The many faces of online learning

Hoeven, D. van der

CHAPTER 6

Exploiting the Surrogate Gap in Online Multiclass Classification

This chapter is based on Van der Hoeven, D. (2020). Exploiting the surrogate gap in online multiclass classification. *To Appear in Advances in Neural Information Processing Systems 33*.

Abstract

We present GAPTRON, a randomized first-order algorithm for online multiclass classification. In the full information setting we show expected mistake bounds with respect to the logistic loss, hinge loss, and the smooth hinge loss with $O(K)$ expected surrogate regret, where the expectation is with respect to the learner's randomness and K is the number of classes. In the bandit classification setting we show that GAPTRON is the first linear time algorithm with $O(K\sqrt{T})$ expected surrogate regret. Additionally, the expected mistake bound of GAPTRON does not depend on the dimension of the feature vector, contrary to previous algorithms with $O(K\sqrt{T})$ surrogate regret in the bandit classification setting. We present a new proof technique that exploits the gap between the zero-one loss and surrogate losses rather than exploiting properties such as exp-concavity or mixability, which are traditionally used to prove logarithmic or constant regret bounds.

6.1 Introduction

In online multiclass classification a learner has to repeatedly predict the label that corresponds to a feature vector. Algorithms in this setting have a wide range of applications ranging from predicting the outcomes of sport matches to recommender systems. In some applications such as sport forecasting the learner obtains the true label regardless of what outcome the learner predicts, but in other applications such as recommender systems the learner only learns whether or not the label he predicted was the true label. The setting in which the learner receives the true label is called the full information multiclass classification setting and the setting in which the learner only receives information about the predicted label is called the bandit multiclass classification setting.

In this chapter we consider both the full information and bandit multiclass classification settings. In both settings the environment chooses the true outcome $y_t \in \{1, \ldots, K\}$ and feature vector $\boldsymbol{x}_t \in \mathbb{R}^d$. The environment then reveals the feature vector to the learner, after which the learner issues a (randomized) prediction $y'_t \in \{1, \ldots, K\}$. The goal of both settings is to minimize the number of expected mistakes the learner makes with respect to the best offline linear predictor $\boldsymbol{U} \in \mathbb{R}^{K \times d}$, where each row of $\boldsymbol{U}$ essentially keeps track of a linear predictor for each class. Standard practice in both settings is to upper bound the non-convex zero-one loss with a convex surrogate loss ℓ_t (see for example Bartlett et al. (2006)). This leads to guarantees of the form

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y'_t \neq y_t]\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\boldsymbol{U})\right] + \tilde{\mathcal{R}}_T,$$

where $\mathbb{1}$ is the indicator function, y_t is the true label, the expectation is taken with respect to the learner's randomness, and $\tilde{\mathcal{R}}_T$ is the surrogate regret after T rounds.

We introduce GAPTRON, which is a randomized first-order algorithm that exploits the gap between the zero-one loss and the surrogate loss. In the full information multiclass classification setting GAPTRON has $O(K)$ surrogate regret with respect various surrogate losses. In the bandit multiclass classification setting we show that GAPTRON has $O(K\sqrt{T})$ surrogate regret with respect to the same surrogate losses as in the full information setting. Importantly, our surrogate regret bounds do not depend on the dimension of the feature vector in either the full or bandit information setting, contrary to previous results with comparable surrogate regret bounds. Furthermore, in the bandit multiclass classification setting GAPTRON is the first $O(dK)$ running time algorithm with $O(K\sqrt{T})$ surrogate regret.

To achieve these results we develop a new proof technique. Standard approaches that lead to small surrogate regret bounds exploit properties of the surrogate loss function such as strong convexity, exp-concavity (Hazan et al., 2007), or mixability (Vovk, 2001). Instead, inspired by the recent success of Neu and Zhivotovskiy (2020) in online classification with abstention[1], we exploit the *gap* between the zero-one loss, which is used to measure the performance of the learner, and the surrogate loss, which is used to measure the performance of the comparator $\boldsymbol{U}$, hence the name GAPTRON.

For an overview of our results and a comparison to previous work see Table 6.1. Here we briefly discuss the most relevant literature to place our results into perspective. A more detailed comparison can be found in the relevant sections. The full information multiclass classification setting is well understood and has been studied by many authors. Perhaps the most well known algorithm in this setting is the PERCEPTRON (Rosenblatt, 1958) and its multiclass versions (Crammer and Singer, 2003; Fink et al., 2006). The PERCEPTRON is a deterministic first-order algorithm which has $O(\sqrt{T})$ surrogate regret with respect to the hinge loss in the worst-case. Variants of the PERCEPTRON such as AROW (Crammer et al., 2009) and the second-order PERCEPTRON (Cesa-Bianchi et al., 2005) are second-order methods which result in a possibly smaller surrogate regret at the cost of longer running time. Online logistic regression (Berkson, 1944) is an alternative to the PERCEPTRON which has been thoroughly studied. For an overview of results for online logistic regression we refer the reader to Shamir (2020). We mention a recent result by Foster et al. (2018a), who use Exponential Weights (Vovk, 1990; Littlestone and Warmuth, 1994) to optimize the logistic loss and obtain a surrogate regret bound of order $O(dK\ln(DT+1))$, where D is an upper bound on the Frobenius norm of $\boldsymbol{U}$, with a polynomial time algorithm.

The bandit multiclass classification setting was first studied by Kakade et al. (2008) and is a special case of the contextual bandit setting (Langford and Zhang, 2008). Kakade et al. (2008) present a first-order algorithm called BANDITRON with a $O((DK)^{1/3}T^{2/3})$ surrogate regret bound with respect to the hinge loss. The impractical EXP4 algorithm (Auer et al., 2002) has a $O(\sqrt{TdK\ln(T+1)})$ surrogate regret bound and Abernethy and Rakhlin (2009) posed the problem of obtaining a practical algorithm which attains an $O(K\sqrt{T})$ surrogate regret bound. Several authors have proposed polynomial running time algorithms that have a surrogate regret bound of order $O(K\sqrt{dT\log(T+1)})$ such as NEWTRON (Hazan and Kale, 2011), SOBA (Beygelzimer et al., 2017), and OBAMA (Foster et al., 2018a).

[1] In fact, in Section 6.10 we slightly generalize the results of Neu and Zhivotovskiy (2020).

[2] These results hold for a family of loss functions parametrized by $\kappa \in [0,1]$, which includes the

Table 6.1: Main results and comparisons with previous work (see Section 6.2 for notation). The references are for the surrogate regret bounds, not necessarily for the first analysis of the algorithm. For this table we assume that $\|\boldsymbol{x}_t\| \leq 1 \; \forall t$ and denote by $L_T = \sum_{t=1}^T \ell_t(\boldsymbol{U})$ the sum of the surrogate losses of the comparator.

Algorithm	Loss	surrogate regret full information setting	surrogate regret bandit setting	Time (per round)
PERCEPTRON(Fink et al., 2006; Kakade et al., 2008)	hinge	$O(\|\boldsymbol{U}\|^2 + \|\boldsymbol{U}\|\sqrt{L_T})$	$O((DK)^{1/3}T^{2/3})$	$O(dK)$
Second-Order PERCEPTRON (Orabona et al., 2012; Beygelzimer et al., 2017)	hinge^2	$O(\frac{\kappa}{2-\kappa}\|\boldsymbol{U}\|^2 + \frac{dK}{\kappa(2-\kappa)}\ln(L_T))$	$O(\|\boldsymbol{U}\|^2 + \frac{K}{\kappa}\sqrt{dT\ln(T)})$	$O((dK)^2)$
ONS (Hazan et al., 2014; Hazan and Kale, 2011)	logistic	$O(\exp(D)dK\ln(T))$	$O(dK^3DT^{2/3})$	$O((dK)^2)$
Vovk's Aggregating Algorithm (Foster et al., 2018a)	logistic	$O(dK\ln(DT))$	$O(K\sqrt{dT\ln(DT)})$	$O(\max\{dK, T\}^{12})$
GAPTRON (This work)	logistic, hinge, smooth hinge	$O(K\|\boldsymbol{U}\|^2)$	$O(KD\sqrt{T})$	$O(dK)$

6.2 Preliminaries

Notation. Let $\mathbf{1}$ and $\mathbf{0}$ denote vectors with only ones and zeros respectively and let $\boldsymbol{e}_k$ denote the basis vector in direction k. The inner product between vectors $\boldsymbol{g} \in \mathbb{R}^d$ and $\boldsymbol{w} \in \mathbb{R}^d$ is denoted by $\langle \boldsymbol{w}, \boldsymbol{g} \rangle$. The rows of matrix $\boldsymbol{W} \in \mathbb{R}^{K \times d}$ are denoted by $\boldsymbol{W}^1, \ldots, \boldsymbol{W}^K$. We will interchangeably use $\boldsymbol{W}$ to denote a matrix and a column vector in $\mathbb{R}^{Kd}$ to avoid unnecessary notation. The vector form of matrix $\boldsymbol{W}$ is $(\boldsymbol{W}^1, \ldots, \boldsymbol{W}^K)^\intercal$. The Frobenius norm of matrix $\boldsymbol{W}$ is denoted by $\|\boldsymbol{W}\| = \sqrt{\sum_{k=1}^K \sum_{i=1}^d W_{k,i}^2}$. Likewise the l_2 norm of vector $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^d x_i^2}$. We denote the Kronecker product between matrices $\boldsymbol{W}$ and $\boldsymbol{U}$ by $\boldsymbol{W} \otimes \boldsymbol{U}$. For a given round t we use $\mathbb{E}_t[\cdot]$ to denote the conditional expectation given the predictions $y_1', y_2', \ldots, y_{t-1}'$.

6.2.1 Multiclass Classification

The multiclass classification setting proceeds in rounds $t = 1, \ldots, T$. In each round t the environment first picks an outcome $y_t \in \{1, \ldots K\}$ and feature vector $\boldsymbol{x}_t$ such that $\|\boldsymbol{x}_t\| \leq X$ for all t. Before the learner makes his prediction y_t' the environment reveals the feature vector $\boldsymbol{x}_t$ which the learner may use to form y_t'. In the full information multiclass classification setting, after the learner has issued y_t', the environment reveals the outcome y_t to the learner. In the bandit multiclass

hinge loss.

Algorithm 13 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 $t = 1 \ldots T$ 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 $y_t' \sim \boldsymbol{p}_t'$
    Obtain $\mathbb{1}[y_t' \neq y_t]$ 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
```

classification setting (Kakade et al., 2008) the environment only reveals whether the prediction of the learner was correct or not, i.e. $\mathbb{1}[y_t' \neq y_t]$. We only consider the adversarial setting, which means that we make no assumptions on how y_t or $\boldsymbol{x}_t$ is generated. In both settings we allow the learner to use randomized predictions. The goal of the multiclass classification setting is to control the number of expected mistakes the learner makes in T rounds: $M_T = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y_t' \neq y_t]\right]$, where the expectation is taken with respect to the learner's randomness.

Since the zero-one loss is non-convex a standard approach is to use a surrogate loss ℓ_t as a function of a weight matrix $\boldsymbol{W}_t \in \mathcal{W}$, where $\mathcal{W} = \{\boldsymbol{W} : \|\boldsymbol{W}\| \leq D\}$. The surrogate loss function is a convex upper bound on the zero-one loss, which is then optimized using an Online Convex Optimization algorithm such as Online Gradient Descent (OGD) (Zinkevich, 2003), Online Newton Step (ONS) (Hazan et al., 2007), or Exponential Weights (EW) (Vovk, 1990; Littlestone and Warmuth, 1994). In this chapter we treat three surrogate loss functions: logistic loss, the hinge loss, and the smooth hinge loss, all of which result in different guarantees on the number of expected mistakes a learner makes.

6.3 GAPTRON

In this section we discuss GAPTRON (Algorithm 13). The prediction y_t' is sampled from

$$\boldsymbol{p}_t' = (1 - \max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\})\boldsymbol{e}_{y_t^\star} + \max\{a(\boldsymbol{W}_t, \boldsymbol{x}_t), \gamma\}\frac{1}{K}\boldsymbol{1},$$

where $\gamma \in [0, 1]$, $a : \mathbb{R}^{K\times d} \times \mathbb{R}^d \to [0, 1]$, $y_t^\star = \arg\max_k \langle \boldsymbol{W}_t^k, \boldsymbol{x}_t \rangle$, and $\boldsymbol{e}_{y_t^\star}$ is the basis vector in direction $y_t^\star$. In the full information setting γ is set to 0 but in the bandit setting γ is used to guarantee that each label is sampled with at least probability $\frac{\gamma}{K}$, which is a common strategy in bandit algorithms (see for example Auer et al. (2002)). The fact that each label is sampled with at least probability $\frac{\gamma}{K}$ is important because in the bandit setting we use importance weighting to form estimated surrogate losses ℓ_t and their gradients $\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{W}_t)$ and we need to control the variance of these estimates. The main difference between GAPTRON and standard algorithms for multiclass classification is the a function, which governs the mixture that forms $\boldsymbol{p}_t'$. In fact, if we set $a(\boldsymbol{W}, \boldsymbol{x}) = 0$, $\gamma = 0$, and choose ℓ_t to be the hinge loss we recover an algorithm that closely resembles the PERCEPTRON (Rosenblatt, 1958), which can be interpreted as OGD on the hinge loss[3]. GAPTRON also uses OGD, which is used to update weight matrix $\boldsymbol{W}_t$, which in turn is used to form distribution $\boldsymbol{p}_t'$. For convenience we will define $a_t = a(\boldsymbol{W}_t, \boldsymbol{x}_t)$.

The role of a, which we will refer to as the gap map, is to exploit the gap between the surrogate loss and the zero-one loss. Before we explain how we exploit said gap we first present the expected mistake bound of GAPTRON in Lemma 25. The proof of Lemma 25 follows from applying the regret bound of OGD and working out the expected number of mistakes. The formal proof can be found in Section 6.7.

Lemma 25. *For any $\boldsymbol{U} \in \mathcal{W}$ Algorithm 13 satisfies*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y_t' \neq y_t]\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\boldsymbol{U})\right] + \frac{\|\boldsymbol{U}\|^2}{2\eta} + \gamma\frac{K-1}{K}T \\
&+ \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right]}_{\text{surrogate gap}}.
\end{aligned}$$

As we mentioned before, standard classifiers such as the PERCEPTRON simply set $a(\boldsymbol{W}, \boldsymbol{x}) = 0$ and upper bound $\mathbb{1}[y_t^\star \neq y_t] - \ell_t(\boldsymbol{W}_t)$ by 0. In the full information setting we can set $\gamma = 0$ and $\eta = \sqrt{\frac{\|\boldsymbol{U}\|^2}{\sum_{t=1}^T \|\boldsymbol{g}_t\|^2}}$ to obtain[4] $M_T \leq \sum_{t=1}^T \ell_t(\boldsymbol{U}) + \|\boldsymbol{U}\|\sqrt{\sum_{t=1}^T \|\boldsymbol{g}_t\|^2}$. However, the gap between the surrogate loss and the zero-one

[3] Other interpretations exist which lead to possibly better guarantees, see for example Beygelzimer et al. (2017).

[4] Although such tuning is impossible due to not knowing $\|\boldsymbol{U}\|$ or $\sum_{t=1}^T \|\boldsymbol{g}_t\|^2$ there exist algorithms that are able to achieve the same guarantee up to logarithmic factors, see for example Cutkosky and Orabona (2018).

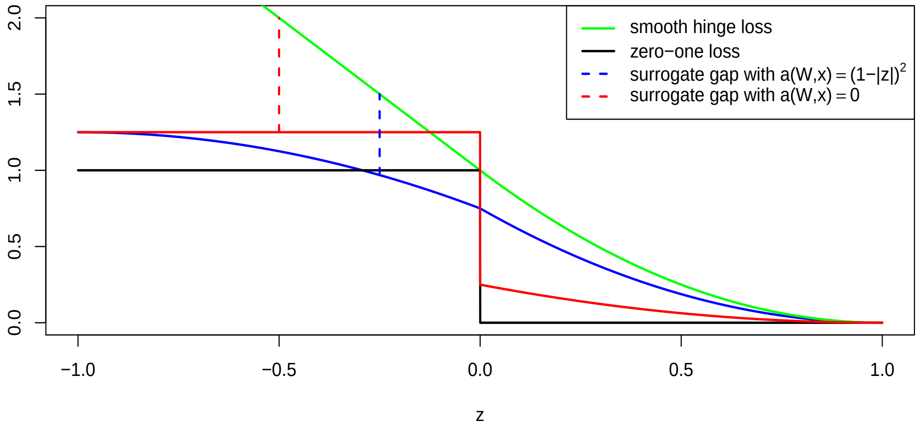

Figure 6.1: The surrogate gap for the smooth hinge loss as a function of margin z with $\eta = \frac{1}{8}$, $\gamma = 0$, and $\|\boldsymbol{x}\| = 1$. The solid red line is given by $\mathbb{1}[z \leq 0] + \frac{\eta}{2}\|\boldsymbol{g}\|^2$, where $\|\boldsymbol{g}\|^2 = 4(1-z)^2$ if $z > 0$ and $\|\boldsymbol{g}\|^2 = 4$ otherwise. The solid blue line is given by $(1-(1-|z|)^2)\mathbb{1}[z \leq 0] + \frac{1}{2}(1-|z|)^2 + \frac{\eta}{2}\|\boldsymbol{g}\|^2$. The surrogate gap is positive whenever the red or blue line is above the green line.

loss can be large. In fact, even with $a(\boldsymbol{W}, \boldsymbol{x}) = 0$, the gap between the zero-one loss and the surrogate loss is large enough to bound $\mathbb{1}[y_t^\star \neq y_t] - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2$ by 0 for some loss functions and values of $\boldsymbol{W}_t$ and $\boldsymbol{x}_t$.

In Figure 6.1 we can see a depiction of the surrogate gap for the smooth hinge loss for $K = 2$ (Rennie and Srebro, 2005) in the full information setting (see Section 6.4.3 for the definition of the smooth multiclass hinge loss). In the case where $K = 2$, $\boldsymbol{W}$ is a vector rather than a matrix and outcomes y_t are coded as $\{-1, 1\}$. We see that with $a(\boldsymbol{W}, \boldsymbol{x}) = 0$, only when margin $z = y\langle \boldsymbol{W}, \boldsymbol{x}\rangle \in [-0.125, 0]$ the surrogate gap is not upper bounded by 0. Decreasing η would increase the range for which the surrogate gap is bounded by zero, but only for $\eta = 0$ the surrogate gap is bounded by 0 everywhere. However, with $a(\boldsymbol{W}, \boldsymbol{x}) = (1-|z|)^2$ the surrogate gap is upper bounded by 0 for all z, which leads to constant surrogate regret. The remainder of the chapter is concerned with deriving different a for different loss functions for which the surrogate gap is bounded by 0. In the following section we start in the full information setting.

6.4 Full Information Multiclass Classification

In this section we derive gap maps that allow us to upper bound the surrogate gap by 0 for the logistic loss, hinge loss, and smooth hinge loss in the full information setting. Throughout this section we will set $\gamma = 0$. We start with the result for logistic loss.

6.4.1 Logistic Loss

The logistic loss is defined as

$$\ell_t(\boldsymbol{W}) = -\log_2(\sigma(\boldsymbol{W}, \boldsymbol{x}_t, y_t)), \tag{6.4.1}$$

where $\sigma(\boldsymbol{W}, \boldsymbol{x}, k) = \frac{\exp(\langle \boldsymbol{W}^k, \boldsymbol{x}\rangle)}{\sum_{\tilde{k}=1}^{K} \exp(\langle \boldsymbol{W}^{\tilde{k}}, \boldsymbol{x}\rangle)}$ is the softmax function. For the logistic loss we will use the following gap map:

$$a(\boldsymbol{W}_t, \boldsymbol{x}_t) = 1 - \mathbb{1}[p_t^\star \geq 0.5]p_t^\star,$$

where $p_t^\star = \max_k \sigma(\boldsymbol{W}_t, \boldsymbol{x}_t, k)$. This means that GAPTRON samples a label uniformly at random as long as $p_t^\star \leq 0.5$. While this may appear counter-intuitive at first sight note that when $p_t^\star < 0.5$ the zero-one loss is upper bounded by the logistic loss regardless of what we play since $-\log_2(p) \geq 1$ for $p \in [0, 0.5]$, which we use to show that the surrogate gap is bounded by 0 whenever $p_t^\star < 0.5$. The mistake bound of GAPTRON can be found in Theorem 30. To prove Theorem 30 we show that the surrogate gap is bounded by 0 and then use Lemma 25. The formal proof can be found in Section 6.8.1.

Theorem 30. *Let* $a(\boldsymbol{W}_t, \boldsymbol{x}_t) = 1 - \mathbb{1}[p_t^\star \geq 0.5]p_t^\star$, $\eta = \frac{\ln(2)}{2KX^2}$, $\gamma = 0$, *and let* ℓ_t *be the logistic loss defined in* (6.4.1). *Then for any* $\boldsymbol{U} \in \mathcal{W}$ *Algorithm 13 satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y_t' \neq y_t]\right] \leq \sum_{t=1}^{T} \ell_t(\boldsymbol{U}) + \frac{KX^2\|\boldsymbol{U}\|^2}{\ln(2)}.$$

Let us compare the mistake bound of GAPTRON with other results for logistic loss. Foster et al. (2018a) circumvent a lower bound for online logistic regression by Hazan et al. (2014) by using an improper learning algorithm and achieve $O(dK\ln(DT + 1))$ surrogate regret. Unfortunately this algorithm is impractical since the running time can be of order $O(D^6 \max\{dK, T\}^{12}T)$. In the case where $K = 2$ Jézéquel et al. (2020) provide a faster improper learning algorithm called AIOLI based on the Vovk-Azoury-Warmuth forecaster (Vovk, 2001; Azoury and Warmuth, 2001) that has running time $O(d^2T)$ and a surrogate regret of order $O(dD\ln(T))$. Unfortunately it is not known if AIOLI can be extended to $K > 2$. An alternative algorithm is ONS, which has running time $O((dK)^2T)$ but a surrogate regret bound of order $O(\exp(D)dK\ln(T + 1))$. With standard OGD we could degrade the dependence on T to improve the dependence on D to find a surrogate regret of order $O(D\sqrt{T})$ with an algorithm that has running time $O(dKT)$. Depending on $\|\boldsymbol{U}\|^2$ the surrogate regret of Gaptron can be significantly smaller than the surrogate regret of the aforementioned algorithms as the surrogate regret of Gaptron is independent of T and d. Furthermore, since Gaptron uses OGD to update $\boldsymbol{W}_t$ the running time is $O(dKT)$, significantly improving upon the running time of previous algorithms with comparable mistake bounds.

6.4.2 Multiclass Hinge Loss

We use a variant of the multiclass hinge loss of Crammer and Singer (2001), which is defined as:

$$\ell_t(\boldsymbol{W}) = \begin{cases} \max\{1 - m_t(\boldsymbol{W}, y_t), 0\} & \text{if } m_t^\star \leq \beta \\ \max\{1 - m_t(\boldsymbol{W}, y_t), 0\} & \text{if } y_t^\star \neq y_t \text{ and } m_t^\star > \beta \\ 0 & \text{if } y_t^\star = y_t \text{ and } m_t^\star > \beta, \end{cases} \tag{6.4.2}$$

where $m_t(\boldsymbol{W}_t, y) = \langle \boldsymbol{W}_t^y, \boldsymbol{x}_t\rangle - \max_{k \neq y}\langle \boldsymbol{W}_t^k, \boldsymbol{x}_t\rangle$ and $m_t^\star = \max_k m_t(\boldsymbol{W}_t, k)$. Note that we set $\ell_t(\boldsymbol{W}) = 0$ when $y_t^\star = y_t$ and $m_t^\star > \beta$. In common implementations of the Perceptron $\ell_t(\boldsymbol{W}) = 0$ whenever $y_t^\star = y_t$ (see for example Kakade et al. (2008)). However, for the surrogate gap to be bounded by zero we need ℓ_t to be positive whenever $a_t > 0$ otherwise there is nothing to cancel out the $a_t\frac{K-1}{K}$ term. The gap map for the hinge loss is $a(\boldsymbol{W}_t, \boldsymbol{x}_t) = 1 - \max\{\mathbb{1}[m_t^\star > \beta], m_t^\star\}$. This means that whenever $m_t^\star > \beta$ the predictions of Gaptron are identical to the predictions of the Perceptron. The mistake bound of Gaptron for the hinge loss can be found in Theorem 31 (its proof is deferred to Section 6.8.2).

Theorem 31. *Set* $a(\boldsymbol{W}_t, \boldsymbol{x}_t) = 1 - \max\{\mathbb{1}[m_t^\star > \beta], m_t^\star\}$, $\eta = \frac{1-\beta}{KX^2}$, $\gamma = 0$, *and let* ℓ_t *be the multiclass hinge loss defined in* (6.4.2) *with* $\beta = \frac{1}{K}$. *Then for any* $\boldsymbol{U} \in \mathcal{W}$ *Algorithm 13 satisfies*

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq \sum_{t=1}^T \ell_t(\boldsymbol{U}) + \frac{K^2X^2\|\boldsymbol{U}\|^2}{2(K-1)}.$$

Let us compare the mistake bound of Gaptron with the mistake bound of the Perceptron. The Perceptron guarantees $M_T \leq \sum_{t=1}^T \ell_t(\boldsymbol{U}) + X^2\|\boldsymbol{U}\|^2 + 2X\|\boldsymbol{U}\|\sqrt{2\sum_{t=1}^T \ell_t(\boldsymbol{U})}$ (see Beygelzimer et al. (2017) for a proof). The factor K in the surrogate regret of Gaptron is due to the cost of exploring uniformly at random. For small K the mistake bound of Gaptron can be significantly smaller in the adversarial case, but for large K the cost of sampling uniformly at random can be too high and the mistake bound of Gaptron can be larger than that of the PERCEPTRON. In the separable case the PERCEPTRON has a strictly better guarantee for any K since then only the $X^2\|\boldsymbol{U}\|^2$ term remains.

Orabona et al. (2012) show that for all loss functions of the form $\ell_t(\boldsymbol{W}) = \max\{1 - \frac{2}{2-\kappa}m_t(\boldsymbol{W}, y_t) + \frac{\kappa}{2-\kappa}m_t(\boldsymbol{W}, y_t)^2, 0\}$ the second-order PERCEPTRON guarantees $M_T \leq \sum_{t=1}^T \ell_t(\boldsymbol{U}) + O(\frac{\kappa}{2-\kappa}X^2\|\boldsymbol{U}\|^2 + \frac{dK}{\kappa(2-\kappa)}\ln(\sum_{t=1}^T \ell_t(\boldsymbol{U}) + 1)$. Thus, for small K GAPTRON always has a smaller surrogate regret term but for larger K the guarantee of GAPTRON can be worse, although this also depends on the performance and norm of the comparator $\boldsymbol{U}$.

6.4.3 Smooth Multiclass Hinge Loss

The smooth multiclass hinge loss (Rennie and Srebro, 2005) is defined as

$$\ell_t(\boldsymbol{W}) = \begin{cases} \max\{1 - 2m_t(\boldsymbol{W}, y_t), 0\} & \text{if } m_t(\boldsymbol{W}, y_t) \leq 0 \\ \max\{(1 - m_t(\boldsymbol{W}, y_t))^2, 0\} & \text{if } m_t(\boldsymbol{W}, y_t) > 0, \end{cases} \tag{6.4.3}$$

where $m_t(\boldsymbol{W}_t, y) = \langle \boldsymbol{W}_t^y, \boldsymbol{x}_t\rangle - \max_{k\neq y}\langle \boldsymbol{W}_t^k, \boldsymbol{x}_t\rangle$ as in Section 6.4.3. This loss function is not exp-concave nor is it strongly-convex. This means that with standard methods from Online Convex Optimization we cannot hope to achieve a better surrogate regret bound than $O(D\sqrt{T})$ in the worst-case. Theorem 32 shows that with gap map $a(\boldsymbol{W}_t, \boldsymbol{x}_t) = (1 - \min\{1, m_t^\star\})^2$, where $m_t^\star = \max_k m_t(\boldsymbol{W}_t, k)$, GAPTRON has a $O(K)$ surrogate regret bound. The proof of Theorem 32 follows from bounding the surrogate gap by zero and can be found in Section 6.8.3.

Theorem 32. *Set* $a(\boldsymbol{W}_t, \boldsymbol{x}_t) = (1 - \min\{1, m_t^\star\})^2$, $\eta = \frac{1}{4KX^2}$, $\gamma = 0$, *and let* ℓ_t *be the smooth multiclass hinge loss defined in* (6.4.3). *Then for any* $\boldsymbol{U} \in \mathcal{W}$ *Algorithm 13 satisfies*

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq \sum_{t=1}^T \ell_t(\boldsymbol{U}) + 2KX^2\|\boldsymbol{U}\|^2.$$

6.5 Bandit Multiclass Classification

In this section we will analyse GAPTRON in the bandit multiclass classification setting. While in the full information setting the fact that GAPTRON is a randomized algorithm can be seen as a drawback, in the adversarial bandit setting it is actually a requirement (see for example chapter 11 by Lattimore and Szepesvári (2018)). We will use the same gap maps as in the full information setting. The only difference is how we feed the surrogate loss to GAPTRON. We will use the same loss functions as in the full information setting but now multiplied by $\mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}$, which is simply importance weighting. This also means that, compared to the full information setting, the gradients that OGD uses to update weight matrix $\boldsymbol{W}_t$ are multiplied by $\mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}$. To control the surrogate gap we set $\gamma > 0$, which allows us to bound the variance of the norm of the gradients. The proofs in this section follow the same structure as in the full information setting, with the notable change that we suffer increased surrogate regret due to the $\gamma\frac{K-1}{K}T$ bias term and the increased $\mathbb{E}[\|\boldsymbol{g}_t\|^2] = O(\frac{K}{\gamma})$ term.

The results in this section provide three new answers to the open problem by Abernethy and Rakhlin (2009), who posed the problem of obtaining an efficient algorithm with $O(K\sqrt{T})$ surrogate regret. Several solutions with various loss function have been proposed. Beygelzimer et al. (2017) solved the open problem using an algorithm called SOBA. SOBA is a second-order algorithm which is analysed using a family of surrogate loss functions introduced by Orabona et al. (2012) ranging from the standard multiclass hinge loss to the squared multiclass hinge loss. The loss functions are parameterized by κ, where $\kappa = 0$ corresponds to the multiclass hinge loss and $\kappa = 1$ corresponds to the squared hinge loss. Simultaneously for all surrogate loss functions in the family of loss functions SOBA suffers a surrogate regret of order $O(\|\boldsymbol{U}\|^2X^2 + \frac{K}{\kappa}\sqrt{dT\ln(T+1)})$ and has a running time of order $O((dK)^2T)$. Hazan and Kale (2011) consider the logistic loss and obtain surrogate regret of order $O(dK^3\min\{\exp(DX)\ln(T+1), DXT^{\frac{2}{3}}\})$. Hazan and Kale (2011) also obtain $DX\sqrt{T}$ surrogate regret for a variant of the logistic loss function we consider in this chapter. Both results of Hazan and Kale (2011) are obtained by running ONS on (a variant of) the logistic loss, which has running time $O((dK)^2T)$. Foster et al. (2018a) introduce OBAMA, which improves the results of Hazan and Kale (2011) and suffers $O(\min\{dK^2\ln(TDX+1), K\sqrt{dT\ln(TDX+1)}\})$ surrogate regret for the logistic loss. Unfortunately, OBAMA has running time $O(D^6\max\{dK, T\}^{12}T)$.

GAPTRON is the first $O(dKT)$ running time algorithm which has $O(DK\sqrt{T})$ surrogate regret in bandit multiclass classification with respect to the logistic, hinge, or smooth hinge loss. GAPTRON also improves the surrogate regret bounds of previous algorithms with $O(DK\sqrt{T})$ surrogate regret by a factor $O(\sqrt{d\log(T+1)})$. The remainder of this section provides the settings for GAPTRON to achieve these results, starting with the logistic loss.

6.5.1 Bandit Logistic Loss

The bandit version of the logistic loss is defined as:

$$\ell_t(\boldsymbol{W}) = -\mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}\log_2(\sigma(\boldsymbol{W}, \boldsymbol{x}_t, y_t)). \tag{6.5.1}$$

A similar definition of the bandit logistic loss is used by Hazan and Kale (2011); Foster et al. (2018a). It is straightforward to verify that $\mathbb{E}_t[\ell_t(\boldsymbol{w})]$ is equivalent to its full information counterpart (6.4.1). This loss is a factor $\frac{1}{\ln(2)}$ larger than the loss used by Hazan and Kale (2011); Foster et al. (2018a), who use the natural logarithm instead of the logarithm with base 2. To stay consistent with the full information setting we opt to use base 2 in the bandit setting. Using GAPTRON with the natural logarithm will give similar results.

The mistake bound of GAPTRON for this loss can be found in Theorem 33 (its proof can be found in Section 6.9.1). Compared to OBAMA, which achieves a surrogate regret bound of order $O(\min\{dK^2\ln(TDX+1), K\sqrt{dT\ln(TDX+1)}\})$, GAPTRON has a larger dependency on D and X. However, the mistake bound of GAPTRON does not depend on d, which can be a significant improvement over the surrogate regret bound of OBAMA. Theorem 33 answers the two questions by Hazan and Kale (2011) affirmatively; GAPTRON is a linear time algorithm with exponentialy improved constants in the surrogate regret bound compared to NEWTRON.

Theorem 33. *Let* $a(\boldsymbol{W}_t, \boldsymbol{x}_t) = 1 - \mathbb{1}[p^\star_t \geq 0.5]p^\star_t$, $\eta = \frac{\ln(2)((1-\gamma)\exp(-2DX)\frac{1}{K}+\gamma)}{2K^2X^2}$, *and let* ℓ_t *be the bandit logistic loss* (6.5.1). *Then there exists a setting of* γ *such that Algorithm 13 satisfies*

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y'_t \neq y_t]\right]\\ &\leq \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right] + KXD\min\left\{\max\left\{\frac{2KXD}{\ln(2)}, 2\sqrt{\frac{T}{\ln(2)}}\right\}, \frac{KXD}{e^{-2DX}\ln(2)}\right\}.\end{aligned}$$

6.5.2 Bandit Multiclass Hinge Loss

We use the following definition of the bandit multiclass hinge loss:

$$\ell_t(\boldsymbol{W}_t) = \begin{cases}\mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}\max\{1 - m_t(\boldsymbol{W}_t, y_t), 0\} & \text{if } m^\star_t \leq \beta\\ \mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}\max\{1 - m_t(\boldsymbol{W}_t, y_t), 0\} & \text{if } y^\star_t \neq y_t \text{ and } m^\star_t > \beta\\ 0 & \text{if } y'_t = y^\star_t = y_t \text{ and } m^\star_t > \beta.\end{cases} \tag{6.5.2}$$

It is straightforward to see that the conditional expectation of the bandit multiclass hinge loss is the full information multiclass hinge loss. Both the BANDITRON algorithm (Kakade et al., 2008) and SOBA (Beygelzimer et al., 2017) use a similar loss function.

As we mentioned before, Beygelzimer et al. (2017) present SOBA, which is a second-order algorithm with surrogate regret $O(\|\boldsymbol{U}\|^2X^2 + \frac{K}{\kappa}\sqrt{dT}\ln(T+1))$. BANDITRON is a first-order algorithm based on the PERCEPTRON algorithm and suffers $O((KDX)^{1/3}T^{2/3})$ surrogate regret. For the more general setting of contextual bandits (Foster and Krishnamurthy, 2018) use continuous Exponential Weights with the hinge loss to also obtain an $O(KDX\sqrt{dT}\ln(T+1))$ surrogate regret bound with a polynomial time algorithm. The expected mistake bound of GAPTRON can be found in Theorem 34 and its proof can be found in Section 6.9.2. Compared to the BANDITRON GAPTRON has larger surrogate regret in terms of D, K, and X, but smaller surrogate regret in terms of T. Compared to the surrogate regret of SOBA the surrogate regret of GAPTRON does not contain a factor $\sqrt{d\ln(T+1)}$.

Theorem 34. *Set* $a(\boldsymbol{W}_t, \boldsymbol{x}_t) = 1 - \max\{\mathbb{1}[m_t^\star > \beta], m_t^\star\}$, $\eta = \frac{\gamma(1-\beta)}{K^2X^2}$, $\gamma = \min\left\{1, \sqrt{\frac{K^3X^2D^2}{2(1-\beta)(K-1)T}}\right\}$, *and let* ℓ_t *be the bandit multiclass hinge loss defined in* (6.5.2) *with* $\beta = \frac{1}{K}$. *Then for any* $\boldsymbol{U} \in \mathcal{W}$ *Algorithm 13 satisfies*

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right] + \max\left\{\frac{K^3X^2D^2}{K-1},\ 2KXD\sqrt{\frac{T}{2}}\right\}.$$

6.5.3 Bandit Smooth Multiclass Hinge Loss

In this section we use the following loss function:

$$\ell_t(\boldsymbol{W}) = \begin{cases} \mathbb{1}[y_t' = y_t]p_t'(y_t')^{-1}\max\{1 - 2m_t(\boldsymbol{W}, y_t), 0\} & \text{if } m_t(\boldsymbol{W}, y_t) \leq 0 \\ \mathbb{1}[y_t' = y_t]p_t'(y_t')^{-1}\max\{(1 - m_t(\boldsymbol{W}, y_t))^2, 0\} & \text{if } m_t(\boldsymbol{W}, y_t) > 0. \end{cases} \tag{6.5.3}$$

This loss function is the bandit version of the smooth multiclass hinge loss that we we used in Section 6.4.3 and its expectation is equivalent to its full information counterpart in equation (6.4.3). The surrogate regret of GAPTRON with this loss function can be found in Theorem 35. The proof of Theorem 35 can be found in Section 6.9.3.

Theorem 35. *Set* $a(\boldsymbol{W}_t, \boldsymbol{x}_t) = (1 - \min\{1, m_t^\star\})^2$, $\eta = \frac{\gamma}{4K^2X^2}$, $\gamma = \min\left\{1, \sqrt{\frac{4K^2X^2D^2}{T}}\right\}$, *and let* ℓ_t *be the bandit smooth multiclass hinge loss* *defined in* (6.5.3). *Then for any* $\boldsymbol{U} \in \mathcal{W}$ *Algorithm 13 satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y_t' \neq y_t]\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\boldsymbol{U})\right] + \max\left\{4K^2X^2D^2,\ 2KXD\sqrt{2T}\right\}.$$

6.6 Conclusion

In this chapter we introduced Gaptron, a randomized first-order algorithm for the full and bandit information multiclass classification settings. Using a new technique we showed that Gaptron has a $O(K)$ surrogate regret bound in the full information setting and a surrogate regret bound of order $O(K\sqrt{T})$ in the bandit setting. One of the main drawbacks of Gaptron is that it is a randomized algorithm. Our bounds only hold in expectation and it would be interesting to show similar bounds also hold with high probability. Another interesting venue to explore is how to extend the ideas in this chapter to the stochastic setting or the more general contextual bandit setting. In future work we would like to conduct experiments to compare Gaptron with other algorithms, particularly in the bandit setting.

6.7 Details of Section 6.3

Proof of Lemma 25. As we said before, the updates of $\boldsymbol{W}_t$ are Online Gradient Descent (Zinkevich, 2003), which guarantees

$$\sum_{t=1}^{T} (\ell_t(\boldsymbol{W}_t) - \ell_t(\boldsymbol{U})) \leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \sum_{t=1}^{T} \frac{\eta}{2}\|\boldsymbol{g}_t\|^2. \tag{6.7.1}$$

Now, by using (6.7.1) and $\mathbb{E}_t[\mathbb{1}[y'_t \neq y_t]] = (1 - \max\{a_t, \gamma\})\mathbb{1}[y^\star_t \neq y_t] + \max\{a_t, \gamma\}\frac{K-1}{K}$ we find

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}[y'_t \neq y_t] - \ell_t(\boldsymbol{U})\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}[y'_t \neq y_t] - \ell_t(\boldsymbol{W}_t)\right) + \sum_{t=1}^{T}\left(\ell_t(\boldsymbol{W}_t) - \ell_t(\boldsymbol{U})\right)\right] \\
&\le \frac{\|\boldsymbol{U}\|^2}{2\eta} + \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}[y'_t \neq y_t] - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right)\right] \\
&= \frac{\|\boldsymbol{U}\|^2}{2\eta} + \mathbb{E}\left[\sum_{t=1}^{T}\Big((1 - \max\{a_t, \gamma\})\mathbb{1}[y^\star_t \neq y_t]\right. \qquad (6.7.2) \\
&\qquad \left. + \max\{a_t, \gamma\}\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\Big)\right] \\
&\le \frac{\|\boldsymbol{U}\|^2}{2\eta} + \gamma\frac{K-1}{K}T + \mathbb{E}\left[\sum_{t=1}^{T}\Big((1 - a_t)\mathbb{1}[y^\star_t \neq y_t]\right. \\
&\qquad \left. + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\Big)\right],
\end{aligned}$$

where in the last inequality we used $(1-\max\{a_t, \gamma\}) \le (1-a_t)$ and $\max\{a_t, \gamma\} \le a_t + \gamma$. Adding $\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\boldsymbol{U})\right]$ to both sides of equation (6.7.2) completes the proof. □

6.8 Details of Full Information Multiclass Classification

6.8.1 Details of Section 6.4.1

Proof of Theorem 30. We will prove the Theorem by showing that the surrogate gap is bounded by 0 and then using Lemma 25. The gradient of the logistic loss evaluated at $\boldsymbol{W}_t$ is given by:

$$\nabla \ell_t(\boldsymbol{W}_t) = \frac{1}{\ln(2)}(\tilde{\boldsymbol{p}}_t - \boldsymbol{e}_{y_t}) \otimes \boldsymbol{x}_t,$$

where $\tilde{\boldsymbol{p}}_t = (\tilde{p}_t(1), \dots, \tilde{p}_t(k))^\intercal$ and $\tilde{p}_t(k) = \sigma(\boldsymbol{W}_t, \boldsymbol{x}_t, k)$.

We continue by writing out the surrogate gap:

$$
\begin{aligned}
&(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2 \\
&\le (1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) - \frac{\eta}{\ln(2)}\|\boldsymbol{x}_t\|^2 \log_2(\tilde{p}_t(y_t)) \\
&\le (1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) - \frac{\eta}{\ln(2)}X^2 \log_2(\tilde{p}_t(y_t)) \\
&= \begin{cases}
0 + \frac{K-1}{K} + \log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)) & \text{if } p_t^\star < 0.5 \\
p_t^\star + (1-p_t^\star)\frac{K-1}{K} + \log_2(\tilde{p}_t(y_t)) & \\
\quad -\frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)) & \text{if } y_t^\star \neq y_t \text{ and } p_t^\star \geq 0.5 \\
(1-p_t^\star)\frac{K-1}{K} + \log_2(p_t^\star) - \frac{\eta}{\ln(2)}X^2\log_2(p_t^\star) & \text{if } y_t^\star = y_t \text{ and } p_t^\star \geq 0.5,
\end{cases}
\end{aligned}
\tag{6.8.1}
$$

where the first inequality is due to Lemma 26 below.

We now split the analysis into the cases in (6.8.1). We start with $p_t^\star < 0.5$. In this case we use $1 \leq -\log_2(x)$ for $x \in [0, \frac{1}{2}]$ and obtain

$$
\begin{aligned}
&\frac{K-1}{K} + \log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)) \\
&\le -\frac{K-1}{K}\log_2(\tilde{p}_t(y_t)) + \log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)) \\
&= \frac{1}{K}\log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)),
\end{aligned}
$$

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

The second case we consider is when $y_t^\star \neq y_t$ and $p_t^\star \geq 0.5$. 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

$$\begin{aligned}
&p_t^\star + (1-p_t^\star)\frac{K-1}{K} + \log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)) \\
&\leq -\tfrac{1}{2}\log_2(1-p_t^\star) - \frac{K-1}{K}\tfrac{1}{2}\log_2(1-p_t^\star) + \log_2(\tilde{p}_t(y_t)) \\
&\quad - \frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)) \\
&= -\tfrac{1}{2}\log_2\left(\sum_{k\neq y_t}^{K}\tilde{p}_t(k)\right) - \frac{K-1}{K}\tfrac{1}{2}\log_2\left(\sum_{k\neq y_t}^{K}\tilde{p}_t(k)\right) + \log_2(\tilde{p}_t(y_t)) \\
&\quad - \frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)) \\
&\leq -\tfrac{1}{2}\log_2\left(\tilde{p}_t(y_t)\right) - \frac{K-1}{K}\tfrac{1}{2}\log_2\left(\tilde{p}_t(y_t)\right) + \log_2(\tilde{p}_t(y_t)))) \\
&\quad - \frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)) \\
&= \frac{1}{2K}\log_2\left(\tilde{p}_t(y_t)\right) - \frac{\eta}{\ln(2)}X^2\log_2(\tilde{p}_t(y_t)),
\end{aligned}$$

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

The last case we need to consider is $y_t^\star = y_t$ and $p_t^\star \geq 0.5$. In this case we use $1 - x \leq -\log_2(x)$ and obtain

$$\begin{aligned}
&(1-p_t^\star)\frac{K-1}{K} + \log_2(p_t^\star) - \frac{\eta}{\ln(2)}X^2\log_2(p_t^\star) \\
&\leq -\frac{K-1}{K}\log_2(p_t^\star) + \log_2(p_t^\star) - \frac{\eta}{\ln(2)}X^2\log_2(p_t^\star),
\end{aligned}$$

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

We now apply Lemma 25, plug in $\gamma = 0$, and use the above to find:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y_t' \neq y_t]\right] \\
&\leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \sum_{t=1}^{T} \ell_t(\boldsymbol{U}) + \gamma\frac{K-1}{K}T \\
&\quad + \sum_{t=1}^{T}\Big((1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\Big) \\
&\leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \sum_{t=1}^{T} \ell_t(\boldsymbol{U}).
\end{aligned}$$

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

□

Lemma 26. *Let ℓ_t be the logistic loss* (6.4.1)*, then*

$$\|\nabla \ell_t(\boldsymbol{W}_t)\|^2 \leq \frac{2}{\ln(2)}\|\boldsymbol{x}_t\|^2 \ell_t(\boldsymbol{W}_t).$$

Proof. We have

$$\begin{aligned}
\|\nabla \ell_t(\boldsymbol{W}_t)\|^2 =& \frac{1}{\ln(2)^2}\|\boldsymbol{x}_t\|^2\left(\sum_{k=1}^{K}(\mathbb{1}[y_t = k] - \tilde{p}_t(k))^2\right) \\
\leq& \frac{1}{\ln(2)^2}\|\boldsymbol{x}_t\|^2\left(\sum_{k=1}^{K}|\mathbb{1}[y_t = k] - \tilde{p}_t(k)|\right)^2 \\
\leq& -2\frac{1}{\ln(2)}\|\boldsymbol{x}_t\|^2\log_2(\tilde{p}_t(y_t)) \\
=& 2\frac{1}{\ln(2)}\|\boldsymbol{x}_t\|^2\ell_t(\boldsymbol{W}_t),
\end{aligned}$$

where the last inquality follows from Pinsker's inequality (Cover and Thomas, 1991, Lemma 12.6.1). □

6.8.2 Details of Section 6.4.2

Proof of Theorem 31. We will prove the Theorem by showing that the surrogate gap is bounded by 0 and then using Lemma 25. Let $\tilde{k} = \arg\max_{k \neq y_t}\langle \boldsymbol{W}_t^k, \boldsymbol{x}_t\rangle$.

The gradient of the smooth multiclass hinge loss is given by

$$\nabla \ell_t(\boldsymbol{W}_t) = \begin{cases} (\boldsymbol{e}_{\tilde{k}} - \boldsymbol{e}_{y_t}) \otimes \boldsymbol{x}_t & \text{if } y_t^\star \neq y_t \\ (\boldsymbol{e}_{\tilde{k}} - \boldsymbol{e}_{y_t}) \otimes \boldsymbol{x}_t & \text{if } y_t^\star = y_t \text{ and } m_t^\star \leq \beta \\ 0 & \text{if } y_t^\star = y_t \text{ and } m_t^\star > \beta. \end{cases}$$

We continue by writing out the surrogate gap:

$$(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t \frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2 =$$
$$\begin{cases} m_t^\star + (1-m_t^\star)\frac{K-1}{K} - (1-m_t(\boldsymbol{W}_t, y_t)) + \eta\|\boldsymbol{x}_t\|^2 & \text{if } y_t^\star \neq y_t \text{ and } m_t^\star \leq \beta \\ (1-m_t^\star)\frac{K-1}{K} - (1-m_t^\star) + \eta\|\boldsymbol{x}_t\|^2 & \text{if } y_t^\star = y_t \text{ and } m_t^\star \leq \beta \\ 1 - (1-m_t(\boldsymbol{W}_t, y_t)) + \eta\|\boldsymbol{x}_t\|^2 & \text{if } y_t^\star \neq y_t \text{ and } m_t^\star > \beta \\ 0 & \text{if } y_t^\star = y_t \text{ and } m_t^\star > \beta. \end{cases} \tag{6.8.2}$$

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

$$\begin{aligned} m_t^\star + m_t(\boldsymbol{W}_t, y_t) =& \langle \boldsymbol{W}_t^{y_t^\star}, \boldsymbol{x}_t\rangle - \max_{k \neq y_t^\star}\langle \boldsymbol{W}_t^k, \boldsymbol{x}_t\rangle + \langle \boldsymbol{W}_t^{y_t}, \boldsymbol{x}_t\rangle - \max_{k \neq y_t}\langle \boldsymbol{W}_t^k, \boldsymbol{x}_t\rangle \\ =& \langle \boldsymbol{W}_t^{y_t}, \boldsymbol{x}_t\rangle - \max_{k \neq y_t^\star}\langle \boldsymbol{W}_t^k, \boldsymbol{x}_t\rangle \\ \leq& \langle \boldsymbol{W}_t^{y_t}, \boldsymbol{x}_t\rangle - \langle \boldsymbol{W}_t^{y_t}, \boldsymbol{x}_t\rangle = 0. \end{aligned} \tag{6.8.3}$$

We now split the analysis into the cases in (6.8.2). We start with $y_t^\star \neq y_t$ and $m_t^\star \leq \beta$, in which case the surrogate gap can be bounded by 0 when $\eta \leq \frac{1}{KX^2}$:

$$\begin{aligned} & m_t^\star + (1-m_t^\star)\frac{K-1}{K} - (1-m_t(\boldsymbol{W}_t, y_t)) + \eta\|\boldsymbol{x}_t\|^2 \\ &= m_t^\star + m_t(\boldsymbol{W}_t, y_t) + (1-m_t^\star)\frac{K-1}{K} - 1 + \eta\|\boldsymbol{x}_t\|^2 \\ &\leq -\frac{1}{K} + \eta X^2 && \text{(by equation (6.8.3))} \\ &\leq 0. \end{aligned}$$

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

$$\begin{aligned} (1-m_t^\star)\frac{K-1}{K} - (1-m_t^\star) + \eta\|\boldsymbol{x}_t\|^2 &= -(1-m_t^\star)\frac{1}{K} + \eta\|\boldsymbol{x}_t\|^2 \\ &\leq -\frac{1-\beta}{K} + \eta X^2, \end{aligned}$$

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

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

$$\begin{aligned} 1 - (1 - m_t(\boldsymbol{W}_t, y_t)) + \eta\|\boldsymbol{x}_t\|^2 =& m_t(\boldsymbol{W}_t, y_t) + \eta\|\boldsymbol{x}_t\|^2 \\ \leq& - m_t^\star + \eta\|\boldsymbol{x}_t\|^2 \qquad \text{(by equation (6.8.3))} \\ \leq& - \beta + \eta X^2, \end{aligned}$$

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

We now apply Lemma 25, plug in $\gamma = 0$, and use the above to find:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq& \frac{\|\boldsymbol{U}\|^2}{2\eta} + \sum_{t=1}^T \ell_t(\boldsymbol{U}) + \gamma T + \\ & \sum_{t=1}^T \left((1 - a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right) \\ \leq& \frac{\|\boldsymbol{U}\|^2}{2\eta} + \sum_{t=1}^T \ell_t(\boldsymbol{U}). \end{aligned}$$

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

6.8.3 Details of Section 6.4.3

Proof of Theorem 32. We will prove the Theorem by showing that the surrogate gap is bounded by 0 and then using Lemma 25. Let $\tilde{k} = \arg\max_{k \neq y_t} \langle \boldsymbol{W}_t^k, \boldsymbol{x}_t \rangle$. The gradient of the smooth multiclass hinge loss is given by

$$\nabla \ell_t(\boldsymbol{W}_t) = \begin{cases} 2(\boldsymbol{e}_{\tilde{k}} - \boldsymbol{e}_{y_t}) \otimes \boldsymbol{x}_t & \text{if } y_t^\star \neq y_t \\ 2(\boldsymbol{e}_{\tilde{k}} - \boldsymbol{e}_{y_t})(1 - m_t^\star) \otimes \boldsymbol{x}_t & \text{if } y_t^\star = y_t \text{ and } m_t^\star < 1 \\ 0 & \text{if } y_t^\star = y_t \text{ and } m_t^\star \geq 1. \end{cases}$$

We continue by writing out the surrogate gap:

$$(1 - a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2 =$$
$$\begin{cases} 2m_t^\star - m_t^{\star 2} + (1 - m_t^\star)^2\frac{K-1}{K} & \\ -(1 - 2m_t(\boldsymbol{W}_t, y_t)) + \eta 4\|\boldsymbol{x}_t\|^2 & \text{if } y_t^\star \neq y_t \text{ and } m_t^\star < 1 \\ (1 - m_t^\star)^2\frac{K-1}{K} - (1 - m_t^\star)^2 + \eta 4\|\boldsymbol{x}_t\|^2(1 - m_t^\star)^2 & \text{if } y_t^\star = y_t \text{ and } m_t^\star < 1 \\ 1 - (1 - 2m_t(\boldsymbol{W}_t, y_t)) + \eta 4\|\boldsymbol{x}_t\|^2 & \text{if } y_t^\star \neq y_t \text{ and } m_t^\star \geq 1 \\ 0 & \text{if } y_t^\star = y_t \text{ and } m_t^\star \geq 1. \end{cases} \tag{6.8.4}$$

We now split the analysis into the cases in (6.8.4). We start with the case where $y_t^\star \neq y_t$ and $m_t^\star < 1$. By using (6.8.3) we can see that with $\eta = \frac{1}{4KX^2}$ the surrogate gap is bounded by 0:

$$\begin{aligned}
&2m_t^\star - m_t^{\star 2} + (1 - m_t^\star)^2 \frac{K-1}{K} - (1 - 2m_t(\boldsymbol{W}_t, y_t)) + \eta 4 \|\boldsymbol{x}_t\|^2 \\
&= 2(m_t^\star + m_t(\boldsymbol{W}_t, y_t)) - m_t^{\star 2} + (1 - m_t^\star)^2 \frac{K-1}{K} - 1 + \eta 4 \|\boldsymbol{x}_t\|^2 \\
&\leq -m_t^{\star 2} + (1 - m_t^\star)^2 \frac{K-1}{K} - 1 + \eta 4 X^2 \qquad \text{(by equation (6.8.3))} \\
&\leq -\frac{1}{K} + \eta 4 X^2 \leq 0.
\end{aligned}$$

The next case we consider is when $y_t^\star = y_t$ and $m_t^\star < 1$. In this case we have

$$\begin{aligned}
&(1 - m_t^\star)^2 \frac{K-1}{K} - (1 - m_t^\star)^2 + \eta 4 \|\boldsymbol{x}_t\|^2 (1 - m_t^\star)^2 \\
&= -(1 - m_t^\star)^2 \frac{1}{K} + \eta 4 \|\boldsymbol{x}_t\|^2 (1 - m_t^\star)^2,
\end{aligned}$$

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

Finally, if $y_t^\star \neq y_t$ and $m_t^\star \geq 1$ then

$$\begin{aligned}
1 - (1 - 2m_t(\boldsymbol{W}_t, y_t)) + \eta 4 \|\boldsymbol{x}_t\|^2 =& 2m_t(\boldsymbol{W}_t, y_t) + \eta 4 \|\boldsymbol{x}_t\|^2 \\
\leq& -2m_t^\star + \eta 4 \|\boldsymbol{x}_t\|^2 \quad \text{(by equation (6.8.3))} \\
\leq& -2 + \eta 4 X^2,
\end{aligned}$$

which is bounded by 0 since $\eta < \frac{1}{2X^2}$. We apply Lemma 25 with $\gamma = 0$ and use the above to find:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq& \frac{\|\boldsymbol{U}\|^2}{2\eta} + \sum_{t=1}^T \ell_t(\boldsymbol{U}) + \gamma \frac{K-1}{K} T + \\
& \sum_{t=1}^T \left((1 - a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t \frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right) \\
\leq& \frac{\|\boldsymbol{U}\|^2}{2\eta} + \sum_{t=1}^T \ell_t(\boldsymbol{U}).
\end{aligned}$$

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

□

6.9 Details of Bandit Multiclass Classification

6.9.1 Details of Section 6.5.1

Proof of Theorem 33. First, by straightforward calculations we can see that $p'_t(y_t) \geq \frac{(1-\gamma)\exp(-2DX)+\gamma}{K} = \delta$. 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 25. By using $\mathbb{E}_t[\ell_t(\boldsymbol{W}_t)] = -\log_2(\tilde{p}_t(y_t))$ and $\mathbb{E}_t\left[\|\boldsymbol{g}_t\|^2\right] = \frac{1}{\ln(2)p'_t(y_t)}\|(\tilde{\boldsymbol{p}}_t - \boldsymbol{e}_{y_t}) \otimes \boldsymbol{x}_t\|^2$ we write out the surrogate gap:

$$
\begin{aligned}
&\mathbb{E}\left[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right] \\
&= \mathbb{E}\Bigg[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} + \log_2(\tilde{p}_t(y_t)) \\
&\quad + \frac{\eta}{2\ln(2)^2 p'_t(y_t)}\|(\tilde{\boldsymbol{p}}_t - \boldsymbol{e}_{y_t}) \otimes \boldsymbol{x}_t\|^2\Bigg] \\
&\leq \mathbb{E}\Bigg[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} + \log_2(\tilde{p}_t(y_t)) \\
&\quad - \frac{\eta}{\ln(2)p'_t(y_t)}X^2\log_2(\tilde{p}_t(y_t))\Bigg] \\
&= \begin{cases}
\frac{K-1}{K} + \mathbb{E}\Big[\log_2(\tilde{p}_t(y_t)) \\
\quad -\frac{\eta}{\ln(2)p'_t(y_t)}X^2\log_2(\tilde{p}_t(y_t))\Big] & \text{if } p_t^\star < 0.5 \\
\mathbb{E}\Big[p_t^\star + (1-p_t^\star)\frac{K-1}{K} + \log_2(\tilde{p}_t(y_t)) \\
\quad -\frac{\eta}{\ln(2)p'_t(y_t)}X^2\log_2(\tilde{p}_t(y_t))\Big] & \text{if } y_t^\star \neq y_t \text{ and } p_t^\star \geq 0.5 \\
\mathbb{E}\Big[(1-p_t^\star)\frac{K-1}{K} + \log_2(p_t^\star) \\
\quad -\frac{\eta}{\ln(2)p'_t(y_t^\star)}X^2\log_2(p_t^\star)\Big] & \text{if } y_t^\star = y_t \text{ and } p_t^\star \geq 0.5,
\end{cases}
\end{aligned}
\tag{6.9.1}
$$

where the first inequality is due to Lemma 26.

We now split the analysis into the cases in (6.9.1). We start with $p_t^\star < 0.5$. In this case we use $1 \leq -\log_2(x)$ for $x \in [0, \frac{1}{2}]$ and obtain

$$\begin{aligned}
&\frac{K-1}{K} + \mathbb{E}[\log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)p'_t(y_t)} X^2 \log_2(\tilde{p}_t(y_t))] \\
&\leq \mathbb{E}\left[-\frac{K-1}{K}\log_2(\tilde{p}_t(y_t)) + \log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)p'_t(y_t)} X^2 \log_2(\tilde{p}_t(y_t))\right] \\
&\leq \mathbb{E}\left[-\frac{K-1}{K}\log_2(\tilde{p}_t(y_t)) + \log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)\delta} X^2 \log_2(\tilde{p}_t(y_t))\right]
\end{aligned}$$

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

The second case we consider is when $y_t^\star \neq y_t$ and $p_t^\star \geq 0.5$. 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

$$\begin{aligned}
&\mathbb{E}\left[p_t^\star + (1-p_t^\star)\frac{K-1}{K} + \log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)p'_t(y_t)} X^2 \log_2(\tilde{p}_t(y_t))\right] \\
&\leq \mathbb{E}\Bigg[-\tfrac{1}{2}\log_2(1-p_t^\star) - \frac{K-1}{K}\tfrac{1}{2}\log_2(1-p_t^\star) + \log_2(\tilde{p}_t(y_t)) \\
&\quad - \frac{\eta}{\ln(2)\delta} X^2 \log_2(\tilde{p}_t(y_t))\Bigg] \\
&= \mathbb{E}\Bigg[-\tfrac{1}{2}\log_2\left(\sum_{k\neq y_t}^{K} \tilde{p}_t(k)\right) - \frac{K-1}{K}\tfrac{1}{2}\log_2\left(\sum_{k\neq y_t}^{K} \tilde{p}_t(k)\right) + \log_2(\tilde{p}_t(y_t)) \\
&\quad - \frac{\eta}{\ln(2)\delta} X^2 \log_2(\tilde{p}_t(y_t))\Bigg] \\
&\leq \mathbb{E}\Bigg[-\tfrac{1}{2}\log_2\left(\tilde{p}_t(y_t)\right) - \frac{K-1}{K}\tfrac{1}{2}\log_2\left(\tilde{p}_t(y_t)\right) + \log_2(\tilde{p}_t(y_t)) \\
&\quad - \frac{\eta}{\ln(2)\delta} X^2 \log_2(\tilde{p}_t(y_t))\Bigg] \\
&= \mathbb{E}\left[\frac{1}{2K}\log_2(\tilde{p}_t(y_t)) - \frac{\eta}{\ln(2)\delta} X^2 \log_2(\tilde{p}_t(y_t))\right],
\end{aligned}$$

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

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

$$\begin{aligned}
&\mathbb{E}\left[(1-p_t^\star)\frac{K-1}{K} + \log_2(p_t^\star) - \frac{\eta}{\ln(2)p_t'(y_t^\star)}X^2\log_2(p_t^\star)\right]\\
&\leq \mathbb{E}\left[-\frac{K-1}{K}\log_2(p_t^\star) + \log_2(p_t^\star) - \frac{\eta}{\ln(2)\delta}X^2\log_2(p_t^\star)\right],
\end{aligned}$$

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

We now apply Lemma 25 and use the above to find:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right]\\
&\leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right] + \gamma\frac{K-1}{K}T\\
&\quad + \sum_{t=1}^T \mathbb{E}\left[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right]\\
&\leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \gamma T + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right].
\end{aligned}$$

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

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq \frac{K^2X^2\|\boldsymbol{U}\|^2}{\ln(2)((1-\gamma)\exp(-2DX)+\gamma)} + \gamma T + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right],$$

Setting $\gamma = 0$ gives us

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq \frac{K^2X^2D^2}{\ln(2)\exp(-2DX)} + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right].$$

If instead we set $\gamma = \min\left\{1, \sqrt{\frac{K^2X^2D^2}{\ln(2)T}}\right\}$ we consider two cases. In the case where $1 \leq \sqrt{\frac{K^2X^2D^2}{T}}$ we have that $T \leq K^2X^2D^2$ and therefore

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq 2\frac{K^2X^2D^2}{\ln(2)} + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right].$$

In the case where $1 > \sqrt{\frac{K^2X^2D^2}{T}}$ we have that

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq 2KXD\sqrt{\frac{T}{\ln(2)}} + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right],$$

which after combining the above completes the proof. □

6.9.2 Details of Section 6.5.2

Proof of Theorem 34. First, note that $p'_t(y_t) \geq \frac{\gamma}{K}$. The proof proceeds in a similar way as in the full information setting (Theorem 31), except now we use that $p'_t(y_t) \geq \frac{\gamma}{K}$ to bound $\mathbb{E}_t[\|\boldsymbol{g}_t\|^2]$. We will prove the Theorem by showing that the surrogate gap is bounded by 0 and then using Lemma 25. By using $\mathbb{E}_t\left[\mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}\right] = 1$ and $\mathbb{E}_t\left[\left(\mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}\right)^2\right] = \mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}$ we start by splitting the surrogate gap in cases:

$$\begin{aligned}
&\mathbb{E}\left[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right] \\
&= \begin{cases}
\mathbb{E}\left[m_t^\star + (1-m_t^\star)\frac{K-1}{K} - (1-m_t(\boldsymbol{W}_t, y_t)) \right. & \\
\left. + \frac{\eta}{p'_t(y_t)}\|\boldsymbol{x}_t\|^2\right] & \text{if } y_t^\star \neq y_t \text{ and } m_t^\star \leq \beta \\
\mathbb{E}\left[(1-m_t^\star)\frac{K-1}{K} - (1-m_t^\star) + \frac{\eta}{p'_t(y_t)}\|\boldsymbol{x}_t\|^2\right] & \text{if } y_t^\star = y_t \text{ and } m_t^\star \leq \beta \\
\mathbb{E}\left[1 - (1-m_t(\boldsymbol{W}_t, y_t)) + \frac{\eta}{p'_t(y_t)}\|\boldsymbol{x}_t\|^2\right] & \text{if } y_t^\star \neq y_t \text{ and } m_t^\star > \beta \\
0 & \text{if } y_t^\star = y_t \text{ and } m_t^\star > \beta.
\end{cases}
\end{aligned} \tag{6.9.2}$$

We now split the analysis into the cases in (6.9.2). We start with $y_t^\star \neq y_t$ and $m_t^\star \leq \beta$. The surrogate gap can now be bounded by 0 when $\eta \leq \frac{\gamma}{K^2X^2}$:

$$\begin{aligned}
&\mathbb{E}\left[m_t^\star + (1-m_t^\star)\frac{K-1}{K} - (1-m_t(\boldsymbol{W}_t, y_t)) + \frac{\eta}{p'_t(y_t)}\|\boldsymbol{x}_t\|^2\right] \\
&= \mathbb{E}\left[m_t^\star + m_t(\boldsymbol{W}_t, y_t) + (1-m_t^\star)\frac{K-1}{K} - 1 + \frac{\eta}{p'_t(y_t)}\|\boldsymbol{x}_t\|^2\right] \\
&\leq -\frac{1}{K} + \frac{K\eta}{\gamma}X^2 && \text{(equation (6.8.3))} \\
&\leq 0.
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[(1-m_t^\star)\frac{K-1}{K} - (1-m_t^\star) + \eta\|\boldsymbol{x}_t\|^2\right] &= \mathbb{E}\left[-(1-m_t^\star)\frac{1}{K} + \frac{\eta}{p'_t(y_t)}\|\boldsymbol{x}_t\|^2\right] \\
&\leq -\frac{1-\beta}{K} + \frac{K\eta}{\gamma}X^2,
\end{aligned}$$

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

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

$$\begin{aligned}
&\mathbb{E}\left[1 - (1 - m_t(\boldsymbol{W}_t, y_t)) + \frac{\eta}{p_t'(y_t)}\|\boldsymbol{x}_t\|^2\right] \\
&= \mathbb{E}\left[m_t(\boldsymbol{W}_t, y_t) + \frac{\eta}{p_t'(y_t)}\|\boldsymbol{x}_t\|^2\right] \\
&\leq \mathbb{E}\left[-m_t^\star + \frac{\eta}{p_t'(y_t)}\|\boldsymbol{x}_t\|^2\right] && \text{(by equation (6.8.3))} \\
&\leq -\beta + \frac{K\eta}{\gamma}X^2,
\end{aligned}$$

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

We now apply Lemma 25 and use the above to find:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \\
&\leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right] + \gamma\frac{K-1}{K}T \\
&\quad + \sum_{t=1}^T \mathbb{E}\left[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right] \\
&\leq \frac{D^2}{2\eta} + \gamma\frac{K-1}{K}T + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right].
\end{aligned}$$

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

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq \frac{K^3X^2D^2}{2\gamma(K-1)} + \gamma\frac{K-1}{K}T + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right].$$

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

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq \frac{K^3X^2D^2}{K-1} + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right].$$

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

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y'_t \neq y_t]\right] \leq 2KXD\sqrt{\frac{T}{2}} + \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\boldsymbol{U})\right],$$

which completes the proof. □

6.9.3 Details of Section 6.5.3

Proof of Theorem 35. First, note that $p'_t(y_t) \geq \frac{\gamma}{K}$. 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 25. By using $\mathbb{E}_t\left[\mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}\right] = 1$ and $\mathbb{E}_t\left[\left(\mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}\right)^2\right] = \mathbb{1}[y'_t = y_t]p'_t(y'_t)^{-1}$ we can expand the surrogate gap:

$$\begin{aligned}
&\mathbb{E}\left[(1-a_t)\mathbb{1}[y^\star_t \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right] \\
&= \begin{cases}
\mathbb{E}\Big[2m^\star_t - m^{\star 2}_t + (1-m^\star_t)^2\frac{K-1}{K} & \\
\quad -(1-2m_t(\boldsymbol{W}_t, y_t)) + \frac{\eta}{p'_t(y_t)}4\|\boldsymbol{x}_t\|^2\Big] & \text{if } y^\star_t \neq y_t \text{ and } m^\star_t < 1 \\
\mathbb{E}\Big[(1-m^\star_t)^2\frac{K-1}{K} - (1-m^\star_t)^2 & \\
\quad +\frac{\eta}{p'_t(y_t)}4\|\boldsymbol{x}_t\|^2(1-m^\star_t)^2\Big] & \text{if } y^\star_t = y_t \text{ and } m^\star_t < 1 \\
\mathbb{E}\left[1-(1-2m_t(\boldsymbol{W}_t, y_t)) + \frac{\eta}{p'_t(y_t)}4\|\boldsymbol{x}_t\|^2\right] & \text{if } y^\star_t \neq y_t \text{ and } m^\star_t \geq 1 \\
0 & \text{if } y^\star_t = y_t \text{ and } m^\star_t \geq 1.
\end{cases}
\end{aligned} \tag{6.9.3}$$

We now split the analysis into the cases in (6.9.3). We start with the case where $y^\star_t \neq y_t$ and $m^\star_t < 1$. By using (6.8.3) we can see that for $\eta = \frac{\gamma}{4K^2X^2}$

$$\begin{aligned}
&\mathbb{E}\left[2m^\star_t - m^{\star 2}_t + (1-m^\star_t)^2\frac{K-1}{K} - (1-2m_t(\boldsymbol{W}_t, y_t)) + \frac{\eta}{p'_t(y_t)}4\|\boldsymbol{x}_t\|^2\right] \\
&= \mathbb{E}\left[2(m^\star_t + m_t(\boldsymbol{W}_t, y_t)) - m^{\star 2}_t + (1-m^\star_t)^2\frac{K-1}{K} - 1 + \frac{\eta}{p'_t(y_t)}4\|\boldsymbol{x}_t\|^2\right] \\
&\leq \mathbb{E}\left[-m^{\star 2}_t + (1-m^\star_t)^2\frac{K-1}{K} - 1 + \frac{\eta}{p'_t(y_t)}4X^2\right] \qquad \text{(by equation (6.8.3))} \\
&\leq -\frac{1}{K} + \frac{K\eta}{\gamma}4X^2 \leq 0.
\end{aligned}$$

The next case we consider is when $y_t^\star = y_t$ and $m_t^\star < 1$. In this case we have

$$\begin{aligned}
&\mathbb{E}\left[(1-m_t^\star)^2\frac{K-1}{K} - (1-m_t^\star)^2 + \frac{\eta}{p_t'(y_t)}4\|\boldsymbol{x}_t\|^2(1-m_t^\star)^2\right] \\
&= \mathbb{E}\left[-(1-m_t^\star)^2\frac{1}{K} + \frac{\eta}{p_t'(y_t)}4\|\boldsymbol{x}_t\|^2(1-m_t^\star)^2\right] \\
&= \mathbb{E}\left[-(1-m_t^\star)^2\frac{1}{K} + \frac{K\eta}{\gamma}4X^2(1-m_t^\star)^2\right],
\end{aligned}$$

which is bounded by 0 since $\eta = \frac{\gamma}{4K^2X^2}$.

Finally, if $y_t^\star \neq y_t$ and $m_t^\star \geq 1$ then

$$\begin{aligned}
&\mathbb{E}\left[1 - (1 - 2m_t(\boldsymbol{W}_t, y_t)) + \frac{\eta}{p_t'(y_t)}4\|\boldsymbol{x}_t\|^2\right] \\
&= \mathbb{E}\left[2m_t(\boldsymbol{W}_t, y_t) + \frac{\eta}{p_t'(y_t)}4\|\boldsymbol{x}_t\|^2\right] \\
&\leq \mathbb{E}\left[-2m_t^\star + \frac{\eta}{p_t'(y_t)}4\|\boldsymbol{x}_t\|^2\right] && \text{(by equation (6.8.3))} \\
&\leq -2 + \frac{K\eta}{\gamma}4X^2,
\end{aligned}$$

which is bounded by 0 since $\eta < \frac{\gamma}{2K^2X^2}$. We apply Lemma 25 and use the above to find:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \\
&\leq \frac{\|\boldsymbol{U}\|^2}{2\eta} + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right] + \gamma T \\
&\quad + \sum_{t=1}^T \mathbb{E}\left[(1-a_t)\mathbb{1}[y_t^\star \neq y_t] + a_t\frac{K-1}{K} - \ell_t(\boldsymbol{W}_t) + \frac{\eta}{2}\|\boldsymbol{g}_t\|^2\right] \\
&\leq \frac{D^2}{2\eta} + \gamma T + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right].
\end{aligned}$$

Plugging in $\eta = \frac{\gamma}{4K^2X^2}$ gives us:

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[y_t' \neq y_t]\right] \leq \frac{2K^2X^2D^2}{\gamma} + \gamma T + \mathbb{E}\left[\sum_{t=1}^T \ell_t(\boldsymbol{U})\right].$$

Algorithm 14 AdaHedge with abstention

```
Input: AdaHedge
for $t = 1 \dots T$ do
    Obtain expert predictions $\boldsymbol{y}_t = (y_t^1, \dots, y_t^d)^\intercal$
    Obtain expert distribution $\hat{\boldsymbol{p}}_t$ from AdaHedge
    Set $\hat{y}_t = \langle \hat{\boldsymbol{p}}_t, \boldsymbol{y}_t \rangle$
    Let $y_t^\star = \mathrm{sign}(\hat{y}_t)$
    Set $b_t = 1 - |\hat{y}_t|$
    Predict $y_t' = y_t^\star$ with probability $1 - b_t$ and predict $y_t' = *$ with probability $b_t$
    Obtain $\ell_t$ and send $\ell_t$ to AdaHedge
end for
```

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

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y_t' \neq y_t]\right] \leq 4K^2X^2D^2 + \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\boldsymbol{U})\right].$$

In the case where $1 > \sqrt{\frac{2K^2X^2D^2}{T}}$ we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[y_t' \neq y_t]\right] \leq 2DKX\sqrt{2T} + \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\boldsymbol{U})\right],$$

which completes the proof.

□

6.10 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 $i = 1, \dots, d$

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.

The algorithm we use can be found in Algorithm 14. A parallel result to Lemma 25 can be found in Lemma 27, which we will use to derive the regret of Algorithm 14.

Lemma 27. *For any expert i, the expected loss of Algorithm 14 satisfies:*

$$\begin{aligned}
&\sum_{t=1}^{T} ((1 - b_t)\ell_t(y_t^\star) + b_t c_t) \\
&\leq \sum_{t=1}^{T} \ell_t(y_t^i) + \inf_{\eta > 0} \left\{ \frac{\ln(d)}{\eta} + \sum_{t=1}^{T} \underbrace{((1 - b_t)\ell_t(y_t^\star) + c_t b_t + \eta v_t - \ell_t(\hat{y}_t))}_{\text{Abstention gap}} \right\} \\
&\quad + \frac{4}{3}\ln(d) + 2,
\end{aligned}$$

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

Before we prove Lemma 27 let us compare Algorithm 14 with GAPTRON. The updates of weight matrix $\boldsymbol{W}_t$ in GAPTRON are performed with OGD. In Algorithm 14 the updates or $\hat{p}_t$ are performed using ADAHEDGE. The roles of a_t in GAPTRON and b_t in Algorithm 14 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 27. First, ADAHEDGE guarantees that

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

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

$$\begin{aligned}
&\sum_{t=1}^{T} ((1-b_t)\ell_t(y_t^\star) + b_t c_t) \\
&= \sum_{t=1}^{T} \big((1-b_t)\ell_t(y_t^\star) + b_t c_t + \ell_t(y_t^i) - \ell_t(\hat{y}_t)\big) + \sum_{t=1}^{T} \big(\ell_t(\hat{y}_t) - \ell_t(y_t^i)\big) \\
&\leq \sum_{t=1}^{T} \big((1-b_t)\ell_t(y_t^\star) + b_t c_t + \ell_t(y_t^i) - \ell_t(\hat{y}_t)\big) + 2\sqrt{\ln(d)\sum_{t=1}^{T} v_t} \\
&\quad + 4/3\ln(d) + 2 \\
&= \sum_{t=1}^{T} \ell_t(y_t^i) + \inf_{\eta>0}\left\{\frac{\ln(d)}{\eta} + \sum_{t=1}^{T} ((1-b_t)\ell_t(y_t^\star) + b_t c_t + \eta v_t - \ell_t(\hat{y}_t))\right\} \\
&\quad + 4/3\ln(d) + 2.
\end{aligned}$$

□

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 η or b_t such that the abstention gap is smaller than 0. The result for abstention can be found in Theorem 36 below.

Theorem 36. *Suppose* $\max_t c_t < \frac{1}{2}$ *for all* T. *Then Algorithm 14 guarantees*

$$\begin{aligned}
&\sum_{t=1}^{T} ((1-b_t)\ell_t(y_t^\star) + b_t c_t) \\
&\leq \sum_{t=1}^{T} \ell_t(y_t^i) + \min\left\{\frac{\ln(d)}{1-2\max_t c_t}, 2\sqrt{\ln(d)\sum_{t=1}^{T} v_t}\right\} + 4/3\ln(d) + 2.
\end{aligned}$$

Proof. We start by upper bounding the v_t term. We have

$$v_t = \frac{1}{4}\mathbb{E}_{\hat{\boldsymbol{p}}_t}\left[(y_t^i - \hat{y}_t)^2\right] \leq \frac{1}{4}(1-\hat{y}_t)(\hat{y}_t+1) \leq \tfrac{1}{2}(1-|\hat{y}_t|),$$

where the first inequality is the Bhatia-Davis inequality (Bhatia and Davis, 2000).

Chapter 6

As with the proofs of GAPTRON we split the abstention gap in cases:

$$\begin{aligned}
&(1-b_t)\ell_t(y_t^\star) + c_t b_t + \eta v_t - \ell_t(\hat{y}_t)\\
&\leq (1-b_t)\ell_t(y_t^\star) + c_t b_t + \eta\tfrac{1}{2}(1-|\hat{y}_t|) - \ell_t(\hat{y}_t)\\
&= \begin{cases} c_t(1-|\hat{y}_t|) + \eta\frac{1}{2}(1-|\hat{y}_t|) - \frac{1}{2}(1-|\hat{y}_t|) & \text{if } y_t^\star = y_t\\ |\hat{y}_t| + c_t(1-|\hat{y}_t|) + \eta\frac{1}{2}(1-|\hat{y}_t|) - \frac{1}{2}(1+|\hat{y}_t|) & \text{if } y_t^\star \neq y_t. \end{cases}
\end{aligned} \tag{6.10.1}$$

Note that regardless of the the true label $(1-b_t)\ell_t(y_t^\star) + c_t b_t - \ell_t(\hat{y}_t) \leq 0$ since $c_t < \frac{1}{2}$. Hence, by using Lemma 27, we can see that as long as $c_t < \frac{1}{2}$

$$\sum_{t=1}^T ((1-b_t)\ell_t(y_t^\star) + b_t c_t) \leq \sum_{t=1}^T \ell_t(y_t^i) + 2\sqrt{\ln(d)\sum_{t=1}^T v_t} + 4/3\ln(d) + 2.$$

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

$$\begin{aligned}
&|\hat{y}_t| + c_t(1-|\hat{y}_t|) + \eta\tfrac{1}{2}(1-|\hat{y}_t|) - \tfrac{1}{2}(1+|\hat{y}_t|)\\
&= c_t(1-|\hat{y}_t|) + \eta\tfrac{1}{2}(1-|\hat{y}_t|) - \tfrac{1}{2}(1-|\hat{y}_t|).
\end{aligned}$$

So as long as $\eta \leq 1 - 2c_t$ the abstention gap is bounded by 0. Applying Lemma 27 now gives us

$$\begin{aligned}
&\sum_{t=1}^T \big((1-b_t)\ell_t(y_t^\star) + b_t c_t - \ell_t(y_t^i)\big)\\
&\leq \inf_{\eta>0}\left\{\frac{\ln(d)}{\eta} + \sum_{t=1}^T ((1-b_t)\ell_t(y_t^\star) + c_t b_t + \eta v_t - \ell_t(\hat{y}_t))\right\} + 4/3\ln(d) + 2\\
&\leq \frac{\ln(d)}{1 - 2\max_t c_t} + 4/3\ln(d) + 2,
\end{aligned}$$

which completes the proof. □

With a slight modification of the proof of Theorem 36 one can also show a similar result as Theorem 8 by Neu and Zhivotovskiy (2020), albeit with slightly worse constants. We leave this as an exercise for the reader.