Ho Far Can Robust Learning Go?

Mohammad Mahmoody

based on joint works from NeurIPS-18, AAAI-19, ALT-19 *with*

Dimitrios Diochnos **Saeed Mahloujifar**

What is Machine Learning?

- Learning from historical data to make decisions about unseen data.

- Traditional Programming

  Data $\longrightarrow$ Computer $\longrightarrow$ Output

  Program $\longrightarrow$

- Machine Learning

  Data $\longrightarrow$ Computer $\longrightarrow$ Program

  Output $\longrightarrow$

Success of Machine Learning

- Machine learning (ML) has changed our lives
  - Health
  - Language processing
  - Finance/Economy
  - Vision and image classification
  - Computer Security
  - Etc. etc.,..

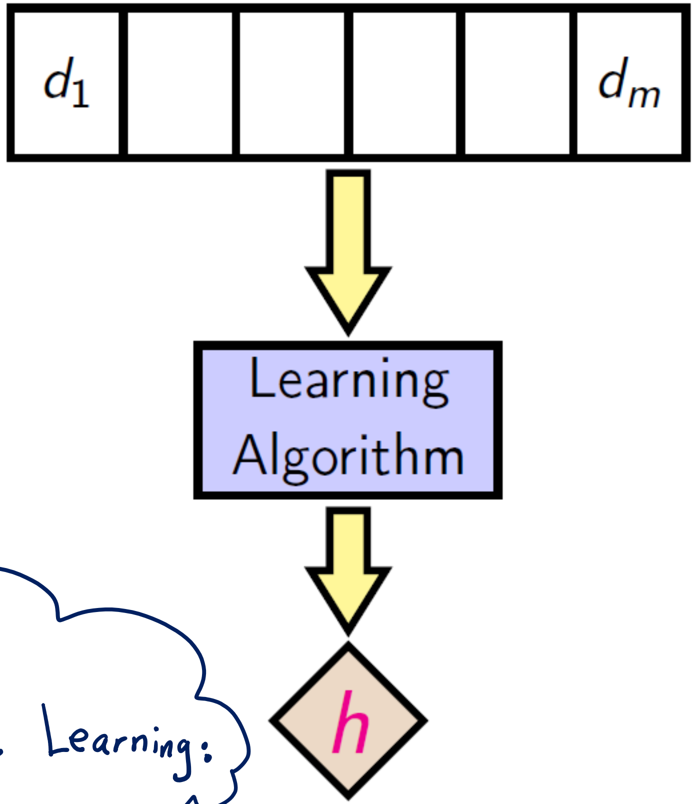
Not primarily designed for **adversarial** contexts!

Classification

Training

$x_i \leftarrow D$

$d_i = (x_i, c(x_i))$



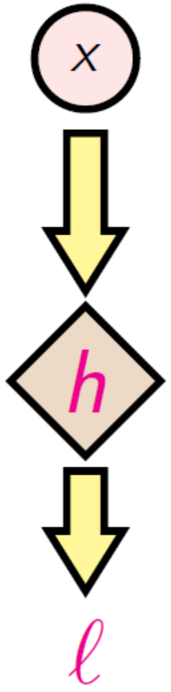
(ε, δ)-PAC Learning:

$1-\delta \leqslant Conf(L) = \mathbf{Pr}\left(\mathbf{Risk}_D(h, c) < \varepsilon\right)$

Testing

$x \leftarrow D$

$d = (x, c(x))$

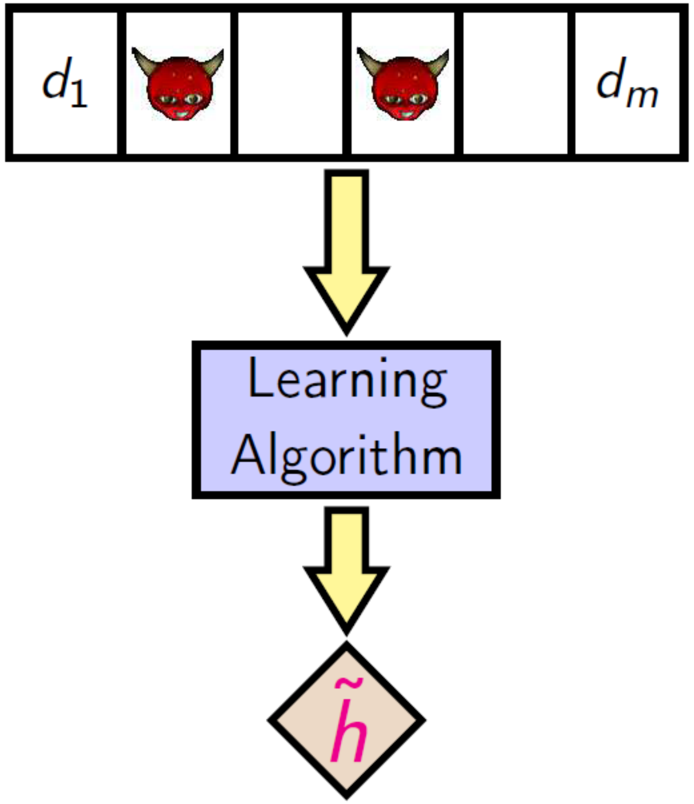


$\mathbf{Risk}_D(h, c) = \mathbf{Pr}_D(\ell \neq c(x))$

Classification under Attack

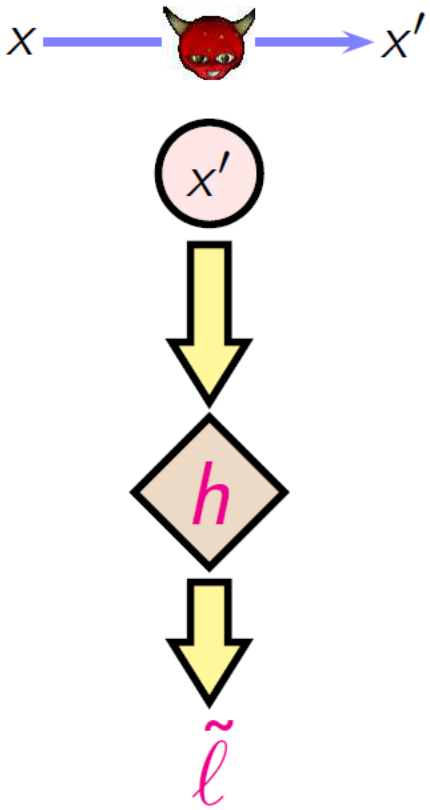
Poisoning Attack

$$x_i \leftarrow D$$
$$d_i = (x_i, c(x_i))$$



Evasion Attack

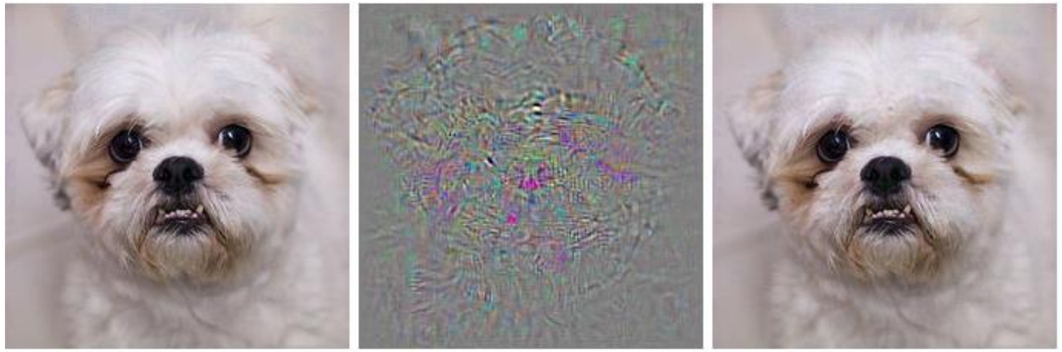
$$x \leftarrow D$$
$$d = (x, c(x))$$



Secure (Adversarially Robust) Machine Learning

- Is achieving low risk still possible in presence of malicious adversaries?
  - Subverting spam filter by poisoning training data [Nelson et. al. 2008]
  - Evading PDF malware detectors [Xu et. al. 2016]
  - Making image classifiers misclassify by adding small perturbations [Szegedy et. al. 2014]

Dog



Camel !

Arms Race of Attacks vs. Defenses

- A repeated cycle of new attacks followed by new defenses:

Nelson et. al. 2008,
Rubinstein et. al. 2009
Kloft et. al. 2010
Biggio et. al. 2012
Xiao et. al. 2012
Kloft et. al. 2012
Biggio et. al. 2014
Newell et. al .2014
Xiao et. al. 2015
Mei et. al. 2015
Burkard et. al. 2017
Koh et. al. 2017
Laishram et. al. 2018
Munoz-Gonz et. al. 2018
…. …. …

Wittel et al. 2004, Dalvi et al. 2004
Lowd et al. 2005, Globerson et al. 2006
Globerson et al. 2008, Dekel et al. 2010
Biggio et al. 2013, Szegedy et al. 2013
Srndic et al. 2014, Goodfellow et al. 2014
Kurakin et al. 2016, Sharma et al. 2017
Kurakin et al. 2016, Carlini et al. 2017
Papernot et al. 2017, Carlini et al. 2017
Tramer et al. 2018, Madry et al. 2018
Raghunathan et al. 2018, Sinha et al. 2018
Na et al. 2018, Gou et al. 2018
Dhillon et al. 2018, Xie et al. 2018
Song et al. 2018,Madry et al. 2018
Samangouei et al. 2018, Athalye et al. 2018
…. …. …

Important Questions in Adversarial Machine Learning

- Formalizing (complexity-theoretic) notions of security.
- What are the inherent powers and limitations of adversaries against ML systems?
- Barriers for provable robustness of ML systems against adversarial attacks, whether poisoning or evasion.
  - information-theoretic, with all-knowing adversaries
  - computationally bounded adversaries
- Can ML systems achieve Probably Approximately Correct (PAC) generalization bounds under adversarial attacks?

Are there inherent reasons enabling adversarial examples and poisoning attacks?

Candidate reason: Concentration of Measure!

Are there inherent reasons enabling
***Polynomial-time** attacks?*

Candidate reason: **Computational** Concentration of Measure!

Related to certain polynomial-time attacks **on coin-tossing** protocols.

Talk Outline

1a. Defining evasion attacks formally

1b. Evasion attacks from measure concentration of instances

2a. Defining poisoning attacks formally

2b. Poisoning attacks from measure concentration of products

3a. Poly-time attacks from computational concentration of products
3b. Connections to attacks on coin-tossing protocols

Talk Outline

1a. Defining evasion attacks formally

1b. Evasion attacks from measure concentration of instances

2a. Defining poisoning attacks formally

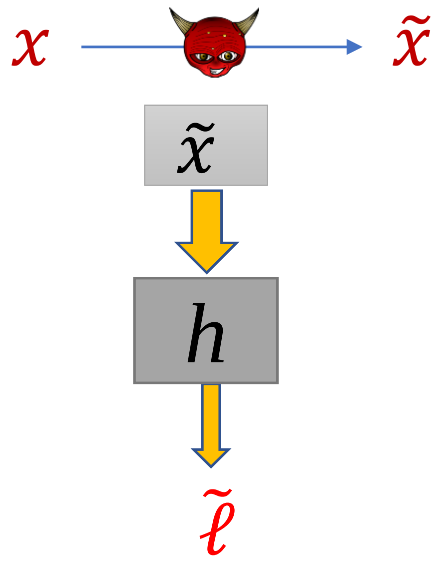
2b. Poisoning attacks from measure concentration of products

3a. Poly-time attacks from computational concentration of products

3b. Connections to attacks on coin-tossing protocols

Evasion Attacks Finding Adversarial Examples

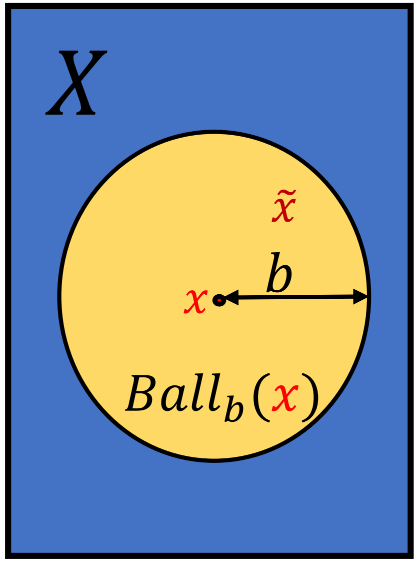
$$x \leftarrow D$$
$$d = (x, c(x))$$



$$Risk(h) = \Pr_{x \leftarrow D}[\tilde{\ell} \neq c(\tilde{x})]$$

- Metric M
  - $\tilde{x}$ close to x w.r.t. M
  - *i.e.* $\tilde{x} \in Ball_b(x)$ for small b

- **Error-region** Adversarial Risk:



$$AdvRisk_b(h) = \Pr_{x \leftarrow D}[\exists \tilde{x} \in Ball_b(x); h(\tilde{x}) \neq c(\tilde{x})]$$

$$AdvRisk_0(h) = Risk(h)$$

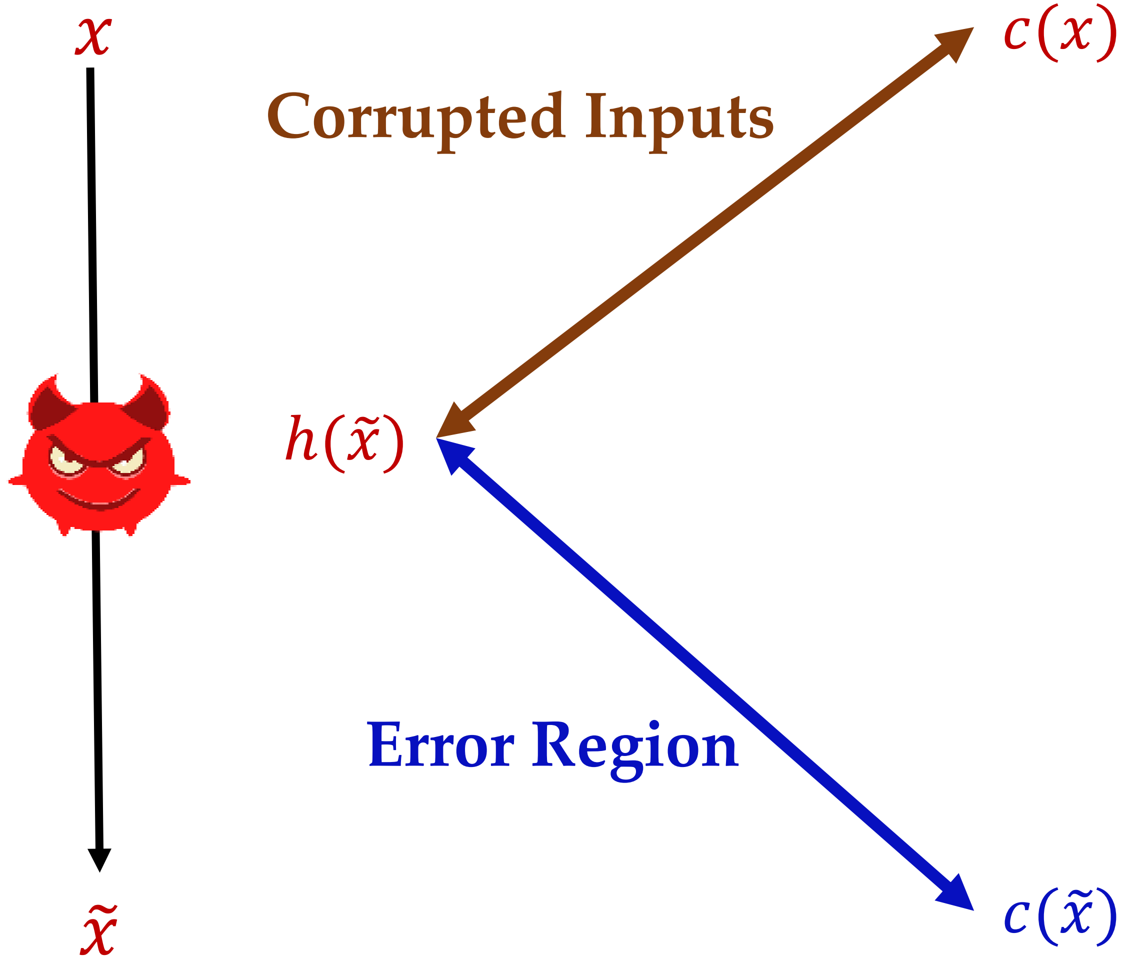
Comparing Definitions of Adversarial Examples

Corrupted inputs

- [Feige Mansour Shapire 15]
- [Madry et al., 17]
- [Feige Mansour Shapire 18]
- [Attias Kontorovich Mansour 19]

Error region

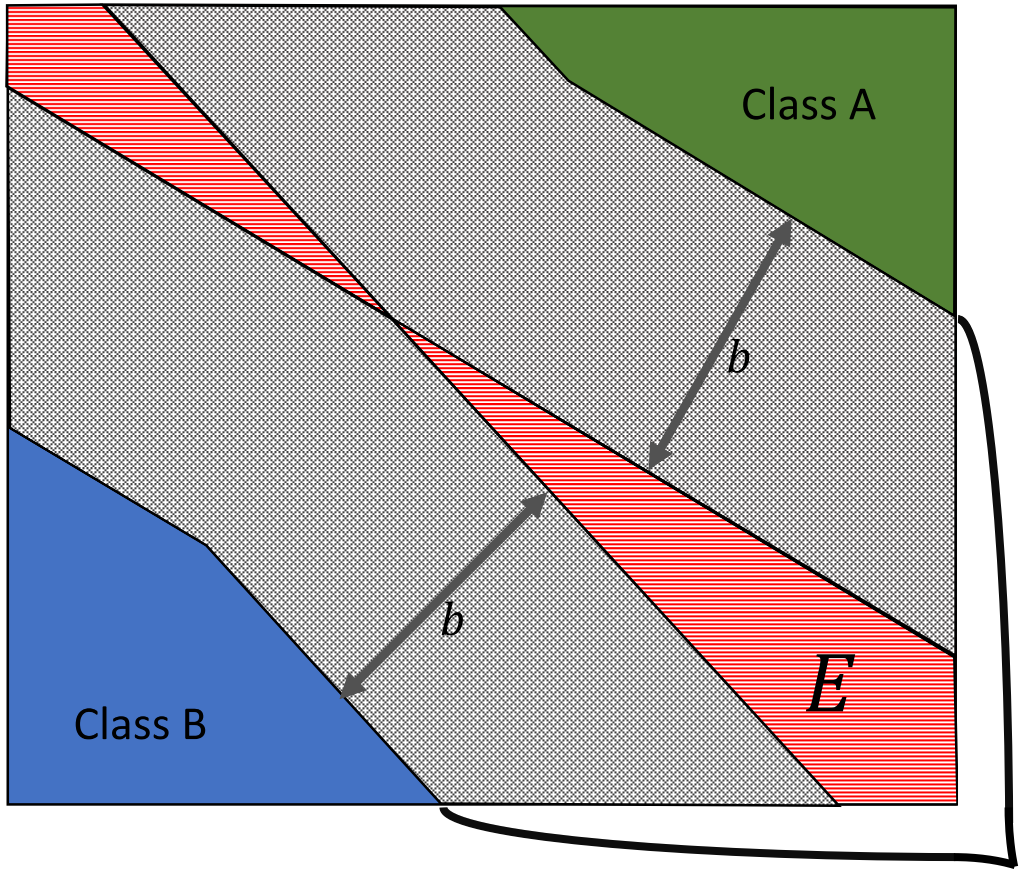
- [Diochnos M Mahmoody 18]
- [Gilmer et al., 18]
- [Bubeck Price Razenshtein 18]
- [Degwekar Vaikuntanatan, 19]



Adversarial Examples from Expansion of Error Region

- Define error region E
  - Error region $E = \{x; h(x) \neq c(x)\}$
  - $\text{Risk}(h) = \Pr[E]$
- $\text{Risk}_b(h) = \Pr[b\text{-expansion } of\ E]$

Adversarial examples almost always exist if the expansion of E covers almost all inputs



b expansion of set E

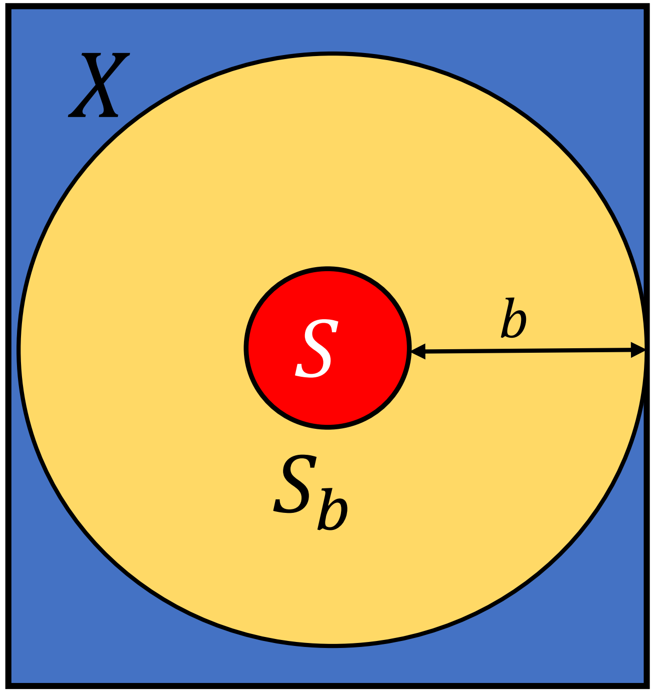
Concentration of Measure

- Metric probability space (M, D) over set X
  - Example: n-dimensional Gaussian with ℓ_2

- b-expansion of set $S \subseteq X$

$$S_b = \left\{x \in D; \min_{s \in S} M(x, s) \leq b\right\}$$

- For any set S with constant probability
  - S_b converges to 1 very fast as b grows
  - i.e. $\Pr[S_b] \approx 1$ for small $b \ll \mathrm{Diam}_M(X)$



Examples of Concentrated Distributions

- Normal Lévy families are concentrated distributions [Lévy 1951]
  - with dimension and diameter n
  - Such that for any S such that $\Pr[S] = 0.01$
  - and for $\mathrm{b} \approx \sqrt{n}$ we have $\Pr[S_b] = 0.99$

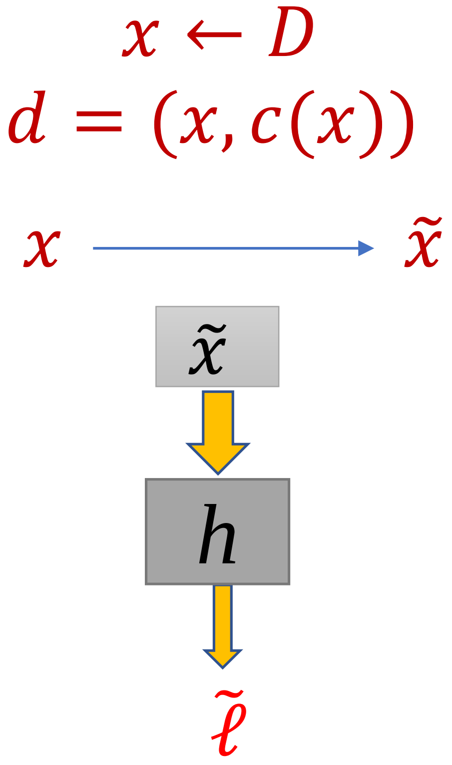
- Examples [Amir & Milman 1980], [Ledoux 2001]:
  - n-dimensional isotropic Gaussian with Euclidean distance
  - n-dimensional Spheres with geodesics distance
  - Any product distribution with Hamming distance (e.g. uniform over Hypercube)
  - And ***many more***…

Main Theorem 1:
Adversarial examples for Lévy families

If (D, M) is Lévy family with both dimension and "typical norm" n:

… then Adversary can add "small" perturbations $b \approx \sqrt{n}$,…

…and increase risk of any classifier with non-negligible (original) risk $\mathrm{Risk}(h) \approx 1/100$ to adversarial risk $\mathrm{AdvRisk}_b(h) \approx 1$,



Previous Work on Provable Evasion Attacks

- Similar attacks using isoperimetric inequalities
  - [Gilmer et al 2017]: Use isoperimetric inequality on n-dimensional spheres
  - [Fawzi et al 2018]: Use isoperimetric inequality on gaussian
  - [Diochnos, Mahloujifar, M 2018]: Use isoperimetric inequality on Hypercube

- Our (Normal Levy) theorem generalizes previous works as special cases and covers many more distributions.

Talk Outline

1a. Defining evasion attacks formally

1b. Evasion attacks from measure concentration of instances

2a. Defining poisoning attacks formally

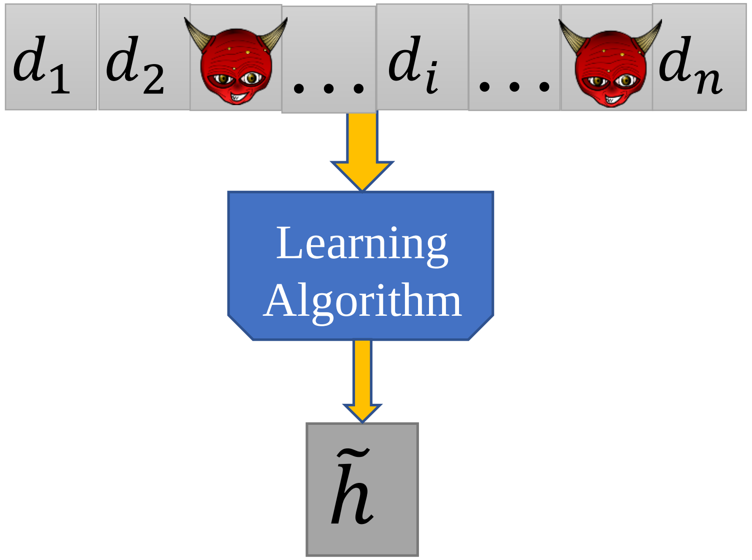
2b. Poisoning attacks from measure concentration of products

3a. Poly-time attacks from computational concentration of products

3b. Connections to attacks on coin-tossing protocols

Poisoning Attacks: Definition

$$x_i \leftarrow D$$
$$d_i = (x_i, c(x_i))$$



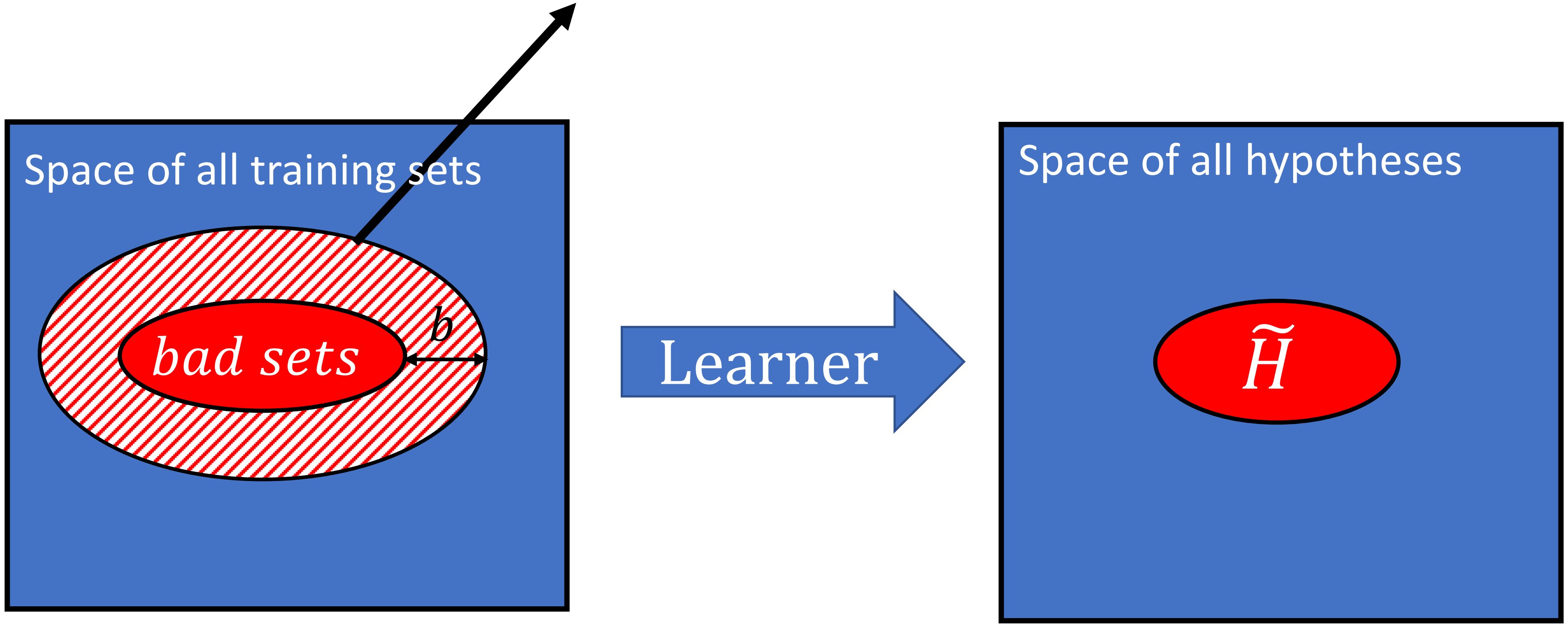
- Hypothesis space H
- $\widetilde{H} \subseteq H$: containing "bad" hypotheses (e.g., those that give me the loan)

Adversary wants to change training set $S = (d_1, \dots, d_n)$ into a "close" (Hamming distance) $\tilde{S}$ such that $\tilde{h} \in \widetilde{H}$

Adversary can depend on D and c (but not on h as it is not produced yet)

Why is concentration also relevant to poisoning?

training sets that are b-close to a bad trainint set



Distribution from which a training set S is sampled is X^m for $X = (D, c(c))$

Recall: Examples of Concentrated Distributions

- Normal Lévy families are concentrated distributions [Lévy 1951]
  - with dimension and diameter n
  - Such that for any S such that $\Pr[S] = 0.01$
  - and for $\mathrm{b} \approx \sqrt{n}$ we have

$$\Pr[S_b] \approx 1$$

- Examples [Amir & Milman 1980], [Ledoux 2001]:
  - n-dimensional isotropic Gaussian with Euclidean distance
  - n-dimensional Spheres with geodesics distance
  - **Any product distribution with Hamming distance**
  - And ***many more***…

Main Theorem 2:
Poisoning attacks from concentration of products

- For any deterministic learner L and any $\widetilde{H}$ where

$$\Pr\left[\widetilde{H}\right] = 1/100$$

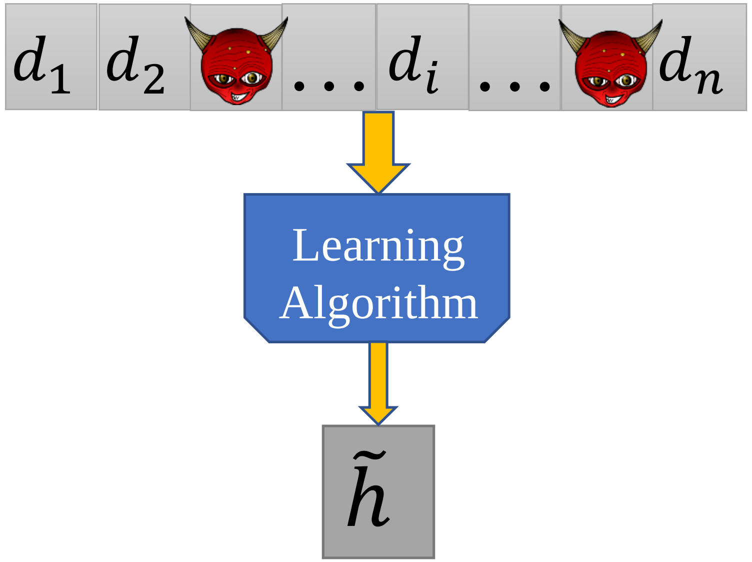
$$x_i \leftarrow D$$

$$d_i = \left(x_i, c(x_i)\right)$$

Adv can change $\approx \sqrt{m}$ fraction of training data

and make probability of getting $\tilde{h} \in \widetilde{H} \approx 1$

while the poisoned data are **still correctly labeled**!



Other works on “clean label” poisoning attacks:

- [Mahloujifar, M TCC-2017] Defined **p-tampering** poisoning attacks, which are Valiant’s malicious noise but only using correct/clean labels.
- [Mahloujifar, Diochnos, M ALT-2018] positive and negative results for PAC-learning under p-tampering attacks
- [Shafahi et al, NeurIPS-2018] practical attacks using clean labels
- [Turner et al, ICLR-2018] backdoor attacks using clean labels

Talk Outline

1a. Defining evasion attacks formally

1b. Evasion attacks from measure concentration of instances

2a. Defining poisoning attacks formally

2b. Poisoning attacks from measure concentration of products

3a. Poly-time attacks from computational concentration of products

3b. Connections to attacks on coin-tossing protocols

Concentration of Products -- a Closer Look

Proposition 2.1.1 in [Talagrand 1994]

- Let $\mathrm{HD}(\cdot,\cdot)$ be Hamming distance and $\mathrm{HD}(x, S) = \min_{s \in S} \mathrm{HD}(x, s)$
  Let D be any distribution and D^n its n-fold product
  Let S be any target set of probability $\mu = \Pr[D^n \in S]$

- Then the probability of being b-far from S is bounded:

$$\Pr_{x \leftarrow D^n}[\mathrm{HD}(x, S) \geq b] \leq \frac{e^{-b^2/n}}{\mu}$$

- Example: if $\mu = 1/poly(n)$ then 99% of samples from D^n **are in $\approx \sqrt{n}$ Hamming Distance from <u>some point</u> in S**

Algorithmically finding such points in S?

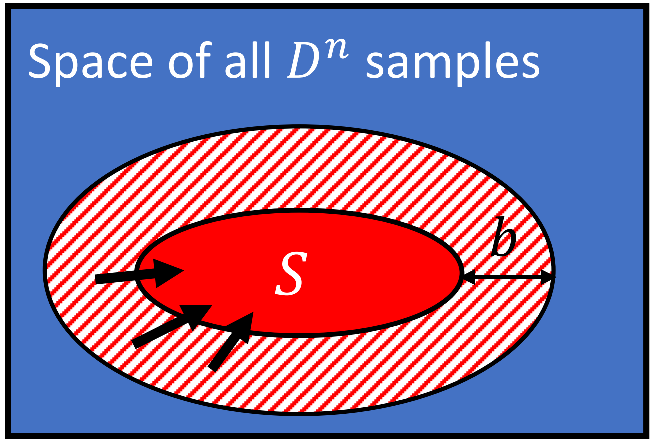
- Recall formal setting:
  Let D be any distribution and D^n its n-fold product
  Let S be any target set of probability $\mu = \Pr[D^n \in S] \geq 1/poly(n)$

- Suppose algorithm A runs in $poly(n)$ while having oracle access to membership in S and to sampler for D

- Question: given input $x \leftarrow D^n$ can A find (with high probability over x) a "close" point $s \in S$ such that

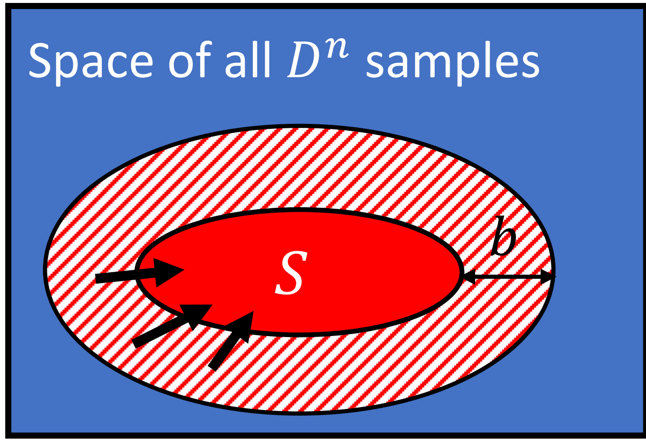
$$\mathrm{HD}(x, s) = \tilde{O}(\sqrt{n})$$

Can we compute the arrow ↗ mapping efficiently?



Man Theorem 3:
Computational Concentration of Products

- Yes we can! compute the arrow mapping efficiently in product distributions under Hamming distance



- More formally:
  If $\Pr[D^n \in S] \geq 1/poly(n)$ → there is a $poly(n)$ time A who finds, with high probability over the input $x \leftarrow D^n$, a “close” point $s \in S$ where

$$\mathrm{HD}(x, s) = \tilde{O}(\sqrt{n})$$

Talk Outline

1a. Defining evasion attacks formally
1b. Evasion attacks from measure concentration of instances

2a. Defining poisoning attacks formally
2b. Poisoning attacks from measure concentration of products

3a. Poly-time attacks from computational concentration of products
3b. Connections to attacks on coin-tossing protocols

A Stronger Result: Attacking Single-Message Coin Tossing Protocols

- Let $P_1, \dots P_n$ run a coin tossing protocol in which P_i sends i^{th} message m_i
- Suppose $\Pr[f(m_1, \dots m_n) = \textbf{heads}] \geq 1/\text{poly}(n)$
- If Adv can corrupt up to b of the parties and it can decide to corrupt or not **by looking at** their locally prepare message m_i
- Then Adv can make $\Pr[f(m_1, \dots m_n) = \textbf{heads}] \approx 1$
- Model is the **strong adaptive** corruption of [Goldwasser,Kalai,Park 2015] who proved a similar **exponential time** attack for 1-round protocols.

Conclusion

- Formalizing security notions in adversarial ML is important.
  Different definitions (though equivalent in some cases) behave differently

- **Concentration of measure** phenomenon can potentially lead to both evasion and poisoning attacks.

- Product distributions are even **computationally** concentrated under Hamming distance due to certain polynomial-time coin-tossing attacks