A Table with Different Settings of GAPTRON

Table 2: Settings of Gaptron

<table>
<thead>
<tr>
<th>Surrogate Loss</th>
<th>Gap map $a$</th>
<th>Learning rate $\eta$</th>
<th>Exploration $\gamma$</th>
<th>Regret</th>
</tr>
</thead>
<tbody>
<tr>
<td>logistic [$1$]</td>
<td>$1 - | p_t^* \geq 0.5 | p_t^*$</td>
<td>$\frac{\ln(2)}{2KX^2}$</td>
<td>$0$</td>
<td>$\frac{KX^2| U |^2}{\ln(2)}$</td>
</tr>
<tr>
<td>bandit logistic [$4$]</td>
<td>$1 - | p_t^* \geq 0.5 | p_t^*$</td>
<td>$\frac{\ln(2)}{2KX^2}$</td>
<td>$0$</td>
<td>$\frac{exp(2DKX^2)X^2D^2}{\ln(2)}$</td>
</tr>
<tr>
<td>bandit logistic [$4$]</td>
<td>$1 - | p_t^* \geq 0.5 | p_t^*$</td>
<td>$\gamma \frac{\ln(2)}{2KX^2}$</td>
<td>$\sqrt{\frac{2KX^2}{I}}$</td>
<td>$2KXD \sqrt{\frac{T}{I}}$</td>
</tr>
<tr>
<td>hinge [$2$]</td>
<td>$1 - \max{| m_t^* &gt; \beta | }$</td>
<td>$\frac{K - 1}{KX^2}$</td>
<td>$0$</td>
<td>$\frac{K^2X^2| U |^2}{2(K - 1)}$</td>
</tr>
<tr>
<td>bandit hinge [$5$]</td>
<td>$1 - \max{| m_t^* &gt; \beta | }$</td>
<td>$\frac{\gamma(K - 1)}{KX^2}$</td>
<td>$\sqrt{\frac{K^2X^2D^2}{2(K - 1)^2T}}$</td>
<td>$2KXD \sqrt{\frac{T}{2}}$</td>
</tr>
<tr>
<td>smooth hinge [$3$]</td>
<td>$(1 - \min(1, m_t^*))^2$</td>
<td>$\frac{1}{KX^2}$</td>
<td>$0$</td>
<td>$2KX^2| U |^2$</td>
</tr>
<tr>
<td>bandit smooth hinge [$6$]</td>
<td>$(1 - \min(1, m_t^*))^2$</td>
<td>$\frac{\gamma}{KX^2}$</td>
<td>$\sqrt{\frac{4K^2X^2D^2}{I}}$</td>
<td>$2DKX \sqrt{2T}$</td>
</tr>
</tbody>
</table>

B Details of Section 3

Proof of Lemma [$7$] As we said before, the updates of $W_t$ are Online Gradient Descent [Zinkevich 2003], which guarantees

$$\sum_{t=1}^{T} \ell_t(W_t) - \ell_t(U) \leq \frac{\| U \|^2}{2\eta} + \sum_{t=1}^{T} \frac{\eta}{2} \| g_t \|^2.$$  \hspace{1cm} (7)

Now, by using (7) we find

$$E \left[ \sum_{t=1}^{T} (1[y_t^* \neq y_t] - \ell_t(U)) \right] = E \left[ \sum_{t=1}^{T} (1[y_t^* \neq y_t]) - \ell_t(W_t)) + \sum_{t=1}^{T} (\ell_t(W_t) - \ell_t(U)) \right] \leq \frac{\| U \|^2}{2\eta} + E \left[ \sum_{t=1}^{T} (1[y_t^* \neq y_t]) - \ell_t(W_t)) + \frac{\eta}{2} \| g_t \|^2 \right] \leq \frac{\| U \|^2}{2\eta} + \frac{K - 1}{K} T + E \left[ \sum_{t=1}^{T} (1 - a_t)1[y_t^* \neq y_t] + a_t \frac{K - 1}{K} - \ell_t(W_t)) + \frac{\eta}{2} \| g_t \|^2 \right],$$  \hspace{1cm} (8)

where in the last inequality we used $(1 - a_t, \gamma) \leq (1 - a_t)$ and $\max(a_t, \gamma) \leq a_t + \gamma$. Adding $E \left[ \sum_{t=1}^{T} \ell_t(U) \right]$ to both sides of equation (8) completes the proof.

C Details of Full Information Multiclass Classification (Section 4)

C.1 Details of Section 4.1

Proof of Theorem [$7$] We will prove the Theorem by showing that the surrogate gap is bounded by 0 and then using Lemma [$1$] The gradient of the logistic loss evaluated at $W_t$ is given by:

$$\nabla \ell_t(W_t) = \frac{1}{\ln(2)} (\bar{p}_t - e_{y_t}) \otimes x_t,$$

where $\bar{p}_t = (\bar{p}_t(1), \ldots, \bar{p}_t(k))^\top$ and $\bar{p}_t(k) = \sigma(W_t, x_t, k)$. 

12
We now split the analysis into the cases in (9). We start with

\[ \eta < 1 \]

where the first inequality is due to Lemma 2 below.

The last case we need to consider is \( y_i^* = y_i \) and \( p_i^* \geq 0.5 \). In this case we use \( 1 - x \leq \log_2 (1 - x) \) for \( x \in [0, 1] \) and obtain

\[
(1 - a_t) I[y_i^* \neq y_i] + a_t \frac{K - 1}{K} - \ell_t(W_{t}) - \frac{\eta}{2} \|g_t\|^2 \\
\leq (1 - a_t) I[y_i^* \neq y_i] + a_t \frac{K - 1}{K} - \ell_t(W_{t}) - \frac{\eta}{\ln(2)} \|x_t\|^2 \log_2(\hat{p}_t(y_i)) \\
\leq (1 - a_t) I[y_i^* \neq y_i] + a_t \frac{K - 1}{K} - \ell_t(W_{t}) - \frac{\eta}{\ln(2)} X^2 \log_2(\hat{p}_t(y_i))
\]

for \( \eta < \frac{\ln(2)}{K X^2} \).

The second case we consider is when \( y_i^* \neq y_i \) and \( p_i^* \geq 0.5 \). In this case we use \( x \leq \frac{1}{2} \log_2 (1 - x) \) for \( x \in [0, 1] \) and obtain

\[
p_i^* + (1 - p_i^*) \frac{K - 1}{K} + \log_2(\hat{p}_t(y_i)) - \frac{\eta}{\ln(2)} X^2 \log_2(\hat{p}_t(y_i)) \\
\leq - \frac{1}{2} \log_2 (1 - p_i^*) - \frac{K - 1}{K} \log_2 (1 - p_i^*) + \log_2(\hat{p}_t(y_i)) - \frac{\eta}{\ln(2)} X^2 \log_2(\hat{p}_t(y_i)) \\
= - \frac{1}{2} \log_2 \left( \sum_{k \neq y_i} \hat{p}_t(k) \right) - \frac{K - 1}{K} \log_2 \left( \sum_{k \neq y_i} \hat{p}_t(k) \right) + \log_2(\hat{p}_t(y_i)) - \frac{\eta}{\ln(2)} X^2 \log_2(\hat{p}_t(y_i)) \\
\leq - \frac{1}{2} \log_2 (\hat{p}_t(y_i)) - \frac{K - 1}{K} \log_2 (\hat{p}_t(y_i)) + \log_2(\hat{p}_t(y_i)) - \frac{\eta}{\ln(2)} X^2 \log_2(\hat{p}_t(y_i)) \\
= \frac{1}{2K} \log_2 (\hat{p}_t(y_i)) - \frac{\eta}{\ln(2)} X^2 \log_2(\hat{p}_t(y_i)),
\]

which is 0 since \( \eta = \frac{\ln(2)}{2K X^2} \).

The last case we need to consider is \( y_i^* = y_i \) and \( p_i^* \geq 0.5 \). In this case we use \( 1 - x \leq - \log_2 (x) \) and obtain

\[
(1 - p_i^*) \frac{K - 1}{K} + \log_2(p_i^*) - \frac{\eta}{\ln(2)} X^2 \log_2(p_i^*) \\
\leq - \frac{K - 1}{K} \log_2(p_i^*) + \log_2(p_i^*) - \frac{\eta}{\ln(2)} X^2 \log_2(p_i^*)
\]

which is bounded by 0 since \( \eta = \frac{\ln(2)}{2K X^2} \).
We now apply Lemma 1, plug in $\gamma = 0$, and use the above to find:

$$
\mathbb{E} \left[ \sum_{t=1}^T \mathbb{I}[y_t^* \neq y_t] \right] \leq \frac{\|U\|^2}{2\eta} + \sum_{t=1}^T \ell_t(U) + \gamma \frac{K - 1}{K} T
$$

\[
+ \sum_{t=1}^T \left((1 - a_t) \mathbb{I}[y_t^* \neq y_t] + a_t \frac{K - 1}{K} - \ell_t(W_t) + \frac{\eta}{2} \|g_t\|^2\right)
\]

\[
\leq \frac{\|U\|^2}{2\eta} + \sum_{t=1}^T \ell_t(U).
\]

Using $\eta = \frac{\ln(2)}{2Rk^2}$ completes the proof.

\[\square\]

**Lemma 2.** Let $\ell_t$ be the logistic loss [1], then

$$
\|\nabla \ell_t(W_t)\|^2 \leq \frac{2}{\ln(2)} \|x_t\|^2 \ell_t(W_t).
$$

**Proof.** We have

$$
\|\nabla \ell_t(W_t)\|^2 = \frac{1}{\ln(2)^2} \|x_t\|^2 \left(\sum_{k=1}^K (\mathbb{I}[y_t = k] - \hat{p}_t(k))^2\right)
$$

\[
\leq \frac{1}{\ln(2)^2} \|x_t\|^2 \left(\sum_{k=1}^K \mathbb{I}[y_t = k] - \hat{p}_t(k)\right)^2
\]

\[
\leq -2 \frac{1}{\ln(2)^2} \|x_t\|^2 \log_2(\hat{p}_t(y_t))
\]

\[
= -2 \frac{1}{\ln(2)^2} \|x_t\|^2 \ell_t(W_t),
\]

where the last inequality follows from Pinsker’s inequality [Cover and Thomas, 1991] Lemma 12.6.1.

\[\square\]

**C.2 Details of Section 4.2**

**Proof of Theorem 2** We will prove the Theorem by showing that the surrogate gap is bounded by 0 and then using Lemma 1. Let $k = \text{arg max}_{k \neq y_t} \langle W_t^k, x_t \rangle$. The gradient of the smooth multiclass hinge loss is given by

$$
\nabla \ell_t(W_t) = \begin{cases} 
(e_k - e_{y_t}) \otimes x_t & \text{if } y_t^* \neq y_t \\
(e_k - e_{y_t}) \otimes x_t & \text{if } y_t^* = y_t \text{ and } m_t^* \leq \beta \\
0 & \text{if } y_t^* = y_t \text{ and } m_t^* > \beta.
\end{cases}
$$

We continue by writing out the surrogate gap:

$$
(1 - a_t) \mathbb{I}[y_t^* \neq y_t] + a_t \frac{K - 1}{K} - \ell_t(W_t) + \frac{\eta}{2} \|g_t\|^2
$$

\[
= \begin{cases} 
m_t^* + (1 - m_t^*) \frac{K - 1}{K} - (1 - m_t(W_t, y_t)) + \eta \|x_t\|^2 & \text{if } y_t^* \neq y_t \text{ and } m_t^* \leq \beta \\
(1 - m_t^*) \frac{K - 1}{K} - (1 - m_t(W_t, y_t)) + \eta \|x_t\|^2 & \text{if } y_t^* = y_t \text{ and } m_t^* \leq \beta \\
1 - (1 - m_t(W_t, y_t)) + \eta \|x_t\|^2 & \text{if } y_t^* \neq y_t \text{ and } m_t^* > \beta \\
0 & \text{if } y_t^* = y_t \text{ and } m_t^* > \beta.
\end{cases}
\]

\[\text{(10)}\]

In the remainder of the proof we will repeatedly use the following useful inequality for whenever $y_t \neq y_t^*$:

$$
m_t^* + m_t(W_t, y_t) = \langle W_t^{y_t^*}, x_t \rangle - \max_{k \neq y_t^*} \langle W_t^k, x_t \rangle + \langle W_t^{y_t^*}, x_t \rangle - \max_{k \neq y_t} \langle W_t^k, x_t \rangle
$$

\[
= \langle W_t^{y_t}, x_t \rangle - \max_{k \neq y_t^*} \langle W_t^k, x_t \rangle \text{ (11)}
\]

\[
\leq \langle W_t^{y_t}, x_t \rangle - \langle W_t^{y_t}, x_t \rangle = 0.
\]
We now split the analysis into the cases in (10). We start with $y_t^* \neq y_t$ and $m_t^* \leq \beta$, in which case the surrogate gap can be bounded by 0 when $\eta \leq \frac{1}{KX^2}$:

$$m_t^* + (1 - m_t^*) \frac{K - 1}{K} - (1 - m_t(W_t, y_t)) + \eta \|x_t\|^2$$

$$= m_t^* + m_t(W_t, y_t) + (1 - m_t^*) \frac{K - 1}{K} - 1 + \eta \|x_t\|^2$$

$$\leq -\frac{1}{K} + \eta X^2$$

(by equation (11))

$$\leq 0.$$

We continue with the case where $y_t^* = y_t$ and $m_t^* \leq \beta$. In this case we have:

$$(1 - m_t^*) \frac{K - 1}{K} - (1 - m_t^*) + \eta \|x_t\|^2 = - (1 - m_t^*) \frac{1}{K} + \eta \|x_t\|^2 \leq -\frac{1}{K} + \eta X^2,$$

which is zero since $\eta = \frac{1 - \beta}{KX^2}$.

Finally, in the case where $y_t^* \neq y_t$ and $m_t^* > \beta$ we have:

$$1 - (1 - m_t(W_t, y_t)) + \eta \|x_t\|^2 = m_t(W_t, y_t) + \eta \|x_t\|^2$$

$$\leq - m_t^* + \eta \|x_t\|^2$$

(by equation (11))

$$\leq - \beta + \eta X^2,$$

which is bounded by zero since $\beta = \frac{1}{K}$ and $\eta \leq \frac{1}{KX^2}$.

We now apply Lemma [1] plug in $\gamma = 0$, and use the above to find:

$$E \left[ \sum_{t=1}^{T} \mathbb{I}[y_t^* \neq y_t] \right] \leq \frac{\|U\|^2}{2\eta} + \sum_{t=1}^{T} \ell_t(U) + \gamma T$$

$$+ \sum_{t=1}^{T} \left( (1 - a_t) \mathbb{I}[y_t^* \neq y_t] + a_t \frac{K - 1}{K} - \ell_t(U) + \eta \frac{1}{2} \|g_t\|^2 \right)$$

$$\leq \frac{\|U\|^2}{2\eta} + \sum_{t=1}^{T} \ell_t(U).$$

Using $\eta = \frac{1 - \beta}{KX^2} = \frac{K - 1}{KX^2}$ completes the proof.

**C.3 Details of Section 4.3**

**Proof of Theorem 4.3** We will prove the Theorem by showing that the surrogate gap is bounded by 0 and then using Lemma 1. Let $\tilde{k} = \arg \max_{k \neq y_t}(W_t^k, x_t)$. The gradient of the smooth multiclass hinge loss is given by

$$\nabla \ell_t(W_t) = \begin{cases} 
2(e_{\tilde{k}} - e_{y_t}) \otimes x_t & \text{if } y_t^* \neq y_t, \\
2(e_{\tilde{k}} - e_{y_t})(1 - m_t^*) \otimes x_t & \text{if } y_t^* = y_t \text{ and } m_t^* < 1, \\
0 & \text{if } y_t^* = y_t \text{ and } m_t^* \geq 1.
\end{cases}$$

We continue by writing out the surrogate gap:

$$\begin{cases} 
2m_t^* - m_t^2 + (1 - m_t^2) \frac{K - 1}{K} - (1 - 2m_t(W_t, y_t)) + \eta 4 \|x_t\|^2 & \text{if } y_t^* \neq y_t \text{ and } m_t^* < 1, \\
(1 - m_t^2) \frac{K - 1}{K} - (1 - m_t^2)^2 + \eta 4 \|x_t\|^2 & \text{if } y_t^* = y_t \text{ and } m_t^* < 1, \\
1 - (1 - 2m_t(W_t, y_t)) + \eta 4 \|x_t\|^2 & \text{if } y_t^* = y_t \text{ and } m_t^* \geq 1.
\end{cases}$$

(12)
We now split the analysis into the cases in (12). We start with the case where \( \eta = \frac{1}{4KX^2} \). By using (11) we can see that with \( \eta \leq \frac{1}{4KX^2} \), the surrogate gap is bounded by 0:

\[
2m_t^* - m_t^{*2} + (1 - m_t^*)^2 \frac{K - 1}{K} - (1 - 2m_t(W_t, y_t)) + \eta 4\|x_t\|^2 \\
= 2(m_t^* + m_t(W_t, y_t)) - m_t^{*2} + (1 - m_t^*)^2 \frac{K - 1}{K} - 1 + \eta 4\|x_t\|^2 \\
\leq -m_t^* + (1 - m_t^*)^2 \frac{K - 1}{K} - 1 + \eta 4X^2 \\
\leq -\frac{1}{K} + \eta 4X^2 \leq 0.
\]

The next case we consider is when \( y_t^* = y_t \) and \( m_t^* < 1 \). In this case we have

\[
(1 - m_t^*)^2 \frac{K - 1}{K} - (1 - m_t^*)^2 + \eta 4\|x_t\|^2(1 - m_t^*)^2 = -(1 - m_t^*)^2 \frac{1}{K} + \eta 4\|x_t\|^2(1 - m_t^*)^2,
\]

which is bounded by 0 since \( \eta = \frac{1}{4KX^2} \).

Finally, if \( y_t^* \neq y_t \) and \( m_t^* \geq 1 \) then

\[
1 - (1 - 2m_t(W_t, y_t)) + \eta 4\|x_t\|^2 = 2m_t(W_t, y_t) + \eta 4\|x_t\|^2 \\
\leq - 2m_t^* + \eta 4\|x_t\|^2 \\
\leq - 2 + \eta 4X^2,
\]

which is bounded by 0 since \( \eta < \frac{1}{4KX^2} \). We apply Lemma 1 with \( \gamma = 0 \) and use the above to find:

\[
E \left[ \sum_{t=1}^{T} \mathbb{I}[y_t^* \neq y_t] \right] \leq \frac{\|U\|^2}{2\eta} + \sum_{t=1}^{T} \ell_t(U) + \frac{\gamma}{2} \|U\|^2 + \frac{\gamma}{2} \|g_t\|^2 \\
+ \sum_{t=1}^{T} \left( (1 - a_t)\mathbb{I}[y_t^* \neq y_t] + a_t \frac{K - 1}{K} - \ell_t(W_t) + \frac{\eta}{2} \|g_t\|^2 \right) \\
\leq \frac{\|U\|^2}{2\eta} + \sum_{t=1}^{T} \ell_t(U).
\]

Using \( \eta = \frac{1}{4KX^2} \) completes the proof.

\[ \square \]

D Details of Bandit Multiclass Classification (Section 5)

D.1 Details of Section 5.1

Proof of Theorem 4. First, by straightforward calculations we can see that \( p_t^*(y_t) \geq \frac{1}{(1-\gamma) \exp(-2DX) + \gamma} = \delta \). As in the full information case we will prove the Theorem by showing that the surrogate gap is bounded by 0 and then using Lemma 1. We start by writing out the surrogate gap:

\[
E \left[ (1 - a_t)\mathbb{I}[y_t^* \neq y_t] + a_t \frac{K - 1}{K} - \mathbb{E}_t[\ell_t(W_t)] + \frac{\eta}{2} \mathbb{E}_t[\|g_t\|^2] \right] \\
= E \left[ (1 - a_t)\mathbb{I}[y_t^* \neq y_t] + a_t \frac{K - 1}{K} + \log_2(\hat{p}_t(y_t)) + \frac{\eta}{2 \ln(2)p_t(y_t)} \|\hat{p}_t - e_{y_t} \otimes x_t\|^2 \right] \\
\leq E \left[ (1 - a_t)\mathbb{I}[y_t^* \neq y_t] + a_t \frac{K - 1}{K} + \log_2(\hat{p}_t(y_t)) - \frac{\eta}{\ln(2)p_t(y_t)} X^2 \log_2(\hat{p}_t(y_t)) \right] \\
\leq \left\{ \begin{array}{ll} \frac{K-1}{K} + E \left[ \log_2(\hat{p}_t(y_t)) - \frac{\eta}{\ln(2)p_t(y_t)} X^2 \log_2(\hat{p}_t(y_t)) \right] & \text{if } p_t^* < 0.5 \\
E \left[ p_t^* + (1 - p_t^*) \frac{K-1}{K} + \log_2(\hat{p}_t(y_t)) - \frac{\eta}{\ln(2)p_t(y_t)} X^2 \log_2(\hat{p}_t(y_t)) \right] & \text{if } p_t^* \geq 0.5 \end{array} \right.
\]

\[
\leq \left\{ \begin{array}{ll} E \left[ (1 - p_t^*) \frac{K-1}{K} + \log_2(p_t^*) - \frac{\eta}{\ln(2)p_t(y_t)} X^2 \log_2(p_t^*) \right] & \text{if } y_t^* = y_t \text{ and } p_t^* \geq 0.5, \end{array} \right.
\]

(13)
where the first inequality is due to Lemma 2.

We now split the analysis into the cases in (13). We start with $p_i^* < 0.5$. In this case we use $1 \leq -\log_2(x)$ for $x \in [0, 1/2]$ and obtain

$$\frac{K - 1}{K} + E\left[\log_2(p_i(y_i)) - \frac{\eta}{\ln(2)p_i(y_i)}X^2 \log_2(p_i(y_i))\right] \leq E\left[-\frac{K - 1}{K} \log_2(p_i(y_i)) + \log_2(p_i(y_i)) - \frac{\eta}{\ln(2)p_i(y_i)}X^2 \log_2(p_i(y_i))\right] \leq E\left[-\frac{K - 1}{K} \log_2(p_i(y_i)) + \log_2(p_i(y_i)) - \frac{\eta}{\ln(2)}X^2 \log_2(p_i(y_i))\right],$$

which is bounded by 0 when $\eta \leq \frac{\ln(2)\delta}{KX^2}$.

The second case we consider is when $y_i^* \neq y_i$ and $p_i^* \geq 0.5$. In this case we use $x \leq \frac{1}{2} \log_2(1 - x)$ for $x \in [0.5, 1]$ and $1 - x \leq \frac{1}{2} \log_2(1 - x)$ for $x \in [0.5, 1]$ and obtain

$$E\left[p_i^* + (1 - p_i^*)\frac{K - 1}{K} + \log_2(p_i(y_i)) - \frac{\eta}{\ln(2)p_i(y_i)}X^2 \log_2(p_i(y_i))\right] \leq E\left[-\frac{1}{2} \log_2(1 - p_i^*) - \frac{K - 1}{K} \log_2(1 - p_i^*) + \log_2(p_i(y_i)) - \frac{\eta}{\ln(2)}X^2 \log_2(p_i(y_i))\right]
= E\left[-\frac{1}{2} \log_2\left(\sum_{k \neq y_i} \hat{p}_i(k)\right) - \frac{K - 1}{K} \frac{1}{2} \log_2\left(\sum_{k \neq y_i} \hat{p}_i(k)\right) + \log_2(p_i(y_i)) - \frac{\eta}{\ln(2)}X^2 \log_2(p_i(y_i))\right]
\leq E\left[-\frac{1}{2} \log_2(p_i(y_i)) - \frac{K - 1}{K} \log_2(p_i(y_i)) + \log_2(p_i(y_i)) - \frac{\eta}{\ln(2)}X^2 \log_2(p_i(y_i))\right]
= E\left[\frac{1}{2K} \log_2(p_i(y_i)) - \frac{\eta}{\ln(2)}X^2 \log_2(p_i(y_i))\right],$$

which is bounded by 0 since $\eta = \frac{\ln(2)\delta}{2KX^2}$.

The last case we need to consider is when $y_i^* = y_i$ and $p_i^* \geq 0.5$. In this case we use $1 - x \leq -\log_2(x)$ and obtain

$$E\left[(1 - p_i^*)\frac{K - 1}{K} + \log_2(p_i(y_i)) - \frac{\eta}{\ln(2)p_i(y_i)}X^2 \log_2(p_i(y_i))\right] \leq E\left[-\frac{K - 1}{K} \log_2(p_i(y_i)) + \log_2(p_i(y_i)) - \frac{\eta}{\ln(2)}X^2 \log_2(p_i(y_i))\right],$$

which is bounded by 0 when $\eta \leq \frac{\ln(2)\delta}{KX^2}$.

We now apply Lemma 1 and use the above to find:

$$E\left[\sum_{t=1}^T \mathbf{1}[y_i^* \neq y_i]\right] \leq \frac{\|U\|^2}{2\eta} + E\left[\sum_{t=1}^T \ell_t(U)\right] + \frac{K - 1}{K} T + E\left[\sum_{t=1}^T \left((1 - a_t)\mathbf{1}[y_i^* \neq y_i] + a_t \frac{K - 1}{K} - \ell_t(W_t) + \frac{\eta}{2}\|g_t\|^2\right)\right]
\leq \frac{\|U\|^2}{2\eta} + \gamma T + E\left[\sum_{t=1}^T \ell_t(U)\right].$$

Using $\eta = \frac{\ln(2)\delta}{2KX^2}$ gives us:

$$E\left[\sum_{t=1}^T \mathbf{1}[y_i^* \neq y_i]\right] \leq \frac{K^2X^2\|U\|^2}{\ln(2)((1 - \gamma) \exp(-2DX) + \gamma)} + \gamma T + E\left[\sum_{t=1}^T \ell_t(U)\right],$$

17
Setting $\gamma = 0$ gives
\[
E \left[ \sum_{t=1}^{T} \mathbb{I}[y'_t \neq y_t] \right] \leq \frac{K^2 X^2 D^2}{\ln(2)} \exp(-2DX) + E \left[ \sum_{t=1}^{T} \ell_t(U) \right].
\]

If instead we set $\gamma = \min \left\{ 1, \sqrt{\frac{K^2 X^2 D^2}{\ln(2)}} \right\}$ we consider two cases. In the case where $1 \leq \sqrt{\frac{K^2 X^2 D^2}{T}}$ we have that $T \leq K^2 X^2 D^2$ and therefore
\[
E \left[ \sum_{t=1}^{T} \mathbb{I}[y'_t \neq y_t] \right] \leq 2 \frac{K^2 X^2 D^2}{\ln(2)} + E \left[ \sum_{t=1}^{T} \ell_t(U) \right].
\]

In the case where $1 > \sqrt{\frac{K^2 X^2 D^2}{T}}$ we have that
\[
E \left[ \sum_{t=1}^{T} \mathbb{I}[y'_t \neq y_t] \right] \leq 2KXD \sqrt{\frac{T}{\ln(2)}} + E \left[ \sum_{t=1}^{T} \ell_t(U) \right],
\]
which after combining the above completes the proof.

D.2 Details of Section 5.2

Proof of Theorem 5.2

First, note that $p'_i(y_t) \geq \frac{\gamma}{X^2}$. The proof proceeds in a similar way as in the full information setting (Theorem 2), except now we use that $p'_i(y_t) \geq \frac{\gamma}{X^2}$ to bound $E_t[||g_t||^2]$. We will prove the Theorem by showing that the surrogate gap is bounded by 0 and then using Lemma 1. We start by splitting the surrogate gap in cases:

\[
E \left[ \left( 1 - a_t \right) \mathbb{I}[y'_t \neq y_t] + a_t \frac{K - 1}{K} - E_t[\ell_t(W_t)] + \frac{\eta}{2} E_t[||g_t||^2] \right]
= \begin{cases} 
E \left[ m'_t + (1 - m'_t) \frac{K - 1}{K} - (1 - m_t(W_t, y_t)) + \frac{\eta}{p'_t(y_t)} ||x_t||^2 \right] & \text{if } y'_t \neq y_t \text{ and } m'_t \leq \beta \\
E \left[ (1 - m'_t) \frac{K - 1}{K} - (1 - m'_t) + \frac{\eta}{p'_t(y_t)} ||x_t||^2 \right] & \text{if } y'_t = y_t \text{ and } m'_t \leq \beta \\
E \left[ 1 - (1 - m_t(W_t, y_t)) + \frac{\eta}{p'_t(y_t)} ||x_t||^2 \right] & \text{if } y'_t \neq y_t \text{ and } m'_t > \beta \\
0 & \text{if } y'_t = y_t \text{ and } m'_t > \beta.
\end{cases}
\]

(14)

We now split the analysis into the cases in (14). We start with $y'_t \neq y_t$ and $m'_t \leq \beta$. The surrogate gap can be now be bounded by 0 when $\eta \leq \frac{K^2 X^2}{\gamma}$:

\[
E \left[ m'_t + (1 - m'_t) \frac{K - 1}{K} - (1 - m_t(W_t, y_t)) + \frac{\eta}{p'_t(y_t)} ||x_t||^2 \right]
= E \left[ m'_t + m_t(W_t, y_t) - 1 - m_t \frac{K - 1}{K} \right] - 1 + \frac{\eta}{p'_t(y_t)} ||x_t||^2 
\leq - \frac{1}{K} + \frac{K \eta}{\gamma} X^2
\]

(equation (11))

\[
\leq 0.
\]

We continue with the case where $y'_t = y_t$ and $m'_t \leq \beta$. In this case we have:

\[
E \left[ (1 - m'_t) \frac{K - 1}{K} - (1 - m'_t) + \eta ||x_t||^2 \right] = E \left[ -1 - m'_t \frac{1}{K} + \frac{\eta}{p'_t(y_t)} ||x_t||^2 \right] \leq - \frac{1 - \beta}{K} + \frac{K \eta}{\gamma} X^2,
\]

which is bounded by zero since $\eta = \frac{\gamma(1-\beta)}{K^2 X^2}$. 

18
Finally, in the case where \( y_t^* \neq y_t \) and \( m_t^* > \beta \) we have:

\[
\mathbb{E} \left[ 1 - \left(1 - m_t(W_t, y_t)\right) + \frac{\eta}{p_t(y_t)} \|x_t\|^2 \right] = \mathbb{E} \left[ m_t(W_t, y_t) + \frac{\eta}{p_t(y_t)} \|x_t\|^2 \right] \\
\leq \mathbb{E} \left[ -m_t^* + \frac{\eta}{p_t(y_t)} \|x_t\|^2 \right] \quad \text{(by equation (11))} \\
\leq - \beta + \frac{K\eta}{\gamma X^2},
\]

which is bounded by zero since \( \eta = \frac{\gamma(1-\beta)}{KX^2} \) and \( \beta \leq 0.5 \).

We now apply Lemma 1 and use the above to find:

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I}[y_t^* \neq y_t] \right] \leq \frac{\|U\|^2}{2\eta} + \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(U) \right] + \gamma \frac{K-1}{K} T
\]

\[
+ \mathbb{E} \left[ \sum_{t=1}^{T} \left( (1 - a_t)\mathbb{I}[y_t^* \neq y_t] + a_t \frac{K-1}{K} \mathbb{I}[x_t, W_t] + \frac{\eta}{2} \|g_t\|^2 \right) \right]
\]

\[
\leq \frac{D^2}{2\eta} + \gamma \frac{K-1}{K} T + \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(U) \right].
\]

Plugging in \( \eta = \frac{\gamma(1-\beta)}{KX^2} \) and \( \beta = \frac{1}{K} \) gives us:

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I}[y_t^* \neq y_t] \right] \leq \frac{K^3X^2D^2}{2\gamma(K-1)} + \gamma \frac{K-1}{K} T + \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(U) \right].
\]

We now set \( \gamma = \min \left\{ 1, \sqrt{\frac{K^3X^2D^2}{2(K-\beta)(K-1)}} \right\} \). In the case where \( 1 \leq \sqrt{\frac{K^3X^2D^2}{2(K-\beta)(K-1)}} \), we have

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I}[y_t^* \neq y_t] \right] \leq \frac{K^3X^2D^2}{K-1} + \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(U) \right].
\]

In the case where \( 1 > \sqrt{\frac{K^3X^2D^2}{2(K-\beta)(K-1)}} \), we have

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{I}[y_t^* \neq y_t] \right] \leq 2KXD\sqrt{\frac{T}{2}} + \mathbb{E} \left[ \sum_{t=1}^{T} \ell_t(U) \right],
\]

which completes the proof. \( \square \)

D.3 Details of Section 5.3

Proof of Theorem 6
First, note that \( p_t'(y_t) \geq \frac{\eta}{K} \). The proof proceeds in a similar way as in the full information case. We will prove the theorem by showing that the surrogate gap is bounded by 0 and then using Lemma 1. We start by writing out the surrogate gap:

\[
\mathbb{E} \left[ (1 - a_t)\mathbb{I}[y_t^* \neq y_t] + a_t \frac{K-1}{K} \mathbb{I}[x_t, W_t] + \frac{\eta}{2} \|g_t\|^2 \right]
\]

\[
= \begin{cases} 
\mathbb{E} \left[ 2m_t^* - m_t^{*2} + \frac{1 - m_t^*}{2} \frac{K-1}{K} - (1 - 2m_t(W_t, y_t)) + \frac{\eta}{2} \|g_t\|^2 \right] 
& \text{if } y_t^* \neq y_t \text{ and } m_t^* < 1 \\
\mathbb{E} \left[ (1 - m_t^*)^2 \frac{K-1}{K} - (1 - m_t^*)^2 + \frac{\eta}{2} \|g_t\|^2 \right] 
& \text{if } y_t^* = y_t \text{ and } m_t^* < 1 \\
\mathbb{E} \left[ 1 - (1 - 2m_t(W_t, y_t)) + \frac{\eta}{2} \|g_t\|^2 \right] 
& \text{if } y_t^* \neq y_t \text{ and } m_t^* \geq 1 \\
0 
& \text{if } y_t^* = y_t \text{ and } m_t^* \geq 1.
\end{cases}
\]
We now split the analysis into the cases in (15). We start with the case where \( y_t^* \neq y_t \) and \( m_t^* < 1 \). By using (11) we can see that for \( \eta = \frac{1}{4KX^2} \)

\[
E \left[ 2m_t^* - m_t^* - (1 - m_t^*)^2 \frac{K - 1}{K} - (1 - 2m_t(W_t, y_t)) + \frac{\eta}{p^t(y_t)} 4||x_t||^2 \right] \\
= E \left[ 2(m_t^* + m_t(W_t, y_t)) - m_t^* - (1 - m_t^*)^2 \frac{K - 1}{K} - 1 + \frac{\eta}{p^t(y_t)} 4||x_t||^2 \right] \\
\leq E \left[ -m_t^* + (1 - m_t^*)^2 \frac{K - 1}{K} - 1 + \frac{\eta}{p^t(y_t)} 4X^2 \right] \\
\leq -\frac{1}{K} + \frac{K\eta}{\gamma} 4X^2 \leq 0.
\]

The next case we consider is when \( y_t^* = y_t \) and \( m_t^* < 1 \). In this case we have

\[
E \left[ (1 - m_t^*)^2 \frac{K - 1}{K} - (1 - m_t^*)^2 + \frac{\eta}{p^t(y_t)} 4||x_t||^2 (1 - m_t^*)^2 \right] \\
= E \left[ - (1 - m_t^*)^2 \frac{1}{K} + \frac{\eta}{p^t(y_t)} 4||x_t||^2 (1 - m_t^*)^2 \right] \\
= E \left[ - (1 - m_t^*)^2 \frac{1}{K} + \frac{K\eta}{\gamma} 4X^2 (1 - m_t^*)^2 \right],
\]

which is bounded by 0 since \( \eta = \frac{1}{4KX^2} \).

Finally, if \( y_t^* \neq y_t \) and \( m_t^* \geq 1 \) then

\[
E \left[ 1 - (1 - 2m_t(W_t, y_t)) + \frac{\eta}{p^t(y_t)} 4||x_t||^2 \right] = E \left[ 2m_t(W_t, y_t) + \frac{\eta}{p^t(y_t)} 4||x_t||^2 \right] \\
\leq E \left[ -2m_t^* + \frac{\eta}{p^t(y_t)} 4||x_t||^2 \right] \quad \text{(by equation (11))} \\
\leq -2 + \frac{K\eta}{\gamma} 4X^2,
\]

which is bounded by 0 since \( \eta < \frac{1}{2K^2X^2} \). We apply Lemma 1 and use the above to find:

\[
E \left[ \sum_{t=1}^T \mathbb{1}[y_t^* \neq y_t] \right] \leq \frac{\|U\|^2}{2\eta} + E \left[ \sum_{t=1}^T \ell_t(U) \right] + \gamma T \\
+ E \left[ \sum_{t=1}^T \left( (1 - a_t)\mathbb{1}[y_t^* \neq y_t] + a_t \frac{K - 1}{K} - \ell_t(W_t) + \frac{\eta}{2} ||g_t||^2 \right) \right] \\
\leq \frac{D^2}{2\eta} + \gamma T + E \left[ \sum_{t=1}^T \ell_t(U) \right].
\]

Plugging in \( \eta = \frac{1}{4KX^2} \) gives us:

\[
E \left[ \sum_{t=1}^T \mathbb{1}[y_t^* \neq y_t] \right] \leq \frac{2K^2X^2D^2}{\gamma} + \gamma T + E \left[ \sum_{t=1}^T \ell_t(U) \right].
\]

Now we set \( \gamma = \min \left\{ 1, \sqrt{\frac{2K^2X^2D^2}{\eta}} \right\} \). In the case where \( 1 \leq \sqrt{\frac{2K^2X^2D^2}{\eta}} \) we have

\[
E \left[ \sum_{t=1}^T \mathbb{1}[y_t^* \neq y_t] \right] \leq 4K^2X^2D^2 + E \left[ \sum_{t=1}^T \ell_t(U) \right].
\]

In the case where \( 1 > \sqrt{\frac{2K^2X^2D^2}{\eta}} \) we have

\[
E \left[ \sum_{t=1}^T \mathbb{1}[y_t^* \neq y_t] \right] \leq 2DKX \sqrt{2T} + E \left[ \sum_{t=1}^T \ell_t(U) \right].
\]
which completes the proof.

\[\square\]

### E Online Classification with Abstention

The online classification with abstention setting was introduced by [Neu and Zhivotovskiy (2020)] and is a special case of the prediction with expert advice setting [Vovk (1990); Littlestone and Warmuth (1994)]. For brevity we only consider the case where there are only 2 labels, -1 and 1. The online classification with abstention setting is different from the standard classification setting in that the learner has access to a third option, abstaining. [Neu and Zhivotovskiy (2020)] show that when the cost for abstaining is smaller than \(\frac{1}{2}\) in all rounds it is possible to tune Exponential Weights such that it suffers constant regret with respect to the best expert in hindsight. [Neu and Zhivotovskiy (2020)] only consider the zero-one loss, but we show that a similar bound also holds for the hinge loss (and also for the zero-one loss as a special case of the hinge loss). We use a different proof technique from [Neu and Zhivotovskiy (2020)], which was the inspiration for the proofs of the mistake bounds of GAPTRON. Instead of vanilla Exponential Weights we use a slight adaptation of ADAHEDGE ([De Rooij et al. (2014)] to prove constant regret bounds when all abstention costs \(c_t\) are smaller than \(\frac{1}{2}\). In online classification with abstention, in each round \(t\):

1. the learner observes the predictions \(y_t^i \in [-1, 1]\) of experts \(i = 1, \ldots, d\)
2. based on the experts’ predictions the learner predicts \(y_t' \in [-1, 1] \cup \ast\), where \(\ast\) stands for abstaining
3. the environment reveals \(y_t \in \{-1, 1\}\)
4. the learner suffers loss \(\ell_t(y_t') = \frac{1}{2}(1 - y_t y_t')\) if \(y_t' \in [-1, 1]\) and \(c_t\) otherwise.

The algorithm we use can be found in Algorithm 2. A parallel result to Lemma 1 can be found in Lemma 3, which we will use to derive the regret of Algorithm 2.

**Lemma 3.** For any expert \(i\), the expected loss of Algorithm 2 satisfies:

\[
\sum_{t=1}^{T} ((1 - b_t)\ell_t(y_t^i) + b_t c_t) \leq \sum_{t=1}^{T} \ell_t(y_t^i) + \inf_{\eta > 0} \left\{ \frac{\ln(d)}{\eta} + \sum_{t=1}^{T} ((1 - b_t)\ell_t(y_t^i) + c_t b_t + \eta v_t - \ell_t(y_t^i)) \right\} + \frac{4}{3} \ln(d) + 2,
\]

where \(v_t = E_{\hat{y}_t \sim p_t}[(\ell_t(\hat{y}_t) - \ell_t(y_t^i))^2].\)

Before we prove Lemma 3, let us compare Algorithm 2 with GAPTRON. The updates of weight matrix \(W_t\) in GAPTRON are performed with OGD. In Algorithm 2 the updates of \(\hat{p}_t\) are performed using ADAHEDGE. The roles of \(a_t\) in GAPTRON and \(b_t\) in Algorithm 2 are similar. The role of \(a_t\) is to ensure that the surrogate gap is bounded by 0, the role of \(b_t\) is to ensure that the abstention gap is bounded by 0.
Proof of Lemma 3. First, ADAHEDGE guarantees that

\[
\sum_{t=1}^{T} \ell_t(\hat{y}_t) - \ell_t(y_t^*) \leq 2 \sqrt{\ln(d) \sum_{t=1}^{T} v_t + 4/3 \ln(d) + 2}.
\]

Using the regret bound of ADAHEDGE we can upper bound the expectation of the loss of the learner as

\[
\sum_{t=1}^{T} ((1 - b_t)\ell_t(y_t^*) + b_tc_t)
\]

\[
= \sum_{t=1}^{T} ((1 - b_t)\ell_t(y_t^*) + b_tc_t + \ell_t(y_t^*) - \ell_t(\hat{y}_t)) + \sum_{t=1}^{T} (\ell_t(\hat{y}_t) - \ell_t(y_t^*))
\]

\[
\leq \sum_{t=1}^{T} ((1 - b_t)\ell_t(y_t^*) + b_tc_t + \ell_t(y_t^*) - \ell_t(\hat{y}_t)) + 2 \sqrt{\ln(d) \sum_{t=1}^{T} v_t + 4/3 \ln(d) + 2}
\]

\[
= \sum_{t=1}^{T} \ell_t(y_t^*) + \inf_{\eta > 0} \left\{ \frac{\ln(d)}{\eta} + \sum_{t=1}^{T} ((1 - b_t)\ell_t(y_t^*) + c_tb_t + \eta v_t - \ell_t(\hat{y}_t)) \right\} + 4/3 \ln(d) + 2.
\]

To upper bound the abstention gap by 0 is more difficult than to upper bound the surrogate gap as the negative term is no longer an upper bound on the zero-one loss. Hence, the abstention cost has to be strictly better than randomly guessing as otherwise there is no \(\eta\) or \(b_t\) such that the abstention gap is smaller than 0. The result for abstention can be found in Theorem 7 below.

Theorem 7. Suppose \(\max_t c_t < \frac{1}{2}\) for all \(T\). Then Algorithm 2 guarantees

\[
\sum_{t=1}^{T} ((1 - b_t)\ell_t(y_t^*) + b_tc_t) \leq \sum_{t=1}^{T} \ell_t(y_t^*) + \min \left\{ \frac{\ln(d)}{1 - 2\max_t c_t}, 2 \sqrt{\ln(d) \sum_{t=1}^{T} v_t} \right\} + 4/3 \ln(d) + 2.
\]

Proof. We start by upper bounding the \(v_t\) term. We have

\[
v_t = \frac{1}{4} \mathbb{E}_{p_t} [(y_t^* - \hat{y}_t)^2] \leq \frac{1}{4}(1 - \hat{y}_t)(\hat{y}_t + 1) \leq \frac{1}{2}(1 - |\hat{y}_t|),
\]

where the first inequality is the Bhatia-Davis inequality [Bhatia and Davis, 2000]. As with the proofs of GAPTRON we split the abstention gap in cases:

\[
(1 - b_t)\ell_t(y_t^*) + c_tb_t + \eta v_t - \ell_t(\hat{y}_t)
\]

\[
\leq (1 - b_t)\ell_t(y_t^*) + c_tb_t + \eta \frac{1}{2}(1 - |\hat{y}_t|) - \ell_t(\hat{y}_t)
\]

\[
= \begin{cases} 
\ell_t(\hat{y}_t) + c_t(1 - |\hat{y}_t|) + \eta \frac{1}{2}(1 - |\hat{y}_t|) - \frac{1}{2}(1 + |\hat{y}_t|) & \text{if } y_t^* = y_t \\
|\hat{y}_t| + c_t(1 - |\hat{y}_t|) + \eta \frac{1}{2}(1 - |\hat{y}_t|) - \frac{1}{2}(1 - |\hat{y}_t|) & \text{if } y_t^* \neq y_t.
\end{cases}
\]

Note that regardless of the true label \((1 - b_t)\ell_t(y_t^*) + c_tb_t - \ell_t(\hat{y}_t) \leq 0\) since \(c_t < \frac{1}{2}\). Hence, by using Lemma 3 we can see that as long as \(c_t < \frac{1}{2}\)

\[
\sum_{t=1}^{T} (1 - b_t)\ell_t(y_t^*) + b_tc_t \leq \sum_{t=1}^{T} \ell_t(y_t^*) + 2 \sqrt{\ln(d) \sum_{t=1}^{T} v_t + 4/3 \ln(d) + 2}.
\]

Now consider the case where \(y_t^* = y_t\). In this case, as long as \(\eta \leq 1 - 2c_t\) the abstention gap is bounded by 0. If \(y_t^* \neq y_t\) then

\[
|\hat{y}_t| + c_t(1 - |\hat{y}_t|) + \eta \frac{1}{2}(1 - |\hat{y}_t|) - \frac{1}{2}(1 + |\hat{y}_t|) = c_t(1 - |\hat{y}_t|) + \eta \frac{1}{2}(1 - |\hat{y}_t|) - \frac{1}{2}(1 - |\hat{y}_t|).
\]
So as long as $\eta \leq 1 - 2c_t$ the abstention gap is bounded by 0. Applying Lemma 3 now gives us

$$\sum_{t=1}^{T} (1 - b_t)\ell_t(y^*_t) + b_t c_t - \ell_t(y^*_t) \leq \inf_{\eta > 0} \left\{ \frac{\ln(d)}{\eta} + \sum_{t=1}^{T} ((1 - b_t)\ell_t(y^*_t) + c_t b_t + \eta v_t - \ell_t(\hat{y}_t)) \right\} + \frac{4}{3} \ln(d) + 2$$

$$\leq \frac{\ln(d)}{1 - 2 \max_t c_t} + \frac{4}{3} \ln(d) + 2,$$

which completes the proof. \[\square\]

With a slight modification of the proof of Theorem 7 one can also show a similar result as Theorem 8 by [Neu and Zhivotovskiy (2020)](https://example.com), albeit with slightly worse constants. We leave this as an exercise for the reader.