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

Anonymous Author
Anonymous Institution

Abstract

It is widely believed that given the same labeling budget, active learning algorithms like uncertainty sampling achieve better predictive performance than passive learning (i.e. uniform sampling), albeit at a higher computational cost. Recent empirical evidence suggests that this added cost might be in vain, as uncertainty sampling can sometimes perform even worse than passive learning. While existing works offer different explanations in the low-dimensional regime, this paper shows that the underlying mechanism is entirely different in high dimensions: we prove for logistic regression that passive learning outperforms uncertainty sampling even for noiseless data and when using the uncertainty of the Bayes optimal classifier. Insights from our proof indicate that this high-dimensional phenomenon is exacerbated when the separation between the classes is small. We corroborate this intuition with experiments on 20 high-dimensional datasets spanning a diverse range of applications, from finance and histology to chemistry and computer vision.

1 Introduction

In numerous machine learning applications, it is often prohibitively expensive to acquire labeled data, even when unlabeled data is readily available. For instance, consider the task of precise cancer diagnosis (e.g. carcinoma, sarcoma etc). Large amounts of unlabeled data such as EKGs, EEGs and blood tests are available for yet undiagnosed patients under monitoring. However, obtaining labels for this data is expensive and risky: to determine the cancer cell type, the patient needs to undergo surgery for a biopsy. Active learning algorithms aim to reduce labeling costs, by collecting a small labeled set that still results in a model with good predictive performance.

A popular family of active learning algorithms relies on uncertainty sampling (U-AL) for collecting the labeled data (Lewis and Gale, 1994). This paradigm proposes to alternate between (i) training a prediction model (e.g. logistic regression, deep neural network) on the currently available labeled set; and (ii) augmenting the labeled set by acquiring labels for the unlabeled points on which the model has a high uncertainty score. The uncertainty score aims to reflect the spread of the predictive distribution $\mathbb{P}(Y|X = x)$. For example, for binary linear predictors under the logistic noise model, $\mathbb{P}(Y|X = x; \theta)$ is proportional to the distance between x and the decision boundary determined by θ. Alternatively, for deep neural networks, the softmax output of the predicted label or calibrated versions thereof are often used to compute the uncertainty score.

Numerous prior works have documented the advantages of uncertainty sampling in various settings (Tong and Koller, 2001; Settles et al., 2007; Schohn and Cohn, 2000; Raj and Bach, 2021). However, recent theoretical and empirical works on both linear and deep learning models (Schein and Ungar, 2007; Lughofer and Pratama, 2017; Yang and Loog, 2018; Mussmann and Liang, 2018; Hacohen et al., 2022) have reported that oftentimes uncertainty sampling "fails" – that is, it leads to worse predictive error than passive learning (i.e. *uniform sampling*). The existing literature identifies two main causes for the failure of uncertainty sampling. First, in low dimensions when the Bayes optimal model has high error (Mussmann and Liang, 2018), uncertainty sampling provably does not improve upon the sample efficiency of passive learning (PL). Further, several works (Huang et al., 2014; Sener and Savarese, 2018; Hacohen et al., 2022) argue that uncertainty sampling fails due to the *cold start problem*: using only a small labeled set, one cannot obtain a meaningful measure of uncertainty.

Perhaps surprisingly, for high-dimensional problems with a low labeling budget, U-AL underperforms uniform sampling *even* when (i) the Bayes error is zero; and (ii) one uses the uncertainty score corresponding to the Bayes optimal classifier for sampling (referred to as *oracle uncertainty*

Preliminary work. Under review by AISTATS 2023. Do not distribute.

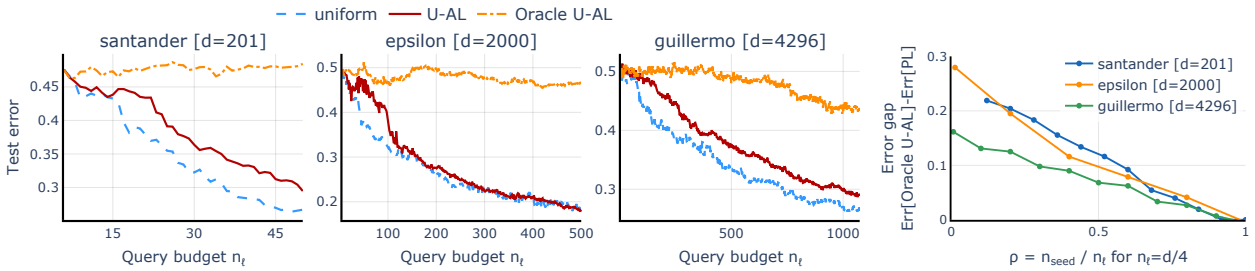


Figure 1: Surprisingly, uncertainty sampling (with or without oracle uncertainty, i.e. *Oracle U-AL* and *U-AL* respectively) leads to worse test error compared to passive learning (*PL*) on a broad range of high-dimensional datasets and for several small query budgets. Moreover, increasing the seed set size n_{seed} leads to a smaller gap between Oracle U-AL and PL. See Appendix E.2 for more datasets.

sampling or *Oracle U-AL*).[1] Our experiments reveal the failure of Oracle U-AL for logistic regression on a wide variety of datasets, a subset of which are presented in Figure 1 (a few works make a similar observation for neural networks (Sener and Savarese, 2018; Hacohen et al., 2022)). As the low-dimensional explanations do not apply to the noiseless scenario, to date, there exists no result that sheds light on this phenomenon.[2] In this paper, we characterize theoretically and empirically, in which settings passive learning outperforms uncertainty sampling for logistic regression. In particular, our contributions are as follows:

1. We observe that uncertainty sampling performs worse than passive learning for logistic regression on numerous high-dimensional datasets from different application domains (e.g. finance, chemistry, histology).
2. We prove non-asymptotic error bounds for logistic regression that directly imply worse performance of uncertainty sampling compared to passive learning – even when using oracle uncertainty (Section 3) and for noiseless data distributions (truncated Gaussian mixture and Gaussian marginal).
3. Distinct from the low-dimensional intuition (Mussmann and Liang, 2018), our proof suggests that in high dimensions uncertainty sampling benefits from a large separation margin between the classes. We confirm this intuition experimentally for logistic regression on 15 real-world datasets (Section 4).

Our results reveal that for high-dimensional data, uncertainty sampling is not only more computationally costly compared to passive learning, but often provably less effective as well. Our paper hence suggests an important avenue for future work: identify active learning algorithms that are provably consistently and substantially outperform passive learning in high-dimensional and low-budget settings.

[1] Oracle uncertainty sampling corresponds to having access to the true predictive posterior distribution $\mathbb{P}(Y|X)$, for both a frequentist and a Bayesian perspective.

[2] The work of Zhang (2018) also focuses on high-dimensional data, but with the purpose of improved computational efficiency.

2 Active learning for classification

We now introduce the active learning framework that we consider throughout most of this paper.

Algorithm 1: Uncertainty sampling

Input: Seed set $\mathcal{D}_{seed}$, unlabeled set $\mathcal{D}_u$, budget n_ℓ, uncertainty function u

Result: Prediction model $f(\cdot;\hat{\theta})$

$\mathcal{D}_\ell \leftarrow \mathcal{D}_{seed}$

$\hat{\theta} \leftarrow \arg\min_\theta \frac{1}{|\mathcal{D}_{seed}|} \sum_{(x,y)\in\mathcal{D}_{seed}} \ell(f(x;\theta),y)$

for $n \in \{n_{seed}+1, ..., n_\ell\}$ **do**

  $x_n \leftarrow \arg\max_{x\in\mathcal{D}_u} u(x;\hat{\theta})$

  $y_n \leftarrow \mathit{AcquireLabel}(x_n)$

  $\mathcal{D}_\ell \leftarrow \mathcal{D}_\ell \cup \{(x_n,y_n)\}; \mathcal{D}_u \leftarrow \mathcal{D}_u \setminus \{x_n\}$

  $\hat{\theta} \leftarrow \arg\min_\theta \frac{1}{|\mathcal{D}_\ell|} \sum_{(x,y)\in\mathcal{D}_\ell} \ell(f(x;\theta),y)$

return $f(\cdot;\hat{\theta})$

Our high-level goal is to train a binary classifier that predicts a label $y \in \{-1,1\}$ from covariates $x \in \mathbb{R}^d$, where $(x,y) \sim \mathbb{P}_{XY}$. More specifically, the task is to find parameters $\theta \in \Theta$ such that the classifier $x \to \text{sgn}(f(x;\theta))$ achieves a low population error $\text{Err}(f(\cdot;\theta)) = \mathbb{E}_{(x,y)\sim\mathbb{P}_{XY}} \mathbb{1}[y \neq \text{sgn}(f(x;\theta))]$. In practice the population error is not available, and hence, one can instead minimize the empirical risk defined by a loss function ℓ on a collection of labeled training points $\hat{\theta} = \arg\min_\theta \frac{1}{|\mathcal{D}_\ell|} \sum_{(x,y)\in\mathcal{D}_\ell} \ell(f(x;\theta),y)$. The goal of active learning is to find a good set $\mathcal{D}_\ell$ which induces a $\hat{\theta}$ that generalizes well.

Collecting the training set via uncertainty-based AL. We consider standard pool-based active learning using an uncertainty score function u like in Algorithm 1 and assume access to a large unlabeled dataset $\mathcal{D}_u$ of size n_u. At first the labeled set $\mathcal{D}_\ell$ consists of a small seed set $\mathcal{D}_{seed}$ containing n_{seed} i.i.d. samples drawn from the training distribution. At each querying step n, we first sample and label the point from the unlabeled data with the highest uncertainty score $u(x;\theta) \in [0,\infty)$[3] and add it to the labeled set. Then we train a classifier on the resulting labeled

[3] In Section E.8 we discuss the implications of our results to strategies that combine an uncertainty and a diversity score (e.g. (Brinker, 2003)).

set.[4] These querying steps are repeated until we exhaust the labeling budget, denoted by n_ℓ (we use labeling or query budget interchangeably). Moreover, we define the seed set proportion $\rho := \frac{n_{\text{seed}}}{n_\ell}$, which effectively captures the fraction of labeled points sampled via uniform sampling. Note that $\rho = 1$ corresponds to passive learning.

Intuition behind uncertainty sampling. Intuitively, in low dimensions (i.e. $d < n_\ell$), uncertainty sampling behaves like binary search (Cohn et al., 1994), and hence, needs significantly fewer samples to find the optimal decision boundary. Following this intuition, it would be natural that using the uncertainty of the Bayes optimal classifier instead, should further improve the sample complexity of active learning. In contrast to this low-dimensional intuition, we show in this work that in fact, both *empirical* and *oracle uncertainty sampling* (using the uncertainty scores of an empirical predictor and of a Bayes optimal classifier, respectively) perform poorly in high-dimensional settings where $d > n_\ell$.

3 Theoretical analysis of uncertainty sampling in high dimensions

In this section we show that, for high-dimensional logistic regression, uncertainty sampling can lead to higher prediction error than passive learning, even for noiseless data. Note that analogous negative results can also be derived for noisy data, as it is an even more challenging setting for uncertainty sampling as we argue in Section 3.2.

Our main theoretical results provide lower bounds on the gap between the population error resulting from uncertainty sampling (U-AL in short) vs. uniform sampling (also called passive learning or PL). We first consider oracle U-AL where we use the oracle uncertainty score determined by the Bayes optimal classifier $\theta^\star$. In what follows, we refer to the max-ℓ_2-margin classifier trained on a labeled dataset acquired (ii) via uniform sampling ($n_{\text{seed}} = n_\ell$) as $\hat{\theta}_{unif}$ and (ii) via oracle and empirical uncertainty sampling respectively as $\hat{\theta}_{oracle}$ and $\hat{\theta}_{uncert}$.

3.1 Uncertainty sampling for logistic regression

We consider linear models of the form $f(x;\theta) = \langle \theta, x \rangle$ with θ in a fixed-norm ball and minimize the logistic loss, i.e. $\ell(z, y) = \log(1 + e^{-zy})$. We note that for linearly separable data, minimizing the logistic loss with gradient descent recovers the max-ℓ_2-margin (interpolating) solution (Soudry et al., 2018; Ji and Telgarsky, 2019). The generalization behavior of this interpolating estimator has been analyzed extensively in recent years in different contexts (Bartlett et al., 2020; Javanmard and Soltanolkotabi, 2020; Muthukumar et al., 2021; Donhauser et al., 2021).

[4]For the theoretical analysis of uncertainty sampling we use the same modification of Chaudhuri et al. (2015); Mussmann and Liang (2018) to slightly change this procedure (see Section 3.5).

Moreover, under the (linear) logistic noise model, for fixed θ, the predictive uncertainty $\mathbb{P}(Y = 1|x, \theta) = \left(1 + e^{-\langle \theta, x \rangle}\right)^{-1}$ is a monotonic function of the inverse distance to the decision boundary. Hence for logistic models, the inverse distance to the decision boundary as an uncertainty score $u(x;\theta) = \left(\frac{|\langle \theta, x \rangle|}{\|\theta\|_2}\right)^{-1}$ directly reflects the predictive uncertainty as desired (Platt, 1999; Mussmann and Liang, 2018; Raj and Bach, 2021).

3.2 Data distribution

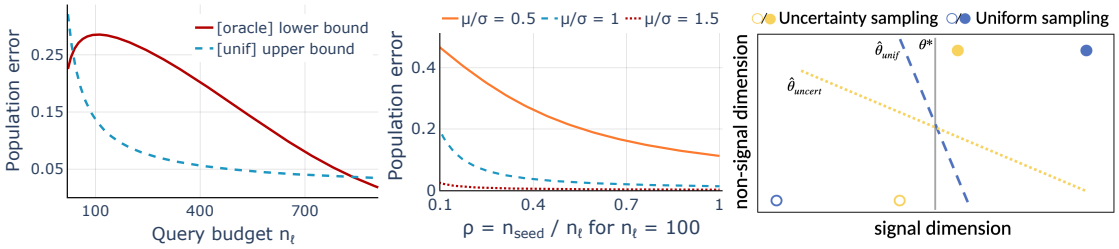
We now introduce the family of joint data distributions $\mathbb{P}$ for our theoretical analysis that includes distributions where the covariates follow a Gaussian or mixture of truncated Gaussian distribution.

Noiseless and balanced observations. Recall that for linear classifiers, the intuitive reason for the effectiveness of uncertainty sampling is that after only a few queries, it selects points in the neighborhood of the optimal decision boundary. Notice that this region is where the label noise is concentrated, for a Gaussian mixture model. Therefore, U-AL is prone to query numerous noisy samples. We wish to show the failure of U-AL even in the most benign setting. Hence, we assume a noiseless binary classification problem where the Bayes error vanishes, i.e. $\text{Err}(\theta^\star) = 0$ for some $\theta^\star$ unique up to scaling and in this section we set $\|\theta^\star\|_2 = 1$. More precisely, we assume that the joint distribution $\mathbb{P}$ is such that the labels $y = \text{sgn}(\langle \theta^\star, x \rangle) \in \{-1, 1\}$ with $\theta^\star = e_1 = [1, 0, ..., 0] \in \mathbb{R}^d$ without loss of generality; if $\theta^\star \neq e_1$ we can rotate and translate the data to get $\theta^\star = e_1$. Further, to disentangle from phenomena stemming from imbalanced data, we consider a distribution where the classes have equal proportions in expectation.

We obtain a family of joint distributions satisfying these conditions by sampling $y \in \{+1, -1\}$ each with probability one half, and then sampling from the class-conditional probability defined by $\mathbb{P}(x_1|y) = \mathcal{N}_{trunc}(x_1; y\mu, \sigma^2, y)$ and $\mathbb{P}(\tilde{x}|y) = \mathcal{N}(\tilde{x}; 0, I_d)$, where $\mathcal{N}_{trunc}(\cdot; y\mu, \sigma^2, y)$ denotes the truncated Gaussian distribution with support $(-\infty, 0)$ if $y = -1$ and, respectively, support $(0, \infty)$ if $y = 1$. The parameters $\mu, \sigma \geq 0$ denote the mean and standard deviation of the non-truncated Gaussian.

Gaussian marginals. Further note that by setting $\mu = 0$, we recover the marginal Gaussian covariate distribution (also known as a discriminative model) – a popular distribution to prove benefits for active learning (Beygelzimer et al., 2010; Hanneke, 2013), as both supervised and semi-supervised learning require large amounts of labeled data to achieve low prediction error, even given infinite unlabeled samples (Scholkopf et al., 2012).

Mixture of truncated Gaussians. For $\mu > 0$, the marginal covariate distribution is a mixture of two truncated Gaussians. Gaussian mixtures have been shown to adequately approximate data generated in many practical



(a) Error vs n_ℓ (fixed n_{seed}). (b) Error vs ρ (fixed n_ℓ). (c) Proof intuition in 2D.

Figure 2: Theoretical population error lower bounds for oracle U-AL and upper bounds for PL in Theorem 3.1 on the truncated Gaussian mixture model. (a) For large d/n_ℓ the lower bound of Theorem 3.1 on the error of oracle U-AL is much larger than the upper bound on the error of PL for fixed $n_{\text{seed}} = 10$ and $\mu/\sigma = 1$. (b) The lower bound on the error is smaller when the seed set proportion ρ and the ratio μ/σ are increasing, for fixed $d = 1000$ and $n_\ell = 100$. (c) Intuition for the failure of U-AL in high dimensions. The classifier assigns higher weight to the non-signal dimension when trained on the yellow points close to the optimal decision boundary.

applications (Bouguila and Fan, 2019). Since we wish to analyze the most benign setting for U-AL, we consider a mixture of truncated Gaussians. Each truncated component has standard deviation $\sigma_{tr} \leq \sigma$ and mean

$$\mu_{tr} := \mathbb{E}[yx_1] = \mu + \sigma \frac{\phi(-\mu/\sigma)}{1 - \Phi(-\mu/\sigma)}, \tag{1}$$

where ϕ and Φ denote the probability density and the cumulative distribution functions of the standard normal distribution, respectively.

3.3 Main result for oracle uncertainty sampling

In this section we present a theorem that rigorously proves how logistic regression with oracle uncertainty sampling leads to worse classifiers in the high-dimensional setting (i.e. $n_\ell \ll d$) where most samples are acquired with U-AL (i.e. $\rho \ll 1$) — a phenomenon observed on real-world data in Figure 1. Moreover, we discuss an insight that directly follows from the proof intuition: in problems where many samples in the unlabeled dataset are close to the optimal decision boundary (i.e. small μ/σ ratio), the error gap between oracle U-AL and PL increases. We confirm this intuition on real-world data in Section 4.5. We state the informal theorem here and refer to Appendix A for the formal statement and proof.

Theorem 3.1 (informal)**.** *Let $n_\ell \ll n_u$ and let $\mathcal{D}_u$ and $\mathcal{D}_{seed}$ be datasets drawn i.i.d. from the joint distribution described in Section 3.2 with $\mu/\sigma < 2$. Then, in the high-dimensional regime when $d \gg n_\ell$, there exist universal constants $0 < c_1, \epsilon \ll 1$ and $t, c_2 > 0$ such that with probability larger than $1 - e^{-c_2 t^2/2}$ it holds that:*

$$Err(\hat{\theta}_{oracle}) - Err(\hat{\theta}_{unif}) > \Psi_{\mu,\sigma}(\alpha_{oracle}) - \Psi_{\mu,\sigma}(\alpha_{unif}),$$

where $\Psi_{\mu,\sigma}$ is a strictly increasing function, defined in Section A.1, and

$$\alpha_{oracle} = \frac{(1-\epsilon)\sqrt{d/n_\ell}}{\rho\left(\mu_{tr} + t\sigma_{tr}/\sqrt{\rho n_\ell}\right) + c_1(1-\rho)\mu_{tr}},$$

$$\alpha_{unif} = \frac{(1+\epsilon)\sqrt{d/n_\ell}}{\mu_{tr} - t\sigma_{tr}/\sqrt{n_\ell}}.$$

In particular, if $\rho < \delta$ for $0 < \delta < 1/2$, then

$$Err(\hat{\theta}_{oracle}) - Err(\hat{\theta}_{unif}) > 0.$$

Note that $\Psi_{\mu,\sigma}$ is similar to the cumulative distribution function of the Gaussian distribution (see Appendix A.1 for the exact definition and more details). We now discuss the assumptions of Theorem 3.1. Observe that if the ratio μ/σ is large, then PL can also be used to achieve low error. Hence, we assume settings where $\mu/\sigma < 2$, which also cover real-world datasets that contain ambiguous samples that lie near the optimal decision boundary. Finally, a small seed set proportion ρ allows for sufficiently many active queries, and is common in situations that employ AL.

Proof intuition. Figure 2c illustrates the intuition behind the failure of U-AL in high dimensions using a 2D cartoon. We depict the samples chosen by oracle uncertainty sampling (yellow), which lie close to the optimal decision boundary (vertical line) and the points selected by uniform sampling (blue), which are farther away from the optimal decision boundary. Note that for both sampling strategies, the selected samples are far apart in the non-signal direction. More specifically, in high dimensions the large distance in the non-signal components $\tilde{x}$ is a consequence of sampling $\tilde{x}$ from a multivariate Gaussian. It follows from these facts that the max-ℓ_2-margin classifier trained on the samples near the optimal decision boundary (yellow dotted) is more tilted (and hence, has larger population error) than the one trained on uniformly sampled points that are further away (blue dashed).

The assumption $\mu/\sigma < 2$ implies that the covariate distribution has sufficient density in the neighborhood of the optimal decision boundary. Hence, with high probability, the labeled data acquired with oracle U-AL lies in this region, leading to an estimator with high population error.

3.4 Interpretation of Theorem 3.1

Theorem 3.1 characterizes when the population error gap between oracle U-AL and PL is positive: for small labeling budgets (leading to a large d/n_ℓ ratio), for a small seed set proportion ($\rho \ll 1$) and for sufficiently many unlabeled samples near the Bayes optimal decision boundary (implied by $\mu/\sigma < 2$). In particular, in the regime $d/n_\ell \to \infty$ that implies $\epsilon \to 0$ (as argued in Appendix A.2), we can observe how for small $\rho \approx 0$, it holds that $\alpha_{\text{oracle}} \gg \alpha_{\text{unif}}$ as $0 < c_1 \ll 1$. In turn, $\alpha_{\text{oracle}} \gg \alpha_{\text{unif}}$ implies

$\text{Err}(\hat{\theta}_{oracle}) - \text{Err}(\hat{\theta}_{unif}) \gg 0$ since $\Psi_{\mu,\sigma}$ is strictly increasing. In Figure 2 we depict how the error bounds, and hence the gap in Theorem 3.1, depend on the three quantities $\rho, d/n_\ell$ and μ/σ.

In Figure 2a, we show the dependence of the bound in Theorem 3.1 on n_ℓ (and hence, d/n_ℓ), for fixed $n_{\text{seed}} = 10, d = 1000$ and $\mu/\sigma = 1$. If the query budget n_ℓ and the ratio ρ are small (the middle region of the plot on the horizontal axis), we have a large error gap between oracle U-AL and PL. This phenomenon is inherently high-dimensional and stops occurring for large sample sizes n_ℓ (the right part of the figure). We also identify these regimes in experiments on real-world data in Section 4.4. Note that the bounds are loose for extremely small budgets (left part of the figure).

In Figure 2b, we vary the seed set size n_{seed} (and hence, the ratio ρ), for fixed $n_\ell = 100, d = 1000$. We observe that increasing the seed set proportion ρ reduces the error of oracle U-AL (note that $\rho = 1$ corresponds to PL). We highlight that the dependence of the error on the seed set proportion ρ is not due to the uncertainty measure becoming more meaningful for larger ρ, as conjectured by some prior works (Huang et al., 2014; Sener and Savarese, 2018): in our case, the uncertainty score is defined using the Bayes optimal classifier $\theta^\star$ at every querying step. Instead, Theorem 3.1 captures another failure case of uncertainty sampling, specific to high-dimensional settings.

Moreover, Figure 2b also illustrates the dependence of the error lower bound on the ratio μ/σ, for oracle U-AL. In Theorem 3.1, the distribution-dependent ratio μ/σ enters the bounds via the quantity c_1 which is strictly increasing in μ/σ (see Lemma A.9 in Appendix A for an upper bound on c_1 as a function of μ/σ). For small μ/σ, the population error gap between oracle U-AL (i.e. $\rho < 1$) and PL (i.e. $\rho = 1$) is large. In Appendix C, we show that the theoretical lines also closely predict the values from simulations.

Finally, this phenomenon is caused by choosing to label samples close to the Bayes optimal decision boundary, which are, by definition, the points queried with oracle U-AL. This suggests that, perhaps surprisingly, oracle U-AL exacerbates this high-dimensional phenomenon. Indeed, as evident from comparing the bounds in Theorems 3.1 and 3.2, oracle U-AL performs even worse than U-AL that uses the uncertainty of an empirical estimator $\hat{\theta}$. We confirm this intuition on real-world datasets in Appendix E.3.

3.5 Main result for empirical uncertainty sampling

For completeness, we prove the counterpart of Theorem 3.1 for uncertainty sampling that uses the empirical estimator $\hat{\theta}$ instead of the optimal classifier $\theta^\star$. We first introduce a modification of the uncertainty sampling algorithm and then state the informal theorem.

Two-stage uncertainty sampling. For the theoretical analysis of empirical uncertainty sampling we modify Algorithm 1 slightly, and consider instead a two-stage procedure similar to (Mussmann and Liang, 2018; Chaudhuri et al., 2015): 1) we obtain $\hat{\theta}_{seed}$ using the initial small seed set; and 2) we use $\hat{\theta}_{seed}$ to select a batch of $(1-\rho)n_\ell$ samples to query from the unlabeled set. This two-stage process allows us to simplify the analysis significantly by getting rid of dependence of the classifier at stage n on the unlabeled dataset. As argued in Chaudhuri et al. (2015), it essentially suffices to analyze two-stage uncertainty sampling to draw conclusions about the iterative algorithm.

We now state the main theorem informally for empirical uncertainty sampling and defer the formal statement and proof to Appendix A.

Theorem 3.2 (informal)**.** *Let $n_\ell \ll n_u$ and let $\mathcal{D}_u$ and $\mathcal{D}_{seed}$ be datasets drawn i.i.d. from the joint distribution described in Section 3.2 with $\mu/\sigma < 2$ and $\sigma > 1$. Then, in the high-dimensional regime when $d \ll n_u$, there exist universal constants $0 < c_1, \epsilon \ll 1$ and $t, c_2 > 0$ such that with probability larger than $1 - e^{-c_2 t^2/2}$ it holds that:*

$$\textit{Err}(\hat{\theta}_{uncert}) - \textit{Err}(\hat{\theta}_{unif}) > \Psi_{\mu,\sigma}\left(\alpha_{\textit{uncert}}\right) - \Psi_{\mu,\sigma}\left(\alpha_{\textit{unif}}\right),$$

where $\Psi_{\mu,\sigma}, \alpha_{\textit{unif}}$ as in Theorem 3.1 and

$$\alpha_{\textit{uncert}} = \frac{(1-\epsilon)\sqrt{\frac{d}{n_\ell}}}{\rho C_{seed} + (1-\rho)\left(c_1\mu_{tr} + \sqrt{\frac{\log n_u}{C_{seed}}}\sqrt[4]{\frac{(1+\epsilon)d}{\rho n_\ell}}\right)},$$

with $|\mu_{tr} - C_{seed}| < t\sigma_{tr}(\rho n_\ell)^{-1/2}$. In particular, for small fixed constants $c_3, c_4, c_5 > 0$ if $\rho n_\ell < c_3$, and

$$\mu_{tr} > c_4\left(\frac{d}{\rho n_\ell}\right)^{1/6}(\log n_u)^{1/3},$$
$$\rho < c_5 \frac{C_{seed}}{\mu_{tr} - \frac{t\sigma_{tr}}{(\rho n_\ell)^{1/2}} - \left(\frac{d}{\rho n_\ell}\right)^{1/6}(\log n_u)^{1/3}},$$

then $\textit{Err}(\hat{\theta}_{uncert}) - \textit{Err}(\hat{\theta}_{unif}) > 0$.

The proof shares the same intuition and key steps as the proof of Theorem 3.1 and differs only in certain technicalities that arise from estimating the classifier $\hat{\theta}_{seed}$. In particular, we additionally need to assume that the mean μ_{tr} is large enough such that the classifier trained on the seed set has non-trivial prediction error. Note that it is possible to obtain a large μ_{tr} even for the marginal Gaussian case when $\mu = 0$, by choosing a large σ in Equation 1.

4 Experiments

In this section, we provide extensive experiments to investigate the ineffectiveness of low-budget uncertainty sampling on high-dimensional real-world data. In particular, we train logistic regression with oracle and empirical uncertainty sampling on a wide variety of tabular datasets.

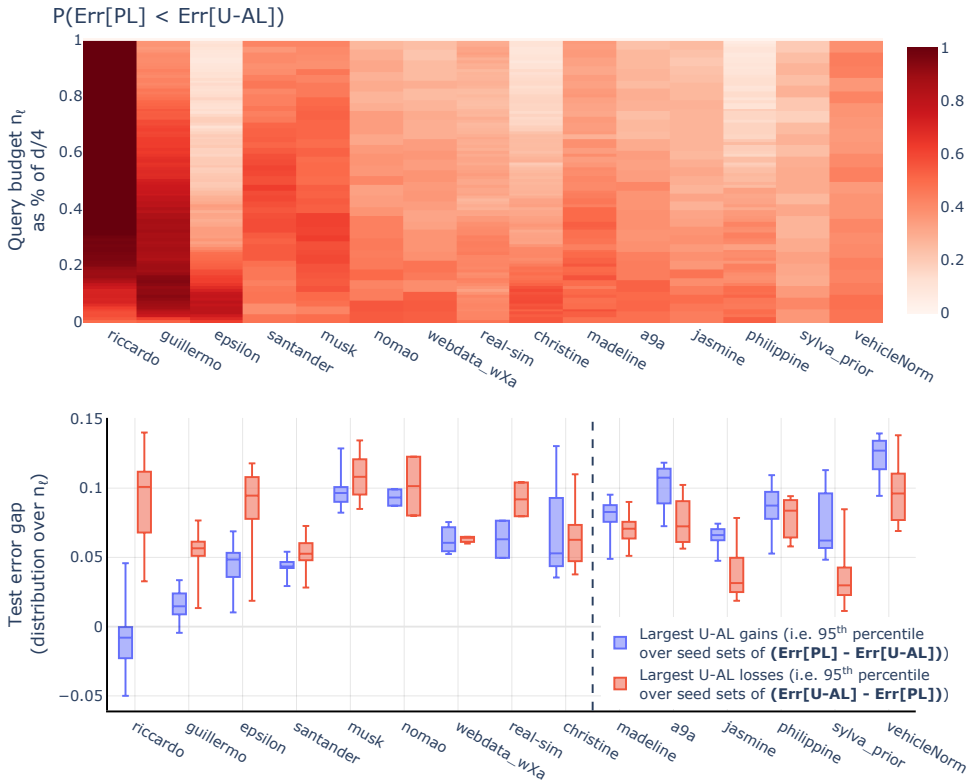


Figure 3: **Top:** The probability that the test error is lower with PL than with U-AL, over 100 draws of the seed set. PL outperforms U-AL, for a significant fraction of $n_\ell \in [n_{\text{seed}}, d/4]$ and for all datasets (i.e. dark red regions). See Appendix E.4 for more precise numerical values. **Bottom:** Largest gains and losses in test error of U-AL compared to PL, over 100 draws of $\mathcal{D}_{seed}$. Box plots show distribution over $n_\ell \in [n_{\text{seed}}, n_{\text{transition}}]$. The sporadic gains of U-AL over PL are generally lower (to left of dashed line) or similar to the losses in test error that it can incur.

4.1 Datasets

We select binary classification datasets from OpenML Vanschoren et al. (2013) and from the UCI data repository Dua and Graff (2017) according to a number of criteria: i) the data should be high-dimensional ($d > 100$) with enough samples that can serve as the unlabeled set ($n_u > \max(1000, 2d)$); ii) linear classifiers trained on the entire data should have high accuracy (which excludes most image or text datasets). Moreover, we discard datasets that have missing values. A total number of 15 datasets satisfy these criteria and cover a broad range of applications from finance and ecology to chemistry and histology. We provide more details about the datasets in Appendix D.1.

Like in Section 3, we wish to isolate the effect of high-dimensionality, and hence, we balance the two classes by subsampling the majority class uniformly at random. In addition, we mimic the noiseless setting considered in Section 3 using the following procedure: after fitting a linear classifier on the entire dataset, we remove the training samples that are not correctly predicted and use the subsequent smaller subset as the new dataset. We show that, even in the more favorable noiseless setting, the performance of U-AL suffers in high dimensions compared to PL. Experiments on the original, uncurated datasets presented in Appendix E.1 reveal a similar trend.

4.2 Methodology

We split each dataset into a test set and a training set. In all experiments, the training set is then randomly divided into a labeled seed set $\mathcal{D}_{seed}$ of fixed size $n_{\text{seed}} = 6$ (see Appendix E.7 for experiments with larger seed sets) and the covariates of the remaining training samples constitute the unlabeled set $\mathcal{D}_u$.

In practice, one seeks to find the sampling strategy that performs best for a fixed seed set $\mathcal{D}_{seed}$ and labeling budget n_ℓ. To provide an extensive experimental analysis, in this work we compare U-AL (Algorithm 1) and PL over a large number of configurations of $(\mathcal{D}_{seed}, n_\ell)$. We repeatedly draw different seed sets uniformly at random (10 or 100 draws, depending on the experiment) and consider all integer values in $[n_{\text{seed}}, d/4]$ as the labeling budget n_ℓ, where d is the ambient dimension of the dataset.[5]

At each querying step, we use L-BFGS (Liu and Nocedal, 1989) to train a linear classifier by minimizing the average logistic loss on the labeled dataset collected until then. Appendix E.6 shows the same high-dimensional phenomenon for ℓ_1- or ℓ_2-regularized classifiers.

[5]The value $d/4$ is chosen only for illustration purposes. Since for the *real-sim* dataset $d > 20,000$, we set the maximum labeling budget to $n_\ell = 3,000 < d/4$ for computational reasons.

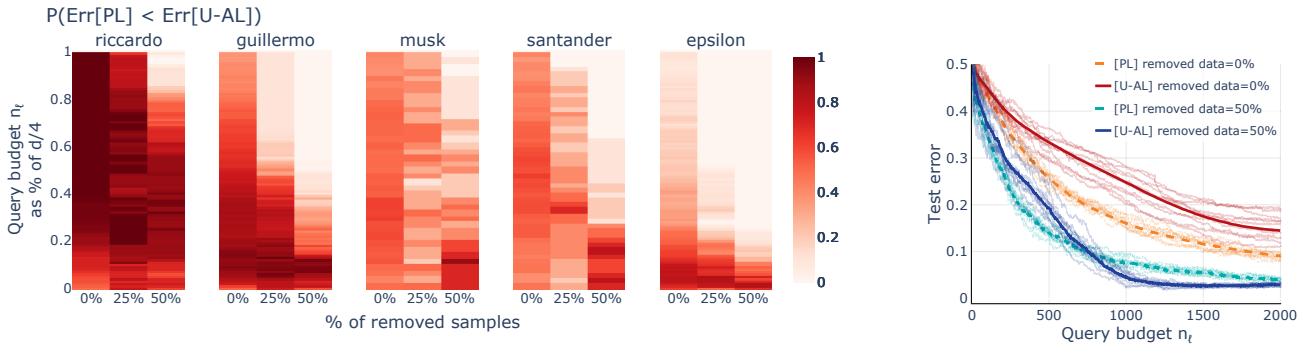


Figure 4: Increasing the separation between the classes in the unlabeled dataset improves the performance of U-AL. Removing the 25% or 50% closest points to the Bayes optimal decision boundary improves U-AL (**left**) which now outperforms PL for many query budgets (i.e. lighter colors), and even on challenging datasets like *riccardo* (**right**).

4.3 Evaluation metrics

We compare uncertainty-based active learning and passive learning with respect to two performance indicators. On the one hand, we measure the probability (over the draws of the seed set) that passive learning leads to a smaller test error than uncertainty sampling. We compute this probability for each labeling budget $n_\ell \in [n_{\text{seed}}, d/4]$. On the other hand, we wish to quantify the magnitude of the failure of U-AL. We compare the most significant gains of U-AL with its most significant losses across small query budgets for which PL outperforms U-AL with high probability. In particular, we focus on query budgets smaller than the dataset-dependent transition point $n_{\text{transition}}$, where $n_{\text{transition}}$ is the largest query budget $n_\ell \in [n_{\text{seed}}, d/4]$ for which the probability of PL outperforming U-AL exceeds 50%. If no such budget exists, $n_{\text{transition}} = d/4$. For each query budget size $n_\ell \in \{n_{\text{seed}}, ..., n_{\text{transition}}\}$, we then compute the gap between the test error obtained with uncertainty sampling and with passive learning, over 100 draws of the seed set. For every labeling budget n_ℓ we report the largest loss of U-AL (i.e. 95^{th} percentiles over the draws of $\mathcal{D}_{seed}$ of the error gap $\text{Err}(\hat{\theta}_{uncert}) - \text{Err}(\hat{\theta}_{unif})$) and the largest gain of U-AL (i.e. 95^{th} percentiles of the error gap $\text{Err}(\hat{\theta}_{unif}) - \text{Err}(\hat{\theta}_{uncert})$). Then, we depict the distribution of these values of extreme losses/gains over $n_\ell \in \{n_{\text{seed}}, ..., n_{\text{transition}}\}$. In Appendix E.2 and E.5, we present more evaluation metrics (e.g. the dependence of the test error on the budget n_ℓ), which provide further evidence that uncertainty sampling fails to be effective in high dimensions.

4.4 Main results

In Figure 3-Top we show the probability (over 100 draws of the seed set) that PL leads to lower test error than U-AL. The observations match the trend predicted by our theoretical results (see Theorem A.4): decreasing the seed set ratio ρ (here, by increasing the budget n_ℓ along the y-axis) leads to a higher probability that PL outperforms U-AL. Analogous to the discussion in Section 3.4, we observe two regimes: for small query budgets n_ℓ, U-AL performs poorly with probability larger than 50% (dark regions). This regime spans a broad range of budgets $n_\ell \in [n_{\text{seed}}, d/4]$ for most datasets. In the second regime, U-AL eventually outperforms PL for large query budgets.[6]

Furthermore, Figure 3-Bottom shows that in 9 out of 15 datasets, the median (over budgets $n_\ell \in [n_{\text{seed}}, n_{\text{transition}}]$) of the largest gain of U-AL is lower than the median loss it can incur, compared to uniform sampling. Intuitively, this indicates that even in the unlikely event that uncertainty sampling leads to better accuracy, the largest gains we can achieve are lower than the potential losses. While the severity of the phenomenon varies with the dataset, we can conclude after this extensive study that uncertainty sampling cannot be used reliably when the dimension of the data exceeds the size of the query budget.

Finally, recall that in Section 3 and in Figure 1 we show theoretically and empirically that using the uncertainty score of the Bayes optimal classifier exacerbates the failure of U-AL in high dimensions. In fact, oracle U-AL performs consistently much worse than both PL and vanilla U-AL in experiments, as indicated in Appendix E.3.

4.5 Verifying trends predicted by the theory for U-AL

The intuition developed in Section 3 suggests that the performance of uncertainty sampling in high dimensions may improve for: i) a larger separation margin between the classes (modeled by μ/σ in Theorem 3.1); or ii) a larger seed set size (leading to a larger ratio $\rho = n_{\text{seed}}/n_\ell$ in Theorem 3.1). We test whether these insights also underlie the phenomenon observed in real-world datasets.

First, we investigate the role of the separation margin. We train a linear classifier on the full labeled dataset. Then, we artificially increase the distance between the classes by removing the 25% or 50% closest samples to the decision boundary determined by the classifier trained on the entire dataset. Indeed, we confirm that removing the 25% or 50% most "difficult" samples improves the performance of uncertainty sampling significantly as we show in Figure 4. In

[6]The *riccardo* dataset is particularly challenging for uncertainty sampling and needs more than $d/4$ labeled samples before closing the gap between uncertainty and uniform sampling.

particular, after removing the points close to the Bayes optimal decision boundary, uncertainty sampling outperforms uniform sampling even on *riccardo*, the most challenging dataset in our benchmark for active learning.

Further, we analyze the impact of a larger seed set size on the performance of uncertainty sampling. For a fixed labeling budget n_ℓ, increasing n_{seed} corresponds to a larger ratio ρ, which leads to a smaller gap in error between U-AL and PL as we explain in Section 3: Indeed, we observe that larger seed set sizes can lead to more effective uncertainty sampling both on synthetic experiments in Figure 8 and on real-world data in experiments presented in Appendix E.7.

4.6 Other AL methods in high dimensions and potential mitigations

AL strategies effectively equivalent to U-AL. Finally, we note that the same failure case occurs for other algorithms, such as margin-based active learning (Scheffer and Wrobel, 2001; Ducoffe and Precioso, 2018; Mayer and Timofte, 2020) or entropy sampling (Settles, 2009), which effectively aim to sample the same points as U-AL.

Combining uncertainty sampling and representativeness. Recall that, by definition, varying the ratio ρ modulates the fraction of the labeling budget selected with uncertainty sampling, with $\rho = 1$ corresponding to PL. Another way to interpolate between uncertainty and uniform sampling is via a strategy that combines informativeness (via uncertainty sampling) and representativeness (via uniform sampling), as proposed by Brinker (2003); Huang et al. (2014); Yang et al. (2015); Gal et al. (2017); Shui et al. (2020); Farquhar et al. (2021) We analyze an ϵ-greedy scheme that also falls in this family of AL algorithms: at each querying step, we sample using uncertainty with probability $1-\epsilon$ and perform uniform sampling with probability ϵ. In Appendix E.8 we provide evidence that PL continues to surpass AL for a large fraction of the labeling budgets, even when using the ϵ-greedy strategy. The intuition for the failure of these algorithms in high dimensions is the same as the one presented in Section 3.

AL with no uncertainty sampling. Could it be that AL algorithms not relying on any form of uncertainty score, such as Sener and Savarese (2018); Gissin and Shalev-Shwartz (2019); Hacohen et al. (2022), mitigate this high-dimensional phenomenon? Indeed, we show experimentally in Figure 21 that coreset-based AL (Sener and Savarese, 2018) outperforms U-AL with high probability in real-world applications. However, compared to *PL*, the coreset method is still often worse (Figure 22), that is, the high-dimensional phenomenon outlined in Section 3 still persists to a large extent. In particular, no mechanism prevents the coreset strategy from selecting points close to the Bayes optimal decision boundary. As highlighted in Figure 4, selecting these "difficult" samples can hurt the performance of active learning.

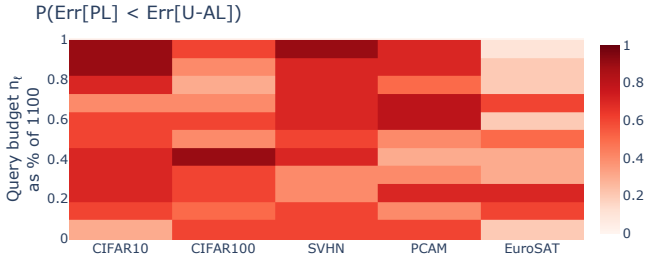


Figure 5: The probability that the test error is lower with PL than with U-AL, when using ResNet18 models on 5 image classification tasks. The probability is computed over 10 runs with different random seeds. PL outperforms U-AL, for a significant fraction of the query budgets and for all datasets (i.e. dark red regions). See Appendix F for detailed experimental details.

Discussion on mitigations. Figure 4 suggests that uncertainty sampling constrained to points far enough from the Bayes optimal decision boundary might outperform passive learning in high dimensions. As the Bayes optimal predictor is not available during training, the closest derived mitigation strategy would be to not allow selecting the points closest to the decision boundary determined by the empirical estimator $\hat{\theta}$ (instead of the optimal $\theta^\star$). We find that this mitigation strategy does not help to alleviate the negative effect of uncertainty sampling, as it does not effectively remove all the difficult points from the set of query candidates. We regard it as important future work to investigate whether an active learning strategy can be provably effective in high-dimensional settings similar to ours.

5 Discussion and future work

In this work we show theoretically and through extensive experiments that active learning, and in particular uncertainty sampling, performs worse than uniform sampling for *linear models* in high dimensions. While we focus on logistic regression and the max-ℓ_2-margin solution, we conjecture that the same intuition outlined in Section 3 holds for other linear predictors like lasso- or ridge-regularized estimators, as indicated by experiments in Appendix E.6.

Moreover, this phenomenon is more general and also occurs for complex non-linear models like deep neural networks. Our experiments suggest that U-AL performs poorly in the context of deep learning on a number of different image classification tasks (see Figure 5 and Appendix F for details and more experiments). An exciting avenue for future work is to investigate whether the insights about linear models revealed by our theoretical analysis transfer to non-linear predictors like deep neural networks.

Furthermore, for practical purposes, an important question for future work is whether it is possible to construct a strategy that improves upon uniform sampling in high dimensions. Based on Figure 4, we believe that imposing certain assumptions on the distributions could allow for improvements via active learning.

References

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, pages 30063–30070, 2020.

A. Beygelzimer, D. J. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems*, 2010.

N. Bouguila and W. Fan. *Mixture Models and Applications*. Springer, 2019.

K. Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

K. Chaudhuri, S. M. Kakade, P. Netrapalli, and S. Sanghavi. Convergence rates of active learning for maximum likelihood estimation. *Proceedings in Advances in Neural Information Processing Systems 28*, 2015.

D. Cohn, R. Ladner, and A. Waibel. Improving generalization with active learning. In *Machine Learning*, 1994.

K. Donhauser, A. Tifrea, M. Aerni, R. Heckel, and F. Yang. Interpolation can hurt robust generalization even when there is no noise. In *Proceedings in Advances in Neural Information Processing Systems 34*, 2021.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

M. Ducoffe and F. Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

S. Ertekin, J. Huang, L. Bottou, and L. Giles. Learning on the border: Active learning in imbalanced data classification. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 2007.

S. Farquhar, Y. Gal, and T. Rainforth. On statistical bias in active learning: How and when to fix it. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Y. Gal, R. Islam, and Z. Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

D. Gissin and S. Shalev-Shwartz. Discriminative active learning. *arXiv preprint arXiv:1907.06347*, 2019.

G. Hacohen, A. Dekel, and D. Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022.

S. Hanneke. *A Statistical Theory of Active Learning*. 2013.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

P. Helber, B. Bischke, A. Dengel, and D. Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *arXiv preprint arXiv:1709.00029*, 2017.

S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. *Proceedings in Advances in Neural Information Processing Systems 27*, 2014.

A. Javanmard and M. Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *arXiv preprint arXiv:2010.11213*, 2020.

Z. Ji and M. Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Conference on Learning Theory (COLT)*, 2019.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the international ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Programming*, 45(3, (Ser. B)), 1989.

E. Lughofer and M. Pratama. Online active learning in data stream regression using uncertainty sampling based on evolving generalized fuzzy models. *IEEE Trans. on fuzzy systems*, 2017.

C. Mayer and R. Timofte. Adversarial sampling for active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

S. Mussmann and P. Liang. On the relationship between data efficiency and error for uncertainty sampling. In *Proceedings of the 34th International Conference on Machine Learning*, 2018.

V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *The Journal of Machine Learning Research*, pages 10104–10172, 2021.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 1999.

A. Raj and F. Bach. Convergence of uncertainty sampling for active learning. *arXiv preprint arXiv:2110.15784*, 2021.

T. Scheffer and S. Wrobel. Active learning of partially Hidden Markov Models. In *In Proceedings of the ECML/PKDD Workshop on Instance Selection*, 2001.

A. I. Schein and L. H. Ungar. Active learning for logistic regression: An evaluation. *Machine Learning*, 2007.

G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

B. Scholkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

B. Settles. Active learning literature survey. 2009.

B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. *Proceedings in Advances in Neural Information Processing Systems 7*, 2007.

C. Shui, F. Zhou, C. Gagné, and B. Wang. Deep active learning: Unified and principled method for query and training. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 2018.

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2001.

J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. OpenML: networked science in machine learning. *SIGKDD Explorations*, 2013.

B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant CNNs for digital pathology. *arXiv preprint arXiv:1806.03962*, 2018.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Y. Yang and M. Loog. A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 2018.

Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 2015.

C. Zhang. Efficient active learning of sparse halfspaces. In *Proceedings of the 31st Conference On Learning Theory*, pages 1856–1880, 2018.

A Formal statements and proofs of the main results

In this section we state and give the proofs of the main results. We first discuss some preliminaries regarding the mixture of truncated Gaussians distribution after which we state the formal results and the proofs.

A.1 Properties of a mixture of truncated Gaussians

Recall that we consider a mixture of two Gaussians, with means $\mu \in \{-\mu, \mu\}$ and standard deviation $\sigma > 0$, each corresponding to one of two classes $y \in \{-1, 1\}$. As detailed in Section 3.2 we truncate both Gaussians to ensure that the data is noiseless. We denote the truncated Gaussian mixture distribution by $\mathbb{P}_{\text{TGMM}}(\mu, \sigma)$, where $\mu, \sigma \geq 0$ are the absolute value of the mean and the standard deviation of the Gaussians before truncation. To state the formal theorems, we now discuss some properties of the distribution $\mathbb{P}_{\text{TGMM}}(\mu, \sigma)$ that will also be used throughout the proofs in this section.

Mean and standard deviation of a truncated Gaussian. We state the known formulas for the mean and standard deviation of a truncated Gaussian random variable. A positive one-sided truncated Gaussian distribution is defined as follows: we restrict the support of a normal random variable with parameters (μ, σ) to the interval $(0, \infty)$. Clearly the mean of the truncated Gaussian is slightly larger than μ and the standard deviation slightly smaller than σ. Let ϕ be the probability density function of the standard normal distribution and denote by Φ the cumulative distribution function. Then we find that the mean of a positive one-sided truncated Gaussian distribution is given by

$$\mu_{tr} = \mu + \frac{\sigma\phi(-\mu/\sigma)}{1 - \Phi(-\mu/\sigma)} \tag{2}$$

and the standard deviation is given by

$$\sigma_{tr} = \sigma\left(1 - \frac{\mu}{\sigma} \cdot \frac{\phi(-\mu/\sigma)}{1 - \Phi(-\mu/\sigma)} - \left(\frac{\phi(-\mu/\sigma}{1 - \Phi(-\mu/\sigma)}\right)^2\right) \tag{3}$$

Error of a linear classifier. We now derive the closed form of the population error of a linear classifier evaluated on data drawn from $\mathbb{P}_{\text{TGMM}}(\mu, \sigma)$. Consider without loss of generality a classifier induced by a vector $\theta \in \mathbb{R}^d$, with $\theta = [1, \alpha\tilde{\theta}]$, where $\|\tilde{\theta}\|_2 = 1$, as introduced in Section 2. Note that the last $d - 1$ coordinates of any sample drawn from the truncated Gaussian distribution are distributed like a standard multivariate normal. By definition, the error of a classifier $\theta = [1, \alpha\tilde{\theta}]$ is given by

$$\text{Err}_{0-1}(\theta) = P\left[y\sum_{i=1}^{d}\theta_i x_i < 0\right] = P\left[yx_1 + y\alpha\sum_{i=2}^{d}\tilde{\theta}_i x_i < 0\right].$$

Because the sum of Gaussian random variables is again a Gaussian random variable, we find that $y\sum_{i=2}^{d}\theta_i x_i$ is Gaussian distributed with a mean of zero and a standard deviation of α. Using the known probability density function of a truncated Gaussian and as all coordinates are independent, we find that the test error is given by

$$\text{Err}_{0-1}(\theta) = \frac{1}{2\pi\alpha\sigma(1 - \Phi(-\mu/\sigma))}\int_0^\infty\int_t^\infty e^{-\frac{(t-\mu)^2}{2\sigma^2}}e^{-\frac{l^2}{2\alpha^2}}dldt =: \Psi_{\mu,\sigma}(\alpha) \tag{4}$$

Note that the expression in Equation 4 can easily be approximated numerically. Moreover, note that $\Psi_{\mu,\sigma}$ is similar to the cumulative density function of a standard normal random variable: for a fixed t in Equation 4, we exactly find the cumulative density function of a Gaussian distributed random variable. Then, we integrate over all possible values for t. See Figure 6b for a visualization of $\Psi_{\mu,\sigma}$. For convenience we state two properties of the error function $\Psi_{\mu,\sigma}$:

1. $\Psi_{\mu,\sigma}$ is a monotonically increasing function of α.
2. $\Psi_{\mu,\sigma}$ is monotonically decreasing in μ.

Therefore, for fixed distributional parameters μ and σ, we have that α fully characterizes the error of the classifier. Hence, proving a gap between the values of α obtained with uncertainty sampling and with uniform sampling is sufficient to show a gap in the population error.

A.2 Formal statement of Theorem 3.1

In this section, we state the formal version of Theorem 3.1. The proof of the theorem can be found in Section A.4. We start with formally introducing the setting and assumptions. Thereafter, we state the theorem that compares oracle uncertainty sampling with passive learning.

Setting. We look at the same setting as in Section 3, namely pool based active learning with uncertainty-based sampling strategies. We consider the typical setting where we start with a small labeled dataset $\mathcal{D}_{seed}$ of n_{seed} samples and a large unlabeled dataset $\mathcal{D}_u$ with n_u samples all i.i.d. drawn from the truncated Gaussian distribution. We denote by $\rho = \frac{n_{\text{seed}}}{n_\ell}$ the fraction of samples we query using uniform sampling. Recall that we define $\hat{\theta}_{unif}$ and $\hat{\theta}_{oracle}$ as the classifiers obtained after querying $(1-\rho)n_\ell$ samples either uniformly or with oracle-based uncertainty, respectively.

We now introduce some assumption on the setting. In practice, an unlabeled dataset is available that is much larger than the number of queries. Moreover, most real-world datasets contain some hard examples that are difficult to classify even by human experts. The equivalent synthetic counterpart is the existence of unlabeled samples close to the Bayes optimal decision boundary. Finally, we consider high-dimensional settings, where the dimensionality is larger than the labeling budget. To state our theorem formally, we make these conditions precise in the following assumption.

Assumption A.1. *We assume that* $n_u > \max(10^3 n_\ell, 10^5)$, $\mu/\sigma < 2$ *and* $d > n_\ell$.

With Assumption A.1, we are now ready to state the theorem.

Theorem A.2 (Oracle uncertainty sampling)**.** *For a small constant* $c > 0$ *independent of the dimension and under Assumption A.1, it holds with probability of at least* $0.99(1-2e^{-t^2/2})^5(1-e^{-c\frac{d}{n_\ell}})$ *that:*

$$\textit{Err}(\hat{\theta}_{oracle}) - \textit{Err}(\hat{\theta}_{unif}) \geq \Psi_{\mu,\sigma}(\alpha_{\textit{oracle}}) - \Psi_{\mu,\sigma}(\alpha_{\textit{unif}}), \tag{5}$$

where

$$\alpha_{\textit{unif}} = \frac{\sqrt{d/n_\ell + 2tn_\ell^{-1/2}}}{\mu_{tr} - t\sigma_{tr}n_\ell^{-1/2}} \qquad \alpha_{\textit{oracle}} = \frac{\sqrt{d/n_\ell} - 1 - t}{\rho(\mu_{tr} + t\sigma_{tr}n_{\textit{seed}}^{-1/2}) + (1-\rho)6.059\cdot 10^{-2}\mu_{tr}} \tag{6}$$

A consequence of Theorem A.2 is that for high-dimensional data (i.e. $d \gg n_\ell$) and for a small ratio of uniformly sampled seed points $\rho \ll 1$, then with high probability $\text{Err}(\hat{\theta}_{oracle}) - \text{Err}(\hat{\theta}_{unif}) > 0$. We state this observation precisely in Corollary A.3 and prove it in Section B.8.

Corollary A.3. *Under the same assumptions and with the same probability as in Theorem A.2 it holds that* $\textit{Err}(\hat{\theta}_{oracle}) - \textit{Err}(\hat{\theta}_{unif}) > 0$ *if*

$$\frac{d}{n_\ell} > 4\left(1 + t + \sqrt{t(4n_\ell)^{-1/2}}\right)^2, \qquad \rho < \frac{1}{2} - \frac{1+\sqrt{2}}{2}\frac{t\sigma_{tr}}{\sqrt{n_\ell}\mu_{tr}}. \tag{7}$$

The condition on ρ ensures that enough samples are queried using oracle uncertainty sampling such that the difference to passive learning is large enough. Finally, the term $(1-e^{-c\frac{d}{n_\ell}})$ lower bounds the probability that all points of each sampling strategy are support points. Explicit expressions for the terms can be found in Lemmas A.10 and A.8 for oracle uncertainty and uniform sampling respectively.

Lastly, note that Theorem A.2 and Corollary A.3 are the formalization of Theorem 3.1: set

$$\epsilon = \max\left(\sqrt{\frac{n_\ell}{d}}(1+t), \sqrt{1 + \frac{2tn_\ell^{1/2}}{d}} - 1\right), \tag{8}$$

then for large d/n_ℓ we retrieve the informal statement. In particular, as claimed in Section 3, for $d/n_\ell \to \infty$ we have $\epsilon \to 0$.

A.3 Formal statement of Theorem 3.2

In this section, we state the formal version of Theorem 3.2 which shows an error gap between passive learning and active learning using the uncertainty of the empirical classifier $\hat{\theta}$. Before we state the theorem, we first discuss two-stage uncertainty sampling, a slight modification of Algorithm 1.

Two-stage uncertainty sampling. We consider the same modification of the uncertainty sampling procedure as (Chaudhuri et al., 2015; Mussmann and Liang, 2018). Instead of the iterative process of labeling a point and updating the estimator $\hat{\theta}$, we use a two-stage procedure: 1) we obtain $\hat{\theta}_{seed}$ using the initial small seed set; and 2) we use $\hat{\theta}_{seed}$ to select a batch of $(1-\rho)n_\ell$ samples to query from the unlabeled set. Without this two-stage strategy, the estimator $\hat{\theta}$ at a certain iteration is not independent of the unlabeled set, which makes the analysis more challenging, as also noted by (Mussmann and Liang, 2018).

We stress that we do not need this simplification for the analysis of oracle uncertainty sampling, since with this strategy the queried points are independent of the estimators $\hat{\theta}$. Moreover, the two-stage procedure is necessary only for one step of the proof highlighted in Section B.2.

We now state the main theorem for empirical uncertainty sampling:

Theorem A.4. *For a small constants $c > 0$ independent of the dimension and under Assumption A.1 with $\sigma > 1$, it holds with probability of at least $0.99(1-2e^{-t^2/2})^5(1-e^{-c\frac{d}{n_\ell}})$ that:*

$$\mathit{Err}(\hat{\theta}_{uncert}) - \mathit{Err}(\hat{\theta}_{unif}) \geq \Psi_{\mu,\sigma}(\alpha_{\mathit{uncert}}) - \Psi_{\mu,\sigma}(\alpha_{\mathit{unif}}), \tag{9}$$

where α_{unif} is defined as in Theorem A.2 and

$$\alpha_{\mathit{uncert}} = \frac{\sqrt{d/n_\ell} - \sqrt{2\log n_u} - 1 - t}{\rho C_{seed} + (1-\rho)\left(0.061\mu_{tr} + \sqrt{\frac{2\log n_u}{C_{seed}}}\left(\frac{d+\sigma_{tr}t}{\rho n_\ell}\right)^{1/4} + t\right)} \tag{10}$$

with C_{seed} a constant that satisfies

$$\mu_{tr} - t\sigma_{tr}n_{\mathit{seed}}^{-1/2} \leq C_{seed} \leq \mu_{tr} + t\sigma_{tr}n_{\mathit{seed}}^{-1/2} \tag{11}$$

Theorem A.4 gives a high probability bound for the error gap between passive learning and two-stage empirical uncertainty sampling. As before, we state precise conditions when this gap is positive in Corollary A.5 and provide the proof in in Section B.9.

Corollary A.5. *Under the same assumptions and with the same probability as in Theorem A.4 it holds that $\mathit{Err}(\hat{\theta}_{oracle}) - \mathit{Err}(\hat{\theta}_{unif}) > 0$ if the following conditions are satisfied:*

1. *(high-dimensional regime) $d/n_\ell > 4\left(\sqrt{2\log n_u} + 1 + t + \sqrt{t}(4n_\ell)^{-1/4}\right)$*
2. *(large signal-to-noise ratio) $\mu_{tr} \geq \left(\frac{d+\sigma_{tr}t}{\rho n_\ell}\right)^{1/6}(\log n_u)^{1/3} + \frac{t\sigma_{tr}}{(\rho n_\ell)^{1/2}}$*
3. *(numerous uncertainty-based queries) $\rho < \frac{2C_{seed}}{0.878\mu_{tr} - \frac{t\sigma_{tr}}{(\rho n_\ell)^{1/2}} - \left(\frac{d+\sigma_{tr}t}{\rho n_\ell}\right)^{1/6}(\log n_u)^{1/3} - t}$.*

The second condition is necessary to ensure that the classifier $\hat{\theta}_{seed}$ trained on the seed set has non-trivial error. Like in Section A.2, the third condition guarantees that the influence of the uniformly sampled seed set is reduced. To get an explicit condition on ρ on the right hand side, we note that $\rho n_\ell \geq 2$ by definition.

Furthermore, similar to Section A.2, the term $(1-e^{-c\frac{d}{n_\ell}})$ lower bounds the probability that all points are support points for each of the two sampling strategies. Explicit expressions for the term can be found in Lemmas A.12 and A.8 for empirical uncertainty and uniform sampling respectively.

Finally, observe that Theorem A.4 and Corollary A.5 are together the formalization of theorem 3.2. More specifically, by setting ϵ similar as in Equation 8 and considering large d/n_ℓ, we find the informal statement.

A.4 Proofs of Theorems A.2 and A.4

Recall that, without loss of generality, we consider predictors $\theta = [1, \alpha\tilde{\theta}]$, with $\|\tilde{\theta}\|_2 = 1$ and we can write the population error of θ as a strictly increasing function of α. Therefore, to prove Theorems A.2 and A.4 we derive bounds on α for uniform and oracle/empirical uncertainty sampling. We split the proof in 3 main steps. In the first step we bound the α-parameter of a dataset obtained using an arbitrary sampling strategy as a function of certain geometric quantities. The second step then bounds these geometric quantities for the specific sampling strategies that we are interested in. Lastly, in the third step we develop these bounds further for the special case of a mixture of truncated Gaussians. Our results also hold for $\mu \to 0$, which recovers the marginal Gaussian distribution that is usually analyzed in the active learning literature (Hanneke, 2013).

We reiterate that we focus on separable data, a setting that benefits active learning. As described in Section 3.2, we consider a Bayes optimal predictor $\theta^\star$ with vanishing population error and choose without loss of generality $\theta^\star = e_1 = [1, 0, ..., 0] \in \mathbb{R}^d$.[7] We can write the covariates as $x = [x_1, \tilde{x}]$, where we explicitly separate the coordinates of x into a signal $x_1 \in \mathbb{R}$ and non-signal component $\tilde{x} \in \mathbb{R}^{d-1}$. The marginal distribution of the covariates takes the form $\mathbb{P}(x) = \mathbb{P}(x_1) \cdot \mathbb{P}(\tilde{x})$, where $\mathbb{P}(\tilde{x}) = \mathcal{N}(\tilde{x}; 0, I_{d-1})$ is the distribution of the non-signal dimensions.

We point out that the first two steps of the proof of Theorems A.2 and A.4 hold for any arbitrary distribution $\mathbb{P}(x_1)$. If $\mathbb{P}(x_1)$ is a mixture of truncated Gaussians, then $\mathbb{P}(x) = \mathbb{P}_{\text{TGMM}}(\mu, \sigma)$, which in turn corresponds to a marginal Gaussian for $\mu \to 0$.

Characterizing $\hat{\theta}$ for arbitrary $\mathbb{P}(x)$. To state the key lemma, we first introduce three geometric quantities. Let $\mathcal{D}_\ell \subset \mathbb{R}^d \times \{-1, 1\}$ be a dataset of n_ℓ labeled samples. Importantly, note that the covariates need not be drawn from the distribution $\mathbb{P}$ described above. Instead, $\mathcal{D}_\ell$ can be acquired using any query strategy from an unlabeled set $\mathcal{D}_u \sim \mathbb{P}$. Therefore, the only condition on $\mathcal{D}_\ell$ is that $x \in \text{supp}(\mathbb{P})$ for any $(x, y) \in \mathcal{D}_\ell$. We define the max-$\ell_2$-margin of $\mathcal{D}_\ell$ in the last $d-1$ coordinates as:

$$\tilde{\gamma} = \max_{\tilde{\theta}} \min_{(x,y)\in\mathcal{D}_\ell} y \frac{\langle \tilde{x}, \tilde{\theta} \rangle}{\|\tilde{\theta}\|_2} \tag{12}$$

Similarly, the max-average-ℓ_2-margin of $\mathcal{D}_\ell$ in the last $d-1$ coordinates is defined as

$$\tilde{\gamma}_{avg} = \max_{\tilde{\theta}} \frac{1}{n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell} y \frac{\langle \tilde{x}, \tilde{\theta} \rangle}{\|\tilde{\theta}\|_2} \tag{13}$$

Lastly, we define the average distance to the decision boundary of the optimal classifier induced by $\theta^\star$ as

$$d^* = \frac{1}{n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell} y x_1 \tag{14}$$

We now state the lemma that bounds the parameter α of the max-ℓ_2-margin classifier trained on an arbitrary labeled set. We provide the proof of the lemma in Section B.1.

Lemma A.6 (Bound on the optimal classifier for active learning)**.** *Let $\mathcal{D}_\ell$ be a labeled dataset with $x \in$ supp$(\mathbb{P})$ for any $(x, y) \in \mathcal{D}_\ell$. If all covariates of the dataset are support vectors of the max-ℓ_2-margin classifier $\hat{\theta}$ of $\mathcal{D}_\ell$, then the α-parameter of $\hat{\theta}$ is bounded as follows:*

$$\frac{\tilde{\gamma}}{d^*} \leq \alpha \leq \frac{\tilde{\gamma}_{avg}}{d^*}$$

Once equipped with Lemma A.6, the next step is to derive bounds on $\tilde{\gamma}$, $\tilde{\gamma}_{avg}$ and d^* for uniform and oracle/empirical uncertainty sampling.

[7] If $\theta^\star \neq e_1$, we can rotate and translate the data in order to get $\theta^\star = e_1$.

Bounding d^*, $\tilde{\gamma}$ and $\tilde{\gamma}_{avg}$ for specific sampling strategies. In this step, we derive further the bounds of Lemma A.6 and consider specific sampling strategies, namely uniform and oracle/empirical uncertainty sampling. We begin by introducing some geometric quantities which we will use to bound the α-parameter. First, we denote by $d_q^\star$ the maximal distance of the newly sampled queries to the decision boundary of $\theta^\star$:

$$d_q^\star = \max_{(x,y)\in\mathcal{D}_\ell\setminus\mathcal{D}_{seed}} yx_1 \tag{15}$$

Furthermore, let $\hat{\theta}_{seed}$ be the parameter vector of the max-ℓ_2-margin classifier of the seed set with α-parameter α_{seed}. We define $\hat{d}_q$ as the maximal distance of the newly queried points to the decision boundary determined by $\hat{\theta}_{seed}$:

$$\hat{d}_q = \max_{(x,y)\in\mathcal{D}_\ell\setminus\mathcal{D}_{seed}} \frac{\left|\langle\hat{\theta}_{seed}, x\rangle\right|}{\|\hat{\theta}_{seed}\|_2} \tag{16}$$

Lastly, we define C_{seed} as the average distance to the decision boundary of $\theta^\star$ of the samples in the seed set:

$$C_{seed} = \frac{1}{n_{\text{seed}}} \sum_{(x,y)\in\mathcal{D}_{seed}} yx_1 \tag{17}$$

Finally, recall that in pool-based active learning one has access to an unlabeled set $\mathcal{D}_u$ drawn i.i.d. from $\mathbb{P}$ and a small labeled seed set $\mathcal{D}_{seed}$ where the covariates are also drawn i.i.d. from $\mathbb{P}$. Moreover, the noiseless label of a sample x is given by $y = \operatorname{sgn}(\langle x, \theta^\star\rangle)$. We collect a labeled set $\mathcal{D}_\ell$ that includes the uniformly sampled $\mathcal{D}_{seed}$ and $(1-\rho)n_\ell$ labeled points whose covariates are selected from $\mathcal{D}_u$ according to a sampling strategy. We are now ready to state the following lemma, which bounds the quantities that show up in Lemma A.6, namely d^*, $\tilde{\gamma}$ and $\tilde{\gamma}_{avg}$. The proof of the lemma is presented in Section B.2.

Lemma A.7 (Bounds on d^*, $\tilde{\gamma}$ and $\tilde{\gamma}_{avg}$)**.** *Consider the standard pool-based active learning setting in which we collect a labeled set $\mathcal{D}_\ell$ and assume $n_\ell < d < n_u$ where $n_\ell = |\mathcal{D}_\ell|$ and $n_u = |\mathcal{D}_u|$. Moreover, we assume that all points in $\mathcal{D}_\ell$ are support points. The following is true about $\hat{\theta} = [1, \alpha\tilde{\theta}]$, i.e. the max-$\ell_2$-margin classifier trained on $\mathcal{D}_\ell$:*

1. *If $\mathcal{D}_\ell$ is collected using **two-stage empirical uncertainty sampling**, then with probability greater than $(1-2e^{-t^2/2})^3$, it holds that*

$$d^* < \rho C_{seed} + (1-\rho)(\hat{d}_q + \sqrt{2\alpha_{seed}\log n_u} + t) \quad \tilde{\gamma} > \sqrt{d/n_\ell} - \sqrt{2\log n_u} - 1 - t$$

2. *If $\mathcal{D}_\ell$ is collected using **oracle uncertainty sampling**, then with probability greater than $(1-2e^{-t^2/2})^2$ it holds that*

$$d^* < \rho C_{seed} + (1-\rho)d_q^\star \quad \tilde{\gamma} > \sqrt{d/n_\ell} - 1 - t$$

3. *If $\mathcal{D}_\ell$ is collected using **uniform sampling**, then with a probability greater than $(1-2e^{-t^2/2})^2$, it holds that*

$$d^* = \frac{1}{n_\ell}\sum_{(x,y)\in\mathcal{D}_\ell} yx_1 \quad \tilde{\gamma}_{avg} < \sqrt{d/n_\ell + 2tn_\ell^{-1/2}}$$

Plugging these bounds into the result of Lemma A.6, we find high-probability bounds on the α-parameter of each of these 3 sampling strategies. As explained in Section A.4, these bounds hold for any arbitrary distribution of the signal component $\mathbb{P}(x_1)$. In what follows we derive the bounds on α further for a mixture of truncated Gaussians.

Bounding α for a mixture of truncated Gaussian. In this step, we bound $d_q^\star, \hat{d}_q, \alpha_{seed}$ and C_{seed} assuming that the signal component is distributed according to a mixture of truncated Gaussians $x_1 \sim \mathbb{P}_{\text{TGMM}}(\mu, \sigma)$, with $\mu \geq 0$ and $\sigma > 0$. Importantly, if we consider $\mu = 0$, the signal component is drawn from a normal distribution and the classification problem has vanishing margin, the setting often analyzed in the active learning literature (Hanneke, 2013). We treat each of the 3 sampling strategies separately.

Uniform sampling. We note that for uniform sampling, d^* is the average of n_ℓ i.i.d. samples from a one-sided truncated Gaussian, which is a sub-Gaussian random variable with mean μ_{tr} and standard deviation σ_{tr}. Hence, with probability larger than $1 - 2e^{-t^2/2}$, it holds that

$$d^* < \mu_{tr} - t\sigma_{tr} n_\ell^{-1/2} \tag{18}$$

It is left to show that all points in $\mathcal{D}_\ell$ are support points with high probability. We use the following lemma for which we provide the proof in Section B.7.

Lemma A.8 (Support points for uniform sampling)**.** *Let $\mathcal{D}_\ell$ be a dataset of n_ℓ samples drawn i.i.d. from the mixture of truncated Gaussians distribution. Then, for a constant $c > 0$ independent of d and n_ℓ with a probability larger than $1 - 2e^{-(\sqrt{d/n_\ell} - \sqrt{\log n_\ell} - c)^2}$ all samples in $\mathcal{D}_\ell$ are support points of the max-ℓ_2-margin classifier of $\mathcal{D}_\ell$.*

Plugging the bounds on d^* (Equation 18) and $\tilde{\gamma}_{avg}$ (Lemma A.7) into Lemma A.6 and multiplying the independent probability statements gives us the expression of α_{unif} which appears in both Theorems A.2 and A.4.

Oracle uncertainty sampling. Since a one-sided truncated Gaussian is a sub-Gaussian random variable with mean μ_{tr} and standard deviation σ_{tr}, it holds with probability greater then $1 - 2e^{-t^2/2}$ that

$$C_{seed} < \mu_{tr} + t\sigma_{tr} n_{\text{seed}}^{-1/2} \tag{19}$$

Furthermore, we bound $d_q^\star$ using the following lemma and give the proof of the lemma in Section B.5.

Lemma A.9 (Bound on $d_q^\star$)**.** *Let $\mathcal{D}_u$ and $\mathcal{D}_{seed}$ be the unlabeled set and the labeled seed set, respectively, with covariates drawn i.i.d. from the mixture of truncated Gaussians distribution. Then, with probability larger than $1 - e^{-t^2}$, we have that*

$$d_q^\star < \sigma \left(\Phi^{-1} \left(\left(t(2n_u)^{-1/2} + (1-\rho) n_\ell / n_u \right) (1 - \Phi(-\mu/\sigma)) + \Phi(-\mu/\sigma) \right) \right) + \mu \tag{20}$$

Moreover, if $n_u > \max(10^5, 10^3 n_\ell)$ and $\mu/\sigma < 2$ then with probability greater than 0.99

$$d_q^\star < 6.059 \cdot 10^{-2} \mu_{tr} \tag{21}$$

We now argue that the conditions required for Equation 21 to hold are not too restrictive. Indeed, it is standard in practical active learning settings that the unlabeled set is orders of magnitude larger than the labeling budget. Moreover, in most real-world datasets there exist ambiguous samples, close to the optimal decision boundary, which can be be difficult to classify even for human experts. The condition $\mu/\sigma < 2$ ensures that that is the case in our setting as well, with high probability.

It is left to consider the probability that all samples in $\mathcal{D}_\ell$ are support points for oracle uncertainty sampling. Note that if a point in the seed set is not a support point, then we can delete it out of the labeled set by plugging in $n_{\text{seed}} - 1$ and $n_\ell - 1$ instead of n_{seed} and n_ℓ in Lemma A.7. Since $d_q^\star < C_{seed}$ by definition, reducing n_{seed} and n_ℓ will only increase the upper bound on $d_q^\star$. Therefore, it suffices to derive the probability that all the newly queried samples are support points. We state the following lemma and prove it in Section B.7.

Lemma A.10 (Support points for oracle uncertainty sampling)**.** *For a small constant $c > 0$ independent of d and n with a probability larger than $1 - 2e^{-\frac{1}{2}(\sqrt{d/n_\ell} - \sqrt{\log n_\ell} - c)^2}$ all newly queried points using oracle uncertainty sampling are support points.*

Plugging the bounds on C_{seed} (Equation 19), $\tilde{\gamma}$ (Lemma A.7) and $d_q^\star$ (Lemma A.9) into Lemma A.6 gives the expression for α_{oracle} that appears in Theorem A.2. Invoking all the probability statements involved and combining this result with the previous derivation of α_{unif} finishes the proof of Theorem A.2.

Two-stage uncertainty sampling. For bounding $\hat{d}_q$ we can use a similar technique as in Lemma A.9, if we assume further that $\sigma \geq 1$. This condition ensures that, with high probability, there exist examples with a high signal component for any $\mu \geq 0$. The following lemma states the bound on $\hat{d}_q$ (see Section B.6 for the proof).

Lemma A.11 (Bound on $\hat{d}_q$)**.** *Let $\mathcal{D}_u$ and $\mathcal{D}_{seed}$ be the unlabeled set and the labeled seed set, respectively, with covariates drawn i.i.d. from the mixture of truncated Gaussians distribution. If $\sigma \geq 1$ then it holds with probability larger than $1 - e^{-t^2}$ that*

$$\hat{d}_q < \sigma \left(\Phi^{-1} \left(\left(t(2n_u)^{-1/2} + (1-\rho)n_\ell/n_u \right) (1 - \Phi(-\mu/\sigma)) + \Phi(-\mu/\sigma) \right) \right) + \mu \tag{22}$$

Moreover, if $n_u > \max(10^5, 10^3 n_\ell)$ and $\mu/\sigma < 2$ then with a probability greater than 0.99

$$\hat{d}_q < 6.059 \cdot 10^{-2} \mu_{tr} \tag{23}$$

Using Lemma A.11, we now derive an upper bound on α_{seed}. We note that $\alpha_{\text{seed}} > \tilde{\gamma}/d^*$ by Lemma A.6, where we take $\mathcal{D}_\ell = \mathcal{D}_{seed}$. Then, by Lemma A.7, we have that $\tilde{\gamma} < \left(\frac{d}{\rho n_\ell} + \frac{2t\sigma_{tr}}{(\rho n_\ell)^{1/2}} \right)^{1/2}$ with probability greater than $1 - 2e^{-t^2/2}$. Moreover, we can lower bound C_{seed} using the expression in the uniform case. We find that $C_{seed} > \mu_{tr} - \frac{t\sigma_{tr}}{(\rho n_\ell)^{1/2}}$. Note that if all uniform samples are support points of the max-ℓ_2-margin classifier, then all samples in the seed set are as well for the max-ℓ_2-margin classifier of the seed set. Putting things together, we find that with a probability greater than $(1 - 2e^{-t^2/2})^2$

$$\alpha_{\text{seed}} \leq \left(\frac{d}{\rho n_\ell} + \frac{2t\sigma}{(\rho n_\ell)^{1/2}} \right)^{1/2} \left(\mu_{tr} - (\rho n_\ell)^{-1/2} \sigma_{tr} t \right)^{-1} \tag{24}$$

We now argue that all points in the labeled dataset are with high probability support points. We state the lemma here and give the proof in Section B.7.

Lemma A.12. *For a constant $c > 0$ independent of d and n, we have with a probability of at least $1 - 2e^{-\frac{1}{2}(\sqrt{d/n_\ell} - \sqrt{\log n_\ell} - c)^2}$ that all newly queried points are support points.*

Plugging the bounds on C_{seed} (Equation 19), $\tilde{\gamma}$ (Lemma A.7) , α_{seed} (Equation 24) and $\hat{d}_q$ (Lemma A.11) into Lemma A.6 gives the expression for α_{uncert} that appears in Theorem A.4. Invoking all the probability statements involved and combining this result with the previous derivation of α_{unif} finishes the proof of Theorem A.4.

B Proofs of Lemmas

In this section, we provide proofs for the lemmas needed to prove the main theoretical results presented in Section A.

B.1 Proof of Lemma A.6

We start by rewriting the max-ℓ_2-margin classifier $\hat{\theta}$ in a more convenient form. Recall that the error of any classifier induced by a vector θ is invariant to scaling θ. Hence, for $b > 0$, we can write the max-ℓ_2-margin classifier in the form

$$\hat{\theta} = [1, b\tilde{\theta}],$$

with $\|\tilde{\theta}\|_2 = 1$. For convenience of notation we define $\bar{a} = \frac{1}{n_\ell} \sum_{(x,y) \in \mathcal{D}_\ell} y \langle \tilde{\theta}, \tilde{x} \rangle$ to be the average over $\mathcal{D}_\ell$ of the distance between the noise dimensions of a sample $\tilde{x}$ and the decision boundary of the max-ℓ_2-margin in the last $d-1$ coordinates. By the definition of support points, the distance of all support points to the max-ℓ_2-margin classifier is equal. Therefore, we can write the max-ℓ_2-margin γ as any of the n_ℓ margins of $\hat{\theta}$, or, equivalently, as the average:

$$\begin{aligned}
\gamma &= \frac{1}{\|\hat{\theta}\|_2 n_\ell} \sum_{(x,y) \in \mathcal{D}_\ell} \left(y x_1 + b y \langle \tilde{\theta}, \tilde{x} \rangle \right) \\
&= \frac{1}{\sqrt{1+b^2}} \left(\frac{1}{n_\ell} \sum_{(x,y) \in \mathcal{D}_\ell} y x_1 + b\bar{a} \right)
\end{aligned}$$

We find that

$$b = \frac{\bar{a}}{\frac{1}{n_\ell} \sum_{(x,y) \in \mathcal{D}_\ell} y x_1} = \frac{\bar{a}}{d^*}$$

maximizes the ℓ_2-margin γ. Hence, the max-ℓ_2-margin classifier can be written as:

$$\hat{\theta} = \left[1, \frac{\bar{a}}{d^*}\tilde{\theta}\right].$$

Now we prove that we can use the maximum average-ℓ_2-margin and the max-ℓ_2-margin in the $d-1$ noise coordinates to bound $\bar{a}$ as follows: $\tilde{\gamma}_{avg} \geq \bar{a} \geq \tilde{\gamma}$.

Note that $\bar{a}$ represents the average-ℓ_2-margin in the $d-1$ coordinates of the predictor $\hat{\theta}$, where $\hat{\theta}$ is the max-ℓ_2-margin classifier of $\mathcal{D}_\ell$ considering both signal and noise coordinates. By the definition of the maximum average-ℓ_2-margin, we have that $\tilde{\gamma}_{\text{avg}} \geq \bar{a}$.

To prove the other inequality, we first introduce some notation. We fix a permutation of the samples in $\mathcal{D}_\ell$ and construct the matrix $\bar{X}_{d-1} \in \mathbb{R}^{(d-1)\times n_\ell}$ where each column holds the last $d-1$ coordinates of each sample multiplied by the corresponding label. The samples are ordered according to the fixed permutation. Analogously, let $\bar{x}_{sig} \in \mathbb{R}^{n_\ell}$ and $\bar{y} \in \{-1, 1\}^{n_\ell}$ be vectors where each entry corresponds to yx_1 and y, respectively, for all $(x, y) \in \mathcal{D}_\ell$. Further, we define

$$\epsilon := \frac{1}{d^*}\bar{x}_{sig}^\top \mathbb{I}_{n_\ell}\bar{y} \in \mathbb{R}^{n_\ell} \text{ and } \eta := \frac{1}{\bar{a}}\tilde{\theta}^\top \bar{X}_{d-1} \in \mathbb{R}^{n_\ell} \tag{25}$$

and note that, by definition, we have that $\frac{1}{n_\ell}\sum_{k=1}^{n_\ell} \epsilon_k = 1$ and $\frac{1}{n_\ell}\sum_{k=1}^{n_\ell} \eta_k = 1$. Using the fact that support points have equal margin to the classifier, it holds for any $(x, y) \in \mathcal{D}_\ell$ that

$$y\langle \hat{\theta}, x\rangle = \langle \bar{x}_{sig}, \epsilon\rangle^2 + \bar{a}^2\langle \eta, \mathbb{1}_{d-1}\rangle = \left(\sum_{(x,y)\in\mathcal{D}_\ell} yx_1\right)^2 + \bar{a}^2.$$

Further, let θ_{d-1} be the max-ℓ_2-margin classifier and $\tilde{\gamma}$ be the max-ℓ_2-margin in the $d-1$ last coordinates of the dataset, with $\|\theta_{d-1}\|_2 = 1$. Then, there exists an orthonormal matrix $Q \in \mathbb{R}^{(d-1)\times(d-1)}$ such that $Q\tilde{\theta} = \theta_{d-1}$. By the definition of Q, we have that

$$\bar{a}\eta = \tilde{\theta}^\top \bar{X}_{d-1} = Q\theta_{d-1}^\top \bar{X}_{d-1} \geq \tilde{\gamma}\mathbb{1}_{n_\ell},$$

where the inequality is element-wise. For the inequality we use that all points are at least $\tilde{\gamma}$ from the decision boundary of the max-ℓ_2-margin classifier. Comparing the norms yields

$$\bar{a}\|\eta\|_2 \geq \sqrt{n_\ell}\tilde{\gamma} \iff \frac{\|\eta\|_2}{\sqrt{n_\ell}} \geq \frac{\tilde{\gamma}}{\bar{a}}$$

Since $\sum_{k=1}^{n_\ell} \eta_k = n_\ell$, we have that $\|\eta\|_2 \leq \sqrt{n_\ell}$. Hence, $\bar{a} > \tilde{\gamma}$ which proves the lemma.

B.2 Proof of Lemma A.7

Lemma A.7 consists of three statements: a lower bound on the value of α for uncertainty sampling and oracle uncertainty sampling, and an upper bound on α for passive learning.

Recall that by Lemma A.6 we have that for a labeled set $\mathcal{D}_\ell$ collected through any sampling strategy, the α-parameter of the max-ℓ_2-margin classifier of $\mathcal{D}_\ell$ is lower and upper bounded by

$$\frac{\tilde{\gamma}}{d^*} \leq \alpha \leq \frac{\tilde{\gamma}_{avg}}{d^*} \tag{26}$$

where $d^* = \frac{1}{n_\ell}\sum_{(x,y)\in\mathcal{D}_\ell} yx_1$. To prove Lemma A.7, we apply Lemma A.6 and bound $\tilde{\gamma}, \tilde{\gamma}_{avg}$ and d^* for each sampling strategy separately.

Oracle uncertainty sampling. Consider the max-ℓ_2-margin classifier trained only on the $d-1$ noise coordinates of a dataset $\mathcal{D}_\ell$. Let us define the $d-1$-dimensional labeled dataset that contains only the noise components of the covariates:

$$\tilde{\mathcal{D}}_\ell := \{(\tilde{x}, y) : (x, y) \in \mathcal{D}_\ell, \text{where } x = [x_1, \tilde{x}]\} \tag{27}$$

Recall that the noise components $\tilde{x}$ are drawn i.i.d. from $\mathbb{N}(0, \mathbb{I}_{d-1})$. We use the following lemma to bound the max-ℓ_2-margin of $\tilde{\mathcal{D}}_\ell$ and provide its proof in Section B.3.

Lemma B.1. *[Non-signal max-ℓ_2-margin of i.i.d. subset of $\mathcal{D}_u$] Let $\tilde{\mathcal{D}}_\ell$ be a labeled dataset of size $n_\ell < d-1$ where $\tilde{x} \sim \mathbb{N}(0, \mathbb{I}_{d-1})$ for all $(\tilde{x}, y) \in \tilde{\mathcal{D}}_\ell$. Then it holds with probability at least $1 - e^{-t^2/2}$ that the max-ℓ_2-margin of $\tilde{\mathcal{D}}_\ell$ is upper and lower bounded by*

$$\sqrt{\frac{d}{n_\ell}} - 1 - t \leq \tilde{\gamma} \leq \sqrt{\frac{d}{n_\ell}} + 1 + t.$$

Note that oracle uncertainty sampling queries the $(1-\rho)n_\ell$ closest points to the optimal decision boundary. Importantly, the Bayes optimal classifier is independent of the $d-1$ noise coordinates of the covariates. Therefore, $\tilde{\mathcal{D}}_\ell$ selected with oracle uncertainty sampling is drawn i.i.d. from a standard normal distribution, and hence, satisfies the conditions of Lemma B.1. By applying the lemma we get that $\tilde{\gamma}_{oracle} \geq \sqrt{d/n_\ell} - 1 - t$ with probability greater than $1 - 2e^{-t^2/2}$.

Next, we bound $d^* = \frac{1}{n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell} yx_1$. The terms of the sum can be partitioned into the ones corresponding to samples in $\mathcal{D}_{seed}$ and the ones corresponding to points in $\mathcal{D}_\ell \setminus \mathcal{D}_{seed}$. Let $C_{seed} = \frac{1}{n_{\text{seed}}} \sum_{(x,y)\in\mathcal{D}_{seed}} yx_1$ be the former. Note that the term C_{seed} is independent of the sampling strategy and can hence be treated as a constant when comparing sampling strategies. Lastly, we observe that $d_q^\star$ is by definition an upper bound on the average over the points in $\mathcal{D}_\ell \setminus \mathcal{D}_{seed}$. Using the definition that $\rho = \frac{n_{\text{seed}}}{n_\ell}$, we find that

$$\frac{1}{n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell} yx_1 \leq \rho C_{seed} + (1-\rho) d_q^\star$$

This concludes the statement about oracle uncertainty sampling.

Uncertainty sampling. We first lower bound the max-ℓ_2-margin in the $d-1$ noise coordinates using the following lemma, where we define $\tilde{\mathcal{D}}_u$ analogously to $\tilde{\mathcal{D}}_\ell$. The proof of the lemma is presented in Section B.3.

Lemma B.2. *[Non-signal max-ℓ_2-margin for arbitrary subset of $\mathcal{D}_u$] Let $\tilde{\mathcal{D}}_u$ be an unlabeled dataset of size $n_u > d-1$ where $\tilde{x} \sim \mathbb{N}(0, \mathbb{I}_{d-1})$ for all $\tilde{x} \in \tilde{\mathcal{D}}_u$. Further, let $\tilde{\mathcal{D}}_\ell$ be any arbitrary labeled dataset of size n_ℓ where $\tilde{x} \in \tilde{\mathcal{D}}_u$ for all $(\tilde{x}, y) \in \tilde{\mathcal{D}}_\ell$. Then, with probability at least $1 - e^{-t^2/2}$, the max-ℓ_2-margin of $\tilde{\mathcal{D}}_\ell$ is upper and lower bounded by*

$$\sqrt{d/n_\ell} + \sqrt{2\log n_u} + 1 + t \geq \tilde{\gamma} \geq \sqrt{d/n_\ell} - \sqrt{2\log n_u} - 1 - t.$$

As Lemma B.2 applies for *any* sampling strategy, we find that with probability greater than $1 - e^{-t^2/2}$, the max-ℓ_2-margin in the $d-1$ noise coordinates is lower bounded by $\sqrt{d/n_\ell} - \sqrt{2\log n_u} - 1 - t$.

Next we bound $d^* = \frac{1}{n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell} yx_1$. We stress that this is the only step in the entire proof of Theorem A.4 where we use the two-stage uncertainty sampling procedure (instead of the iterative process described in Algorithm 1).

As in the case of oracle uncertainty sampling, we need to derive a bound for the sum of the terms that correspond to samples from $\mathcal{D}_\ell \setminus \mathcal{D}_{seed}$. Recall that $\hat{\theta}_{seed} = [1, \alpha_{\text{seed}}\tilde{\theta}_{seed}]$ with $\|\tilde{\theta}_{seed}\|_2 = 1$ is the max-ℓ_2-margin classifier of $\mathcal{D}_{seed}$. Moreover, by definition it holds that $\hat{d}_q \geq y\langle\hat{\theta}_{seed}, x\rangle = yx_1 + y\alpha_{\text{seed}}\langle\tilde{\theta}_{seed}, \tilde{x}\rangle$ for all $(x, y) \in \mathcal{D}_\ell \setminus \mathcal{D}_{seed}$.

Due to the two-stage procedure, $\tilde{\theta}_{seed}$ is independent of all the samples in the unlabeled dataset. Using this fact together with the union bound, we find that $\max_{(x,y)\in\mathcal{D}_u} \alpha_{\text{seed}}\langle\tilde{\theta}_{seed}, \tilde{x}\rangle < \sqrt{2\alpha_{\text{seed}}\log n_u} + t$ with a probability greater than $1 - 2e^{-t^2/2}$. Putting everything together we find that the following holds with probability greater than $1 - 2e^{-t^2/2}$:

$$\frac{1}{(1-\rho)n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell\setminus\mathcal{D}_{seed}} yx_1 < \hat{d}_q + \sqrt{2\alpha_{seed}\log n_u} + t$$

Hence we find that

$$\frac{1}{n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell} yx_1 \leq \rho C_{seed} + (1-\rho)(\hat{d}_q + \sqrt{2\alpha_{seed}\log n_u} + t)$$

with probability at least $1 - 2e^{-t^2/2}$, which concludes the proof for two-stage uncertainty sampling.

Uniform sampling. In the case of uniform sampling we need to upper bound the maximum average-ℓ_2-margin of $\tilde{\mathcal{D}}_\ell$, namely $\tilde{\gamma}_{avg}$. We use the following lemma which we prove in Section B.4.

Lemma B.3 (Upper bound on $\tilde{\gamma}_{avg}$ for uniform sampling)**.** *Let $\tilde{\mathcal{D}}_\ell$ be a labeled dataset with $\tilde{x} \sim \mathbb{N}(0, \mathbb{I}_{d-1})$ for all $(\tilde{x}, y) \in \tilde{\mathcal{D}}_\ell$. Then, for $n_\ell < d-1$, it holds with probability at least $1 - 2e^{-t^2/2}$ that the maximum average-ℓ_2-margin $\tilde{\gamma}_{avg}$ is upper bounded by*

$$\tilde{\gamma}_{avg} \leq \sqrt{d/n_\ell + 2tn_\ell^{-1/2}}$$

Simply applying this lemma concludes this part of the proof.

B.3 Proof of Lemmas B.1 and B.2

The proofs of Lemmas B.1 and B.2 consist of two parts. In the first part we lower bound the extremal singular values of the data matrix of $\tilde{\mathcal{D}}_\ell = \{(\tilde{x}, y) : (x, y) \in \mathcal{D}_\ell$, where $x = [x_1, \tilde{x}]\}$ using a classical result on random Gaussian matrices and the union bound. In the second part we use the bounds on the minimal/maximal non-zero singular value of the data matrix to obtain lower and upper bounds on the max-l_2-margin of $\tilde{\mathcal{D}}_\ell$, which concludes the proof. In the sequel we assume for convenience that $\tilde{x}$ is d-dimensional and invoke the lemmas with $d \to d-1$ in Section B.2.

Bounding the singular values of the data matrix. First, we recall bounds on the extremal singular values of a normal random matrix. Then we use the union bound to compute bounds on the extremal singular values of a fixed subset of columns of a normal random matrix. For convenience, let $\tilde{X}_u \in \mathbb{R}^{d\times n_u}$ be the data matrix of the dataset $\tilde{\mathcal{D}}_u$. Observe that by definition all entries of $\tilde{X}_u$ are drawn i.i.d. from a standard normal distribution. Recall that $\tilde{\mathcal{D}}_\ell$ consists of an arbitrary subset of size $n_\ell < n_u$ of $\tilde{\mathcal{D}}_u$, labeled using $\theta^\star$.

We use the following result on the maximal and minimal singular values of a matrix with i.i.d. standard normal distributed entries. By Corollary 5.35 of Vershynin (2010), the singular values of any i.i.d. normal random matrix $\tilde{X}_\ell \in \mathbb{R}^{d\times n_\ell}$ with $d > n_\ell$ are lower and upper bounded by

$$\sqrt{d} - \sqrt{n_\ell} - t \leq s_{\min}(\tilde{X}_\ell) \leq s_{\max}(\tilde{X}_\ell) \leq \sqrt{d} + \sqrt{n_\ell} + t \tag{28}$$

with probability greater than $1 - 2e^{-t^2/2}$. We define the set of data matrices corresponding to all possible subsets of $\tilde{\mathcal{D}}_u$ as

$$\Theta := \{\tilde{X}_\ell \in \mathbb{R}^{d\times n_\ell} : \text{ columns of } \tilde{X}_\ell \text{ are a subset of the columns of } \tilde{X}_u\} \tag{29}$$

Note that there are exactly $m = \frac{n_u!}{(n_u-n_\ell)!n_\ell!} \leq n_u^{n_\ell}$ ways in which a sampling strategy can choose a subset of n_ℓ columns from $\tilde{X}_u$. Using the union bound we get

$$\begin{aligned} P\left[\max_{\tilde{X}_\ell\in\Theta} s_{\max}(\tilde{X}_\ell) > (\sqrt{2\log n_u}+1)\sqrt{n_\ell} + \sqrt{d} + t\right] \\ \leq mP\left[s_{\max}(\tilde{X}_\ell) > \left(\sqrt{2\log n_u}+1\right)\sqrt{n_\ell} + \sqrt{d} + t\right] \end{aligned} \tag{30}$$

We can simplify the expression using Equation 28 as follows

$$
\begin{aligned}
mP\left[s_{\max}(\tilde{X}_\ell) > (\sqrt{2\log n_u}+1)\sqrt{n_\ell} + \sqrt{d} + t\right] &\leq e^{\log 2m} e^{-(\sqrt{2n_\ell \log n_u}+t)^2/2} \\
&= e^{\log(m)+\log(2)-n_\ell \log(n_u) - \sqrt{2\log n_u}\sqrt{n_\ell} - t^2/2} \\
&\leq e^{-t^2/2}
\end{aligned} \tag{31}
$$

This concludes the upper-bound on the maximum singular value of the data matrix corresponding to an arbitrary $\tilde{\mathcal{D}}_\ell$. Observe that by symmetry of the random variable, the same derivation holds for the minimal singular value as well, which concludes the first part of the proof.

Bounding the max-ℓ_2-margin of $\tilde{\mathcal{D}}_\ell$. We now use the upper and lower bounds on the extremal singular values of the data matrices to derive upper and lower bounds on the max-ℓ_2-margin of the dataset. The max-ℓ_2-margin of $\tilde{\mathcal{D}}_\ell$ is given by

$$
\tilde{\gamma} = \max_{\theta \in \mathcal{S}^{d-1}} \min_{(\tilde{x},y)\in\tilde{\mathcal{D}}_\ell} y\langle \theta, \tilde{x}\rangle. \tag{32}
$$

We can rewrite the max-ℓ_2-margin using the data matrix $\tilde{X}_\ell$.

$$
\begin{aligned}
&\tilde{\gamma} = \max_{\theta \in \mathcal{S}^{d-1}} b \\
&\text{subject to } \theta^\top \tilde{X}_\ell \geq b\mathbb{1}_{n_\ell},
\end{aligned} \tag{33}
$$

where the inequality is element-wise and $\mathbb{1}_{n_\ell}$ denotes the all ones vector. Since $n_\ell < d$ the max-ℓ_2-margin is larger than 0, i.e. there exists a $b > 0$ that satisfies the constraint. Hence, there exists a vector $v \in \mathbb{R}^{n_u}$, such that every entry is larger or equal than 1 and $\theta^\top \tilde{X}_\ell \geq bv$. By the Eckart-Young-Mirsky theorem, the optimal θ must lie in the space spanned by the singular vectors corresponding to the non-zero singular values of $\tilde{X}_\ell$. Hence, as $\|\theta\|_2 = 1$, we have that $s_{\min}(\tilde{X}_\ell) \leq \|\tilde{\gamma} v\|_2 \leq s_{\max}(\tilde{X}_\ell)$. Since all entries of v are larger or equal to 1, we find that

$$
\tilde{\gamma} \leq \frac{s_{\max}(\tilde{X}_\ell)}{\sqrt{n_\ell}}, \tag{34}
$$

which, using the upper bound on the singular values derived earlier, yields the upper bound on the max-ℓ_2-margin. For the minimal margin, note that $\tilde{\gamma}\|v\|_2 = s_{\min}(\tilde{X}_\ell)$ implies a minimal margin of

$$
\tilde{\gamma} \geq \frac{s_{\min}(\tilde{X}_\ell)}{\sqrt{n_\ell}} \tag{35}
$$

Plugging in the bounds on the minimal singular values of the data matrix $\tilde{\mathcal{D}}_\ell$ concludes the proof of the lemma.

B.4 Proof of Lemma B.3

The maximum average-ℓ_2-margin is defined as

$$
\begin{aligned}
\tilde{\gamma}_{\text{avg}} &= \max_{\theta \in \mathcal{S}^{d-1}} \frac{1}{n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell} y\langle \theta, x\rangle \\
&= \max_{\theta \in \mathcal{S}^{d-1}} \frac{\theta_1}{n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell} yx_1 + ... + \frac{\theta_d}{n_\ell} \sum_{(x,y)\in\mathcal{D}_\ell} yx_d \\
&= \frac{1}{\sqrt{n_\ell}} \max_{\theta \in \mathcal{S}^{d-1}} \theta^\top X,
\end{aligned} \tag{36}
$$

where in the last equation X is a d-dimensional vector distributed according to a standard Gaussian (note that we consider the samples in $\mathcal{D}_\ell$ to be random variables). By Cauchy-Schwarz, the maximum is found by setting $\theta = \frac{X}{\|X\|_2}$. Using

Chernoff's bound, we find that $\|X\|_2 < \sqrt{d(1+t)}$ with probability larger than $1 - 2e^{-dt^2/8}$. Multiplying by $1/\sqrt{n_\ell}$ yields the lemma.

B.5 Proof of Lemma A.9

Define $n_q := (1-\rho)n_\ell$ to be the number of queries made with uncertainty sampling. Recall that $d_q^\star$ is defined as the distance to the decision boundary determined by $\theta^\star$ of the n_q^{th} closest sample from the unlabeled dataset. Note that the unlabeled dataset is drawn i.i.d. from the mixture of truncated Gaussians distribution described in Section 3.2. Let x_q be the n_q^{th} closest sample to the decision boundary determined by $\theta^\star$. Let Φ_{tr} denote the cumulative distribution function of a Gaussian with mean μ and standard deviation σ truncated to the interval $(0, \infty)$:

$$\Phi_{tr}(t) = \frac{\Phi\left(\frac{t-\mu}{\sigma}\right) - \Phi\left(-\frac{\mu}{\sigma}\right)}{1 - \Phi\left(-\frac{\mu}{\sigma}\right)} \tag{37}$$

We find that for some $t > 0$ the following holds:

$$P[d_q^\star < t] = 1 - \sum_{i=1}^{n_q - 1} \binom{n_u}{i} \Phi_{tr}(t)^i (1 - \Phi_{tr}(t)^{n_u - i}) \tag{38}$$

Using Hoeffding's inequality, we can bound the probability as

$$P[d_q^\star < t] \geq 1 - e^{-2n_u\left(\Phi_{tr}(t) - \frac{n_q}{n_u}\right)^2}$$

After the change of variable $\tilde{t} = \Phi_{tr}^{-1}\left(\frac{t}{\sqrt{2n_u}} + \frac{n_q}{n_u}\right)$ we arrive at:

$$P[d_q^\star < \Phi_{tr}^{-1}(\tilde{t}/\sqrt{2n_u} + n_q/n_u)] \geq 1 - e^{-\tilde{t}^2}$$

Now plugging in the definition of the inverse of the CDF of the positive-sided truncated Gaussian yields that, with probability of at least $1 - e^{-t^2}$, the following holds:

$$d_q^\star < \sigma\left(\Phi^{-1}\left(\left(\frac{t}{\sqrt{2n_u}} + \frac{n_q}{n_u}\right)\left(1 - \Phi\left(-\frac{\mu}{\sigma}\right)\right) + \Phi\left(-\frac{\mu}{\sigma}\right)\right)\right) + \mu \tag{39}$$

Observe that the right-hand side of Equation 39 is monotonically increasing in $\frac{1}{n_u}$ and $\frac{n_q}{n_u}$. From the assumptions required for Lemma A.9 we have that $\frac{n_q}{n_u} < \frac{n_\ell}{n_u} < 10^{-3}$ and $\frac{1}{n_u} < 10^{-5}$. Fixing t such that the probability is 0.99, we can further write the upper bound as follows:

$$d_q^\star < \sigma\left(\Phi^{-1}\left(c\left(1 - \Phi\left(-\frac{\mu}{\sigma}\right)\right) + \Phi\left(-\frac{\mu}{\sigma}\right)\right)\right) + \mu$$

where c is a small positive constant that depends on t and which can be computed numerically for fixed t. For convenience, we define $\beta := \frac{\mu}{\sigma}$. Then, using the formula for μ_{tr} from Equation 2 we arrive at:

$$\frac{d_q^\star}{\mu_{tr}} < (1 - \Phi(-\beta)) \frac{\left(\Phi^{-1}\left(c\left(1 - \Phi(-\beta)\right) + \Phi(-\beta)\right)\right) + \beta}{\beta\left(1 - \Phi(-\beta)\right) + \phi(-\beta)} = \frac{\beta + \Phi^{-1}\left(c\left(1 - \Phi(-\beta)\right) + \Phi(-\beta)\right)}{\beta + \frac{\phi(-\beta)}{1 - \Phi(-\beta)}}$$

Taking the derivative and an algebraic exercise shows that the right-hand side is an increasing function of β. Hence, we can plug in the numerical value of c and the condition $\beta = \mu/\sigma < 2$ to find the desired upper bound and conclude the proof.

B.6 Proof of Lemma A.11

Define $n_q := (1-\rho)n_\ell$ to be the number of queries made with uncertainty sampling. Recall that $\hat{d}_q$ is defined as the distance to the decision boundary determined by $\hat{\theta}_{seed}$ of the n_q^{th} closest sample from the unlabeled dataset. Note that the unlabeled dataset is drawn i.i.d. from the mixture of truncated Gaussians distribution described in Section 3.2. Let x_q be the n_q^{th} closest sample to the decision boundary determined by $\hat{\theta}_{seed}$ and define $p_t = P[|\langle \hat{\theta}_{seed}, x\rangle| < t]$ for a sample x drawn from the mixture of truncated Gaussians distribution and a constant $t > 0$. Clearly

$$P[\hat{d}_q < t] = 1 - \sum_{i=1}^{n_q-1} \binom{n_u}{i} p_t^i (1-p_t)^{n_u-i}$$

Using Hoeffding's inequality, we can bound the probability as

$$P[\hat{d}_q < t] \geq 1 - e^{-2n_u(p_t - n_q/n_u)^2}$$

Now using the definition of the mixture of truncated Gaussians distribution and recalling that $\hat{\theta}_{seed} = [1, \alpha_{\text{seed}}\tilde{\theta}_{seed}]$, we find that

$$p_t = P\left[\left|x_1 + \alpha_{\text{seed}}\langle \tilde{\theta}_{seed}, \tilde{x}\rangle\right| < t\sqrt{1+\alpha_{\text{seed}}^2}\right]$$

We note that $\langle \tilde{\theta}_{seed}, \tilde{x}\rangle$ is distributed according to a standard normal and x_1 according to a mixture of univariate Gaussians truncated at 0 with mean $y\mu$ and variance σ^2. Denote by Φ_{tr} the cumulative distribution function of the truncated Gaussian distribution. Then $x_1 < t$ with probability $\Phi_{tr}(t)$. If $\sigma > 1$, then $P[x_1 < t] < P[\langle \tilde{\theta}_{seed}, \tilde{x}\rangle < t]$ for all $t > 0$. In that case, we find that

$$p_t \leq P\left[|x_1| < t\right]$$

Hence, we can take $p = P\left[|x_1| < t\right]$ as an upper bound and use the derivation in the proof of Lemma A.9 from Equation 38 onwards.

B.7 Proof of Lemmas A.8, A.10 and A.12

We prove Lemma A.10, since the proofs of Lemmas A.8 and A.12 are almost identical. For the proof of Lemma A.12, the only difference comes from swapping $d_q^\star$ for $\hat{d}_q$ and using the respective bound on $\tilde{\gamma}$. In the case of uniform sampling, we do not exclude the seed set and consider instead of $d_q^\star$, the quantities $\mu_{tr} + t\sigma_{tr}$ and $\mu_{tr} + tn_\ell^{-1/2}\sigma_{tr}$ as bounds on the first coordinate of a sample and mean of the first coordinate over the dataset respectively. We now prove Lemma A.10.

Recall that, by definition, all support points have the same ℓ_2-distance to the decision boundary of the max-ℓ_2-margin classifier, denoted by γ. By a similar argument to the one in Section B.1, we have that the max-ℓ_2-margin γ is lower bounded by

$$\gamma \geq \sqrt{\left(\frac{1}{|D_s|}\sum_{(x,y)\in D_s} yx_1\right)^2 + \tilde{\gamma}^2} \tag{40}$$

where $D_s \subseteq \mathcal{D}_\ell$ is the subset containing all support points. Moreover, by Lemma A.6, the normalized max-ℓ_2-margin classifier can be written as follows:

$$\hat{\theta} = \frac{1}{\sqrt{\hat{\theta}_1^2 + \tilde{\gamma}^2}}\left[\hat{\theta}_1, \tilde{\gamma}\tilde{\theta}\right]$$

where $\|\tilde{\theta}\|_2 = 1$ and we use the notation $\hat{\theta}_1 = \frac{1}{|D_s|}\sum_{(x,y)\in D_s} yx_1$. Therefore, it holds that $\gamma = y\langle\hat{\theta}, x\rangle \geq \sqrt{\hat{\theta}_1^2 + \tilde{\gamma}^2}$ for all $(x, y) \in D_s$. After rewriting it follows that $\mathcal{D}_\ell \setminus \mathcal{D}_{seed} \subseteq D_s$, if the following condition is satisfied:

$$\max_{(x,y)\in\mathcal{D}_\ell\setminus\mathcal{D}_{seed}} \tilde{\gamma} y\langle\tilde{\theta}, \tilde{x}\rangle \leq y\langle\hat{\theta}, x\rangle\sqrt{\hat{\theta}_1^2 + \tilde{\gamma}^2} - y\hat{\theta}_1 x_1 \tag{41}$$

We now take steps to give a more restrictive sufficient condition that implies the one in Equation 41, and hence, also implies that $\mathcal{D}_\ell \setminus \mathcal{D}_{seed} \subseteq D_s$. For an arbitrary pair $(x, y) \in \mathcal{D}_\ell \setminus \mathcal{D}_{seed}$ the following chain of inequalities holds:

$$\begin{aligned} y\langle\hat{\theta}, x\rangle\sqrt{\hat{\theta}_1^2 + \tilde{\gamma}^2} - y\hat{\theta}_1 x_1 &\overset{(i)}{\geq} \gamma^2 - y\hat{\theta}_1 x_1 \\ &\overset{(ii)}{\geq} \gamma^2 - \hat{\theta}_1 d_q^\star \end{aligned}$$

where inequality (i) follows from Equation 40 and (ii) holds by the definition of $d_q^\star$. Now, note that

$$\gamma^2 - \hat{\theta}_1 d_q^\star + y\hat{\theta}_1 x_1 \geq \min\{\gamma^2, \gamma^2 - (d_q^\star)^2\} = \gamma^2 - (d_q^\star)^2$$

Hence, for all samples in $\mathcal{D}_\ell \setminus \mathcal{D}_{seed}$ to be support points it is sufficient that the following holds:

$$\max_{(x,y)\in\mathcal{D}_\ell\setminus\mathcal{D}_{seed}} \tilde{\gamma} y\langle\tilde{\theta}, \tilde{x}\rangle \leq \tilde{\gamma}^2 - (d_q^\star)^2 \tag{42}$$

Since $\tilde{x}$ is distributed according to a multivariate standard Gaussian and $\|\tilde{\theta}\|_2 = 1$, we know that

$$\max_{(x,y)\in\mathcal{D}_\ell\setminus\mathcal{D}_{seed}} y\langle\tilde{\theta}, \tilde{x}\rangle \leq \sqrt{2\log n_\ell} + t \tag{43}$$

with probability at least $1 - 2e^{-t^2/2}$. By combining Equations 43 and 42, we get that $\mathcal{D}_\ell \setminus \mathcal{D}_{seed} \subseteq D_s$ with probability at least $1 - 2e^{-0.5\left(\tilde{\gamma} - (d_q^\star)^2\tilde{\gamma}^{-1} - \sqrt{2\log n_\ell}\right)^2}$.

Recall that $d_q^\star$ is independent of d and small by Lemma A.9. Moreover, by Lemma B.1 we have that $\tilde{\gamma} > \sqrt{\frac{d}{n_\ell}} - 1 - t$. Hence for a small constant $0 < c$, we have that $\mathcal{D}_\ell \setminus \mathcal{D}_{seed} \subset D_s$, with a probability greater than $1 - 2e^{-0.5\left(\sqrt{d/n_\ell} - \sqrt{2\log n_\ell} - c\right)^2}$.

B.8 Proof of Corollary A.3

In Theorem A.2, let us denote the numerators of the expressions of α_{oracle} and α_{unif} by γ_{oracle} and γ_{unif}, respectively. Similarly, we use the notation M_{oracle}, M_{unif} for the denominators of α_{oracle} and α_{unif}, respectively.

Since the function $\Psi_{\mu,\sigma}$ defined in Equation 4 is monotonic in α, it follows from Theorem A.2 with high probability that oracle uncertainty sampling performs worse than passive learning if $\alpha_{\text{oracle}} = \frac{\gamma_{oracle}}{M_{oracle}} > \frac{\gamma_{unif}}{M_{unif}} = \alpha_{\text{unif}}$. We observe that

$$\frac{\gamma_{oracle}}{M_{oracle}} > \frac{\gamma_{unif}}{M_{unif}} \iff \frac{\gamma_{oracle}}{\gamma_{unif}} > \frac{M_{oracle}}{M_{unif}}$$

For an $\eta \in (0, 1)$ and using the expressions for γ_{oracle} and γ_{unif} from Theorem A.2, we have that

$$\frac{\gamma_{oracle}}{\gamma_{unif}} > \eta \iff d > n_\ell \frac{\left(1 + t + \eta\sqrt{2tn_\ell^{-1/2}}\right)^2}{(1-\eta)^2}$$

Moreover, for $\eta > \frac{M_{oracle}}{M_{unif}}$, we find that

$$\eta > \frac{M_{oracle}}{M_{unif}} \iff \mu_{tr} > \frac{\rho(\mu_{tr} + t\sigma_{tr}n_{\text{seed}}^{-1/2}) + 6.059(1-\rho)\cdot 10^{-2}\mu_{tr} + t\sigma_{tr}n_\ell^{-1/2}\eta}{\eta}$$

Hence, by plugging in $\eta = 0.5$, we have that $\frac{\gamma_{oracle}}{M_{oracle}} > \frac{\gamma_{unif}}{M_{unif}}$ if

$$\frac{d}{n_\ell} > 4\left(1 + t + \sqrt{t(4n_\ell)^{-1/2}}\right)^2 \tag{44}$$

$$\mu_{tr} > 2\rho(\mu_{tr} + t\sigma_{tr}n_{\text{seed}}^{-1/2}) + 2 \cdot 6.059(1-\rho)\cdot 10^{-2}\mu_{tr} + 2t\sigma_{tr}n_\ell^{-1/2} \tag{45}$$

Using $\rho < 0.5$ we arrive at

$$\mu_{tr} > \frac{t\sigma_{tr}(1 + 2\rho^{1/2})n_\ell^{-1/2}}{1 - 2\rho}$$

We now solve for ρ and find

$$\sqrt{\rho} < q\left(\sqrt{1 + \frac{1}{2q^2} - \frac{1}{q}} - 1\right)$$

with $q = \frac{t\sigma_{tr}}{2\mu_{tr}\sqrt{n_\ell}}$. Using the fact that $\sqrt{\rho} < \frac{1}{\sqrt{2}}$ and ignoring negligible terms, we get the following bound on ρ

$$\rho < \frac{1}{2} - \frac{1+\sqrt{2}}{2}\frac{t\sigma_{tr}}{\sqrt{n_\ell}\mu_{tr}}$$

which concludes the proof.

B.9 Proof of Corollary A.5

Similar to Section B.8, let us denote in Theorem A.4 the numerators of the expressions of α_{uncert} and α_{unif} by γ_{uncert} and γ_{unif}, respectively. Similarly, we use the notation M_{uncert}, M_{unif} for the denominators of α_{uncert} and α_{unif}, respectively.

By Theorem A.4 uncertainty sampling leads to a classifier with a lower test error than uniform sampling, if $\alpha_{\text{uncert}} = \frac{\gamma_{uncert}}{M_{uncert}} \geq \frac{\gamma_{unif}}{M_{unif}} = \alpha_{\text{unif}}$. Similar to the proof of Corollary A.3, we find that

$$\frac{\gamma_{uncert}}{M_{uncert}} \geq \frac{\gamma_{unif}}{M_{unif}} \iff \frac{\gamma_{uncert}}{\gamma_{unif}} \geq \frac{M_{uncert}}{M_{unif}}$$

Let $\eta \in (0,1)$. Then it holds that

$$\begin{aligned}
\frac{\gamma_{uncert}}{\gamma_{unif}} \geq \eta &\iff \sqrt{d/n_\ell} - \sqrt{2\log n_u} - 1 - t \geq \eta\left(\sqrt{d/n_\ell + 2tn_\ell^{-1/2}}\right) \\
&\Rightarrow \sqrt{d/n_\ell} - \sqrt{2\log n_u} - 1 - t \geq \eta\left(\sqrt{d/n_\ell} + \sqrt{2tn_\ell^{-1/2}}\right) \\
&\iff d/n_\ell > \frac{\sqrt{2\log n_u} + 1 + t + \eta\sqrt{2tn_\ell^{-1/2}}}{(1-\eta)^2}
\end{aligned}$$

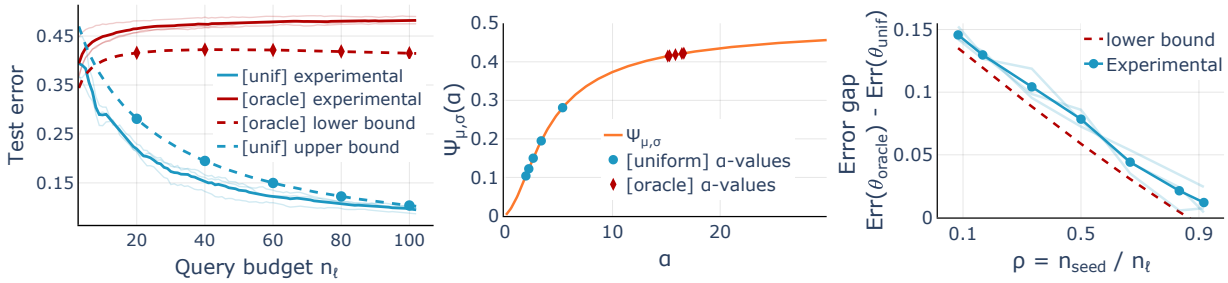
Choosing $\eta = 0.5$ yields

$$d/n_\ell > 4\left(\sqrt{2\log n_u} + 1 + t + \sqrt{t(4n_\ell)^{-1/2}}\right)$$

Similarly, we have that $\frac{M_{uncert}}{M_{unif}} < \eta$. Plugging in the expressions given in Theorem A.4, we find that

$$\rho C_{seed} + (1-\rho)\left(6.059\cdot 10^{-2}\mu_{tr} + \left(\frac{2\log n_u}{C_{seed}}\right)^{1/2}\left(\frac{d}{\rho n_\ell} + \frac{2\sigma_{tr}t}{\rho n_\ell}\right)^{1/4} + t\right) < \eta\left(\mu_{tr} - \frac{t\sigma_{tr}}{n_\ell^{1/2}}\right)$$

Recalling the bound on C_{seed} in Equation 19, we find that if



(a) Test error gap vs query budget

(b) $\Psi_{\mu,\sigma}(\alpha)$

(c) Test error gap vs ρ

Figure 6: For large d/n_ℓ the bounds of Theorem A.2 are close to tight. (a) The bounds in Theorem A.2 (dashed lines) show that U-AL leads to lower test error compared to uniform sampling. The lighter color lines correspond to one of 3 runs with different draws of the seed set, while the solid line indicates their mean. (b) The function $\Psi_{\mu,\sigma}(\alpha)$ is monotonically increasing in α. The markers show the $(\alpha, \Psi_{\mu,\sigma}(\alpha))$ values corresponding to the query budgets indicated in figure (a). (c) By the bound in Theorem 3.2 (dashed line), the error gap between U-AL and PL decreases as the proportion of seed samples grows.

$$\mu_{tr} \geq \left(\frac{d}{\rho n_\ell} + \frac{2\sigma_{tr} t}{\rho n_\ell}\right)^{1/6} (2\log n_u)^{1/3} + \frac{t\sigma_{tr}}{(\rho n_\ell)^{1/2}} \tag{46}$$

then the following condition on ρ suffices in order to guarantee that $\alpha_{\text{unif}} < \alpha_{\text{uncert}}$:

$$\rho < \frac{0.878\mu_{tr} - t\sigma_{tr}(\rho n_\ell)^{-1/2} - \left(\frac{d}{\rho n_\ell} + \frac{\sigma_{tr} t}{\rho n_\ell}\right)^{1/6} (2\log n_u)^{1/3} - t}{C_{seed}}$$

C Synthetic experiments on the mixture of truncated Gaussians distribution

In this section, we give the experimental details to the synthetic experiments in Figures 2a and 2b. Further, we show empirically that for large d/n_ℓ the theoretical bounds closely predict the experimental values. Lastly, we further empirically discuss the dependency on the distributional parameters σ and μ of the truncated Gaussian mixture model for **empirical** uncertainty sampling.

C.1 Experimental details to Figures 2a and 2b

In both Figures 2a and 2b, we plot the theoretical upper and lower bounds of Theorem A.2 with $n_u = 10^5$, $t = 3$ and compute $\Psi_{\mu,\sigma}$ by integrating using Sklearns function "dblquad".

In Figure 2a, we set $d = 1000$, $\mu = 2$, $\sigma = 2$, $n_{\text{seed}} = 10$ and vary n_ℓ from n_{seed} to 1000. On the other hand, in Figure 2b we set $d = 1000$, $\sigma = 2$ and vary the mean-parameter μ in $\{1, 2, 3\}$ and the seed set size n_{seed} in $[1, ..., n_\ell]$.

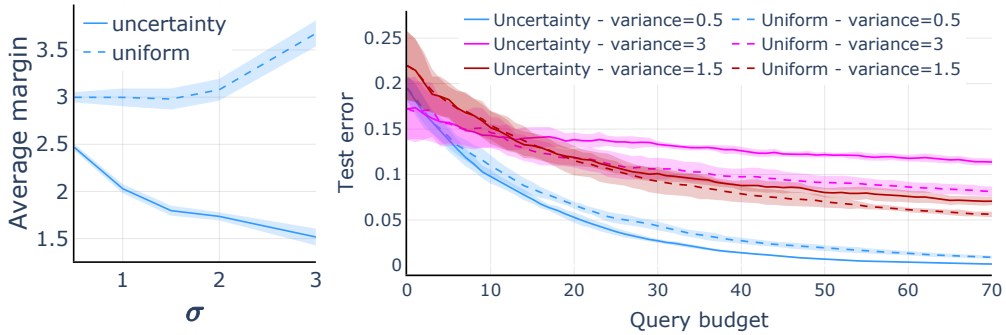
C.2 Verifying the bounds in Theorem A.2 on synthetic data

We now experimentally confirm the bounds in Theorem A.2. Recall that for large d/n_ℓ, the bounds on $\tilde{\gamma}$ of Lemma B.1 are tight. Therefore we consider two settings where d/n_ℓ is large.

First, in Figure 2a, we set $d = 3k$, $n_u = 10^5$, $\sigma = 2$ and $\mu = 3$. Then we vary n_ℓ from n_{seed} to 100. We plot the results of 3 independent experiments for each setting along with the theoretical lower and upper bounds given in Theorem A.2. Observe that the theoretical bounds closely predict both the test error of passive learning as well as the test error of oracle U-AL.

Further, for completeness, in Figure 6b we also plot the function $\Psi_{\mu,\sigma}$ with corresponding α-values from the setting in Figure 2a. Observe that for small α the function $\Psi_{\mu,\sigma}$ increases fast.

Lastly, in Figure 6c, we set $d = 10k$,$n_u = 10^5$, $\mu = 0$, $\sigma = 3$, $n_\ell = 60$ and vary n_{seed} from 5 to 55. Observe that the theoretical bound closely predicts the test error gap. Moreover, observe that the test error gap monotonically decreases in ρ both experimentally and according to the theoretical bound.



(a) Average margin in σ

(b) Test error for varying σ

Figure 7: (a) Uncertainty sampling collects labeled sets with a smaller average margin in the signal component as σ increases. (b) As the average margin of a set acquired via uncertainty sampling is smaller for increasing σ, the test error deteriorates, as predicted by Lemma A.7. The test error of uncertainty sampling can even be larger than that of uniform sampling, for large enough σ. We use $d = 1000$, $\mu = 3$ and $n_u = 10^5$ for all the experiments in this figure. The shaded areas indicate one standard deviation bands around the mean error, computed over 5 random draws of the seed set.

Logistic regression implementation. In all synthetic experiments, we use the SGDClassifier of the Scikit-learn library (Pedregosa et al., 2011) with the following settings: we set the learning rate to be a constant of 10^{-4} and train for at least 10^4 epochs without regularization. Moreover, we set the tolerance parameter to 10^{-5} and the maximum number of epochs to 10^6. In all experiments, we consider regular uncertainty sampling as defined in Algorithm 1.

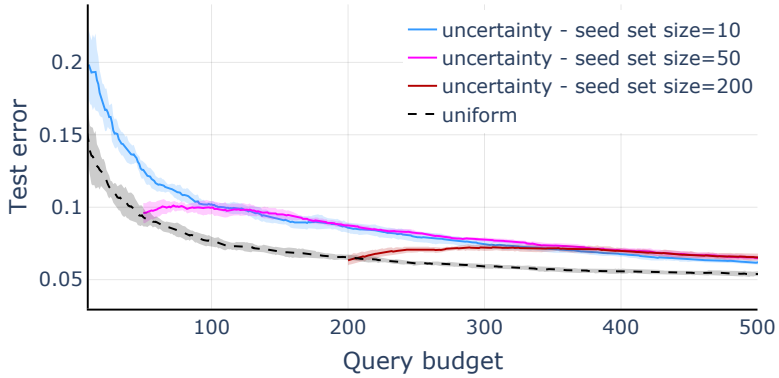


Figure 8: We set $d = 1000$, $\mu = 3$ and $n_u = 10^5$. The shaded areas indicate one standard deviation bands around the mean error, computed over 5 random draws of the seed set. Observe that for increasing seed size the gap between uncertainty and uniform sampling closes, but does not vanish. Note that we study the high-dimensional regime and hence only consider seed sizes up to $d/4$. Therefore, seed size larger than d may fully close the test error gap between uniform and standard sampling.

C.3 Dependence on the standard deviation σ and seed size n_{seed} for regular uncertainty sampling

For completeness, we illustrate the dependence on the variance and the seed size of regular uncertainty sampling with following experiments. To simulate realistic settings, we set $d = 1000$, $n_u = 10^5$, $n_{\text{seed}} = 10$, $\sigma = 3$ and $\mu = 3$.

First, we perform a set of experiments to analyze the dependence on the variance σ and to also confirm to main intuition empirically. We compute the average distance to the decision boundary of the ground truth $\theta^\star$ of a labeled set acquired through uncertainty and uniform sampling. Indeed, in Figure 7a we see that the average margin of uncertainty sampling decreases with increasing σ. Moreover, we note that uncertainty sampling indeed queries points close to the optimal decision boundary. In Figure 7b we observe that, as predicted by Lemma A.7, the decrease of the average margin gap is directly correlated with an increase of the error gap between uncertainty and uniform sampling. Hence, our main intuition is also here empirically verifiable: uncertainty sampling queries points relatively close to the ground truth, which causes in high dimensions the max-ℓ_2-margin classifier to rely more on the non-signal components to classify the training data.

Secondly, we perform a set of experiments to an analyze the dependence on the seed size n_{seed}. In Figure 8, we see that the test error gap between uncertainty and uniform sampling closes slowly for an increasing seed size. However, we note that the gap remains non-zero for all seed sizes up to $d/4$.

Dataset name	d	Training set size	Test set size	Majority/minority ratio	Linear classif. training error
a9a	123	39074	9768	3.17	0.1789
vehicleNorm	100	78823	19705	1.00	0.1415
nomao	118	27572	6893	2.50	0.0531
santander	200	160000	40000	8.95	0.2188
webdata_wXa	123	29580	7394	3.16	0.1813
sylva_prior	108	11516	2879	15.24	0.0011
real-sim	20958	57848	14461	2.25	0.0027
riccardo	4296	16000	4000	3.00	0.0007
guillermo	4296	16000	4000	1.49	0.2536
jasmine	144	2388	596	1.00	0.1867
madeline	259	2512	628	1.01	0.3405
philippine	308	4666	1166	1.00	0.2445
christine	1636	4335	1083	1.00	0.1408
musk	166	5279	1319	5.48	0.0438
epsilon	2000	48000	12000	1.00	0.0947

Table 1: Some characteristics of the uncurated datasets considered in our experimental study.

D Experiment details for tabular data

D.1 Datasets

To assess how suitable uncertainty sampling is for high-dimensional data, we conduct experiments on a wide variety of real-world datasets.

We select datasets from OpenML (Vanschoren et al., 2013) and from the UCI data repository (Dua and Graff, 2017) according to a number of criteria. In particular, we focus on datasets for binary classification that are high-dimensional ($d > 100$) and which have enough samples that can serve as the unlabeled set ($n_u > \max(1000, 2d)$). We do not consider text or image datasets where the features are sequences of characters or raw pixels as estimators other than linear models are better suited for these data modalities (e.g. CNNs, transformers etc). Instead we want to analyze uncertainty sampling in a simple setting and thus focus on datasets that are (approximately) linearly separable. Moreover, we discard datasets that have missing values. Finally, we are left with 15 datasets that cover a broad range of applications from finance and ecology to chemistry and histology. We provide more details about the selected datasets in Appendix D.1.

To disentangle the effect of high-dimensionality from other factors such as class imbalance, we subsample uniformly at random the examples of the majority class, in order to balance the two classes. In addition, to ensure that the data is noiseless, we fit a linear classifier on the entire dataset, and remove the samples that are not interpolated by the linear estimator. This noiseless setting is advantageous for active learning, since we are guaranteed to not waste the limited labeling budget on noisy samples. However, as we show later, even in this favorable scenario, the performance of uncertainty sampling suffers in high-dimensions. For completeness, we also compare uncertainty sampling and passive learning on the original, uncurated datasets in Appendix E.1 and observe similar trends as in this section.

More dataset statistics. Table 1 summarizes some important characteristics of the datasets. The datasets span a wide range of applications (e.g. ecology, finance, chemistry, histology etc). All datasets are high-dimensional ($d \geq 100$) and have sufficiently many training samples that will serve as the unlabeled set. The test error is computed on a holdout set, whose size we report in Table 1. We also present the class-imbalance of the original, uncurated datasets and the training error of a linear classifier trained on the entire dataset, which indicates the degree of linear separability of the data.

D.2 Methodology

We split each dataset in a test set and a training set. The covariates of the training samples constitute the unlabeled set. We assume that the labels are known for a small seed set of size $n_{\text{seed}} = 6$ (see Appendix E.7 for experiments with larger seed sets). For each experiment and each dataset, we repeat the draw of the seed set several times (10 or 100, depending on the experiment).

For illustration purposes, we set the labeling budget to be equal to a quarter of the number of dimensions.[8] We query one point at a time and select the sample whose label we want to acquire either via uniform sampling (i.e. passive learning) or using uncertainty sampling (i.e. active learning).

[8]Since the real-sim dataset has over 20,000 features, we set a labeling budget lower than $d/4$, namely of only 3,000 queries, for computational reasons.

We use L-BFGS (Liu and Nocedal, 1989) to train linear classifiers by minimizing the logistic loss on the labeled dataset. In Appendix E.6 we show that the same high-dimensional phenomenon occurs when using ℓ_1- or ℓ_2-regularized classifiers.

E Additional experiments on tabular data

E.1 Experiments on uncurated data

For completeness, in this section we provide experiments on the original, uncurated datasets. We distinguish two scenarios: 1) balanced data, but not necessarily linearly separable; and 2) possibly imbalanced and not linearly separable data. In both cases, we use the same methodology described in Section 4 to plot the probability (over draws of the seed set) that the error with PL is lower than with U-AL and the losses/gains of U-AL compared to PL.

Balanced, but non-linearly separable data. As indicated in Appendix D.1, not all datasets are originally linearly separable. For clarity, in the experiments in the main text we curate the data such that a linear classifier can achieve vanishing training error. This is provides a clean test bed for comparing uncertainty and uniform sampling in high-dimensions.

In Figure 9 we keep the datasets class-balanced, but allow them to be potentially not linearly separable. We observe similar trends as the ones illustrated in Figure 3 for the noiseless versions of the datasets.

Imbalanced and non-linearly separable data. Uncertainty sampling brings about surprising benefits when applied on high-dimensional imbalanced data. In particular, Figure 10 shows that for a broad range of query budgets uncertainty sampling leads to better test error than uniform sampling. For these experiments we did not alter the original datasets in any way, and kept all the training samples.

These results reveal a perhaps unexpected phenomenon. When the unlabeled data is imbalanced (see Appendix D.1 for the exact imbalance ratio of each dataset), uncertainty sampling tends to achieve better predictive performance compared to passive learning. This phenomenon has also been previously observed by Ertekin et al. (2007).

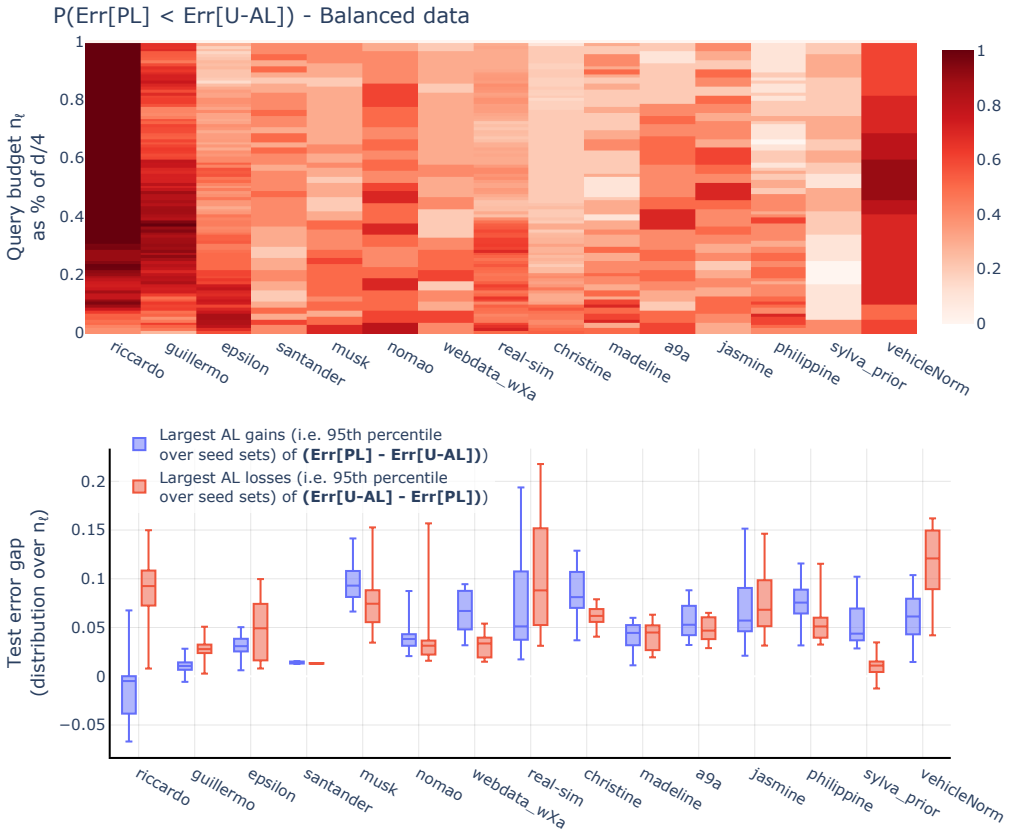


Figure 9: **Top:** The probability that the test error is lower with PL than with U-AL, over 10 draws of the seed set. Data is class-balanced, but potentially not linearly separable. **Bottom:** For the range of budgets where U-AL does poorly with high probability, its sporadic gains over PL are generally similar or lower than the losses it can incur in terms of increased test error. Data is class-balanced, but potentially not linearly separable.

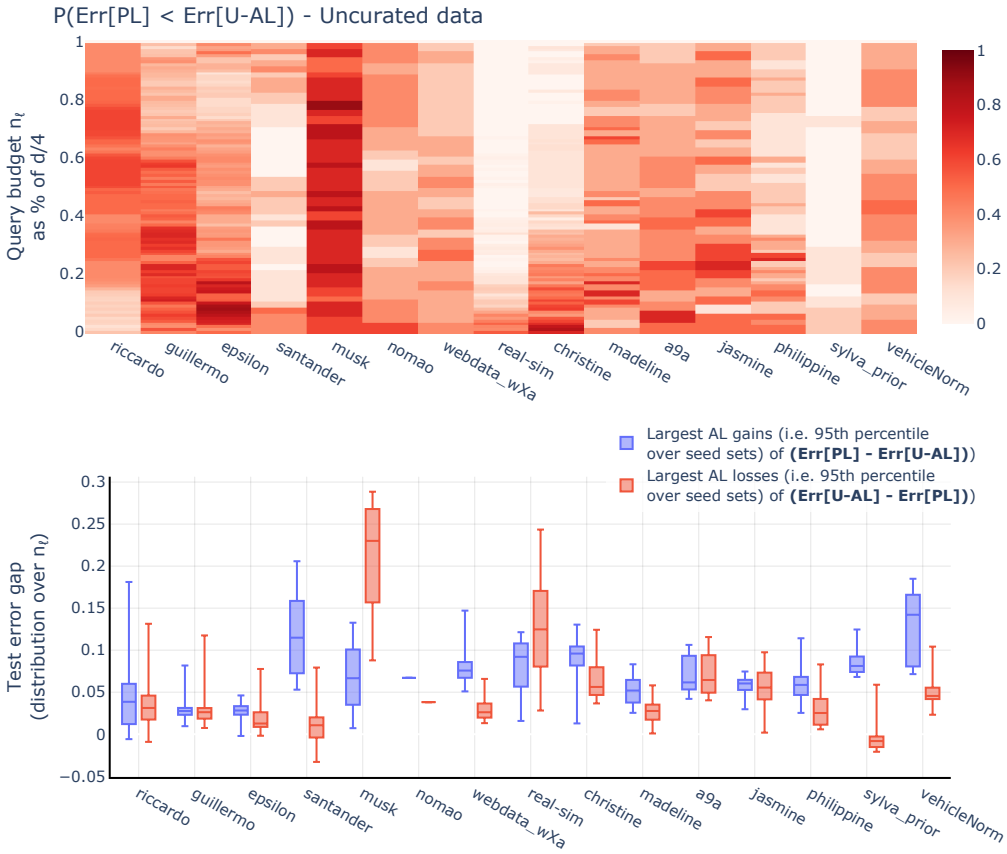


Figure 10: **Top:** The probability that the test error is lower with PL than with U-AL, over 10 draws of the seed set. Data is potentially class-imbalanced and not linearly separable. **Bottom:** For the range of budgets where U-AL does poorly with high probability, its sporadic gains over passive learning are generally similar or lower than the losses it can incur in terms of increased test error. Data is potentially class-imbalanced and not linearly separable.

E.2 Test error at different query budgets – more datasets

We compare the test error of uniform and uncertainty sampling, similar to Figure 1, but for more real-world datasets. For uncertainty sampling, we use both oracle uncertainty and the uncertainty of $f(\cdot; \hat{\theta})$ as shown in Algorithm 1. Figure 11 show that oracle uncertainty sampling consistently leads to larger test error compared to passive learning on all datasets. In addition, using the uncertainty determined by the max-ℓ_2-margin classifier also leads to worse prediction performance, in particular on the high-dimensional datasets and for small query budgets. For illustration and computational purposes, we limit the query budget to $\min(3000, d/4)$.

E.3 Uniform versus oracle uncertainty sampling

In this section we provide the counterpart of Figure 3, but now we use an oracle uncertainty estimate for the active learning algorithm. Recall that for oracle U-AL we first train a classifier on the entire labeled dataset (this estimator will act as a stand-in for the Bayes optimal predictor). Then we use the predictive uncertainty of this approximation of the Bayes optimal classifier to select points to query.

Figure 12 reveals that the gap between uncertainty and uniform sampling is even more significant when using the oracle U-AL, which is in line with the intuition provided in Section 3. Oracle uncertainty sampling will select samples close to the Bayes optimal decision boundary (i.e. the yellow points in Figure 2c). Hence, the decision boundary of the classifier trained on the labeled set collected with active learning will be tilted compared to the optimal predictor, as long as the query budget is significantly smaller than the dimensionality.

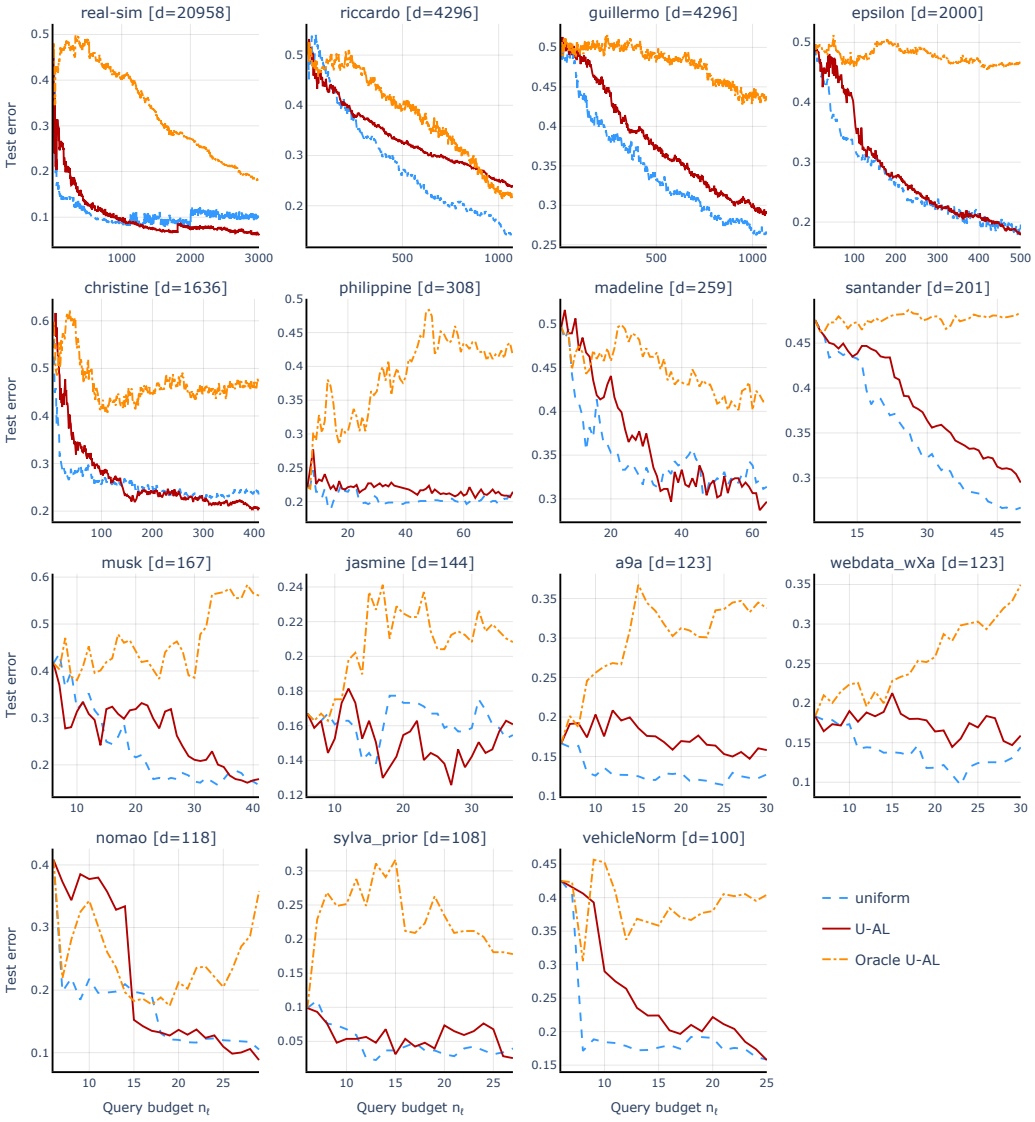


Figure 11: The test error using uncertainty sampling (with or without an oracle uncertainty estimate) is often higher than what is achieved with uniform sampling, for all the datasets that we consider.

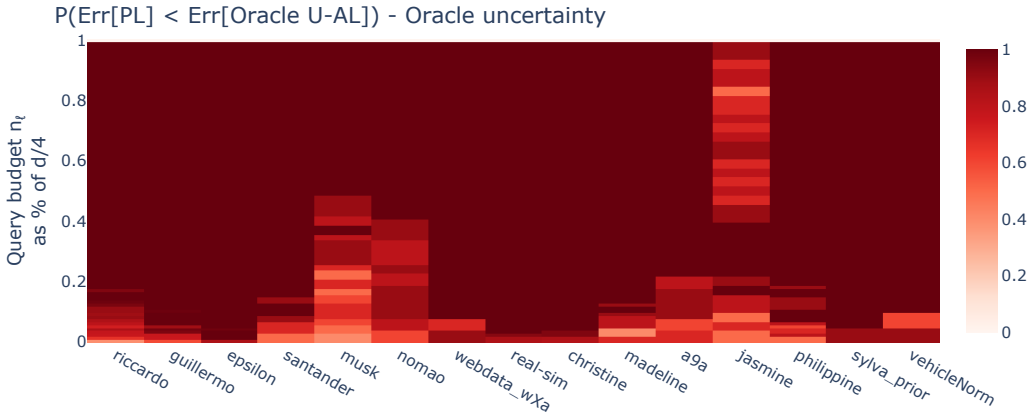


Figure 12: The probability that the test error is lower with PL than with **oracle U-AL**, over 10 draws of the seed set. Oracle U-AL performs consistently worse than PL (the dark red regions).

E.4 Another perspective on Figure 3

In Figure 3-Top we provide an overview of the gap that exists between uncertainty and uniform sampling in high-dimensions. Here, we provide a more detailed perspective of the same evaluation metric. Each panel in Figure 13 corresponds to one column in Figure 3. The horizontal dashed line indicates the 50% threshold at which the event that uncertainty sampling performs better is equally likely to its complement. Notice that in all figures the solid line starts at 0, since before any queries are made, both uniform sampling and uncertainty sampling yield the same test error, namely the error of the max-ℓ_2-margin classifier trained on the seed set.

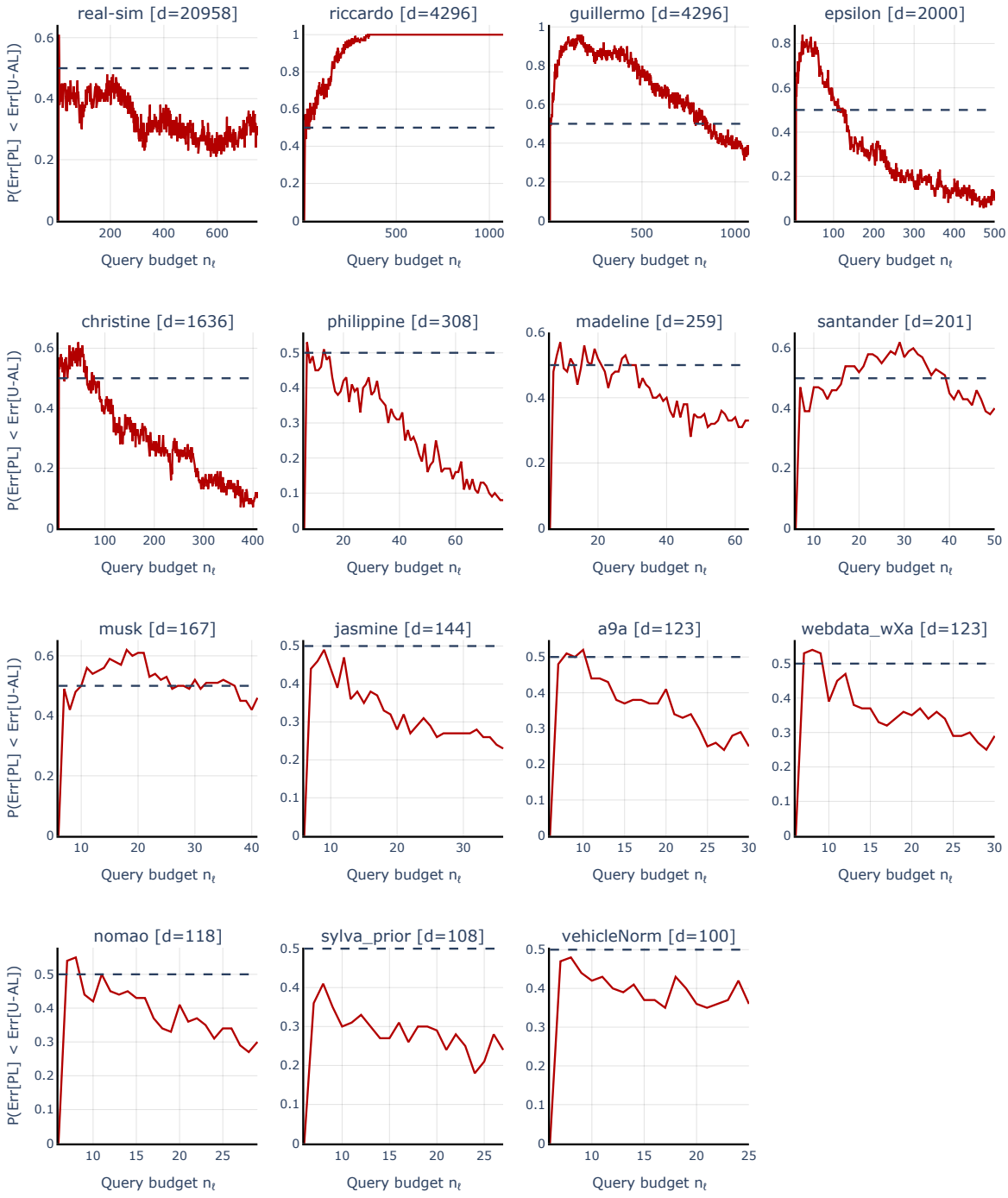


Figure 13: The probability that active learning via uncertainty sampling performs worse than passive learning at different query budgets. The empirical probability is computed over 100 draws of the initial seed set.

We note that the spikes in the lines in Figure 13 come from the fact that for different seed sets, uncertainty sampling may start to underperform at different iterations. Hence, aggregating over several seed sets can lead to non-smooth lines like in the figure.

In addition, in Figure 14 we summarize each of the panels in Figure 13 in a box plot that offers yet another perspective on this experiment. Notably, the boxes are fairly concentrated for all datasets, confirming that the gap between the test error with uniform and uncertainty sampling stays roughly the same for any query budget $n_q \in \{n_{\text{seed}}, ..., d/4\}$. Note that here the probability is over the draws of the seed set, and the box plots show percentiles of the distribution over query budgets for each dataset.

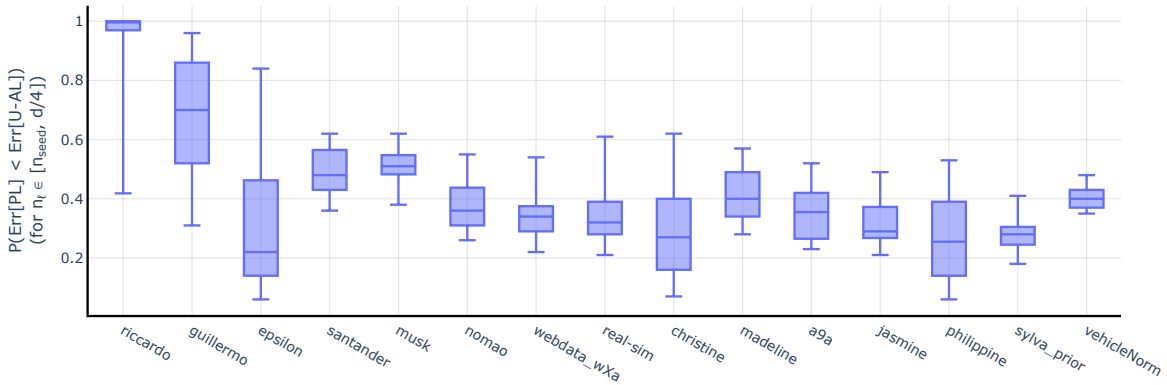


Figure 14: Box plot of the distribution of P(Err[PL] $<$ Err[AL]) over query budgets $n_q \in \{n_{\text{seed}}, ..., d/4\}$.

For these experiments we use the predictive uncertainty of $f(\cdot; \hat{\theta})$ as shown in Algorithm 1. In Figure 3-Bottom we show the largest gains and losses of uncertainty sampling for query budgets $n_q \in \{n_{\text{seed}}, ..., n_{\text{transition}}\}$, where $n_{\text{transition}}$ is defined as the budget after which uncertainty sampling is always better than uniform sampling with probability at least 50%. In other words, one can read $n_{\text{transition}}$ off Figure 13 as the leftmost point on the horizontal axis for which the solid line intersects the horizontal dashed line. For datasets that never intersect the 50% dashed line, we take $n_{\text{transition}} = d/4$ conservatively. This is more advantageous for uncertainty sampling, as larger query budgets tend to lead to larger gains over uniform sampling.

E.5 Fraction of budgets for which active learning underperforms

An alternative to using the metric illustrated in Figure 3-Top and in Appendix E.4 is to instead compute the fraction of the query budgets for which active learning performs worse than passive learning. In Figure 15 we present this evaluation metric for all the datasets that we consider. The box plot indicates the distribution over 100 draws of the initial seed set. For all datasets and with high probability over the draws of the seed data uncertainty sampling underperforms on a large fraction of the query budgets between n_{seed} and $d/4$.

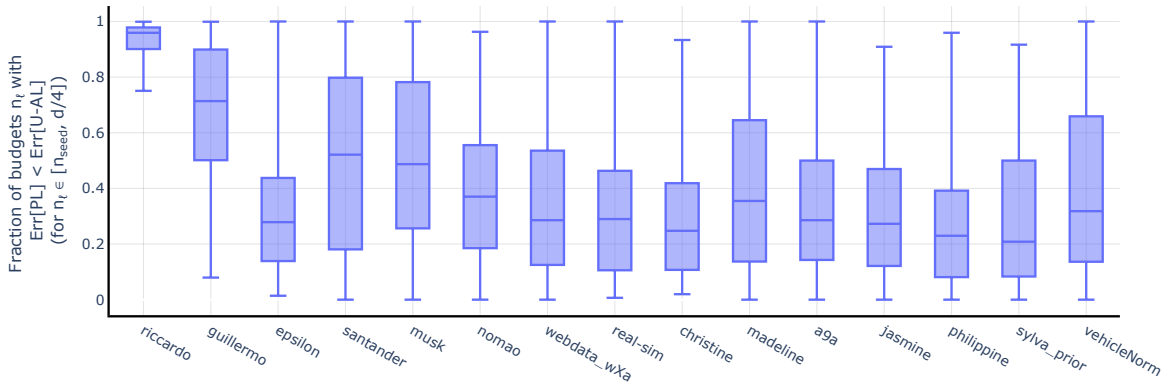


Figure 15: Fraction of the query budgets between n_{seed} and $d/4$ for which the error with U-AL is worse than with PL. The box plot indicates the distribution over 100 draws of the seed set (median, lower and upper quartiles).

Note that the fences of the box plots that almost cover the entire $[0, 1]$ range are a consequence of having a large number of runs (i.e. 100). The whiskers indicate the minimum and maximum values and they will be more extreme, the larger the set over which we take the minimum/maximum is.

E.6 Experiments with regularized estimators

The failure case of uncertainty-based active learning that we discuss in this paper is not limited to the situation when we use interpolating estimators. Indeed, as we show here, even regularization in the form of an ℓ_2 or ℓ_1 penalty still leads to classifiers with high test error when the data is collected using uncertainty sampling.

We not that, in what follows, a small coefficient C corresponds to stronger regularization, since we employ the scikit-learn (Pedregosa et al., 2011) implementation of penalized logistic regression. Therefore $C \to 0$ implies the predictive error term in the loss is ignored, while $C \to \infty$ leads to no regularization (note that unless otherwise specified, all results throughout the paper are reported for the unregularized max-ℓ_2-margin classifier).

Figures 16 and 17 indicate that for strong enough regularization, the gap between the test error of uncertainty and uniform sampling vanishes. This outcome is expected since stronger regularization leads to a poorer fit of the data, and hence, classifiers trained on different data sets (e.g. one collected with uncertainty sampling and another collected with uniform sampling) will tend to be similar. The downside of increasing regularization is, of course, worse predictive performance. For instance, for an ℓ_1 penalty and a coefficient of 0.01, the test error is close to that of a random predictor (i.e. 50%) an all datasets for both uniform and uncertainty sampling. For moderate regularization, there continue to exist broad ranges of query budgets for which uncertainty sampling underperforms compared to passive learning.

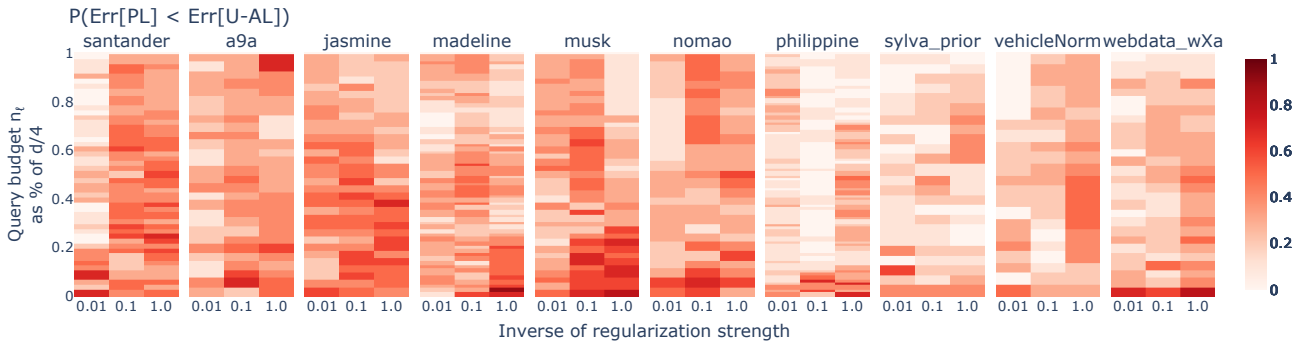


Figure 16: The probability that the test error is lower with PL than with U-AL, over 10 draws of the seed set. We use an ℓ_2-**regularized** classifier for both prediction and uncertainty sampling. Note that smaller values along the x-axis correspond to stronger regularization. If we regularize too much (e.g. for a coefficient of 0.01), the prediction error is poor for both PL and U-AL, which explains the light-colored regions.

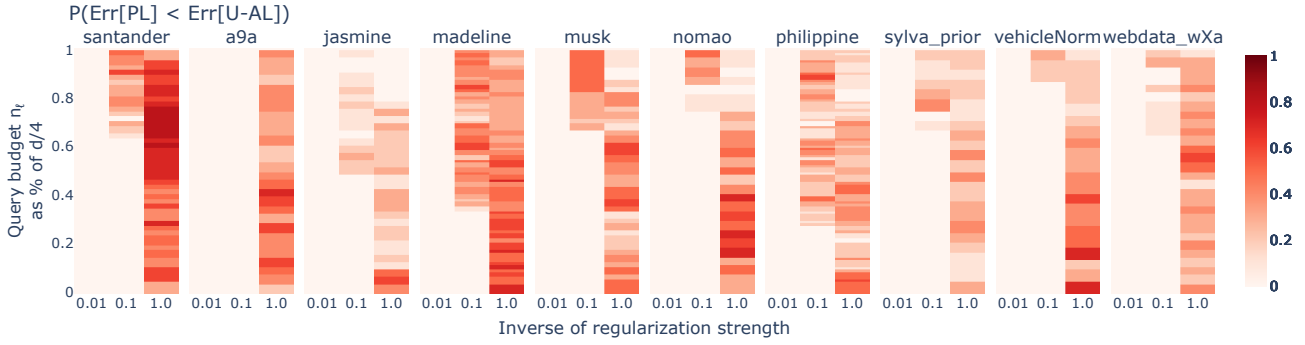


Figure 17: The probability that the test error is lower with PL than with U-AL, over 10 draws of the seed set. We use an ℓ_1-**regularized** classifier for both prediction and uncertainty sampling. Note that smaller values along the x-axis correspond to stronger regularization. If we regularize too much (e.g. for a coefficient of 0.01), the prediction error is poor for both PL and U-AL, which explains the light-colored regions.

E.7 Experiments with different seed set sizes

Our theory predicts that for a fixed labeling budget and an increasing see set size, the gap between the error with uncertainty sampling and uniform sampling vanishes.[9] We verify this insight experimentally in Figures 18 and 19. Our empirical findings confirm the trend predicted by our theory: uncertainty sampling leads to better performance for large seed set sizes, but underperforms for small seed sets.

Perhaps surprisingly, the same trend occurs even for oracle uncertainty sampling. This is noteworthy, since prior work suggests that the failure of uncertainty sampling for small seed set sizes is due to the usage of a meaningless measure of uncertainty, obtained by training a predictor on the small seed set. Instead, our results show that uncertainty sampling fails even when using the uncertainty of the Bayes optimal predictor, which highlights a novel failure case of this sampling strategy.

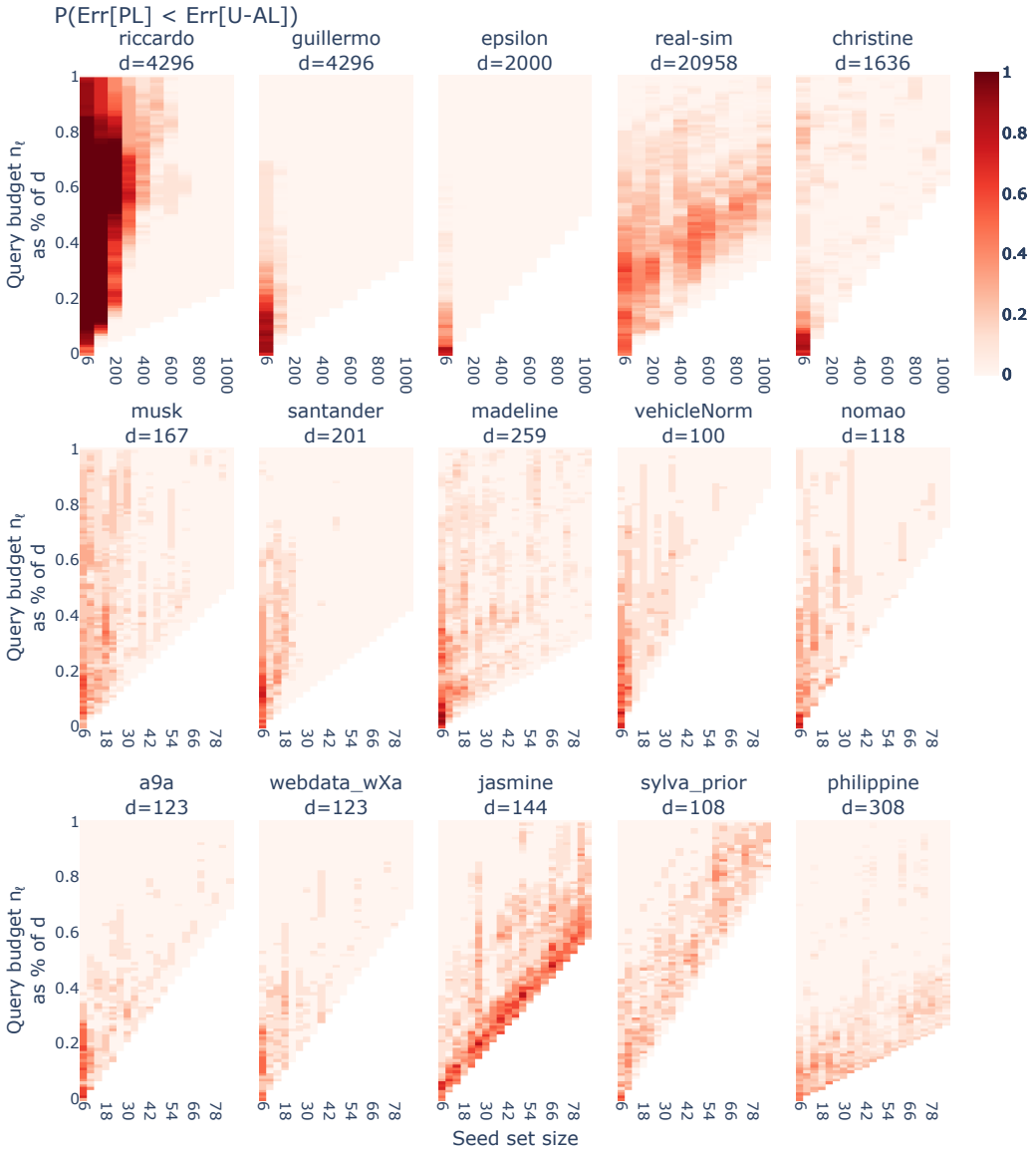


Figure 18: As predicted by our theory, increasing the seed size leads to improved performance when using uncertainty sampling to acquire new labeled samples.

[9]Note that if the seed set size matches the labeling budget uncertainty sampling is trivially equivalent to uniform sampling, since no queries are issued.

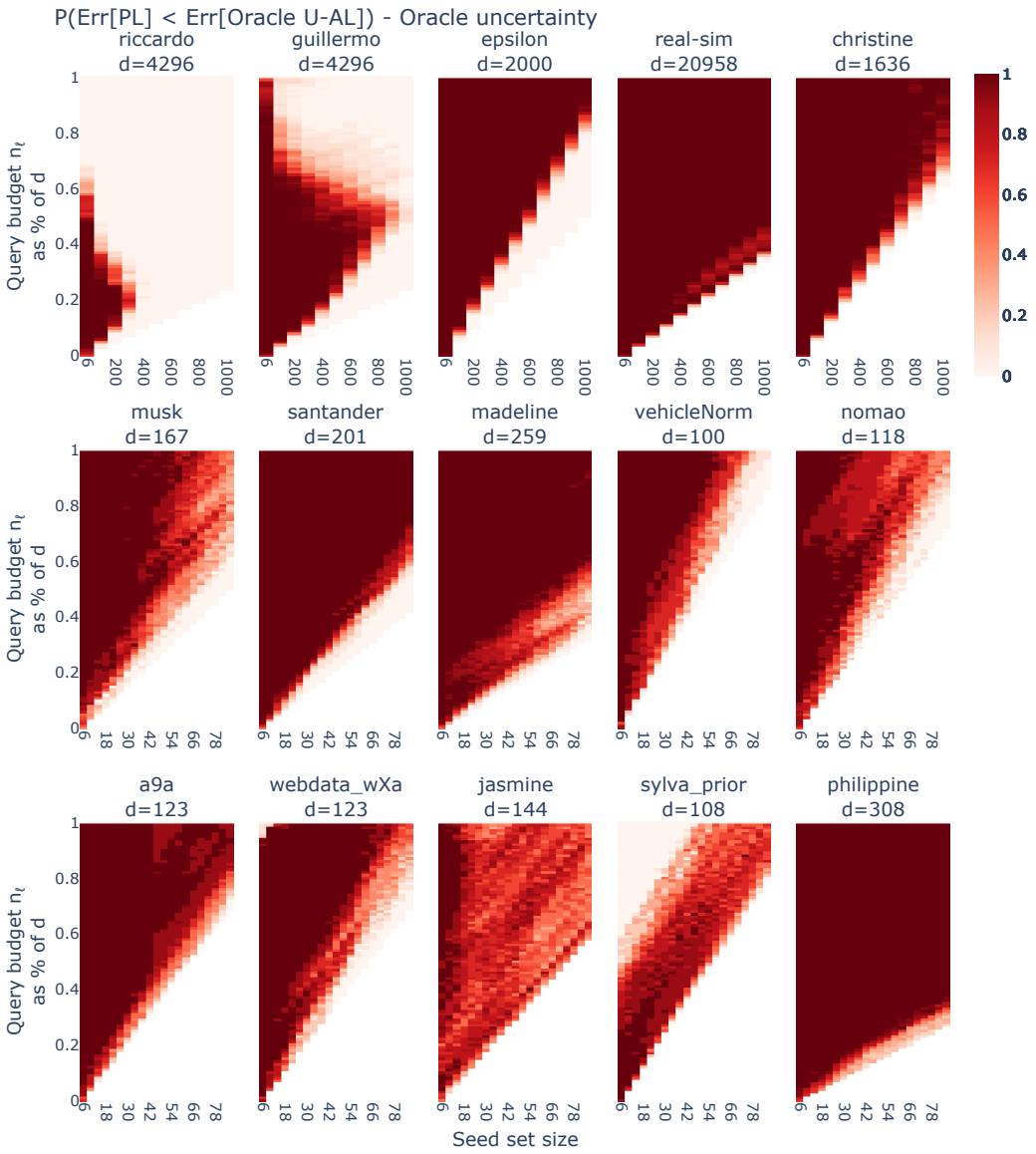


Figure 19: Surprisingly, increasing the seed set size also benefits oracle uncertainty sampling. This trend is not predicted by prior work and underlines a novel failure case of uncertainty sampling for high-dimensional data.

E.8 Combining uncertainty sampling and representativeness

In this section we provide evidence that the shortcoming of uncertainty sampling that we identify in this paper also extends to other active learning strategies that try to balance exploration and exploitation. In particular, we focus on an ϵ-greedy strategy which samples points using uncertainty with probability $1 - \epsilon$, and samples points uniformly at random with probability ϵ. Hence, this approach combines selecting informative samples via uncertainty sampling with collecting a labeled set that is representative of the training distribution. This strategy resembles the works of Brinker (2003); Huang et al. (2014); Yang et al. (2015); Gissin and Shalev-Shwartz (2019); Shui et al. (2020).

First, we note that for oracle uncertainty sampling, the ϵ-greedy strategy is equivalent to simply selecting a larger uniform seed set, since the queries are independent of each other when we use oracle uncertainty. Therefore, for a fixed query budget n_ℓ, the ϵ-greedy strategy with oracle uncertainty is identical to regular oracle uncertainty sampling where $n_{\text{seed}} = \epsilon \cdot n_\ell$. We conclude that the results in Section E.7, and more specifically Figure 19, show that the ϵ-greedy strategy performs worse than uniform sampling when using oracle uncertainty.

Finally, we check whether the ϵ-greedy strategy using the uncertainty of the empirical predictor $\hat{\theta}$ is also detrimental compared to passive learning. We notice in Figure 20 that for different values of ϵ, active learning continues to perform worse than passive learning. Varying ϵ between 0 and 1 effectively interpolates between vanilla uncertainty sampling and uniform sampling. This explains why the test error gap between the ϵ-greedy strategy and uniform sampling gets smaller as ϵ increases (e.g. for $\epsilon = 1$ the gap will always be 0, i.e. all cells would be white).

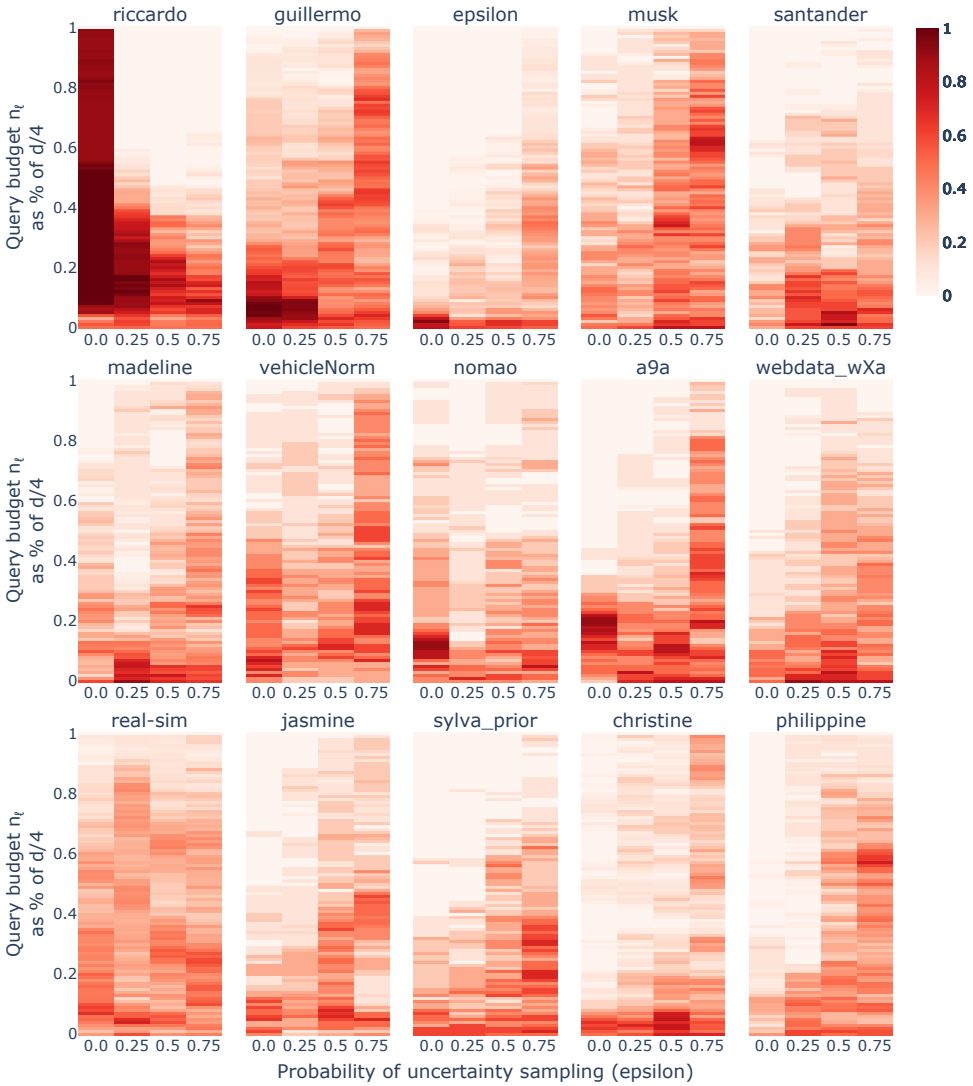


Figure 20: The probability that the test error is lower with uniform sampling than with an ϵ-greedy sampling approach, over 10 draws of the seed set. The active learning strategy performs uncertainty sampling, with probability $1 - \epsilon$ and samples uniformly at random with probability ϵ.

E.9 Coreset-based active learning

In this section we investigate whether the coreset-based sampling strategy proposed in Sener and Savarese (2018) can be a viable alternative to uncertainty sampling in low-sample regimes. We follow the same active learning methodology as described in Section 4, but use the greedy algorithm from Sener and Savarese (2018) to select queries. We use the Euclidean distance for our experiments.

Figure 21 shows that for a large fraction of query budgets, coreset-based active learning outperforms uncertainty sampling with high probability (dark red areas). However, for some datasets (e.g. *vehicleNorm, a9a, philippine*), coreset-based sampling can still lead to larger error than passive learning, as illustrated in Figure 22. We hypothesize that this behavior is due to not constraining the queried points to lie far from the Bayes optimal decision boundary. Hence, the high-dimensional phenomenon that we describe in Section 3 still occurs.

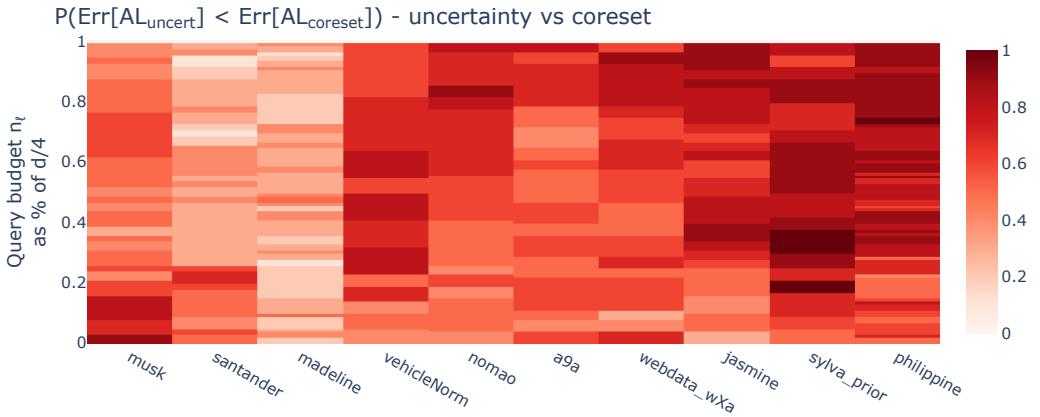


Figure 21: The coreset strategy of Sener and Savarese (2018) often outperforms U-AL with high probability (dark red).

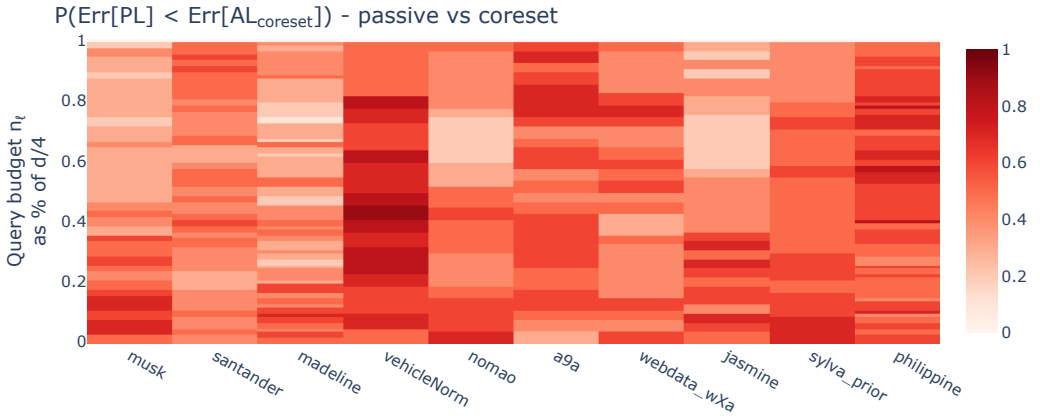


Figure 22: On some datasets (e.g. *vehicleNorm*), PL outperforms coreset-based active learning in the low-sample regime.

F Experiments on image datasets

In this section we describe our experiments on image datasets in which we explore the limitations of uncertainty sampling for low query budgets.

F.1 Experiment details

We consider 3 standard image datasets: CIFAR10 (Krizhevsky, 2009), CIFAR100 (Krizhevsky, 2009), SVHN (Netzer et al., 2011). In addition to these, we also run experiments on a binary classification task for medical images (PCAM (Veeling et al., 2018)) and on a 10-class task on satellite images (EuroSAT (Helber et al., 2017)). For prediction and for the uncertainty estimates we use ResNet18 networks (He et al., 2016) and start from weights pretrained on ImageNet. To get the oracle uncertainty estimates, we train on the entire labeled training set for each dataset until the training error reaches 0. We consider batch active learning, as usual in the context of deep learning, and set the batch size to 20 (experiments with larger batch sizes lead to similar results). For each dataset, we start from an initial seed set of 100 labeled examples and perform 50 queries. Hence, the largest query budget that we consider is of 1100 labeled samples. After each query step, we fine-tune the ResNet18 model for 20 epochs, and achieve 0 training error. For fine-tuning we use SGD with a learning rate of 0.001 and momentum coefficient of 0.9.

F.2 Summary of results

As illustrated in Figure 23, uncertainty sampling leads to significantly larger test error compared to passive learning. This phenomenon persists even when we use an oracle uncertainty (Figure 24). Moreover, the gains that uncertainty sampling can produce, are often dominated by the losses that it can incur. Note that for Figure 23-Bottom and Figure 24-Bottom we take $n_{\text{transition}} = 1100$, namely the maximum query budget n_ℓ that we consider.

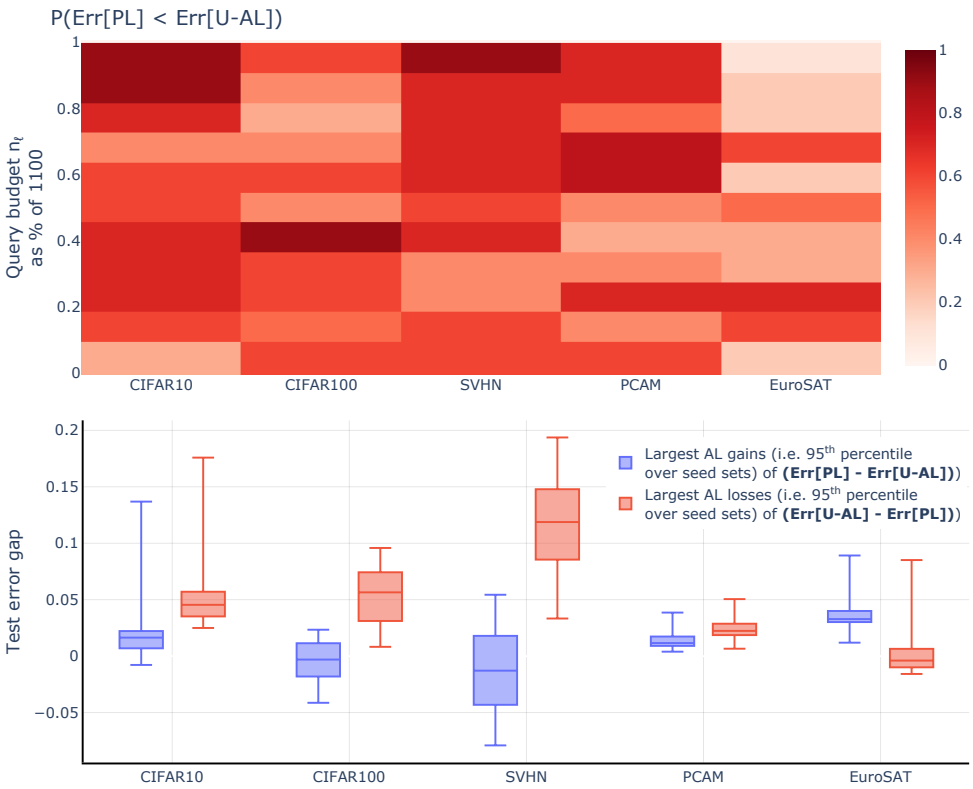


Figure 23: **Top:** The probability that the test error is lower with PL than with U-AL, over 10 different random seeds. PL outperforms U-AL, for a significant fraction of the query budgets and for all datasets (i.e. dark red regions). **Bottom:** The sporadic gains of U-AL over PL are generally similar or lower than the losses it can incur in terms of increased test error (negative values indicate that PL is always better than AL).

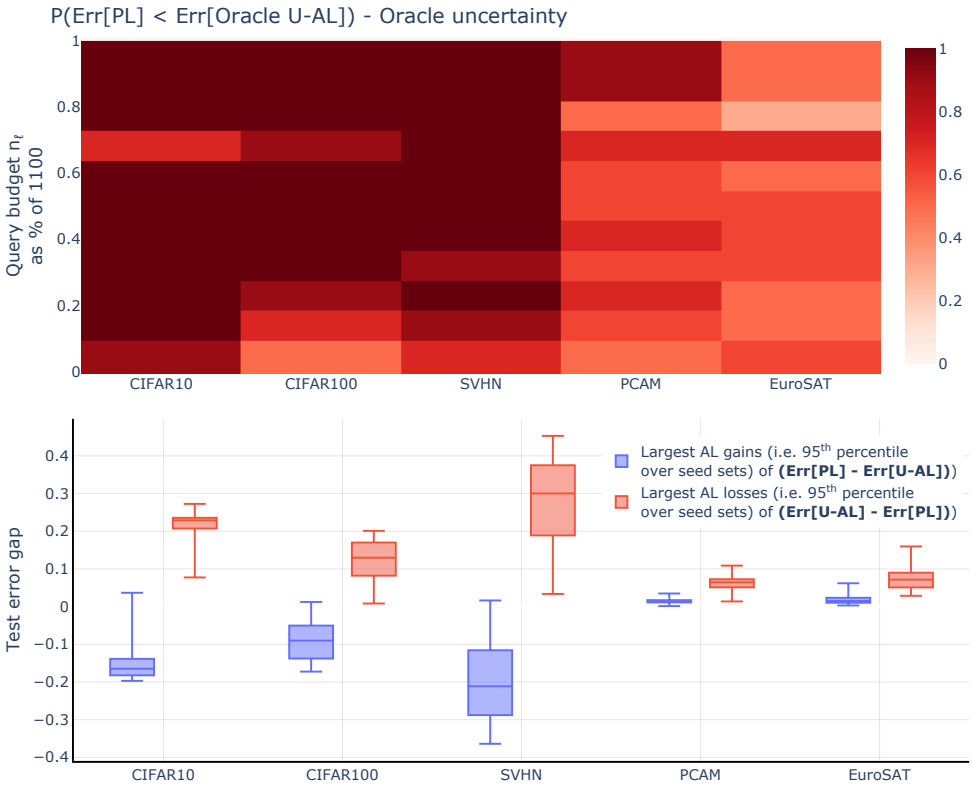


Figure 24: Same experiment as in Figure 23, but this time using **oracle U-AL**. Similar to the logistic regression experiments, U-AL leads to even worse error when using oracle uncertainty, as predicted by our theory.