Uniform versus uncertainty sampling: When being active is less efficient than staying passive

Alexandru Țifrea*, Jacob Clarysse*, Fanny Yang

Department of Computer Science, ETH Zurich

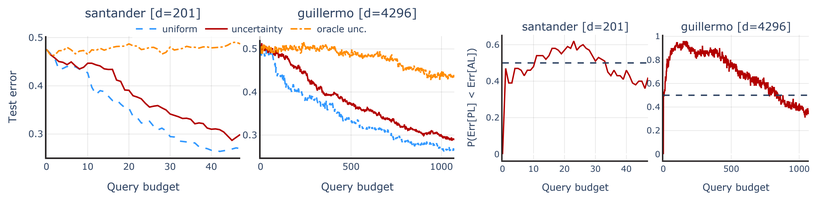
High-dimensional active learning (AL)

Goal: Get same error as passive learning with less labeled data.

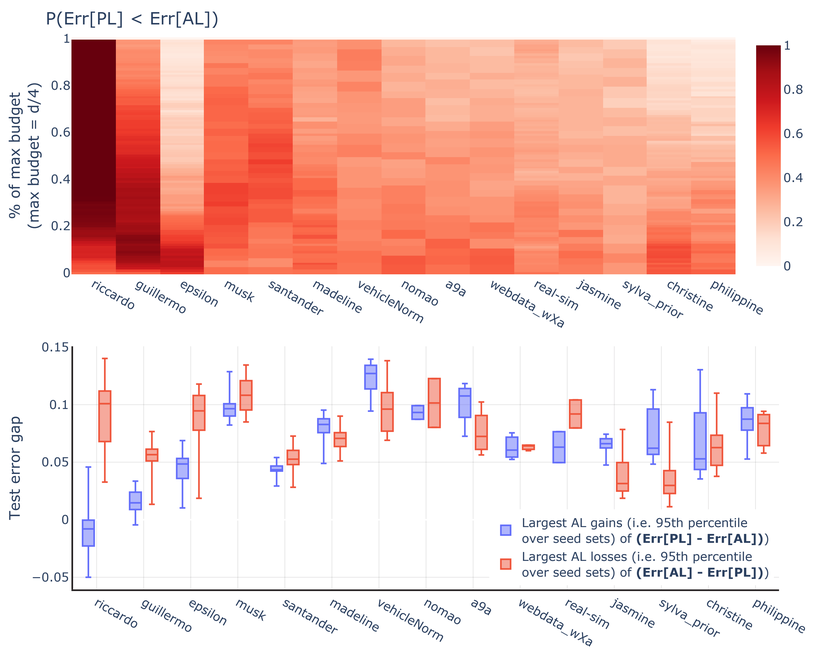
- high dimensions i.e. $n_{\text{labeled}} \ll d \ll n_{\text{unlabeled}}$
- common strategy: **uncertainty sampling for AL (U-AL)**
- U-AL outperforms PL for low-dimensional and noiseless data

Does uncertainty-based active learning outperform passive learning for high-dimensional data?

Logistic regression experiments



Extensive experiments on 15 datasets:



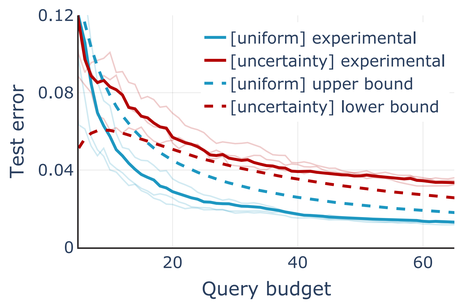
- **PL outperforms U-AL** for many query budgets with $n_{\text{labeled}} < d$.
- The (rare) **gains are lower than the losses** incurred by U-AL.

Theoretical result

Noiseless data: Truncated Gaussian mixture; $d \gg n_{\text{labeled}} > n_{\text{seed}}$

Classifier: Logistic regression; uncertainty given by $|\hat{\theta}^\top x|/\|\hat{\theta}\|_2$

Theorem (informal): Under mild conditions on $n_{\text{unlabeled}}$, μ and σ passive learning leads to lower prediction error than uncertainty sampling, w.h.p. over the draws of data.



- Similar result for **oracle uncertainty sampling** (i.e. using $\theta^\star$)!

Proof intuition and insights

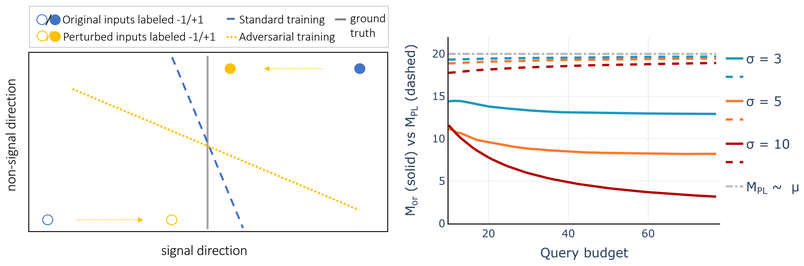
1. Sampling points close to ground truth (GT) ⇒ poor test error

Left: Querying ambiguous points leads to poor $\hat{\theta}$ for $n_{\text{labeled}} \ll d$.

2. U-AL queries points closer to the GT than PL

Right: Distance between GT and the points in the labeled set $\mathcal{D}_\ell$

- $\mathcal{D}_\ell$ is collected with oracle U-AL (M_{or}) or passive learning (M_{PL})



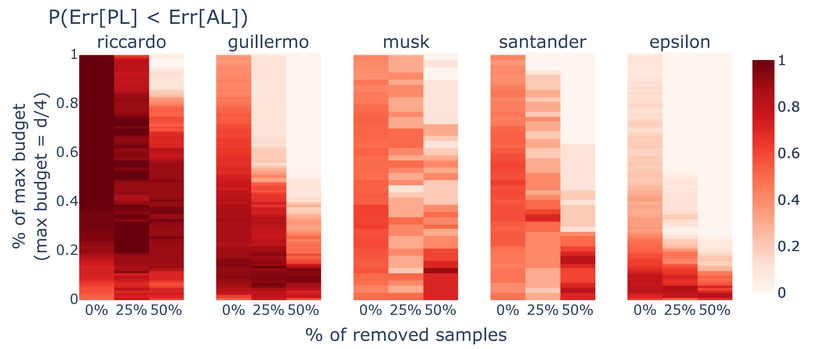
Takeaways: U-AL consistently outperforms PL **if**

1. data is well separated by large margin
2. the initial seed set is large

Experimental validation of insights

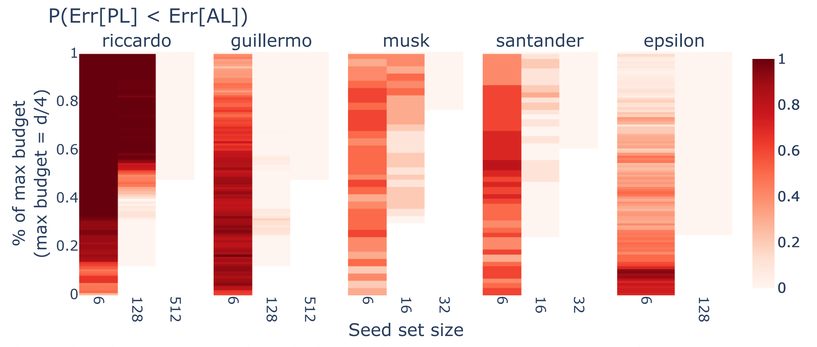
1. **Remove the samples closest to the true decision boundary.**

- Compute oracle by training on the whole dataset.
- The same curated unlabeled set is used for both PL and U-AL.



Conclusion: Large class separation ⇒ U-AL outperforms PL

2. **Increase the size of the initial seed set.**



Conclusion: Large initial seed set ⇒ U-AL outperforms PL

Extension to non-linear models

Same phenomenon for ResNet18 models on image data.