GAPTRON: Exploiting the Surrogate Gap in Online Multiclass Classification

K expected regret

Full information setting

$K\sqrt{T}$ expected regret

Bandit setting

Dirk van der Hoeven

Universiteit Leiden
The Netherlands

Setting: Online Multiclass Classification

The 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. - **Full Information Setting:** the environment reveals **true outcome** y_t
   - **Bandit setting:** the environment reveals **loss** $\mathbb{1}[y_t \neq \hat{y}_t]$
4. the learner suffers $\mathbb{E}[\mathbb{1}[y_t \neq \hat{y}_t]]$

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

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

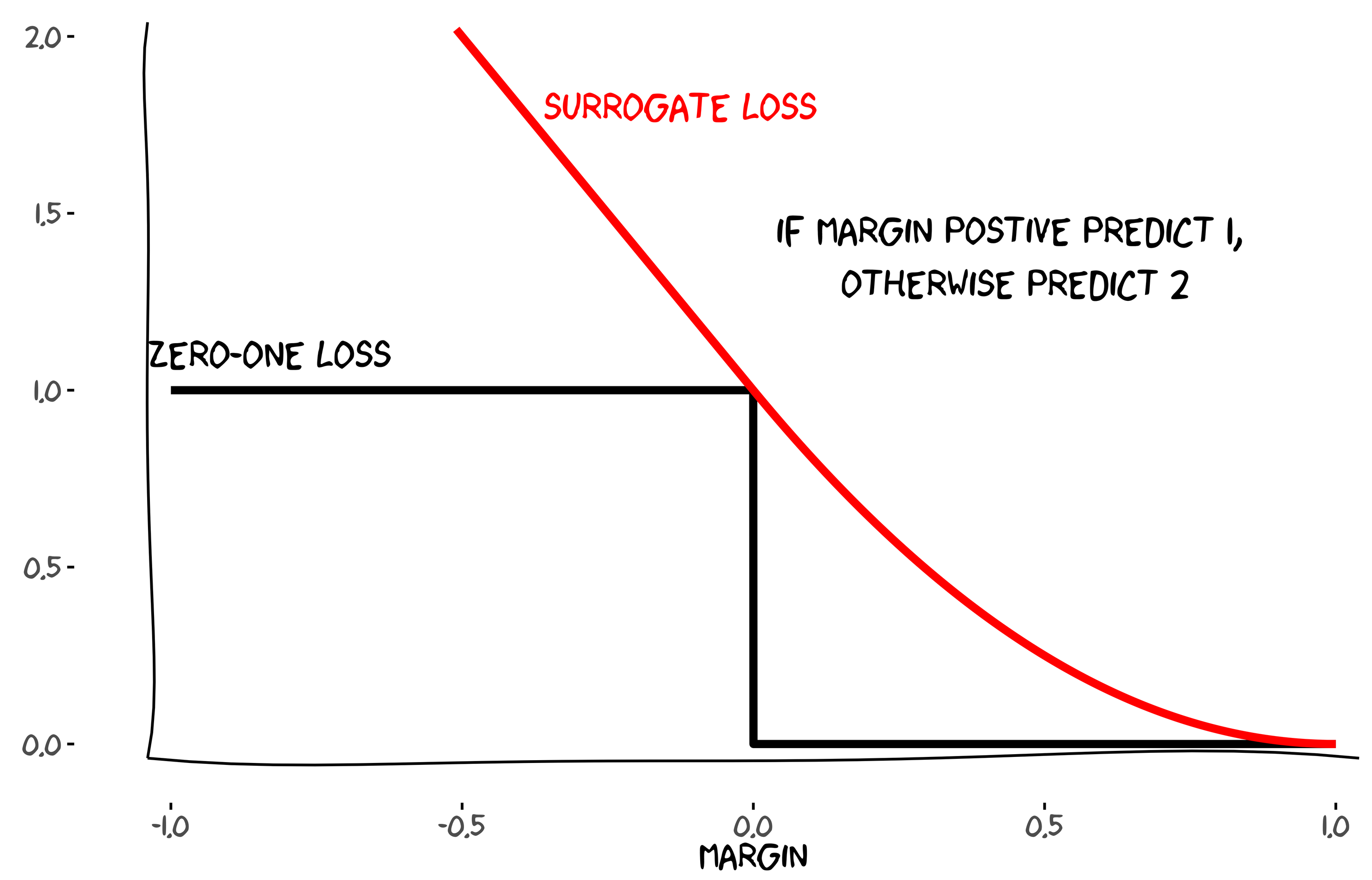
Results

Algorithm	**Full information** $\mathcal{R}_T$	**Bandit** $\mathcal{R}_T$	Time (per round)
Standard first-order	$O(\lVert \boldsymbol{U} \rVert \sqrt{T})$	$O((K)^{1/3}T^{2/3})$	$O(dK)$
Standard second-order	$O(\mathbf{e}^{\lVert \boldsymbol{U} \rVert} dK \ln(T))$	$O(K\sqrt{dT\ln(T)})$	$O((dK)^2)$
Gaptron	$O(K \lVert \boldsymbol{U} \rVert^2)$	$O(K\sqrt{T})$	$O(dK)$

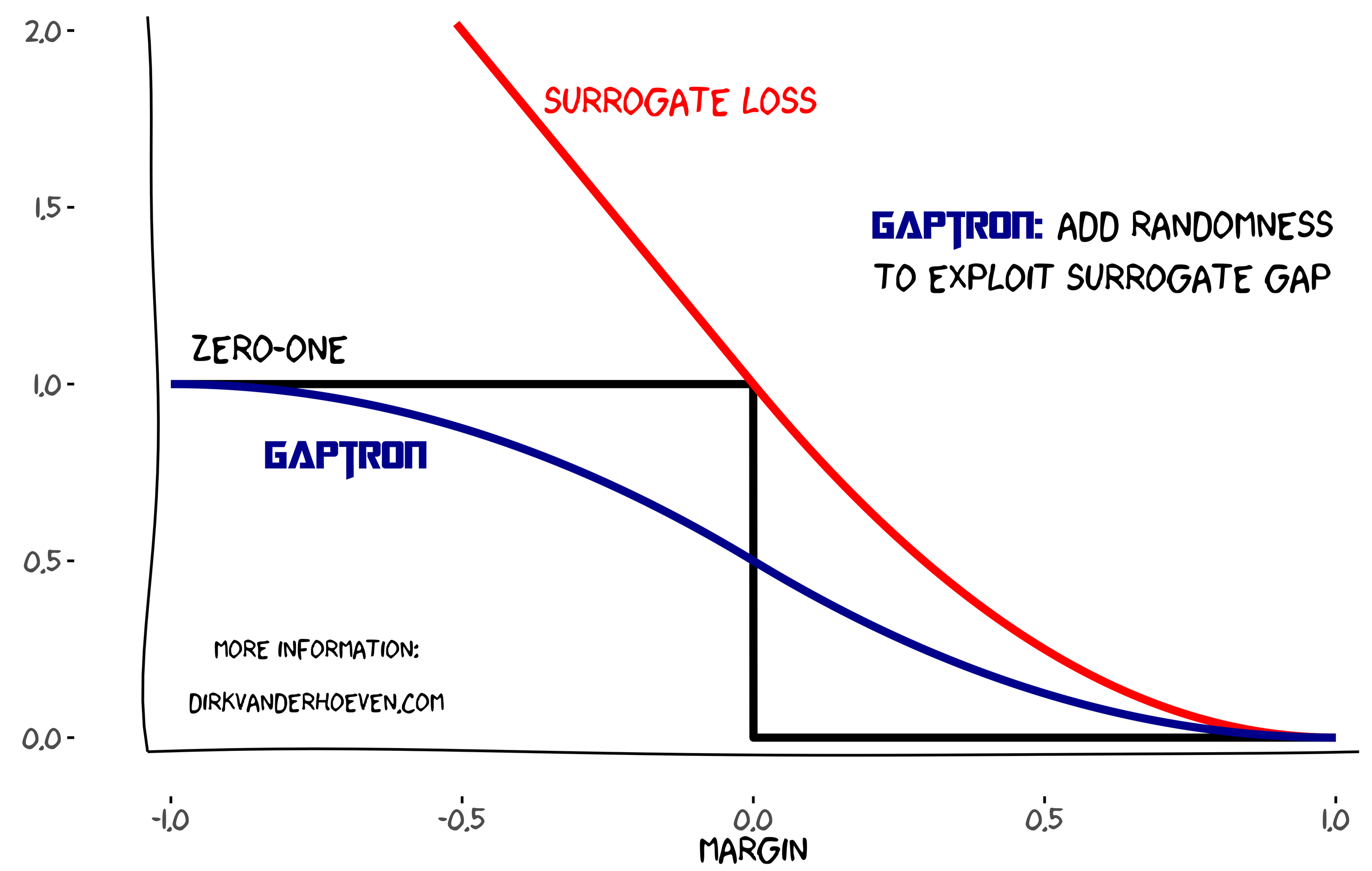
Standard Analysis

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

Standard Analysis: Figure

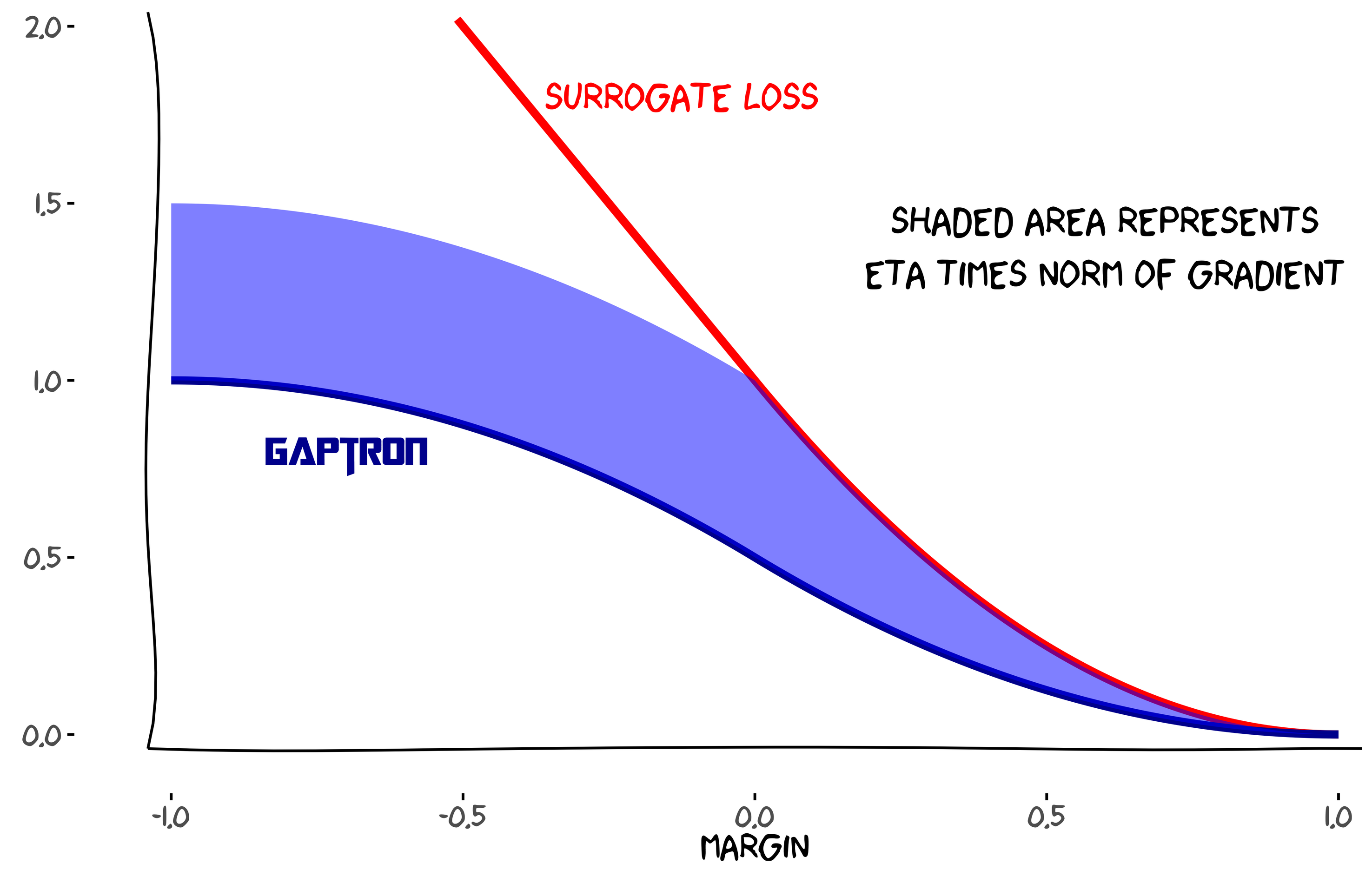


Gaptron Key Idea



Gaptron Analysis

$$\begin{aligned}
&\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right]}_{\text{Gaptron is random}} - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t) \\
&= \left(\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)\right) + \underbrace{\sum_{t=1}^{T} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) - \ell(\langle \boldsymbol{U}, \boldsymbol{x}_t \rangle, y_t)}_{\text{controlled by OGD}} \\
&\leq \frac{\lVert \boldsymbol{U} \rVert^2}{2\eta} + \left(\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\mathbb{1}[y_t \neq \hat{y}_t]\right] + \frac{\eta}{2}\lVert \nabla_{\boldsymbol{W}_t} \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t) \rVert^2} - \ell(\langle \boldsymbol{W}_t, \boldsymbol{x}_t \rangle, y_t)\right)
\end{aligned}$$



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