A Characterization of Semi-Supervised Adversarially Robust PAC Learnability

Idan Attias

Ben-Gurion University

Joint work with:

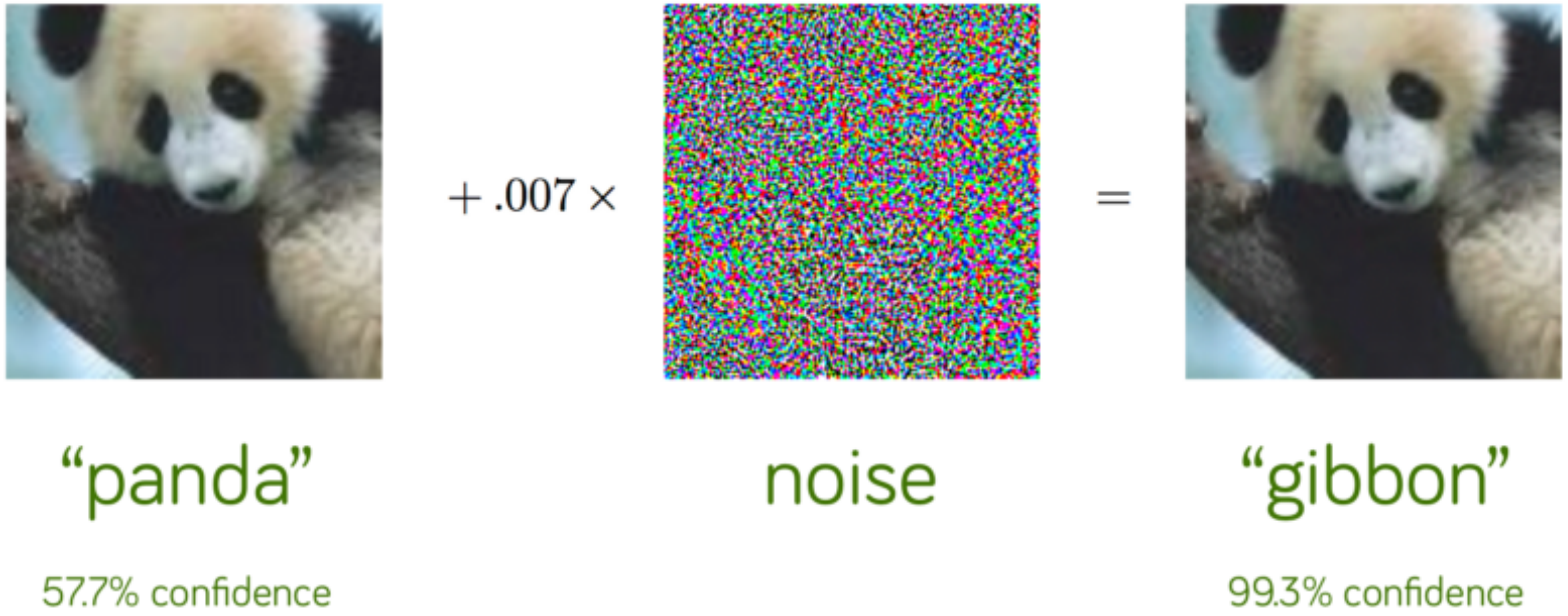
Steve Hanneke

Purdue University

Yishay Mansour

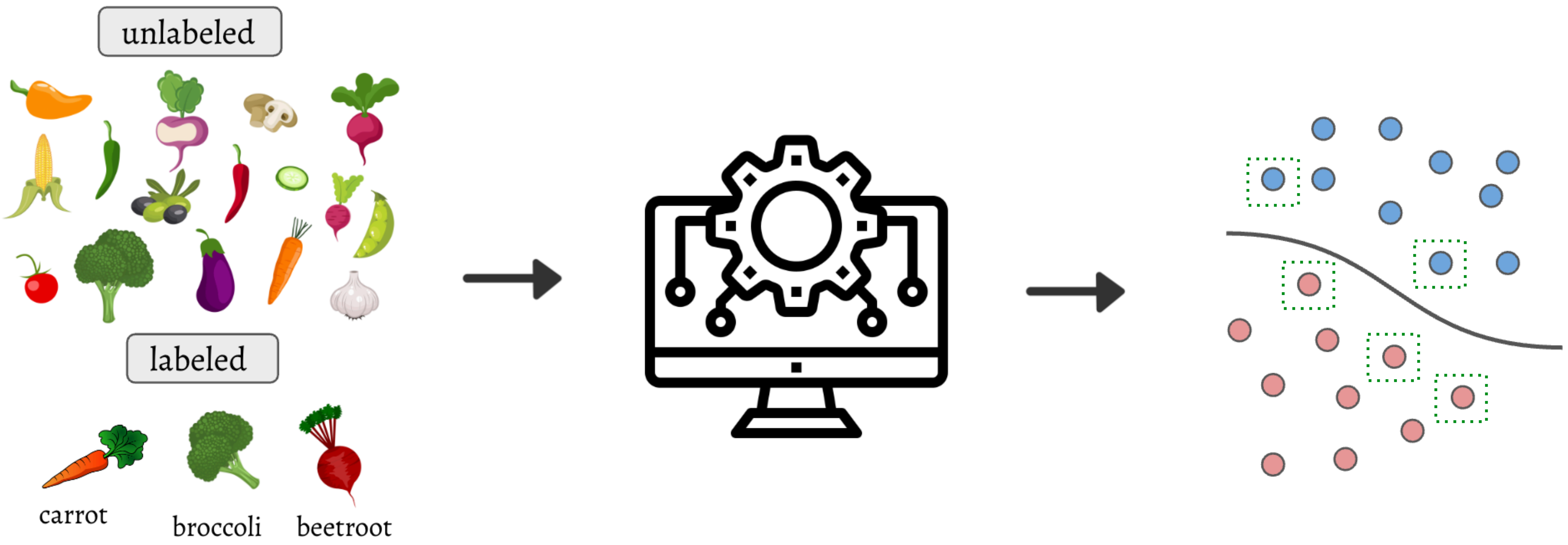
Tel Aviv University and Google Research

Adversarial Examples



Goodfellow, Shlens, Szegedy, ICLR ‘15

Main Question



Question:
How many labeled and unlabeled samples are sufficient for learning a robust classifier in the PAC model?

Answer (informal):
The labeled sample size can be arbitrarily smaller than the unlabeled one, and controlled by a different complexity measure.

Semi-Supervised Robust PAC Learning

- Unknown distribution D over $X \times \{0,1\}$.
- Perturbation function $U : X \to 2^X$.
- Robust error of classifier $h : X \to \{0,1\}$: $\mathrm{err}_U(h) = \mathbb{E}_{(x,y)\sim D}\left[\sup_{z\in U(x)} \mathbb{I}\{h(z) \neq y\}\right]$.
- Semi-Supervised learning algorithm A^{ss}:

  Input: $S^l = \{(x_i, y_i)\}_{i=1}^n$ and $S^u = \{x_j\}_{j=n+1}^m$, $(x_i, y_i) \sim D$, and $x_j \sim D_X$.

  Output: $\hat{h}_{n,m-n}$.

- Definition (semi-supervised learning): $H \subseteq \{0,1\}^X$ is robustly learnable in the realizable case, $\inf_{h\in H} \mathrm{err}_U(h) = 0$, if $\exists$ algorithm A^{SS}, s.t. $\forall \epsilon, \delta$, $\forall D$, with probability $1-\delta$, $\mathrm{err}_U(A^{SS}) \leq \epsilon$, using $M^l(\epsilon,\delta) < \infty$ labeled examples and $M^u(\epsilon,\delta) < \infty$ unlabeled examples.

- $M^l(\epsilon,\delta)$ and $M^u(\epsilon,\delta)$ are called the Sample Complexity.
- Agnostic case: $\mathrm{err}_U(A^{SS}) \leq \inf_{h\in H} \mathrm{err}_U(h) + \epsilon$.

Supervised Robust PAC Learning

- Definition (supervised learning):

$H \subseteq \{0,1\}^X$ is robustly learnable in the realizable case, $\inf_{h \in H} \mathrm{err}_U(h) = 0$, if $\exists$ algorithm A^S, s.t. $\forall \epsilon, \delta$, $\forall D$, with probability $1 - \delta$, $\mathrm{err}_U(A^S) \le \epsilon$, using $\Lambda^S(\epsilon, \delta) < \infty$ labeled examples.

- Montasser, Hanneke, Srebro (COLT '19):

$$\frac{RS_U(H)}{\epsilon} + \frac{\log 1/\delta}{\epsilon} \lesssim \Lambda^S(\epsilon, \delta) \lesssim \frac{VC(H)VC^*(H)}{\epsilon} + \frac{\log 1/\delta}{\epsilon}.$$

  - $VC^*(H) \le 2^{VC(H)}$.
  - $\exists H$, s.t. $RS_U(H) \ll VC(H)$.

Main Result

- Realizable:

$$M^l(\epsilon, \delta) \lesssim \frac{VC_U(H)}{\epsilon} + \frac{\log 1/\delta}{\epsilon}.$$

  - VC_U dimension d is the largest number s.t. $\exists x_1, \ldots, x_d$ and all 2^d classifications of the entire perturbation set $U(x_1), \ldots, U(x_1)$ are realized by a function in H.
  - $\exists H,$ s.t. $VC_U(H) \ll RS_U(H) \to M^l(\epsilon, \delta) \ll \Lambda^S(\epsilon, \delta)$.

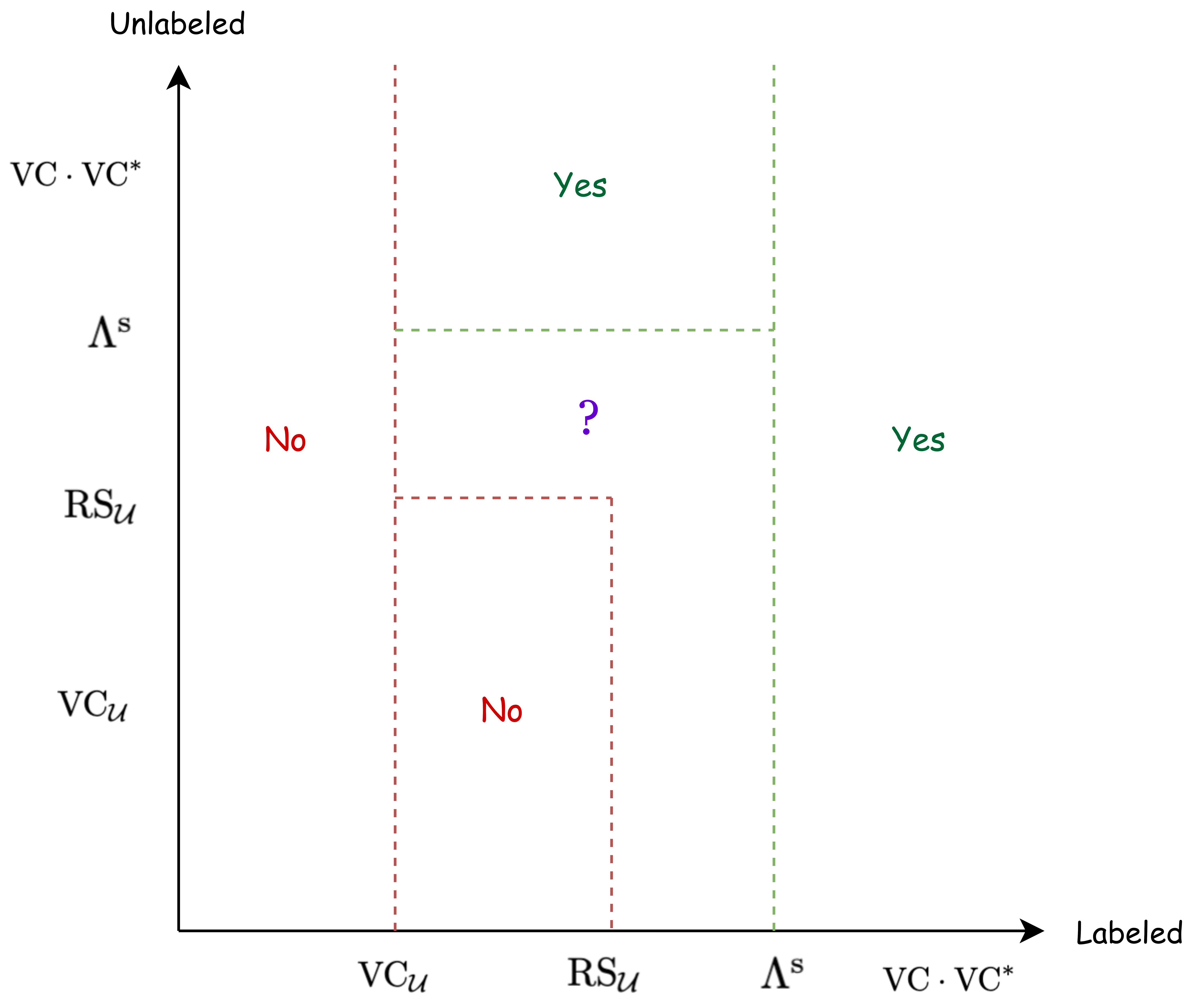
$$M^u(\epsilon, \delta) \lesssim \Lambda^S(\epsilon, \delta) = \text{ sample size required for fully supervised learning .}$$

- Agnostic: Improved labeled sample complexity with error $3\text{OPT(H)} + \epsilon$.

  Impossible to improve for $\text{OPT(H)} + \epsilon$.

Summary

Sample complexity for semi-supervised adversarially-robust learning



Algorithmic Idea

Generic semi-supervised learner:

1. Preprocess step: keep only functions in H that are robustly self-consistent.
2. Learn the new class with the 0-1 loss.
3. Use the output of step 2 to label an unlabeled sample.
4. Execute a fully-supervised robust learner.