8 Proofs

8.1 Proof of Theorem 1

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

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

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

(15)

Proof. The drift is equivalent to expected importance weights,

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

(16)

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

\[ 1 - \mathbb{E}_{X,Y \sim P_{\text{med}}} \left[ \frac{P_{\text{test}}(y)}{P_{\text{med}}(y)} \right] = 1 - \int_{X,Y} C P_{\text{src}}(x,y) P_{\text{ss}}(h_0(x)) \frac{P_{\text{test}}(y)}{P_{\text{med}}(y)} \]
\[ = 1 - C \mathbb{E}_{X,Y \sim P_{\text{src}}} \left[ P_{\text{ss}}(h_0(x)) \frac{P_{\text{test}}(y)}{P_{\text{med}}(y)} \right] \]
\[ = 1 - C \mathbb{E}_{X,Y \sim P_{\text{src}}} \left[ P_{\text{ss}}(y) \frac{P_{\text{test}}(y)}{P_{\text{med}}(y)} \right] + C \mathbb{E}_{X,Y \sim P_{\text{src}}} \left[ (P_{\text{ss}}(h_0(x)) - P_{\text{ss}}(y)) \frac{P_{\text{test}}(y)}{P_{\text{med}}(y)} \right] \]
\[ = 1 - \sum_{y} P_{\text{med}}(y) \frac{P_{\text{test}}(y)}{P_{\text{med}}(y)} + C \mathbb{E}_{X,Y \sim P_{\text{src}}} \left[ (P_{\text{ss}}(h_0(x)) - P_{\text{ss}}(y)) \frac{P_{\text{test}}(y)}{P_{\text{med}}(y)} \right] \]  

(17)

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

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

(18)

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

Plugging Lemma 1 into Theorem 2 in [Azizzadenesheli et al., 2019] yields a generalization of Theorem 1 where the number of unlabeled datapoints from the test distribution is \( n' \).
Theorem 4. With probability $1 - \delta$, for all $n \geq 1$:

$$|\Delta| \leq \mathcal{O} \left( \frac{2}{\sigma_{\min}} \left( \|\theta_{m-1}\|_2 \sqrt{\frac{\log \left( \frac{\|s\|}{\delta} \right)}{n}} + \sqrt{\frac{\log \left( \frac{n}{\delta} \right)}{n}} + \sqrt{\frac{\log \left( \frac{n}{\delta} \right)}{n'}} + \|\theta_{s-m}\|_\infty \text{err}(h_0, r_{m-1}) \right) \right)$$

(19)

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

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

8.2 Theorem 2 and Theorem 3 Proofs

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

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

$$h^* := \text{argmin}_{h \in H} \text{err}_{P_{test}}(h),$$

$$h_k := \text{argmin}_{h \in H} \text{err}_{S_{k-1}}(h),$$

$$h_k' := \text{argmin}\{|\text{err}_{S_{k-1}}(h)|: h \in H \land h(D^{(k)}_{unlab}) \neq h_k(D^{(k)}_{unlab})\}$$

$$G_k := \text{err}_{S_{k-1}}(h_k') - \text{err}_{S_{k-1}}(h_k)$$

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

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

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

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

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

$$|\text{err}(h, Z_{1:n}) - \text{err}(h^*, Z_{1:n}) - \text{err}(h) + \text{err}(h^*)|$$

$$\leq \mathcal{O} \left( (2 + \|\theta\|_2) \sqrt{\frac{\varepsilon_n}{P_{\text{min,n}}(h)}} + \sqrt{\frac{\log \left( \frac{n}{\delta} \right)}{\lambda n}} + \sqrt{\frac{\log \left( \frac{n}{\delta} \right)}{n'}} + \sqrt{\frac{\log \left( \frac{n}{\delta} \right)}{n}} + \|\theta_{s-m}\|_\infty \text{err}(h_0, r_{m-1}) \right)$$

(20)

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

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

We borrow notation for the terms $W, Q$ from [Beygelzimer et al., 2010], where $Q_i$ is an indicator random variable indicating whether the $i$th datapoint is labeled and $W := Q_i \tilde{Q}_i r_{m-i}^{(i)} f(x_i, y_i)$. We use the shorthand $r^{(i)}_m$ for the $y_i$th component of importance weight $r$. Similarly, the indicator random variable $\tilde{Q}_i$ indicates whether the $i$th data sample is retained by the subsampler. The expectation $\mathbb{E}_h[W]$ is taken over the randomness of $Q$ and $Q$. We
also borrow [Azizzadenesheli et al., 2019]’s label shift notation and define $k$ as the size of the output space (finite) and denote estimated importance weights with hats, e.g. $\hat{r}$. We also introduce a variant of $W$ using estimated importance weights $r$: $\tilde{W} := Q_1 \hat{Q}_1 \hat{r}_{m-t}^{(i)} f_i(x_i, y_i)$. Finally, we follow [Cortes et al., 2010] and use $d_{\alpha}(P || P')$ to denote $2 D_{\alpha}(P || P')$ where $D_{\alpha}(P || P') := \log \frac{P}{P'}$ is the Renyi divergence of distributions $P$ and $P'$.

We seek to bound with high probability,

$$\left| \Delta \right| := \frac{1}{n} \left( \sum_{i=1}^{n} \tilde{W}_i \right) - E_{x,y \sim P_{trg}} [f(x,y)] \leq |\Delta_1| + |\Delta_2| + |\Delta_3| + |\Delta_4|$$  \hspace{2cm} (21)

where,

$$\Delta_1 := E_{x,y \sim P_{trg}} [f(x,y)] - E_{x,y \sim P_{src}} [W_i],$$

$$\Delta_2 := E_{x,y \sim P_{src}} [W_i] - \frac{1}{n} \sum_{i=1}^{n} E_i [W_i],$$

$$\Delta_3 := \frac{1}{n} \sum_{i=1}^{n} E_i [W_i] - \hat{W}_i$$

$$\Delta_4 := \frac{1}{n} \sum_{i=1}^{n} E_i [\hat{W}_i] - \hat{W}_i$$

$\Delta_1$ corresponds to the drift from label shift introduced by subsampling, $\Delta_2$ to finite-sample variance, and $\Delta_3$ to label shift estimation errors. The final $\Delta_4$ corresponds to the variance from randomly sampling.

We bound $\Delta_4$ using a Martingale technique from [Zhang, 2005] also adopted by [Beygelzimer et al., 2010]. We take Lemmas 1, 2 from [Zhang, 2005] as given. We now proceed in a fashion similar to the proof of Theorem 1 from [Beygelzimer et al., 2010]. We begin with a generalization of Lemma 6 in [Beygelzimer et al., 2010].

**Lemma 2.** If $0 < \lambda < 3 \frac{P}{r_{m-t}}$, then

$$\log E_i [\exp (\lambda (\hat{W}_i - E_i[\hat{W}_i]))] \leq \frac{\hat{r}_i \hat{r}_{m-t}^{(i)} \lambda^2}{2 P_i (1 - \frac{\hat{r}_i \hat{r}_{m-t}^{(i)}}{r_{m-t}})}$$  \hspace{2cm} (22)

where $\hat{r}_i := \hat{r}_{m-t}^{(i)} E_i[\hat{Q}_i]$. If $E_i[\hat{W}_i] = 0$ then

$$\log E_i [\exp (\lambda (\hat{W}_i - E_i[\hat{W}_i]))] = 0$$  \hspace{2cm} (23)

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

$$|\hat{W}_i| \leq \frac{|Q_1 \hat{Q}_1 \hat{r}_{m-t}^{(i)}|}{P_1} \leq \frac{\hat{r}_{m-t}^{(i)}}{P_1}$$  \hspace{2cm} (24)

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

$$E_i [(\hat{W}_i - E_i[\hat{W}_i])^2] \leq \frac{\hat{r}_{m-t}^{(i)}}{P_1} f(x_i, y_i) - 2 \frac{\hat{r}_i}{P_1} f(x_i, y_i)^2 + \frac{\hat{r}_i^2}{P_1} f(x_i, y_i)^2 \leq \frac{\hat{r}_{m-t}^{(i)}}{P_1}$$  \hspace{2cm} (25)

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

$$E_i [\exp (\lambda (\hat{W}_i - E_i[\hat{W}_i]))] = E_i [1 + \lambda (\hat{W}_i - E_i[\hat{W}_i]) + \lambda^2 (\hat{W}_i - E_i[\hat{W}_i])^2 g(\lambda (\hat{W}_i - E_i[\hat{W}_i]))]$$

$$= 1 + \lambda^2 E_i [(\hat{W}_i - E_i[\hat{W}_i])^2 g(\lambda (\hat{W}_i - E_i[\hat{W}_i]))]$$

$$\leq 1 + \lambda^2 E_i [(\hat{W}_i - E_i[\hat{W}_i])^2 g(\hat{r}_{m-t}^{(i)}/P_i)]$$

$$= 1 + \lambda^2 E_i [(\hat{W}_i - E_i[\hat{W}_i])^2 g(\hat{r}_{m-t}^{(i)}/P_i)]$$

$$\leq 1 + \frac{\lambda^2 \hat{r}_{m-t}^{(i)}}{P_i} g(\hat{r}_{m-t})$$  \hspace{2cm} (26)
where the first inequality follows from our range bound and the second follows from our variance bound. The first claim then follows from the definition of \( g(x) \) and the facts that \( \exp(x) - x - 1 \leq x^2/(2(1-x/3)) \) for \( 0 \leq x < 3 \) and \( \log(1+x) \leq x \). The second claim follows from definition of \( \hat{W}_i \) and the fact that \( E_i[\hat{W}_i] = \hat{r}f(X_i, Y_i) \). \( \square \)

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

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

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{W}_i - \frac{1}{n} \sum_{i=1}^{n} E_i[\hat{W}_i] \geq (1 + M) \sqrt{\frac{t}{2np_{\min}}} + \frac{t}{3np_{\min}}
\]

and

\[
\min\{ \frac{P_i}{P_{m-1}} : 1 \leq i \leq n \land E_i[W_i] \neq 0 \} \geq p_{\min}
\]

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

\[
\lambda := 3p_{\min} \sqrt{\frac{2t}{9np_{\min}}} + \frac{t}{3np_{\min}}
\]

Note that \( 0 < \lambda < 3p_{\min} \). By Lemma 2, we know that if \( \min\{ \frac{P_i}{P_{m-1}} : 1 \leq i \leq n \land E_i[W_i] \neq 0 \} \geq p_{\min} \) then

\[
\frac{1}{n} \sum_{i=1}^{n} \log E_i[exp(\lambda(W_i - E_i[\hat{W}_i]))] \leq \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{r}_i}{2P_i(1 - \frac{\epsilon(i)^2}{3\lambda^2})} \leq M \sqrt{\frac{t}{2np_{\min}}}
\]

and

\[
\frac{t}{n\lambda} = \sqrt{\frac{t}{2np_{\min}}} + \frac{t}{3np_{\min}}
\]

Let \( E' \) be the event that

\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{W}_i - E_i[\hat{W}_i]) - \frac{1}{n\lambda} \sum_{i=1}^{n} \log E_i[exp(\lambda(W_i - E_i[\hat{W}_i]))] \geq \frac{t}{n\lambda}
\]

and let \( E'' \) be the event \( \min\{ \frac{P_i}{P_{m-1}} : 1 \leq i \leq n \land E_i[W_i] \neq 0 \} \geq p_{\min} \). Together, the above two equations imply \( E \subseteq E' \cap E'' \). By [Zhang, 2005]'s lemmas 1 and 2, \( \Pr(E) \leq \Pr(E' \cap E'') \leq \Pr(E') \leq e^{-t} \). \( \square \)

The following is an immediate consequence of the previous lemma.

**Lemma 4.** Pick any \( t \geq 0 \) and \( n \geq 1 \). Assume \( 1 \leq \frac{\epsilon(i)^2}{P_i} \leq r_{\max} \) for all \( 1 \leq i \leq n \), and let \( R_n := \max\{ \frac{\epsilon(i)^2}{P_i} : 1 \leq i \leq n \land E_i[\hat{W}] \neq 0 \} \cup \{1\} \). We have

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

**Proof.** This proof follows identically to [Beygelzimer et al., 2010]'s lemma 8. \( \square \)

We can finally bound \( \Delta_4 \) by bounding the remaining free quantity \( M \).

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

\[
|\Delta_4| \leq (2 + \ell_2) \sqrt{\frac{\epsilon_n}{P_{\min,n}(h)}} + \frac{\epsilon_n}{P_{\min,n}(h)}
\]

where \( \epsilon_n := \frac{16 \log(2(2+n \log_2 n)n(n+1)|H|/\delta)}{n} \) and \( P_{\min,n}(h) = \min\{ P_i : 1 \leq i \leq n \land h(X_i) \neq h^*(X_i) \} \cup \{1\} \).
Applying the Cauchy-Schwarz inequality, we have that Theorem 2 follows by replacing the deviation bound in [Beygelzimer et al., 2010]'s Theorem 2 with our Theorem where $n$ is the number of datapoints. Both of which disappear in the limit where $n >> m$. For the remainder of this proof, we continue to set $m = 0$.

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

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

$$|\Delta_2| \leq \frac{2d_2(P_{\text{test}}, P_{\text{arc}}) \log(\frac{2n|H|}{\delta})}{3n} + \frac{\sqrt{2d_2(P_{\text{test}}, P_{\text{arc}}) \log(\frac{2n|H|}{\delta})}}{n}$$

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

The following lemma bounds the remaining term $\Delta_1$.

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

$$|\Delta_1| \leq \|r_{x-m}\|_{\infty} \text{err}(h_0, r_{x-m})$$

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

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

$$|\text{err}(h, Z_{1:n}) - \text{err}(h^*, Z_{1:n}) - \text{err}(h) + \text{err}(h^*)|$$

$$\leq C \left( (2 + \frac{n \|\theta_{w-t}\|_2 + m \|\theta_{w-t}\|_2}{n + m}) \sqrt{\frac{\varepsilon_n}{P_{\text{min,n}}(h)}} + \frac{\varepsilon_n}{P_{\text{min,n}}(h)} + \frac{2d_2(P_{\text{test}}, P_{\text{arc}}) \log(\frac{2n|H|}{\delta})}{3(n + m)} 
+ \frac{n \|r_{x-m}\|_{\infty} \text{err}(h_0, r_{x-m})}{n + m} 
+ \frac{n}{\sigma_{\text{min}}} \left( \frac{\|\theta_{m-t}\|_2}{\lambda n} + \sqrt{\frac{\log(\frac{n \|\theta_{m-t}\|_2}{\lambda n})}{\lambda n}} + \sqrt{\frac{\log(\frac{\|\theta_{m-t}\|_2}{\lambda n})}{n'}} + \frac{\|r_{x-m}\|_{\infty} \text{err}(h_0, r_{x-m})}{n'} \right) \right)$$

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

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

$$\Pr\left(n \geq \frac{n}{\|r_{x-m}\|_{\infty}} + \log_2\left(\frac{1}{\epsilon}\right)\right) \leq \epsilon$$

**Proof.** The number of subsampled datapoints is sum of independent Bernoulli trials with mean $\mu$,

$$\mu = \mathbb{E}_{y \sim P_{\text{arc}}} [P_{\text{arc}}(y)] = \mathbb{E}_{y \sim P_{\text{arc}}} \left[ \frac{C_{P_{\text{arc}}}(y)}{P_{\text{arc}}(y)} \right] = \mathbb{E}_{y \sim P_{\text{arc}}} [C] = C$$

where $C$ is a constant such that $C_{P_{\text{arc}}}(y) \leq 1$ for all labels $y$. Thus, $\mu = C \leq 1/\|r_{x-m}\|_{\infty}$. 

9 Supplementary Experiments

9.1 NABirds Regional Species Experiment

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

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Acc (854 Labels)</th>
<th>Acc (1708)</th>
<th>Acc (3416)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALLS (MC-D)</td>
<td>0.51</td>
<td>0.53</td>
<td>0.56</td>
</tr>
<tr>
<td>Vanilla (MC-D)</td>
<td>0.46</td>
<td>0.48</td>
<td>0.50</td>
</tr>
<tr>
<td>Random</td>
<td>0.38</td>
<td>0.40</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Table 1: NABirds (species) Experiment Average Accuracy

9.2 Change in distribution

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

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

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th># Datapoints</th>
<th>Epochs (init/fine)</th>
<th># Batches</th>
<th># Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NABirds1</td>
<td>Resnet-34</td>
<td>30,000</td>
<td>60/10</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>NABirds2</td>
<td>Resnet-34</td>
<td>30,000</td>
<td>60/10</td>
<td>20</td>
<td>228</td>
</tr>
<tr>
<td>CIFAR</td>
<td>Resnet-18</td>
<td>40,000</td>
<td>80/10</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>Resnet-18</td>
<td>40,000</td>
<td>80/10</td>
<td>40</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2: Dataset-wide statistics and parameters

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

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