Supplementary Materials

8 Proofs

8.1 Proof of Theorem 1

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

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

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

Proof. The drift is equivalent to expected importance weights,

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

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

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

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

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

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

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

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

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

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

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

8.2 Theorem 2 and Theorem 3 Proofs

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

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

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

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

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

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

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

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

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

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

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

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

We seek to bound with high probability,

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

where,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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*

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

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

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

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

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

and

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

Let E' be the event that

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

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

The following is an immediate consequence of the previous lemma.

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

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

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

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

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

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

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

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 4 and a union bound over hypotheses and datapoints. □

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

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

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

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

The following lemma bounds the remaining term Δ_1.

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

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

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

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

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

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

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

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

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

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

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

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

9 Supplementary Experiments

9.1 NABirds Regional Species Experiment

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

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

Table 1: NABirds (species) Experiment Average Accuracy

9.2 Change in distribution

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

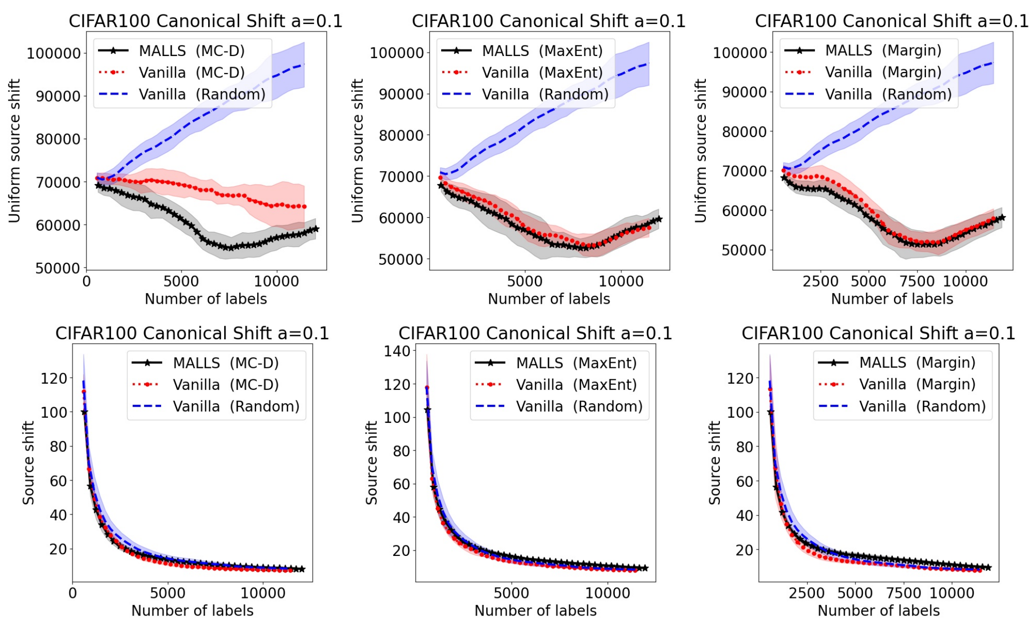

Figure 8: Average L2 distance between labeled class distribution and uniform/target distribution with 95% confidence intervals on 10 runs of experiments on CIFAR100 in the *canonical label shift* setting. MALLS (denoted by ALLS) converges to the target label distribution slower than vanilla active learning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

10 Experiment Details

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

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

Table 2: Dataset-wide statistics and parameters

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

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