Supplementary Material to "On Multilabel Classification and Ranking with Partial Feedback"

Claudio Gentile
DiSTA, Università dell'Insubria, Italy
claudio.gentile@uninsubria.it

Francesco Orabona
TTI Chicago, USA
francesco@orabona.com

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 $Y_s = \{j_1, j_2, \dots, j_s\}$, we have $\mathbb{E}_t[\ell_{a,c}(Y^*_t, Y_s)] = (1-a) \sum_{i \in Y_s} \left(c(j_i, s) - \left(\frac{a}{1-a} + c(j_i, s) \right) p_{i,t} \right)$. 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 nonicreasing order of $p_{i,t}$. By the sake of contradiction, assume there are i and k in $Y^*_{s,t}$ such that i preceeds 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

$$\begin{aligned}
& c(j_1, s) - \left(\tfrac{a}{1-a} + c(j_1, s) \right) p_{i,t} + c(j_2, s) - \left(\tfrac{a}{1-a} + c(j_2, s) \right) p_{k,t} \\
& - \left(c(j_1, s) - \left(\tfrac{a}{1-a} + c(j_1, s) \right) p_{k,t} \right) - \left(c(j_2, s) - \left(\tfrac{a}{1-a} + c(j_2, s) \right) p_{i,t} \right) \\
& = (p_{k,t} - p_{i,t}) \left(c(j_1, s) - c(j_2, s) \right) > 0 \, .
\end{aligned}$$

Hence switching would get a smaller loss which leads as a consequence to $Y^*_{s,t} = (j_1, j_2, \dots, j_s)$. □

The algorithm in Figure 1 works by updating through the gradients $\nabla_{i,t}$ of a modular margin-based loss function $\sum_{i=1}^K L(\boldsymbol{w}_i^\top \boldsymbol{x})$ associated with the label generation model (2) so as to make the parameters $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_K) \in \mathcal{R}^{dK}$ therein achieve the Bayes optimality condition

$$(\boldsymbol{u}_1, \dots, \boldsymbol{u}_K) = \arg \min_{\boldsymbol{w}_1, \dots, \boldsymbol{w}_K \,:\, \boldsymbol{w}_i^\top \boldsymbol{x}_t \in D} \mathbb{E}_t \left[\sum_{i=1}^K L(s_{i,t} \, \boldsymbol{w}_i^\top \boldsymbol{x}_t) \right] \, , \tag{4}$$

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 $\hat{Y}_t$). 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 $\boldsymbol{w}_1, \dots, \boldsymbol{w}_K \in \mathcal{R}^{dK}$ be arbitrary weight vectors such that $\boldsymbol{w}_i^\top \boldsymbol{x}_t \in D$, $i \in [K]$, $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_K) \in \mathcal{R}^{dK}$ 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*

$$\mathbb{E}_t \left[\sum_{i=1}^K L(s_{i,t} \, \boldsymbol{w}_i^\top \boldsymbol{x}_t) \right] \geq \mathbb{E}_t \left[\sum_{i=1}^K L(s_{i,t} \, \boldsymbol{u}_i^\top \boldsymbol{x}_t) \right] ,$$

i.e., (4) holds.

Proof: Let us introduce the shorthands $\Delta_i = \boldsymbol{u}_i^\top \boldsymbol{x}_t$, $\widehat{\Delta}_i = \boldsymbol{w}_{i,t}^\top \boldsymbol{x}_t$, $s_i = s_{i,t}$, and $p_i = \mathbb{P}(y_{i,t} = 1 \,|\, \boldsymbol{x}_t) = \frac{L'(-\Delta_i)}{L'(\Delta_i)+L'(-\Delta_i)}$. Moreover, let $\mathbb{P}_t(\cdot)$ be an abbreviation for the conditional probability $\mathbb{P}(\cdot \,|\, (y_1, \boldsymbol{x}_1), \ldots, (y_{t-1}, \boldsymbol{x}_{t-1}), \boldsymbol{x}_t)$. Recalling the way $s_{i,t}$ is constructed (Figure 1), we can write

$$\begin{aligned}\mathbb{E}_t\left[\sum_{i=1}^K L(s_{i,t}\,\widehat{\Delta}_i)\right] &= \sum_{i\in\hat{Y}_t}\Big(\mathbb{P}_t(s_{i,t}=1)\,L(\widehat{\Delta}_i) + \mathbb{P}_t(s_{i,t}=-1)\,L(-\widehat{\Delta}_i)\Big) + (K-|\hat{Y}_t|)\,L(0)\\ &= \sum_{i\in\hat{Y}_t}\Big(p_i\,L(\widehat{\Delta}_i) + (1-p_i)\,L(-\widehat{\Delta}_i)\Big) + (K-|\hat{Y}_t|)\,L(0)\;,\end{aligned}$$

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 $\mathbb{E}_t\left[\sum_{i=1}^K L(s_{i,t}\,\widehat{\Delta}_i)\right]$ when viewed as a function of the $\widehat{\Delta}_i$. We have that $\frac{\partial\,\mathbb{E}_t\left[\sum_{i=1}^K L(s_{i,t}\,\widehat{\Delta}_i)\right]}{\partial\widehat{\Delta}_i} = 0$ if and only if for all $i \in \hat{Y}_t$ we have that $\widehat{\Delta}_i$ satisfies

$$p_i = \frac{L'(-\widehat{\Delta}_i)}{L'(\widehat{\Delta}_i) + L'(-\widehat{\Delta}_i)}\;.$$

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

Let now $Var_t(\cdot)$ be a shorthand for $Var(\cdot \,|\, (y_1, \boldsymbol{x}_1), \ldots, (y_{t-1}, \boldsymbol{x}_{t-1}), \boldsymbol{x}_t)$. 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 $(\boldsymbol{w}'_{1,t}, \ldots, \boldsymbol{w}'_{K,t}) \in \mathcal{R}^{dK}$ 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 5) of time t, and $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K) \in \mathcal{R}^{dK}$ be the comparison vectors defined through (2). Let $L : D = [-R, R] \subseteq \mathcal{R} \to \mathcal{R}^+$ be a $C^2(D)$ convex function of its argument, with $g(\Delta) = -L'(\Delta)$ and such that there are positive constants c'_L and c''_L with $(L'(\Delta))^2 \le c'_L$ and $L''(\Delta) \ge c''_L$ for all $\Delta \in D$. Then for any $i \in \hat{Y}_t$*

$$0 \le Var_t\left(L(s_{i,t}\,\boldsymbol{x}_t^\top \boldsymbol{w}'_{i,t}) - L(s_{i,t}\,\boldsymbol{u}_i^\top \boldsymbol{x}_t)\right) \le \frac{2c'_L}{c''_L}\,\mathbb{E}_t\left[L(s_{i,t}\,\boldsymbol{x}_t^\top \boldsymbol{w}'_{i,t}) - L(s_{i,t}\,\boldsymbol{u}_i^\top \boldsymbol{x}_t)\right]\;.$$

Proof: Let us introduce the shorthands $\Delta_i = \boldsymbol{x}_t^\top \boldsymbol{u}_i$, $\widehat{\Delta}_i = \boldsymbol{x}_t^\top \boldsymbol{w}'_{i,t}$, $s_i = s_{i,t}$, and $p_i = \mathbb{P}(y_{i,t} = 1 \,|\, \boldsymbol{x}_t) = \frac{L'(-\Delta_i)}{L'(\Delta_i)+L'(-\Delta_i)}$. Then, for any $i \in [K]$,

$$Var_t\left(L(s_{i,t}\,\boldsymbol{x}_t^\top \boldsymbol{w}'_{i,t}) - L(s_{i,t}\,\boldsymbol{u}_i^\top \boldsymbol{x}_t)\right) \le \mathbb{E}_t\left(\Big(L(s_i\,\widehat{\Delta}_i) - L(s_i\,\Delta_i)\Big)^2\right) \le c'_L\,(\widehat{\Delta}_i - \Delta_i)^2\;. \tag{5}$$

Moreover, for any $i \in \hat{Y}_t$ we can write

$$\begin{aligned}
\mathbb{E}_t\left[L(s_i\,\widehat{\Delta}_i) - L(s_i\,\Delta_i)\right] &= p_i\,(L(\widehat{\Delta}_i) - L(\Delta_i)) + (1-p_i)\,(L(-\widehat{\Delta}_i) - L(-\Delta_i)) \\
&\geq p_i\,\left(L'(\Delta_i)(\widehat{\Delta}_i - \Delta_i) + \frac{c''_L}{2}(\widehat{\Delta}_i - \Delta_i)^2\right) \\
&\quad + (1-p_i)\,\left(L'(-\Delta_i)(\Delta_i - \widehat{\Delta}) + \frac{c''_L}{2}(\widehat{\Delta}_i - \Delta_i)^2\right) \\
&= p_i\,\frac{c''_L}{2}(\widehat{\Delta}_i - \Delta_i)^2 + (1-p_i)\,\frac{c''_L}{2}(\widehat{\Delta}_i - \Delta_i)^2 \\
&= \frac{c''_L}{2}(\widehat{\Delta}_i - \Delta_i)^2, \qquad (6)
\end{aligned}$$

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

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 $L : D = [-R, R] \subseteq \mathcal{R} \to \mathcal{R}^+$ be a convex, twice differentiable, and nonincreasing function of its argument. Let $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K) \in \mathcal{R}^{dK}$ be defined in (2) with $g(\Delta) = -L'(\Delta)$ for all $\Delta \in D$. Let also c_L be a positive constant such that*

$$\frac{L'(\Delta)\,L''(-\Delta) + L''(\Delta)\,L'(-\Delta)}{(L'(\Delta) + L'(-\Delta))^2} \geq -c_L$$

holds for all $\Delta \in D$. Finally, let $\Delta_{i,t}$ denote $\boldsymbol{u}_i^\top \boldsymbol{x}_t$, and $\widehat{\Delta}'_{i,t}$ denote $\boldsymbol{x}_t^\top \boldsymbol{w}'_{i,t}$, where $\boldsymbol{w}'_{i,t}$ is the i-the weight vector computed by the algorithm at the beginning (Step 2) of time t. If time t is such that $|\Delta_{i,t} - \widehat{\Delta}'_{i,t}| \leq \epsilon_{i,t}$ for all $i \in [K]$, then

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

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

$$\begin{aligned}
&\mathbb{E}_t[\ell_{a,c}(Y_t, \hat{Y}_t)] - \mathbb{E}_t[\ell_{a,c}(Y_t, Y_t^*)] \\
&\qquad = (1-a) \sum_{i \in \hat{Y}_t} \left(c(\hat{j}_i, |\hat{Y}_t|) - \left(\tfrac{a}{1-a} + c(\hat{j}_i, |\hat{Y}_t|)\right)\,p(\Delta_{i,t})\right) \\
&\qquad\quad - (1-a) \sum_{i \in Y_t^*} \left(c(j_i^*, |Y_t^*|) - \left(\tfrac{a}{1-a} + c(j_i^*, |Y_t^*|)\right)\,p(\Delta_{i,t})\right),
\end{aligned}$$

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,

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

since $g(\Delta) = -L'(\Delta)$, and $L(\cdot)$ is convex and nonincreasing. Hence $p(\Delta)$ is itself a nondecreasing function of Δ. 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 $|\Delta_{i,t} - \widehat{\Delta}'_{i,t}| \leq \epsilon_{i,t}$ and the definition of $\hat{Y}_t$ we

can write

$$
\begin{aligned}
&\mathbb{E}_t[\ell_{a,c}(Y_t,\hat{Y}_t)] - \mathbb{E}_t[\ell_{a,c}(Y_t,Y_t^*)] \\
&\qquad \le (1-a)\sum_{i\in\hat{Y}_t}\Big(c(\hat{j}_i,|\hat{Y}_t|) - \Big(\tfrac{a}{1-a} + c(\hat{j}_i,|\hat{Y}_t|)\Big)\, p([\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D)\Big) \\
&\qquad\quad - (1-a)\sum_{i\in Y_t^*}\Big(c(j_i^*,|Y_t^*|) - \Big(\tfrac{a}{1-a} + c(j_i^*,|Y_t^*|)\Big)\, p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D)\Big) \\
&\qquad \le (1-a)\sum_{i\in\hat{Y}_t}\Big(c(\hat{j}_i,|\hat{Y}_t|) - \Big(\tfrac{a}{1-a} + c(\hat{j}_i,|\hat{Y}_t|)\Big)\, p([\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D)\Big) \\
&\qquad\quad - (1-a)\sum_{i\in\hat{Y}_t}\Big(c(\hat{j}_i,|\hat{Y}_t|) - \Big(\tfrac{a}{1-a} + c(\hat{j}_i,|\hat{Y}_t|)\Big)\, p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D)\Big) \\
&\qquad = (1-a)\sum_{i\in\hat{Y}_t}\Big(c(\hat{j}_i,|\hat{Y}_t|)\Big(p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D) - p([\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D)\Big)\Big) \\
&\qquad \le 2\,(1-a)\,c_L\sum_{i\in\hat{Y}_t}\epsilon_{i,t}\ ,
\end{aligned}
$$

the last inequality deriving from $c(i,s) \le 1$ for all $i \le s \le K$, and

$$
p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D) - p([\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D) \le c_L\big([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D - [\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D\big) \le 2\,c_L\,\epsilon_{i,t}. \quad \square
$$

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 $\boldsymbol{w}'_{i,t}$ be the i-th weight vector computed by this algorithm at the beginning (Step 2) of time t. If this algorithm ranks classes as $\widehat{p}_{j_1,t} \ge \ldots \ge \widehat{p}_{j_{S_t},t} \ge 0$, and time t is such that $|\Delta_{i,t} - \widehat{\Delta}'_{i,t}| \le \epsilon_{i,t}$ for all $i \in [K]$, then*

$$
\begin{aligned}
\mathbb{E}_t[\ell_{rank,t}(Y_t,(\widehat{p}_{j_1,t},\ldots,\widehat{p}_{j_{S_t},t},0,\ldots,0))] - \mathbb{E}_t[\ell_{rank,t}(Y_t,(p_{i_1,t},\ldots,p_{i_{S_t},t},0,\ldots,0))] \\
\le 2\,S_t\,c_L\sum_{i\in\hat{Y}_t}\epsilon_{i,t}\ ,
\end{aligned}
$$

where the $p_{i,t} = \mathbb{P}_t(y_{i,t} = 1 \,|\, \boldsymbol{x}_t)$ are sorted as $p_{i_1,t} \ge \ldots \ge p_{i_{S_t},t} \ge 0$, and $\hat{Y}_t = (j_1, j_2, \ldots, j_{S_t})$.

Proof: Recall the notation $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \,|\, \boldsymbol{x}_t)$, 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 $(\widehat{p}_{1,t},\ldots,\widehat{p}_{K,t})$ and $(p_{1,t},\ldots,p_{K,t})$ one has

$$
\begin{aligned}
&\mathbb{E}_t[\ell_{rank}(Y_t,(\widehat{p}_{1,t},\ldots,\widehat{p}_{K,t}))] - \mathbb{E}_t[\ell_{rank}(Y_t,(p_{1,t},\ldots,p_{K,t}))] \\
&\qquad = \sum_{i,j\in[K],\, i<j}\left(\widehat{r}(i,j) - r(i,j) + \widehat{r}(j,i) - r(j,i)\right), \qquad (7)
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{r}(i,j) &= \mathbb{P}_t(y_{i,t} > y_{j,t})\ \big(\{\widehat{p}_{i,t} < \widehat{p}_{j,t}\} + \tfrac{1}{2}\{\widehat{p}_{i,t} = \widehat{p}_{j,t}\}\big) \\
r(i,j) &= \mathbb{P}_t(y_{i,t} > y_{j,t})\ \big(\{p_{i,t} < p_{j,t}\} + \tfrac{1}{2}\{p_{i,t} = p_{j,t}\}\big)
\end{aligned}
$$

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

$$
\widehat{r}(i,j) - r(i,j) + \widehat{r}(j,i) - r(j,i) = \begin{cases}
\frac{1}{2}\,(p_{i,t} - p_{j,t}) & \text{If } \widehat{p}_{i,t} < \widehat{p}_{j,t},\ p_{i,t} = p_{j,t} \text{ or } \widehat{p}_{i,t} = \widehat{p}_{j,t},\ p_{i,t} > p_{j,t} \\
\frac{1}{2}\,(p_{j,t} - p_{i,t}) & \text{If } \widehat{p}_{i,t} = \widehat{p}_{j,t},\ p_{i,t} < p_{j,t} \text{ or } \widehat{p}_{i,t} > \widehat{p}_{j,t},\ p_{i,t} = p_{j,t} \\
p_{i,t} - p_{j,t} & \text{If } \widehat{p}_{i,t} < \widehat{p}_{j,t},\ p_{i,t} > p_{j,t} \\
p_{j,t} - p_{i,t} & \text{If } \widehat{p}_{i,t} > \widehat{p}_{j,t},\ p_{i,t} < p_{j,t}\ ,
\end{cases}
$$

which can be uniformly upper bounded by $|p_{i,t} - \widehat{p}_{i,t}| + |p_{j,t} - \widehat{p}_{j,t}|$.

We now specialize the above to the two sequences $(\widehat{p}_{j_1,t}, \ldots, \widehat{p}_{j_{S_t},t}, 0, \ldots, 0)$ and $(p_{i_{1,t}}, \ldots, p_{i_{S_t},t}, 0, \ldots, 0)$, and use $\ell_{rank,t}$ instead of ℓ_{rank}. Setting $\hat{Y}_t = \{j_1, j_2, \ldots, j_{S_t}\}$ and $Y_t^* = \{i_1, i_2, \ldots, i_{S_t}\}$, and proceeding as in Lemma 7 we can write

$$
\begin{aligned}
&\mathbb{E}_t[\ell_{rank,t}(Y_t, (\widehat{p}_{j_1,t}, \ldots, \widehat{p}_{j_{S_t},t})] - \mathbb{E}_t[\ell_{rank,t}(Y_t, (p_{i_1,t}, \ldots, p_{i_{S_t},t}))] \\
&\quad \le \sum_{i,j \in \hat{Y}_t,\, i<j} |p_{i,t} - \widehat{p}_{i,t}| + |p_{j,t} - \widehat{p}_{j,t}| + \sum_{i \in \hat{Y}_t} \Big(p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D) - p([\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D) \Big) \\
&\quad = (S_t - 1) \sum_{i \in \hat{Y}_t} |p_{i,t} - \widehat{p}_{i,t}| + \sum_{i \in \hat{Y}_t} \Big(p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D) - p([\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D) \Big) \\
&\quad = (S_t - 1) \sum_{i \in \hat{Y}_t} |p(\Delta_{i,t}) - p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D)| + \sum_{i \in \hat{Y}_t} \Big(p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D) - p([\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D) \Big) \\
&\quad \le (S_t - 1) \sum_{i \in \hat{Y}_t} c_L \, |\Delta_{i,t} - [\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D| + \sum_{i \in \hat{Y}_t} \Big(p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D) - p([\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D) \Big) \\
&\quad \le (S_t - 1) \sum_{i \in \hat{Y}_t} c_L \, \Big(|\Delta_{i,t} - \widehat{\Delta}'_{i,t}| + \epsilon_{i,t} \Big) + \sum_{i \in \hat{Y}_t} \Big(p([\widehat{\Delta}'_{i,t} + \epsilon_{i,t}]_D) - p([\widehat{\Delta}'_{i,t} - \epsilon_{i,t}]_D) \Big) \\
&\quad \le 2\, S_t\, c_L \sum_{i \in \hat{Y}_t} \epsilon_{i,t},
\end{aligned}
$$

as claimed. □

Lemma 9. *Let $L : D = [-R, R] \subseteq \mathcal{R} \to \mathcal{R}^+$ be a $C^2(D)$ convex and nonincreasing function of its argument, $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K) \in \mathcal{R}^{dK}$ be defined in (2) with $g(\Delta) = -L'(\Delta)$ for all $\Delta \in D$, and such that $\|\boldsymbol{u}_i\| \le U$ for all $i \in [K]$. Assume there are positive constants c'_L and c''_L with $(L'(\Delta))^2 \le c'_L$ and $L''(\Delta) \ge c''_L$ for all $\Delta \in D$. With the notation introduced in Figure 1, we have that*

$$
(\boldsymbol{x}^\top \boldsymbol{w}'_{i,t} - \boldsymbol{u}_i^\top \boldsymbol{x})^2 \le \boldsymbol{x}^\top A_{i,t-1}^{-1} \boldsymbol{x} \left(U^2 + \frac{d\, c'_L}{(c''_L)^2} \ln\left(1 + \frac{t-1}{d}\right) + \frac{12}{c''_L} \left(\frac{c'_L}{c''_L} + 3L(-R) \right) \ln \frac{K(t+4)}{\delta} \right)
$$

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

Proof: For any given class i, the time-t update rule $\boldsymbol{w}'_{i,t} \to \boldsymbol{w}_{i,t+1} \to \boldsymbol{w}'_{i,t+1}$ in Figure 1 allows us to start off from [7] (proof of Theorem 2 therein), from which one can extract the following inequality

$$
\begin{aligned}
&d_{i,t-1}(\boldsymbol{u}_i, \boldsymbol{w}'_{i,t}) \\
&\quad \le U^2 + \frac{1}{(c''_L)^2} \sum_{k=1}^{t-1} r_{i,k} - \frac{2}{c''_L} \sum_{k=1}^{t-1} \left(\nabla_{i,k}^\top (\boldsymbol{w}'_{i,k} - \boldsymbol{u}_i) - \frac{c''_L}{2} \, \big(s_{i,k}\, \boldsymbol{x}_k^\top (\boldsymbol{w}'_{i,k} - \boldsymbol{u}_i) \big)^2 \right),
\end{aligned} \tag{8}
$$

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

$$
\begin{aligned}
&L(s_{i,k}\, \boldsymbol{x}_k^\top \boldsymbol{w}'_{i,k}) - L(s_{i,k}\, \boldsymbol{u}_i^\top \boldsymbol{x}_k) \\
&\quad \le L'(s_{i,k}\, \boldsymbol{x}_k^\top \boldsymbol{w}'_{i,k})(s_{i,k} \boldsymbol{x}_k^\top \boldsymbol{w}'_{i,k} - s_{i,k}\, \boldsymbol{u}_i^\top \boldsymbol{x}_k) - \frac{c''_L}{2} (s_{i,k}\, \boldsymbol{x}_k^\top \boldsymbol{w}'_{i,k} - s_{i,k}\, \boldsymbol{u}_i^\top \boldsymbol{x}_k)^2 \\
&\quad = \nabla_{i,k}^\top (\boldsymbol{w}'_{i,k} - \boldsymbol{u}_i) - \frac{c''_L}{2} \, \big(s_{i,k}\, \boldsymbol{x}_k^\top (\boldsymbol{w}'_{i,k} - \boldsymbol{u}_i) \big)^2 \ .
\end{aligned}
$$

Plugging back into (8) yields

$$
d_{i,t-1}(\boldsymbol{u}_i, \boldsymbol{w}'_{i,t}) \le U^2 + \frac{1}{(c''_L)^2} \sum_{k=1}^{t-1} r_{i,k} - \frac{2}{c''_L} \sum_{k=1}^{t-1} \big(L(s_{i,k}\, \boldsymbol{x}_k^\top \boldsymbol{w}'_{i,k}) - L(s_{i,k}\, \boldsymbol{u}_i^\top \boldsymbol{x}_k) \big) \tag{9}
$$

We now borrow a proof technique from [4] (see also [1, 5] and references therein). Define $L_{i,k} = L(s_{i,k}\, \boldsymbol{x}_k^\top \boldsymbol{w}'_{i,k}) - L(s_{i,k}\, \boldsymbol{u}_i^\top \boldsymbol{x}_k)$ and $L'_{i,k} = \mathbb{E}_k[L_{i,k}] - L_{i,k}$. Notice that the sequence of random variables $L'_{i,1}, L'_{i,2}, \ldots$, forms a martingale difference sequence such that, for any $i \in \hat{Y}_k$:

i. $\mathbb{E}_k[L_{i,k}] \geq 0$, by Lemma 6;

ii. $|L'_{i,k}| \leq 2L(-R)$, since $L(\cdot)$ is nonincreasing over D, and $s_{i,k}\, \boldsymbol{x}_k^\top \boldsymbol{w}'_{i,k}$, $s_{i,k}\, \boldsymbol{u}_i^\top \boldsymbol{x}_k \in D$;

iii. $Var_k(L'_{i,k}) = Var_k(L_{i,k}) \leq \frac{2c'_L}{c''_L}\, \mathbb{E}_k[L_{i,k}]$ (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 trivally 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

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

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

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

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

$$(\boldsymbol{x}^\top \boldsymbol{w}'_{i,t} - \boldsymbol{u}_i^\top \boldsymbol{x})^2 \leq \boldsymbol{x}^\top A_{i,t-1}^{-1}\, \boldsymbol{x}\, d_{i,t-1}(\boldsymbol{u}_i, \boldsymbol{w}'_{i,t})$$

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

$$(\boldsymbol{x}^\top \boldsymbol{w}'_{i,t} - \boldsymbol{u}_i^\top \boldsymbol{x})^2 \leq \boldsymbol{x}^\top A_{i,t-1}^{-1} \boldsymbol{x} \left(U^2 + \frac{1}{(c''_L)^2} \sum_{k=1}^{t-1} r_{i,k} + \frac{12}{c''_L} \left(\frac{c'_L}{c''_L} + 3L(-R) \right) \ln \frac{K(t+4)}{\delta} \right) \tag{10}$$

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

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 $\nabla_{i,k} = L'(s_{i,k}\, \boldsymbol{x}_k^\top \boldsymbol{w}'_{i,k})\, s_{i,k} \boldsymbol{x}_k$ we have

$$\begin{aligned}
\sum_{k=1}^{t-1} r_{i,k} &= \sum_{k=1}^{t-1} |s_{i,j}|\, (L'(s_{i,k}\, \boldsymbol{x}_k^\top \boldsymbol{w}'_{i,k}))^2\, \boldsymbol{x}_k^\top A_{i,k}^{-1} \boldsymbol{x}_k \\
&\leq c'_L \sum_{k=1}^{t-1} |s_{i,k}| \boldsymbol{x}_k^\top A_{i,k}^{-1} \boldsymbol{x}_k \\
&\leq c'_L \sum_{k=1}^{t-1} \ln \frac{|A_{i,k}|}{|A_{i,k-1}|} \\
&= c'_L \ln \frac{|A_{i,t-1}|}{|A_{i,0}|} \\
&\leq d\, c'_L \ln \left(1 + \frac{t-1}{d} \right) .
\end{aligned}$$

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

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 $1-\delta$,

$$R_T \leq 2\,(1-a)\,c_L \sum_{t=1}^{T} \sum_{i \in \hat{Y}_t} \epsilon_{i,t}, \tag{11}$$

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 $\sum_{t=1}^{T} \sum_{i \in \hat{Y}_t} \epsilon_{i,t}^2$, 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

$$\begin{aligned}
\epsilon_{i,t}^2 &\leq C\, \boldsymbol{x}_t^\top A_{i,t-1}^{-1} \boldsymbol{x}_t \\
&= C\, \frac{\boldsymbol{x}_t^\top (A_{i,t-1} + |s_{i,t}|\, \boldsymbol{x}_t \boldsymbol{x}_t^\top)^{-1} \boldsymbol{x}_t}{1 - |s_{i,t}| \boldsymbol{x}_t^\top (A_{i,t-1} + |s_{i,t}|\, \boldsymbol{x}_t \boldsymbol{x}_t^\top)^{-1} \boldsymbol{x}_t} \\
&= C\, \frac{\boldsymbol{x}_t^\top A_{i,t}^{-1} \boldsymbol{x}_t}{1 - |s_{i,t}| \boldsymbol{x}_t^\top (A_{i,t-1} + |s_{i,t}|\, \boldsymbol{x}_t \boldsymbol{x}_t^\top)^{-1} \boldsymbol{x}_t} \\
&\leq C\, \frac{\boldsymbol{x}_t^\top A_{i,t}^{-1} \boldsymbol{x}_t}{1 - |s_{i,t}| \boldsymbol{x}_t^\top (A_0 + |s_{i,t}|\, \boldsymbol{x}_t \boldsymbol{x}_t^\top)^{-1} \boldsymbol{x}_t} \\
&= C\, \frac{\boldsymbol{x}_t^\top A_{i,t}^{-1} \boldsymbol{x}_t}{1 - \frac{1}{2}} \\
&= 2\, C\, \boldsymbol{x}_t^\top A_{i,t}^{-1} \boldsymbol{x}_t\ .
\end{aligned}$$

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

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

as claimed. □

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

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

First, notice that the $4R^2$ cap seamlessly applies, since $(\boldsymbol{x}^\top \boldsymbol{w}_{i,t}' - \boldsymbol{u}_i^\top \boldsymbol{x})^2$ 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

$$\begin{aligned} R_T &\leq 2\, c_L \sum_{t=1}^{T} S_t \sum_{i \in \hat{Y}_t} \epsilon_{i,t} \\ &\leq 2\, S\, c_L \sum_{t=1}^{T} \sum_{i \in \hat{Y}_t} \epsilon_{i,t} \\ &\leq 2\, S\, c_L \sum_{t=1}^{T} \sum_{i \in [K]} \epsilon_{i,t} \\ &= 2\, S\, c_L \sum_{i \in [K]} \sum_{t=1}^{T} \epsilon_{i,t}, \end{aligned}$$

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

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

as claimed. □

References

[1] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *25th NIPS*, 2011.

[2] K. S. Azoury and M. K. Warmuth. Relative loss bounds for online density estimation with the exponential family of distributions. *Machine Learning*, 43, 2001.

[3] N. Cesa-Bianchi, A. Conconi, and C. Gentile. A second-order perceptron algorithm. In *15th Annual Conference on Computational Learning Theory (COLT 2002)*, pages 121–137, 2002.

[4] K. Crammer and C. Gentile. Multiclass classification with bandit feedback using adaptive regularization. In *28th ICML*, 2011.

[5] O. Dekel, C. Gentile, and K. Sridharan. Robust selective sampling from single and multiple teachers. In *23rd Colt*, 2010.

[6] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hullermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, to appear.

[7] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[8] S. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithm. In *Nips*, 2008.

[9] L.J. Savage. Elicitation of personal probabilities and expectations. *J. of the American Statistical Association*, 336:783–801, 1973.