

A Details of the Hyperparameter Study

A.1 Hyperparameters Definitions

Apart from major hyperparameters described in Section 2.1, some additional hyperparameters are considered in our hyperparameter sensitivity study. We provide their definitions as following:

- **normalization** (TracIn) [28]: We may normalize the per-example gradients in TracIn by dividing their L2-norm. In other words, **normalization**=True means that $\tau_{\text{TracIn}}(z', z_i) := \sum_t \eta_t \frac{\nabla_{\theta} f(z', \theta_t)^{\top}}{\|\nabla_{\theta} f(z', \theta_t)^{\top}\|} \frac{\nabla_{\theta} L(z_i, \theta_t)}{\|\nabla_{\theta} L(z_i, \theta_t)\|}$.
- **checkpoint-selection** (TracIn) [28]: The checkpoints used to calculate τ_{TracIn} can be designed in various way. In this paper, we select the last 10 epochs for different maximum training epoch numbers.
- **max-iteration** (IF (CG)) [20]: This hyperparameter specifies the maximum number of steps the conjugate gradient algorithm will take to attempt convergence. In the experiment, we do not set stop criteria so that the algorithm will stop at the maximum number of iteration.
- **scaling & recursion-depth & batch-size** (IF (LiSSA)) [20]: The LiSSA algorithm for inverse Hessian vector product (IHVP) is an iterative algorithm where each iteration t applies the formula $v^t = g + (I - \frac{1}{\eta}(H^t + \lambda I))v^{t-1}$, and g is the target vector. The hyperparameters **scaling** refers to η ; **recursion-depth** indicates the maximum t where the iteration stops; **batch-size** indicates how many data points are used to calculate the batch-wise hessian matrix H^t .

Default values. In the following table, we list the default value of each hyperparameter. The default value is used when other hyperparameters are searched. They are selected according to the default value in original paper.

TDA method	Hyperparameter Name	Default values
TRAK-10	regularization	0
	pojection-dimension	512 (2048 for WikiText2+GPT2)
	training-epoch	50 (3 for WikiText2+GPT2)
LoGra	regularization	1e-3
	pojection-dimension	64 ²
	training-epoch	3
IF (explicit)	regularization	1e-5
	training-epoch	50
IF (CG)	regularization	1e-2
	max-iteration	10
	training-epoch	50
IF (LiSSA)	regularization	1e-3
	scaling	5
	recursion-depth	1000
	batch-size	50
	training-epoch	50
TracIn	normalization	False
	pojection-dimension	None (i.e., no projection)
	checkpoint-selection	last 10 checkpoints

Table 1: Default values of hyperparameters.

752 A.2 Additional Results

753 In Figure 4, we present additional results not included in Figure 1 due to space limit in the main text.
 754 These additional results confirm that most TDA methods are sensitive to certain hyperparameters.
 755 Intriguingly, **training-epoch** is one of the most sensitive hyperparameter for most TDA methods.

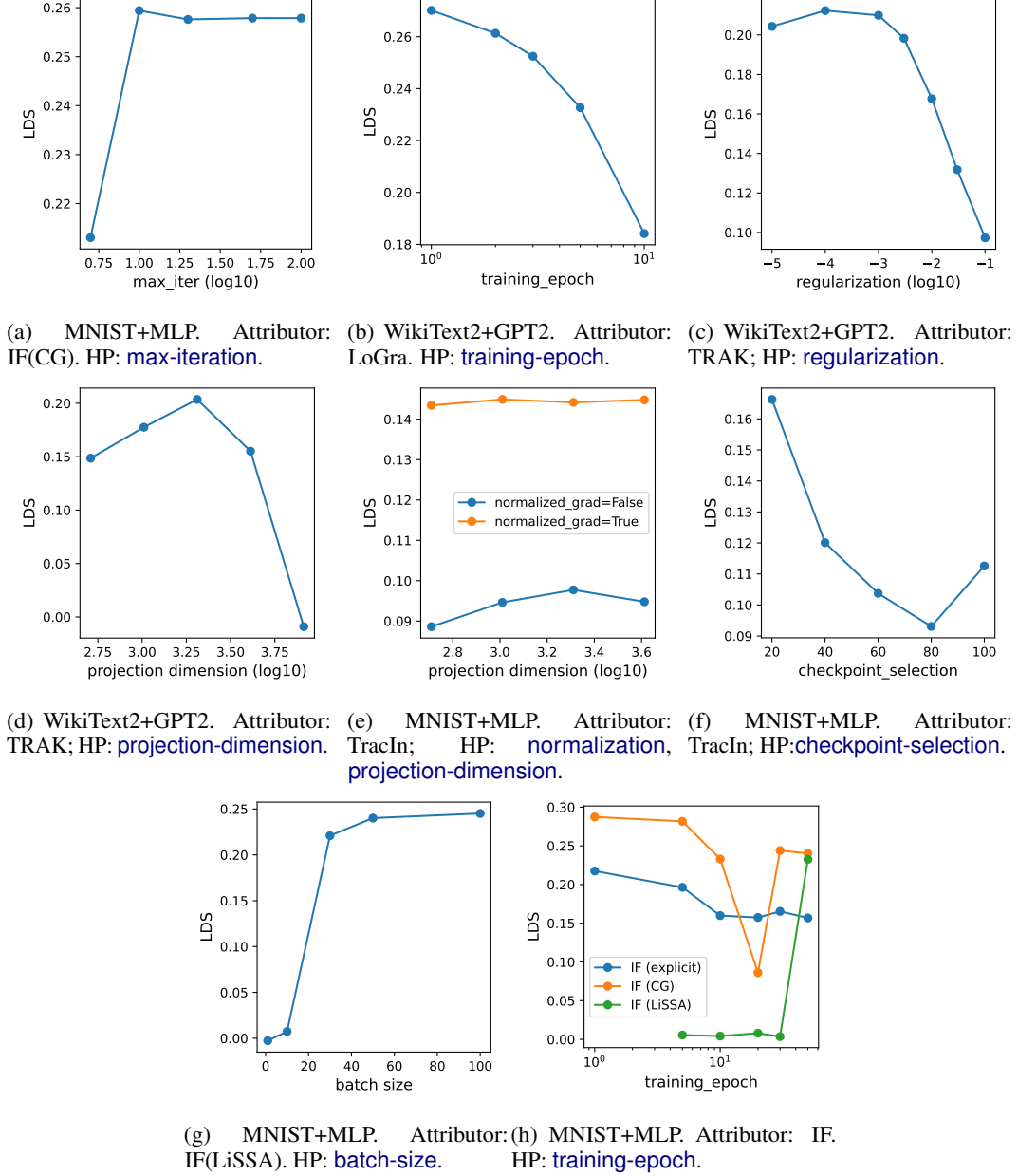


Figure 4: Study of hyperparameter sensitivity additional results.

756 A.3 Computational Resources and Dataset Licenses

757 The experiments for the hyperparameter sensitivity analysis are done on 4 A100 GPUs in around 100
 758 hours, excluding model retraining (we reused some model checkpoints provided by the dattri library
 759 to avoid extensive model retraining). For the dataset we use: MNIST-10 dataset holds CC BY-SA 3.0
 760 license; CIFAR-10 dataset holds CC-BY 4.0 license; WikiText2 dataset holds CC BY-SA 3.0 license.

B Omitted Details of the Theoretical Analysis in Section 4

B.1 Relationship between TRAK and IFFIM

As noted in Section 2.1, TRAK can be viewed as a variant of IFFIM with additional computational tricks. This section demonstrates their similarity and distinctions in more details.

We first formally introduce the TRAK attributor without gradient projection as follows⁸.

$$\tau_{\text{TRAK}}(z', z_i) := \nabla_{\theta} f(z', \theta_S^*)^{\top} (\Phi^{\top} R \Phi)^{-1} \nabla_{\theta} f(z_i, \theta_S^*) \cdot (1 - p_i),$$

where $\Phi \in \mathbb{R}^{n \times p}$ has its i th row being $\nabla_{\theta} f(z_i, \theta_S^*)^{\top}$, $R := \text{diag}\{p_i(1 - p_i)\}_{i=1}^n$, and $p_i := p(z_i, \theta_S^*)$.

Compared to the IFFIM attributor in Eq. (2), there are two main differences. First, the right most gradient term in TRAK is $\nabla_{\theta} f(z_i, \theta_S^*)(1 - p_i)$ while its counterpart in IFFIM is $-\nabla_{\theta} L(z_i, \theta_S^*)$. Second, IFFIM has the empirical FIM F_S in the middle while TRAK has $\Phi^{\top} R \Phi$.

The right gradient term difference. We first show that $\nabla_{\theta} f(z_i, \theta_S^*)(1 - p_i)$ is equivalent to $-\nabla_{\theta} L(z_i, \theta_S^*)$ as

$$-\nabla_{\theta} L(z_i, \theta_S^*) = - \left. \frac{\partial \ln(1 + e^{-f})}{\partial f} \right|_{f=f(z_i, \theta_S^*)} \nabla_{\theta} f(z_i, \theta_S^*) = (1 - p(z_i, \theta_S^*)) \nabla_{\theta} f(z_i, \theta_S^*).$$

The middle term difference. The middle term in TRAK can be written as

$$\frac{1}{n} \Phi^{\top} R \Phi = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} f(z_i, \theta_S^*)(p_i(1 - p_i)) \nabla_{\theta} f(z_i, \theta_S^*)^{\top}.$$

Noting that $p_i(1 - p_i)$ is the Hessian of cross-entropy loss with respect to model output, and the whole $\frac{1}{n} \Phi^{\top} R \Phi$ is known as the Generalized Gauss Newton (GGN) matrix [24].

Furthermore, Martens [24] has shown that $\frac{1}{n} J^{\top} J$ (the empirical FIM) and $\frac{1}{n} \Phi^{\top} R \Phi$ (the GGN) both reduce to the true FIM when S converges to the true underlying distribution of \mathcal{Z} .

This comparison indicates that there is fundamental similarity between TRAK and IFFIM despite superficial algorithmic differences.

B.2 Proof of Theorem 4.3

We prove Theorem 4.3 before Lemma 4.2 for better readability. We first establish several intermediate results for Eq. (8).

Lemma B.1. *We have*

$$\text{Var}_{A \sim D_a} [\nabla_{\theta} R_A(\theta_S^*)] = \frac{1}{a} F_S,$$

where D_a is the distribution of uniformly sampled size a subsets of S .

Proof. Because A is uniformly sampled size a subsets of S ,

$$\text{Var}_{A \sim D_a} [\nabla_{\theta} R_A(\theta_S^*)] = \frac{1}{a} \text{Var}_{z \sim S} [\nabla_{\theta} L(z, \theta_S^*)].$$

Further, since $\mathbb{E}_{z \sim S} [\nabla_{\theta} L(z, \theta_S^*)] = \nabla_{\theta} R_S(\theta_S^*) = 0$,

$$\text{Var}_{z \sim S} [\nabla_{\theta} L(z, \theta_S^*)] = \mathbb{E}_{z \sim S} [\nabla_{\theta} L(z, \theta_S^*) \nabla_{\theta} L(z, \theta_S^*)^{\top}] = F_S.$$

This completes the proof. \square

Lemma B.1 facilitates analyzing the variance of $\sum_{z \in A} \tau_{\text{IFFIM}, \lambda}(z', z)$ in Eq. (6), thereby enabling verification of the following equivalent condition of $\dot{c}_p(\lambda; z') > 0$.

⁸In practical implementation, Park et al. [27] dropped the diagonal matrix R due to slightly improved empirical performance, and projected Φ and the gradients to lower dimension for computational efficiency.

787 **Proposition B.2.** For a test example z' , $\dot{c}_p(\lambda; z') > 0$ is equivalent to

$$r_{z',\lambda} \cdot t_{3,z',\lambda} > (-\nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-2} g_{z'}) \cdot t_{2,z',\lambda}. \quad (10)$$

788

789 *Proof.* We first simplify $\sum_{z \in A} \tau_{\text{IFFIM},\lambda}(z', z)$ in Eq.(6):

$$\begin{aligned} \sum_{z \in A} \tau_{\text{IFFIM},\lambda}(z', z) &= - \sum_{z \in A} \nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} \nabla_{\theta} L(z, \theta_S^*) \\ &= -a \cdot \nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} \nabla_{\theta} R_A(\theta_S^*). \end{aligned}$$

790 Three terms are involved when we expand the definition of $c_p(\tau_{\text{IFFIM},\lambda})$. The first is the numerator:

$$\begin{aligned} &\mathbb{E}_{A \sim D_a} \left[\sum_{z \in A} \tau_{\text{IFFIM},\lambda}(z', z) \cdot (f(z', \theta_A^*) - \mathbb{E}_{A' \sim D_a} [f(z', \theta_{A'}^*)]) \right] \\ &= \mathbb{E}_{A \sim D_a} [-a \nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} \nabla_{\theta} R_A(\theta_S^*) \cdot (f(z', \theta_A^*) - \mathbb{E}_{A' \sim D_a} [f(z', \theta_{A'}^*)])], \end{aligned}$$

791 where the component that depends on A is

$$\begin{aligned} &\mathbb{E}_{A \sim D_a} [\nabla_{\theta} R_A(\theta_S^*) \cdot (f(z', \theta_A^*) - \mathbb{E}_{A' \sim D_a} [f(z', \theta_{A'}^*)])] \\ &= \frac{1}{\binom{n}{a}} \sum_A (f(z', \theta_A^*) - \mathbb{E}_{A' \sim D_a} [f(z', \theta_{A'}^*)]) \frac{1}{a} \sum_{z \in A} \nabla_{\theta} L(z, \theta_S^*) \\ &= \frac{1}{a \binom{n}{a}} \sum_{i=1}^n \left(\sum_A \mathbb{I}[z_i \in A] (f(z', \theta_A^*) - \mathbb{E}_{A' \sim D_a} [f(z', \theta_{A'}^*)]) \right) \nabla_{\theta} L(z_i, \theta_S^*) \\ &= \frac{1}{a \binom{n}{a}} \sum_{i=1}^n \binom{n-1}{a-1} \mathbb{E}_{A \sim D_a} [f(z', \theta_A^*) - \mathbb{E}_{A' \sim D_a} [f(z', \theta_{A'}^*)] | z_i \in A] \nabla_{\theta} L(z_i, \theta_S^*) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{A \sim D_a} [f(z', \theta_A^*) - \mathbb{E}_{A' \sim D_a} [f(z', \theta_{A'}^*)] | z_i \in A] \nabla_{\theta} L(z_i, \theta_S^*) \\ &= \frac{1}{n} \sum_{i=1}^n \alpha_{z',i} \nabla_{\theta} L(z_i, \theta_S^*) = g_{z'}. \end{aligned}$$

792 Note that we let \sum_A sums over each subset A of S with size a and we apply $a \binom{n}{a} = n \binom{n-1}{a-1}$.

793 The second term is the variance of $\sum_{z \in A} \tau_{\text{IFFIM},\lambda}(z', z)$:

$$\begin{aligned} &\text{Var}_{A \sim D_a} \left[\sum_{z \in A} \tau_{\text{IFFIM},\lambda}(z', z) \right] \\ &= \text{Var}_{A \sim D_a} [-a \cdot \nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} \nabla_{\theta} R_A(\theta_S^*)] \\ &= a^2 \cdot \nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} \text{Var}_{A \sim D_a} [\nabla_{\theta} R_A(\theta_S^*)] (F_S + \lambda I_p)^{-1} \nabla_{\theta} f(z', \theta_S^*) \\ &= a \cdot \nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} F_S (F_S + \lambda I_p)^{-1} \nabla_{\theta} f(z', \theta_S^*), \end{aligned}$$

794 where we apply Lemma B.1.

The third term is the variance of $f(z', \theta_A^*)$, which, together with a in the numerator and the variance of $\sum_{z \in A} \tau_{\text{IFFIM},\lambda}(z', z)$, are omitted because they do not depend on λ . To summarize, so far we have shown that $\dot{c}_p(\lambda; z') > 0$ is equivalent to

$$\frac{\partial}{\partial \lambda} \frac{-\nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} g_{z'}}{\sqrt{\nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} F_S (F_S + \lambda I_p)^{-1} \nabla_{\theta} f(z', \theta_S^*)}} \Big|_{\lambda} > 0.$$

By direct calculation, the left hand side is

$$\frac{(-\nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} g_{z'}) (\nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-3} F_S \nabla_{\theta} f(z', \theta_S^*)) - (-\nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-2} g_{z'}) (\nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-2} F_S \nabla_{\theta} f(z', \theta_S^*))}{(\nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} F_S (F_S + \lambda I_p)^{-1} \nabla_{\theta} f(z', \theta_S^*))^{3/2}}.$$

When the positive denominator is dropped, we obtain the desired formula.

796

□

Before we move to the proof of next result, we first introduce a useful matrix transformation lemma.

Lemma B.3. For matrix $X \in \mathbb{R}^{n \times p}$, non-negative integer k , and $\lambda > 0$, we have

$$(X^\top X + \lambda I_p)^{-k} X^\top = X^\top (X X^\top + \lambda I_n)^{-k}.$$

799

Proof. Note that

$$X^\top (X X^\top + \lambda I_n) = X^\top X X^\top + \lambda X^\top = (X^\top X + \lambda I_p) X^\top.$$

If we left multiply $(X^\top X + \lambda I_p)^{-1}$ and right multiply $(X X^\top + \lambda I_n)^{-1}$ on both sides, we derive

$$(X^\top X + \lambda I_p)^{-1} X^\top = X^\top (X X^\top + \lambda I_n)^{-1}.$$

By applying the equality k times on $(X^\top X + \lambda I_p)^{-k} X^\top$ to “push X^\top through” the inverted matrix, we complete the proof. □

Proof of Theorem 4.3. Because F_S is a positive semi-definite matrix, so is $(F_S + \lambda I_p)^{-k} F_S$ for any positive integer k . As a results,

$$t_{k,z',\lambda} = \nabla_\theta f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-k} F_S \nabla_\theta f(z', \theta_S^*) \geq 0.$$

In fact, $t_{k,z',\lambda} = 0 \Leftrightarrow F_S \nabla_\theta f(z', \theta_S^*) = 0$ implies $t_{2,z',\lambda} = 0$ which results in an undefined c_p . Therefore, $t_{k,z',\lambda} > 0$. Further, by the premise, we have $r_{z',\lambda} > 0$ as $t_{2,z',\lambda} > 0$. Without loss of generality, we only consider the case where $-\nabla_\theta f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-2} g_{z'} > 0$, because otherwise Eq.(10) holds automatically.

We proceed by rewriting Eq.(10) with Lemma B.3. By setting $X = J/\sqrt{n}$,

$$\begin{aligned} t_{3,z',\lambda} &= \frac{1}{n} \nabla_\theta f(z', \theta_S^*)^\top J^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-3} J \nabla_\theta f(z', \theta_S^*), \\ -\nabla_\theta f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-2} g_{z'} &= -\frac{1}{n} \nabla_\theta f(z', \theta_S^*)^\top J^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-2} \alpha_{z'}, \end{aligned}$$

where we apply the fact that $F_S = \frac{1}{n} J^\top J = X^\top X$ and $g_{z'} = \frac{1}{n} J^\top \alpha_{z'}$ by their definitions. Then, note that

$$\begin{aligned} &\sqrt{t_{3,z',\lambda} \cdot o_{z',\lambda}} \\ &= \sqrt{\left(\frac{1}{n} \nabla_\theta f(z', \theta_S^*)^\top J^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-3} J \nabla_\theta f(z', \theta_S^*) \right) \left(\frac{1}{n} \alpha_{z'}^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-1} \alpha_{z'} \right)} \quad (11) \\ &\geq -\frac{1}{n} \nabla_\theta f(z', \theta_S^*)^\top J^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-2} \alpha_{z'} = -\nabla_\theta f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-2} g_{z'} > 0, \end{aligned}$$

by applying the generalized Cauchy-Schwarz inequality on $-\left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-1} J \nabla_\theta f(z', \theta_S^*)$ and $\alpha_{z'}$ with inner product defined by positive definite matrix $\frac{1}{n} \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-1}$. Now, we finish the proof by observing that Eq.(10) is derived directly by multiplying Eq.(8) with Eq.(11) and rearranging terms. □

B.3 Assumptions and Proof of Lemma 4.2

We base the discussion of Lemma 4.2 on Eq.(8). For simplicity, we first eliminate $\sqrt{t_{1,z',\lambda}}$ on both sides to state the following sufficient condition for $\dot{c}_p(\lambda; z') > 0$:

$$\frac{r_{z',\lambda}}{\sqrt{o_{z',\lambda}}} > \frac{t_{2,z',\lambda}}{\sqrt{t_{3,z',\lambda}}}. \quad (12)$$

Now, we let μ_i denote the i th eigenvalue of F_S . To better establish a relationship between the spectrum of F_S and $c_p(\tau_{\text{IFFIM},\lambda})$, we first make the following definition for convenience.

817 **Definition B.4.** For a vector $v \in \mathbb{R}^p$ and $i \in [p]$, let \tilde{v}_i denote the i th component of v under a
818 chosen eigenbasis of F_S , where eigenvalues are sorted in descending order. Let μ_i denote the i th
819 eigenvalue. Further, let \tilde{v}_{\min} denote the component of v under the eigenbasis corresponding to the
820 minimal non-zero eigenvalue μ_{\min} of F_S .

821 We then provide a result on the expectation of $\widetilde{\nabla_{\theta} L_i}(z, \theta_S^*)^2$, the square of the i th eigen-component
822 of $\nabla_{\theta} L(z, \theta_S^*)$, when z ranges over S .

823 **Lemma B.5.** For $i = 1, 2, \dots, p$, we have

$$\mathbb{E}_{z \sim S}[\widetilde{\nabla_{\theta} L_i}(z, \theta_S^*)^2] = \mu_i.$$

824

Proof. We adapt the proof from [5]. Let's write the eigen-decomposition of F_S :

$$F_S = V \Lambda V^{\top}$$

825 where $\Lambda = \text{diag}\{\mu_i\}_{i=1}^p$. Then

$$\begin{aligned} \Lambda &= V^{\top} F_S V \\ &= V^{\top} \mathbb{E}_{z \sim S}[\nabla_{\theta} L(z, \theta_S^*) \nabla_{\theta} L(z, \theta_S^*)^{\top}] V \\ &= \mathbb{E}_{z \sim S}[V^{\top} \nabla_{\theta} L(z, \theta_S^*) \nabla_{\theta} L(z, \theta_S^*)^{\top} V] \\ &= \mathbb{E}_{z \sim S}[(\widetilde{\nabla_{\theta} L_i}(z, \theta_S^*) \widetilde{\nabla_{\theta} L_j}(z, \theta_S^*))_{i,j=1}^p]. \end{aligned}$$

826 We obtain the desired equality by comparing diagonal terms. □

827 Next, we introduce an assumption regarding the concentration of distribution of $\widetilde{\nabla_{\theta} L_i}(z', \theta_S^*)^2$ when
828 the test example z' is sampled.

829 **Assumption B.6.** Assume there exist constants $0 < C_1 < C_2$ such that for $i = 1, 2, \dots, n$,

$$C_1 \mathbb{E}_{z \sim S}[\widetilde{\nabla_{\theta} L_i}(z, \theta_S^*)^2] \leq \widetilde{\nabla_{\theta} L_i}(z', \theta_S^*)^2 \leq C_2 \mathbb{E}_{z \sim S}[\widetilde{\nabla_{\theta} L_i}(z, \theta_S^*)^2],$$

830 with high probability over the choice of z' .

831 Assumption B.6 basically assumes that the distribution given by S can well represent the true
832 underlying distribution regarding the relative size of eigen-components. Then, we are able to derive
833 an upper bound for the RHS (right hand side) of Eq.(12).

834 **Proposition B.7.** Under Assumption B.6, for all $\lambda > 0$,

$$\text{RHS} < \frac{C_2 \sqrt{n}}{(1 - p(z', \theta_S^*)) \sqrt{C_1}} \frac{(\mu_{\min} + \lambda)^{3/2}}{\mu_{\min}}, \quad (13)$$

835 with high probability.

Proof. It is known that the eigenvalues of $\frac{1}{n} J J^{\top}$ are exactly the top n eigenvalues of F_S . Additionally,
the i th component of $J \nabla_{\theta} f(z', \theta_S^*)$ under the eigenbasis of $\frac{1}{n} J J^{\top}$ is precisely

$$\sqrt{\mu_i} \widetilde{\nabla_{\theta} f_i}(z', \theta_S^*) = \frac{\sqrt{\mu_i}}{p(z', \theta_S^*) - 1} \widetilde{\nabla_{\theta} L_i}(z', \theta_S^*),$$

836 for $i = 1, 2, \dots, n$.

837 For the numerator, with high probability

$$\begin{aligned}
t_{2,z',\lambda} &= \frac{1}{n} \nabla_{\theta} f(z', \theta_S^*)^{\top} J^{\top} \left(\frac{1}{n} J J^{\top} + \lambda I_n \right)^{-2} J \nabla_{\theta} f(z', \theta_S^*) \\
&= \frac{1}{n(1-p(z', \theta_S^*))^2} \sum_{i=1}^n \frac{(\sqrt{\mu_i} \widetilde{\nabla_{\theta} L_i}(z', \theta_S^*))^2}{(\mu_i + \lambda)^2} \\
&\leq \frac{1}{n(1-p(z', \theta_S^*))^2} \sum_{i=1}^n \frac{C_2 \mu_i^2}{(\mu_i + \lambda)^2} \\
&< \frac{1}{n(1-p(z', \theta_S^*))^2} \sum_{i=1}^n \mathbb{I}[\mu_i \neq 0] C_2 \\
&\leq \frac{1}{(1-p(z', \theta_S^*))^2} C_2.
\end{aligned}$$

838 For the denominator, with high probability,

$$\begin{aligned}
t_{3,z',\lambda} &= \frac{1}{n} \nabla_{\theta} f(z', \theta_S^*)^{\top} J^{\top} \left(\frac{1}{n} J J^{\top} + \lambda I_n \right)^{-3} J \nabla_{\theta} f(z', \theta_S^*) \\
&= \frac{1}{n(1-p(z', \theta_S^*))^2} \sum_{i=1}^n \frac{(\sqrt{\mu_i} \widetilde{\nabla_{\theta} L_i}(z', \theta_S^*))^2}{(\mu_i + \lambda)^3} \\
&\geq \frac{1}{n(1-p(z', \theta_S^*))^2} \sum_{i=1}^n \frac{C_1 \mu_i^2}{(\mu_i + \lambda)^3} \\
&\geq \frac{1}{n(1-p(z', \theta_S^*))^2} \frac{C_1 \mu_{\min}^2}{(\mu_{\min} + \lambda)^3}.
\end{aligned}$$

839 We finish the proof by combining these two bounds. \square

840 We move to analyze the LHS (left hand side) of Eq. (12). Before that, we need two additional technical
841 assumptions.

842 **Assumption B.8.** Assume there exist constants $C_3 > 0$ and $\varepsilon_g > 0$ such that for $i = 1, 2, \dots, n$,

$$\tilde{g}_{z',i}^2 \leq C_3 \mu_i^{1+\varepsilon_g}.$$

843

844 **Remark B.9.** Because $g_{z'} = \frac{1}{n} J^{\top} \alpha_{z'}$, for μ_i equal to 0, $\tilde{g}_{z',i}$ must also be 0, which can be easily
845 shown through the singular value decomposition of J . Here with Assumption B.8, intuitively we want
846 to make sure $\tilde{g}_{z',i}^2$ is super-linear with respect to μ_i .

847 **Assumption B.10.** Assume $\alpha_{z'}$ is in the column space of J .

848 **Remark B.11.** A special case for this assumption is when J has rank $n - 1$, where the deducted 1
849 rank is because $\sum_{i=1}^n \nabla_{\theta} L(z_i, \theta_S^*) = 0$. In this case, since $\sum_{i=1}^n \alpha_{z',i} = 0$ by simple computation,
850 $\alpha_{z'}$ is in the column space of J .

851 Now we are ready to derive a bound for the LHS of Eq. (12).

852 **Proposition B.12.** Under Assumption B.6, Assumption B.8, and Assumption B.10, for all $\lambda > 0$,

$$\text{LHS} \geq \frac{r_{z',\lambda}}{n\sqrt{C_3}} \mu_{\min}^{(1-\varepsilon_g)/2}. \quad (14)$$

853

854 *Proof.* Before we bound the denominator $\sqrt{O_{z',\lambda}}$, we first note that because $g_{z'} = \frac{1}{n} J^{\top} \alpha_{z'}$, the
855 corresponding component of $\alpha_{z'}$ under the eigenbasis of $\frac{1}{n} J J^{\top}$ can be computed as $n \tilde{g}_{z',i} \cdot \mu_i^{-1/2}$
856 for $\mu_i \neq 0$. For $\mu_i = 0$, Assumption B.10 guarantees that the component of $\alpha_{z'}$ is 0. These results
857 can be derived directly from the singular value decomposition of J . Therefore,

$$o_{z',\lambda} = \frac{1}{n} \alpha_{z'}^\top \left(\frac{1}{n} J^\top J + \lambda I_n \right)^{-1} \alpha_{z'} = \sum_{i=1}^n \mathbb{I}[\mu_i \neq 0] \frac{n \tilde{g}_{z',i}^2 / \mu_i}{\mu_i + \lambda} \leq \sum_{i=1}^n \mathbb{I}[\mu_i \neq 0] \frac{n C_3 \mu_i^{\varepsilon_g}}{\mu_i + \lambda} \leq \frac{C_3 n^2}{\mu_{\min}^{1-\varepsilon_g}}.$$

858 Then we directly derive the desired formula. \square

859 Finally, by combining all the results, we state the following formal version of Lemma 4.2.

860 **Lemma B.13** (Formal version of Lemma 4.2). *Assume Assumption B.6, Assumption B.8, and*
861 *Assumption B.10 hold. With high probability over the choice of z' , $r_{z',0^+} := \lim_{\lambda \rightarrow 0^+} r_{z',\lambda}$ exists.*
862 *Further assume $r_{z',0^+} > 0$. Then if $\mu_{\min} < C_\mu$ where C_μ is some positive value depending on z' ,*
863 *there exists some $C > 0$ such that for $0 < \lambda < C$,*

$$\dot{c}_p(\lambda; z') > 0.$$

864

Proof. We first show that with high probability the limit exists. By expanding $r_{z',\lambda} = -\nabla_{\theta} f(z', \theta_S^*)^\top (F_S + \lambda I_p)^{-1} g_{z'}$ under the eigenbasis of F_S , with high probability, the absolute value of the summation term corresponding to μ_i for $i = 1, 2, \dots, n$ is

$$\left| \frac{-\widetilde{\nabla_{\theta} f_i(z', \theta_S^*)} \tilde{g}_{z',i}}{\mu_i + \lambda} \right| \leq \frac{\sqrt{C_2 C_3}}{1 - p(z', \theta_S^*)} \frac{\mu_i^{1/2 + (1+\varepsilon_g)/2}}{\mu_i + \lambda} \leq \frac{\sqrt{C_2 C_3}}{1 - p(z', \theta_S^*)} \mu_i^{\varepsilon_g/2}.$$

865 Additionally, since $g_{z'} = \frac{1}{n} J^\top \alpha_{z'}$, we know $\tilde{g}_{z',i} = 0$ for $i > n$. As a result, $\lim_{\lambda \rightarrow 0^+} r_{z',\lambda}$ exists.

866 As $r_{z',0^+} > 0$, there exists $C^* > 0$ such that for all $0 < \lambda < C^*$, $r_{z',\lambda} > \frac{1}{2} r_{z',0^+}$. Now, let

$$C_\mu := \left(\frac{r_{z',0^+} (1 - p(z', \theta_S^*)) \sqrt{C_1}}{2n^{3/2} C_2 \sqrt{C_3}} \right)^{2/\varepsilon_g} > 0,$$

867 and

$$C := \min \left\{ C^*, \left(\frac{r_{z',0^+} (1 - p(z', \theta_S^*)) \sqrt{C_1}}{2n^{3/2} C_2 \sqrt{C_3}} \right)^{2/3} \mu_{\min}^{1-\varepsilon_g/3} - \mu_{\min} \right\}.$$

868 By direct calculation, if $\mu_{\min} < C_\mu$, then $C > 0$. Further, when $0 < \lambda < C$,

$$\text{LHS} \geq \frac{r_{z',\lambda}}{n \sqrt{C_3}} \mu_{\min}^{(1-\varepsilon_g)/2} > \frac{r_{z',0^+}}{2n \sqrt{C_3}} \mu_{\min}^{(1-\varepsilon_g)/2} > \frac{C_2 \sqrt{n}}{(1 - p(z', \theta_S^*)) \sqrt{C_1}} \frac{(\mu_{\min} + \lambda)^{3/2}}{\mu_{\min}} > \text{RHS},$$

869 which implies $\dot{c}_p(\lambda; z') > 0$. \square

870 **Remark B.14.** *Our analysis relies on the condition that μ_{\min} , the smallest non-zero eigenvalue*
871 *of F_S , remains small. This assumption aligns with established analyses demonstrating eigenvalue*
872 *concentration near zero for both the FIM and Hessian in deep neural networks near convergence*
873 *[30, 18]. These results empirically justify our treatment of μ_{\min} .*

874 We subsequently focus on a discussion of the positivity condition $r_{z',0^+} > 0$ which constitutes the
875 remainder of this section.

876 **Discussion of $r_{z',0^+} > 0$.** Here we show that, in the special case where $z' = z_i \in S$ for some
877 $i \in [n]$, we have $r_{z',0^+} > 0$. We first write the singular value decomposition $J = U \Sigma V^\top$, and denote
878 the j th singular value of J as σ_j . Then,

$$\begin{aligned} \lim_{\lambda \rightarrow 0^+} r_{z',\lambda} &= \lim_{\lambda \rightarrow 0^+} -\frac{1}{n} \nabla_{\theta} f(z', \theta_S^*)^\top J^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-1} \alpha_{z'} \\ &= \lim_{\lambda \rightarrow 0^+} \frac{1}{n(1-p_i)} \nabla_{\theta} L(z_i, \theta_S^*) J^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-1} \alpha_{z'} \\ &= \frac{1}{1-p_i} \lim_{\lambda \rightarrow 0^+} e_i^\top \frac{1}{n} J J^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-1} \alpha_{z'} \\ &= \frac{1}{1-p_i} \lim_{\lambda \rightarrow 0^+} \sum_{j=1}^n \frac{\sigma_j^2/n}{\sigma_j^2/n + \lambda} (U^\top e_i)_j (U^\top \alpha_{z'})_j \\ &= \frac{1}{1-p_i} \sum_{j=1}^n \mathbb{I}[\sigma_j \neq 0] (U^\top e_i)_j (U^\top \alpha_{z'})_j, \end{aligned}$$

where e_i is the i th standard basis vector. By Assumption B.10, $\alpha_{z'}$ is in the column space of J , so $\sigma_j = 0$ implies $(U^\top \alpha_{z'})_j = 0$. Therefore,

$$r_{z',0+} = \lim_{\lambda \rightarrow 0+} r_{z',\lambda} = \frac{1}{1-p_i} e_i^\top \alpha_{z'} = \frac{\alpha_{z',i}}{1-p_i} > 0,$$

because $p_i < 1$ and $\alpha_{z',i} = \alpha_{z_i,i} = \mathbb{E}_{A \sim D_a}[f(z_i, \theta_A^*) | z_i \in A] - \mathbb{E}_{A \sim D_a}[f(z_i, \theta_A^*)]$ is the expected change in model output of z_i itself when z_i is included in the training set, which is positive.

B.4 Effects of Gradient Projection

Our derivation does not rely on specific properties of J . When incorporating gradient projection via Eq. (4), as outlined in Remark 4.5, the projection matrix P can be systematically applied to all gradient terms. We now analyze how gradient projection affects μ_{\min} , the smallest non-zero eigenvalue of F_S (equivalently, of $\frac{1}{n} J J^\top$). Under projection, this eigenvalue corresponds to the smallest non-zero eigenvalue of $\frac{1}{n} J P P^\top J^\top$.

In practical applications, gradient projection approximately preserves the dominant eigenvalues of F_S . Consequently, when F_S exhibits vanishingly small (but non-zero) eigenvalues near index $\min\{n, \tilde{p}\}$, the projected matrix $\frac{1}{n} J P P^\top J^\top$ retains a comparably small non-zero eigenvalue, validating the critical condition in Lemma B.13.

Unlike Arnoldi-based method [31] that aims to preserve top eigenvalues by design, we resort to the standard subspace embedding results [34] for random projection based methods [5, 27]: For common choices of P (e.g. Gaussian random projection) and $0 < \varepsilon_{\text{rp}}, \delta < 1$, if $\tilde{p} = \Theta((n + \ln(1/\delta))\varepsilon_{\text{rp}}^{-2})$, then with probability $1 - \delta$, P satisfies that for any $v \in \mathbb{R}^n$,

$$(1 - \varepsilon_{\text{rp}}) \|J^\top v\|_2 \leq \|P^\top J^\top v\|_2 \leq (1 + \varepsilon_{\text{rp}}) \|J^\top v\|_2.$$

This implies, by min-max theorem,

$$(1 - \varepsilon_{\text{rp}}) \mu_i(J J^\top) \leq \mu_i(J P P^\top J^\top) \leq (1 + \varepsilon_{\text{rp}}) \mu_i(J J^\top),$$

where $\mu_i(\cdot)$ stands for the i th biggest eigenvalue. This guarantees an approximation of top eigenvalues.

B.5 Proof of Proposition 4.7

Proof of Proposition 4.7. We utilize Lemma B.3 to obtain

$$\begin{aligned} t_{k,z',\lambda} &= \frac{1}{n} \nabla_\theta f(z', \theta_S^*)^\top J^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-k} J \nabla_\theta f(z', \theta_S^*), \\ r_{z',\lambda} &= -\frac{1}{n} \nabla_\theta f(z', \theta_S^*)^\top J^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-1} \alpha_{z'}, \end{aligned}$$

where $k = 1, 2, 3$. Further,

$$o_{z',\lambda} = \frac{1}{n} \alpha_{z'}^\top \left(\frac{1}{n} J J^\top + \lambda I_n \right)^{-1} \alpha_{z'}.$$

We point out that LHS of Eq.(8) lying in $[0, 1]$ is the direct result of applying the generalized Cauchy-Schwarz inequality on $-J \nabla_\theta f(z', \theta_S^*)$ and $\alpha_{z'}$ with inner product defined by positive definite matrix $\frac{1}{n} (\frac{1}{n} J J^\top + \lambda I_n)^{-1}$, assuming $r_{z',\lambda} > 0$. Similarly, the RHS of Eq.(8) is shown bounded when the inequality is applied on $-(\frac{1}{n} J J^\top + \lambda I_n)^{-1} J \nabla_\theta f(z', \theta_S^*)$ and $-J \nabla_\theta f(z', \theta_S^*)$ with the same inner product. We use the fact that $t_{2,z',\lambda} \geq 0$ as $(\frac{1}{n} J J^\top + \lambda I_n)^{-2}$ is positive semi-definite. \square

C Details and Extra Experiments of the Surrogate Indicator

C.1 Visualization of the Average Surrogate Indicator

To provide better insights about the proposed surrogate indicator, we present the curve of $\bar{\xi}_{T,\lambda}$ as a function of λ in Figure 5. In all experiment settings, empirically, $\bar{\xi}_{T,\lambda}$ is a monotonic function of λ with a *transitional phase* near $\bar{\xi}_{T,\lambda} = 0.5$, indicating a good sensitivity to λ around this value, and supporting our choice of threshold in Algorithm 1.

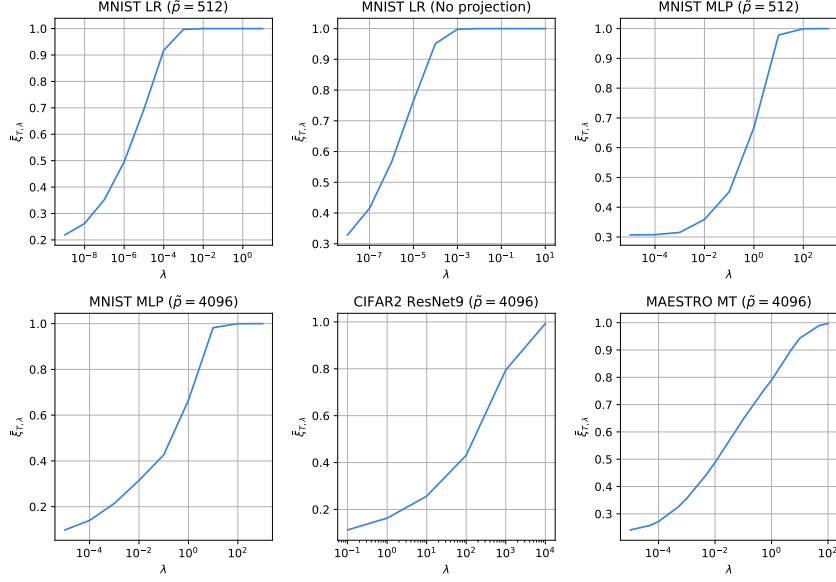


Figure 5: The curve of $\bar{\xi}_{T,\lambda}$ as a function of λ . Each subfigure corresponds to an experiment setting outlined in Section 4.3.

C.2 Computational Resources and Dataset Licenses

The experiments for the surrogate indicator are done on an A40 GPU in around 10 hours, excluding model retraining (we reused some model checkpoints provided by the dattri library to avoid extensive model retraining). For the datasets we use: MNIST-10 dataset holds CC BY-SA 3.0 license; CIFAR-10 dataset holds CC-BY 4.0 license; MAESTRO dataset holds CC BY-NC-SA 4.0 license.

C.3 Effects of Different Subset Fractions

We provide an extended analysis of how the subset fraction $a/|S|$ influences the surrogate indicator, considering values in $\{0.25, 0.5, 0.75\}$, where 0.5 corresponds to the setting used in earlier experiments. Figure 6 illustrates the surrogate indicator’s behavior for each subset fraction, shown in green (0.25), blue (0.5), and violet (0.75). While the magnitude of the LDS varies with the subset fraction, its overall trend as a function of λ remains similar across different values of subset fraction. This consistency suggests that the surrogate indicator is robust to changes in subset size across different experimental settings.

D Code Availability

Our code is publicly available at <https://github.com/data-attribution-hp/data-attribution-hp>.

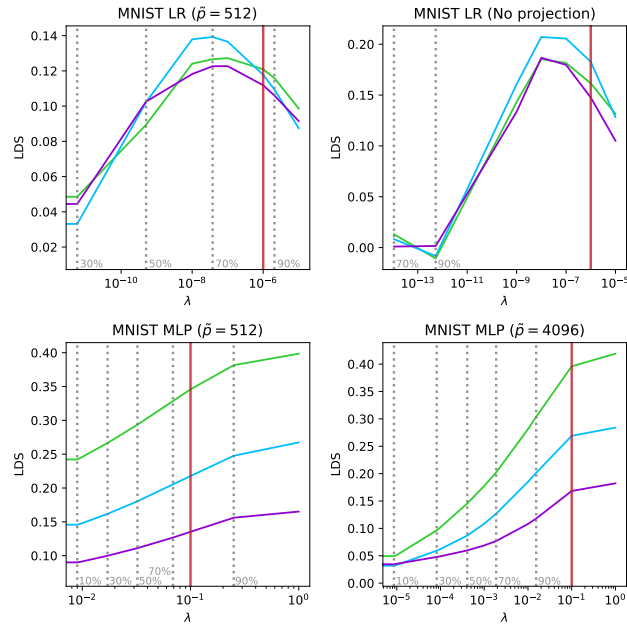


Figure 6: Experiment results with different values of subset fraction. Green, blue, and violet curves illustrate LDS- λ relationship with subset fraction 0.25, 0.5, and 0.75, respectively.