arXiv:2007.08479v1 [cs.LG] 16 Jul 2020

Active Learning under Label Shift

Eric Zhao **Anqi Liu** **Animashree Anandkumar** **Yisong Yue**
{elzhao,anqiliu,anima,yyue}@caltech.edu
Department of Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91126, USA

Abstract

Distribution shift poses a challenge for active data collection in the real world. We address the problem of active learning under *label shift* and propose ALLS, the first framework for active learning under label shift. ALLS builds on label shift estimation techniques to correct for label shift with a balance of importance weighting and class-balanced sampling. We show a bias-variance trade-off between these two techniques and prove error and sample complexity bounds for a disagreement-based algorithm under ALLS. Experiments across a range of label shift settings demonstrate ALLS consistently improves performance, often reducing sample complexity by more than half an order of magnitude. Ablation studies corroborate the bias-variance trade-off revealed by our theory.

1 Introduction

Distribution shift poses a significant challenge for traditional active learning techniques. We study how to effectively perform active learning under *label shift*, an important but often overlooked form of distribution shift. Label shift arises when class proportions differ between training and testing distributions but the feature distributions of each class are unchanged. For instance, the problem of training a bird classifier using data from a different geographical region poses a label shift problem: while the likelihood of observing a subspecies (i.e. $p(y)$) varies by region, members of a subspecies look the same (i.e. $p(x \mid y)$) regardless of location. The problem of active learning under label shift is particularly important for adapting existing machine learning models to new domains or addressing under-represented classes in imbalanced datasets [1, 2]. This problem is also relevant to the correction of societal bias in datasets, such as the important concern of minority under-representation in computer vision datasets [3]. Label shift is also a helpful heuristic for addressing general distribution shift. For instance, an importance weighting function for addressing data shift in classification problems can be approximated with a finite set of importance weights by reducing to a label shift problem [1].

Current techniques for active learning under distribution shift, sometimes termed "active domain adaptation", either rely on heuristics for correcting general forms of distribution shift [4, 5] or build on the assumption of *covariate shift* [6–8]. Unlike the covariate shift setting, active learning under label shift is complicated by the fact that the underlying distribution shift is associated with labels, which cannot be observed in unlabeled datapoints.

Algorithm contributions We build on (1) importance weight estimation methods from prior literature on learning under label shift [1, 9] and (2) *subsampling* heuristics from literature on active learning under class imbalance [10]. We pose and address an important question of whether to correct for label shift by importance weighting or sampling from under-represented classes. We present a novel framework for active learning under label shift (ALLS) which unifies and balances the use of importance weighting and subsampling to adapt to label shifts of varying forms and strengths. To the

Preprint. Under review.

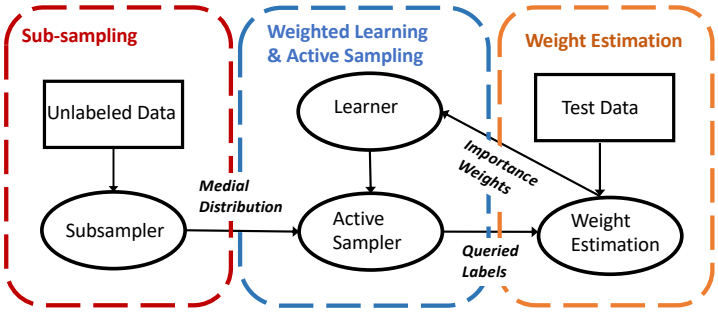


Figure 1: The ALLS Framework. ALLS consists of 3 routines: (1) subsample from the unlabeled set, (2) estimate importance weights for correcting label shift, and (3) actively query for labels and update learner. Details are illustrated in Sec. 3 and 4.

best of our knowledge, ALLS is the first active learning framework for general label shift settings. Figure 1 shows a flow chart of our framework.

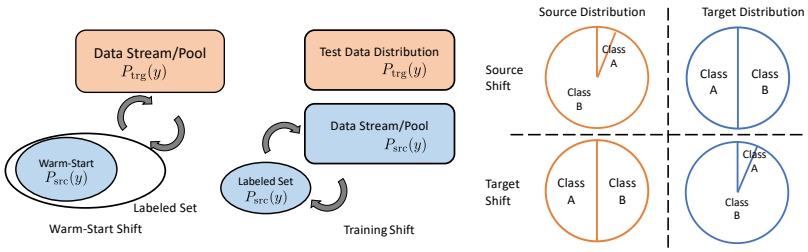
Theoretical contributions We derive label complexity and generalization PAC bounds for ALLS, the first such guarantees for this setting, by instantiating our framework on a well-studied disagreement-based active learning algorithm (IWAL-CAL). To this end, we formalize the practice of subsampling through the concept of a *medial distribution*. Our analysis shows that label shift estimation and importance weighting techniques preserve the provable consistency of IWAL-CAL with similar asymptotic sample complexity bounds. Our analysis also reveals a bias-variance trade-off between using importance weighting and subsampling. Specifically, we show that subsampling introduces a bias which scales with the minimum achievable error while importance weighting introduces variance which scales with label shift magnitude.

Empirical contributions We further instantiate our framework with various uncertainty sampling algorithms and empirically demonstrate that our framework outperforms both the original active learning algorithm and random sampling—even when the original active learning algorithm fails under strong label shift. We show the effectiveness of ALLS across both synthetic label-shift settings in CIFAR10, CIFAR100 [11], and under natural label-shift settings in the NABirds dataset [12]. To help close the gap between the theory and practice of learning under label shift, we present best practices which scale label shift estimation techniques to deep learning settings. We also present extensive ablation studies which corroborate our theoretical insights into the trade-off between importance weighting and subsampling.

2 Preliminaries

Active Learning under Distribution Shift In an active learning problem, a learner L actively queries an oracle for labels on strategically selected datapoints. The learner L begins with a prelabeled pool $\mathbf{D}_{\text{warm}}$ of m "warm start" datapoints sampled IID from distribution P_{warm}. The learner seeks to maximize its performance on a test distribution, P_{test}. In a pool-based setting, L accesses a pool $\mathbf{D}_{\text{ulb}}$ of n unlabeled datapoints drawn IID from distribution P_{ulb}. L can view all unlabeled datapoints and progressively selects datapoints from $\mathbf{D}_{\text{ulb}}$ to be labeled and added to a labeled set S. In an online setting, L only observes one (or a batch) of datapoints from $\mathbf{D}_{\text{ulb}}$ at a time and must immediately decide to discard or label.

Traditional active learning settings assume $P_{\text{ulb}} = P_{\text{warm}} = P_{\text{test}}$, which is rarely the case in practice. The setting where $P_{\text{ulb}} = P_{\text{test}}$ but $P_{\text{warm}} \neq P_{\text{test}}$ is well-studied and known as the *active domain adaptation* problem. We refer to this as the *canonical label shift* setting. The more *general label shift* setting where the assumption that $P_{\text{ulb}} = P_{\text{test}}$ is lifted has received comparatively little attention despite its practical relevance: soliciting labels from the test distribution is often impractical. We investigate both the *canonical label shift* and this more challenging *general label shift* setting, as illustrated in figure 2 (a). In either shift setting, the task of adapting to the test distribution necessarily assumes access to an unlabeled pool of samples, $\mathbf{D}_{\text{test}}$, sampled IID from P_{test}.



(a) (b)

Figure 2: (a) Diagram of *canonical label shift* and *general label shift* settings; (b): Example of *imbalanced source* and *imbalanced target* settings in a binary classification problem.

Label Shift The distribution shift problem concerns training and evaluating models on different distributions, termed the source (P_{src}) and target (P_{trg}) respectively. Due to the difficulty of the distribution shift problem in a *general label shift* setting, assumptions on the nature of the distribution shift are necessary. Covariate shift, the most popular and widely analyzed form of distribution shift, assumes that the underlying distribution shift arises solely from a change in the input distribution where $P_{\text{trg}}(x) \neq P_{\text{src}}(x)$, while conditional label probabilities are unaffected: $P_{\text{trg}}(y|x) = P_{\text{src}}(y|x)$. In contrast, label shift—the subject of this paper—assumes distribution shift arises solely from a change in label marginals where $P_{\text{trg}}(y) \neq P_{\text{src}}(y)$ but the anti-causal conditionals are unaffected: $P_{\text{trg}}(x|y) = P_{\text{src}}(x|y)$. These shift assumptions are illustrated in Figure 3.

Importance weighting provides an important shift correction method for both covariate and label shift settings. Under label shift, weighting datapoints by their likelihood ratio $\frac{P_{\text{trg}}(y)}{P_{\text{src}}(y)}$ produces asymptotically unbiased importance weighted estimators.

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

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 $\frac{P_{\text{trg}}(y)}{P_{\text{src}}(y)}$ with only labeled data from the source distribution, unlabeled data from the target distribution, and a blackbox hypothesis h [1]. Let C_h denote a finite sample confusion matrix for h on P_{src} where $\mathbb{E}[C_h[i,j]] := P_{\text{src}}(h(X) = y_i, Y = y_j)$ and define vector q_h where $q_h[i] := \widehat{P_{\text{trg}}}(h(X) = y_i)$. Assuming $\forall y : P_{\text{trg}}(y) > 0 \implies P_{\text{src}}(y) > 0$, it holds that,

$$P_{\text{trg}}(h(X) = y_i) = \sum_{j=1}^{k} P_{\text{src}}(h(X) = y_i, Y = y_j)\frac{P_{\text{trg}}(y_j)}{P_{\text{src}}(y_j)} \tag{1}$$

$$r := \frac{P_{\text{trg}}(y)}{P_{\text{src}}(y)} = C_h^{-1} q_h \tag{2}$$

For instance, Regularized Learning under Label Shift (RLLS) finds importance weights r through convex optimization of equation 3 where λ is some regularization constant [9]:

$$C_h^{-1} q_h \approx \text{argmin}_r \left\| C^{-1} r - b \right\| + \lambda \left\| r - 1 \right\|, \tag{3}$$

Subsampling by Class The class imbalance problem arises when the label distributions of a dataset are highly imbalanced. Prior literature on active learning under class imbalance prescribe a variety of class-based sampling techniques which adjust the sampling likelihood of datapoints associated with rare classes. We build on an intuitive and simple strategy for class-based sampling which filters out each datapoint x_t according to their label with probability $1 - \frac{k\sum_{i=1}^{n}\mathbb{1}[y_i = y_t]}{n}$. For batch-mode pool-based settings, we build on an analogous—but lower variance—strategy of mandating that c datapoints of each class are labeled from every batch. In practice, since labels are hidden, a classifier ϕ is necessary to guess labels. To generalize these tools for our setting, we frame the class imbalance problem as a form of label shift with a target distribution known a-priori. We can accordingly

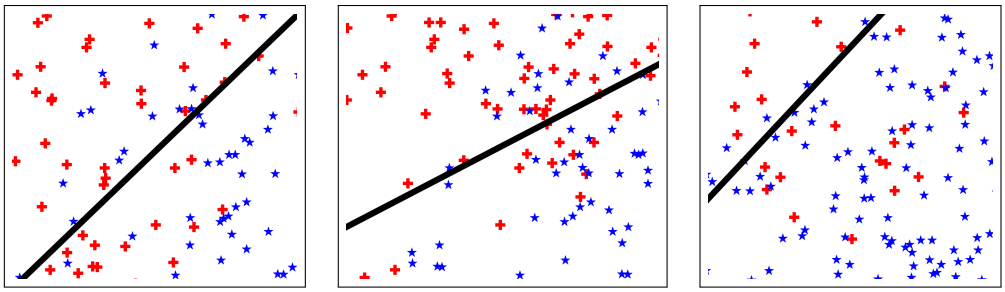


Figure 3: (a) Binary classification data separated by optimal hypothesis (black line); (b) Covariate shift featuring dense top-right corner; (c) Label shift featuring higher density of the blue class.

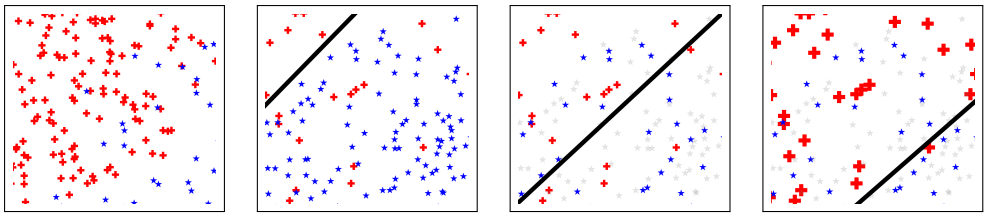


Figure 4: (a) Target data, red class dominant; (b) Source data, blue class dominant; (c) Apply subsampling for uniform medial. (d) Importance weighting for red dominance. Black line denotes ERM. Larger markers indicate larger importance weights.

generalize class-balanced sampling techniques for arbitrary target distributions, a practice we term *subsampling* and depict in Algorithms 1, 2. Here, filter distribution P_{ss} is defined as $P_{\text{ss}} := \frac{P_{\text{tar}}}{P_{\text{src}}}$ for some choice of target and source distributions $P_{\text{tar}}, P_{\text{src}}$. Note that a choice of filter distribution $P_{ss}[y] = \frac{k \sum_{i=1}^{n} \mathbb{1}[y_i = y]}{n}$ and target distribution $P_{tar}[y] := \frac{\mathbf{1}}{\mathbf{k}}$ in Algorithms 1, 2 respectively exactly coincide with their class-imbalance counterparts.

Algorithm 1 Subsampling (General)

Input: Hypothesis $\phi \in H$, unlabeled $x \in X$, filter distribution P_{ss}, policy $\pi : X \to [0, 1]$.

Guess label $y \in Y$ for x as $y := \phi(x)$.
With probability $1 - P_{\text{ss}}(y)$, exit.
Otherwise, sample according to $\pi(x)$.

Algorithm 2 Subsampling (Batch-mode)

Input: Hypothesis $\phi \in H$, unlabeled pool $\mathbf{D} \subset X$, distribution P_{tar}, batch size B, policy $\pi_m : (m', X^m) \to [0, 1]^m$.
For $y \in Y$
  Create sub-batch $\mathbf{D}_y := \{x \in \mathbf{D} \mid \phi(x) = y\}$
  Sample according to $\pi_{|\mathbf{D}_y|}(BP_{\text{tar}}(y), \mathbf{D}_y)$.

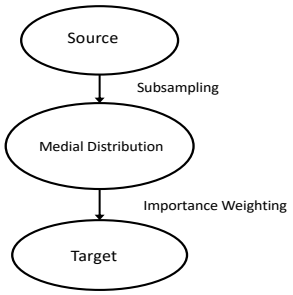
3 The ALLS Framework

In this section, we present a new learning framework: Active Learning under Label Shift (ALLS). We first present a unified view of subsampling and importance-weighting as shift-correction techniques and pose a key question regarding the trade-off between the two strategies. We then detail our proposed framework and discuss its online and pool-based versions.

Both importance-weighting and subsampling serve to "correct" label shift. As previously noted, importance weights r as defined in Equation 2 provide for asymptotically unbiased estimators. Subsampling functions similarly; in the expectation over the randomness of subsampling, subsampling (Algorithm 1) with a label oracle as ϕ is equivalent to importance weighting. To see this, note that the effect of subsampling on estimators is equivalent to multiplying samples with a random bit. Then,

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

where $Q_i \in \{0, 1\}$ is an independent random variable with conditional expectation $\mathbb{E}[Q_i \mid y_i] = P_{\text{ss}}(y_i)$. Here, Q_i captures the function of subsampling with label oracle ϕ. Due to their similarity, importance weighting and subsampling can be combined to correct label shift. Figure 4 depicts a label shift scenario where subsampling partially corrects for label shift and importance weighting corrects the remaining label shift.



Medial Distribution It is helpful to conceptually frame the use of subsampling as a form of intentional label shift applied to the source data which induces a new distribution. We refer to this implicit distribution as the *medial distribution* P_{med}, as the choice of this distribution mediates the balance between subsampling and importance weighting. We can then formalize the "amount" of subsampling versus importance weighting used by the distance of P_{med} to P_{ulb} and to P_{test}. To formalize an intuition for label shift distance, we follow [1] and set $\theta := r - \mathbf{1}$; then, $\|\theta\|$ corresponds to the shift magnitude between whatever source and target distribution that r is defined for. We define $\theta_{u \to m}$ and $\theta_{m \to t}$ accordingly for the shift between the unlabeled and medial, and the medial and test distributions respectively.

Proposed Algorithm Our proposed framework, ALLS, is depicted in Figure 1 and detailed in Algorithm 3. In addition to a primary active learning loop, ALLS diverts a fraction (λ) of datapoints in $\mathbf{D}_{\text{ulb}}$ to accumulate an independent holdout dataset O_t, which is used for estimating label shift weights r and training a hypothesis ϕ for subsampling. In the primary active learning loop, ALLS first subsamples according to ϕ and then samples according to an active learning policy π, weighting any empirical risk or uncertainty estimates with the importance weights r. We use Regularized Learning under Label Shift (RLLS) [9] for label shift estimation, but any blackbox method (e.g. BBSE [1]) can plug into ALLS.

Holdout Set Label shift estimation techniques such as BBSE and RLLS [1,9] require independence between the data used for estimating importance weights and the data used for the main learning task. This poses a challenge as we thus require a holdout dataset O'_t drawn IID from P_{med} but which is independent of S_t and ϕ. To this end, we use a trick for mimicking IID draws from P_{med} using a buffer O_t of IID draws from P_{ulb}. Hence, the motivation for the use of holdout set O_t for label shift estimation and training ϕ is purely theoretical [9]. While necessary for our theoretical analysis, the use of O_t renders our practically-motivated version of ALLS intractable due to a need for knowledge of both P_{ss} and P_{med}, only one of which can be known at a time. Thus, in practice, this ALLS variant forgoes the use of a holdout set for label shift estimation, as suggested by [9], and learns r and ϕ on the main labeled set S instead.

Algorithm 3 Active Learning under Label Shift (ALLS)

Input: $\mathbf{D}_{\text{warm}}$, $\mathbf{D}_{\text{ulb}}$, $\mathbf{D}_{\text{test}}$, P_{ss}, P_{med}, λ, blackbox h_0, initial holdout set O_0, max timestep T, label oracle C, policy π
Initialize: $r_0 \leftarrow \text{RLLS}(O_0, P_{\text{test}}, h_0)$, $S_0 \leftarrow \{(x_i, y_i, r_0(y_i))\}$ for $(x_i, y_i) \in P_{\text{warm}}$
For $x_t, y_t \in P_{\text{warm}} \setminus O_0$ append $S_0 \leftarrow \{(x_t, y_t, r_0(y_t))\}$
For $t < T$
  Label λn datapoints into holdout set: $O_t \leftarrow O_{t-1} \bigcup \{x_i, y_i, P_{\text{ss}}(y_i)\}_{i=1}^{\lambda n}$;
  Populate O'_t with $(x_i, y_i, p_i) \in O_t$ sampled w.p. $\frac{p_i}{\max_{j \in [1, |O_t|]} p_j}$;
  Train ϕ on O_t and $r \leftarrow \text{RLLS}(O'_t, h_0)$;
  $\{x_t, P_t\} \leftarrow \text{Subsample}(\phi, x_t, P_{\text{ss}}, \pi, r)$; # Also pass weights r to sampling subroutine.
  Label and add to set $S_t \leftarrow S_{t-1} \bigcup \{x_t, C(x_t), P_t\}$;
Output: $h_T = \text{argmin}\{\text{err}(h, S_T, r_T) : h \in \mathcal{H}\}$

4 Theoretical Analysis

We now analyze label complexity and generalization bounds for Algorithm 3 by instantiating ALLS on IWAL-CAL [13], an agnostic active learning algorithm with rigorous guarantees. Let $\text{err}(h, S_i) \to [0, 1]$ denote the error of $h \in H$ as estimated on S_i while $\text{err}(h)$ denote the expected error of h on

P_{test}. We next define,

$$\begin{aligned} h^* &:= \text{argmin}_{h \in H} \text{err}(h), \quad h_k := \text{argmin}_{h \in H} \text{err}(h, S_{k-1}), \\ h'_k &:= \text{argmin}\{\text{err}(h, S_{k-1}) \mid h \in H \wedge h(\mathbf{D}^{(k)}_{\text{unlab}}) \neq h_k(\mathbf{D}^{(k)}_{\text{unlab}})\} \\ G_k &:= \text{err}(h'_k, S_{k-1}) - \text{err}(h_k, S_{k-1}) \end{aligned} \tag{4}$$

IWAL-CAL corresponds to selecting a sampling policy π which outputs sampling probability $P_t = \min\{1, s\}$ for the $s \in (0, 1)$ which solves,

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

where C_0 is as defined in Theorem 1 and $c_1 := 5 + 2\sqrt{2}, c_2 := 5$. For the remainder of this section, we work in the challenging *general label shift* setting. As the presence of warm start data is not particularly interesting in our analysis, we set $m = 0$ for reading convenience and defer the case where $m > 0$ to Appendix C for interested readers.

4.1 Theoretical Guarantees

Let $\sigma_{\min}$ denote the smallest singular value of blackbox hypothesis h_0, and $P_{\min,n}(h) := \min_{h(x_i) \neq h^*(x_i)} P_i$ the minimum sampling probability in the disagreement region of h and h'. We also denote the noise rate of the subsampling problem with $\text{err}_W(h^*) := \min_{h \in H} \mathbb{E}_{x,y \in P_{\text{ulb}}}[\frac{P_{\text{med}}(y)}{P_{\text{ulb}}(y)} - \frac{P_{\text{med}}(h(x))}{P_{\text{ulb}}(h(x))}]$. We now present generalization and sample complexity bounds for IWAL-CAL on ALLS.

Theorem 1. *With at least probability $1 - \delta$, for all $n \geq 1$,*

$$err(h_n) \leq err(h^*) + \sqrt{\frac{2C_0 \log n}{n-1}} + \frac{2C_0 \log n}{n-1} + \mathcal{O}\left((\|\theta_{m \to t}\|_2 + 1) err_W(h^*_{online})\right) \tag{5}$$

where

$$\begin{aligned} C_0 \in \mathcal{O}\Bigg(\log\left(\frac{|H|}{\delta}\right)\Big(d_\infty(P_{test}||P_{ulb}) + d_2(P_{test}||P_{ulb}) + 1 + \|\theta_{u \to t}\|_2^2\Big) \\ + \frac{\log\left(\frac{k}{\delta}\right)}{\sigma^2_{\min}} d_\infty(P_{test}||P_{med}) \|\theta_{m \to t}\|_2^2 \left(err_W(h^*_{online}) + 1\right)\Bigg) \end{aligned} \tag{6}$$

Our generalization bound differs from the original IWAL-CAL bound in two key aspects. (1) The use of subsampling introduces a new constant term which scales with the noise rate of the subsampling estimation task: $\text{err}_W(h^*_{\text{online}})$. (2) Most terms are now scaled by magnitude of the label shift; the largest such label shift terms arise from the variance of importance weighting. Aside from the constant noise rate term, however, ALLS preserves the $\log(n)/n + \sqrt{\log(n)/n}$ asymptotic bound of IWAL-CAL. In addition, when only importance weighting is used ($P_{\text{med}} = P_{\text{ulb}}$), the subsampling learning problem is trivial. Accordingly, the subsampling noise rate is zero: $\text{err}_W(h^*_{\text{online}}) = 0$. In this case, ALLS preserves the consistency guarantee of IWAL-CAL even under *general label shift*.

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

$$1 + (\lambda + \Theta \cdot (2err(h^*) + \|\theta_{m \to t}\|_2 err_W(h^*))) \cdot (n-1) + \mathcal{O}\left(\Theta\sqrt{C_0 n \log n} + \Theta C_0 \log^3 n\right), \tag{7}$$

where Θ denotes the disagreement coefficient [14].

Besides the changes to C_0 noted in our discussion of the generalization bound, we note two differences with the sample complexity given in traditional IWAL-CAL. First, we introduce two additional linear terms into the sample complexity: one corresponding to the bias of subsampling (again proportional to noise rate of the subsampling problem) and one corresponding to the accumulation of holdout set H_t (proportional to λ). These accompany a linear term proportional to the noise rate of the original learning problem, which is also present in the original IWAL-CAL bounds and unavoidable in agnostic active learning.

4.2 Bias-Variance Trade-off

We note a key bias-variance trade-off in the use of importance weighting and subsampling. In agnostic learning problems, labels cannot be predicted with certainty and hence a non-trivial subsampling strategy will always incur errors. This introduces a constant bias term into both generalization and sample complexity bounds, a term which linearly scales with the subsampling noise rate $\text{err}_W(h^*)$. In contrast, importance weighting suffers from high-variance as importance weights can easily grow to large values. Thus, importance weighting introduces a multiplicative factor into our bounds which scales quadratically with importance weight magnitudes: $\|\theta_{m\to t} + 1\|_\infty \|\theta_{m\to t}\|_2^2$. The key to addressing this trade-off is minimizing some combination of $\text{err}_W(h^*)$ and $\|\theta_{m\to t}\|_2$. This requires striking a balance between the two as, assuming a reasonable choice of P_{med}, decreasing $\|\theta_{m\to t}\|$ increases $\|\theta_{u\to m}\|$ and thus $\text{err}_W(h^*)$.

To inform a choice of medial distribution, we now analyze two common label shift regimes which we term *imbalanced source* and *imbalanced target* and depict in Figure 2 (b). 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 $P_{\text{test}} := D_1$, $n-2$ additional samples from the under-represented class are necessary for negating label shift. Under *imbalanced target*, where $P_{\text{src}} := D_1$ and $P_{\text{test}} := D_2$, $(n-2)\frac{n}{2} \in \mathcal{O}(n^2)$ additional samples from the under-represented class are necessary for negating label shift. While these two scenarios feature label shifts of identical magnitude, subsampling is more efficient under *imbalanced source*. This suggests a simple heuristic to subsample when under *imbalanced source* and importance weight when under *imbalanced target*. We can extend this heuristic to the generic label shift case by noting that a choice of uniform medial distribution precisely decomposes every label shift problem into the combination of a *imbalanced source* and *imbalanced target* problem. Thus, as we verify experimentally, a uniform distribution serves as a reliable baseline choice for medial distributions.

5 Experiments

We now present empirical evaluations of our pool-based ALLS framework on real-world species recognition dataset NABirds [12] and benchmark datasets CIFAR10 & CIFAR100 [11]. We demonstrate that ALLS improves active learning performance under a diverse range of label shift scenarios.

Scaling Label Shift Estimation Due to the largely theoretical focus of prior literature on label shift estimation, existing label shift estimation techniques often fail to scale when used out-of-the-box. To scale these techniques for use on deep neural networks on high-dimensional datasets, we introduce two techniques: posterior regularization (PR) and iterative reweighting (ITIW). Posterior regularization avoids applying importance weights to the loss function by instead applying importance weights during inference time: $p'(y) = p(y)\frac{r(y)}{\sum_i r(y_i)}$. This reduces variance while preserving the use of the label shift information to correct uncertainty estimation. This technique bears relation to the expectation-maximization algorithm described in [15]. Iterative reweighting uses hypotheses trained with importance weights to estimate better importance weights, an iterative process which helps steer its finite sample confusion matrix estimates away from singularity.

Methods We evaluate our ALLS framework on instantiations of uncertainty sampling algorithms: (1) Monte Carlo dropout (MC-D), where uncertainty is given by disagreement between forward passes due to dropout; [16] (2) Maximum Entropy sampling (MaxEnt), given by the entropy of predictive distributions; and (3) Maximum Margin (Margin): given by the gap in logits of the most and second-most likely classes. As baselines, we compare against the original active learning algorithm (marked *Vanilla*) and random sampling. In ablation studies, we also compare against partial applications of ALLS which only use either importance weighting or shift correction. Further results and experiment details can be found in the appendix, including a link to source code.

5.1 Primary Results

We present our primary experimental results in Figure 5. In these experiments, we apply ALLS to training Resnet18 models with batch-mode pool-based active learning on three datasets: CIFAR10,

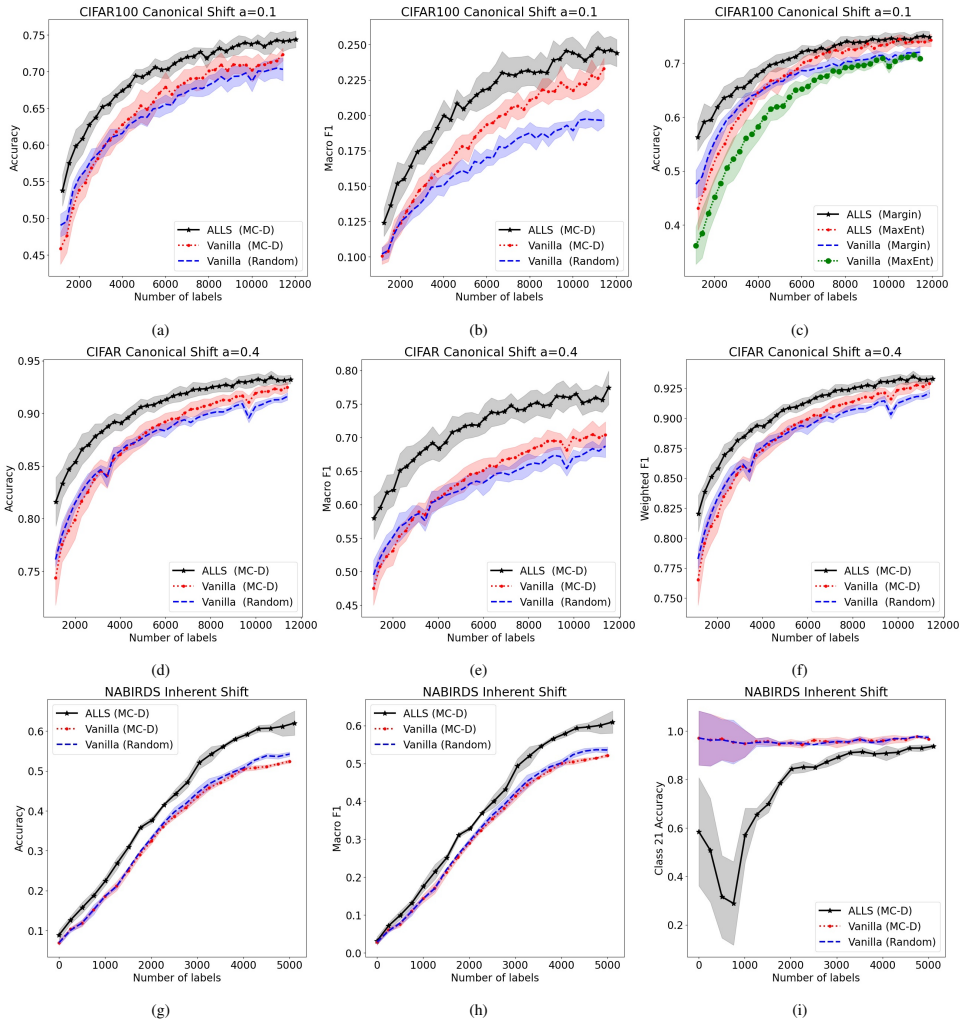


Figure 5: Average performance and 95% confidence intervals of 10 runs on CIFAR10, CIFAR100 and 4 runs on NABirds. Plots (a)-(h) demonstrate ALLS consistently improves both accuracy, macro F1, and weighted F1 scores on the CIFAR10, CIFAR100 and NABirds datasets, for both synthetic and natural label shift settings. Plot (c) demonstrates these gains generalize to other uncertainty sampling algorithms. Plot (i) depicts the learning dynamics of ALLS and verifies a suppression of the over-represented class during learning.

CIFAR100, and NABirds. In the NABirds experiment, we apply ALLS to a naturally occurring class imbalance problem in the NABirds dataset. As noted by [17], the coarsest available set of 22 bird labels in NABirds features strong class imbalance with a dominant class constituting almost a majority of available training data. We adopt this imbalance and evaluate on a uniformly sampled test distribution; this is a *imbalanced source* problem. In the CIFAR10 and CIFAR100 experiments, we artificially induce *canonical label shift* settings by applying [1]'s *Dirichlet Shift* procedure individually to each of the source and target data splits. In all experiments, we observe a significant gain in performance from the application of ALLS, both in accuracy and macro F1 scores. In the synthetic shift experiments, ALLS reduces sample complexity by up to half an order of magnitude. In Figure 5(c), we demonstrate the gains introduced by ALLS are similarly realized on other uncertainty sampling algorithms. In Figure 5(i), we can observe the learning dynamics of ALLS by tracking the accuracy of the dominant class. Observe that the ALLS curve features a drop in accuracy followed by a recovery in performance. The initial period of decline in accuracy can be attributed to a period of improvement in label shift estimation. As label shift estimation improves, dominant class accuracy is suppressed by shrinking importance weights. Accuracy then recovers as the label shift is diluted and the importance weight on the dominant class increases.

5.2 Ablation Studies

Source vs Target Shift To investigate the trade-off between subsampling and importance weighting suggested by theory, we induce synthetic *imbalanced source* and *imbalanced target* scenarios on CIFAR100 in an ablation study depicted in Figure 6. To compare the strengths of these strategies, we compare ALLS against the use of subsampling or importance weighting alone. We again use [1]'s *Dirichlet Shift* procedure to induce synthetic shifts. While Figure 6(a) demonstrates that subsampling accounts for ALLS's performance gains under *imbalanced source*, Figure 6(b) demonstrates that importance weighting leads to gains under *imbalanced target*. This corroborates our previous analysis. Although the strengths of the importance weighting and subsampling appear complementary, figure 6(c) demonstrates that, when properly balanced under ALLS, the joint usage of importance weighting and subsampling outperforms the individual use of either technique.

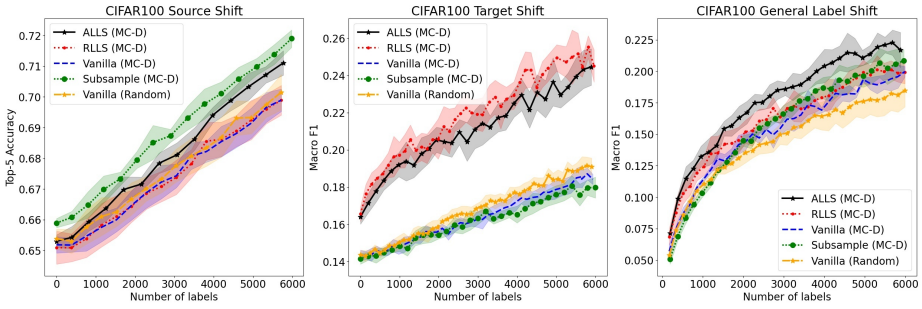


Figure 6: Average performance and 95% confidence intervals on 10 runs in various *general label shift* settings. (a) Top-5 accuracy under *imbalanced source*, subsampling outperforms importance weighting; (b) Macro F1 under *imbalanced target*, importance weighting outperforms subsampling; (c) Macro F1 under generic label shift. Importance weighting and subsampling feature complimentary strengths and yield additional gains when unified in ALLS.

Label Shift Estimation Heuristics We analyze the effects of these techniques on performance and learning behavior in an ablation study on the CIFAR100 dataset, as depicted in Figure 7. Figure 7 demonstrates that a combination of these techniques provides performance gains in both accuracy and macro F1 score. In addition, the study verifies the high variance associated with importance weighting and the cumulative gains afforded by iteratively reweighting.

6 Related Work

Active Learning Active learning has been investigated extensively from theoretical and practical perspectives. Disagreement-based active learning and its variants [18–24] enjoy rigorous learning

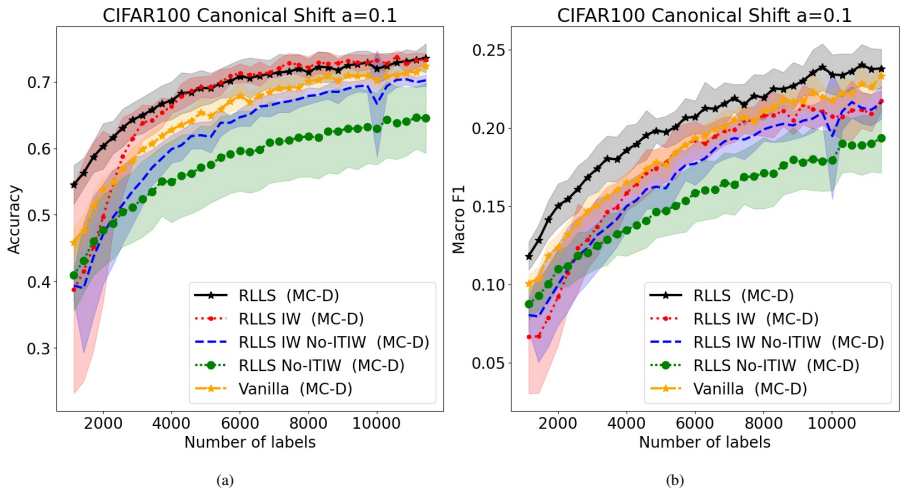


Figure 7: Average performance and 95% confidence intervals on 10 runs of experiments on CIFAR100 in a *canonical label shift* setting. (a) Accuracy using MC-D; (b) Macro F1 using MC-D. Posterior regularization lowers variance (versus importance weighting) and especially improves early-stage performance. Iterative reweighting similarly introduces consistent performance gains. Combining the two provides additional gains in macro F1 scores.

guarantees and focus on the stream-based active learning setting. On the other hand, active learning has been widely studied in natural language processing [25], computer vision [26], and even robotics [27]. Instead of the theoretically derived strategies, the most prevalent label solicitation method is uncertainty sampling. In this paper, we provide both theoretical analysis and practical algorithmic framework that is easy to implement.

Distribution Shift General domain adaptation theory [28–31] looks at joint distribution shift. Covariate shift [32–34] and label shift [35] are two special cases of distribution shift when more specific assumptions are made regarding which distribution is variant and which is invariant in the joint data distribution. Importance weighting methods under covariate shift and label shift are asymptotically unbiased. Density ratio estimation [36–38] on the input distribution is challenging due to the high-dimension nature of features in many application. In contrast, label shift assumptions make the weight estimation more tractable using black-box predictors [1, 39]. In our work, we utilize RLLS [9] for label shift correction and take advantage of its theoretical properties to help prove our guarantees in active learning.

Active Learning under Distribution Shift Active domain adaptation [40–43] has been studied under the general joint distribution shift assumption. Even though their problem settings are similar to ours, these methods focus on the practical side. Active learning from covariate-shifted warm-start set [7] has guaranteed label complexity but requires known importance weights beforehand. Instead of the covariate shift, we focus on the label shift case. On the other hand, active learning for imbalanced data [10, 44] proposes useful sampling heuristics, like diverse sampling or class-balanced sampling, without theoretical justification. Class-balance sampling has also been investigated extensively in self-training for unsupervised domain adaptation [45]. We focus on the general label shift and leverage subsampling to construct a *medial distribution*, which help achieve empirical and theoretical trade-off between importance weighting and subsampling.

7 Conclusion

In this paper, we propose ALLS, a novel framework for active learning under label shift. Our framework utilizes both importance weighting and subsampling to correct for label shift when active learning. We derive a rigorously guaranteed online active learning algorithm and prove its label complexity and the generalization bound. Our analysis shed light on the trade-off between importance

weighting and subsampling under label shift. We show the effectiveness of our method on both real-world inherent-shift data and large-scale benchmark synthetic-shift data.

Data distribution bias in training and testing has a huge impact on model behaviors in machine learning [3]. Our work generally tackles this problem by incorporating active data collection to correct distribution shift. In many applications that require manually labeling of data, like natural language processing and computer vision, an extension of the techniques we explore in ALLS may help mitigate bias in the data collection process. We also believe this approach can be extended to new settings, include cost-sensitive, multi-domain, and Neyman-Pearson settings.

Acknowledgement

Anqi Liu is supported by PIMCO Postdoctoral Fellowship at Caltech. Prof. Anandkumar is supported by Bren endowed Chair, faculty awards from Microsoft, Google, and Adobe, and LwLL grants.

References

[1] Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and Correcting for Label Shift with Black Box Predictors. February 2018.

[2] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, January 2002.

[3] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 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, 2020.

[4] Yee Seng Chan and Hwee Tou Ng. 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, June 2007. Association for Computational Linguistics.

[5] Piyush Rai, Avishek Saha, Hal Daumé, and Suresh Venkatasubramanian. 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, June 2010. Association for Computational Linguistics.

[6] Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L. DuVall. Active Supervised Domain Adaptation. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 97–112, Berlin, Heidelberg, 2011. Springer.

[7] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active Learning with Logged Data. *arXiv:1802.09069 [cs, stat]*, June 2018. arXiv: 1802.09069.

[8] Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *30th International Conference on Machine Learning, ICML 2013*, pages 1290–1298. International Machine Learning Society (IMLS), 2013.

[9] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized Learning for Domain Adaptation under Label Shifts. *arXiv:1903.09734 [cs, stat]*, March 2019. arXiv: 1903.09734.

[10] U. Aggarwal, A. Popescu, and C. Hudelot. Active learning for imbalanced datasets. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1417–1426, 2020.

[11] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009.

[12] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. 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, 2015.

[13] Alina Beygelzimer, Daniel Hsu, John Langford, and Tong Zhang. Agnostic Active Learning Without Constraints. *arXiv:1006.2588 [cs]*, June 2010. arXiv: 1006.2588.

[14] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, January 2009.

[15] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv:1506.02142 [cs, stat]*, October 2016. arXiv: 1506.02142.

[17] Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed Elgammal. Link the head to the "beak": Zero shot learning from noisy text description at part precision, 2017.

[18] Steve Hanneke. 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, June 2007. Association for Computing Machinery.

[19] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

[20] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56, 2009.

[21] Steve Hanneke. Activized Learning: Transforming Passive to Active with Improved Label Complexity. *arXiv:1108.1766 [cs, math, stat]*, August 2011. arXiv: 1108.1766.

[22] Alina Beygelzimer, Daniel J Hsu, John Langford, and Tong Zhang. Agnostic active learning without constraints. In *Advances in neural information processing systems*, pages 199–207, 2010.

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

[24] Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. *Journal of Machine Learning Research*, 20(65):1–50, 2019.

[25] Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*, 2017.

[26] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015.

[27] Sanjiban Choudhury and Siddhartha S Srinivasa. A bayesian active learning approach to adaptive motion planning. In *Robotics Research*, pages 33–40. Springer, 2020.

[28] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of Representations for Domain Adaptation. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press, 2007.

[29] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

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

[31] Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

[32] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[33] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[34] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

[35] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

[36] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[37] Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

[38] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in neural information processing systems*, pages 594–602, 2011.

[39] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C Lipton. A unified view of label shift estimation. *arXiv preprint arXiv:2003.07554*, 2020.

[40] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32. Association for Computational Linguistics, 2010.

[41] Giona Matasci, Devis Tuia, and Mikhail Kanevski. 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, 2012.

[42] Cheng Deng, Xianglong Liu, Chao Li, and Dacheng Tao. Active multi-kernel domain adaptation for hyperspectral image classification. *Pattern Recognition*, 77:306–315, 2018.

[43] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 739–748, 2020.

[44] Christopher H Lin, Mausam Mausam, and Daniel S Weld. Active learning with unbalanced classes and example-generation queries. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[45] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training. *arXiv preprint arXiv:1810.07911*, 2018.

[46] Tong Zhang. Data Dependent Concentration Bounds for Sequential Prediction Algorithms. pages 173–187, June 2005.

[47] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance Weighted Active Learning. *arXiv:0812.4952 [cs]*, May 2009. arXiv: 0812.4952.

A Theorem 1 and Theorem 2 Proofs

A.1 Deviation Bound

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

Theorem 3. *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 $1 - \delta$:*

$$
\begin{aligned}
&|err(h, Z_{1:n}) - err(h^*, Z_{1:n}) - err(h) + err(h^*)| \\
&\leq \mathcal{O}\left(d_\infty(P_{test}, P_{src}) \frac{\log(n|H|/\delta)}{n} + \sqrt{d_2(P_{test}, P_{src}) \frac{\log(n|H|/\delta)}{n}} + \sqrt{\frac{\log(n|H|/\delta)}{nP_{\min,n}(h)}} + \frac{\log(n|H|/\delta)}{nP_{\min,n}(h)} \right. \\
&+ \left(1 + err_W(h^*_{online}) + \frac{\log(\lambda n/\delta)}{\lambda n} + \sqrt{\frac{err_W(h^*_{online}) \log(\lambda n/\delta)}{\lambda n}} + \|\theta_{src \to med}\| \sqrt{\frac{\log(n|H|/\delta)}{nP_{\min,n}(h)}} \right) \\
&\cdot \left(\frac{\left\|\tilde{\theta}\right\|_2 + 1}{\sigma_{\min}} \right) \sqrt{\frac{d_\infty(P_{test}, P_{med}) \log(nk/\delta)}{\lambda n - \sqrt{n d_\infty(P_{test}, P_{med}) \log(n/\delta)\lambda}}} \\
&\left. + (\left\|\tilde{\theta}\right\|_2 + 1) \left(err_W(h^*_{online}) + \frac{\log(\lambda n/\delta)}{\lambda n} + \sqrt{\frac{err_W(h^*_{online}) \log(\lambda n/\delta)}{\lambda n}} \right) + \|\theta\|_2 \sqrt{\frac{\log(n|H|/\delta)}{nP_{\min,n}(h)}} \right)
\end{aligned}
$$

The corresponding bound for the case where only importance weighting is used can be recovered by setting $P_{\text{med}} = P_{\text{src}}$. Since we are ignoring warm starts in this section, we use $P_{\text{src}} := P_{\text{ulb}}$. 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 3. We proceed by expressing theorem 3 in a more general form with a bounded function $f : X \times Y \to [-1, 1]$ which will eventually represent $\text{err}(h) - \text{err}(h^*)$.

We borrow notation for the terms W, Q from [13], where Q_i is an indicator random variable indicating whether the ith datapoint is labeled and $W := Q_i \tilde{Q}_i \tilde{r}_i f(x_i, y_i)$. Our notation convention for the accented letters is denoting the estimated (from data) version with *hat* and denoting the medial distribution version with *tilde*. For example, $\tilde{Q}_i$ denotes whether the ith data sample in the medial data set is labeled or not. We introduce the accented variants $\tilde{W} := Q_i \hat{\tilde{Q}}_i \tilde{r}_i f(x_i, y_i)$ and $\hat{\tilde{W}} := Q_i \hat{\tilde{Q}}_i \hat{\tilde{r}}_i f(x_i, y_i)$. We also borrow [9]'s label shift notation and define k as the size of the output space (finite) and denote estimated importance weights with hats $\hat{\cdot}$. We introduce $\tilde{r} := r_{\text{med} \to \text{tar}}$ apply these same semantics to accents on $\theta := r - 1$. Finally, we follow [30] 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 P and P'.

We now arrive at a general form for the left-hand-side of theorem 3. To prove the theorem, we seek to bound with high probability,

$$
\Delta := \frac{1}{n} \left(\sum_{i=1}^{n} \hat{\tilde{W}}_i \right) - \mathbb{E}[rf(X, Y)] \tag{8}
$$

We eventually individually bound the following terms,

$$\begin{aligned}
\Delta_1 &:= \mathbb{E}[rf(X,Y)] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i[W_i] \\
\Delta_2 &:= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i[W_i] - \mathbb{E}_i[\hat{W}_i] \\
\Delta_3 &:= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i[\hat{W}_i] - \mathbb{E}_i[\hat{\tilde{W}}_i] \\
\Delta_4 &:= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i[\hat{\tilde{W}}_i] - \hat{\tilde{W}}_i
\end{aligned} \tag{9}$$

where Δ_1 corresponds to the variance associated with inherent stochasticity in datapoints. Δ_2 corresponds to label inference error during subsampling. Δ_3 corresponds to label shift estimation errors. Δ_4 corresponds to the stochasticity of the IWAL-CAL sampling policy. Using repeated applications of triangle inequalities, a bound on Δ is given by:

$$|\Delta| \leq |\Delta_1| + |\Delta_2| + |\Delta_3| + |\Delta_4| \tag{10}$$

A.2 Bounding Active Learning Stochasticity

We bound Δ_4 using a Martingale technique from [46] also adopted by [13]. We take Lemmas 1, 2 from [46] as given. We now proceed in a fashion similar to the proof of Theorem 1 from [13]. We begin with a generalization of Lemma 6 in [13].

Lemma 1. *If* $0 < \lambda < 3\frac{P_i}{\hat{\tilde{r}}_i}$, *then*

$$\log \mathbb{E}_i[\exp(\lambda(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i]))] \leq \frac{\hat{r}_i\hat{\tilde{r}}_i\lambda^2}{2P_i(1 - \frac{\hat{\tilde{r}}_i\lambda}{3P_i})} \tag{11}$$

If $\mathbb{E}_i[\hat{\tilde{W}}_i] = 0$ *then*

$$\log \mathbb{E}_i[\exp(\lambda(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i]))] = 0 \tag{12}$$

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

$$|\hat{\tilde{W}}_i| \leq \left|\frac{Q_i\hat{\tilde{Q}}_i\hat{\tilde{r}}_i}{P_i}\right| \leq \frac{\hat{\tilde{r}}_i}{P_i} \tag{13}$$

To bound variance, note that $\hat{r}_i = \hat{\tilde{r}}_i\mathbb{E}_i[\hat{\tilde{Q}}_i]$ by definition. In other words, when combined, subsampling and importance weighting should fully correct for any (perception of) underlying label shift. Therefore

$$\mathbb{E}_i[(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i])^2] \leq \frac{\hat{r}_i\hat{\tilde{r}}_i}{P_i}f(x_i,y_i)^2 - 2\hat{r}_i^2 f(x_i,y_i)^2 + \hat{r}_i^2 f(x_i,y_i)^2 \leq \frac{\hat{r}_i\hat{\tilde{r}}_i}{P_i} \tag{14}$$

Following [13], 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,

$$\begin{aligned}
\mathbb{E}_i[\exp(\lambda(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i]))] &= \mathbb{E}_i[1 + \lambda(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i]) + \lambda^2(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i])^2 g(\lambda(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i]))] \\
&= 1 + \lambda^2\mathbb{E}_i[(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i])^2 g(\lambda(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i]))] \\
&\leq 1 + \lambda^2\mathbb{E}_i[(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i])^2 g(\lambda\hat{\tilde{r}}_i/P_i)] \\
&= 1 + \lambda^2\mathbb{E}_i[(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i])^2]g(\lambda\hat{\tilde{r}}_i/P_i) \\
&\leq 1 + \frac{\lambda^2\hat{r}_i\hat{\tilde{r}}_i}{P_i}g(\frac{\hat{\tilde{r}}_i\lambda}{P_i})
\end{aligned} \tag{15}$$

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 $\exp(x) - x - 1 \leq x^2/(2(1-x/3))$ for $0 \leq x < 3$ and $\log(1+x) \leq x$. The second claim follows from definition of $\hat{\tilde{W}}_i$ and the fact that $\mathbb{E}_i[\hat{\tilde{W}}_i] = \hat{r} f(X_i, Y_i)$. □

The following lemma is an analogue of Lemma 7 in [13].

Lemma 2. *Pick any $t \geq 0, p_{\min} > 0$ and let E be the joint event*

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\tilde{W}}_i - \sum_{i=1}^{n} \mathbb{E}_i[\hat{\tilde{W}}_i] \geq (1+M)\sqrt{\frac{t}{2np_{\min}}} + \frac{t}{3np_{\min}} \quad \textit{and} \quad \min\{\frac{P_i}{\hat{\tilde{r}}_i} : 1 \leq i \leq n \wedge \mathbb{E}_i[W_i] \neq 0\} \geq p_{\min} \tag{16}$$

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

Proof. We follow [13] and let

$$\lambda := 3p_{\min} \frac{\sqrt{\frac{2t}{9np_{\min}}}}{1 + \sqrt{\frac{2t}{9np_{\min}}}} \tag{17}$$

Note that $0 < \lambda < 3p_{\min}$. By Lemma 1, we know that if $\min\{\frac{P_i}{\hat{\tilde{r}}_i} : 1 \leq i \leq n \wedge \mathbb{E}_i[\hat{\tilde{W}}_i] \neq 0\} \geq p_{\min}$ then

$$\frac{1}{n\lambda}\sum_{i=1}^{n} \log \mathbb{E}_i[\exp(\lambda(W_i - \mathbb{E}_i[W_i]))] \leq \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{r}_i \hat{\tilde{r}}_i \lambda}{2P_i(1 - \frac{\hat{\tilde{r}}_i \lambda}{3P_i})} \leq M\sqrt{\frac{t}{2np_{\min}}} \tag{18}$$

and

$$\frac{t}{n\lambda} = \sqrt{\frac{t}{2np_{\min}}} + \frac{t}{3np_{\min}} \tag{19}$$

Let E' be the event that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i]) - \frac{1}{n\lambda}\sum_{i=1}^{n} \log \mathbb{E}_i[\exp(\lambda(\hat{\tilde{W}} - \mathbb{E}_i[\hat{\tilde{W}}]))] \geq \frac{t}{n\lambda} \tag{20}$$

and let E'' be the event $\min\{\frac{P_i}{\hat{\tilde{r}}_i} : 1 \leq i \leq n \wedge \mathbb{E}_i[\hat{\tilde{W}}_i] \neq 0\} \geq p_{\min}$. Together, the above two equations imply $E \subseteq E' \bigcap E''$. By [46]'s lemmas 1 and 2, $\Pr(E) \leq \Pr(E' \bigcap E'') \leq Pr(E') \leq e^{-t}$. □

The following is an immediate consequence of the previous lemma.

Lemma 3. *Pick any $t \geq 0$ and $n \geq 1$. Assume $1 \leq \frac{\hat{\tilde{r}}_i}{P_i} \leq r_{\max}$ for all $1 \leq i \leq n$, and let $R_n := \max\{\frac{\hat{\tilde{r}}_i}{P_i} : 1 \leq i \leq n \wedge \mathbb{E}_i[\hat{\tilde{W}}] \neq 0\} \bigcup \{1\}$. We have*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} \hat{\tilde{W}}_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_i[\hat{\tilde{W}}_i]\right| \geq (1+M)\sqrt{\frac{R_n t}{2n}} + \frac{R_n t}{3n}\right) \leq 2(2 + \log_2 r_{\max})e^{-t/2} \tag{21}$$

Proof. This proof follows identically to [13]'s lemma 8. □

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

Lemma 4. *With probability at least $1-\delta$, the following holds over all $n \geq 1$ and $h \in H$:*

$$|\Delta_4| \leq (2 + \left\|\hat{\theta}\right\|_2)\sqrt{\frac{\varepsilon_n}{P_{\min,n}(h)}} + \frac{\varepsilon_n}{P_{\min,n}(h)} \tag{22}$$

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

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 3 and a union bound over hypotheses and datapoints. □

A.3 Bounding Subsampling Error

We now bound Δ_3, the error associated with the inference in subsampling. It holds that

$$|\Delta_3| = \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_i[\tilde{Q}_i] - \mathbb{E}_i[\hat{\tilde{Q}}_i])\hat{\tilde{r}}_i f(x_i, y_i)\right| \leq \left|\frac{1}{n}\sum_{j=1}^{k}\hat{\tilde{r}}_i\tilde{\ell}(j)\right| \tag{23}$$

where we define $\tilde{\ell} \in \mathcal{R}^k$ such that $\tilde{\ell}(j) = \sum_{i=1}^{n}\mathbb{1}_{y(i)=j}(\mathbb{E}_i[\tilde{Q}_i] - \mathbb{E}_i[\hat{\tilde{Q}}_i])f(x_i, y_i)$. Recall this inequality follows similarly to the proof in the previous lemma and simply concerns a change in perspective: summing over labels rather than datapoints. We can then apply Cauchy Schwarz inequality,

$$|\Delta_3| \leq \frac{1}{n}\left\|\tilde{\ell}\right\|_2\left\|\hat{\tilde{r}}\right\|_2 \tag{24}$$

Intuitively, the quantity $\left\|\tilde{\ell}\right\|_2$ represents an intuitive measure of the error of the model used for subsampling. For instance, a classifier with zero error drives $\tilde{\ell}$ to 0. Similarly, a trivial subsampling strategy where all labels are assigned the same subsampling probability drives $\tilde{\ell}$ to 0. Note that $\left\|\tilde{\ell}\right\|_2$ is simply the regret of an online agnostic learner in a standard supervised setting over an L1 (absolute) error loss. We can thus plug in the standard bound of $\mathcal{O}(\text{err}_W(h^*_{\text{online}}) + \frac{\log(n/\delta)}{n} + \sqrt{\frac{\text{err}_W(h^*_{\text{online}})\log(n/\delta)}{n}})$ to hold with probability at least $1-\delta$. Here, err_W denotes the absolute error and $\text{err}_W(h^*_{\text{online}})$ denotes the best achievable loss of a subsampling weight estimator on the source distribution. Note that $\text{err}_W(h^*_{\text{online}}) = 0$ if the medial distribution is simply the source distribution as the subsampling learning problem is trivial.

Lemma 5. *With probability at least* $1-\delta$*,*

$$|\Delta_3| \leq \left\|\hat{\tilde{r}}\right\|_2 \mathcal{O}\left(err_W(h^*_{online}) + \frac{\log(\lambda n/\delta)}{\lambda n} + \sqrt{\frac{err_W(h^*_{online})\log(\lambda n/\delta)}{\lambda n}}\right) \tag{25}$$

Proof. Follows immediately by noting that $\left\|\tilde{\ell}\right\|_1 \geq \left\|\tilde{\ell}\right\|_2$ and recalling that the subsampling model is only trained on the holdout buffer. □

The subsampling problem may be seperable with $\text{err}_W(h^*_{\text{online}}) = 0$, even in an agnostic learning setting where the original learning problem is non-seperable. This is because labels may share the same subsampling probability. In practice, this is often a consequence of label shift estimation via RLLS, where L2 regularization drives uncertain labels to similar label shift weights. When the subsampling problem is separable, $\text{err}_W(h^*_{\text{online}}) = 0$.

A.4 Bounding Label Shift Error

We now bound Δ_2: the label shift error. If the medial distribution is known, label shift estimation is straight-forward—simply estimate the label shift from the source to the target. We can then compensate for the label shift correction already performed through subsampling by adjusting the importance weight according to the medial distribution. However, as we do not assume knowledge of the source label distribution, the user's knowledge of the subsampling distribution does not afford knowledge of the medial distribution.

Hence, we require the use of a special buffer as prescribed in Algorithm 1 to enable correct usage of RLLS [9] label shift estimation. Specifically, we sample already-labeled source datapoints from a holdout set independent of the data used for the rest of the learning procedure, with the notable exception of the subsampling model. The following lemma bounds the number of samples we can draw from the buffer, and hence the effective size of our RLLS holdout set.

Lemma 6. *With probability at least $1-\delta$, the number of source samples is bounded below by*

$$n_p \geq \frac{\lambda n}{d_\infty(P_{med}||P_{src})} - \sqrt{-2\frac{\lambda n}{d_\infty(P_{med}||P_{src})}\log(\delta)} \tag{26}$$

Proof. We seek to bound the number of datapoints we sample as a holdout set, which is a random variable in itself. We directly apply Chernoff's inequality. To use Chernoff's, we first seek a lower bound on the expectation of n_p, which we denote by μ. By linearity of expectation,

$$\mu := \mathbb{E}[n_p] = \mathbb{E}[\frac{\sum_{i=1}^{\lambda n} P_{\text{ss}}(y_i)}{\max_i P_{\text{ss}}(y_i)}] \geq \frac{\sum_{i=1}^{\lambda n}\mathbb{E}[P_{\text{ss}}(y_i)]}{d_\infty(P_{\text{med}}||P_{\text{src}})} = \frac{\lambda n}{d_\infty(P_{\text{med}}||P_{\text{src}})} \tag{27}$$

Hence, with probability at most $\exp(-\mu\delta^2/2)$, we have that

$$n_p \leq (1-\delta)\mu \tag{28}$$

and with probability at most δ that

$$\begin{aligned} n_p &\leq \mu(1-\sqrt{-2\log(\delta)/\mu}) \\ &= \frac{\lambda n}{d_\infty(P_{\text{med}}||P_{\text{src}})} - \sqrt{-2\frac{\lambda n}{d_\infty(P_{\text{med}}||P_{\text{src}})}\log(\delta)} \\ &= \frac{1}{d_\infty(P_{\text{med}}||P_{\text{src}})}\left(\lambda n - \sqrt{-2\lambda n d_\infty(P_{\text{med}}||P_{\text{src}})\log(\delta)}\right) \end{aligned} \tag{29}$$

□

With a lower bound on the size of the RLLS holdout set, we can now bound label shift estimation error directly.

Lemma 7. *With probability $1-2\delta$, for all $n \geq 1$:*

$$|\Delta_2| \leq \frac{2}{\sigma_{\min}}\mathcal{O}\left(\left\|\tilde{\theta}\right\|_2\sqrt{\frac{d_\infty(P_{med}||P_{src})\log\left(\frac{nk}{\delta}\right)}{\lambda n - \sqrt{2\lambda n d_\infty(P_{med}||P_{src})\log\left(\frac{n}{\delta}\right)}}} + \sqrt{\frac{d_\infty(P_{med}||P_{src})\log\left(\frac{n}{\delta}\right)}{\lambda n - \sqrt{2\lambda n d_\infty(P_{med}||P_{src})\log\left(\frac{n}{\delta}\right)}}}\right) \tag{30}$$

Proof. We seek a bound on the label shift estimation error for importance weights which correct from the medial distribution to the target distribution. We apply Bernstein's inequality as demonstrated by RLLS Appendix B.6. The following holds as the simple re-indexing of a summation

$$|\Delta_2| = \left|\frac{1}{n}\sum_{i=1}^{n}(\tilde{r}_i - \hat{\tilde{r}}_i)f(x_i,y_i)\right| \leq \left|\frac{1}{n}\sum_{j=1}^{k}(\tilde{r}(j) - \hat{\tilde{r}}(j))\tilde{\ell}(j)\right| \tag{31}$$

where we define $\tilde{\ell} \in \mathcal{R}^k$ as $\tilde{\ell}(j) = \sum_{i=1}^{n}\mathbb{1}_{y(i)=j}f(x_i,y_i)$. We can then apply the Cauchy Schwarz inequality:

$$\left|\frac{1}{n}\sum_{j=1}^{k}(\tilde{r}(j) - \hat{\tilde{r}}(j))\tilde{\ell}(j)\right| \leq \frac{1}{n}\left\|\tilde{r}(j) - \hat{\tilde{r}}(j)\right\|_2\left\|\tilde{\ell}\right\|_2 \tag{32}$$

Since $f(x,y) \in [-1,1]$, we can bound $\left\|\tilde{\ell}\right\|_2$ by $2n$. Then, $|\Delta_2| \leq 2\left\|\tilde{\theta} - \hat{\tilde{\theta}}\right\|_2$. [9]'s (RLLS) lemma 1 then gives the following bound on $\left\|\tilde{\theta} - \hat{\tilde{\theta}}\right\|_2$ which holds with probability $1-\delta$:

$$\left\|\tilde{\theta} - \hat{\tilde{\theta}}\right\|_2 \leq \mathcal{O}\left(\frac{1}{\sigma_{\min}}(\|\theta\|_2\sqrt{\frac{\log(k/\delta)}{n_p}} + \sqrt{\frac{\log(1/\delta)}{n_p}})\right) \tag{33}$$

where n_p denote the number of datapoints used in the holdout dataset for RLLS. In our above application of lemma 1, we drop terms associated with RLLS regularization (i.e. we choose not to regularize) and assume free access to unlabeled target samples.

Similarly, with probability at least $1-\delta$:

$$|\Delta_2| \leq \mathcal{O}\left(\frac{2}{\sigma_{\min}}\left(\|\tilde{r}-1\|_2\sqrt{\frac{\log(k/\delta)}{n_p}}+\sqrt{\frac{\log(1/\delta)}{n_p}}\right)\right) \tag{34}$$

The bound then follows immediately by lemma 6 and a union bound over H and n. For sufficiently large label shift magnitude, the first term dominates and so we discard the second term in subsequent Big-O expressions, such as Theorem 1. □

A.5 Remaining Terms

We now bound the remaining term, Δ_1. This is a simple generalization bound of an importance weighted estimate of f.

Lemma 8. *For any $\delta > 0$, with probability at least $1-\delta$, then for all $n \geq 1$, $h \in H$:*

$$|\Delta_1| \leq \frac{2d_\infty(P_{test}, P_{src})\log(\frac{2n|H|}{\delta})}{3(n+m)} + \sqrt{\frac{2d_2(P_{test}, P_{src})\log(\frac{2n|H|}{\delta})}{n+m}} \tag{35}$$

Proof. This inequality is a direct application of Theorem 2 from [30]. □

We now combine our bounded terms to bound Δ. Recall that our bounds on Δ_3, Δ_4 still rely on the norm of the estimated label shift weights $\hat{\theta}$ or $\hat{\tilde{\theta}}$. We remove these terms using our known bounds on Δ_2 through a simple triangle inequality. Specifically, $\left\|\hat{\theta}\right\| \leq \|\theta\| + \left\|\hat{\theta}-\theta\right\|$ where we have already bounded the latter term in the proof of lemma 7. Theorem 3 follows by applying a triangle inequality over $\Delta_1, \Delta_2, \Delta_3, \Delta_4$.

To highlight trade-offs in distributions and for simplicity of reading, we assume the distribution shift is sufficiently large to dominate constant terms.

B Correctness and Sample Complexity Corollaries

As in [13], we define a C_0 such that ϵ_n is bounded as $\epsilon_n \leq C_0\log(n+1)/n$ where ϵ_n is defined as follows. With probability at least $1-\delta$, for all $n \geq 1$ and all $h \in H$:

$$|err(h, Z_{1:n}) - err(h^*, Z_{1:n}) - err(h) + err(h^*)| \leq (\left\|\tilde{\theta}\right\|_2 + 1)\text{err}_W(h^*_{\text{online}}) + \sqrt{\frac{\epsilon_n}{P_{\min,n}(h)}} + \frac{\epsilon_n}{P_{\min,n}(h)} \tag{36}$$

We simply base C_0 off the deviation bound from Theorem 3. For readibility, we aggressively drop terms from the asymptotic in Equation 3 to bound:

$$\begin{aligned} C_0 \in \mathcal{O}\Bigg(\log\left(\frac{|H|}{\delta}\right)\Big(d_\infty(P_{\text{test}}||P_{\text{src}}) + d_2(P_{\text{test}}||P_{\text{src}}) + 1 + \|\theta\|_2^2\Big) \\ + \frac{\log\left(\frac{k}{\delta}\right)}{\sigma^2_{\min}} d_\infty(P_{\text{test}}||P_{\text{med}})\left\|\tilde{\theta}\right\|_2^2 (\text{err}_W(h^*_{\text{online}}) + 1)\Bigg) \end{aligned} \tag{37}$$

In the literal algorithm specification, many terms in C_0 may be unknown—in practice, we simply guess a convenient value for C_0 that provides the desired amount of "mellowness" in sampling.

We now proceed almost identically to [13], noting that our ϵ_n is asymptotically equivalent to the ϵ_n in the original IWAL-CAL derivations of [13], differing only in the choice of constant C_0 and the presence of an additional bias term, $\text{err}(h^*_{\text{online}})$, in Equation 36. Hence, our proof of Theorem 1 follows immediately from Lemma 2 and Theorem 2 in [13]. Substituting our Theorem 1 into Theorem 3 from [13] similarly immediately yields 2 minus the λn labels necessary for accumulating a holdout set for RLLS and subsampling.

C Deviation Bound with a Warmstart Set

We now extend our deviation bound to a generalized setting where a warm start dataset is available to the learner. We substitute $P_{\text{src}} = \frac{nP_{\text{ulb}}+mP_{\text{warm}}}{n+m}$ and $P_{\text{lab}} = \frac{nP_{\text{med}}+mP_{\text{warm}}}{n+m}$. We redefine $\tilde{\theta}$ as $\theta_{\text{lab}\to\text{tar}}$. Our bound on Δ_4 from lemma 4 holds as is (there is no active learning associated with the warm start datapoints). Our bound on Δ_3 simply scales by a factor of $\frac{n}{n+m}$. The following lemmas are trivial extensions of their no-warm-start counterparts.

Lemma 9. *With probability* $1-2\delta$*, for all* $n \geq 1$*,* $h \in H$*:*

$$|\Delta_2| \leq \mathcal{O}\left(\frac{2}{\sigma_{\min}}\left(\left\|\tilde{\theta}\right\|_2 \sqrt{\frac{d_\infty(P_{med}||P_{src})\log\left(\frac{nk}{\delta}\right)}{\lambda(n+m) - \sqrt{2\lambda(n+m)d_\infty(P_{med}||P_{src})\log\left(\frac{n}{\delta}\right)}}} + \sqrt{\frac{d_\infty(P_{med}||P_{src})\log\left(\frac{n}{\delta}\right)}{\lambda(n+m) - \sqrt{2\lambda(n+m)d_\infty(P_{med}||P_{src})\log\left(\frac{n}{\delta}\right)}}}\right)\right) \tag{38}$$

Lemma 10. *For any* $\delta > 0$*, with probability at least* $1-\delta$*, then for all* $n \geq 1$*,* $h \in H$*:*

$$|\Delta_1| \leq \frac{2d_\infty(P_{test}, P_{src})\log(\frac{2n|H|}{\delta})}{3(n+m)} + \sqrt{\frac{2d_2(P_{test}, P_{src})\log(\frac{2n|H|}{\delta})}{n+m}} \tag{39}$$

Substituting these additional constants into Δ gives the analogous deviation bound under a, potentially shifted, warm start. This yields a modified version of the ϵ_n derived in the previous section:

$$C_0 \in \mathcal{O}\left(\log\left(\frac{|H|}{\delta}\right)\left(d_\infty(P_{\text{test}}||P_{\text{src}}) + d_2(P_{\text{test}}||P_{\text{src}}) + 1 + \|\theta\|_2^2\right) + \frac{n\log\left(\frac{k}{\delta}\right)}{(n+m)\sigma_{\min}^2} d_\infty(P_{\text{test}}||P_{\text{lab}})\left\|\tilde{\theta}\right\|_2^2 (\text{err}_W(h^*_{\text{online}}) + 1)\right) \tag{40}$$

The corresponding generalization and sample complexity bounds follow accordingly.

D Additional Experiment Settings

D.1 NABirds Regional Species Experiment

We conduct an additional experiment on the NABirds dataset using the grandchildren 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 5. 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)
ALLS (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

D.2 Online IWAL on CIFAR10

We evaluate a bootstrap approximation of IWAL [47] using a version space of 8 Resnet-18 models on the CIFAR dataset. We observe modest gains due to ALLS despite no observable gains with vanilla IWAL.

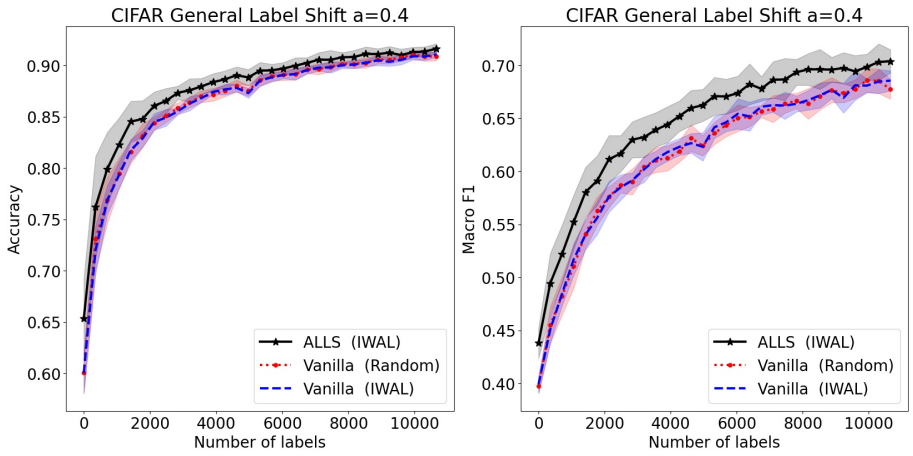


Figure 8: Average performance and 95% confidence intervals on 10 runs of experiments on CIFAR in a *general label shift* setting. (a) Accuracy using IWAL; (b) Macro F1 using IWAL. ALLS leads to modest gains even in difficult online learning settings.

D.3 Change in distribution

To further analyze the learning behavior of ALLS, we can analyze the label distribution of datapoints selected by the active learner. In Figure 9, MC-Dropout, Max-Margin and Max-Entropy strategies are evaluated on CIFAR100 under *canonical label shift*. By analyzing the uniformity bias and rate of convergence to the target distribution, we can observe that ALLS exhibits a unique sampling bias which cannot be explained away as simply a class-balancing bias. This indicates ALLS may be successful in recovering information from distorted uncertainty estimates.

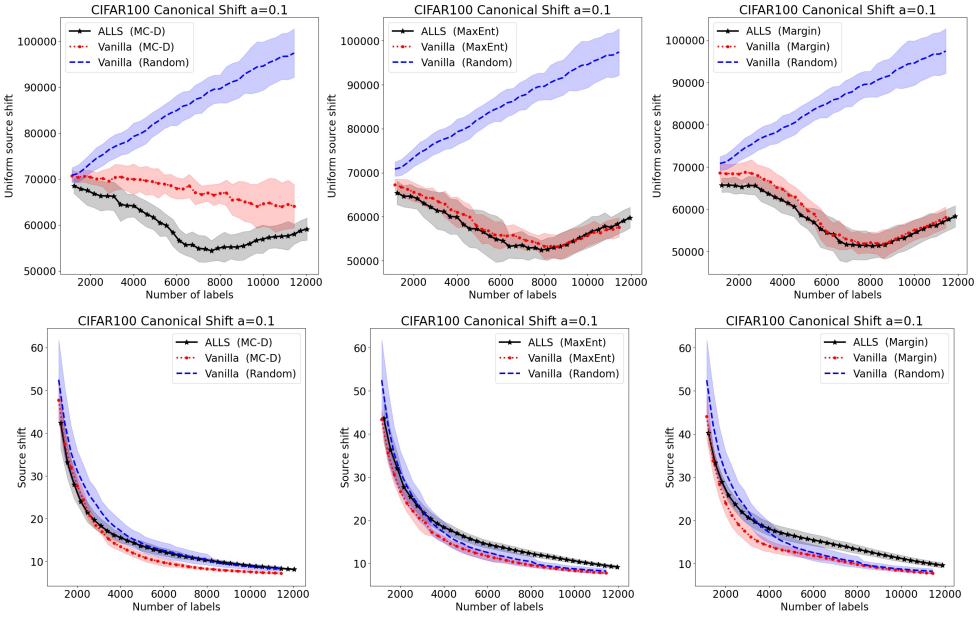


Figure 9: Average L2 distance between labeled class distribution and uniform/target distribution with 95% confidence intervals on 10 runs of experiments on CIFAR100 in the *canonical label shift* setting. ALLS converges to the target label distribution slower than vanilla active learning but with a similar uniform sampling bias. This suggests ALLS leverages a sampling bias different from that of vanilla active learning or naive class-balanced sampling.

D.4 Different label shift magnitudes

These experiments evaluate ALLS 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 10 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, ALLS does not perform significantly worse than vanilla active learning.

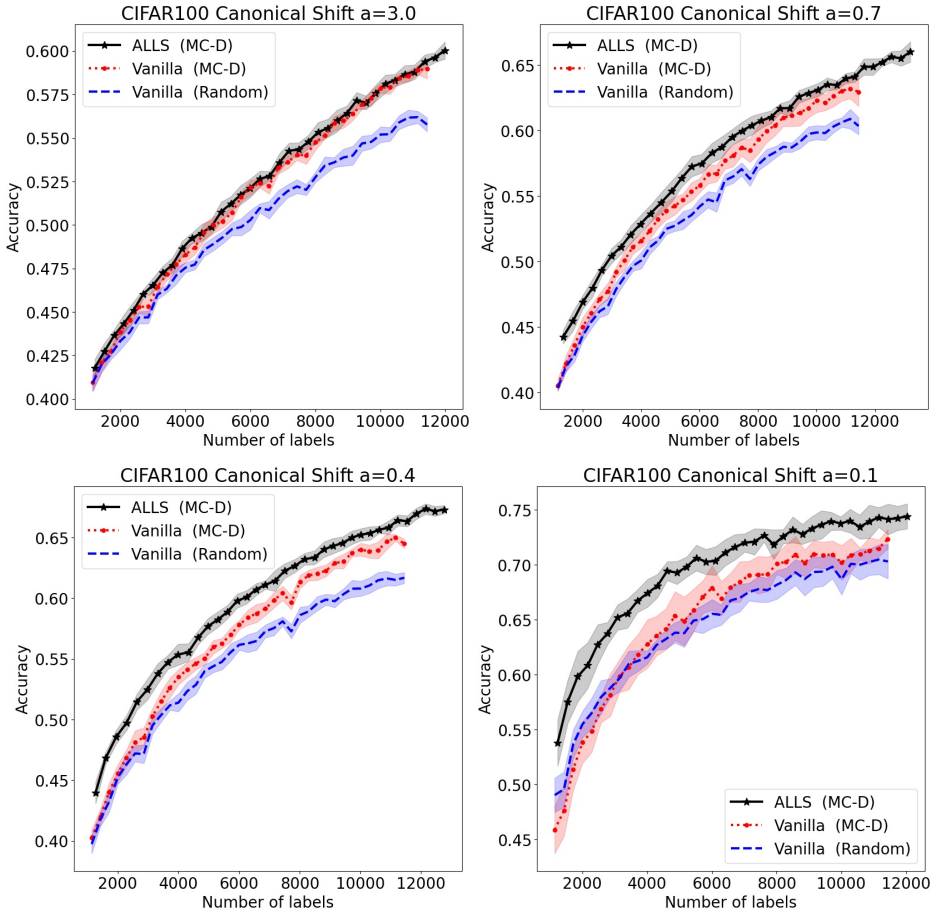


Figure 10: Average performance and 95% confidence intervals on 10 runs of experiments on CIFAR100 in the *general label shift* setting. In order of increasing label shift magnitude: (a), (b), (c), (d). ALLS performance gains scale by label shift magnitude.

E 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 $5\text{e}{-4}$. We also share the same batch size of 128 and RLLS [9] regularization constant of $2\text{e}{-6}$ 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-c)	CIFAR100	0.4	Dir	Dir	Yes	0.1
5(d-f)	CIFAR	0.3	Dir	Dir	Yes	0.7
5(g-i)	NABirds1	1.0	Inh	Inh	No	N/A
6(a)	CIFAR100	0.4	Dir	Uni	No	1.0
6(b)	CIFAR100	0.3	Uni	Dir	No	0.1
6(c)	CIFAR100	0.3	Dir	Dir	No	0.7
7	CIFAR100	0.4	Dir	Dir	Yes	0.1
T1(g-i)	NABirds1	1.0	N/A	Dir	No	0.1
8	CIFAR	0.4	Dir	Dir	No	0.4
9	CIFAR100	0.4	Dir	Dir	Yes	0.1
10(a)	CIFAR100	0.4	Dir	Dir	Yes	3.0
10(b)	CIFAR100	0.4	Dir	Dir	Yes	0.7
10(d)	CIFAR100	0.4	Dir	Dir	Yes	0.4
10(e)	CIFAR100	0.4	Dir	Dir	Yes	0.1

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

The complete source code for replicating and expanding our experiment base is released anonymously at `https://anonymous.4open.science/r/1133eed9-b6c0-4e64-82f1-ab48e5c03109/#`.