Unified Taxonomy in AI Safety: Watermarks, Adversarial Defenses, and Transferable Attacks

Grzegorz Gluch [1] **Sai Ganesh Nagarajan** [*2] **Berkant Turan** [*2 3]

Abstract

As AI becomes omnipresent in today's world, it is crucial to study the safety aspects of learning, such as guaranteed watermarking capabilities and defenses against adversarial attacks. In prior works, these properties were generally studied separately and empirically barring a few exceptions. Meanwhile, strong forms of adversarial attacks that are *transferable* had been developed (empirically) for discriminative DNNs (Liu et al., 2016) and LLMs (Zou et al., 2023). In this ever-evolving landscape of attacks and defenses, we initiate the formal study of watermarks, defenses, and transferable attacks for classification, under a *unified framework*, by having two time-bounded players participate in an interactive protocol. Consequently, we show that for every learning task, at least one of the three schemes exists. Importantly, our results cover regimes where VC theory is not necessarily applicable. Finally we provide provable examples of the three schemes and show that transferable attacks exist only in regimes beyond bounded VC dimension. The example we give is a nontrivial construction based on cryptographic tools, i.e. homomorphic encryption.

1. Introduction

The increasing deployment of deep learning models amplifies the need for secure and robust models. In autonomous driving, ensuring the integrity of perception algorithms against adversarial manipulations is crucial for public safety (Amodei et al., 2016; Deng et al., 2020; Goodfellow et al., 2015; Kurakin et al., 2018). Additionally, in the competitive landscape of AI companies, where innovative models can represent significant intellectual and commercial value, it is crucial to have mechanisms to assert and protect ownership.

The concept of *watermarking* has emerged as a strategy for asserting content authenticity and ownership of both discriminative (Adi et al., 2018; Uchida et al., 2017; Fan et al., 2019; Zhang et al., 2018; Namba and Sakuma, 2019) and generative models (Kirchenbauer et al., 2023; Kuditipudi et al., 2023; Zhao et al., 2023a). These techniques embed unique signals into a model or its outputs, acting as digital fingerprints that safeguard intellectual property. Notably, in discriminative models, watermarking often utilizes backdoor-like methods (Zhang et al., 2018; Merrer et al., 2017; Adi et al., 2018; Namba and Sakuma, 2019), where developers embed specific mechanisms in the models that can be triggered by unusual phrases or image patterns.

Another aspect to consider is *transferable* adversarial attacks. These attacks involve creating adversarial examples that are effective across multiple models within the same class, such as deep neural networks (DNNs). Research has shown that these attacks can be highly effective against both discriminative and generative models (Liu et al., 2016; Zou et al., 2023). This underscores the need for defenses that can withstand attacks crafted not just for a specific model but for a broad set of models within the domain.

Provable guarantees for strong adversarial defenses and undetectable watermarks are rare and often hold only in special cases. For example, (Goldwasser et al., 2020) shows a defense that is secure against "all" adversarial examples for learning tasks with bounded VC-dimension.

This situation leads us to question which learning tasks, if any, can have provably *unremovable* watermarks and defenses that are robust against "all" adversaries. Ideally, we want every task to have these features. Early studies, like the one in (Goldwasser et al., 2022), show that some common adversarial defenses, such as those outlined in (Cohen et al., 2019), can remove undetectable backdoors, while methods like (Merrer et al., 2017) use adversarial samples to implant watermarks. These findings suggest a natural connection between the two areas.

Our contributions. We initiate a formal study of the taxonomy of discriminative learning tasks based on whether they have a watermark, an adversarial defense, or a trans-

[*]Equal contribution. [1]EPFL [2]Zuse Institute Berlin [3]Technische Universität Berlin. Correspondence to: Grzegorz Gluch <grzegorz.gluch@epfl.ch>, Sai Ganesh Nagarajan <nagarajan@zib.de>, Berkant Turan <turan@zib.de>.

Work presented at TF2M workshop at ICML 2024, Vienna, Austria. PMLR 235, 2024.

ferable attack. Building upon (Adi et al., 2018; Goldwasser et al., 2022; Merrer et al., 2017) we give a new formal definition of a watermark. Similarly, inspired by (Goldwasser et al., 2020) we introduce a new formal definition of an adversarial defense. Thirdly, we give, as far as we know, the first *formal* definition of a transferable attack.

Using this new formalism, our main result shows that:

Every learning task has at least one of the three: a Watermark, an Adversarial Defense, or a Transferable Attack.

Our results prove that this is, remarkably, a property of the learning task and the amount of resources available to learners—such as compute, memory, and data—rather than our ability to design algorithms.

In addition, we *provably* show:

1. The existence of a Adversarial Defense for all learning tasks with bounded VC-dimension, thereby ruling out Transferable Attacks in this regime.

2. An example of a Watermark for some class of learning tasks with bounded VC-dimension.

3. An Example of a Transferable Attack, albeit in the case with multiple valid labels for one input. Interestingly, the example uses tools from cryptography such as Fully Homomorphic Encryption (FHE). Thus ruling out Watermarks and Adversarial Defenses for this task.

Implications for Adversarial Attacks: Finally, we would like to point out some implications of our results, particularly for the well-studied notion of adversarial attacks, which can be seen as the negation of our notion of Adversarial Defense. Our main result states that if an Adversarial Defense does not exist, there must be either a Watermark or a Transferable Attack that fools all resource-bounded learners. Moreover, if a Transferable Attack does not exist, an adversarial attack can, in principle, be turned into a Watermark.

Comments on the Learning Task. Our definitions and the main results are phrased with respect to a *fixed* learning task (Section 2), while VC-theory takes an alternate viewpoint that tries to show guarantees on the risk (mostly sample complexity-based) for any distribution. However, it is known that for DNNs and other modern architectures, moving beyond classical VC theory would be necessary (Zhang et al., 2021; Nagarajan and Kolter, 2019). In our case due to the requirements of our schemes (e.g. unremovability and undetectability) it might not be possible to obtain this formalization that works for all distributions as is the case in classical VC theory.[1] We defer the discussion of related works to the Appendix A.

2. Preliminaries

Learning Task. For $n \in \mathbb{N}$ we define $[n] := \{0, 1, \ldots, n-1\}$. A *learning task* $\mathcal{L}$ is a pair (D, h) of a distribution D, $\text{supp}(D) \subseteq \mathcal{X}$ and a ground truth map $h : \mathcal{X} \to \mathcal{Y} \cup \{\perp\}$, where $\mathcal{Y}$ is a finite space of labels and $\perp$ represents a situation where h is not defined. To every $f : \mathcal{X} \to \mathcal{Y}$, we associate $\text{err}(f) := \mathbb{E}_{x \sim D}[f(x) \neq h(x)]$. We implicitly assume h does not map to $\perp$ on $\text{supp}(D)$. We assume all parties have access to i.i.d. samples $(x, h(x))$, where $x \sim D$, although D and h are unknown to the parties. For $q \in \mathbb{N}, \mathbf{x} \in \mathcal{X}^q, \mathbf{y} \in \mathcal{Y}^q$ we define $\text{err}(\mathbf{x}, \mathbf{y}) := \frac{1}{q} \sum_{i \in [q]} \mathbb{1}_{\{h(x(i)) \neq y(i), h(x(i)) \neq \perp\}}$, i.e. we count (x, y) as an error if h is well-defined on x and $h(x) \neq y$.

Distinguishability. For $q \in \mathbb{N}$ and distributions D_0, D_1 over $\mathcal{X}^q$ and an algorithm $\mathbf{A}$ accepting $\mathbf{x} \in \mathcal{X}^q$ and returning $\hat{b} \in \{0, 1\}$, we say that the probability of $\mathbf{A}$ for distinguishing D_0, D_1 is $p \in [0, 1]$ if $p = \mathbb{P}_{b \sim U(\{0,1\}), \mathbf{x} \sim D_b}\left[b = \mathbf{A}(\mathbf{x})\right]$, where $b \sim U(\{0, 1\})$ is a uniformly random bit.

3. Watermarks, Adversarial Defenses and Transferable Attacks

Here we define the three schemes and discuss the design choices we made. To do that we use the language of interactive protocols (Goldwasser and Sipser, 1986), where a verifier ($\mathbf{V}$) and a prover ($\mathbf{P}$) communicate according to certain rules. Watermarks and Transferable Attacks are initiated by $\mathbf{V}$, and Adversarial Defense by $\mathbf{P}$. Both $\mathbf{V}$ and $\mathbf{P}$ are time-bounded algorithms, but meaningful settings occur when either has sufficient time to train a model to a given accuracy. This setup enables us to generalize processes for watermarks and defenses, allowing richer interaction while capturing the essence of the process. For more details on the properties of each scheme, please refer to Appendix C.1.

Watermark is "an efficient algorithm that computes a low-error classifier $f : \mathcal{X} \to \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}^q$ such that fast adversaries are not able to find low-error answers $\mathbf{y} \in \mathcal{Y}^q$ nor distinguish $\mathbf{x}$ from a sample from D^q. Moreover, there is an efficient adversary that can find a low-error $\mathbf{y}$. "

Definition 1 (Watermark (Informal)). *Let $\mathcal{L} = (D, h)$ be a learning task, $\epsilon \in \left(0, \frac{1}{2}\right), t, T \in \mathbb{N}$, where t bounds the running time of $\mathbf{P}$, and T the running time of $\mathbf{V}$.*

[1] Since we are considering time-bounded parties, this implicitly restricts what can be learned. But at the same time, it allows the freedom of appropriate priors to be chosen by the parties, as long as they respect the time.

We say that a succinctly representable $\mathbf{V}_{\text{WATERMARK}}$ ***running in time*** T *implements a watermarking scheme for* $\mathcal{L}$*, denoted by* $\mathbf{V}_{\text{WATERMARK}} \in \text{WATERMARK}(\mathcal{L}, \epsilon, T, t)$*, if it computes* $(f, \mathbf{x})$*,* $f : \mathcal{X} \to \mathcal{Y}, \mathbf{x} \in \mathcal{X}^q$*, and* $\mathbf{P}$ *that on input* $(f, \mathbf{x})$ *returns* $\mathbf{y} \in \mathcal{Y}^q$ *satisfies the following.*

- ***Correctness (*** f ***has low error).*** *With high probability,* $err(f) \leq \epsilon$.
- ***Uniqueness (models trained from scratch give low-error answers).*** *There exists succinctly representable* $\mathbf{P}$ *running in time* T *such that with high probability,* $err(\mathbf{x}, \mathbf{y}) \leq 2\epsilon$.
- ***Unremovability (fast*** $\mathbf{P}$ ***return high-error answers).*** *For every succinctly representable* $\mathbf{P}$ ***running in time*** t *we have that with high probability,* $err(\mathbf{x}, \mathbf{y}) > 2\epsilon$.
- ***Undetectability (fast*** $\mathbf{P}$ ***cannot detect that they are tested).*** *For every succinctly representable* $\mathbf{P}$ ***running in time*** t *the advantage of* $\mathbf{P}$ *for distinguishing* $\mathbf{x} \sim D^q$ *from* $\mathbf{x} := \mathbf{V}$ *is small.*

Adversarial Defense is "is an efficient algorithm that computes a low-error $f : \mathcal{X} \to \mathcal{Y}$ *such that fast adversaries are not able to find* $\mathbf{x} \in \mathcal{X}^q$ *where* f *makes mistakes and which looks indistinguishable from a sample from* D^q*."*

Definition 2 (Adversarial Defense (Informal)). *Let* $\mathcal{L} = (D, h)$ *be a learning task. Let* $\epsilon \in (0, \frac{1}{2}), t, T \in \mathbb{N}$*, where* t *bounds the* $\mathbf{V}$*'s running time, and* T *the* $\mathbf{P}$*'s running time.*

We say that a succinctly representable $\mathbf{P}_{\text{DEFENSE}}$ ***running in time*** T *implements an adversarial defense for* $\mathcal{L}$*, denoted by* $\mathbf{P}_{\text{DEFENSE}} \in \text{DEFENSE}(\mathcal{L}, \epsilon, t, T)$*, if* $\mathbf{P}$ *computes* $f : \mathcal{X} \to \mathcal{Y}$*,* $\mathbf{V}$ *replies with* $\mathbf{x} := \mathbf{V}(f), \mathbf{x} \in \mathcal{X}^q$*, and* $\mathbf{P}$ *outputs* $b = \mathbf{P}(f, \mathbf{x}), b \in \{0, 1\}$ *satisfying the following.*

- ***Correctness (*** f ***has low error).*** *With high probability,* $err(f) \leq \epsilon$.
- ***Completeness (if*** $\mathbf{x}$ ***came from the right distribution*** $\mathbf{P}$ ***does not signal it is attacked).*** *When* $\mathbf{x} \sim D^q$ *then with high probability,*[2] $b = 0$.
- ***Soundness (fast attacks creating*** $\mathbf{x}$ ***on which*** f ***makes mistakes are detected).*** *For every succinctly representable* $\mathbf{V}$ ***running in time*** t *we have that with high probability,* $err(\mathbf{x}, f(\mathbf{x})) \leq 7\epsilon$ *or* $b = 1$.

Transferable Attack is "an efficient algorithm to compute $\mathbf{x} \in \mathcal{X}^q$ *for which fast adversaries are not able to find low error answers and that looks indistinguishable from a sample from* D^q*."*

[2] Correctness implies $\text{err}(\mathbf{x}, f(\mathbf{x})) \leq 2\epsilon$ with high probability.

Definition 3 (Transferable Adversarial Attack (Informal)). *Let* $\mathcal{L} = (D, h)$ *be a learning task. Let* $\epsilon \in (0, \frac{1}{2}), T \in \mathbb{N}$*,* T *bounds the running time of* $\mathbf{V}$ *and* $\mathbf{P}$*.*

We say that a succinctly representable $\mathbf{V}$ ***running in time*** T *is a transferable adversarial attack, denoted by* $\mathbf{V} \in \text{TRANSFATTACK}(\mathcal{L}, \epsilon, T, t)$*, if it computes* $\mathbf{x} \in \mathcal{X}^q$*, and the interaction with* $\mathbf{P}$ *that on input* $\mathbf{x}$ *returns* $\mathbf{y} = \mathbf{P}(\mathbf{x}), \mathbf{y} \in \mathcal{Y}^q$ *satisfies the following.*

- ***Transferability (fast provers return high error answers).*** *For every succinctly representable* $\mathbf{P}$ ***running in time*** t *we have that with high probability,* $err(\mathbf{x}, \mathbf{y}) > 2\epsilon$.
- ***Undetectability (fast provers cannot detect that they are tested).*** *For every succinctly representable* $\mathbf{P}$ ***running in time*** t *the advantage of* $\mathbf{P}$ *for distinguishing* $\mathbf{x} \sim D^q$ *from* $\mathbf{x} := \mathbf{V}$ *is small.*

On Error Oracles. We imagine the interaction is judged by an external party that potentially knows the distribution and h and can compute the necessary errors and provide the final decision. Basically, we imagine the "transcript" of the interaction is sent to this judge. It is an interesting future work for the parties to have access to *restricted* versions of error oracles, but this is beyond the scope of this work.

Comments on Succinct Representation for $\mathbf{V}$ and $\mathbf{P}$. We require the algorithms run by $\mathbf{V}$ and $\mathbf{P}$ to be succinctly representable, i.e. their code should be much smaller than their running time. For details see Appendix C.1.

4. Main Result

We are ready to state an informal version of our main result. Please refer to Theorem 4 for the details and full proof. The key idea here is to define a zero-sum game between $\mathbf{V}$ and $\mathbf{P}$, where the "actions" of each player are the possible algorithms/circuits that can be implemented in the given time bound. Notably, this game is finite, but there are exponentially many such "actions" for each player. For our result, we rely on some key properties of such large zero-sum games (Lipton and Young, 1994b; Lipton et al., 2003).

Theorem 1 (Informal). *For every learning task* $\mathcal{L}$ *and* $\epsilon \in (0, \frac{1}{2}), T \in \mathbb{N}$*, such that there exists a learner running in time* T *that, with high probability, learns* f *such that* $err(f) \leq \epsilon$*, at least one of the following exists:*

$$\text{WATERMARK}\left(\mathcal{L}, \epsilon, T, T^{1/\sqrt{\log(T)}}\right),$$
$$\text{DEFENSE}\left(\mathcal{L}, \epsilon, T^{1/\sqrt{\log(T)}}, O(T)\right),$$
$$\text{TRANSFATTACK}\left(\mathcal{L}, \epsilon, T, T\right).$$

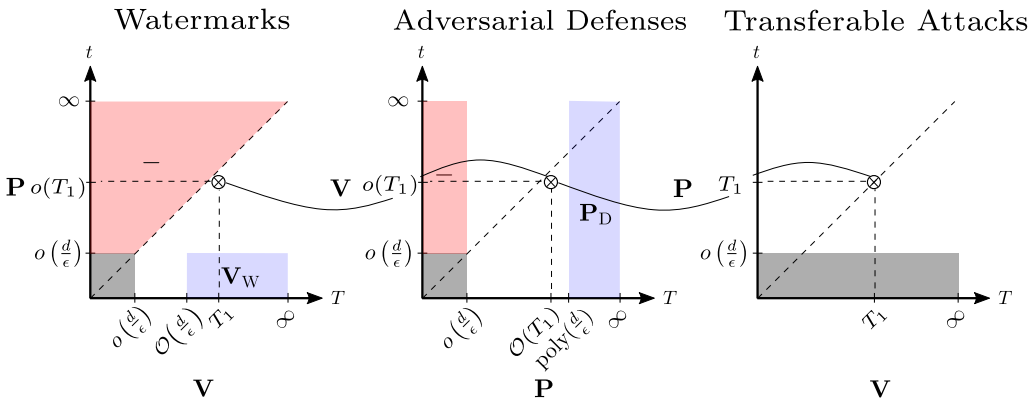

Figure 1: Taxonomy of learning tasks with bounded VC-dimension. The axes represent the time bound for the parties in the corresponding schemes. The blue regions depict positive results, the red negative, and the gray regimes of parameters which are not of interest. See Section 5 for details about the blue regions. The curved line represents an application of Theorem 1, which says that at least one of the three points should be blue.

The role of T, ϵ and specific constants and probabilities in Theorem 1. Defining the protocols with respect to the time parameter and then looking at the appropriately sized circuits allows us to argue about the existence of one of these properties across *all* algorithms running in the specified time bound. Additionally, we want to emphasize that Theorem 1 qualitatively also holds with other settings of parameters like acceptance and rejection probabilities as well as constants present in definitions of *unremovability, soundness, and transferability*. We chose specific values of parameters, rather than giving respective ranges for which the theorem holds, for simplicity. One could hope to generalize our result beyond the setting of fixed time (circuit size) T similarly to how it is done in the field of computational learning theory, where one defines an efficient learner as running in time polynomial in $1/\epsilon$ and a size parameter of the learning problem. We leave this question for future work.

We end this section with a simple observation that if a Transferable Attack exists then neither a Watermark nor a Defense exists. Indeed, a Transferable Attack is a strong notion of an attack so it rules out a Defense. Secondly, a Transferable Attack against adversaries running in time T rules out a Watermark because it conflicts with the *uniqueness* property.

5. Instantiations of our Defintions

In this section, we give three instantiations of our definitions, which demonstrate why the upper bounds on the running time of $\mathbf{V}$ and $\mathbf{P}$ are crucial parameters that distinguish between tasks having Watermarks, Adversarial Defenses, and Transferable Attacks. The full statements are given in Appendix E. Figure 1 summarizes the results visually.

We show an Adversarial Defense $\mathbf{P}_{\mathrm{D}}$ for *all* learning tasks $\mathcal{L}$ with VC-dimension bounded by d, i.e. for all ϵ

$$\mathbf{P}_{\mathrm{D}} \in \textsc{Defense}\left(\mathcal{L}, \epsilon, t = \infty, T = \mathtt{poly}\left(d/\epsilon\right)\right).$$

We also show a Watermark $\mathbf{V}_{\mathrm{W}}$ for a *class* of learning tasks $\mathcal{L}$ with the VC-dimension bounded by d, i.e. for all ϵ

$$\mathbf{V}_{\mathrm{W}} \in \textsc{Watermark}\left(\mathcal{L}, \epsilon, T = O\left(d/\epsilon\right), t = d/100\right).$$

Theorem 2 (Cryptography based Transferable Attack (Informal)). *There exists a family of distributions D, hypothesis class $\mathcal{H} = \{h^k\}_k$, distribution $D^{\mathcal{L}}$ over k and $\mathbf{V}_{\mathrm{TA}}$ such that for all sufficiently small ϵ if $k \sim D^{\mathcal{L}}_n$ then*

$$\mathbf{V}_{\mathrm{TA}} \in \textsc{TransfAttack}\left(\left(D, h^k\right), \epsilon, T = O\left(\frac{1}{\epsilon^{1.3}}\right), t = \frac{1}{\epsilon^2}\right).$$

The learning task from Theorem 2 considers an extension of the setup from Section 2, where there are multiple equally valid outputs for most inputs. In this sense, it is closer to generative than classification models (see also discussion in Section 6). Interestingly, this task is such that learning a good model is easy ($\mathtt{poly}(n, 1/\epsilon)$ time is enough) but evaluating an error of an input/output pair is hard (requires $2^{\Omega(n)}$ time).

6. Future Work - Beyond Classification

Inspired by Theorem 2 we conjecture a possibility of generalizing our results to develop a taxonomy for generative learning tasks. Instead of a ground truth function, one could consider a ground truth quality oracle Q, which measures the quality of every input/output pair. This model introduces new phenomena *not* present in the case of classification. For

example, the task of *generation*, i.e. producing a high-quality output y on input x, is decoupled from the task of *verification*, i.e. evaluating the quality of y as output for x. By, decoupled we mean that there is no clear formal reduction of one task to the other. Conversely, for classification, where the space of possible labels is small, the two tasks are equivalent. Without going into details, this decoupling is the reason why the proof of Theorem 1 does not automatically transfer to the generative case.

This decoupling introduces new complexities, but it also suggests that considering new definitions may be beneficial. For example, because generation and verification are equivalent for classification tasks, we allowed neither $\mathbf{V}$ nor $\mathbf{P}$ access to h, as it would trivialize the definitions. However, a modification of the Watermark definition, where access to Q is given to $\mathbf{P}$ could be investigated in the generative case. Interestingly, such a setting was considered in (Zhang et al., 2023b), where access to Q was crucial for mounting a provable attack on "all" strong watermarks. As we alluded to, Theorem2 can be seen as an example of a task, where generation is easy but verification is hard - the opposite to what (Zhang et al., 2023b) posits.

We hope that careful formalizations of the interaction and capabilities of all parties might give insights into not only the schemes considered in this work, but also problems like weak-to-strong generalization or scalable oversight.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

References

Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1615–1631, 2018.

Noga Amit, Shafi Goldwasser, Orr Paradise, and Guy Rothblum. Models that prove their own correctness. *arXiv preprint arXiv:2405.15722*, 2024.

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

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks, 2024.

Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 309–325, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311520. doi: 10.1145/2090236.2090262. URL https://doi.org/10.1145/2090236.2090262.

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Scalable ai safety via doubly-efficient debate. *arXiv preprint arXiv:2311.14125*, 2023.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *ArXiv*, abs/2306.15447, 2023. URL https://api.semanticscholar.org/CorpusID:259262181.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2023.

Jiefeng Chen, Yang Guo, Xi Wu, Tianqi Li, Qicheng Lao, Yingyu Liang, and Somesh Jha. Towards adversarial robustness via transductive learning. *arXiv preprint arXiv:2106.08387*, 2021.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*, 2023.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems*, pages 485–497, 2019.

Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou, and Miryung Kim. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*, pages 1–10. IEEE, 2020.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples

by translation-invariant attacks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4307–4316, 2019. URL https://api.semanticscholar.org/CorpusID:102350868.

Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *ArXiv*, abs/1712.02779, 2017. URL https://api.semanticscholar.org/CorpusID:21929206.

Yousof Erfani, Ramin Pichevar, and Jean Rouat. Audio watermarking using spikegram and a two-dictionary approach. *IEEE Transactions on Information Forensics and Security*, 12(4):840–852, 2017. doi: 10.1109/TIFS.2016.2636094.

Lixin Fan, Kam Woh Ng, and Chee Seng Chan. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. *Advances in neural information processing systems*, 32, 2019.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023.

A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, page 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

Craig Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, page 169–178, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585062. doi: 10.1145/1536414.1536440. URL https://doi.org/10.1145/1536414.1536440.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian J. Goodfellow. Adversarial spheres. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018. URL https://openreview.net/forum?id=SkthlLkPf.

S Goldwasser and M Sipser. Private coins versus public coins in interactive proof systems. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, STOC '86, page 59–68, New York, NY, USA, 1986. Association for Computing Machinery. ISBN 0897911938. doi: 10.1145/12130.12137. URL https://doi.org/10.1145/12130.12137.

Shafi Goldwasser, Yael Kalai, Raluca Ada Popa, Vinod Vaikuntanathan, and Nickolai Zeldovich. Reusable garbled circuits and succinct functional encryption. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 555–564, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290. doi: 10.1145/2488608.2488678. URL https://doi.org/10.1145/2488608.2488678.

Shafi Goldwasser, Adam Tauman Kalai, Yael Tauman Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. *ArXiv*, abs/2204.06974, 2022. URL https://api.semanticscholar.org/CorpusID:248177888.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. Evading watermark based detection of ai-generated content. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023. URL https://api.semanticscholar.org/CorpusID:258557682.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR, 23–29 Jul

2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-verifier games improve legibility of llm outputs. *arXiv preprint arXiv:2407.13692*, 2024.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *CoRR*, abs/2307.15593, 2023. doi: 10.48550/ARXIV.2307.15593. URL https://doi.org/10.48550/arXiv.2307.15593.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. CRC Press, 2018.

Richard J. Lipton and Neal E. Young. Simple strategies for large zero-sum games with applications to complexity theory. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '94, page 734–740, New York, NY, USA, 1994a. Association for Computing Machinery. ISBN 0897916638. doi: 10.1145/195058.195447. URL https://doi.org/10.1145/195058.195447.

Richard J Lipton and Neal E Young. Simple strategies for large zero-sum games with applications to complexity theory. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 734–740, 1994b.

Richard J Lipton, Evangelos Markakis, and Aranyak Mehta. Playing large games using simple strategies. In *Proceedings of the 4th ACM Conference on Electronic Commerce*, pages 36–41, 2003.

Chang Liu, Jie Zhang, Han Fang, Zehua Ma, Weiming Zhang, and Nenghai Yu. Dear: A deep-learning-based audio re-recording resilient watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13201–13209, 2023.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically, 2024.

Rimon Melamed, Lucas H. McCabe, Tanay Wakhare, Yejin Kim, H. Howie Huang, and Enric Boix-Adsera. Prompt have evil twins, 2024.

Erwan Le Merrer, Patrick Pérez, and Gilles Trédan. Adversarial frontier stitching for remote neural network watermarking. *Neural Computing and Applications*, 32:9233 – 9244, 2017. URL https://api.semanticscholar.org/CorpusID:11008755.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Transductive robust learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 11461–11471. PMLR, 2022.

Yuki Nagai, Yusuke Uchida, Shigeyuki Sakazawa, and Shin'ichi Satoh. Digital watermarking for deep neural networks. *International Journal of Multimedia Information Retrieval*, 7:3–16, 2018.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Ryota Namba and Jun Sakuma. Robust watermarking of neural network with exponential weighting. *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 2019. URL https://api.semanticscholar.org/CorpusID:58028915.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Bys4ob-Rb.

Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 84–93. ACM, 2005.

R. Rivest, L. Adleman, and M. Dertouzos. On data banks and privacy homomorphisms. In *Foundations of Secure Computation*, page 169–179, New York, NY, USA, 1978. Academic Press.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua

Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Stuart A. Thompson Tiffany Hsu. Disinformation researchers raise alarms about a.i. chatbots. https://scottaaronson.blog/?p=6823, 2023. Accessed: March 2024.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 269–277, 2017.

Vinod Vaikuntanathan. Computing blindfolded: New developments in fully homomorphic encryption. In *Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, FOCS '11, page 5–16, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 9780769543001. doi: 10.1109/FOCS.2011.98. URL https://doi.org/10.1109/FOCS.2011.98.

Stephan Wäldchen, Kartikey Sharma, Berkant Turan, Max Zimmer, and Sebastian Pokutta. Interpretability guarantees with merlin-arthur classifiers. In *International Conference on Artificial Intelligence and Statistics*, pages 1963–1971. PMLR, 2024.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *ArXiv*, abs/2307.02483, 2023. URL https://api.semanticscholar.org/CorpusID:259342528.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 51008–51025. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a00548031e4647b13042c97c922fadf1-Paper-Conference.pdf.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *ArXiv*, abs/2305.20030, 2023b. URL https://api.semanticscholar.org/CorpusID:258987524.

Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5283–5292. PMLR, 2018. URL http://proceedings.mlr.press/v80/wong18a.html.

Yi-Hsuan Wu, Chia-Hung Yuan, and Shan-Hung Wu. Adversarial robustness via runtime masking and cleansing. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10399–10409. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wu20f.html.

Cihang Xie, Zhishuai Zhang, Jianyu Wang, Yuyin Zhou, Zhou Ren, and Alan Loddon Yuille. Improving transferability of adversarial examples with input diversity. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2725–2734, 2018. URL https://api.semanticscholar.org/CorpusID:3972825.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiV*, abs/2311.04378, 2023a. doi: 10.48550/ARXIV.2311.04378. URL https://doi.org/10.48550/arXiv.2311.04378.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiV*, abs/2311.04378, 2023b. doi: 10.48550/ARXIV.2311.04378. URL https://doi.org/10.48550/arXiv.2311.04378.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, page 159–172, New York, NY, USA, 2018. Association for Computing

Machinery. ISBN 9781450355766. doi: 10.1145/3196494.3196550. URL https://doi.org/10.1145/3196494.3196550.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *CoRR*, abs/2306.17439, 2023a. doi: 10.48550/ARXIV.2306.17439. URL https://doi.org/10.48550/arXiv.2306.17439.

Xuandong Zhao, Kexun Zhang, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. 2023b. URL https://api.semanticscholar.org/CorpusID:259075167.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *ArXiv*, abs/2303.10137, 2023c. URL https://api.semanticscholar.org/CorpusID:257622907.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *ArXiv*, abs/2307.15043, 2023. URL https://api.semanticscholar.org/CorpusID:260202961.

A. Related Work

This section provides an overview of the main areas relevant to our work: Watermarking techniques, adversarial defenses, and transferable attacks on deep neural networks (DNNs). Each subsection outlines important contributions and the current state of research in these areas.

A.1. Watermarking

Watermarking techniques are crucial for protecting the intellectual property of machine learning models. These techniques can be broadly categorized based on the type of model they target. We review watermarking schemes for both discriminative and generative models, with a primary focus on discriminative models, as our work builds upon these methods.

A.1.1. Watermarking Schemes for Discriminative Models

Discriminative models, which are designed to categorize input data into predefined classes, have been a major focus of watermarking research. The key approaches in this domain can be divided into black-box and white-box approaches.

Black-Box Setting: In the black-box setting, the model owner does not have access to the internal parameters or architecture of the model, but can query the model to observe its outputs. This setting has seen the development of several watermarking techniques, primarily through backdoor-like methods.

Adi et al. and Zhang et al. proposed frameworks that embed watermarks using specifically crafted input data (e.g., unique patterns) with predefined outcomes. These watermarks can be verified by feeding these special inputs into the model and checking for the expected outputs, thereby confirming ownership.

Another significant contribution in this domain is by Merrer et al., who introduced a method that employs adversarial examples to embed the backdoor. Adversarial examples are perturbed inputs that cause the model to produce specific outputs, thus serving as a watermark.

Namba and Sakuma further enhanced the robustness of black-box watermarking schemes by developing techniques that withstand various model modifications and attacks. These methods ensure that the watermark remains intact and detectable even when the model undergoes transformations.

Provable undetectability of backdoors was achieved in the context of classification tasks by Goldwasser et al.. Unfortunately, it is known ((Goldwasser et al., 2022)) that some undetectable watermarks are easily removed by simple mechanisms similar to randomized smoothing (Cohen et al., 2019).

The popularity of black-box watermarking is due to its practical applicability, as it does not require access to the model's internal workings. This makes it suitable for scenarios where models are deployed as APIs or services. Our framework builds upon these black-box watermarking techniques, extending their robustness and applicability in adversarial environments.

White-Box Setting: In contrast, the white-box setting assumes that the model owner has full access to the model's parameters and architecture, allowing for direct examination to confirm ownership. The initial methodologies for embedding watermarks into the weights of DNNs were introduced by Uchida et al. and Nagai et al.. Uchida et al. presented a framework for embedding watermarks into the model weights, which can be examined to confirm ownership.

An advancement in white-box watermarking is provided by Darvish Rouhani et al., who developed a technique to embed an N-bit ($N \geq 1$) watermark in DNNs. This technique is both *data-* and *model-dependent*, meaning the watermark is activated only when specific data inputs are fed into the model. For revealing the watermark, activations from intermediate layers are necessary in the case of white-box access, whereas only the final layer's output is needed for black-box scenarios.

Our work does not focus on white-box watermarking techniques. Instead, we concentrate on exploring the interaction between backdoor-like watermarking techniques, adversarial defenses, and transferable attacks. Overall, watermarking through backdooring has become more popular due to its applicability in the black-box setting.

A.1.2. Watermarking Schemes for Generative Models

Watermarking techniques for generative models have attracted considerable attention with the advent of large language models (LLMs) and other advanced generative models. This increased interest has led to a surge in research and diverse

contributions in this area.

Large Language Models: Watermarking LLMs is critical for mitigating potential harms associated with generated text. Significant contributions in this domain include (Kirchenbauer et al., 2023), who proposed a watermarking framework that embeds signals into generated text that are invisible to humans but detectable algorithmically. This method promotes the use of a randomized set of "green" tokens during text generation, and detects the watermark without access to the language model API or parameters.

Kuditipudi et al. introduced robust distortion-free watermarks for language models. Their method ensures that the watermark does not distort the generated text, providing robustness against various text manipulations while maintaining the quality of the output.

Zhao et al. presented a provable, robust watermarking technique for AI-generated text. This approach offers strong theoretical guarantees for the robustness of the watermark, making it resilient against attempts to remove or alter it without significantly changing the generated text.

However, Zhang et al. highlighted vulnerabilities in these watermarking schemes. Their work demonstrates that current watermarking techniques can be effectively broken, raising important considerations for the future development of robust and secure watermarking methods for LLMs.

Image Generation Models: Various watermarking techniques have been developed for image generation models to address ethical and legal concerns. Fernandez et al. introduced a method combining image watermarking with Latent Diffusion Models, embedding invisible watermarks in generated images for future detection. This approach is robust against modifications such as cropping. Wen et al. proposed Tree-Ring Watermarking, which embeds a pattern into the initial noise vector during sampling, making the watermark robust to transformations like convolutions and rotations. Jiang et al. highlighted vulnerabilities in watermarking schemes, showing that human-imperceptible perturbations can evade watermark detection while maintaining visual quality. Zhao et al. provided a comprehensive analysis of watermarking techniques for Diffusion Models, offering a recipe for efficiently watermarking models like Stable Diffusion, either through training from scratch or fine-tuning. Additionally, Zhao et al. demonstrated that invisible watermarks are vulnerable to regeneration attacks that remove watermarks by adding random noise and reconstructing the image, suggesting a shift towards using semantically similar watermarks for better resilience.

Audio Generation Models: Watermarking techniques for audio generators have been developed for robustness against various attacks. Erfani et al. introduced a spikegram-based method embedding watermarks in high-amplitude kernels, robust against MP3 compression and other attacks while preserving quality. Liu et al. proposed DeAR, a deep-learning-based approach resistant to audio re-recording (AR) distortions.

A.2. Adversarial Defense

The field of adversarial robustness has a rich and extensive literature (Szegedy et al., 2014; Gilmer et al., 2018; Raghunathan et al., 2018; Wong and Kolter, 2018; Engstrom et al., 2017). Adversarial defenses are essential for ensuring the security and reliability of machine learning models against adversarial attacks that aim to deceive them with carefully crafted inputs.

For discriminative models, there has been significant progress in developing adversarial defenses. Techniques such as adversarial training (Madry et al., 2018), which involves training the model on adversarial examples, have shown promise in improving robustness. Certified defenses (Raghunathan et al., 2018) provide provable guarantees against adversarial attacks, ensuring that the model's predictions remain unchanged within a specified perturbation bound. Additionally, methods like randomized smoothing (Cohen et al., 2019) offer robustness guarantees.

A particularly relevant work for our study is (Goldwasser et al., 2020), which considers a different model for generating adversarial examples. This approach has significant implications for the robustness of watermarking techniques in the face of adversarial attacks.

In the context of Large Language Models (LLMs), there is a rapidly growing body of research focused on identifying adversarial examples (Zou et al., 2023; Carlini et al., 2023; Wen et al., 2023a). This research is closely related to the notion of *jailbreaking* (Andriushchenko et al., 2024; Chao et al., 2023; Mehrotra et al., 2024; Wei et al., 2023), which involves manipulating models to bypass their intended constraints and protections.

Furthermore, by employing the Greedy Coordinate Gradient (GCG) technique, Melamed et al. (2024) were able to develop prompts that, despite being incomprehensible to humans, achieved similar outcomes as the original natural-language prompts. These so-called *evil twins* can be exploited by malicious users. For instance, it is feasible for a malicious user to leverage this framework to construct prompts that generate a corpus of toxic or harmful documents, while not appearing malicious at surface level. However, these risks can be mitigated by implementing pre-processing defenses like paraphrasing or retokenization (Jain et al., 2023).

A.3. Transferable Attacks and Transductive Learning

Transferable attacks refer to adversarial examples that are effective across multiple models. Moreover, transductive learning has been explored as a means to enhance adversarial robustness, and since our Definition 3 captures some notion of transductive learning in the context of transferable attacks, we highlight significant contributions in these areas.

Adversarial Robustness via Transductive Learning: Transductive learning (Gammerman et al., 1998) has shown promise in improving the robustness of models by utilizing both training and test data during the learning process. This approach aims to make models more resilient to adversarial perturbations encountered at test time.

One significant contribution is by Goldwasser et al., which explores learning guarantees in the presence of arbitrary adversarial test examples, providing a foundational framework for transductive robustness. Another notable study by Chen et al. formalizes transductive robustness and proposes a bilevel attack objective to challenge transductive defenses, presenting both theoretical and empirical support for transductive learning's utility.

Additionally, Montasser et al. introduce a transductive learning model that adapts to perturbation complexity, achieving a robust error rate proportional to the VC dimension. The method by Wu et al. improves robustness by dynamically adjusting the network during runtime to mask gradients and cleanse non-robust features, validated through experimental results. Lastly, Tramer et al. critique the standard of adaptive attacks, demonstrating the need for specific tuning to effectively evaluate and enhance adversarial defenses.

Transferable Attacks on DNNs: Transferable attacks exploit the vulnerability of models to adversarial examples that generalize across different models. For discriminative models, significant works include (Liu et al., 2016), which investigates the transferability of adversarial examples and their effectiveness in black-box attack scenarios, (Xie et al., 2018), who propose input diversity techniques to enhance the transferability of adversarial examples across different models, and (Dong et al., 2019), which presents translation-invariant attacks to evade defenses and improve the effectiveness of transferable adversarial examples.

In the context of generative models, including large language models (LLMs) and other advanced generative architectures, relevant research is rapidly emerging, focusing on the transferability of adversarial attacks. This area is crucial as it aims to understand and mitigate the risks associated with adversarial examples in these powerful models. Notably, Zou et al. explored universal and transferable adversarial attacks on aligned language models, highlighting the potential vulnerabilities and the need for robust defenses in these systems.

B. Additional Preliminaries

Interactive protocols. We will model the interaction in the language of *interactive protocols* (Goldwasser and Sipser, 1986). These were initially introduced in the context of computational complexity and cryptography, but later found applications in a broad spectrum of fields. There are two parties $\mathbf{V}$, as in verifier, and $\mathbf{P}$, as in prover. $\mathbf{V}$ plays the role of a watermarking party or an adversary trying to break an adversarial defense and $\mathbf{P}$ plays the role of an adversarial defense or an adversary trying to break a watermarking scheme. In addition, $\mathbf{V}$ plays the role of the attacking party for the transferable attack, and $\mathbf{P}$ is the defending party here.

The concept of *interative protocols* has found a lot of interest in recent years for various domains in AI safety (Brown-Cohen et al., 2023; Amit et al., 2024; Khan et al., 2024) and interpretability (Wäldchen et al., 2024; Kirchner et al., 2024). In this paper we use the verifier-prover interactions to define the primary safety properties studied in this paper (as defined in Section 3) and the induced game to guarantee the existence of the aforementioned properties.

		Undetectability	Unremovability	Uniqueness
Classification	(Goldwasser et al., 2022)	✔	robust to some smoothing attacks	✔(E)
	(Adi et al., 2018; Zhang et al., 2018)	✔(E)	✗	✔(E)
	(Merrer et al., 2017)	✔(E)	robust to fine tunning attacks	✔(E)
LLMs	(Christ et al., 2023; Kuditipudi et al., 2023)	✔	✗	✔
	(Zhao et al., 2023a)	✗	robust to edit distance attacks only	✔
	(Tiffany Hsu, 2023)	✔(E)	✗	✔
	(Kirchenbauer et al., 2023)	✗	✗	✔

Table 1: Overview of properties across various watermarking schemes. The symbol ✔ denotes properties with formal guarantees or where proof is plausible, whereas ✗ indicates the absence of such guarantees. Entries marked with ✔(E) represent properties observed empirically; these lack formal proof in the corresponding literature, suggesting that deriving such proof may present substantial challenges.

	Fast ($t \ll T_{\mathcal{L}}$) or Slow ($t \gtrsim T_{\mathcal{L}}$)	Robust against
Arbitrary test examples (Goldwasser et al., 2020)	**Slow** – runs in time $t \gtrsim T_{\mathcal{L}}$ for bounded VC-dim classes	**ALL** - adversarial examples works for bounded VC-dim classes
Randomized smoothing (Cohen et al., 2019)	**Fast**- runs in time $O(1)$	**ℓ_p-bounded perturbations**
Adversarial training (Madry et al., 2018)	**Slow**- runs in time $t \gtrsim T_{\mathcal{L}}$	**"ℓ_p-bounded perturbations"**

Table 2: Properties of some adversarial defenses. For a learning task $\mathcal{L}$ we denote by $T_{\mathcal{L}}$ the computational cost needed to learn $\mathcal{L}$, by which we need the time needed to learn $\mathcal{L}$. "·" signifies that the property holds empirically but the corresponding paper provides no proof.

C. Formal Definitions

As mentioned in the main paper we are interested in succinct circuits.

Definition 4 (Succint circuits). *Let C be a circuit of width w and depth d. We will denote size$(C) := w \cdot d$. We say that C is **succinctly representable** if there exists a circuit of size $100\log(\text{size}(C))$[3] that accepts as input $i \in [w], j, j_1, j_2 \in [d], g \in [O(1)]$, where g represents a gate from a universal constant-sized gate set, and returns 0 or 1, depending if g appears in location (i, j) in C and if it is connected to gates in locations $(i-1, j_1)$ and $(i-1, j_2)$.*

We are ready to formally define our notion of a Watermark, Adversarial Defense and Transferable Attack.

Definition 5 (Watermark). *Let $\mathcal{L} = (D, h)$ be a learning task. Let $T, t, q \in \mathbb{N}, \epsilon \in \left(0, \frac{1}{2}\right), l, c, s \in (0, 1), s < c$, where t bounds the running time of $\mathbf{P}$, and T the running time of $\mathbf{V}$, q the number of queries, ϵ the risk level, c probability that uniqueness holds, s probability that unremovability and undetectability holds, l the learning probability.*

*We say that a succinctly representable circuit $\mathbf{V}_{\text{WATERMARK}}$ **of size** T implements a watermarking scheme for $\mathcal{L}$, denoted by $\mathbf{V}_{\text{WATERMARK}} \in \text{WATERMARK}(\mathcal{L}, \epsilon, q, T, t, l, c, s)$, if it computes $(f, \mathbf{x})$, and $\mathbf{P}$ that on input $(f, \mathbf{x})$ returns $\mathbf{y} \in \mathcal{Y}^q$ satisfies the following.*

[3]Constant 100 is chosen arbitrarily. One often considers circuits representable by `polylog`-sized circuits. But for us, the constants play a role and this is why we chose this definition.

- ***Correctness (f has low error).*** *With probability at least l*

$$err(f) \leq \epsilon.$$

- ***Uniqueness (models trained from scratch give low-error answers).*** *There exists a succinctly representable circuit $\mathbf{P}$ of size T such that with probability at least c*

$$err(\mathbf{x}, \mathbf{y}) \leq 2\epsilon.$$

- ***Unremovability (fast $\mathbf{P}$ give high-error answers).*** *For every succinctly representable circuit $\mathbf{P}$ of size at most t we have that with probability at most s*

$$err(\mathbf{x}, \mathbf{y}) \leq 2\epsilon.$$

- ***Undetectability (fast $\mathbf{P}$ cannot detect that they are tested).*** *For every succinctly representable circuit $\mathbf{P}$ of size at most t the probability of $\mathbf{P}$ for distinguishing $\mathbf{x} \sim D^q$ from $\mathbf{x} := \mathbf{V}$ is at most $\frac{1}{2} + \frac{s}{2}$.*

Definition 6 (**Adversarial Defense**). *Let $\mathcal{L} = (D, h)$ be a learning task. Let $T, t, q \in \mathbb{N}, \epsilon \in \left(0, \frac{1}{2}\right), l, c, s \in (0, 1), s < c$, where t bounds the running time of $\mathbf{V}$, and T the running time of $\mathbf{P}$, q the number of queries, ϵ the risk level, c the completeness, s the soundness, l the learning probability.*

We say that a succinctly representable circuit $\mathbf{P}_{\text{DEFENSE}}$ of size T implements an adversarial defense for $\mathcal{L}$, denoted by $\mathbf{P}_{\text{DEFENSE}} \in \text{DEFENSE}(\mathcal{L}, \epsilon, q, t, T, l, c, s)$, if $\mathbf{P}$ sends f, $\mathbf{V}$ replies with $\mathbf{x} \in \mathcal{X}^q$, and $\mathbf{P}$ outputs $b = \mathbf{P}(f, \mathbf{x}), b \in \{0, 1\}$ satisfying the following.

- ***Correctness (f has low error).*** *With probability at least l*

$$err(f) \leq \epsilon.$$

- ***Completeness (if $\mathbf{x}$ came from the right distribution $\mathbf{P}$ does not signal it is attacked).*** *When $\mathbf{x} \sim D^q$ then with probability at least c*

$$b = 0.$$

- ***Soundness (fast attacks creating $\mathbf{x}$ on which f makes mistakes are detected).*** *For every succinctly representable circuit $\mathbf{P}$ of size at most t we have that with probability at most s*

$$err(\mathbf{x}, f(\mathbf{x})) > 7\epsilon \text{ and } b = 0.$$

Definition 7 (**Transferable Attack**). *Let $\mathcal{L} = (D, h)$ be a learning task. Let $T, t, q \in \mathbb{N}, \epsilon \in \left(0, \frac{1}{2}\right), c, s \in (0, 1)$, where T bounds the running time of $\mathbf{V}$ and $\mathbf{P}$, q the number of queries, ϵ the risk level, c the transferability probability, s the undetectability probability.*

We say that a succinctly representable circuit $\mathbf{V}$ running in time T is a transferable adversarial attack, denoted by $\mathbf{V} \in \text{TRANSFATTACK}(\mathcal{L}, \epsilon, q, T, t, c, s)$, if it computes $\mathbf{x} \in \mathcal{X}^q$, and the interaction with $\mathbf{P}$ that on input $\mathbf{x}$ returns $\mathbf{y} = \mathbf{P}(\mathbf{x}), \mathbf{y} \in \mathcal{Y}^q$ satisfies the following.

- ***Transferability (fast provers return high error answers).*** *For every succinctly representable circuit $\mathbf{P}$ of size at most t we have that with probability at least c*

$$err(\mathbf{x}, \mathbf{y}) > 2\epsilon.$$

- ***Undetectability (fast provers cannot detect that they are tested).*** *For every succinctly representable circuit $\mathbf{P}$ of size at most t we have that the probability of $\mathbf{P}$ for distinguishing $\mathbf{x} \sim D^q$ from $\mathbf{x} := \mathbf{V}$ is at most $\frac{1}{2} + \frac{s}{2}$.*

C.1. Discussion of the Definitions

Watermark. *Unremovability* and *Undectability* are standard but desirable properties of a watermark. But some detailed remarks are in order.

Capabilities of the watermarking party and the adversaries. Naturally, for the protocol to be useful, the watermarking party and the honest adversary should be efficient algorithms. On the other hand, unremovability requires that dishonest adversaries should *not* be accepted. This shows that for a watermark to be possible, the class of allowed dishonest adversaries *has to be strictly smaller* than the class of honest adversaries. To model the intuition that it should be *hard* to remove a watermark, we chose to define the allowed classes of algorithms based on their *running time*.[4] By the above discussion, we can't model both honest and dishonest adversaries as running in polynomial time, as then the classes would be equal. To address this, we chose to use a more granular approach, and we assume that there are $t, T \in \mathbb{N}, t < T$ such that the honest adversary runs in time T and the cheating adversaries are limited to running in time t. We chose to model the watermarking party as also running in time T, although other choices are possible. Our running time restrictions are similar to those in (Adi et al., 2018). We will see later that t, T will play a crucial role for the existence of watermarks.

Uniqueness. We enforce the property that, if an honest adversary does *not* use f and trains a model f_{H} from scratch, then it should be accepted. Formally, we require that there exists an algorithm running in time T that produces $\mathbf{y}_{\mathrm{H}}$, which has an error at most 2ϵ with high probability. The value 2ϵ is the previously mentioned threshold for acceptance of the adversary.

Unremovability. The watermark should be hard to remove. In our definition, this is modeled by requiring that every adversary running in time t is *not* able to produce $\mathbf{y}$ that has an error lower than 2ϵ. Note the separation that it gives from uniqueness.

Undetectability. Lastly, we require *undetectability*. It should be *hard* for the adversary to detect that it is being tested, i.e. $\mathbf{x} \in \mathcal{X}^q$ should be indistinguishable for the adversary from a sample from D^q. Finally, a dishonest adversary should not be able to detect that it is being tested. Similarly to unremovability, we say that *no* adversary running in time t can distinguish $\mathbf{x}$ from a sample from D^q with high probability.

Adversarial Defense. Similar to our definition of watermark, we enforce that the defending party sends a low-error classifier that it has learned. The main property of our defense is that a successful defense must be able to detect that it is being *tested*. This forces the attacker to provide samples that are an adversarial attack, where the low-error classifier makes mistakes, and also that these examples must be *indistingushable* from the data distribution D.

Transferable Attack. The main remark here for our definition of Transferable Attack is that $\mathbf{P}$ can learn f after seeing $\mathbf{x}$ sent by $\mathbf{V}$ and is still not able to provide low error answers nor detect that it is being tested.

Comments on *Correctness* Requirement. Note that we enforce *Correctness* in the protocol for both Watermarks and Adversarial Defenses. This is to mimic the idea that trained models are generally available publicly and one aims to exhibit good quality models. Also, without enforcing this the protocols are trivial to satisfy.

Comments on Succinct Representation for $\mathbf{V}$ and $\mathbf{P}$. This is in accordance with how learning takes place in practice, for instance, consider DNNs and learning algorithms for those DNNs. The code representing gradient descent algorithms is almost always much shorter than the time of the optimization of weights DNNs for which it is run. Additionally, the requirement that $\mathbf{V}$'s and $\mathbf{P}$'s algorithms are succinct forbids a trivial way to circumvent learning by hard-coding f in the description of the Watermark or Adversarial Defense algorithms.

D. Main Theorem

Before proving our main theorem we recall a result from (Lipton and Young, 1994a) about simple strategies for large zero-sum games.

Game theory. A *two-player zero-sum game* is specified by a payoff matrix $\mathcal{G}$. $\mathcal{G}$ is an $r \times c$ matrix. MIN, the row player, chooses a probability distribution p_1 over the rows. MAX, the column player, chooses a probability distribution p_2 over the columns. A row i and a column j are drawn from p_1 and p_2 and MIN pays $\mathcal{G}_{ij}$ to MAX. MIN tries to minimize the expected payment; MAX tries to maximize it.

[4] It is possible to consider data-limited adversaries also.

By the Min-Max Theorem, there exist optimal strategies for both MIN and MAX. Optimal means that playing first and revealing one's mixed strategy is not a disadvantage. Such a pair of strategies is also known as a Nash equilibrium. The expected payoff when both players play optimally is known as the value of the game and is denoted by $\mathcal{V}(\mathcal{G})$.

We will use the following theorem from (Lipton and Young, 1994a), which says that optimal strategies can be approximated by uniform distributions over sets of pure strategies of size $O(\log(c))$.

Theorem 3 ((Lipton and Young, 1994a)). *Let $\mathcal{G}$ be an $r \times c$ payoff matrix for a two-player zero-sum game. For any $\eta \in (0,1)$ and $k \geq \frac{\log(c)}{2\eta^2}$ there exists a multiset of pure strategies for the* MIN *(row player) of size k such that a mixed strategy p_1 that samples uniformly from this multiset satisfies*

$$\max_j \sum_i p_1(i)\mathcal{G}_{ij} \leq \mathcal{V}(\mathcal{G}) + \eta(\mathcal{G}_{max} - \mathcal{G}_{min}),$$

where $\mathcal{G}_{max}, \mathcal{G}_{min}$ denote the maximum and minimum entry of $\mathcal{G}$ respectively. The symmetric result holds for the MAX *player.*

We are ready to prove our main theorem.

Theorem 4. *For every learning task $\mathcal{L} = (D, h)$; and $\epsilon \in (0,1)$, $T, q \in \mathbb{N}$, such that there exists a succinctly representable circuit of size T that learns $\mathcal{L}$ up to error ϵ with probability $1 - \frac{1}{48}$, at least one of*

$$\begin{aligned}
\textsc{Watermark}&\left(\mathcal{L}, \epsilon, q, T, T^{\frac{1}{2^{10}\sqrt{\log(T)}}}, l = \frac{10}{24}, c = \frac{21}{24}, s = \frac{19}{24}\right),\\
\textsc{Defense}&\left(\mathcal{L}, \epsilon, q, T^{\frac{1}{2^{10}\sqrt{\log(T)}}}, 2T, l = 1 - \frac{1}{48}, c = \frac{13}{24}, s = \frac{11}{24}\right),\\
\textsc{TransfAttack}&\left(\mathcal{L}, \epsilon, q, T, T, c = \frac{3}{24}, s = \frac{19}{24}\right)
\end{aligned}$$

exists.

Proof of Theorem 4. Let $\mathcal{L} = \Big(D, h\Big)$ be a learning task. Let $T, q, C \in \mathbb{N}, \epsilon \in \left(0, \frac{1}{2}\right)$.

Let $\mathfrak{Candidate}_{\mathfrak{W}}$ be a set of $T^{\frac{1}{2^{10}\sqrt{\log(T)}}}$-sized succinctly representable circuits computing $(f, \mathbf{x})$, where $f : \mathcal{X} \to \mathcal{Y}$. Similarly, let $\mathfrak{Candidate}_{\mathfrak{D}}$ be a set of $T^{\frac{1}{2^{10}\sqrt{\log(T)}}}$-sized succinctly representable circuits accepting as input $(f, \mathbf{x})$ and outputing $(\mathbf{y}, b)$, where $\mathbf{y} \in \mathcal{Y}^q, b \in \{0,1\}$. We interpret $\mathfrak{Candidate}_{\mathfrak{W}}$ as candidate algorithms for a watermark, and $\mathfrak{Candidate}_{\mathfrak{D}}$ as candidate algorithms for attacks on watermarks.

Define a zero-sum game $\mathcal{G}$ between $(\mathbf{V}, \mathbf{P}) \in \mathfrak{Candidate}_{\mathfrak{W}} \times \mathfrak{Candidate}_{\mathfrak{D}}$. The payoff is given by

$$\begin{aligned}
\mathcal{G}(\mathbf{V}, \mathbf{P}) = &\frac{1}{2}\,\mathbb{P}_{(f,\mathbf{x}):=\mathbf{V}}\Big[\text{err}(f) > \epsilon \text{ or } \text{err}(\mathbf{x}, \mathbf{y}) \leq 2\epsilon \text{ or } b = 1\Big]\\
&+ \frac{1}{2}\,\mathbb{P}_{f:=\mathbf{V}, \mathbf{x}\sim D^q}\Big[\text{err}(f) > \epsilon \text{ or } \Big(\text{err}(\mathbf{x}, \mathbf{y}) \leq 2\epsilon \text{ and } b = 0\Big)\Big],
\end{aligned}$$

where $\mathbf{V}$ tries to minimize and $\mathbf{P}$ maximize the payoff.

Applying Theorem 3 to $\mathcal{G}$ with $\eta = 2^{-5}$ we get two probability distributions, p over a multiset of pure strategies in $\mathfrak{Candidate}_{\mathfrak{W}}$ and r over a multiset of pure strategies in $\mathfrak{Candidate}_{\mathfrak{D}}$ that lead to a 2^{-5}-approximate Nash equilibrium.

The size k of the multisets is bounded

$$\begin{aligned}
k &\leq 2^6 \log\left(|\mathfrak{Candidate}_{\mathfrak{W}}|\right)\\
&\leq 2^6 \log\left(2^{100\log\left(T^{\frac{1}{2^{10}\sqrt{\log(T)}}}\right)}\right) && \text{Because of the number of possible succinct circuits}\\
&\leq 2^{13} \log\left(T^{\frac{1}{2^{10}\sqrt{\log(T)}}}\right)\\
&\leq 2^3\sqrt{\log(T)}. && (1)
\end{aligned}$$

Next, observe that the mixed strategy corresponding to the distribution p can be represented by a succinct circuit of size

$$k \cdot 100 \log\left(T^{\frac{1}{2^{10}\sqrt{\log(T)}}}\right) \leq \frac{k}{2^3}\sqrt{\log(T)}, \tag{2}$$

because we can create a circuit that is a collection of k circuits corresponding to the multiset of p, where each one is of size $100 \log\left(T^{\frac{1}{2^{10}\sqrt{\log(T)}}}\right)$. Combining (1) and (2) we get that the size of the circuit succinctly representing strategy p is bounded by

$$\begin{aligned}
&\frac{k}{2^3}\sqrt{\log(T)} \\
&\leq 2^3\sqrt{\log(T)} \cdot \frac{1}{2^3}\sqrt{\log(T)} \\
&\leq \log(T).
\end{aligned}$$

This implies that p can be implemented by a T-sized succinctly representable circuit. The same hold for r. Let's call the strategy corresponding to p, $\mathbf{V}_{\text{Nash}}$, and the strategy corresponding to r, $\mathbf{P}_{\text{Nash}}$.

Consider cases:

Case $\mathcal{G}(\mathbf{V}_{\text{NASH}}, \mathbf{P}_{\text{NASH}}) \geq \frac{19}{24}$. Define $\mathbf{P}_{\text{DEFENSE}}$ to work as follows:

1. Simulate $f := \mathbf{V}_{\text{LEARN}}$, where $\mathbf{V}_{\text{LEARN}}$ is a circuit of size T, such that
$$\mathbb{P}\Big[\text{err}(f) \leq \epsilon\Big] \geq 1 - \frac{1}{48}.$$
2. Send f to $\mathbf{V}$.
3. Receive $\mathbf{x}$ from $\mathbf{V}$.
4. Simulate $(\mathbf{y}, b) := \mathbf{P}_{\text{NASH}}(f, \mathbf{x})$.
5. Return $b' = 1$ if $b = 1$ or $d(f(\mathbf{x}), \mathbf{y}) > 3\epsilon q$ and $b' = 0$ otherwise,

where $d(\cdot, \cdot)$ is the Hamming distance. Note that $\mathbf{P}_{\text{DEFENSE}}$ runs in time $2T$ and not T because it first simulates $\mathbf{V}_{\text{LEARN}}$ and then simulates $\mathbf{P}_{\text{NASH}}$.

We claim that

$$\mathbf{P}_{\text{DEFENSE}} \in \text{DEFENSE}\left(\mathcal{L}, \epsilon, q, T^{\frac{1}{2^{10}\sqrt{\log(T)}}}, 2T, l = 1 - \frac{1}{48}, c = \frac{13}{24}, s = \frac{11}{24}\right). \tag{3}$$

Assume towards contradiction that completeness or soundness of $\mathbf{P}_{\text{DEFENSE}}$ as defined in Definition 6 does not hold.

If completeness of $\mathbf{P}_{\text{DEFENSE}}$ does not hold then

$$\mathbb{P}_{\mathbf{x} \sim D^q}\Big[b' = 0\Big] < \frac{13}{24}. \tag{4}$$

Let's compute the payoff of $\mathbf{V}$, which first runs $f := \mathbf{V}_{\text{LEARN}}$ and sets $\mathbf{x} \sim D^q$, in the game $\mathcal{G}$, when playing against $\mathbf{P}_{\text{NASH}}$.

$$
\begin{aligned}
&\mathcal{G}(\mathbf{V}, \mathbf{P}_{\text{NASH}}) \\
&= \frac{1}{2}\mathbb{P}_{(f,\mathbf{x}):=\mathbf{V}}\Big[\text{err}(f) > \epsilon \text{ or } \text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon \text{ or } b' = 1\Big] \\
&+ \frac{1}{2}\mathbb{P}_{f:=\mathbf{V},\mathbf{x}\sim D^q}\Big[\text{err}(f) > \epsilon \text{ or } \Big(\text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon \text{ and } b' = 0\Big)\Big] \\
&\le \delta + \frac{1}{2}\mathbb{P}_{f:=\mathbf{V}_{\text{LEARN}},\mathbf{x}\sim D^q}\Big[\text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon \text{ or } b' = 1\Big] \\
&+ \frac{1}{2}\mathbb{P}_{f:=\mathbf{V}_{\text{LEARN}},\mathbf{x}\sim D^q}\Big[\text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon \text{ and } b' = 0\Big] && \text{By def. of } \mathbf{V}, \mathbf{P}_{\text{DEFENSE}} \text{ and } \mathbb{P}\Big[\text{err}(f) \le \epsilon\Big] \ge 1 - \frac{1}{48} \\
&< \frac{1}{48} + \frac{1}{2} + \frac{\frac{13}{24}}{2} && \text{By (4)} \\
&= \frac{38}{48} \\
&\le \mathcal{G}(\mathbf{V}_{\text{NASH}}, \mathbf{P}_{\text{NASH}}), \ \lightning
\end{aligned}
$$

where the contradiction is with the properties of Nash equilibria.

Assume that $\mathbf{V}$ breaks the soundness of $\mathbf{P}_{\text{DEFENSE}}$, which translates to

$$
\mathbb{P}_{\mathbf{x}:=\mathbf{V}(f)}\Big[\text{err}(\mathbf{x}, f(\mathbf{x})) > 7\epsilon \ \text{ and } \ b = 0 \ \text{ and } \ d(f(\mathbf{x}),\mathbf{y})) > 3\epsilon q\Big] > \frac{11}{24}. \tag{5}
$$

Let $\mathbf{V}'$ first simulate $f := \mathbf{V}_{\text{LEARN}}$, then runs $\mathbf{x} := \mathbf{V}(f)$, and returns $(f, \mathbf{x})$. We have

$$
\begin{aligned}
&\mathcal{G}(\mathbf{V}', \mathbf{P}_{\text{NASH}}) \\
&= \frac{1}{2}\,\mathbb{P}_{(f,\mathbf{x}):=\mathbf{V}'}\Big[\text{err}(f) > \epsilon \text{ or } \text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon \text{ or } b' = 1\Big] \\
&+ \frac{1}{2}\,\mathbb{P}_{f:=\mathbf{V}',\mathbf{x}\sim D^q}\Big[\text{err}(f) > \epsilon \text{ or } \Big(\text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon \text{ and } b' = 0\Big)\Big] \\
&= \frac{1}{2}\,\mathbb{P}_{f:=\mathbf{V}_{\text{LEARN}},\mathbf{x}=\mathbf{V}(f)}\Big[\text{err}(f) > \epsilon \text{ or } \text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon \text{ or } b' = 1\Big] \\
&+ \frac{1}{2}\,\mathbb{P}_{f:=\mathbf{V}_{\text{LEARN}},\mathbf{x}\sim D^q}\Big[\text{err}(f) > \epsilon \text{ or } \Big(\text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon \text{ and } b' = 0\Big)\Big] && \text{By def. of } \mathbf{V}' \\
&< \frac{1}{2} + \frac{1 - \frac{11}{24}}{2} && \text{By (5)} \\
&= \frac{37}{48} \\
&\le \mathcal{G}(\mathbf{V}_{\text{NASH}}, \mathbf{P}_{\text{NASH}}), \ \lightning
\end{aligned}
$$

where the contradiction is with the properties of Nash equilibria. Thus (3) holds.

Case $\mathcal{G}(\mathbf{V}_{\text{NASH}}, \mathbf{P}_{\text{NASH}}) < \frac{19}{24}$. Consider $\mathbf{P}$ that returns $(f(\mathbf{x}), b)$ for a uniformly random b. We have

$$
\begin{aligned}
&\mathcal{G}(\mathbf{V}_{\text{NASH}}, \mathbf{P}) \\
&\ge \Big(1 - \mathbb{P}_{f:=\mathbf{V}_{\text{NASH}}}\Big[\text{err}(f) \le \epsilon\Big]\Big) + \mathbb{P}_{f:=\mathbf{V}_{\text{Nash}}}\Big[\text{err}(f) \le \epsilon\Big] \cdot \frac{1}{2},
\end{aligned}
$$

because when $\mathbf{x} \sim D^q$ and $\text{err}(f) \le \epsilon$ the probability that $\text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon$ and $b = 0$ is $\frac{1}{2}$, and similarly when $\mathbf{x} := \mathbf{V}_{\text{NASH}}$ then the probability that $b = 1$ is equal $\frac{1}{2}$. The assumption that $\mathcal{G}(\mathbf{V}_{\text{Nash}}, \mathbf{P}) < \frac{19}{24}$ and properties of Nash equilibria imply that $\mathbb{P}_{f:=\mathbf{V}_{\text{Nash}}}[\text{err}(f) \le \epsilon] \ge \frac{10}{24}$. This implies that *correctness* holds for $\mathbf{V}_{\text{Nash}}$ with $l = \frac{10}{24}$.

Next, assume towards contradiction that *unremovability* of $\mathbf{V}_{\text{NASH}}$ does not hold, i.e. there is $\mathbf{P}$ running in time t such that $\mathbb{P}\big[\text{err}(\mathbf{x},\mathbf{y}) \le 2\epsilon\big] > \frac{19}{24}$. Consider $\mathbf{P}'$ that on input $(f, \mathbf{x})$ returns $(\mathbf{P}(f, \mathbf{x}), 0)$. Then by definition of $\mathcal{G}$, $\mathcal{G}(\mathbf{V}_{\text{NASH}}, \mathbf{P}') > \frac{19}{24}$, which is a contradiction $\lightning$.

Next, assume towards contradiction that *undetectability* of $\mathbf{V}_{\text{NASH}}$ does not hold, i.e. there exists $\mathbf{P}$ such that it distinguishes $\mathbf{x} \sim D^q$ from $\mathbf{x} := \mathbf{V}_{\text{NASH}}$ with probability higher than $\frac{19}{24}$. Consider $\mathbf{P}'$ that on input $(f, \mathbf{x})$ returns $(f(\mathbf{x}), \mathbf{P}(f, \mathbf{x}))$.[5] Then by definition of $\mathcal{G}$, $\mathcal{G}(\mathbf{V}_{\text{NASH}}, \mathbf{P}') > \frac{19}{24}$, which is a contradiction $\lightning$

There are two further subcases. If $\mathbf{V}_{\text{NASH}}$ satisfies *uniqueness* then

$$\mathbf{V}_{\text{NASH}} \in \text{WATERMARK}\left(\mathcal{L}, \epsilon, q, T, T^{\frac{1}{2^{10}\sqrt{\log(T)}}}, l = \frac{10}{24}, c = \frac{21}{24}, s = \frac{19}{24}\right).$$

If $\mathbf{V}_{\text{NASH}}$ does not satisfy *uniqueness* then, by definition, every succinctly representable circuit $\mathbf{P}$ of size T satisfies $\text{err}(\mathbf{x}, \mathbf{y}) \leq 2\epsilon$ with probability at most $\frac{21}{24}$. Consider the following $\mathbf{V}$. It computes $(f, \mathbf{x}) := \mathbf{V}_{\text{Nash}}$, ignores f and sends $\mathbf{x}$ to $\mathbf{P}$. By the assumption that *uniqueness* is not satisfied for $\mathbf{V}_{\text{NASH}}$ we have that *transferability* of Definition 3 holds for $\mathbf{V}$ with $c = \frac{3}{24}$. Note that $\mathbf{P}$ in the transferable attack does not receive f but it makes it no easier for it to satisfy the properties. Note that *undetectability* still holds with the same parameter. Thus

$$\mathbf{V}_{\text{NASH}} \in \text{TRANSFATTACK}\left(\mathcal{L}, \epsilon, q, T, T, c = \frac{3}{24}, s = \frac{19}{24}\right).$$

□

E. Missing Proofs of Section 5

E.1. Adversarial Defenses exist

Our result is based on (Goldwasser et al., 2020). Before we state and prove our result we give an overview of the learning model considered in (Goldwasser et al., 2020).

E.2. Transductive learning with rejections.

In (Goldwasser et al., 2020) the authors consider a model, where a learner $\mathbf{L}$ receives a training set of labeled samples from the original distribution $(\mathbf{x}_D, \mathbf{y}_D = h(\mathbf{x}_D)), \mathbf{x} \sim D^N, \mathbf{y}_D \in \{-1, +1\}^N$, where h is the ground truth, together with a test set $\mathbf{x}_T \in \mathcal{X}^q$. Next, $\mathbf{L}$ uses $(\mathbf{x}_D, \mathbf{y}_D, \mathbf{x}_T)$ to compute $\mathbf{y}_T \in \{-1, +1, \square\}^q$, where $\square$ represents that $\mathbf{L}$ abstains (rejects) from classifying the corresponding x.

Before we define when learning is successful, we will need some notation. For $q \in \mathbb{N}, \mathbf{x} \in \mathcal{X}^q, \mathbf{y} \in \{-1, +1, \square\}^q$ we define

$$\text{err}(\mathbf{x}, \mathbf{y}) := \frac{1}{q} \sum_{i \in [q]} \mathbb{1}_{\left\{h(x(i)) \neq y(i), y(i) \neq \square, h(x(i)) \neq \perp\right\}}, \quad \square(\mathbf{y}) := \frac{1}{q}\left|\left\{i \in [q] : y(i) = \square\right\}\right|,$$

which means that we count $(x, y) \in \mathcal{X} \times \{-1, +1, \square\}$ as an error if h is well defined on x, y is not an abstantion and $h(x) \neq y$.

Learning is successful if it satisfies two properties.

- If $\mathbf{x}_T \sim D^q$ then with high probability $\text{err}(\mathbf{x}_T, \mathbf{y}_T)$ and $\square(\mathbf{y}_T)$ are small.
- For **every** $\mathbf{x}_T \in \mathcal{X}^q$ with high probability $\text{err}(\mathbf{x}_T, \mathbf{y}_T)$ is small.[6]

The formal guarantee of a result from (Goldwasser et al., 2020) are given in Theorem 5. Let's call this model Transductive Learning with Rejections (TLR).

Note the differences between TLR and our definition of Adversarial Defenses. To compare the two models we associate the learner $\mathbf{L}$ from TLR with $\mathbf{P}$ in our setup, and the party producing $\mathbf{x}_T$ with $\mathbf{V}$ in our definition. First, in TLR, $\mathbf{P}$ does not send f to $\mathbf{V}$. Secondly, and most importantly, we don't allow $\mathbf{P}$ to reply with rejections ($\square$) but instead require that $\mathbf{P}$ can "distinguish" that it is being tested (see soundness of Definition 6). Finally, there are no apriori time bounds on either $\mathbf{V}$ or $\mathbf{P}$ in TLR. The models are similar but a priori incomparable and any result for TLR needs to be carefully analyzed before being used to prove that it is an Adversarial Defense.

[5] Formally $\mathbf{P}$ receives as input $(f, \mathbf{x})$ and not only $\mathbf{x}$.

[6] Note that, crucially, in this case $\square(\mathbf{y}_T)$ might be very high, e.g. equal to 1.

E.3. Formal guarantee for Transductive Learning with Rejections (TLR)

Theorem 5.3 from (Goldwasser et al., 2020) adapted to our notation reads.

Theorem 5 (**TLR guarantee ((Goldwasser et al., 2020))**). *For any $N \in \mathbb{N}, \epsilon \in (0,1), h \in \mathcal{H}$ and distribution D over $\mathcal{X}$:*

$$\mathbb{P}_{\mathbf{x}_D, \mathbf{x}'_D \sim D^N}\left[\forall\, \mathbf{x}_T \in \mathcal{X}^N : err(\mathbf{x}_T, f(\mathbf{x}_T)) \leq \epsilon^* \wedge \square\left(f\left(\mathbf{x}'_D\right)\right) \leq \epsilon^*\right] \geq 1-\epsilon,$$

where $\epsilon^* = \sqrt{\frac{2d}{N}\log(2N) + \frac{1}{N}\log\left(\frac{1}{\epsilon}\right)}$ and $f = \textsc{Rejectron}(\mathbf{x}_D, h(\mathbf{x}_D), \mathbf{x}_T, \epsilon^*)$, where $f : \mathcal{X} \to \{-1, +1, \square\}$ and d denotes the VC-dimension on $\mathcal{H}$. Rejectron *is defined in Figure 2. in (Goldwasser et al., 2020).*

Rejectron is an algorithm that accepts a labeled training set $(\mathbf{x}_D, h(\mathbf{x}_D))$ and a test set $\mathbf{x}_T$ and returns a classifier f, which might reject some inputs. The learning is successful if with a high probability f rejects a small fraction of D^N and for every $\mathbf{x}_T \in \mathcal{X}^N$ the error on labeled x's in $\mathbf{x}_T$ is small.

E.4. Adversarial Defense for bounded VC-dimension

We are ready to state the main result of this section.

Lemma 1 (**Adversarial Defense for bounded VC-dimension**). *Let $d \in \mathbb{N}$ and $\mathcal{H}$ be a binary hypothesis class on input space $\mathcal{X}$ of VC-dimension bounded by d. There exists an algorithm $\mathbf{P}$ such that for every $\epsilon \in \left(0, \frac{1}{8}\right)$, D over $\mathcal{X}$ and $h \in \mathcal{H}$ we have*

$$\mathbf{P} \in \textsc{Defense}\Bigg((D,h), \epsilon, q = \frac{d\log^2(d)}{\epsilon^3}, t = \infty, T = \texttt{poly}\left(\frac{d}{\epsilon}\right), l = 1-\epsilon, c = 1-\epsilon, s = \epsilon \Bigg).$$

Proof. The proof is based on an algorithm from (Goldwasser et al., 2020).

Construction of $\mathbf{P}$. Let $\epsilon \in (0,1)$ and

$$N := \frac{d\log^2(d)}{\epsilon^3}.$$

Let $q := N$. First, $\mathbf{P}$, draws N labeled samples $(\mathbf{x}_{\text{FRESH}}, h(\mathbf{x}_{\text{FRESH}}))$. Next, it finds $f \in \mathcal{H}$ consistent with them and sends f to $\mathbf{V}$. Importantly this computation is the same as the first step of Rejectron.

Next, $\mathbf{P}$ receives as input $\mathbf{x} \in \mathcal{X}^q$ from $\mathbf{V}$. $\mathbf{P}$. Let $\epsilon^* := \sqrt{\frac{2d}{N}\log(2N) + \frac{1}{N}\log\left(\frac{1}{\epsilon}\right)}$. Next $\mathbf{P}$ runs $f' = \textsc{Rejectron}(\mathbf{x}_{\text{FRESH}}, h(\mathbf{x}_{\text{FRESH}}), \mathbf{x}, \epsilon^*)$, where Rejectron is starting from the second step of the algorithm (Figure 2 (Goldwasser et al., 2020)). Importantly, for every $x \in \mathcal{X}$, if $f'(x) \neq \square$ then $f(x) = f'(x)$. In words, f' is equal to f everywhere where f' doesn't reject.

Finally $\mathbf{P}$ returns 1 if $\square(f'(\mathbf{x})) > \frac{2}{3}\epsilon$, and returns 0 otherwise.

$\mathbf{P}$ is a defense. First, by the standard PAC theorem we have that with probability at least $1-\epsilon$, $\text{err}(f) \leq \frac{\epsilon}{2}$. This means that *correctness* holds with probability $l = 1-\epsilon$.

Note that with our setting of N, we have that

$$\epsilon^* \leq \frac{\epsilon}{2}.$$

Theorem 5 guarantees that

- if $\mathbf{x} \in D^q$ then with probability at least $1-\epsilon$ we have that

  $$\square(f'(\mathbf{x})) \leq \frac{\epsilon}{2}.$$

  which in turn implies that with the same probability $\mathbf{P}$ returns $b = 0$. This implies that *completeness* holds with probability $1-\epsilon$.

- for every $\mathbf{x} \in \mathcal{X}^q$ with probability at least $1-\epsilon$ we have that

$$\text{err}(\mathbf{x}, f'(\mathbf{x})) \leq \frac{\epsilon}{2}.$$

To compute soundness we want to upper bound the probability that $\text{err}(\mathbf{x}, f(\mathbf{x})) > 2\epsilon$[7] and $b = 0$. By construction of $\mathbf{P}$ if $b = 0$ then $\square(f'(\mathbf{x})) \leq \frac{2\epsilon}{3}$, which means that with probability at least $1-\epsilon$

$$\text{err}(\mathbf{x}, \mathbf{y}) \leq \frac{2\epsilon}{3} + \frac{\epsilon}{2} < 2\epsilon \text{ or } b = 1.$$

This translates to *soundness* holding with $s = \epsilon$.

REJECTRON runs in polynomial time in the number of samples and makes $O(\frac{1}{\epsilon})$ calls to an Empirical Risk Minimizer on $\mathcal{H}$ (that we assume runs in time polynomial in d), which implies the promised running time. □

E.5. Watermark Example

Lemma 2 (Watermark for bounded VC-dimension against fast adversaries). *For every $d \in \mathbb{N}$ there exists a distribution D and a binary hypothesis class $\mathcal{H}$ of VC-dimension d there exists $\mathbf{V}$ such that for any $\epsilon \in \left(\frac{10000}{d}, \frac{1}{8}\right)$ if $h \in \mathcal{H}$ is taken uniformly at random from $\mathcal{H}$ then*

$$\mathbf{V} \in \text{WATERMARK}\Bigg((D,h), \epsilon, q = O\left(\frac{1}{\epsilon}\right), T = O\left(\frac{d}{\epsilon}\right), t = \frac{d}{100}, l = 1 - \frac{1}{100}, c = 1 - \frac{2}{100}, s = \frac{56}{100} \Bigg).$$

Proof. Let $\mathcal{X} = \mathbb{N}$. Let D be the uniform distribution over $[N]$ for $N = 100d^2$. Let $\mathcal{H}$ be the concept class of functions that have exactly d $+1$'s in $[N]$. Note $\mathcal{H}$ has VC-dimension d. Let $h \in \mathcal{H}$ be the ground truth.

Construction of $\mathbf{V}$. $\mathbf{V}$ works as follows. It draws $n = O\left(\frac{d}{\epsilon}\right)$ samples from D labeled with h. Let's call them $\mathbf{x}_{\text{TRAIN}}$. Let

$$A := \{x \in [N] : \mathbf{x}_{\text{TRAIN}}, h(x) = +1\}, B := \{x \in [N] : x \in \mathbf{x}_{\text{TRAIN}}, h(x) = -1\}.$$

$\mathbf{V}$ takes a uniformly random subset $A_w \subseteq A$ of size q. It defines sets

$$A' := A \setminus A_w,\ B' := B \cup A_w.$$

$\mathbf{V}$ computes f consistent with the training set $\{(x,+1) : x \in A'\} \cup \{(x,-1) : x \in B'\}$. $\mathbf{V}$ samples $S \sim D^q$. It defines the watermark to be $\mathbf{x} := A_w$ with probability $\frac{1}{2}$ and $\mathbf{x} := S$ with probability $\frac{1}{2}$.

$\mathbf{V}$ sends $(f, \mathbf{x})$ to $\mathbf{P}$. $\mathbf{V}$ can be implemented in time $O\left(\frac{d}{\epsilon}\right)$.

$\mathbf{V}$ is a watermark. We claim that $(f, \mathbf{x})$ constitutes a watermark.

It is possible to construct a watermark of prescribed size, i.e. find a subset A_w of a given size, only if $|A| \geq q$. The probability that a single sample from D is labeled $+1$ is $\frac{d}{N}$, so by the Chernoff bound (Fact 1) $|A|, |B| > \frac{dn}{2N} \geq q$ with probability $1 - \frac{1}{100}$, where we used that $n = O\left(\frac{d}{\epsilon}\right), N = 100d^2, q = O(\frac{1}{\epsilon})$.

Correctness property. Let $h'(x) := h(x)$ if $x \in [N] \setminus A_w$ and $h'(x) := -h(x)$ otherwise. Note that h' has exactly $d - q$ $+1$'s in $[N]$. By construction, f is a classifier consistent with h'. By the PAC theorem we know that with probability $1 - \frac{1}{100}$, f has an error at most ϵ wrt to h' (because the hypothesis class of functions with *at most* d $+1$'s has a VC dimension of $O(d)$). h' differs from h on q points, so

$$\text{err}(f) \leq \epsilon + q/N = O\left(\epsilon + \frac{1}{\epsilon d^2}\right) = O(\epsilon). \tag{6}$$

with probability $1 - \frac{1}{100}$, which implies that *correctness* is satisfied with $l = 1 - \frac{1}{100}$.

[7] Note that we measure the error of f not f'.

Distinguishing of $\mathbf{x}$ and D^q. Note that the distribution of A_w is the same as the distribution of a uniformly random subset of $[N]$ of size q (when taking into account the randomness of the choice of $h \sim U(\mathcal{H})$). Observe that the probability that drawing q i.i.d. samples from $U([N])$ we encounter repetitions is at most

$$\frac{1}{N} + \frac{2}{N} + \cdots + \frac{q}{N} \leq \frac{3q^2}{N} \leq \frac{1}{100},$$

because $q < \frac{d}{100} < \frac{\sqrt{N}}{10}$. This means that $\frac{1}{100}$ is an information-theoretic upper bound on the distinguishing advantage between $\mathbf{x} = A_w$ and D^q.

Moreover, $\mathbf{P}$ has access to at most t samples and the probability that the set of samples $\mathbf{P}$ draws from D^t and A_w have empty intersection is at least $1 - \frac{1}{100}$. It is because it is at least $(1 - \frac{t}{N})^t \geq (1 - \frac{1}{\sqrt{N}})^{\sqrt{N}/10} \geq 1 - \frac{1}{100}$, where we used that $t < \frac{\sqrt{N}}{10}$.[8]

Note that by construction f maps all elements of A_w to -1. The probability over the choice of $F \sim D^q$ that $F \subseteq h^{-1}(\{-1\})$, i.e. all elements of F have true label -1, is at least

$$\left(1 - \frac{d}{N}\right)^q \geq 1 - \frac{1}{100}.$$

The three above observations and the union bound imply that the distinguishing advantage for distinguishing $\mathbf{x}$ from D^q of $\mathbf{P}$ is at most $\frac{4}{100}$ and so the *undetectability* holds with $s = \frac{8}{100}$.

Unremovability property. Assume, towards contradiction with *unremovability*, that $\mathbf{P}$ can find $\mathbf{y}$ that with probability $s' = \frac{1}{2} + \frac{6}{100}$ satisfies $\text{err}(\mathbf{x}, \mathbf{y}) \leq 2\epsilon$. Notice, that $\text{err}(A_w, f(A_w)) = 1$ by construction.

Consider an algorithm $\mathbf{A}$ for distinguishing A_w from D^q. Upon receiving $(f, \mathbf{x})$ it first runs $\mathbf{y} = \mathbf{P}(f, \mathbf{x})$ and returns 1 iff $d(\mathbf{y}, f(\mathbf{x})) \geq \frac{q}{2}$. We know that the distinguishing advantage is at most $\frac{1}{2} + \frac{4}{100}$, so

$$\frac{1}{2}\mathbb{P}_{\mathbf{x}:=A_w}[\mathbf{A}(f, \mathbf{x}) = 1] + \frac{1}{2}\mathbb{P}_{\mathbf{x}\sim D^q}[\mathbf{A}(f, \mathbf{x}) = 0] \leq \frac{1}{2} + \frac{4}{100}.$$

But also note that

$$\begin{aligned}
s' &\leq \mathbb{P}_{\mathbf{x}\sim\mathbf{V}}[\text{err}(\mathbf{x}, \mathbf{y}) \leq 2\epsilon] \\
&\leq \frac{1}{2}\mathbb{P}_{\mathbf{x}:=A_w}[d(\mathbf{y}, f(\mathbf{x})) \geq (1 - 2\epsilon)q] + \frac{1}{2}\mathbb{P}_{\mathbf{x}\sim D^q}[d(\mathbf{y}, f(\mathbf{x})) \leq (2\epsilon + \text{err}(f))q] \\
&\leq \frac{1}{2}\mathbb{P}_{\mathbf{x}:=A_w}[d(\mathbf{y}, f(\mathbf{x})) \geq q/2] + \frac{1}{2}\mathbb{P}_{\mathbf{x}\sim D^q}[d(\mathbf{y}, f(\mathbf{x})) \leq q/2] + \frac{1}{100} \\
&\leq \frac{1}{2}\mathbb{P}_{\mathbf{x}:=A_w}[\mathbf{A}(f, \mathbf{x}) = 1] + \frac{1}{2}\mathbb{P}_{\mathbf{x}\sim D^q}[\mathbf{A}(f, \mathbf{x}) = 0] + \frac{1}{100}.
\end{aligned}$$

Combining the two above equations we get a contradiction and thus the *unremovability* holds with $s' = \frac{1}{2} + \frac{6}{100}$.

Uniqueness property. The following $\mathbf{P}$ certifies uniqueness. It draws $O\left(\frac{d}{\epsilon}\right)$ samples from D, let's call them $\mathbf{x}'_{\text{TRAIN}}$ and trains f' consistent with it. By the PAC theorem $\text{err}(f') \leq \epsilon$ with probability at least $1 - \frac{1}{100}$. Next upon receiving $\mathbf{x} \in \mathcal{X}^q = [N]^q$ it returns $y = f'(\mathbf{x})$. By the fact that $\mathbf{x}$ is a random subset of $[N]$ of size q by the Chernoff bound, the union bound we know that $\text{err}(\mathbf{x}, \mathbf{y}) = \text{err}(\mathbf{x}, f'(\mathbf{x})) \leq 2\epsilon$ with probability at least $1 - \frac{2}{100}$ over the choice of h. This proves *uniqueness*. □

E.6. Transferable Attacks exist

E.6.1. Fully Homomorphic Encryption (FHE)

We include a definition of fully holomorphic encryption based on the definition from (Goldwasser et al., 2013). The notion of fully homomorphic encryption was first proposed by Rivest, Adleman and Dertouzos (Rivest et al., 1978) in 1978. The first fully homomorphic encryption scheme was proposed in a breakthrough work by Gentry in 2009 (Gentry, 2009). A history and recent developments on fully homomorphic encryption is surveyed in (Vaikuntanathan, 2011).

[8] If the sets were not disjoint then $\mathbf{P}$ could see it as suspicious because f makes mistakes on all of A_w.

E.6.2. PRELIMINARIES

We say that a function f is *negligible* in an input parameter λ, if for all $d > 0$, there exists K such that for all $\lambda > K$, $f(\lambda) < \lambda^{-d}$. For brevity, we write: for all sufficiently large λ, $f(\lambda) = \text{negl}(\lambda)$. We say that a function f is *polynomial* in an input parameter λ, if there exists a polynomial p such that for all λ, $f(\lambda) \leq p(\lambda)$. We write $f(\lambda) = \text{poly}(\lambda)$. A similar definition holds for $\text{polylog}(\lambda)$. For two polynomials p, q, we say $p \leq q$ if for every $\lambda \in \mathbb{N}$, $p(\lambda) \leq q(\lambda)$.

When saying that a Turing machine $\mathbf{A}$ is p.p.t. we mean that $\mathbf{A}$ is a non-uniform probabilistic polynomial-time machine.

E.6.3. DEFINITIONS

Definition 8 ((Goldwasser et al., 2013))**.** *A homomorphic (public-key) encryption scheme* FHE *is a quadruple of polynomial time algorithms (*FHE.KEYGEN, FHE.ENC, FHE.DEC, FHE.EVAL*) as follows:*

- FHE.KEYGEN(1^λ) *is a probabilistic algorithm that takes as input the security parameter* 1^λ *and outputs a public key* pk *and a secret key* sk*.*
- FHE.ENC$(pk, x \in \{0,1\})$ *is a probabilistic algorithm that takes as input the public key* pk *and an input bit* x *and outputs a ciphertext* ψ*.*
- FHE.DEC(sk, ψ) *is a deterministic algorithm that takes as input the secret key* sk *and a ciphertext* ψ *and outputs a message* $x^* \in \{0,1\}$*.*
- FHE.EVAL$(pk, C, \psi_1, \psi_2, \ldots, \psi_n)$ *is a deterministic algorithm that takes as input the public key* pk*, some circuit* C *that takes* n *bits as input and outputs one bit, as well as* n *ciphertexts* $\psi_1, \ldots, \psi_n$*. It outputs a ciphertext* ψ_C*.*

Compactness: *For all security parameters* λ*, there exists a polynomial* $p(\cdot)$ *such that for all input sizes* n*, for all* $x_1, \ldots, x_n$*, for all* C*, the output length of* FHE.EVAL *is at most* $p(n)$ *bits long.*

Definition 9 (C-homomorphism, (Goldwasser et al., 2013))**.** *Let* $C = \{C_n\}_{n\in\mathbb{N}}$ *be a class of boolean circuits, where* C_n *is a set of boolean circuits taking* n *bits as input. A scheme FHE is* C*-homomorphic if for every polynomial* $n(\cdot)$*, for every sufficiently large security parameter* λ*, for every circuit* $C \in C_n$*, and for every input bit sequence* $x_1, \ldots, x_n$*, where* $n = n(\lambda)$*,*

$$\mathbb{P}\left[\begin{array}{c}(pk, sk) \leftarrow \text{FHE.KEYGEN}(1^\lambda);\\ \psi_i \leftarrow \text{FHE.ENC}(pk, x_i) \text{ for } i = 1 \ldots n;\\ \psi \leftarrow \text{FHE.EVAL}(pk, C, \psi_1, \ldots, \psi_n) :\\ \text{FHE.DEC}(sk, \psi) \neq C(x_1, \ldots, x_n)\end{array}\right] = \textit{negl}(\lambda).$$

where the probability is over the coin tosses of FHE.KEYGEN *and* FHE.ENC*.*

Definition 10 (Fully homomorphic encryption)**.** *A scheme FHE is fully homomorphic if it is homomorphic for the class of all arithmetic circuits over* $\mathbb{GF}(2)$*.*

Definition 11 (Leveled fully homomorphic encryption)**.** *A leveled fully homomorphic encryption scheme is a homomorphic scheme where* FHE.KEYGEN *receives an additional input* 1^d *and the resulting scheme is homomorphic for all depth-*d *arithmetic circuits over* $\mathbb{GF}(2)$*.*

Definition 12 (IND-CPA security)**.** *A scheme FHE is IND-CPA secure if for any p.p.t. adversary* $\mathbf{A}$*,*

$$\Big|\, \mathbb{P}\left[(pk, sk) \leftarrow \text{FHE.KEYGEN}(1^\lambda) : \mathbf{A}(pk, \text{FHE.ENC}(pk, 0)) = 1\right] +$$
$$- \mathbb{P}\left[(pk, sk) \leftarrow \text{FHE.KEYGEN}(1^\lambda) : \mathbf{A}(pk, \text{FHE.ENC}(pk, 1)) = 1\right] \Big| = \textit{negl}(\lambda).$$

We now state the result of Brakerski, Gentry, and Vaikuntanathan (Brakerski et al., 2012) that shows a leveled fully homomorphic encryption scheme based on a standard assumption in cryptography called Learning with Errors ((Regev, 2005)):

Theorem 6 (Fully Homomorphic Encryption, definition from (Goldwasser et al., 2013)). *Assume that there is a constant* $0 < \epsilon < 1$ *such that for every sufficiently large* ℓ*, the approximate shortest vector problem gapSVP in* ℓ *dimensions is hard to approximate to within a* $2^{O(\ell^\epsilon)}$ *factor in the worst case. Then, for every* n *and every polynomial* $d = d(n)$*, there is an*

IND-CPA secure d-leveled fully homomorphic encryption scheme where encrypting n bits produces ciphertexts of length poly$(n, \lambda, d^{1/\epsilon})$, the size of the circuit for homomorphic evaluation of a function f is size$(C_f) \cdot$ poly$(n, \lambda, d^{1/\epsilon})$ and its depth is depth$(C_f) \cdot$ poly$(\log n, \log d)$.

Learning theory preliminaries. For the next lemma, we will consider a slight generalization of learning tasks to the case where there are many valid outputs for a given input. This can be understood as the case of generative tasks. We call a function $h : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ an error oracle for a learning task (D, h) if the error of $f : \mathcal{X} \to \mathcal{Y}$ is defined as

$$\text{err}(f) := \mathbb{E}_{x\sim D}[h(x, f(x))],$$

where the randomness of expectation includes the potential randomness of f. We assume that all parties have access to samples $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $x \sim D$ and $y \in \mathcal{Y}$ is some y such that $h(x, y) = 0$.

Definition 13 (Learning lines on a circle). *The input space is $\mathcal{X} = \{x \in \mathbb{R}^2 \mid \|x\|_2 = 1\}$, and the output space $\mathcal{Y} = \{-1, +1\}$. The hypothesis class is $\mathcal{H} = \{h_w \mid w \in \mathbb{R}^2, \|w\|_2 = 1\}$, where $h_w(x) := sgn(\langle w, x\rangle)$. Let $D = U(\mathcal{X})$ and $\mathcal{L} = (D, \mathcal{H})$. Note that $\mathcal{H}$ has VC-dimension equal to 2 so $\mathcal{L}$ is learnable to error ϵ with $O(\frac{1}{\epsilon})$ samples.*

Moreover, define $B_w(\alpha) := \{x \in \mathcal{X} \mid |\measuredangle(x, w)| \leq \alpha\}$.

Lemma 3 (Learning lower bound). *Let $\mathbf{L}$ be an algorithm for $(D, \mathcal{H})$ that uses K samples and returns a classifier f. Then*

$$\mathbb{P}_{w\sim U(\mathcal{X}), f\leftarrow\mathbf{L}}\left[\mathbb{P}_{x\sim U(\mathcal{X})}[f(x) \neq h_w(x)] \leq \frac{1}{2K}\right] \leq \frac{3}{100}.$$

Proof. Consider the following algorithm $\mathbf{A}$. It first simulates $\mathbf{L}$ on K samples to compute f. Next, it performs a smoothing of f, i.e. computes

$$f_\eta(x) := \begin{cases} +1, & \text{if } \mathbb{P}_{x'\sim U(B_x(2\pi\eta))}[f(x') = +1] > \mathbb{P}_{x'\sim U(B_x(2\pi\eta))}[f(x') = -1] \\ -1, & \text{otherwise.} \end{cases}$$

Note that if $\text{err}(f) \leq \eta$ for a ground truth h_w then for every $x \in \mathcal{X} \setminus B_x(2\pi\eta)$ we have $f_\eta(x) = h_w(x)$. This implies that $\mathbf{A}$ can be adapted to an algorithm that with probability 1 finds w' such that $|\measuredangle(w, w')| \leq \text{err}(f)$.

Assuming towards contradiction that the statement of the lemma doesn't hold it means that there is an algorithm using K samples that with probability $\frac{3}{100}$ locates w up to angle $\frac{1}{2K}$.

Consider any algorithm $\mathbf{A}$ using K samples. Probability that $\mathbf{A}$ doesn't see any sample in $B_w(2\pi\eta)$ is at least

$$(1 - 4\eta)^K \geq \left((1 - 4\eta)^{\frac{1}{4\eta}}\right)^{4\eta K} \geq \left(\frac{1}{2e}\right)^{4\eta K},$$

which is bigger than $1 - \frac{3}{100}$ if we set $\eta = \frac{1}{2K}$. But note that if there is no sample in $B_w(2\pi\eta)$ then $\mathbf{A}$ cannot locate w up to η with certainty. This proves the lemma. □

Lemma 4 (Boosting). *Let $\eta, \nu \in (0, \frac{1}{4})$, $\mathbf{L}$ be a learning algorithm for $(D, \mathcal{H})$ that uses K samples and outputs $f : \mathcal{X} \to \{-1, +1\}$ such that with probability δ*

$$\mathbb{P}_{w\sim U(\mathcal{X}), x\sim U(B_w(2\pi\eta))}[f(x) \neq h_w(x)] \leq \nu. \tag{7}$$

Then there exists a learning algorithm $\mathbf{L}'$ that uses $\max\left(K, \frac{9}{\eta}\right)$ samples such that with probability $\delta - \frac{1}{1000}$ returns f' such that

$$\mathbb{P}_{w\sim U(\mathcal{X}), x\sim U(\mathcal{X})}[f'(x) \neq h_w(x)] \leq 4\eta\nu.$$

Proof. Let $\mathbf{L}'$ first draws $\max\left(K, \frac{9}{\eta}\right)$ samples Q and defines $g : \mathcal{X} \to \{-1, +1, \perp\}$ as, g maps to -1 the smallest continuous interval containing all samples from Q with label -1. Similarly g maps to $+1$ the smallest continuous interval containing all samples from Q with label $+1$. The intervals are disjoined by construction. Unmapped points are mapped to

$\perp$. Next, $\mathbf{L}'$ simulates $\mathbf{L}$ with K samples and gets a classifier f that with probability δ satisfies the assumption of the lemma. Finally, it returns

$$f'(x) := \begin{cases} g(x), & \text{if } g(x) \neq \perp \\ f(x), & \text{otherwise.} \end{cases}$$

Consider 4 arcs defined as the 2 arcs constituting $B_w(2\pi\eta)$ divided into 2 parts each by the line $\{x \in \mathbb{R}^2 \mid \langle w, x\rangle = 0\}$. Let E be the event that some of these intervals don't contain a sample from Q. Observe that

$$\mathbb{P}[E] \leq 4(1-\eta)^{\frac{9}{\eta}} \leq \frac{1}{1000}.$$

By the union bound with probability $\delta - \frac{1}{1000}$, f satisfies (7) and E doesn't happen. By definition of f' this gives the statement of the lemma. □

Lemma 5 (Transferable Attack for a Cryptography based learning task). *There exists a polynomial p such that for every polynomial $r \geq p$[9] and for every sufficiently large security parameter $\lambda \in \mathbb{N}$ there exists a family of distributions $\mathcal{D}_\lambda = \{D^k_\lambda\}_k$, hypothesis class of error oracles $\mathcal{H}_\lambda = \{h^k_\lambda\}_k$, distribution $D_{\mathcal{L}}$ over k such that the following conditions are satisfied.*

1. *There exists $\mathbf{V}$ such that for all $\epsilon \in \left(\frac{1}{r(\lambda)}, \frac{1}{p(\lambda)}\right)$ if $k \sim D_{\mathcal{L}}$ then*

$$\mathbf{V} \in \textsc{TransfAttack}\Big(\left(D^k_\lambda, h^k_\lambda\right), \epsilon, q = \frac{16}{\epsilon}, T = \frac{10^3}{\epsilon^{1.3}}, t = \frac{1}{\epsilon^2}, c = 1 - \frac{1}{10}, s = \mathit{negl}(\lambda)\Big).$$

2. *There exists a learner $\mathbf{L}$ such that for every $\epsilon \in \left(\frac{1}{r(\lambda)}, \frac{1}{p(\lambda)}\right)$, with probability $1 - \frac{1}{10}$ over the choice of k and the internal randomness of $\mathbf{L}$, $\mathbf{L}$ returns a classifier of error at most ϵ. Additionally, $\mathbf{L}$ runs in time $\frac{10^3}{\epsilon^{1.3}}$ and uses $\frac{900}{\epsilon}$ samples.*

3. *For every $\epsilon \in \left(\frac{1}{r(\lambda)}, \frac{1}{p(\lambda)}\right)$, every learner $\mathbf{L}$ using at most $\frac{1}{\epsilon}$ samples (and in particular time) the probability over the choice of k and the internal randomness of $\mathbf{L}$ that it returns a classifier of error at most ϵ is smaller than $\frac{1}{10}$.*

We start with a proof sketch before moving to the formal proof.

Proof Sketch. We base the learning task on Definition 13. Let $w \in \mathcal{X}$. We define the distribution as an equal mixture of two parts $D = \frac{1}{2}D_{\textsc{Clear}} + \frac{1}{2}D_{\textsc{Enc}}$. The first part, i.e. $D_{\textsc{Clear}}$, is equal to $x \sim U(\mathcal{X})$ with label $y = h_w(x)$. The second part, i.e. $D_{\textsc{Enc}}$, is equal to $x' \sim U(\mathcal{X}), y' = h_w(x'), (x, y) = (\text{FHE.}\textsc{Enc}(x), \text{FHE.}\textsc{Enc}(y))$, which can be thought of as $D_{\textsc{Clear}}$ under an FHE encryption.

Note that ignoring samples from $D_{\textsc{Enc}}$, $\mathbf{V}$ can learn, using $O(\frac{1}{\epsilon})$ samples from $D_{\textsc{Clear}}$, a classifier $f_{w'}$ of error ϵ for $D_{\textsc{Clear}}$. Moreover, having access to the public key of the FHE, $\mathbf{V}$ can also evaluate $f_{w'}$ homomorphically on $\text{FHE.}\textsc{Enc}(x)$ to obtain $\text{FHE.}\textsc{Enc}(y)$ of error ϵ on $D_{\textsc{Enc}}$ also. This means that $\mathbf{V}$ is able to learn a low-error classifier on D and thus point 2 of the lemma is satisfied.

To compute $\mathbf{x}$, $\mathbf{V}$ works as follows.[10] It samples a uniformly random point $x_{\textsc{Bnd}}$ from $B_{w'}(O(\epsilon))$ (see Definition 13). Next, $\mathbf{V}$ encrypts it, i.e. $x' := \text{FHE.}\textsc{Enc}(x_{\textsc{Bnd}})$. Next, it flips a uniformly random bit $b \in \{0, 1\}$. If $b = 0$ it sends to $\mathbf{P}$ the encrypted x', and if $b = 1$ it sends to $\mathbf{P}$ a uniformly random $x' \sim U(\mathcal{X})$.

We claim that x' satisfies two crucial properties:

[9] This is only a formal requirement so that the interval $(1/r(\lambda), 1/p(\lambda))$ is non-empty.

[10] In this proof sketch we will assume that $q = 1$.

1. **$\mathbf{P}$ is likely to misclassify x'.** Observe that by the fact that $\mathbf{V}$ learns a good $f_{w'}$ with high probability, x_{BND} is a uniformly random point from an arc that contains the boundary of h_w. $\mathbf{P}$'s running time is upper-bounded by $\frac{1}{\epsilon^2}$, which means that it can only learn (see Lemma 3 for intuition) a classifier of error, say, $\geq 10\epsilon^2$. Taking these two facts together we expect $\mathbf{P}$ to misclassify x' with probability

$$\approx \frac{1}{2} \cdot 10\epsilon^2/\epsilon = 5\epsilon > 2\epsilon,$$

   where the factor $\frac{1}{2}$ takes into account that we send an encrypted x' only half of the time. If this property held it would imply the *transferability* of the scheme.

2. **$\mathbf{P}$ cannot distinguish x' from a sample from the original distribution.** First note that x' is encrypted with the same probability as in the original distribution because we send an encryption of x_{BND} with probability $\frac{1}{2}$ and a uniformly random $x' \sim U(\mathcal{X})$ also with probability $\frac{1}{2}$.

   Crucially, if we send $\text{FHE.ENC}(x_{\text{BND}})$ then it is indistinguishable for $\mathbf{P}$ from $\text{FHE.ENC}(x)$ for any other $x \in \mathcal{X}$. This follows, up to technical details, essentially automatically from the security of the FHE scheme. This implies *undetectability*.

Finally, essentially automatically, Lemma 3 implies point 3 of the lemma.

It is important to note that we need to augment the distribution with the public key of FHE so that $\mathbf{V}$ can execute FHE.ENC and FHE.EVAL.

Note 1. *We want to emphasize that it is crucial (for our construction) that the distribution has both an encrypted and an unencrypted part.*

*If there was no D_{CLEAR} then $\mathbf{V}$ would see only samples of the form $(\text{FHE.ENC}(x), \text{FHE.ENC}(y))$ and would not know which of them lie close to the boundary of h_w, and so it would not be able to choose tricky samples. $\mathbf{V}$ would be able to learn a low-error classifier, but **only** under the encryption. More concretely, $\mathbf{V}$ would be able to evaluate homomorphically a circuit that given a training set and a test point learns a good classifier and classifies the test point with it. However, it would **not** be able to, with high probability, generate $\text{FHE.ENC}(x)$, for x close to the boundary as it would not know (in the clear) where the decision boundary is.*

If there was no D_{ENC} then everything would happen in the clear and so $\mathbf{P}$ would be able to distinguish x's that appear too close to the boundary.

□

Next, we give a formal proof.

Proof. The learning task is based on the task from Definition 13.

Setting of parameters for FHE. Let FHE be a fully homomorphic encryption scheme from Theorem 6. We will use the scheme for constant leveled circuits $d = O(1)$. Let $s(n, \lambda)$ be the polynomial bounding the size of the encryption of inputs of length n with λ security as well as bounding size of the circuit for holomorphic evaluation, which is guaranteed to exist by Theorem 6. Let $\beta \in (0, 1)$ and p be a polynomial such that

$$s(n^\beta, \lambda, d) \leq (n \cdot p(\lambda))^{0.1}, \tag{8}$$

which exist because s is a polynomial. Let $\lambda \in \mathbb{N}$ and $n := p^{1/\beta}(\lambda)$[11] for the length of inputs in the FHE scheme. Observe

$$\begin{aligned} s(n, \lambda, d)) &\leq (p(\lambda) \cdot p(\lambda))^{0.1} && \text{By (8)} \\ &\leq \frac{1}{\epsilon^{0.2}} && \text{By } \epsilon \in \left(\frac{1}{r(\lambda)}, \frac{1}{p(\lambda)}\right). \end{aligned} \tag{9}$$

[11]Note that this setting allows to represent points on $\mathcal{X}$ up to $2^{-p^{1/\beta}(\lambda)}$ precision and this precision is better than $\frac{1}{r(\lambda)}$ for every polynomial r for sufficiently large λ. This implies that this precision is enough to allow for learning up to error ϵ, because of the setting $\epsilon \geq \frac{1}{q(\lambda)}$.

Definition of the learning task. We will omit λ from indexes of D and h for simplicity of notation. Let $\mathcal{D} = \{D^{(\text{pk,sk})}\}_{(\text{pk,sk})}, \mathcal{H} = \{h^{(\text{pk,sk,w})}\}_{(\text{pk,sk,w})}$ indexed by valid public/secret key pairs of FHE and $w \in \mathcal{X}$, with $\mathcal{X}$ as in Definition 13. Let $D_{\mathcal{L}}$ over $(\text{pk,sk}, w)$ be equal to $\text{FHE.KeyGen}(1^\lambda) \times U(\mathcal{X})$.

For a valid (pk,sk) pair we define $D^{(\text{pk,sk})}$ as the result of the following process: $x \sim D = U(\mathcal{X})$, with probability $\frac{1}{2}$ return $(0, x, \text{pk})$ and with probability $\frac{1}{2}$ return $(1, \text{FHE.Enc}(x), \text{pk})$, where the first element of the triple describes if the x is encrypted or not. x is represented as a number $\in (0,1)$ using n bits.[12]

For a valid (pk,sk) pair and $w \in \mathcal{X}$ we define $h^{(\text{pk,sk,w})}((b, x, \text{pk}), y)$ as a result of the following process: if $b = 0$ return $\mathbb{1}_{h_w(x)=y}$, otherwise let $x_{\text{Dec}} \leftarrow \text{FHE.Dec}(\text{sk}, x), y_{\text{Dec}} \leftarrow \text{FHE.Dec}(\text{sk}, y)$ and if $x_{\text{Dec}}, y_{\text{Dec}} \neq \perp$ (decryption is succesful) return $\mathbb{1}_{h_w(x_{\text{Dec}})=y_{\text{Dec}}}$ and return 1 otherwise.

Note 2 ($\Omega(\frac{1}{\epsilon})$-sample learning lower bound.). *Note, that by construction any learner using K samples for learning task $\{D^{(pk,sk)}\}_{(pk,sk)}, \{h^{(pk,sk,w)}\}_{(pk,sk,w)}$ can be transformed (potentially computationally inefficiently) into a learner using K samples for the task from Defnition 13 that returns a classifier of at most the same error. This together with a lower bound for learning from Lemma 3 proves point 3 of the lemma.*

Definition of $\mathbf{V}$. $\mathbf{V}$ draws N samples $Q = \{((b_i, x_i, \text{pk}), y_i)\}_{i \in [N]}$ for $N := \frac{900}{\epsilon}$.

Next, $\mathbf{V}$ chooses a subset $Q_{\text{CLEAR}} \subseteq Q$ of samples for which $b_i = 0$. It trains a classifier $f_{w'}(\cdot) := \text{sgn}(\langle w', \cdot \rangle)$ on Q_{CLEAR} by returning any $f_{w'}$ consistent with Q_{CLEAR}. This can be done in time

$$N \cdot n \leq \frac{900}{\epsilon} \cdot p^{1/\beta}(\lambda) \leq \frac{900}{\epsilon^{1.1}} \tag{10}$$

by keeping track of the smallest interval containing all samples in Q_{CLEAR} labeled with $+1$ and then returning any $f_{w'}$ consistent with this interval.

Note 3 ($O(\frac{1}{\epsilon^{1.3}})$-time learning upper bound.). *First note that $\mathbf{V}$ learns well, i.e. with probability at least $1 - 2\left(1 - \frac{\epsilon}{100}\right)^{\frac{900}{\epsilon}} \geq 1 - \frac{1}{1000}$ we have that*

$$|\measuredangle(w, w')| \leq \frac{2\pi\epsilon}{100} \tag{11}$$

Moreover, $f_{w'}(x)$ can be implemented by a circuit $C_{f_{w'}}$ that compares x with the endpoints of the interval. This can be done by a constant leveled circuit. Moreover $C_{f_{w'}}$ can be evaluated with FHE.Eval *in time*

$$\textit{size}(C_{f_{w'}})s(n, \lambda, d) \leq 10n \cdot s(n, \lambda, d) \leq 10p^{1/\beta}(\lambda)s(n, \lambda, d) \leq \frac{10}{\epsilon^{0.3}},$$

where the last inequality follows from (9). This implies that $\mathbf{V}$ can, in time T, return a classifier of error $\leq \epsilon$ for $(D^{(pk,sk)}, h^{(pk,sk,w)})$. This proves point 2. of the lemma.

Next, $\mathbf{V}$ prepares $\mathbf{x}$ as follows. It samples $q = \frac{16}{\epsilon}$ points $\{x'_i\}_{i \in [q]}$ from $\mathcal{X}$ uniformly at random. It chooses a uniformly random subset $S \subseteq [q]$. Next, $\mathbf{V}$ generates $q - |S|$ inputs using the following process: $x_{\text{BND}} \sim U(B_{w'}(2\pi(\epsilon + \frac{\epsilon}{100})))$ (x_{BND} is close to the decision boundary of $f_{w'}$), return $\text{FHE.Enc}(\text{pk}, x_{\text{BND}})$. Call the set of $q - |S|$ points E_{BND}. $\mathbf{V}$ defines:

$$\mathbf{x} := \{(0, x'_i, \text{pk}) \mid i \in [q] \setminus S\} \cup \{(1, x', \text{pk}) \mid x' \in E_{\text{BND}}\}.$$

The running time of this phase is dominated by evaluations of FHE.Eval, which takes

$$q \cdot s(n, \lambda, d) \leq \frac{16}{\epsilon} \cdot \frac{1}{\epsilon^{0.2}} \leq \frac{16}{\epsilon^{1.2}}, \tag{12}$$

where the first inequality follows from (9). Taking the sum of (10) and (12) we get that the running time of $\mathbf{V}$ is smaller than the required $T = \frac{10^3}{\epsilon^{1.3}}$.

[12] Note that the space over which $D^{(\text{pk,sk})}$ is defined on is *not* $\mathcal{X}$.

$\mathbf{V}$ is a transferable attack. Now, consider $\mathbf{P}$ that runs in time $t = \frac{1}{\epsilon^2}$. By the assumption $t \leq r(\lambda)$, which implies that the security guarantees of FHE hold for $\mathbf{P}$.

We first claim that $\mathbf{x}$ is indistinguishable from $D^{(\text{pk,sk})}$ for $\mathbf{P}$. Observe that by construction the distribution of ratio of encrypted and not encrypted x's in $\mathbf{x}$ is identical to that of $D^{(\text{pk,sk})}$. Moreover, the distribution of unencrypted x's is identical to that of $D^{(\text{pk,sk})}$ by construction. Finally, by the IND-CPA security of FHE and the fact that the running time of $\mathbf{P}$ is bounded by $q(\lambda)$ for some polynomial q we have that $\text{FHE.}\textsc{Enc}(\text{pk}, x_{\text{BND}})$ is distinguishable from $x \sim \mathcal{X}, \text{FHE.}\textsc{Enc}(\text{pk}, x)$ with advantage at most $\text{negl}(\lambda)$. Thus undetectability holds with near perfect soundness $s = \frac{1}{2} + \text{negl}(\lambda)$.

Next, we claim that $\mathbf{P}$ can't return low-error answers on $\mathbf{x}$.

Assume towards contradiction that with probability $\frac{5}{100}$

$$\mathbb{P}_{w\sim U(\mathcal{X}),x\sim U(B_w(2\pi\epsilon))}[f(x) \neq h_w(x)] \leq 10\epsilon. \tag{13}$$

We can apply Lemma 4 to get that there exists a learner using $t + \frac{9}{\epsilon}$ samples that with probability $\frac{4}{100}$ returns f' such that

$$\mathbb{P}_{w\sim U(\mathcal{X}),x\sim U(\mathcal{X})}[f'(x) \neq h_w(x)] \leq 40\epsilon^2. \tag{14}$$

Applying Lemma 3 to (14) we know that

$$40\epsilon^2 \geq \frac{1}{2(t + \frac{9}{\epsilon})},$$

which implies

$$t \geq \frac{10}{\epsilon^2},$$

which is a contradiction with the assumed running time of $\mathbf{P}$. Thus (13) doesn't hold and in consequence using (11) we have that with probability $1 - \frac{6}{100}$

$$\mathbb{P}_{w\sim U(\mathcal{X}),x\sim U(B_{w'}(2\pi(\epsilon+\frac{\epsilon}{10})))}[f(x) \neq h_w(x)] \geq \frac{10}{14} \cdot 10\epsilon \geq 7\epsilon, \tag{15}$$

where crucially x is sampled from $U(B_{w'})$ and not $U(B_w)$. By Fact 1 we know that $|S| \geq \frac{q}{3}$ with probability at least

$$1 - 2e^{-\frac{q}{72}} = 1 - 2e^{-\frac{1}{8\epsilon}} \geq 1 - \frac{1}{1000}.$$

Another application of the Chernoff bound and the union bound we get from (15) that with probability at least $1 - \frac{1}{10}$ we have that $\text{err}(\mathbf{x}, \mathbf{y})$ is larger than 2ϵ by the setting of $q = \frac{16}{\epsilon}$.

□

Fact 1 (Chernoff-Hoeffding)**.** *Let $X_1, \ldots, X_k$ be independent Bernoulli variables with parameter p. Then for every $0 < \epsilon < 1$*

$$\mathbb{P}\left[\left|\frac{1}{k}\sum_{i=1}^{k} X_i - p\right| > \epsilon\right] \leq 2e^{-\frac{\epsilon^2 k}{2}}$$

and

$$\mathbb{P}\left[\frac{1}{k}\sum_{i=1}^{k} X_i \leq (1-\epsilon)p\right] \leq e^{-\frac{\epsilon^2 kp}{2}}.$$