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

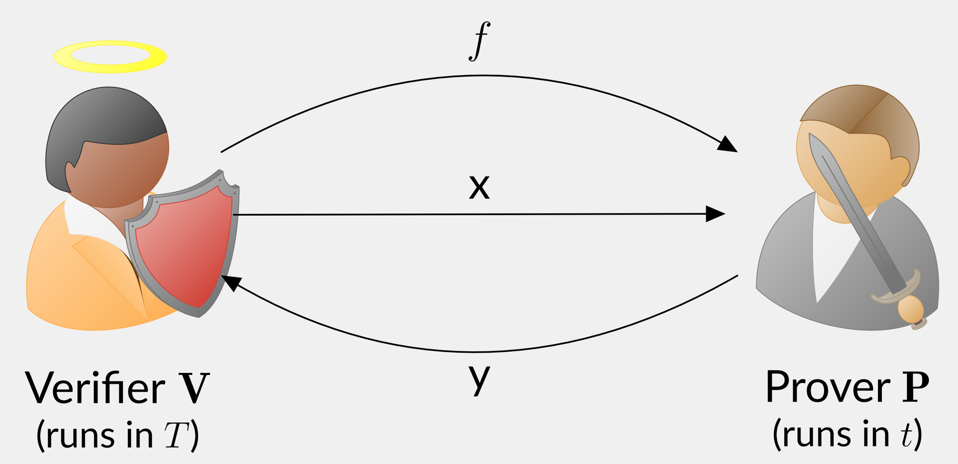
Grzegorz Głuch, Sai Ganesh Nagarajan*, and Berkant Turan*
EPFL, TU Berlin & Zuse Institute Berlin
* Equal contribution.

Learning Task

- *Learning Task:* A classification based learning task $\mathcal{L}$ is a pair $(\mathcal{D}, h)$ of a distribution $\mathcal{D}$, supp$(\mathcal{D}) \subseteq \mathcal{X}$, and a ground truth map to a set of labels $h : \mathcal{X} \to \mathcal{Y} \cup \{\perp\}$.
- *Risk Measure:* To every $f : \mathcal{X} \to \mathcal{Y}$, we associate $\text{err}(f) := \mathbb{E}_{x \sim \mathcal{D}}[f(x) \neq h(x)]$.
- *Information Access:* We assume all parties have access to i.i.d. samples $(x, h(x))$, where $x \sim \mathcal{D}$, although $\mathcal{D}$ and h are unknown to the parties.

Every learning task has at least one of the three:

Watermark



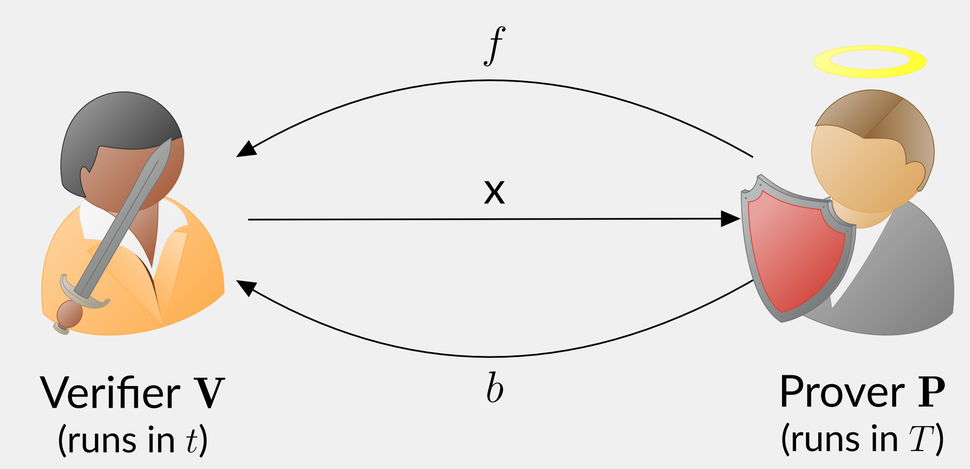
*Watermark is "an undetectable trigger set that protects your model from **all** thieves."*

Properties of Watermark$(\mathcal{L}, \epsilon, T, t)$

- **Correctness (f has low error):** W.h.p., $\text{err}(f) \leq \epsilon$.
- **Uniqueness (training from scratch):** There exists succinctly representable **P** running in time T such that w.h.p., $\text{err}(\mathbf{x}, \mathbf{y}) \leq 2\epsilon$.
- **Unremovability (fast P give high-error):** For every succinctly representable **P** running in time t, w.h.p., $\text{err}(\mathbf{x}, \mathbf{y}) > 2\epsilon$.
- **Undetectability (fast P accept tests):** For every succinctly representable **P** running in time t, the advantage in distinguishing $\mathbf{x} \sim \mathcal{D}^q$ from $\mathbf{x} := \mathbf{V}$ is small.

Note that in the case of Uniqueness, **P** runs in time T.

Adversarial Defense



*Adversarial Defense is "a way to detect **all** adversarial examples."*

Properties of Defense$(\mathcal{L}, \epsilon, t, T)$

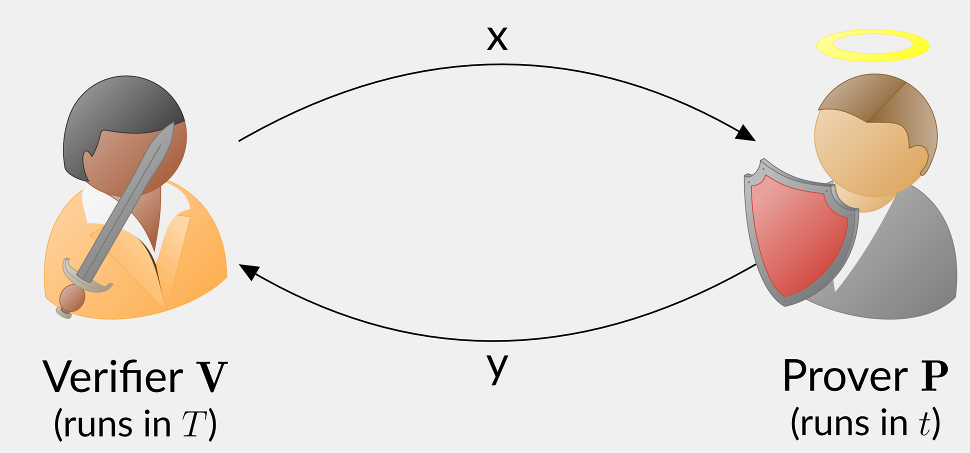
- **Correctness (f has low error):** W.h.p., $\text{err}(f) \leq \epsilon$.
- **Completeness (if x is from correct distribution, P does accept the test):** When $\mathbf{x} \sim \mathcal{D}^q$, then w.h.p.

$$b = 0.$$

- **Soundness (fast attacks creating x on which f makes mistakes are detected):** For every succinctly representable **V** running in time t, we have that w.h.p.,

$$\text{err}(\mathbf{x}, f(\mathbf{x})) \leq 7\epsilon \ \text{ or } \ b = 1.$$

Transferable Attack



*Transferable Attack is "a way to undetectably attack **all** models."*

Properties of TransfAttack$(\mathcal{L}, \epsilon, T, t)$

- **Transferability (fast P give high-error answers):** For every succinctly representable **P** running in time t, w.h.p.,

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

- **Undetectability (fast P accept tests):** For every succinctly representable **P** running in time t, the advantage in distinguishing $\mathbf{x} \sim \mathcal{D}^q$ from $\mathbf{x} := \mathbf{V}$ is small.

Theorem 1 (Unified Taxonomy)

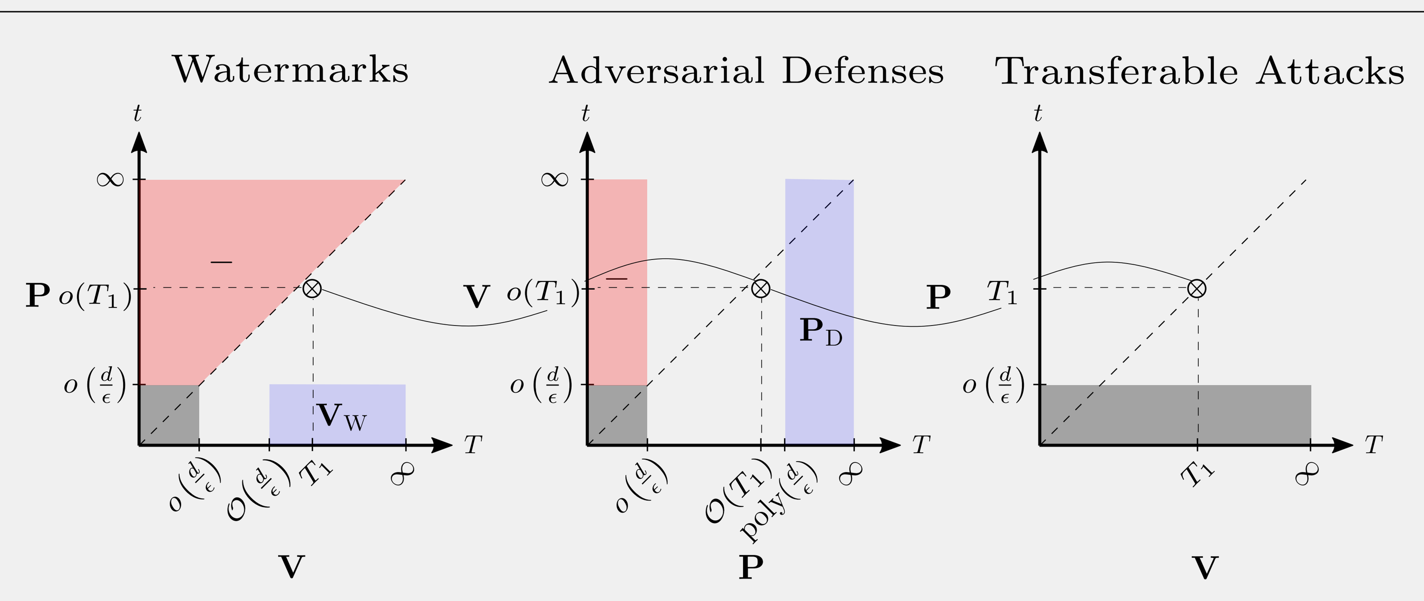
*For **every learning task** $\mathcal{L}$ and $\epsilon \in \left(0, \frac{1}{2}\right)$, $T \in \mathbb{N}$, such that there exists a learner running in time T that, w.h.p., learns f such that $\text{err}(f) \leq \epsilon$, at least one of*

$$\text{Watermark}\left(\mathcal{L}, \epsilon, T, T^{1/\sqrt{\log(T)}}\right),$$
$$\text{Defense}\left(\mathcal{L}, \epsilon, T^{1/\sqrt{\log(T)}}, O(T)\right),$$
$$\text{TransfAttack}\left(\mathcal{L}, \epsilon, T, T\right)$$

exists.

Notably, when a **Defense does not exist**, there **must be a Watermark or a Transferable Attack**, which goes beyond the prior understanding of the existence of adversarial attacks.

Examples (Bounded VC-Dimension)



Overview of learning tasks with *Watermarks*, *Adversarial Defenses*, and *Transferable Attacks* for **bounded VC dimension**.

Example 1 (Adversarial Defense for bounded VC-dimension). *There exists an algorithm $\mathbf{P}_\text{D}$ that is an Adversarial Defense for every hypothesis class $\mathcal{H}$ of VC-dimension d, i.e. for every $h \in \mathcal{H}$ and a distribution $\mathcal{D}$*

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

$\mathbf{P}_\text{D}$ is an adaptation of the defense from [Goldwasser et al. 2020].

Example 2 (Watermark for bounded VC-dimension against fast adversaries). *For every $d \in \mathbb{N}$ there exists a learning task $\mathcal{L}$ with a hypothesis class of VC-dimension d for which there is a Watermark $\mathbf{V}_\text{W}$, i.e.*

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

Open Questions

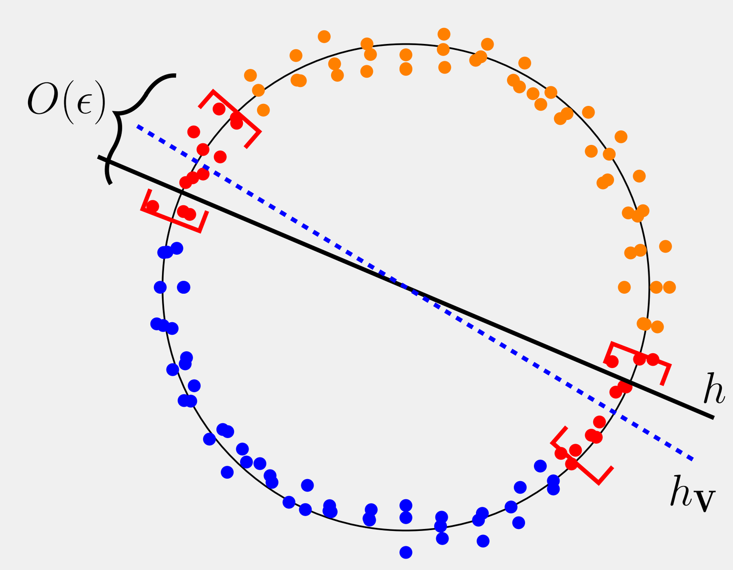
Is it possible to generalize the definitions and obtain a similar taxonomy for generative learning tasks?
Key challenges: verification vs. generation, quality oracles [Zhang et al., 2023], self-evaluation.

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 $\mathbf{V}_\text{TA}$ such that if h is sampled uniformly from $\mathcal{H}$, then

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

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.

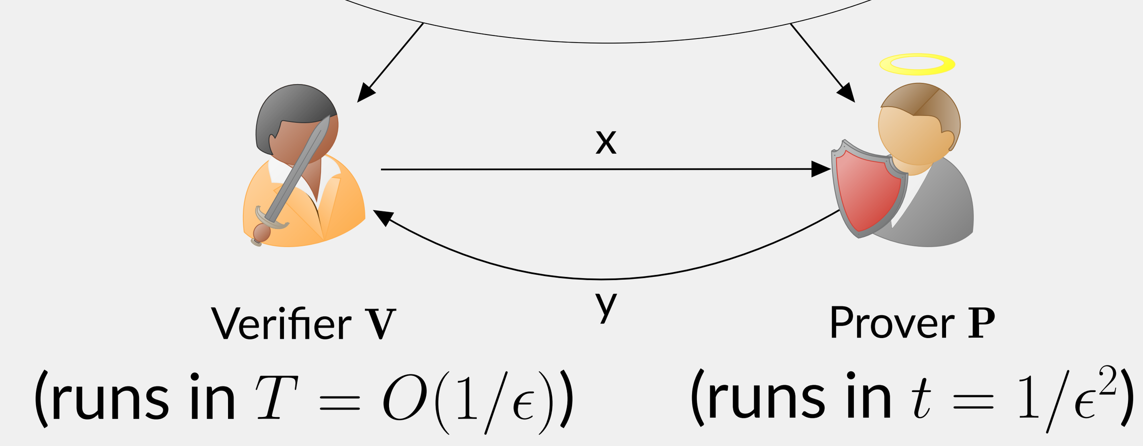


Fully Homomorphic Encryption (FHE)
Cryptographic primitive allowing computation on encrypted data without decrypting it.

- pk,sk = KeyGen(1^n): Samples public and secret key.
- ψ = Enc(pk, x): Encrypts x with public key pk.
- ψ_C = Eval(pk, C, ψ): Given public key pk, encrypted input ψ, and circuit C, it returns an encryption of an evaluation of C on the input encrypted to ψ.
- y = Dec(sk, ψ_C): Given secret key sk and an encrypted evaluation of C, it returns the result in the clear.

Data Generation Process $\mathcal{D}$

1. $x \sim U_\text{Circle}$, $b \sim \text{Ber}(1/2)$
2. If $b = 0$: return $(x, h(x))$
3. Else: return $(\text{Enc}(\text{pk}, x), \text{Enc}(\text{pk}, h(x)))$



Algorithm V

1. $S \sim \mathcal{D}^{O(1/\epsilon)}$
2. $h_\mathbf{V}$: Learns a line consistent with all unencrypted samples from S
3. $\mathbf{x}_B \sim$ uniformly random point from points $O(\epsilon)$-close to the decision boundary of $h_\mathbf{V}$
4. $b \sim \text{Ber}(1/2)$
5. If: $b = 0$ return Enc$(\mathbf{x}_B)$
6. Else: return $\mathbf{x} \sim U_\text{circle}$

Why V is a Tranferable Attack?

- **Transferability: P** runs in time $\frac{1}{\epsilon^2}$, so its error on $\mathcal{D}$ is $\Omega(\epsilon^2)$, which implies that the error on $\mathbf{x}_B$ is at least $\Omega(\epsilon)$
- **Undetectability:** By the properties of FHE, Enc$(\mathbf{x}_B)$ is indistinguishable from Enc$(\mathbf{x})$, where $\mathbf{x} \sim U_\text{circle}$