# Data Attribution for Model/Learning Based Control (Extended Abstract)

**Anonymous Author(s)**
Affiliation
Address
`email`

## Abstract

Ensuring safety of learning-based robotic controllers requires understanding which training data influences performance, yet identifying influential trajectories through exhaustive retraining is computationally prohibitive. We introduce a framework using influence functions to efficiently approximate the impact of individual training trajectories on learned dynamics and control performance. We formulate IF1 to estimate effects on model accuracy and IF2 to quantify impacts on LQR control cost—a proxy for tracking error and stability. Empirical validation demonstrates strong correlations between influence predictions and ground truth.

## 1 Introduction

The deployment of learning-based control systems in safety-critical robotic applications—from autonomous vehicles Levinson et al. [2011] to surgical robots Shademan et al. [2016]—requires rigorous quality assurance of training data. Corrupted or low-quality training trajectories can propagate through the learning pipeline, resulting in controllers that exhibit unsafe behaviors: excessive tracking errors, oscillations, or instability Dean et al. [2020]. Understanding which training data most influences controller performance is critical for ensuring reliable deployment Amodei et al. [2016].

Traditional LOO retraining for assessing data importance is computationally prohibitive for realistic datasets with thousands of trajectories. Recent advances in machine learning and large language models demonstrate remarkable success using influence functions for efficient data attribution Grosse et al. [2023], including identifying mislabeled examples Koh and Liang [2017], detecting memorization Feldman and Zhang [2020], and guiding data curation Hammoudeh and Lowd [2022]. This motivates applying influence functions to learning-based control, where understanding how training trajectories impact controller performance is critical for safety-critical robotic applications.

**Our Contribution.** We introduce influence functions for efficient data attribution in learning-based control, enabling practitioners to identify influential training trajectories in seconds rather than hours. Specifically:

- **IF1 (Dynamics-level)**: Estimates how removing a trajectory affects model predictive accuracy—critical for ensuring accurate state estimation in safety-critical regions.
- **IF2 (Control-level)**: Quantifies impact on LQR control cost, a proxy for tracking error and stability margins. This directly connects training data to controller safety performance.

To our knowledge, this is the first work connecting influence functions to control performance, enabling efficient quality assurance for safety-critical learning-based controllers. This extended abstract summarizes key contributions from our recent work Li et al. [2025], with emphasis on safety implications for robotic systems. Full mathematical derivations and extended experimental results appear in Li et al. [2025].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

## 2 Problem Setup

**System.** Consider a discrete-time system $\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t, \boldsymbol{u}_t) + \boldsymbol{w}_t$ with state $\boldsymbol{x}_t \in \mathbb{R}^{n_x}$, control $\boldsymbol{u}_t \in \mathbb{R}^{n_u}$, and Gaussian noise $\boldsymbol{w}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma}_w)$.

**Learning.** We learn a linear model $\hat{\boldsymbol{x}}_{t+1} = \boldsymbol{A}_{\boldsymbol{\theta}}\boldsymbol{x}_t + \boldsymbol{B}_{\boldsymbol{\theta}}\boldsymbol{u}_t$ from dataset $\mathcal{D} = \{\tau_1, \ldots, \tau_N\}$ by minimizing:

$$L(\boldsymbol{\theta}, \mathcal{D}) = \sum_{k=1}^{N} \mathcal{L}_k(\boldsymbol{\theta}) = \sum_{k=1}^{N} \sum_{s \in \tau_k} \|\boldsymbol{x}_{t+1,s} - \boldsymbol{\Phi}_s \boldsymbol{\theta}\|_2^2$$

where $\mathcal{L}_k(\boldsymbol{\theta})$ is the loss on trajectory $\tau_k$.

**Control.** Using learned dynamics $(\boldsymbol{A}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{B}_{\hat{\boldsymbol{\theta}}})$, we design an LQR controller minimizing $J(\boldsymbol{\theta}) = \mathrm{Tr}(\boldsymbol{P}(\boldsymbol{\theta}))$ where $\boldsymbol{P}(\boldsymbol{\theta})$ solves the Discrete Algebraic Riccati Equation (DARE). The cost $J(\boldsymbol{\theta})$ represents expected tracking error and state deviation—key safety metrics.

**Goal.** Efficiently estimate how removing trajectory $\tau_k$ affects: (1) model accuracy $L_{pred}(\boldsymbol{\theta})$ on test data, and (2) control performance $J(\boldsymbol{\theta})$, without exhaustive retraining.

## 3 Method Overview

### 3.1 IF1: Influence on Model Accuracy

Removing $\tau_k$ changes optimal parameters from $\hat{\boldsymbol{\theta}}$ to $\hat{\boldsymbol{\theta}}_{\setminus k}$. Using influence functions Koh and Liang [2017], we approximate:

$$\Delta\hat{\boldsymbol{\theta}}_k \approx \boldsymbol{H}_{\hat{\boldsymbol{\theta}}}^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}_k(\hat{\boldsymbol{\theta}})$$

where $\boldsymbol{H}_{\hat{\boldsymbol{\theta}}} = \nabla^2_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}, \mathcal{D})$ is the Hessian. The influence on test set loss is:

$$IF1(\tau_k, L_{pred}) = (\nabla_{\boldsymbol{\theta}} \mathcal{L}_k(\hat{\boldsymbol{\theta}}))^T \boldsymbol{H}_{\hat{\boldsymbol{\theta}}}^{-1} \nabla_{\boldsymbol{\theta}} L_{pred}(\hat{\boldsymbol{\theta}}) \tag{1}$$

**Safety relevance:** IF1 identifies trajectories whose removal degrades model accuracy in safety-critical state regions, potentially leading to poor state estimation and unsafe control decisions.

### 3.2 IF2: Influence on Control Performance

More critically for safety, we quantify how $\tau_k$ affects LQR cost:

$$IF2(\tau_k, J) = (\nabla_{\boldsymbol{\theta}} \mathcal{L}_k(\hat{\boldsymbol{\theta}}))^T \boldsymbol{H}_{\hat{\boldsymbol{\theta}}}^{-1} \nabla_{\boldsymbol{\theta}} J(\hat{\boldsymbol{\theta}}) \tag{2}$$

The key challenge is computing $\nabla_{\boldsymbol{\theta}} J(\hat{\boldsymbol{\theta}})$, which requires tracing sensitivities through the DARE solution. For each parameter $\theta_m$, we compute $S_m = \frac{\partial \boldsymbol{P}}{\partial \theta_m}$ by solving a Lyapunov equation:

$$S_m - \boldsymbol{A}_{CL}^T S_m \boldsymbol{A}_{CL} = -\frac{\partial \mathcal{R}}{\partial \theta_m}$$

where $\mathcal{R}(\boldsymbol{P}, \boldsymbol{\theta}) = 0$ is the DARE and $\boldsymbol{A}_{CL} = \boldsymbol{A}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{B}_{\hat{\boldsymbol{\theta}}}\boldsymbol{K}(\hat{\boldsymbol{\theta}})$ is the closed-loop system matrix. This yields $\nabla_{\boldsymbol{\theta}} J(\hat{\boldsymbol{\theta}}) = [\mathrm{Tr}(S_1), \ldots, \mathrm{Tr}(S_p)]^T$.

**Safety relevance:** IF2 identifies trajectories that degrade control performance—manifesting as increased tracking errors, larger state deviations, or reduced stability margins in deployed systems.

## 4 Empirical Validation

**Systems.** We validate on two linear systems: Single-link manipulator analogue (2 states, 1 input),Two-link manipulator analogue (4 states, 2 inputs).

**Setup.** Training sets: 30-50 trajectories of 25-30 steps. Test sets: 20 trajectories. LQR weights: $\boldsymbol{Q}_c = \boldsymbol{I}$, $\boldsymbol{R}_c = 0.1\boldsymbol{I}$. For each trajectory $\tau_k$, we computed IF1/IF2 predictions and compared against ground truth via explicit retraining on $\mathcal{D} \setminus \{\tau_k\}$.

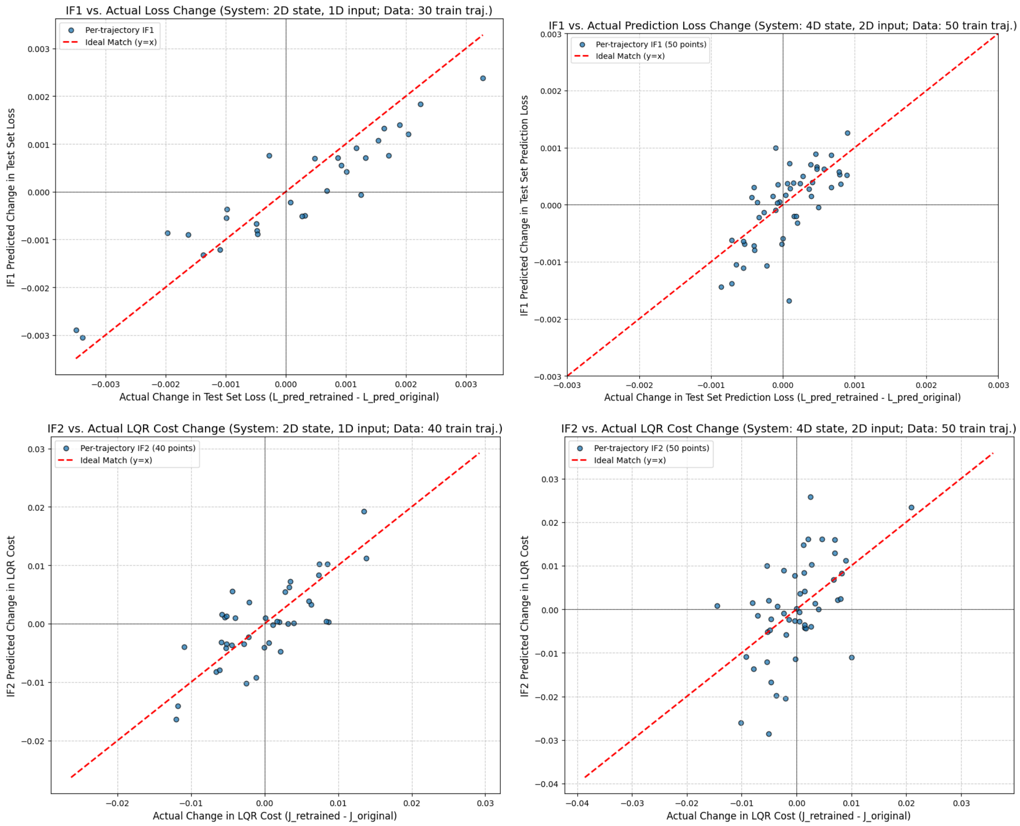


Figure 1: **Validation of influence functions.** Top: IF1 predictions vs. actual changes in test set prediction loss for S1 (left, $r = 0.93$) and S2 (right, $r = 0.69$). Bottom: IF2 predictions vs. actual changes in LQR control cost for S1 (left, $r = 0.71$) and S2 (right, $r = 0.64$). Strong positive correlations demonstrate that influence functions efficiently identify training trajectories that impact controller reliability.

**Results.** Figure 1 demonstrates strong agreement between influence predictions and ground truth. Critically, IF2 successfully identifies trajectories that affect downstream control performance—enabling efficient quality assurance for safety-critical applications. The moderate scatter in higher-dimensional S2 reflects approximation error from first-order Taylor expansion; second-order corrections could improve accuracy (future work).

**Computational Efficiency.** Computing IF1/IF2 for all $N = 50$ trajectories required <1 second, versus 8.3 minutes for exhaustive LOO retraining—a 500× speedup.

# 5 Implications for Safe Robotic Control

## 5.1 Quality Assurance Applications

Our framework enables several safety-critical capabilities:

**Corrupted Data Detection.** Trajectories with anomalously high IF2 scores may indicate corrupted or adversarial data that degrades controller performance through systematic model bias. Such data could arise from sensor failures, communication errors, or malicious attacks. Practitioners can flag these for manual review and potential removal to improve controller safety margins.

**Critical Data Identification.** Trajectories with high positive IF2 influence are critical for maintaining performance in key operating regions. Their removal would increase tracking error and potentially compromise stability— these should be preserved, verified for accuracy, and prioritized during data storage and backup procedures.

**Coverage Verification.** Low-influence regions may indicate under-represented operating regimes where the learned model lacks sufficient training support. In safety-critical applications (autonomous driving near obstacles, surgical robots near anatomical boundaries), ensuring adequate data coverage across all operational conditions is essential for reliable deployment.

**Active Data Collection.** IF scores can guide additional data gathering in regions where current training data provides weak support for safe control. This enables targeted data acquisition strategies that efficiently improve controller robustness in underrepresented or high-risk scenarios.

### 5.2 Limitations and Future Work

**Linear Dynamics and LQR Control.** The current framework is restricted to linear systems and LQR control. Extension to nonlinear dynamics (neural network models) and nonlinear controllers (MPC, iLQR) requires tractable gradient computation through nonlinear Riccati equations and handling influence propagation through neural network parameters.

**Stochastic Modeling.** Our formulation treats process noise $\boldsymbol{w}_t$ as given but does not model how training data influences noise covariance estimation. Incorporating influence functions for jointly learned dynamics and noise models would enable more complete data attribution in stochastic control settings.

**End-to-End Data-Driven Control.** Modern methods like Data-Enabled Predictive Control (DeePC) and Direct Policy Control (DPC) bypass explicit system identification. Extending influence functions to these frameworks would require tracing data influence directly through Hankel matrices or policy parameterizations to closed-loop performance.

**Hessian Computation.** Computing the full Hessian $\boldsymbol{H}_{\hat{\boldsymbol{\theta}}}$ becomes prohibitive for high-dimensional systems. Future work should explore Gauss-Newton or Fisher information approximations, implicit Hessian-vector products via automatic differentiation, and Hessian-free methods using conjugate gradient or Neumann series to enable scaling to large neural network models.

**Approximation Error.** First-order Taylor approximation may be inaccurate for large parameter perturbations. Deriving theoretical error bounds or developing adaptive second-order corrections would strengthen reliability guarantees.

## 6 Conclusion

We introduced influence functions (IF1, IF2) for efficient data attribution in learning-based control systems. IF1 quantifies how training trajectories affect model predictive accuracy, while IF2 represents the first method to efficiently trace training data influence through system identification to control performance—directly connecting individual trajectories to LQR cost and controller reliability. This framework enables quality assurance without exhaustive retraining. It provides practitioners with a computationally tractable tool for identifying influential data in safety-critical applications.

## References


Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20:633–679, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

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 preprint arXiv:2308.03296*, 2023.

Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 157–172. Springer, 2022.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. *IEEE Intelligent Vehicles Symposium*, pages 163–168, 2011.

Jiachen Li, Shihao Li, Jiamin Xu, Soovadeep Bakshi, and Dongmei Chen. Influence functions for data attribution in linear system identification and lqr control, 2025. URL `https://arxiv.org/abs/2506.11293`.

Azad Shademan, Ryan S Decker, Justin D Opfermann, Simon Leonard, Axel Krieger, and Peter CW Kim. Supervised autonomous robotic soft tissue surgery. *Science Translational Medicine*, 8(337): 337ra64–337ra64, 2016.