

# Supervised Machine Learning with Plausible Deniability

Stefan Rass<sup>\*†</sup>   Sandra König<sup>‡</sup>   Jasmin Wachter<sup>§</sup>  
 Manuel Egger<sup>¶</sup>   Manuel Hobisch<sup>||</sup>

June 9, 2021

## Abstract

We study the question of how well machine learning (ML) models trained on a certain data set provide privacy for the training data, or equivalently, whether it is possible to reverse-engineer the training data from a given ML model. While this is easy to answer negatively in the most general case, it is interesting to note that the protection extends over non-recoverability towards *plausible deniability*: Given an ML model  $f$ , we show that one can take a set of purely random training data, and from this define a suitable “learning rule” that will produce a ML model that is exactly  $f$ . Thus, any speculation about which data has been used to train  $f$  is deniable upon the claim that any other data could have led to the same results. We corroborate our theoretical finding with practical examples, and open source implementations of how to find the learning rules for a chosen set of training data.

## 1 Introduction

Imagine a situation in which training data has been used to fit a ML model, which Alice gives away to Bob for his own use. Alice’s training data, however, shall remain her own private property, and Bob should be unable to recover this

---

<sup>\*</sup>Universitaet Klagenfurt, Institut of Artificial Intelligence and Cybersecurity, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria, [stefan.rass@aau.at](mailto:stefan.rass@aau.at)

<sup>†</sup>Johannes Kepler University, Secure and Correct Systems Lab, Altenberger Straße 69, 4040 Linz, Austria, [stefan.rass@jku.at](mailto:stefan.rass@jku.at)

<sup>‡</sup>AIT Austrian Institute of Technology, Center for Digital Safety and Security, Giefinggasse 4, 1210 Vienna, Austria, [sandra.koenig@ait.ac.at](mailto:sandra.koenig@ait.ac.at)

<sup>§</sup>Universitaet Klagenfurt, Doctoral School for Responsible Safe and Secure Robotic Systems Engineering, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria, [jawachte@edu.aau.at](mailto:jawachte@edu.aau.at)

<sup>¶</sup>Universitaet Klagenfurt, Institut of Artificial Intelligence and Cybersecurity, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria, [megger@edu.aau.at](mailto:megger@edu.aau.at)

<sup>||</sup>Universitaet Klagenfurt, Institut of Artificial Intelligence and Cybersecurity, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria, [mahobisch@edu.aau.at](mailto:mahobisch@edu.aau.at)

information from the ML model in his possession. For example, Alice could be a provider of a critical infrastructure, having trained a digital twin to emulate the behavior of her system, which Bob, as a risk analyst, shall assess on Alice’s behalf. To this end, however, Alice must not disclose all the details of her infrastructure, since this is highly sensitive information and Bob, as an external party, may not be sufficiently trustworthy to open up to him. Still, Alice needs Bob’s expertise on risk management and risk assessment to help her protect her assets, and therefore needs to involve Bob to some extent.

We cannot prevent Bob from “guessing”, i.e., Bob can always try to reverse-engineer the data that Alice used to create the model. This comes to a perhaps high-dimensional, yet conceptually simple, optimization problem, which may indeed be tractable with today’s computing power. Our goal here is the proof of two statements about this possibility: First, if the training data set is “sufficiently large” (where the term “sufficient” will be quantified more precisely), Bob cannot unambiguously recover the training data. Second, and more importantly, Alice can deny any proposal training data that Bob thinks to have recovered, by exposing a set of random data along with a certificate that this random decoy data has been used to train the model (although it was not). Alice can do so by adapting her optimization problem accordingly to give a desired result (the ML model that Bob has) from any a priori (randomly chosen) training data set.

Note that Bob, since he can “use” the ML model, has no difficulties to evaluate it on a given dataset to produce data upon which a re-training of the model would reproduce what Bob received from Alice. This trivial possibility cannot be eliminated. Our question, however, is whether Bob cannot just produce “any” dataset, but find Alice’s original dataset that way used to produce the model in his possession. In other words, does an ML model leak out private information of Alice? The answer obtained in this work is “no”, by leveraging a degree of freedom in how an artificial intelligence (AI) model is trained: Alice can provide Bob with decoy data that she claims to underly what Bob has as the ML model; however, Alice can plausibly claim the model to have come up as the optimum under some optimization problem that she can craft to her wishes.

The key observation reported in this paper is the fact that we can “utilize” non-explainability for the purpose of privacy of data embodied in an ML model. More specifically, we will show how to define an error metric that makes the learning algorithm converge to any target output that we like. We state this intuition more rigorously in Section 4, after some necessary preliminary considerations. In a way, such a designed error metric acts similar to a “secret key” in encryption, only that it accomplishes plausible deniability in our context. A numerical proof-of-concept is given in Section 5. Section 6 embeds ours in the landscape of related work and links the results with issues of the General Data Protection Regulation (GDPR). Section 7 is devoted to further uses, limitations, ethical considerations and possible extensions (further expanded in the Appendix).

## 1.1 Problem Setting

Throughout this work, scalars will appear in regular font, while bold printing will indicate vectors (lower case letters) or matrices (uppercase letters); for example, the symbols  $\mathbf{A} \in \mathbb{R}^{n \times m}$  means an  $(n \times m)$ -matrix over  $\mathbb{R}$ . Uppercase letters in normal font will denote sets, vector spaces, and random variables. Probability distributions appear as calligraphic letters, like  $\mathcal{F}$ . The symbol  $X \sim \mathcal{F}$  indicates the random variable  $X$  to have the distribution  $\mathcal{F}$ .

Let the ML model training be the problem to find a best function  $f$  to approximate a given set of  $n$  points, called *training data*  $(\mathbf{x}_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$  by “minimizing” the error vector  $\mathbf{e} = (y_1 - f(\mathbf{x}_1), y_2 - f(\mathbf{x}_2), \dots, y_n - f(\mathbf{x}_n)) \in \mathbb{R}^n$ . The resulting goodness of fit is later assessed by evaluating  $f$  on a (distinct) set of *validation data*, often providing some error measure to quantify the approximation quality<sup>1</sup>.

The best function  $f$  is usually found by fixing its algebraic form, and tuning some parameters therein by sophisticated optimization methods. Let us postpone the formal optimization problem until Section 4, to first state the problem: assume that we *are given* a trained (fitted) model  $f$ , but not the training data. Is there a way to reverse-engineer the training data from  $f$  alone? For example, if we are given access and insight to a trained neural network (NN), can we use the weights that we see therein to learn something about the data that the NN has been trained with?

An obvious answer is “yes”, if we have the training samples at least partly, since it is straightforward to evaluate  $f$  on given values  $\mathbf{x}_i$  to recover at least an approximate version of the target value  $y_i$ , if it is the only unknown quantity. To avoid such triviality, let us assume that the training data is *not available* but that we have *white-box access* to the machine learning model  $f$ . This means that we can look into how  $f$  is constructed (i.e., see the weights if it is a NN, regression model, etc.), but have no clue about the data or any parts of it, on which the model has been trained. This is what we are after, and wish to reverse-engineer. The case of partial knowledge of the attacker is revisited and discussed in Section 7.

## 1.2 Some (selected) Applications

**Making Community Knowledge Securely Available:** Suppose that we want to release data not directly, but “functionally useable” by fitting an ML model so that everyone can produce artificial data from  $f$ , but we do not hereby disclose the original data that  $f$  was trained from. This is to retain intellectual property, while still making the knowledge publicly available.

**Co-Simulation:** simulations are in many cases domain-specific, e.g., water networks are described using different (physical) mechanisms as traffic or energy networks. Combining these in a co-simulation framework, such as brought up in [14], raises compatibility issues between different simulation models. Fitting

---

<sup>1</sup>We will hereafter have no need for the distinction of training and validation data, since our concern is exclusively on the training here.

ML models, say, NNs, to emulate the outputs of different simulations provides a simple compatibility layer for co-simulation. Plausible deniability is here good for privacy, say, if the physical structure of the simulated process is sensitive information (e.g., a critical infrastructure, uses data related to persons, etc.)

## 2 Definitions

Our formalization of security distinguishes *deniability* from *plausible deniability*, where the latter notion is stronger. Informally, deniability of a hypothesis about training data can be understood as the possibility that there may be another set of training records that have produced the same result. To formalize these notions, we first introduce a generic representation of the machine learning problem. The following section is not meant as an introduction to the general field, but to settle the context and symbols in terms of which we state the main results of this work.

### 2.1 Fitting ML Models

We will consider only supervised training in this work. Specifically, we will view an algorithm to train an ML model as a *function* that returns a parameterized function  $f(\cdot; \mathbf{p})$  upon input of the training data set  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ , together with a set  $\Omega$  of parameters to configure the training (optimizer). We assume this configuration to be arbitrary, but admit an unambiguous string representation, i.e.,  $\Omega \subseteq \{0, 1\}^*$ . The variable inputs to  $f$  herein take the same structure as the training data. Viewing the training algorithm as a mapping, it is natural to ask for invertibility of it, and *deniability* then turns out as non-invertibility. This brings us to the first definition:

**Definition 1** (Machine Learning Model and Training Algorithms). *A machine learning model is a set  $ML$  of functions  $f : \mathbb{R}^m \times \mathbb{R}^d \rightarrow \mathbb{R}$ , mapping an input  $\mathbf{x} \in \mathbb{R}^m$  and parameter vector  $\mathbf{p} \in \mathbb{R}^d$  into  $\mathbb{R}$ .*

*A training algorithm for a machine learning model  $ML$  is a function fit:  $\mathbb{R}^{n \times (m+1)} \times \Omega \rightarrow ML$ . This function takes a training data matrix  $\mathbf{T}$  composed from  $n$  instances of input/output pairs  $(\mathbf{x}_i, y_i) \in \mathbb{R}^{m+1}$  for  $i = 1, 2, \dots, n$ , and auxiliary information  $\omega \in \Omega$ , to output a (concrete) element  $f \in ML$ .*

The temporary assumption of  $f$  outputting only scalars is here adopted only for simplicity, and later dropped towards ML models with many outputs in Section 4.2 as Corollary 2.

The set  $ML$  can contain functions of various shape, and is not constrained to have all functions of the same algebraic structure, although in most practical cases, the functions will have a homogeneous form. For example,  $ML$  could be (among many more possibilities)

- the set of all linear regression models  $f(\mathbf{x}, \mathbf{p}) = \mathbf{p}^\top \mathbf{x}$ , where the vector  $\mathbf{p}$  is the coefficients in the linear model. We will use this example in Section 5.

- the set of (deep) neural networks with a fixed topology and number of layers. The entirety of synaptic weights and node biases then defines the vector  $\mathbf{p}$ .
- the set of support vector machines, in which  $\mathbf{p}$  is the normal vector and bias for the classifying (separating) hyperplane,
- and many more.

In Definition 1, an implicit consistency between the set of machine learning models  $ML$  and the training algorithm is implied by the (obvious) requirement that (i) the training data needs to have the proper form and dimension to be useful with the functions in  $f$ , and (ii) that the particular element  $f$  is specified by an *admissible* parameterization  $\mathbf{p} \in \mathbb{R}^d$  for the functions in  $ML$ , since not all settings for  $\mathbf{p}$  may be meaningful to substitute in the general function  $f$ .

The inclusion of the auxiliary information  $\omega$  in the training models the fact that different models may require different techniques of training, essentially meaning the application of different optimization techniques. In particular,  $\omega$  will in practical cases (among others) include a *specification of the error metric* to be used with the training, which is the goal function to optimize. The core of a training algorithm is a “learning rule”, being a prescription of how to update the ML model parameterization (iteratively). We will hereafter simplify matters by abstracting from the detailed optimization technique, and confining ourselves to look only at the error metric to be used with the optimization, and going into the training as part of the training algorithm configuration  $\omega$ .

## 2.2 Supervised Training by Optimization

Generally, we will let the error metric measure the approximation error in a supervised learning strategy. This learning is based on a set of  $n$  samples  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^m \times \mathbb{R}$ . In general, the machine learning problem then takes the generic form of a minimization problem

$$\min \|((\mathbf{x}_i, y_i) - f(\mathbf{x}_i, \mathbf{p}))_{i=1}^n\| \text{ over } \mathbf{p} \in P, \quad (1)$$

where the set  $P \subseteq \mathbb{R}^d$  optionally constrains parameters to feasible ranges and combinations. We let  $\mathbf{p}^*$  denote an (arbitrary) optimum to this problem, which then pins down a specific  $f^* \in ML$ . In (1),  $\|\cdot\|$  is a topological norm, specified via the *auxiliary information*  $\omega$ . Since all norms on  $\mathbb{R}^n$  are equivalent (Theorem 4), choosing a different norm/error metric only amounts to a scaling of the (absolute) error bound. Popular error metrics like root mean squared error (RMSE), mean absolute error (MAE), etc., are all expressible by norms (see Appendix A for details omitted here), so that their use here in place of RMSE, MAE, or others, goes without loss of much generality. Appendix A defines norms, induced metrics and pseudometrics rigorously, for convenience of the reader.

### 2.3 Deniability and Plausible Deniability

Returning to our view of ML training as a mere function that, under a given configuration  $\omega$  maps training data to a concrete function  $f \in ML$ , we can consider invertibility of this process as the problem of reverse-engineering the training data from a given model  $f$ . If this is not possible, in the sense of (normal) function inversion, then the recovery of training data from  $f$  will fail. Since invertibility is equivalent of simultaneous injectivity and surjectivity of the training function, the recovery can fail in two cases:

1. the given  $f \in ML$  simply does not correspond to *any* possible training data under any (or a given) configuration  $\omega$ . In that case, the training algorithm “fit” was not surjective, as a function.
2. the given  $f \in ML$  may arise identically from several different sets of training data, in which case the fitting, as a function, was apparently not injective.

It is the latter incident that we will use to define *deniability*, understood as the *possibility* of alternative training data sets, besides what we have recovered. Formally:

**Definition 2** (Deniability). *Let a (fixed)  $f_0 \in ML$  be given that has been trained from some unknown data set under a configuration  $\omega$ . We call a given (proposed) training data set  $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$  deniable, if another set  $T' \neq T$  exists, upon which the training algorithm fit would have produced the same function  $f_0$ , possibly under a different configuration  $\omega'$  that can depend on  $T'$ .*

*Intuitively: plausibility holds if there is another quantity of training data that would have lead to the same  $f_0$ .*

Obviously, the non-invertibility of the training as a mapping implies deniability, but the converse is not true, since if the training function/algorithm is not surjective, no alternative training data  $T'$  would exist. To keep the data recovery problem interesting, however, let us in the following assume that the model has really been trained from existing yet unknown information, so that the parameterization is guaranteed to be admissible.

Even if there is an alternate set of training data, one may question its validity on perhaps semantic grounds. For example, if the training data is known to obey certain numeric bounds, or coming from physical processes with a known distribution, we could perhaps judge an alternative proposal as implausible, since it *may* produce the same ML model, but the underlying data is arguably not meaningful in the application context. The stronger notion of *plausible deniability* demands that the alternative training data should also “statistically agree” with the expectations, or more formally:

**Definition 3** (Plausible Deniability). *Let a (fixed)  $f_0 \in ML$  be given that has, under a configuration  $\omega$ , been trained from some unknown data set. Let, in addition, be a distribution family  $\mathcal{F}$  be given to describe the context/source of*

the training data. We call a given (proposed) training data set  $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$  plausibly deniable, if another set  $T' \neq T$  exists that has the same statistical distribution  $\mathcal{F}$ , and upon which the training algorithm would have produced the same function  $f_0$ , possibly under a different configuration  $\omega'$  that can depend on  $T'$ .

Intuitively: *plausible deniability holds if it cannot be demonstrated that the alternative proposal data is purely artificial.*

Definition 3 differs from Definition 2 only in the fact that a proposal training data should not look “too much different” from what we would expect about the unknown training data, formalized by imposing a given distribution  $\mathcal{F}$ . The important point here is the order of quantifiers, demanding that the distribution family  $\mathcal{F}$  is given a priori, as a specification of what sort of training data *can be plausible* in the given context. It is important to observe here that this *does not* require the unknown data, upon which the given ML model  $f_0$  has been trained, needs to have a distribution from  $\mathcal{F}$ ; this can hold in practical instances, but the denial may indeed be a claim that  $f_0$  has been trained from data coming from an entirely different source, not having the distribution  $\mathcal{F}$ . Let us briefly expand on the intuition by giving an example:

**Example 1.** *Suppose that in a social network, somebody uses the data from a user to predict upcoming messages concerning a certain topic, or just trains a model to predict a persons overall activity in posting news on the network. If the model is, for simplicity, about the inter-arrival times of a posting on the media, we can model the event of postings as a Poisson process, having an exponential distribution for the time between two activities with a rate parameter  $\lambda > 0$ . Letting  $\lambda$  vary over  $(0, \infty)$  yields the family  $\mathcal{F}$  in Definition 3.*

*Now, suppose that the provider aggregates some statistics about the community’s activity (say, for advertising purposes), and releases the concrete distribution of inter-arrival times between postings to the public (e.g., underpinning the empirical findings by releasing artificial data coming out of a Generative Adversarial Networks (GAN) for others to confirm the data science independently). This would come to the publication of a specific distribution  $F_\lambda \in \mathcal{F}$  from the aforementioned family of distributions.*

*Now, to have a need for deniability, one may suspect the provider to have profiled a particular network user  $X$ , and suppose that the activity prediction model  $f_0$  is about user  $X$  specifically. This would be yet another member  $F_{\lambda_X} \in \mathcal{F}$ .*

*The point behind plausible deniability is that the provider, facing accusal of having released an activity model  $f_0$  for user  $X$ , can deny this upon admitting that the model was trained from social network data, but not specifically the data of user  $X$ , having had the distribution  $F_{\lambda_X}$ , but rather from the data for the entire community, having the (different) distribution  $F_\lambda$ . The fact that the underlying data is admitted to have an exponential distribution is for plausibility, while the claim that it was not user  $X$ ’s data is the denial.*

While Example 1 used the same distribution shape as the underlying unknown data may have had, a denial may be argued even stronger by claiming

that the distribution used to train  $f_0$  may have come from an entirely different source, having a distinct distribution at all. Definition 3 allows this by not constraining the distribution family to include only distributions of a particular shape or algebraic structure (e.g., gamma distribution or more general exponential family), but allowing it to be any shape that is “believable” in the given context. Our experimental results shown later in Section 5 demonstrate that this possibility also practically works.

Since this is a much stronger notion than the previous, it comes somewhat unexpected that it is satisfiable under some conditions, in the sense that we can even *freely* choose the alternative training data, if we (heavily) exploit the freedom to change the configuration  $\omega$  for the optimization. In particular, we can modify the error metric, as part of  $\omega$ , to let us attain the optimum at the given function  $f$  (more specifically its parameterization  $\mathbf{p}$ ) for any a priori chosen training data. This will be Theorem 2. Before proving this main result, let us briefly return to the weaker notion of deniability first. Proving the possibility to deny is in fact an easy matter of information-theoretic arguments, as we show in Section 3

### 3 Deniability by Non-Unique Recovery

Suppose that we are given a model with a (fixed) number of  $d$  parameters. The number  $d$  can be large, but still much smaller than the training sample size, so that there is intuitively no unique recovery possible. In fact, we have a simple result, whose proof appears in Appendix B.1:

**Theorem 1.** *For a given ML model (according to Definition 1) with  $d$  parameters. Let the (unknown) training data come from a random source  $Z$  with entropy  $H(Z)$  bits, and let the function  $f$  require (at least)  $k$  bits to encode, and assume that  $f$  has been trained from  $n$  unknown records.*

*If the number  $n$  exceeds*

$$n > \frac{k}{H(Z)}, \quad (2)$$

*then any candidate training data extracted from  $f$  is deniable (in the sense of Definition 2).*

A suitable number  $k$  as used in the above result is practically easy to find, since it suffices to find *any* number  $k$  of bits that encodes  $f$ , and if this number is not the minimum, the bound (2) only becomes coarser<sup>2</sup>. In the simplest case,  $k$  can be found by saving the ML model to a file, and taking the file size to approximate  $k$  from above. Expressed boldly, we cannot hope to extract a “uniquely defined” Giga-byte of training data from a 100 kbit sized model  $f$ .

---

<sup>2</sup>finding a tight bound in (2) would require to replace  $k$  by the entropy of the parameter vector  $\mathbf{p}$  or the Kolmogorov complexity of the random  $f_0$  as emitted by the training algorithm. Either quantity appears hardly possible to get in practice.

## 4 Plausible Deniability

To formalize and prove plausible deniability of the training, imagine an adversary to have a given model  $f_0 = f(\cdot, \mathbf{p}^*)$  in its possession, looking to recover the unknown training data  $(\mathbf{x}_i, y_i)_{i=1}^n$  from it. For feasibility, let us even assume that the model contains “enough” information to let the attacker expect a successful such recovery. Specifically, the training has lead to the vector  $\mathbf{p}^*$ , from which the recovery of the data is attempted.

Generically, the recovery is the solution of an inverse (optimization) problem with  $\mathbf{p}^*$  as fixed input, and using a norm  $\|\cdot\|$  of the adversarial reverse-engineer’s choice:

$$\operatorname{argmin} \|((\mathbf{x}_i, y_i) - f(\mathbf{x}_i, \mathbf{p}^*))_{i=1}^n\| \quad (3)$$

over  $(\mathbf{x}_i, y_i)_{i=1}^n \in \mathbb{R}^{n \times (m+1)}$  (here being unconstrained for simplicity and to be clear on the dimensions). Once confronted with the adversary’s proposal solution, the original trainer can deny the result’s correctness by plausibly claiming that the training algorithm in (1) used a norm that is *different* from the adversary’s choice in (1). Theorem 2 gives conditions under which this claim is possible; more precisely, it lets the trainer construct a norm from a randomly chosen training data set according to a desired distribution  $\mathcal{F}$ , which recovers the model  $f$  upon training with this hand-crafted norm.

Like in encryption, the norm herein takes the role of a “secret key” to train the model, and the plausibility is by exposing a different “secret” (norm) to claim that the training was done from entirely different data, and only coincidentally produced the model in the adversary’s hands (Figure 4 in the Appendix graphically shows the flow as an analogy to the secrecy of contemporary encryption; the concept is comparable).

### 4.1 The Main Result

The bottom line of our previous considerations is that we are thus free to define our error metric in any way we like, without changing the results of the training in a substantial way, by crafting our own norm as we desire, and define a distance metric as the norm of the absolute error vector. In a nutshell, our construction will use the semi-norm  $\|\mathbf{x}\|_A := \sqrt{\mathbf{x}^\top \cdot \mathbf{A} \cdot \mathbf{x}}$ , induced by any positive semi-definite matrix  $\mathbf{A}$ . The trick will be choosing  $\mathbf{A}$  so that the semi-norm becomes zero at a desired error vector, i.e., point in  $\mathbb{R}^n$ . Given any decoy training data  $T'$ , it is not difficult to find such a matrix  $\mathbf{A}$  by computing the error vector  $\mathbf{e} = (f(\mathbf{x}_i, \mathbf{p}^*) - y_i)_{i=1}^n \in \mathbb{R}^n$ , and picking  $\mathbf{A}$  such that  $\mathbf{A} \cdot \mathbf{e} = 0$ . Lemma 2 in Appendix A.1 describes how to do this step-by-step.

This is almost one half of the construction, culminating in Lemma 1, which adds conditions to ensure the local optimality of the desired error vector  $\mathbf{e}$ . The other half is the extension of this semi-norm into a norm, which is Theorem 2.

**Lemma 1.** *Let  $f : \mathbb{R}^m \times \mathbb{R}^d \rightarrow \mathbb{R}$  be parameterized by a vector  $\mathbf{p} \in \mathbb{R}^d$  and map an input value vector  $\mathbf{x}$  to a vector  $\mathbf{y} = f(\mathbf{x}, \mathbf{p})$ . Let  $\mathbf{p}^* \in \mathbb{R}^d$  be given as fixed,*

and let us pick arbitrary training data  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ . Finally, define the error vector  $\mathbf{e} = (y_i - f(\mathbf{x}_i, \mathbf{p}^*))_{i=1}^n \in \mathbb{R}^n$ .

Let for all  $\mathbf{x}_i$  the functions  $f(\mathbf{x}_i, \cdot)$  be totally differentiable w.r.t.  $\mathbf{p}$  at  $\mathbf{p} = \mathbf{p}^*$  with derivative  $\mathbf{d}_i = D_{\mathbf{p}}(f(\mathbf{x}_i, \mathbf{p}))(\mathbf{p}^*) \in \mathbb{R}^d$ . Put all  $\mathbf{d}_i^\top$  for  $i = 1, 2, \dots, n$  as rows into a matrix  $\mathbf{M} \in \mathbb{R}^{n \times d}$  and assume that it satisfies the rank condition

$$\text{rank}(\mathbf{M}|\mathbf{e}) \neq \text{rank}(\mathbf{M}). \quad (4)$$

Then, there exists a semi-norm  $\|\cdot\|$  on  $\mathbb{R}^n$  such that  $\mathbf{p}^*$  locally minimizes  $\|e(\mathbf{p}^*)\|$ , i.e., there is an open neighborhood  $U$  of  $\mathbf{p}^*$  inside which  $\|e(\mathbf{p}^*)\| \leq \|e(\mathbf{p})\|$  for all  $\mathbf{p} \in U$ .

**Remark 1.** The perhaps more convenient condition to work with is assuming  $f$  to be partially differentiable w.r.t. all parameters  $p_1, \dots, p_d$ , and to assume the derivatives  $\partial f / \partial p_i$  to be continuous at all training data points  $\mathbf{x}_i$ . In that case,  $\mathbf{d}_i$  is just the gradient  $\nabla_{\mathbf{p}} f(\mathbf{x}_i, \mathbf{p})$  and  $\mathbf{M}$  is nothing else than the Jacobian of the function  $g : \mathbb{R}^d \rightarrow \mathbb{R}^n$ , sending  $\mathbf{p}$  to the vector of values  $(f(\mathbf{x}_1, \mathbf{p}), \dots, f(\mathbf{x}_n, \mathbf{p}))$ , where all  $\mathbf{x}_i$  are fixed, and the result depends only on  $\mathbf{p}$ . The general condition stated in Lemma 1 is just total differentiability of  $g$ , or, in a slightly stronger version,  $g$  having all continuous partial derivatives.

The proof of Lemma 1, as well as the proof for the stronger Theorem 2 are both given in the Appendix.

**Theorem 2.** Under the hypotheses of Lemma 1, there exists a norm  $\|\cdot\|$  on  $\mathbb{R}^N$  such that  $\mathbf{p}^*$  locally minimizes  $\|e(\mathbf{p})\|$  as a function of  $\mathbf{p}$ .

Now, let us go back and remember the order of specification: given the model by its parameters  $\mathbf{p}^*$ , and – independently of that – given an arbitrary probability distribution family  $\mathcal{F}$ , we can sample decoy training data from  $\mathcal{F}$ , and construct the norm from it. Thm. 2 thus makes Def. 3 of plausible deniability straightforwardly satisfiable.

It is natural to ask whether the norm that Theorem 2 asserts can be replaced by a “more common” choice of error metric, such as MSE or MAE. This is in fact possible for MAE; see Appendix B.4 for the proof of this Corollary:

**Corollary 1.** Under the hypotheses of Theorem 2, there is a matrix  $\mathbf{C}$  such that  $\mathbf{p}^*$  locally minimizes the mean average error  $\text{MAE}(\mathbf{C} \cdot \mathbf{e})$  of the error vector  $\mathbf{e}$ .

## 4.2 Multi-Output ML Models

Let us now drop the assumption of our ML model to output only numbers, and look at vectors as output. This transforms the error vector into an error matrix, and we have the following result, stated again in full detail, and proven in Appendix B.5.

**Corollary 2.** Take  $k, m, d \geq 1$  and let  $f : \mathbb{R}^m \times \mathbb{R}^d \rightarrow \mathbb{R}^k$  be parameterized by a vector  $\mathbf{p} \in \mathbb{R}^d$ , and write  $f_j$  for  $j = 1, \dots, k$  to denote the  $j$ -th coordinate function. For a fixed parameter vector  $\mathbf{p}^*$  and arbitrary training

data  $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n) \in \mathbb{R}^m \times \mathbb{R}^k$ , define the error matrix  $\mathbf{E}$  row-wise as  $\mathbf{E} = (\mathbf{y}_i^\top - f(\mathbf{x}_i, \mathbf{p}^*))_{i=1}^n \in \mathbb{R}^{n \times k}$ . In this matrix, let  $\mathbf{e}_j \in \mathbb{R}^n$  be the  $j$ -th column.

For all  $j = 1, 2, \dots, k$  and all training points  $\mathbf{x}_i$ , assume that each  $f_j(\mathbf{x}_i, \mathbf{p})$  is totally differentiable w.r.t.  $\mathbf{p}$  at (the same point)  $\mathbf{p} = \mathbf{p}^*$ , with derivative  $\mathbf{d}_{i,j} = D_{\mathbf{p}}(f_j(\mathbf{x}_i, \mathbf{p}))(\mathbf{p}^*) \in \mathbb{R}^d$ . For each  $j$ , define the matrix  $\mathbf{M}_j = (\mathbf{d}_{i,j})_{i=1}^n \in \mathbb{R}^{n \times d}$  and let the rank condition  $\text{rank}(\mathbf{M}_j | \mathbf{e}_j) \neq \text{rank}(\mathbf{M}_j)$  hold.

Then, there exists a matrix-norm  $\|\cdot\|$  on  $\mathbb{R}^{n \times k}$  such that  $\mathbf{p}^*$  locally minimizes  $\|\mathbf{E}(\mathbf{p}^*)\|$ , i.e., there is an open neighborhood  $U$  of  $\mathbf{p}^*$  s.t.  $\|\mathbf{E}(\mathbf{p}^*)\| \leq \|\mathbf{E}(\mathbf{p})\|$  for all  $\mathbf{p} \in U$ .

Equipped with Theorem 2 and its corollaries, we can now finally state a result about plausible deniability, similar to Theorem 1. The proof is by a direct application of the respective results as stated above.

**Theorem 3.** *For a given ML model  $f$ , let the (unknown) training data come from a random source with known distribution  $\mathcal{F}$ . Then, for every choice of alternative training data  $T'$ , randomly sampled from the same distribution  $\mathcal{F}$ , we can find an error metric induced by a (properly crafted) norm  $\|\cdot\|$  so that the training algorithm, upon receiving the training data  $T'$  and error metric (through the configuration  $\omega$ ), reproduces the given model  $f$  exactly. Thus, any data recovered from  $f$  is plausibly deniable in the sense of Def. 3.*

The case where the distribution  $\mathcal{F}$  is unknown is even simpler, since plausibility can only be argued if there is a ground truth known as the distribution  $\mathcal{F}$ . If this ground truth is not available, there is nothing to argue regarding plausibility.

## 5 Numerical Evaluation and Validation

We demonstrate a proof-of-concept for our plausible deniability concept in machine learning in the context of a fictional scenario of fitting a regression model, delegating the (lengthier) details to Appendix C. The experiment was conducted as follows: we picked a random vector  $\mathbf{p}$  and defined the ML model  $f(\mathbf{x}) = \mathbf{p}^T \cdot \mathbf{x}$  from it. Next, this model was evaluated on randomly chosen vectors  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , computing the responses  $y_i = f(\mathbf{x}_i) + \varepsilon_i$  with a random error term on it. This mimics the model  $f$  to have been fitted from the so-constructed training data  $T = (\mathbf{x}_i, y_i)_{i=1}^n$ .

Then, towards a denial of the (correct!) training data set, we randomly sampled a fresh set  $T' = (\mathbf{x}'_i, y'_i)_{i=1}^n$ , in which the values  $y'_i$  were also drawn stochastically independent (of their  $\mathbf{x}'_i$ 's). From this set  $T'$ , we constructed the norm as Theorem 2 prescribes (see Figure 3 in Appendix B.5 for the algorithmic details), and re-fitted the regression model. Plausible deniability is then the expectation of finding approximately the vector  $\mathbf{p}$  again, and indeed, an example execution of this program delivered the following results for a six-dimensional regression model (small enough for a visual inspection):

original vector $\mathbf{p}$	$\mathbf{p}$ as trained from decoy data $T'$
-0.57104	-0.56936
-1.53456	-1.53402
-2.45770	-2.45657
-2.12341	-2.12261
-1.26093	-1.25992
-1.91170	-1.91082

This experiment is repeatable (with comparably good results) using our implementation<sup>3</sup> of the construction behind Theorem 2 in GNU Octave (version 5.2.0) [7], with the `optim` package (version 1.6.0) [16], and for the particular application to a regression model. We stress that the algorithms used to fit the ML model were hereby taken “off the shelf” that `optim` provides, with no modification to the inner code (or its default configuration).

## 6 Related Work

The conflicting interests of available data and data privacy have long been understood. It has been shown that the problem of minimizing information loss under given privacy constraints is NP-hard [17]. An overview on threats and solutions of privacy preserving machine learning is provided in [1] to close the gap between the communities of ML and privacy.

Legal requirements such as the GDPR put limitations on any kind of method that uses personal data, including ML applications. The regulation aims at preventing any discrimination, so critical data such as health data now require protection [2]. Approaches such as the privacy-aware machine learning model provisioning platform AMNESIA [15] make sure that ML models only remember data they are supposed to remember. A new method to preserve privacy for classification methods in distributed systems prevents that data or the learned models are directly revealed [10] and can even be extended to hierarchical distributed systems [9]. The vulnerabilities ML methods induce in software systems can also be analysed based on known attacks [13]. A recent survey on privacy-preserving ML is given in [11], showing that the majority of new approaches focus on specific domains. In social networks, systems are developed that decide (semi-)automatically whether to share information with others [3]. Frameworks for privacy-preserving methods in healthcare are also in development [8]. Classification protocols that ensure confidentiality of both data and classifier are described in [5] and implemented by modification of existing protocols. In 2017, Google presented a protocol that enables deep learning from user data without learning about the individual user [4]. An algorithm for privacy-preserving logistic regression was designed to address the trade-off between privacy and learnability and to learn from private databases [6].

---

<sup>3</sup>code will be released if this paper receives positive reviews

## 7 Conclusions

### 7.1 Suspicion by “non-standard” error metrics

Obviously, it may be suspicious if the norm used for the training is not released a priori as part of the description of the ML model, and our proposed mechanism of deniability works only if the norm used for the training is kept secret initially. Furthermore, the honest creator of the model cannot later come out with a strangely crafted norm to claim having done the training with this, if the more natural choice would have been MAE, RMSE or others. So, to make the denial “work”, the process would require the model creator to initially state that the training will be done with a norm that has a “certain algebraic structure”, namely that which Theorem 2 prescribes. This lets the honest owner of the norm later change the appearance of the norm for a denial, without creating suspicion by coming out with something completely different. Since all vector norms, and hence also all matrix norms are topologically equivalent, such an a priori vote for a certain class of norms is not precluded by theory, and a legitimate design choice up to the model trainer.

### 7.2 Accounting for Partial Knowledge

If the attacker has partial knowledge of the training data, say, a few columns / variables are known, but not all of them, the situation with plausible deniability is unchanged: the denying party can simply include this knowledge in the decoy training data (as this can be chosen freely anyway), and construct the norm from the remaining variables. This even works when the attacker knows *all* variables in the training records  $\mathbf{x}_i$ , in which case the resulting responses  $y_i$  are uniquely recoverable by a mere evaluation of the function  $f$ . This is the trivial case of recovery, against which no countermeasure can be given. However, if there is at least some uncertainty about a variable in the training data, and the model is “sufficiently dependent” on this unknown inputs, then plausible deniability becomes applicable again.

Overall, the finding in this work is that *privacy by non-recoverability* essentially holds without much ado, provided that there is lot more data used for the training than the model can embody via its parameters. Additional precautions for plausible deniability are only required by announcing the error metric prior to any training, or as part of the description of the model upon its release.

The important point here is *not* that the training on a suitably crafted norm is algorithmically feasible, but instead that *it is possible*. While we do not claim the norm from Theorem 2 to lend itself to an efficient optimization in high-dimensional cases (such as neural networks), but the existence assertion made by the theorem may already be enough, since it is arguable that one has taken the decoy data and went through very lengthy and time-consuming training to have produced the model in discussion.

The lesson learned here to escape the plausible deniability issue is to go for maximum transparency of the learning process, which includes in particular an

*a priori* and *publicly documented* specification of the error metric and training algorithm before deniability arguments are made. In this way, one cannot later silently change the error metric towards consistency with faked training data.

## Acknowledgments

This work was supported by the research Project ODYSSEUS (“Simulation und Analyse kritischer Netzwerk Infrastrukturen in Städten”) funded by the Austrian Research Promotion Agency under Grant No. 873539.

## References

- [1] Mohammad Al-Rubaie and J. Morris Chang. Privacy-preserving machine learning: Threats and solutions. 17(2):49–58, 2019.
- [2] C.-A. Azencott. Machine learning and genomics: precision medicine versus patient privacy. 376(2128):20170350, 2018.
- [3] Igor Bilogrevic, Kévin Huguenin, Berker Agir, Murtuza Jadliwala, Maria Gazaki, and Jean-Pierre Hubaux. A machine-learning based approach to privacy-aware information-sharing in mobile social networks. 25:125–142, 2016.
- [4] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191. ACM, 2017.
- [5] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. Machine learning classification over encrypted data. In *Proceedings 2015 Network and Distributed System Security Symposium*. Internet Society, 2015.
- [6] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21, pages 289–296. Curran Associates, Inc., 2009.
- [7] John W. Eaton, David Bateman, Søren Hauberg, and Rik Wehbring. *GNU Octave version 5.2.0 manual: a high-level interactive language for numerical computations*, 2020.
- [8] Kyle Fritchman, Keerthanaa Saminathan, Rafael Dowsley, Tyler Hughes, Martine De Cock, Anderson Nascimento, and Ankur Teredesai. Privacy-preserving scoring of tree ensembles : a novel framework for {AI} in health-care. pages 2413–2422. IEEE, 2018.

- [9] Qi Jia, Linke Guo, Yuguang Fang, and Guirong Wang. Efficient privacy-preserving machine learning in hierarchical distributed system. 6(4):599–612, 2019.
- [10] Qi Jia, Linke Guo, Zhanpeng Jin, and Yuguang Fang. Preserving model privacy for machine learning in distributed systems. 29(8):1808–1822, 2018.
- [11] Liu Junxu and Meng Xiaofeng. Survey on privacy-preserving machine learning. 57(2):346, 2020. Publisher: Journal of Computer Research and Development.
- [12] Keras Team. Keras documentation: Losses, 2020. <https://keras.io/api/losses/>.
- [13] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman. SoK: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414. IEEE, 2018.
- [14] Stefan Schauer, Sandra König, Thomas Schaberreiter, Stefan Rass, Klaus Steinnocher, and Gerald Quirchmayr. Cross-Domain Risk Analysis to Strengthen City Resilience: the ODYSSEUS Approach. In *A.L. Hughes, F. McNeill and C. Zobel (eds.): ISCRAM 2020 Conference Proceedings - 17th International Conference on Information Systems for Crisis Response and Management*, pages 652–662. ISCRAM Association, 2020.
- [15] Christoph Stach, Corinna Giebler, Manuela Wagner, Christian Weber, and Bernhard Mitschang. {AMNESIA}: A technical solution towards {GDPR}-compliant machine learning. volume Proceedings of the 6th International Conference on Information Systems Security and Privacy, pages 21–32, 2020.
- [16] Olaf Till. The 'optim' package, 2019.
- [17] S.A. Vinterbo. Privacy: a machine learning view. 16(8):939–948, 2004.
- [18] Wolfgang Walter. *Analysis 2*. Grundwissen Mathematik. Springer, Berlin, 4., durchges. und erg. aufl edition, 1995. OCLC: 263611766.

## A Error Measures from Topological Norms

A norm on  $\mathbb{R}^n$  is a mapping  $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$  with the following properties:

1. positive definiteness:  $\|\mathbf{x}\| \geq 0$  for all  $\mathbf{x}$ , with  $\|\mathbf{x}\| = 0$  if and only if  $\mathbf{x} = 0$ .
2. homogeneity:  $\|\lambda \cdot \mathbf{x}\| = |\lambda| \cdot \|\mathbf{x}\|$  for all  $\lambda \in \mathbb{R}$ .
3. triangle inequality:  $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$  for all  $\mathbf{x}, \mathbf{y}$ .

If one allows  $\|\mathbf{x}\| = 0$  for some  $\mathbf{x} \neq 0$ , then we call  $\|\cdot\|$  a *semi-norm*. Every norm induces a *metric*  $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ , or a *pseudometric* if we use a semi-norm.

At least the following popular choices for error measures are directly expressible via norms. For the description, let us put  $\hat{y}_i := f(\mathbf{x}_i, p)$  be the ML model's estimate on the training data  $(\mathbf{x}_i, y_i)$  for a total of  $i = 1, 2, \dots, n$  training samples. For abbreviation, put  $\mathbf{y} = (y_1, \dots, y_n)$ ,  $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n) \in \mathbb{R}^n$ , and recall that a general  $p$ -norm for  $p \geq 1$  on  $\mathbb{R}^n$  is defined by

$$\|\mathbf{y}\|_p = \left[ \sum_{i=1}^n |y_i|^p \right]^{\frac{1}{p}},$$

with the practically most important special cases of the 1-norm  $\|\mathbf{y}\|_1 = \sum_{i=1}^n |y_i|$ , Euclidian norm  $\|\mathbf{y}\|_2 = \sqrt{y_1^2 + y_2^2 + \dots + y_n^2}$ , and maximum-norm  $\|\mathbf{y}\|_\infty = \max_i |y_i|$ .

1. Mean squared error

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \quad (5)$$

2. Root mean squared error

$$RMSE = \sqrt{MSE} = \frac{1}{\sqrt{n}} \|\mathbf{y} - \hat{\mathbf{y}}\|_2 \quad (6)$$

3. Mean absolute error

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| = \frac{1}{n} \cdot \|\mathbf{y} - \hat{\mathbf{y}}\|_1 \quad (7)$$

We will not go into discussions about pros and cons of these choices (or alternatives thereto), beyond remarking that the squared errors can be easier to handle for their differentiability properties. The MAE is on the contrary more robust against outliers, which the (R)MSE penalize more, so that the fitting is more sensitive to training data that has not been cleaned from outliers before.

Defining an error metric from a norm as yet another appeal, since (topologically) all norms over finite-dimensional real vector-spaces are equivalent. Since we will make implicit use of that in the following, we state this well known result for vector-norms, whose canonical version for matrix-norms holds likewise:

**Theorem 4** (see, e.g., [18, p.17]). *Let any two norms  $\|\cdot\|'$  and  $\|\cdot\|''$  on  $\mathbb{R}^n$  be given. Then there are constants  $\alpha, \beta > 0$  such that*

$$\alpha \cdot \|\mathbf{x}\|' \leq \|\mathbf{x}\|'' \leq \beta \cdot \|\mathbf{x}\|'.$$

By symmetry, this is an equivalence relation on the set of norms on  $\mathbb{R}^n$ , and topologically speaking, they all induce the same topology. For optimization, it means that once the distance  $\|\mathbf{x}_i - \mathbf{y}\| \rightarrow 0$  as  $i \rightarrow \infty$  for a point sequence  $\mathbf{x}_i$  towards approximating a (fixed) target vector  $\mathbf{y}$ , this convergence would occur in the same way (though not necessarily at the same speed) in *every other norm* on  $\mathbb{R}^n$ .

Practically, this means that fitting a ML model to a training data set by optimizing the norm of the error vector as in (1), will eventually lead to results within a spherical neighborhood (ball) whose radius changes only by a constant factor upon switching from  $\|\cdot\|'$  to  $\|\cdot\|''$ . Moreover, if an approximation with zero error is possible, both norms will admit finding this optimum point.

## A.1 Pseudometrics for the Training

Picking up on the outline started in Section 4.1, a flexible construction for a norm is  $\|\mathbf{x}\|_A := \sqrt{\mathbf{x}^\top \cdot \mathbf{A} \cdot \mathbf{x}}$  with any positive definite matrix  $\mathbf{A}$ . If  $\mathbf{A}$  is not positive definite, we can still get a semi-norm as  $\mathbf{x} \mapsto \|\mathbf{A} \cdot \mathbf{x}\|$ , with only the property  $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$  being violated in case that  $\mathbf{A}$  has a nontrivial nullspace  $N(\mathbf{A}) = \{\mathbf{x} : \mathbf{A} \cdot \mathbf{x} = \mathbf{0}\}$ , where by nontrivial we mean  $N(\mathbf{A}) \neq \{\mathbf{0}\}$ .

We will proceed by constructing a semi-norm that vanishes only for the given error vector  $e(\mathbf{p}^*)$  or scalar multiples thereof, under the chosen parameter  $\mathbf{p}^*$ . Let us call this particular matrix  $\mathbf{B}$ , whose existence and construction is not difficult to describe:

**Lemma 2.** *Let  $\mathbf{e} \in \mathbb{R}^n$  be a vector, then there exists a matrix  $\mathbf{B}$  having the nullspace  $N(\mathbf{B}) = \text{span}\{\mathbf{e}\}$ . Geometrically, this matrix is a projection on a  $(n-1)$ -dimensional subspace of  $\mathbb{R}^n$ , corresponding to the orthogonal complement of  $\text{span}\{\mathbf{e}\}$  within  $\mathbb{R}^n$ .*

*Proof.* Compute a Singular Value Decomposition (SVD)  $\mathbf{e} = \mathbf{U} \cdot \Sigma \cdot \mathbf{V}$  for the error vector  $\mathbf{e}$ , and construct  $\mathbf{B}$  with the same rows taken from  $\mathbf{U}^\top$  that correspond to all-zero rows (i.e., zero diagonal elements) in  $\Sigma$ . The nullspace and geometric properties then directly follow from this construction.  $\square$

Using the matrix  $\mathbf{B}$ , we can define the semi-norm

$$b(\mathbf{x}) := \|\mathbf{B} \cdot \mathbf{x}\| \tag{8}$$

in which  $\|\cdot\|$  is an arbitrary (full) norm on  $\mathbb{R}^n$ . This is a well-defined semi-norm, with the properties that

- $b(e(\mathbf{p}^*)) = 0$ ,
- and  $b(\mathbf{x}) > 0$  whenever  $\mathbf{x} \notin \text{span}\{e(\mathbf{p}^*)\}$ .

The function  $b$  induces a pseudometric on  $\mathbb{R}^n$ , as lacking only the identity of indiscernible elements  $d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$ , but still satisfying  $b(\mathbf{x}) \geq 0$  for all  $\mathbf{x}$ , so that  $\mathbf{x} = \mathbf{p}^*$  is already an optimum. For later reference, let us capture the matrix  $\mathbf{B}$  more explicitly:

The vector  $\mathbf{e}$  in Lemma 2 will be our error vector  $e(\mathbf{p})$  for the parameterization  $\mathbf{p}$ , and the subspace that  $\mathbf{B}$  projects on will be called  $V$  throughout all other proofs appearing hereafter.

Note that, in principle, we could directly use this pseudometric to train our function  $f$  towards taking a minimum error for the parameter  $\mathbf{p}^*$ . The necessary assumption is that upon a change from  $\mathbf{p}^*$  to another  $\mathbf{p} \neq \mathbf{p}^*$ , we would leave the nullspace of  $\mathbf{B}$ , thus making the function  $b$  take on strictly positive values.

## B Proofs

### B.1 Proof of Theorem 1

This is a simple information-theoretic argument: call  $X$  the random variable representing the (entirety) of the training data that went into the ML model. Suppose this is a set of  $n$  records containing values that are sampled from a random vector  $Z$  in a stochastically independent manner. Then,  $X$  is a matrix of  $n$  rows, and has the entropy  $H(X) = n \cdot H(Z)$ , where  $H(Z)$  is the entropy of the joint distribution over the attributes in the training data record. From here on, let all logarithms have base 2.

The trained model is, from the adversary’s perspective, a sample of another random variable  $Y$ , representing the collection of parameters that define the model. The recovery problem is the unique reconstruction of  $X$ , given  $Y$ , and, information-theoretically speaking, solvable if and only if  $H(X|Y) = 0$ . First, note that  $H(X|Y) = H(X, Y) - H(Y)$ , and that  $H(X, Y) \geq H(X)$ , giving  $H(X|Y) \geq H(X) - H(Y)$ . Similarly, the information extractable from the trained model cannot be more than the shortest encoding of the model itself. So, suppose that the model  $f$ , as a realization of the random variable  $Y$ , comes with a string description of length at least  $K(Y) = \min\{\ell \in \mathbb{N} : f \sim Y \text{ has an } \ell \text{ bit string representation}\}$  bits. Then, the uncertainty reduction by  $-H(Y)$  cannot exceed the bit count to represent  $f$ , hence  $H(X) - H(Y) \geq H(X) - K(Y)$ . The maximum additional knowledge of  $K(Y)$  bits, contributed by  $Y$ , is increasing in  $d$ , since the parameters at some point must be encoded within the string representation of  $f$ . Using this and the fact that  $H(X) = n \cdot H(Z)$ , with  $H(Z)$  being constant (and determined by the uncertainty in the attributes of the data that were used for training), we find

$$H(X|Y) = H(X, Y) - H(Y) \geq H(X) - H(Y) \geq n \cdot H(Z) - K(Y) > 0, \quad (9)$$

if the number  $n$  of training records grows sufficiently large over the number  $d$  of parameters in the model. Once  $H(X|Y) > 0$ , we have no hope for a unique recovery of the training data from a model. To be precise, it means that the distribution is non-degenerate, meaning that there is at least another *possibility*

(i.e., element in the support) to appear with nonzero probability. This completes the proof of Theorem 1.

Theorem 1 *does not* imply any claim about the possibility or impossibility to single out a most plausible among the possible solutions. This would be more likely or easy, the smaller the conditional or residual entropy comes out, so making  $n$  large over  $d$  is practically desirable. Quantifying the chances of guessing is another story, calling for conditional min-entropies here, and left as a direction of future research.

While this already positively answers the question of *privacy* of the data embodied in a ML model, this does not rule out a “lucky guess” of the correct training data. This guess becomes more likely, the smaller the residual uncertainty  $H(X|Y)$  is.

Irrespective of the residual uncertainty, the stronger possibility of denying a lucky guess *even if it is correct* is what plausible deniability is about.

## B.2 Proof of Lemma 1

Let  $e(\mathbf{p}^*)$  be a vector spanning the nullspace of a matrix  $\mathbf{B}$ , and let  $b$  be defined by (8). Since  $f$  is differentiable, we can locally write the error term as

$$e(\mathbf{p}) = e(\mathbf{p}^*) + (J_p(f))(\mathbf{p}^*) \cdot (\mathbf{p} - \mathbf{p}^*) + o(\|\mathbf{p} - \mathbf{p}^*\|)$$

for all  $\mathbf{p}$  in some neighborhood of  $\mathbf{p}^*$ . Abbreviating our notation by writing  $\mathbf{M} := (J_p(f))(\mathbf{p}^*)$ , i.e., calling  $\mathbf{M}$  the Jacobian of  $f$  evaluated at  $\mathbf{p}^*$ , and rearranging terms, we get

$$e(\mathbf{p}) - e(\mathbf{p}^*) = \mathbf{M} \cdot (\mathbf{p} - \mathbf{p}^*) + o(\|\mathbf{p} - \mathbf{p}^*\|). \quad (10)$$

Towards a contradiction, assume  $e(\mathbf{p}) \in N(\mathbf{B})$ . By construction, we have  $e(\mathbf{p}^*) \in N(\mathbf{B})$ , so the difference  $e(\mathbf{p}) - e(\mathbf{p}^*)$  of the two is also in  $N(\mathbf{B})$ . Likewise must thus be the right hand of (10) in  $N(\mathbf{B})$ , and we can find a sequence  $(\mathbf{p}_i)_{i \in \mathbb{N}}$  inside  $N(\mathbf{B})$  that satisfies (10). Because  $N(\mathbf{B}) = \text{span}\{e(\mathbf{p}^*)\}$ , we can write this sequence as  $\mathbf{p}_i := \mathbf{p}^* + h_i \cdot \mathbf{v}$ , using another null-sequence  $(h_i)_{i \in \mathbb{N}}$  of values in  $\mathbb{R}$  and the unit vector  $\mathbf{v} := e(\mathbf{p}^*)/\|e(\mathbf{p}^*)\|$  (the norm is herein the one from (10), and has nothing to do with the one asserted by Theorem 2). Since the sequence  $h_i \rightarrow 0$  is arbitrary (as is the sequence  $\mathbf{p}_i$ ), let us just write  $h \rightarrow 0$  to define the sequence of points in  $N(\mathbf{B})$ .

This lets us rewrite (10) as

$$e(\mathbf{p}^* + h \cdot \mathbf{v}) - e(\mathbf{p}^*) = \mathbf{M} \cdot h \cdot \mathbf{v} + o(h),$$

which we can divide by  $h > 0$  to get the quotient

$$\frac{e(\mathbf{p}^* + h \cdot \mathbf{v}) - e(\mathbf{p}^*)}{h} = \mathbf{M} \cdot \mathbf{v} + \frac{o(h)}{h}.$$

Therein, we have  $\frac{o(h)}{h} \rightarrow 0$  as  $h \rightarrow 0$  by the definition of the small-o, and on the left hand side, we get the directional derivative along  $\mathbf{v}$  by taking  $h \rightarrow 0$ , since  $f$  was assumed to be totally differentiable.

Before, we noted the left side of (10) to be in  $N(\mathbf{B})$ , and since subspaces are topologically closed, the limit, i.e., the directional derivative must also be in  $N(\mathbf{B})$ . Accordingly, this puts the right side  $\mathbf{M} \cdot \mathbf{v} \in N(\mathbf{B})$ , implying that there is some number  $\lambda \in \mathbb{R}$  so that  $\mathbf{M} \cdot \mathbf{v} = \lambda \cdot e(\mathbf{p}^*)$ . But this means that  $e(\mathbf{p}^*)$  must be in the column space of  $\mathbf{M}$ , which contradicts our hypothesis (4) on the rank and refutes the assumption that  $e(\mathbf{p})$  can be in  $N(\mathbf{B})$ .

We thus have  $e(\mathbf{p}) \notin N(\mathbf{B})$  in a neighborhood of  $\mathbf{p}^*$ , but  $e(\mathbf{p}^*) \in N(\mathbf{B})$ . Now, using the semi-norm  $b(\mathbf{x}) = \|\mathbf{B} \cdot \mathbf{x}\|$ , we see that  $\|e(\mathbf{p}^*)\| = 0$ , while  $\|e(\mathbf{p})\| > 0$ , so  $\mathbf{p}^*$  is locally optimal under this semi-norm.

### B.3 Proof of Theorem 2

The norm as claimed to exist above will be

$$\|\mathbf{x}\| := \|\mathbf{x}\|_e + b(\mathbf{x}), \quad (11)$$

with  $b$  as we had so far, and another norm  $\|\cdot\|_e$ , to be designed later (the subscript  $\mathbf{e}$  to the norm is hereafter a reminder that this norm will depend on the error vector  $\mathbf{e}$ ). Intuitively, one may think of  $b$  as a ‘‘penalty term’’ to increase the norm upon any deviation from the desired error vector (hence making this point a minimum).

At  $\mathbf{p}^*$ , we have

$$\|e(\mathbf{p}^*)\| = \|e(\mathbf{p}^*)\|_e + \underbrace{b(e(\mathbf{p}^*))}_{=0} = \|e(\mathbf{p}^*)\|_e,$$

by our choice of the semi-norm  $b$ . Our goal is showing that

$$\|e(\mathbf{p}^*)\| \leq \|e(\mathbf{p})\|. \quad (12)$$

From the triangle inequality that  $\|\cdot\|_e$  must satisfy, we get for any  $\mathbf{p} \neq \mathbf{p}^*$ ,  $\|e(\mathbf{p}^*)\|_e = \|e(\mathbf{p}^*) - e(\mathbf{p}) + e(\mathbf{p})\|_e \leq \|e(\mathbf{p})\|_e + \|e(\mathbf{p}^*) - e(\mathbf{p})\|_e$ , and by rearranging terms, we find  $\|e(\mathbf{p})\|_e \geq \|e(\mathbf{p}^*)\|_e - \|e(\mathbf{p}^*) - e(\mathbf{p})\|_e$ . Substituting this into (11), we get

$$\begin{aligned} \|e(\mathbf{p})\| &= \|e(\mathbf{p})\|_e + b(e(\mathbf{p})) \\ &\geq \|e(\mathbf{p}^*)\|_e - \|e(\mathbf{p}^*) - e(\mathbf{p})\|_e + b(e(\mathbf{p})). \end{aligned} \quad (13)$$

To prove (12), it suffices to construct a norm  $\|\cdot\|_e$  that satisfies

$$\|e(\mathbf{p}^*) - e(\mathbf{p})\|_e \leq b(e(\mathbf{p})), \quad (14)$$

for all  $\mathbf{p}$  for which  $e(\mathbf{p})$  is *outside* of  $N(\mathbf{B})$  (otherwise, for  $e(\mathbf{p}) \in N(\mathbf{B})$  distinct from  $\mathbf{p}^*$  we would have  $\|e(\mathbf{p}^*) - e(\mathbf{p})\| > 0$  but  $b(e(\mathbf{p})) = 0$ , invalidating (14)). The assurance that  $e(\mathbf{p}) \notin N(\mathbf{B})$  is hereby implied by the hypothesis and arguments of Lemma 1, which we included in the theorem’s hypothesis and hence not repeat here.

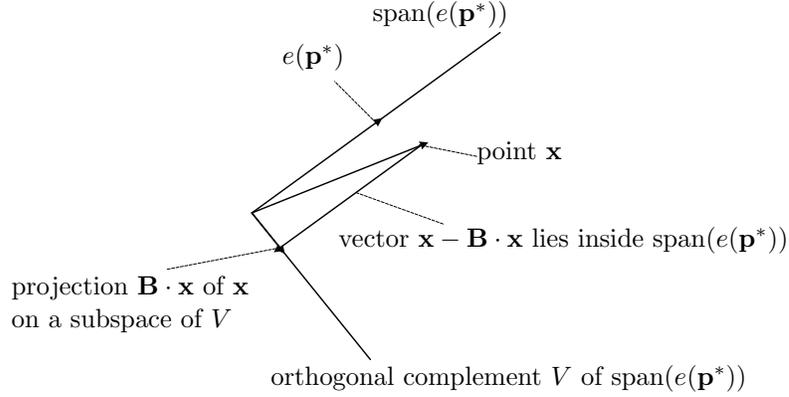


Figure 1: Illustration of the projection norm  $\|\cdot\|_V$

So we can continue (13) as

$$\begin{aligned} \|e(\mathbf{p})\| &\geq \|e(\mathbf{p}^*)\|_e - \underbrace{\|e(\mathbf{p}^*) - e(\mathbf{p})\|_e}_{\geq 0} + b(e(\mathbf{p})) \\ &\geq \|e(\mathbf{p}^*)\|_e. \end{aligned}$$

With that accomplished, and recalling that  $b$  was constructed towards  $b(e(\mathbf{p}^*)) = 0$ , we would find  $\|e(\mathbf{p})\| \geq \|e(\mathbf{p}^*)\|_e = \|e(\mathbf{p}^*)\|_e + b(e(\mathbf{p}^*)) = \|e(\mathbf{p}^*)\|$ , which is exactly our goal (12).

Thus, we are left with the task of finding a norm  $\|\cdot\|_e$  that satisfies (14). To this end, recall that the semi-norm  $b$  becomes a (full) norm on the factor space  $\mathbb{R}^n/\sim$ , modulo the equivalence relation  $\mathbf{x} \sim \mathbf{y} \iff (\mathbf{x} - \mathbf{y}) \in N(\mathbf{B})$ . By the dimension formula, we have  $\dim(\mathbb{R}^n) = \dim(\mathbb{R}^n/\sim) + \dim(N(\mathbf{B}))$ , and since  $\dim(N(\mathbf{B})) = 1$ , we find  $\dim(\mathbb{R}^n/\sim) = n - 1$ . Since the factor space is a vector space over the reals, it is isomorphic to the  $(n-1)$ -dimensional orthogonal complement  $V := N(\mathbf{B})^\perp \subset \mathbb{R}^n$  of  $N(\mathbf{B}) \simeq \mathbb{R}^1$ . On  $V$ , we can define a norm, e.g.  $\|\cdot\|_2$ . By Lemma 2,  $\text{proj}_V = \mathbf{B}$  is the projection of a vector onto  $V$ , then (taking the same norm as in (8)),

$$\|\mathbf{x}\|_V := \frac{1}{2} \|\text{proj}_V(\mathbf{x})\| = \frac{1}{2} \cdot b(\mathbf{x})$$

is a semi-norm on  $\mathbb{R}^n$ . This semi-norm trivially satisfies  $\|\mathbf{x}\|_V \leq \frac{1}{2}b(\mathbf{x})$  for all  $x \in \mathbb{R}^n$ . Figure 1 provides an illustration.

Now, for an intermediate wrap-up,  $\|\cdot\|_V$  is a semi-norm obeying the desired bounds for all vectors, especially those in the orthogonal complement of  $N(\mathbf{B})$ , as desired. We now need to extend it to a full norm on the entire space  $\mathbb{R}^n$  using the following idea: the sum of two semi-norms over the same vector space is again a semi-norm and it is a full norm, if and only if the intersection of kernels of the two semi-norms is exactly  $\{0\}$ . So we can construct a full norm by adding another semi-norm, that is a full norm on a 1-dimensional space (isomorphic to  $N(\mathbf{B})$ ), which retains (14) on  $\mathbb{R}^n \setminus N(\mathbf{B})$ .

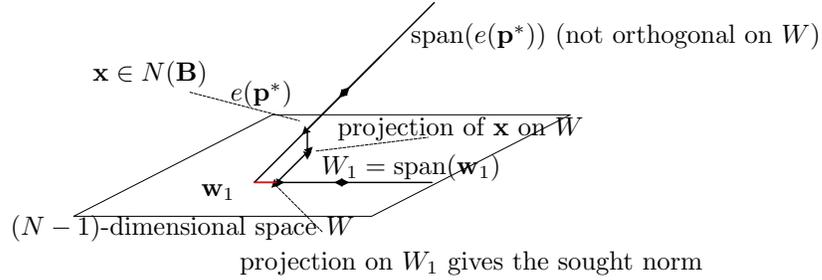


Figure 2: Illustration of the construction of  $\|\cdot\|_W$

The idea is to project a vector in  $N(\mathbf{B})$  to the exterior of  $N(\mathbf{B})$  and take the norm of the projection there. To materialize this plan, let  $\{\mathbf{v}_1, \dots, \mathbf{v}_{n-1}\}$  be an orthonormal basis of  $V$ . Furthermore, pick any vector  $\mathbf{w}_1 \in \mathbb{R}^n$  with two properties: (1) it is not a scalar multiple of  $e(\mathbf{p}^*)$ , and (2) it is linearly independent of all  $\{\mathbf{v}_1, \dots, \mathbf{v}_{n-1}\}$ . In other words, we want both sets  $\{\mathbf{w}_1, e(\mathbf{p}^*)\}$  and  $\{\mathbf{w}_1, \mathbf{v}_1, \dots, \mathbf{v}_{n-1}\}$  to be linearly independent<sup>4</sup>. An easy choice for  $\mathbf{w}_1$  is to rotate the vector  $e(\mathbf{p}^*)$  enough to become linearly independent of it, but not far enough to become lying in the orthogonal complement. Figure 2 graphically sketches the idea formalized now.

Call  $W_1 := \text{span}\{\mathbf{w}_1\}$  the linear hull of  $\mathbf{w}_1$ , and pick another  $n-2$  pairwise orthogonal vectors  $\mathbf{w}_2, \dots, \mathbf{w}_{n-1}$ , whose entirety spans the space  $W_{n-2}^\perp = \text{span}\{\mathbf{w}_2, \dots, \mathbf{w}_{n-1}\}$  (the subscript and superscript are here serving as reminders about the dimensionality and the orthogonality of this space relative to  $W_1$ ). Clearly, we have

$$\mathbb{R}^n \simeq W_1 \oplus W_{n-2}^\perp \oplus \underbrace{\text{span}\{e(\mathbf{p}^*)\}}_{=N(\mathbf{B})}.$$

Now, let any  $\mathbf{x} \in N(\mathbf{B})$  be given. We can project  $\mathbf{x}$  on the spaces  $W_1$  and  $W_{n-2}^\perp$ . Since the space  $W := W_1 \oplus W_{n-2}^\perp$  is also over  $\mathbb{R}$  and has dimension  $n-1$ , we have the isomophy

$$W_1 \oplus W_{n-2}^\perp \simeq \mathbb{R}^n / \sim,$$

so that the function  $b$  is again a norm on  $W$ . Now, let us take the 1-norm (an arbitrary choice here) to define another norm on  $W$  as

$$\|\mathbf{x}\|_W := \|\text{proj}_{W_1}(\mathbf{x})\|_1 + \|\text{proj}_{W_{n-2}^\perp}(\mathbf{x})\|_1.$$

Since all norms over  $\mathbb{R}^d$  are equivalent by Theorem 4 (for all  $d$ , especially  $d = n$  or  $d = n-1$ ), there is a constant  $\alpha > 0$  such that  $\alpha \cdot \|\mathbf{x}\|_W < b(\mathbf{x})$ . By definition

<sup>4</sup>note that  $\mathbf{w}_1$  is in any case *non-orthogonal* to  $e(\mathbf{p}^*)$ , which assures that the projection of any element in  $\text{span}(e(\mathbf{p}^*))$  onto the subspace spanned by  $\mathbf{w}_1$  is nontrivial; if  $w_1$  were orthogonal to  $e(\mathbf{p}^*)$ , it would necessarily be a scalar multiple of some vector among  $\mathbf{v}_1, \dots, \mathbf{v}_{n-1}$ , in which case it cannot be linearly independent of them, as we required too.

of  $\|\mathbf{x}\|_W$ , we also have  $\alpha \cdot \|\mathbf{x}\|_1 \leq \alpha \cdot \|\mathbf{x}\|_W \leq b(\mathbf{x})$ . This lets us define a norm on the subspace  $W_1 \subset W$  as

$$\|\mathbf{x}\|_{W_1} := \frac{\alpha}{2} \cdot \|\text{proj}_{W_1}(\mathbf{x})\|_1,$$

which satisfies the desired inequality  $\|\mathbf{x}\|_{W_1} \leq \frac{1}{2}b(\mathbf{x})$ .

Now, let us put together the pieces: define the sought norm  $\|\cdot\|_e$  as

$$\|\mathbf{x}\|_e := \|\mathbf{x}\|_V + \|\mathbf{x}\|_{W_1} \leq \frac{1}{2}b(\mathbf{x}) + \frac{1}{2}b(\mathbf{x}) = b(\mathbf{x}),$$

where the inequality is only demanded to hold for  $\mathbf{x} \notin N(\mathbf{B})$ . Observe that this is indeed a (full) norm on  $\mathbb{R}^n$ , since:

- if  $\mathbf{x} = 0$ , then  $\|\mathbf{x}\|_V = \|\mathbf{x}\|_{W_1} = 0$
- if  $\mathbf{x} \neq 0$  and  $\mathbf{x} \notin N(\mathbf{B})$ , then there is a nonzero projection  $\mathbf{x}_V$  on the orthogonal complement of  $N(\mathbf{B})$ , on which  $\|\mathbf{x}_V\|_V > 0$ , and hence  $\|\mathbf{x}\|_e > 0$ . Likewise, if  $\mathbf{x} \neq 0$  and  $\mathbf{x} \in N(\mathbf{B})$  ( $\iff \mathbf{x} \notin N(\mathbf{B})^\perp$ ), then there is a nonzero projection on  $W_1$ , making the other part of the norm  $> 0$ .
- Homogeneity and the triangle inequality hold by construction and are obvious to check.

Substituting this into (11), we finally get

$$\begin{aligned} \|e(\mathbf{p}) - e(\mathbf{p}^*)\|_e &\leq \|e(\mathbf{p})\|_e + \|e(\mathbf{p}^*)\|_e \\ &\leq b(e(\mathbf{p})) + b(e(\mathbf{p}^*)) \\ &= b(e(\mathbf{p})), \end{aligned}$$

thus satisfying (14), and yielding the final norm from (11) as

$$\|\mathbf{x}\| = \frac{3}{2}b(\mathbf{x}) + \|\mathbf{x}\|_{W_1}.$$

This completes the proof of Theorem 2. So far, this argument is not entirely constructive, but can be made so by reconsidering the construction in a little more detail, to which we devote the next paragraph.

### B.3.1 Computing the Projections and the Value $\alpha$

As stated, the proof of Theorem 2 is not constructive at the point where it claims the *existence* of the constant  $\alpha$  to make  $\alpha \cdot \|\mathbf{x}\|_W \leq b(\mathbf{x})$ . Working out a suitable constant  $\alpha$  explicitly is not difficult: every  $\mathbf{x} \in W_1 = \text{span}(\mathbf{w}_1)$  takes the form  $\mathbf{x} = \lambda \cdot \mathbf{w}_1$  for some  $\lambda \in \mathbb{R}$ , and we can, w.l.o.g., assume  $\mathbf{w}_1$  to have unit length w.r.t.  $\|\cdot\|_1$  on  $\mathbb{R}^n$ . Then,  $\|\text{proj}_{W_1}(\mathbf{x})\|_1 = |\lambda|$ , and  $b(\mathbf{x}) = b(\lambda \cdot \mathbf{w}_1) = |\lambda| \cdot b(\mathbf{w}_1)$ . So, it suffices to choose any  $\alpha \in (0, b(\mathbf{w}_1))$  to accomplish  $\alpha \cdot \|\text{proj}_{W_1}(\mathbf{x})\|_1 < b(\mathbf{x})$  for  $\mathbf{x} \in W_1$ , as desired. If  $\mathbf{x} \in \mathbb{R}^n$  is arbitrary, its projection is directly obtained

Input: Let  $\mathbf{e} = f(\mathbf{x}, \mathbf{p}^*) - \mathbf{y} \in \mathbb{R}^n$  be the error vector of the ML model  $f$  using the parameters  $\mathbf{p}^*$ , on the training/validation data  $(\mathbf{x}, \mathbf{y})$ .

Output: The norm that Theorem 2 speaks about.

1. Compute  $\mathbf{B}$  as shown in the proof of Lemma 2.
2. Pick a random vector  $\mathbf{w}_1 \in \mathbb{R}^n$  with  $\|\mathbf{w}_1\|_1 = 1$ . With probability 1, this will deliver a vector that is linearly independent of all rows in  $\mathbf{B}$ , and also not a scalar multiple of  $\mathbf{e}$  (but this should nonetheless be checked by checking if the  $\mathbf{w}_1 \neq \mathbf{B} \cdot \mathbf{w}_1$  is fulfilled. Otherwise sample another vector  $\mathbf{w}_1$  and repeat). The probability assurance follows from the fact that any lower-dimensional subspace of  $\mathbb{R}^n$  has zero Lebesgue measure in  $\mathbb{R}^n$ .
3. Put  $\alpha := \frac{1}{2} \cdot b(\mathbf{w}_1)$ , with the function  $b$  defined from the matrix  $\mathbf{B}$  via (8).
4. Given any vector  $\mathbf{x} \in \mathbb{R}^n$ , compute the norm  $\|\mathbf{x}\|_e = \|\mathbf{x}\|_V + \|\mathbf{x}\|_{W_1}$ , utilizing that  $\|\mathbf{x}\|_V = \|\text{proj}_V(\mathbf{x})\| := \frac{1}{2} \cdot b(\mathbf{x})$ , and  $\|\mathbf{x}\|_{W_1} = \frac{\alpha}{2} \cdot |\mathbf{x}^\top \mathbf{w}_1|$ , to obtain  $\|\mathbf{x}\|$  from (11) as

$$\|\mathbf{x}\| = \frac{3}{2}b(\mathbf{x}) + \frac{\alpha}{2} \cdot |\mathbf{x}^\top \cdot \mathbf{w}_1| \quad (15)$$

Figure 3: Computation of the norm asserted by Theorem 2

from the standard scalar product  $\text{proj}_{W_1}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w}_1 \rangle \cdot \mathbf{w}_1 = (\mathbf{x}^\top \cdot \mathbf{w}_1) \cdot \mathbf{w}_1$  with  $\lambda = \langle \mathbf{x}, \mathbf{w}_1 \rangle$ .

Computing the projection of a vector  $\mathbf{x} \in \mathbb{R}^n$  on the subspace  $V$  is simply the mapping  $\mathbf{x} \mapsto \mathbf{B} \cdot \mathbf{x}$ , if  $\mathbf{B}$  is constructed as Lemma 2 prescribes.

Putting together the pieces, given the parameter set  $\mathbf{p}^*$  and the resulting residual error vector  $\mathbf{e}$ , the norm as told by Theorem 2 is explicitly computable along the steps summarized in Figure 3.

## B.4 Proof of Corollary 1

A re-inspection of the proof of Theorem 2 in Section B.3 quickly shows that it nowhere depends on the algebraic structure of the function  $b$  as given by (8), and we only used the fact that  $b$  is a semi-norm. With that in mind, we can investigate special cases:

Define  $b$  as

$$b(\mathbf{x}) := \|\mathbf{B} \cdot \mathbf{x}\|_1, \quad (16)$$

which has the kernel  $N(\mathbf{B})$ , and is also a semi-norm. However, it lets us express the final norm that Theorem 2 concludes with by a more elegant algebraic expression. Upon re-arriving at (15) (see Figure 3) using the function  $b$  as

defined by (16), we can expand towards

$$\|\mathbf{x}\| = \frac{3}{2} \|\mathbf{B}\mathbf{x}\|_1 + \frac{\alpha}{2} |\mathbf{x}^\top \mathbf{w}_1|,$$

and, recalling that adding the right term to the 1-norm on the left is the same as taking the 1-norm on a vector with merely one additional coordinate, we see with a block matrix  $\mathbf{C} = \begin{pmatrix} (3/2)\cdot\mathbf{B} \\ (\alpha/2)\cdot\mathbf{w}_1^\top \end{pmatrix}$

$$\begin{aligned} \|\mathbf{C} \cdot \mathbf{x}\|_1 &= \left\| \begin{pmatrix} \frac{3}{2}\mathbf{B} \cdot \mathbf{x} \\ \frac{\alpha}{2}\mathbf{w}_1^\top \cdot \mathbf{x} \end{pmatrix} \right\|_1 \\ &= \frac{3}{2} \|\mathbf{B}\mathbf{x}\|_1 + \frac{\alpha}{2} |\mathbf{x}^\top \mathbf{w}_1| \\ &= \|\mathbf{x}\|, \end{aligned} \tag{17}$$

so that  $\|\mathbf{e}\| = \|\mathbf{C} \cdot \mathbf{e}\|_1 = n \cdot MAE(\mathbf{C} \cdot \mathbf{e})$  on the error  $\mathbf{e}$ .

This means that the error measured by the norm from Theorem 2 is “just” the *mean absolute error*, except for a linear transformation of the error vector. Contemporary machine learning libraries often provide the possibility to define custom loss functions, such as, e.g., `keras` [12].

## B.5 Proof of Corollary 2

If  $f$  is vector-valued with  $k$  coordinates, we can apply Theorem 2 to each coordinate function  $f_j$  for  $j = 1, \dots, k$  to obtain a vector norm  $\|\cdot\|_{e_j}$  on  $\mathbb{R}^N$  that depends on  $\mathbf{e}_j(\mathbf{p}^*)$  and satisfies

$$\|\mathbf{e}_j(\mathbf{p}^*)\|_{e_j} \leq \|\mathbf{e}_j(\mathbf{p})\|_{e_j} \tag{18}$$

for the parameterization  $\mathbf{p}^*$  that is the same for all  $k$ , and all  $\mathbf{p}$  in a neighborhood of  $\mathbf{p}^*$ . From these vector norms, we can define

$$\|\mathbf{A}\| = \sum_{j=1}^k \|\mathbf{a}_j\|_{e_j}, \tag{19}$$

with  $\mathbf{a}_j$  being the  $j$ -th column in the matrix  $\mathbf{A}$ . This is readily checked to be a matrix-norm, but now works on the multivariate error  $\mathbf{E} = (\mathbf{e}_1(\mathbf{p}), \dots, \mathbf{e}_k(\mathbf{p}))$ . The optimality of  $\mathbf{p}^*$  under this norm then directly follows by summing up (18) over  $j = 1, 2, \dots, k$ . This completes the proof.

The practical evaluation of the norm in the multivariate case thus boils down to an  $k$ -fold evaluation of norms from Theorem 2 using the algorithm from Figure 3, and summing up the results. Since all matrix norms are likewise to Theorem 4 equivalent, the previous remarks on the freedom to choose any matrix norm for fitting the ML model remains valid.

## C Example: Regression Model

Let us first illustrate the application of Theorem 2 on a simple linear regression model. This choice is convenient for both, a closed-form expressibility of objects like the Jacobian, as well as it can be designed with only a few number of parameters for a manual check that the resulting model really comes up almost identical, whether it has been trained with real or decoy data.

The overall experiment went as follows, where we let the data hereafter be *purely artificial* for the mere sake of easy visual inspection during the computations and in particular regarding the results:

1. The overall regression model is given by a function with parameter  $\mathbf{p} = (\beta_0, \beta_1, \dots, \beta_{d-1})$

$$f(\mathbf{x}, \mathbf{p}) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_{d-1} \cdot x_{d-1} + \varepsilon, \quad (20)$$

in which  $\varepsilon$  is a random error term with assumed zero mean. From the model, it is evident that  $d = m + 1$ , so that the input vector  $\mathbf{x} \in \mathbb{R}^m$  has one dimension less than  $\mathbf{p}$ . For the experiment, we took a uniformly random vector  $\mathbf{p} \in [-6, +6]^d$  of reals, to define an incoming model  $f_0$  “at random”. The magnitude  $\pm 6$  is herein an arbitrary choice, to keep the numbers feasibly small for a manual visual inspection later.

2. Equation (20) was then evaluated on a total of  $n = 10$  uniformly random samples  $\mathbf{X}_i \sim \mathcal{U}(\{1, 2, \dots, 8\}^m)$ , adding stochastically independent error terms  $\varepsilon$ , each with an exponential distribution with rate parameter  $\lambda = 5$  (to, say, let the data be inter-arrival times, with an eye back on Example 1). Again, the choice of  $x$ -values in the integer range  $1, \dots, 8$  is arbitrary, and only to keep the numbers small for a visual checkup. This computation delivers the values  $y_i \leftarrow f(\mathbf{x}_i) + \varepsilon$  for  $i = 1, 2, \dots, 10$ , which, together with the  $\mathbf{x}_i$  form the *training data*.

3. Next, we “forget” about the underlying model (that we know here) and fit a regression model of the same structure, given only the training data. Since this data originally came out of a regression model, this lets us expect a quite good fit, and an approximate re-discovery of the same parameter vector  $\hat{\mathbf{p}}$  as we had for producing the training data. Deviations are equally natural (yet at small scale), since the training data is not overly extensive.

The resulting model  $f_0$  is obtained by invoking a nonlinear optimization via a call to `nonlin_min`, to minimize the functional  $\|(f(\mathbf{x}_i, \mathbf{p}) - y_i)_{i=1}^{10}\|_2$  using vectorization in GNU Octave. The minimization using the 2-norm has, in our case, the appeal of making the resulting model a best linear unbiased estimator by the Gauss-Markov theorem, whose hypotheses are here satisfied by construction. Thus, the trained model  $f_0$  is indeed a “good” ML model, as could be expected in real-life applications.

4. Now, for a plausible denial, we took a fresh set of (stochastically independent) samples of *decoy* training data  $\mathbf{X}'_i \sim \mathcal{U}(\{1, \dots, 8\}^m)$ , and another

set of random, and hence unrelated, response values  $\mathbf{Y}'_i \sim \mathcal{U}(\{1, \dots, 8\}^m)$ . Two things are important to note here:

- The decoy data is picked stochastically independent and at random, so the experiment was repeatable with different instances of all ingredients (only retaining fixed numeric ranges for the values),
  - and, more importantly, the response values  $y_i$  are *independent* of the inputs  $x_i$ , so any underlying functional relation between  $\mathbf{x}_i$  and the corresponding  $y_i$  is most likely not a linear regression model. Thus, the decoy data is completely different from the true training data.
5. Given the set of decoy samples  $(\mathbf{x}_i, y_i)_{i=1}^{10}$ , we proceed by implementing the steps as shown in Figure 3, producing the GNU Octave local variables `B`, `w1` corresponding to  $\mathbf{B}$  and  $\mathbf{w}_1$  from the text, and implementing the norm that Theorem 2 constructs as a function `crafted_norm`. All these computations take less than 10 lines of code<sup>5</sup>.

For checking the hypothesis of Lemma 1, i.e., the rank condition (4), the regression model comes in handy once more: it allows for a closed form expression of the Jacobian at  $\mathbf{p}$ , given directly by the data matrix, augmented with a mere column of all ones, i.e., for our model  $f(\mathbf{x}, (\beta_0, \dots, \beta_{d-1})) = \beta_0 + (\beta_1, \dots, \beta_{d-1}) \cdot \mathbf{x}$ , we find the Jacobian to be constant<sup>6</sup>, and given as

$$\mathbf{J} = \begin{pmatrix} 1 & \mathbf{x}_1 \\ 1 & \mathbf{x}_2 \\ \vdots & \vdots \\ 1 & \mathbf{x}_n \end{pmatrix},$$

in which each row  $\mathbf{x}_i$  is the  $i$ -th data sample used to train the model. This is the matrix against we check the rank to change when attaching the vector  $\mathbf{e}$ .

6. With these items, we then go back into the nonlinear optimization, again using the same function `nonlin_min`, but this time minimizing our designed norm implemented in the function `crafted_norm`, and formally found as Figure 3 tells.

The results, quite satisfyingly, demonstrated that the model fitted to the decoy data but using the specially constructed norm comes up approximately equal to the original model. Notably, it does so with the decoy data having no relation

<sup>5</sup>In Octave only, but a port to Python or other languages is not expected to become considerably more complex.

<sup>6</sup>More complex models would require a manual approximation of the Jacobian (unless analytic expressions are obtainable), but this amounts to nested for loop over  $i = 1 \dots n$  and over  $j = 1 \dots d$  to approximate the derivative  $\partial f_i / \partial p_j \approx \frac{1}{h} \cdot (f(\mathbf{x}_i, \mathbf{p} + h \cdot \mathbf{u}_j) - f(\mathbf{x}_i, \mathbf{p}))$ , in which  $\mathbf{u}_j$  is the  $j$ -th unit vector, and  $h > 0$  is some (very) small constant. This requires the ML model, as a programming object, has access routines to get and set the model parameters as we wish (the regression model is again convenient here, since it is easy to implement).

to the training data whatsoever, not even necessarily sharing its original distribution (the original data was a linear combination of uniform distributions, which is no longer uniform for two or more terms, while the decoy data had an overall uniform distribution). The numeric discrepancies between the newly fitted model and the original model can partly be attributed to our lack of fine-tuning in the optimization process; indeed, we invoked `nonlin_min` with all *default* settings, except for the starting point to be inside a neighborhood of the given parameter vector  $\mathbf{p}$ , known from the given model  $f_0$ . Indeed, even in the default configuration, the model fitted under the true and the decoy data came up quite “close” to each other, indicating potentially higher accuracy upon careful fine-tuning of the optimization. In addition, the choice of  $\mathbf{w}_1$  may also have an impact on the numeric behavior of the optimizer, as does any randomness that the optimization algorithm may employ internally. We leave both possibilities for numeric accuracy gains aside here, leaving the demonstration with the pointer towards the observation that higher dimensionality of the model (and we conducted further experiments with larger values for  $d$ ) made the approximation worse. Again, this is not unexpected in light of higher-dimensional optimization problems generally behaving less nice than lower-dimensional ones. Our choice of  $d = 6$ , however, makes a manual check of equality among 6 pairs of model parameters quick and simple to show in Section 5.

## D A “Cryptographic” View

The flow in Figure 4 resembles an analogous situation as for probabilistic encryption, where the norm is *playing the role of a random auxiliary input* to the encryption function: let  $E_{pk}(m_0, \omega)$  denote the probabilistic encryption of a message  $m_0$  under a public key  $pk$  and a random string (random coins)  $\omega$ . Given a ciphertext  $c$ , one could deny the validity of any proposed plaintext  $m_1$  if  $\forall c \exists m, \omega : E_{pk}(m, \omega) = c$ . This is indeed the case for ElGamal encryption (for example). This is the common way of defining security of encryption (see any of the standard cryptography textbooks), and our notion of plausible deniability is completely analogue to this.

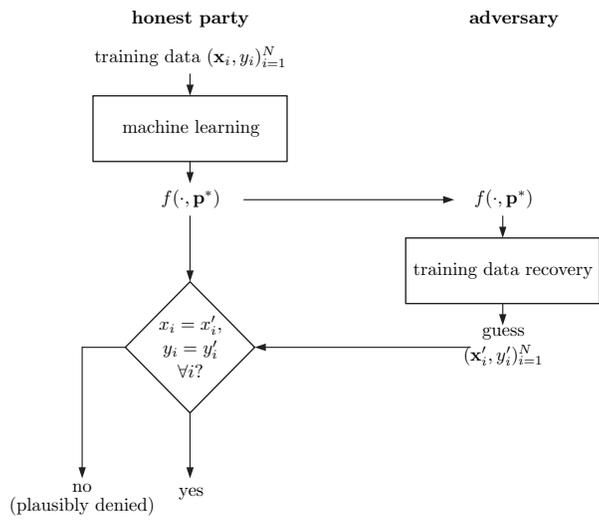


Figure 4: Plausible Deniability Experiment