Berlin Mathematics Research Center MATH+

The Good, the Bad and the Ugly: Watermarks, Transferable Attacks and Adversarial Defenses

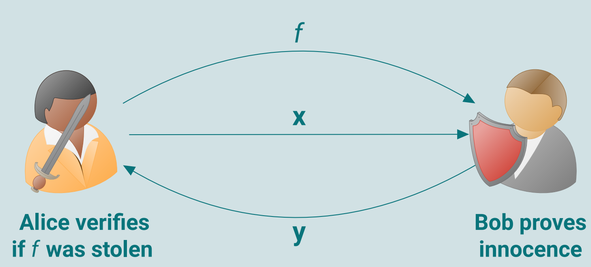
Grzegorz Głuch (EPFL), Berkant Turan (ZIB, TUB), Sai Ganesh Nagarajan (ZIB), Sebastian Pokutta (ZIB, TUB)

Motivation

A nonprofit organization plans to open-source a classifier f but wants to detect its use by embedding a watermark directly into the model. Alice is tasked with creating this watermark. Bob aims to make f adversarially robust, i.e., to ensure that it is hard to find queries that appear unsuspicious but cause f to make mistakes. Both face challenges: Alice struggles to create a watermark that cannot be removed, and Bob's defenses become increasingly complex. They discover their projects are connected. Alice's idea was to plant a backdoor [1, 2] in f, enabling her to craft queries with a hidden trigger that activates the backdoor, causing f to misclassify, thus detecting the usage of f. Bob's approach involved smoothing f to enhance robustness, which inadvertently removes such backdoors [2]. They realized that their challenges are two sides of the same coin: the impossibility of one task might guarantee the success of the other.

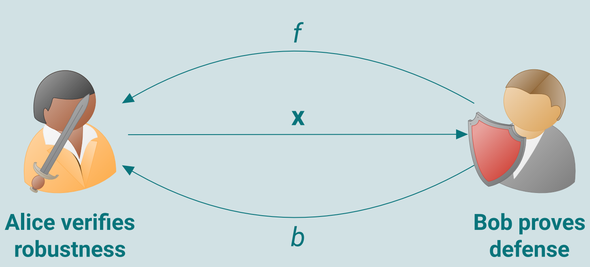
Every learning task has at least one of the three:

Backdoor-Based Watermarks



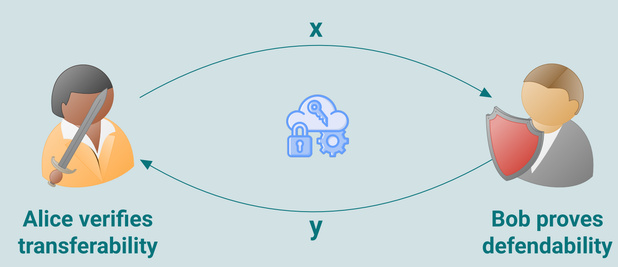
A **Watermark** is an efficient algorithm that computes a low-error classifier f and a set of queries **x** such that (fast) defenders are unable to find low-error answers **y** nor distinguish **x** from the data distribution.

Adversarial Defenses



An **Adversarial Defense** is an efficient algorithm that computes a low-error classifier f and a detection bit b, such that (fast) adversaries are unable to find queries **x**, which look indistinguishable from the data distribution and where f is incorrect.

Transferable Attacks



A **Transferable Attack** is an efficient algorithm that computes queries **x** that look indistinguishable from the data distribution that fool all efficient defenders.

Definition 1 (*Watermark*)

An algorithm $\mathbf{A}_{\text{WATERMARK}}$, running in time $T_{\mathbf{A}}$, implements a *watermarking scheme* for the learning task $\mathcal{L}$, with error parameter $\epsilon > 0$, if an interactive protocol in which $\mathbf{A}_{\text{WATERMARK}}$ computes a classifier $f\colon \mathcal{X} \to \mathcal{Y}$ and a sequence of queries $\mathbf{x} \in \mathcal{X}^q$, and a prover **B** outputs $\mathbf{y} = \mathbf{B}(f, \mathbf{x}) \in \mathcal{Y}^q$ satisfies the following properties:

1. **Correctness:** f has low error, i.e.,
$$\text{err}(f) \leq \epsilon\,.$$
2. **Uniqueness:** There exists a prover **B**, running in time bounded by $T_{\mathbf{A}}$, which provides low-error answers, such that
$$\text{err}(\mathbf{x}, \mathbf{y}) \leq 2\epsilon\,.$$
3. **Unremovability:** For every prover **B** running in time $T_{\mathbf{B}}$, it holds that
$$\text{err}(\mathbf{x}, \mathbf{y}) > 2\epsilon\,.$$
4. **Undetectability:** For every prover **B** running in time $T_{\mathbf{B}}$, the advantage of **B** in distinguishing the queries **x** generated by $\mathbf{A}_{\text{WATERMARK}}$ from random queries sampled from $\mathcal{D}^q$ is small.

Main Result

Theorem 1 (*Main Theorem* [3]) *For every learning task $\mathcal{L}$ and $\epsilon \in \left(0, \frac{1}{2}\right)$, $T \in \mathbb{N}$, where a learner exists that runs in time T and, with high probability, learns f satisfying $\text{err}(f) \leq \epsilon$, at least one of these three exists:*

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

Proof (Sketch).

The proof is based on the *complementary nature* of watermarking and defense definitions. Any attempt to remove a Watermark leads to a potential Defense, and vice versa. We define a *zero-sum game* $\mathcal{G}$ between a watermarking algorithm **A** and a removal algorithm **B**, both succinctly representable to avoid trivial solutions. The strategies and payoffs are determined by whether the error and rejection criteria are met. Nash's theorem guarantees an equilibrium $(\mathbf{A}_{\text{NASH}}, \mathbf{B}_{\text{NASH}})$, where the value of the game implies one of three outcomes: a Watermark, an Adversarial Defense, or a Transferable Attack.

- Adversarial Defense: If the Nash equilibrium payoff exceeds a threshold, there exists a Defense. Here, $\mathbf{B}_{\text{DEFENSE}}$ learns a classifier f, engages with an attacker by simulating $(\mathbf{y}, b) = \mathbf{B}_{\text{NASH}}(f, \mathbf{x})$, and returns $b = 1$ when an attack is being detected.
- Watermark or Transferable Attack: If the payoff is below the threshold, we either have a Watermark or a Transferable Attack, depending on $\mathcal{G}$'s details.

□

Definition 2 (*Adversarial Defense*)

An algorithm $\mathbf{B}_{\text{DEFENSE}}$, running in time $T_{\mathbf{B}}$, implements an *adversarial defense* for the learning task $\mathcal{L}$, with error parameter $\epsilon > 0$, if an interactive protocol in which $\mathbf{B}_{\text{DEFENSE}}$ computes a classifier $f\colon \mathcal{X} \to \mathcal{Y}$, a verifier **A** replies with $\mathbf{x} = \mathbf{A}(f)$, where $\mathbf{x} \in \mathcal{X}^q$, and $\mathbf{B}_{\text{DEFENSE}}$ outputs $b = \mathbf{B}_{\text{DEFENSE}}(f, \mathbf{x}) \in \{0, 1\}$ satisfies the following properties:

1. **Correctness:** f has low error, i.e.,
$$\text{err}(f) \leq \epsilon\,.$$
2. **Completeness:** When $\mathbf{x} \sim \mathcal{D}^q$, then
$$b = 0\,.$$
3. **Soundness:** For every **A** running in time $T_{\mathbf{A}}$, we have
$$\text{err}(\mathbf{x}, f(\mathbf{x})) \leq 7\epsilon \;\text{ or }\; b = 1\,.$$

Tasks with Watermarks and Adversarial Defenses (for bounded VC-Dimension)

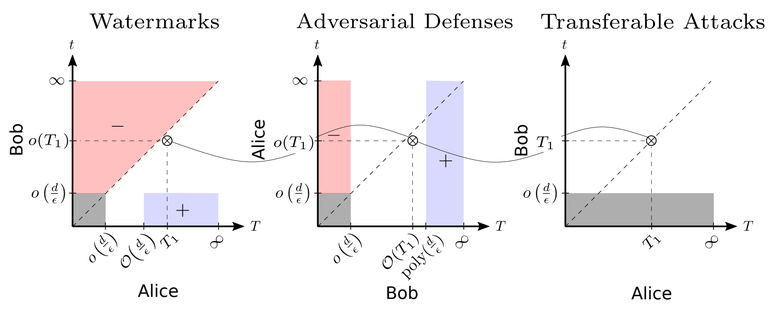


Figure 1. Overview of the taxonomy of learning tasks, illustrating the presence of Watermarks, Adversarial Defenses, and Transferable Attacks for learning tasks of bounded VC dimension. The axes represent the time bound for the parties in the corresponding schemes. The blue regions depict positive results, the red negative, and the gray regimes of parameters which are not of interest. See Lemma 1 and 2 for details about blue regions. The curved line represents a potential application of Theorem 1, which says that at least one of the three points should be blue.

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 **B** such that for every $\epsilon \in \left(0, \frac{1}{8}\right)$, $\mathcal{D}$ over $\mathcal{X}$ and $h \in \mathcal{H}$ we have*

$$\mathbf{B} \in \text{DEFENSE}\left((\mathcal{D}, h), \epsilon, T_{\mathbf{A}} = \infty, T_{\mathbf{B}} = \texttt{poly}\left(\frac{d}{\epsilon}\right)\right).$$

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

$$\mathbf{A} \in \text{WATERMARK}\left((\mathcal{D}, h), \epsilon, T_{\mathbf{A}} = O\left(\frac{d}{\epsilon}\right), T_{\mathbf{B}} = \frac{d}{100}\right).$$

Definition 3 (*Transferable Attack*)

An algorithm $\mathbf{A}_{\text{TRANSFATTACK}}$, running in time $T_{\mathbf{A}}$, implements a *transferable attack* for the learning task $\mathcal{L}$, with error parameter $\epsilon > 0$, if an interactive protocol in which $\mathbf{A}_{\text{TRANSFATTACK}}$ computes $\mathbf{x} \in \mathcal{X}^q$, and **B** outputs $\mathbf{y} = \mathbf{B}(\mathbf{x}) \in \mathcal{Y}^q$ satisfies the following properties:

1. **Transferability:** For every prover **B** running in time $T_{\mathbf{A}}$, we have
$$\text{err}(\mathbf{x}, \mathbf{y}) > 2\epsilon\,.$$
2. **Undetectability:** For every prover **B** running in time $T_{\mathbf{B}}$, the advantage of **B** in distinguishing the queries **x** generated by $\mathbf{A}_{\text{TRANSFATTACK}}$ from random queries sampled from $\mathcal{D}^q$ is small.

Transferable Attack for Cryptography based Learning Task

Theorem 2 (*Transferable Attack for Cryptography based Learning Task*) *There exists a distribution $\mathcal{D}$ and a hypothesis class $\mathcal{H}$ for which there is a Transferable Attack $\mathbf{A}_{TA}$ such that if h is sampled uniformly from $\mathcal{H}$, then*

$$\mathbf{A}_{TA} \in \textit{TRANSFATTACK}\left((\mathcal{D}, h), \epsilon, T = O\left(1/\epsilon\right), t = 1/\epsilon^2\right).$$

Moreover, for every ϵ, $O\left(1/\epsilon\right)$ time and $O\left(1/\epsilon\right)$ samples are sufficient, while $\Omega\left(1/\epsilon\right)$ samples (and time) are necessary to, on average, learn w.h.p. a classifier of error ϵ.

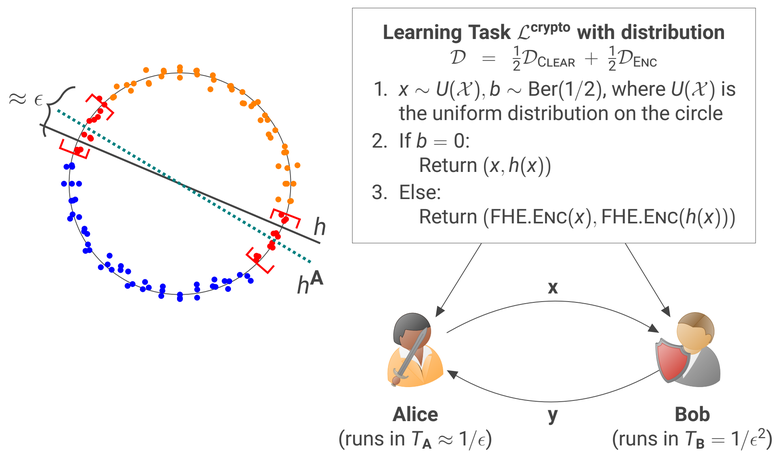


Figure 2. The left part of the figure represents a *Lines on Circle Learning Task* $\mathcal{L}^{\circ}$ with a ground truth function denoted by h. On the right, we define a *cryptography-augmented* learning task derived from $\mathcal{L}^{\circ}$. In its distribution, a "clear" or an "encrypted" sample is observed with equal probability. Given their respective times, both **A** and **B** are able to learn a low-error classifier $h^{\mathbf{A}}$, $h^{\mathbf{B}}$ respectively, by learning only on the *clear samples*. **A** is able to compute a Transferable Attack by computing an encryption of a point close to the decision boundary of her classifier $h^{\mathbf{A}}$.

References

[1] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. 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.

[2] S. Goldwasser, M. P. Kim, V. Vaikuntanathan, and O. Zamir.
Planting Undetectable Backdoors in Machine Learning Models.
ArXiv, abs/2204.06974, 2022.
URL: `https://api.semanticscholar.org/CorpusID:248177888`.

[3] G. Głuch, B. Turan, S. G. Nagarajan, and S. Pokutta.
The Good, the Bad and the Ugly: Watermarks, Transferable Attacks and Adversarial Defenses.
ArXiv, abs/2410.08864, 2024.
URL: `https://arxiv.org/pdf/2410.08864`.

ZIB ZUSE INSTITUTE BERLIN · IOI · TU berlin