Appendix A SPECIFIC BOUNDS FOR OnlineBMR

We begin this section by introducing a *random walk framework* to compute potentials. Suppose $\mathbf{X}^i := (X_1, \cdots, X_k)$ is a random vector that tracks the number of draws of each label among i i.i.d. random draws w.r.t. $\mathbf{u}_\gamma^{Y_t}$. Then according to (1), we may write

$$\phi_t^i(\mathbf{s}) = \mathbb{E}L^{Y_t}(\mathbf{s} + \mathbf{X}).$$

This framework will appear frequently throughout the proofs. We start from rank loss.

Lemma 6. *Under the same setting as in Theorem 2 but with potentials built upon rank loss, we may bound $\phi_t^N(\boldsymbol{0})$ as following:*

$$\phi_t^N(\boldsymbol{0}) \le e^{-\frac{\gamma^2 N}{2}}.$$

Proof. For simplicity, we drop t in the proof. Let $\mathbf{X}^N$ be the aforementioned random vector. Then we may write the potential by

$$\begin{aligned}
\phi^N(\mathbf{0}) &= \mathbb{E}L^Y_{\text{rnk}}(\mathbf{X}^N) \\
&\le w_Y \sum_{l \in Y} \sum_{r \notin Y} \mathbb{E}\mathbb{1}(X_r \ge X_l) \\
&= w_Y \sum_{l \in Y} \sum_{r \notin Y} \mathbb{P}(X_r - X_l \ge 0).
\end{aligned}$$

Fix $l \in Y$ and $r \notin Y$. By definition of $\mathbf{u}_\gamma^Y$, we have

$$a := \mathbf{u}_\gamma^Y[l] = \mathbf{u}_\gamma^Y[r] + \gamma =: b.$$

Now suppose we draw 1 with probability a, -1 with probability b, and 0 otherwise. Then $\mathbb{P}(X_r - X_l \ge 0)$ equals the probability that the summation of N i.i.d. random numbers is non-negative. Then we can apply the Hoeffding's inequality to get

$$\mathbb{P}(X_r - X_l \ge 0) \le e^{-\frac{\gamma^2 N}{2}}.$$

Since w_Y is the inverse of the number of pairs (l, r), this proves our assertion. □

Lemma 7. *Under the same setting as in Theorem 2 but with potentials built upon rank loss, we can show that $\forall i,\ w^{i*} \le O(\frac{1}{\sqrt{N-i}})$.*

Proof. First we fix t and i. We also fix $l^* \in Y_t$ and $r^* \in Y_t^c$. Then write $\mathbf{s}_1 := \mathbf{s}_t^{i-1} + \mathbf{e}_{l^*}$ and $\mathbf{s}_2 := \mathbf{s}_t^{i-1} + \mathbf{e}_{r^*}$. Again we introduce $\mathbf{X}^{N-i}$. Then we may write

$$\begin{aligned}
\mathbf{c}_t^i[r^*] - \mathbf{c}_t^i[l^*] &= \phi_t^{N-i}(\mathbf{s}_2) - \phi_t^{N-i}(\mathbf{s}_1) \\
&= \mathbb{E}[L^{Y_t}_{\text{rnk}}(\mathbf{s}_2 + \mathbf{X}^{N-i}) - L^{Y_t}_{\text{rnk}}(\mathbf{s}_1 + \mathbf{X}^{N-i})] \\
&\le w_{Y_t} \sum_{l \in Y_t} \sum_{r \notin Y_t} f(r, l),
\end{aligned}$$

where

$$\begin{aligned}
f(r, l) := \mathbb{E}[&\mathbb{1}(\mathbf{s}_2[r] + X_r \ge \mathbf{s}_2[l] + X_l) \\
&- \mathbb{1}(\mathbf{s}_1[r] + X_r > \mathbf{s}_1[l] + X_l)].
\end{aligned}$$

Here we intentionally include and exclude equality for the ease of computation. Changing the order of terms, we can derive

$$\begin{aligned}
f(r, l) &\le \mathbb{P}(\mathbf{s}_1[l] - \mathbf{s}_1[r] \ge X_r - X_l \ge \mathbf{s}_2[l] - \mathbf{s}_2[r]) \\
&\le 3 \max_n \mathbb{P}(X_r - X_l = n),
\end{aligned}$$

where the last inequality is deduced from the fact that

$$(\mathbf{s}_1[l] - \mathbf{s}_1[r]) - (\mathbf{s}_2[l] - \mathbf{s}_2[r]) \in \{0, 1, 2\}.$$

Using Berry-Esseen theorem, it is shown by Jung et al. [2017, Lemma 10] that $\max_n \mathbb{P}(X_r - X_l = n) \le O(\frac{1}{\sqrt{N-i}})$, which implies that

$$\mathbf{c}_t^i[r^*] - \mathbf{c}_t^i[l^*] \le O(\frac{1}{\sqrt{N-i}}).$$

Since l^* and r^* are arbitrary, and the bound does not depend on t, the last inequality proves our assertion. □

Now we provide similar bounds when the potentials are computed from hinge loss.

Lemma 8. *Under the same setting as in Theorem 2 but with potentials built upon hinge loss, we may bound $\phi_t^N(\boldsymbol{0})$ as following:*

$$\phi_t^N(\boldsymbol{0}) \le (N+1)e^{-\frac{\gamma^2 N}{2}}.$$

Proof. Again we drop t in the proof and introduce $\mathbf{X}^N$. Then we may write the potential by

$$\begin{aligned}
\phi^N(\mathbf{0}) &= \mathbb{E}L^Y_{\text{hinge}}(\mathbf{X}^N) \\
&= w_Y \sum_{l \in Y} \sum_{r \notin Y} \mathbb{E}(1 + X_r - X_l)_+ \\
&= w_Y \sum_{l \in Y} \sum_{r \notin Y} \sum_{n=0}^{N} \mathbb{P}(X_r - X_l \ge n) \\
&\le w_Y \sum_{l \in Y} \sum_{r \notin Y} (N+1)\mathbb{P}(X_r - X_l \ge 0).
\end{aligned}$$

We already checked in Lemma 6 that

$$\mathbb{P}(X_r - X_l \ge 0) \le e^{-\frac{\gamma^2 N}{2}},$$

which concludes the proof. □

Lemma 9. *Under the same setting as in Theorem 2 but with potentials built upon hinge loss, we can show that $\forall i,\ w^{i*} \le 2$.*

Proof. First we fix t and i. We also fix $l^* \in Y_t$ and $r^* \in Y_t^c$. Then write $\mathbf{s}_1 := \mathbf{s}_t^{i-1} + \mathbf{e}_{l*}$ and $\mathbf{s}_2 := \mathbf{s}_t^{i-1} + \mathbf{e}_{r*}$. Again with $\mathbf{X}^{N-i}$, we may write

$$\begin{aligned}\mathbf{c}_t^i[r^*] - \mathbf{c}_t^i[l^*] &= \phi_t^{N-i}(\mathbf{s}_2) - \phi_t^{N-i}(\mathbf{s}_1)\\ &= \mathbb{E}[L_{\text{hinge}}^{Y_t}(\mathbf{s}_2 + \mathbf{X}^{N-i}) - L_{\text{hinge}}^{Y_t}(\mathbf{s}_1 + \mathbf{X}^{N-i})]\\ &= w_{Y_t} \sum_{l \in Y_t} \sum_{r \notin Y_t} f(r, l),\end{aligned}$$

where

$$\begin{aligned}f(r, l) := \mathbb{E}[&(1 + (\mathbf{s}_2 + \mathbf{X}^{N-i})[r] - (\mathbf{s}_2 + \mathbf{X}^{N-i})[l])_+\\ &- (1 + (\mathbf{s}_1 + \mathbf{X}^{N-i})[r] - (\mathbf{s}_1 + \mathbf{X}^{N-i})[l])_+].\end{aligned}$$

It is not hard to check that the term inside the expectation is always bounded above by 2. This fact along with the definition of w_{Y_t} provides that $\mathbf{c}_t^i[r^*] - \mathbf{c}_t^i[l^*] \le 2$. Since our choice of l^* and r^* are arbitrary, this proves $\mathbf{w}^i[t] \le 2$, which completes the proof. □

Appendix B COMPLETE PROOF OF THEOREM 4

Proof. We assume that an adversary draws a label Y_t uniformly at random from $2^{[k]} - \{\emptyset, [k]\}$, and the weak learners generate single-label predictions w.r.t. $\mathbf{p}_t \in \Delta[k]$. Any boosting algorithm can only make a final decision by weighted cumulative votes of N weak learners. We manipulate $\mathbf{p}_t$ such that weak learners satisfy OnlineWLC (δ, γ, S) but the best possible performance is close to (6).

As we are assuming single-label predictions, $\mathbf{h}_t = \mathbf{e}_{l_t}$ for some $l_t \in [k]$ and $\mathbf{c}_t \cdot \mathbf{h}_t = \mathbf{c}_t[l_t]$. Furthermore, the bounded condition of $\mathcal{C}_0^{eor}$ ensures $\mathbf{c}_t[l_t]$ is contained in $[0, 1]$. The Azuma-Hoeffding inequality provides that with probability $1 - \delta$,

$$\begin{aligned}\sum_{t=1}^T w_t \mathbf{c}_t[l_t] &\le \sum_{t=1}^T w_t \mathbf{c}_t \cdot \mathbf{p}_t + \sqrt{2||\mathbf{w}||_2^2 \ln(\frac{1}{\delta})}\\ &\le \sum_{t=1}^T w_t \mathbf{c}_t \cdot \mathbf{p}_t + \frac{\gamma||\mathbf{w}||_2^2}{k} + \frac{k \ln(\frac{1}{\delta})}{2\gamma} \qquad (18)\\ &\le \sum_{t=1}^T w_t \mathbf{c}_t \cdot \mathbf{p}_t + \frac{\gamma||\mathbf{w}||_1}{k} + \frac{k \ln(\frac{1}{\delta})}{2\gamma},\end{aligned}$$

where the second inequality holds by arithmetic mean and geometric mean relation and the last inequality holds due to $w_t \in [0, 1]$.

We start from providing a lower bound on the number of weak learners. Let $\mathbf{p}_t = \mathbf{u}_{2\gamma}^{Y_t}$ for all t. This can be done by the constraint $\gamma < \frac{1}{4k}$. From the condition of $\mathcal{C}_0^{eor}$ that $\min_l \mathbf{c}[l] = 0, \max_l \mathbf{c} = 1$ along with the fact that $Y \notin \{\emptyset, [k]\}$, we can show that $\mathbf{c} \cdot (\mathbf{u}_\gamma^Y - \mathbf{u}_{2\gamma}^Y) \ge \frac{\gamma}{k}$. Then the last line of (18) becomes

$$\begin{aligned}\sum_{t=1}^T &w_t \mathbf{c}_t \cdot \mathbf{u}_{2\gamma}^{Y_t} + \frac{\gamma||\mathbf{w}||_1}{k} + \frac{k \ln(\frac{1}{\delta})}{2\gamma}\\ &\le \sum_{t=1}^T (w_t \mathbf{c}_t \cdot \mathbf{u}_\gamma^{Y_t} - \frac{\gamma w_t}{k}) + \frac{\gamma||\mathbf{w}||_1}{k} + \frac{k \ln(\frac{1}{\delta})}{2\gamma}\\ &\le \sum_{t=1}^T w_t \mathbf{c}_t \cdot \mathbf{u}_\gamma^{Y_t} + S,\end{aligned}$$

which validates that weak learners indeed satisfy OnlineWLC (δ, γ, S). Following the argument of Schapire and Freund [2012, Section 13.2.6], we can also prove that the optimal choice of weights over the learners is $(\frac{1}{N}, \cdots, \frac{1}{N})$.

Now we compute a lower bound for the booster's loss. Let $\mathbf{X} := (X_1, \cdots, X_k)$ be a random vector that tracks the number of labels drawn from N i.i.d. random draws w.r.t. $\mathbf{u}_{2\gamma}^Y$. Then the expected rank loss of the booster can be written as:

$$\mathbb{E}L_{\text{rnk}}^Y(\mathbf{X}) \ge w_Y \sum_{l \in Y} \sum_{r \notin Y} \mathbb{P}(X_l < X_r).$$

Adopting the arguments in the proof by Jung et al. [2017, Theorem 4], we can show that

$$\mathbb{P}(X_l < X_r) \ge \Omega(e^{-4Nk^2\gamma^2}).$$

This shows $\mathbb{E}L_{\text{rnk}}^Y(\mathbf{X}) \ge \Omega(e^{-4Nk^2\gamma^2})$. Setting this value equal to ϵ, we have $N \ge \Omega(\frac{1}{\gamma^2} \ln \frac{1}{\epsilon})$, considering k as a fixed constant. This proves the first part of the theorem.

Now we move on to the optimality of sample complexity. We record another inequality that can be checked from the conditions of $\mathcal{C}_0^{eor}$: $\mathbf{c} \cdot (\mathbf{u}_0^Y - \mathbf{u}_\gamma^Y) \le \gamma$. Let $T_0 := \frac{S}{4\gamma}$ and define $\mathbf{p}_t = \mathbf{u}_0^{Y_t}$ for $t \le T_0$ and $\mathbf{p}_t = \mathbf{u}_{2\gamma}^{Y_t}$ for $t > T_0$. Then for $T \le T_0$, (18) implies

$$\begin{aligned}\sum_{t=1}^T &w_t \mathbf{c}_t[l_t]\\ &\le \sum_{t=1}^T w_t \mathbf{c}_t \cdot \mathbf{u}_0^{Y_t} + \frac{\gamma||\mathbf{w}||_1}{k} + \frac{k \ln(\frac{1}{\delta})}{2\gamma}\\ &\le \sum_{t=1}^T w_t \mathbf{c}_t \cdot \mathbf{u}_\gamma^{Y_t} + \gamma(1 + \frac{1}{k})||\mathbf{w}||_1 + \frac{k \ln(\frac{1}{\delta})}{2\gamma} \qquad (19)\\ &\le \sum_{t=1}^T w_t \mathbf{c}_t \cdot \mathbf{u}_\gamma^{Y_t} + S.\end{aligned}$$

where the last inequality holds because $||\mathbf{w}||_1 \le T_0 =$

$\frac{S}{4\gamma}$. For $T > T_0$, again (18) implies

$$\begin{aligned}\sum_{t=1}^{T} w_t \mathbf{c}_t[l_t] &\leq \sum_{t=1}^{T_0} w_t \mathbf{c}_t \cdot \mathbf{u}_0^{Y_t} + \sum_{t=T_0+1}^{T} w_t \mathbf{c}_t \cdot \mathbf{u}_{2\gamma}^{Y_t} \\ &\quad + \frac{\gamma ||\mathbf{w}||_1}{k} + \frac{k \ln(\frac{1}{\delta})}{2\gamma} \\ &\leq \sum_{t=1}^{T} w_t \mathbf{c}_t \cdot \mathbf{u}_\gamma^{Y_t} + \frac{k+1}{k}\gamma T_0 + \frac{k \ln(\frac{1}{\delta})}{2\gamma} \\ &\leq \sum_{t=1}^{T} w_t \mathbf{c}_t \cdot \mathbf{u}_\gamma^{Y_t} + S.\end{aligned} \tag{20}$$

(19) and (20) prove that the weak learners indeed satisfy OnlineWLC (δ, γ, S). Observing that weak learners do not provide meaningful information for $t \leq T_0$, we can claim any online boosting algorithm suffers a loss at least $\Omega(T_0)$. Therefore to get the certain accuracy, the number of instances T should be at least $\Omega(\frac{T_0}{\epsilon}) = \Omega(\frac{S}{\epsilon\gamma})$, which completes the second part of the proof. □