Active Learning under Label Shift

Eric Zhao **Anqi Liu** **Anima Anandkumar** **Yisong Yue**
California Institute of Technology

Abstract

We address the problem of active learning under *label shift*: when the class proportions of source and target domains differ. We introduce a "medial distribution" to incorporate a tradeoff between importance weighting and class-balanced sampling and propose their combined usage in active learning. Our method is known as Mediated Active Learning under Label Shift (MALLS). It balances the bias from class-balanced sampling and the variance from importance weighting. We prove sample complexity and generalization guarantees for MALLS which show active learning reduces asymptotic sample complexity even under arbitrary label shift. We empirically demonstrate MALLS scales to high-dimensional datasets and can reduce the sample complexity of active learning by 60% in deep active learning tasks.

1 Introduction

Label Shift In many real-world applications, the target (testing) distribution of a model can differ from the source (training) distribution. Label shift arises when class proportions differ between the source and target, but the feature distributions of each class do not. For example, the problems of bird identification in San Francisco (SF) versus New York (NY) exhibit label shift. While the likelihood of observing a snowy owl may differ, snowy owls should look similar in New York and San Francisco. The well-known class-imbalance problem is a specific form of label shift where the target label distribution is uniform but the source is not.

Active Learning under Label Shift Label shift poses a problem for active learning in the real world.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

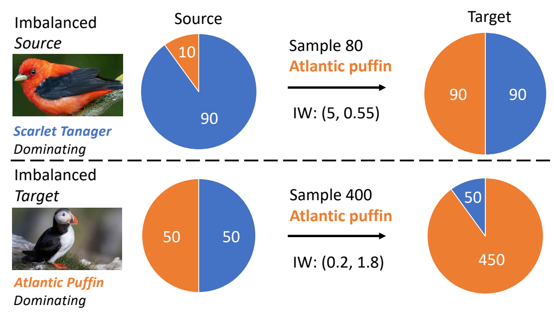


Figure 1: Extreme label shift examples of binary classification on 100 datapoints. The arrows in the example illustrate two ways of correcting *imbalanced source* and *imbalanced target*. It requires larger importance weights to correct *imbalanced source* than *imbalanced target*, but correcting *imbalanced target* requires more additional samples than *imbalanced source*. A uniform *medial distribution* can decompose any label shift into an *imbalanced source* and *imbalanced target*.

For example, we can train a bird classifier for New York by labeling bird images off Google. However, due to the label shift between New York and Google Images, naive active learning algorithms will fail to collect data on birds relevant in New York. The correction of minority underrepresentation in computer vision datasets [Yang et al., 2020] similarly poses an active learning under label shift problem. Proper label shift correction must be incorporated into active learning techniques to avoid inefficient and biased data collection.

Importance Weighting & Subsampling There are two ways to correct label shift, as shown in the two extreme cases depicted in Figure 1. The arrows demonstrate the required additional samples and importance weights for the correction of *imbalanced source* and *imbalanced target*. Importance weighting can correct for label shift with rigorous theoretical guarantees [Lipton et al., 2018, Azizzadenesheli et al., 2019]. However, under large label shift, the estimation and use of importance weights result in high variance. Class-balanced sampling (subsampling) can also correct for label shift and, although lacking strong theoretical guarantees, is practical and effective [Aggarwal et al., 2020].

However, in active learning settings, subsampling is imprecise as only label predictions—not true labels—can be used to assign subsampling probabilities to unlabeled datapoints.

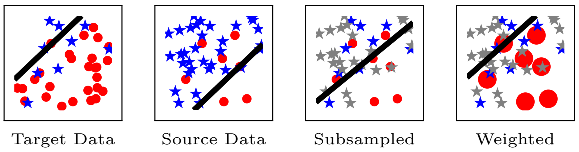


Figure 2: Noisy linear classification of stars and circles with a star-heavy source and circle-heavy target. Black lines depict the empirical risk minimizer (ERM). Ignored data are light grey. Our proposed algorithms first subsample a *medial distribution*—in this case, equal parts circle and star—then importance weighting produces a circle-dominant ERM. See sec. 5-6 for details.

In this paper, we answer the question: how can we use both importance weighting and subsampling for active learning—and how much should we use each? We answer this question by introducing a *medial distribution* (Figure 2). Rather than active learning on datapoints from the source distribution, datapoints are instead sampled from a *medial distribution* by subsampling. Importance weighting corrects the remaining label shift between the *medial* and target distributions.

Our contributions:

1. Introduction of a *medial distribution* to describe a bias-variance trade-off in label shift correction.
2. Mediated Active Learning under Label Shift (MALLS), a principled algorithm with theoretical guarantees even under label shift.
3. A batched variant of MALLS for practitioners which integrates best practices and uncertainty sampling.

Aggressive use of subsampling can reduce the need for, and thus variance of, importance weighting. However, subsampling also introduces bias from its use of proxy labels. We derive a bias-variance tradeoff that formalizes this trade-off and guides algorithm design. In particular, we show subsampling can mitigate the effect of label shift on importance weighting variance and label complexity—but at the cost of introducing bias. We further propose a choice of uniform medial distribution, as we illustrate in Figure 1.

To the best of our knowledge, MALLS is the first active learning framework for general *label shift* settings. We also derive label complexity and generalization PAC bounds for MALLS, the first such guarantees for this setting. We present experiments of MALLS which corroborate our theoretical insights into the trade-off between importance weighting and subsampling. In particular, batched MALLS improves the sample efficiency of popular active learning algorithms by up to 60% in the CIFAR10, CIFAR100 [Krizhevsky, 2009], and NABirds datasets [Van Horn et al., 2015]. We share the source code for the implementation of our method in this repository: `https://github.com/ericzhao28/alls`.

2 Related Works

Active Learning Active learning has been investigated extensively from both theoretical and practical perspectives. Disagreement-based active learning and its variants enjoy rigorous learning guarantees and focus on the stream-based active learning setting [Hanneke, 2007, Hanneke, 2011, Balcan et al., 2009, Hanneke, 2014, Beygelzimer et al., 2010, Krishnamurthy et al., 2019]. On the other hand, uncertainty sampling techniques are popular practical algorithms which have been successfuly applied to natural language processing [Shen et al., 2018], computer vision [Yang et al., 2015], and even robotics [Choudhury and Srinivasa, 2020]. We can incorporate our medial distribution design principle to arrive at both a streaming disagreement-based MALLS approach, as well as a Batched MALLS for uncertainty sampling.

Distribution Shift General domain adaptation theory [Ben-David et al., 2007, Ben-David et al., 2010, Cortes et al., 2010, Cortes and Mohri, 2014] looks at joint distribution shift. Covariate shift is the most popular refinement of joint distribution shift [Shimodaira, 2000, Gretton et al., 2009, Sugiyama et al., 2007]. However, density estimation for joint distribution shift or covariate shift is challenging due to the high-dimension nature of input features in many applications [Sugiyama et al., 2012, Tsuboi et al., 2009, Yamada et al., 2011]. The label shift setting is comparatively less popular, but has received increased attention in recent years [Lipton et al., 2018, Azizzadenesheli et al., 2019, Garg et al., 2020]. Density estimation under label shift is comparatively simpler than under covariate shift: label spaces are simpler and often finite [Lipton et al., 2018].

Active Learning under Distribution Shift Active learning [Rai et al., 2010, Matasci et al., 2012, Deng et al., 2018, Su et al., 2019] has been studied under joint distribution shift and covariate shift. Existing literature, which sometimes term the problem "active domain adaptation", rely on heuristics for correcting joint distribution shift [Chan and Ng, 2007, Rai et al., 2010] or build on the assumption of *covariate shift* [Saha et al., 2011, Yan et al., 2018, Chattopadhyay et al., 2013]. While active learning with a covariate-shifted warm start guarantees label complexity bounds, it requires importance weights known a priori [Yan et al., 2018]. Label shift is a particularly difficult setting as, unlike covariate shift, label shift cannot be estimated from unlabeled data.

With few exceptions [Huang and Chen, 2016], existing literature assume active learners can query datapoints from the test domain (our *canonical label shift* setting). To the best of our knowledge, MALLS provides the first guarantees for where test data is limited or labels cannot be queried in the test domain.

The closest existing work to active learning under label shift is active learning for imbalanced data [Aggarwal et al., 2020, Lin et al., 2018], which can be formalized as an instance of label shift with a uniform test distribution. While existing work have proposed useful heuristics like diverse sampling and class-balanced sampling, theoretical results are scarce.

3 Preliminaries

Active Learning under Distribution Shift In an active learning problem, a learner L actively collects a labeled dataset S with the goal of maximizing the performance of the hypothesis $h \in H$ learned from S. m labeled datapoints sampled from some distribution P_{warm} initially populate S and constitute the "warm start" dataset D_{warm}. L samples unlabeled datapoints D_{ulb} from some distribution P_{ulb}, and may select up to n for labeling and appending to S. The learned hypothesis h is evaluated on a test distribution P_{test}. Traditional active learning assumes,

$$P_{\text{ulb}} = P_{\text{warm}} = P_{\text{test}}. \tag{1}$$

In contrast, *active domain adaptation* does not assume the warm start is sampled from the test distribution:

$$P_{\text{ulb}} = P_{\text{test}} \text{ and } P_{\text{warm}} \neq P_{\text{test}}. \tag{2}$$

This setting, which we term *canonical label shift*, is well-studied but assumes active learning occurs in the test domain. We address the more challenging *general label shift* setting (Figure 3) which drops this assumption. In the worst case, all distributions could be different:

$$P_{\text{warm}} \neq P_{\text{ulb}}, P_{\text{warm}} \neq P_{\text{test}}, P_{\text{ulb}} \neq P_{\text{test}}. \tag{3}$$

For instance, the problem of creating a bird classifier for New York by actively labeling data off Google is *general label shift*. This setting has received comparatively little attention despite its practical relevance [Huang and Chen, 2016]: there may be a scarcity of unlabeled target data or practical issues with labeling target data, such as patient privacy or ownership rights.

Label Shift The distribution shift problem concerns training and evaluating models on different distributions, termed the source (P_{src}) and target (P_{trg}) respectively. We refer to a source and target in the abstract. For instance, in the *canonical label shift* setting, the source is the warm start $P_{\text{src}} = P_{\text{warm}}$, and the target is the test $P_{\text{trg}} = P_{\text{test}}$. Unlike covariate shift, which assumes the underlying distribution shift arises solely from a change in the input distribution while conditional label probabilities are unaffected[1], label shift assumes distribution shift arises solely from a change in label marginals:

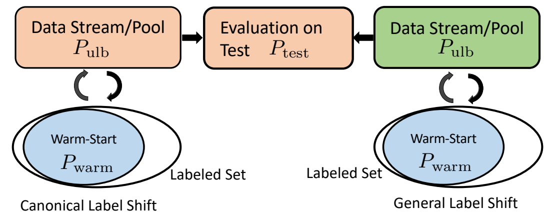


Figure 3: Diagram of active learning under label shift settings. The *canonical label shift* actively samples from the test distribution. The *general label shift setting* actively samples from elsewhere. Identical colors indicate identical distributions.

$$P_{\text{trg}}(Y) \neq P_{\text{src}}(Y), P_{\text{trg}}(X|Y) = P_{\text{src}}(X|Y). \tag{4}$$

Importance Weighting (IW) Importance weighting is a straight-forward solution to label shift. Weighting datapoints by likelihood ratio produces asymptotically unbiased importance weighted estimators.

$$\frac{1}{n}\sum_{i=1}^{n} \frac{P_{\text{trg}}(y_i)}{P_{\text{src}}(y_i)} f(x_i, y_i) \to \mathbb{E}_{x,y\sim P_{\text{src}}}\left[\frac{P_{\text{trg}}(y)}{P_{\text{src}}(y)} f(x,y)\right] = \mathbb{E}_{x,y\sim P_{\text{trg}}}\left[f(x,y)\right]. \tag{5}$$

Following existing label shift literature, we restrict our learning problems to those with a finite k-class label space. We can estimate these importance weights with only labeled data from the source distribution, unlabeled data from the target distribution, and a blackbox hypothesis h_0 [Lipton et al., 2018]. Let C_h denote the confusion matrix for hypothesis h on P_{src} where $\mathbb{E}[C_h[i,j]] := P_{\text{src}}(h(X)=y^{(i)}, Y=y^{(j)})$ and q_h denote a k-vector with $q_h[i] := P_{\text{trg}}(h(X)=y^{(j)})$. Assuming for all labels $\forall y : P_{\text{trg}}(y) > 0 \implies P_{\text{src}}(y) > 0$, [Lipton et al., 2018] shows importance weights r are,

$$r := \frac{P_{\text{trg}}(y)}{P_{\text{src}}(y)} = C_{h_0}^{-1} q_{h_0}. \tag{6}$$

For instance, Regularized Learning under Label Shift (RLLS) [Azizzadenesheli et al., 2019] finds r through

[1] We abuse notation and define $P(\cdot)$ as $P(x) := P(X=x)$ or $P(y) := P(Y=y)$ depending on the context.

convex optimization of:

$$C_{h_0}^{-1} q_{h_0} \approx \operatorname{argmin}_r \left\| C_{h_0} r - q_{h_0} \right\|_2 + \lambda \left\| r - 1 \right\|_2, \quad (7)$$

where λ is some regularization constant.

Class-balanced Sampling (Subsampling) A popular heuristic for addressing class imbalance in active learning is adjusting the probability of labeling a datapoint by its predicted label [Yang and Ma, 2010, Park, 2011]. Traditionally, class-balanced sampling aims to ensure equal representation of each label and can be framed as a form of label shift with a uniform target label distribution. We generalize class-balanced sampling to general label shift problems with potentially non-uniform targets, a practice we term *subsampling*. We now describe two methods of subsampling. In these examples, we subsample a user-defined distribution P_{med} from a source P_{src} using predictor ϕ for predicting proxy labels.

1. **Subsampling with a filter** $P_{\mathbf{ss}}$, where $P_{\text{med}}(y) \propto P_{\text{ss}}(y{=}\phi(x))P_{\text{src}}(y{=}\phi(x))$. Repeat until a sample is yielded: sample datapoint x from P_{src} and, with probability $P_{\text{ss}}(Y{=}\phi(x))$, yield x.

2. **Subsampling with the target** $P_{\mathbf{med}}$. Collect N datapoints from P_{src} into a buffer S', where N is large. For each label $y \in Y$, randomly add $NP_{\text{med}}(y)$ datapoints from $\{x \in S' \mid \phi(x) = y\}$ into a buffer S. To sample from P_{med}, draw from S.

While in finite settings only the former yields IID samples, the two are identical in the limit by the law of large numbers. Since subsampling strictly concerns proxy labels as thus does not require labeled samples, we assume subsampling occurs at the limit and use the two interchangeably.

In the expectation, subsampling is equivalent to importance weighting with proxy labels predicted by ϕ:

$$\mathbb{E}_Q \left[\frac{1}{n} \sum_{i=1}^{n} Q_i f(x_i, y_i) \right] = \frac{1}{n} \sum_{i=1}^{n} \frac{P_{\text{trg}}(y_i{=}\phi(x_i))}{P_{\text{src}}(y_i{=}\phi(x_i))} f(x_i, y_i),$$

where $Q_i \in \{0, 1\}$ is an indicator variable for whether the ith datapoint is subsampled and has conditional expectation $\mathbb{E}[Q_i \mid y_i] = P_{\text{ss}}(y_i)$.

4 Medial Distribution

In this section, we propose the concept of a *medial distribution*. We conceptually frame subsampling as the importance sampling of an alternative distribution from the source distribution. We term this alternative distribution the *medial distribution* P_{med}. As we will show, P_{med} mediates a trade-off between subsampling and importance weighting (IW).

IW-Subsampling Trade-off In this section, we adopt domain adaptation notation and denote source, medial and target distributions as $P_{\text{src}}, P_{\text{med}}, P_{\text{trg}}$. Let $r_{s\to t} := P_{\text{trg}}(y)/P_{\text{src}}(y)$ denote the importance weights which shift the source to the target. Similarly, let $r_{s\to m} := P_{\text{med}}(y)/P_{\text{src}}(y)$ and $r_{m\to t} := P_{\text{trg}}(y)/P_{\text{med}}(y)$ denote the importance weights *to* and *from* the medial distribution. Note that $P_{\text{med}}(y), P_{\text{src}}(y), P_{\text{trg}}(y)$ denote the likelihood of a ground-truth label y. Estimated weights are accented with a hat: $\hat{r}$. We follow [Lipton et al., 2018] and formalize label shift magnitude as $\|\theta\|$: some norm of $\theta := r - \mathbf{1}$, usually the L2 norm $\|\cdot\|_2$. A large $\|\theta\|$ corresponds to a larger label shift and harder learning problem. $\|\theta_{s\to m}\|$ is the amount of label shift corrected by subsampling and $\|\theta_{m\to t}\|$ is the amount corrected by importance weighting. We can analyze this *medial distribution* trade-off by introducing the following bound on the accuracy of empirical loss estimates under label shift where subsampling and importance weighting are used. This theorem is a modification of a common error bound for offline supervised learning under label shift.

Theorem 1. *Let Δ denote the subsampling and importance weighting (trained on n datapoints) estimation error of the empirical loss of N datapoints:*

$$\Delta := \frac{1}{N} \sum_{i=1}^{N} \left(r_i P_{ss}(y_i) - \hat{r}_i P_{ss}(h(x_i)) \right) \ell(h(x_i), y_i), \quad (8)$$

where $\ell : Y \times Y \to [0, 1]$ is a loss function. With probability $1 - 2\delta$, for all $n \geq 1$:

$$|\Delta| \leq \mathcal{O}\left(\frac{2}{\sigma_{\min}} \left(\|\theta_{m\to t}\|_2 \sqrt{\frac{\log\left(\frac{nk}{\delta}\right)}{n}} + \sqrt{\frac{\log\left(\frac{n}{\delta}\right)}{n}} + \|\theta_{s\to m}\|_\infty \, err(h_0, r_{m\to t}) \right) \right), \quad (9)$$

where $\sigma_{\min}$ denotes the smallest singular value of the confusion matrix and $err(h_0, r)$ denotes the importance weighted $0/1$-error of a blackbox predictor h_0 on P_{src}.

The error bound in theorem 1 shows that the use of subsampling versus importance weighting results in different error bounds with different trade-offs. In particular, the trade-off lies between the first summand, $\|\theta_{m\to t}\|_2 \sqrt{\frac{\log\left(\frac{nk}{\delta}\right)}{n}}$, and the third summand, $\|\theta_{s\to m}\|_\infty \operatorname{err}(h_0, r_{m\to t})$. The former term corresponds to the error introduced by the use of importance weights—in particular, the variance that arises from importance weight estimation. This variance is sensitive to the magnitude of the ground-truth importance weights, $\|r_{m\to t}\|_2$. Recall that in our medial distribution framework, importance weights correct the label shift between P_{med} and P_{trg}. The latter term corresponds to the subsampling estimation error—in particular, the bias introduced by the use of proxy labels for data weighting. This bias is sensitive to the magnitude of subsampling, $\|\theta_{s\to m}\|_\infty$, and the accuracy of the blackbox hypothesis $\text{err}(h_0, r_{m\to t})$. Hence, subsampling mitigates sensitivity to label shift magnitude by splitting the norm of total label shift, $\|\theta_{s\to t}\|$, into the sum of two factors which scale with $\|\theta_{s\to m}\|$ and $\|\theta_{m\to t}\|$.

The key to addressing this bias-variance trade-off is choosing a medial distribution which balances the quality of the blackbox hypothesis $\text{err}(h_0, r_{m\to t})$ and the label shift magnitude $\|r_{s\to t}\|$. In effect, subsampling turns a difficult label shift problem, requiring large importance weights $r_{s\to t}$, into an easier label shift problem, with smaller importance weights $r_{m\to t}$.

Uniform Medial Distribution We motivate a particular choice of medial distribution, a uniform label distribution, with an example. Figure 1 depicts two fundamental label shift regimes which we term *imbalanced source* and *imbalanced target*. *Imbalanced target* requires smaller importance weights to correct than *imbalanced source* and is hence more efficient for IW to correct. *Imbalanced source* requires fewer additional examples than *imbalanced target* and is more efficient for subsampling to correct. This holds more broadly. Consider a binary classification problem with n datapoints and two possible label distributions: balanced distribution D_1 with $n/2$ datapoints in each class, and imbalanced distribution D_2 with $n-1$ datapoints in the majority class. Under *imbalanced source*, where $P_{\text{src}} := D_2$ and $P_{\text{trg}} := D_1$, $n-2$ additional samples from the under-represented class are necessary for negating label shift. Under *imbalanced target*, where $P_{\text{src}} := D_1$ and $P_{\text{trg}} := D_2$, $(n-2)\frac{n}{2} \in \mathcal{O}(n^2)$ additional samples are necessary.

This suggests subsampling under *imbalanced source* and importance weighting under *imbalanced target*. A uniform medial distribution decomposes every label shift problem into the two settings: subsample *to* a uniform label distribution (imbalanced source) then importance weight away *from* uniform (imbalanced target). As we will show, this affords a convenient upper bounds on the sample complexity of active learning with a uniform medial distribution. We will also show, experimentally, that uniform distributions serve as a reliable choice for medial distributions and perform similarly to "square root" medial distributions that are optimal in simple cases, e.g., singleton X.

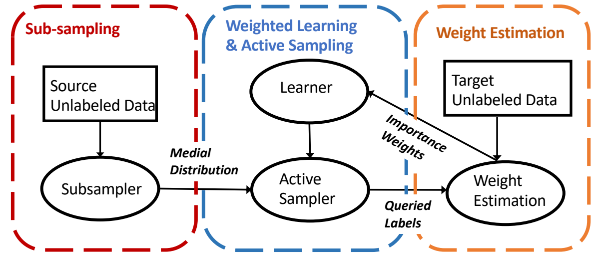


Figure 4: The MALLS Algorithm. MALLS consists of 3 routines: (1) class-balanced sampling from the unlabeled set, (2) actively querying for labels, and (3) importance weight estimation for correcting label shift.

5 Streaming MALLS

In this section, we present a streaming active learning algorithm: Mediated Active Learning under Label Shift (MALLS). We analyze the generalization error and the label complexity of steaming MALLS and validate the theory with experiments. We present a practical batched MALLS approach in Sec. 6. We also open-source an implementation of MALLS.

Algorithm 1 Mediated Active Learning under Label Shift

Input: Warm start set (D_{warm}), unlabeled set (D_{ulb}), test set (D_{test}), active learning budget n, label shift budget λ, blackbox predictor h_0, medial distribution P_{med}, hypothesis class $\mathcal{H}$
Initialize the dataset $S \leftarrow$ warm start set;
Subsample the unlabeled set using h_0 to induce P_{med}.
Estimate importance weights:
  Obtain r with RLLS [Azizzadenesheli et al., 2019] using h_0, unlabeled test data, and λn labeled datapoints from the unlabeled set;
While $|S| < n$
  Calculate IWAL-CAL [Beygelzimer et al., 2010] sampling probability P_t for x_t using S weighted by r;
  Label and append (x_t, y_t) to S with probability P_t;
Output: $h_T := \text{argmin}_{h\in H}\text{err}(h, r)$ where err is estimated on S.

Proposed Algorithm We build on a popular importance-weighted agnostic active learning algorithm IWAL-CAL [Beygelzimer et al., 2010]. We refer to IWAL-CAL as a subprocedure and defer its details to the Appendix. IWAL-CAL takes as input a datapoint x_t and returns a sampling probability P_t. MALLS modifies the computation of P_t by applying importance weights to correct for label shift in empirical loss estimates. Specifically, IWAL-CAL depends on estimating hypothesis loss on the actively labeled dataset S:

$$\text{err}(h) = \frac{1}{|S|}\sum_{t=1}^{|S|}\ell(h(x_t), y_t), \tag{10}$$

where x_i, y_i are drawn from S. MALLS instead computes empirical loss estimates as:

$$\mathrm{err}(h, r) = \frac{1}{|S|}\sum_{t=1}^{|S|} r(y_t)\ell(h(x_t), y_t), \tag{11}$$

where $r(y_t)$ denotes an importance weight for datapoints of label y_t. MALLS computes these importance weights r by calling a blackbox label shift estimator (e.g. BBSE [Lipton et al., 2018]). Our derivations use Regularized Learning under Label Shift (RLLS) [Azizzadenesheli et al., 2019]. Since label shift estimation algorithms require an independent holdout set for estimating importance weights, MALLS estimates importance weights on a holdout set of λn labeled datapoints sampled from P_{med} through subsampling. MALLS also adds subsampling as a preprocessing step to IWAL-CAL, re-using the blackbox hypothesis used in label shift estimation as a predictor. Thus, instead of directly sampling points from D_{ulb}, IWAL-CAL instead interacts with datapoints subsampled from D_{ulb} and distributed according to P_{med}. We detail the high-level flow of MALLS in Figure 4 and provide pseudocode in Algorithm 1.

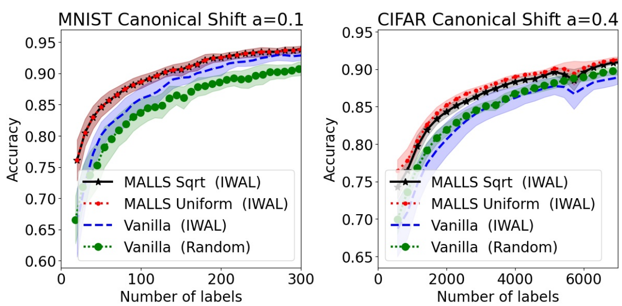


Figure 5: Average performance and 95% confidence intervals on 10 runs of experiments on MNIST, 5 runs on CIFAR in a *canonical label shift* setting (defined in Preliminaries). Accuracy on (a) MNIST, (b) CIFAR. MALLS leads to sample efficiency gains in both settings. A "uniform" medial performs on par with a "square root" medial.

Theoretical Analysis We now analyze label complexity and generalization bounds for Algorithm 1. In the canonical label shift setting, label shift naturally disappears asymptotically as the warm start dataset is diluted. For the remainder of this section, we instead work in the more challenging *general label shift* setting. As the presence of warm start data is not particularly interesting in our analysis, we set the warm start budget $m = 0$ for reading convenience and defer the case where $m > 0$ to the Appendix for interested readers. We also defer the case where the quantity of unlabeled test data is bounded to the Appendix.

The derivation of theoretical guarantees for MALLS builds off our Theorem 1 and existing results from IWAL-CAL. The proof consists of two primary steps. First, new deviation bounds are derived for IWAL-CAL to compensate for the additional variance introduced by subsampling and importance weighting. Second, triangle inequalities plug in results from Section 3 on the bias-variance tradeoff. The resulting deviation bound (see Appendix) yields the following guarantees for MALLS.

Theorem 2. *With probability $> 1-\delta$, for all $n \geq 1$,*

$$\begin{aligned} err_Q(h_n) \leq & \mathcal{O}\Bigg((1+\frac{1}{\sigma_{\min}})\,\|r_{s\to t}\|_\infty\, err(h_0, r_{m\to t}) \\ & + err_Q(h^*) + \sqrt{\frac{2C_0 \log n}{n-1}} + \frac{2C_0 \log n}{n-1}\Bigg), \end{aligned} \tag{12}$$

where err_Q denotes hypothesis error in the target domain, n denotes observed datapoints including those not labeled or subsampled, and the constant C_0 is,

$$\begin{aligned} C_0 \in \mathcal{O}\Bigg(&\frac{2}{\lambda\sigma_{\min}}\left(\|\theta_{m\to t}\|_2^2 \log\left(\frac{k}{\delta}\right) + \log\left(\frac{1}{\delta}\right)\right) \\ &+ \log\left(\frac{|H|}{\delta}\right)(1+\|\theta_{s\to t}\|_2^2)\Bigg). \end{aligned} \tag{13}$$

Our generalization bound differs from the original IWAL-CAL bound in two key aspects. (1) The use of subsampling introduces bias related to the performance of the blackbox hypothesis: $\frac{1}{\sigma_{\min}}\mathrm{err}(h_0, r_{m\to t})$. (2) In the original IWAL-CAL algorithm $C_0 \in \mathcal{O}(\log(|H|/\delta)$. However label shift inevitably introduces, to the constant C_0, a dependence on the number of label classes k and label shift magnitudes $\|\theta_{s\to t}\|_2^2$ and $\|\theta_{m\to t}\|_2^2$. When the subsampling error is high, Theorem 2 shows importance weighting can be used alone to preserve a consistency guarantee even under *general label shift*.

Theorem 3. *With high probability*[2] *at least $1-\delta$, the number of labels queried is at most:*

$$\begin{aligned} &\mathcal{O}\Bigg(1+\log\left(\frac{1}{\delta}\right) + \Theta\sqrt{C_0\frac{n}{\|r_{s\to m}\|_\infty}\log n} + \Theta C_0 \log^3 n + \lambda n \\ &+\Theta\cdot(n-1)\cdot\left(\frac{err_Q(h^*)}{\|r_{s\to m}\|_\infty} + (1+\frac{1}{\sigma_{\min}})err(h_0, r_{m\to t})\right)\Bigg), \end{aligned} \tag{14}$$

where Θ denotes the disagreement coefficient [Balcan et al., 2009].

Subsampling effectively increases the noise rate of the underlying problem. This increases the linear noise rate term $O(n)$ inevitable in agnostic active learning labeling complexities. However, subsampling also reduces sample complexity by a factor of $\frac{1}{\|r_{s\to m}\|_\infty}$. Importance

[2] Where $\delta < 2^{(-2e-1)/\|r_{s\to m}\|_\infty}$.

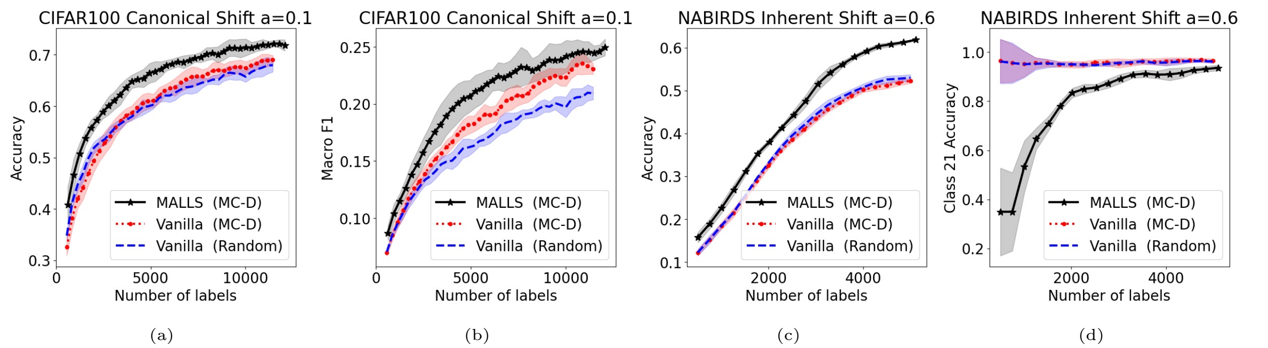


Figure 6: Average performance and 95% confidence intervals of 10 runs on CIFAR100, and 4 runs on NABirds. Plots (a)-(c) demonstrate MALLS consistently improves accuracy and macro F1 scores. Plot (d) depicts the learning dynamics of MALLS and verifies a suppression of the over-represented class (Class 21) during learning.

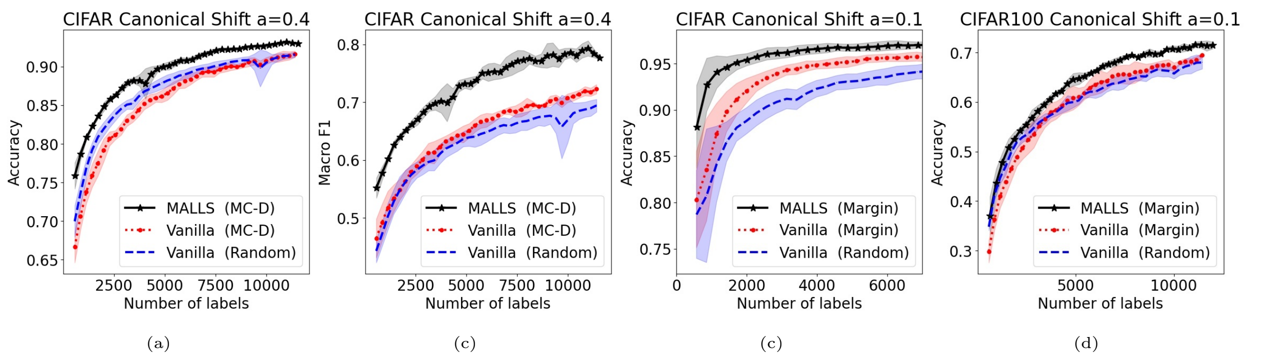


Figure 7: Average performance and 95% confidence intervals of 10 runs on CIFAR10 and CIFAR100. MALLS consistently improves accuracy, macro F1, and weighted F1 scores.

weighting introduces a new linear label complexity term $O(\lambda n)$. This is used to collect a holdout set for label shift estimation. Thus, when the blackbox hypothesis is bad and strong importance weighting is necessary, the sample complexity improvements of active learning are lost. However, given a good blackbox hypothesis, the medial distribution can be set closer to the target (small $\|\theta_{m\to t}\|$) and λ can be set small so MALLS retains the sample complexity gains of active learning.

Experiments We empirically validate MALLS with experiments on synthetic label shift problems on the MNIST and CIFAR benchmark datasets. These experiments employ a bootstrap approximation of IWAL-CAL recommended in [Beygelzimer et al., 2009] using a version space of 8 Resnet-18 models. The blackbox hypothesis is obtained by training a standalone model on the warm start data split. Random sampling and vanilla active learning (IWAL-CAL) are compared against MALLS for two choices of medial distributions:

1. A "square root" medial distribution where $r_{s\to m} = r_{m\to t} = \sqrt{r_{s\to t}}$. This is a bare optimization of the error tradeoff in Theorem 1.
2. A "uniform" medial distribution motivated by Figure 1 and intuition of *imbalanced sources/targets*.

The results, shown in Figure 5 demonstrate significant sample efficiency gains due to MALLS, even when vanilla IWAL no longer beats random sampling. Despite its simplicity, the performance of the uniform medial distribution is indistinguishable from the theoretically motivated "square root" medial distribution.

6 Batched MALLS

We present a variant of MALLS for practitioners which integrates best practices for scaling label shift estimation. This variant, depicted in Algorithm 2, is a framework for batched active learning that supports any blackbox uncertainty sampling algorithm.

Best Practices Batched MALLS incorporates five important techniques for scaling the real world practice of label shift correction.

1. Forgo use of independent holdout sets and instead learn importance weights r on the main dataset S.
2. Motivated by Theorem 1, Batched MALLS uses the current active learning predictor for subsampling.
3. Approximate subsampling with a generalization of

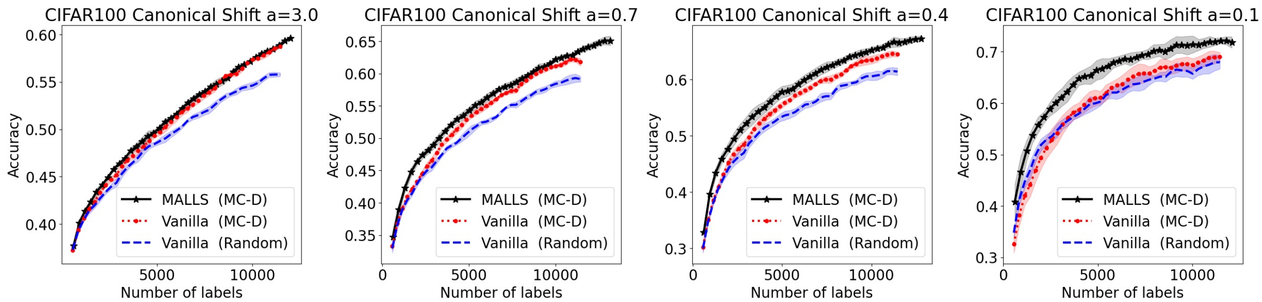


Figure 8: Average performance and 95% confidence intervals on 10 runs of experiments on CIFAR100 in the *canonical label shift* setting (defined in Preliminaries). In order of increasing label shift magnitude: (a), (b), (c), (d). MALLS performance gains scale by label shift magnitude.

class-balanced sampling that is compatible with batch-mode active learning [Aggarwal et al., 2020].

4. Apply importance weights during inference time. Batched MALLS replaces the importance weighting of empirical loss estimates with posterior regularization, a practice closely related to the expectation-maximization algorithm in [Saerens et al., 2002].
5. Use hypotheses learned with importance weights as blackbox predictors to learn better importance weights. We term this iterative reweighting.

Algorithm 2 Batched MALLS

Input: Warm start set, unlabeled pool D_{ulb}, test set, number of batches T, medial distribution P_{med}, uncertainty quantifier π, batch size B; hypothesis class $\mathcal{H}$
Initialize the labeled dataset $S_0 \leftarrow$ the warm start set;
Initialize hypothesis ϕ by training on warm start set;
For $t \in 1, \dots, T$
  Find importance weights $r_t \leftarrow \text{RLLS}(S_t, \phi, P_{\text{med}})$;
  Train hypothesis ϕ on S weighted by r_t;
  For $y \in Y$
    Number of datapoints to collect $k := B \times P_{\text{med}}(y)$
    Find top-k most uncertain datapoints of label y:
      $D_y := \text{top-}k(\pi, \{x \in D_{\text{ulb}} \setminus S_{t-1} \mid \phi(x) = y\})$
    Label and append the top-k datapoints, D_y, to S_t
Output: $h_T = \text{argmin}\{\text{err}(h, S_T, r_T) : h \in \mathcal{H}\}$

Experiments We demonstrate the Batched MALLS framework on the ornithology dataset NABirds [Van Horn et al., 2015] and the benchmark datasets CIFAR10 & CIFAR100 [Krizhevsky, 2009]. Our experiments show MALLS improves active learning performance under a diverse range of label shift scenarios.

Methods We evaluate our Batched MALLS framework on several uncertainty sampling algorithms: (1) Monte Carlo dropout (MC-D) [Gal and Ghahramani, 2016]; (2) maximum entropy sampling (MaxEnt); and (3) maximum margin (Margin). We compare against random sampling and active learning without MALLS (marked *Vanilla*). In ablation studies, we also compare against only importance weighting or subsampling. As in Section 5, the blackbox hypothesis is obtained by training a model on the warm start data split.

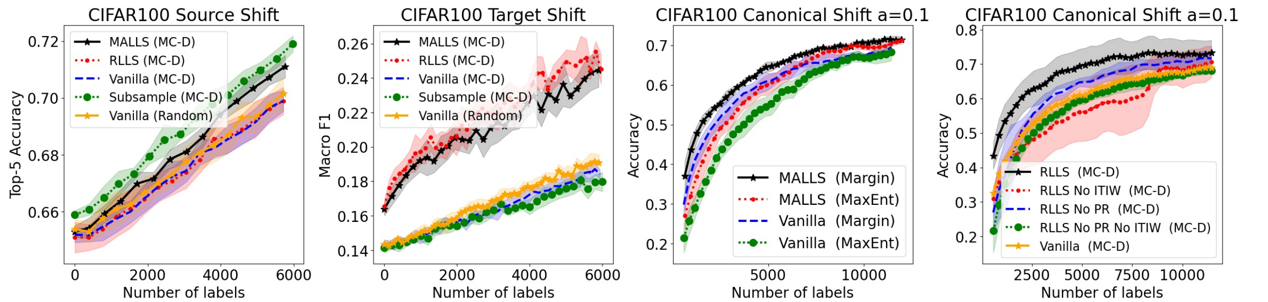
Primary Results We present our primary results in Figures 6-7. These experiments apply MALLS to the batch-mode pool-based active learning of Resnet18 models. The label shift in the NABirds dataset arises from a naturally occurring class imbalance where a dominant class constitutes a near majority of all data [Aggarwal et al., 2020]. We adopt this imbalance and assume a uniform test label distribution. We artificially induce *canonical label shift* in the CIFAR10 and CIFAR100 experiments by applying [Lipton et al., 2018]'s *Dirichlet Shift* procedure to the unlabeled D_{ulb} and test D_{test} datasets.

In all experiments, MALLS significantly improves both accuracy and macro F1 scores. In synthetic shift experiments, MALLS reduces sample complexity by up to half an order of magnitude.

Learning Dynamics of MALLS Figure 6(d) details the learning evolution of MALLS by depicting a dominant class's accuracy over training time. The class's accuracy initially declines due to the class's low importance weights, but recovers as the label shift is corrected and the dominant class's importance weight grows.

Uncertainty Measures Figure 7(c)(d) and 9(c) demonstrates the performance improvements from using Batched MALLS generalize to several popular uncertainty sampling algorithms. Importantly, the gains realized by using Batched MALLS is largely independent of the choice of uncertainty sampling.

Imbalanced Source v.s. Imbalanced Target Figures 9(a)(b) depicts synthetic *general label shift* problems under *imbalanced source* and *imbalanced target* settings on CIFAR100. We compare MALLS against the use of subsampling or importance weighting alone to investigate the trade-off implied by theory. While Figure 9(a) demonstrates that subsampling accounts for



(a) (b) (c) (d)

Figure 9: Average and 95% confidence intervals on 10 runs of experiments on CIFAR100. Figures (a)(b) depict *general label shift* problems, (c)(d) depict *canonical label shift* (defined in Preliminaries). (a) Top-5 accuracy under *imbalanced source*, subsampling outperforms importance weighting; (b) Macro F1 under *imbalanced target*, importance weighting outperforms subsampling; (c) Batched MALLS provides performance gains for multiple popular uncertainty sampling methods; (d) Batched MALLS's best practices provides significant gains on performance.

MALLS's performance gains under *imbalanced source*, Figure 9(b) demonstrates that importance weighting accounts for MALLS's performance gains under *imbalanced target*. This corroborates our theoretical analysis.

Label Shift Magnitude These experiments evaluate MALLS on different magnitudes of label shift, where label shift is induced according to Dirichlet distributions for varying choices of α. Note that shift magnitude is inversely correlated with α—smaller α denotes a larger shift. Figure 8 demonstrates that the performance gains introduced by RLLS scale with the magnitude of the label shift. The results also confirm that the effectiveness of active learning drops under strong label shift. Plot (a) confirms that even when label shift is negligible, MALLS does not perform significantly worse than vanilla active learning.

Best Practices Figures 9(d) compares performance when Batched MALLS's heuristics of posterior regularization (PR) and iterative reweighting (ITIW) are not used. Posterior regularization lowers variance (versus importance weighting) and especially improves early-stage performance. Iterative reweighting similarly introduces consistent performance gains. Combining them provides additional gains.

7 Conclusion

In this paper, we propose an algorithm for active learning under label shift, MALLS, with strong label complexity and generalization bounds. We also introduce a framework, Batched MALLS, for practitioners to address label shift in real world uncertainty sampling applications. In many applications that require manually labeling of data, like natural language processing and computer vision, an extension of the techniques we explore in MALLS may help mitigate bias in the data collection process. Many problems of theoretical importance—such as cost-sensitive, multi-domain, and Neyman-Pearson settings—share a fundamental connection with the label shift problem. We believe MALLS can be extended to provide novel results in these settings as well.

Acknowledgements

Anqi Liu is supported by the PIMCO Postdoctoral Fellowship. Prof. Anandkumar is supported by Bren endowed Chair, faculty awards from Microsoft, Google, and Adobe, Beyond Limits, and LwLL grants. This work is also supported by funding from Raytheon and NASA TRISH.

References

[Aggarwal et al., 2020] Aggarwal, U., Popescu, A., and Hudelot, C. (2020). Active Learning for Imbalanced Datasets. pages 1428–1437.

[Azizzadenesheli et al., 2019] Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. (2019). Regularized Learning for Domain Adaptation under Label Shifts. *arXiv:1903.09734 [cs, stat]*. arXiv: 1903.09734.

[Balcan et al., 2009] Balcan, M.-F., Beygelzimer, A., and Langford, J. (2009). Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89.

[Ben-David et al., 2010] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175.

[Ben-David et al., 2007] Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). Analysis of Representations for Domain Adaptation. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press.

[Beygelzimer et al., 2009] Beygelzimer, A., Dasgupta, S., and Langford, J. (2009). Importance Weighted Active Learning. *arXiv:0812.4952 [cs]*. arXiv: 0812.4952.

[Beygelzimer et al., 2010] Beygelzimer, A., Hsu, D., Langford, J., and Zhang, T. (2010). Agnostic Active Learning Without Constraints. *arXiv:1006.2588 [cs]*. arXiv: 1006.2588.

[Chan and Ng, 2007] Chan, Y. S. and Ng, H. T. (2007). Domain Adaptation with Active Learning for Word Sense Disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.

[Chattopadhyay et al., 2013] Chattopadhyay, R., Fan, W., Davidson, I., Panchanathan, S., and Ye, J. (2013). Joint transfer and batch-mode active learning. In *30th International Conference on Machine Learning, ICML 2013*, pages 1290–1298. International Machine Learning Society (IMLS).

[Choudhury and Srinivasa, 2020] Choudhury, S. and Srinivasa, S. S. (2020). A Bayesian Active Learning Approach to Adaptive Motion Planning. In *Robotics Research*, pages 33–40. Springer.

[Cortes et al., 2010] Cortes, C., Mansour, Y., and Mohri, M. (2010). Learning Bounds for Importance Weighting. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 442–450. Curran Associates, Inc.

[Cortes and Mohri, 2014] Cortes, C. and Mohri, M. (2014). Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126. Publisher: Elsevier.

[Deng et al., 2018] Deng, C., Liu, X., Li, C., and Tao, D. (2018). Active multi-kernel domain adaptation for hyperspectral image classification. *Pattern Recognition*, 77:306–315. Publisher: Elsevier.

[Gal and Ghahramani, 2016] Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv:1506.02142 [cs, stat]*. arXiv: 1506.02142.

[Garg et al., 2020] Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. C. (2020). A Unified View of Label Shift Estimation. *arXiv:2003.07554 [cs, stat]*. arXiv: 2003.07554.

[Gretton et al., 2009] Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B., Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. (2009). Covariate Shift by Kernel Mean Matching. *Dataset Shift in Machine Learning, 131-160 (2009).*

[Hanneke, 2007] Hanneke, S. (2007). A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 353–360, Corvalis, Oregon, USA. Association for Computing Machinery.

[Hanneke, 2011] Hanneke, S. (2011). Activized Learning: Transforming Passive to Active with Improved Label Complexity. *arXiv:1108.1766 [cs, math, stat]*. arXiv: 1108.1766.

[Hanneke, 2014] Hanneke, S. (2014). Theory of Disagreement-Based Active Learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309. Publisher: Now Publishers, Inc.

[Huang and Chen, 2016] Huang, S.-J. and Chen, S. (2016). Transfer learning with active queries from source domain. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1592–1598, New York, New York, USA. AAAI Press.

[Krishnamurthy et al., 2019] Krishnamurthy, A., Agarwal, A., Huang, T.-K., Daume III, H., and Langford, J. (2019). Active Learning for Cost-Sensitive Classification. *arXiv:1703.01014 [cs, stat]*. arXiv: 1703.01014.

[Krizhevsky, 2009] Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images.

[Lin et al., 2018] Lin, C. H., Mausam, M., and Weld, D. S. (2018). Active Learning with Unbalanced Classes and Example-Generation Queries. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[Lipton et al., 2018] Lipton, Z. C., Wang, Y.-X., and Smola, A. (2018). Detecting and Correcting for Label Shift with Black Box Predictors.

[Matasci et al., 2012] Matasci, G., Tuia, D., and Kanevski, M. (2012). SVM-based boosting of active learning strategies for efficient domain adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(5):1335–1343. Publisher: IEEE.

[Park, 2011] Park, W. J. (2011). An Improved Active Learning in Unbalanced Data Classification. In Lee, C., Seigneur, J.-M., Park, J. J., and Wagner, R. R., editors, *Secure and Trust Computing, Data Management, and Applications*, Communications in Computer and Information Science, pages 84–93, Berlin, Heidelberg. Springer.

[Rai et al., 2010] Rai, P., Saha, A., Daumé, H., and Venkatasubramanian, S. (2010). Domain Adaptation meets Active Learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, Los Angeles, California. Association for Computational Linguistics.

[Saerens et al., 2002] Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41. Publisher: MIT Press.

[Saha et al., 2011] Saha, A., Rai, P., Daumé, H., Venkatasubramanian, S., and DuVall, S. L. (2011). Active Supervised Domain Adaptation. In Gunopulos, D., Hofmann, T., Malerba, D., and Vazirgiannis, M., editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 97–112, Berlin, Heidelberg. Springer.

[Shen et al., 2018] Shen, Y., Yun, H., Lipton, Z. C., Kronrod, Y., and Anandkumar, A. (2018). Deep Active Learning for Named Entity Recognition. *arXiv:1707.05928 [cs]*. arXiv: 1707.05928.

[Shimodaira, 2000] 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. Publisher: Elsevier.

[Su et al., 2019] Su, J.-C., Tsai, Y.-H., Sohn, K., Liu, B., Maji, S., and Chandraker, M. (2019). Active Adversarial Domain Adaptation. *arXiv:1904.07848 [cs]*. arXiv: 1904.07848 version: 1.

[Sugiyama et al., 2007] Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate Shift Adaptation by Importance Weighted Cross Validation. *The Journal of Machine Learning Research*, 8:985–1005.

[Sugiyama et al., 2012] Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

[Tsuboi et al., 2009] Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155. Publisher: Information Processing Society of Japan.

[Van Horn et al., 2015] Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., and Belongie, S. (2015). Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604.

[Yamada et al., 2011] Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. (2011). Relative density-ratio estimation for robust distribution comparison. In *Advances in neural information processing systems*, pages 594–602.

[Yan et al., 2018] Yan, S., Chaudhuri, K., and Javidi, T. (2018). Active Learning with Logged Data. *arXiv:1802.09069 [cs, stat]*. arXiv: 1802.09069.

[Yang et al., 2020] Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. (2020). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558.

[Yang and Ma, 2010] Yang, Y. and Ma, G. (2010). Ensemble-based active learning for class imbalance problem. *Journal of Biomedical Science and Engineering*, 3(10):1022–1029. Number: 10 Publisher: Scientific Research Publishing.

[Yang et al., 2015] Yang, Y., Ma, Z., Nie, F., Chang, X., and Hauptmann, A. G. (2015). Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127. Publisher: Springer.

[Zhang, 2005] Zhang, T. (2005). Data Dependent Concentration Bounds for Sequential Prediction Algorithms. pages 173–187.

Supplementary Materials

8 Proofs

8.1 Proof of Theorem 1

We formalize the violation of label shift assumptions resulting from subsampling as label shift drift [Azizzadenesheli et al., 2019].

Lemma 1. *The drift from label shift is bounded by:*

$$\left|1 - \mathbb{E}_{X,Y\sim P_{test}}\left[\frac{P_{med}(x|y)}{P_{test}(x|y)}\right]\right| \leq \|r_{s\to m}\|_\infty \, err(h_0, r_{s\to m}) \tag{15}$$

Proof. The drift is equivalent to expected importance weights,

$$\begin{aligned}\left|1 - \mathbb{E}_{X,Y\sim P_{\text{test}}}\left[\frac{P_{\text{med}}(x|y)}{P_{\text{test}}(x|y)}\right]\right| &= \left|1 - \int_{X,Y} P_{\text{med}}(x|y)P_{\text{test}}(y)\right| \\ &= \left|1 - \int_{X,Y} P_{\text{med}}(x,y)\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right| \\ &= \left|1 - \mathbb{E}_{X,Y\sim P_{\text{med}}}\left[\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right]\right| \end{aligned} \tag{16}$$

Drift can therefore be estimated in practice by randomly labeling subsampled points and measuring the average importance weight value. We can further expand the value of drift as:

$$\begin{aligned}\left|1 - \mathbb{E}_{X,Y\sim P_{\text{med}}}\left[\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right]\right| &= \left|1 - \int_{X,Y} CP_{\text{src}}(x,y)P_{\text{ss}}(h_0(x))\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right| \\ &= \left|1 - C\mathbb{E}_{X,Y\sim P_{\text{src}}}\left[P_{\text{ss}}(h_0(x))\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right]\right| \\ &= \left|1 - C\mathbb{E}_{X,Y\sim P_{\text{src}}}\left[P_{\text{ss}}(y)\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right]\right| + \left|C\mathbb{E}_{X,Y\sim P_{\text{src}}}\left[\left(P_{\text{ss}}(h_0(x)) - P_{\text{ss}}(y)\right)\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right]\right| \\ &= \left|1 - \sum_{Y}\left[P^*_{\text{med}}(y)\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right]\right| + \left|C\mathbb{E}_{X,Y\sim P_{\text{src}}}\left[\left(P_{\text{ss}}(h_0(x)) - P_{\text{ss}}(y)\right)\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right]\right| \end{aligned} \tag{17}$$

where C is a constant where $P_{\text{ss}} = \frac{1}{C}\frac{P_{\text{med}}}{P_{\text{src}}}$ and P^*_{med} denotes the target medial distribution. The second term corresponds to a weighted L1 error on P_{src}.

$$\begin{aligned}\left|C\mathbb{E}_{X,Y\sim P_{\text{src}}}\left[\left(P_{\text{ss}}(h_0(x)) - P_{\text{ss}}(y)\right)\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right]\right| &\leq \|r_{s\to m}\|_\infty \mathbb{E}_{X,Y\sim P_{\text{src}}}\left[\left|\mathbb{1}[h_0(x) \neq y]\right|\frac{P_{\text{test}}(y)}{P_{\text{med}}(y)}\right] \\ &= \|r_{s\to m}\|_\infty \operatorname{err}(h_0, r_{s\to m}) \end{aligned} \tag{18}$$

where $\operatorname{err}(h_0, r)$ denotes the importance weighted 0/1-error of a blackbox predictor h_0 on Ps. As the first term is thus dominated, we have that drift is bounded by the accuracy of the blackbox hypothesis. □

Plugging Lemma 1 into Theorem 2 in [Azizzadenesheli et al., 2019] yields a generalization of Theorem 1 where the number of unlabeled datapoints from the test distribution is n'.

Theorem 4. *With probability $1-\delta$, for all $n \geq 1$:*

$$|\Delta| \leq \mathcal{O}\left(\frac{2}{\sigma_{\min}}\left(\|\theta_{m\to t}\|_2\sqrt{\frac{\log\left(\frac{nk}{\delta}\right)}{n}}+\sqrt{\frac{\log\left(\frac{n}{\delta}\right)}{n}}+\sqrt{\frac{\log\left(\frac{n}{\delta}\right)}{n'}}+\|\theta_{s\to m}\|_\infty\, err(h_0, r_{m\to t})\right)\right) \tag{19}$$

where $\sigma_{\min}$ denotes the smallest singular value of the confusion matrix and $err(h_0, r)$ denotes the importance weighted $0/1$-error of a blackbox predictor h_0 on P_{src}.

Theorem 1 follows by setting $n' \to \infty$.

8.2 Theorem 2 and Theorem 3 Proofs

We will prove Theorem 2 and Theorem 3 for the general case where the number of unlabeled datapoints from the test distribution is n'. For the case depicted in the main paper, set $n' \to \infty$.

First, we review the IWAL-CAL active learning algorithm [Beygelzimer et al., 2010]. Let $\mathrm{err}_{S_i}(h) \to [0,1]$ denote the error of hypothesis $h \in H$ as estimated on S_i while $\mathrm{err}_{P_{\text{test}}}(h)$ denote the expected error of h on P_{test}. We next define,

$$\begin{aligned}
h^* &:= \mathrm{argmin}_{h\in H}\mathrm{err}_{P_{\text{test}}}(h),\\
h_k &:= \mathrm{argmin}_{h\in H}\mathrm{err}_{S_{k-1}}(h),\\
h'_k &:= \mathrm{argmin}\{\mathrm{err}_{S_{k-1}}(h) \mid h \in H \wedge h(\mathbf{D}^{(k)}_{\text{unlab}}) \neq h_k(\mathbf{D}^{(k)}_{\text{unlab}})\}\\
G_k &:= \mathrm{err}_{S_{k-1}}(h'_k) - \mathrm{err}_{S_{k-1}}(h_k)
\end{aligned}$$

IWAL-CAL employs a sampling probability $P_t = \min\{1, s\}$ for the $s \in (0,1)$ which solves the equation,

$$G_t = \left(\frac{c_1}{\sqrt{s}} - c_1 + 1\right)\sqrt{\frac{C_0 \log t}{t-1}} + \left(\frac{c_2}{s} - c_2 + 1\right)\frac{C_0 \log t}{t-1}$$

where C_0 is a constant bounded in Theorem 2 and $c_1 := 5 + 2\sqrt{2}, c_2 := 5$.

The most involved step in deriving generalization and sample complexity bounds for MALLS is bounding the deviation of empirical risk estimates. This is done through the following theorem.

Theorem 5. *Let $Z_i := (X_i, Y_i, Q_i)$ be our source data set, where Q_i is the indicator function on whether (X_i, Y_i) is sampled as labeled data. The following holds for all $n \geq 1$ and all $h \in \mathcal{H}$ with probability $1-\delta$:*

$$\begin{aligned}
&|err(h, Z_{1:n}) - err(h^*, Z_{1:n}) - err(h) + err(h^*)|\\
&\leq \mathcal{O}\left((2 + \|\theta\|_2)\sqrt{\frac{\varepsilon_n}{P_{\min,n}(h)}} + \frac{\varepsilon_n}{P_{\min,n}(h)} + \frac{2d_\infty(P_{test}, P_{src})\log(\frac{2n|H|}{\delta})}{3n} + \sqrt{\frac{2d_2(P_{test}, P_{src})\log(\frac{2n|H|}{\delta})}{n}}\right.\\
&\left.+ \|r_{s\to m}\|_\infty\, err(h_0, r_{s\to m}) + \frac{2}{\sigma_{\min}}\left(\|\theta_{m\to t}\|_2\sqrt{\frac{\log\left(\frac{nk}{\delta}\right)}{\lambda n}} + \sqrt{\frac{\log\left(\frac{n}{\delta}\right)}{\lambda n}} + \sqrt{\frac{\log\left(\frac{n}{\delta}\right)}{n'}} + \|\theta_{s\to m}\|_\infty\, err(h_0, r_{m\to t})\right)\right)
\end{aligned} \tag{20}$$

where $\varepsilon_n := \frac{16\log(2(2+n\log_2 n)n(n+1)|H|/\delta)}{n}$.

For reading convenience, we set $P_{\text{src}} := P_{\text{ulb}}$. This deviation bound will plug in to IWAL-CAL for generalization and sample complexity bounds. In the remainder of this appendix section, we detail our proof of Theorem 5. We proceed by expressing Theorem 5 in a more general form with a bounded function $f : X \times Y \to [-1, 1]$ which will eventually represent $\mathrm{err}(h) - \mathrm{err}(h^*)$.

We borrow notation for the terms W, Q from [Beygelzimer et al., 2010], where Q_i is an indicator random variable indicating whether the ith datapoint is labeled and $W := Q_i\tilde{Q}_i r^{(i)}_{m\to t} f(x_i, y_i)$. We use the shorthand $r^{(i)}$ for the y_ith component of importance weight r. Similarly, the indicator random variable $\tilde{Q}_i$ indicates whether the ith data sample is retained by the subsampler. The expectation $\mathbb{E}_i[W]$ is taken over the randomness of Q and $\tilde{Q}$. We also borrow [Azizzadenesheli et al., 2019]'s label shift notation and define k as the size of the output space (finite) and denote estimated importance weights with hats, e.g. $\hat{r}$. We also introduce a variant of W using estimated importance weights r: $\hat{W} := Q_i \tilde{Q}_i \hat{r}^{(i)}_{m\to t} f(x_i, y_i)$. Finally, we follow [Cortes et al., 2010] and use $d_\alpha(P||P')$ to denote $2^{D_\alpha(P||P')}$ where $D_\alpha(P||P') := \log(\frac{P_i}{P'_i})$ is the Renyi divergence of distributions P and P'.

We seek to bound with high probability,

$$|\Delta| := \left| \frac{1}{n} \left(\sum_{i=1}^{n} \hat{W}_i \right) - \mathbb{E}_{x,y\sim P_{\text{trg}}}[f(x,y)] \right| \leq |\Delta_1|+|\Delta_2|+|\Delta_3|+|\Delta_4| \tag{21}$$

where,

$$\begin{aligned}
\Delta_1 &:= \mathbb{E}_{x,y\sim P_{\text{trg}}}[f(x,y)] - \mathbb{E}_{x,y\sim P_{\text{src}}}[W_i], \\
\Delta_2 &:= \mathbb{E}_{x,y\sim P_{\text{src}}}[W_i] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_i\left[W_i\right], \\
\Delta_3 &:= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_i\left[W_i\right] - \mathbb{E}_i\left[\hat{W}_i\right] \\
\Delta_4 &:= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_i[\hat{W}_i] - \hat{W}_i
\end{aligned}$$

Δ_1 corresponds to the drift from label shift introduced by subsampling, Δ_2 to finite-sample variance. and Δ_3 to label shift estimation errors. The final Δ_4 corresponds to the variance from randomly sampling.

We bound Δ_4 using a Martingale technique from [Zhang, 2005] also adopted by [Beygelzimer et al., 2010]. We take Lemmas 1, 2 from [Zhang, 2005] as given. We now proceed in a fashion similar to the proof of Theorem 1 from [Beygelzimer et al., 2010]. We begin with a generalization of Lemma 6 in [Beygelzimer et al., 2010].

Lemma 2. *If* $0 < \lambda < 3\frac{P_i}{\hat{r}^{(i)}_{m\to t}}$*, then*

$$\log \mathbb{E}_i[\exp(\lambda(\hat{W}_i - \mathbb{E}_i[\hat{W}_i]))] \leq \frac{\hat{r}_i \hat{r}^{(i)}_{m\to t} \lambda^2}{2P_i(1 - \frac{\hat{r}^{(i)}_{m\to t}\lambda}{3P_i})} \tag{22}$$

where $\hat{r}_i := \hat{r}^{(i)}_{m\to t} \mathbb{E}_i[\tilde{Q}_i]$*. If* $\mathbb{E}_i[\hat{W}_i] = 0$ *then*

$$\log \mathbb{E}_i[\exp(\lambda(\hat{W}_i - \mathbb{E}_i[\hat{W}_i]))] = 0 \tag{23}$$

Proof. First, we bound the range and variance of $\hat{W}_i$. The range is trivial

$$|\hat{W}_i| \leq \left| \frac{Q_i \tilde{Q}_i \hat{r}^{(i)}_{m\to t}}{P_i} \right| \leq \frac{\hat{r}^{(i)}_{m\to t}}{P_i} \tag{24}$$

Since subsampling and importance weighting ideally corrects underlying label shift, we can simplify the variance as,

$$\mathbb{E}_i[(\hat{W}_i - \mathbb{E}_i[\hat{W}_i])^2] \leq \frac{\hat{r}_i \hat{r}^{(i)}_{m\to t}}{P_i} f(x_i,y_i)^2 - 2\hat{r}_i^2 f(x_i,y_i)^2 + \hat{r}_i^2 f(x_i,y_i)^2 \leq \frac{\hat{r}_i \hat{r}^{(i)}_{m\to t}}{P_i} \tag{25}$$

Following [Beygelzimer et al., 2010], we choose a function $g(x) := (\exp(x) - x - 1)/x^2$ for $x \neq 0$ so that $\exp(x) = 1 + x + x^2 g(x)$ holds. Note that $g(x)$ is non-decreasing. Thus,

$$\begin{aligned}
\mathbb{E}_i[\exp(\lambda(\hat{W}_i - \mathbb{E}_i[\hat{W}_i]))] &= \mathbb{E}_i[1 + \lambda(\hat{W}_i - \mathbb{E}_i[\hat{W}_i]) + \lambda^2(\hat{W}_i - \mathbb{E}_i[\hat{W}_i])^2 g(\lambda(\hat{W}_i - \mathbb{E}_i[\hat{W}_i]))] \\
&= 1 + \lambda^2 \mathbb{E}_i[(\hat{W}_i - \mathbb{E}_i[\hat{W}_i])^2 g(\lambda(\hat{W}_i - \mathbb{E}_i[\hat{W}_i]))] \\
&\leq 1 + \lambda^2 \mathbb{E}_i[(\hat{W}_i - \mathbb{E}_i[\hat{W}_i])^2 g(\lambda \hat{r}^{(i)}_{m\to t}/P_i)] \\
&= 1 + \lambda^2 \mathbb{E}_i[(\hat{W}_i - \mathbb{E}_i[\hat{W}_i])^2] g(\lambda \hat{r}^{(i)}_{m\to t}/P_i) \\
&\leq 1 + \frac{\lambda^2 \hat{r}_i \hat{r}^{(i)}_{m\to t}}{P_i} g(\frac{\hat{r}^{(i)}_{m\to t}\lambda}{P_i})
\end{aligned} \tag{26}$$

where the first inequality follows from our range bound and the second follows from our variance bound. The first claim then follows from the definition of $g(x)$ and the facts that $\exp(x) - x - 1 \le x^2/(2(1-x/3))$ for $0 \le x < 3$ and $\log(1+x) \le x$. The second claim follows from definition of $\hat{W}_i$ and the fact that $\mathbb{E}_i[\hat{W}_i] = \hat{r} f(X_i, Y_i)$. □

The following lemma is an analogue of Lemma 7 in [Beygelzimer et al., 2010].

Lemma 3. *Pick any $t \ge 0, p_{\min} > 0$ and let E be the joint event*

$$\frac{1}{n}\sum_{i=1}^{n} \hat{W}_i - \sum_{i=1}^{n} \mathbb{E}_i[\hat{W}_i] \ge (1+M)\sqrt{\frac{t}{2np_{\min}}} + \frac{t}{3np_{\min}}$$
$$\text{and } \min\{\frac{P_i}{\hat{r}^{(i)}_{m\to t}} : 1 \le i \le n \wedge \mathbb{E}_i[W_i] \ne 0\} \ge p_{\min} \tag{27}$$

Then $\Pr(E) \le e^{-t}$ *where* $M := \frac{1}{n}\sum_{i=1}^{n} \hat{r}_i$.

Proof. We follow [Beygelzimer et al., 2010] and let

$$\lambda := 3p_{\min} \frac{\sqrt{\frac{2t}{9np_{\min}}}}{1 + \sqrt{\frac{2t}{9np_{\min}}}} \tag{28}$$

Note that $0 < \lambda < 3p_{\min}$. By Lemma 2, we know that if $\min\{\frac{P_i}{\hat{r}^{(i)}_{m\to t}} : 1 \le i \le n \wedge \mathbb{E}_i[\hat{W}_i] \ne 0\} \ge p_{\min}$ then

$$\frac{1}{n\lambda}\sum_{i=1}^{n} \log \mathbb{E}_i[\exp(\lambda(W_i - \mathbb{E}_i[W_i]))] \le \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{r}_i \hat{r}^{(i)}_{m\to t}\lambda}{2P_i(1 - \frac{\hat{r}^{(i)}_{m\to t}\lambda}{3P_i})} \le M\sqrt{\frac{t}{2np_{\min}}} \tag{29}$$

and

$$\frac{t}{n\lambda} = \sqrt{\frac{t}{2np_{\min}}} + \frac{t}{3np_{\min}} \tag{30}$$

Let E' be the event that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{W}_i - \mathbb{E}_i[\hat{W}_i]) - \frac{1}{n\lambda}\sum_{i=1}^{n} \log \mathbb{E}_i[\exp(\lambda(\hat{W} - \mathbb{E}_i[\hat{W}]))] \ge \frac{t}{n\lambda} \tag{31}$$

and let E'' be the event $\min\{\frac{P_i}{\hat{r}^{(i)}_{m\to t}} : 1 \le i \le n \wedge \mathbb{E}_i[\hat{W}_i] \ne 0\} \ge p_{\min}$. Together, the above two equations imply $E \subseteq E' \bigcap E''$. By [Zhang, 2005]'s lemmas 1 and 2, $\Pr(E) \le \Pr(E' \bigcap E'') \le Pr(E') \le e^{-t}$. □

The following is an immediate consequence of the previous lemma.

Lemma 4. *Pick any $t \ge 0$ and $n \ge 1$. Assume $1 \le \frac{\hat{r}^{(i)}_{m\to t}}{P_i} \le r_{\max}$ for all $1 \le i \le n$, and let $R_n := \max\{\frac{\hat{r}^{(i)}_{m\to t}}{P_i} : 1 \le i \le n \wedge \mathbb{E}_i[\hat{W}] \ne 0\} \bigcup \{1\}$. We have*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} \hat{W}_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_i[\hat{W}_i]\right| \ge (1+M)\sqrt{\frac{R_n t}{2n}} + \frac{R_n t}{3n}\right) \le 2(2 + \log_2 r_{\max})e^{-t/2} \tag{32}$$

Proof. This proof follows identically to [Beygelzimer et al., 2010]'s lemma 8. □

We can finally bound Δ_4 by bounding the remaining free quantity M.

Lemma 5. *With probability at least $1-\delta$, the following holds over all $n \ge 1$ and $h \in H$:*

$$|\Delta_4| \le (2 + \left\|\hat{\theta}\right\|_2)\sqrt{\frac{\varepsilon_n}{P_{\min,n}(h)}} + \frac{\varepsilon_n}{P_{\min,n}(h)} \tag{33}$$

where $\varepsilon_n := \frac{16\log(2(2+n\log_2 n)n(n+1)|H|/\delta)}{n}$ *and* $P_{\min,n}(h) = \min\{P_i : 1 \le i \le n \wedge h(X_i) \ne h^*(X_i)\} \bigcup \{1\}$.

Proof. We define the k-sized vector $\tilde{\ell}(j) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i=j}\hat{\theta}(j)$. Here, $v(j)$ is an abuse of notation and denotes the jth element of a vector v. Note that we can write M by instead summing over labels, $M = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i = \sum_{j=1}^{k}\tilde{\ell}(j)$. Applying the Cauchy-Schwarz inequality, we have that $\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i \leq \frac{1}{n}\left\|\hat{\theta}\right\|_2 \left\|\dot{\ell}\right\|_2$ where $\dot{\ell}(j)$ is another k-sized vector where $\dot{\ell}(j) := \sum_{i=1}^{n}\mathbb{1}_{y_i=j}$. Since $\left\|\dot{\ell}\right\|_2 \leq n$, we have that $M \leq 1 + \left\|\hat{\theta}\right\|_2$. The rest of the claim follows by lemma 4 and a union bound over hypotheses and datapoints. □

The term Δ_1 is be bounded with Theorem 1. We now bound Δ_2. This is a simple generalization bound of an importance weighted estimate of f.

Lemma 6. *For any $\delta > 0$, with probability at least $1-\delta$, then for all $n \geq 1$, $h \in H$:*

$$|\Delta_2| \leq \frac{2d_\infty(P_{test}, P_{src})\log(\frac{2n|H|}{\delta})}{3n} + \sqrt{\frac{2d_2(P_{test}, P_{src})\log(\frac{2n|H|}{\delta})}{n}} \tag{34}$$

Proof. This inequality is a direct application of Theorem 2 from [Cortes et al., 2010]. □

The following lemma bounds the remaining term Δ_1.

Lemma 7. *For all $n \geq 1, h \in H$:*

$$|\Delta_1| \leq \|r_{s\to m}\|_\infty \, err(h_0, r_{s\to m}) \tag{35}$$

Proof. This inequality follows from our Lemma 1 and [Azizzadenesheli et al., 2019]'s Theorem 2. □

Theorem 5 follows by applying a triangle inequality over $\Delta_1, \Delta_2, \Delta_3, \Delta_4$. If a warm start of m datapoints sampled from P_{warm} is used, the deviation bound is instead:

$$\begin{aligned}
&|err(h, Z_{1:n}) - err(h^*, Z_{1:n}) - err(h) + err(h^*)| \\
&\leq \mathcal{O}\left((2 + \frac{n\|\theta_{u\to t}\|_2 + m\|\theta_{w\to t}\|_2}{n+m})\sqrt{\frac{\varepsilon_n}{P_{\min,n}(h)}} + \frac{\varepsilon_n}{P_{\min,n}(h)} + \frac{2d_\infty(P_{\text{test}}, P_{\text{src}})\log(\frac{2n|H|}{\delta})}{3(n+m)} \right. \\
&+ \sqrt{\frac{2d_2(P_{\text{test}}, P_{\text{src}})\log(\frac{2n|H|}{\delta})}{n+m}} + \frac{n}{n+m}\|r_{s\to m}\|_\infty \text{err}(h_0, r_{s\to m}) \\
&\left. + \frac{n}{\sigma_{\min}}\left(\|\theta_{m\to t}\|_2 \sqrt{\frac{\log\left(\frac{nk}{\delta}\right)}{\lambda n}} + \sqrt{\frac{\log\left(\frac{n}{\delta}\right)}{\lambda n}} + \sqrt{\frac{\log\left(\frac{n}{\delta}\right)}{n'}} + \|\theta_{s\to m}\|_\infty \text{err}(h_0, r_{m\to t}) \right) \right)
\end{aligned}$$

The only change is that variance and subsampling terms are scaled by $\frac{n}{n+m}$, both of which disappear in the limit where $n >> m$. For the remainder of this proof, we continue to set $m = 0$.

Theorem 2 follows by replacing the deviation bound in [Beygelzimer et al., 2010]'s Theorem 2 with our Theorem 5. Theorem 3 similarly follows from [Beygelzimer et al., 2010]'s Theorem 3 but with two additions. First, λn datapoints are sampled for label shift estimation. Second, the number of datapoints which are either accepted or rejected by the active learning algorithm can be much smaller than the number of datapoints sampled from P_{src} due to subsampling. We can determine this proportion with an upper-tail Chernoff bound.

Lemma 8. *When $\epsilon < 2^{(-2e-1)/\|r_{s\to m}\|_\infty}$, given n datapoints from P_{src}, subsampling will yield $\boldsymbol{n}$ where,*

$$\Pr\left(\boldsymbol{n} \geq \frac{n}{\|r_{s\to m}\|_\infty} + \log_2\left(\frac{1}{\epsilon}\right)\right) \leq \epsilon \tag{36}$$

Proof. The number of subsampled datapoints is sum of independent Bernoulli trials with mean μ,

$$\mu = \mathbb{E}_{y\sim P_{\text{src}}}[P_{\text{ss}}(y)] = \mathbb{E}_{y\sim P_{\text{src}}}\left[C\frac{P_{\text{med}}(y)}{P_{\text{src}}(y)}\right] = \mathbb{E}_{y\sim P_{\text{med}}}[C] = C \tag{37}$$

where C is a constant such that $C\frac{P_{\text{med}}(y)}{P_{\text{src}}(y)} \leq 1$ for all labels y. Thus, $\mu = C \leq 1/\|r_{s\to m}\|_\infty$. □

9 Supplementary Experiments

9.1 NABirds Regional Species Experiment

We conduct an additional experiment on the NABirds dataset using the grandchild level of the class label hierarchy, which results in 228 classes in total. These classes correspond to individual species and present a significantly larger output space than considered in Figure 6. For realism, we retain the original training distribution in the dataset as the source distribution; sampling I.I.D. from the original split in the experiment. To simulate a scenario where a bird species classifier is adapted to a new region with new bird frequencies, we induce an imbalance in the target distribution to render certain birds more common than others. Table 1 demonstrates the average accuracy of our framework at different label budgets. We observe consistent gains in accuracy at different label budgets.

Strategy	Acc (854 Labels)	Acc (1708)	Acc (3416)
MALLS (MC-D)	**0.51**	**0.53**	**0.56**
Vanilla (MC-D)	0.46	0.48	0.50
Random	0.38	0.40	0.42

Table 1: NABirds (species) Experiment Average Accuracy

9.2 Change in distribution

To further analyze the learning behavior of MALLS, we can analyze the label distribution of datapoints selected by the active learner. In Figure 8, MC-Dropout, Max-Margin and Max-Entropy strategies are evaluated on CIFAR100 under *canonical label shift*. By analyzing the uniformity bias and the rate of convergence to the target distribution, we can observe that MALLS exhibits a unique sampling bias which cannot be explained away as simply a class-balancing bias. This indicates that MALLS may be successful in recovering information from distorted uncertainty estimates.

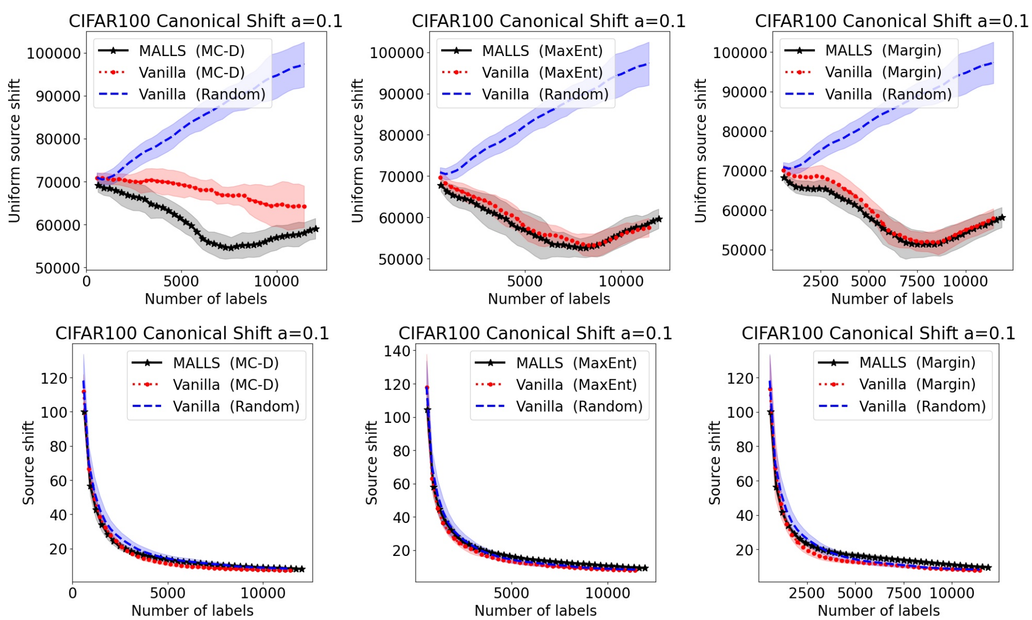


Figure 8: Average L2 distance between labeled class distribution and uniform/target distribution with 95% confidence intervals on 10 runs of experiments on CIFAR100 in the *canonical label shift* setting. MALLS (denoted by ALLS) converges to the target label distribution slower than vanilla active learning but with a similar uniform sampling bias. This suggests MALLS leverages a sampling bias different from that of vanilla active learning or naive class-balanced sampling.

10 Experiment Details

We list our detailed experimental settings and hyperparameters which are necessary for reproducing our results. Across all experiments, we use a stochastic gradient descent (SGD) optimizer with base learning rate 0.1, finetune learning rate 0.02, momentum rate 0.9 and weight decay 5e−4. We also share the same batch size of 128 and RLLS [Azizzadenesheli et al., 2019] regularization constant of 2e−6 across all experiments. As suggested in our analysis, we employ a uniform medial distribution to achieve a balance between distance to the target and distance to the source distributions. For computational efficiency, all experiments are conducted with minibatch-mode active learning. In other words, rather than retraining models upon each additional label, multiple labels are queried simultaneously. Table 2 lists the specific hyperparameters for each experiment, categorized by dataset. Table 3 lists the specific parameters of simulated label shifts (if any) created for individual experiments. Figure numbers reference figures in the main paper and appendix. "Dir" is short for Dirichlet distribution, "Inh" is short for inherent distribution, and "Uni" is short for uniform distribution.

Dataset	Model	# Datapoints	Epochs (init/fine)	# Batches	# Classes
NABirds1	Resnet-34	30,000	60/10	20	21
NABirds2	Resnet-34	30,000	60/10	20	228
CIFAR	Resnet-18	40,000	80/10	40	10
CIFAR100	Resnet-18	40,000	80/10	40	100

Table 2: Dataset-wide statistics and parameters

Figure	Dataset	Warm Ratio	Source Dist	Target Dist	Canonical?	Dirichlet α
5(a)	MNIST	0.1	Dir	Dir	Yes	0.1
5(b)	CIFAR	0.4	Dir	Dir	Yes	0.4
6(a-b)	CIFAR100	0.4	Dir	Dir	Yes	0.1
6(c-d)	NABirds1	1.0	Inh	Inh	No	N/A
7(a-b)	CIFAR	0.3	Dir	Dir	Yes	0.7
7(c)	CIFAR	0.3	Dir	Dir	Yes	0.7
7(d)	CIFAR100	0.4	Dir	Dir	Yes	0.1
8(a)	CIFAR100	0.4	Dir	Dir	Yes	3.0
8(b)	CIFAR100	0.4	Dir	Dir	Yes	0.7
8(c)	CIFAR100	0.4	Dir	Dir	Yes	0.4
8(d)	CIFAR100	0.4	Dir	Dir	Yes	0.1
9(a)	CIFAR100	0.4	Dir	Uni	No	1.0
9(b)	CIFAR100	0.3	Uni	Dir	No	0.1
9(c-d)	CIFAR100	0.4	Dir	Dir	Yes	0.1
T1(g-i)	NABirds1	1.0	N/A	Dir	No	0.1
8	CIFAR100	0.4	Dir	Dir	Yes	0.1

Table 3: Label Shift Setting Parameters (in order of paper)