# SHERLOC: STATISTICAL HYPOTHESIS EVALUATION AND RECOVERY FOR LEARNING OPPONENT CHARACTERISTICS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

When used in automated decision-making systems, machine learning (ML) models are vulnerable to data-manipulation attacks. Some defense mechanisms (e.g., adversarial regularization) directly affect the ML models while others (e.g., anomaly detection) act within the broader system. However, many defenses operate under a predefined threat model that assumes characteristics of the adversary, including their knowledge, capabilities, and objectives, are known and fixed. In practice, these characteristics may be unknown or evolve over time, creating a mismatch between the assumptions of the defender and the true attacker.

In this paper, we study a new perspective on adversarial defense: focusing on the attack*er*, rather than the attack. We present and demonstrate **SHERLOC**, a framework for adversarial learning forensics – namely, a *detective* that learns about the attacker from observed attack(s).

We prove that, without additional knowledge, the attacker is non-identifiable — multiple potential attackers would perform the same observed attack. This reveals an important distinction between *reconstructing* an attacker's behavior and *recovering* the attacker that generated it. We demonstrate empirically that SHERLOC is able to reconstruct a hypothesis attacker that is consistent with the observed attacks against various defender models (i.e., generalized least-squares linear regressors, logistic regressors, and polynomial classifiers) across various security-critical domains. Finally, we show that these inferred attackers need not match the ground truth attacker exactly to have downstream utility. Defenses using the SHERLOC-inferred attackers achieve approximately $90-100\%$ of robustness improvement that would be achieved if the defender had access to the ground-truth attacker.

## 1 INTRODUCTION

Despite their capabilities, machine learning (ML) systems are vulnerable to adversarial attacks — small perturbations to input data that cause the model to predict incorrectly (Szegedy et al., 2014; Moosavi-Dezfooli et al., 2015; Dabouei et al., 2019). However, many defenses are designed around a fixed threat model that makes assumptions about the adversary, including their knowledge of the target system, their capabilities and limitations, and their objective (Athalye et al., 2018; Cohen et al., 2019). When these assumptions do not match the adversary, the resulting defense provides limited protection, contributing to an ongoing arms race between increasingly sophisticated attacks and defenses (Tramer et al., 2020).

This reliance on a predefined attacker presents a fundamental challenge in real-world settings, where an adversary's knowledge, capabilities, and objectives are non-stationary and/or unknown. Robustness against such adversaries therefore requires moving beyond defenses that assume these characteristics about the attacker. Instead, we ask whether information about an attack*er* can be inferred from the attacks themselves, enabling the defender to reason about an otherwise unknown or evolving threat model.

To this end, this paper pursues two distinct but complementary goals.



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

First, we seek to understand what can be learned about an attacker from the attacks they leave behind. The importance of "knowing one's enemy" has long been recognized in adversarial settings (Tzu, Circa 500 B.C.E.), and in real-world ML systems, knowledge about the attacker can inform risk assessment, attack detection, and robust design of future systems. In this paper, we focus on attackers against simple learners such as linear and polynomial models, which are often favored in safety-critical applications for their interpretability, transparency, and ease of validation. Under these settings, we (i) highlight the limitations of exact attacker recovery and (ii) investigate the ability to reconstruct hypothesis attackers that behave consistently with the observed attacks via optimization (i.e., SHERLOC-inferred attackers). Thus, our first goal is to characterize what information about an attacker can be **recovered** vs. **reconstructed** from their observed behavior, and under what conditions.

Second, we ask whether the information inferred about an attacker can be used to construct a more informed defense. In practice, the attacker's beliefs, objective, and capabilities are almost never known by the defender; the only information that is reliably known by the defender is the actual input to the defender's model. Hence, rather than assuming fixed attacker parameters, the defender should infer them from the attacks they observe and adapt their defensive strategy accordingly. As one concrete instantiation, we use the attacker parameters inferred by our framework to construct an attacker-specific adversarial regularizer.

Concretely, we make the following contributions:

1. We establish a fundamental limitation of attacker forensics: distinct attackers can induce the same attack behavior and are therefore non-identifiable from attack observations alone. For linear learners, we mathematically illustrate this phenomenon, yielding non-trivial equivalence classes of behaviorally indistinguishable attackers.
2. We introduce SHERLOC, a general framework and method for inferring attacker hypotheses behaviorally consistent with the observed attack(s). Motivated by non-identifiability, SHERLOC distinguishes between reconstructing attacker behavior and recovering the ground-truth attacker parameters.
3. We empirically demonstrate SHERLOC on linear and polynomial learning methods, then apply the inferred attacker to defense. We show that SHERLOC can accurately reconstruct attacker behavior even when the ground-truth attacker is not exactly recovered, and that the inferred attackers remain useful for downstream defense, achieving approximately 90–100% of the robustness improvement obtained with access to the ground-truth attacker.

In total these contributions help to solidify the foundation of the RED (Reverse Engineering of Deception) research area.

## 2 Threat Model

We consider two players, the defender/learner `DFDR` and the attacker/adversary `ATKR`. At deployment time, `ATKR` manipulates inputs to `DFDR`, while `DFDR` observes the resulting attack(s) and aims to infer properties of the attacker that generated them. Based on the classical taxonomy widely used in adversarial machine learning (Barreno et al., 2006; 2010; Huang et al., 2011), an attacker is defined by three fundamental aspects: what it *knows* about the target model, what actions it *can* perform, and what it *aims* to achieve. In our framework, we model these aspects of `ATKR` as the tuple

$$(\mathcal{K}, \mathcal{C}, \mathcal{O})$$

where:

1. $\mathcal{K}$: Their knowledge about `DFDR`. What does `ATKR` think `DFDR`'s prediction function is?
2. $\mathcal{C}$: Their capability. What perturbations to the data can `ATKR` make?
3. $\mathcal{O}$: Their objective. What function does `ATKR` try to optimize?

This decomposition is central to SHERLOC. Beyond corresponding to the standard conceptual dimensions of an attacker, $\mathcal{K}, \mathcal{C}, \mathcal{O}$ play distinct roles in the process by which an attack is generated: $\mathcal{K}$ determines the model whose behavior the attacker reasons about, $\mathcal{C}$ determines which manipulations

are feasible, and $\mathcal{O}$ determines which outcomes the attacker prefers. Consequently, inferring each component constitutes a qualitatively different inverse problem.

In the forward problem, ATKR uses $(\mathcal{K}, \mathcal{C}, \mathcal{O})$ to generate an attack. ATKR aims to perturb the input within its capability constraint $\mathcal{C}$ such that the output of its knowledge model $\mathcal{K}$ shifts as far as possible in the directions weighted by its objective $\mathcal{O}$.

We denote the optimal attack on a benign input $\boldsymbol{x}$ as the following function of the input and ATKR's parameters:

$$\boldsymbol{\alpha}_{opt}(\boldsymbol{x}; \texttt{ATKR}) = \boldsymbol{\alpha}_{opt}(\boldsymbol{x}; \mathcal{K}, \mathcal{C}, \mathcal{O}) \tag{1}$$

$$= \arg\max_{\boldsymbol{\alpha} \in \mathcal{A}_{\mathcal{C}}} g_{\mathcal{O}}(f_{\mathcal{K}}, \boldsymbol{x}, \boldsymbol{\alpha}) \tag{2}$$

where:

- $f_{\mathcal{K}}$ is the prediction function of the model used by ATKR,
- $\mathcal{A}_{\mathcal{C}}$ is the set of attacks ATKR is capable of,
- and $g_{\mathcal{O}}$ is the gain function defined by ATKR's objective.

In the inverse problem, which is the focus of SHERLOC, DFDR observes a series of resulting attack(s):

$$\boldsymbol{A}_{obs}(\boldsymbol{X}_k; \texttt{ATKR}) = \{\boldsymbol{\alpha}_{obs}(\boldsymbol{x}; \texttt{ATKR}) \ : \ \boldsymbol{x} \in \boldsymbol{X}_k\} \tag{3}$$

where $\boldsymbol{X}_k = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ is a set of distinct benign inputs and $\boldsymbol{\alpha}_{obs}(\cdot)$ may or may not be optimal.

Then, it reasons backward about the $(\mathcal{K}, \mathcal{C}, \mathcal{O})$ that could have produced it. Importantly, solving the inverse problem need not uniquely identify the true attacker ATKR$^*$. In Section 3, we show that an attacker is non-identifiable — distinct parameterizations may induce identical attacks. We therefore distinguish throughout the paper between *reconstructing* an attacker's behavior (i.e., finding parameters that explain the observed attacks) and *recovering* parameters that are close to the attacker's true $(\mathcal{K}^*, \mathcal{C}^*, \mathcal{O}^*)$.

We begin by instantiating this framework for linear learners, including linear regression and logistic regression. The linear setting is useful for analysis, but exposes a fundamental limitation of reverse-engineering exact attacker parameters from observed attacks — linearity severely restricts the information exposed by an observed attack. We then extend our study to polynomial learners, where observing multiple attacks against different inputs can provide repeated evidence about the attacker. The specific instantiations and corresponding ATKR parameterizations in these three settings are presented in Appendix A. Finally, we ask whether exact parameter recovery is necessary for a downstream defender: even when DFDR cannot uniquely identify ATKR, the recovered representation may capture enough of the attacker's behavior to support effective defense.

In Section 3, we discuss the key challenges of reverse-engineering ATKR and characterize the resulting inference problem. In Section 4, we demonstrate SHERLOC's empirical behavior and applicability to defense across linear and polynomial learners. Together, these settings illustrate both the theoretical foundation and the practical applicability of our approach.

Note that we assume DFDR knows the parameterized form of the attacker, but not the specific parameter values, which are inferred from the observed attack using our framework. The flexibility of these parameters allows our framework to be domain-agnostic, freely admitting other attackers, defenders, and parameterizations thereof.

## 3 Recovering Attacker Parameters

Suppose DFDR has no prior knowledge (other than the parameterization) of ATKR's knowledge ($\mathcal{K}$), capability[1] ($\mathcal{C}$), or objective ($\mathcal{O}$). We now consider the inverse problem faced by DFDR: Given one or more observed attacks, what can DFDR infer about the $(\mathcal{K}, \mathcal{C}, \mathcal{O})$ that generated them?

At first glance, one might formulate this problem by searching for $(\mathcal{K}, \mathcal{C}, \mathcal{O})$ for which the observed attacks are optimal. However, reproducing an observed attack does not necessarily identify the exact

[1]Except that the observed attack is clearly something ATKR is capable of.

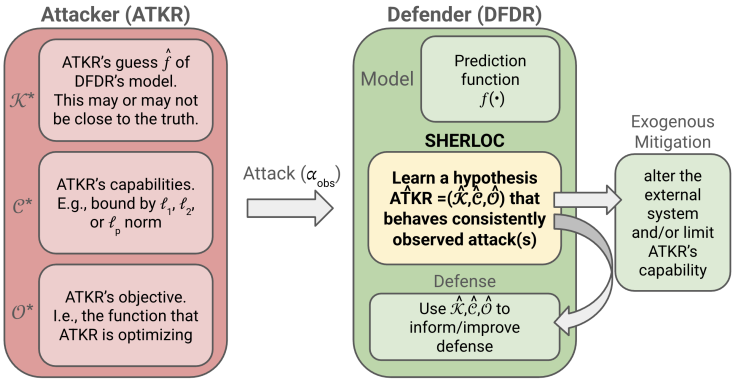


Figure 1: Schematic overview of our framework within the overall attacker-defender system. In this paper, we consider the example cases in which $f$ is a linear regression, logistic regression, or polynomial regression model.

attacker that generated it. Multiple values of $\mathcal{K}, \mathcal{C}, \mathcal{O}$ may induce exactly the same optimal attack. This distinction between *explaining an attack* and *identifying its source* ATKR is fundamental to the reverse-engineering problem.

In this section, we aim to answer the question: **What can be learned about an attacker from the attacks they leave behind?** We first formalize this problem and show that, even in the controlled linear setting, exact attacker parameters are generally non-identifiable from an observed attack. We then introduce **SHERLOC**, which treats reverse engineering as inference over the equivalence class of attacker hypotheses consistent with the observed behavior. This perspective also motivates two distinct notions of success that we evaluate empirically: whether the inferred attacker hypothesis $\hat{\text{ATKR}}$ can *reconstruct* the observed attack behavior, and whether its inferred parameters actually *recover* those of the ground-truth attacker.

### 3.1 Non-identifiability Property of Attacker

The most canonical question is: is the attacker identifiable? That is, does the observed attack uniquely determine the parameters of ATKR? Intuitively (and as we show mathematically below) the answer is, in general, no — two different attackers (e.g., suppose both their objectives and capabilities differ) may end up performing the same attack. While one can easily construct particular examples to show instances of this phenomenon, in what follows we show a much stronger result for linear attackers. That is, for *any* attack $\boldsymbol{\alpha}$ and any fixed two-of-the-three parameters $(\mathcal{K}, \mathcal{C}, \mathcal{O})$, there is a set of attackers that are behaviorally indistinguishable without additional information.

We define the equivalence class $\mathcal{S}_{\boldsymbol{\alpha}}$ of attackers for an attack vector $\boldsymbol{\alpha}$ as follows:

**Definition 3.1** (Attacker equivalence class)**.** Let ATKR be defined by parameters $(\mathcal{K}, \mathcal{C}, \mathcal{O})$. The optimal attack for ATKR is denoted $\boldsymbol{\alpha}_{opt}(\mathcal{K}, \mathcal{C}, \mathcal{O})$. We define the set $\mathcal{S}_{\boldsymbol{\alpha}}$ as:

$$\text{ATKR} \in \mathcal{S}_{\boldsymbol{\alpha}} \textbf{ iff } \boldsymbol{\alpha} = \boldsymbol{\alpha}_{opt}(\mathcal{K}, \mathcal{C}, \mathcal{O}) \tag{4}$$

**Theorem 3.2** (Non-identifiablity of attacker)**.** *For a linear DFDR, the following hold for any attack* $\boldsymbol{\alpha}$*:*

1. $\forall \mathcal{K}, \mathcal{C},\ \exists \mathcal{O}$ *s.t.* $ATKR \in \mathcal{S}_{\boldsymbol{\alpha}}$,
2. $\forall \mathcal{K}, \mathcal{O},\ \exists \mathcal{C}$ *s.t.* $ATKR \in \mathcal{S}_{\boldsymbol{\alpha}}$,
3. $\forall \mathcal{C}, \mathcal{O},\ \exists \mathcal{K}$ *s.t.* $ATKR \in \mathcal{S}_{\boldsymbol{\alpha}}$

The full proof of Theorem 3.2 is presented in Appendix B.

Theorem 3.2 states that the tuple of ATKR parameters $(\mathcal{K}, \mathcal{C}, \mathcal{O})$ is **non-identifiable**. It is a fundamental limitation on what can be inferred from attack observations alone. Even if DFDR knows two components of the attacker exactly, observing a single optimal attack does not, in general, uniquely

determine the third. For any fixed values of two components, the remaining component can be chosen so that the same attack remains optimal. Thus, an observed attack identifies an *equivalence class* of possible attackers rather than a unique point in the space of attacker parameters.

This observation has an important consequence for how reverse engineering should be evaluated. Suppose an inference procedure returns parameters $(\hat{\mathcal{K}}, \hat{\mathcal{C}}, \hat{\mathcal{O}})$ that reproduce the observed attack. Theorem 3.2 says that there is no guarantee that $(\hat{\mathcal{K}}, \hat{\mathcal{C}}, \hat{\mathcal{O}}) = (\mathcal{K}^*, \mathcal{C}^*, \mathcal{O}^*)$, where $(\mathcal{K}^*, \mathcal{C}^*, \mathcal{O}^*)$ is the tuple of true $\texttt{ATKR}^*$ parameters.

We address this challenge with SHERLOC, a novel framework in which DFDR infers ATKR parameters given the observed attack via optimization. SHERLOC aims to choose a hypothesis $\hat{\texttt{ATKR}}$ that *behaves consistently with the true* $\texttt{ATKR}^*$ *based on the observed attacks*. We evaluate our SHERLOC-inferred attacker $\hat{\texttt{ATKR}}$ with three distinct measures:

1. **Attack reconstruction:** Does the inferred attacker $\hat{\texttt{ATKR}}$ explain the behavior that DFDR observed and does it produce attack perturbations similar to the true attacker on unseen points?
2. **Attacker recovery:** Among attackers that explain the observations, how much can DFDR actually learn about the ground-truth attacker $\texttt{ATKR}^*$?
3. **Downstream defense:** Can a learned attacker hypothesis $\hat{\texttt{ATKR}}$ that is consistent with the observed attacks be useful for downstream defense?

### 3.2 SHERLOC: Inference Under Non-Identifiability

The non-identifiability result does not imply that observed attacks contain no information about their source; rather, they are insufficient for choosing a single unique attacker from the equivalence class. SHERLOC formalizes this distinction by treating attacker recovery as an inference problem within that equivalence class. Given a set of $k$ observed attacks $\boldsymbol{A}_{obs}$ and their corresponding benign inputs $\boldsymbol{X}_k$, DFDR aims to infer the attacker parameters $(\mathcal{K}, \mathcal{C}, \mathcal{O})$ that explain the observed behavior:

$$\hat{\mathcal{K}}, \hat{\mathcal{C}}, \hat{\mathcal{O}} = \arg\min_{\mathcal{K},\mathcal{C},\mathcal{O}} \sum_{i=1}^{k} \left\| \boldsymbol{\alpha}_{obs}^{(i)} - \boldsymbol{\alpha}_{opt}(\boldsymbol{x}_i; \mathcal{K}, \mathcal{C}, \mathcal{O}) \right\|_2 \tag{5}$$

For each $\boldsymbol{\alpha}_{obs}^{(i)}$, we compute its distance from the optimal attack $\boldsymbol{\alpha}_{opt}(\boldsymbol{x}_i; \mathcal{K}, \mathcal{C}, \mathcal{O})$, which depends on both the corresponding benign input $\boldsymbol{x}_i$ and the current attacker parameters $(\mathcal{K}, \mathcal{C}, \mathcal{O})$. Consequently, each optimization step requires solving $k$ inner problems.

When additional information about the attacker is available, SHERLOC can easily incorporate it to further constrain the set of plausible attacker hypotheses. For example, prior knowledge about $\mathcal{K}$ can be encoded through regularization toward models considered plausible by DFDR, while structural knowledge about $\mathcal{C}$ or $\mathcal{O}$ can restrict their admissible parameterizations. In Section 4, we consider examples of such additional structure, as well as evaluate how they affect attacker recovery.

## 4 Experiments

Our proposed framework is the foundation of a general tool for reverse engineering attacker characteristics. Our theoretical analysis establishes that a series of observed attacks cannot, on its own, uniquely identify the attacker $\texttt{ATKR}^*$ that generated it. In what follows, we use SHERLOC to empirically investigate *when* observed attack behavior contains sufficient information to recover an attacker's knowledge, capability, or objective, and when fundamental ambiguities remain.

Our evaluation is organized around three key questions. **First**, can SHERLOC find attacker parameters that explain the observed attacks, even when those parameters are not uniquely identifiable? **Second**, when does such reconstruction translate into recovery of the ground-truth attacker parameters $(\mathcal{K}^*, \mathcal{C}^*, \mathcal{O}^*)$? **Third**, when exact recovery is difficult, can the inferred attacker nevertheless provide useful information for downstream defenses?

To answer these questions, we evaluate SHERLOC across linear and polynomial learners, synthetic and real datasets, and inference of each attacker component $\mathcal{K}$, $\mathcal{C}$, and $\mathcal{O}$. Across these settings,

we evaluate two versions of SHERLOC: the generic SHERLOC and SHERLOC+, in which we incorporated additional information as discussed in Section 4.1. The resulting SHERLOC-inferred attackers are evaluated under 3 lenses: reconstruction, recovery, and defense. Full details of the experimental setup, including data generation, attack strengths, optimization configurations, and computational resources are provided in Appendix C.

### 4.1 Incorporating Prior Knowledge

SHERLOC does not prescribe a particular prior, but can incorporate application-specific information to constrain plausible attacker hypotheses. To demonstrate this flexibility, we additionally evaluate SHERLOC+, which uses an iid-training assumption for $K$ and known matrix spectra (but unknown orientations) for $C$ and $O$. We provide more details on these assumptions in Appendix D and compare SHERLOC against SHERLOC+ below.

### 4.2 Attack Reconstruction and Attacker Recoverability

Motivated by the distinction established in Section 2, we separately measure *attack reconstruction* and *attacker recovery*. Reconstruction measures whether the inferred attacker reproduces the attacks observed by `DFDR`. We report the Normalized Attack Reconstruction Score (`NARS`), where 1 indicates perfect reconstruction and 0 indicates no improvement over the prior attacker. We consider an inferred attacker to adequately reconstruct the observations when $\texttt{NARS} \geq 0.95$.

We separately measure recovery of the underlying attacker parameters using the Normalized Knowledge, Capability, and Objective Recovery Scores (`NKRS`, `NCRS`, and `NORS`, respectively). A recovery score of 1 indicates performance equivalent to knowing the ground-truth parameter, 0 indicates no improvement over the prior, and a negative value indicates an estimate farther from the ground truth than the prior. A detailed definition of these evaluation metrics is provided in Appendix E.

**Linear learners.** Linear models (linear regression, logistic regression) provide the controlled setting predicted by our theoretical analysis. For these learners, there is often a fixed optimal attack (Alfeld et al. (2017)). Consequently, SHERLOC may be able to determine how an attacker behaves without having sufficient information to determine why it behaves that way.

This phenomenon appears clearly in our results. Across linear settings, SHERLOC routinely finds attacker hypotheses that reproduce the observed attacks nearly perfectly ($\texttt{NARS} \approx 1$). For example, on the `uci_energy` and `breast_cancer` datasets, SHERLOC achieves perfect $\texttt{NARS} = 1$, yet the corresponding recovery scores of the parameters themselves are strongly negative (see Table 1).

Thus, successful reconstruction of the observed attack does not imply identification of the true underlying attacker. The same limitation worsens quickly with increasing dimension. On synthetic linear problems, capability recovery with the best-case spectrum information decreases from $1.00$ at $d = 2$, to $0.17$ at $d = 8$, to $0.07$ at $d = 16$. Objective recovery likewise falls from $1.00$ at $d = 2$ to 0.1–0.30 for $d = 4$–16. These results empirically confirm the attacker equivalence class discussed in Section 3, and additionally suggest that the order of that equivalence class increases as dimensionality increases.

Table 1: **Attack reconstruction and parameter recovery on synthetic data.** We report `NARS`, followed by `NKRS`, `NCRS`, and `NORS` for linear defenders and degree-2 polynomial defenders. ‡ denotes SHERLOC+. Note that blank cells reflect the experimental design — the linear dimensionality sweep was evaluated through $d = 16$, while polynomial were evaluated only at selected dimensions.

| | `NARS` | | $\mathcal{C}$ (`NCRS`) | | $\mathcal{O}$ (`NORS`) | | $\mathcal{K}$ (`NKRS`) | |
|---|---|---|---|---|---|---|---|---|
| $d$ | Linear | Polynomial | Linear | Polynomial | Linear | Polynomial | Linear | Polynomial |
| 2 | $\geq 0.95$ | $\geq 0.95$ | 1.00‡ | 1.00 | 1.00 | 1.00 | 0.29 | 0.02–0.34 |
| 4 | $\geq 0.95$ | $\geq 0.95$ | 0.47‡ | 1.00 | 0.15 | 1.00 | 0.20 | — |
| 8 | $\geq 0.95$ | — | 0.17‡ | — | 0.20 | — | 0.02 | — |
| 12 | $\geq 0.95$ | $\geq 0.95$ | 0.14‡ | 1.00 | 0.25 | 1.00 | 0.13 | 0.46–0.50 |
| 16 | $\geq 0.95$ | $\geq 0.95$ | 0.07‡ | 0.99 | 0.30 | 1.00 | 0.40 | 0.64–0.72 |

Table 2: **Attacker recovery as the number of observed attacks increases.** We report NKRS, NCRS, and NORS achieved by SHERLOC for degree-2 polynomial defenders using increasingly many observed attacks. Capability and knowledge results use the $d = 12$ `make_classification` setting; objective results summarize the three evaluated settings.

| Target | $n = 100$ | $n = 1000$ | $n = 5000$ |
|---|---|---|---|
| $\mathcal{C}$ | 0.996 | 0.999 | 0.948 |
| $\mathcal{O}$ | 1.000 | 1.000 | 1.000 |
| $\mathcal{K}$ | 0.46–0.48 | 0.49–0.50 | 0.49–0.50 |

**Polynomial learners.** The linear results motivate a natural question: does recovery improve when multiple observed attacks provide additional information about the same attacker? To answer this question, we must look to non-linear models, out of which we look at the polynomial classifier. Although these models remain linear in their learned parameters, their gradients with respect to the original input vary across points. Multiple attack observations can therefore impose different constraints on the latent attacker.

The contrast is strongest for $\mathcal{O}$ recovery. As shown in Table 1, NORS $\approx 1.0$ at every tested synthetic polynomial dimension. $\mathcal{C}$ also becomes highly recoverable: NCRS $\approx 1.0$ through $d = 16$, including $0.99$ on held-out trials at $d = 16$.

However, SHERLOC uniformly struggles at recovering $\mathcal{K}$, even in the non-linear setting. As shown in Table 1, NKRS is approximately $0.46$–$0.50$ on synthetic data with $d = 12$ dimensions and $0.64$–$0.72$ at $d = 16$. On `pendigits`, NKRS is $0.38$. Increasing the number of attacked points from $100$ to $1,000$ to $5,000$ shows that knowledge recovery plateaus, improving only slightly as shown in Table 2. Thus, observing more attacks can reduce ambiguity for SHERLOC learning $\mathcal{K}$ when applied to non-linear learners, but neither removes it completely nor uniformly across parameters.

### 4.3 Downstream Defense Applicability

The preceding experiments treat recovery partially as a forensic objective: how accurately can DFDR infer the latent parameters of ATKR? However, exact identification need not be necessary for every downstream task. In this section, we aim to answer whether an inferred attacker hypothesis that captures the observed attack behavior can still be useful for downstream defense, despite the latent attacker parameters being only partially recovered.

We compare adversarial regularization using (i) a baseline attacker model $\text{ATKR}_0$ representing the information available to DFDR without observing the attacks, (ii) the attacker $\hat{\text{ATKR}}$ inferred by SHERLOC, and (iii) the ground-truth attacker $\text{ATKR}^*$. The ground-truth condition represents the benefit available under perfect attacker information. Because this comparison is only meaningful when perfect attacker information improves the defense over the prior, we restrict interpretation to settings with positive *defense headroom*, defined as the difference between the ground-truth and baseline defense effectiveness.

Table 3: **Attack reconstruction and parameter recovery on real data.** We report NARS, followed by NKRS, NCRS, and NORS. Linear defenders are evaluated on `uci_energy` ($d = 8$), `breast_cancer` ($d = 30$), and `pendigits` ($d = 16$); degree-2 polynomial defenders are evaluated on `pendigits` and `breast_cancer`. Again, blank cells reflect the experimental design — a target is omitted when it was not evaluated in that learner-dataset configuration.

| | | $\mathcal{C}$ | | $\mathcal{O}$ | | $\mathcal{K}$ | |
|---|---|---|---|---|---|---|---|
| Learner | Dataset | NARS | NCRS | NARS | NORS | NARS | NKRS |
| Linear | `uci_energy` | 1.000 | $-5.0$–$-6.4$ | — | 0.60 | — | $-0.22$–$-0.70$ |
| Linear | `breast_cancer` | 1.000 | $-11.7$ | — | — | — | 0.13 |
| Linear | `pendigits` | — | $-1.9$–$-8.9$ | — | 0.00–$-16.7$ | — | 0.07–0.13 |
| Poly-2 | `pendigits` | — | 0.19–0.49 | — | 0.87–0.95 | — | 0.31 |
| Poly-2 | `breast_cancer` | 0.98 | $-4.4$ | — | — | — | — |

Table 4: **Downstream defense using SHERLOC-inferred attackers.** We report GAIN for adversarial regularization on real datasets in settings with measurable defensive headroom. Only settings with interpretable headroom are reported.

| Setting | Headroom | Defense gain |
|---|---|---|
| $\mathcal{K}$, linear uci_energy | $+0.14\ R^2$ | 0.99–1.00 |
| $\mathcal{K}$, linear pendigits | $+0.065\ p_{\text{true}}$ | 0.95–0.97 |
| $\mathcal{K}$, linear breast_cancer | $+0.04\ p_{\text{true}}$ | 0.89–0.99 |
| $\mathcal{K}$, poly-2 pendigits | $+0.03\ p_{\text{true}}$ | 1.00 |
| $\mathcal{O}$, poly-2 pendigits | $+0.02$–$0.03\ p_{\text{true}}$ | 0.72–1.05 |

We measure downstream utility of SHERLOC with *defense gain*:

$$\texttt{GAIN} = \frac{\texttt{DEF}(\hat{\texttt{ATKR}}) - \texttt{DEF}(\texttt{ATKR}_0)}{\texttt{DEF}(\texttt{ATKR}^*) - \texttt{DEF}(\texttt{ATKR}_0)} \tag{6}$$

When $\texttt{GAIN} = 1$ indicates that the inferred attacker $\hat{\texttt{ATKR}}$ fully achieves the benefit of defending with the knowledge of the ground-truth attacker $\texttt{ATKR}^*$, and $\texttt{GAIN} = 0$ indicates no improvement over the baseline.

Where measurable headroom exists, adversarial regularization with the inferred attacker $\hat{\texttt{ATKR}}$ captures most of the benefit obtained using the ground-truth attacker. Across $\mathcal{K}$ recovery trials with a linear DFDR, GAIN is $0.99 - 1.00$ on uci_energy, 0.95–0.97 on pendigits, and $0.89 - 0.99$ on breast_cancer (see Table 4). For the polynomial DFDR, GAIN is $1.00$ on pendigits. The linear pendigits result is especially illustrative. Although NKRS is approximately zero, the inferred attacker captures $95 - 97\%$ of the available defensive benefit. Thus, poor recovery does not imply poor downstream utility.

For objective inference on pendigits, defense gain ranges from $0.72$ to $1.05$, while a control that receives the same structural information but no observed attacks achieves only $0.15 - 0.20$. This comparison indicates that the observed attacks contribute information beyond the supplied structure. Capability-based defense cannot be interpreted reliably in our current experiments because the available headroom for defense is generally below $0.02$, meaning there is not adequate room for improving robustness.

Overall, these results expose a useful separation between forensic identification and defensive utility. Recovering the exact latent attacker parameters requires distinguishing among behaviorally equivalent hypothesis attackers, whereas the defense may only require the attack-relevant behavior that those hypotheses share with the ground-truth. Therefore, SHERLOC can provide useful defensive information even in settings where the attacker non-identifiability discussed in Section 3 prevents exact parameter recovery.

## 5 Related Work

Adversarial learning is a well-established research area. We direct the reader to Biggio & Roli (2018), Vorobeychik & Kantarcioglu (2018), and other surveys (Chakraborty et al., 2018; Machado et al., 2021) for an overview of the field. In this paper, we focus on the deployment-time attacker, which is an especially well-studied class of adversarial attacks. The deployment-time attacker is particularly relevant in the field of computer vision, where imperceptible input perturbations have been shown to reliably induce misclassification in high-capacity models (Szegedy et al., 2014; Moosavi-Dezfooli et al., 2018). In contrast to much of this literature, our contribution is not the proposal of a new attack or defense mechanism; rather, we develop a framework that enables the defender to infer key properties of the attacker from observed attacks.

Reverse Engineering of Deception (RED) is an emerging area of adversarial learning that aims to extract insights into an adversary by examining observed attacks. Prior work in this area has shown that adversarial perturbations can exhibit structured, attacker-dependent patterns (characteristic geometric or statistical properties of the perturbation) that can be exploited to identify or classify the underlying attack mechanism (Guo et al., 2023; Souri et al., 2024). These approaches demonstrate

that observed attacks encode informative signals about the attacker. However, they are typically tailored to specific attack families ($\ell_p$-constrained) and rely on strong assumptions about the attacker's optimization procedure (Thaker et al., 2022).

Many RED approaches are additionally limited to particular input domains, such as images. For instance, prior work has shown that it is possible to reverse engineer the exact perturbations made in order to produce an adversarial image (Gong et al., 2022). It is also possible to attribute an adversarially attacked image to a particular attack toolchain or family (Goebel et al., 2021). In contrast, our framework is domain-agnostic and does not assume, but *learns* the attacker's perturbation structure. This abstraction allows our approach to generalize across attack types and learning models.

# 6 Conclusions

We introduced SHERLOC, a novel framework for reasoning backward from observed attacks to an attacker that could have produced them. SHERLOC characterizes an attacker through their knowledge ($\mathcal{K}$), capability ($\mathcal{C}$), and objective ($\mathcal{O}$), then answers the question of what an adversary reveals through their actions.

SHERLOC does not recover the exact ground-truth attacker — however, it is able to successfully and reliably reconstruct a hypothesis attacker that behaves consistently to the observed attacks. For linear learners, SHERLOC can reconstruct attacks nearly perfectly while recovering very little about the underlying attacker. When attacks provide richer, input-dependent behavioral information, as with polynomial learners, capability and objective become substantially more recoverable, while knowledge remains comparatively difficult to identify. These results empirically demonstrate the effect of attacker non-identifiability — a non-trivial equivalence class of distinct attackers that produce the same attacks prevent exact recovery.

However, exact identification is not necessary for downstream utility. Even when SHERLOC recovers an attacker that differs substantially from the ground-truth, the SHERLOC-inferred attacker proves useful in downstream defense. In settings with measurable defensive headroom, adversarial regularization constructed using SHERLOC-inferred attackers captured most of the robustness improvement attainable using the ground-truth attacker. Thus, a defender may not need to know exactly *who* the attacker is to learn enough about *how* they attack to respond effectively.

SHERLOC provides both a practical tool and a new outlook on Reverse Engineering of Deception (RED). It reframes adversarial learning from only asking how to withstand an attack to also asking what can be learned from the attack that occurred. The resulting fundamental question of *what information is leaked by an attacker when they manipulate data* opens a broader direction for understanding and responding to dynamic adversaries.

## 6.1 Future Work

SHERLOC opens a broad space of questions for the field of Reverse Engineering of Deception (RED). Rather than prescribing a single attacker model, prior, or inference setting, the framework provides a common language for asking what an adversary reveals through their behavior, what remains fundamentally ambiguous, and what additional information can resolve that ambiguity. The settings studied here therefore represent only a small portion of the possible design space.

One particularly important direction is the development of effective *priors* for attacker inference. Our experiments demonstrate that additional defender knowledge can substantially alter recoverability, but the appropriate prior will necessarily depend on the application and domain. Future work can investigate other forms of prior knowledge. More broadly, understanding the interaction between information revealed by an attack and information supplied by the defender is itself a fundamental problem for RED. The framework can also be extended well beyond the deployment-time attacks considered in this paper. In particular, training-time attacks provide a natural next setting.

These directions are illustrative rather than exhaustive. Different learners, attacker parameterizations, and downstream uses each induce new reverse-engineering questions. We view SHERLOC as a starting point for systematically asking these questions, and hope that it enables a broader community effort around RED towards not only developing new defenses, but understanding what adversaries reveal about themselves by acting in the first place.

## AI USE STATEMENT

In this work, we used generative AI tools for the implementation of methods and providing feedback on a portion of our experiments. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, proving mathematical claims, writing proofs, proposing or refining hypotheses, designing research methodology or experiments, qualitative and thematic data analysis, or interpret results, and neither assisting with translation nor cleaning/reformatting datasets are applicable to this work. Additionally, we used generative AI tools for creating some scientific figures, creating/editing software code, and editing parts of the paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. AI-edited writing was read over by 3 authors, and AI-generated figures were verified by 1 author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Scott Alfeld, Xiaojin Zhu, and Paul Barford. Explicit defense actions against test-set attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. doi: 10.1609/aaai.v31i1.10767.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/athalye18a.html.

Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, ASIACCS '06, pp. 16–25, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932720. doi: 10.1145/1128817.1128824. URL https://doi.org/10.1145/1128817.1128824.

Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. The security of machine learning. *Mach. Learn.*, 81(2):121–148, November 2010. ISSN 0885-6125. doi: 10.1007/s10994-010-5188-5. URL https://doi.org/10.1007/s10994-010-5188-5.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, December 2018. ISSN 0031-3203. doi: 10.1016/j.patcog.2018.07.023. URL http://dx.doi.org/10.1016/j.patcog.2018.07.023.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey, 2018. URL https://arxiv.org/abs/1810.00069.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019.

Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, Jeremy Dawson, and Nasser M. Nasrabadi. Smoothfool: An efficient framework for computing smooth adversarial perturbations, 2019. URL https://arxiv.org/abs/1910.03624.

Michael Goebel, Jason Bunk, Srinjoy Chattopadhyay, Lakshmanan Nataraj, Shivkumar Chandrasekaran, and B. S. Manjunath. Attribution of gradient based adversarial attacks for reverse engineering of deceptions. *Electronic Imaging*, 33(4):300–1–300–1, 2021. doi: 10.2352/ISSN.2470-1173.2021.4.MWSF-300. URL https://library.imaging.org/ei/articles/33/4/art00011.

Yifan Gong, Yuguang Yao, Yize Li, Yimeng Zhang, Xiaoming Liu, Xue Lin, and Sijia Liu. Reverse engineering of imperceptible adversarial image perturbations. 2022. URL https://arxiv.org/abs/2203.14145.

Zhongyi Guo, Keji Han, Yao Ge, Yun Li, and Wei Ji. Attribution adversarial attacks multi-task learning. In *Neural Information Processing: 30th International Conference, ICONIP 2023, Changsha, China, November 20–23, 2023, Proceedings, Part II*, pp. 81–94, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-981-99-8081-9. doi: 10.1007/978-981-99-8082-6_7. URL https://doi.org/10.1007/978-981-99-8082-6_7.

Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, AISec '11, pp. 43–58, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450310031. doi: 10.1145/2046684.2046692. URL https://doi.org/10.1145/2046684.2046692.

Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. Adversarial machine learning in image classification: A survey toward the defender's perspective. *ACM Comput. Surv.*, 55(1), November 2021. ISSN 0360-0300. doi: 10.1145/3485133. URL https://doi.org/10.1145/3485133.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, abs/1511.04599, 2015.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa, 2018. URL https://arxiv.org/abs/1811.09716.

Hossein Souri, Pirazh Khorramshahi, Chun Pong Lau, Micah Goldblum, and Rama Chellappa. Identifying attack-specific signatures in adversarial examples. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7050–7054, 2024. doi: 10.1109/ICASSP48485.2024.10446989.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *Proceedings of the International Conference on Learning Representations*, 2014.

Darshan Thaker, Paris Giampouras, and Rene Vidal. Reverse engineering $\ell_p$ attacks: A block-sparse optimization approach with recovery guarantees. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21253–21271. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/thaker22a.html.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *A sdvances in Neural Information Processing Systems*, volume 33, pp. 1633–1645. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf.

Sun Tzu. *The Art of War*. Circa 500 B.C.E.

Yevgeniy Vorobeychik and Murat Kantarcioglu. Adversarial machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:52100908.

## A Parameterization of ATKR for Linear Models

Suppose DFDR is a generalized least-squares linear model (i.e., linear regression, logistic regression) with the prediction function $f : \mathbb{R}^d \to \mathbb{R}^q$:

$$f(\boldsymbol{x}) = \boldsymbol{M}\boldsymbol{x} \tag{7}$$

where $\boldsymbol{M} \in \mathbb{R}^{q \times d}$ is the weight matrix of the model.

Suppose ATKR performs an *evasion attack under Mahalanobis constraints*. ATKR aims to solve the following optimization problem:

$$\boldsymbol{\alpha}_{opt}(\boldsymbol{x}; \texttt{ATKR}) = \arg\max_{\boldsymbol{\alpha}} \|\boldsymbol{M}'(\boldsymbol{x}+\boldsymbol{\alpha}) - \boldsymbol{M}'\boldsymbol{x}\|_{\boldsymbol{W}}^2 \tag{8}$$

$$= \arg\max_{\boldsymbol{\alpha}} \|\boldsymbol{M}'\boldsymbol{\alpha}\|_{\boldsymbol{W}}^2 \tag{9}$$

$$\text{s.t. } ||\boldsymbol{\alpha}||_{\boldsymbol{C}} \leq c \tag{10}$$

where

- $\boldsymbol{M}' \in \mathbb{R}^{q\times d}$ is ATKR's estimate of DFDR's model $\boldsymbol{M}$.
- $\boldsymbol{W} \in \mathbb{R}^{q\times q}$ is a positive definite matrix defining the Mahalanobis loss of ATKR.
- $\boldsymbol{C} \in \mathbb{R}^{d\times d}$ is a positive definite matrix defining (with positive scalar $c$) the Mahalanobis constraint of ATKR.

In this case, we parameterize ATKR as follows:

$$\mathcal{K} = \{\boldsymbol{M}'\}, \mathcal{C} = \{\boldsymbol{C}, c\}, \mathcal{O} = \{\boldsymbol{W}\} \tag{11}$$

Let $d$ denote the input dimension and $q$ denote the output dimension of the model. Hence, we have the following ATKR parameters:

$$\boldsymbol{M} \in \mathbb{R}^{q\times d} \tag{12}$$

$$\boldsymbol{C} \in \mathbb{R}^{d\times d} \tag{13}$$

$$c \in \mathbb{R} \tag{14}$$

$$\boldsymbol{W} \in \mathbb{R}^{q\times q} \tag{15}$$

## B Proof of Theorem 3.2

**Theorem 3.2. General non-identifiablity of attacker.**

In the case of a linear DFDR, the following hold for any attack $\boldsymbol{\alpha}$:

- $\forall \mathcal{K}, \mathcal{C},\ \exists \mathcal{O}$ s.t. ATKR $\in \mathcal{S}_{\boldsymbol{\alpha}}$
- $\forall \mathcal{K}, \mathcal{O},\ \exists \mathcal{C}$ s.t. ATKR $\in \mathcal{S}_{\boldsymbol{\alpha}}$
- $\forall \mathcal{C}, \mathcal{O},\ \exists \mathcal{K}$ s.t. ATKR $\in \mathcal{S}_{\boldsymbol{\alpha}}$

**Proof.**

We will prove each of the three statements below.

**1.** $\forall \mathcal{K}, \mathcal{C},\ \exists \mathcal{O}$ s.t. ATKR $\in \mathcal{S}_{\boldsymbol{\alpha}}$

This statement can be proved for any invertible model matrix $\boldsymbol{M}$.

We can construct $\boldsymbol{V}$ such that the largest singular value of $\boldsymbol{V}\boldsymbol{M}\boldsymbol{G}^{-1}$ corresponds to the vector $\frac{1}{c}\boldsymbol{G}\boldsymbol{\alpha}$.

First, construct a placeholder matrix $\boldsymbol{V}$ (utilizing singular value decomposition):

$$\boldsymbol{V} = \hat{\mathbf{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^\top \tag{16}$$

$$= \begin{bmatrix} \frac{1}{c}\boldsymbol{G}\boldsymbol{\alpha}^* & \dots \end{bmatrix} \hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^\top \tag{17}$$

$$\tag{18}$$

Then, adjust $\boldsymbol{V}$ to account for $\boldsymbol{M}$ and $\boldsymbol{G}^{-1}$.

$$\boldsymbol{V}' = \boldsymbol{V}\boldsymbol{G}\boldsymbol{M}^{-1} \tag{19}$$

Then, we have:

$$\|\boldsymbol{V}'\boldsymbol{M}\boldsymbol{G}^{-1}\| = \|\boldsymbol{V}\boldsymbol{G}\boldsymbol{M}^{-1}\boldsymbol{M}\boldsymbol{G}^{-1}\| \tag{20}$$

$$= \|\boldsymbol{V}\boldsymbol{G}\boldsymbol{G}^{-1}\| \tag{21}$$

$$= \|\boldsymbol{V}\| \tag{22}$$

Thus, the largest singular vector $s_1$ of $\boldsymbol{V}'\boldsymbol{M}\boldsymbol{G}^{-1}$ is the largest singular vector of $\boldsymbol{V}$, which is clearly $\frac{1}{c}\boldsymbol{G}\boldsymbol{\alpha}$. Hence, the optimal attack is $\boldsymbol{\alpha}$.

Let $\boldsymbol{W} = \boldsymbol{V}'^{\top}\boldsymbol{V}'$. Then, we have picked the objective $\mathcal{O} = \{\boldsymbol{W}\}$ such that the attacker constraints and model belief result in the attack $\boldsymbol{\alpha}$. Hence, $\forall \mathcal{K}, \mathcal{C},\ \exists \mathcal{O}$ s.t. $\texttt{ATKR} \in \mathcal{S}_{\boldsymbol{\alpha}}$. □

**2.** $\forall \mathcal{K}, \mathcal{O},\ \exists \mathcal{C}$ s.t. $\texttt{ATKR} \in \mathcal{S}_{\boldsymbol{\alpha}}$

Let $\boldsymbol{M}$ be the model that the attacker assumed of the defender.

Let $\boldsymbol{W}$ = some positive definite matrix. This defines the goal of the attacker, i.e., $\mathcal{O}$.

We know that $||\boldsymbol{\alpha}||_{\boldsymbol{C}} \leq c$. That is, the attack must be within the capabilities of the attacker.

Consider the goal of the attacker — to maximize the defender loss on the adversarial example. Assume that the attacker's attack was approximately optimal. Then,

$$\begin{aligned} &\|\boldsymbol{M}\boldsymbol{\alpha}\|_{\boldsymbol{W}}^2 \\ &= \boldsymbol{\alpha}^{\top}\boldsymbol{M}^{\top}\boldsymbol{W}\boldsymbol{M}\boldsymbol{\alpha} \quad \text{was maximized.} \end{aligned} \tag{23}$$

Let $\boldsymbol{C} = \boldsymbol{M}^{\top}\boldsymbol{W}\boldsymbol{M}$. Then,

$$\|\boldsymbol{\alpha}\|_{\boldsymbol{C}} = \boldsymbol{\alpha}^{\top}\boldsymbol{C}\boldsymbol{\alpha} \tag{24}$$

$$= \boldsymbol{\alpha}^{\top}\boldsymbol{M}^{\top}\boldsymbol{W}\boldsymbol{M}\boldsymbol{\alpha} \tag{25}$$

$$= (\boldsymbol{M}\boldsymbol{\alpha})^{\top}\boldsymbol{W}(\boldsymbol{M}\boldsymbol{\alpha}) \tag{26}$$

$$= \|\boldsymbol{M}\boldsymbol{\alpha}\|_{\boldsymbol{W}}^2 \tag{27}$$

Let $c = \|\boldsymbol{M}\boldsymbol{\alpha}\|_{\boldsymbol{W}}^2$.

Then, the attacker is constrained by the exact value it is maximizing. The upper bound is also the maximum value of the objective. Hence, we have found a $\mathcal{C}$ such that for any $\mathcal{K},\ \mathcal{O}$, the attacker would learn exactly $\boldsymbol{\alpha}$, i.e., $\texttt{ATKR} \in \mathcal{S}_{\boldsymbol{\alpha}}$. □

**3.** $\forall \mathcal{C}, \mathcal{O},\ \exists \mathcal{K}$ s.t. $\texttt{ATKR} \in \mathcal{S}_{\boldsymbol{\alpha}}$

The proof follows similarly to the proof of statement 1. We can construct $\boldsymbol{M}$ such that the largest singular value of $\boldsymbol{V}\boldsymbol{M}\boldsymbol{G}^{-1}$ corresponds to the vector $\frac{1}{c}\boldsymbol{G}\boldsymbol{\alpha}^*$.

First, construct a placeholder model $\boldsymbol{M}$ (utilizing singular value decomposition):

$$\boldsymbol{M} = \hat{\mathbf{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\top} \tag{28}$$

$$= \begin{bmatrix} \frac{1}{c}\boldsymbol{G}\boldsymbol{\alpha}^* & \dots \end{bmatrix} \hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\top} \tag{29}$$

$$\tag{30}$$

Then, adjust $\boldsymbol{M}$ to account for $\boldsymbol{V}$ and $\boldsymbol{G}^{-1}$.

$$\boldsymbol{M}' = \boldsymbol{V}^{-1}\boldsymbol{M}\boldsymbol{G} \tag{31}$$

Then, we have:

$$\|\boldsymbol{V}\boldsymbol{M}'\boldsymbol{G}^{-1}\| = \|\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{M}\boldsymbol{G}\boldsymbol{G}^{-1}\| \tag{32}$$

$$= \|\boldsymbol{M}\| \tag{33}$$

Thus, the largest singular vector $s_1$ of $\boldsymbol{V}\boldsymbol{M}\boldsymbol{G}^{-1}$ is the largest singular vector of $\boldsymbol{M}$, which is clearly $\frac{1}{c}\boldsymbol{G}\boldsymbol{\alpha}$. Hence, the optimal attack is $\boldsymbol{\alpha}$.

Then, we have picked the attacker's assumed defender model $\mathcal{K} = \{\boldsymbol{M}\}$ such that the attacker constraints and objective result in the attack $\boldsymbol{\alpha}$. Hence, $\forall \mathcal{C}, \mathcal{O},\ \exists \mathcal{K}$ s.t. $\texttt{ATKR} \in \mathcal{S}_{\boldsymbol{\alpha}}$. □

## C Experimental Setup

SHERLOC is evaluated across both linear and polynomial learners. Linear experiments use logistic and ridge regression. Polynomial experiments use logistic regression over degree-2 or degree-3 polynomial features. Both model families are implemented via PyTorch.

This evaluation is performed across several datasets. First, synthetic experiments are used to systematically vary input dimensionality and the amount of observed attack behavior. Linear experiments cover dimensions $d \in 2, 4, 8, 12, 16$ for the reported results in Section 4. Polynomial experiments extend selected configurations through $d = 32$, with reported degree-2 experiments at $d \in 2, 4, 12, 16, 24, 32$ depending on the inference target.

To test whether recovery is limited by the quantity of observed behavior, we additionally varied the number of attacked points for degree-2 polynomial learners. We evaluate $k \in 100, 1000, 5000$ observed attacks. For $\mathcal{K}$ and $\mathcal{C}$, the reported comparison uses the $d = 12$ `make_classification` setting; the $\mathcal{O}$ comparison aggregates the three evaluated settings.

In addition to this synthetic setting, we evaluate SHERLOC on three real datasets: `uci_energy` ($d = 8$, regression), `breast_cancer` ($d = 30$), and `pendigits` ($d = 16$). Linear learners are evaluated on all three datasets; degree-2 polynomial learners are evaluated on `pendigits` and `breast_cancer`. Real-data results are reported as medians over trials.

Detective hyperparameters (optimizer settings/configuration) are selected using synthetic experiments and subsequently applied to the real datasets without dataset-specific tuning. For each dataset, we evaluate attack strengths $\epsilon$ for which the attack measurably affects the learner. Settings in which the attack is too weak i.e., does not change model performance are excluded from the reported results.

Our experiments include both generic SHERLOC and informed SHERLOC+ detectives. The generic configuration uses six optimization restarts. For $\mathcal{K}$, we additionally evaluate an iid-training assumption formalized in Appendix D: the attacker's proxy model is trained on separate data drawn from the same underlying distribution as the defender's data.

For $\mathcal{C}$ and $\mathcal{O}$, we additionally evaluate spectrum-informed detectives. These detectives are supplied with the eigenvalues of the corresponding ground-truth matrix but not its eigenvectors. We also use a no-data control for the spectrum-informed setting. This detective receives the spectrum but no observed attacks, allowing us to distinguish information supplied directly by the spectrum from information learned from attack behavior.

To test whether incomplete recovery is caused simply by insufficient optimization, we perform a heavier-optimization diagnostic for $\mathcal{K}$, increasing the optimization budget to three restarts of 1000 iterations. We compare this configuration against the standard detective rather than treating additional computation as a distinct inference assumption. Together with the experiment varying the number of observed attacks, this diagnostic separates limitations attributable to search effort or sample quantity from ambiguity in the information exposed by the attacks.

Unless otherwise noted, reported real-data values are medians over all trials. Synthetic summary tables report the **best** detective satisfying the reconstruction criterion for each setting, summarized

by the median across the corresponding settings. Linear synthetic values are held-out. Polynomial high-dimensional cells are based on fewer trials, with held-out confirmation for the reported (d=16) and (d=24) synthetic experiments. These differences in replication are noted when interpreting high-dimensional polynomial results.

## D INCORPORATING PRIOR KNOWLEDGE

To demonstrate SHERLOC's ability to incorporate additional information in inferring $\hat{\texttt{ATKR}}$, we evaluate SHERLOC+ in addition to generic SHERLOC.

To learn $\mathcal{K}$, SHERLOC+ encodes the assumption that ATKR constructs its proxy model using data drawn from the same underlying distribution as DFDR. DFDR knows only its own model $\boldsymbol{M}$ and the dataset $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{Y})$ it was trained on. SHERLOC+ assumes ATKR performs a traditional transfer attack using a surrogate model $\boldsymbol{M}'$ trained on some dataset $\boldsymbol{D}' = (\boldsymbol{X}', \boldsymbol{Y}')$ drawn i.i.d. from the same underlying data distribution $\mathcal{P}_{data}$.

Rather than assuming that DFDR knows ATKR's model parameters, this induces a prior over plausible knowledge models: models that could reasonably have resulted from training on another sample from the same distribution are preferred over arbitrary models. Since $\boldsymbol{M}'$ depends only on the training dataset $\boldsymbol{D}'$, we have

$$P(\boldsymbol{M}' \mid \mathcal{P}_{data}) = \int P(\boldsymbol{M}' \mid \boldsymbol{D}') P(\boldsymbol{D}' \mid \mathcal{P}_{data})\, d\boldsymbol{D}'. \tag{34}$$

Suppose both DFDR and ATKR use a generalized least-squares linear regressor. We assume the data-generating process is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{M}^{*\top} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon) \tag{35}$$

where $\boldsymbol{M}^* \in \mathbb{R}^{q \times d}$ is the unknown true weight matrix.

The generalized least-squares estimator is linear in the observed responses and can therefore be written as:

$$\boldsymbol{M}' = \boldsymbol{Y}'^{\top} A(\boldsymbol{X}'), \tag{36}$$

$$A(\boldsymbol{X}') = \Omega^{-1}\boldsymbol{X}'(\boldsymbol{X}'^{\top}\Omega^{-1}\boldsymbol{X}')^{-1} \tag{37}$$

and $\Omega \in \mathbb{R}^{\times}$ is the covariance matrix of the observation errors. Substituting the data-generating model,

$$\boldsymbol{M}' = (\boldsymbol{X}'\boldsymbol{M}^{*\top} + \varepsilon)^{\top} A(\boldsymbol{X}') \tag{38}$$

$$= \boldsymbol{M}^*(\boldsymbol{X}'^{\top} A(\boldsymbol{X}')) + \varepsilon^{\top} A(\boldsymbol{X}') \tag{39}$$

Since $A(\boldsymbol{X}')$ is deterministic and $\varepsilon$ is Gaussian, $\boldsymbol{M}'$ is an affine transformation of a Gaussian random matrix and is therefore itself Gaussian. Therefore, we have:

$$P(\boldsymbol{M}'|\mathcal{P}_{data}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{40}$$

where the mean and covariance are determined by the true $\boldsymbol{M}^*, \Sigma_\varepsilon$, and the estimator$A(\boldsymbol{X}')$.

In practice, $\boldsymbol{M}^*, \Sigma_\varepsilon$, and the underlying data distribution $\mathcal{P}_{data}$ are unknown. Hence, DFDR approximates them using its own training data $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{Y})$. Specifically, we substitute DFDR's learned model $\boldsymbol{M}$ for $\boldsymbol{M}^*$, estimate the noise covariance $\hat{\Sigma}_\varepsilon$ from the residuals on $\boldsymbol{D}$, and compute the corresponding estimator matrix $A(\boldsymbol{X})$. This yields the following approximate prior:

$$P(\boldsymbol{M}') = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{41}$$

We use this distribution as the prior over ATKR's knowledge parameter $\mathcal{K}$ to enable inference for non-linear models.

For nonlinear models such as polynomial regressors and classifiers, an analogous prior is generally intractable. Instead, we define an implicit prior. We introduce a *regularizer* that addresses

the attacker non-identifiability problem by favoring surrogate models whose predictions are more consistent with DFDR's model.

We estimate the disagreement between DFDR's model and an estimated attacker surrogate empirically via the mean squared output difference over $n$ samples drawn i.i.d. from the true data distribution $\mathcal{P}_{data}$:

$$\mathbf{DIS}_{\texttt{DFDR}}(\mathcal{K}) = \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - f_{\mathcal{K}'}(x_i)|^2 \tag{42}$$

where $f(\cdot)$ and $f_{\mathcal{K}}(\cdot)$ denote the outputs of DFDR's model and the attacker surrogate model, respectively.

The defender now solves the following optimization problem.

**Definition D.1.** Given the set of $k$ observed attacks $\boldsymbol{A}_{obs}$, the MLP DFDR that knows ATKR's true $\mathcal{C}^*, \mathcal{O}^*$ and aims to infer ATKR's $\mathcal{K}^*$ has the following objective function:

$$\hat{\mathcal{K}} = \arg\max_{\mathcal{K}} \sum_{i=1}^{k} \left\| \boldsymbol{\alpha}_{obs}^{(i)} - \boldsymbol{\alpha}_{opt}^{(i)} \right\|_2 - \lambda \cdot \mathbf{DIS}_{\texttt{DFDR}}(\mathcal{K}) \tag{43}$$

$$\text{s.t. } \boldsymbol{\alpha}_{opt}^{(i)} = \arg\max_{\boldsymbol{\alpha} \in \mathcal{A}_{\mathcal{C}^*}} g_{\mathcal{O}^*}(f_{\mathcal{K}}, \boldsymbol{x}_i, \boldsymbol{\alpha}), \quad i = 1, \ldots, k \tag{44}$$

For $\mathcal{C}$, SHERLOC+ utilizes a *spectrum-informed* setting in which DFDR knows the eigenvalues of the attacker's capability matrix, but not its eigenvectors. Equivalently, DFDR knows the *shape and scale* of the attacker's capability ellipsoid, but not its orientation in feature space. Thus, DFDR may know how strongly the attacker's ability to manipulate inputs varies across directions without knowing *which* feature directions correspond to those capabilities. SHERLOC+ must infer this unknown orientation from the observed attacks.

For $\mathcal{O}$, SHERLOC+ analogously considers a spectrum-informed setting. That is, DFDR is given the objective spectrum, but not its orientation. Intuitively, DFDR knows the relative scale with which the attacker values different directions in the output space, but does not know which output directions receive those weights. SHERLOC+ uses the observed attacks to infer this missing orientation.

Together, these assumptions demonstrate information that can be supplied to SHERLOC to improve inference. The experiments in Section 4 separately evaluate SHERLOC and SHERLOC+, allowing us to measure how much attacker information is revealed by the attacks themselves and how much reconstruction and recovery respectively depend on additional defender knowledge.

## E Evaluation Metrics

We first evaluate whether the inferred attacker reproduces the observed attack behavior. The evaluation is done on both the seen data (attacked points) and unseen data (held-out set). This allows a separation of attackers that can reconstruct the exact observed attacks vs. the same behavior as the ground-truth attacker on unseen data.

First, we look at objective reduction, computed entirely on seen data (the attack set the detective fit to):

$$\text{reduction} = 1 - \frac{L(\hat{\theta};, X_{\text{attack}})}{L(\theta_0;, X_{\text{attack}})}$$

This is the literal quantity the outer loop minimizes, evaluated at the final SHERLOC-inferred attacker $\hat{\texttt{ATKR}}$.

Then, we compute Normalized Attack Reconstruction Score (NARS) and the component-wise scores (NKRS, NCRS, NORS) on held-out data.

The reconstruction score NARS, is normalized relative to a naive baseline: 1 represents perfect reconstruction of the observed attacks and 0 represents no improvement over the baseline. We use

$\texttt{NARS} \geq 0.95$ as the criterion for successful reconstruction. When comparing detective configurations for recovery, reported synthetic summaries select among configurations satisfying this reconstruction criterion.