Blind Attacks on Machine Learners

Alex Beatson
Department of Computer Science
Princeton University
abeatson@princeton.edu

Zhaoran Wang
Department of Operations Research
and Financial Engineering
Princeton University
zhaoran@princeton.edu

Han Liu
Department of Operations Research
and Financial Engineering
Princeton University
hanliu@princeton.edu

Abstract

The importance of studying the robustness of learners to malicious data is well established. While much work has been done establishing both robust estimators and effective data injection attacks when the attacker is omniscient, the ability of an attacker to provably harm learning while having access to little information is largely unstudied. We study the potential of a "blind attacker" to provably limit a learner's performance by data injection attack without observing the learner's training set or any parameter of the distribution from which it is drawn. We provide examples of simple yet effective attacks in two settings: firstly, where an "informed learner" knows the strategy chosen by the attacker, and secondly, where a "blind learner" knows only the proportion of malicious data and some family to which the malicious distribution chosen by the attacker belongs. For each attack, we analyze minimax rates of convergence and establish lower bounds on the learner's minimax risk, exhibiting limits on a learner's ability to learn under data injection attack even when the attacker is "blind".

1 Introduction

As machine learning becomes more widely adopted in security and in security-sensitive tasks, it is important to consider what happens when some aspect of the learning process or the training data is compromised [1–4]. Examples in network security are common and include tasks such as spam filtering [5, 6] and network intrusion detection [7, 8]; examples outside the realm of network security include statistical fraud detection [9] and link prediction using social network data or communications metadata for crime science and counterterrorism [10].

In a training set attack, an attacker either adds adversarial data points to the training set ("data injection") or preturbs some of the points in the dataset so as to influence the concept learned by the learner, often with the aim of maximizing the learner's risk. Training-set data injection attacks are one of the most practical means by which an attacker can influence learning, as in many settings an attacker which does not have insider access to the learner or its data collection or storage systems may still be able to carry out some activity which is monitored and the resulting data used in the learner's training set [2, 6]. In a network security setting, an attacker might inject data into the training set for an anomaly detection system so that malicious traffic is classified as normal, thus making a network vulnerable to attack, or so that normal traffic is classified as malicious, thus harming network operation.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

A growing body of research focuses on game-theoretic approaches to the security of machine learning, analyzing both the ability of attackers to harm learning and effective strategies for learners to defend against attacks. This work often makes strong assumptions about the knowledge of the attacker. In a single-round game it is usually assumed that the attacker knows the algorithm used by the learner (e.g. SVM or PCA) and has knowledge of the training set either by observing the training data or the data-generating distribution [2, 5, 11]. This allows the construction of an optimal attack to be treated as an optimization problem. However, this assumption is often unrealistic as it requires insider knowledge of the learner or for the attacker to solve the same estimation problem the learner faces to identify the data-generating distribution. In an iterated-game setting it is usually assumed the attacker can query the learner and is thus able to estimate the learner's current hypothesis in each round [12–14]. This assumption is reasonable in some settings, but in other scenarios the attacker may not receive immediate feedback from the learner, making the iterated-game setting inappropriate. We provide analysis which makes weaker assumptions than either of these bodies of work by taking a probabilistic approach in tackling the setting where a "blind attacker" has no knowledge of the training set, the learner's algorithm or the learner's hypothesis.

Another motivation is provided by the field of privacy. Much work in the field of statistical privacy concerns disclosure risk: the probability that an entry in a dataset might be identified given statistics of the dataset released. This has been formalized by "differential privacy", which provides bounds on the maximum disclosure risk [15]. However, differential privacy hinges on the benevolence of an organization to which you give your data: the privacy of individuals is preserved as long as organizations which collect and analyze data take necessary steps to enforce differential privacy. Many data are gathered without users' deliberate consent or even knowledge. Organizations are also not yet under legal obligation to use differentially-private procedures.

A user might wish to take action to preserve their own privacy without making any assumption of benevolence on the part of those that collect data arising from the user's actions. For example, they may wish to prevent an online service from accurately estimating their income, ethnicity, or medical history. The user may have to submit some quantity of genuine data in order to gain a result from the service which addresses a specific query, and may not even observe all the data the service collects. They may wish to enforce the privacy of their information by also submitting fabricated data to the service or carrying out uncharacteristic activity. This is a data injection training set attack, and studying such attacks thus reveals the ability of a user to prevent a statistician or learner from making inferences from the user's behavior.

In this paper we address the problem of a one-shot data injection attack carried out by a blind attacker who does not observe the training set, the true distribution of interest, or the learner's algorithm. We approach this problem from the perspective of minimax decision theory to provide an analysis of the rate of convergence of estimators on training sets subject to such attacks. We consider both an "informed learner" setting where the learner is aware of the exact distribution used by the attacker to inject malicious data, and a "blind learner" setting where the learner is unaware of the malicious distribution. In both settings we suggest attacks which aim to minimize an upper bound on the pairwise KL divergences between the distributions conditioned on particular hypotheses, and thus maximize a lower bound on the minimax risk of the learner. We provide lower bounds on the rate of convergence of any estimator under these attacks.

2 Setting and contributions

2.1 Setting

A learner attempts to learn some parameter θ of a distribution of interest F_θ with density f_θ and belonging to some family $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$, where Θ is a set of candidate hypotheses for the parameter. "Uncorrupted" data $X_1, ..., X_n \in \mathcal{X}$ are drawn i.i.d. from F_θ. The attacker chooses some malicious distribution G_ϕ with density g_ϕ and from a family $\mathcal{G} = \{G_\phi : \phi \in \Phi\}$, where Φ is a parameter set representing candidate attack strategies. "Malicious" data $X'_1, .., X'_n \in \mathcal{X}$ are drawn i.i.d from the malicious distribution. The observed dataset is made up of a fraction α of true examples and $1 - \alpha$ of malicious examples. The learner observes a dataset $Z_1, ..., Z_n \in \mathcal{Z}$, where

$$Z_i = \begin{cases} X_i & \text{with probability} \quad \alpha \\ X'_i & \text{with probability} \quad 1 - \alpha. \end{cases} \tag{1}$$

We denote the distribution of Z with P. P is clearly a mixture distribution with density:

$$p(z) = \alpha f_\theta(z) + (1-\alpha) g_\phi(z).$$

The distribution of Z conditional on X is:

$$p(z|x) = \alpha \mathbb{1}\{z = x\} + (1-\alpha) g_\phi(z).$$

We consider two distinct settings based on the knowledge of the attacker and of the learner. First we consider the scenario where the learner knows the malicious distribution, G_ϕ and the fraction of inserted examples ("informed learner"). Second we consider the scenario where the learner knows only the family $\mathcal{G}$ to which G_ϕ belongs and fraction of inserted examples ("blind learner"). Our work assumes that the attacker knows only the family of distributions $\mathcal{F}$ to which the true distribution belongs ("blind attacker"). As such, the attacker designs an attack so as to maximally lower bound the learner's minimax risk. We leave as future work a probabilistic treatment of the setting where the attacker knows the true F_θ but not the training set drawn from it ("informed attacker"). To our knowledge, our work is the first to consider the problem of learning in a setting where the training data is distributed according to a mixture of a distribution of interest and a malicious distribution chosen by an adversary *without knowledge of the distribution of interest*.

2.2 Related work

Our paper has very strong connections to several problems which have previously been studied in the minimax framework.

First is the extensive literature on robust statistics. Our framework is very similar to Huber's ϵ-contamination model, where the observed data follows the distribution:

$$(1-\epsilon)P_\theta + \epsilon Q.$$

Here ϵ controls the degree of corruption, Q is an arbitrary corruption distribution, and the learner attempts to estimate θ robust to the contamination. A general estimator which achieves the minimax optimal rate under Huber's ϵ-contamination model was recently proposed by Chen, Gao and Ren[16]. Our work differs from the robust estimation literature in that rather than designing optimal estimators for the learner, we provide concrete examples of attack strategies which harm the learning rate of any estimator, even those which are optimal under Huber's model. Unlike robust statistics, our attacker does not have complete information on the generating distribution, and must select an attack which is effective for any data-generating distribution drawn from some set. Our work has similar connections to the literature on minimax rates of convergence of estimators for mixture models [17] and minimax rates for mixed regression with multiple components [18], but differs in that we consider the problem of designing a corrupting distribution.

There are also connections to the work on PAC learning with contaminated data [19]. Here the key difference, beyond the fact that we focus on strategies for a blind attacker as discussed earlier, is that we use information-theoretic proof techniques rather than reductions to computational hardness. This means that our bounds restrict all learning algorithms, not just polynomial-time learning algorithms.

Our work has strong connections to the analysis of minimax lower bounds in local differential privacy. In [20] and [21], Duchi, Wainwright and Jordan establish lower bounds in the local differential privacy setting, where $P(Z_i|X_i = x)$, the likelihood of an observed data point Z_i given X_i takes any value x, is no more than some constant factor greater than $P(Z_i|X_i = x')$, the likelihood of Z_i given X_i takes any other value x'. Our work can be seen as an adaptation of those ideas to a new setting: we perform very similar analysis but in a data injection attack setting rather than local differential privacy setting. Our analysis for the blind attacker, informed learner setting and our examples in Section 5 for both settings draw heavily from [21].

In fact, the blind attack setting is by nature locally differentially private with the likelihood ratio upper bounded by $\max_z \frac{\alpha f_\theta(z) + (1-\alpha) g_\phi(z)}{(1-\alpha) g_\phi(z)}$, as in the blind attack setting only α of the data points are drawn from the distribution of interest F. This immediately suggests bounds on the minimax rates of convergence according to [20]. However, the rates we obtain by appropriate choice of G_ϕ by the attacker obtain lower bounds on the rate of convergence which are often much slower than the bounds due to differential privacy obtained by arbitrary choice of G_ϕ.

The rest of this work proceeds as follows. Section 3.1 formalizes our notation. Section 3.2 introduces our minimax framework and the standard techniques of lower bounding the minimax risk by reduction

from parameter estimation to testing. Sections 3.3 and 3.4 discuss the "blind attacker; informed learner" and "blind attacker; blind learner" settings in this minimax framework. Section 3.5 briefly proposes how this framework could be extended to consider an "informed attacker" which observes the true distribution of interest F_θ. Section 4 provides a summary of the main results. In Section 5 we give examples of estimating a mean under blind attack in both the informed and blind learner setting and performing linear regression in the informed learner setting. In Section 6 we conclude. Proof of the main results is presented in the appendix.

3 Background and problem formulation

3.1 Notation

We denote the "uncorrupted" data with the random variables $X_{1:n}$. F_i is the distribution and f_i the density of each X_i conditioning on $\theta = \theta_i \in \Theta$; F_θ and f_θ are the generic distribution and density parametrized by θ. We denote malicious data with the random variables $X'_{1:n}$. In the "informed learner" setting, G is the distribution and g the density from which each X'_i is drawn. In the "blind learner" setting, G_j and g_j are the distribution and density of X'_i conditioning on $\phi = \phi_j \in \Phi$; G_ϕ and g_ϕ are the generic distribution and density parametrized by ϕ. We denote the observed data $Z_{1:n}$, which is distributed according to (1). P_i is the distribution and p_i the density of each Z_i, conditioning on $\theta = \theta_i$ and $\phi = \phi_i$. P_θ or $P_{\theta,\phi}$ is the parametrized form. We say that $P_i = \alpha F_i + (1-\alpha)G_i$, or equivalently $p_i(z) = \alpha f_i(z) + (1-\alpha)g_i(z)$, to indicate that P_i is a weighted mixture of the distributions F_i and G_i. We assume that X, X' and Z have the same support, denoted $\mathcal{Z}$. $\mathfrak{M}_n$ is the minimax risk of a learner. $D_{\mathrm{KL}}(P_1||P_2)$ is the KL-divergence. $||P_1 - P_2||_{\mathrm{TV}}$ is the total variation distance. $I(Z,V)$ is the mutual information between the random variables Z and V. $\hat{\theta}_n : \mathcal{Z}^n \to \Theta$ denotes an arbitrary estimator for θ with a sample size of n; $\hat{\psi}_n : \mathcal{Z}^n \to \Psi$ denotes an arbitrary estimator for an arbitrary parameter vector ψ with a sample size of n.

3.2 Minimax framework

The minimax risk of estimating a parameter $\psi \in \Psi$ is equal to the risk of the estimator $\hat{\psi}_n$ which achieves smallest maximal risk across all $\psi \in \Psi$:

$$\mathfrak{M}_n = \inf_{\hat{\psi}} \sup_{\psi \in \Psi} \mathbb{E}_{Z_{1:n} \sim P_\psi^n} L(\psi, \hat{\psi}_n).$$

The minimax risk thus provides a strong guarantee: the population risk of an estimator can be no worse than the minimax risk, no matter which $\psi \in \Psi$ happens to be the true parameter. Our analysis aims to build insight into how the minimax risk increases when the training set is subjected to blind data injection attacks. In the informed learner setting we fix some ϕ and G_ϕ, and consider $\Psi = \Theta$, letting $L(\theta, \hat{\theta}_n)$ be the squared $\ell 2$ distance $||\theta - \hat{\theta}_n||_2^2$. In the blind learner setting we account for there being two parameters unknown to the learner ϕ and θ by letting $\Psi = \Phi \times \Theta$ and considering a loss function which depends only on the value of θ and its estimator, $L(\psi, \hat{\psi}_n) = ||\theta - \hat{\theta}_n||_2^2$

We follow the standard approach to lower bounding the minimax risk [22], reducing the problem of estimating θ to that of testing the hypothesis $H : V = V_j$ for $V_j \in \mathcal{V}$, where $V \sim \mathcal{U}(\mathcal{V})$, a uniform distribution across $\mathcal{V}$. $\mathcal{V} \subset \Psi$ is an appropriate finite packing of the parameter space.

The Le Cam method provides lower bound on the minimax risk of the learner in terms of the KL divergence $D_{\mathrm{KL}}(P_{\psi_1}||P_{\psi_2})$ for $\psi_1, \psi_2 \in \Psi$ [22]:

$$\mathfrak{M}_n \geq L(\psi_1, \psi_2)\Big[\frac{1}{2} - \frac{1}{2\sqrt{2}}\sqrt{nD_{\mathrm{KL}}(P_{\phi_1}||P_{\phi_2})}\Big]. \tag{2}$$

The Fano method provides lower bounds on the minimax risk of the learner in terms of the mutual information $I(Z,V)$ between the observed data and V chosen uniformly at random from $\mathcal{V}$, where $L(V_i, V_j) \geq 2\delta \; \forall V_i, V_j \in \mathcal{V}$ [22]:

$$\mathfrak{M}_n \geq \delta\Big[1 - \frac{I(Z_{1:n}; V) + \log 2}{\log |\mathcal{V}|}\Big]. \tag{3}$$

The mutual information is upper bounded by the pariwise KL divergences as

$$I(Z_{1:n}, V) \leq \frac{n}{|\mathcal{V}|^2} \sum_i \sum_j D_{\mathrm{KL}}(P_{V_i} || P_{V_j}). \tag{4}$$

3.3 Blind attacker, informed learner

In this setting we assume the attacker does not know F_θ but does know $\mathcal{F}$. The learner knows both G_ϕ and α prior to picking an estimator. In this case, as G_ϕ is known, we do not need to consider a distribution over possible values of ϕ; instead, we consider some fixed $p(z|x)$. The attacker chooses G_ϕ to attempt to maximally lower bound the minimax risk of the learner:

$$\phi^* = \mathrm{argmax}_\phi \mathfrak{M}_\mathrm{n} = \mathrm{argmax}_\phi \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\mathrm{Z}_{1:\mathrm{n}} \sim \mathrm{P}_{\theta,\psi}} \mathrm{L}(\theta, \hat{\theta}_\mathrm{n}),$$

where $L(\theta, \theta')$ is the learner's loss function; in our case the squared $\ell 2$ distance $||\theta - \theta'||_2^2$.

The attacker chooses a malicious distribution $G_{\hat{\phi}}$ which minimizes the sum of KL-divergences between the distributions indexed by $\mathcal{V}$:

$$\hat{\phi} = \mathrm{argmin}_\phi \sum_{\theta_\mathrm{i} \in \mathcal{V}} \sum_{\theta_\mathrm{j} \in \mathcal{V}} \mathrm{D}_{\mathrm{KL}}(\mathrm{P}_{\theta_\mathrm{i},\phi} || \mathrm{P}_{\theta_\mathrm{j},\phi}) \geq \frac{|\mathcal{V}|^2}{\mathrm{n}} \mathrm{I}(\mathrm{Z}^\mathrm{n}; \theta),$$

$$\text{where} \quad P_{\theta_i,\phi} = \alpha F_{\theta_i} + (1-\alpha) G_\phi.$$

This directly provides lower bounds on the minimax risk of the learner via (2) and (3).

3.4 Blind attacker, blind learner

In this setting, the learner does not know the specific malicious distribution G_ϕ used to inject points into the training set, but is allowed to know the family $\mathcal{G} = \{G_\phi : \phi \in \Phi\}$ from which the attacker picks this distribution. We propose that the minimax risk is thus with respect to the worst-case choice of both the true parameter of interest θ and the parameter of the malicious distribution ϕ:

$$\mathfrak{M}_n = \inf_{\hat{\theta}} \sup_{(\phi,\theta) \in \Phi \times \Theta} \mathbb{E}_{Z_{1:n} \sim P_{\theta,\psi}} L(\theta, \hat{\theta}_n).$$

That is, the minimax risk in this setting is taken over worst-case choice of the parameter pair $(\phi, \theta) \in \Phi \times \Theta$, but the loss $L(\theta, \hat{\theta})$ is with respect to only the true value of of θ and its estimator $\hat{\theta}$. The attacker thus designs a family of malicious distributions $\mathcal{G} = \{G_\phi : \phi \in \Phi\}$ so as to maximally lower bound the minimax risk:

$$\mathcal{G}^* = \mathrm{argmax} \inf_{\hat{\theta}} \sup_{(\mathrm{F}_\theta, \mathrm{G}_\phi) \in \mathcal{F} \times \mathcal{G}} \mathbb{E}_{\mathrm{Z}_{1:\mathrm{n}}} \mathrm{L}(\theta, \hat{\theta}).$$

We use the Le Cam approach (2) in this setting. To accommodate the additional set of parameters Φ we consider nature picking (ϕ, θ) from $\Phi \times \Theta$. The loss function is $L\big((\psi_i, \theta_i), (\psi_j, \theta_j)\big) = ||\theta_i - \theta_j||_2^2$, and thus only depends on θ. Therefore when constructing our hypothesis set we must choose well-separated θ but may arbitrarily pick each element ϕ. The problem reduces from that of estimating θ to that of testing the hypothesis $H : (\phi, \theta) = (\phi, \theta)_j$ for $(\phi, \theta)_j \in \mathcal{V}$, where nature chooses $(\phi, \theta) \sim \mathcal{U}(\mathcal{V})$.

The attacker again lower bounds the minimax risk by choosing $\mathcal{G}$ to minimize an upper bound on the pairwise KL divergences. Unlike the informed learner setting where the KL divergence was between the distributions indexed by θ_i and θ_j with ϕ fixed, here the KL divergence is between the distributions indexed by appropriate choice of pairings (θ_i, ϕ_i) and (θ_j, ϕ_j):

$$\hat{\mathcal{G}} = \mathrm{argmin}_\mathcal{G} \sum_{(\theta_\mathrm{i},\phi_\mathrm{i}) \in \mathcal{V}} \sum_{(\theta_\mathrm{j},\phi_\mathrm{j}) \in \mathcal{V}} \mathrm{D}_{\mathrm{KL}}(\mathrm{P}_{\theta_\mathrm{i},\phi_\mathrm{i}} || \mathrm{P}_{\theta_\mathrm{j},\phi_\mathrm{j}}) \geq \frac{|\mathcal{V}|^2}{\mathrm{n}} \mathrm{I}(\mathrm{Z}^\mathrm{n}; \theta),$$

$$\text{where} \quad P_{\theta_i,\phi_i} = \alpha F_{\theta_i} + (1-\alpha) G_{\phi_i}.$$

3.5 Informed attacker

We leave this setting as future work, but briefly propose a formulation so as to complete the description of settings where a minimax learner is and is not aware of G_ϕ and where the attacker is and is not aware of F_θ. In this setting the attacker knows F_θ prior to picking G_ϕ. We assume that the learner picks some $\hat{\theta}$ which is either minimax-optimal over $\mathcal{F}$ and $\mathcal{G}$ (blind learner) or minimax-optimal over $\mathcal{F}$ with respect to a fixed G_ϕ (informed learner) as defined in Section 1.5 and 1.6 respectively. We denote the appropriate set of such estimators as $\hat{\Theta}$. The attacker picks G_ϕ so as to maximally lower bound the risk for any $\hat{\theta} \in \Theta$:

$$R_{\theta,\phi}(\hat{\theta}) = \mathbb{E}_{Z_{1:n} \sim P_{\theta,\phi}} L(\theta, \hat{\theta}_n).$$

4 Main results

4.1 Informed learner, blind attacker

In the informed learner setting, the attacker chooses a single malicious distribution (known to the learner) from which to draw malicious data.

Theorem 1 (Uniform attack)**.** *The attacker picks $g_\phi(z) := g$ uniform over $\mathcal{Z}$ in the informed learner setting. We assume that $\mathcal{Z}$ is compact and that $G \ll F_i \ll F_j \; \forall \theta_i, \theta_j \in \Theta$. Then:*

$$D_{\mathrm{KL}}(P_i||P_j) + D_{\mathrm{KL}}(P_j||P_i) \leq \frac{\alpha^2}{(1-\alpha)} ||F_i - F_j||_{\mathrm{TV}}^2 \mathrm{Vol}(\mathcal{Z}) \quad \forall \theta_{\mathrm{i}}, \theta_{\mathrm{j}} \in \Theta.$$

The proof modifies the analysis used to prove Theorem 1 in [21] and is presented in the appendix. By applying Le Cam's method to P_1 and P_2 as described in the theorem, we find:

Corollary 1.1 (Le Cam bound with uniform attack)**.** *Given a data injection attack as described in Theorem 1, the minimax risk of the learner is lower bounded by*

$$\mathfrak{M}_n \geq L(\theta_1, \theta_2)\Big(\frac{1}{2} - \frac{1}{2\sqrt{2}} \sqrt{\frac{\alpha^2}{(1-\alpha)} n ||F_1 - F_2||_{\mathrm{TV}}^2 \mathrm{Vol}(\mathcal{Z})}\Big).$$

We turn to the local Fano method. Consider the traditional setting ($P_\theta = F_\theta$), and consider a packing set $\mathcal{V}$ of Θ which obeys $L(\theta_i, \theta_j) \geq 2\delta \; \forall \theta_i, \theta_j \in \mathcal{V}$, and where the KL divergences are bounded such that there exists some fixed τ fulfilling $D_{\mathrm{KL}}(F_i||F_j) \leq \delta\tau \; \forall \theta_i, \theta_j \in \mathcal{V}$. We can use this inequality and the bound on mutual information in (4) to rewrite the Fano bound in (3) as:

$$\mathfrak{M}_n \geq \delta\Big[1 - \frac{n\delta\tau + \log 2}{\log |\mathcal{V}|}\Big].$$

If we consider the uniform attack setting with the same packing set $\mathcal{V}$ of Θ, then by applying Theorem 1) in addition to the bound on mutual information in (4) to the standard fano bound in (3), we obtain:

Corollary 1.2 (Local Fano bound with uniform attack)**.** *Given a data injection attack as described in Theorem 1, and given any packing $\mathcal{V}$ of Θ so such $L(\theta_i, \theta_j) \geq 2\delta \; \forall \theta_i, \theta_j \in \mathcal{V}$ and $D_{\mathrm{KL}}(F_i||F_j) \leq \delta\tau$ $\forall \theta_i, \theta_j \in \mathcal{V}$, then the minimax risk of the learner is lower bounded by*

$$\mathfrak{M}_n \geq \delta\Big(1 - \frac{\frac{\alpha^2}{(1-\alpha)} \mathrm{Vol}(\mathcal{Z}) \mathrm{n}\tau\delta + \log 2}{\log |V|}\Big).$$

Remarks. For $\alpha \in [0, \frac{\sqrt{5}-1}{2}]$, comparing the two corollaries to the standard form of the Le Cam and Fano bounds shows that a uniform attack has the effect of reducing the effective sample size from n to $n\frac{\alpha^2}{(1-\alpha)}\mathrm{Vol}(\mathcal{Z})$. We illustrate the consequences of these corollaries for some classical estimation problems in Section 3.

4.2 Blind learner, blind attacker

We begin with a lemma that shows that for $\alpha \leq \frac{1}{2}$ the attacker can make learning impossible beyond permutation for higher rates of injection. Similar results have been shown in [18] among others, and this is included for completeness.

Lemma 1 (Impossibility of learning beyond permutation for $\alpha \leq 0.5$)**.** *Consider any hypotheses θ_1 and θ_2, with $F_1 \ll F_2$ and $F_2 \ll F_1$. We construct $\mathcal{V} = \{F, G\}^2 = \{(F_1, G_1), (F_2, G_2)\}$. For all $\alpha \leq 0.5$, there exist choices of G_1 and G_2 such that $D_{\mathrm{KL}}(P_1||P_2) + D_{\mathrm{KL}}(P_2||P_1) = 0$.*

The proof progresses by considering $g_1(z) = \frac{\alpha f_2(z)}{(1-\alpha)} + c$, $g_2(z) = \frac{\alpha f_1(z)}{(1-\alpha)} + c$, such that $||P_1 - P_2||_{\mathrm{TV}} = 0$. Full proof is provided in the appendix.

It is unnecessary to further consider values of α less than 0.5. We proceed with an attack where the attacker chooses a family of malicious distributions $\mathcal{G}$ which mimics the family of candidate distributions of interest $\mathcal{F}$, and show that this increases the lower bound on the learner's minimax risk for $0.5 < \alpha < \frac{3}{4}$.

Theorem 2 (Mimic attack)**.** *Consider any hypotheses θ_1 and θ_2, with $F_1 \ll F_2$ and $F_2 \ll F_1$. The attacker picks $\mathcal{G} = \mathcal{F}$. We construct $\mathcal{V} = \{F, G\}^2 = \{(F_1, G_1), (F_2, G_2)\}$ where $G_1 = F_2$ and $G_2 = F_1$. Then:*

$$D_{\mathrm{KL}}(P_1||P_2) + D_{\mathrm{KL}}(P_2||P_1) \leq \frac{(2\alpha-1)^2}{1-\alpha}||F_1 - F_2||_{\mathrm{TV}} \leq 4\frac{\alpha^4}{1-\alpha}||F_1 - F_2||^2_{\mathrm{TV}}.$$

The proof progresses by upper bounding $|\log \frac{p_1(z)}{p_2(z)}|$ by $\log \frac{\alpha}{1-\alpha}$, and consequently upper bounding the pairwise KL divergence in terms of the total variation distance. It is presented in the appendix. By applying the standard Le Cam bound with the the bound on KL divergence provided by the theorem, we obtain:

Corollary 2.1 (Le Cam bound with mimic attack)**.** *Given a data injection attack as described in Theorem 2, the minimax risk of the learner is lower bounded by*

$$\mathfrak{M}_n \geq L(\theta_1, \theta_2)\Big(\frac{1}{2} - \frac{1}{\sqrt{2}}\sqrt{\frac{(2\alpha-1)^2}{1-\alpha}n||F_1 - F_2||^2_{\mathrm{TV}}}\Big).$$

Remarks. For $\alpha \in [0, \frac{3}{4}]$, comparing the corollary to the standard form of the Le Cam bound shows that this attack reduces the effective sample size from n to $\frac{(2\alpha-1)^2}{1-\alpha}n$. We illustrate the consequences of this corollary for estimating a mean in Section 3. There are two main differences in the result from the bound for the uniform attack. Firstly, the dependence on $(2\alpha - 1)^2$ instead of α^2 means that the KL divergence rapidly approaches zero as $\alpha \to \frac{1}{2}$, rather than as $\alpha \to 0$ as in the uniform attack. Secondly, there is no dependence on the volume of the support of the data.

5 Minimax rates of convergence under blind attack

We analyze the minimax risk in the settings of mean estimation and of fixed-design linear regression by showing how the blind attack forms of the Le Cam and Fano bounds modify the lower bounds on the minimax risk for each model.

5.1 Mean estimation

In this section we address the simple problem of estimating a one-dimensional mean when the training set is subject to a blind attack. Consider the following family, where Θ is the interval $[-1, 1]$:

$$\mathcal{F} = \{F_\theta : \mathbb{E}_{F_\theta} X = \theta; \mathbb{E}_{F_\theta} X^2 \leq 1; \theta \in \Theta\}.$$

We apply Theorems 1 and 2 and the associated Le Cam bounds to obtain:

Proposition 1 (Mean estimation under uniform attack — blind attacker, informed learner)**.** *If the attacker carries out a uniform attack as presented in theorem 1, then there exists a universal constant $0 < c < \infty$ such that the minimax risk is bounded as:*

$$\mathfrak{M}_n \geq c \min\Big[1, \sqrt{2\frac{1-\alpha}{\alpha^2 n}}\Big].$$

The proof is direct by using the uniform-attack form of the Le Cam lower bound on minimax risk presented in corollary 1.1 in the proof of (20) in [21] in place of the differentially private form of the lower bound in equation (16) of that paper.

Proposition 2 (Mean estimation under mimic attack — blind attacker, blind learner)**.** *If the attacker carries out a mimic attack as presented in theorem* 2*, then there exists a universal constant* $0 < c < \infty$ *such that the minimax risk is bounded as:*

$$\mathfrak{M}_n \geq c \min\Big[1, \frac{1}{4-2\alpha}\sqrt{\frac{1-\alpha}{n}}\Big].$$

The proof is direct by using the mimic-attack form of the Le Cam lower bound on minimax risk presented in corollary 2.1 in the proof of (20) in [21] in place of the differentially private form of the lower bound in equation (16) of that paper.

5.2 Linear regression with fixed design

We now consider the minimax risk in a standard fixed-design linear regression problem. Consider a fixed design matrix $X \in \mathbb{R}^{n\times d}$, and the standard linear model

$$Y = X\theta^* + \epsilon,$$

where $\epsilon \in \mathbb{R}^n$ is a vector of independent noise variables with each entry of the noise vector upper bounded as $|\epsilon_i| \leq \sigma < \infty \; \forall i$. We assume that the problem is appropriately scaled so that $||X||_\infty \leq 1$, $||Y||_\infty \leq 1$, and so that it suffices to consider $\theta^* \in \Theta$, where $\Theta = S_d$ is the d-dimensional unit sphere. The loss function is the squared $\ell 2$ loss with respect to θ^*: $L(\hat{\theta}_n, \theta^*) = ||\hat{\theta}_n - \theta^*||_2^2$. It is also assumed that X is full rank to make estimation of θ possible.

Proposition 3 (Linear regression under uniform attack - blind attacker, informed learner)**.** *If the attacker carries out a uniform attack per Theorem 1, and* $s_i(A)$ *is the* i*th singular value of* A*, then the minimax risk is bounded by*

$$\mathfrak{M}_n \geq \min\Big[1, \frac{\sigma^2 d(1-\alpha)}{n\alpha^2 s_{\max}^2(X/\sqrt{n})}\Big].$$

The proof is direct by using the uniform-attack form of the Fano lower bound on minimax risk presented in corollary 1.2 in the proof of (22) in [21] in place of the differentially private form of the lower bound in equation (19) of that paper, noting that $\mathrm{Vol}(\mathcal{Z}) \leq 1$ by construction. If we consider the orthonormal design case such that $s_{\max}^2(X/\sqrt{n}) = 1$, and recall that lower bounds on the minimax risk in linear regression in traditional settings is $\mathcal{O}(\frac{\sigma^2 d}{n})$, we see a clear reduction in effective sample size from n to $\frac{\alpha^2}{1-\alpha}n$.

6 Discussion

We have approached the problem of data injection attacks on machine learners from a statistical decision theory framework, considering the setting where the attacker does not observe the true distribution of interest or the learner's training set prior to choosing a distribution from which to draw malicious examples. This has applications to the theoretical analysis of both security settings, where an attacker attempts to compromise a machine learner through data injection, and privacy settings, where a user of a service aims to protect their own privacy by sumbitting some proportion of falsified data. We identified simple attacks in settings where the learner is and is not aware of the malicious distribution used which reduce the effective sample size when considering rates of convergence of estimators. These attacks maximize lower bounds on the minimax risk. These lower bounds may not be tight, and we leave as future work thorough exploration of optimality of attacks in this setting and the establishing of optimal estimation procedures in the presence of such attacks. Exploration of attacks on machine learners in the minimax framework should lead to better understanding of the influence an attacker might have over a learner in settings where the attacker has little information.

References

(1) M. Barreno, B. Nelson, R. Sears, A. D. Joseph and J. D. Tygar, ACM Symposium on Information, computer and communications security, 2006.

(2) M. Barreno, B. Nelson, A. D. Joseph and J. Tygar, *Machine Learning*, 2010, **81**, 121–148.

(3) P. Laskov and M. Kloft, ACM workshop on security and artificial intelligence, 2009.
(4) P. Laskov and R. Lippmann, *Machine learning*, 2010, **81**, 115–119.
(5) H. Xiao, H. Xiao and C. Eckert, European Conference on Artificial Intelligence, 2012, pp. 870–875.
(6) B. Biggio, B. Nelson and P. Laskov, *arXiv preprint arXiv:1206.6389*, 2012.
(7) B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, N. Taft and D. Tygar, *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2008-73*, 2008.
(8) R. Sommer and V. Paxson, IEEE Symposium on Security and Privacy, 2010.
(9) R. J. Bolton and D. J. Hand, *Statistical science*, 2002, 235–249.
(10) M. Al Hasan, V. Chaoji, S. Salem and M. Zaki, SDM Workshop on Link Analysis, Counter-terrorism and Security, 2006.
(11) S. Mei and X. Zhu, Association for the Advancement of Artificial Intelligence, 2015.
(12) W. Liu and S. Chawla, IEEE International Conference on Data Mining, 2009.
(13) S. Alfeld, X. Zhu and P. Barford, Association for the Advancement of Artificial Intelligence, 2016.
(14) M. Bruckner and T. Scheffer, ACM SIGKDD, 2011.
(15) C. Dwork, in *Automata, languages and programming*, Springer, 2006, pp. 1–12.
(16) M. Chen, C. Gao and Z. Ren, *arXiv preprint arXiv:1511.04144*, 2015.
(17) M. Azizyan, A. Singh and L. Wasserman, Neural Information Processing Systems, 2013.
(18) Y. Chen, X. Yi and C. Caramanis, *arXiv preprint arXiv:1312.7006*, 2013.
(19) M. Kearns and M. Li, *SIAM Journal on Computing*, 1993, **22**, 807–837.
(20) J. Duchi, M. J. Wainwright and M. I. Jordan, Neural Information Processing Systems, 2013.
(21) J. Duchi, M. Wainwright and M. Jordan, *arXiv preprint arXiv:1302.3203v4*, 2014.
(22) A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer Publishing Company, Incorporated, 1st, 2008.