5 Appendix

This appendix contains the proofs of all lemmas and theorems presented in the main text.

**Proof:**  [Lemma 1] First observe that, for any given size $s$, the sequence $Y^*_{s,t}$ must contain the $s$ top-ranked classes in the sorted order of $p_{i,t}$. This is because, for any candidate sequence $Y_s = \{j_1, j_2, \ldots, j_s\}$, we have $E_t[\ell_{c_i}(Y^*_t, Y_s)] = (1 - a) \sum_{i \in Y_s} c(j_i, s) - \left( \frac{a}{1 - a} + c(j_i, s) \right) p_{i,t}$.

If there exists $i \in Y_s$ which is not among the $s$-top ranked ones, then we could replace class $i$ in position $j_i$ within $Y_s$ with class $k \notin Y_s$ such that $p_{k,t} > p_{i,t}$ obtaining a smaller loss.

Next, we show that the optimal ordering within $Y^*_{s,t}$ is precisely ruled by the nonincreasing order of $p_{i,t}$. By the sake of contradiction, assume there are $i$ and $k$ in $Y^*_{s,t}$ such that $i$ precedes $k$ in $Y^*_{s,t}$ but $p_{k,t} > p_{i,t}$. Specifically, let $i$ be in position $j_1$ and $k$ be in position $j_2$ with $j_1 < j_2$ and such that $c(j_1, s) > c(j_2, s)$. Then, disregarding the $(1 - a)$-factor, switching the two classes within $Y^*_{s,t}$ yields an expected loss difference of

$$c(j_1, s) - \left( \frac{a}{1 - a} + c(j_1, s) \right) p_{i,t} + c(j_2, s) - \left( \frac{a}{1 - a} + c(j_2, s) \right) p_{k,t}$$

$$- \left( c(j_1, s) - \left( \frac{a}{1 - a} + c(j_1, s) \right) p_{k,t} \right) - \left( c(j_2, s) - \left( \frac{a}{1 - a} + c(j_2, s) \right) p_{i,t} \right)$$

$$= (p_{k,t} - p_{i,t}) (c(j_1, s) - c(j_2, s)) > 0.$$  

Hence switching would get a smaller loss which leads as a consequence to $Y^*_{s,t}$ is precisely ruled by the nonincreasing order of $p_{i,t}$.

The algorithm in Figure 1 works by updating through the gradients $\nabla w_{i,t}$ of a modular margin-based loss function $\sum_{i=1}^{K} L(w_i^\top x_i)$ associated with the label generation model (2) so as to make the parameters $(u_1, \ldots, u_K) \in \mathcal{R}^{dk}$ therein achieve the Bayes optimality condition

$$(u_1, \ldots, u_K) = \underset{w_1, \ldots, w_K : w_i^\top x_i \in D}{\arg\min} \mathbb{E}_t \left[ \sum_{i=1}^{K} L(s_{i,t} w_i^\top x_i) \right],$$

where $\mathbb{E}_t[\cdot]$ above is over the generation of $Y_t$ in producing the sign value $s_{i,t} \in \{-1, 0, +1\}$, conditioned on the past (in particular, conditioned on $Y_{t-1}$). The requirement in (4) is akin to the classical construction of proper scoring rules in the statistical literature (e.g., [9]).

The following lemma faces the problem of hand-crafting a convenient loss function $L(\cdot)$ such that (4) holds.

**Lemma 5.** Let $w_1, \ldots, w_K \in \mathcal{R}^{dk}$ be arbitrary weight vectors such that $w_i^\top x_i \in D$, $i \in [K]$, $(u_1, \ldots, u_K) \in \mathcal{R}^{dk}$ be defined in (2), $s_{i,t}$ be the updating signs computed by the algorithm at the end (Step 5) of time $t$, $L : D = [-R, R] \subseteq \mathcal{R} \to \mathcal{R}^+$ be a convex and differentiable function of its argument, with $g(\Delta) = -L'(\Delta)$. Then for any $t$ we have

$$\mathbb{E}_t \left[ \sum_{i=1}^{K} L(s_{i,t} w_i^\top x_i) \right] \geq \mathbb{E}_t \left[ \sum_{i=1}^{K} L(s_{i,t} u_i^\top x_i) \right],$$

Supplementary Material to “On Multilabel Classification and Ranking with Partial Feedback”
For similar reasons, all i

Since rates of convergence contained in the subsequent Lemma 9.

Proof: Let us introduce the shorthands \( \Delta_i = u_i^\top x_t, \hat{\Delta}_i = w_{i,t}^\top x_t, s_i = s_{i,t}, \) and \( p_i = P(y_{i,t} = 1 \mid x_t) = \frac{L'(\Delta_i)}{L(\Delta_i) + L(-\Delta_i)}. \) Moreover, let \( P_i(\cdot) \) be an abbreviation for the conditional probability \( P(\cdot \mid (y_1, x_1), \ldots, (y_{t-1}, x_{t-1}), x_t). \) Recalling the way \( w_{i,t} \) is constructed (Figure 1), we can write

\[
\mathbb{E}_t \left[ \sum_{i=1}^K L(s_{i,t} \hat{\Delta}_i) \right] = \sum_{i \in \hat{Y}_t} \left( P_i(s_{i,t} = 1) L(\hat{\Delta}_i) + P_i(s_{i,t} = -1) L(-\hat{\Delta}_i) \right) + (K - |\hat{Y}_t|) L(0)
\]

\[
= \sum_{i \in \hat{Y}_t} \left( p_i L(\hat{\Delta}_i) + (1 - p_i) L(-\hat{\Delta}_i) \right) + (K - |\hat{Y}_t|) L(0),
\]

For similar reasons,

\[
\mathbb{E}_t \left[ \sum_{i=1}^K L(s_{i,t} \Delta_i) \right] = \sum_{i \in \hat{Y}_t} \left( p_i L(\Delta_i) + (1 - p_i) L(-\Delta_i) \right) + (K - |\hat{Y}_t|) L(0).
\]

Since \( L(\cdot) \) is convex, so is \( \mathbb{E}_t \left[ \sum_{i=1}^K L(s_{i,t} \hat{\Delta}_i) \right] \) when viewed as a function of the \( \hat{\Delta}_i. \) We have that

\[
\frac{\partial \mathbb{E}_t}{\partial \hat{\Delta}_i} \left[ \sum_{i=1}^K L(s_{i,t} \hat{\Delta}_i) \right] = 0 \text{ if and only if for all } i \in \hat{Y}_t \text{ we have that } \hat{\Delta}_i \text{ satisfies }
\]

\[
p_i = \frac{L'(-\hat{\Delta}_i)}{L'(\hat{\Delta}_i) + L'(-\hat{\Delta}_i)}.
\]

Since \( p_i = \frac{L'(-\hat{\Delta}_i)}{L'(\hat{\Delta}_i) + L'(-\hat{\Delta}_i)}, \) we have that \( \mathbb{E}_t \left[ \sum_{i=1}^K L(s_{i,t} \hat{\Delta}_i) \right] \) is minimized when \( \hat{\Delta}_i = \Delta_i \) for all \( i \in [K]. \) The claimed result immediately follows.

Let now \( Var(\cdot) \) be a shorthand for \( Var(\cdot \mid (y_1, x_1), \ldots, (y_{t-1}, x_{t-1}), x_t). \) The following lemma shows that under additional assumptions on the loss \( L(\cdot), \) we are afforded to bound the variance of a difference of losses \( L(\cdot) \) by the expectation of this difference. This will be key to proving the fast rates of convergence contained in the subsequent Lemma 9.

**Lemma 6.** Let \( (u_{1,t}, \ldots, u_{K,t}) \in \mathcal{R}^{dK} \) be the weight vectors computed by the algorithm in Figure 1 at the beginning (Step 2) of time \( t, \) \( s_{i,t} \) be the updating signs computed at the end (Step 3) of time \( t, \) and \( (u_1, \ldots, u_K) \in \mathcal{R}^{dK} \) be the comparison vectors defined through (2). Let \( L : D = [-R, R] \subseteq \mathcal{R} \rightarrow \mathcal{R}^+ \) be a \( C^2 \) convex function of its argument, with \( g(\Delta) = -L'(\Delta) \) and such that there are positive constants \( c'_L \) and \( c''_L \) with \( (L'(\Delta))^2 \leq c'_L \) and \( L''(\Delta) \geq c''_L \) for all \( \Delta \in D. \) Then for any \( i \in \hat{Y}_t \)

\[
0 \leq Var_t \left( L(s_{i,t} x_t^\top u_{i,t}) - L(s_{i,t} u_{i,t}^\top x_t) \right) \leq \frac{2c'_L}{c''_L} \mathbb{E}_t \left[ L(s_{i,t} x_t^\top u_{i,t}) - L(s_{i,t} u_{i,t}^\top x_t) \right]
\]

Proof: Let us introduce the shorthands \( \Delta_i = x_t^\top u_i, \hat{\Delta}_i = x_t^\top u_{i,t}, s_i = s_{i,t}, \) and \( p_i = P(y_{i,t} = 1 \mid x_t) = \frac{L'(-\Delta_i)}{L'(\Delta_i) + L(-\Delta_i)}. \) Then, for any \( i \in [K], \)

\[
Var_t \left( L(s_{i,t} x_t^\top u_{i,t}) - L(s_{i,t} u_{i,t}^\top x_t) \right) \leq \mathbb{E}_t \left( \left( L(s_i \hat{\Delta}_i) - L(s_i \Delta_i) \right)^2 \right) \leq c'_L (\hat{\Delta}_i - \Delta_i)^2.
\]
Moreover, for any \( i \in \check{Y}_t \) we can write

\[
\begin{align*}
E_t \left[ L(s_i \hat{\Delta}_t) - L(s_i \Delta_t) \right] &= p_i \left( L(\hat{\Delta}_t) - L(\Delta_t) \right) + (1 - p_i) \left( L(-\hat{\Delta}_t) - L(-\Delta_t) \right) \\
&\geq p_i \left( L'(\Delta_t)(\hat{\Delta}_t - \Delta_t) + \frac{c''}{2}(\hat{\Delta}_t - \Delta_t)^2 \right) \\
&\quad + (1 - p_i) \left( L'(-\Delta_t)(\hat{\Delta}_t - \Delta) + \frac{c''}{2}(\hat{\Delta}_t - \Delta_t)^2 \right) \\
&= p_i \frac{c''}{2}(\hat{\Delta}_t - \Delta_t)^2 + (1 - p_i) \frac{c''}{2}(\hat{\Delta}_t - \Delta_t)^2 \\
&= \frac{c''}{2}(\hat{\Delta}_t - \Delta_t)^2,
\end{align*}
\]

where the second equality uses the definition of \( p_i \). Combining (5) with (6) gives the desired bound.

\( \square \)

We continue by showing a one-step regret bound for our original loss \( \ell_{a,c} \). The precise connection to loss \( L(\cdot) \) will be established with the help of a later lemma (Lemma 9).

**Lemma 7.** Let \( L : D = [-R, R] \subseteq \mathbb{R} \to \mathbb{R}^+ \) be a convex, twice differentiable, and nonincreasing function of its argument. Let \((u_1, \ldots, u_K) \in \mathbb{R}^{dK}\) be defined in (2) with \( g(\Delta) = -L'(\Delta) \) for all \( \Delta \in D \). Let also \( c_L \) be a positive constant such that

\[
\frac{L'(\Delta) L''(-\Delta) + L''(\Delta) L'(-\Delta)}{(L'(\Delta))^2} \geq -c_L
\]

holds for all \( \Delta \in D \). Finally, let \( \Delta_{i,t} \) denote \( u_i^\top x_t \), and \( \hat{\Delta}_{i,t} \) denote \( \hat{\Delta}_t \top \hat{\Delta}_t \), where \( \hat{\Delta}_{i,t} \) is the \( i \)-th weight vector computed by the algorithm at the beginning (Step 2) of time \( t \). If time \( t \) is such that \( |\Delta_{i,t} - \hat{\Delta}_{i,t}| \leq \epsilon_{i,t} \) for all \( i \in [K] \), then

\[
E_t[\ell_{a,c}(Y_t, \hat{Y}_t)] - E_t[\ell_{a,c}(Y_t, Y^*_t)] \leq 2 (1 - a) c_L \sum_{i \in \check{Y}_t} \epsilon_{i,t}.
\]

**Proof:** Introduce the shorthand notation \( p(\Delta) = \frac{g(-\Delta)}{g(\Delta) + g(-\Delta)} \). We can write

\[
\begin{align*}
E_t[\ell_{a,c}(Y_t, \hat{Y}_t)] - E_t[\ell_{a,c}(Y_t, Y^*_t)] &= (1 - a) \sum_{i \in \check{Y}_t} \left( c(\hat{j}_i, |\hat{Y}_t|) - \left( \frac{\alpha}{1 - a} + c(\hat{j}_i, |\hat{Y}_t|) \right) p(\Delta_{i,t}) \right) \\
&\quad - (1 - a) \sum_{i \in \check{Y}_t} \left( c(j^*_i, |Y^*_t|) - \left( \frac{\alpha}{1 - a} + c(j^*_i, |Y^*_t|) \right) p(\Delta_{i,t}) \right).
\end{align*}
\]

where \( \hat{j}_i \) denotes the position of class \( i \) in \( \hat{Y}_t \) and \( j^*_i \) is the position of class \( i \) in \( Y^*_t \). Now,

\[
p'(\Delta) = \frac{-g'(-\Delta) g(\Delta) - g'(\Delta) g(-\Delta)}{(g(\Delta) + g(-\Delta))^2} = \frac{-L'(\Delta) L''(-\Delta) - L'(-\Delta) L''(\Delta)}{(L'(\Delta))^2} \geq 0
\]

since \( g(\Delta) = -L'(\Delta) \), and \( L(\cdot) \) is convex and nonincreasing. Hence \( p(\Delta) \) is itself a nondecreasing function of \( \Delta \). Moreover, the extra condition on \( L \) involving \( L' \) and \( L'' \) is a Lipschitz condition on \( p(\Delta) \) via a uniform bound on \( p'(\Delta) \). Hence, from \( |\Delta_{i,t} - \hat{\Delta}_{i,t}| \leq \epsilon_{i,t} \) and the definition of \( \hat{Y}_t \) we
can write
\[
\mathbb{E}_t[\ell_{a,c}(Y_t, \hat{Y}_t)] - \mathbb{E}_t[\ell_{a,c}(Y_t, Y_t^*)]
\]
\[
\leq (1 - a) \sum_{i \in Y_t} \left( c(j_i, |\hat{Y}_t|) - \left( \frac{a}{1 - a} + c(j_i, |\hat{Y}_t|) \right) p([\Delta_{i,t} - \epsilon_{i,t}]_{D}) \right)
\]
\[
- (1 - a) \sum_{i \in Y_t^*} \left( c(j_i^*, |Y_t^*|) - \left( \frac{a}{1 - a} + c(j_i^*, |Y_t^*|) \right) p([\Delta_{i,t}^* + \epsilon_{i,t}]_{D}) \right)
\]
\[
\leq (1 - a) \sum_{i \in Y_t} \left( c(j_i, |\hat{Y}_t|) - \left( \frac{a}{1 - a} + c(j_i, |\hat{Y}_t|) \right) p([\Delta_{i,t} - \epsilon_{i,t}]_{D}) \right)
\]
\[
- (1 - a) \sum_{i \in Y_t^*} \left( c(j_i^*, |Y_t^*|) - \left( \frac{a}{1 - a} + c(j_i^*, |Y_t^*|) \right) p([\Delta_{i,t}^* + \epsilon_{i,t}]_{D}) \right)
\]
\[
= (1 - a) \sum_{i \in Y_t} \left( c(j_i, |\hat{Y}_t|) \left( p([\Delta_{i,t} - \epsilon_{i,t}]_{D}) - p([\Delta_{i,t}^* + \epsilon_{i,t}]_{D}) \right) \right)
\]
\[
\leq 2 (1 - a) c_L \sum_{i \in Y_t} \epsilon_{i,t}
\]
the last inequality deriving from \(c(i, s) \leq 1\) for all \(i \leq s \leq K\), and
\[
p([\Delta_{i,t} - \epsilon_{i,t}]_{D}) - p([\Delta_{i,t}^* + \epsilon_{i,t}]_{D}) \leq c_L [p(\Delta_{i,t} - \epsilon_{i,t}]_{D} - [\Delta_{i,t}^* - \epsilon_{i,t}]_{D}] \leq 2 c_L \epsilon_{i,t}.
\]

Likewise, we provide a similar bound for the ranking loss.

**Lemma 8.** Under the same assumptions and notation as in Lemma 7, let the Algorithm in Figure 1 be working with \(a \to 1\) and strictly decreasing cost values \(c(i, s)\). Let \(w_{t,i}^*\) be the \(i\)-th weight vector computed by this algorithm at the beginning (Step 2) of time \(t\). If this algorithm ranks classes by \(\tilde{p}_{j_1,t} \geq \ldots \geq \tilde{p}_{j_{|K|},t} \geq 0\), and time \(t\) is such that \(|\Delta_{i,t} - \Delta_{t,i}^*| \leq \epsilon_{i,t}\) for all \(i \in [K]\), then
\[
\mathbb{E}_t[\ell_{\text{rank},t}(Y_t, (\tilde{p}_{j_1,t}, \ldots, \tilde{p}_{j_{|K|},t}, 0, \ldots, 0))] - \mathbb{E}_t[\ell_{\text{rank},t}(Y_t, (p_{i_1,t}, \ldots, p_{i_{|S|},t}, 0, \ldots, 0))]
\]
\[
\leq 2 S_t c_L \sum_{i \in Y_t} \epsilon_{i,t},
\]
where the \(p_{i,t} = \mathbb{P}_t(y_{i,t} = 1 \mid x_t)\) are sorted as \(p_{i_1,t} \geq \ldots \geq p_{i_{|S|},t} \geq 0\), and \(\hat{Y}_t = (j_1, j_2, \ldots, j_{|K|})\).

**Proof:** Recall the notation \(P_t(\cdot) = \mathbb{P}(\cdot \mid x_t)\), and \(p_{i,t} = p(\Delta_{i,t}) = \frac{g(\Delta_{i,t})}{g(\Delta_{i,t}) + g(\Delta_{i,t}^*)}\). Following [6] (proof of Theorem 2 therein), one can see that for generic sequences \((\tilde{p}_1,t, \ldots, \tilde{p}_{|K|},t)\) and \((p_1,t, \ldots, p_{|K|},t)\) one has
\[
\mathbb{E}_t[\ell_{\text{rank}}(Y_t, (\tilde{p}_1,t, \ldots, \tilde{p}_{|K|},t))] - \mathbb{E}_t[\ell_{\text{rank}}(Y_t, (p_1,t, \ldots, p_{|K|},t))]
\]
\[
= \sum_{i,j \in [K], i < j} \left( \tilde{r}(i, j) - r(i, j) + \hat{r}(j, i) - r(j, i) \right),
\]
where
\[
\tilde{r}(i, j) = \mathbb{P}_t(y_{i,t} > y_{j,t}) \left( \{\tilde{p}_{i,t} < \tilde{p}_{j,t}\} + \frac{1}{2} \{\tilde{p}_{i,t} = \tilde{p}_{j,t}\} \right)
\]
\[
r(i, j) = \mathbb{P}_t(y_{i,t} > y_{j,t}) \left( \{p_{i,t} < p_{j,t}\} + \frac{1}{2} \{p_{i,t} = p_{j,t}\} \right)
\]
Since
\[
\mathbb{P}_t(y_{i,t} > y_{j,t}) - \mathbb{P}_t(y_{j,t} > y_{i,t}) = \mathbb{P}_t(y_{i,t} = 1) - \mathbb{P}_t(y_{j,t} = 1) = p_{i,t} - p_{j,t},
\]
a simple case analysis reveals that
\[
\tilde{r}(i, j) - r(i, j) + \hat{r}(j, i) - r(j, i) = \begin{cases} 
\frac{1}{2} (p_{i,t} - p_{j,t}) & \text{if } \tilde{p}_{i,t} < \tilde{p}_{j,t}, \text{ and } p_{i,t} = p_{j,t} \text{ or } \tilde{p}_{i,t} = \tilde{p}_{j,t}, \text{ and } p_{i,t} > p_{j,t} \\
\frac{1}{2} (p_{j,t} - p_{i,t}) & \text{if } \tilde{p}_{i,t} = \tilde{p}_{j,t}, \text{ and } p_{i,t} < p_{j,t} \text{ or } \tilde{p}_{i,t} > \tilde{p}_{j,t}, \text{ and } p_{i,t} = p_{j,t} \\
p_{i,t} - p_{j,t} & \text{if } \tilde{p}_{i,t} < \tilde{p}_{j,t}, \text{ and } p_{i,t} > p_{j,t} \\
p_{j,t} - p_{i,t} & \text{if } \tilde{p}_{i,t} > \tilde{p}_{j,t}, \text{ and } p_{i,t} < p_{j,t} 
\end{cases}
\]
which can be uniformly upper bounded by $|p_{i,t} - \hat{p}_{i,t}| + |p_{j,t} - \hat{p}_{j,t}|$.

We now specialize the above to the two sequences $(\hat{p}_{j_1,t}, \ldots, \hat{p}_{j_{S_t},t}, 0, \ldots, 0)$ and $(p_{i_1,t}, \ldots, p_{i_{S_t},t}, 0, \ldots, 0)$, and use $\ell_{\text{rank},t}$ instead of $\ell_{\text{rank}}$. Setting $Y_t = \{j_1, j_2, \ldots, j_{S_t}\}$ and $Y_t^* = \{i_1, i_2, \ldots, i_{S_t}\}$, and proceeding as in Lemma 7 we can write

$$
E_t[\ell_{\text{rank},t}(Y_t, (\hat{p}_{j_1,t}, \ldots, \hat{p}_{j_{S_t},t}))] - E_t[\ell_{\text{rank},t}(Y_t, (p_{i_1,t}, \ldots, p_{i_{S_t},t}))]
$$

as claimed.

\[ \square \]

**Lemma 9.** Let $L : D \subseteq \mathbb{R} \to \mathbb{R}^+$ be a $C^2(D)$ convex and nonincreasing function of its argument, $(u_1, \ldots, u_K) \in \mathbb{R}^K$ be defined in (2) with $g(\Delta) = -L'(\Delta)$ for all $\Delta \in D$, and such that $\|u_i\| \leq U$ for all $i \in [K]$. Assume there are positive constants $c'_L$ and $c''_L$ with $(L'(\Delta))^2 \leq c'_L$ and $L''(\Delta) \geq c''_L$ for all $\Delta \in D$. With the notation introduced in Figure 1, we have that

$$
(x^T u_{i,t} - u_t^T x)^2 \leq x^T A_{i,t-1} x \left( U^2 + \frac{d c''_L}{(c'_L)^2} \ln \left( 1 + \frac{1 - 1}{d} \right) + \frac{12}{c'_L} \left( c''_L + 3L(-R) \right) \ln \frac{K(t+4)}{\delta} \right)
$$

holds with probability at least $1 - \delta$ for any $\delta < 1/e$, uniformly over $i \in [K]$, $t = 1, 2, \ldots$, and $x \in \mathbb{R}^d$.

**Proof:** For any given class $i$, the time-$t$ update rule $u_{i,t} \to u_{i,t+1} \to u_{i,t+1}$ in Figure 1 allows us to start off from [7] (proof of Theorem 2 therein), from which one can extract the following inequality

$$
d_{i,t-1}(u_i, u_{i,t}) \leq U^2 + \frac{1}{(c'_L)^2} \sum_{k=1}^{t-1} r_{i,k} - \frac{2}{c'_L} \sum_{k=1}^{t-1} \left( \nabla_{i,k}^T (w_{i,k}^T - u_i) - \frac{c''_L}{2} (s_{i,k} x_k^T (w_{i,k}^T - u_i))^2 \right),
$$

where we set $r_{i,k} = \nabla_{i,k} A_{i,k}^{-1} \nabla_{i,k}$. Using the lower bound on the second derivative of $L$ we have

$$
L(s_{i,k} x_k^T w_{i,k}^T) - L(s_{i,k} u_i^T x_k)
$$

where

$$
L'(s_{i,k} x_k^T w_{i,k}^T)(s_{i,k} x_k^T w_{i,k}^T - s_{i,k} u_i^T x_k) - \frac{c''_L}{2} (s_{i,k} x_k^T w_{i,k}^T - s_{i,k} u_i^T x_k)^2
$$

Plugging back into (8) yields

$$
d_{i,t-1}(u_i, u_{i,t}) \leq U^2 + \frac{1}{(c'_L)^2} \sum_{k=1}^{t-1} r_{i,k} - \frac{2}{c'_L} \sum_{k=1}^{t-1} \left( L(s_{i,k} x_k^T w_{i,k}^T) - L(s_{i,k} u_i^T x_k) \right)
$$

We now borrow a proof technique from [4] (see also [1, 5] and references therein). Define $L_{i,k} = L(s_{i,k} x_k^T w_{i,k}^T)$ and $L'_{i,k} = E_{\delta} [ 1_{i \in Y_k} ]$. Notice that the sequence of random variables $L_{i,k}'$, $L'_{i,2}$, \ldots, forms a martingale difference sequence such that, for any $i \in \hat{Y}_k$: 

12
i. \( \mathbb{E}[L_{i,k}] \geq 0 \), by Lemma 6;

ii. \( |L'_{i,k}| \leq 2L(-R) \), since \( L() \) is nonincreasing over \( D \), and \( s_{i,k} x_k^\top w_{i,k}' \), \( s_{i,k} u_i^\top x_k \in D \);

iii. \( \text{Var}_k(L'_{i,k}) = \text{Var}_k(L_{i,k}) \leq \frac{2c'_k}{c_L} \mathbb{E}[L_{i,k}] \) (again, because of Lemma 6).

On the other hand, when \( i \notin \hat{Y}_k \) then \( s_{i,k} = 0 \), and the above three properties are trivially satisfied. Under the above conditions, we are in a position to apply any fast concentration result for bounded martingale difference sequences. For instance, setting for brevity \( B = B(t, \delta) = 3 \ln \frac{K(t+4)}{\delta} \), [8] allows us derive the inequality

\[
\sum_{k=1}^{t-1} \mathbb{E}[L_{i,k}] - \sum_{k=1}^{t-1} L_{i,k} \geq \max \left\{ \sqrt{\frac{8c'_k}{c_L} B \sum_{k=1}^{t-1} \mathbb{E}[L_{i,k}], 6L(-R) B} \right\},
\]

that holds with probability at most \( \frac{\delta}{K(t+1)} \) for any \( t \geq 1 \). We use the inequality \( \sqrt{cB} \leq \frac{1}{2}(c + b) \) with \( c = \frac{4c'_k}{c_L} B \), and \( b = 2 \sum_{k=1}^{t-1} \mathbb{E}[L_{i,k}] \), and simplify. This gives

\[
- \sum_{k=1}^{t-1} L_{i,k} \leq \left( \frac{2c'_k}{c_L} + 6L(-R) \right) B
\]

with probability at least \( 1 - \frac{\delta}{K(t+1)} \). Using the Cauchy-Schwarz inequality

\[
(x^\top w'_{i,t} - u_i^\top x)^2 \leq x^\top A^{-1}_{i,t-1} x d_{i,t-1}(u_i, w'_{i,t})
\]

holding for any \( x \in \mathcal{R}^d \), and replacing back into (9) allows us to conclude that

\[
(x^\top w'_{i,t} - u_i^\top x)^2 \leq x^\top A^{-1}_{i,t-1} x \left( U^2 + \frac{1}{(c'_d)^2} \sum_{k=1}^{t-1} r_{i,k} + \frac{12}{c'_d} \left( \frac{c'_d}{c_L} + 3L(-R) \right) \ln \frac{K(t+4)}{\delta} \right)
\]

holds with probability at least \( 1 - \frac{\delta}{K(t+1)} \), uniformly over \( x \in \mathcal{R}^d \).

The bounds on \( \sum_{k=1}^{t-1} r_{i,k} \) can be obtained in a standard way. Applying known inequalities (e.g., [2, 3, 5, 7]), and using the fact that \( \nabla_{i,k} = L'(s_{i,k} x_k^\top w_{i,k}') s_{i,k} x_k \) we have

\[
\sum_{k=1}^{t-1} r_{i,k} = \sum_{k=1}^{t-1} |s_{i,j}| \left( L'(s_{i,k} x_k^\top w_{i,k}') \right)^2 x_k^\top A^{-1}_{i,k} x_k
\]

\[
\leq c'_k \sum_{k=1}^{t-1} |s_{i,k}| x_k^\top A^{-1}_{i,k} x_k
\]

\[
\leq c'_k \sum_{k=1}^{t-1} \ln \frac{|A_{i,k}|}{|A_{i,k-1}|}
\]

\[
= c'_k \ln \frac{|A_{i,t-1}|}{|A_{i,0}|}
\]

\[
\leq d c'_k \ln \left( 1 + \frac{t-1}{d} \right)
\]

Piecing together as in (10) and stratifying over \( t = 1, 2, \ldots \), and \( i \in [K] \) concludes the proof. \( \square \)

We are now ready to put all pieces together.
Proof: [Theorem 2] From Lemma 7 and Lemma 9, we see that with probability at least $1 - \delta$,

$$R_T \leq 2 (1 - a) c_L \sum_{t=1}^{T} \sum_{i \in Y_t} \epsilon_{i,t},$$

(11)

when $\epsilon_{i,t}^2$ is the one given in Figure 1. We continue by proving a pointwise upper bound on the sum in the RHS. More in detail, we will find an upper bound on $\sum_{t=1}^{T} \sum_{i \in Y_t} \epsilon_{i,t}^2$, and then derive a resulting upper bound on the RHS of (11).

From Lemma 9 and the update rule (Step 5) of the algorithm we can write

$$\epsilon_{i,t}^2 \leq C x_i^T A_{i,t-1} x_t$$

$$= C \frac{x_i^T (A_{i,t-1} + |s_{i,t}| x_i x_i^T)^{-1} x_t}{1 - |s_{i,t}| x_i^T (A_{i,t-1} + |s_{i,t}| x_i x_i^T)^{-1} x_t}$$

$$= C \frac{x_i^T A_{i,t-1} x_t}{1 - |s_{i,t}| x_i^T (A_0 + |s_{i,t}| x_i x_i^T)^{-1} x_t}$$

$$= C \frac{x_i^T A_{i,t-1} x_t}{1 - \frac{3}{2}}$$

$$= 2 C x_i^T A_{i,t-1} x_t.$$

Hence, if we set $r_{i,t} = x_i^T A_{i,t-1} x_t$ and proceed as in the proof of Lemma 9, we end up with the upper bound $\sum_{t=1}^{T} \epsilon_{i,t}^2 \leq 2 C d \ln (1 + \frac{T}{d})$, holding for all $i \in [K]$. Denoting by $M$ the quantity $2 C d \ln (1 + \frac{T}{d})$, we conclude from (11) that

$$R_T \leq 2 (1 - a) c_L \max \left\{ \sum_{i \in [K]} \sum_{t=1}^{T} \epsilon_{i,t} \left| \sum_{t=1}^{T} \epsilon_{i,t}^2 \leq M, \ i \in [K] \right. \right\} = 2 (1 - a) c_L K \sqrt{T M},$$

as claimed. □

Proof: [Theorem 3] As we said, we change the definition of $\epsilon_{i,t}^2$ in the Algorithm in Figure 1 to

$$\epsilon_{i,t}^2 =$$

$$\max \left\{ x^T A_{i,t-1} x \left( \frac{2 d c'_L}{(c'_L)^2} \ln \left( 1 + \frac{t - 1}{d} \right) + 12 \frac{c'_L}{c_L} \left( \frac{c'_L}{c_L} + 3 L (-R) \right) \ln \frac{K (t + 4)}{\delta} \right), 4 R^2 \right\}. $$

First, notice that the $4 R^2$ cap seamlessly applies, since $(x^T w_{i,t} - u_i^T x)^2$ in Lemma 9 is bounded by $4 R^2$ anyway. With this modification, we have that Theorem 2 only holds for $t$ such that $\frac{d c'_L}{(c'_L)^2} \ln \left( 1 + \frac{t - 1}{d} \right) \geq U^2$, i.e., for $t \geq d \left( \exp \left( \frac{(c'_L)^2 U^2}{c_L d} \right) - 1 \right) + 1$, while for $t < d \left( \exp \left( \frac{(c'_L)^2 U^2}{c_L d} \right) - 1 \right) + 1$ we have in the worst-case scenario the maximum amount of regret at each step. From Lemma 7 we see that this maximum amount (the cap on $\epsilon_{i,t}^2$ is needed here) can be bounded by $4 (1 - a) c_L |\hat{Y}_t| R \leq 4 (1 - a) c_L K R$. □
Proof: [Theorem 4] We start from the one step-regret delivered by Lemma 8, and proceed as in the proof of Theorem 2. This yields

\[ R_T \leq 2c_L \sum_{t=1}^{T} S_t \sum_{i \in \hat{Y}_t} \epsilon_{i,t} \]

\[ \leq 2S \sum_{t=1}^{T} \sum_{i \in \hat{Y}_t} \epsilon_{i,t} \]

\[ \leq 2S \sum_{t=1}^{T} \sum_{i \in [K]} \epsilon_{i,t} \]

\[ = 2S \sum_{i \in [K]} \sum_{t=1}^{T} \epsilon_{i,t} \]

with probability at least \(1 - \delta\), where \(\epsilon_{i,t}^2\) is the one given in Figure 1. Let \(M\) be as in the proof of Theorem 2. If \(N_{i,T}\) denotes the total number of times class \(i\) occurs in \(\hat{Y}_t\), we have that \(\sum_{t=1}^{T} \epsilon_{i,t}^2 \leq M\), implying \(\sum_{t=1}^{T} \epsilon_{i,t} \leq \sqrt{N_{i,T} M}\) for all \(i \in [K]\). Moreover, \(\sum_{i \in [K]} N_{i,T} \leq ST\). Hence

\[ R_T \leq 2S \sum_{i \in [K]} \sqrt{N_{i,T} M} \leq 2c_L \sqrt{MSKT} , \]

as claimed. \(\square\)

References


