A Proof of EIWAL

The key step in proving Theorem 1 for EIWAL is using a martingale concentration bound for

$$Z_t = \frac{Q_t}{p_t}\big(\ell(f(x_t), y_t) - \ell(g(x_t), y_t)\big) - \big(R(f) - R(g)\big),$$

where $Z_1, Z_2, \cdots$ is a martingale difference sequence for any pair $f, g \in \mathtt{H}_T$. Instead of using Azuma's inequality as in [Beygelzimer et al., 2009], we rely on a Berstein-like inequality for martingales Freedman [1975]).

The following result is adapted from Lemma 3 of Kakade and Tewari [2009], which is derived from Freedman [1975]). We denote by $\mathcal{F}_t = \{(x_1, y_1, Q_1), \cdots, (x_t, y_t, Q_t\}$ the observations up to time t.

Lemma 4. *For any $0 < \delta < 1$, and $T \geq 3$, with probability at least $1 - \delta$,*

$$\left|\sum_{t=1}^{T} Z_t\right| \leq \max\left\{2\sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_t}[p_t|\mathcal{F}_{t-1}]}, 6\sqrt{\log\left(\frac{8\log(T)}{\delta}\right)}\right\} \times \sqrt{\log\left(\frac{8\log(T)}{\delta}\right)}.$$

Proof. We use Lemma 3 in Kakade and Tewari [2009]. First, observe that variables Z_t are bounded, in particular, $|Z_t| \leq 2$. Furthermore,

$$\begin{aligned}
&\mathrm{var}[Z_t|\mathcal{F}_{t-1}] \\
&\quad = \mathrm{var}\left[\frac{Q_t}{p_t}\big(\ell(f(x_t), y_t) - \ell(g(x_t), y_t)\big)|\mathcal{F}_{t-1}\right] \\
&\quad \leq \mathop{\mathbb{E}}_{x_t, Q_t}\left[\frac{Q_t^2}{p_t^2}\big(\ell(f(x_t), y_t) - \ell(g(x_t), y_t)\big)^2|\mathcal{F}_{t-1}\right] \\
&\quad \leq \mathop{\mathbb{E}}_{x_t, Q_t}\left[\frac{Q_t\, p_t^2}{p_t^2}|\mathcal{F}_{t-1}\right] \\
&\quad = \mathop{\mathbb{E}}_{x_t, Q_t}\left[Q_t|\mathcal{F}_{t-1}\right] \\
&\quad = \mathop{\mathbb{E}}_{x_t}\left[p_t|\mathcal{F}_{t-1}\right].
\end{aligned}$$

A union bound over Z_t and $-Z_t$ concludes the proof. □

Given Lemma 4 above, we can adapt Lemma 3 of [Beygelzimer et al., 2009] to using the Berstein-like inequality. Specifically, let us define

$$\Delta_T = \frac{2}{T}\left(\sqrt{\sum_{t=1}^{T} p_t} + 6\sqrt{\log\left(\frac{(3+T)T^2}{\delta}\right)}\right) \times \sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)}.$$

Then we have the following high-probability statement for the risk of the hypothesis h_T returned by EIWAL after T rounds.

Lemma 5. *Given any hypothesis class $\mathcal{H}$, for all $\delta > 0$, for all $T \geq 3$ and all $f, g \in \mathtt{H}_T$, with probability at least $1 - 2\delta$,*

$$|\widehat{R}_T(f) - \widehat{R}_T(g) - R(f) + R(g)| \leq \Delta_T.$$

In particular, if we let $f = h^*$ and $g = h_T$, it follows that

$$R(h_T) \leq R(h^*) + \Delta_T.$$

Proof. Apply Lemma 4 to time $T \geq 3$ and any pair $f, g \in \mathtt{H}_T$, with error probability $\delta/(T^2|\mathcal{H}|^2)$ for round T. A union bound over $T \geq 3$ and (f, g) gives, with probability at least $1 - \delta$,

$$\begin{aligned}
&|\widehat{R}_T(f) - \widehat{R}_T(g) - R(f) + R(g)| \\
&\leq \frac{1}{T}\max\left\{2\sqrt{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{x_t}[p_t|\mathcal{F}_{t-1}]}, 6\sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)}\right\} \\
&\quad \times \sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)}. \qquad (8)
\end{aligned}$$

Next, according to Proposition 2 of Cesa-Bianchi and Gentile [2008], with probability at least $1 - \delta$, for all $T \geq 3$,

$$\begin{aligned}
&\sum_{t=1}^{T}\mathop{\mathbb{E}}_{x_t}\left[p_t|\mathcal{F}_{t-1}\right] \leq \left(\sum_{t=1}^{T} p_t\right) + 36\log\left(\frac{(3+\sum_{t=1}^{T} p_t)T^2}{\delta}\right) \\
&+ 2\sqrt{\left(\sum_{t=1}^{T} p_t\right)\log\left(\frac{(3+\sum_{t=1}^{T} p_t)T^2)}{\delta}\right)} \\
&\leq \left(\sqrt{\sum_{t=1}^{T} p_t} + 6\sqrt{\log\left(\frac{(3+T)T^2}{\delta}\right)}\right)^2. \qquad (9)
\end{aligned}$$

Combining (8) and (9), we get with probability at least $1 - 2\delta$, for all $T \geq 3$,

$$\begin{aligned}
&|\widehat{R}_T(f) - \widehat{R}_T(g) - R(f) + R(g)| \\
&\leq \frac{2}{T}\left(\sqrt{\sum_{t=1}^{T} p_t} + 6\sqrt{\log\left(\frac{(3+T)T^2}{\delta}\right)}\right) \\
&\quad \times \sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)},
\end{aligned}$$

as claimed. □

The next lemma gives a label complexity bound for EIWAL.

Lemma 6. *Given any hypothesis class $\mathcal{H}$, and distribution $\mathcal{D}$, with $\theta(\mathcal{D}, \mathcal{H}) = \theta$, for all $\delta > 0$, for all $T \geq 3$, with probably at least $1 - \delta$, we have*

$$\begin{aligned}
&\sum_{t=1}^{T}\mathop{\mathbb{E}}_{x_t, Q_t}[Q_t|\mathcal{F}_{t-1}] \\
&\quad \leq 4\theta K_l\left(R(h^*)T + O(\sqrt{R(h^*)T\log(T|\mathcal{H}|/\delta)})\right) \\
&\qquad + O(\log^3(T|\mathcal{H}|/\delta)),
\end{aligned}$$

where K_ℓ is a constant that depends on ℓ.

Proof. From Theorem 11 of Beygelzimer et al. [2009], for $t \geq 3$,

$$\mathop{\mathbb{E}}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right] \leq 4\theta K_l (R^* + \Delta_{t-1}), \tag{10}$$

where $R^* = R(h^*)$ is the risk of best-in-class. Plugging in the expression for Δ_{t-1}, and applying again a similar concentration inequality as before to relate $\sum_{t=1}^T p_t$ to $\sum_{t=1}^T \mathbb{E}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right]$, we end up with a recursion on $\mathbb{E}_{x_t}\left[p_t | \mathcal{F}_{t-1}\right]$:

$$\mathop{\mathbb{E}}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right] \leq 4\theta K_l R^* + \frac{4\theta K_l c_1}{t-1}\sqrt{\sum_{s=1}^{t-1} \mathop{\mathbb{E}}_{x_t}\left[p_s \middle| \mathcal{F}_{s-1}\right]} + c_2\left(\frac{\log\left[(t-1)|\mathcal{H}|/\delta\right]}{t-1}\right), \tag{11}$$

where $c_1 = 2\sqrt{\log\left(\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right)} = O\left(\sqrt{\log\left(\frac{T|\mathcal{H}|}{\delta}\right)}\right)$, and c_2 is a constant.

For simplicity, denote by $4\theta K_l = c_0$. We show by induction that for all $t \geq 3$,

$$\mathop{\mathbb{E}}_{x_t}\left[p_t | \mathcal{F}_{t-1}\right] \leq c_0 R^* + c_4\sqrt{\frac{R^*}{t-1}} + \frac{c5}{t-1}, \tag{12}$$

for some constants c_4, c_5. Assume by induction that (12) holds for all $s \leq t-1$. Thus, from (11), we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t}\left[p_t \middle| \mathcal{F}_{t-1}\right] \leq c_0 R^* \\
&+ \frac{c_0 c_1}{t-1}\sqrt{c_0 R^*(t-1) + 2c_4\sqrt{R^*(t-1)} + c_5\log(t-1)} \\
&+ c_2\left(\frac{\log\left[(t-1)|\mathcal{H}|/\delta\right]}{t-1}\right) \\
&\leq c_0 R^* + \frac{c_0 c_1}{t-1}\left[\sqrt{c_0 R^*(t-1) + 2c_4\sqrt{R^*(t-1)}}\right. \\
&\left. + \sqrt{c_5\log(t-1)}\right] + c_2\left(\frac{\log\left[(t-1)|\mathcal{H}|/\delta\right]}{t-1}\right) \\
&\leq c_0 R^* + \frac{c_0 c_1}{t-1}\left[\sqrt{c_0 R^*(t-1)} + \frac{c_4}{\sqrt{c_0}}\right] \\
&+ \frac{c_0 c_1\sqrt{c_5\log(t-1)} + c_2\log[(t-1)|\mathcal{H}|/\delta]}{t-1} \\
&= c_0 R^* + \frac{c_0 c_1\sqrt{c_0 R^*}}{\sqrt{t-1}} \\
&+ \frac{\sqrt{c_0}c_1 c_4 + c_0 c_1\sqrt{c_5\log(t-1)} + c_2\log[(t-1)|\mathcal{H}|/\delta]}{t-1},
\end{aligned}$$

where we use the fact that $\sqrt{a+b} \leq \sqrt{a} + \frac{b}{2\sqrt{a}}$ for $a, b > 0$.

To complete the induction, we need to show that

$$\begin{aligned}
&c_0 c_1\sqrt{c_0}\sqrt{\frac{R^*}{t-1}} \\
&+ \frac{\sqrt{c_0}c_1 c_4 + c_0 c_1\sqrt{c_5\log(t-1)} + c_2\log[(t-1)|\mathcal{H}|/\delta]}{t-1} \\
&\leq c_4\sqrt{\frac{R^*}{t-1}} + \frac{c5}{t-1}.
\end{aligned}$$

Thus, $c_4 = c_0 c_1\sqrt{c_0} = O(\sqrt{\log(T|\mathcal{H}|/\delta)})$, and

$$\begin{aligned}
&c_5 \geq c_0^2 c_1^2 + c_0 c_1\sqrt{c_5\log(t-1)} + c_2\log[(t-1)|\mathcal{H}|/\delta] \\
&\Rightarrow \sqrt{c_5} = O(c_0 c_1\sqrt{\log T}) \\
&\Rightarrow c_5 = O(c_0^2 c_1^2\log T) = O(\log^2(T|\mathcal{H}|/\delta)).
\end{aligned}$$

Thus,

$$\mathop{\mathbb{E}}_{x_t}\left[p_t | \mathcal{F}_{t-1}\right] \leq c_0 R^* + O(\sqrt{\log(T|\mathcal{H}|/\delta)})\sqrt{\frac{R^*}{(t-1)}} + \frac{O(\log^2(T|\mathcal{H}|/\delta))}{t-1}.$$

Finally,

$$\begin{aligned}
&\sum_{t=1}^T \mathop{\mathbb{E}}_{x_t, Q_t}[Q_t | \mathcal{F}_{t-1}] = \sum_{t=1}^T \mathop{\mathbb{E}}_{x_t}[p_t | \mathcal{F}_{t-1}] \\
&\leq 4\theta K_l[R(h^*)T + O(\sqrt{R(h^*)T\log(T|\mathcal{H}|/\delta)})] \\
&+ O(\log^3(T|\mathcal{H}|/\delta)).
\end{aligned} \tag{13}$$

□

Proof of Theorem 1. The bound of generalization error $R(h_T)$ follows from Lemma 5. To get the bound on the number of labels τ_T, we relate $\sum_{t=1}^T \mathbb{E}_{x_t, Q_t}[Q_t | \mathcal{F}_{t-1}]$ in Lemma 6 to $\tau_T = \sum_{t=1}^T Q_t$ through a Bernstein-like inequality for martingales. Again, from Lemma 3 of Kakade and Tewari [2009], we see that with probability at least $1-\delta$ we have

$$\begin{aligned}
&\sum_{t=1}^T Q_t - \sum_{t=1}^T \mathop{\mathbb{E}}_{x_t, Q_t}[Q_t | \mathcal{F}_{t-1}] \\
&\leq 2\sqrt{\left(\sum_{t=1}^T \mathrm{var}[Q_t | \mathcal{F}_{t-1}]\right)\log\left(\frac{4\log T}{\delta}\right)} \\
&+ 6\log\left(\frac{4\log T}{\delta}\right) \\
&\leq 2\sqrt{\left(\sum_{t=1}^T \mathop{\mathbb{E}}_{x_t, Q_t}[Q_t | \mathcal{F}_{t-1}]\right)\log\left(\frac{4\log T}{\delta}\right)} \\
&+ 6\log\left(\frac{4\log T}{\delta}\right) \\
&\leq \sum_{t=1}^T \mathop{\mathbb{E}}_{x_t, Q_t}[Q_t | \mathcal{F}_{t-1}] + 7\log\left(\frac{4\log T}{\delta}\right).
\end{aligned}$$

Combining with (13) completes the proof. □

B Proofs of ORIWAL

Proof of Theorem 2. We first expand the bound in Theorem 1 and get rid of $\sum_{t=1}^{T} p_t$. Lemma 6 states that with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_t,Q_t}[Q_t|\mathcal{F}_{t-1}] \\
&\quad \le 4\theta K_l \left(R(h^*)T + O(\sqrt{R(h^*)T\log(T|\mathcal{H}|/\delta)})\right) \\
&\qquad + O(\log^3(T|\mathcal{H}|/\delta))\,,
\end{aligned}$$

which implies that

$$\begin{aligned}
&\sqrt{\sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_t,Q_t}[Q_t|\mathcal{F}_{t-1}]} \\
&\le \sqrt{4\theta K_l \left(R(h^*)T + O(\sqrt{R(h^*)T\log(T|\mathcal{H}|/\delta)})\right)} \\
&+ O(\log^{\frac{3}{2}}(T|\mathcal{H}|/\delta)) \\
&\le \sqrt{4\theta K_l R(h^*)T} + O(\log^{\frac{1}{2}}(T|\mathcal{H}|/\delta)) \\
&+ O(\log^{\frac{3}{2}}(T|\mathcal{H}|/\delta)) \\
&\le \sqrt{4\theta K_l R(h^*)T} + O(\log^{\frac{3}{2}}(T|\mathcal{H}|/\delta)).
\end{aligned}$$

Thus, by Theorem 1, with probability at least $1-2\delta$,

$$\begin{aligned}
&R(h_T) \\
&\le R(h^*) + \frac{2}{T}\left[\sqrt{\textstyle\sum_{t=1}^{T} p_t} + 6\sqrt{\log\left[\frac{(3+T)T^2}{\delta}\right]}\right] \\
&\quad \times \sqrt{\log\left[\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right]} \\
&\le R(h^*) + \frac{2}{T}\left[\sqrt{4\theta K_l R(h^*)T} + O(\log^{\frac{3}{2}}(T|\mathcal{H}|/\delta))\right. \\
&\left. + 6\sqrt{\log\left[\frac{(3+T)T^2}{\delta}\right]}\right] \times \sqrt{\log\left[\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right]} \\
&= R(h^*) + 2\sqrt{\frac{4\theta K_l R(h^*)}{T}\log\left[\frac{8T^2|\mathcal{H}|^2\log(T)}{\delta}\right]} \\
&+ \frac{O(\log^2(T|\mathcal{H}|/\delta))}{T}.
\end{aligned}$$

Thus, for each region $\mathcal{X}_k$, with probability at least $1-\frac{\delta}{n}$, for any $T_k > 0$ the following holds:

$$\begin{aligned}
&R(h_{k,T}) \\
&\le R_k^* + 2\sqrt{\frac{4\theta_k K_l R_k^*}{T_k}\log\left[\frac{8T_k^2|\mathcal{H}_k|^2\log(T_k)2n}{\delta}\right]} \\
&+ \frac{O(\log^2(T_k|\mathcal{H}_k|n/\delta))}{T_k}.
\end{aligned}$$

Recall that

$$R(h_T) = \sum_{k=1}^{n} \mathtt{p}_k R_k(h_{k,T}), \qquad R(h^*) = \sum_{k=1}^{n} \mathtt{p}_k R_k^*\,.$$

A union bound over the n regions gives the result for $R(h_T)$:

$$\begin{aligned}
&R(h_T) \le R(h^*) \\
&+ \sum_{k=1}^{n} 2\mathtt{p}_k\sqrt{\frac{4\theta_k K_l R_k^*}{T_k}\log\left[\frac{8T_k^2|\mathcal{H}_k|^2\log(T_k)2n}{\delta}\right]} \\
&+ \Big(\sum_{k=1}^{n}\frac{\mathtt{p}_k}{T_k}\Big)O\Big(\log^2\big(\max_{k\in[n]} T_k|\mathcal{H}_k|n/\delta\big)\Big).
\end{aligned}$$

Furthermore, from Theorem 1 we have for each region $\mathcal{X}_k$, with probability at least $1-\frac{\delta}{n}$, for any $T_k > 0$:

$$\begin{aligned}
\tau_{k,T} &\le 8\theta_k K_l\left(R_k^* T_k + O(\sqrt{R_k^* T_k \log(T_k|\mathcal{H}_k|n/\delta)})\right) \\
&\quad + O\Big(\log^3\left(T_k|\mathcal{H}_k|n/\delta\right)\Big)\,,
\end{aligned}$$

where $\tau_{k,T}$ denotes the number of labels requested in region k up to time T. Again, a union bound over the n regions gives the result for $\tau_T = \sum_{k=1}^{T}\tau_{k,T}$. □

In order to prove Corollary 3 we need the following standard multiplicative Chernoff bounds.

Theorem 7 (Chernoff)**.** *Let $X_1,\cdots,X_m$ be independent random variables drawn according to some distribution $\mathcal{D}$ with mean p and support included in $[0,1]$. Then, for any $\gamma\in[0,\frac{1}{p}-1]$, the following holds for $\widehat{p}=\frac{1}{m}\sum_{i=1}^{m}X_i$:*

$$\begin{aligned}
\mathbb{P}[\widehat{p}\ge(1+\gamma)p] &\le e^{-\frac{mp\gamma^2}{3}}, \\
\mathbb{P}[\widehat{p}\le(1-\gamma)p] &\le e^{-\frac{mp\gamma^2}{2}}.
\end{aligned}$$

Proof of Corollary 3. Given a total of T queries over all regions, we have $\mathbb{E}[T_k] = T\mathtt{q}_k$, where

$$\mathtt{q}_k = \frac{\mathtt{p}_k\alpha_k}{\sum_{k'}^{n}\mathtt{p}_{k'}\alpha_{k'}}$$

is the probability of querying IWAL_k, conditioned on a query being made. By Theorem 7, with probability at least $1-\frac{\delta}{2}$, for all $k\in[n]$,

$$\frac{T_k}{T} \ge \mathtt{q}_k\Big(1-\sqrt{\frac{2\log(2n/\delta)}{T\mathtt{q}_k}}\Big).$$

It follows that with probability at least $1-\frac{\delta}{2}$, for all $k\in[n]$,

$$\frac{\mathtt{q}_k}{\sqrt{T_k}} = \sqrt{\frac{\mathtt{q}_k}{T}}\sqrt{\frac{\mathtt{q}_k}{(T_k/T)}} \le \sqrt{\frac{\mathtt{q}_k}{T}}\frac{1}{\sqrt{1-\sqrt{\frac{2\log(2n/\delta)}{T\mathtt{q}_k}}}}.$$

When $T \geq \frac{4\log(2n/\delta)}{\min_{k\in[n]} \mathtt{q}_k}$, we have $\frac{2\log(2n/\delta)}{T\mathtt{q}_k} < \frac{1}{2}$. Since $\frac{1}{\sqrt{1-\sqrt{x}}} \leq 1 + 2\sqrt{x}$ for any $x \leq \frac{1}{2}$, we can write

$$\begin{aligned}\frac{\mathtt{q}_k}{\sqrt{T_k}} &\leq \sqrt{\frac{\mathtt{q}_k}{T}}\left(1 + 2\sqrt{\frac{2\log(2n/\delta)}{T\mathtt{q}_k}}\right) \\ &= \sqrt{\frac{\mathtt{q}_k}{T}} + \frac{2\sqrt{2\log(2n/\delta)}}{T}.\end{aligned}$$

Plugging into Theorem 2, a union bound implies that with probability at least $1-\delta$,

$$\begin{aligned}&R(h_T) \leq R(h^*) \\ &+ 2\sum_{k=1}^{n} \mathtt{p}_k \sqrt{\frac{4\theta_k K_l R_k^*}{T_k} \log\left[\frac{8T^2|\mathcal{H}_k|^2\log(T)4n}{\delta}\right]} \\ &+ \Big(\sum_{k=1}^{n} \frac{\mathtt{p}_k}{T_k}\Big) O\Big(\log^2\big(\max_{k\in[n]} T|\mathcal{H}_k|n/\delta\big)\Big) \\ &\leq R(h^*) + 2\sum_{k=1}^{n} \mathtt{p}_k \sqrt{\frac{4\theta_k K_l R_k^*}{T\mathtt{q}_k} \log\left[\frac{8T^2|\mathcal{H}_k|^2\log(T)4n}{\delta}\right]} \\ &+ \Big(\sum_{k=1}^{n} \frac{\mathtt{p}_k}{T\mathtt{q}_k}\Big) O\Big(\log^2\big(\max_{k\in[n]} T|\mathcal{H}_k|n/\delta\big)\Big).\end{aligned}$$

Furthermore, by Chernoff bound, with probability at least $1-\delta$, for all $k \in [n]$,

$$\begin{aligned}&T_k \leq T\mathtt{q}_k + \sqrt{3T\mathtt{q}_k\log(n/\delta)} \\ \Rightarrow\ &\sqrt{T_k} \leq \sqrt{T\mathtt{q}_k} + \frac{\sqrt{3T\mathtt{q}_k\log(n/\delta)}}{2\sqrt{T\mathtt{q}_k}} \\ &\qquad (\text{ using the inequality } \sqrt{a+b} \leq \sqrt{a} + b/(2\sqrt{a}) \,) \\ &\qquad \leq \sqrt{T\mathtt{q}_k} + \sqrt{\log(n/\delta)}\,.\end{aligned}$$

Plugging into Theorem 2, with probability at least $1-2\delta$, for any $T > 0$,

$$\begin{aligned}\tau_T &\leq \sum_{k=1}^{n}\Big(8\theta_k K_l\Big[R_k^* T_k + O\Big(\sqrt{R_k^* T_k \log(T_k|\mathcal{H}_k|n/\delta)}\Big)\Big] \\ &\quad + O\Big(\log^3\big(T_k|\mathcal{H}_k|n/\delta\big)\Big)\Big) \\ &\leq 8K_\ell\Big[\sum_{k=1}^{n}\theta_k R_k^* T\mathtt{q}_k\Big] \\ &\quad + \sum_{k=1}^{n} O\Big(\sqrt{R_k^* T\mathtt{q}_k \log\Big[\frac{T|\mathcal{H}_k|n}{\delta}\Big]}\Big) \\ &\quad + O\Big(n\log^3\big(T\max_{k\in[n]}|\mathcal{H}_k|n/\delta\big)\Big)\Big)\,.\end{aligned}$$

This concludes the proof. □

C Two Natural Baselines for Region-Based Active Learning

In Section C.1 and Section C.2 below, we analyze two natural extensions of the IWAL algorithm to the region-based setting, called NAIVE-IWAL and RIWAL, that use the composite hypothesis set $\mathcal{H}_{[n]}$ in two different ways. In Section C.3, we then discuss the advantage of RIWAL over NAIVE-IWAL.

The two region-based baselines NAIVE-IWAL and RIWAL can use either IWAL or EIWAL as their underlying subroutines. To avoid clutter in the notation and to simplify the presentation, we proceed with the original version of IWAL, but a similar (though more involved) analysis can be carried out for the enhanced version EIWAL.

C.1 NAIVE-IWAL

NAIVE-IWAL consists of simply running the IWAL algorithm with the composite hypothesis set $\mathcal{H}_{[n]}$. This algorithm will find a model in this set without explicitly taking into account the structure of the set. Despite its simplicity, NAIVE-IWAL admits theoretical guarantees, since the guarantees from the classical IWAL (see Equation (2) and Equation (3)) directly apply. In particular, when $\mathcal{H}_k$s have the same number of hypotheses across k, the complexity terms in these bounds are multiplied by a factor of $\sqrt{n}$. This is because $|\mathcal{H}_{[n]}| = \prod_{k=1}^{n}|\mathcal{H}_k| = |\mathcal{H}_1|^n$. Thus, as the number of regions increases, the complexity term in the bound increases, while the generalization error of the best in class $R(h^*)$ decreases.

C.2 RIWAL

RIWAL consists of running n separate IWAL algorithms independently for each region. It works exactly in the same way as ORIWAL, except that it simply passes on all points to the subroutines, that is $\alpha_k = 1$ for all $k \in [n]$. Given T_k, which is the number of samples falling into region $\mathcal{X}_k$, RIWAL admits the same generalization error guarantees as that of ORIWAL (Theorem 2). Both results are derived from IWAL for a single region, along with a union bound over n regions. We can also apply a multiplicative Chernoff bound to the empirical quantities T_k to obtain a learning guarantee that only depends on T. The result is in fact a special case of Corollary 3, and is obtained by simply replacing therein $\mathtt{q}_k$ with $\mathtt{p}_k$.

C.3 Comparing NAIVE-IWAL and RIWAL

Even though NAIVE-IWAL and RIWAL learn from the same hypothesis set $\mathcal{H}_{[n]}$, and essentially use the same policy (the disagreement-based policy of IWAL) for requesting labels, the two algorithms are not equivalent. In fact, the two algorithms deliver final hypotheses with comparable generalization error after T rounds but, as we will show

momentarily, NAIVE-IWAL request more labels than RIWAL in expectation.

The following definitions will be useful. Let $\widehat{R}_{k,t}(h)$ and $\widehat{R}_t(h)$ denote the importance weighted empirical error of any hypothesis h after t rounds on region $\mathcal{X}_k$ and over all regions, respectively:

$$\widehat{R}_{k,t}(h) = \frac{\sum_{s=1}^{t} 1_{x_s \in \mathcal{X}_k} \frac{Q_s}{p_s} \ell(h(x_s), y_s)}{\sum_{s=1}^{t} 1_{x_s \in \mathcal{X}_k}},$$
$$\widehat{R}_t(h) = \frac{\sum_{s=1}^{t} \frac{Q_s}{p_s} \ell(h(x_s), y_s)}{t}.$$

Let $\widehat{h}_{k,t}$ and $\widehat{h}_t$ be the respective weighted empirical risk minimizers:

$$\widehat{h}_{k,t} = \operatorname*{argmin}_{h \in \mathcal{H}_k} \widehat{R}_{k,t}(h), \ \widehat{h}_t = \operatorname*{argmin}_{h \in \mathcal{H}_{[n]}} \widehat{R}_t(h).$$

Similar to Equation (1), we have $\widehat{h}_t = \sum_{k=1}^{n} 1_{x \in \mathcal{X}_k} \widehat{h}_{k,t}$.

Recall that for NAIVE-IWAL and RIWAL, the probability of requesting label y_t depends on the "disagreement" among their version spaces on x_t. A larger version space implies a larger disagreement value, and therefore a larger probability of requesting the label. Thus, at a high level, NAIVE-IWAL requests more labels than RIWAL because the version space of NAIVE-IWAL is larger than that of RIWAL. More precisely, assume for now that NAIVE-IWAL and RIWAL have been requesting the same labels up to time $t-1$, thus for any h and k, $\widehat{R}_{k,t}(h)$ has the same value under either algorithm, and the region-specific empirical risk minimizer is $\widehat{h}_{k,t}$. At time t, assume without loss of generality, that the unlabeled x_t lies in region $\mathcal{X}_1$. Given a slack term Δ, the version space is defined as the set of hypotheses whose importance weighted empirical error is Δ-close to the minimal empirical error. Assume there exists a hypothesis $h_1 \in \mathcal{H}_1$ such that

$$\Delta \leq \widehat{R}_{1,t}(h_1) - \widehat{R}_{1,t}(\widehat{h}_{1,t}) \leq \left[\frac{t}{\sum_{s=1}^{t} 1_{x_s \in \mathcal{X}_1}}\right] \Delta.$$

Since $\Delta \leq \widehat{R}_{1,t}(h_1) - \widehat{R}_{1,t}(\widehat{h}_{1,t})$, h_1 will not be included in the current version space of IWAL_1, which is the subroutine associated with $\mathcal{X}_1$ under the RIWAL algorithm. However, the version space of NAIVE-IWAL will include the hypothesis that takes the value of h_1 on region $\mathcal{X}_1$. To see why, let

$$h' = \sum_{k \in [n], k \neq 1} 1_{x \in \mathcal{X}_k} \widehat{h}_{kt} + 1_{x \in \mathcal{X}_1} h_1,$$

that is, the hypothesis that takes the value of the region-specific weighted empirical risk minimizers ($\widehat{h}_{k,t}$) on region $\mathcal{X}_k$, and takes the value of h_1 on region $\mathcal{X}_1$. Since

$$\widehat{R}(h') - \widehat{R}(\widehat{h}_t) = \left[\frac{\sum_{s=1}^{t} 1_{x_s \in \mathcal{X}_1}}{t}\right] \left(\widehat{R}_{1,t}(h_1) - \widehat{R}_{1,t}(\widehat{h}_{1,t})\right) \leq \Delta,$$

h' will be included in the version space of NAIVE-IWAL under the slack term Δ, even though h_1 is not included in the version space of RIWAL on region $\mathcal{X}_1$ under the same slack term. This suggests that NAIVE-IWAL is less efficient at shrinking the version space, and as a result it requests more labels.

We formalize this idea with Lemma 8 and Theorem 9. Lemma 8 relates the region-specific disagreement coefficients $\theta(\mathcal{D}_k, \mathcal{H}_k)$ to the overall disagreement coefficient $\theta(\mathcal{D}, \mathcal{H}_{[n]})$. Theorem 9 compares the learning guarantees of NAIVE-IWAL and RIWAL under certain assumptions.

Lemma 8. *The generalized disagreement coefficient $\theta(\mathcal{D}, \mathcal{H}_{[n]})$ satisfies $\theta(\mathcal{D}, \mathcal{H}_{[n]}) \leq \sum_{k=1}^{n} \theta(\mathcal{D}_k, \mathcal{H}_k)$.*

Proof. Denote $h^* = \operatorname{argmin}_{h \in \mathcal{H}_{[n]}} R(h)$, and $h_k^* = \operatorname{argmin}_{h \in \mathcal{H}_k} R_k(h)$. For simplicity, we denote by $\mathcal{D}_k = \mathcal{D}|\mathcal{X}_k$ the conditional distribution of x on $\mathcal{X}_k$. Recall that $h^* = \sum_{k=1}^{n} 1_{x \in X_k} h_k^*$. Extending the definitions in Section 3, we define

$$\rho_k(f, g) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \max_y |\ell(f(x), y) - \ell(g(x), y)|.$$

Given the hypothesis set $\mathcal{H}_k$ and any real $r > 0$, define

$$B_k(f, r) = \{g \in \mathcal{H}_k \colon \rho_k(f, g) \leq r\}.$$

For a set of non-negative values $\lambda = \{\lambda_1, \ldots, \lambda_n\}$, let

$$G_\lambda(h^*, r) = \left\{ \sum_{k=1}^{n} 1_{x \in X_k} g_k \colon g_k \in B_k(h_k^*, \lambda_k r) \right\}.$$

We first show that, for any λ satisfying $\sum_{k=1}^{n} \mathtt{p}_k \lambda_k \leq 1$, $G_\lambda(h^*, r) \subseteq B(h^*, r)$. Let $g = \sum_{k=1}^{n} 1_{x \in X_k} g_k$, where $g_k \in B_k(h_k^*, \lambda_k r)$. Then,

$$\begin{aligned}
&\rho(h^*, g) \\
&= \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \max_y |\ell(h^*(x), y) - \ell(g(x), y)| \\
&= \sum_{k=1}^{n} \mathtt{p}_k \mathop{\mathbb{E}}_{x \sim \mathcal{D}_k} \max_y |\ell(h_k^*(x), y) - \ell(g_k(x), y)| \\
&\leq \sum_{k=1}^{n} \mathtt{p}_k \lambda_k r \leq r.
\end{aligned}$$

Thus, $\left\{ \cup_{\lambda \colon \sum_{k=1}^{n} \mathtt{p}_k \lambda_k \leq 1} G_\lambda(h^*, r) \right\} \subseteq B(h^*, r)$. On the other hand, if there exits a hypothesis h such that

$$h \in B(h^*, r) \Big\backslash \left\{ \cup_{\lambda \colon \sum_{k=1}^{n} \mathtt{p}_k \lambda_k \leq 1} G_\lambda(h^*, r) \right\},$$

let $h = \sum_{k=1}^{n} 1_{x \in X_k} h_k$. Then,

$$\rho(h^*, h) = \sum_{k=1}^{n} \mathtt{p}_k \rho_k(h_k^*, h_k) \leq r \Rightarrow \sum_{k=1}^{n} \mathtt{p}_k \frac{\rho_k(h_k^*, h_k)}{r} \leq 1.$$

Obviously, $h_k \in B_k(h_k^*, \rho_k(h_k^*, h_k))$. Thus, let $\lambda = \{\frac{\rho_1(h_1^*,h_1)}{r}, \ldots, \frac{\rho_p(h_n^*,h_n)}{r}\}$, then $\sum_{k=1}^n \mathtt{p}_k \lambda_k \leq 1$, and $h \in G_\lambda(h^*, r)$ by definition. We have a contradiction. Therefore,

$$\left\{ \cup_{\lambda\colon \sum_{k=1}^n \mathtt{p}_k \lambda_k \leq 1} G_\lambda(h^*, r) \right\} = B(h^*, r) \,.$$

Given the equivalence above, for any $k \in [n]$,

$$\begin{aligned} \mathcal{H}_k \cap B(h^*, r) &= \mathcal{H}_k \cap \{\cup_{\lambda\colon \sum_{k=1}^n \mathtt{p}_k \lambda_k \leq 1} G_\lambda(h^*, r)\} \\ &= \mathcal{H}_k \cap \{\cup_{\lambda_k \leq 1/\mathtt{p}_k} B_k(h_k^*, \lambda_k r)\} \quad (14) \\ &= B_k(h_k^*, r/\mathtt{p}_k) \,. \quad (15) \end{aligned}$$

Equation (14) holds by the definition of $G_\lambda(h^*, r)$. Putting everything together, we have for any $r \geq 0$,

$$\begin{aligned} &\mathop{\mathbb{E}}_{x\sim D} \sup_{h\in B(h^*,r)} \sup_y |\ell(h(x), y) - \ell(h^*(x), y)| \\ &= \sum_{k=1}^n \mathtt{p}_k \mathop{\mathbb{E}}_{x\sim \mathcal{D}_k} \sup_{h\in B(h^*,r)} \sup_y |\ell(h(x), y) - \ell(h^*(x), y)| \\ &= \sum_{k=1}^n \mathtt{p}_k \mathop{\mathbb{E}}_{x\sim \mathcal{D}_k} \sup_{y, h_k \in B_k(h_k^*, \frac{r}{\mathtt{p}_k})} |\ell(h_k(x), y) - \ell(h_k^*(x), y)| \quad (16) \\ &\leq \sum_{k=1}^n \mathtt{p}_k \theta(\mathcal{D}_k, \mathcal{H}_k) r/\mathtt{p}_k \quad (17) \\ &= \Big(\sum_{k=1}^n \theta(\mathcal{D}|\mathcal{X}_k, \mathcal{H}_k)\Big) r. \end{aligned}$$

Equation (16) holds due to the equivalence in (15), and inequality (17) holds by the definition of $\theta(\mathcal{D}_k, \mathcal{H}_k)$.

Finally, recall the definition of $\theta(\mathcal{D}, \mathcal{H}_{[n]})$:

$$\begin{aligned} \theta(\mathcal{D}, \mathcal{H}) = \inf\Big\{ &\theta\colon \forall r \geq 0, \\ &\mathop{\mathbb{E}}_{x\sim \mathcal{D}} \sup_{h\in B(h^*,r)} \sup_y |\ell(h(x), y) - \ell(h^*(x), y)| \leq \theta r \Big\}. \end{aligned}$$

Therefore $\theta(\mathcal{D}, \mathcal{H}_{[n]}) \leq \sum_{k=1}^n \theta(\mathcal{D}_k, \mathcal{H}_k)$, which conclues the proof. □

In fact, one can show that there exist r, $\mathcal{D}$ and $\mathcal{H}_k$ such that equality is achieved in Lemma 8, thus the upper bound is tight.

Combining Lemma 8 with the learning guarantee of IWAL, we obtain the following result for the case when $|\mathcal{H}_k|$ is the same across all regions $\mathcal{X}_k$.

Theorem 9. *Assume $|\mathcal{H}_k|$ is the same across all regions $\mathcal{X}_k, k \in [n]$, and assume the same holds for $\theta(\mathcal{D}_k, \mathcal{H}_k)$. Then, the hypothesis returned by* NAIVE-IWAL *and* RIWAL *admit comparable generalization error guarantees, but on average* NAIVE-IWAL *would request up to n times more labels than* RIWAL.

Proof. Let $N = |\mathcal{H}_1|$, and $\theta_1 = \theta(\mathcal{D}_1, \mathcal{H}_1)$, so that $|\mathcal{H}_{[n]}| = N^n$ and, from Lemma 8, $\theta(\mathcal{D}, \mathcal{H}_{[n]}) \leq n\theta_1$. According to the learning guarantee of IWAL, with probability at least $1-\delta$, NAIVE-IWAL satisfies

$$R(h_T^{\text{NAIVE-IWAL}}) \leq R(h^*) + O\Big(\sqrt{\frac{\ln(TN^{2n}/\delta)}{T}}\Big), \quad (18)$$

$$\tau_T^{\text{NAIVE-IWAL}} \leq 4n\theta_1 K_\ell \Big[R(h^*)T + O(\sqrt{T\ln(TN^{2n}/\delta)})\Big]. \quad (19)$$

Meanwhile according to Theorem 2, with probability at least $1-\delta$, RIWAL satisfies

$$\begin{aligned} &R(h_T^{\text{RIWAL}}) \\ &\leq R(h^*) + \sum_{k=1}^n \mathtt{p}_k O\Big(\sqrt{\frac{\ln(T|N|^2 n/\delta)}{T_k}}\Big), \quad (20) \\ &\tau_T^{\text{RIWAL}} \\ &\leq \sum_{k=1}^n 4\theta_1 K_\ell \Big[R_k(h^*)T\mathtt{p}_k + O(\sqrt{2T\mathtt{p}_k \ln(2TN^2 n/\delta)})\Big] \\ &= 4\theta_1 K_\ell \Big[R(h^*)T + \sum_{k=1}^n O(\sqrt{2T\mathtt{p}_k \ln(2TN^2 n/\delta)})\Big]. \quad (21) \end{aligned}$$

Replacing T_k with $T\mathtt{p}_k + O(\sqrt{T})$ in the RHS of (20) we obtain

$$R(h_T^{\text{RIWAL}}) \leq R(h^*) + O\Big(\sqrt{\frac{n\ln(T|N|^2 n/\delta)}{T}}\Big). \quad (22)$$

Comparing the upper bound on the generalization error of RIWAL (22) to that of NAIVE-IWAL (18), we conclude that the two algorithms admit comparable learning guarantees.

On the other hand, comparing the proportion of labels requested per round, we have

$$\begin{aligned} \tau_T^{\text{NAIVE-IWAL}}/T &\leq 4n\theta_1 K_\ell R(h^*) + O\Big(\frac{1}{\sqrt{T}}\Big), \\ \tau_T^{\text{RIWAL}}/T &\leq 4\theta_1 K_\ell R(h^*) + O\Big(\frac{1}{\sqrt{T}}\Big). \end{aligned}$$

Thus, NAIVE-IWAL may request up to n times more labels than RIWAL.

□

D More Experimental Results

In this section, we provide results for all the datasets described in Table 1 in the main body of the paper.

Figures 3 show for 10 disjoint regions the misclassification error rate by three region-based algorithms, ORIWAL, RIWAL, and RPASSIVE, against number of labels requested (on

$\log_{10}$ scale), for all datasets. ORIWAL displays a consistent advantage over RIWAL and RPASSIVE.

Figures 4 and Figure 5 compares, for 10 and 20 disjoint regions respectively, the misclassification error rate of our algorithm, ORIWAL, to that of non region-based IWAL, and to non region-based passive learning PASSIVE. IWAL performs comparably to PASSIVE and stops improving early on, while ORIWAL significantly outperforms PASSIVE and continues to reduce the error rate while requesting more labels.

Figures 6 show the misclassification error rate by ORIWAL using 10 regions and 20 regions, respectively, against number of labels requested (on $\log_{10}$ scale), for all datasets. With randomly generated regions, it is unclear whether more regions would be helpful, as sometimes 20 regions admit higher misclassification error compared to 10 regions, given the same amount of requested labels. This observation leads to the following questions: How should the regions be chosen? How would the partitioning method affect the performance of ORIWAL? These are interesting directions for future work.

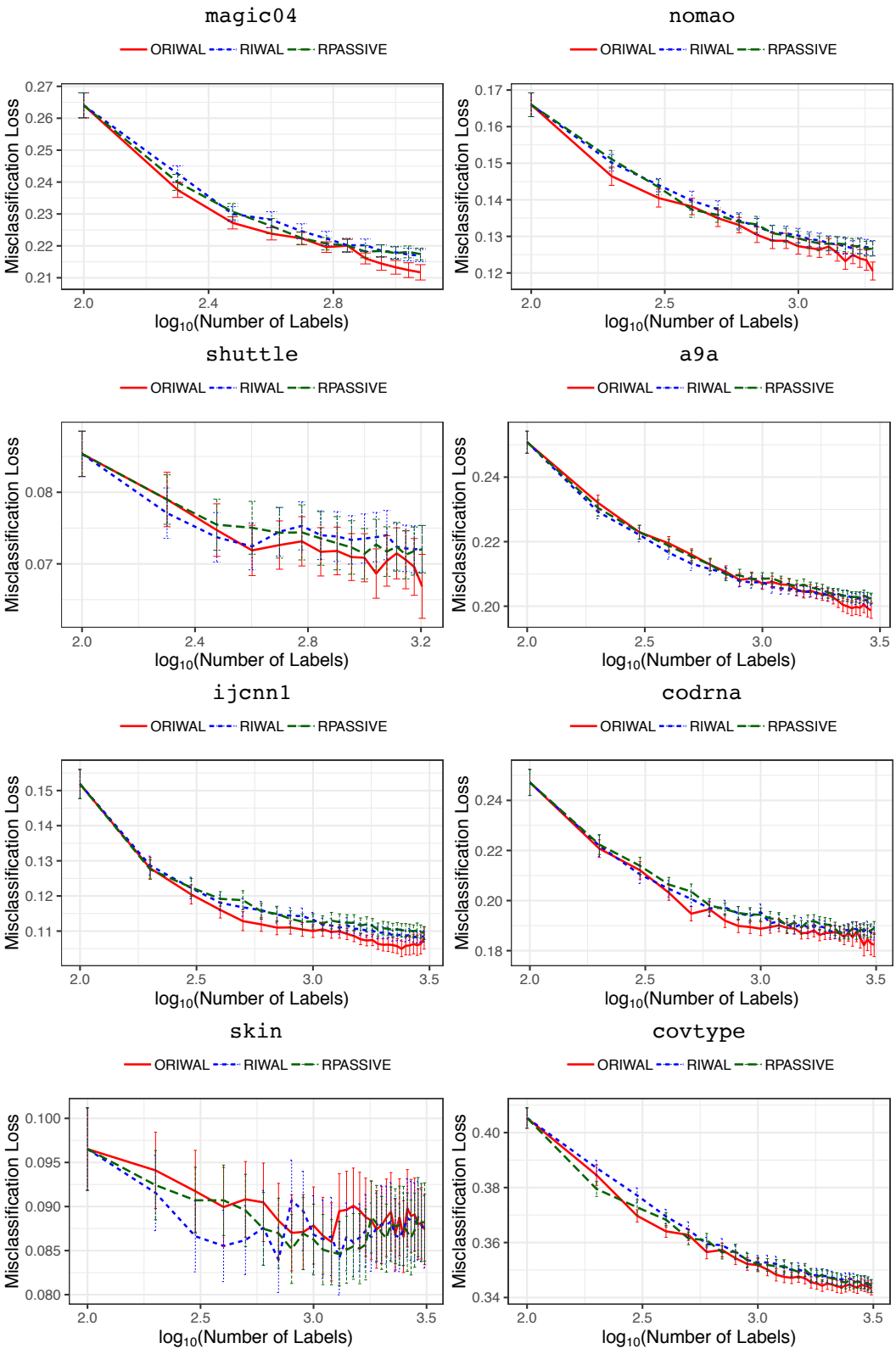

Figure 3: Misclassification loss of ORIWAL, RIWAL, and RPASSIVE on hold out test data vs. number of labels requested ($\log_{10}$ scale). The input space has **10** regions.

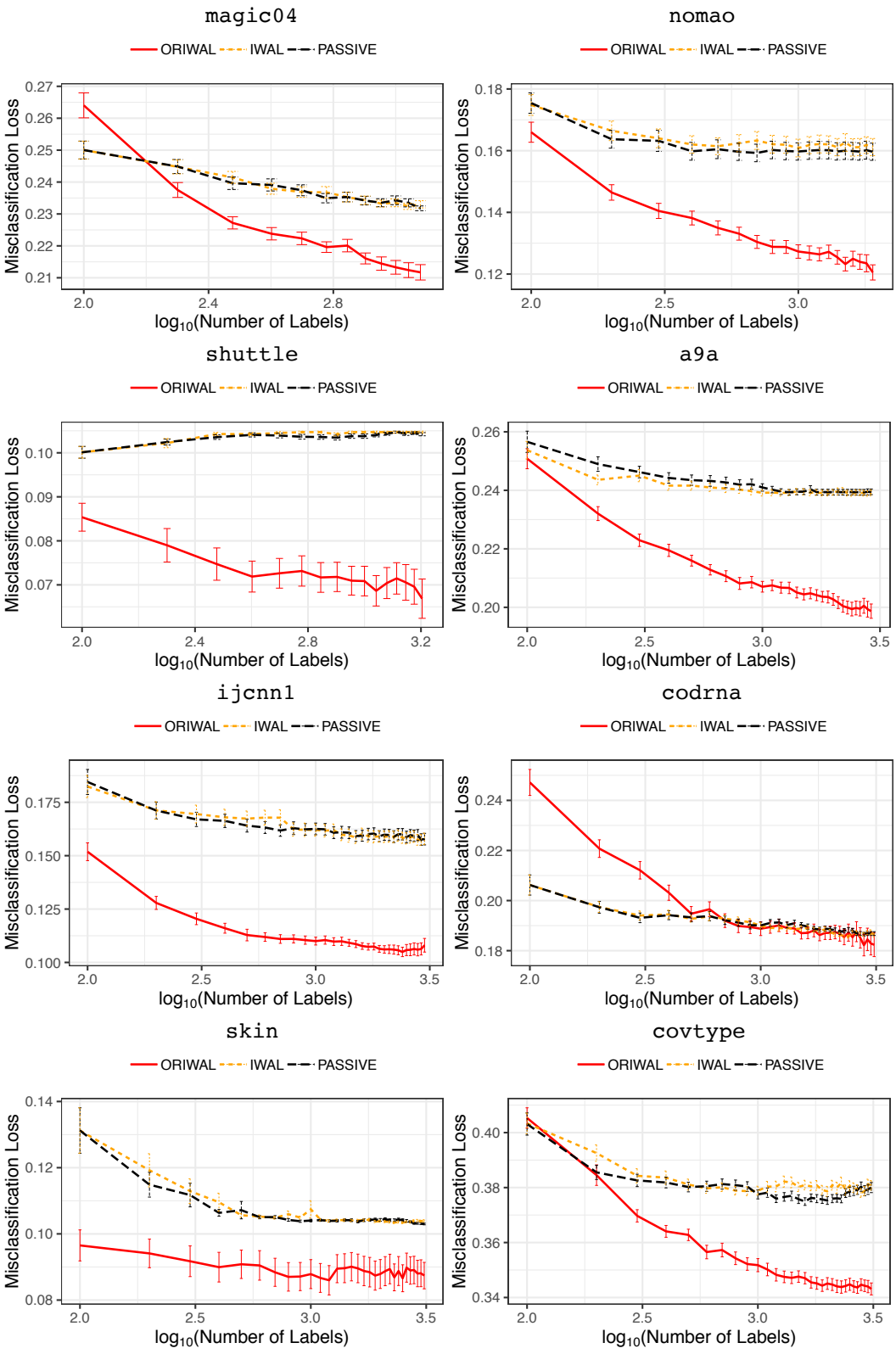

Figure 4: Misclassification loss of non region-based IWAL, non region-based passive learning PASSIVE, and ORIWAL (ours) on hold out test data vs. number of labels requested ($\log_{10}$ scale). The input space has **10** regions.

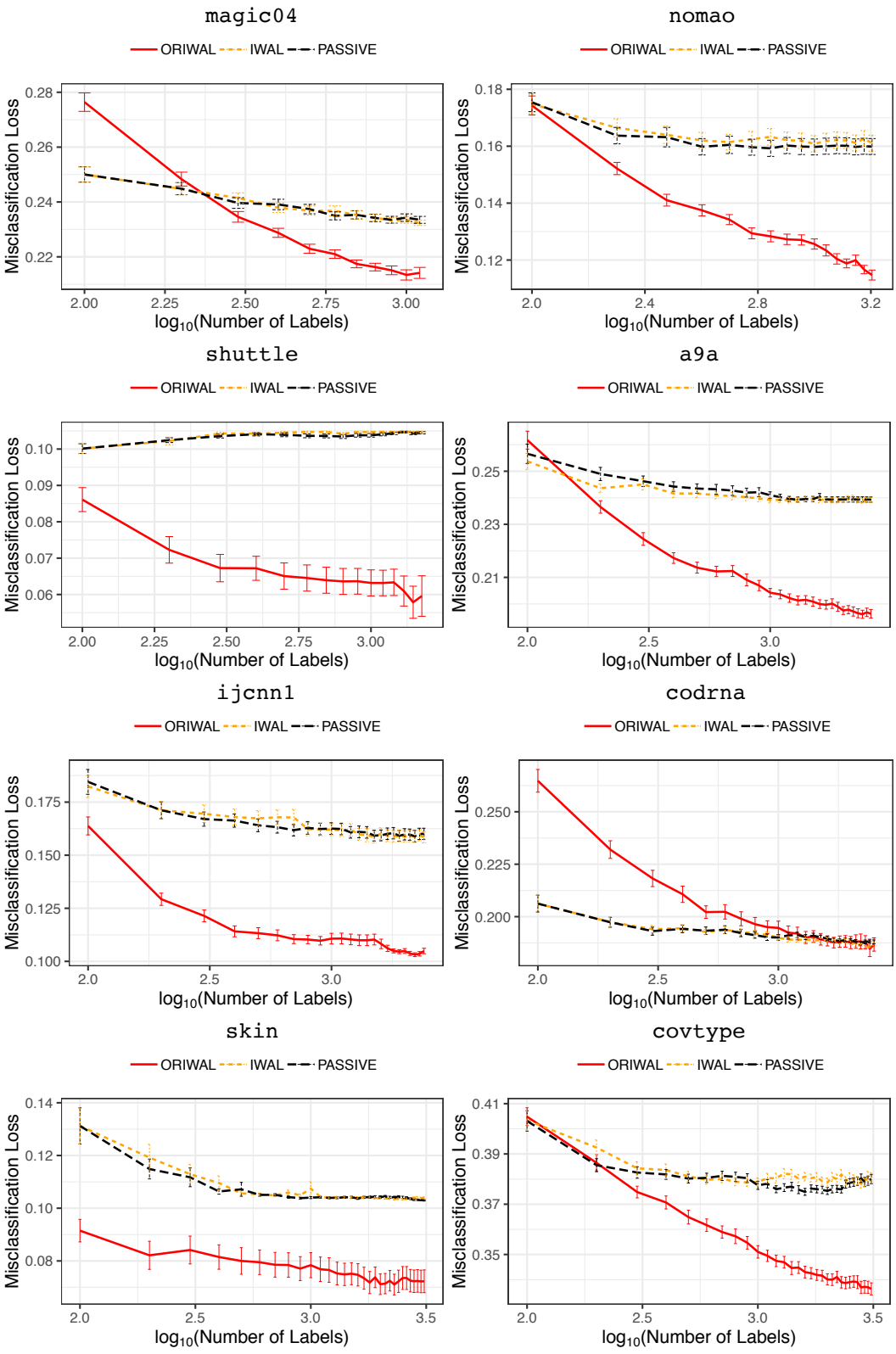

Figure 5: Misclassification loss of non region-based IWAL, non region-based passive learning PASSIVE, and ORIWAL (ours) on hold out test data vs. number of labels requested ($\log_{10}$ scale). The input space has **20** regions.

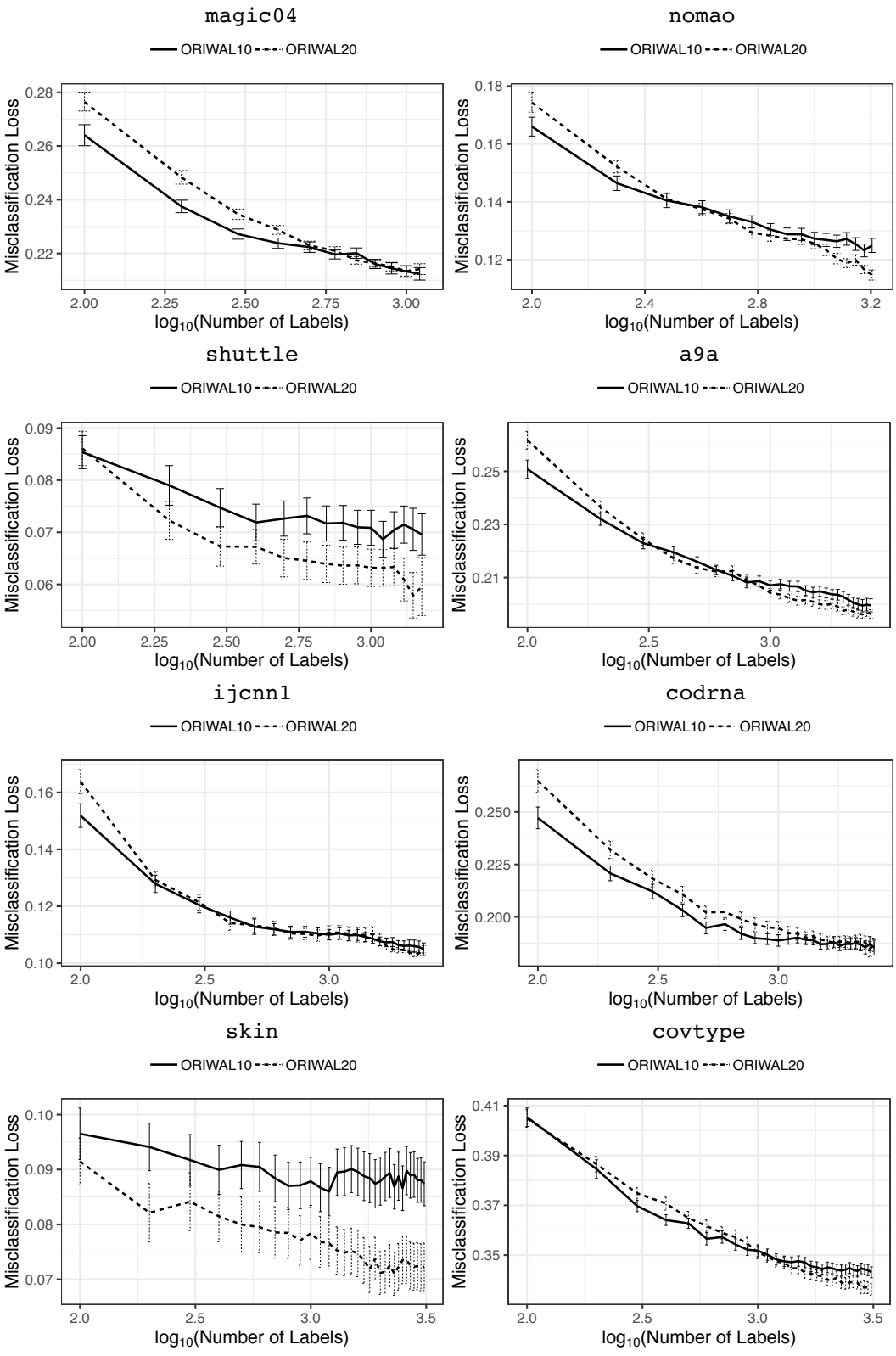

Figure 6: Misclassification loss of ORIWAL, using 10 regions, vs. 20 regions, on hold out test data vs. number of labels requested ($\log_{10}$ scale).