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.

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.

![Image of data distribution](https://github.com/ericzhao28/)

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 trade-off 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/all.

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 successfully 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 $P_{\text{warm}}$ initially populate $S$ and constitute the “warm start” dataset $D_{\text{warm}}$. L samples unlabeled datapoints $D_{\text{ulb}}$ from some distribution $P_{\text{ulb}}$, and may select up to $n$ for labeling and appending to $S$. The learned hypothesis $h$ is evaluated on a test distribution $P_{\text{test}}$.

Traditional active learning assumes,

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

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

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

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:

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

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 ($P_{\text{src}}$) and target ($P_{\text{trg}}$) 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:

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

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

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

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 $P_{\text{src}}$ where $\mathbb{E}[C_h[i,j]] := P_{\text{src}}(h(X)=y^{(j)}|X=x^{(i)})$ and $q_h$ denote a $k$-vector with $q_{h[i]} := P_{\text{src}}(h(X)=y^{(j)})$. 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,

$$r := \frac{P_{\text{trg}}(y)}{P_{\text{src}}(y)} = C_h^{-1} q_h.$$  

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.
While in finite settings only the former yields IID samples, we assume subsampling occurs at the limit and use the 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 \(P_{\text{med}}\) from a source \(P_{\text{src}}\) using predictor \(\phi\) for predicting proxy labels.

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

2. **Subsampling with the target** \(P_{\text{med}}\). Collect \(N\) datapoints from \(P_{\text{src}}\) 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 \(P_{\text{med}}\), 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 \(\phi\):

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

where \(Q_i \in \{0, 1\}\) is an indicator variable for whether the \(i\)th 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* \(P_{\text{med}}\). As we will show, \(P_{\text{med}}\) 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 \(r_{x-r} := P_{\text{trg}}(y) / P_{\text{src}}(y)\) denote the importance weights which shift the source to the target. Similarly, let \(r_{x-m} := P_{\text{med}}(y) / P_{\text{src}}(y)\) and \(r_{m-t} := P_{\text{trg}}(y) / P_{\text{med}}(y)\) 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 \(\theta := r - 1\), usually the L2 norm \(\|\cdot\|_2\). A large \(\|\theta\|\) corresponds to a larger label shift and harder learning problem. \(\|\theta_{m-t}\|\) is the amount of label shift corrected by subsampling and \(\|\theta_{m-t}\|\) 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 \(\Delta\) denote the subsampling and importance weighting (trained on \(n\) datapoints) estimation error of the empirical loss of \(N\) datapoints:

$$\Delta := \frac{1}{N} \sum_{i=1}^{N} (r_x P_{\text{ss}}(y_i) - \hat{r}_x P_{\text{ss}}(h(x_i))) \ell(h(x_i), y_i),$$

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

$$|\Delta| \leq O \left( \frac{2}{\sigma_{\min}} \left( \frac{\|\theta_{m-t}\|_2}{\sqrt{n}} \sqrt{\log \left( \frac{2n}{\delta} \right)} + \frac{\sqrt{\log (\frac{2n}{\delta})}}{\sqrt{n}} + \|\theta_{m-t}\|_{\infty} \text{err}(h_0, r_{m-t}) \right) \right),$$

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

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, \(\|\theta_{m-t}\|_2 \sqrt{\frac{\log (\frac{2n}{\delta})}{n}}\), and the third summand, \(\|\theta_{m-t}\|_{\infty} \text{err}(h_0, r_{m-t})\). 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, \( \|r_{st}\| \). 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, \( \|\theta_{st}\| \), and the accuracy of the blackbox hypothesis error \( \text{err}(h_0, r_{st}) \). Hence, subsampling mitigates sensitivity to label shift magnitude by splitting the norm of total label shift, \( \|\theta_{st}\| \), into the sum of two factors which scale with \( \|\theta_{st}\| \) and \( \|\theta_{m}\| \).

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

**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 *imbalance source* and *imbalance 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 \( n \) data points 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 \( P_{src} := D_2 \) and \( P_{trg} := D_1 \), \( n - 2 \) additional samples from the under-represented class are necessary for negating label shift. Under imbalanced target, where \( P_{src} := D_1 \) and \( P_{trg} := D_2 \), \( (n - 2)2^n \in \mathcal{O}(n^2) \) 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 \( X \).

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 \( \lambda \), blackbox predictor \( h_0 \), medial distribution \( P_{med} \), hypothesis class \( \mathcal{H} \)

**Initialize** the dataset \( S \leftarrow \text{warm start} \)

**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 \( \lambda n \) labeled datapoints from the unlabeled set;
- While \( |S| < n \)
  - Calculate IWAL-CAL [Beygelzimer et al., 2010] sampling probability \( P_t \) for \( x_t \), using \( S \), weighting by \( r \);
- Label and append \( (x_t, y_t) \) to \( S \) with probability \( P_t \);

**Output:** \( h_T := \text{argmin}_{h \in \mathcal{H}} \text{err}(h, r) \) where \( \text{err} \) 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 the label complexity of the dataset \( S \):

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

(10)
where \( x_i, y_i \) are drawn from \( S \). MALLS instead computes empirical loss estimates as:

\[
err(h, r) = \frac{1}{|S|} \sum_{i=1}^{|S|} r(y_i) \ell(h(x_i), y_i),
\]

(11)

where \( r(y_i) \) denotes an importance weight for datapoints of label \( y_i \). 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) [Azizadehshesl 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 \( \lambda n \) labeled datapoints sampled from \( P_{\text{med}} \) 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 \( D_{\text{aux}} \), IWAL-CAL instead interacts with datapoints subsampled from \( D_{\text{aux}} \) and distributed according to \( P_{\text{med}} \). We detail the high-level flow of MALLS in Figure 4 and provide pseudocode in Algorithm 1.

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 \( n \geq 1 \),

\[
err_Q(h_n) \leq O \left( \left( 1 + \frac{1}{\sigma_{\text{min}}} \| r_{s-r} \|_\infty \right) err(h_0, r_{m-r}) \right)
\]

\[
+ err_Q(h^*) + \sqrt{\frac{2C_0 \log n}{n - 1}} + \frac{2C_0 \log n}{n - 1} \right).
\]

(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,

\[
C_0 \in O \left( \frac{2}{\lambda \sigma_{\text{min}}} \left( \| \theta_{m-r} \|_2^2 \log \left( \frac{k}{\delta} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) + \log \left( \frac{H}{\delta} \right) (1 + \| \theta_{s-r} \|_2^2).
\]

(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: \( \frac{1}{\sigma_{\text{min}}} err(h_0, r_{m-r}) \). (2) In the original IWAL-CAL algorithm \( C_0 \in O(\log(|H|/\delta)) \). However label shift inevitably introduces, to the constant \( C_0 \), a dependence on the number of label classes \( k \) and label shift magnitudes \( \| \theta_{s-r} \|_2^2 \) and \( \| \theta_{m-r} \|_2^2 \). 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:

\[
O \left( 1 + \log \left( \frac{1}{\delta} \right) + \Theta \sqrt{\frac{C_0 n}{\| \theta_{s-r} \|_\infty^2} \log n} + \Theta C_0 \log^3 n + \lambda n + \Theta \cdot (n - 1) \cdot \left( \frac{err_Q(h^*)}{\| \theta_{s-r} \|_\infty^2} + \left( 1 + \frac{1}{\sigma_{\text{min}}} \right) err(h_0, r_{m-r}) \right) \right).
\]

(14)

where \( \Theta \) 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 \( \frac{1}{\| \theta_{s-r} \|_\infty} \). Importance

\[\text{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.
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 ($\|\theta_m\|$) and $\lambda$ 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 $r_{m-t} = \sqrt{r_{s-t}}$. 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
Active Learning under Label Shift

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 $\pi$, batch size $B$; hypothesis class $\mathcal{H}$

Initialize the labeled dataset $S_0 \leftarrow$ the warm start set;

For $t \in \{1, \ldots, T\}$
- Find importance weights $r_t \leftarrow \text{RLLS}(S_t, \phi, P_{med})$;
- Train hypothesis $\phi$ on $S_t$ weighted by $r_t$;
- For $y \in \mathcal{Y}$
  - Number of datapoints to collect $k := B \times P_{med}(y)$
  - Find top-$k$ most uncertain datapoints of label $y$: $D_y := \text{top-}k\left(\pi, \{x \in D_{ulb} \setminus S_{t-1} \mid \phi(x) = y\}\right)$
  - 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
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 $\alpha$. Note that shift magnitude is inversely correlated with $\alpha$—smaller $\alpha$ 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


