# Shift-Pessimistic Active Learning Using Robust Bias-Aware Prediction

**Anqi Liu**
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607
aliu33@uic.edu

**Lev Reyzin**
Department of Mathematics,
Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607
lreyzin@math.uic.edu

**Brian D. Ziebart**
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607
bziebart@uic.edu

## Abstract

Existing approaches to active learning are generally optimistic about their certainty with respect to data shift between labeled and unlabeled data. They assume that unknown datapoint labels follow the inductive biases of the active learner. As a result, the most useful datapoint labels—ones that refute current inductive biases—are rarely solicited. We propose a shift-pessimistic approach to active learning that assumes the worst-case about the unknown conditional label distribution. This closely aligns model uncertainty with generalization error, enabling more useful label solicitation. We investigate the theoretical benefits of this approach and demonstrate its empirical advantages on probabilistic binary classification tasks.

## Introduction

Obtaining large amounts of labeled data is prohibitively expensive for many learning tasks. Producing each label could require an invasive medical test and expensive expert knowledge, for example. Active learning (Settles 2012) aims to alleviate this burden by soliciting labels from datapoints that reduce prediction error (loss) by the greatest amount. This has the potential to significantly improve data-efficiency beyond what is possible with randomly provided labels (Angluin 1988). However, data produced from an active learner violates the independent and identically distributed (IID) data property broadly assumed by supervised machine learning techniques (Sugiyama and Kawanabe 2012). This poses serious pitfalls for active learning methods both in theory and in practice that have not yet been resolved.

Existing active learning methods are generally optimistic about their own uncertainty with respect to the sample bias in the labeled data, which is often an intentional byproduct of the active learner's label solicitation strategy. They employ an underlying supervised machine learning model and assume that all unlabeled datapoints' labels are distributed according to the model's (often strong) inductive biases. This approach has theoretical justification in IID settings where the inductive biases are shaped by increasing amounts of *representative* data. However, the potential of improved data-efficiency benefits from active learning is only realized by biasing label solicitation towards *non-representative* data that is the most informative (Settles 2012). Unfortunately, the combination of optimistic extrapolation based on IID assumptions and intentionally non-IID data collection often leads not only to inefficient learning, but also to extreme inaccuracies. Even advocates of popular active learning methods suggest that "random sampling ... may be more advisable than taking one's chances on active learning with an inappropriate learning model" (Settles 2012).

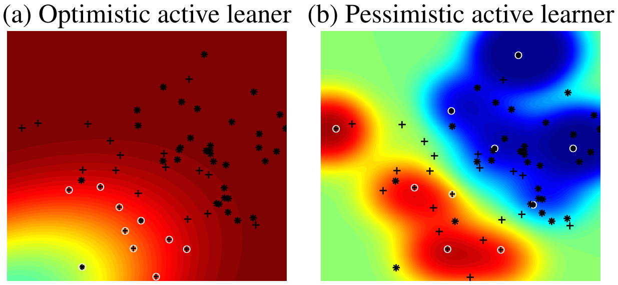

Figure 1: A binary prediction task with + and * classes. Probabilistic predictions ranging from dark red (+) to dark blue (*) are shown after 10 examples solicited (white circles) from active learning using: (a) a standard optimistic approach—uncertainty sampling (Lewis and Gale 1994) with logistic regression; and (b) uncertainty sampling using our more pessimistic robust bias-aware (RBA) active learner. The optimistic active learner exhaustively explores a false decision boundary and extrapolates its predictions to other portions of the input space, incurring large prediction loss. The shift-active learner avoids being misled, and provides predictions with significantly smaller loss.

The perils of optimistic active learning extend far beyond the more benign risks of model misspecification. For example, a logistic regression model fits the synthetic dataset of Figure 1 *in its entirety* with fairly small average prediction loss. However, the optimistic active learner solicits a sequence of labels that does not uncover this appropriately fit model, as shown in Figure 1a. This is primarily because it solicits labels for examples that would be the most useful *if* its current inductive biases were correct. After obtaining an initial '+'-class label and a noisy second '*'-class label from the bottom-left-most datapoint, the active learner forms an incorrect inductive bias—that a decision boundary for the dataset as a whole exists between those two datapoints—and exhaustively solicits labels to better define the belief contours of this incorrect decision boundary. As shown in this example, the active learner typically avoids soliciting the most informative labels, which would strongly refute the active learner's current inductive biases.

In this paper, by explicitly considering active learning as a special case of covariate shift, we develop a pessimistic approach to active learning that avoids inefficiencies created by the combination of optimism and non-representative label solicitation. Our approach leverages a recently developed model for learning from biased source sample data by assuming the worst-case about the unknown conditional label distribution (Liu and Ziebart 2014). Under this approach, we show that model uncertainty is closely calibrated to generalization loss. Thus, common label solicitation strategies guided by model uncertainty tend to directly improve the model's predictive performance. In addition to these theoretical properties, we evaluate and compare the effectiveness of our approach on a range of classification tasks. Figure 1b shows the key difference from previous methods: the limited, more *pessimistic* extrapolation from available labeled data providing smaller prediction loss.

## Background and Previous Work

### Active Learning

A pool-based active learner (Lewis and Gale 1994) sequentially chooses datapoint labels to solicit from a set (pool) of unlabeled datapoints, $(x_i) \in \mathcal{U}$. It constructs an estimate of the conditional label distribution, $\hat{P}(y|x)$, from its labeled dataset $(x_j, y_j) \in \mathcal{L}$. It uses this estimate to select the next datapoint label to solicit. We denote the entire set of labeled and unlabeled datapoints as $\mathcal{D} = \mathcal{U} \cup \mathcal{L}$.

Numerous metrics have been developed to assess the expected utility of a datapoint. The most common, *uncertainty sampling* (Lewis and Gale 1994; Settles 2012), solicits datapoint labels for which the active learner is least certain. The value-conditioned entropy, $H(Y|X = x_i) \triangleq E_{\hat{P}(y|x)}[-\log \hat{P}(Y|X)|x_i] = -\sum_{y\in\mathcal{Y}} \hat{P}(y|x_i) \log \hat{P}(y|x_i)$, often measures this uncertainty. Other metrics assess how a datapoint label: (a) is expected to change the prediction model (Settles and Craven 2008); (b) reduces an upper bound on the generalization error in expectation (Mackay 1992); or (c) represents the input patterns of remaining unlabeled data (Settles 2012).

Yet, as illustrated in Figure 1, the pool-based active learning algorithm often performs poorly in practice (Attenberg and Provost 2011). Ad-hoc modifications to the algorithm that limit the power of the active learner—undermining the purported benefits of active learning—are often required for existing active learners to be competitive with random sampling. These modifications decrease the potential for bias in the labeled dataset by making the label solicitation strategy more similar to random sampling. One modification is to "seed" the learner with a set of randomly drawn datapoint labels (Schein and Ungar 2007; Dligach and Palmer 2011). In other words, the active learner is restricted to sampling uniformly for its first $n$ datapoints. A second modification solicits labels from a very small random subset of the unlabeled dataset (e.g., a pool of 10 examples (Schein and Ungar 2007)) rather than the entire unlabeled dataset, $\mathcal{U}$. These modifications treat the symptoms resulting from optimistic modeling and non-IID label solicitation rather than its cause. Our approach differs from methods that are pessimistic about active learning from a bandit learning perspective (Rokach, Naamani, and Shmilovici 2008).

### Learning under sample selection bias

Inherent sample selection bias exists in active learning because examples for label solicitation are not chosen uniformly at random (Sugiyama and Kawanabe 2012). However, since the active learner can only select examples based on the input values, $x_i$, independently from the unknown label, $y_i$, this corresponds to a special case of sample selection bias known as *covariate shift* (Shimodaira 2000). This setting requires that a common conditional label distribution, $P(y|x)$, is shared in both source, $P_{\text{src}}(y, x) = P(y|x)P_{\text{src}}(x)$, and target, $P_{\text{trg}}(y, x) = P(y|x)P_{\text{trg}}(x)$, distributions. Learning under covariate shift estimates this shared conditional label distribution from sample source data, $\tilde{P}_{\text{src}}(y, x)$, using predictor $\hat{P}(y|x)$, with the goal of minimizing target distribution loss, $\mathbb{E}_{P_{\text{trg}}(x)P(y|x)}[-\log \hat{P}(Y|X)]$. In the pool-based active learning setting, the source distribution represents available labeled data $\mathcal{L}$ and the target distribution represents the combination of labeled and unlabeled data[1] $\mathcal{D}$.

Minimizing the prediction loss for a target distribution that differs from the source data's distribution is difficult. In IID settings, researchers typically minimize the empirical loss of labeled sample data (Sugiyama and Kawanabe 2012). With increasing data, this sample-based approximation converges linearly to the source/target distribution's error. Unfortunately, target sample data is not available to directly measure the empirical loss in the biased setting. Importance sampling (Hammersley and Morton 1954) is a prevalent approach for this setting that estimates the target distribution loss by reweighting the source samples according to the target-source density ratio, $P_{\text{trg}}(x)/P_{\text{src}}(x)$ (Shimodaira 2000; Zadrozny 2004). In the asymptotic limit (infinite amounts of source data), the predictor (with parameters $\theta$) minimizing the reweighted loss is equivalent to the predictor that minimizes the target loss (Shimodaira 2000),

$$\lim_{m\to\infty} \min_{\theta} \mathbb{E}_{\tilde{P}^{(m)}_{\text{src}}(x)\tilde{P}(y|x)} \left[\frac{P_{\text{trg}}(X)}{P_{\text{src}}(X)} \text{loss}\left(Y, \hat{f}_\theta(X)\right)\right]$$
$$= \min_{\theta} \mathbb{E}_{P_{\text{trg}}(x)P(y|x)} \left[\text{loss}\left(Y, \hat{f}_\theta(X)\right)\right]. \quad (1)$$

Despite this asymptotic guarantee, sample reweighting under large sample selection bias can converge very slowly to the target loss as the number of source datapoints ($m$) grows.

[1] The remaining unlabeled data alone could be used for the target distribution if generalization beyond the pool is not intended.

In fact, finite second moments on the target-source density ratio, $\mathbb{E}_{P_{\text{src}}(x)}\left[(P_{\text{trg}}(X)/P_{\text{src}}(X))^2\right] < \infty$, are required for any finite-sample generalization bounds (Cortes, Mansour, and Mohri 2010). Not satisfying this requirement leads to estimates with high variance and target distribution predictions with overly optimistic certainty for any finite amount of source data.

Active learning using sample reweighting has been investigated in a handful of learning tasks (Kanamori and Shimodaira 2003; Sugiyama 2005; Bach 2007). Unfortunately, common label solicitation strategies often produce labeled datapoint distributions that are highly non-representative, as shown in Figure 1. These source distributions lack finite second moment density ratios and produce high-variance predictions from small amounts of data. Extensions to the streaming active learning setting (Beygelzimer, Dasgupta, and Langford 2009) randomize the label solicitation strategy to improve sample complexity bounds.

# Approach

## Robust bias-aware prediction

We employ a recently-developed approach for robust prediction in settings with dataset shift (Liu and Ziebart 2014). It provides *robust bias-aware* (RBA) predictions in the active learning setting for data distributed according to the full data distribution $P_{\mathcal{D}}(x)$, given labeled data samples (denoted with empirical measure $\tilde{P}_{\mathcal{L}}(x)$) treated as being drawn from a labeled data distribution $P_{\mathcal{L}}(x)$. A minimax estimation problem (the primal) and a regularized maximum likelihood estimation problem (the dual) provide equivalent solutions to the formulation.

The primal problem is a minimax game between estimator choosing $\hat{P}(y|x)$ and constrained adversary, choosing evaluation distribution $\check{P}(y|x)$:

$$\min_{\hat{P}(y|x)} \max_{\check{P}(y|x)\in\tilde{\Xi}} \mathbb{E}_{P_{\mathcal{D}}(x)\check{P}(y|x)}\overbrace{[-\log \hat{P}(Y|X)]}^{\text{logarithmic loss}}. \qquad (2)$$

The set $\tilde{\Xi}$ constrains the adversary to (approximately) match a set of its statistics, $\mathbb{E}_{\tilde{P}_{\mathcal{L}}(x)\check{P}(y|x)}[\mathbf{f}(X,Y)]$ with sample statistics $\mathbb{E}_{\tilde{P}_{\mathcal{L}}(x)\tilde{P}(y|x)}[\mathbf{f}(X,Y)]$ from the labeled data distribution. The dual problem selects model parameters, $\theta$, by maximizing a regularized full data distribution likelihood:

$$\theta^* = \operatorname*{argmax}_{\theta} \mathbb{E}_{P_{\mathcal{D}}(x)P(y|x)}[\log \hat{P}_\theta(Y|X)] - \lambda||\theta||, \qquad (3)$$

with the conditional label distribution estimate's form as:

$$\hat{P}_\theta(y|x) = e^{\frac{P_{\mathcal{L}}(x)}{P_{\mathcal{D}}(x)}\theta\cdot\mathbf{f}(x,y)} \Big/ \sum_{y'\in\mathcal{Y}} e^{\frac{P_{\mathcal{L}}(x)}{P_{\mathcal{D}}(x)}\theta\cdot\mathbf{f}(x,y')}. \qquad (4)$$

The density ratio, $\frac{P_{\mathcal{L}}(x)}{P_{\mathcal{D}}(x)}$, moderates the predictions to be less certain wherever the labeled data underrepresents the full data distribution and more certain wherever the labeled data overrepresents it. The uncertainty of this distribution closely matches to its generalization error (Theorem 1).

**Theorem 1.** *Assuming that the actual label distribution $P(y|x)$ is within the set $\tilde{\Xi}$, the full data entropy of the RBA predictor upper bounds its generalization loss:*

$$\begin{aligned} H_{\mathcal{D}}(Y|X) &\triangleq \mathbb{E}_{P_{\mathcal{D}}(x)\hat{P}(y|x)}\left[-\log \hat{P}(Y|X)\right] \qquad (5)\\ &\geq \mathbb{E}_{P_{\mathcal{D}}(x)P(y|x)}[-\log \hat{P}(Y|X)]. \end{aligned}$$

*Proof.* The proof follows from Grünwald and Dawid (2004) and Topsøe (1979) using: (a) strong duality; (b) the equivalence of the logloss minimizer to its evaluation distribution when given; and (c) the assumption that $P(y|x)$ is in set $\tilde{\Xi}$:

$$\begin{aligned} &\min_{\hat{P}(y|x)} \max_{\check{P}(y|x)\in\tilde{\Xi}} \mathbb{E}_{P_{\mathcal{D}}(x)\check{P}(y|x)}[-\log \hat{P}(Y|X)]\\ &\overset{(a)}{=} \max_{\check{P}(y|x)\in\tilde{\Xi}} \min_{\hat{P}(y|x)} \mathbb{E}_{P_{\mathcal{D}}(x)\check{P}(y|x)}[-\log \hat{P}(Y|X)]\\ &\overset{(b)}{=} \max_{\hat{P}(y|x)\in\tilde{\Xi}} H_{\mathcal{D}}(Y|X) \overset{(c)}{\geq} \mathbb{E}_{P_{\mathcal{D}}(x)P(y|x)}[-\log \hat{P}(Y|X)]. \end{aligned}$$

□

Constructing a constraint set $\tilde{\Xi}$ from finite sample data to satisfy the premise of Theorem 1 is overly restrictive. Instead, we can relax the guarantee to be probabilistic based on finite sample error bounds in Corollary 1.

**Corollary 1.** *When $\delta$ defining the $\ell_1$-norm or $\ell_2$-norm constraint set,*

$$\left|\left|\mathbb{E}_{\tilde{P}_{\mathcal{L}}(x)\hat{P}(y|x)}[\mathbf{f}(X,Y)] - \mathbb{E}_{\tilde{P}_{\mathcal{L}}(x)\tilde{P}(y|x)}[\mathbf{f}(X,Y)]\right|\right| \leq \delta,$$

*is chosen using sample error bounds between the labeled data distribution's sample statistics and expected statistics,*

$$P\left(\left|\left|\mathbb{E}_{\tilde{P}_{\mathcal{L}}(x)\hat{P}(y|x)}[\mathbf{f}(X,Y)] - \mathbb{E}_{\tilde{P}_{\mathcal{L}}(x)\tilde{P}(y|x)}[\mathbf{f}(X,Y)]\right|\right| \geq \delta\right) \leq \alpha,$$

*then the bound* (5) *of Theorem 1 holds with probability at least* $(1-\alpha)$.

The constraint set slack $\delta$ corresponds to $\ell_1$ or $\ell_2$ regularization weight $\lambda$ in the dual optimization problem (3) (Dudík and Schapire 2006).

Training the RBA predictor appears difficult because the dual objective function (3) maximizes the log likelihood of the full data, which is partially unlabeled. However, that objective function's gradient is based on labeled data distribution statistics,

$$\begin{aligned} &\mathbb{E}_{P_{\mathcal{L}}(x)P(y|x)}[\mathbf{f}(X,Y)] - \mathbb{E}_{P_{\mathcal{L}}(x)\hat{P}(y|x)}[\mathbf{f}(X,Y)] - \lambda\nabla_\theta||\theta||\\ \approx\ &\mathbb{E}_{\tilde{P}_{\mathcal{L}}(x)\tilde{P}(y|x)}[\mathbf{f}(X,Y)] - \mathbb{E}_{\tilde{P}_{\mathcal{L}}(x)\hat{P}(y|x)}[\mathbf{f}(X,Y)] - \lambda\nabla_\theta||\theta||, \end{aligned}$$

which can be safely approximated using labeled data distribution samples from $\tilde{P}_{\mathcal{L}}(x)$.

## Density estimation

The degree that information from labeled data generalizes to other portions of the input space in the RBA approach is controlled by the density estimates of the labeled data distribution, $P_{\mathcal{L}}(x)$, and the entire data distribution, $P_{\mathcal{D}}(x)$. If the labeled data distribution estimate provides minimal support beyond the labeled data samples, density predictions outside of the labeled samples will tend to be overly conservative and maximally uncertain. If the labeled data distribution estimate provides too broad of support for the full data distribution, the guarantees of Corollary 1 will be improbable (i.e., a large $\alpha$ will be required). If the full data distribution is misestimated, the prediction guarantees (Corollary 1) will not apply to actual full data distribution samples.

Density estimation methods have been investigated extensively in the importance weighting approach (1) to sample selection bias (Shimodaira 2000; Dudík, Schapire, and Phillips 2005; Huang et al. 2006; Sugiyama et al. 2008; Bickel, Brückner, and Scheffer 2009; Yu and Szepesvári 2012; Wen, Yu, and Greiner 2014). In that model, these importance weights, formed from density estimates, shape the strong inductive biases of the approach. Rather than investigate the benefits of each density estimation technique for RBA prediction, we employ a widely accepted density estimation technique when generalized estimates are needed: kernel density estimation with Gaussian kernels.

We leverage specific properties of the active learning setting to help alleviate some of the potentially negative consequences of inaccurate density estimation. We narrow our focus to minimizing the loss on a specific set of full dataset distribution samples (i.e. all labeled and unlabeled datapoints). Thus, we employ the uniform distribution of datapoints,

$$P_{\mathcal{D}}(x) = \begin{cases} \frac{1}{|\mathcal{D}|} & \text{if } x \in \mathcal{D} \\ 0 & \text{otherwise,} \end{cases}$$

to represent the full data distribution density. This would be ill-advised for general covariate shift prediction, because it would make the density ratio $\frac{P_{\mathcal{L}}(x)}{P_{\mathcal{D}}(x)}$ infinite (i.e, no "penalty" for overly certain predictions) at many labeled sample datapoints in $\mathcal{L}$. However, for the active learning setting, all labeled data samples will have support in the full distribution, since $\mathcal{L} \subseteq \mathcal{D}$, so this situation does not occur.

## Active learning using robust predictions

Conditional label distribution estimates guide label solicitation within an active learner as shown in Algorithm 1.

**Algorithm 1** Label solicitation for pool-based active learner with covariate shift correction

```
Input: unlabeled pool dataset $\mathcal{U}$, labeled dataset $\mathcal{L}$
Output: example $x_i \in \mathcal{U}$ to solicit label
  Estimate labeled distribution density $P_{\mathcal{L}}(x)$
  Estimate full data distribution density $P_{\mathcal{D}}(x)$ ($\mathcal{D} = \mathcal{U} \cup \mathcal{L}$)
  Estimate $\hat{P}(y|x)$ from dataset $\mathcal{L}$, $P_{\mathcal{L}}(x)$, and $P_{\mathcal{D}}(x)$.
  Compute value$_i \leftarrow$ metric($\hat{P}, x_i, \mathcal{D}, \mathcal{U}$) for each $x_i \in \mathcal{U}$
  return $x_{\text{argmax}_i \text{ value}_i}$ (example label to solicit)
```

Using uncertainty sampling and kernel density estimation for $P_{\mathcal{L}}(x)$, the complexity of each label solicitation is $O(|\mathcal{L}|^2 + Tk|\mathcal{L}| + k|\mathcal{U}|)$, where $k$ is the number of features and $T$ is the number of optimization steps.

The metrics used to evaluate different unlabeled datapoints for many label solicitation strategies, including uncertainty sampling, are heuristic/greedy methods for minimizing model uncertainty. Theorem 1 and Corollary 1 provide theoretical guarantees that this is an appropriate objective when using the RBA predictor.

**Corollary 2.** *An active learning strategy for RBA prediction that efficiently reduces RBA model uncertainty also efficiently reduces prediction loss.*

In contrast, we can consider the importance reweighting approach to estimating the conditional label distribution within an active learning algorithm. Cortes et al. (2010) show that even when importance weights are not bounded, under this approach, for any hypothesis $h$ the generalization loss $R(h)$ can be bounded as a function of the empirical importance-weighted loss $\hat{R}_w(h)$ as follows:

$$R(h) \le \hat{R}_w(h) + O\left(\sqrt{\mathbb{E}_{P_{\mathcal{L}}(x)}\left[w(X)^2\right]}\ \sqrt[3/8]{\frac{p\log(\frac{m}{p}) + \log(\frac{1}{\delta})}{m}}\right),$$

where $w(x) = P_{\mathcal{L}}(x)/P_{\mathcal{D}}(x)$, $\mathbb{E}_{P_{\mathcal{L}}(x)}\left[w(X)^2\right]$ is finite, $p$ is an upper bound on the pseudo-dimension, a notion of dimension for the hypothesis space, and $1-\delta$ is the bound confidence.

Popular label solicitation strategies (Settles 2012) tend to choose labels with the goal of greedily or approximately minimizing the (importance-weighted) empirical loss $\hat{R}_w(h)$. Often they do so without appropriately bounding the density ratio, $\mathbb{E}_{P_{\mathcal{L}}(x)}\left[(P_{\mathcal{D}}(X)/P_{\mathcal{L}}(X))^2\right]$, which is needed for the empirical loss to generalize from the subset of labeled datapoints to the rest of the dataset. For reasonably large confidence values $1-\delta$, this bound can be looser than agnostic predictions (uniform over labels with logloss of $\log_2|\mathcal{Y}|$) in such cases.

Unfortunately, minimizing the prediction loss for a non-representative labeled data distribution provides no guarantees for the prediction loss on the broader data distribution (Sugiyama and Kawanabe 2012). Thus, active learners that minimize the uncertainty of the logistic regression model should instead solicit labels from representative datapoints to provide any theoretical performance guarantees.

# Experiments

## Classification tasks

We evaluate the performance of different active learning approaches using four datasets from the UCI repository (Bache and Lichman 2013). We consider datasets with real-valued features to simplify density estimation for methods that address covariate shift. We reduce multi-class datasets to binary classification tasks by merging classes (typically plurality class versus other) as detailed in Table 1.

Table 1: Datasets for empirical evaluation

| Dataset | Features | Examples | Positive labels | Negative labels |
|---|---|---|---|---|
| Iris | 4 | 150 | *Setosa* | all others |
| Seed | 7 | 210 | *Type "1"* | all others |
| Banknote | 4 | 1372 | *Class "0"* | *Class"1"* |
| E. coli | 8 | 336 | *Cytoplasm* | all others |

In each of our experiments, we divide the dataset into a training set (80% of data) and a testing set (the remaining 20%).

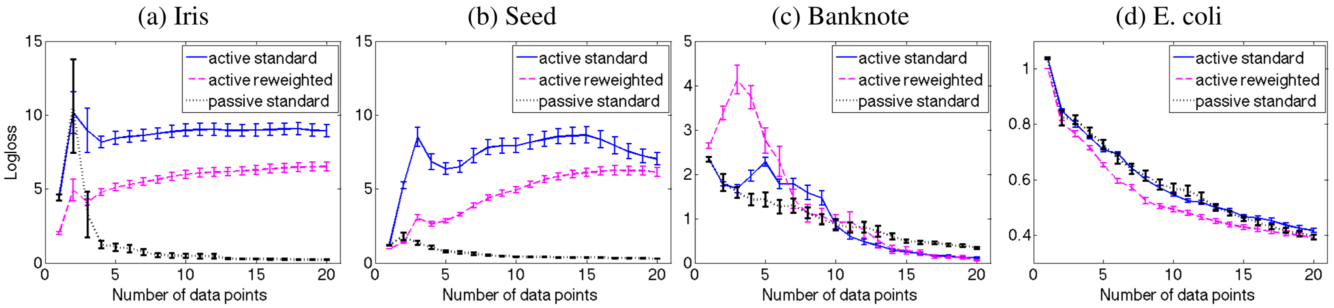

Figure 2: **Logloss of optimistic active learning versus passive (IID) learning** for the first 20 datapoints of learning averaged over 30 randomized withheld evaluation dataset splits with 95% confidence intervals.

## Learning methods

We apply three different models for estimating the conditional label distribution: **Standard logistic regression** (abbreviated as **standard** in this section) uses the Boltzmann distribution $P(y|x) = e^{\theta\cdot\mathbf{f}(x,y)}/(\sum_{y'\in\mathcal{Y}} e^{\theta\cdot\mathbf{f}(x,y)})$ and minimizes the logloss of the labeled distribution samples, $\min_\theta \mathbb{E}_{\tilde{P}_\mathcal{L}(x)\tilde{P}(y|x)}[-\log \hat{P}_\theta(Y|X)] + \lambda||\theta||$; **Sample reweighted logistic regression** (abbreviated as **reweighted**) uses the same logistic regression model, but with parameters estimated to minimize the importance weighted estimate of the target loss, $\min_\theta \mathbb{E}_{\tilde{P}_\mathcal{L}(x)\tilde{P}(y|x)}\left[-\frac{P_\mathcal{D}(X)}{P_\mathcal{L}(X)}\log\hat{P}_\theta(Y|X)\right] + \lambda||\theta||$ ; and **Robust bias-aware prediction** (abbreviated as **robust**) uses the conditional label distribution of (4) trained by maximizing target likelihood (3) (approximating the gradient with labeled datapoints).

We employ two label solicitation strategies for each model: **Uncertainty sampling** (abbreviated as **active**) selects the example with the largest value-conditioned entropy from the unlabeled dataset. The first datapoint label solicited is selected uniformly at random (the same first datapoint as passive learners); and **Random sampling** (abbreviated as **passive**) selects each datapoint uniformly at random from the unlabeled dataset. In addition, we apply a density-ratio-based strategy with our robust approach: **Density-ratio sampling** (abbreviated as **active density**) selects the example with the highest $\frac{P_\mathcal{D}(x)}{P_\mathcal{L}(x)}$ under the estimated distribution.

We conduct 30 experiments with each learner on randomized training/testing splits of each dataset and report the mean and the 95% confidence interval of the predictive performance after every data point solicited in the first 20 steps, corresponding to 0.05 significance level in student t-test. We focus on the first 20 examples because real applications require good predictive performance with limited labeled data.

## Density estimation and dimension reduction

We apply Gaussian kernel density estimation (KDE) on the labeled examples to estimate the labeled data density,

$$P_\mathcal{L}(x) = \frac{1}{|\mathcal{L}|}\sum_{x_i\in\mathcal{L}} K_\mathbf{H}(x - x_i)$$

with a bandwidth that minimizes the logloss on the whole dataset, $\mathbf{H}{=}\text{argmin}\,\mathbb{E}_{P_\mathcal{D}(x)}[-\log\hat{P}(X)]$, from a restricted set of bandwidths proportionate to a covariance estimate of the entire data, $\mathbf{H} = \alpha\hat{\Sigma}(\mathcal{D})$. For higher dimensional data (*Seed* and *E. coli*), we first apply principal component analysis to reduce the dimensionality to a space that covers at least 95% of the input variance, before applying Gaussian KDE. We use the uniform distribution over training and testing datapoints for the full data distribution density.

## Features and regularization

For all methods, we use first-order and second-order statistics of the inputs as features: $x_1^2y$, $x_2^2y, \ldots, x_K^2y$, $x_1x_2y, x_1x_3y, \ldots, x_{K-1}x_Ky$, $x_1y, x_2y, \ldots, x_Ky, y$. Since the regularization weight $\lambda$ corresponds to slack in the constraints (Corollary 1) and feature scales differ, we use a different regularization weight for each feature corresponding with the 95% confidence interval of the feature's mean, $2\sigma(\phi(x,y))/\sqrt{|\mathcal{L}|}$. However when the scale of the density ratio $(P_\mathcal{L}(x)/P_\mathcal{D}(x))$ is overwhelmingly large (*E.coli*) or small (*Banknote*), we reweight each feature's mean using the learning model's density ratio before taking the standard deviation in the reweighted and robust algorithms.

## Optimistic active learning versus IID learning

We first investigate how the optimistic active learning methods (active standard and active reweighted) compare to IID logistic regression (passive standard). In Figures 2a and 2b, the logloss of active standard and active reweighted are worse than passive learners for the entire 20 steps of learning with statistical significance. This is similarly the case for the active standard and active reweighted algorithms in the first 10 steps of learning in *Banknote* (Figure 2d). This frequent poor performance results from the active learners getting "stuck" soliciting labels suggested by its optimistic biases to be useful rather than labels that would correct its incorrect beliefs. Further prediction improvements often require first exhausting from the pool of examples that conform to the learner's incorrect beliefs. Only when the inductive biases of labeled data match those of the unlabeled data, as in active methods for the *E. coli* dataset, will the optimistic active learner not provide high logloss in the initial steps of active learning.

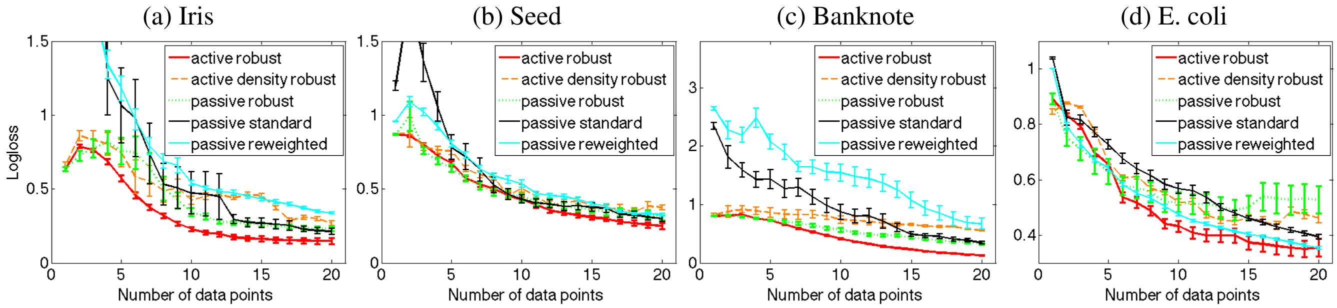

Figure 3: **Logloss of shift-pessimistic active learning versus passive (IID) learning** for the first 20 datapoints of learning averaged over 30 randomized withheld evaluation dataset splits with 95% confidence intervals.

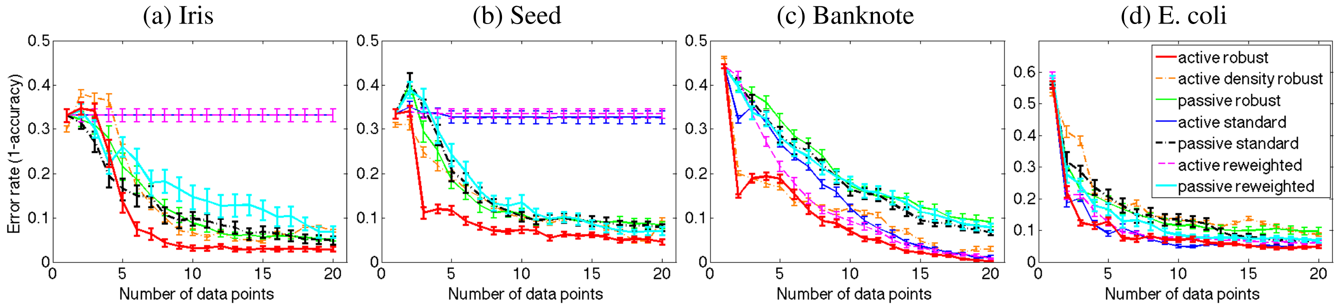

Figure 4: **Classification error rate of all learning methods** for the first 20 datapoints of learning averaged over 30 randomized withheld evaluation dataset splits. The legend is shared for all datasets. Active standard and active reweighted overlap in (a).

## Pessimistic active learning versus IID learning

We next compare the performance of our shift-pessimistic active learning method (active robust and active density robust) to several passive learning methods. As shown in Figure 3, active robust and active density robust perform better from the very beginning than agnostic baseline, which would provide a logloss of *1*, and are better than, or at least comparable to any other methods for all amounts of available data. Small error bars reflect high stability compared to other methods. In contrast, IID learning methods are quite unstable especially at the beginning due to the bias of a small, randomly chosen sample. Active density robust cannot significantly compete with passive robust because it only considers densities when soliciting labels. Passive robust outperforms passive standard and reweighted, which shows that robust bias-aware prediction effectively controls the extent to which the prediction should generalize. However, since the inductive biases from labeled data tend to generalize accurately using the passive standard and reweighted methods on the *E.coli* dataset, they exceed the passive robust method given 20 labeled examples.

## Comparing classification accuracy

Though all the algorithms do not minimize classification error directly, the log loss upper bounds the non-convex classification error (0-1 loss). Thus, one might expect that efficiently reducing log loss in the active learning setting will lead to low classification error. We investigate this in Figure 4, comparing the classification error rate of all seven methods on each dataset. The active robust approach provides the highest prediction accuracy for almost all amounts of available labeled data. In contrast, the high log loss predictions of active standard and active reweighted in *Iris* and *Seed* translate to poor classification error rates.

## Conclusions

We have focused on the often detrimental combination of: machine learning models that are optimistic in the face of their own uncertainty; and active learning strategies that unintentionally avoid soliciting labels that would refute optimistic extrapolations. We propose a new approach to constructing prediction models within an active learning algorithm by considering the problem as a covariate shift prediction task and adopting pessimism about all uncertain properties of the conditional label distribution. Theoretically, this aligns model uncertainty with prediction loss on remaining unlabeled datapoints, better justifying the use of the model's label estimates within active learning label solicitation strategies. We have shown that our RBA approach encounters lower prediction loss and accuracy than IID learning in settings where previous optimistic active learning methods are often not competitive with IID learning.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. #1227495.

## References

Angluin, D. 1988. Queries and concept learning. *Machine learning* 2(4):319–342.

Attenberg, J., and Provost, F. 2011. Inactive learning? Difficulties employing active learning in practice. *ACM SIGKDD Explorations Newsletter* 12(2):36–41.

Bach, F. R. 2007. Active learning for misspecified generalized linear models. In *Advances in Neural Information Processing Systems*, 65–72. MIT Press.

Bache, K., and Lichman, M. 2013. UCI machine learning repository.

Beygelzimer, A.; Dasgupta, S.; and Langford, J. 2009. Importance weighted active learning. In *Proceedings of the International Conference on Machine Learning*, 49–56. ACM.

Bickel, S.; Brückner, M.; and Scheffer, T. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research* 10:2137–2155.

Cortes, C.; Mansour, Y.; and Mohri, M. 2010. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, 442–450.

Dligach, D., and Palmer, M. 2011. Good seed makes a good crop: accelerating active learning using language modeling. In *Proceedings Annual Meeting of the Association for Computational Linguistics*, 6–10. Association for Computational Linguistics.

Dudík, M., and Schapire, R. E. 2006. Maximum entropy distribution estimation with generalized regularization. In *Learning Theory*. Springer Berlin Heidelberg. 123–138.

Dudík, M.; Schapire, R. E.; and Phillips, S. J. 2005. Correcting sample selection bias in maximum entropy density estimation. In *Advances in Neural Information Processing Systems*, 323–330.

Grünwald, P. D., and Dawid, A. P. 2004. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics* 32:1367–1433.

Hammersley, J. M., and Morton, K. W. 1954. Poor man's Monte Carlo. *Journal of the Royal Statistical Society. Series B (Methodological)* 23–38.

Huang, J.; Smola, A. J.; Gretton, A.; Borgwardt, K. M.; and Schlkopf, B. 2006. Correcting sample selection bias by unlabeled data. In *Neural Information Processing Systems*, 601–608.

Kanamori, T., and Shimodaira, H. 2003. Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of Statistical Planning and Inference* 116(1):149–162.

Lewis, D. D., and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *Proc. of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3–12. Springer-Verlag New York, Inc.

Liu, A., and Ziebart, B. D. 2014. Robust classification under sample selection bias. In *Advances in Neural Information Processing Systems*.

Mackay, D. 1992. The evidence framework applied to classification networks. *Neural Computation* 4(5):720–736.

Rokach, L.; Naamani, L.; and Shmilovici, A. 2008. Pessimistic cost-sensitive active learning of decision trees for profit maximizing targeting campaigns. *Data Mining and Knowledge Discovery* 17(2):283–316.

Schein, A. I., and Ungar, L. H. 2007. Active learning for logistic regression: an evaluation. *Machine Learning* 68(3):235–265.

Settles, B., and Craven, M. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1070–1079. Association for Computational Linguistics.

Settles, B. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6(1):1–114.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90(2):227–244.

Sugiyama, M., and Kawanabe, M. 2012. *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press.

Sugiyama, M.; Nakajima, S.; Kashima, H.; Buenau, P. V.; and Kawanabe, M. 2008. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 1433–1440.

Sugiyama, M. 2005. Active learning for misspecified models. In *Advances in Neural Information Processing Systems*, 1305–1312.

Topsøe, F. 1979. Information theoretical optimization techniques. *Kybernetika* 15(1):8–27.

Wen, J.; Yu, C.-N.; and Greiner, R. 2014. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, 631–639.

Yu, Y., and Szepesvári, C. 2012. Analysis of kernel mean matching under covariate shift. In *Proc. of the International Conference on Machine Learning*, 607–614.

Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proc. of the International Conference on Machine Learning*, 903–910.