A. Preliminaries

A.1. Summary of Key Notations

Data Partitions $T_k = \{(X_t, Y_t, Z_t)\}_{t=m+n_1+\cdots+n_{k-1}+1}^{t=m+n_1+\cdots+n_k}$ ($1 \leq k \leq K$) is the online data collected in k-th iteration of size $n_k = 2^{k-1}$. $n = n_1 + \cdots + n_K$, $\alpha = 2m/3n$. We define $n_0 = 0$. $T_0 = \{(X_t, Y_t, Z_t)\}_{t=1}^{t=m}$ is the logged data and is partitioned into $K+1$ parts $T_0^{(0)}, \cdots, T_0^{(K)}$ of sizes $m_0 = m/3, m_1 = \alpha n_1, m_2 = \alpha n_2, \cdots, m_K = \alpha n_K$. $S_k = T_0^{(k)} \cup T_k$.

Recall that $\tilde{S}_k$ and $\tilde{T}_k$ contain inferred labels while S_k and T_k are sets of examples with original labels. The algorithm only observes $\tilde{S}_k$ and $\tilde{T}_k$.

For $(X, Z) \in T_k$ $(0 \leq k \leq K)$, $Q_k(X) = \Pr(Z = 1 \mid X)$.

Disagreement Regions The candidate set V_k and its disagreement region D_k are defined in Algorithm 1. $\hat{h}_k = \arg\min_{h \in V_k} l(h, \tilde{S}_k)$. $\nu = l(h^\star)$.

$B(h, r) := \{h' \in \mathcal{H} \mid \rho(h, h') \leq r\}$, $\text{DIS}(V) := \{x \in \mathcal{X} \mid \exists h_1 \neq h_2 \in V \text{ s.t. } h_1(x) \neq h_2(x)\}$. $S(A, \alpha) = \bigcup_{A' \subseteq A} \left(A' \cap \left\{x : Q_0(x) \leq \inf_{x \in A'} Q_0(x) + \frac{1}{\alpha}\right\}\right)$. $\tilde{\theta}(r_0, \alpha) = \sup_{r > r_0} \frac{1}{r} \Pr(S(\text{DIS}(B(h^\star, r)), \alpha))$.

$\text{DIS}_0 = \mathcal{X}$. For $k = 1, \ldots, K$, $\epsilon_k = \gamma_2 \sup_{x \in \text{DIS}_{k-1}} \frac{\log(2|\mathcal{H}|/\delta_k)}{m_{k-1} Q_0(x) + n_{k-1}} + \gamma_2 \sqrt{\sup_{x \in \text{DIS}_{k-1}} \frac{\log(2|\mathcal{H}|/\delta_k)}{m_{k-1} Q_0(x) + n_{k-1}} l(h^\star)}$, $\text{DIS}_k = \text{DIS}(B(h^\star, 2\nu + \epsilon_k))$.

Other Notations $\rho(h_1, h_2) = \Pr(h_1(X) \neq h_2(X))$, $\rho_S(h_1, h_2) = \frac{1}{|S|} \sum_{X \in S} \mathbb{1}\{h_1(X) \neq h_2(X)\}$.

For $k \geq 0$, $\sigma(k, \delta) = \sup_{x \in D_k} \frac{\log(|\mathcal{H}|/\delta)}{m_k Q_0(x) + n_k}$, $\delta_k = \frac{\delta}{(k+1)(k+2)}$. $\xi_k = \inf_{x \in D_k} Q_0(x)$. $\zeta = \sup_{x \in \text{DIS}_1} \frac{1}{\alpha Q_0(x) + 1}$.

A.2. Elementary Facts

Proposition 4. *Suppose $a, c \geq 0$,$b \in \mathbb{R}$. If $a \leq b + \sqrt{ca}$, then $a \leq 2b + c$.*

Proof. Since $a \leq b + \sqrt{ca}$, $\sqrt{a} \leq \frac{\sqrt{c} + \sqrt{c + 4b}}{2} \leq \sqrt{\frac{c + c + 4b}{2}} = \sqrt{c + 2b}$ where the second inequality follows from the Root-Mean Square-Arithmetic Mean inequality. Thus, $a \leq 2b + c$. □

A.3. Facts on Disagreement Regions and Candidate Sets

Lemma 5. *For any $k = 0, \ldots, K$, any $x \in \mathcal{X}$, any $h_1, h_2 \in V_k$, $\frac{\mathbb{1}\{h_1(x) \neq h_2(x)\}}{m_k Q_0(X) + n_k Q_k(X)} \leq \sup_{x'} \frac{\mathbb{1}\{x' \in D_k\}}{m_k Q_0(x') + n_k}$.*

Proof. The $k = 0$ case is obvious since $D_0 = \mathcal{X}$ and $n_0 = 0$.

For $k > 0$, since $\text{DIS}(V_k) = D_k$, $\mathbb{1}\{h_1(x) \neq h_2(x)\} \leq \mathbb{1}\{x \in D_k\}$, and thus $\frac{\mathbb{1}\{h_1(x) \neq h_2(x)\}}{m_k Q_0(X) + n_k Q_k(X)} \leq \frac{\mathbb{1}\{x \in D_k\}}{m_k Q_0(X) + n_k Q_k(X)}$.

For any x, if $Q_0(x) \leq \xi_k + 1/\alpha$, then $Q_k(x) = 1$, so $\frac{\mathbb{1}\{x \in D_k\}}{m_k Q_0(X) + n_k Q_k(X)} = \frac{\mathbb{1}\{x \in D_k\}}{m_k Q_0(x) + n_k} \leq \sup_{x'} \frac{\mathbb{1}\{x' \in D_k\}}{m_k Q_0(x') + n_k}$.

If $Q_0(x) > \xi_k + 1/\alpha$, then $Q_k(x) = 0$, so $\frac{\mathbb{1}\{x \in D_k\}}{m_k Q_0(X) + n_k Q_k(X)} = \frac{\mathbb{1}\{x \in D_k\}}{m_k Q_0(x)} \leq \frac{\mathbb{1}\{x \in D_k\}}{m_k \xi_k + n_k} \leq \sup_{x'} \frac{\mathbb{1}\{x' \in D_k\}}{m_k Q_0(x') + n_k}$ where the first inequality follows from the fact that $Q_0(x) > \xi_k + 1/\alpha$ implies $m_k Q_0(x) > m_k \xi_k + n_k$ □

Lemma 6. *For any $k = 0, \ldots, K$, if $h_1, h_2 \in V_k$, then $l(h_1, S_k) - l(h_2, S_k) = l(h_1, \tilde{S}_k) - l(h_2, \tilde{S}_k)$.*

Proof. For any $(X_t, Y_t, Z_t) \in S_t$ that $Z_t = 1$, if $X_t \in \text{DIS}(V_k)$, then $Y_t = \tilde{Y}_t$, so $\mathbb{1}\{h_1(X_t) \neq Y_t\} - \mathbb{1}\{h_2(X_t) \neq Y_t\} = \mathbb{1}\{h_1(X_t) \neq \tilde{Y}_t\} - \mathbb{1}\{h_2(X_t) \neq \tilde{Y}_t\}$. If $X_t \notin \text{DIS}(V_k)$, then $h_1(X_t) = h_2(X_t)$, so $\mathbb{1}\{h_1(X_t) \neq Y_t\} - \mathbb{1}\{h_2(X_t) \neq Y_t\} = \mathbb{1}\{h_1(X_t) \neq \tilde{Y}_t\} - \mathbb{1}\{h_2(X_t) \neq \tilde{Y}_t\} = 0$. □

The following lemma is immediate from definition.

Lemma 7. *For any $r \geq 2\nu$, any $\alpha \geq 1$, $\Pr(S(\mathit{DIS}(B(h^\star, r)), \alpha)) \leq r\tilde{\theta}(r, \alpha)$.*

A.4. Facts on Multiple Importance Sampling Estimators

We recall that $\{(X_t, Y_t)\}_{t=1}^{n_0+n}$ is an i.i.d. sequence. Moreover, the following fact is immediate by our construction that $S_0, \cdots, S_K$ are disjoint and that Q_k is determined by $S_0, \cdots, S_{k-1}$.

Fact 8. *For any $0 \leq k \leq K$, conditioned on Q_k, examples in S_k are independent, and examples in T_k are i.i.d.. Besides, for any $0 < k \leq K$, Q_k, $T_0^{(k)}, \ldots, T_0^{(K)}$ are independent.*

Unless otherwise specified, all probabilities and expectations are over the random draw of all random variables (including $S_0, \cdots, S_K, Q_1, \cdots, Q_K$).

The following lemma shows multiple importance estimators are unbiased.

Lemma 9. *For any $h \in \mathcal{H}$, any $0 \leq k \leq K$, $\mathbb{E}[l(h, S_k)] = l(h)$.*

The above lemma is immediate from the following lemma.

Lemma 10. *For any $h \in \mathcal{H}$, any $0 \leq k \leq K$, $\mathbb{E}[l(h, S_k) \mid Q_k] = l(h)$.*

Proof. The $k = 0$ case is obvious since $S_0 = T_0^{(0)}$ is an i.i.d. sequence and $l(h, S_k)$ reduces to a standard importance sampling estimator. We only show proof for $k > 0$.

Recall that $S_k = T_0^{(k)} \cup T_k$, and that $T_0^{(k)}$ and T_k are two i.i.d. sequences conditioned Q_k. We denote the conditional distributions of $T_0^{(k)}$ and T_k given Q_k by P_0 and P_k respectively. We have

$$\begin{aligned}\mathbb{E}[l(h, S_k) \mid Q_k] &= \mathbb{E}\left[\sum_{(X,Y,Z)\in T_0^{(k)}} \frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] + \mathbb{E}\left[\sum_{(X,Y,Z)\in T_k} \frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] \\ &= m_k \mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] + n_k \mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right]\end{aligned}$$

where the second equality follows since $T_0^{(k)}$ and T_k are two i.i.d. sequences given Q_k with sizes m_k and n_k respectively.

Now,

$$\begin{aligned}\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] &= \mathbb{E}_{P_0}\left[\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid X, Q_k\right] \mid Q_k\right] \\ &= \mathbb{E}_{P_0}\left[\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Q_0(X)}{m_k Q_0(X) + n_k Q_k(X)} \mid X, Q_k\right] \mid Q_k\right] \\ &= \mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Q_0(X)}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right]\end{aligned}$$

where the second equality uses the definition $\Pr_{P_0}(Z \mid X) = Q_0(X)$ and the fact that $T_0^{(k)}$ and Q_k are independent.

Similarly, we have $\mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X)\neq Y\}Z}{m_k Q_0(X)+n_k Q_k(X)} \mid Q_k\right] = \mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X)\neq Y\}Q_k(X)}{m_k Q_0(X)+n_k Q_k(X)} \mid Q_k\right]$.

Therefore,

$$\begin{aligned}& m_k \mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] + n_k \mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Z}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] \\ = \; & m_k \mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Q_0(X)}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] + n_k \mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h(X) \neq Y\}Q_k(X)}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] \\ = \; & \mathbb{E}_{P_0}\left[\mathbb{1}\{h(X) \neq Y\}\frac{m_k Q_0(X) + n_k Q_k(X)}{m_k Q_0(X) + n_k Q_k(X)} \mid Q_k\right] \\ = \; & \mathbb{E}_D\left[\mathbb{1}\{h(X) \neq Y\}\right] = l(h)\end{aligned}$$

where the second equality uses the fact that distribution of (X, Y) according to P_0 is the same as that according to P_k, and the third equality follows by algebra and Fact 8 that Q_k is independent with $T_0^{(k)}$. □

The following lemma will be used to upper-bound the variance of the multiple importance sampling estimator.

Lemma 11. *For any $h_1, h_2 \in \mathcal{H}$, any $0 \leq k \leq K$,*

$$\mathbb{E}\left[\sum_{(X,Y,Z)\in S_k}\left(\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{m_kQ_0(X)+n_kQ_k(X)}\right)^2 \mid Q_k\right] \leq \rho(h_1,h_2)\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}}{m_kQ_0(x)+n_kQ_k(x)}.$$

Proof. We only show proof for $k > 0$. The $k = 0$ case can be proved similarly.

We denote the conditional distributions of $T_0^{(k)}$ and T_k given Q_k by P_0 and P_k respectively. Now, similar to the proof of Lemma 10, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{(X,Y,Z)\in S_k}\left(\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{m_kQ_0(X)+n_kQ_k(X)}\right)^2 \mid Q_k\right]\\
=&\sum_{(X,Y,Z)\in S_k}\mathbb{E}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{(m_kQ_0(X)+n_kQ_k(X))^2} \mid Q_k\right]\\
=&m_k\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{(m_kQ_0(X)+n_kQ_k(X))^2} \mid Q_k\right]+n_k\mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Z}{(m_kQ_0(X)+n_kQ_k(X))^2} \mid Q_k\right]\\
=&m_k\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Q_0(X)}{(m_kQ_0(X)+n_kQ_k(X))^2} \mid Q_k\right]+n_k\mathbb{E}_{P_k}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}Q_k(X)}{(m_kQ_0(X)+n_kQ_k(X))^2} \mid Q_k\right]\\
=&\mathbb{E}_{P_0}\left[\mathbb{1}\{h_1(X)\neq h_2(X)\}\frac{m_kQ_0(X)+n_kQ_k(X)}{(m_kQ_0(X)+n_kQ_k(X))^2} \mid Q_k\right]\\
=&\mathbb{E}_{P_0}\left[\frac{\mathbb{1}\{h_1(X)\neq h_2(X)\}}{m_kQ_0(X)+n_kQ_k(X)} \mid Q_k\right]\\
\leq&\mathbb{E}_{P_0}\left[\mathbb{1}\{h_1(X)\neq h_2(X)\} \mid Q_k\right]\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}}{m_kQ_0(x)+n_kQ_k(x)}\\
=&\rho(h_1,h_2)\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}}{m_kQ_0(x)+n_kQ_k(x)}.
\end{aligned}$$

□

B. Deviation Bounds

In this section, we demonstrate deviation bounds for our error estimators on S_k. Again, unless otherwise specified, all probabilities and expectations in this section are over the random draw of all random variables, that is, $S_0, \cdots, S_K$, $Q_1, \cdots, Q_K$.

We use following Bernstein-style concentration bound:

Fact 12. *Suppose $X_1, \ldots, X_n$ are independent random variables. For any $i = 1, \ldots, n$, $|X_i| \leq 1$, $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 \leq \sigma_i^2$. Then with probability at least $1-\delta$,*

$$\left|\sum_{i=1}^n X_i\right| \leq \frac{2}{3}\log\frac{2}{\delta}+\sqrt{2\sum_{i=1}^n\sigma_i^2\log\frac{2}{\delta}}.$$

Theorem 13. *For any $k = 0, \ldots, K$, any $\delta > 0$, with probability at least $1-\delta$, for all $h_1, h_2 \in \mathcal{H}$, the following statement holds:*

$$|(l(h_1,S_k)-l(h_2,S_k))-(l(h_1)-l(h_2))| \leq 2\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}\frac{2\log\frac{4|\mathcal{H}|}{\delta}}{3}}{m_kQ_0(x)+n_kQ_k(x)}+\sqrt{2\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}\log\frac{4|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)}\rho(h_1,h_2)}. \tag{4}$$

Proof. We show proof for $k > 0$. The $k = 0$ case can be proved similarly. When $k > 0$, it suffices to show that for any $k = 1, \dots, K$, $\delta > 0$, conditioned on Q_k, with probability at least $1 - \delta$, (4) holds for all $h_1, h_2 \in \mathcal{H}$.

For any $k = 1, \dots, K$, for any fixed $h_1, h_2 \in \mathcal{H}$, define $A := \sup_{x\in\mathcal{X}} \frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}}{m_kQ_0(x)+n_kQ_k(x)}$. Let $N := |S_k|$, $U_t := \frac{\mathbb{1}\{h_1(X_t)\neq Y_t\}Z_t}{m_kQ_0(X_t)+n_kQ_k(X_t)} - \frac{\mathbb{1}\{h_2(X_t)\neq Y_t\}Z_t}{m_kQ_0(X_t)+n_kQ_k(X_t)}$, $V_t := (U_t - \mathbb{E}[U_t|Q_k])/2A$.

Now, conditioned on Q_k, $\{V_t\}_{t=1}^N$ is an independent sequence by Fact 8. $|V_t| \leq 1$, and $\mathbb{E}[V_t|Q_k] = 0$. Besides, we have

$$
\begin{aligned}
\sum_{t=1}^{N} \mathbb{E}[V_t^2|Q_k] &\leq \frac{1}{4A^2}\sum_{t=1}^{N}\mathbb{E}[U_t^2|Q_k] \\
&\leq \frac{1}{4A^2}\sum_{t=1}^{N}\mathbb{E}\left(\frac{\mathbb{1}\{h_1(X_t)\neq h_2(X_t)\}Z_t}{m_kQ_0(X_t)+n_kQ_k(X_t)}\right)^2 \\
&\leq \frac{\rho(h_1,h_2)}{4A}
\end{aligned}
$$

where the second inequality follows from $|U_t| \leq \frac{\mathbb{1}\{h_1(X_t)\neq h_2(X_t)\}Z_t}{m_kQ_0(X_t)+n_kQ_k(X_t)}$, and the third inequality follows from Lemma 11.

Applying Bernstein's inequality (Fact 12) to $\{V_t\}$, conditioned on Q_k, we have with probability at least $1 - \delta$,

$$
\left|\sum_{t=1}^{m} V_t\right| \leq \frac{2}{3}\log\frac{2}{\delta} + \sqrt{\frac{\rho(h_1,h_2)}{2A}\log\frac{2}{\delta}}.
$$

Note that $\sum_{t=1}^m U_t = l(h_1, S_k) - l(h_2, S_k)$, and $\sum_{t=1}^m \mathbb{E}[U_t \mid Q_k] = l(h_1) - l(h_2)$ by Lemma 10, so $\sum_{t=1}^m V_t = \frac{1}{2A}(l(h_1,S_k) - l(h_2,S_k) - l(h_1) + l(h_2))$. (4) follows by algebra and a union bound over $\mathcal{H}$. □

Theorem 14. *For any* $k = 0, \dots, K$*, any* $\delta > 0$*, with probability at least* $1 - \delta$*, for all* $h_1, h_2 \in \mathcal{H}$*, the following statements hold simultaneously:*

$$
\rho_{S_k}(h_1,h_2) \leq 2\rho(h_1,h_2) + \frac{10}{3}\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}\log\frac{4|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)}; \tag{5}
$$

$$
\rho(h_1,h_2) \leq 2\rho_{S_k}(h_1,h_2) + \frac{7}{6}\sup_{x\in\mathcal{X}}\frac{\mathbb{1}\{h_1(x)\neq h_2(x)\}\log\frac{4|\mathcal{H}|}{\delta}}{m_kQ_0(x)+n_kQ_k(x)}. \tag{6}
$$

Proof. Let $N = |S_k|$. Note that for any $h_1, h_2 \in \mathcal{H}$, $\rho_{S_k}(h_1,h_2) = \frac{1}{N}\sum_t \mathbb{1}\{h_1(X_t)\neq h_2(X_t)\}$, which is the empirical average of an i.i.d. sequence. By Fact 12 and a union bound over $\mathcal{H}$, with probability at least $1 - \delta$,

$$
|\rho(h_1,h_2) - \rho_{S_k}(h_1,h_2)| \leq \frac{2}{3N}\log\frac{4|\mathcal{H}|}{\delta} + \sqrt{\frac{2\rho(h_1,h_2)}{N}\log\frac{4|\mathcal{H}|}{\delta}}.
$$

On this event, by Proposition 4, $\rho(h_1,h_2) \leq 2\rho_{S_k}(h_1,h_2) + \frac{4}{3N}\log\frac{4|\mathcal{H}|}{\delta} + \frac{2}{N}\log\frac{4|\mathcal{H}|}{\delta} \leq 2\rho_{S_k}(h_1,h_2) + \frac{10}{3N}\log\frac{4|\mathcal{H}|}{\delta}$.

Moreover,

$$
\begin{aligned}
\rho_{S_k}(h_1,h_2) &\leq \rho(h_1,h_2) + \frac{2}{3N}\log\frac{4|\mathcal{H}|}{\delta} + \sqrt{\frac{2\rho(h_1,h_2)}{N}\log\frac{4|\mathcal{H}|}{\delta}} \\
&\leq \rho(h_1,h_2) + \frac{2}{3N}\log\frac{4|\mathcal{H}|}{\delta} + \frac{1}{2}(2\rho(h_1,h_2) + \frac{1}{N}\log\frac{4|\mathcal{H}|}{\delta}) \\
&\leq 2\rho(h_1,h_2) + \frac{7}{6N}\log\frac{4|\mathcal{H}|}{\delta}
\end{aligned}
$$

where the second inequality uses the fact that $\forall a, b > 0$, $\sqrt{ab} \leq \frac{a+b}{2}$.

The result follows by noting that $\forall x \in \mathcal{X}$, $N = |S_k| = m_k + n_k \geq m_kQ_0(x) + n_kQ_k(x)$. □

Corollary 15. *There are universal constants $\gamma_0, \gamma_1 > 0$ such that for any $k = 0, \dots, K$, any $\delta > 0$, with probability at least $1 - \delta$, for all $h, h_1, h_2 \in \mathcal{H}$, the following statements hold simultaneously:*

$$|(l(h_1, S_k) - l(h_2, S_k)) - (l(h_1) - l(h_2))| \leq \gamma_0 \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h_1(x) \neq h_2(x)\} \log \frac{|\mathcal{H}|}{2\delta}}{m_k Q_0(x) + n_k Q_k(x)} + \gamma_0 \sqrt{\sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h_1(x) \neq h_2(x)\} \log \frac{|\mathcal{H}|}{2\delta}}{m_k Q_0(x) + n_k Q_k(x)} \rho_S(h_1, h_2)}; \tag{7}$$

$$l(h) - l(h^\star) \leq 2(l(h, S_k) - l(h^\star, S_k)) + \gamma_1 \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \log \frac{|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} + \gamma_1 \sqrt{\sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \log \frac{|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} l(h^\star)}. \tag{8}$$

Proof. Let event E be the event that (4) and (6) holds for all $h_1, h_2 \in \mathcal{H}$ with confidence $1 - \frac{\delta}{2}$ respectively. Assume E happens (whose probability is at least $1 - \delta$).

(7) is immediate from (4) and (6).

For the proof of (8), apply (4) to h and $h^\star$, we get

$$l(h) - l(h^\star) \leq l(h, S_k) - l(h^\star, S_k) + 2 \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \frac{2 \log \frac{4|\mathcal{H}|}{\delta}}{3}}{m_k Q_0(x) + n_k Q_k(x)} + \sqrt{2 \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \log \frac{4|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} \rho(h, h^\star)}.$$

By triangle inequality, $\rho(h, h^\star) = \Pr_D(h(X) \neq h^\star(X)) \leq \Pr_D(h(X) \neq Y) + \Pr_D(h^\star(X) \neq Y) = l(h) - l(h^\star) + 2l(h^\star)$. Therefore, we get

$$\begin{aligned} l(h) - l(h^\star) \quad &\leq \quad l(h, S_k) - l(h^\star, S_k) + 2 \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \frac{2 \log \frac{4|\mathcal{H}|}{\delta}}{3}}{m_k Q_0(x) + n_k Q_k(x)} \\ &\quad + \sqrt{2 \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x))\} \log \frac{4|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} (l(h) - l(h^\star) + 2l(h^\star))} \\ &\leq \quad l(h, S_k) - l(h^\star, S_k) + \sqrt{2 \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \log \frac{4|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} (l(h) - l(h^\star))} \\ &\quad + 2 \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \frac{2 \log \frac{4|\mathcal{H}|}{\delta}}{3}}{m_k Q_0(x) + n_k Q_k(x)} + \sqrt{4 \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{h(x) \neq h^\star(x)\} \log \frac{4|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} l(h^\star)} \end{aligned}$$

where the second inequality uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$.

(8) follows by applying Proposition 4 to $l(h) - l(h^\star)$. □

C. Technical Lemmas

For any $0 \leq k \leq K$ and $\delta > 0$, define event $\mathcal{E}_{k,\delta}$ to be the event that the conclusions of Theorem 13 and Theorem 14 hold for k with confidence $1 - \delta/2$ respectively. We have $\Pr(\mathcal{E}_{k,\delta}) \geq 1 - \delta$, and that $\mathcal{E}_{k,\delta}$ implies inequalities (4) to (8).

We first present a lemma which can be used to guarantee that $h^\star$ stays in candidate sets with high probability by induction..

Lemma 16. *For any $k = 0, \dots K$, any $\delta > 0$. On event $\mathcal{E}_{k,\delta}$, if $h^\star \in V_k$ then,*

$$l(h^\star, \tilde{S}_k) \leq l(\hat{h}_k, \tilde{S}_k) + \gamma_0 \sigma(k, \delta) + \gamma_0 \sqrt{\sigma(k, \delta) \rho_{\tilde{S}_k}(\hat{h}_k, h^\star)}.$$

Proof.

$$\begin{aligned}
&l(h^\star, \tilde{S}_k) - l(\hat{h}_k, \tilde{S}_k)\\
=&l(h^\star, S_k) - l(\hat{h}_k, S_k)\\
\le&\gamma_0 \sup_x \frac{\mathbb{1}\{h^\star(x) \neq \hat{h}_k(x)\} \log \frac{|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} + \gamma_0 \sqrt{\sup_x \frac{\mathbb{1}\{h^\star(x) \neq \hat{h}_k(x)\} \log \frac{|\mathcal{H}|}{\delta}}{m_k Q_0(x) + n_k Q_k(x)} \rho_{S_k}(\hat{h}_k, h^\star)}\\
\le&\gamma_0 \sigma(k,\delta) + \sqrt{\gamma_0 \sigma(k,\delta) \rho_{\tilde{S}_k}(\hat{h}_k, h)}.
\end{aligned}$$

The equality follows from Lemma 6. The first inequality follows from (7) of Corollary 15 and that $l(h^\star) \le l(\hat{h}_k)$. The last inequality follows from Lemma 5 and that $\rho_{\tilde{S}_k}(\hat{h}_k, h^\star) = \rho_{S_k}(\hat{h}_k, h^\star)$. □

Next, we present two lemmas to bound the probability mass of the disagreement region of candidate sets.

Lemma 17. *For any $k = 0, \dots, K$, any $\delta > 0$, let $V_{k+1}(\delta) := \{h \in V_k \mid l(h, \tilde{S}_k) \le l(\hat{h}_k, \tilde{S}_k) + \gamma_0 \sigma(k,\delta) + \gamma_0 \sqrt{\sigma(k,\delta) \rho_{\tilde{S}_k}(\hat{h}_k, h)}\}$. Then there is an absolute constant $\gamma_2 > 1$ such that for any $0, \dots, K$, any $\delta > 0$, on event $\mathcal{E}_{k,\delta}$, if $h^\star \in V_k$, then for all $h \in V_{k+1}(\delta)$,*

$$l(h) - l(h^\star) \le \gamma_2 \sigma(k,\delta) + \gamma_2 \sqrt{\sigma(k,\delta) l(h^\star)}.$$

Proof. For any $h \in V_{k+1}(\delta)$, we have

$$\begin{aligned}
&l(h) - l(h^\star)\\
&\le 2(l(h, S_k) - l(h^\star, S_k)) + \gamma_1 \sigma(k, \frac{\delta}{2}) + \gamma_1 \sqrt{\sigma(k, \frac{\delta}{2}) l(h^\star)}\\
&= 2(l(h, \tilde{S}_k) - l(h^\star, \tilde{S}_k)) + \gamma_1 \sigma(k, \frac{\delta}{2}) + \gamma_1 \sqrt{\sigma(k, \frac{\delta}{2}) l(h^\star)}\\
&= 2(l(h, \tilde{S}_k) - l(\hat{h}_k, \tilde{S}_k) + l(\hat{h}_k, \tilde{S}_k) - l(h^\star, \tilde{S}_k)) + \gamma_1 \sigma(k, \frac{\delta}{2}) + \gamma_1 \sqrt{\sigma(k, \frac{\delta}{2}) l(h^\star)}\\
&\le 2(l(h, \tilde{S}_k) - l(\hat{h}_k, \tilde{S}_k)) + \gamma_1 \sigma(k, \frac{\delta}{2}) + \gamma_1 \sqrt{\sigma(k, \frac{\delta}{2}) l(h^\star)}\\
&\le (2\gamma_0 + \gamma_1) \sigma(k, \frac{\delta}{2}) + 2\gamma_0 \sqrt{\sigma(k, \frac{\delta}{2}) \rho_{\tilde{S}_k}(h, \hat{h}_k)} + \gamma_1 \sqrt{\sigma(k, \frac{\delta}{2}) l(h^\star)}\\
&\le (2\gamma_0 + \gamma_1) \sigma(k, \frac{\delta}{2}) + 2\gamma_0 \sqrt{\sigma(k, \frac{\delta}{2})(\rho_{S_k}(h, h^\star) + \rho_{S_k}(\hat{h_k}, h^\star))} + \gamma_1 \sqrt{\sigma(k, \frac{\delta}{2}) l(h^\star)} \qquad (9)
\end{aligned}$$

where the first inequality follows from (8) of Corollary 15 and Lemma 5, the first equality follows from Lemma 6, the third inequality follows from the definition of $V_k(\delta)$, and the last inequality follows from $\rho_{\tilde{S}_k}(h, \hat{h}_k) = \rho_{S_k}(h, \hat{h}_k) \le \rho_{S_k}(h, h^\star) + \rho_{S_k}(\hat{h_k}, h^\star)$.

As for $\rho_{S_k}(h, h^\star)$, we have $\rho_{S_k}(h, h^\star) \le 2\rho(h, h^\star) + \frac{16}{3}\sigma(k, \frac{\delta}{8}) \le 2(l(h) - l(h^\star)) + 4l(h^\star) + \frac{16}{3}\sigma(k, \frac{\delta}{8})$ where the first inequality follows from (5) of Theorem 14 and Lemma 5, and the second inequality follows from the triangle inequality.

For $\rho_{S_k}(\hat{h}_k, h^\star)$, we have

$$\begin{aligned}
\rho_{S_k}(\hat{h}_k, h^\star) &\leq 2\rho(\hat{h}_k, h^\star) + \frac{16}{3}\sigma(k, \frac{\delta}{8}) \\
&\leq 2(l(\hat{h}_k) - l(h^\star) + 2l(h^\star)) + \frac{16}{3}\sigma(k, \frac{\delta}{8}) \\
&\leq 2(2(l(\hat{h}_k, S_k) - l(h^\star, S_k)) + \gamma_1\sigma(k, \frac{\delta}{2}) + \gamma_1\sqrt{\sigma(k, \frac{\delta}{2})l(h^\star)} + 2l(h^\star)) + \frac{16}{3}\sigma(k, \frac{\delta}{8}) \\
&\leq (2\gamma_1 + \frac{16}{3})\sigma(k, \frac{\delta}{8}) + 2\gamma_1\sqrt{\sigma(k, \frac{\delta}{2})l(h^\star)} + 4l(h^\star) \\
&\leq (4 + \gamma_1)l(h^\star) + (3\gamma_1 + \frac{16}{3})\sigma(k, \frac{\delta}{8})
\end{aligned}$$

where the first inequality follows from (5) of Theorem 14 and Lemma 5, the second follows from the triangle inequality, the third follows from (8) of Theorem 15 and Lemma 5, the fourth follows from the definition of $\hat{h}_k$, the last follows from the fact that $2\sqrt{ab} \leq a + b$ for $a, b \geq 0$.

Continuing (9) and using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we have:

$$l(h) - l(h^\star) \leq (2\gamma_0 + \gamma_1 + 2\gamma_0\sqrt{3\gamma_1 + \frac{32}{3}})\sigma(k, \frac{\delta}{8}) + (2\gamma_0\sqrt{8 + \gamma_1} + \gamma_1)\sqrt{\sigma(k, \frac{\delta}{8})l(h^\star)} + 2\sqrt{2}\gamma_0\sqrt{\sigma(k, \frac{\delta}{8})(l(h) - l(h^\star))}.$$

The result follows by applying Proposition 4 to $l(h) - l(h^\star)$. □

Lemma 18. *On event $\bigcap_{k=0}^{K-1} \mathcal{E}_{k,\delta_k/2}$, for any $k = 0, \dots K$, $D_k \subseteq$ DIS$_k$.*

Proof. Recall that $\delta_k = \frac{\delta}{(k+1)(k+2)}$. On event $\bigcap_{k=0}^{K-1} \mathcal{E}_{k,\delta_k/2}$, $h^\star \in V_k$ for all $0 \leq k \leq K$ by Lemma 16 and induction.

The $k = 0$ case is obvious since $D_0 = \text{DIS}_0 = \mathcal{X}$. Now, suppose $0 \leq k < K$, and $D_k \subseteq \text{DIS}_k$. We have

$$\begin{aligned}
D_{k+1} &\subseteq \text{DIS}\left(\left\{h : l(h) \leq \nu + \gamma_2\left(\sigma(k, \delta_k/2) + \sqrt{\sigma(k, \delta_k/2)\nu}\right)\right\}\right) \\
&\subseteq \text{DIS}\left(B\left(h^\star, 2\nu + \gamma_2\left(\sigma(k, \delta_k/2) + \sqrt{\sigma(k, \delta_k/2)\nu}\right)\right)\right)
\end{aligned}$$

where the first line follows from Lemma 17 and the definition of D_k, and the second line follows from triangle inequality that $\Pr(h(X) \neq h^\star(X)) \leq l(h) + l(h^\star)$ (recall $\nu = l(h^\star)$).

To prove $D_{k+1} \subseteq \text{DIS}_{k+1}$ it suffices to show $\gamma_2\left(\sigma(k, \delta_k/2) + \sqrt{\sigma(k, \delta_k/2)\nu}\right) \leq \epsilon_{k+1}$.

Note that $\sigma(k, \delta_k/2) = \sup_{x \in D_k} \frac{\log(2|\mathcal{H}|/\delta_k)}{m_k Q_0(x) + n_k} \leq \sup_{x \in \text{DIS}_k} \frac{\log(2|\mathcal{H}|/\delta_k)}{m_k Q_0(x) + n_k}$ since $D_k \subseteq \text{DIS}_k$. Consequently, $\gamma_2\left(\sigma(k, \delta_k/2) + \sqrt{\sigma(k, \delta_k/2)\nu}\right) \leq \epsilon_{k+1}$. □

D. Proof of Consistency

Proof. (of Theorem 1) Define event $\mathcal{E}^{(0)} := \bigcap_{k=0}^{K} \mathcal{E}_{k,\delta_k/2}$. By a union bound, $\Pr(\mathcal{E}^{(0)}) \geq 1 - \delta$. On event $\mathcal{E}^{(0)}$, by induction and Lemma 16, for all $k = 0, \dots, K$, $h^\star \in V_k$. Observe that $\hat{h} = \hat{h}_K \in V_{K+1}(\delta_K/2)$. Applying Lemma 17 to $\hat{h}$, we have

$$l(\hat{h}) \leq l(h^\star) + \gamma_2\left(\sup_{x \in D_K} \frac{\log(2|\mathcal{H}|/\delta_K)}{m_K Q_0(x) + n_K} + \sqrt{\sup_{x \in D_K} \frac{\log(2|\mathcal{H}|/\delta_K)}{m_K Q_0(x) + n_K} l(h^\star)}\right).$$

The result follows by noting that $\sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{x \in D_K\}}{m_K Q_0(x) + n_K} \leq \sup_{x \in \mathcal{X}} \frac{\mathbb{1}\{x \in \text{DIS}_K\}}{m_K Q_0(x) + n_K}$ by Lemma 18. □

E. Proof of Label Complexity

Proof. (of Theorem 3) Recall that $\xi_k = \inf_{x \in D_k} Q_0(x)$ and $\zeta = \sup_{x \in \text{DIS}_1} \frac{1}{\alpha Q_0(x)+1}$.

Define event $\mathcal{E}^{(0)} := \bigcap_{k=0}^{K} \mathcal{E}_{k,\delta_k/2}$. On this event, by induction and Lemma 16, for all $k = 0, \dots, K$, $h^\star \in V_k$, and consequently by Lemma 18, $D_k \subseteq \text{DIS}_k$.

For any $k = 0, \dots K-1$, let the number of label queries at iteration k to be $U_k := \sum_{t=n_0+\cdots+n_k+1}^{n_0+\cdots+n_{k+1}} Z_t \mathbb{1}\{X_t \in D_{k+1}\}$.

$$\begin{aligned} Z_t \mathbb{1}\{X_t \in D_{k+1}\} &= \mathbb{1}\{X_t \in D_{k+1} \wedge Q_0(X_t) \le \inf_{x \in D_{k+1}} Q_0(x) + \frac{1}{\alpha}\} \\ &\le \mathbb{1}\{X_t \in S(D_{k+1}, \alpha)\} \\ &\le \mathbb{1}\{X_t \in S(\text{DIS}_{k+1}, \alpha)\}. \end{aligned}$$

Thus, $U_k \le \sum_{t=n_0+\cdots+n_k+1}^{n_0+\cdots+n_{k+1}} \mathbb{1}\{X_t \in S(\text{DIS}_{k+1}, \alpha)\}$, where the RHS is a sum of i.i.d. Bernoulli($\Pr(S(\text{DIS}_{k+1}, \alpha))$) random variables, so a Bernstein inequality implies that on an event $\mathcal{E}^{(1,k)}$ of probability at least $1 - \delta_k/2$, $\sum_{t=n_0+\cdots+n_k+1}^{n_0+\cdots+n_{k+1}} \mathbb{1}\{X_t \in S(\text{DIS}_{k+1}, \alpha)\} \le 2n_{k+1} \Pr(S(\text{DIS}_{k+1}, \alpha)) + 2 \log \frac{4}{\delta_k}$.

Therefore, it suffices to show that on event $\mathcal{E}^{(2)} := \cap_{k=0}^{K} (\mathcal{E}^{(1,k)} \cap \mathcal{E}_{k,\delta_k/2})$, for some absolute constant c_1,

$$\sum_{k=0}^{K-1} n_{k+1} \Pr(S(\text{DIS}_{k+1}, \alpha)) \le c_1 \tilde{\theta}(2\nu + \epsilon_K, \alpha)(n\nu + \zeta \log n \log \frac{|\mathcal{H}| \log n}{\delta} + \log n \sqrt{n\nu\zeta \log \frac{|\mathcal{H}| \log n}{\delta}}).$$

Now, on event $\mathcal{E}^{(2)}$, for any $k < K$, $\Pr(S(\text{DIS}_{k+1}, \alpha)) = \Pr(S(\text{DIS}(B(h^\star, 2\nu + \epsilon_{k+1})), \alpha)) \le (2\nu + \epsilon_{k+1}) \tilde{\theta}(2\nu + \epsilon_{k+1}, \alpha)$ where the last inequality follows from Lemma 7.

Therefore,

$$\begin{aligned} &\sum_{k=0}^{K-1} n_{k+1} \Pr(S(\text{DIS}_{k+1}, \alpha)) \\ &\le n_1 + \sum_{k=1}^{K-1} n_{k+1} (2\nu + \epsilon_{k+1}) \tilde{\theta}(2\nu + \epsilon_{k+1}, \alpha) \\ &\le 1 + \tilde{\theta}(2\nu + \epsilon_K, \alpha)(2n\nu + \sum_{k=1}^{K-1} n_{k+1}\epsilon_{k+1}) \\ &\le 1 + \tilde{\theta}(2\nu + \epsilon_K, \alpha) \left(2n\nu + 2\gamma_2 \sum_{k=1}^{K-1} (\sup_{x \in \text{DIS}_1} \frac{\log \frac{|\mathcal{H}|}{\delta_k/2}}{(\alpha Q_0(x) + 1)} + \sqrt{n_k \nu \sup_{x \in \text{DIS}_1} \frac{\log \frac{|\mathcal{H}|}{\delta_k/2}}{(\alpha Q_0(x) + 1)}}) \right) \\ &\le 1 + \tilde{\theta}(2\nu + \epsilon_K, \alpha)(2n\nu + 2\gamma_2 \zeta \log n \log \frac{|\mathcal{H}|(\log n)^2}{\delta} + 2\gamma_2 \log n \sqrt{n\nu\zeta \log \frac{|\mathcal{H}|(\log n)^2}{\delta}}). \end{aligned}$$

□

F. Experiment Details

F.1. Implementation

All algorithms considered require empirical risk minimization. Instead of optimizing 0-1 loss which is known to be computationally hard, we approximate it by optimizing a squared loss. We use the online gradient descent method in (Karampatziakis & Langford, 2011) for optimizing importance weighted loss functions.

For IDBAL, recall that in Algorithm 1, we need to find the empirical risk minimizer $\hat{h}_k \leftarrow \arg\min_{h \in V_k} l(h, \tilde{S}_k)$, update the candidate set $V_{k+1} \leftarrow \{h \in V_k \mid l(h, \tilde{S}_k) \le l(\hat{h}_k, \tilde{S}_k) + \Delta_k(h, \hat{h}_k)\}$, and check whether $x \in \text{DIS}(V_{k+1})$.

Table 5: Dataset information.

Dataset	# of examples	# of features
synthetic	6000	30
letter (U vs P)	1616	16
skin	245057	3
magic	19020	10
covtype	581012	54
mushrooms	8124	112
phishing	11055	68
splice	3175	60
svmguide1	4000	4
a5a	6414	123
cod-rna	59535	8
german	1000	24

In our experiment, we approximately implement this following Vowpal Wabbit (vw). More specifically,

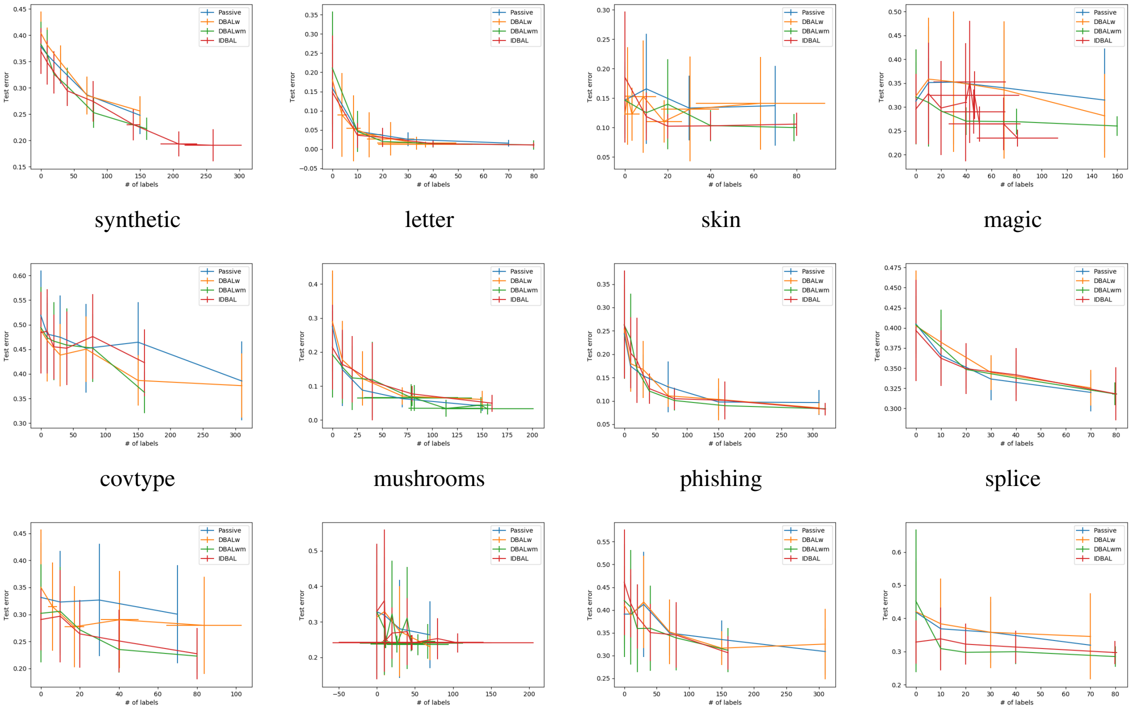
1. Instead of optimizing 0-1 loss which is known to be computationally hard, we use a surrogate loss $l(y, y') = (y - y')^2$.
2. We do not explicitly maintain the candidate set V_{k+1}.
3. To solve the optimization problem $\min_{h \in V_k} l(h, \tilde{S}_k) = \sum_{(X,\tilde{Y},Z) \in \tilde{S}_k} \frac{\mathbb{1}\{h(X) \neq \tilde{Y}\} Z}{m_k Q_0(X) + n_k Q_k(X)}$, we ignore the constraint $h \in V_k$, and use online gradient descent with stepsize $\sqrt{\frac{\eta}{t+\eta}}$ where η is a parameter. The start point for gradient descent is set as $\hat{h}_{k-1}$ the ERM in the last iteration, and the step index t is shared across all iterations (i.e. we do not reset t to 1 in each iteration).
4. To approximately check whether $x \in \text{DIS}(V_{k+1})$, when the hypothesis space $\mathcal{H}$ is linear classifiers, let w_k be the normal vector for current ERM $\hat{h}_k$, and a be current stepsize. We claim $x \in \text{DIS}(V_{k+1})$ if $\frac{|2 w_k^\top x|}{a x^\top x} \leq \sqrt{\frac{C \cdot l(\hat{h}_k, \tilde{S}_k)}{m_k \xi_k + n_k}} + \frac{C \log(m_k + n_k)}{m_k \xi_k + n_k}$ (recall $|\tilde{S}_k| = m_k + n_k$ and $\xi_k = \inf_{x \in \text{DIS}(V_k)} Q_0(x)$) where C is a parameter that captures the model capacity. See (Karampatziakis & Langford, 2011) for the rationale of this approximate disagreement test.
5. $\xi_k = \inf_{x \in \text{DIS}(V_k)} Q_0(x)$ can be approximately estimated with a set of unlabeled samples. This estimate is always an upper bound of the true value of ξ_k.

DBALw and DBALwm can be implemented similarly.

G. Additional Experiment Results

In this section, we present a table of dataset information and plots of test error curves for each algorithm under each policy and dataset.

We remark that the high error bars in test error curves are largely due to the inherent randomness of training sets since in practice active learning is sensitive to the order of training examples. Similar phenomenon can be observed in previous work (Huang et al., 2015).



synthetic

letter

skin

magic

covtype

mushrooms

phishing

splice

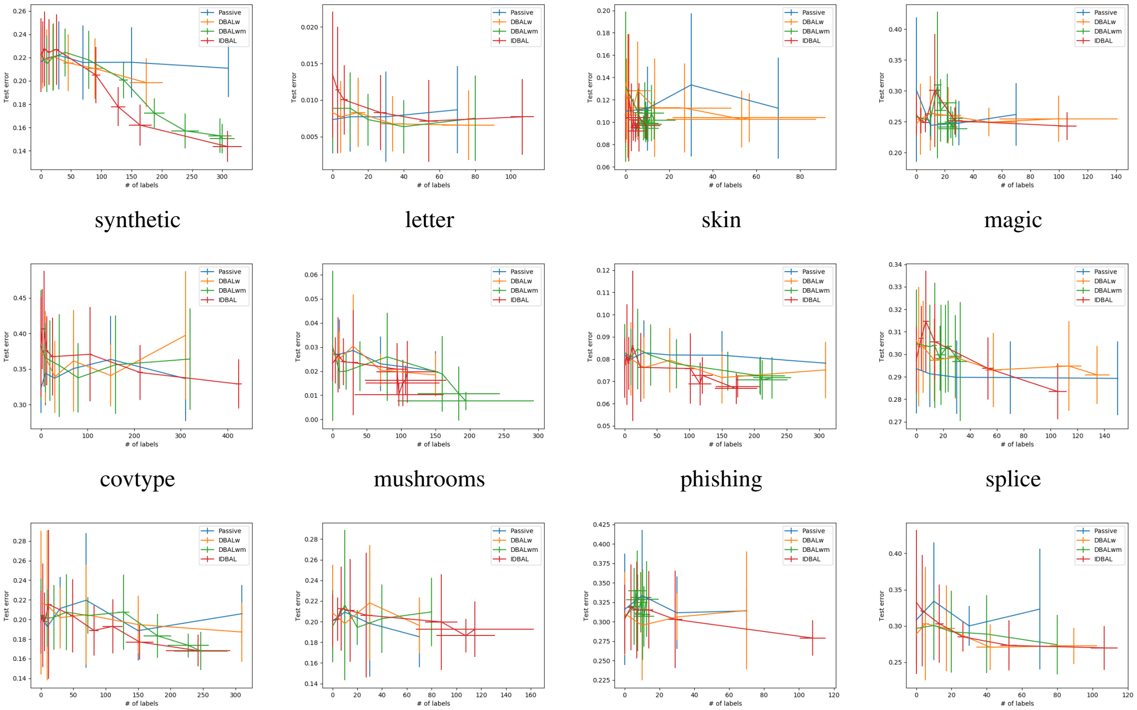
svmguide1

a5a

cod-rna

german

Figure 1: Test error vs. number of labels under the Identical policy



synthetic

letter

skin

magic

covtype

mushrooms

phishing

splice

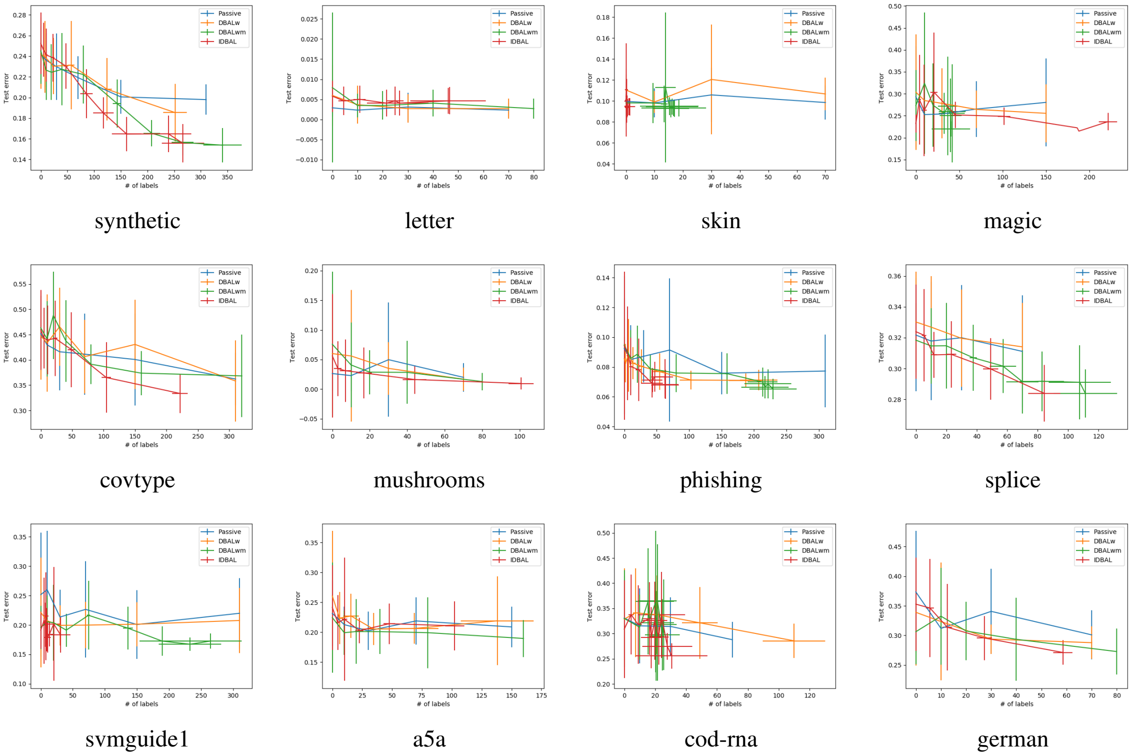
svmguide1

a5a

cod-rna

german

Figure 2: Test error vs. number of labels under the Uniform policy

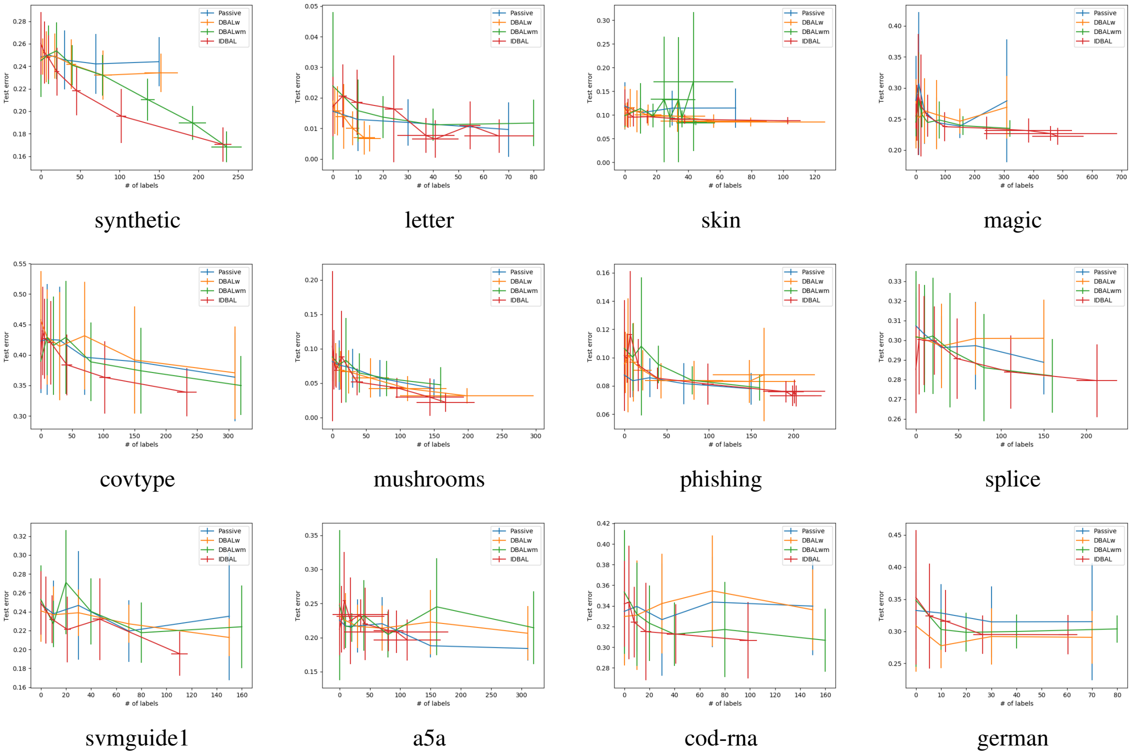


synthetic letter skin magic

covtype mushrooms phishing splice

svmguide1 a5a cod-rna german

Figure 3: Test error vs. number of labels under the Uncertainty policy



synthetic letter skin magic

covtype mushrooms phishing splice

svmguide1 a5a cod-rna german

Figure 4: Test error vs. number of labels under the Certainty policy