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

Alexandru Ţifrea* TIFREAA@INF.ETHZ.CH
Department of Computer Science, ETH Zürich

Jacob Clarysse* JACOB.CLARYSSE@INF.ETHZ.CH
Department of Computer Science, ETH Zürich

Fanny Yang FAN.YANG@INF.ETHZ.CH
Department of Computer Science, ETH Zürich

Abstract

In applications where labeling data is prohibitively expensive, active learning algorithms like uncertainty sampling can lead to remarkable improvements compared to a passive sampling strategy. However, prior works show both theoretically and experimentally that sometimes uncertainty sampling is actually worse than passive learning. Despite a vast literature analyzing the low-dimensional regime, very little is known about how uncertainty sampling behaves in high dimensions. In this work, we study high-dimensional logistic regression and show that passive learning often outperforms uncertainty-based active learning for low label budgets. Our proof suggests that this high-dimensional phenomenon happens primarily when the separation between the classes is small. We corroborate this intuition with experiments on 15 high-dimensional data sets spanning a diverse range of applications, from finance and ecology to chemistry and histology.

Keywords: uncertainty sampling, active learning, logistic regression

1. Introduction

In numerous machine learning applications, it is often prohibitively expensive to acquire labeled data, even if unlabeled data may be 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, labeling all 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 allows for training a model with good predictive performance. Some of the most popular active learning algorithms are based on *uncertainty sampling* (Lewis and Gale, 1994). This paradigm proposes to train a prediction model (e.g. logistic regression, deep neural network) on the labeled set at each step. The algorithm then selects the samples on which the model has high predictive uncertainty and queries their label. Despite a plethora of promising results, recent theoretical and empirical works (Schein and Ungar, 2007; Lughofer and Pratama, 2017; Yang and Loog, 2018; Mussmann and Liang, 2018; Hacohen et al., 2022) have uncovered a number of problems with uncertainty sampling for

∗. Equal contribution.

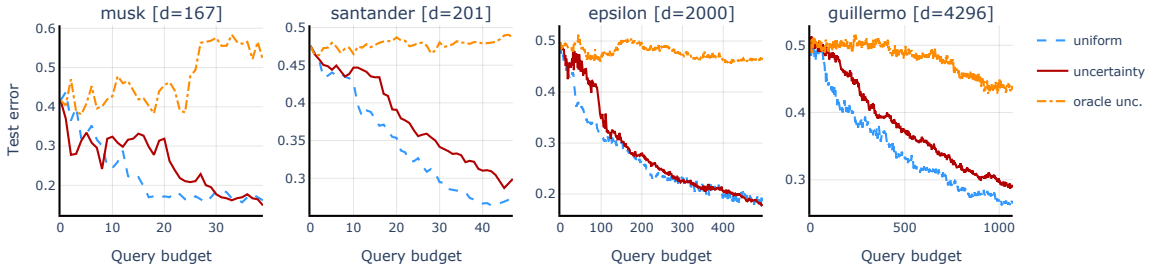


Figure 1: Surprisingly, uncertainty sampling (with or without oracle uncertainty) leads to worse test error compared to passive learning on a broad range of high-dimensional datasets and for several small query budgets. See Appendix H.2 for more datasets.

both linear and deep learning models. These shortcomings question whether the additional computational cost of uncertainty sampling is indeed justified.

The literature identifies two main causes for the failure of uncertainty sampling. First, several works claim that using the uncertainty of a bad predictive model can inconveniently skew the labeled data (Huang et al., 2014; Sener and Savarese, 2018; Hacohen et al., 2022). Second, in low dimensions, uncertainty sampling does not improve upon the sample efficiency of passive learning, if the Bayes optimal model has high error (Mussmann and Liang, 2018).

These prior results suggest that for noiseless data with vanishing Bayes error, an oracle-based uncertainty estimate should consistently be more beneficial than uniform sampling. In contrast, we show that for small query budget, active learning based on either empirical or oracle uncertainty sampling often performs worse than passive learning. As shown in Figure 1, we observe this phenomenon for logistic regression for a wide variety of high-dimensional datasets. This failure case of uncertainty sampling is not explained by previous theoretical analyses. Our proof reveals insights about this phenomenon, which we verify in logistic regression experiments on real-world data.

2. Theoretical analysis of uncertainty sampling for high-dimensional data

In this section, we analyze uncertainty sampling for linear classifiers on a mixture of truncated Gaussians, such that the data is guaranteed to be noiseless. We show that, in high dimensions, uncertainty sampling can lead to higher test error than passive learning.

Prediction task and training data. In this paper we consider binary classification, where the goal is to predict a label $y \in \{-1, 1\}$ from covariates $x \in \mathbb{R}^d$. We are interested in finding a linear classifier $x \to \operatorname{sgn}(x^\top \hat{\theta})$ with low test error $\operatorname{Err}(\hat{\theta}) = \frac{1}{n_{\text{test}}} \sum_{i=0}^{n_{\text{test}}} \mathbb{1}[y_i \neq \operatorname{sgn}(x_i^\top \hat{\theta})]$ on a holdout set $\mathcal{D}_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{test}}}$.

Given a labeled training set $\mathcal{D}$, a standard procedure for finding the parameters $\hat{\theta}$ of a linear classifier is to minimize the average logistic loss, i.e. $\hat{\theta} = \arg\min_\theta \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \ell(x^\top \theta, y)$ where $\ell(z, y) = \log(1 + e^{-zy})$. For linearly separable data, minimizing the logistic loss with stochastic gradient descent recovers the max-ℓ_2-margin (interpolating) solution (Soudry et al., 2018; Ji and Telgarsky, 2019) which has been analyzed extensively in recent years (Hastie et al., 2019; Javanmard and Soltanolkotabi, 2020; Donhauser et al., 2021, 2022).

Collecting the training set via uncertainty-based active learning. We consider standard pool-based active learning and assume access to a finite but large unlabeled dataset

$\mathcal{D}_u = \{x_i\}_{i=1}^{n_u}$ of size n_u (see Algorithm 1 in Appendix B). We start with a small seed set $\mathcal{D}_{seed} = \{(x_i, y_i)\}_{i=1}^{n_{\text{seed}}}$ of n_{seed} labeled points sampled i.i.d. from the training distribution. At each step n, we i) sample a point from the unlabeled data and add it to the labeled set with its true label; and then ii) train a classifier on the current labeled set. The querying steps are repeated n_q times until we exhaust the labeling budget, i.e. $n_\ell = n_{\text{seed}} + n_q$.

For querying, we focus on *uncertainty sampling* (see Appendix H.8 for other uncertainty-based strategies). Uncertainty sampling selects the sample on which the classifier has the highest predictive uncertainty. For linear models, we measure uncertainty using the inverse distance to the decision boundary determined by $\hat{\theta}$ (Platt, 1999; Mussmann and Liang, 2018; Raj and Bach, 2021), i.e. $\frac{\|\hat{\theta}\|_2}{|\hat{\theta}^\top x|}$. We also analyze *oracle uncertainty sampling* which uses the ground truth $\theta^\star$ to compute uncertainty.

Data distribution. We consider a class-balanced problem with a mixture of two truncated Gaussians as the marginal distribution of the covariates $x \in \mathbb{R}^d$. The true labels are given by $y = sgn(\theta^{\star\top} x) \in \{-1, 1\}$, with $\theta^\star = e_1 = [1, 0, ..., 0] \in \mathbb{R}^d$.[1] The ground truth $\theta^\star$ achieves the optimal error of $\text{Err}_{0-1}(\theta^\star) = 0$. We can write the covariates like $x = [x_1, \tilde{x}]$, where we explicitly separate the coordinates of x into a signal $x_1 \in \mathbb{R}$ and non-signal component $\tilde{x} \in \mathbb{R}^{d-1}$. We then write the class-conditional distribution of the covariates $x = [x_1, \tilde{x}]$ as:

$$\mathbb{P}_1(x_1|y) = \mathcal{N}_{trunc}(x_1; y\mu, \sigma^2, y) \text{ and } \tilde{\mathbb{P}}(\tilde{x}|y) = \mathcal{N}(\tilde{x}; 0, I_d),$$

where $\mu, \sigma > 0$ and $\mathcal{N}_{trunc}(\cdot; y\mu, \sigma^2, y)$ denotes the Gaussian distribution with mean $y\mu$ and variance σ^2 truncated to the interval $(-\infty, 0)$ if $y = -1$ and $(0, \infty)$ if $y = 1$. We note that GMMs have often been used to prove various surprising behaviors of machine learning algorithms (Tsipras et al., 2019; Frei et al., 2022).

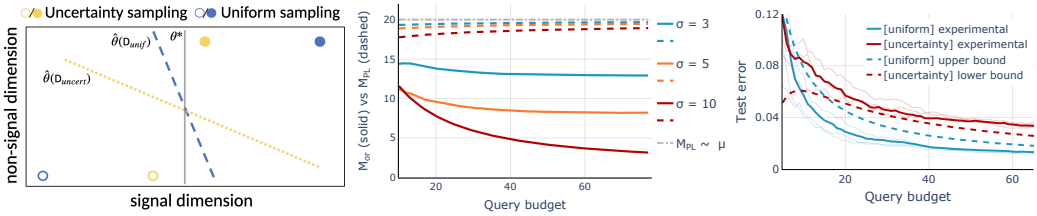
Main results. Our main theoretical result provides a lower bound on the test error that uncertainty and oracle uncertainty sampling incurs compared to uniform sampling.

We begin by noting that the error of a classifier parameterized by a vector $\theta \in \mathbb{R}^d$ is invariant to scaling θ by a positive constant. Therefore, we consider the set of parameters with the first dimension normalized to one, i.e. the set $\{\theta : \theta = [1, \alpha\theta_{2:d}]$, with $\|\theta_{2:d}\|_2 = 1$ and $\alpha \geq 0\}$. Assuming the data distribution in Section 2, the error of a linear classifier θ is fully determined by the relative weight of its first component, that is $\frac{1}{\alpha}$.

Recall that $\hat{\theta}(\mathcal{D})$ denotes the max-ℓ_2-margin classifier trained on $\mathcal{D}$. Let $\hat{\theta}(\mathcal{D}_{seed}) = [1, \alpha_{seed}\hat{\theta}_{2:d}]$ represent the classifier trained only on the seed set. We denote by $\hat{d}_q$ the distance between the decision boundary of $\hat{\theta}(\mathcal{D}_{seed})$ and the n_q^{th} closest point in the unlabeled set $\mathcal{D}_u$. Similarly, we define $d_q^\star$ as the distance between the decision boundary determined by the ground truth $\theta^\star$ and the n_q^{th} closest point in $\mathcal{D}_u$. Note that $\alpha_{seed}, \hat{d}_q$ and $d_q^\star$ are deterministic quantities for fixed $\mathcal{D}_{seed}$ and $\mathcal{D}_u$. We now state our main result (see Appendix C for formal statement and proof).

Theorem 1 (informal) *Let $n_\ell < d < n_u$ Let $\mathcal{D}_u$ and $\mathcal{D}_{seed}$ be datasets following the distribution in Section 2. The error of the max-ℓ_2-margin classifier $\hat{\theta}(\mathcal{D}_\ell)$ trained on a labeled set $\mathcal{D}_\ell$ is monotonically increasing in α. Moreover, for small fixed constants $C_{seed} > 0$ and $t > 0$, we have that the following hold with high probability:*

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



(a) Proof intuition in 2D. (b) M_{US} vs M_{PL} from Thm 1. (c) AL vs PL on GMMs.

Figure 2: (a) 2D intuition for the failure of uncertainty-based active learning in high dimensions. The classifier puts more weight on the non-signal dimension when trained on points close to the ground truth (yellow). (b) M_{US} can be much lower than M_{PL} for large enough σ, leading to worse prediction error for high-dimensional settings with $d \gg n_\ell$. (c) The bounds in Thm 1 (dashed lines) show that uniform sampling leads to lower test error compared to uncertainty sampling. The trend is confirmed by our simulations. We show the average (solid lines) over 3 different draws of the seed set (transparent lines). See Appendix F for more experimental details.

1. *the classifier obtained with* ***uncertainty sampling*** *has* $\alpha_{uncert} \geq \frac{\gamma_{US}}{M_{US}}$, *with* $\gamma_{US} = \Theta(\sqrt{d/n_\ell})$ *and* $M_{US} = (n_{seed}C_{seed} + n_q(\hat{d}_q + \sqrt{2\alpha_{seed}\log n_u} + t))/n_\ell$
2. *the classifier obtained with* ***oracle uncertainty sampling*** *has* $\alpha_{or} \geq \frac{\gamma_{or}}{M_{or}}$, *with* $\gamma_{or} = \Theta(\sqrt{d/n_\ell})$ *and* $M_{or} = (n_{seed}C_{seed} + n_q d_q^\star)/n_\ell$
3. *the classifier obtained by* ***uniform sampling*** *has* $\alpha_{unif} \leq \frac{\gamma_{PL}}{M_{PL}}$, *with* $\gamma_{PL} = \Theta(\sqrt{d/n_\ell})$ *and* $M_{PL} = \mu - t\sigma/\sqrt{n_\ell}$

Lastly, if $\sigma > \mu/2$, $\mathcal{O}(n_q) <= n_u < e^{\Theta(d/n_\ell)}$ *and* $\mu > \Theta((\log n_u)^{1/3}(d/n_{seed})^{1/6})$, *then with non-trivial probability* $\alpha_{uncert} > \alpha_{unif}$.

We now interpret the result of the theorem for oracle uncertainty sampling (a similar reasoning can be carried out for uncertainty sampling with empirical estimates). The theorem indicates that uncertainty sampling has a higher test error than passive learning if $\alpha_{\text{or}} > \alpha_{\text{unif}}$. Observe that in high dimensional settings (large d/n_ℓ ratio), it suffices to compare the denominators M_{US}, M_{or}, M_{PL} to analyze the gap between uncertainty and uniform sampling. The denominator of α has a geometric meaning: it upper bounds the average distance of the points in the labeled dataset to the decision boundary determined by the ground truth $\theta^\star$. Therefore, sampling strategies that query points close to the true decision boundary, like oracle uncertainty sampling, lead to a max-ℓ_2-margin classifier with larger error.

Figure 2a illustrates the intuition for this phenomenon. The max-ℓ_2-margin classifier puts more weight on the signal component when trained on points close to the ground truth (blue) compared to points far from the ground truth (yellow). Hence, the test error in the latter case is higher. Clearly, oracle uncertainty sampling performs particularly poorly as it samples by definition points close to the true decision boundary. In contrast, uniform sampling queries points with a signal component close to μ, on average.

For oracle uncertainty sampling, the dominating term in M_{or} is $n_q d_q^\star$, which can be significantly lower than $M_{PL} \approx \mu$ for a small enough $d_q^\star$. For the distribution that we consider, increasing the variance of the Gaussian components in the mixture leads to more points being close to the decision boundary. Hence, this leads to a smaller $d_q^\star$, which in turn

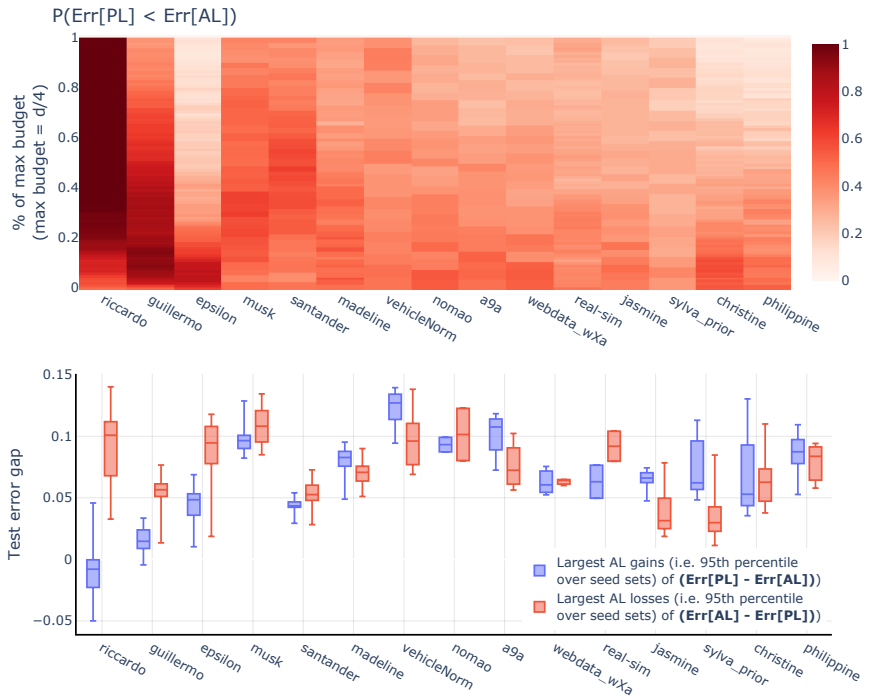


Figure 3: **Top:** The probability that the test error is lower with uniform sampling than with uncertainty sampling, over 100 draws of the seed set. Uniform sampling outperforms uncertainty sampling, for a significant fraction of the querying budgets and for all datasets (i.e. dark red regions). See Appendix H.4 for more precise numerical values. **Bottom:** For the range of budgets where uncertainty sampling 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.

causes M_{or} to be small. Indeed, we see in Figure 2b, that M_{or} decreases for larger σ, while M_{PL} is always close to $\mu \gg M_{or}$. It follows then that for a large d/n_u ratio, $\alpha_{\text{or}} > \alpha_{\text{unif}}$. Finally, by the monotonicity of the error as a function of α, we get that uncertainty sampling leads to higher error than uniform sampling, as illustrated in Figure 2c.

3. Experiments

We present logistic regression experiments on real-world datasets that confirm the insights of our theory (see Appendix I for neural network experiments on image datasets).

Evaluation metrics. [2] 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, for each budget size between n_{seed} and $d/4$. As revealed by Figure 3 (and Appendix H.4), the probability of uncertainty sampling outperforming uniform sampling is oftentimes small in the low-sample regime, before it eventually becomes larger than 50%, given a large enough query budget. The exact point where this transition occurs, which we denote by $n_{\text{transition}}$, differs from dataset to dataset.

Furthermore, we compare the most significant gains of uncertainty sampling with its most significant losses to gauge the magnitude of its failure compared to passive learning.

2. See also Appendix H.2 and H.5 for more evaluation metrics.

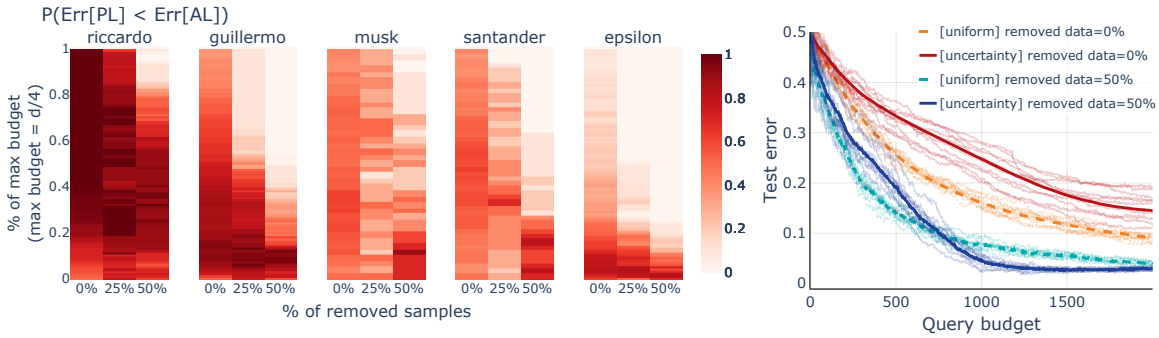


Figure 4: Increasing the separation of the classes in the unlabeled dataset improves the performance of uncertainty sampling. Removing the 25% or 50% closest points to the true decision boundary improves uncertainty sampling (**left**) which now outperforms passive learning for many query budgets (i.e.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

In particular, we focus on query budgets smaller than the transition point $n_{\text{transition}}$. For each budget size $n_\ell \in \{n_{\text{seed}}, ..., n_{\text{transition}}\}$ we compute the gap between the test error obtained with uncertainty sampling and the test error of passive learning. We repeat this procedure for several draws of the seed set and report the 95^{th} percentile of the largest positive and smallest negative gap between uncertainty and uniform sampling.

3.1 Main results

For most datasets, uncertainty sampling leads to a significant fraction of the labeling budgets leads to poorer results, indicated by the red regions that occur especially in the lower part of Figure 3-Top, for labeling budgets much smaller than the dimensionality. This phenomenon is particularly pronounced for very high-dimensional datasets (the right hand side of Figure 3-Top). We also observe the same trend for an ϵ-greedy uncertainty-based approach (see Appendix H.8). Moreover, we notice that the test error gains of uncertainty sampling are oftentimes much lower compared to the losses it can incur (for 9 datasets, the median gain is lower than the median loss).

Theorem 1 suggests that having a larger separation margin between the classes can improve the performance of uncertainty sampling in high dimensions. We verify this insight on real-world datasets. To increase the separation between the classes we remove the 25% or 50% closest samples to the decision boundary determined by $\theta^\star$. We estimate the ground truth by training a linear classifier on the full labeled dataset. Figure 4 shows that indeed, if data is well-separated, uncertainty sampling performs significantly better in high dimensions, even on challenging datasets e.g. riccardo. In addition, Appendix H.7 shows that increasing the size of the seed set also improves performance, as predicted by our theory.

4. Conclusions

When acquiring new labeled data is costly, it is particularly important to not waste the limited budget on samples that will only deteriorate predictive performance. Our results suggest that uniform sampling is oftentimes a more beneficial strategy than uncertainty-based active learning in high dimensions. It remains an open challenge to design a strategy that outperforms uniform sampling in high dimensions.

Appendix A. Related work

Prior analyses of uncertainty sampling. As already pointed out in Section 1, several prior works show that uncertainty sampling performs remarkably well for a broad variety of datasets and applications (Tong and Koller, 2001; Settles et al., 2007; Schohn and Cohn, 2000; Raj and Bach, 2021). Complementing these experimental observations, a number of recent theoretical works analyze uncertainty sampling. For instance, (Raj and Bach, 2021) show that a linear classifier trained on data collected via uncertainty sampling converges to a solution with good generalization. Despite these positive results, a number of works (Schein and Ungar, 2007; Lughofer and Pratama, 2017; Yang and Loog, 2018) show experimentally that uncertainty sampling is sometimes not better than uniform sampling (i.e. passive learning). Indeed, Mussmann and Liang (2018) prove that the benefits of uncertainty sampling are proportional to the Bayes error for a given low-dimensional data distribution. Furthermore, Huang et al. (2014) discuss the issues that arise due to the fact that the data acquired via uncertainty sampling is not representative of the training distribution.

While these works paint a detailed picture of uncertainty sampling in low dimensions, very little is known about how effective this sampling strategy is for high dimensional data. In fact, experiments in Ducoffe and Precioso (2018); Sener and Savarese (2018); Shui et al. (2020); Mahmood et al. (2021) suggest that margin-based sampling strategies for neural networks are rarely better than uniform sampling for image data. Despite these first indications of a serious shortcoming of uncertainty sampling in high dimensions, to the best of our knowledge, no prior work has tried to investigate this phenomenon further (either theoretically or empirically).

Related querying strategies. Not only is uncertainty sampling one of the most popular sampling strategies (Settles, 2009), but it is also related to a number of other active learning algorithms. For instance margin-based active learning (Scheffer and Wrobel, 2001; Ducoffe and Precioso, 2018; Mayer and Timofte, 2020) or entropy sampling (Settles, 2009) are both flavors of the same sampling procedure as uncertainty sampling. In addition, several works propose to query points using a score that combines uncertainty and a measure of how representative of the training distribution the samples are. These strategies too suffer from the problems we expose in Section 2, as long as they sample points close to the true decision boundary.

Relationship to semi-supervised learning. Both active learning and semi-supervised learning (SSL) are suitable in similar settings, namely when both a small labeled set and a large unlabeled set are available. Therefore, it is natural to wonder whether SSL could in fact entirely solve the problems shown by active learning in high dimensions. The issue with SSL, however, is that it requires strong assumptions to work well, while active learning can decrease the labeled sample complexity exponentially compared to passive (supervised) learning for a very broad family of distributions (Dasgupta, 2005; Beygelzimer et al., 2010; Hanneke, 2013). For instance, Schölkopf et al. (2012) argues that SSL cannot improve over supervised learning in causal learning problems (i.e. covariates x are a causal parent of the predictor y), while active learning can still provide benefits in such scenarios. In particular, we highlight that the shortcomings of uncertainty sampling that we reveal in this paper also apply for causal learning problems, where SSL cannot be used either.

Appendix B. Pseudocode for uncertainty sampling

Algorithm 1: Uncertainty sampling

Input: Seed set $\mathcal{D}_{seed}$, Unlabeled set $\mathcal{D}_u$, Budget n_ℓ, Uncertainty function u
Result: Prediction model $\hat{\theta}$

$\mathcal{D}_\ell \leftarrow \mathcal{D}_{seed}$
$\hat{\theta} \leftarrow \arg\min_\theta \frac{1}{|\mathcal{D}_{seed}|} \sum_{(x,y)\in\mathcal{D}_{seed}} \ell(\hat{\theta}^\top x, y)$
for $n \in \{n_{seed}+1, ..., n_\ell\}$ **do**
  $x_n \leftarrow \arg\max_{x\in\mathcal{D}_u} u(x;\hat{\theta})$
  $y_n \leftarrow AcquireTrueLabel(x_n)$
  $\mathcal{D}_\ell \leftarrow \mathcal{D}_\ell \cup \{(x_n, y_n)\}; \mathcal{D}_u \leftarrow \mathcal{D}_u \setminus \{x_n\}$
  $\hat{\theta} \leftarrow \arg\min_\theta \frac{1}{|\mathcal{D}_\ell|} \sum_{(x,y)\in\mathcal{D}_\ell} \ell(\hat{\theta}^\top x, y)$
return $\hat{\theta}$

Appendix C. Theorem 1 formal

In this section, we state and give the proof of the formal version of Theorem 1. We first discuss some basic properties of the truncated Gaussian distribution and two stage uncertainty sampling after which we state the formal version of Theorem 1. Thereafter, we state the lemmas that we use for the proof. Lastly, we give the proof itself. The proofs of the lemmas are given in Appendix E.

C.1 Properties of the truncated Gaussian distribution

Accuracy of a classifier To gain intuition, we start with a discussion on the accuracy of a given classifier induced by a vector $\theta \in \mathbb{R}^d$, with $\theta = [1, \alpha\tilde{\theta}]$, where $\|\tilde{\theta}\|_2 = 1$. Note that the last $d-1$ coordinates of any sample drawn from the truncated Gaussian distribution are standard normal distributed. By definition, the error of the classifier θ is given by

$$\mathrm{Err}_{0-1}(\theta) = P\left[y\theta_1 x_1 + y\sum_{i=2}^{d}\theta_i x_i < 0\right] = P\left[yx_1 + y\sum_{i=2}^{d}\frac{\theta_i}{\theta_1}x_i < 0\right],$$

where w.l.o.g. we assumed that $\theta_1 > 0$. Because the sum of Gaussian random variables is again a Gaussian random variable, we find that $y\sum_{i=2}^{d}\frac{\theta_i}{\theta_1}x_i$ is Gaussian distributed with a mean of zero and a standard deviation of $\tilde{\sigma} = \sqrt{\alpha^2+1}$. Denote by Φ the cumulative density function of the standard normal distribution. 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

$$\mathrm{Err}_{0-1}(\theta) = \Psi(\tilde{\sigma}(\alpha), \mu, \sigma) = \frac{1}{2\pi\tilde{\sigma}\sigma(1-\Phi(-\mu/\sigma))}\int_0^\infty\int_t^\infty e^{-\frac{(t-\mu)^2}{2\sigma^2}}e^{-\frac{l^2}{2\tilde{\sigma}^2}}dldt \qquad (1)$$

We can easily compute the expression in Equation 1 numerically. For convenience we state two properties of Ψ here:

- Ψ is a monotonically increasing function of α.
- Ψ is monotonically decreasing for increasing μ.

Therefore, for fixed distributional parameters, μ and σ, we have that α fully characterizes the error of the classifier. Indeed, in Theorem 1 we give upper bounds on α for uncertainty sampling and oracle uncertainty sampling, and a lower bound for passive learning.

Mean and standard deviation Next, we state the known formulas for the mean and standard deviation of a positive one sided truncated Gaussian random variable. The positive one-sided truncated Gaussian distribution is defined as follows: we cut off a Gaussian with mean μ and standard deviation σ to the interval $[0, \infty]$. Clearly the mean of the truncated Gaussian is slightly larger than the original Gaussian and the standard deviation slightly smaller. Let ϕ be the probability density function of the standard normal distribution. Then we find that the mean of the 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

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

We now discuss two stage uncertainty sampling and how the result distinguishes from regular uncertainty sampling.

C.2 Two stage uncertainty sampling

For oracle uncertainty sampling, we consider regular uncertainty sampling. However, as in (Chaudhuri et al., 2015; Mussmann and Liang, 2018), we formally analyse the following scheme, which Mussmann and Liang (2018) call two stage uncertainty sampling. First, we uniformly sample a seed dataset of size n_{seed}. Thereafter, we sample the $n_\ell - n_{\text{seed}}$ closest points to the classifier obtained by the seed, i.e. one step uncertainty sampling. We only restrict to the two stage uncertainty sampling scheme for one step of the proof. Hence, the bounds given in Theorem 1 and 2 may still provide good estimates for regular uncertainty sampling. Indeed, in Figure 2c we see that the theorem also provides reasonable bounds for regular uncertainty sampling.

C.3 Formal statement of Theorem 1

For ease of interpretation, we state the theorem in function of the parameters d_q and α_{seed} that only depend on the small seed set. Moreover, we assume all points in $\mathcal{D}_\ell$ to be support points of the classifier $\hat{\theta}(\mathcal{D}_\ell)$. Recall that we define d_q to be the distance of the n_q closest samples to the decision boundary of the classifier obtained using the seed set and let α_{seed} be the α-parameter of $\hat{\theta}(\mathcal{D}_{seed})$. We discuss the properties of d_q, α_{seed} and why the assumption of support points holds in Appendix D. We are now able to state the formal version of the main theorem.

Theorem 2 *Assume we sample a dataset $\mathcal{D}_\ell$ of size n_ℓ from an i.i.d. unlabeled dataset of size n_u drawn from the truncated Gaussian distribution, starting from a seed dataset of size n_{seed} also drawn from the truncated Gaussian distribution. Moreover, let $n_u > d > n_\ell$. Then, the error of a given classifier is in the form of $\Psi(\tilde{\sigma}(\alpha), \mu, \sigma)$, where Ψ is monotonically*

increasing in α. Lastly, we assume that all points in $\mathcal{D}_\ell$ are support points. We find for $C_{seed} = \mu_{tr} + t\frac{\sigma_{tr}}{\sqrt{n_{seed}}}$ the following.

1. *If $\mathcal{D}_\ell$ is sampled using **two-stage uncertainty sampling**, then with a probability greater than $(1 - 2e^{-t^2/2})^3$, we have that $\alpha \geq \frac{\gamma_{US}}{M_{US}}$ with*

$$M_{US} = \left(n_{seed}C_{seed} + n_q(\hat{d}_q + \sqrt{2\alpha_{seed}\log n_u} + t)\right)/n_\ell \quad \gamma_{US} = \sqrt{\frac{d}{n_\ell}} - \sqrt{2\log n_u} - 1 - t$$

2. *If $\mathcal{D}_\ell$ is sampled using **oracle uncertainty sampling**, then with a probability greater than $(1 - e^{-t^2/2})^2$ we have that $\alpha \geq \frac{\gamma_{or}}{M_{or}}$ with*

$$M_{or} = (n_{seed}C_{seed} + n_q d_q^\star)/n_\ell \quad \gamma_{or} = \sqrt{\frac{d}{n_\ell}} - 1 - t$$

3. *If $\mathcal{D}_\ell$ is sampled using **uniform sampling**, then with a probability greater than $(1 - e^{-t^2/2})^2$, we have that $\alpha \leq \frac{\gamma_{PL}}{M_{PL}}$ with*

$$M_{PL} = \mu - t\sigma/\sqrt{n_\ell} \quad \gamma_{PL} = \sqrt{\frac{d}{n_\ell}} + 1 + t$$

Note that Theorem 2 is the direct formalization of Theorem 1, besides for the direct comparison of α_{uncert} and α_{unif}. We extent the Theorem with an informal corollary here, for which we give the proof in Appendix C.6

Corollary 3 *For the same setting as in Theorem 2. For an $\eta > 0$. If $\eta_2 > \frac{n_\ell}{n_q}\eta$, $\sigma > \mu\eta_2$, $n_q/(0.8\eta_2) < n_u < e^{\mathcal{O}(\eta^2 d/n_\ell)}$ and*

$$\mu > \frac{1}{(n_q\eta_2 - \eta n_\ell)^{2/3}} n_q^{2/3}(2\log n_u)^{1/3}(d/n_{seed} + 1)^{1/6},$$

then with non-trivial probability two-stage uncertainty sampling has a higher test error than passive learning.

Generalization to regular uncertainty sampling While we do not prove the statement for regular uncertainty sampling, our bounds intuitively approximately hold for this setting as well. In the case of regular uncertainty sampling, we replace d_q with the largest distance of a sample taken to the respective decision boundary, which by definition of the query rule is likely small. Moreover, we then also replace α_{seed} with the α for at each iteration. For all queries where the classifier is not trivial α is reasonably small.

C.4 Lemma statements

We now state the three lemmas that we use to prove Theorem 2. The first lemma provides upper and lower bounds on the max-ℓ_2-margin in the last $d-1$ coordinates of a labeled dataset $\mathcal{D}_\ell$ acquired by querying samples using any sample strategy and uniform sampling. In particular, the lemma is tailored to the high dimensional regime, i.e. $d > n_\ell$. Note that the lemmas do not assume a sampling method and are hence also applicable to regular uncertainty sampling.

Lemma C.1 *[Non-signal margin of active learning] Let $\mathcal{D}_u = \{(x_i, y_i)\}_{i=1}^{n_u}$ be a dataset with $n_u > d$ samples, where each entry of any x_i is independent normal distributed. Moreover, let y_i be Bernouilli distributed with probability $1/2$. Then, for $n_q < d-1$, any dataset $\mathcal{D}_\ell = \{(x_i, y_i)\}_{i=1}^{n_q}$ obtained by sampling from $\mathcal{D}_u$ has with a probability of at least $1 - e^{-\frac{t^2}{2}}$ an max l_2-margin greater than*

$$\gamma \geq \sqrt{\frac{d}{n_q}} - \sqrt{2 \log n_u} - 1 - t.$$

Moreover, if we subsample uniformly, than with a probability of greater than $1 - 2e^{-t^2/2}$, the margin is upper and lower bounded by

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

In the second lemma we look at the max average-ℓ_2-margin in the last $d-1$ coordinates of a dataset acquired using uniform sampling. The max average-ℓ_2-margin of a dataset $\mathcal{D}_\ell = \{(x_i, y_i)\}_{i=1}^{n_\ell}$ is defined as

$$\gamma_{avg} = \max_{\theta \in \mathbb{R}^d, \|\theta\|_2 = 1} \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} y_i \theta^\top x_i \tag{4}$$

Lemma C.2 (Upper bound on the average margin for uniform sampling) *Let $\mathcal{D}_\ell = \{(x_i, y_i)\}_{i=1}^{n_\ell}$ with $d > n_\ell$ be a dataset of i.i.d. random vectors with random binary labels. Then, with a probability greater than $1 - 2e^{-\left(\frac{\sqrt{d}t}{\sqrt{n_\ell}} - 1/d\right)^2}$, the average margin γ_{avg} is upper bounded by*

$$\gamma_{avg} \leq \sqrt{\frac{d}{n_\ell}} + t$$

Lastly, the third lemma gives concrete expression of the max-ℓ_2-margin classifier of a dataset $\mathcal{D}_\ell$ in function of the mean distance of the samples to the ground truth e_1 and the max-ℓ_2-margin classifier in the $d-1$ last coordinates. The lemma formalizes the intuition that query strategies which sample points close to the ground truth may lead to high test errors.

Lemma C.3 (Bound on the optimal classifier for active learning) *Let $\mathcal{D}_\ell = \{(x_i, y_i)\}_{i=1}^{n_\ell}$ be a dataset with a max-ℓ_2-margin of γ and a max average-ℓ_2-margin of γ_{avg} in the subsequent $d-1$ coordinates. Moreover, assume that the labels are given by the sign of the first coordinate and that the first coordinate is almost surely non-zero. Lastly, assume that all covariates of the dataset are support vectors of the max-ℓ_2-margin classifiers. Then, the max-ℓ_2-margin solution is for a $\gamma_{avg} \geq b \geq \gamma$ given by*

$$\hat{\theta} = \left[1, \frac{b}{\frac{1}{n_\ell}\sum_{i=1}^{n_\ell} y_i x_{1,i}} \tilde{\theta}\right],$$

with $\|\tilde{\theta}\|_2 = 1$.

C.5 Proof of Theorem 2

In this section, we prove the main theorem. The main theorem 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.

Let all points in the labeled dataset $\mathcal{D}_\ell = \{(x_i, y_i)\}_{i=1}^{n_\ell}$ be support points: removing any point from $\mathcal{D}_\ell$ changes the max-ℓ_2-margin classifier. Then Lemma C.3 gives the following bounds on α:

$$\frac{\gamma}{\bar{x}} \leq \alpha \leq \frac{\gamma_{avg}}{\bar{x}}, \tag{5}$$

where $\bar{x} = \frac{1}{n_\ell}\sum_{i=1}^{n_\ell} y_i x_{1,i}$ and γ, γ_{avg} are the the max-ℓ_2-margin and average max-ℓ_2-margin respectively. Using Equation 5 and invoking Lemmas C.1 and C.2, we solely need to lower bound $\bar{x}$ for uniform sampling and upper bound it for two-stage and oracle uncertainty sampling.

Seed set Note that the labeled dataset is given by the union of the seed set and the new queried points, i.e. $\mathcal{D}_\ell = \mathcal{D}_{seed} \cup D_q$. For convenience of notation, we define $\bar{x}_{seed}$ and $\bar{x}_q$ as the mean distance of the samples to the optimal decision boundary (the plane defined by the normal e_1) of the seed and queried set respectively. Using these definitions, we find that

$$\bar{x} = (n_{\text{seed}}\bar{x}_{seed} + n_q\bar{x}_q)/n_\ell \tag{6}$$

We now compute bounds on $\bar{x}_{seed}$ and then consider bounds on $\bar{x}_q$ for each case separately. Recall that the CDF of the positive one-sided truncated Gaussian is given by

$$\Phi_{tr}(x) = \frac{1}{1-\Phi(-\mu/\sigma)}\left(\Phi\left(\frac{x-\mu}{\sigma}\right) - \Phi(-\mu/\sigma)\right) \tag{7}$$

Let ϕ be the standard normal density function. Recall that a variable distributed according to the one-sided truncated Gaussian distribution is a sub-gaussian random variable with a mean and a standard deviation as defined in Equations 2 and 3 respectively. Hence, with a probability larger than $1 - e^{-t^2/2}$, we have that

$$\mu - t\frac{\sigma}{\sqrt{n_{\text{seed}}}} \leq \bar{x}_{seed} \leq \mu_{tr} + t\frac{\sigma_{tr}}{\sqrt{n_{\text{seed}}}} \tag{8}$$

Note that we have not used the approximation introduced by two-stage uncertainty sampling yet. In the following, we aim to bound $\bar{x}_q$ for which look at each sampling strategy individually.

Oracle uncertainty sampling We start with oracle uncertainty sampling. By definition, oracle uncertainty sampling queries the n_q closest points to the optimal classifier independently of the $d-1$ last (non-signal) coordinates. As oracle uncertainty sampling queries points independently of the last $d-1$ coordinates, we can invoke Lemma C.1 for independent uniformly drawn samples; we find that the max-ℓ_2-margin in the $d-1$ last coordinates is lower bounded by $\sqrt{d/n_\ell} - 1 - t$ with a probability larger than $1 - 2e^{-t^2/2}$. Observe that $\bar{x}_q$ is in this case trivially upper bounded by d_q. Hence, by Lemma C.3 and C.1, Equation 8 and 6 and the trivial bound $d_q \geq \bar{x}_q$, we have proven the first statement of the theorem.

Uncertainty sampling Next, we consider two-stage uncertainty sampling. First, we find that the max-ℓ_2-margin in the last $d-1$ coordinates of the labeled dataset $\mathcal{D}_\ell$ is by Lemma C.1 upper bounded by $\sqrt{\frac{d}{n_\ell}} - \sqrt{2\log n_u} - 1 - t$ with a probability greater than $1 - 2e^{-t^2/2}$. We now bound the value of $\bar{x}_q$ with high probability. Let $\hat{\theta} = [1, \alpha_{seed}\tilde{\theta}]$ with $\|\tilde{\theta}\|_2 = 1$ be the classifier obtained on the seed set $(\hat{\theta}(\mathcal{D}_{seed}))$. By definition $d_q > y\hat{\theta}^\top x$ for all $(x, y) \in \mathcal{D}_\ell \setminus \mathcal{D}_{seed}$. To circumvent the dependence of the classifiers on the unlabeled dataset, we use the approximation of two-stage uncertainty sampling in the next step. By the independence of $\tilde{\theta}$ on all the samples in the unlabeled dataset and the union bound, we find that $\max_{(x,y)\in\mathcal{D}_u} \alpha_{seed}\tilde{\theta}^\top x_{2:d} < \sqrt{2\alpha_{seed}\log n_u} + t$ with a probability greater than $1 - 2e^{-t^2/2}$. Taking the sum $d_q + \alpha_{seed}\max_{(x,y)\in\mathcal{D}_u} y\tilde{\theta}x$, we find that

$$\bar{x}_q < d_q + \sqrt{2\alpha_{seed}\log n_u} + t \tag{9}$$

with a probability greater than $1 - 2e^{-t^2/2}$. Therefore, by Lemma C.3 and C.1, Equation 8, 6 and 9, we have proven the second statement of the theorem.

Uniform sampling Lastly, we lower bound $\bar{x}$ for uniform sampling. Note that the seed set is also uniformly sampled. Hence, the left side of Equation 8 directly gives us the high probability lower bound. Using

C.6 Proof of Corollary 3

We compare α_{uncert} and α_{unif} using Theorem 2. For the corollary, we aim for a statement in the form of: "with non-trivial probability uncertainty sampling is worse than passive learning". Hence, we can set $t = 0$ and compute when $\alpha_{US} > \alpha_{PL}$ in this case. Clearly, by Theorem 2:

$$\frac{\gamma_{US}}{M_{US}} > \frac{\gamma_{PL}}{M_{PL}} \iff \frac{M_{US}}{M_{PL}} < \frac{\sqrt{d/n_\ell} - \sqrt{2\log n_u} - 1}{\sqrt{d/n_\ell} + 1} = 1 - \frac{\sqrt{2\log n_u} + 2}{\sqrt{d/n_\ell} + 1}$$

Note that the right-hand side is approximately 1. We see that for some $\eta > 0$,

$$\frac{\sqrt{2\log n_u} + 2 + t}{\sqrt{d/n_\ell} + 1} < \eta \iff n_u < e^{(\eta\sqrt{d/n_\ell}+1)^2/2 - 1} = e^{\mathcal{O}(\eta^2 d/n_\ell + \eta\sqrt{d/n_\ell})}$$

Hence, if $\frac{M_{US}}{M_{PL}} < 1-\eta$ and $n_u < e^{\mathcal{O}(\eta^2 d/n_\ell + \eta\sqrt{d/n_\ell})}$ then $\alpha_{PL} < \alpha_{US}$. Now, we have that

$$\frac{M_{US}}{M_{PL}} = \left(n_{\text{seed}} C_{seed} + n_q(\hat{d}_q + \sqrt{2\alpha_{seed}\log n_u})\right)/(\mu n_\ell) < 1-\eta$$

Further, $\hat{d}_q$ is by definition the distance of the n_q closest point to the decision boundary of $\hat{\theta}(\mathcal{D}_{seed})$. We have that $\hat{\theta}(\mathcal{D}_{seed}) = [1, \alpha_{seed}\tilde{\theta}]$. As $\|\tilde{\theta}\|_2 = 1$, any random point (x, y) in the unlabeled dataset satisfies $\tilde{\theta}^\top x_{2,d} = X$ for $X \sim \mathcal{N}(0,1)$. Hence, since the first coordinate is distributed according to a truncated Gaussian with mean $y\mu$ and a standard deviation of σ truncated at 0, we have that $\hat{d}_q$ is largest for $\hat{\theta}(\mathcal{D}_{seed}) = \theta^\star = e_1$. Then, for an $\eta_2 > 0$ if $n_u > n_q/(0.8\eta_2)$ and $\sigma > \eta_2\mu + t$, then $\hat{d}_q < (1-\eta_2)\mu$ with a probability larger than 0.8. Using the bound on $\hat{d}_q$, we find that

$$\frac{M_{US}}{M_{PL}} < 1-\eta \iff n_\ell\mu + n_q(-\eta_2\mu + \sqrt{2\alpha_{seed}\log n_u}) < (1-\eta)n_\ell\mu$$

Inserting the uniform bound on α_{seed}, we get

$$\frac{M_{US}}{M_{PL}} < 1-\eta \iff n_q(2\log n_u)^{1/2}(d/n_{\text{seed}}+1)^{1/4} < (-\eta n_\ell + n_q\eta_2)\mu^{3/2}$$

Hence, if $\eta_2 > \frac{n_\ell}{n_q}\eta$, $\sigma > \mu\eta_2$, $n_q/(0.8\eta_2) < n_u < e^{\mathcal{O}(\eta^2 d/n_\ell + \eta\sqrt{d/n_\ell})}$ and

$$\mu > \frac{1}{(-\eta n_\ell + n_q\eta_2)^{2/3}} n_q^{2/3}(2\log n_u)^{1/3}(d/n_{\text{seed}}+1)^{1/6}$$

then with non-trivial probability uncertainty sampling performs worse than uniform sampling.

Appendix D. Dependence on the seed set

In this Section, we discuss the dependence on the seed set. In Theorem 2, we characterize the test error of uncertainty, oracle uncertainty and uniform sampling in function of the seed set and the unlabeled set. While the theorem hence explains the choice of the practitioner, i.e. to use uncertainty sampling or not, we want to understand how the distributional parameters σ and μ affect the seed set and the choice.

There are exactly two parameters in Theorem 2 that are fully determined by the seed set: $d_q^\star, \hat{d}_q$ and α_{seed}. We first define $d_q^\star, \hat{d}_q$ and α_{seed} after which we discuss each parameter for both (two-stage) uncertainty and oracle uncertainty sampling. We can define $d_q^\star, \hat{d}_q$ equivalently as follows. Suppose we use the classifier induced by θ for the uncertainty estimation. Then $\hat{d}_q$ is the largest distance of the queried sample to the decision boundary defined by θ. For θ we have three cases:

1. For oracle uncertainty estimation, we have $\theta = e_1$.
2. For two-stage uncertainty estimation, we have $\theta = \hat{\theta}(\mathcal{D}_{seed})$.
3. For regular uncertainty estimation, we have that θ changes at any iteration and is at iteration n given by $\hat{\theta}(\mathcal{D}_\ell^n)$.

Secondly, α_{seed} is defined as the α-parameter of the classifier based on the seed set, i.e. $\hat{\theta}(\mathcal{D}_{seed})$.

Dependence on the seed set in the case of oracle uncertainty sampling Note that the expression for oracle uncertainty sampling in Theorem 2 is independent of α_{seed}. We have that $d_q^\star$ is exactly the maximum distance to e_1 of the queried samples. The distance to e_1 is fully defined by the first coordinates, which are truncated Gaussian distributed. Clearly, the larger σ the smaller $d_q^\star$ as more samples in $\mathcal{D}_u$ are close to e_1. Recall that we define the CDF of the one-sided Gaussian Φ_{tr} as in Equation 7. Then, for a $\beta > 0$, with a probability of at least

$$1 - \sum_{i=0}^{n_q-1} \binom{n_u}{i} \Phi_{tr}(\beta)^i (1 - \Phi_{tr}(\beta))^{n_u - i}$$

we have that $d_q^\star < \beta$.

Dependence on the seed set in the case of two-stage uncertainty sampling In the case of two-stage uncertainty sampling, the expression is dependent on both α_{seed} and $\hat{d}_q$. We first discuss α_{seed} after which we discuss $\hat{d}_q$.

Observe that if $\alpha_{seed} \to \infty$, then the classifier induced by $\hat{\theta}(\mathcal{D}_{seed})$ is independent of the first coordinate. In that case, two-stage uncertainty sampling samples points uniformly at random in the first coordinate. However, if $\hat{\theta}(\mathcal{D}_{seed})$ achieves reasonable accuracy, then α_{seed} must be rather small. Note that the upper bound on passive learning also holds for α_{seed} where one sets $n_\ell = n_{\text{seed}}$. We now consider $d_q^\star$.

We note that for a unlabeled large dataset, where the points are spread out (σ not too small), we can always find points near the decision boundary of $\hat{\theta}(\mathcal{D}_{seed})$. Indeed for preciseness, we find that for an $(x, y) \in \mathcal{D}_u$

$$P[|\hat{\theta}(\mathcal{D}_{seed})^\top x| < \beta] = \frac{1}{2\pi\tilde{\sigma}(\alpha_{seed})\sigma(1 - \Phi(-\mu/\sigma))} \int_0^\infty \int_{-t-\beta}^{-t+\beta} e^{-\frac{(t-\mu)^2}{2\sigma^2}} e^{-\frac{l^2}{2\tilde{\sigma}(\alpha_{seed})^2}} dl dt$$

Using the expression on $P[|\hat{\theta}(\mathcal{D}_{seed})^\top x| < \beta]$ we note that with a probability of

$$1 - \sum_{i=0}^{n_q-1} \binom{n_u}{i} P[|\hat{\theta}(\mathcal{D}_{seed})^\top x| < \beta]^i (1 - P[|\hat{\theta}(\mathcal{D}_{seed})^\top x| < \beta])^{n_u - i}$$

we have that $\hat{d}_q < \beta$. We see clearly that for increasing n_u the probability exponentially increases. Moreover, for larger σ, we have that $P[|\hat{\theta}(\mathcal{D}_{seed})^\top x| < \beta]$ is larger as well.

Appendix E. Proofs of Lemmas

In this section, we give the proofs of Lemmas C.1, C.3 and C.2.

E.1 Proof of Lemma C.1

The proof of Lemma C.1 consists of two parts. In the first part we lower bound the extremal singular values of the data matrix of $\mathcal{D}_\ell$ 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 bounds on the minimal/maximal max l_2-margin, which concludes the proof.

Part 1: bounding the singular values We prove part 1 in two steps: first, we recall bounds on the extremal singular values of random normal matrices. Then, we use the union bound to compute bounds on the extremal singular values of non-random subsets of columns of random matrices. For convenience, let $X_u \in \mathbb{R}^{d \times n_u}$ be the data matrix of the dataset $\mathcal{D}_u$. Observe that by definition all entries of X_u are independent standard normal distributed.

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

$$\sqrt{d} - \sqrt{n_\ell} - t \leq s_{\min}(X_\ell) \leq s_{\max}(X_\ell) \leq \sqrt{d} + \sqrt{n_\ell} + t \tag{10}$$

with a probability greater than $1 - 2e^{-t^2/2}$. Note that there are exactly $m = \frac{n_u!}{(n_u - n_\ell)! n_\ell!} \leq n_u^{n_\ell}$ ways how a sampling strategy can choose a set of n_ℓ columns from X_u. Denote by $X_{\ell,i}$ the standard normal matrix of size $\mathbb{R}^{d \times n_\ell}$ induced by the $i - th$ subset. Using the union bound we get

$$\begin{aligned} P\left[\max_{i \in [m]} s_{\max}(X_{\ell,i}) > (\sqrt{2 \log n_u} + 1)\sqrt{n_\ell} + \sqrt{d} + t\right] \\ \leq m P\left[s_{\max}(X_\ell) > \left(\sqrt{2 \log n_u} + 1\right)\sqrt{n_\ell} + \sqrt{d} + t\right] \end{aligned} \tag{11}$$

We can simplify the expression using Equation 10 as follows

$$\begin{aligned} m P\left[s_{\max}(X_\ell) > (\sqrt{2 \log n_u} + 1)\sqrt{n_\ell} + \sqrt{d} + t\right] &\leq e^{\log 2m} e^{|(\sqrt{2 \log n_u}\sqrt{n_\ell} + t)^2/2} \\ &= e^{\log(m) + \log(2) - n_\ell \log(n_u) - \sqrt{2 \log n_u}\sqrt{n_\ell} - t^2/2} \\ &\leq e^{-t^2/2} \end{aligned} \tag{12}$$

This concludes the upper-bound on the maximum singular value of the data matrix. 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.

Part 2: Bounding the max-ℓ_2-margin 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 the dataset is given by

$$\gamma = \max_{\theta \in \mathcal{S}^d} \min_{(x,y) \in \mathcal{D}_\ell} y \theta^\top x. \tag{13}$$

We can rewrite the max-ℓ_2-margin using the data matrix X_ℓ.

$$\begin{aligned} \gamma = &\max_{\theta \in \mathcal{S}^d} b \\ &\text{subject to } \theta^\top X_\ell \geq b \mathbb{1}_{n_\ell}, \end{aligned} \tag{14}$$

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 larger than 0, i.e. there exists a $b > 0$. Hence, there exists a vector $v \in \mathbb{R}^{n_u}$, such that every entry is larger or equal than 1 and $\hat{\theta}^\top X_\ell \geq bv$. By the Eckat-Young-Mirsky theorem, the solution of $\hat{\theta}$ must lie in the space spanned by the singular vectors corresponding to the nonzero singular values. Hence, as $\|\hat{\theta}\|_2 = 1$, we have that $s_{\min}(X_\ell) \leq \|\gamma v\|_2 \leq s_{\max}(X_\ell)$. Since all entries of v are larger or equal to 1, we find that

$$\gamma \leq \frac{s_{\max}(X_\ell)}{\sqrt{n_\ell}}, \tag{15}$$

which using the upper bound on the singular value of Part 1 of the theorem yields the upper bound on the max-ℓ_2-margin. For the minimal margin, note that $\gamma\|v\|_2 = s_{\min}(X_\ell)$ implies a minimal margin of

$$\gamma \geq \frac{s_{\min}(X_\ell)}{\sqrt{n_\ell}} \tag{16}$$

Plugging in the bounds on the minimal singular values of the data matrix from Part 1 of the proof yields the Lemma.

E.2 Proof of Lemma C.3

We split the proof of Lemma C.3 in two parts. In the first part we characterize the form of the max-ℓ_2-margin classifier and in the second part we bound b.

The error of any classifier induced by a vector θ is invariant to a scaling. Hence, we can write the max-ℓ_2-margin classifier in the form

$$\hat{\theta} = [\hat{\theta}_1, b\tilde{\theta}], \tag{17}$$

with $\|\tilde{\theta}\|_2 = 1$. For convenience of notation we define $a_i = y_i\tilde{\theta}^\top x_{i,2:d}$ to be the distance of each sample x_i to the decision boundary of the max-ℓ_2-margin in the last $d-1$ coordinates. By definition of support points, the distance of all points in the dataset to the max-ℓ_2-margin classifier is equal. Therefore, the max-ℓ_2-margin is given by

$$\begin{aligned}\gamma_t &= \frac{1}{\|\hat{\theta}\|_2 n_\ell}\sum_{i=1}^{n_\ell} y_i\hat{\theta}_1 x_{1,i} + by_i\tilde{\theta}x_{i,2:d} \\ &= \frac{1}{\sqrt{\hat{\theta}_1^2 + b^2}}\bar{x}\hat{\theta}_1 + b\bar{a}\end{aligned} \tag{18}$$

where we defined $\bar{x}$ as the average signal component in the first coordinate of the dataset and $\bar{a}$ the average distance to the decision boundary of the max-ℓ_2-margin classifier in the $d-1$ last coordinates. As γ_t is the max-ℓ_2-margin, we need to optimize b and $\hat{\theta}_1$ for γ_t. We get that $\hat{\theta}_1 = \bar{x}$ and $b = \bar{a}$. Hence, the max-ℓ_2-margin is in the form of

$$\hat{\theta} = [\bar{x}, \bar{a}\tilde{\theta}]. \tag{19}$$

It is left to prove that $\gamma_{\text{avg}} \geq \bar{a} \geq \gamma_{d-1}$.

By definition the maximum average-ℓ_2-margin γ_{avg} is bigger than the average margin of the max-ℓ_2-margin $\bar{a}$. Let $\epsilon_i > 0$ be such that $x_{1,i} = \bar{x}\epsilon_i$. By definition of the average, we

find that $\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\epsilon_i = 1$. Analogously, define $\eta_i \in \mathbb{R}$ such that $y_i\tilde{\theta}^\top x_{i;2:d} = \bar{a}\eta_i$, which also yields that $\frac{1}{n_\ell}\sum \eta_i = 1$. Then, using that support points have equal margin to the classifier, we find for any $x \in \mathcal{D}_\ell$ that

$$y\hat{\theta}^\top x = \bar{x}^2\epsilon_i + \bar{a}^2\eta_i = \bar{x}^2 + \bar{a}^2. \tag{20}$$

Further, let θ_{d-1} be the max-ℓ_2-margin classifier and γ_{d-1} the max-ℓ_2-margin in the $d-1$ last coordinates of the dataset. Then, there exists an orthonormal matrix $Q \in \mathbb{R}^{(d-1)\times(d-1)}$ such that $Q\tilde{\theta} = \theta_{d-1}$. Moreover define the vector $\eta = [\eta_1, ..., \eta_{n_\ell}]$ and let $X_{d-1} \in \mathbb{R}^{(d-1)\times n_\ell}$ be the data matrix of the dataset in the last $d-1$ coordinates where each column is multiplied by the respective label of the sample. By the definition of Q, we have that

$$Q\tilde{\theta}^\top X_{d-1} = Q\bar{a}\eta = \gamma_{d-1}\mathbb{1}_{n_\ell}. \tag{21}$$

Comparing the norms yields

$$\bar{a}\|Q\eta\|_2 = \sqrt{n_\ell}\gamma_{d-1} \iff \frac{\|\eta\|_2}{\sqrt{n_\ell}} = \frac{\gamma_{d-1}}{\bar{a}} \tag{22}$$

Since $\sum_{i=1}^{n_\ell}\eta_i = n_\ell$, we have that $\|\eta\|_2 \geq \sqrt{n_\ell}$. Hence, $\bar{a} > \gamma_{d-1}$ and the Lemma is proven.

E.3 Proof of Lemma C.2

The max average-ℓ_2-margin is defined as

$$\begin{aligned}\gamma_{\text{avg}} &= \max_{\theta\in\mathcal{S}^{d-1}} \frac{1}{n_\ell}\sum_{i=1}^{n_\ell} y_i\theta^\top x_i \\ &= \max_{\theta\in\mathcal{S}^{d-1}} \frac{\theta_1}{n_\ell}\sum_{i=1}^{n_\ell} y_i x_{i,1} + ... + \frac{\theta_d}{n_\ell}\sum_{i=1}^{n_\ell} y_i x_{i,d}\end{aligned} \tag{23}$$

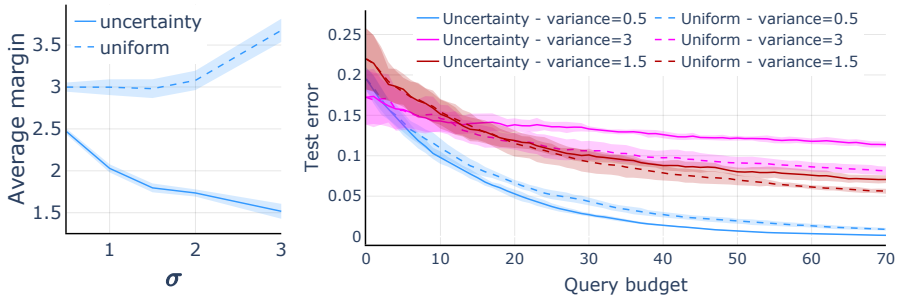
Since all x_i are independent standard normal distributed random variables, we have that $\frac{1}{n_\ell}\sum_{i=1}^{n_\ell} x_{i,j}$ is a normal distributed random variable with a variance of $\frac{1}{\sqrt{n_\ell}}$ and mean 0. Using standard probability theory, we find that for every sum we the absolute value is greater than $\frac{1}{\sqrt{n_\ell}} + \frac{t}{\sqrt{d}}$ with a probability smaller than $2\Phi(-\sqrt{\frac{n_\ell}{d}}t - \frac{1}{d})$. Now, recall that $\|\theta\|_2 = 1$, then we find that with a probability smaller than $2d\Phi(-\sqrt{\frac{n_\ell}{d}}t - \frac{1}{d})$ the average margin is larger than $\sqrt{\frac{d}{n_\ell}} + t$, which concludes the proof.

Appendix F. Further synthetic experiments

In this section, we give the experimental details to the synthetic experiments in Figures 2 and 2b. Moreover, we further empirically discuss the dependency on the standard deviation parameter σ and the μ for the truncated mixture model.

F.1 Experimental details

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}



(a) Average margin in σ

(b) Test error for varying σ

Figure 5: (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 Theorem 1. 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.

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.

We now give the experimental details for the experiments on the mixture of Gaussians. For the experiments in Figure 2c on the truncated , we take $d = 20000$, $n_u = 10^4$, $n_{\text{seed}} = 4$, $\sigma = 5$ and $\mu = 10$ and query for 70 samples. For the theoretical lines, we take $\hat{d}_q$ small enough. To estimate M_{or} in Figure 2b, we set $d = 10k$, $\mu = 20$, $\sigma = 10$ and estimate $d_q^\star$ by taking the average over 10 sets. In Figures 5 and 6, we set $d = 1000$, $\sigma = 3$,$\mu = 3$, $n_u = 10^5$ and $n_{\text{seed}} = 10$ unless specified otherwise on the figure.

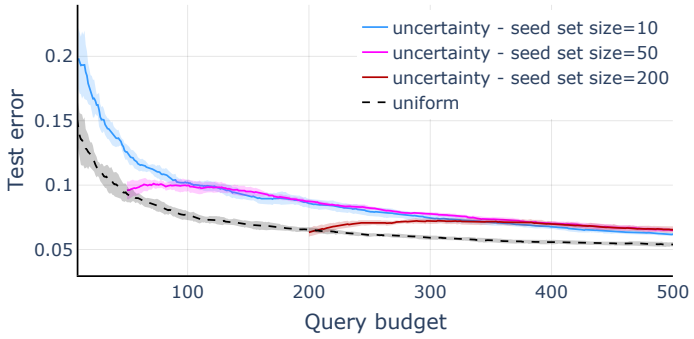


Figure 6: We set $d = 1000$, $\mu = 3$ and $n_u = 10^5$. The shaded areas indicate one standard deviation bands around the mean error, computed over 5 random draws of the seed set. Observe that for increasing seed size the gap between uncertainty and uniform sampling closes, but does not vanish. Note 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.

F.2 Dependence on the standard deviation σ and seed size n_{seed}

For the truncated Gaussian mixture model, we illustrate the dependence on the variance and the seed size 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 5a 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 5b we observe that, as predicted by Theorem 1, 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 6, 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 $d/4$.

Appendix G. Experiment details

G.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 G.3.

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 under these favorable circumstances, the performance of uncertainty sampling suffers in high dimensions. For completeness, we also compare uncertainty sampling and passive

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.

learning on the original, uncurated datasets in Appendix H.1 and observe similar trends as in this section.

G.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 H.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.[3] 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).

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

G.3 More dataset statistics

In this section we provide details about the real-world datasets that we consider in our experimental study. 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,

3. 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.

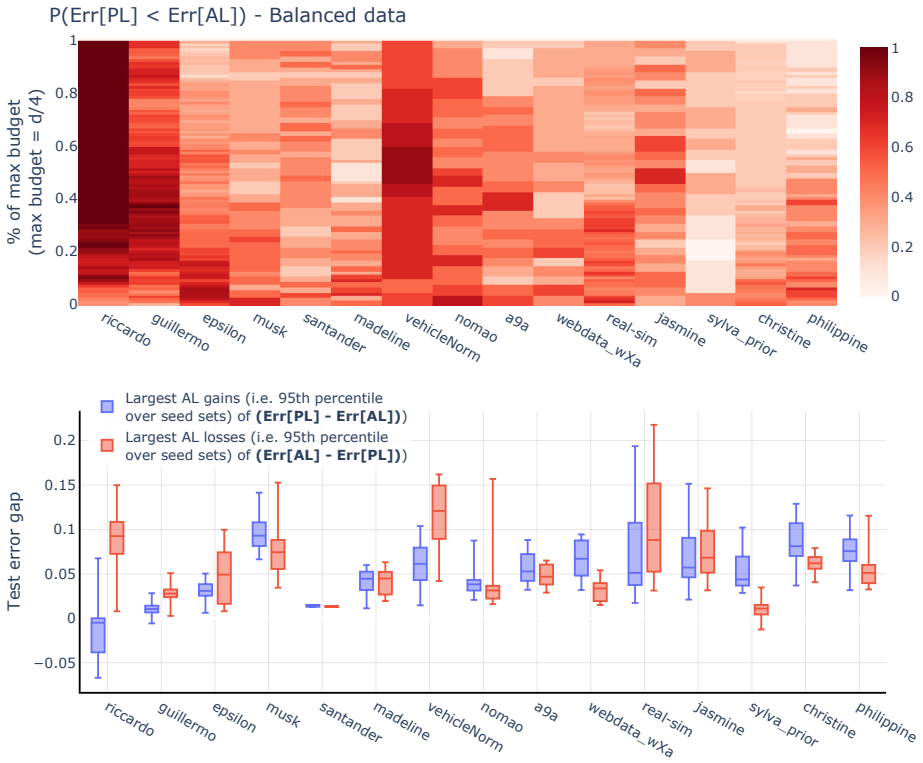


Figure 7: **Top:** The probability that the test error is lower with uniform sampling than with uncertainty sampling, over 10 draws of the seed set. Data is class-balanced, but potentially not linearly separable. **Bottom:** For the range of budgets where uncertainty sampling does poorly with high probability, 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 class-balanced, but potentially not linearly separable.

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.

Appendix H. Additional logistic regression experiments

H.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.

Balanced, but non-linearly separable data. As indicated in Appendix G.3, 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.

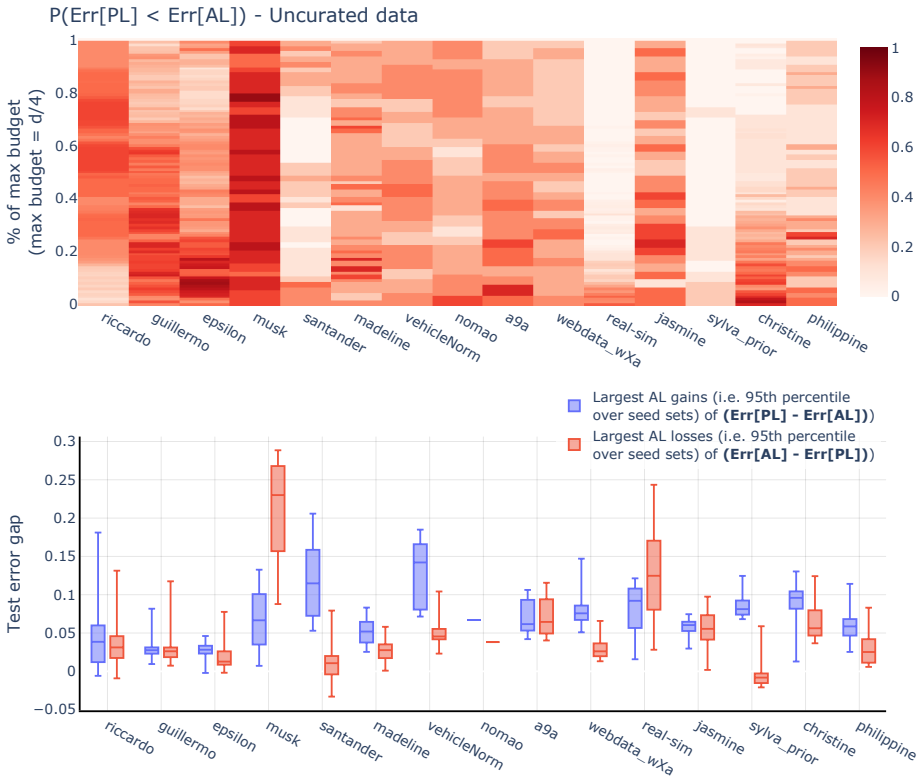


Figure 8: **Top:** The probability that the test error is lower with uniform sampling than with uncertainty sampling, over 10 draws of the seed set. Data is potentially class-imbalanced and not linearly separable. **Bottom:** For the range of budgets where uncertainty sampling does poorly with high probability, 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.

In Figure 7 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 8 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. Even when the unlabeled data is imbalanced (see Appendix G.3 for the exact imbalance ratio of each dataset), uncertainty sampling tends to select more points from the minority class, and hence, it collects a more balanced labeled set which then allows for training a classifier with better accuracy. We leave as future work a more thorough analysis of this phenomenon.

H.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 9 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 $\min(3000, d/4)$.

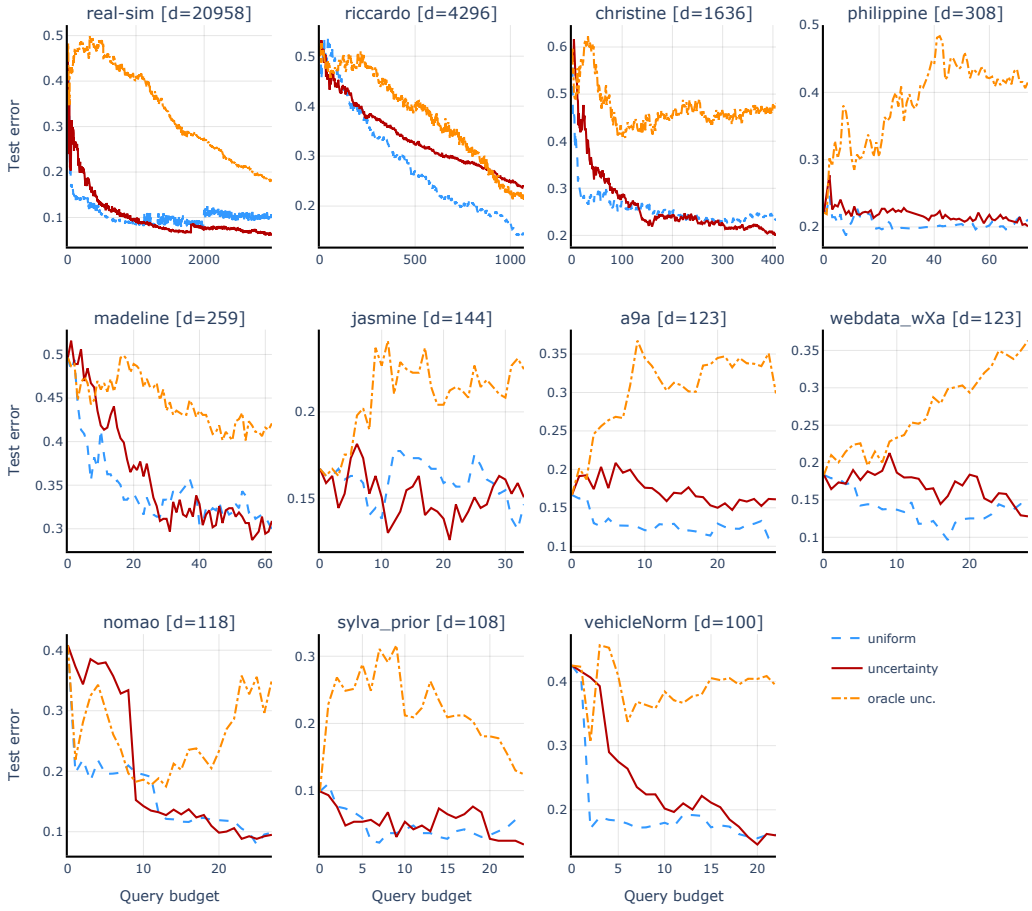


Figure 9: The test error using uncertainty sampling (with or without an oracle uncertainty estimate) is often higher than what is achieved with uniform sampling, for all the datasets that we consider.

H.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 (Figure 10). The gap between uncertainty and uniform sampling is even more significant when using the oracle uncertainty estimate, which

is in line with the intuition provided in Section 2. Oracle uncertainty sampling will select samples close to the ground truth (i.e. the yellow points in Figure 2a). Hence, the decision boundary of the classifier trained on the labeled set collected with active learning will be tilted compared to the ground truth, as long as the query budget is significantly smaller than the dimensionality.

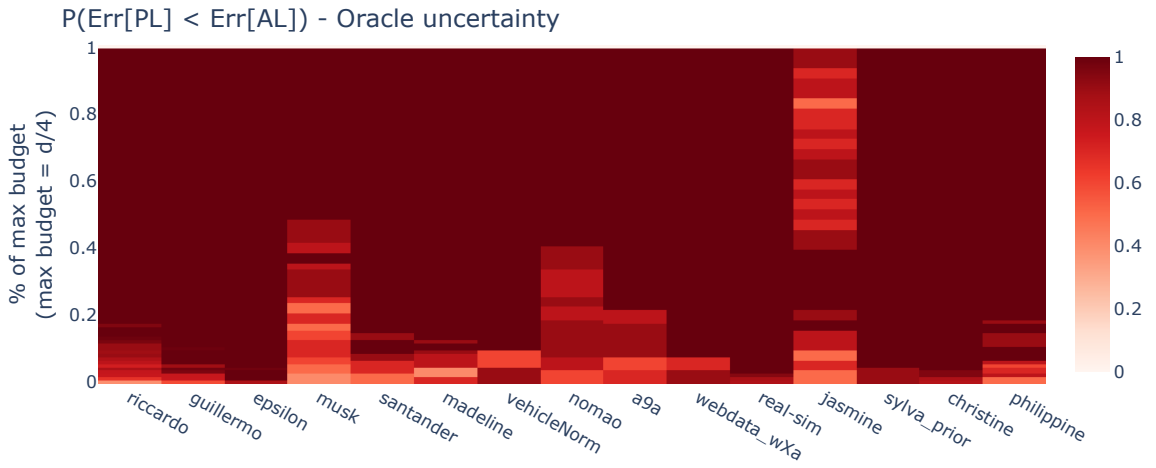


Figure 10: The probability that the test error is lower with uniform sampling than with **oracle** uncertainty sampling, over 10 draws of the seed set. Oracle uncertainty sampling performs consistently worse than passive learning (the dark red regions).

H.4 More details regarding 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 11 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. We note that the spikes in the lines in Figure 11 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 12 we summarize each of the panels in Figure 11 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.

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

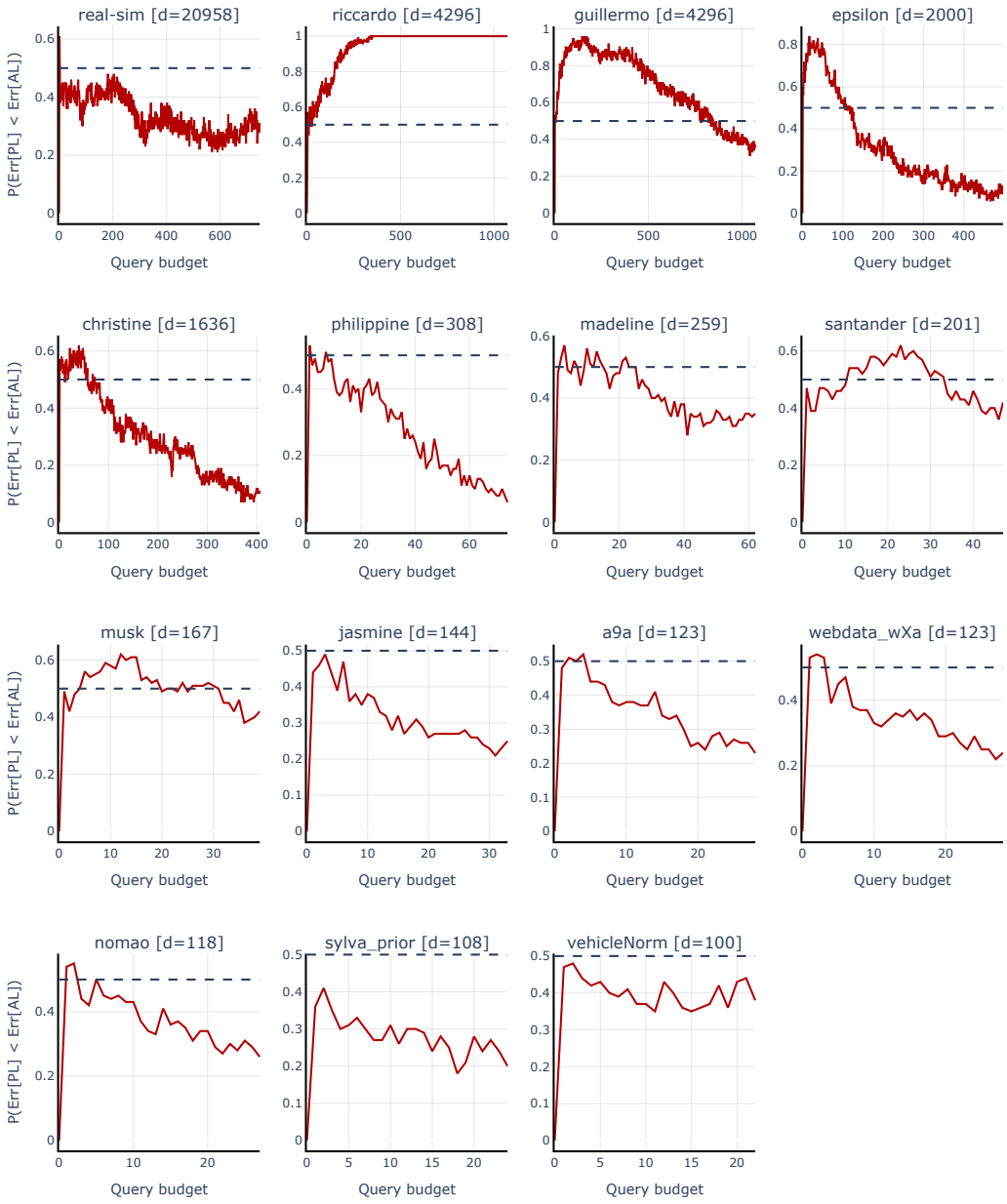


Figure 11: The probability that active learning via uncertainty sampling performs worse than passive learning at different query budgets. The empirical probability is computed over 100 draws of the initial seed set.

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 11 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

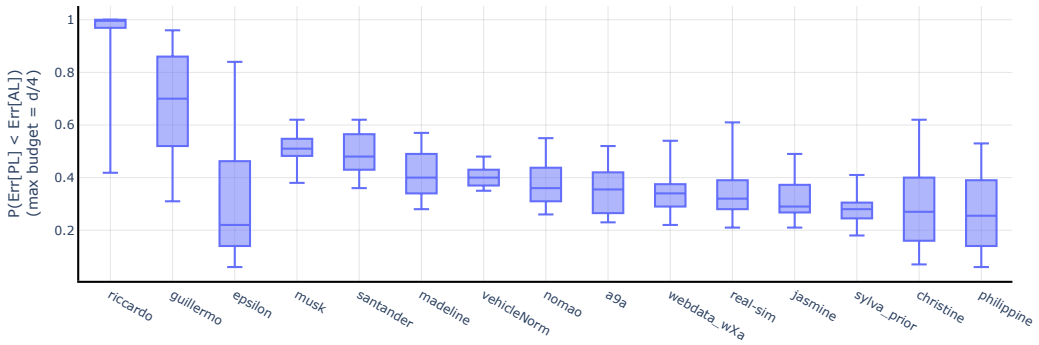


Figure 12: Box plot of the distribution of P(Err[PL] $<$ Err[AL]) over query budgets $n_q \in \{n_{\text{seed}}, ..., d/4\}$.

more advantageous for uncertainty sampling, as larger query budgets tend to lead to larger gains over uniform sampling.

H.5 Fraction of budgets for which active learning underperforms

An alternative to using the metric illustrated in Figure 3-Top and in Appendix H.4 is to instead compute the fraction of the query budgets for which active learning performs worse than passive learning. In Figure 13 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 $d/4$.

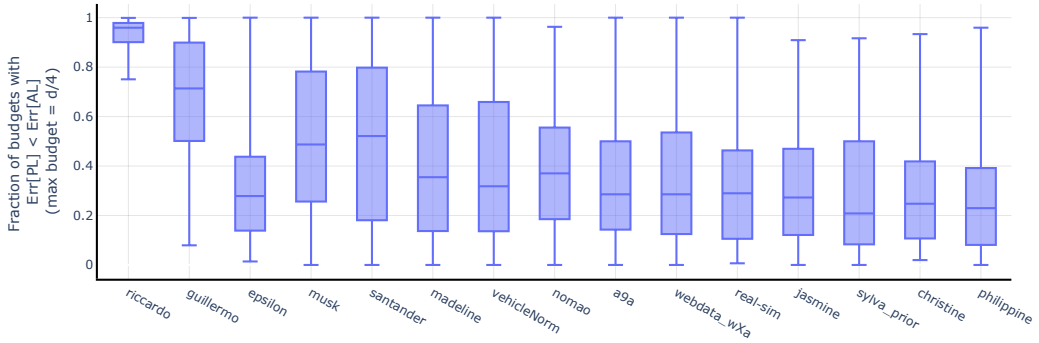


Figure 13: Fraction of the query budgets between n_{seed} and $d/4$ for which the error with uncertainty sampling is worse than with uniform sampling. The box plot indicates the distribution over 100 draws of the seed set (median, lower and upper quartiles).

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.

H.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. Note that a smaller coefficient C corresponds to stronger regularization.

Figures 14 and 15 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.

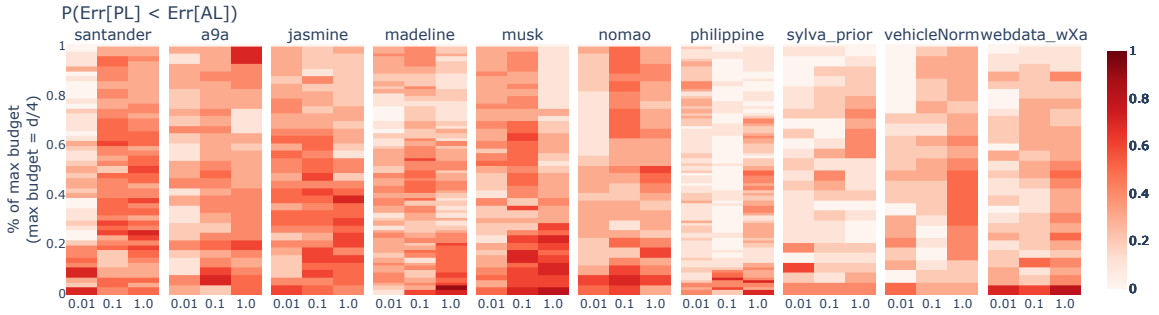


Figure 14: The probability that the test error is lower with uniform sampling than with uncertainty sampling, over 10 draws of the seed set. We use an ℓ_2**-regularized** classifier for both prediction and uncertainty estimation.

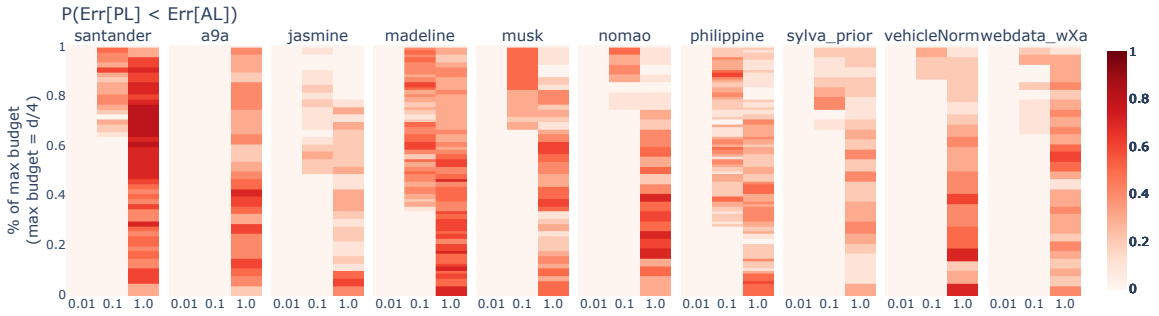


Figure 15: The probability that the test error is lower with uniform sampling than with uncertainty sampling, over 10 draws of the seed set. We use an ℓ_1**-regularized** classifier for both prediction and uncertainty estimation.

H.7 Experiments with different seed set sizes

In Figure 16 we present experiments on real-world datasets where we vary the size of the initial seed set. 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.

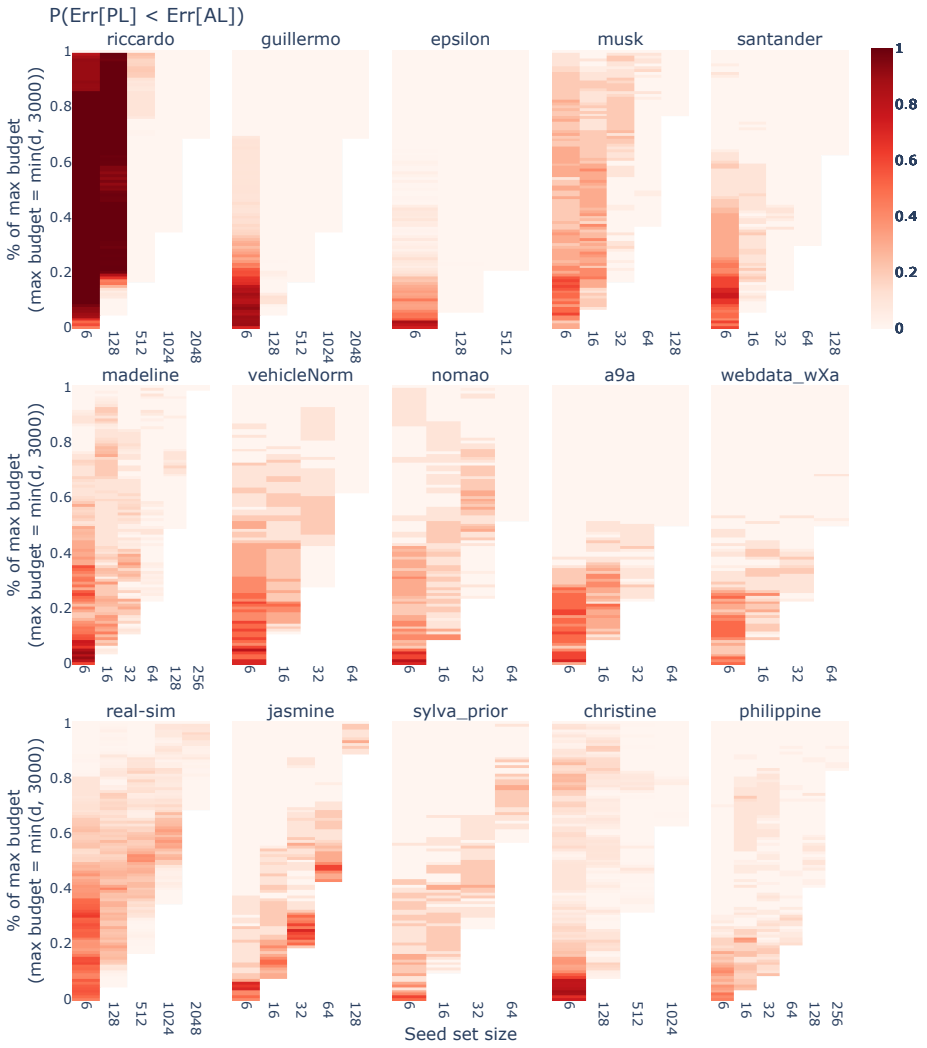


Figure 16: As predicted by our theory, increasing the seed size leads to improved performance when using uncertainty sampling to acquire new labeled samples.

H.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 Huang et al. (2014); Yang et al. (2015); Gissin and Shalev-Shwartz (2019); Shui et al. (2020).

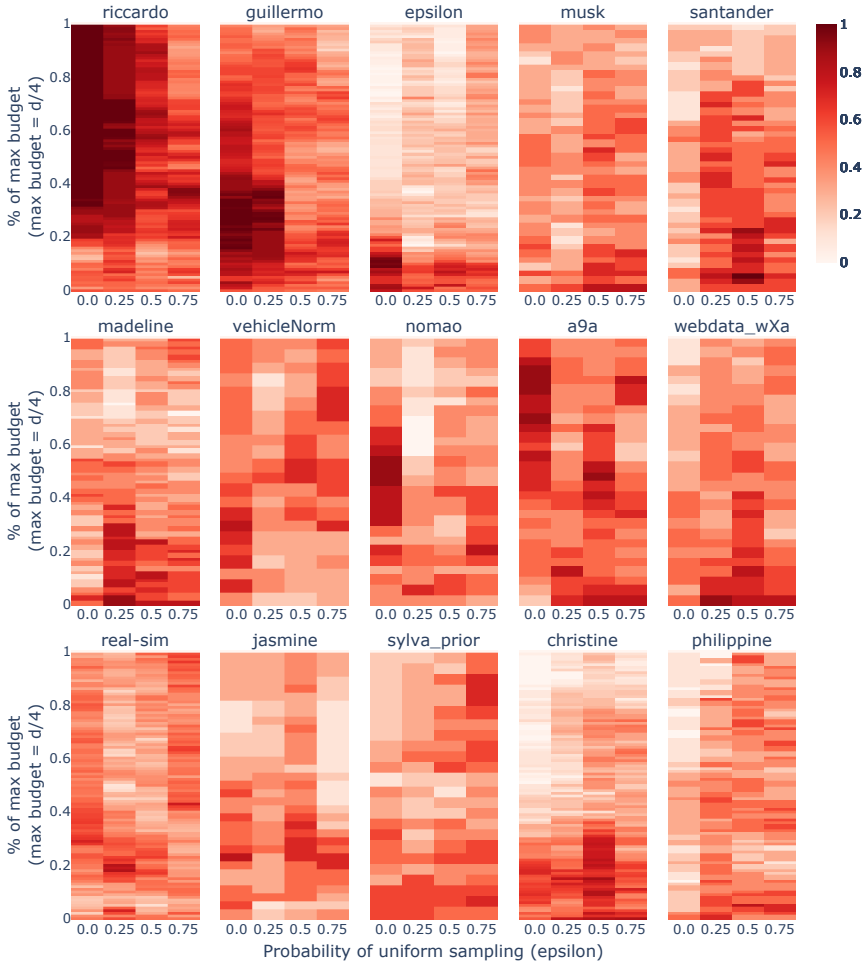


Figure 17: The probability that the test error is lower with uniform sampling than with an ϵ-greedy sampling approach, over 10 draws of the seed set. The active learning strategy preforms uncertainty sampling, with probability $1-\epsilon$ and samples uniformly at random with probability ϵ.

We notice in Figure 17 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).

H.9 Coreset-based active learning

In this section we investigate whether the coreset-based sampling strategy proposed in Sener and Savarese (2018) can perform better than uniform sampling in low-sample regimes. We follow the same active learning methodology as described in Section 3, but use the greedy algorithm from Sener and Savarese (2018) to select queries. We use the Euclidean distance for our experiments.

Figure 18 shows that for a large fraction of query budgets, passive learning tends to have lower error than coreset-based active learning. We hypothesize that this behavior is due to not constraining the queried points to lie far from the ground truth decision boundary. Hence, the high-dimensional phenomenon that we describe in Section 2 still occurs.

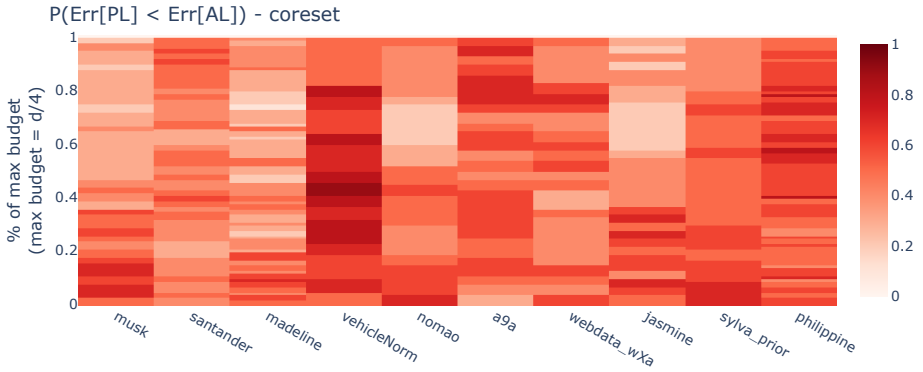


Figure 18: The coreset strategy of Sener and Savarese (2018) also fails to outperform passive learning consistently in the low-sample regime.

Appendix I. 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.

I.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.

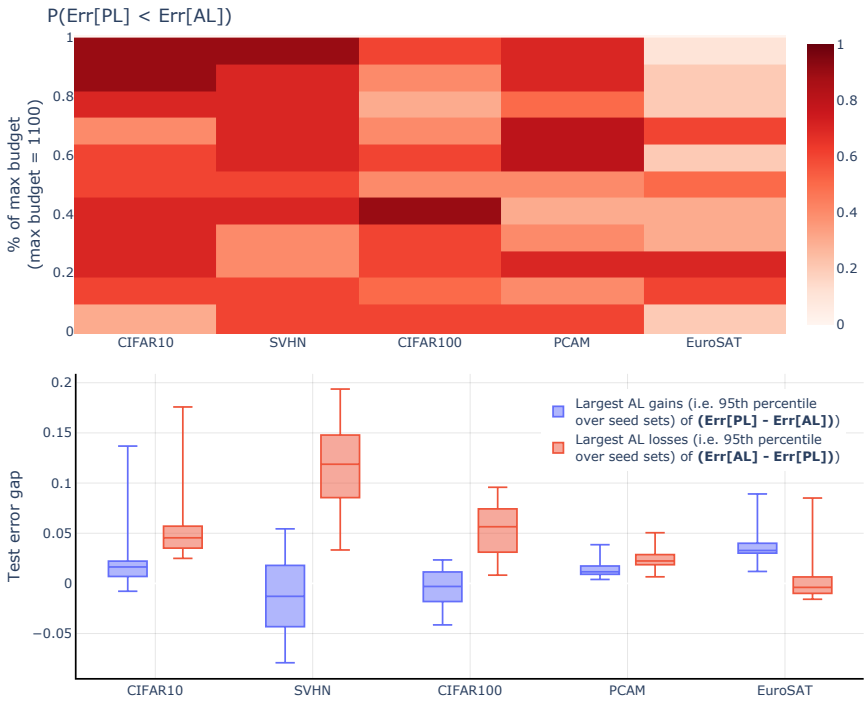


Figure 19: **Top:** The probability that the test error is lower with uniform sampling than with uncertainty sampling, over for 10 different random seeds. Uniform sampling outperforms uncertainty sampling, for a significant fraction of the querying budgets and for all datasets (i.e. dark red regions). **Bottom:** For the range of budgets where uncertainty sampling 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 (negative values indicate that PL is always better than AL).

I.2 Summary of results

As illustrated in Figure 19, uncertainty sampling leads to significantly larger test error compared to passive learning. This phenomenon persists even when we use an oracle uncertainty (Figure 20). Moreover, the gains that uncertainty sampling can produce, are often dominated by the losses that it can incur.

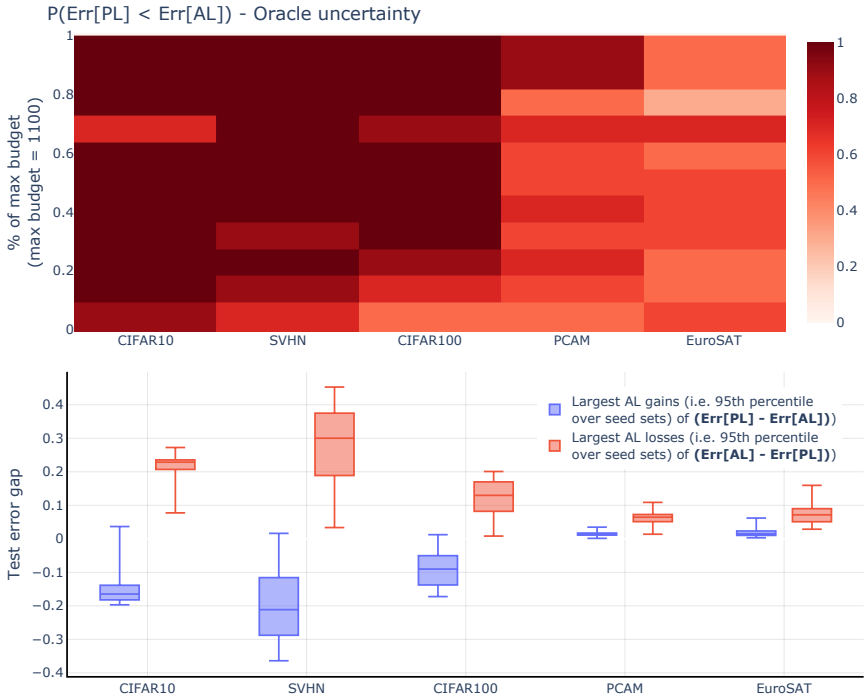


Figure 20: Same experiment as in Figure 19, but this time using **oracle uncertainty**. Similar to the logistic regression experiments, uncertainty sampling leads to even worse error when using oracle uncertainty, as also predicted by our theory.

References

Alina Beygelzimer, Daniel J Hsu, John Langford, and Tong Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems*, 2010.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. *Proceedings in Advances in Neural Information Processing Systems 28*, 2015.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, 2005.

Konstantin Donhauser, Alexandru Tifrea, Michael Aerni, Reinhard Heckel, and Fanny Yang. Interpolation can hurt robust generalization even when there is no noise. In *Proceedings in Advances in Neural Information Processing Systems 34*, 2021.

Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effects of inductive bias. *arXiv preprint arXiv:2203.03597*, 2022.

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

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

Spencer Frei, Difan Zou, Zixiang Chen, and Quanquan Gu. Self-training converts weak learners to strong learners in mixture models. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

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

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022.

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

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *arXiv preprint arXiv:1709.00029*, 2017.

Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *Proceedings in Advances in Neural Information Processing Systems 27*, 2014.

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

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

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

David D. Lewis and William 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.

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

Rafid Mahmood, Sanja Fidler, and Marc T Law. Low budget active learning via Wasserstein distance: An integer programming approach. *arXiv preprint arXiv:2106.02968*, 2021.

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

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

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew 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.

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

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

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

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

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

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 2012.

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

Burr Settles. Active learning literature survey. 2009.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Proceedings in Advances in Neural Information Processing Systems 7*, 2007.

Changjian Shui, Fan Zhou, Christian Gagné, and Boyu 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.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 2018.

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

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. *SIGKDD Explorations*, 2013.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. *arXiv preprint arXiv:1806.03962*, 2018.

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

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

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 2015.