Exploiting the Surrogate Gap in Online Multiclass Classification

Dirk van der Hoeven

Leiden University

This talk

1. Brief introduction Online Multiclass Classification
2. My contributions
3. Some intuition about my contributions
4. Future work

Example Application Full Information

Football prediction. ADO Den Haag versus AFC Ajax.
We know that

- ADO plays at home
- There are 0 supporters for either side
- Players of ADO Den Haag are valued **11.35 million** euros
- Players of AFC Ajax are valued **288.85 million** euros

Who will win?

Example Application Full Information

Football prediction. ADO Den Haag versus AFC Ajax.
We know that

- ADO plays at home
- There are 0 supporters for either side
- Players of ADO Den Haag are valued **11.35 million** euros
- Players of AFC Ajax are valued **288.85 million** euros

Who will win? Probably AFC Ajax, but not absolutely certain.

If we gather this information for all eredivisie games, can we **predict perfectly?**

Example Application Full Information

Football prediction. ADO Den Haag versus AFC Ajax.
We know that

- ADO plays at home
- There are 0 supporters for either side
- Players of ADO Den Haag are valued **11.35 million** euros
- Players of AFC Ajax are valued **288.85 million** euros

Who will win? Probably AFC Ajax, but not absolutely certain.

If we gather this information for all eredivisie games, can we **predict perfectly?** Probably not

Important: regardless of what we predict, we will **see the true outcome**

Setting: Full Information

The online multiclass classification setting proceeds in rounds $t = 1, \ldots, T$. In each round t

1. the environment picks an outcome $y_t \in \{1, \ldots K\}$ and reveals a feature vector $\boldsymbol{x}_t \in \mathbb{R}^d$ to the learner
2. the learner issues a (randomized) prediction $\hat{y}_t$
3. **the environment reveals** y_t
4. the learner suffers $\mathbb{1}[y_t \neq \hat{y}_t]$

Setting: Full Information

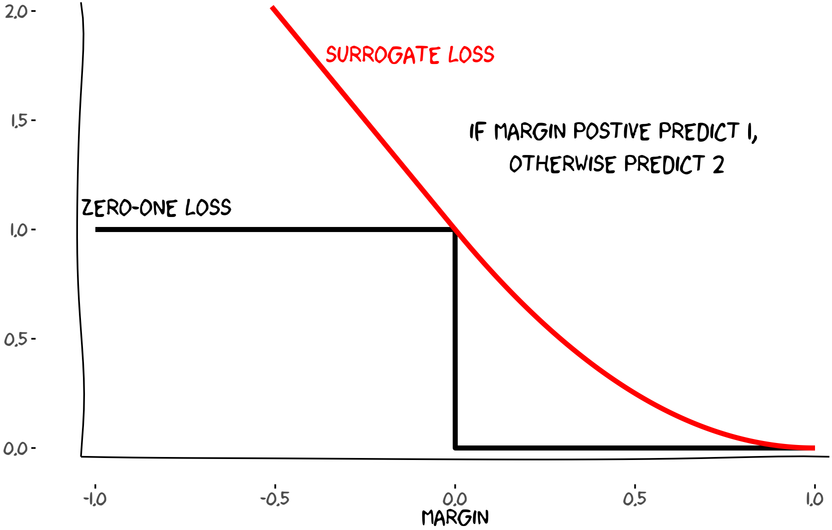
The online multiclass classification setting proceeds in rounds $t = 1, \ldots, T$. In each round t

1. the environment picks an outcome $y_t \in \{1, \ldots K\}$ and reveals a feature vector $\boldsymbol{x}_t \in \mathbb{R}^d$ to the learner
2. the learner issues a (randomized) prediction $\hat{y}_t$
3. **the environment reveals y_t**
4. the learner suffers $\mathbb{1}[y_t \neq \hat{y}_t]$

Goal: minimize the expected surrogate regret $\mathcal{R}_T$

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y_t \neq \hat{y}_t]\right] - \left(\min_{\boldsymbol{U}} \sum_{t=1}^{T} \ell(\underbrace{\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle}_{\text{margin}}, y_t)\right)$$

Surrogate Loss with $K = 2$



Is regret a reasonable measure of performance?

Goal: minimize the expected surrogate regret $\mathcal{R}_T$

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y_t \neq \hat{y}_t]\right] - \left(\underbrace{\min_{\boldsymbol{U}} \sum_{t=1}^{T} \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t)}_{= \textbf{0 If model predicts perfectly}} \right)$$

Translation: I want to be close to the performance of the best offline version of the model.

Example application bandit setting

You have to suggest a movie to friend from a list of movies you like (and suppose there is only 1 movie in this list your friend will like). You know:

- the length of your friend
- the weight of your friend

Can you give the perfect suggestion?

Example application bandit setting

You have to suggest a movie to friend from a list of movies you like (and suppose there is only 1 movie in this list your friend will like). You know:

- the length of your friend
- the weight of your friend

Can you give the perfect suggestion? probably not.

What is your strategy for the suggestion?

Example application bandit setting

You have to suggest a movie to friend from a list of movies you like (and suppose there is only 1 movie in this list your friend will like). You know:

- the length of your friend
- the weight of your friend

Can you give the perfect suggestion? probably not.

What is your strategy for the suggestion? mine: **randomly select** one movie, I don't have any information

Important: You don't get feedback about what you **should have suggested**

Setting: Bandit

The bandit online multiclass classification setting proceeds in rounds $t = 1, \dots, T$. In each round t

1. the environment picks an outcome $y_t \in \{1, \dots K\}$ and reveals a feature vector $\boldsymbol{x}_t$ to the learner
2. the learner issues a (randomized) prediction $\hat{y}_t$
3. **the environment reveals** $\mathbb{1}[y_t \neq \hat{y}_t]$
4. the learner suffers $\mathbb{1}[y_t \neq \hat{y}_t]$

Goal: minimize the expected surrogate regret $\mathcal{R}_T$

$$\mathcal{R}_T = \mathbb{E}\left[\left(\sum_{t=1}^{T} \mathbb{1}[y_t \neq \hat{y}_t]\right) - \left(\min_{\boldsymbol{U}} \sum_{t=1}^{T} \ell(\underbrace{\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle}_{\text{margin}}, y_t)\right)\right]$$

Results 1

Results for the **full information setting**

Algorithm	$\mathcal{R}_T$	Time (per round)
Standard first-order	$O(\|\boldsymbol{U}\|\sqrt{T})$	$O(dK)$
Standard second-order	$O(e^{\|\boldsymbol{U}\|}dK\ln(T))$	$O((dK)^2)$
GAPTRON	$O(K\|\boldsymbol{U}\|^2)$	$O(dK)$

Results 2

Results for the **bandit setting**

Algorithm	$\mathbb{E}[\mathcal{R}_T]$	Time (per round)
Standard first-order	$O((K)^{1/3}T^{2/3})$	$O(dK)$
Standard second-order	$O(K\sqrt{dT\ln(T)})$	$O((dK)^2)$
GAPTRON	$O(K\sqrt{T})$	$O(dK)$

Old Analysis: Upper Bound Zero-One Loss with Surrogate

$$\sum_{t=1}^{T} \mathbb{1}[y_t \neq \hat{y}_t] - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) =$$

Old Analysis: Upper Bound Zero-One Loss with Surrogate

$$\sum_{t=1}^{T} \mathbb{1}[y_t \neq \hat{y}_t] - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) = \left(\sum_{t=1}^{T} \underbrace{\mathbb{1}[y_t \neq \hat{y}_t] - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)}_{\textbf{Very wastefull: bound by 0}} \right) + \left(\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) \right)$$

Old Analysis: Upper Bound Zero-One Loss with Surrogate

$$
\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}[y_t \neq \hat{y}_t] - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) = & \left(\sum_{t=1}^{T} \underbrace{\mathbb{1}[y_t \neq \hat{y}_t] - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)}_{\text{Very wastefull: bound by 0}} \right) \\
& + \left(\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) \right) \\
\leq & \underbrace{\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t)}_{\text{controlled by OGD}}
\end{aligned}
$$

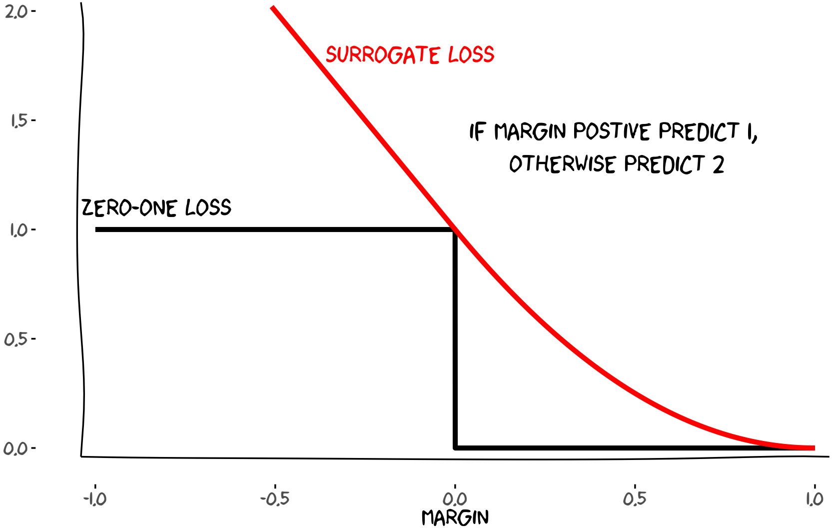
Old Analysis: Upper Bound Zero-One Loss with Surrogate

$$
\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}[y_t \neq \hat{y}_t] - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t\rangle, y_t) &= \left(\sum_{t=1}^{T} \underbrace{\mathbb{1}[y_t \neq \hat{y}_t] - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t\rangle, y_t)}_{\textbf{Very wastefull: bound by 0}}\right) \\
&\quad + \left(\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t\rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t\rangle, y_t)\right) \\
&\leq \underbrace{\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t\rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t\rangle, y_t)}_{\textbf{controlled by OGD}} \\
&\leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\nabla_{\boldsymbol{W}_t} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t\rangle, y_t)\|^2
\end{aligned}
$$

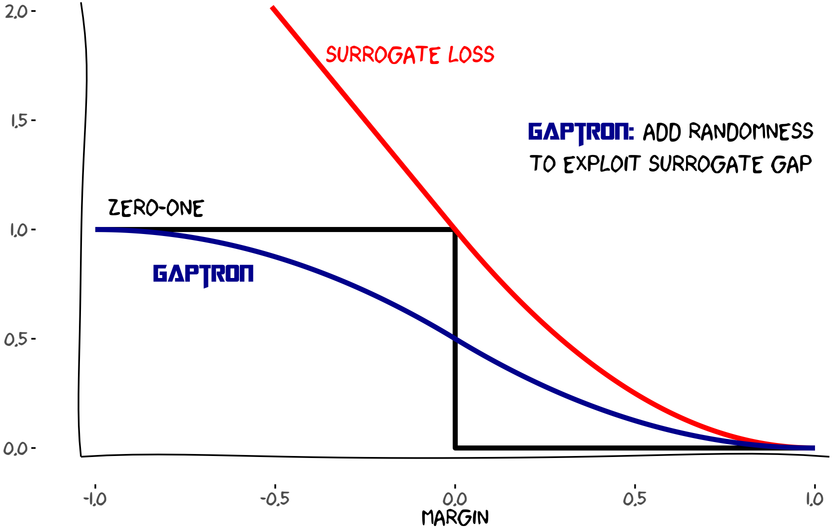
Old Analysis: Upper Bound Zero-One Loss with Surrogate

$$
\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}[y_t \neq \hat{y}_t] - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) = & \left(\sum_{t=1}^{T} \underbrace{\mathbb{1}[y_t \neq \hat{y}_t] - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)}_{\text{Very wastefull: bound by 0}} \right) \\
& + \left(\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) \right) \\
\leq & \underbrace{\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t)}_{\text{controlled by OGD}} \\
\leq & \frac{\|\boldsymbol{U}\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla_{\boldsymbol{W}_t} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)\|^2 \\
\leq & \|\boldsymbol{U}\| X \sqrt{T}
\end{aligned}
$$

How to improve upon standard methods?



Key Idea: when uncertain, randomize.



Gaptron Analysis: first steps

$$\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right]}_{\textbf{Gaptron is random}} - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) =$$

Gaptron Analysis: first steps

$$\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right]}_{\textbf{Gaptron is random}} - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) = \left(\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) \right)$$

$$+ \underbrace{\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t)}_{\textbf{controlled by OGD}}$$

Gaptron Analysis: first steps

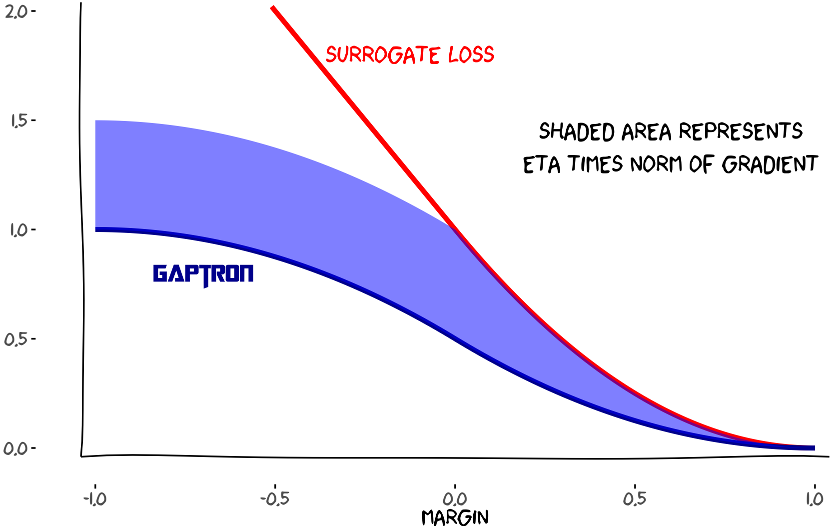
$$\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right]}_{\text{Gaptron is random}} - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) = \left(\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) \right)$$

$$+ \underbrace{\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t)}_{\text{controlled by OGD}}$$

$$\leq \frac{\|\boldsymbol{U}\|^2}{2\eta}$$

$$+ \left(\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] + \frac{\eta}{2} \|\nabla_{\boldsymbol{W}_t} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)\|^2}_{\text{Smaller than surrogate loss!}} - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) \right)$$

Gaptron Analysis: is that really smaller than surrogate loss?



Gaptron Analysis: wow, that is very useful!

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) \leq \frac{\|\boldsymbol{U}\|^2}{2\eta}$$

$$+ \left(\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] + \frac{\eta}{2}\|\nabla_{\boldsymbol{W}_t} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)\|^2}_{\textbf{Smaller than surrogate loss!}} - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) \right)$$

$$\leq KX^2 \|\boldsymbol{U}\|^2$$

Gaptron Analysis: wow, that is very useful!

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) \leq \frac{\|\boldsymbol{U}\|^2}{2\eta}$$

$$+ \left(\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] + \frac{\eta}{2}\|\nabla_{\boldsymbol{W}_t}\ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)\|^2}_{\textbf{Smaller than surrogate loss!}} - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) \right)$$

$$\leq KX^2\|\boldsymbol{U}\|^2$$

Full information before: $\|\boldsymbol{U}\|X\sqrt{T}$ regret

Full information now: $KX^2\|\boldsymbol{U}\|^2$ regret

Gaptron Analysis: wow, that is very useful!

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) \leq \frac{\|\boldsymbol{U}\|^2}{2\eta}$$
$$+ \left(\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] + \frac{\eta}{2}\|\nabla_{\boldsymbol{W}_t}\ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)\|^2}_{\text{Smaller than surrogate loss!}} - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) \right)$$
$$\leq KX^2\|\boldsymbol{U}\|^2$$

Bandit before: $O(K\sqrt{dT\ln(T)})$ regret with $O((dK)^2)$ runtime

Bandit now: $O(K\sqrt{T})$ regret with $O(dK)$ runtime (very cool in high-dimensional applications)

A closer look at GAPTRON

Input: Learning rate $\eta > 0$, exploration rate $\gamma \in [0, 1]$, and gap map $a : \mathbb{R}^{K\times d} \times \mathbb{R}^d \to [0, 1]$

Initialize $\boldsymbol{W}_1 = \boldsymbol{0}$
for $t = 1 \dots T$ **do**
  Obtain $\boldsymbol{x}_t$
  Let $y_t^\star = \arg\max_k \langle \boldsymbol{W}_t^k, \boldsymbol{x}_t \rangle$
  Set $\boldsymbol{p}_t' = (1 - \max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\})\boldsymbol{e}_{y_t^\star} + \max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\}\frac{1}{K}\boldsymbol{1}$
  Predict with label $\hat{y}_t \sim \boldsymbol{p}_t'$
  Obtain $\mathbb{1}[\hat{y}_t \neq y_t]$ and set $\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{W}_t)$
  Update $\boldsymbol{W}_{t+1} = \arg\min_{\boldsymbol{W} \in \mathcal{W}} \eta\langle \boldsymbol{g}_t, \boldsymbol{W} \rangle + \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{W}_t\|^2$
end for

A closer look at GAPTRONs predictions

$$\boldsymbol{p}_t' = \underbrace{(1 - \max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\})\boldsymbol{e}_{y_t^\star}}_{\textbf{I think the outcome is } y_t^\star} + \underbrace{\max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\}\frac{1}{K}\mathbf{1}}_{\textbf{But I am not certain}}$$

A closer look at GAPTRONs predictions

$$\boldsymbol{p}'_t = \underbrace{(1 - \max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\})\boldsymbol{e}_{y_t^\star}}_{\text{I think the outcome is } y_t^\star} + \underbrace{\max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\}\frac{1}{K}\mathbf{1}}_{\text{But I am not certain}}$$

Choosing the right a ensures that the expected loss of GAPTRON plus the norm of the gradient is smaller than the surrogate loss.
If $a(\boldsymbol{W}, \boldsymbol{x}) = 0$ we recover standard algorithms such as the PERCEPTRON.

A closer look at GAPTRONs predictions

$$\boldsymbol{p}_t' = \underbrace{(1 - \max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\})\boldsymbol{e}_{y_t^\star}}_{\textbf{I think the outcome is } y_t^\star} + \underbrace{\max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\}\frac{1}{K}\mathbf{1}}_{\textbf{But I am not certain}}$$

Choosing the right a ensures that the expected loss of GAPTRON plus the norm of the gradient is smaller than the surrogate loss.
If $a(\boldsymbol{W}, \boldsymbol{x}) = 0$ we recover standard algorithms such as the PERCEPTRON.
If $\gamma > 0$ we sample any outcome with probability at least $\gamma\frac{1}{K}$, which is important for the **bandit setting**

Main Lemma of the paper

Lemma

For any $\boldsymbol{U} \in \mathcal{W}$ GAPTRON *satisfies*

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \neq y_t] - \sum_{t=1}^{T} \ell_t(\boldsymbol{U})\right] \\
&\leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \gamma \frac{K-1}{K} T \\
&+ \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t \frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right]}_{\textbf{surrogate gap}}.
\end{aligned}
$$

Main Lemma of the paper

Lemma

For any $\boldsymbol{U} \in \mathcal{W}$ Gaptron *satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \neq y_t] - \sum_{t=1}^{T} \ell_t(\boldsymbol{U})\right]$$
$$\leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \gamma\frac{K-1}{K}T$$
$$+ \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right]}_{\textbf{surrogate gap}}.$$

Rest of the paper: finding the correct a, η, and γ to bound the surrogate gap by 0

Choosing a_t and η for smooth losses in 2 dimensions

In 2 dimensions $y_t \in \{-1, +1\}$ and $\ell_t(\boldsymbol{W}) = \ell(\langle \boldsymbol{W}, \boldsymbol{x}_t \rangle y_t)$.

Choosing a_t and η for smooth losses in 2 dimensions

In 2 dimensions $y_t \in \{-1, +1\}$ and $\ell_t(\boldsymbol{W}) = \ell(\langle \boldsymbol{W}, \boldsymbol{x}_t \rangle y_t)$. A function f is H-smooth if

$$f(\boldsymbol{x} + \boldsymbol{z}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{z} \rangle + \frac{H}{2} \|\boldsymbol{z}\|_2^2.$$

Let $\boldsymbol{U}_t^\star = \arg\min_{\boldsymbol{W}} \ell_t(\boldsymbol{W})$. For smooth surrogate losses we have

$$\|\nabla \ell_t(\boldsymbol{W}_t)\|_2^2 \leq H(\ell_t(\boldsymbol{W}_t) - \ell_t(\boldsymbol{U}_t^\star)) = H\ell_t(\boldsymbol{W}_t)$$

Choosing a_t and η for smooth losses

For smooth surrogate losses we have:

$$
\begin{aligned}
&(1 - a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t \frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2 \\
&\leq (1 - a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t \frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta H}{2}\ell_t(\boldsymbol{W}_t).
\end{aligned}
$$

Picking $a_t = \ell_t^\star(\boldsymbol{W}_t) = \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle y_t^\star)$ and $\eta = \frac{2}{HK}$ we have

$$
\begin{aligned}
&(1 - a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t \frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2 \\
&\leq (1 - \ell_t^\star(\boldsymbol{W}_t))\mathbb{1}[y_t^\star \neq y_t] + \ell_t^\star(\boldsymbol{W}_t)\frac{K-1}{K} - \frac{K-1}{K}\ell_t(\boldsymbol{W}_t).
\end{aligned}
$$

Choosing a_t and η for smooth losses

If $y_t^\star = y_t$:

$$\begin{aligned}
&(1 - a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2 \\
&\leq (1 - \ell_t^\star(\boldsymbol{W}_t))\mathbb{1}[y_t^\star \neq y_t] + \ell_t^\star(\boldsymbol{W}_t)\frac{K-1}{K} - \frac{K-1}{K}\ell_t(\boldsymbol{W}_t) \\
&= \ell_t(\boldsymbol{W}_t)\frac{K-1}{K} - \frac{K-1}{K}\ell_t(\boldsymbol{W}_t) \\
&= 0
\end{aligned}$$

Choosing a_t and η for smooth losses

If $y_t^\star \neq y_t$:

$$
\begin{aligned}
&(1 - a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t \frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2 \\
&\leq (1 - \ell_t^\star(\boldsymbol{W}_t))\mathbb{1}[y_t^\star \neq y_t] + \ell_t^\star(\boldsymbol{W}_t)\frac{K-1}{K} - \frac{K-1}{K}\ell_t(\boldsymbol{W}_t) \\
&= 1 - \frac{1}{K}\ell_t^\star(\boldsymbol{W}_t) - \frac{K-1}{K}\ell_t(\boldsymbol{W}_t) \\
&= 1 - \frac{1}{K}\ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t\rangle y_t^\star) - \frac{K-1}{K}\ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t\rangle y_t) \\
&\leq 1 - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t\rangle(\underbrace{\frac{1}{K}y_t^\star + \frac{K-1}{K}y_t}_{\text{opposite sign of } \langle \boldsymbol{W}_t, \boldsymbol{x}_t\rangle})) \\
&\leq 0
\end{aligned}
$$

Inspiration from classification with abstention

Setting:

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 $y_t' \in [-1, 1] \cup *$, where $*$ 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.

Inspiration from classification with abstention

Algorithm:

Input: AdaHedge

1: **for** $t = 1 \dots T$ **do**
2: Obtain expert predictions $\boldsymbol{y}_t = (y_t^1, \dots, y_t^d)^\top \in [-1, 1]^d$
3: Obtain expert distribution $\hat{\boldsymbol{p}}_t$ from AdaHedge
4: Set $\hat{y}_t = \langle \hat{\boldsymbol{p}}_t, \boldsymbol{y}_t \rangle$
5: Let $y_t^\star = \text{sign}(\hat{y}_t)$
6: Set $b_t = 1 - |\hat{y}_t|$
7: Predict $\boldsymbol{y}_t' = y_t^\star$ with probability $1 - b_t$ and predict $y_t' = *$ with probability b_t
8: Obtain ℓ_t and send ℓ_t to AdaHedge
9: **end for**

Inspiration from classification with abstention

Lemma

For any expert i, the expected loss of satisfies:

$$\sum_{t=1}^{T}(1-b_t)\ell_t(y_t^\star)+b_t c_t$$

$$\leq \sum_{t=1}^{T}\ell_t(y_t^i)+\inf_{\eta>0}\left\{\frac{\ln(d)}{\eta}+\sum_{t=1}^{T}\underbrace{(1-b_t)\ell_t(y_t^\star)+c_t b_t+\eta v_t-\ell_t(\hat{y}_t)}_{\text{Abstention gap}}\right\}$$

$$+\frac{4}{3}\ln(d)+2,$$

where $v_t=\mathbb{E}_{i\sim\hat{\boldsymbol{p}}_t}[(\ell_t(\hat{y}_t)-\ell_t(y_t^i))^2]$.

Inspiration from classification with abstention

Abstention gap

$$(1 - b_t)\ell_t(y_t^\star) + b_t c_t + \eta v_t - \ell_t(\hat{y}_t)$$

Surrogate gap:

$$(1 - a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t c_t' + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2 - \ell_t(\boldsymbol{W}_t)$$

c_t' is the cost for guessing rather than abstaining, although if abstaining costs strictly less than 1 we could replace c_t' with abstention cost c_t and obtain surrogate regret that satisfies

$$\mathcal{R}_T(\boldsymbol{U}) = O\left(\frac{\|X\|_2^2\|U\|_2^2}{1 - \max_t c_t}\right)$$

Future work

- High probability bounds
- Can we exploit curvature to improve regret bound?
- Empirical performance?
- Can we improve regret when the model can predict perfectly (both bandit and full information)?
  - For full information, I think yes.
  - For bandit setting, I am not sure.
- Can we apply this idea in other problems such as ranking?

Where to find the paper?

- my website: `dirkvanderhoeven.com/research`
- arxiv: `https://arxiv.org/abs/2007.12618`
- NeurIPS 2020 proceedings
- By googling "gaptron" (the twitter and instagram accounts are not mine).