Active Learning under Label Shift

Eric Zhao ELZHAO@CALTECH.EDU
Anqi Liu ANQILIU@CALTECH.EDU
Animashree Anandkumar ANIMA@CALTECH.EDU
Yisong Yue YYUE@CALTECH.EDU
Department of Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91126, USA

Editor:

Abstract

Distribution shift poses a challenge for active data collection in the real world. We address the problem of active learning under *label shift* assumptions 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 analyze the 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. We further highlight the interplay between the components of ALLS with a series of ablation studies.

Keywords: Active Learning, Label Shift, Importance Weighting, Subsampling

1. Introduction

Distribution shift poses a significant challenge for active learning algorithms. 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. 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 (Lipton et al., 2018; Saerens et al., 2002b). 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 (Yang et al., 2020).

While previous works have offered a formal treatment of supervised learning under label shift (Lipton et al., 2018; Azizzadenesheli et al., 2019), active learning under label shift remains an open problem. To this end, we present a novel framework for Active Learning under Label Shift (ALLS). Our framework, ALLS, corrects for label shift with the use of importance weighting and a generalization of class-balanced sampling. We derive generalization and label complexity bounds for ALLS—the first theoretical results for this active learning setting—and demonstrate that algorithms instantiated under our framework improve performance across a variety of label shift settings.

2. Background

Learning under Label Shift The problem of distribution shift in supervised learning, *domain adaptation*, is typically analyzed under covariate shift assumptions where observational distributions shift ($P_{\text{trg}}(x) \neq P_{\text{src}}(x)$) but output conditionals do not ($P_{\text{trg}}(y|x) = P_{\text{src}}(y|x)$) (Shimodaira, 2000; Gretton et al., 2009; Sugiyama et al., 2007). Label shift—historically underrepresented in domain adaptation literature (Lipton et al., 2018)—assumes that label likelihoods shift ($P_{\text{trg}}(y) \neq P_{\text{src}}(y)$) while conditional feature distributions do not ($P_{\text{trg}}(x|y) = P_{\text{src}}(x|y)$) (Schölkopf et al., 2012). Importance weighting by class likelihood ratio $\frac{P_{\text{trg}}(y)}{P_{\text{src}}(y)}$ enables consistent label shift learning by producing asymptotically unbiased estimators by weighting datapoints. These importance weights can be found with a blackbox hypothesis h, a finite sample estimate of confusion matrix C_h on P_{src}, and $\hat{q}$ the label prediction proportions on P_{trg} (Lipton et al., 2018; Garg et al., 2020). We select RLLS (Azizzadenesheli et al., 2019) for label shift estimation, solving importance weights r as $\text{argmin}_r||C_h^{-1}r - \hat{q} - \mathbf{1} + 1|| + \lambda||r - 1||$, regularized by λ (Azizzadenesheli et al., 2019).

Active Learning under Distribution Shift While active learning under distribution shift (Rai et al., 2010; Matasci et al., 2012; Deng et al., 2018; Su et al., 2020) have been approached without specific covariate/label-shift assumptions, said results forgo formal treatment in favor of practical heuristics. In contrast, active domain adaptation leverages covariate shift (Yan et al., 2018) for guaranteed label complexity but assumes importance weights known a-priori. The only prior active learning literature related to label shift cover useful heuristics for active learning for imbalanced data (Aggarwal et al., 2020), a specific case of label shift, but without theoretical foundations or the formal framework of label shift. We build on one such method, subsampling, which "filters" the datapoints visible to an active learner according to their (estimated) labels so as to influence sampling likelihoods. Canonical active learning under distribution shift settings assume a *Warm-Start Shift* setting where label shift lies between a warm start dataset (D_{warm}) and the target (D_{tar}) distribution, while active learning occurs directly in the target domain. We consider the more general—and difficult— *Training Shift* setting, a important but often overlooked scenario (Huang and Chen, 2016) where the unlabeled pool/stream (D_{src}) and target distributions may differ.

3. The ALLS Framework

Algorithm Description A key design choice for a framework addressing label shift is the technique through which to correct for label shift. ALLS jointly employs importance weighting and subsampling, with a balance mediated through the choice of a *medial distribution*. We frame the joint strategy of ALLS as first inducing a *medial* distribution by subsampling from the source distribution and then applying importance weights to close the label shift between the medial and target distributions. Hence, the closer the medial distribution is to the target distribution, the less importance weighting is used.

Our proposed framework, as depicted in Figure 1, iteratively accumulates an independent holdout dataset O_t for training a classifier ϕ on D_{src} and estimating label shift weights r. Then, in the primary active learning loop, ALLS subsamples according to ϕ and then samples according to an active learning policy, which, in the pool-based setting, depends on the uncertainty estimates weighted according to the label shift weights r. We detail a general version of our framework in Algorithm 1, where n denotes the active learning budget, m the warmstart budget, λn the eventual size of the holdout set, and π some sampling probability rule (online) or uncertainty quantifier (pool).

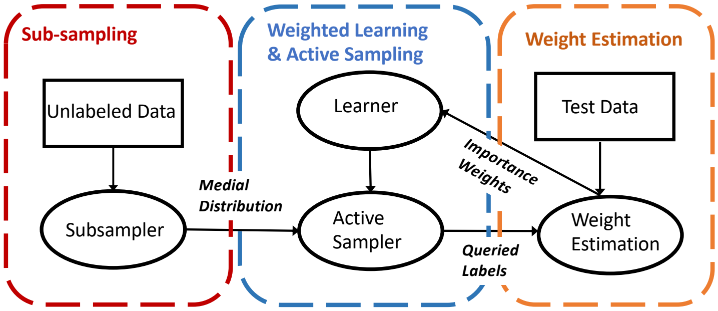


Figure 1: Illustration of ALLS Framework depicting active learning process.

Algorithm 1 Active Learning under Label Shift (ALLS)

```
**Input:** warmstart set $D_{\text{warm}}$ , unlabeled pool/stream $D_{\text{src}}$, subsampling distribution $D_{\text{ss}}$, target data $D_{\text{tar}}$, blackbox predictor $h_0$, regularizer for RLLS $\lambda$, initial holdout set $O_0$, maximum timestep $T$, label oracle $C$, active sampling criterion $\pi$
**Initialize:** $r_0 \leftarrow \text{RLLS}(O_0, D_{\text{tar}}, h_0)$, $S_0 \leftarrow \{(x_i, y_i, r_0(y_i))\}$ for $(x_i, y_i) \in D_{\text{warm}}$
**For** $x_t, y_t \in D_{\text{warm}} \setminus O_0$ append $S_0 \leftarrow \{(x_t, y_t, r_0(y_t))\}$
**For** $t < T$
    Label $\lambda n$ datapoints into holdout set: $O_t \leftarrow O_{t-1} \bigcup \{x_i, y_i, D_{ss}(y_i)\}_{i=1}^{\lambda n}$;
    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 $\phi$ on $O_t$ and $r \leftarrow \text{RLLS}(O'_t, h_0)$.
    $\{x_t, P_t\} \leftarrow \text{ActiveSample}(\pi, \phi, x_t, r_t, D_{\text{src}}, D_{\text{ss}})$;      # Sample data using $\pi$, $\phi$ and weighted predictor
    $y_t \leftarrow C(x_t)$                                                   # Obtain label from oracle
    Update $S_t \leftarrow S_{t-1} \bigcup \{x_t, y_t, P_t\}$.                           # Update labeled set
**Output:** $h_T = \text{argmin}\{\text{err}(h, S_T, r_T) : h \in \mathcal{H}\}$
```

In a online setting, the ActiveSample subroutine labels datapoint x_t with probability $P_t D_{\text{ss}}(\phi(x_t))$ where P_t is the sampling probability given by $\pi(x_t)$. The analogous ActiveSample subroutine for mini-batch pool-based settings labels the top k data points in each class ranked in terms of the active sampling criterion π, where k is $BD_{\text{med}}(h_{t-1}(D_{\text{src}} \setminus S_{t-1}))$ and B is the batch size.

The medial distribution To inform a choice of medial distribution, we dichotomize label shift problems into two regimes: *source shift* and *target shift*. Source shift is characterized by an under-representation of certain classes in available data and is associated with label imbalance in the source distribution. Target shift is characterized by a change in priors and is associated with label imbalance in the target distribution. To intuit the effectiveness of subsampling under source versus target shift, we first observe that source shift typically results in a significantly larger label shift magnitude (evaluated by $||r - 1||$) than seen in target shift. For instance, for n datapoints with binary labels, the number of subsampled points necessary to correct extreme source shift (only 1 datapoint in the minority class) is $O(n)$ while the number of sbusampled points necessary to correct extreme target shift is $O(n^2)$. This difference in subsampling efficiency for two problems with an identical label shift magnitude suggests subsampling to a uniform label distribution before applying importance weighting. This exactly coincides with the choice of a uniform medial distribution.

4. Theoretical Analysis

We now analyze the instantiation of ALLS on IWAL-CAL, a disagreement-based active learning algorithm (Beygelzimer et al., 2010). To adapt the algorithm to our new setting, we substitute IWAL-CAL's original choice of a constant C_0 for our C_0 as given in Theorem 1. We omit a review of IWAL-CAL for brevity. In this section, we assume no warm start and instead defer the setting, along with complete proofs for the Theorems 1, 2 to the Appendix.

We now establish notation for the section. Let r denote the importance weights from the source to the target distribution ($r[i] = \frac{D_{\text{tar}}(y_i)}{D_{\text{src}}(y_i)}$) and $\theta := r - \mathbf{1}$. Similarly define $\theta_{s\to m}$ and $\theta_{m\to t}$ as θ's counterparts from source to medial and from medial to target distributions respectively. Also let $\sigma_{\min}$ denote the smallest singular value of blackbox hypothesis h_0. Finally, we define h_n as the ERM hypothesis after n unlabeled datapoints, err the target error, and $P_{\min,n}(h) := \min_{h(x_i)\neq h^*(x_i)} P_i$ the minimum sampling probability in the disagreement region of hypotheses h and h'.

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

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

where

$$\begin{aligned} C_0 \in \mathcal{O}\Bigg(&\log\left(\frac{|H|}{\delta}\right)\Big(d_\infty(D_{\mathit{tar}}||D_{\mathit{src}}) + d_2(D_{\mathit{tar}}||D_{\mathit{src}}) + 1 + \|\theta_{u\to t}\|_2^2\Big) \\ &+ \frac{\log\left(\frac{k}{\delta}\right)}{\sigma^2_{\min}} d_\infty(D_{\mathit{tar}}||D_{\mathit{med}})\, \|\theta_{m\to t}\|_2^2\, (\mathit{err}_W(h^*_{\mathit{online}}) + 1)\Bigg) \end{aligned} \tag{2}$$

Our generalization bound differs from the original IWAL-CAL bound in two key aspects. (1) 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 scale by 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 ($D_{\text{med}} = D_{\text{src}}$), 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 *training shift*.

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

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

where Θ *denotes the disagreement coefficient (Balcan et al., 2009).*

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 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.

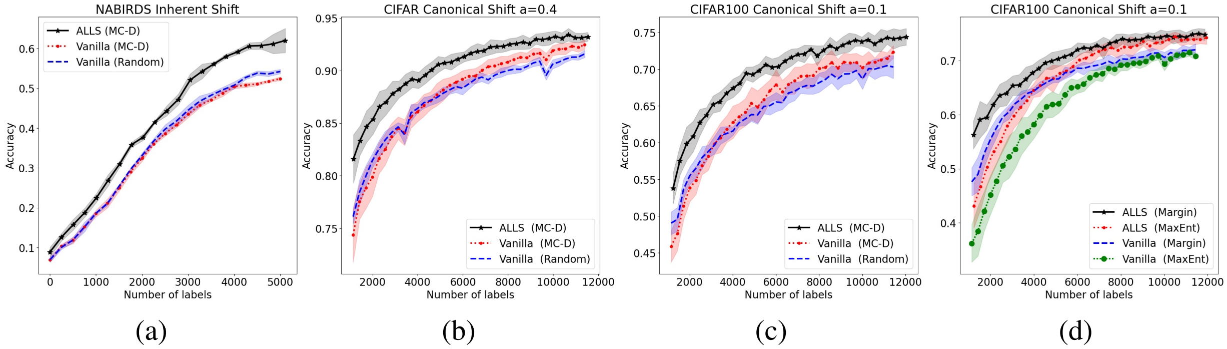


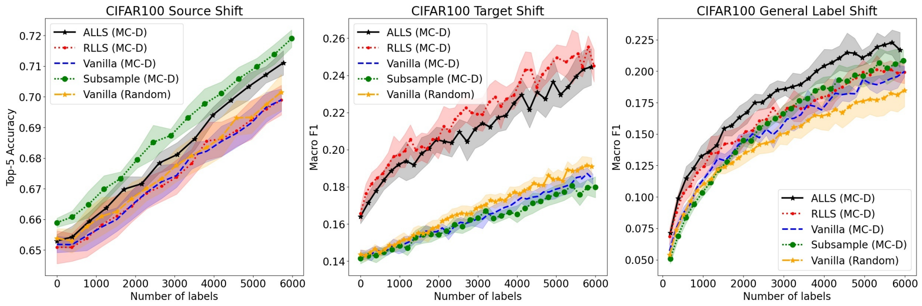
Figure 2: Average performance and 95% confidence intervals on 10 runs of experiments. (a) Accuracy on NABirds using MC-D; (b) Accuracy on CIFAR10 using MC-D; (c) Accuracy on CIFAR100 using Margin; (d) Accuracy on CIFAR100 using MaxEnt and MC-D as uncertainty estimates.

5. Experiments

We evaluate ALLS on real-world species recognition dataset NABirds (Van Horn et al., 2015) and benchmark datasets CIFAR10 & CIFAR100 (Krizhevsky, 2009). In these experiments, we instantiate ALLS on the uncertainty sampling algorithms: (1) Monte Carlo dropout (MC-D) (Gal and Ghahramani, 2016); (2) Maximum Entropy sampling (MaxEnt); and (3) Maximum Margin (Margin). In Figure 2, we present results on NABirds, CIFAR10, and CIFAR100. In the NABirds experiment, we apply ALLS to a naturally occuring class imbalance in the NABirds label hierarchy, where a single class constitutes a near-majority (Elhoseiny et al., 2017), and evaluate on a uniform label distribution. On each of the CIFAR10 and CIFAR100 datasets, we induce synthetic *warm-start* label shift by applying Lipton et al. (2018)'s *Dirichlet Shift* procedure to source and target data independently. In Figure 3, we similarly induce a diverse set of synthetic *training* label shift scenarios—a source shift setting, a target shift setting, and a mixed (general) shift setting—on CIFAR100 to highlight empirical evidence for the tradeoff between the importance weighting and sub-sampling. Further results and experiment details can be found in the appendix.

ALLS Practices Recommendations To scale label shift estimation to deep neural networks on high-dimensional datasets, we introduce the following techniques. Rather than applying importance weights to the loss function, we apply importance weights to the prediction distribution of the model and predict $p(y) = p(y)\frac{r(y)}{\sum_i r(y_i)}$. This reduces variance while preserving the use of the label shift information to correct uncertainty estimation, and bears some relation to posterior regularization (Saerens et al., 2002a). We also use the latest active learned hypothesis for label shift estimation instead of the static blackbox hypothesis demanded by theory; this heuristic steers our finite sample confusion matrices away from singularity (Lipton et al., 2018). We also forgo the use of a holdout set for RLLS, as suggested by Azizzadenesheli et al. (2019) and use actively sampled data for estimating the subsampling hypothesis ϕ and label shift weights r.

Performance Analysis Figure 2 exhibits our primary results and demonstrate that ALLS instantiations consistently introduce sample efficiency gains, outperforming random sampling even when vanilla counterparts dramatically underperform random sampling. Figures 2 and 3 demon-



(a) (b) (c)

Figure 3: Average performance and 95% confidence intervals on 10 runs of ablation studies. (a) Top-5 accuracy of the first ablation (source shift) experiment; (b) Macro F1 of the second ablation (target shift) experiment; (c) Macro F1 of the third ablation (general shift) experiment.

strate that ALLS gains can be realized both under "warmstart" and under "target" shift. In CIFAR10 and CIFAR100, instantiations of our framework required less than half the number of labels for achieving equivalent performance with their vanilla counterparts. Figure 3 corroborates a target-shift/source-shift trade-off in the respective strengths of importance weighting and subsampling. Under source shift, subsampling is the dominant strategy while under target shift, importance weighting is the dominant strategy. Although the strengths of the importance weighting and subsampling appear complementary, Figure 3 (c) demonstrates that, when properly balanced under ALLS, combined, the methods can realize additional gains. Overall, ALLS consistently demonstrates performance gains across active learning algorithms (Figure 2(c)), natural and synthetic label shift settings (Figures 2 (a),2(b)(c)), and *warmstart* and *training* shift types (Figures 2, 3(a) (b)).

6. 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.

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

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.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized Learning for Domain Adaptation under Label Shifts. *arXiv:1903.09734 [cs, stat]*, March 2019. URL `http://arxiv.org/abs/1903.09734`. arXiv: 1903.09734.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, January 2009. ISSN 0022-0000. doi: 10.1016/j.jcss.2008.07.003. URL `http://www.sciencedirect.com/science/article/pii/S0022000008000652`.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance Weighted Active Learning. *arXiv:0812.4952 [cs]*, May 2009. URL `http://arxiv.org/abs/0812.4952`. arXiv: 0812.4952.

Alina Beygelzimer, Daniel Hsu, John Langford, and Tong Zhang. Agnostic Active Learning Without Constraints. *arXiv:1006.2588 [cs]*, June 2010. URL `http://arxiv.org/abs/1006.2588`. arXiv: 1006.2588.

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. URL `http://papers.nips.cc/paper/4156-learning-bounds-for-importance-weighting.pdf`.

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

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.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv:1506.02142 [cs, stat]*, October 2016. URL `http://arxiv.org/abs/1506.02142`. arXiv: 1506.02142.

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

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.

Sheng-Jun Huang and Songcan Chen. Transfer learning with active queries from source domain. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1592–1598, New York, New York, USA, July 2016. AAAI Press. ISBN 978-1-57735-770-4.

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

Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and Correcting for Label Shift with Black Box Predictors. February 2018. URL `https://arxiv.org/abs/1802.03916v3`.

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.

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.

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, 2002a.

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 2002b. ISSN 0899-7667. doi: 10.1162/089976602753284446.

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

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.

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.

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.

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.

Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active Learning with Logged Data. *arXiv:1802.09069 [cs, stat]*, June 2018. URL `http://arxiv.org/abs/1802.09069`. arXiv: 1802.09069.

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.

Tong Zhang. Data Dependent Concentration Bounds for Sequential Prediction Algorithms. pages 173–187, June 2005. doi: 10.1007/11503415_12.

Appendix 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 first 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 $1 - \delta$:*

$$
\begin{aligned}
&|err(h, Z_{1:n}) - err(h^*, Z_{1:n}) - err(h) + err(h^*)| \\
&\leq \mathcal{O}\Bigg(d_\infty(D_{\textit{tar}}, D_{\textit{src}}) \frac{\log(n|H|/\delta)}{n} + \sqrt{d_2(D_{\textit{tar}}, D_{\textit{src}}) \frac{\log(n|H|/\delta)}{n}} + \sqrt{\frac{\log(n|H|/\delta)}{nP_{\min,n}(h)}} + \frac{\log(n|H|/\delta)}{nP_{\min,n}(h)} \\
&+ \left(1 + \textit{err}(h^*_{\textit{online}}) + \frac{\log(\lambda n/\delta)}{\lambda n} + \sqrt{\frac{\textit{err}(h^*_{\textit{online}}) \log(\lambda n/\delta)}{\lambda n}} + ||\theta_{\textit{src}\to\textit{med}}|| \sqrt{\frac{\log(n|H|/\delta)}{nP_{\min,n}(h)}}\right) \\
&\cdot \left(\frac{||\tilde{\theta}||_2 + 1}{\sigma_{\min}}\right) \sqrt{\frac{d_\infty(D_{\textit{tar}}, D_{\textit{med}}) \log(nk/\delta)}{\lambda n - \sqrt{n d_\infty(D_{\textit{tar}}, D_{\textit{med}}) \log(n/\delta) \lambda}}} \\
&+ ||\tilde{\theta}||_2 \left(\textit{err}(h^*_{\textit{online}}) + \frac{\log(\lambda n/\delta)}{\lambda n} + \sqrt{\frac{\textit{err}(h^*_{\textit{online}}) \log(\lambda n/\delta)}{\lambda n}}\right) + ||\theta||_2 \sqrt{\frac{\log(n|H|/\delta)}{nP_{\min,n}(h)}} \Bigg)
\end{aligned}
\tag{4}
$$

The corresponding bound for the case where only importance weighting is used can be recovered by setting $D_{\text{medial}} := D_{\text{src}}$. Our deviation bound approaches 0 as $n \to \infty$ at the same $\log(n)/n + \sqrt{\log(n)/n}$ asymptotic rate as IWAL-CAL. This deviation bound will plug in to IWAL-CAL for generalization and sample complexity bounds.

Through sections 7.1-7.6, we detail a proof of theorem 3. Let $f : X \times Y \to [-1, 1]$ be a bounded function; f will eventually represent $\text{err}(h) - \text{err}(h^*)$. We adopt Beygelzimer et al. (2010)'s notation where W denotes $Q_i \tilde{Q}_i \tilde{r}_i f(x_i, y_i)$ and Q_i is an indicator random variable indicating whether the ith datapoint is labeled. We also introduce $\tilde{W} := Q_i \hat{\tilde{Q}}_i \tilde{r}_i f(x_i, y_i)$ and $\hat{\tilde{W}} := Q_i \hat{\tilde{Q}}_i \hat{\tilde{r}}_i f(x_i, y_i)$, and analogous accented variants of Q. 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 adopt Azizzadenesheli et al. (2019)'s label shift notation and define k as the size of the output space (finite) and denote estimated importance weights with hats $\hat{\cdot}$. We further introduce $\tilde{r} := r_{\text{med}\to\text{tar}}$. These same semantics apply to accents on $\theta := r - 1$. We follow (Cortes et al., 2010) and use $d_\alpha(P||P')$ to denote $2^{D_\alpha(P||P')}$ where $D_\alpha(P||P') := \log(\frac{P_i}{P'_i})$ is the Renyi divergence of P and P'. We assume that for all $y \in Y$, $D_{\text{warm}}(y) \neq 0$, $D_{\text{src}}(y) \neq 0$. To prove theorem 3, we thus seek to bound with high probability

$$
\Delta := \frac{1}{n} \left(\sum_{i=1}^{n} \hat{\tilde{W}}_i \right) - \mathbb{E}[rf(X, Y)] \tag{5}
$$

We will individually bound the following terms,

$$\begin{aligned}
\Delta_1 &:= \mathbb{E}[rf(X,Y)] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i[W_i] \\
\Delta_2 &:= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i[W_i] - \mathbb{E}_i[\hat{W}_i] \\
\Delta_3 &:= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i[\hat{W}_i] - \mathbb{E}_i[\hat{\tilde{W}}_i] \\
\Delta_4 &:= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i[\hat{\tilde{W}}_i] - \hat{\tilde{W}}_i
\end{aligned} \tag{6}$$

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:

$$|\Delta| \leq |\Delta_1| + |\Delta_2| + |\Delta_3| + |\Delta_4| \tag{7}$$

A.2 Bounding Active Learning Stochasticity

We bound Δ_4 using a Martingale technique from (Zhang, 2005) also adopted by (Beygelzimer et al., 2010). We take Lemmas 1, 2 in (Zhang, 2005) as given and proceed in a fashion similar to the proof of Beygelzimer et al. (2010)'s Theorem 1.

The following lemma is a slightly modified analogue of Lemma 6 in (Beygelzimer et al., 2010).

Lemma 4 *If* $0 < \lambda < 3\frac{P_i}{\hat{\tilde{r}}_i}$, *then*

$$\log \mathbb{E}_i[\exp(\lambda(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i]))] \leq \frac{\hat{r}_i\hat{\tilde{r}}_i\lambda^2}{2P_i(1-\frac{\hat{\tilde{r}}_\lambda}{3P_i})} \tag{8}$$

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

$$\log \mathbb{E}_i[\exp(\lambda(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i]))] = 0 \tag{9}$$

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

$$|\hat{\tilde{W}}_i| \leq \left|\frac{Q_i\hat{\tilde{Q}}_i\hat{\tilde{r}}_i}{P_i}\right| \leq \frac{\hat{\tilde{r}}_i}{P_i} \tag{10}$$

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

$$\mathbb{E}_i[(\hat{\tilde{W}}_i - \mathbb{E}_i[\hat{\tilde{W}}_i])^2] \leq \frac{\hat{r}_i\hat{\tilde{r}}_i}{P_i}f(x_i,y_i)^2 - 2\hat{r}_i^2 f(x_i,y_i)^2 + \hat{r}_i^2 f(x_i,y_i)^2 \leq \frac{\hat{r}_i\hat{\tilde{r}}_i}{P_i} \tag{11}$$

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

$$\begin{aligned}\mathbb{E}_i[\exp(\lambda(\hat{\hat{W}}_i - \mathbb{E}_i[\hat{\hat{W}}_i]))] &= \mathbb{E}_i[1 + \lambda(\hat{\hat{W}}_i - \mathbb{E}_i[\hat{\hat{W}}_i]) + \lambda^2(\hat{\hat{W}}_i - \mathbb{E}_i[\hat{\hat{W}}_i])^2 g(\lambda(\hat{\hat{W}}_i - \mathbb{E}_i[\hat{\hat{W}}_i]))] \\ &= 1 + \lambda^2 \mathbb{E}_i[(\hat{\hat{W}}_i - \mathbb{E}_i[\hat{\hat{W}}_i])^2 g(\lambda(\hat{\hat{W}}_i - \mathbb{E}_i[\hat{\hat{W}}_i]))] \\ &\leq 1 + \lambda^2 \mathbb{E}_i[(\hat{\hat{W}}_i - \mathbb{E}_i[\hat{\hat{W}}_i])^2 g(\lambda \hat{\hat{r}}_i / P_i)] \\ &= 1 + \lambda^2 \mathbb{E}_i[(\hat{\hat{W}}_i - \mathbb{E}_i[\hat{\hat{W}}_i])^2] g(\lambda \hat{\hat{r}}_i / P_i) \\ &\leq 1 + \frac{\lambda^2 \hat{r}_i \hat{\hat{r}}_i}{P_i} g(\frac{\hat{\hat{r}}_i \lambda}{P_i})\end{aligned} \tag{12}$$

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 $\hat{\hat{W}}_i$ and the fact that $\mathbb{E}_i[\hat{\hat{W}}_i] = \hat{r} f(X_i, Y_i)$. ■

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

Lemma 5 *Pick any $t \geq 0, p_{\min} > 0$ and let E be the joint event*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \hat{\hat{W}}_i - \sum_{i=1}^{n} \mathbb{E}_i[\hat{\hat{W}}_i] \geq (1 + M)\sqrt{\frac{t}{2np_{\min}}} + \frac{t}{3np_{\min}} \\ \textit{and } \min\{\frac{P_i}{\hat{\hat{r}}_i} : 1 \leq i \leq n \wedge \mathbb{E}_i[W_i] \neq 0\} \geq p_{\min}\end{aligned} \tag{13}$$

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

Proof We follow (Beygelzimer et al., 2010) and let

$$\lambda := 3p_{\min} \frac{\sqrt{\frac{2t}{9np_{\min}}}}{1 + \sqrt{\frac{2t}{9np_{\min}}}} \tag{14}$$

Note that $0 < \lambda < 3p_{\min}$. By lemma 4, we know that if $\min\{\frac{P_i}{\hat{\hat{r}}_i} : 1 \leq i \leq n \wedge \mathbb{E}_i[\hat{\hat{W}}_i] \neq 0\} \geq p_{\min}$ then

$$\frac{1}{n\lambda}\sum_{i=1}^{n} \log \mathbb{E}_i[\exp(\lambda(W_i - \mathbb{E}_i[W_i]))] \leq \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{r}_i \hat{\hat{r}}_i \lambda}{2P_i(1 - \frac{\hat{\hat{r}}_i \lambda}{3P_i})} \leq M\sqrt{\frac{t}{2np_{\min}}} \tag{15}$$

and

$$\frac{t}{n\lambda} = \sqrt{\frac{t}{2np_{\min}}} + \frac{t}{3np_{\min}} \tag{16}$$

Let E' be the event that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\hat{W}}_i - \mathbb{E}_i[\hat{\hat{W}}_i]) - \frac{1}{n\lambda}\sum_{i=1}^{n} \log \mathbb{E}_i[\exp(\lambda(\hat{\hat{W}} - \mathbb{E}_i[\hat{\hat{W}}]))] \geq \frac{t}{n\lambda} \tag{17}$$

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 Zhang (2005)'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 6 *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*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\tilde{W}}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_i[\hat{\tilde{W}}_i]\right| \geq (1+M)\sqrt{\frac{R_n t}{2n}} + \frac{R_n t}{3n}\right) \leq 2(2+\log_2 r_{\max})e^{-t/2} \tag{18}$$

Proof This proof follows identically to (Beygelzimer et al., 2010)'s lemma 8. ■

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

Lemma 7 *With probability at least* $1 - \delta$, *the following holds over all* $n \geq 1$ *and* $h \in H$:

$$|\Delta_4| \leq (2 + ||\hat{\theta}||_2)\sqrt{\frac{\varepsilon_n}{P_{\min,n}(h)}} + \frac{\varepsilon_n}{P_{\min,n}(h)} \tag{19}$$

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 $\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i \leq \frac{1}{n}||\hat{\theta}||_2||\dot{\ell}||_2$ where $\dot{\ell}(j)$ is another k-sized vector where $\dot{\ell}(j) := \sum_{i=1}^{n}\mathbb{1}_{y_i=j}$. Since $||\dot{\ell}||_2 \leq n$, we have that $M \leq 1 + ||\hat{\theta}||_2$. The rest of the claim follows by lemma 6 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

$$|\Delta_3| = \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_i[\tilde{Q}_i] - \mathbb{E}_i[\hat{\tilde{Q}}_i])\hat{\tilde{r}}_i f(x_i, y_i)\right| \leq \left|\frac{1}{n}\sum_{j=1}^{k}\hat{\tilde{r}}_i\tilde{\ell}(j)\right| \tag{20}$$

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,

$$|\Delta_3| \leq \frac{1}{n}||\tilde{\ell}||_2||\hat{\tilde{r}}||_2 \tag{21}$$

Intuitively, the quantity $||\tilde{\ell}||_2$ 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 $||\tilde{\ell}||_2$ 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 $\mathcal{O}(\text{err}(h^*_{\text{online}}) + \frac{\log(n/\delta)}{n} + \sqrt{\frac{\text{err}(h^*_{\text{online}})\log(n/\delta)}{n}})$ to hold with probability at least $1-\delta$. Here, err denotes the absolute error and $\text{err}(h^*_{\text{online}})$ denotes the best achievable loss of a subsampling weight estimator on the source distribution.

Lemma 8 *With probability at least* $1-\delta$*,*

$$|\Delta_3| \leq ||\hat{\hat{r}}||_2 \mathcal{O}\left(err(h^*_{online}) + \frac{\log(\lambda n/\delta)}{\lambda n} + \sqrt{\frac{err(h^*_{online})\log(\lambda n/\delta)}{\lambda n}}\right) \tag{22}$$

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

A key observation is that $\text{err}(h^*_{\text{online}})$ is often 0, even in an agnostic learning setting. 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. Consistency is achievable when $\text{err}(h^*_{\text{online}}) = 0$, which is the setting we assume in the main paper.

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 (Azizzadenesheli et al., 2019) 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 9 *With probability at least* $1-\delta$*, the number of source samples is bounded below by*

$$n_p \geq \frac{\lambda n}{d_\infty(D_{med}||D_{src})} - \sqrt{-2\frac{\lambda n}{d_\infty(D_{med}||D_{src})}\log(\delta)} \tag{23}$$

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,

$$\mu := \mathbb{E}[n_p] = \mathbb{E}[\frac{\sum_{i=1}^{\lambda n} D_{ss}(y_i)}{\max_i D_{ss}(y_i)}] \geq \frac{\sum_{i=1}^{\lambda n} \mathbb{E}[D_{ss}(y_i)]}{d_\infty(D_{med}||D_{src})} = \frac{\lambda n}{d_\infty(D_{med}||D_{src})} \tag{24}$$

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

$$n_p \leq (1-\delta)\mu \tag{25}$$

and with probability at most δ that

$$\begin{aligned} n_p &\leq \mu(1 - \sqrt{-2\log(\delta)/\mu}) \\ &= \frac{\lambda n}{d_\infty(D_{med}||D_{src})} - \sqrt{-2\frac{\lambda n}{d_\infty(D_{med}||D_{src})}\log(\delta)} \\ &= \frac{1}{d_\infty(D_{med}||D_{src})}\left(\lambda n - \sqrt{-2\lambda n d_\infty(D_{med}||D_{src})\log(\delta)}\right) \end{aligned} \tag{26}$$

■

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

Lemma 10 *With probability* $1 - 2\delta$*, for all* $n \geq 1$*:*

$$|\Delta_2| \leq \frac{2}{\sigma_{\min}}\mathcal{O}\left(||\tilde{\theta}||_2\sqrt{\frac{d_\infty(D_{med}||D_{src})\log\left(\frac{nk}{\delta}\right)}{\lambda n - \sqrt{2\lambda n d_\infty(D_{med}||D_{src})\log\left(\frac{n}{\delta}\right)}}} + \sqrt{\frac{d_\infty(D_{med}||D_{src})\log\left(\frac{n}{\delta}\right)}{\lambda n - \sqrt{2\lambda n d_\infty(D_{med}||D_{src})\log\left(\frac{n}{\delta}\right)}}}\right) \tag{27}$$

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

$$|\Delta_2| = \left|\frac{1}{n}\sum_{i=1}^{n}(\tilde{r}_i - \hat{\hat{r}}_i)f(x_i, y_i)\right| \leq \left|\frac{1}{n}\sum_{j=1}^{k}(\tilde{r}(j) - \hat{\hat{r}}(j))\tilde{\ell}(j)\right| \tag{28}$$

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:

$$\left|\frac{1}{n}\sum_{j=1}^{k}(\tilde{r}(j) - \hat{\hat{r}}(j))\tilde{\ell}(j)\right| \leq \frac{1}{n}||(\tilde{r}(j) - \hat{\hat{r}}(j))||_2||\tilde{\ell}||_2 \tag{29}$$

Since $f(x,y) \in [-1,1]$, we can bound $||\tilde{\ell}||_2$ by $2n$. Then, $|\Delta_2| \leq 2||\tilde{\theta} - \hat{\hat{\theta}}||_2$. Azizzadenesheli et al. (2019)'s (RLLS) lemma 1 then gives the following bound on $||\tilde{\theta} - \hat{\hat{\theta}}||_2$ which holds with probability $1 - \delta$:

$$||\tilde{\theta} - \hat{\hat{\theta}}||_2 \leq \mathcal{O}\left(\frac{1}{\sigma_{\min}}(||\theta||_2\sqrt{\frac{\log(k/\delta)}{n_p}} + \sqrt{\frac{\log(1/\delta)}{n_p}})\right) \tag{30}$$

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 $1-\delta$:

$$|\Delta_2| \leq \mathcal{O}\left(\frac{2}{\sigma_{\min}}\left(||\tilde{r}-1||_2\sqrt{\frac{\log(k/\delta)}{n_p}}+\sqrt{\frac{\log(1/\delta)}{n_p}}\right)\right) \tag{31}$$

The bound then follows immediately by lemma 9 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 11 *For any $\delta > 0$, with probability at least $1-\delta$, then for all $n \geq 1$, $h \in H$:*

$$|\Delta_1| \leq \frac{2d_\infty(D_{tar}, D_{src})\log(\frac{2n|H|}{\delta})}{3(n+m)} + \sqrt{\frac{2d_2(D_{tar}, D_{src})\log(\frac{2n|H|}{\delta})}{n+m}} \tag{32}$$

Proof This inequality is a direct application of Theorem 2 from (Cortes et al., 2010). ■

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 $\hat{\hat{\theta}}$. We remove these terms using our known bounds on Δ_2 through a simple triangle inequality. Specifically, $||\hat{\theta}|| \leq ||\theta|| + ||\hat{\theta}-\theta||$ where we have already bounded the latter term in the proof of lemma 10. 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.

Appendix B. Correctness and Sample Complexity Corollaries

As in (Beygelzimer et al., 2010), 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 $h \in H$:

$$|err(h, Z_{1:n}) - err(h^*, Z_{1:n}) - err(h) + err(h^*)| \leq (\left\|\tilde{\theta}\right\|_2 + 1)\text{err}_W(h^*_{\text{online}}) + \sqrt{\frac{\epsilon_n}{P_{\min,n}(h)}} + \frac{\epsilon_n}{P_{\min,n}(h)} \tag{33}$$

We simply base C_0 off the deviation bound from Theorem 3. For readibility, we aggressively drop terms from the asymptotic in Equation 4 to bound:

$$\begin{aligned} C_0 \in \mathcal{O}\Bigg(&\log\left(\frac{|H|}{\delta}\right)\left(d_\infty(D_{\text{tar}}||D_{\text{src}}) + d_2(D_{\text{tar}}||D_{\text{src}}) + 1 + \|\theta\|_2^2\right) \\ &+ \frac{\log\left(\frac{k}{\delta}\right)}{\sigma_{\min}^2} d_\infty(D_{\text{tar}}||D_{\text{med}})\left\|\tilde{\theta}\right\|_2^2(\text{err}_W(h^*_{\text{online}})+1)\Bigg) \end{aligned} \tag{34}$$

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 (Beygelzimer et al., 2010), noting that our ϵ_n is asymptotically equivalent to the ϵ_n in the original IWAL-CAL derivations of (Beygelzimer et al., 2010), differing only in the choice of constant C_0 and the presence of an additional bias term, $\text{err}(h^*_{\text{online}})$, in Equation 33. Hence, our proof of Theorem 1 follows immediately from Lemma 2 and Theorem 2 in (Beygelzimer et al., 2010). Substituting our Theorem 1 into Theorem 3 from (Beygelzimer et al., 2010) similarly immediately yields 2 minus the λn labels necessary for accumulating a holdout set for RLLS and subsampling.

Appendix 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 $D_{\text{src}} = \frac{nD_{\text{src}}+mD_{\text{warm}}}{n+m}$ and $D_{\text{lab}} = \frac{nD_{\text{med}}+mD_{\text{warm}}}{n+m}$. We redefine $\tilde{\theta}$ as $\theta_{\text{lab}\to\text{tar}}$. Our bound on Δ_4 from lemma 7 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 12 *With probability* $1 - 2\delta$*, for all* $n \geq 1$*,* $h \in H$*:*

$$|\Delta_2| \leq \mathcal{O}\left(\frac{2}{\sigma_{\min}}\left(\left\|\tilde{\theta}\right\|_2 \sqrt{\frac{d_\infty(D_{med}||D_{src})\log\left(\frac{nk}{\delta}\right)}{\lambda(n+m) - \sqrt{2\lambda(n+m)d_\infty(D_{med}||D_{src})\log\left(\frac{n}{\delta}\right)}}} + \sqrt{\frac{d_\infty(D_{med}||D_{src})\log\left(\frac{n}{\delta}\right)}{\lambda(n+m) - \sqrt{2\lambda(n+m)d_\infty(D_{med}||D_{src})\log\left(\frac{n}{\delta}\right)}}}\right)\right) \tag{35}$$

Lemma 13 *For any* $\delta > 0$*, with probability at least* $1 - \delta$*, then for all* $n \geq 1$*,* $h \in H$*:*

$$|\Delta_1| \leq \frac{2d_\infty(D_{tar}, D_{src})\log(\frac{2n|H|}{\delta})}{3(n+m)} + \sqrt{\frac{2d_2(D_{tar}, D_{src})\log(\frac{2n|H|}{\delta})}{n+m}} \tag{36}$$

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:

$$C_0 \in \mathcal{O}\left(\log\left(\frac{|H|}{\delta}\right)\left(d_\infty(D_{\text{tar}}||D_{\text{src}}) + d_2(D_{\text{tar}}||D_{\text{src}}) + 1 + \|\theta\|_2^2\right) + \frac{n\log\left(\frac{k}{\delta}\right)}{(n+m)\sigma_{\min}^2} d_\infty(D_{\text{tar}}||D_{\text{lab}})\left\|\tilde{\theta}\right\|_2^2 (\text{err}_W(h^*_{\text{online}}) + 1)\right) \tag{37}$$

The corresponding generalization and sample complexity bounds follow accordingly.

Appendix 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 **??**. 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 setting 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. Table 1 also breaks out the L2 distance between the label distribution of the target distribution, and the label distribution of data labeled by the active learning (or random) policy. In line with intuition, active learning exhibits an inherent bias towards uniformity while ALLS exaggerates this effect due to our choice of a uniform medial distribution.

Strategy	Acc (854 Labels)	Acc (1708)	Acc (3416)	L2 Dist to Uniform (854)
ALLS (MC-D)	**0.51**	**0.53**	**0.56**	7.922k
Vanilla (MC-D)	0.46	0.48	0.50	8.168k
Random	0.38	0.40	0.42	8.358k

Table 1: NABirds (species) Experiment Average Accuracy

D.2 Additional Ablation Studies

We also include additional ablations studies regarding different shift magnitudes and different practices for scaling label shift estimation algorithms to large output spaces and deep learning settings.

D.2.1 Different levels of α

We evaluate our framework on different magnitude of artificial label shift, where label shift is induced according to a Dirichlet distribution of parameter α as described in (Lipton et al., 2018). Figure 4 demonstrates that ALLS improves over the vanilla active learning and random sampling across all the shift magnitudes. Moreover, the improvement is more significant with larger shift. Note that shift magnitude is inversely correlated with α—smaller α denotes a larger shift.

D.2.2 Different RLLS modifications

We introduce two major modifications to the RLLS (Azizzadenesheli et al., 2019) label shift estimation procedure to scale the practice of importance weighted learning under label shift to the deep neural networks setting: posterior regularization (PR) and iterative reweighting (ITIW v.s. No-ReW). The former provides an alternative to the direct use of label shift weights as importance weights, and is inspired by the expectation-maximization algorithm described in (Saerens et al., 2002a). The latter dispenses with theoretical concerns of statistical independence, and use a label-shift-corrected estimator as the black-box for label shift estimation. We conduct ablation studies

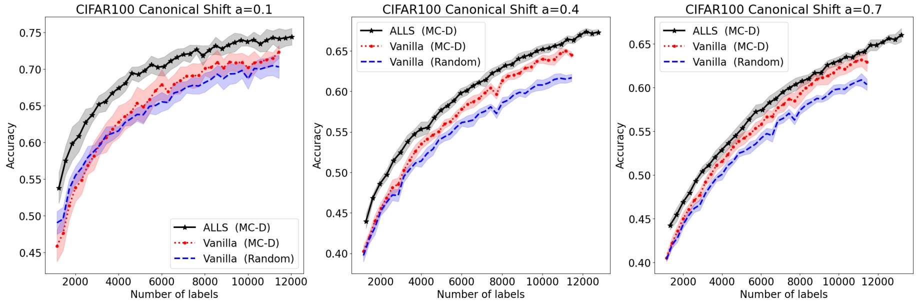


Figure 4: Average performance and 95% confidence intervals on 10 runs of experiments on CIFAR100 in the general shift setting. ALLS gains scale by label shift magnitude.

on strategies incorporating either—or both—of these approaches on the CIFAR100 dataset, which features a moderately sized output space. Figure 5 demonstrates that, in the absence of these heuristics, the label shift corrected active learning policy remains on par with random sampling but either heuristic realizes a statistically significant improvement in performance. Importantly, our results demonstrate that applying both further boosts performance.

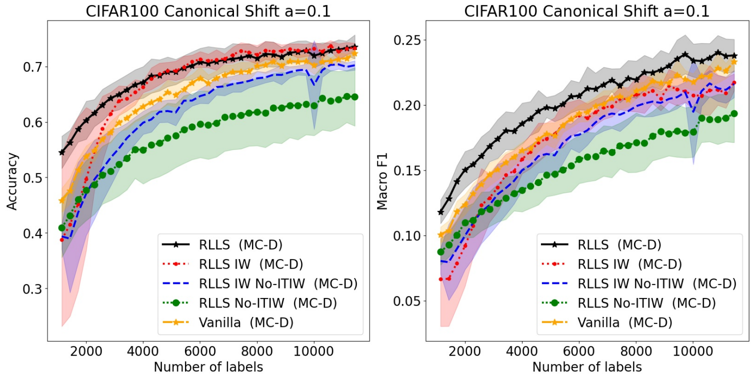


Figure 5: Average performance and 95% confidence intervals on 10 runs of experiments on CIFAR100 in warm-start shift setting. **Left:** Accuracy using MC-D; **Right:** Macro F1 using MC-D; Posterior regularization, iterative reweighting allow for additional gains on top of ALLS by improving the stability and performance of label shift estimation on high-dimensional data.

D.3 Online ALLS

We also evaluate a bootstrap approximation of IWAL (Beygelzimer et al., 2009) using a version space of 8 Resnet-18 models on the CIFAR dataset. As with the more practical pool-based instantiations, we see modest gains due to ALLS.

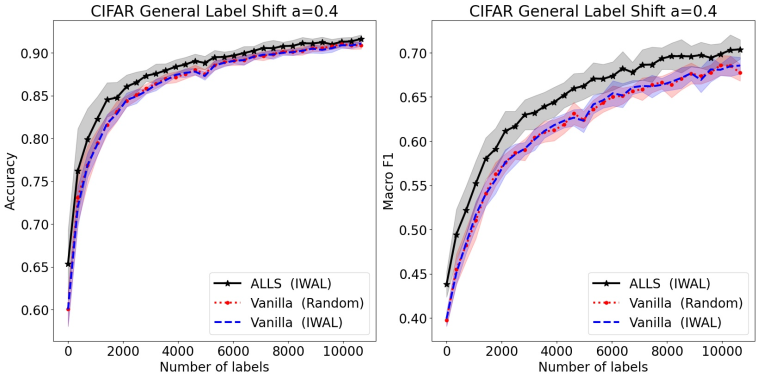


Figure 6: Average performance and 95% confidence intervals on 10 runs of experiments on CIFAR in warm-start shift setting. **Left:** Accuracy using IWAL-CAL; **Right:** Macro F1 using IWAL-CAL. ALLS leads to modest gains even in difficult online learning settings.

Appendix 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 (Azizzadenesheli et al., 2019) 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 in a batched setting. 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; we use "warm" to denote "warmstart", and figure numbers reference figures in the main paper and appendix.

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

Table 2: Dataset-wide statistics and parameters

The complete source code for replicating and expanding our experiment base is released anonymously at `https://bit.ly/2UVr1bb`.

Setting	Warm Ratio	Source Dist	Target Dist	Warm Shift?	Dirichlet α
NABirds (Figure 2)	1.0	Inherent	Inherent	No	N/A
CIFAR (Figure 2)	0.3	Dirichlet	Dirichlet	Yes	0.7
CIFAR100 (Figure 2)	0.4	Dirichlet	Dirichlet	Yes	0.1
CIFAR100 (Figure 3 Left)	0.4	Dirichlet	Uniform	No	1.0
CIFAR100 (Figure 3 Mid)	0.3	Uniform	Dirichlet	No	0.1
NABirds (Table 1)	1.0	N/A	Dirichlet	No	0.1
CIFAR100 (Figure 4)	0.4	Dirichlet	Dirichlet	No	0.1/0.4/0.7
CIFAR100 (Figure 5)	0.4	Dirichlet	Dirichlet	Yes	0.1
CIFAR (Figure 6)	0.5	Dirichlet	Dirichlet	Yes	0.4

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