Surprising Failures of Standard Practices in ML

When the Sample Size is Small

December 3rd 2022, ICBINB NeurIPS Workshop

Fanny Yang, joint work with **J. Clarysse, A. Tifrea**

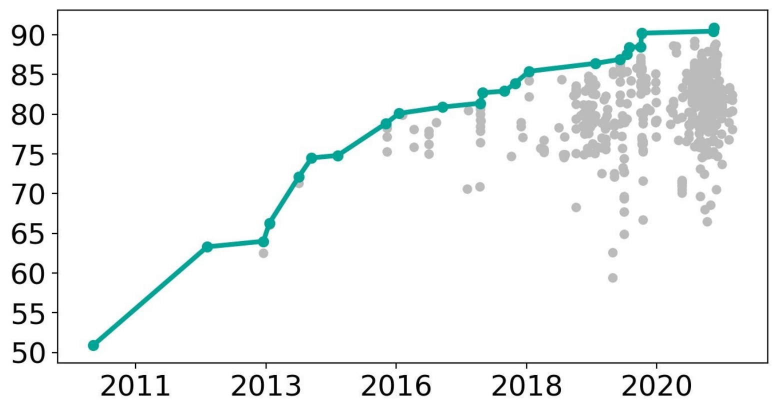
Statistical Machine Learning group, CS department, ETH Zurich

Reliability crisis in modern supervised learning

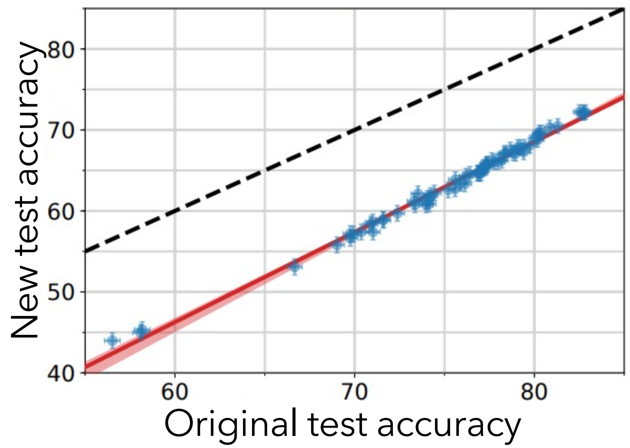
Modern ML works well …

sometimes maybe not so much…

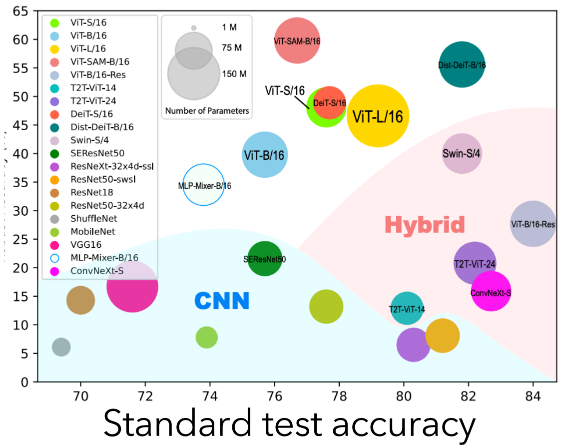
Top1 test accuracy on ImageNet



Top1-accuracy on new ImageNet



Robust accuracy on ImageNet-1k



But how can we know **when** a new method fails to perform well?

Sources: [Recht et al. '19], [Shao et al. '22]

One role of theory: failure case characterization

Empirical "wisdom:

Fancy method better than vanilla method

Empirical failure hypothesis:

- Fancy method worse than vanilla method
- Failure worse under key properties

Theory: Prove in (simplified) setting

- Fancy method worse than vanilla method
- Failure worse under key properties

Two examples in this talk:

- Failure I: Uncertainty sampling worse than Uniform sampling
- Failure II: Adversarial training worse than Standard training

Failure I: When uncertainty sampling is worse than uniform sampling

joint work with Alexandru Tifrea, Jacob Clarysse

Active learning via uncertainty sampling 😊

Goal: Find model θ with low test error Err(θ)= $\mathbb{E}_{x,y}\ell(y, f_\theta(x'))$ using fixed labeling budget n_ℓ

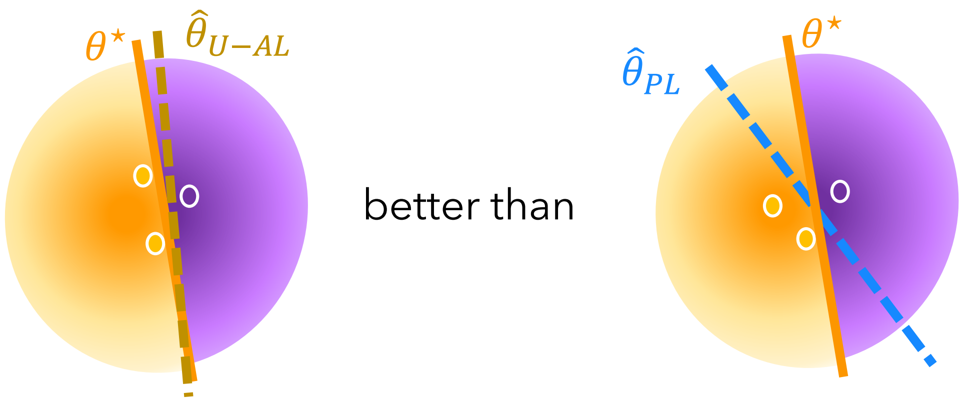
Simple and hence often used: Uncertainty based active learning (U-AL)

Given uncertainty score, large unlabeled dataset D_u, labeled seedset D_ℓ of size n_{seed}

At iteration t:

- Query label y^t for sample in D_u with highest uncertainty score for model θ^{t-1}
- Remove sample from D_u, add labeled sample to D_ℓ, train θ^t on D_ℓ

uncertainty based active learning (U-AL)

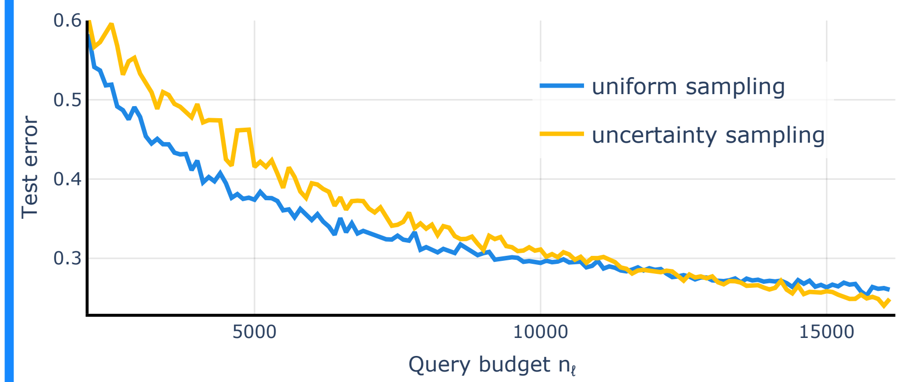


uniform sampling / passive learning (PL)

Failure of uncertainty sampling

Empirically often reported to fail!

e.g. ResNet18 on CIFAR-100



Theoretically grounded explanations

- "cold start" & bad uncertainty estimates

  e.g. [Huang et al. '14], [Sener et al. '18]

- large noise / high Bayes error

  [Mussmann et al. '18]

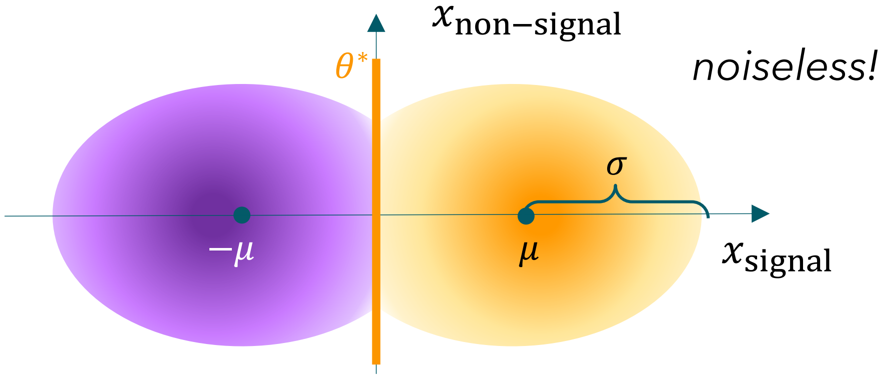
Our work: Different reason why U-AL fails, even with "optimal" uncertainty & noiseless data

Theoretical results → new failure hypothesis

n_ℓ labeled samples from d-dimensional covariates

- $x_{\text{signal}} \sim$ truncated Gaussian mixture

  $x_{\text{non-signal}} \sim$ isotropic Normal $N(0, I)$



Theorem [TCY '22] (informal):

For $n_\ell \ll d$, large enough unlabeled dataset

$$\text{Err}(\hat{\theta}_{\text{U-AL}}) - \text{Err}(\hat{\theta}_{\text{PL}}) > 0 \text{ w.h.p.}$$

Further, the error gap increases for smaller

1. $\frac{n_\ell}{d}$ (query budget)
2. $\frac{\mu}{\sigma}$ (class separation).

- $\hat{\theta}$: linear SVM solution on labeled dataset
- Uncertainty score: distance to decision boundary of current (or optimal) model

Theoretical results → new failure hypothesis

Empirical hypothesis: For test accuracy

U-AL may be worse than PL even for

noiseless data and oracle uncertainty if

1. budget is small
2. a lot of unlabeled data near optimal decision boundary

←

Theorem [TCY '22] (informal):

For $n_\ell \ll d$, large enough unlabeled dataset

$$\mathbf{Err}(\hat{\theta}_{\text{U-AL}}) - \mathbf{Err}(\hat{\theta}_{\text{PL}}) > 0 \text{ w.h.p.}$$

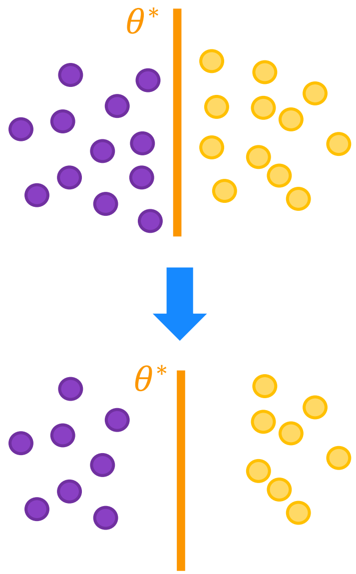
Further, the error gap increases for smaller

1. $\frac{n_\ell}{d}$ (query budget)
2. $\frac{\mu}{\sigma}$ (class separation).

Key property ② : class separation

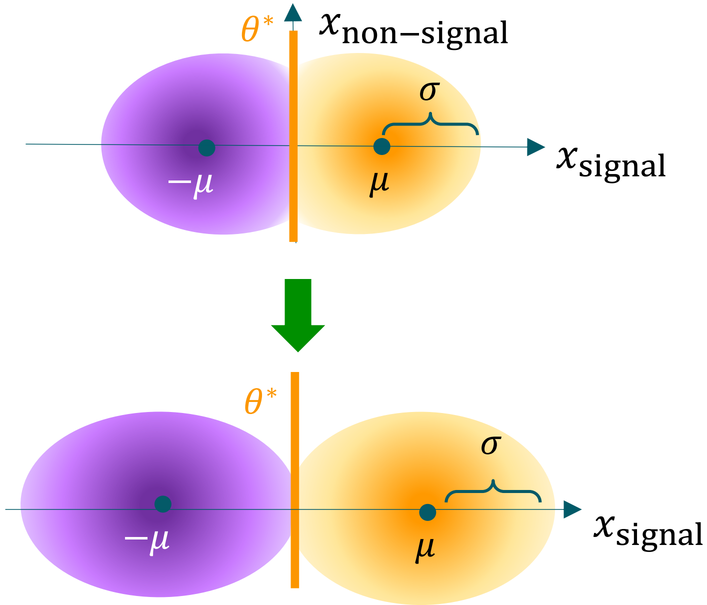
More class separation on empirical dataset:

Removing % samples closest to decision boundary $\theta^\star$ trained on whole dataset



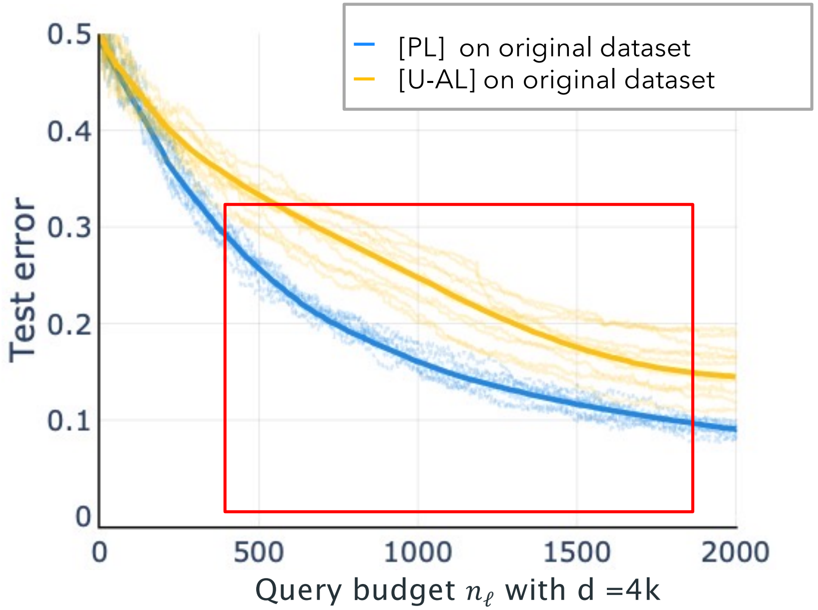
More class separation in theory

Larger mean separation μ in signal direction

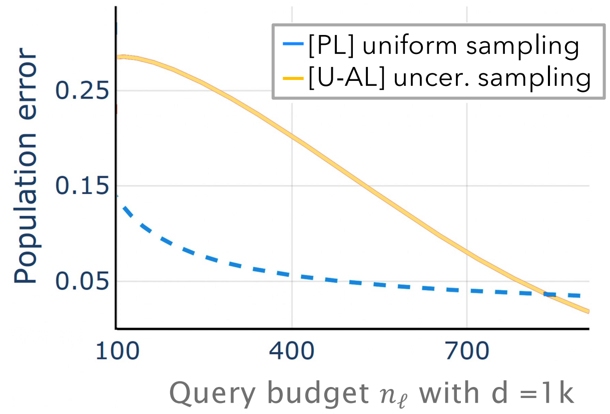


Empirical validation: ① Failure of small label budget

Empirical confirmation



Theory



Gap between U-AL and PL increases with smaller

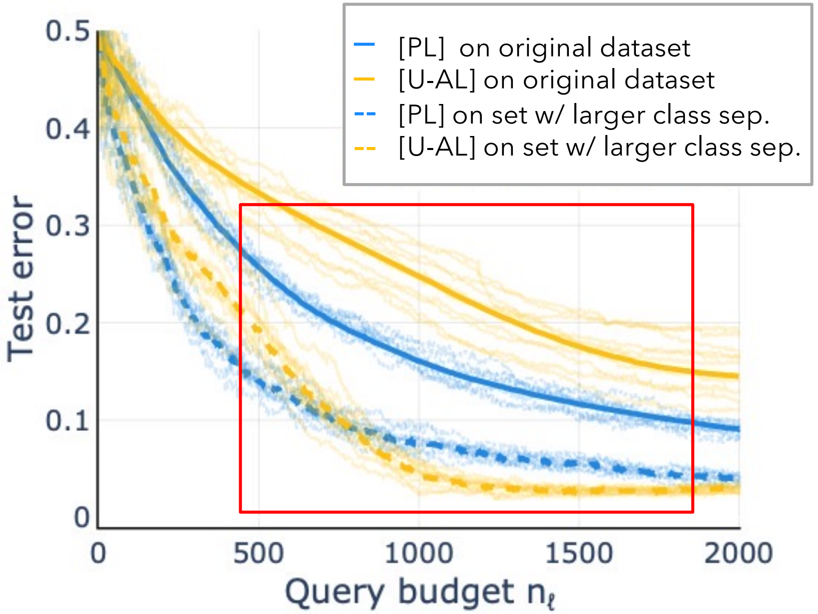
① $\frac{n_\ell}{d}$ (label budget)

Happens in a small-sample regime that is still relevant (test accuracy ~ 80%)

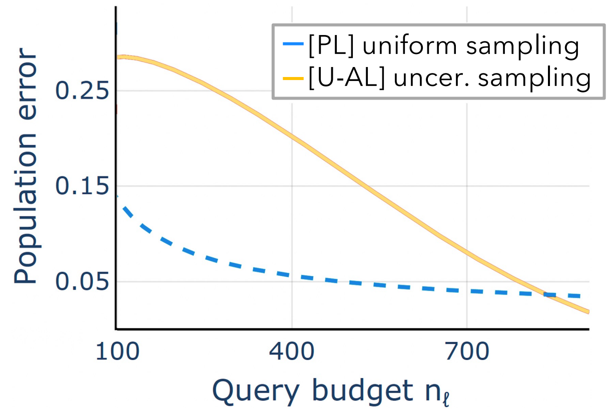
Binary classification dataset: Riccardo [OpenML]

Empirical validation: ② Failure for small separation

Empirical confirmation



Theory



Gap between U-AL and PL increases with smaller

① $\frac{n_\ell}{d}$ (label budget) ② $\frac{\mu}{\sigma}$ (class separation).

Happens in a small-sample regime that is still relevant (test accuracy ~ 80%)

Binary classification dataset: Riccardo [OpenML]

Failure II: When adversarial training hurts robust generalization

joint work with Jacob Clarysse, Julia Hörrmann

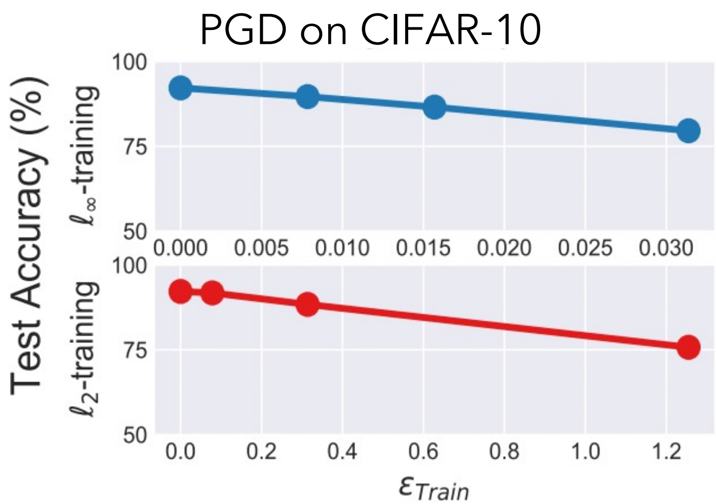
Adversarial robustness and adversarial training

Goal: Low robust error RobErr(θ) = $\mathbb{E}_{x,y} \max_{x' \in T(x,\epsilon)} \ell(y, f(x';\theta))$ w/ $T(x,\epsilon)$: set of ϵ-perturbed versions of x

Adversarial training (AT)
At iteration t: ◦ for each x_i in mini-batch, find adversarial example $x_i' = \mathrm{argmax}_{x \in T(x_i,\epsilon)} \ell(y_i, f(x;\theta^t))$ ◦ SGD step on loss w.r.t. θ^t at adversarial points x_i'

But: Known caveat of adversarial training (AT) 🙁

Adversarial training hurts **std. accuracy**



Theoretically grounded explanations:

- optimal classifiers not robust (inherent tradeoff), e.g. [Tsipras et al. '19, Zhang et al. '19…]
- robust model more complex [Nakkiran et al. '19]
- wrong inductive bias [Raghunathan et al. '20]

Our work: AT may have worse **adv. robust accuracy** even w/o inherent tradeoff in well-specified setting

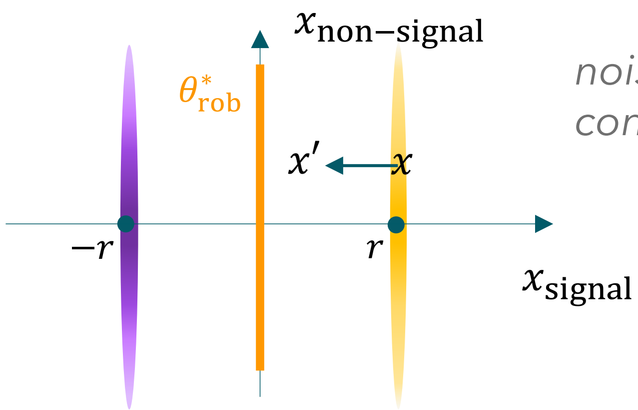
Figure source: [Tsipras et al. '19]

Theoretical results → new failure hypothesis

n samples from d-dimensional covariates

- $x_{\text{signal}} = r \cdot y\, \theta^\star$ for $y \sim U(\{-1, +1\})$

  $x_{\text{non-signal}} \sim$ isotropic Normal $N(0, I)$



noiseless!
consistent!

- Perturbation set: $T(x;\epsilon) = \{x + \delta\theta^\star \text{ with } |\delta| \leq \epsilon\}$
- $\hat{\theta}$: GD until convergence on (robust) logistic loss

Theorem [CHY '22] (informal):

For $n < d$, almost surely

RobErr($\widehat{\theta}_{AT}$) – RobErr($\widehat{\theta}_{ST}$) > 0

Further, the error gap increases for

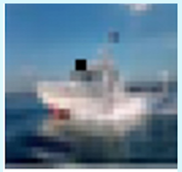
1. smaller $\frac{n}{d}$ (sample size)
2. if attack always reduces signal

Theoretical results → new failure hypothesis

Empirical hypothesis: For robust accuracy

AT may be worse than ST

1. budget is small
2. attacks directed to object, such as masks, illumination, motion blur



Theorem [CHY '22] (informal):

For $n < d$, almost surely

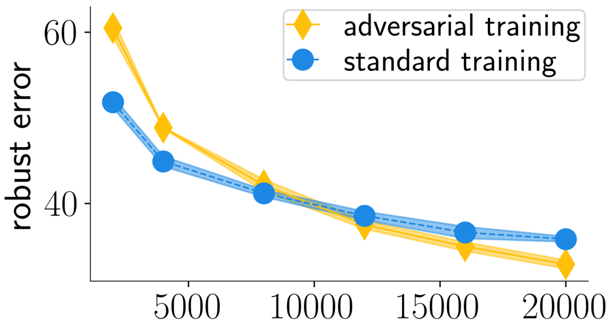
$$\mathbf{RobErr}(\widehat{\theta}_{AT}) - \mathbf{RobErr}(\widehat{\theta}_{ST}) > 0$$

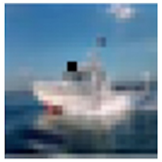
Further, the error gap increases for

1. smaller $\frac{n}{d}$ (sample size)
2. if attack always reduces signal

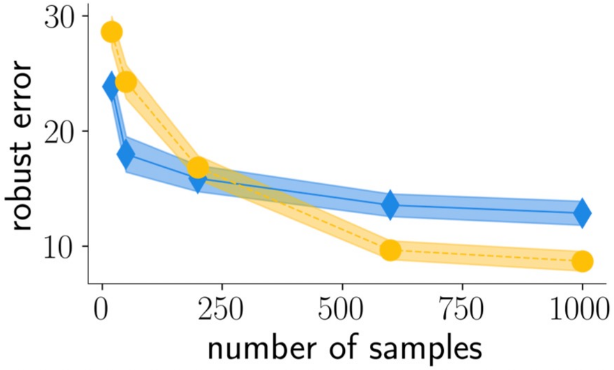
Empirical validation: Failure for small sample directed attacks

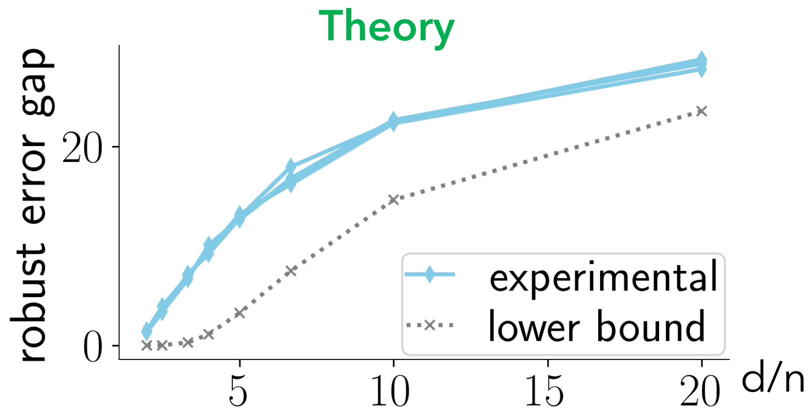
AT worse for mask attack on CIFAR-10





… and illumination attacks on Waterbirds





AT worse than ST for

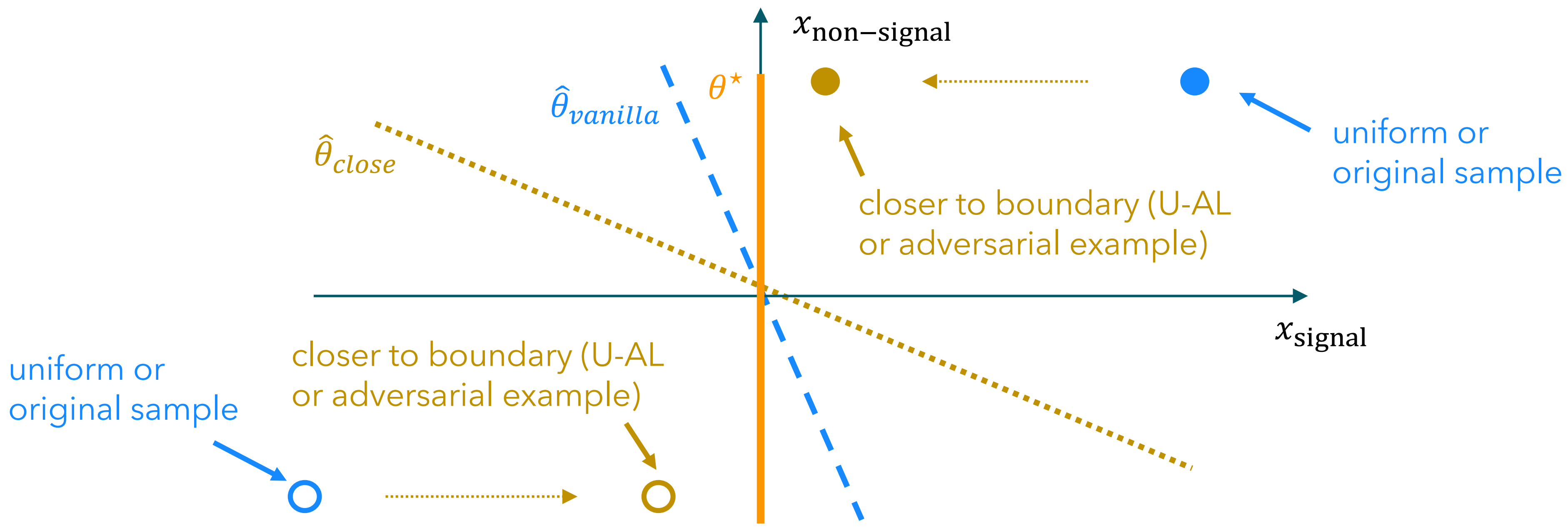
1. small $\frac{n}{d}$ (sample size)
2. if attack always reduces signal

Happens in a small-sample regime that is still relevant (standard accuracy ~ 80%)

Common proof intuition for both failure cases

What do AT and (oracle) U-AL have in common?

⇒ Models trained on points closer to good dec. boundary ($\hat{\theta}_{close}$)



Summary: Theory-guided failure case hypotheses

Empirical ”wisdom:

Fancy method better than vanilla method

Empirical take-away for practitioner

Well-established practical methods

should be used with care, in particular

1. in the small sample size regime
2. when another key property holds

Theory

For logistic regression,

Fancy method worse than vanilla method

1. in the small-sample regime
2. when another key property holds

References, also to more failure cases in modern ML

SML group: sml.inf.ethz.ch

Papers discussed in this talk

- Clarysse, Hörmann, Yang "**Why adversarial training can hurt robust accuracy**", arxiv preprint '22
- Tifrea, Clarysse, Yang "**Uncertainty vs. uniform sampling: When being passive is better than being active**", arxiv preprint '22

Further "failures" identified in our group:

- Bartolomeis, Clarysse, Yang, Sanyal "**Certified defenses hurt generalization**", this workshop
- Sanyal*, Hu*, Yang "**How unfair is private learning?",** UAI 2022
- Aerni*, Milanta*, Donhauser, Yang **"Strong inductive biases provably prevent harmless interpolation"**, on OpenReview