-Prints*, 2023.

Thomas Steinke and Lydia Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. In *Conference on Learning Theory*, 2020.

Weiwei Sun, Haokun Liu, Nikhil Kandpal, Colin Raffel, and Yiming Yang. Enhancing Training Data Attribution with Representational Optimization. *ArXiv e-Prints*, 2025.

Sellamanickam Sundararajan and Sathiya Keerthi. Predictive app roaches for choosing hyperparameters in gaussian processes. *Advances in neural information processing systems*, 12, 1999.

Dennis Wei, Inkit Padhi, Soumya Ghosh, Amit Dhurandhar, Karthikeyan Natesan Ramamurthy, and Maria Chang. Final-Model-Only Data Attribution with a Unifying View of Gradient-Based Methods. *ArXiv e-Prints*, 2024.

Max Welling and Yee Whye Teh. Linear Response Algorithms for Approximate Inference in Graphical Models. *Neural Computation*, 2004.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 2017.

Xinnuo Xu, Minyoung Kim, Royson Lee, Brais Martinez, and Timothy Hospedales. A Bayesian Approach to Data Point Selection. *Advances in Neural Information Processing Systems*, 2024.

Greg Yang. Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation. *ArXiv e-Prints*, 2019.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Sections 2 and 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] See Section 3 (Computational footprint paragraph, Lemma 1) and Appendices A–D.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] Will be released at a later stage.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Lemma 1 in Section 3 and Appendices A–D.
   (b) Complete proofs of all theoretical results. [Yes] See Appendices A–D.
   (c) Clear explanations of any assumptions. [Yes] See Section 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] See supplementary material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Section 4 and Appendix E.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] All experiments are repeated 5 times with different initializations and seeds; we report standard error (see Section 4 and Appendix E).
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See supplementary material.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] All datasets (CIFAR-10, Fashion-MNIST, GLUE-RTE) and external codebases (TRAK, KronInfluence) are cited.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Bayesian Information-Theoretic Approach to Data Attribution – Supplementary Materials –

A HIGH-NOISE ASYMPTOTIC EQUIVALENCE BETWEEN INFORMATION LOSS AND INFORMATION GAIN

The goal of this section is to prove the following Lemma:

Lemma 1. *Fix a query $\mathbf{x}_*$ and subset size M, and let $\mathcal{S}_M := \{S \subseteq \mathcal{D} : |S| = M\}$. Then, as $\sigma^2 \to \infty$,*

$$\mathcal{I}_{\mathrm{IG}}(S; \mathbf{x}_*) = \frac{\|\mathbf{k}_{S*}\|^2}{2k_{**}\sigma^2} + \mathcal{O}(\sigma^{-4}),$$
$$\mathcal{I}_{\mathrm{IL}}(S; \mathbf{x}_*) = \frac{\|\mathbf{k}_{S*}\|^2}{2k_{**}\sigma^2} + \mathcal{O}(\sigma^{-4}),$$

uniformly over $S \in \mathcal{S}_M$. In particular, if $\|\mathbf{k}_{S*}\|^2$ has a unique maximizer $S^\dagger$ over $\mathcal{S}_M$, then both criteria are maximized by $S^\dagger$ for sufficiently large σ^2.

Proof. Fix the subset size M and define $\mathcal{S}_M := \{S \subseteq \mathcal{D} : |S| = M\}$. For notational convenience, write $a(S) := \|\mathbf{k}_{S*}\|^2$, use $v_*(\cdot)$ for the marginal variance of $f_* = f(\mathbf{x}_*)$ as a function of the conditioning set, and define $\bar{S} = \mathcal{D} \setminus S$ as the complement of the training subset. Then we can write the criteria as,

$$\mathcal{I}_{\mathrm{IG}}(S; \mathbf{x}_*) = -\frac{1}{2} \log \frac{v_*(S)}{k_{**}}, \tag{16}$$
$$\mathcal{I}_{\mathrm{IL}}(S; \mathbf{x}_*) = \frac{1}{2} \log \frac{v_*(\bar{S})}{v_*(\mathcal{D})}. \tag{17}$$

where we consider the marginal variances $v_*(S)$ and $v_*(\bar{S})$ under an underlying GP model without prescribing the kernel further. Using the Neumann series $(\mathbf{I}+\mathbf{A})^{-1} = \sum_{k=0}^{\infty}(-1)^k \mathbf{A}^k$ for which convergence is assured when $\rho(\mathbf{A}) < 1$ where $\rho(\cdot)$ is the spectral radius, we can show:

$$\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1} = \sigma^{-2}\left(\mathbf{I} + \sigma^{-2}\mathbf{K}\right)^{-1} \tag{18}$$
$$= \sigma^{-2}\left(\mathbf{I} - \sigma^{-2}\mathbf{K} + \sigma^{-4}\mathbf{K}^2 + \ldots\right) \tag{19}$$
$$= \sigma^{-2}\mathbf{I} + \mathcal{O}(\sigma^{-4}) \tag{20}$$

where the remainder is entrywise of order σ^{-4} and the expansion is valid whenever $\rho(\sigma^{-2}\mathbf{K}) < 1$, which holds for sufficiently large σ^2. Because $\mathcal{S}_M$ is finite, all $\mathcal{O}(\sigma^{-4})$ terms below can be taken uniformly over $S \in \mathcal{S}_M$. Using Eq. (20), we can write the marginal variance $v_*(S)$ as,

$$v_*(S) = k_{**} - \mathbf{k}_{S*}^\top \left(\mathbf{K}_S + \sigma^2\mathbf{I}_M\right)^{-1} \mathbf{k}_{S*} \tag{21}$$
$$= k_{**} - \sigma^{-2}a(S) + \mathcal{O}(\sigma^{-4}). \tag{22}$$

Hence,

$$\mathcal{I}_{\mathrm{IG}}(S; \mathbf{x}_*) = -\frac{1}{2} \log \left(1 - \frac{a(S)}{k_{**}\sigma^2} + \mathcal{O}(\sigma^{-4})\right) \tag{23}$$
$$= \frac{a(S)}{2k_{**}\sigma^2} + \mathcal{O}(\sigma^{-4}), \tag{24}$$

where the second line uses the Taylor expansion $\log(1+u) = u + \mathcal{O}(u^2)$ together with $u = \mathcal{O}(\sigma^{-2})$. For $v_*(\bar{S})$, we first turn to the formula for the leave-M-out marginal variance (see derivation in App. A.1),

$$v_*(\bar{S}) = v_*(\mathcal{D}) + [\mathbf{K}_y^{-1}\mathbf{k}_{\mathcal{D}*}]_S^\top \left([\mathbf{K}_y^{-1}]_{S,S}\right)^{-1} [\mathbf{K}_y^{-1}\mathbf{k}_{\mathcal{D}*}]_S \tag{25}$$

where we denote $\mathbf{K}_y := \mathbf{K}_{\mathcal{D}\mathcal{D}} + \sigma^2\mathbf{I}_N$ and overload S as an index in order to extract entries corresponding to rows/columns of S in $\mathcal{D}$. Again using Eq. (20), we can expand the two factors on the right-hand side of Eq. (25). First,

$$[\mathbf{K}_y^{-1}]_{S,S} = \sigma^{-2}\mathbf{I}_M - \sigma^{-4}\mathbf{K}_{SS} + \mathcal{O}(\sigma^{-6}) \tag{26}$$

$$\implies \left([\mathbf{K}_y^{-1}]_{S,S}\right)^{-1} = \sigma^2\left(\mathbf{I}_M - \sigma^{-2}\mathbf{K}_{SS} + \mathcal{O}(\sigma^{-4})\right)^{-1} \tag{27}$$

$$= \sigma^2\mathbf{I}_M + \mathbf{K}_{SS} + \mathcal{O}(\sigma^{-2}). \tag{28}$$

and second,

$$[\mathbf{K}_y^{-1}\mathbf{k}_{\mathcal{D}*}]_S = \sigma^{-2}\left[\mathbf{I}_N - \sigma^{-2}\mathbf{K}_{\mathcal{D}\mathcal{D}} + \mathcal{O}(\sigma^{-4})\right]_{S,:}\mathbf{k}_{\mathcal{D}*} \tag{29}$$

$$= \sigma^{-2}\left(\mathbf{k}_{S*} - \sigma^{-2}\mathbf{K}_{S\mathcal{D}}\mathbf{k}_{\mathcal{D}*} + \mathcal{O}(\sigma^{-4})\right) \tag{30}$$

$$= \sigma^{-2}\mathbf{k}_{S*} + \mathcal{O}(\sigma^{-4}). \tag{31}$$

Substituting Eqs. (28) and (31) into Eq. (25) we obtain,

$$v_*(\bar{S}) - v_*(\mathcal{D}) = \left(\sigma^{-2}\mathbf{k}_{S*} + \mathcal{O}(\sigma^{-4})\right)^\top \left(\sigma^2\mathbf{I}_M + \mathbf{K}_{SS} + \mathcal{O}(\sigma^{-2})\right)\left(\sigma^{-2}\mathbf{k}_{S*} + \mathcal{O}(\sigma^{-4})\right) \tag{32}$$

$$= \sigma^{-2}\mathbf{k}_{S*}^\top\mathbf{k}_{S*} + \mathcal{O}(\sigma^{-4}) \tag{33}$$

$$= \sigma^{-2}a(S) + \mathcal{O}(\sigma^{-4}). \tag{34}$$

Moreover, applying Eq. (22) with $S = \mathcal{D}$ yields

$$v_*(\mathcal{D}) = k_{**} + \mathcal{O}(\sigma^{-2}). \tag{35}$$

Therefore,

$$\mathcal{I}_{\mathrm{IL}}(S;\mathbf{x}_*) = \frac{1}{2}\log\left(1 + \frac{v_*(\bar{S}) - v_*(\mathcal{D})}{v_*(\mathcal{D})}\right) \tag{36}$$

$$= \frac{1}{2}\log\left(1 + \frac{a(S)}{k_{**}\sigma^2} + \mathcal{O}(\sigma^{-4})\right) \tag{37}$$

$$= \frac{a(S)}{2k_{**}\sigma^2} + \mathcal{O}(\sigma^{-4}), \tag{38}$$

where in the second line we used Eqs. (34) and (35). Eqs. (24) and (38) prove the shared leading-order expansion.

For the final claim, suppose $a(S)$ has a unique maximizer $S^\dagger$ over $\mathcal{S}_M$ and define the gap

$$\Delta := a(S^\dagger) - \max_{S\in\mathcal{S}_M,\ S\neq S^\dagger} a(S) > 0. \tag{39}$$

Since $k_{**} > 0$, by uniformity of Eqs. (24) and (38) there exist constants $C > 0$ and $\sigma^2_{\mathrm{asym}} > 0$ such that for both criteria $\mathcal{J} \in \{\mathcal{I}_{\mathrm{IG}}, \mathcal{I}_{\mathrm{IL}}\}$ we may write,

$$\mathcal{J}(S;\mathbf{x}_*) = \frac{a(S)}{2k_{**}\sigma^2} + r_{\mathcal{J},\sigma}(S), \qquad |r_{\mathcal{J},\sigma}(S)| \le C\sigma^{-4} \tag{40}$$

for all $S \in \mathcal{S}_M$ and $\sigma^2 \ge \sigma^2_{\mathrm{asym}}$. Now fix any $S \neq S^\dagger$. Then for either criterion $\mathcal{J}$,

$$\mathcal{J}(S^\dagger;\mathbf{x}_*) - \mathcal{J}(S;\mathbf{x}_*) = \frac{a(S^\dagger) - a(S)}{2k_{**}\sigma^2} + r_{\mathcal{J},\sigma}(S^\dagger) - r_{\mathcal{J},\sigma}(S) \tag{41}$$

$$\ge \frac{a(S^\dagger) - a(S)}{2k_{**}\sigma^2} - 2C\sigma^{-4} \tag{42}$$

$$\ge \frac{\Delta}{2k_{**}\sigma^2} - 2C\sigma^{-4} \tag{43}$$

The lower bound in Eq. (43) is positive whenever

$$\frac{\Delta}{2k_{**}\sigma^2} - 2C\sigma^{-4} > 0 \qquad \iff \qquad \sigma^2 > \frac{4Ck_{**}}{\Delta}. \tag{44}$$

Therefore, if

$$\sigma^2 > \sigma^2_{\text{crit}} := \max\left\{\sigma^2_{\text{asym}}, \frac{4Ck_{**}}{\Delta}\right\}, \tag{45}$$

then $\mathcal{J}(S^\dagger; \mathbf{x}_*) - \mathcal{J}(S; \mathbf{x}_*) > 0$ for every $S \neq S^\dagger$ and either criterion $\mathcal{J}$. Hence both $\mathcal{I}_{\text{IG}}(\cdot; \mathbf{x}_*)$ and $\mathcal{I}_{\text{IL}}(\cdot; \mathbf{x}_*)$ are uniquely maximized by $S^\dagger$ for all $\sigma^2 > \sigma^2_{\text{crit}}$. This proves the lemma.

A.1 Derivation of leave-M-out marginal variance

This is a generalization of the leave-one-out expression commonly used in cross-validation of GPs (Sundararajan and Keerthi, 1999) (also see Ch. 5.4.2 in Rasmussen and Williams (2006)) but in addition to leaving out multiple points, here we are concerned with evaluating variance at a test point rather than the omitted point(s). The derivation proceeds by exploiting the partitioned matrix inversion lemma to relate the marginal variance on the full data $v_*(\mathcal{D})$ to the leave-M-out marginal variance $v_*(\bar{S})$. To recap the marginal variance on the full data is given by,

$$v_*(\mathcal{D}) = k_{**} - \mathbf{k}_{\mathcal{D}*}^\top \mathbf{K}_y^{-1} \mathbf{k}_{\mathcal{D}*}, \quad \text{where } \mathbf{K}_y := \mathbf{K}_{\mathcal{D}\mathcal{D}} + \sigma^2 \mathbf{I}_N. \tag{46}$$

We consider a partition of $\mathbf{K}_y$ into the following block matrix with the first block corresponding to the retained points $\bar{S}$ and the second block corresponding to the omitted points S (the expression for marginal variance allows for arbritrary ordering of the points so we can re-order the points as necessary):

$$\mathbf{K}_y = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix} \quad \text{with} \quad \begin{aligned} \mathbf{A} &= \mathbf{K}_{\bar{S}\bar{S}} + \sigma^2 \mathbf{I}_{N-M} \\ \mathbf{B} &= \mathbf{K}_{\bar{S}S} \\ \mathbf{C} &= \mathbf{K}_{SS} + \sigma^2 \mathbf{I}_M \end{aligned} \tag{47}$$

Then using the partitioned matrix inversion lemma, we have,

$$\mathbf{K}_y^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{M}\mathbf{B}^\top\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{M} \\ -\mathbf{M}\mathbf{B}^\top\mathbf{A}^{-1} & \mathbf{M} \end{bmatrix} \tag{48}$$

where $\mathbf{M}^{-1} = \mathbf{C} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}$ is the Schur complement of $\mathbf{A}$ in $\mathbf{K}_y$. Substituting Eq. (48) into the expression for $v_*(\mathcal{D})$ given in Eq. (46) and expanding $\mathbf{k}_{\mathcal{D}*}^\top = [\mathbf{k}_{\bar{S}*}\ \mathbf{k}_{S*}]^\top$ we have,

$$\begin{aligned} v_*(\mathcal{D}) &= k_{**} - \mathbf{k}_{\bar{S}*}^\top \mathbf{A}^{-1}\mathbf{k}_{\bar{S}*} - (\mathbf{k}_{S*} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{k}_{\bar{S}*})^\top \mathbf{M}(\mathbf{k}_{S*} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{k}_{\bar{S}*}) && (49) \\ &= v_*(\bar{S}) - (\mathbf{k}_{S*} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{k}_{\bar{S}*})^\top \mathbf{M}(\mathbf{k}_{S*} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{k}_{\bar{S}*}) && (50) \end{aligned}$$

where we noticed that the first two terms on the right-hand side of the first line are exactly $v_*(\bar{S})$. Next, we consider the following,

$$\begin{aligned} [\mathbf{K}_y^{-1}]_{S,\mathcal{D}}\, \mathbf{k}_{\mathcal{D}*} &= \begin{bmatrix} -\mathbf{M}\mathbf{B}^\top\mathbf{A}^{-1} & \mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{k}_{\bar{S}*} \\ \mathbf{k}_{S*} \end{bmatrix} = \mathbf{M}(\mathbf{k}_{S*} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{k}_{\bar{S}*}) && (51) \\ \implies \mathbf{k}_{S*} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{k}_{\bar{S}*} &= \mathbf{M}^{-1}[\mathbf{K}_y^{-1}]_{S,\mathcal{D}}\, \mathbf{k}_{\mathcal{D}*} && (52) \end{aligned}$$

where the expression to the left of the equality in Eq. (52) is identical to that in the quadratic form in Eq. (50). We proceed to substitute Eq. (52) into Eq. (50) and rearrange it in terms of $v_*(\bar{S})$:

$$\begin{aligned} v_*(\bar{S}) &= v_*(\mathcal{D}) + ([\mathbf{K}_y^{-1}]_{S,\mathcal{D}}\, \mathbf{k}_{\mathcal{D}*})^\top \mathbf{M}^{-1} [\mathbf{K}_y^{-1}]_{S,\mathcal{D}}\, \mathbf{k}_{\mathcal{D}*} && (53) \\ &= v_*(\mathcal{D}) + [\mathbf{K}_y^{-1}\mathbf{k}_{\mathcal{D}*}]_S^\top \left([\mathbf{K}_y^{-1}]_{S,S}\right)^{-1} [\mathbf{K}_y^{-1}\mathbf{k}_{\mathcal{D}*}]_S && (54) \end{aligned}$$

where in the second line we used a simple manipulation to write $[\mathbf{K}_y^{-1}]_{S,\mathcal{D}}\, \mathbf{k}_{\mathcal{D}*}$ as $[\mathbf{K}_y^{-1}\mathbf{k}_{\mathcal{D}*}]_S$ and write $\mathbf{M}$ as a subblock of $\mathbf{K}_y^{-1}$ (see Eq. (48)).

B DERIVATION OF THE GREEDY RULES UNDER GP SURROGATE

B.1 Information Gain

At each of the M rounds we pick the example whose inclusion most reduces the marginal variance at $\mathbf{x}_*$ conditioned on the set of acquired points so far $\mathcal{A}$ and the candidate point $\mathbf{x}$. This is given in Eq. (7) which is repeated here for convenience:

$$\begin{aligned}\mathbf{x}^{(m)} &\leftarrow \underset{\mathbf{x}\in\mathcal{D}\setminus\mathcal{A}}{\arg\min}\, v_*^{\mathcal{A}\cup\{\mathbf{x}\}} \\ \text{where}\quad v_*^{\mathcal{A}\cup\{\mathbf{x}\}} &= k_{**} - \mathbf{k}_{\mathcal{A}\cup\{\mathbf{x}\},*}^\top \left(\mathbf{K}_{\mathcal{A}\cup\{\mathbf{x}\}} + \sigma^2\mathbf{I}\right)^{-1} \mathbf{k}_{\mathcal{A}\cup\{\mathbf{x}\},*}.\end{aligned} \tag{55}$$

Similar to App. A.1, we exploit the partitioned matrix inversion lemma to relate the "add-one-in" marginal variance $v_*^{\mathcal{A}\cup\{\mathbf{x}\}}$ to $v_*^{\mathcal{A}}$. We consider a partition of $\mathbf{K}_{\mathcal{A}\cup\{\mathbf{x}\}} + \sigma^2\mathbf{I}$ into a block matrix with the first block corresponding to the set of acquired points $\mathcal{A}$ and the second block corresponding to the candidate point $\mathbf{x}$:

$$\mathbf{K}_{\mathcal{A}\cup\{\mathbf{x}\}} + \sigma^2\mathbf{I} = \begin{bmatrix}\mathbf{A} & \mathbf{b} \\ \mathbf{b}^\top & c\end{bmatrix} \quad \text{with} \quad \begin{aligned}\mathbf{A} &= \mathbf{K}_{\mathcal{A}} + \sigma^2\mathbf{I} \\ \mathbf{b} &= \mathbf{k}_{\mathcal{A},\mathbf{x}} \\ c &= k_{\mathbf{x},\mathbf{x}} + \sigma^2\end{aligned} \tag{56}$$

Using the partitioned matrix inversion lemma, we have:

$$\left(\mathbf{K}_{\mathcal{A}\cup\{\mathbf{x}\}} + \sigma^2\mathbf{I}\right)^{-1} = \begin{bmatrix}\mathbf{A}^{-1} + m\mathbf{A}^{-1}\mathbf{b}\mathbf{b}^\top\mathbf{A}^{-1} & -m\mathbf{A}^{-1}\mathbf{b} \\ -m\mathbf{b}^\top\mathbf{A}^{-1} & m\end{bmatrix} \tag{57}$$

where $m^{-1} = c - \mathbf{b}^\top\mathbf{A}^{-1}\mathbf{b}$ is the Schur complement of $\mathbf{A}$ in $\left(\mathbf{K}_{\mathcal{A}\cup\{\mathbf{x}\}} + \sigma^2\mathbf{I}\right)$. Substituting Eq. (57) into Eq. (55) and expanding $\mathbf{k}_{\mathcal{A}\cup\{\mathbf{x}\},*}^\top = [\mathbf{k}_{\mathcal{A},*}\ \mathbf{k}_{\mathbf{x},*}]^\top$ we have:

$$v_*^{\mathcal{A}\cup\{\mathbf{x}\}} = k_{**} - \mathbf{k}_{\mathcal{A},*}^\top\mathbf{A}^{-1}\mathbf{k}_{\mathcal{A},*} - m\left(k_{\mathbf{x},*} - \mathbf{b}^\top\mathbf{A}^{-1}\mathbf{k}_{\mathcal{A},*}\right)^2. \tag{58}$$

We recognise the first two terms on the right-hand side of Eq. (58) as $v_*^{\mathcal{A}}$ and simplifying we obtain:

$$v_*^{\mathcal{A}\cup\{\mathbf{x}\}} = v_*^{\mathcal{A}} - \frac{\left(v_{\mathbf{x},*}^{\mathcal{A}}\right)^2}{\sigma^2 + v_{\mathbf{x}}^{\mathcal{A}}} \quad \text{with} \quad \begin{aligned}v_*^{\mathcal{A}} &= k_{**} - \mathbf{k}_{\mathcal{A},*}^\top\left(\mathbf{K}_{\mathcal{A}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{\mathcal{A},*} \\ v_{\mathbf{x},*}^{\mathcal{A}} &= k_{\mathbf{x},*} - \mathbf{k}_{\mathcal{A},\mathbf{x}}^\top\left(\mathbf{K}_{\mathcal{A}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{\mathcal{A},*} \\ v_{\mathbf{x}}^{\mathcal{A}} &= k_{\mathbf{x},\mathbf{x}} - \mathbf{k}_{\mathcal{A},\mathbf{x}}^\top\left(\mathbf{K}_{\mathcal{A}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{\mathcal{A},\mathbf{x}}.\end{aligned} \tag{59}$$

Since $v_*^{\mathcal{A}}$ is independent of $\mathbf{x}$, we can ignore it in the greedy rule thereby obtaining Eq. (8) – the minus sign in front of the remaining term changes the problem from a minimization to a maximization.

B.2 Information Loss

Whilst the information loss criterion is not sub-modular, it can nevertheless be optimized greedily (just without guarantees) and therefore we compare against this in our experiments. The greedy rule corresponds to choosing the candidate point whose removal maximally increases the marginal variance at $\mathbf{x}_*$. For notational convenience let $\mathcal{A}^+ = \mathcal{A}\cup\{\mathbf{x}\}$ then,

$$\begin{aligned}\mathbf{x}^{(m)} &\leftarrow \underset{\mathbf{x}\in\mathcal{D}\setminus\mathcal{A}}{\arg\max}\, v_*^{\mathcal{D}\setminus\mathcal{A}^+} \\ \text{where}\quad v_*^{\mathcal{D}\setminus\mathcal{A}^+} &= k_{**} - \mathbf{k}_{\mathcal{D}\setminus\mathcal{A}^+,*}^\top\left(\mathbf{K}_{\mathcal{D}\setminus\mathcal{A}^+} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{\mathcal{D}\setminus\mathcal{A}^+,*}.\end{aligned} \tag{60}$$

We can use a special case of App. A.1, that is the *leave-one-out* setting, to express $v_*^{\mathcal{D}\setminus\mathcal{A}^+}$ with $\mathcal{D}\setminus\mathcal{A}$ as the conditioning set throughout. This leads to,

$$v_*^{\mathcal{D}\setminus\mathcal{A}^+} = v_*^{\mathcal{D}\setminus\mathcal{A}} + \frac{\left(v_{\mathbf{x},*}^{\mathcal{D}\setminus\mathcal{A}}\right)^2}{\sigma^2 - v_{\mathbf{x}}^{\mathcal{D}\setminus\mathcal{A}}} \quad \text{with} \quad \begin{aligned}v_*^{\mathcal{D}\setminus\mathcal{A}} &= k_{**} - \mathbf{k}_{\mathcal{D}\setminus\mathcal{A},*}^\top\left(\mathbf{K}_{\mathcal{D}\setminus\mathcal{A}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{\mathcal{D}\setminus\mathcal{A},*} \\ v_{\mathbf{x},*}^{\mathcal{D}\setminus\mathcal{A}} &= k_{\mathbf{x},*} - \mathbf{k}_{\mathcal{D}\setminus\mathcal{A},\mathbf{x}}^\top\left(\mathbf{K}_{\mathcal{D}\setminus\mathcal{A}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{\mathcal{D}\setminus\mathcal{A},*} \\ v_{\mathbf{x}}^{\mathcal{D}\setminus\mathcal{A}} &= k_{\mathbf{x},\mathbf{x}} - \mathbf{k}_{\mathcal{D}\setminus\mathcal{A},\mathbf{x}}^\top\left(\mathbf{K}_{\mathcal{D}\setminus\mathcal{A}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_{\mathcal{D}\setminus\mathcal{A},\mathbf{x}}.\end{aligned} \tag{61}$$

Thus, the greedy rule can be stated as,

$$\mathbf{x}^{(m)} \leftarrow \underset{\mathbf{x}\in\mathcal{D}\setminus\mathcal{A}}{\arg\max} \frac{\left(v_{\mathbf{x},*}^{\mathcal{D}\setminus\mathcal{A}}\right)^2}{\sigma^2 - v_{\mathbf{x}}^{\mathcal{D}\setminus\mathcal{A}}} \tag{62}$$

C DERIVATION OF LINEAR-RESPONSE VARIANCE CORRECTION

We consider the "weight-space" form (via Woodbury formula) of the marginal variance given in Eq. (7):

$$v_*^{\mathcal{A}\cup\{\mathbf{x}\}} = \boldsymbol{\phi}_*^\top \left(\mathbf{S}_\mathcal{A} + \tfrac{1}{\sigma^2}\boldsymbol{\phi}_\mathbf{x}\boldsymbol{\phi}_\mathbf{x}^\top\right)^{-1}\boldsymbol{\phi}_* \tag{63}$$

where we use $\boldsymbol{\phi}_\mathbf{x}$ to denote the tangent features (with slight abuse of notation) and $\mathbf{S}_\mathcal{A}$ is the posterior precision matrix. We can construct a perturbative variant of this expression by introducing a "weight" ϵ on the term corresponding to the candidate point. For notational convenience let us denote $\bar{v}_* = v_*^{\mathcal{A}\cup\{\mathbf{x}\}}$ then,

$$\bar{v}_*^\epsilon = \boldsymbol{\phi}_*^\top \left(\mathbf{S}_\mathcal{A} + \epsilon\tfrac{1}{\sigma^2}\boldsymbol{\phi}_\mathbf{x}\boldsymbol{\phi}_\mathbf{x}^\top\right)^{-1}\boldsymbol{\phi}_* \tag{64}$$

$$= \boldsymbol{\phi}_*^\top \left(\mathbf{S}_\mathcal{A}^{-1} - \frac{\epsilon\mathbf{S}_\mathcal{A}^{-1}\boldsymbol{\phi}_\mathbf{x}\boldsymbol{\phi}_\mathbf{x}^\top\mathbf{S}_\mathcal{A}^{-1}}{\sigma^2 + \epsilon\boldsymbol{\phi}_\mathbf{x}^\top\mathbf{S}_\mathcal{A}^{-1}\boldsymbol{\phi}_\mathbf{x}}\right)\boldsymbol{\phi}_* \tag{65}$$

$$= v_*^\mathcal{A} - \frac{\epsilon\left(v_{\mathbf{x},*}^\mathcal{A}\right)^2}{\sigma^2 + \epsilon v_\mathbf{x}^\mathcal{A}} \tag{66}$$

where in the last line we defined $v_*^\mathcal{A} = \boldsymbol{\phi}_*^\top\mathbf{S}_\mathcal{A}^{-1}\boldsymbol{\phi}_*$, $v_\mathbf{x}^\mathcal{A} = \boldsymbol{\phi}_\mathbf{x}^\top\mathbf{S}_\mathcal{A}^{-1}\boldsymbol{\phi}_\mathbf{x}$ and $v_{\mathbf{x},*}^\mathcal{A} = \boldsymbol{\phi}_\mathbf{x}^\top\mathbf{S}_\mathcal{A}^{-1}\boldsymbol{\phi}_*$. In the second line above, we used the Sherman-Morrison formula giving a form typical of recursive least-squares or Kalman filtering. The resulting expression can be interpreted as a weight-dependent functional that we can differentiate with respect to ϵ and evaluate at $\epsilon = 0$ giving the following sensitivity-based variance estimate,

$$\frac{d\bar{v}_*^\epsilon}{d\epsilon} = -\frac{\sigma^2\left(v_{\mathbf{x},*}^\mathcal{A}\right)^2}{\left(\sigma^2 + \epsilon v_\mathbf{x}^\mathcal{A}\right)^2}, \qquad \left.\frac{d\bar{v}_*^\epsilon}{d\epsilon}\right|_{\epsilon=0} = -\frac{1}{\sigma^2}\left(v_{\mathbf{x},*}^\mathcal{A}\right)^2 \tag{67}$$

Using this we can construct a first-order perturbative update to the predictive variance $\bar{v}_*^\epsilon \approx v_*^\mathcal{A} + d\bar{v}_*^\epsilon/d\epsilon|_{\epsilon=0}\cdot\epsilon$. Evaluating this at $\epsilon = 1$ we have,

$$\bar{v}_* = \bar{v}_*^{\epsilon=1} \approx \underbrace{v_*^\mathcal{A} - \frac{1}{\sigma^2}\left(v_{\mathbf{x},*}^\mathcal{A}\right)^2}_{\bar{v}_*^{\text{approx}}} \tag{68}$$

This leads to the following (approximate) greedy rule,

$$\mathbf{x}^{(m)} \leftarrow \underset{\mathbf{x}\in\mathcal{D}\setminus\mathcal{A}}{\arg\max} \left(\boldsymbol{\phi}_\mathbf{x}^\top\mathbf{S}_\mathcal{A}^{-1}\boldsymbol{\phi}_*\right)^2 \tag{69}$$

Alternative view of Eq. (68) when in the regime of large observation noise Starting from the marginal variance in Eq. (59) we can show,

$$v_*^{\mathcal{A}\cup\{\mathbf{x}\}} = v_*^\mathcal{A} - \frac{\left(v_{\mathbf{x},*}^\mathcal{A}\right)^2}{\sigma^2 + v_\mathbf{x}^\mathcal{A}} = v_*^\mathcal{A} - \sigma^{-2}\left(v_{\mathbf{x},*}^\mathcal{A}\right)^2\left(\frac{1}{1+\sigma^{-2}v_\mathbf{x}^\mathcal{A}}\right) \tag{70}$$

$$= v_*^\mathcal{A} - \sigma^{-2}\left(v_{\mathbf{x},*}^\mathcal{A}\right)^2\left(1 - \sigma^{-2}v_\mathbf{x}^\mathcal{A} + \mathcal{O}(\sigma^{-4})\right) \tag{71}$$

$$= v_*^\mathcal{A} - \sigma^{-2}\left(v_{\mathbf{x},*}^\mathcal{A}\right)^2 + \mathcal{O}(\sigma^{-4}) \tag{72}$$

where in the second line we used the geometric series $(1+a)^{-1} = \sum_{k=0}^\infty(-1)^k a^k$. Convergence is assured when $v_\mathbf{x}^\mathcal{A} < \sigma^2$ where we used the fact that variance is non-negative. Terms that are of order $\mathcal{O}(\sigma^{-4})$ or smaller are neglected similar to App. A. The resulting expression in Eq. (72) has leading terms identical to the linear-response variance correction of Eq. (68).

Approximation error of the linear-response variance correction Let us denote $\Delta\bar{v}_*^{\text{exact}} = v_*^{\mathcal{A}\cup\{\mathbf{x}\}} - v_*^{\mathcal{A}}$ and $\Delta\bar{v}_*^{\text{approx}} = \bar{v}_*^{\text{approx}} - v_*^{\mathcal{A}}$ then we can show the relative error is given by,

$$\frac{|\Delta\bar{v}_*^{\text{approx}} - \Delta\bar{v}_*^{\text{exact}}|}{\Delta\bar{v}_*^{\text{exact}}} = \frac{1}{\sigma^2} v_{\mathbf{x}}^{\mathcal{A}} \tag{73}$$

where we dropped the absolute value since the difference is non-negative by construction. This indicates that the quality of the approximation is determined by the marginal variance of the candidate point. When $v_{\mathbf{x}}^{\mathcal{A}}$ is large then $\bar{v}_*^{\text{approx}}$ overestimates the actual value.

We can write the approximation error exactly as,

$$v_*^{\mathcal{A}\cup\{\mathbf{x}\}} - \bar{v}_*^{\text{approx}} = \frac{1}{\sigma^2}\left(v_{\mathbf{x},*}^{\mathcal{A}}\right)^2 \cdot \frac{v_{\mathbf{x}}^{\mathcal{A}}}{\sigma^2 + v_{\mathbf{x}}^{\mathcal{A}}} \tag{74}$$

For the sake of exposition, we can also give this error in Lagrange's Form as follows,

$$R_2^{\epsilon=1} = \frac{1}{2}\frac{\partial^2 \bar{v}_*^{\epsilon}}{\partial \epsilon^2}\bigg|_{\epsilon=\varepsilon} \qquad \text{for some } 0 < \varepsilon < 1 \tag{75}$$

This second derivative is analytically given by,

$$\frac{\partial^2 \bar{v}_*^{\epsilon}}{\partial \epsilon^2} = \frac{2\sigma^2\left(v_{\mathbf{x},*}^{\mathcal{A}}\right)^2 v_{\mathbf{x}}^{\mathcal{A}}}{(\sigma^2 + \epsilon v_{\mathbf{x}}^{\mathcal{A}})^3} \tag{76}$$

Then we can bound the remainder in Eq. (75) as follows,

$$|R_2^{\epsilon=1}| \leq \frac{1}{2}\max_{\varepsilon\in[0,1]}\left|\frac{\partial^2 \bar{v}_*^{\epsilon}}{\partial \epsilon^2}\bigg|_{\epsilon=\varepsilon}\right| \tag{77}$$

$$= \sigma^2\left(v_{\mathbf{x},*}^{\mathcal{A}}\right)^2 v_{\mathbf{x}}^{\mathcal{A}} \underbrace{\max_{\varepsilon\in[0,1]}\left[\frac{1}{(\sigma^2 + \varepsilon v_{\mathbf{x}}^{\mathcal{A}})^3}\right]}_{=1/(\sigma^2)^3} \tag{78}$$

$$\leq \frac{1}{\sigma^4}\left(v_{\mathbf{x},*}^{\mathcal{A}}\right)^2 v_{\mathbf{x}}^{\mathcal{A}} \tag{79}$$

where we drop the absolute value since all the terms in Eq. (76) are non-negative. In the second line, we factor out terms independent of ε from the max and observe that the maximum value is attained when $\varepsilon = 0$.

D CONNECTIONS BETWEEN OUR INFORMATION LOSS CRITERION AND INFLUENCE-BASED TDA

Let us consider the singleton case, *i.e.* $S = \{(\mathbf{x}, y)\}$, of Eq. (4) with the GP surrogate:

$$\mathcal{I}_{\text{IL}}(\{(\mathbf{x}, y)\}; \mathbf{x}_*) = \frac{1}{2}\log\left(\frac{v_*^{\mathcal{D}\setminus\{\mathbf{x}\}}}{v_*^{\mathcal{D}}}\right) \tag{80}$$

▷ substitute $v_*^{\mathcal{D}\setminus\{\mathbf{x}\}} = v_*^{\mathcal{D}} + (v_{\mathbf{x},*}^{\mathcal{D}})^2/(\sigma^2 - v_{\mathbf{x}}^{\mathcal{D}})$

$$= \frac{1}{2}\log\left(1 + \frac{\left(v_{\mathbf{x},*}^{\mathcal{D}}\right)^2}{v_*^{\mathcal{D}}(\sigma^2 - v_{\mathbf{x}}^{\mathcal{D}})}\right) \tag{81}$$

where in the second line we substituted the first-round score derived in Eq. (61). Similar to App. C, we write the variance terms in weight-space form and use $\boldsymbol{\phi}_{\mathbf{x}}$ to denote tangent features, $v_*^{\mathcal{D}} = \boldsymbol{\phi}_*^\top \mathbf{S}_{\mathcal{D}}^{-1}\boldsymbol{\phi}_*$, $v_{\mathbf{x},*}^{\mathcal{D}} = \boldsymbol{\phi}_{\mathbf{x}}^\top \mathbf{S}_{\mathcal{D}}^{-1}\boldsymbol{\phi}_*$ and $v_{\mathbf{x}}^{\mathcal{D}} = \boldsymbol{\phi}_{\mathbf{x}}^\top \mathbf{S}_{\mathcal{D}}^{-1}\boldsymbol{\phi}_{\mathbf{x}}$, where $\mathbf{S}_{\mathcal{D}} = \frac{1}{\sigma^2}\sum_{\mathbf{x}'\in\mathcal{D}}\boldsymbol{\phi}_{\mathbf{x}'}\boldsymbol{\phi}_{\mathbf{x}'}^\top + \mathbf{I}$. Now we will show that Eq. (81) has resemblance to a cross-influence term normalized by self-influence terms.

Let us consider (1^{st}-order) influence function, the canonical approach to influence-based TDA, which performs attribution by the locally approximating the counterfactual change in test loss:

$$\ell_*(\hat{\boldsymbol{\theta}}^{\setminus \mathbf{z}}) - \ell_*(\hat{\boldsymbol{\theta}}) \approx \mathbf{g}_*^\top \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_{\mathbf{z}} \tag{82}$$

where we define $\mathbf{z} = (\mathbf{x}, y)$, denote the gradients of the per-example train and test loss by $\mathbf{g}_{\mathbf{z}}$ and $\mathbf{g}_*$ respectively, and let $\mathbf{H}_{\hat{\boldsymbol{\theta}}}$ be the Hessian evaluated at $\hat{\boldsymbol{\theta}}$ (assuming convergence). Now Barshan et al. (2020) claims that when the dataset contains outliers or mislabelled examples, such examples are often most influential for test examples but these are poor explanatory examples for TDA. They reinterpret influence function from a geometric lens and propose to use cosine similarity instead to reduce the impact of "globally" influential examples. This takes the form of a cross-influence term normalized by self-influence terms:

$$\cos(\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-\frac{1}{2}} \mathbf{g}_*, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-\frac{1}{2}} \mathbf{g}_{\mathbf{z}}) = \frac{\mathbf{g}_*^\top \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_{\mathbf{z}}}{\sqrt{\mathbf{g}_*^\top \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_*} \cdot \sqrt{\mathbf{g}_{\mathbf{z}}^\top \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_{\mathbf{z}}}} \tag{83}$$

▷ expand gradients by chain rule: $\mathbf{g}_* = e_* \boldsymbol{\phi}_*$ and $\mathbf{g}_{\mathbf{z}} = e_{\mathbf{z}} \boldsymbol{\phi}_{\mathbf{x}}$

$$= \frac{e_* \left(\boldsymbol{\phi}_*^\top \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \boldsymbol{\phi}_{\mathbf{z}} \right) e_{\mathbf{z}}}{\sqrt{e_* \left(\boldsymbol{\phi}_*^\top \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \boldsymbol{\phi}_* \right) e_*} \cdot \sqrt{e_{\mathbf{z}} \left(\boldsymbol{\phi}_{\mathbf{z}}^\top \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \boldsymbol{\phi}_{\mathbf{z}} \right) e_{\mathbf{z}}}} \tag{84}$$

▷ approximate Hessian by the Gauss-Newton matrix $\mathbf{H}_{\hat{\theta}} \approx \mathbf{S}_{\mathcal{D}}$ and substitute variance terms

$$\approx \frac{\cancel{e_*} v_{\mathbf{x},*}^{\mathcal{D}} \cancel{e_{\mathbf{z}}}}{\cancel{e_*} \sqrt{v_*^{\mathcal{D}}} \cdot \cancel{e_{\mathbf{z}}} \sqrt{v_{\mathbf{x}}^{\mathcal{D}}}} \tag{85}$$

▷ observe the residuals cancel

$$= \frac{v_{\mathbf{x},*}^{\mathcal{D}}}{\sqrt{v_*^{\mathcal{D}}} \cdot \sqrt{v_{\mathbf{x}}^{\mathcal{D}}}} \tag{86}$$

where $e_{\mathbf{z}} = f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) - y$ and $e_* = f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_*) - y_*$ are the train and test residuals respectively (assuming single-output case). The resulting expression is independent of the training label (and test label), similar to Eq. (81). An important difference however is that this score is *signed* allowing one to distinguish positively-influential examples from negatively-influential examples, sometimes called "proponents"/"opponents" (Pruthi et al., 2020). Another difference is that whilst cosine similarity downweights train examples with high (marginal) variance, our criterion does the opposite.

E ADDITIONAL EXPERIMENTAL DETAILS

Datasets and models. We used a ResNet-9 architecture for CIFAR-10, a two-hidden-layer MLP with ReLU activations for Fashion-MNIST, and BERT for RTE. For the vision models in brittleness/backdoor, we apply NTK reparameterization *post hoc* for attribution: we do not modify the ResNet-9 default initialization and simply apply the transformation assuming standard Normal initialization; for the MLP this follows Lee et al. (2020) where Kaiming-Normal initialization is used without any bias terms. For the CIFAR-10 coreset experiment, we instead keep the original parameterization. For BERT on RTE, we do not apply NTK parameterization and use linear layers only for Jacobian-based attribution. For training, the ResNet-9 was optimized using SGD with momentum (0.9), weight decay (5e-4), a batch size of 512, over 24 epochs. We employed a triangular learning rate schedule that linearly increased to a peak learning rate of 0.5 by epoch 5 and then decayed linearly to zero by the final epoch. For Fashion-MNIST, we used SGD with momentum (0.9), batch size 64, learning rate 0.03, weight decay 1e-3, and trained for 20 epochs. For RTE, we fine-tuned BERT for 3 epochs with batch size 32, learning rate 2×10^{-5}, and weight decay 0.01. The experiments were run on an internal cluster of GPUs of the following type: Tesla V100 SXM2 16 GB.

Baselines. We provide additional implementation details for the TDA baseline methods adapted from Bae et al. (2024). All baselines produce a ranking of training examples via additive pointwise scores, and attributed subsets are obtained by the rank-and-pick rule described in Sec. 2. In addition, all baselines are evaluated on the final checkpoint only, without any model retraining. Unless otherwise stated, all reported metrics are accompanied by standard error across five random seeds.

- RANDOM: Selects examples uniformly at random.
- REPSIM: Ranks examples by cosine similarity between penultimate-layer features of the query and training examples.
- GRADDOT: Ranks examples by the dot-product between query and train gradients (*i.e.* TracIn (Pruthi et al., 2020) at the final checkpoint only). We use standard cross-entropy loss for train gradients and the same measurement function as TRAK for query gradients (detailed below).
- TRAK: An extension of generalized linear model (GLM) influence (Pregibon, 1981) to neural networks using random projections and the Generalized Gauss-Newton approximation. For multi-class classification (as in our setting), TRAK uses a measurement function that reduces multi-class outputs to a single logistic regression (see App. E.5 in Park et al. (2023)). We use a Rademacher sketch with projection dimension 4096 for CIFAR-10/Fashion-MNIST and 20480 for RTE (matching our methods per dataset), but do not use ensembling as originally proposed. This is implemented via the `trak` package (Park et al., 2023) with a fast CUDA extension that avoids materializing the projection matrix.
- KRONINFLUENCE: An influence function implementation using the EK-FAC approximation to the Hessian. Computation is restricted to fully-connected and convolutional layers, the layers supported by the `kroninfluence` package (Grosse et al., 2023). We use the same measurement function as TRAK for query gradients and adopt the damping factor heuristic proposed in Grosse et al. (2023).

Hyperparameter tuning (observation noise). For each method we tune σ^2 over the log-spaced grid $\{10^{-6}, \ldots, 10^{6}\}$ using three random repeats per candidate value. For the brittleness experiments, we use a reduced set of 25 queries and obtain a held-out validation set by partitioning the original test set in half; we use margin change as the validation score and fix the subset size to 500 for CIFAR-10, 150 for Fashion-MNIST and 50 for RTE. For the coreset experiment, we use the same held-out validation split and fix the coreset size to 500. Tuned values are reused for the full runs reported in the main text.

Brittleness protocol details. *Queries.* We evaluate 100 correctly classified queries for CIFAR-10/Fashion-MNIST and 50 for RTE. *Candidate pool.* For each query we restrict training candidates to the same class as the query (Singla et al., 2023). *Subset sizes.* We vary M over a grid (CIFAR-10: $M \in \{200, 400, 600, 800, 1000, 1200\}$; Fashion-MNIST: $M \in \{50, 100, 150, 200, 250, 300\}$; RTE: $M \in \{20, 40, 60, 80, 100, 120\}$. *Retraining.* After removing the top-M points, we retrain from the same initialization with identical data ordering and per-epoch shuffling (to isolate the effect of removal).

Backdoor protocol details. Let $\mathcal{Q}$ denote the set of 100 backdoored test queries. For each query $q \in \mathcal{Q}$, let $\mathcal{G}(q)$ be its ground-truth attribution set, *i.e.* the 50 poisoned training examples with the same base class as q. For each query $q \in \mathcal{Q}$, let $\pi_q = (i_1, i_2, \ldots)$ denote the ranked list of training examples produced by a given method, and let $\mathrm{rank}_q(i)$ denote the position of training example i in π_q. For TDA baselines this ranking is obtained by sorting training examples by descending attribution score, whereas for our greedy information-theoretic methods it is given by the order of greedy selection. Let $\mathcal{R}_K(q) = \{i : \mathrm{rank}_q(i) \le K\}$ denote the top-K retrieved set. We report,

$$\text{Recall@50} \;=\; \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{|\mathcal{R}_{50}(q) \cap \mathcal{G}(q)|}{|\mathcal{G}(q)|}, \tag{87}$$

which measures the fraction of ground-truth poisoned examples recovered in the top 50, averaged over queries. Let $r_q^\star = \min_{i \in \mathcal{G}(q)} \mathrm{rank}_q(i)$ denote the rank of the highest-ranked ground-truth poisoned example for query q. We also report the cutoff-based mean reciprocal rank,

$$\text{MRR@100} \;=\; \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \begin{cases} \dfrac{1}{r_q^\star}, & \text{if } r_q^\star \le 100, \\ 0, & \text{otherwise}, \end{cases} \tag{88}$$

which is the reciprocal rank of the first ground-truth poisoned example for each query, truncated to zero if no such example appears in the top 100.

Coreset protocol details. We extract 500 examples from the test set as the query set for attribution and report test accuracy on the remaining test examples after retraining on the selected coreset. We vary coreset sizes over $M \in \{100, 200, 500, 1000, 2000, 5000\}$. For each seed, the coreset-trained model is retrained from the same initialization as the full-data model, using the same training procedure (optimizer, learning rate schedule, number of epochs, and data augmentation). For this experiment we use the same scalar multiclass measurement function as TRAK for all Jacobian-based methods and do not apply NTK reparameterization. We additionally report a class-balanced variant of this experiment in Fig. 6, where per-class quotas are enforced during selection; as discussed in Sec. 4.3, this substantially improves all TDA baselines, yet both InfoGain variants remain strongest overall.

For the TDA baselines, attribution is defined pointwise, so we aggregate over the query set by averaging the per-query scores for each training example and then taking the top-M examples. Our greedy methods jointly optimise over all N_{test} query points at each step, so the single-query greedy rules do not directly apply. Below we derive a tractable multi-query selection criterion for InfoGain; the corresponding InfoLoss derivation is analogous and is therefore omitted. Let $\mathbf{\Phi}_* \in \mathbb{R}^{N_{\text{test}} \times K}$ stack the query features $\{\boldsymbol{\phi}_{*q}\}_{q=1}^{N_{\text{test}}}$ row-wise. The exact AOI greedy rule for InfoGain is then,

$$\mathbf{x}^{(m)} \leftarrow \underset{\mathbf{x} \in \mathcal{D} \setminus \mathcal{A}}{\arg\min} \log\det\left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}\cup\{\mathbf{x}\}}^{-1} \mathbf{\Phi}_*^\top\right). \tag{89}$$

The objective depends on $\mathbf{x}$ through the AOI precision $\mathbf{S}_{\mathcal{A}\cup\{\mathbf{x}\}}$. We isolate the $\mathbf{x}$-dependent terms via Sherman-Morrison and the matrix determinant lemma, then approximate the expensive query-query interaction:

$$\log\det\left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}\cup\{\mathbf{x}\}}^{-1} \mathbf{\Phi}_*^\top\right) \tag{90}$$

▷ expand $\mathbf{S}_{\mathcal{A}\cup\{\mathbf{x}\}}^{-1}$ via Sherman-Morrison

$$= \log\det\left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \mathbf{\Phi}_*^\top - \frac{\left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \boldsymbol{\phi}_{\mathbf{x}}\right)\left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \boldsymbol{\phi}_{\mathbf{x}}\right)^\top}{\sigma^2 + v_{\mathbf{x}}^{\mathcal{A}}}\right) \tag{91}$$

▷ matrix determinant lemma

$$= \underbrace{\log\det\left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \mathbf{\Phi}_*^\top\right)}_{\text{const. in } \mathbf{x}} + \log\left(1 - \frac{\left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \boldsymbol{\phi}_{\mathbf{x}}\right)^\top \left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \mathbf{\Phi}_*^\top\right)^{-1} \left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \boldsymbol{\phi}_{\mathbf{x}}\right)}{\sigma^2 + v_{\mathbf{x}}^{\mathcal{A}}}\right) \tag{92}$$

▷ first term is constant in $\mathbf{x}$; approximate $\left(\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \mathbf{\Phi}_*^\top\right)^{-1} \approx \mathbf{I}$ (drop query-query interactions)

$$\approx \text{const.} + \log\left(1 - \frac{\|\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \boldsymbol{\phi}_{\mathbf{x}}\|_2^2}{\sigma^2 + v_{\mathbf{x}}^{\mathcal{A}}}\right). \tag{93}$$

This yields our tractable multi-query InfoGain criterion,

$$\mathbf{x}^{(m)} \leftarrow \underset{\mathbf{x} \in \mathcal{D} \setminus \mathcal{A}}{\arg\max} \frac{\|\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \boldsymbol{\phi}_{\mathbf{x}}\|_2^2}{\sigma^2 + v_{\mathbf{x}}^{\mathcal{A}}}. \tag{94}$$

Applying the same linear-response variance correction as in App. C removes the candidate-dependent denominator and yields the approximate variant,

$$\mathbf{x}^{(m)} \leftarrow \underset{\mathbf{x} \in \mathcal{D} \setminus \mathcal{A}}{\arg\max} \|\mathbf{\Phi}_* \mathbf{S}_{\mathcal{A}}^{-1} \boldsymbol{\phi}_{\mathbf{x}}\|_2^2. \tag{95}$$

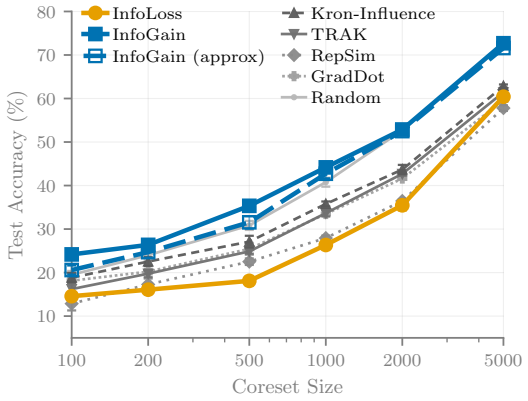


Figure 6: CIFAR-10 coreset selection with class-balanced TDA baselines. Enforcing equal per-class quotas substantially improves TRAK, KronInfluence, GradDot and RepSim, often lifting the strongest baselines above InfoLoss; however, both InfoGain variants remain strongest overall.

F VISUALIZATION OF TRAINING DATA ATTRIBUTION METHODS

Figure 7: Visualization of training data attribution using our Bayesian information-theoretic methods on the CIFAR-10 dataset. Top: InfoLoss, Middle: InfoGain, Bottom: InfoGain (approx).

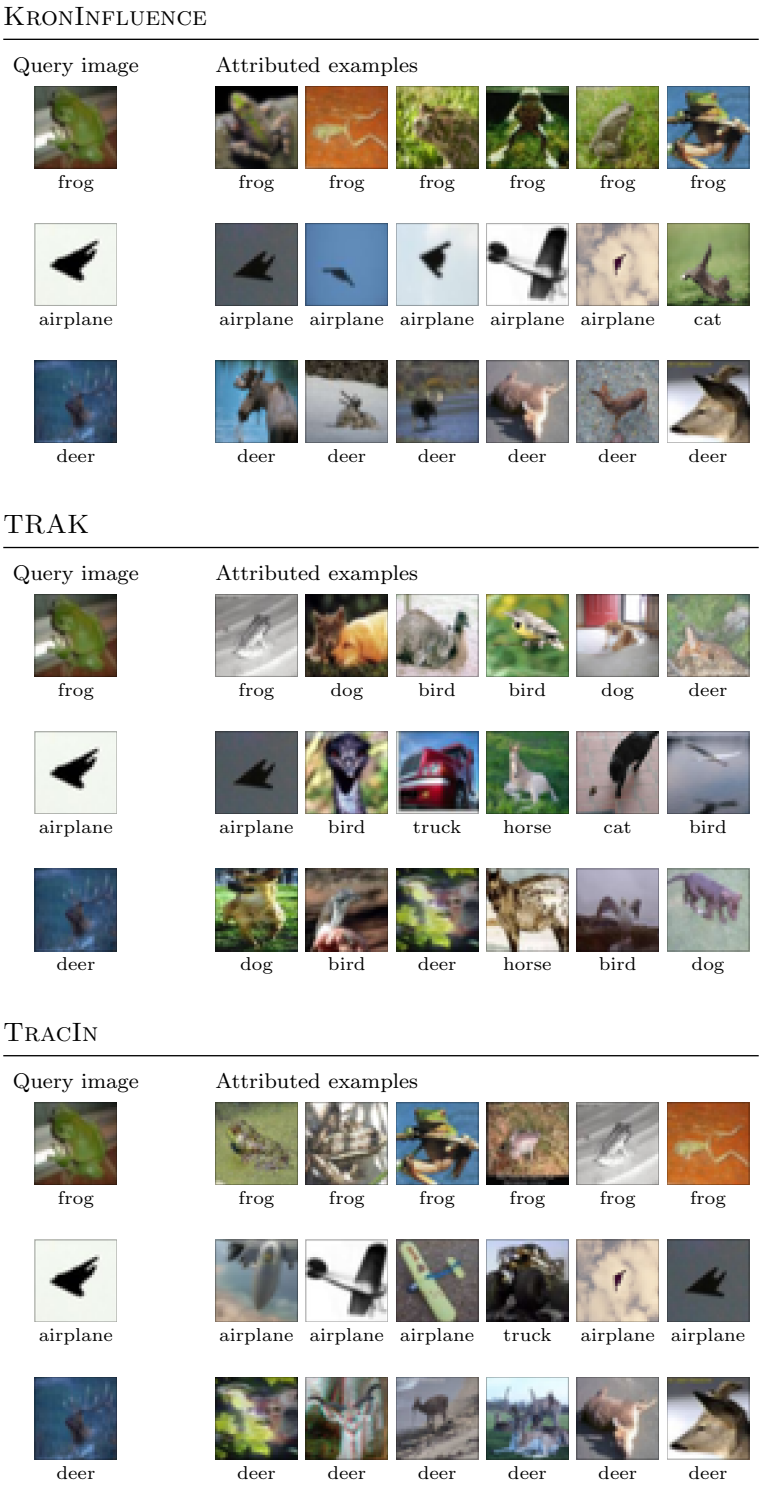


Figure 8: Visualization of training data attribution using baseline influence-based methods on the CIFAR-10 dataset. Top: KronInfluence, Middle: TRAK, Bottom: TracIn.

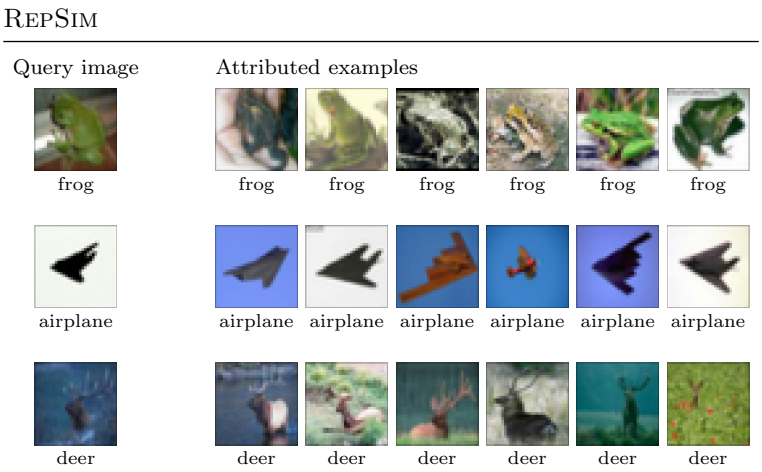


Figure 9: Visualization of training data attribution using representational similarity (RepSim) on the CIFAR-10 dataset.