# Constants Matter: The Performance Gains of Active Learning

ICML 2022 Spotlight

**Stephen Mussmann**, Sanjoy Dasgupta

UNIVERSITY of WASHINGTON

UC San Diego

# Active learning

- Acquiring labels for machine learning systems can be **very expensive** (medical images require doctors)
- Active learning attempts to **reduce the number of labels** needed to learn by adaptively choosing data points to label.
- Empirically, active learning sometimes helps but **often doesn't**.
- Theoretical active learning often focuses on the type of **dependence of error on the number of samples** (e.g., excess error decays as $2^{-cn}$).

# Tsybakov Noise

- Tsybakov noise: $\kappa \in [1, \infty)$ function of data distribution (Tsybakov, 2004)
  - Intuitively, "shape" of noise near decision boundary
- Exponential gains possible for $\kappa = 1$ (e.g., linear classifiers on the uniform sphere (Balcan et al., 2007)).
- Existing lower bounds (for any algorithm) show that the excess error for fixed Tsybakov noise $\kappa$ on the uniform sphere is $\Omega\left(\left(\frac{1}{n}\right)^{\frac{\kappa}{2(\kappa-1)}}\right)$ (Wang & Singh, 2016) for $n$ labeled points.

# Motivation

- What is a “practical” value for $\kappa$?
- Are there lower bounds (for any algorithm) for a simple, benign setting?
- In addition to the dependence on the number of samples $n$, what are the problem-dependent constants?

# Our setting

- Inputs **uniformly** drawn from $d$-dimensional unit sphere: $X \in S^{d-1}$
- True parameters with fixed norm: $w^* \in M \cdot S^{d-1}$
- Binary labels

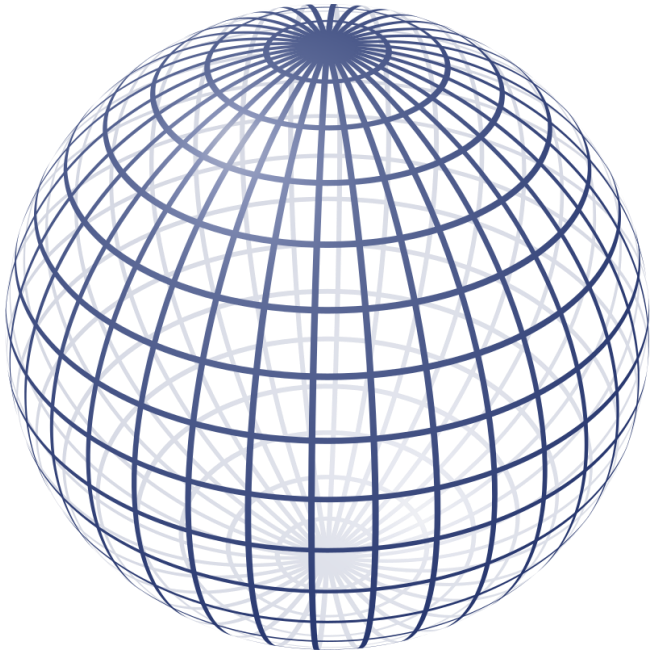
$$Y \in \{-1,1\}$$

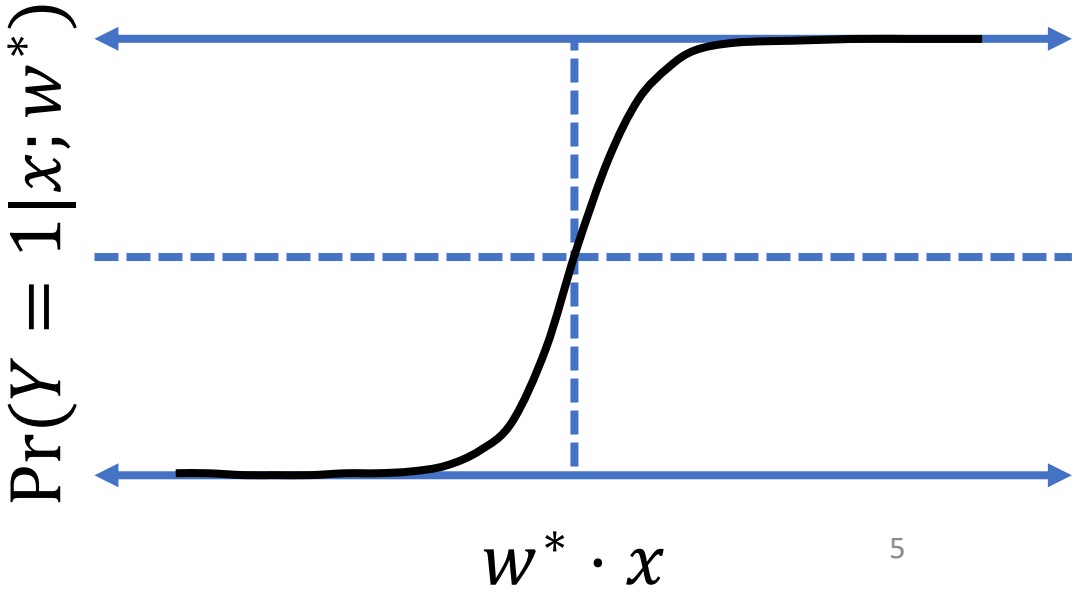
- Logistic conditional label distribution:

$$\Pr(Y = 1|\mathrm{x}; \mathrm{w}^*) = \sigma(\mathrm{w}^* \cdot x)$$

$$\sigma(u) = \frac{1}{1+e^{-u}}$$

- For this setting, $\kappa = 2$



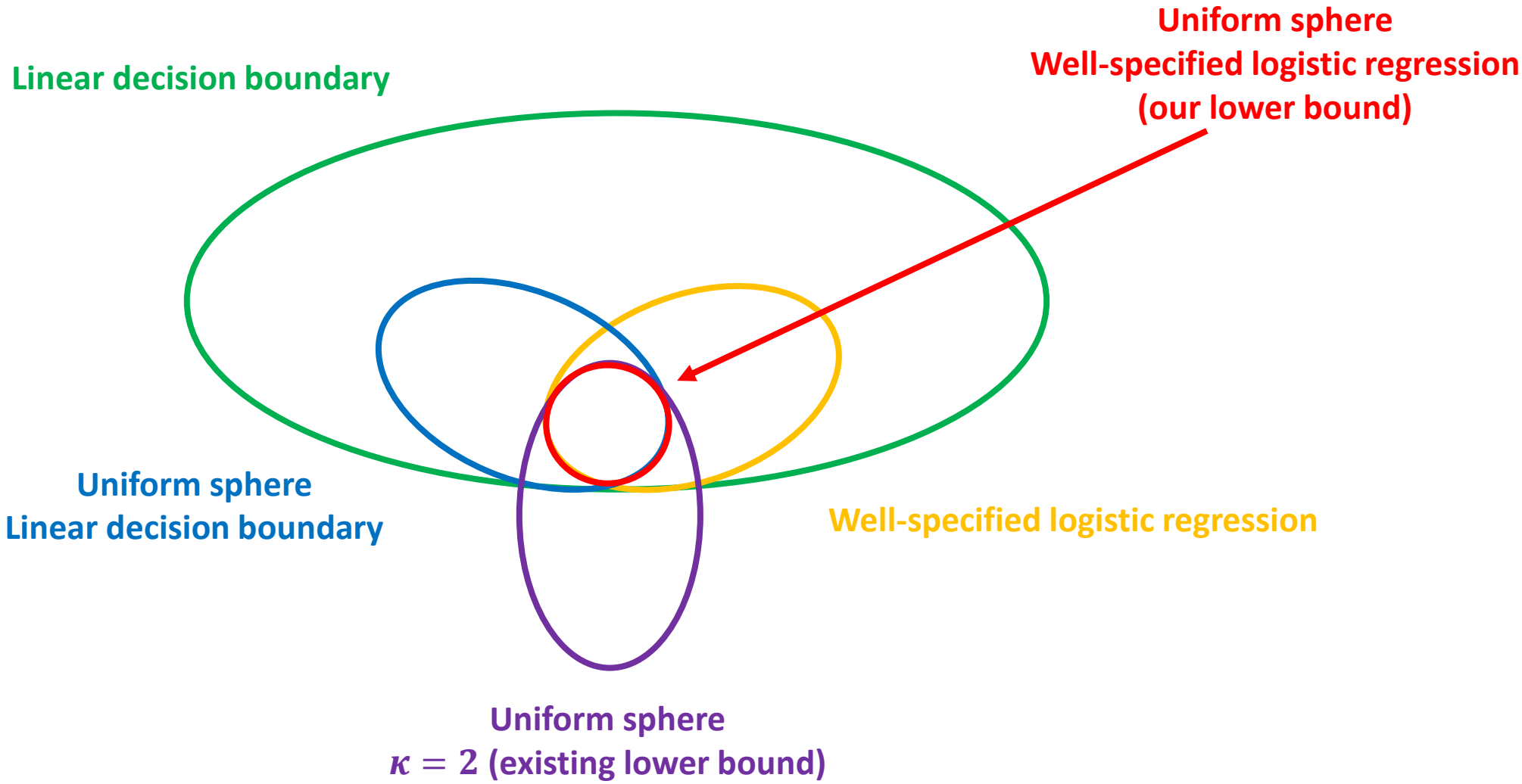


# Our results

- Define $\mathrm{err}^*$ as the Bayes error.
- Up to **universal constants**, the expected excess (w.r.t. Bayes) error is:

| | Random Sampling | Adaptive Selection |
|---|---|---|
| Upper Bound | $O\left(\frac{d \log d}{n}\right)$ | $O\left(\mathrm{err}^* \frac{d}{n}\right)$ |
| Lower Bound | $\Omega\left(\frac{d}{n}\right)$ | $\Omega\left(\mathrm{err}^* \frac{d}{n}\right)$ |

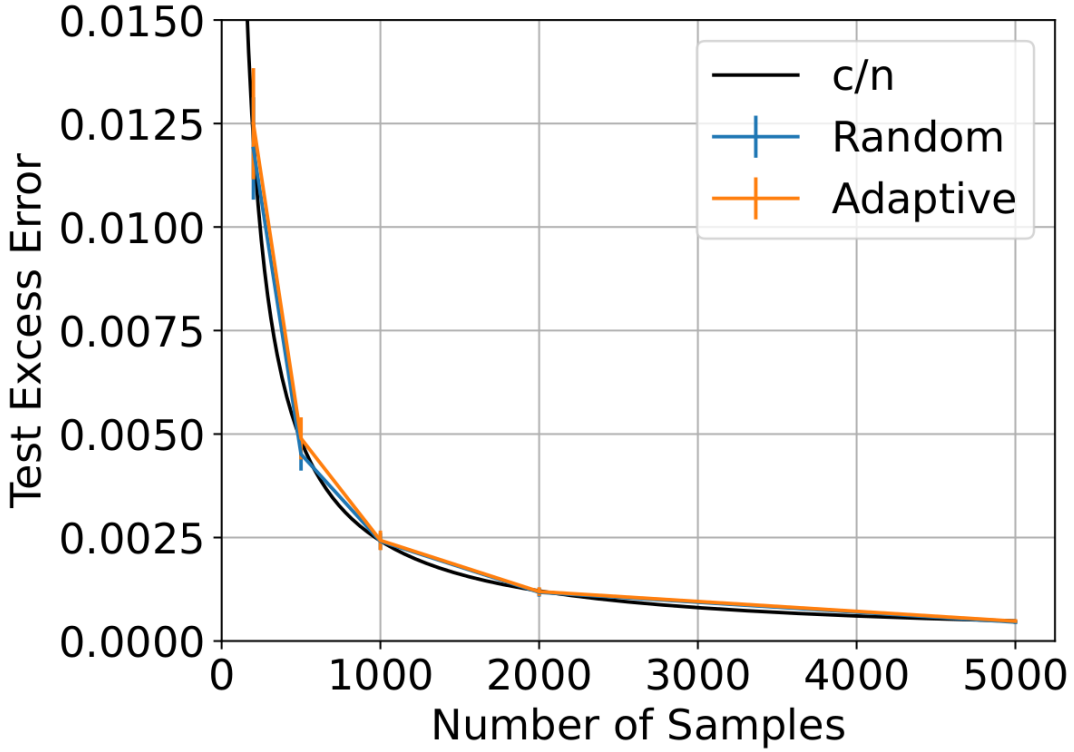
# More specific lower bounds are stronger!

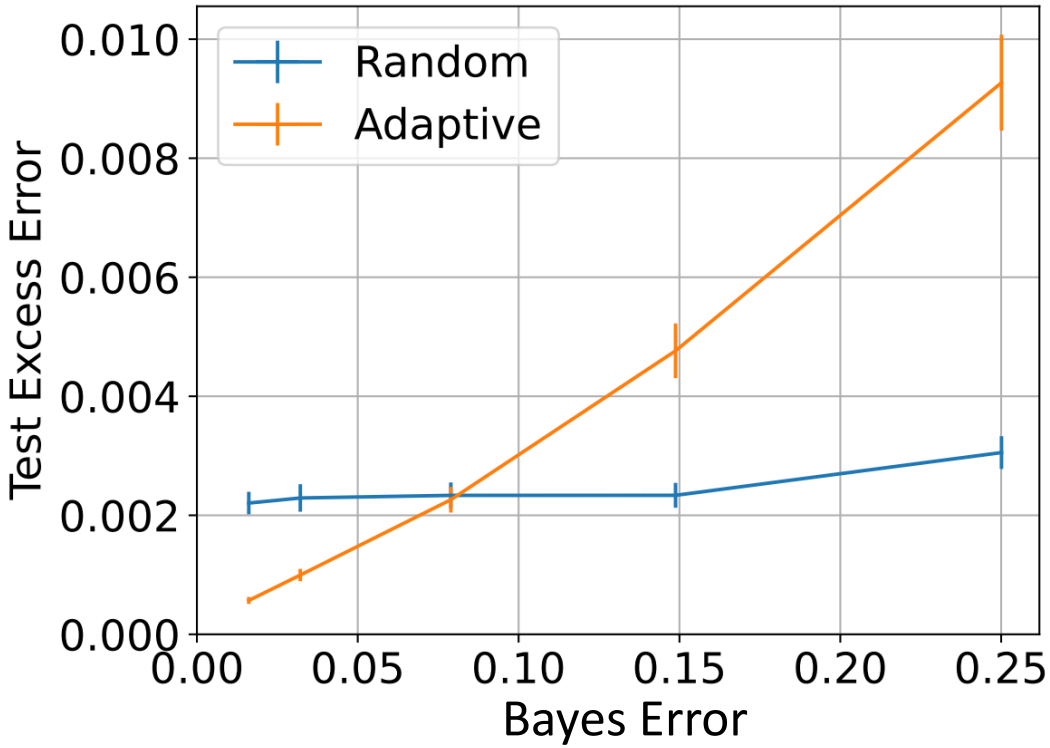


# Synthetic Experiments

Random sampling: ERM with logistic loss
Adaptive selection: uncertainty sampling variant

| | Random Sampling | Adaptive Selection |
|---|---|---|
| Upper Bound | $O\left(\frac{d \log d}{n}\right)$ | $O\left(\mathrm{err}^* \frac{d}{n}\right)$ |
| Lower Bound | $\Omega\left(\frac{d}{n}\right)$ | $\Omega\left(\mathrm{err}^* \frac{d}{n}\right)$ |




Unless otherwise specified: $d = 10,\ n = 1000,\ \mathrm{err}^* \approx 0.08$

# Conclusion

| | Random Sampling | Adaptive Selection |
|---|---|---|
| Upper Bound | $O\left(\frac{d \log d}{n}\right)$ | $O\left(\mathrm{err}^* \frac{d}{n}\right)$ |
| Lower Bound | $\Omega\left(\frac{d}{n}\right)$ | $\Omega\left(\mathrm{err}^* \frac{d}{n}\right)$ |

- For practical distribution classes, adaptive selection **cannot achieve** a better rate than $\Theta\left(\frac{1}{n}\right)$.
- For our specific setting, we show that the ratio of the optimal excess error of **random sampling** to the optimal excess error of **adaptive selection** is between $\Omega(1/\mathrm{err}^*)$ and $O((\log \mathrm{d})/\mathrm{err}^*)$**.**
- Perhaps, it is not possible to always achieve improvements with active learning, but instead, the improvement depends on problem dependent quantities.

# Thank you for your attention

Come by poster **#1213** at the poster session later today!


Graduate Research Fellowship Program