

Efficient Methods for Online Multiclass Logistic Regression

Naman Agarwal

Satyen Kale

Julian Zimmert

Google Research

NAMANAGARWAL@GOOGLE.COM

SATYENKALE@GOOGLE.COM

ZIMMERT@GOOGLE.COM

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

Multiclass logistic regression is a fundamental task in machine learning with applications in classification and boosting. Previous work (Foster et al., 2018) has highlighted the importance of improper predictors for achieving “fast rates” in the online multiclass logistic regression problem without suffering exponentially from secondary problem parameters, such as the norm of the predictors in the comparison class. While Foster et al. (2018) introduced a statistically optimal algorithm, it is in practice computationally intractable due to its run-time complexity being a large polynomial in the time horizon and dimension of input feature vectors. In this paper, we develop a new algorithm, FOLKLORE, for the problem which runs significantly faster than the algorithm of Foster et al. (2018) – the running time per iteration scales quadratically in the dimension – at the cost of a linear dependence on the norm of the predictors in the regret bound. This yields the first practical algorithm for online multiclass logistic regression, resolving an open problem of Foster et al. (2018). Furthermore, we show that our algorithm can be applied to online bandit multiclass prediction and online multiclass boosting, yielding more practical algorithms for both problems compared to the ones in (Foster et al., 2018) with similar performance guarantees. Finally, we also provide an online-to-batch conversion result for our algorithm.

Keywords: multiclass logistic regression, online learning

1. Introduction

Logistic regression is a classical model in statistics used for estimating conditional probabilities that dates back to (Berkson, 1944). The model has been extensively studied in statistical and online learning and has been widely used in practice both for binary classification and multi-class classification in a variety of applications.

In recent years, motivated by applications in clickthrough-rate prediction in large scale advertisement systems, the problem of *online* logistic regression has been intensively studied. This is a sequential decision-making problem where examples consist of a feature vector and a categorical class label. In each time step an online learner is provided a feature vector representing the example to be classified, and the learner is required to output a prediction of probabilities¹ of the label of the example. The learner’s performance is measured by their *regret*, which is the difference between their cumulative log-loss, and the log-loss of the best possible linear predictor of bounded norm for the online sequence of examples computed in hindsight.

This problem is an instance of online convex optimization and hence the standard tools for deriving low regret algorithms, viz. Online Gradient Descent (Zinkevich, 2003) (OGD) and Online Newton Step (Hazan et al., 2007) (ONS) apply. However, the regret bounds derived using these

1. To be precise, the learner outputs logits which induce a probability distribution via softmax.

methods are suboptimal. To discuss these bounds, let d be the dimension of the input feature vectors (assumed to have length at most 1), $B > 0$ be a bound on (an appropriate) norm of the comparator linear predictor, and T be the number of prediction rounds. Then OGD has a regret bound of $O(B\sqrt{T})$, whereas ONS has a regret bound of $O(de^B \log(T))$. While the latter bound has much better dependence on T than the former, the exponential dependence on the B parameter makes it worse for most practical settings of interest.

In COLT 2012, [McMahan and Streeter \(2012\)](#) posed the open problem of whether it is possible to develop an algorithm with regret scaling polynomially in B but logarithmically in T . This question was answered in the *negative* by [Hazan et al. \(2014\)](#) who proved a lower bound showing that no *proper* online learning algorithm (i.e. one that generates a linear predictor to use in each round *before* observing the feature vector in that round) can achieve these desiderata. [Foster et al. \(2018\)](#) broke this lower bound via an *improper* online learning algorithm based on Vovk’s Aggregating Algorithm ([Vovk, 1998](#)) which enjoys a regret bound of $O(d \log(BT))$, thus achieving a *doubly-exponential* improvement in the dependence on B compared to ONS. A similar result was also obtained by [Kakade and Ng \(2005\)](#) for the binary classification case. [Foster et al. \(2018\)](#) also proved a lower bound showing that this regret bound is *optimal*, even for improper predictors. Despite the optimality of the regret bound, this algorithm is unfortunately not very practical since its running time is a polynomial in d , B , and T with high degrees. [Foster et al. \(2018\)](#) recognized the impracticality of their method, and state that “obtaining a truly practical algorithm with a modest polynomial dependence on the dimension is a significant open problem.” [Jézéquel et al. \(2020\)](#) solved this open problem for the *binary* classification setting, and developed a new improper online learning algorithm called AIOLI that has a suboptimal regret bound of $O(dB \log(BT))$, but much better running time: each round can be implemented to run in $O(d^2 + \log(T))$ time. Thus, AIOLI is a practical algorithm for online binary logistic regression. Similarly, for the offline (batch statistical, i.i.d.) setting for binary logistic regression, [Marteau-Ferey et al. \(2019\)](#) and [Mourtada and Gaiffas \(2019\)](#) have developed learning algorithms with fast rates. The former develops a proper learning algorithm via generalized self-concordance techniques along with additional assumptions on the data distribution, whereas the latter develops an improper learning algorithm based on empirical risk minimization with an improper regularizer.

With the exception of ([Foster et al., 2018](#)), none of the papers mentioned above giving algorithms with fast rates and non-exponential dependence on B extend to the practically important online *multiclass* logistic regression setting. In fact, it is possible to formally show that the style of analysis for the AIOLI algorithm ([Jézéquel et al., 2020](#)) using Hessian dominance does not work for even 3 classes, see Section F. [Jézéquel et al. \(2020\)](#) also note the difficulty of extending the analysis to the multiclass setting and explicitly list developing a *practical* algorithm for the multiclass setting, and applying it to problems such as online bandit multiclass prediction and online multiclass boosting, as open problems.

In this paper, we solve these open problems. Specifically, if K denotes the number of classes, we devise a new algorithm, FOLKLORE (Fast OnLine K-class LOgistic REgression), that has a regret bound of $O(d(B + \log(K))K \log(T))$, and runs in $\tilde{O}(d^2 K^3 + BK^2 \log(T))$ time per round. Similarly to AIOLI, FOLKLORE is also based on Follow-The-Regularized-Leader (FTRL) paradigm (see ([Hazan, 2019](#)) for background on FTRL) with quadratic surrogate losses and an improper regularizer which, in part, penalizes all the K possible labels equally. The main distinction from AIOLI is the introduction of a *linear* term in the regularizer which is key to the analysis. The linear term is designed to ensure that a certain notion of instantaneous regret in any given time step is minimized.

Algorithm	Regret	Running time (per step)
OGD	$B\sqrt{T}$	dK
ONS	$de^B \log(T)$	$d^2 K^2$
Foster et al. (2018)	$d \log(T)$	$\text{poly}(d, K, B, T)$
AIOLI (Jézéquel et al., 2020) [K=2]	$dB \log(T)$	$d^2 + B \log(T)$
GAF (Jézéquel et al., 2021)	$dBK \log(T)$	$d^2 K^3 + K^2 T^3$
FOLKLORE (This paper)	$dBK \log(T)$	$d^2 K^3 + K^2 B \log(T)$

Table 1: Regret bounds and running time (in $\tilde{O}(\cdot)$) for relevant algorithms for online logistic regression. The running time bound for Foster et al. (2018) depends on the sampling algorithm and is a high-degree polynomial in d, K, B, T .

This crucially relies on the fact that there are only K possible labels, and therefore, only K possible values of the instantaneous regret. We choose the regularizer to be the solution to a minimax problem: minimize the maximum of the K possible values of the instantaneous regret. It turns out that this minimax problem has a closed-form solution, which is used to design FOLKLORE. Since the minimax problem relies on the knowledge of the feature vector in the current time step, FOLKLORE becomes an *improper* online learning algorithm. An intriguing fact is that although FOLKLORE does *not* reduce to AIOLI in the binary setting, the *probabilities* predicted by FOLKLORE in the binary case exactly match those predicted by AIOLI (with some minor adjustments) (see Section D).

We then employ FOLKLORE as a subroutine in two problems of interest, also previously considered by Foster et al. (2018): online bandit multiclass prediction, and online multiclass boosting. For the online bandit multiclass problem, we give a new reduction which transforms a regret bound in the full-information setting to the bandit setting. Using FOLKLORE as the full-information online multiclass prediction algorithm, the resulting algorithm has a regret bound of $O(\sqrt{dK^2(B + \log(K))\log(BT)\bar{T}})$, which is roughly an $O(\sqrt{B})$ factor worse than the regret bound of the OBAMA algorithm of Foster et al. (2018), but is significantly more practical. For the online multiclass boosting problem, we give a new reduction to *binary* logistic regression. Using FOLKLORE (or AIOLI) as the online prediction algorithm results in a sample complexity of $O(T \log(K)/(N\gamma^2) + \log^2(T)/(\gamma^2) + KS/\gamma)$. Again, this is a factor $\log(T)$ worse than the AdaBoost.OML++ algorithm of Foster et al. (2018), but significantly more practical.

After this manuscript initially appeared on arXiv, we were made aware of the independent and concurrent work (Jézéquel et al., 2021) which provides an efficient low-regret algorithm called Gaussian Aggregating Forecaster (GAF) for 1-mixable losses satisfying certain assumptions. In particular for multiclass logistic regression, GAF has the same regret bound of $O(dBK \log(T))$ as FOLKLORE, but a per-step running time of $O(d^2 K^3 + K^2 T^3)$. GAF is based on the same lower-bound approximation as our algorithm, however it still requires sampling which leads to a T -dependent running time complexity per step. Our algorithm on the other hand is purely based on the FTRL paradigm and enjoys a much better running time. This and other relevant results for multiclass logistic regression are summarized in Table 1.

2. Problem setting and notation

Let $\|\cdot\|_p$ denote the ℓ_p norm on \mathbb{R}^d and $\|\cdot\|_F$ as the Frobenius norm of a matrix. Further let $\|\cdot\|_{2,\infty}$ denote the $2 \rightarrow \infty$ norm of a matrix defined as

$$\|A\|_{2,\infty} \triangleq \sup_{x: \|x\|_2 \leq 1} \|Ax\|_\infty$$

for any $K \in \mathbb{N}$, let $[K]$ be the set $[1, \dots, K]$. Define $\Delta_K \in \mathbb{R}^K$ be the standard K -dimensional simplex representing distributions over $[K]$. For any positive definite matrix A , the norm $\|\cdot\|_A$ on \mathbb{R}^d is given by $\|x\|_A = \sqrt{\langle x, Ax \rangle}$. We will denote by \otimes the standard Kronecker product between matrices. When using vectors $u \in \mathbb{R}^{j_1}, v \in \mathbb{R}^{j_2}, u \otimes v \in \mathbb{R}^{j_1 j_2}$ is defined as the Kronecker product obtained by interpreting u, v as $j_1 \times 1, j_2 \times 1$ dimensional matrices respectively.

In the multiclass learning problem the input is assumed to be from the set $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_2 \leq R\}$ for some $R \geq 0$. The number of output labels is K . A linear predictor parameterized by $W \in \mathbb{R}^{K \times d}$, given an input $x \in \mathcal{X}$, assigns a score to every class as given by the vector $Wx \in \mathbb{R}^K$. Given a $K \times d$ sized matrix W , we will denote by $\vec{W} \in \mathbb{R}^{Kd}$ the Kd -dimensional vectorization of W in a canonical manner. Let W_k represent the k^{th} row of the matrix W . The set of linear predictors we will be dealing with will be assumed to be $\mathcal{W} = \{W \in \mathbb{R}^{d \times K} | \|W\|_{2,\infty} \leq B\}$ for some $B \geq 0$. Under these restrictions it is readily observed that $\forall x \in \mathcal{X}$ and $\forall W \in \mathcal{W}$, we have that $\|Wx\|_\infty \leq BR$.

Given a vector $z \in \mathbb{R}^K$ define the softmax function $\sigma(z) : \mathbb{R}^K \rightarrow \Delta_K$ as

$$[\sigma(z)]_k \triangleq \frac{e^{[z]_k}}{\sum_{j \in [K]} e^{[z]_j}} \quad \forall k \in [K].$$

With the above notation we can define the multiclass logistic loss $\ell(\cdot, \cdot)$ function given a vector $z \in \mathbb{R}^K$ and a distribution $y \in \Delta_K$, as

$$\ell(z, y) \triangleq \sum_{k=1}^K -y_k \log([\sigma(z)]_k).$$

In several contexts the label y will denote a class in $[K]$ rather a distribution in Δ_K . In these cases, we will overload notation and use $\ell(z, y)$ to denote the logistic loss: $-\log([\sigma(z)]_y)$.

Online multiclass logistic regression. We define the learning problem as follows. The learner makes predictions over T rounds with individual rounds indexed by t . In each round t , nature provides an input $x_t \in \mathcal{X}$ to the learner and the learner predicts $z_t \in \mathbb{R}^K$ in response. Nature then provides a label $y_t \in \Delta_K$ and the learner suffers the loss $\ell(z_t, y_t)$. The aim of the learner is to minimize regret with respect to any $W \in \mathcal{W}$, defined as

$$\text{Regret}(W) \triangleq \sum_{t=1}^T \ell(z_t, y_t) - \ell(Wx_t, y_t).$$

For notational convenience, we define

$$\ell_t(W) \triangleq \ell(Wx_t, y_t).$$

An online learning algorithm for the problem is said to be *proper* if it generates a linear predict $W_t \in \mathcal{W}$ before seeing x_t , and then setting $z_t = W_t x_t$. An *improper* predictor is allowed to output any prediction z_t taking into account the observation x_t . In particular, the learner we propose will produce a W_t at every step, and predict using $z_t = W_t x_t$, but W_t will depend on the observed x_t , and the learner is thus improper.

Finally, we make the convention that the gradient $\nabla \ell_t(W) \in \mathbb{R}^{Kd}$ and the Hessian $\nabla^2 \ell_t(W) \in \mathbb{R}^{Kd \times Kd}$ are defined with respect to the canonical vectorization of the the matrix W .

3. Main result

The algorithm we propose FOLKLORE (Algorithm 1), is an improper FTRL over a sequence of quadratic surrogate losses. At any time step, the algorithm decides on an improper regularizer ϕ_t which adjusts the prediction matrix depending on x_t . The regularizer ϕ_t is defined as follows:

$$\phi_t(W) \triangleq \frac{1}{K} \sum_{k=1}^K \ell(W x_t, k) + \langle \vec{W}, B_t \rangle. \quad (3.1)$$

The vector $B_t \in \mathbb{R}^{Kd}$ defines a linear term in the regularizer and is key to the analysis. It will be specified in the course of the analysis. We use the following surrogate loss:

$$\hat{\ell}_t(W) \triangleq \ell_t(W_t) + \langle \vec{W} - \vec{W}_t, \nabla \ell_t(W_t) \rangle + \frac{1}{BR + \ln(K)/2} \|\vec{W} - \vec{W}_t\|_{\nabla^2 \ell_t(W_t)}^2 \quad (3.2)$$

By definition $\hat{\ell}_t(W_t) = \ell_t(W_t)$. The following lemma shows that the surrogate loss provides a lower bound over $W \in \mathcal{W}$.

Lemma 1 *For any x_t and W such that $\|W x_t\|_\infty \leq BR$, the surrogate function defined in equation (3.2) is a lower bound on ℓ_t , i.e.*

$$\hat{\ell}_t(W) \leq \ell_t(W).$$

The above lemma is the multiclass analogue of Lemma 5 in [Jézéquel et al. \(2020\)](#) (although Lemma 5 is not stated in the Hessian form above, it can be readily seen to be the same in the appropriate parametrization). The proof is provided in Section B.

Remark 2 *Given the quadratic form of the surrogate loss, it is natural to suspect that the above lemma is a by-product of a self-concordance like property of logistic regression ([Bach, 2010](#); [Tran-Dinh et al., 2015](#)). Indeed a similar lower bound can be obtained via generalized self-concordance, however it requires both $\|W_t\|_{2,\infty}$ and $\|W\|_{2,\infty}$ to be bounded. Instead we provide an alternative full proof to a slightly stronger version requiring boundedness of only W .*

The following theorem provides a regret bound for FOLKLORE. We prove this theorem in the next section.

Theorem 3 *The regret of Algorithm 1 with surrogate losses defined by (3.2), improper regularization defined by (3.1) and $\lambda = 2 \frac{R}{B}$ against any sequence of contexts is bounded by*

$$\text{Regret} = \mathcal{O}((BR + \ln(K))dK \ln(T)).$$

Algorithm 1 FOLKLORE

input: regularization λ , improper regularization ϕ_t

Set $A_0 = \lambda \cdot \mathbf{I}_{Kd}$, $G_0 = 0$.

for $t = 1, \dots, T$ **do**

 Receive x_t and play $z_t = W_t x_t$ where

$$W_t = \underset{W \in \mathbb{R}^{d \times k}}{\operatorname{argmin}} \lambda \|W\|_F^2 + \sum_{s=1}^{t-1} \hat{\ell}_s(W) + \phi_t(W) = \underset{W \in \mathbb{R}^{d \times k}}{\operatorname{argmin}} \|\vec{W}\|_{A_{t-1}}^2 + \langle \vec{W}, G_{t-1} \rangle + \phi_t(W). \quad (3.3)$$

 Receive y_t and suffer loss ℓ_t .

 Update $A_t = A_{t-1} + \frac{1}{BR + \ln(K)/2} \nabla^2 \ell_t(W_t)$.

 Update $G_t = G_{t-1} + \nabla \ell_t(W_t)$.

The next theorem specifies the per-step running time of the algorithm. The proof is delegated to the Appendix (Section E).

Theorem 4 *For every step t of Algorithm 1, for any $\epsilon > 0$, we can compute a vector \hat{z}_t such that $\|\hat{z}_t - W_t x_t\|^2 \leq \epsilon$ in total time*

$$O \left(d^2 K^3 + K^2 \left(1 + \frac{R^2}{\lambda} \right) \log \left(\epsilon^{-1} \left(1 + \frac{R^2}{\lambda} \right) \right) \right).$$

Compared to [Foster et al. \(2018\)](#), our algorithm is significantly faster, but the improved running time comes at the price of a linear dependence on BR instead of logarithmic.

Remark 5 *Note that, FOLKLORE is stated with W_t being the true minimizer in (3.3). However it can be seen from our analysis that it is sufficient to solve the problem upto additive error $\text{poly}(1/T)$. Since Theorem 4 shows a logarithmic dependence on this error for running time, exact minimization is assumed for brevity through the rest of the paper.*

3.1. Analysis

Our analysis proceeds by first recalling the decomposition of the regret into a sum of *instantaneous regret* terms obtained by [Jézéquel et al. \(2020\)](#). While this fact was proved by [Jézéquel et al. \(2020\)](#), for completeness we provide a concise proof in Appendix (Section B).

Lemma 6 *Let $\hat{\ell}_t$ be the surrogate quadratic lower bound of the loss function which ensure $\hat{\ell}_t(W_t) = \ell_t(W_t)$ and $\hat{\ell}_t(W) \leq \ell_t(W)$ for all $W \in \mathcal{W}$. Then for any $W \in \mathcal{W}$, the regret of Algorithm 1 for any sequence of regularizers $(\phi_t)_{t=1}^T$ is bounded by*

$$\text{Regret}(W) \leq \lambda \|W\|_F^2 + \frac{1}{4} \sum_{t=1}^T \left[\|\nabla \phi_t(W_t) - \nabla \ell_t(W_t)\|_{A_t^{-1}}^2 - \|\nabla \phi_t(W_t)\|_{A_{t-1}^{-1}}^2 \right].$$

Given this decomposition, the goal of selecting a suitable ϕ_t is to minimize the *instantaneous regret* defined as

$$\|\nabla \phi_t(W_t) - \nabla \ell_t(W_t)\|_{A_t^{-1}}^2 - \|\nabla \phi_t(W_t)\|_{A_{t-1}^{-1}}^2.$$

For the case of $K = 2$, Jézéquel et al. (2020) choose the regularizer $\phi_t(W)$ as the function $\ell(Wx_t, 0) + \ell(Wx_t, 1)$. A simple calculation shows that a similar approach with the natural extension $\sum_k \ell(Wx_t, k)$ does not quite work (see also Section F for a more general discussion). We propose to take an alternative approach. Since we do not know the loss ℓ_t ahead of time, we aim at finding the ϕ_t that minimizes the worst case instantaneous regret. For arbitrary values of $\nabla \ell_t(W_t)$, this minimax problem does not have a finite solution because the instantaneous regret is quadratic in $\nabla \ell_t(W_t)$. The key insight for logistic regression is that the loss function is a convex combination over the finite set $\{\ell(\cdot; x_t, \mathbf{e}_k) \mid k \in [K]\}$. We will focus over these K “base” loss functions first. For these K functions, there are only K possible values of $\nabla \ell_t(W_t)$. This allows to recast the quadratic problem in $\nabla \ell_t(W_t)$ as a linear problem, for which a minimax solution exists. The following lemma summarizes this idea of linearization:

Lemma 7 *For any W , any positive definite matrix $A \succ 0$, any regularization function ϕ_t , and any $\ell_t(W) \in \{\ell(Wx_t, y) \mid y \in \Delta_K\}$, the instantaneous regret decomposes as*

$$\begin{aligned} \|\nabla \phi_t(W) - \nabla \ell_t(W)\|_{A^{-1}}^2 - \|\nabla \phi_t(W)\|_{A^{-1}}^2 \\ \leq -2\langle A^{-1} \nabla \ell_t(W), \nabla \phi_t(W) - b_t(W, A) \rangle + \text{Tr}(A^{-1} \nabla^2 \ell_t(W)), \end{aligned}$$

where

$$b_t(W; A) = \sigma(Wx_t) \otimes x_t - \frac{1}{2} A \text{diag}_{\otimes}(A^{-1})(\mathbf{1}_K \otimes x_t),$$

and diag_{\otimes} denotes the operator that sets all matrix entries besides those corresponding to the K -many $d \times d$ blocks on the diagonal to 0.

The proof is delegated to the Appendix (Section B). Given the above lemma, the natural minimax solution for ϕ_t is to ensure for all W

$$\nabla \phi_t(W) - b_t(W; A) = 0, \quad (3.4)$$

in which case the instantaneous regret reduces to $\text{Tr}(A^{-1} \nabla^2 \ell_t(W_t))$. Further note that the LHS in Lemma 7 upper bounds the instantaneous regret with both choices $A = A_{t-1}$ and $A = A_t$. In terms of regret, the better choice is selecting $A = A_t$ due to the telescoping sum. Unfortunately, A_t depends on W_t which makes finding the right ϕ_t computationally expensive². Therefore we continue by choosing $A = A_{t-1}$. The following lemma (proved in the Appendix (Section B)) shows how to choose ϕ_t to satisfy (3.4).

Lemma 8 *Setting*

$$B_t = \frac{1}{K} \mathbf{1}_K \otimes x_t - \frac{1}{2} A_{t-1} \text{diag}_{\otimes}(A_{t-1}^{-1})(\mathbf{1}_K \otimes x_t),$$

in the definition of ϕ_t (3.1) ensures that for all W we have that $\nabla \phi_t(W) = b_t(W; A_{t-1})$.

Combining the ideas above we are now ready to prove the main regret bound provided in Theorem 3.

2. This approach requires solving a fixed point equation.

Proof of Theorem 3. Combining Lemma 6,7,8 and the fact that for all $W \in \mathcal{W}$, we have that $\|W\|_F^2 \leq KB^2$ directly leads to the following conclusion that for all $W \in \mathcal{W}$,

$$\text{Regret}(W) \leq \lambda KB^2 + \sum_{t=1}^T \text{Tr}(A_{t-1}^{-1} \nabla^2 \ell_t(W_t)).$$

We have by definition

$$A_t = \lambda \mathbf{I} + \frac{1}{BR + \ln(K)/2} \sum_{s=1}^t \nabla^2 \ell_s(W_s).$$

For convenience of analysis we define a slightly different series

$$\tilde{A}_t = \frac{\lambda}{2} \mathbf{I} + \frac{1}{BR + \ln(K)/2} \sum_{s=1}^t \nabla^2 \ell_s(W_s).$$

In Lemma 14 in the appendix, we prove that for any t

$$\nabla^2 \ell_t(W_t) \preceq R^2 \mathbf{I},$$

and therefore by using $\lambda = \frac{2R}{B}$, we have

$$A_{t-1} \succeq A_t - \frac{\lambda}{2} \mathbf{I} = \tilde{A}_t.$$

Using the concavity of the log det function, which in particular implies for any two positive definite matrices A, B , $\text{Tr}(A^{-1}(A - B)) \leq \log \det(A) - \log \det(B)$, we get that

$$\begin{aligned} \text{Regret} &\leq \lambda KB^2 + \sum_{t=1}^T \text{Tr}(A_{t-1}^{-1} \nabla^2 \ell_t(W_t)) \\ &\leq \lambda KB^2 + \sum_{t=1}^T \text{Tr}(\tilde{A}_t^{-1} \nabla^2 \ell_t(W_t)) \\ &= \lambda KB^2 + (BR + \ln(K)/2) \sum_{t=1}^T \text{Tr} \left(\tilde{A}_t^{-1} \frac{\nabla^2 \ell_t(W_t)}{BR + \ln(K)/2} \right) \\ &\leq \lambda KB^2 + (BR + \ln(K)/2) \sum_{t=1}^T \left(\log \det(\tilde{A}_t) - \log \det(\tilde{A}_{t-1}) \right) \\ &\leq \lambda KB^2 + (BR + \ln(K)/2) \left(\log \det(\tilde{A}_T) - \log \det(\tilde{A}_0) \right) \\ &= \lambda KB^2 + (BR + \ln(K)/2) \log \det \left(I + \frac{2}{\lambda(BR + \ln(K)/2)} \sum_{t=1}^T \nabla^2 \ell_t(W) \right) \\ &\leq \lambda KB^2 + (BR + \ln(K)/2) \log \det((1+T)\mathbf{I}) \\ &\leq K(2BR + (BR + \ln(K)/2)(d \ln(1+T))) \end{aligned}$$

■

4. Applications

In this section, we show that all applications of logistic loss with linear function classes listed in (Foster et al., 2018) are solvable with our algorithm.

4.1. Bandit Multiclass Learning

The next application of our techniques is the bandit multiclass problem. This problem, first studied by Kakade et al. (2008), considers the protocol of online multiclass learning in Section 2 with nature choosing $y_t \in [K]$ in each round, but with the added twist of bandit feedback: in each round, the learner predicts a class \hat{y}_t sampled from some probability distribution p_t that it computes, and receives feedback on whether the prediction was correct or not, i.e. $\mathbb{I}[\hat{y}_t \neq y_t]$. The goal is to minimize regret with respect to a reference class of linear predictors, using some appropriate surrogate loss function for the 0-1 loss. In most recent papers, regret bounds for this problem have been replaced by the weaker notion of relative mistake bounds (i.e. a bound on the number of mistakes discounted by the minimal logistic loss of a linear predictor).

In COLT 2009, Abernethy and Rakhlin (2009) posed the open problem of obtaining an *efficient* algorithm for the problem with $O(\sqrt{T})$ regret using the multiclass logistic loss as the surrogate loss. Hazan and Kale (2011) solved the open problem and obtained an algorithm, Newtron, based on the Online Newton Step algorithm (Hazan et al., 2007) with $\tilde{O}(\sqrt{T})$ regret for the case when norm of the linear predictors scales at most logarithmically in T . This result was improved by Foster et al. (2018) who developed an algorithm called OBAMA based on the Aggregating Algorithm that obtains $\tilde{O}(\sqrt{T})$ relative mistake bound across all parameter ranges, but has poor running time. In a beautiful paper, van der Hoeven (2020) managed to get a $O(\sqrt{T})$ relative mistake bound for this problem with excellent running time – $O(dK)$ per iteration – by exploiting the gap between the 0-1 loss and the multiclass logistic loss. Relative mistake bounds were also obtained for other surrogate losses: e.g. multiclass hinge loss by Kakade et al. (2008), and for a special family of quadratic loss functions by Beygelzimer et al. (2017).

We now give a generic reduction showing that any regret bound for the (full-information) online logistic regression implies a relative mistake bound for bandit multiclass learning. By using FOLKLORE as the base algorithm for this reduction, we obtain another bandit multiclass prediction algorithm with $\tilde{O}(\sqrt{T})$ relative mistake bound with a practical running time. Algorithm 2 gives the reduction. It assumes that it is given an algorithm for online logistic regression, Alg , with the interface $\text{Alg}.\text{Initialize}()$ to initialize the algorithm, $\text{Alg}.\text{Update}(x, y)$ to update its internal state with the example (x, y) , and $\text{Alg}.\text{Predict}(x)$ to predict logits for class probabilities given an input x with the current internal state. The reduction is very simple and based on an exploration-exploitation split: in each round the algorithm samples a Bernoulli random variable with parameter γ (denoted $\text{Ber}(\gamma)$), and if the outcome is 1, it chooses a class uniformly at random, and if the predicted class is correct, updates Alg . Otherwise, the algorithm predicts using the current internal state of Alg . The following theorem (proved in the Appendix (Section B) provides the regret bound for this reduction.

Algorithm 2 Bandit multiclass reduction

input: logistic regression algorithm $\text{Alg}, \gamma \in [0, 1]$
 $\text{Alg}.\text{Initialize}()$
for $t = 1, \dots, T$ **do**
if $Ber(\gamma) = 1$ **then**

 Predict $\hat{y}_t \sim \text{unif}_K$.

if $\hat{y}_t = y_t$ **then**
 $\text{Alg}.\text{Update}(x_t, y_t)$
else

 Predict $\hat{y}_t \sim \sigma(\text{Alg}.\text{Predict}(x_t))$

Theorem 9 For an online multiclass logistic regression algorithm which satisfies for all loss sequences ℓ_t and comparators $W \in \mathcal{W}$:

$$\sum_{t=1}^T \ell_t(W_t) - \ell_t(W) \leq \mathcal{R}(T),$$

where $\mathcal{R}(T)$ is a monotonous upper bound function, Algorithm 2 with input $\gamma = \sqrt{\frac{K\mathcal{R}(T)}{T}}$ satisfies

$$\mathbb{E} \left[\sum_{t=1}^T \mathbb{I}\{\hat{y}_t \neq y_t\} \right] \leq \min \left\{ \min_{W \in \mathcal{W}} \sum_{t=1}^T \ell(W x_t, y_t) + \sqrt{K T \mathcal{R}(T)}, T \right\}.$$

Applying Theorem 9 to the regression algorithm of [Foster et al. \(2018\)](#) directly recovers their Theorem 6, but our reduction holds in more generality. Combining the wrapper with Algorithm 1 directly leads to a relative mistake bound of

$$\mathbb{E} \left[\sum_{t=1}^T \hat{y}_t \neq y_t \right] \leq \min_{W \in \mathcal{W}} \sum_{t=1}^T \ell_t(W) + \mathcal{O} \left(\min \{ T, \sqrt{d K^2 (B R + \ln(K)) T \ln(B R T)} \} \right).$$

We can now compare this result with that of [van der Hoeven \(2020\)](#). Their algorithm gets a relative mistake bound of $O(K B R \sqrt{T})$ with $O(dK)$ running time. Our algorithm gets a better dependence on $B R$, but comes at a cost of $O(\sqrt{d \log(T)})$ factor, and somewhat worse running time.

4.2. Online Multiclass Boosting

The next application of our techniques is the problem of online multiclass boosting (OMB). We begin by describing first the basic setup of OMB, following closely the notation of [Jung et al. \(2017\)](#). In this problem, over a sequence of T rounds indexed by t , the learner receives an instance $x_t \in \mathcal{X}$, then selects a class $\hat{y}_t \in [K]$, and finally observes the true class $y_t \in [K]$. The goal is to minimize the total number of mistakes $\sum_{t=1}^T \mathbb{I}[\hat{y}_t \neq y_t]$. Assisting the algorithm in this task are N copies of a *weak learner* which abide by the following protocol. At any time step $t = 1, \dots, T$, the weak learners receive a feature vector x_t and a cost-matrix $C_t \in \mathcal{C}$ (from some class of matrices $C \in \mathbb{R}^{K \times K}$), and returns a prediction $l_t \in [K]$. Then the true label y_t is revealed to the learner and the weak learner suffers the loss $C_t(y_t, l_t)$. Following [Jung et al. \(2017\)](#), we restrict $\mathcal{C} = \{C \in$

$\mathbb{R}_+^{K \times K} \mid \forall y \in [K] : C(y, y) = 0 \text{ and } \|C(y, \cdot)\|_1 \leq 1\}$. Weak learners are characterized by the following property.

Definition 10 (Weak learning condition (Jung et al., 2017)) *An environment and a learner outputting $(l_t)_{t=1}^T$ satisfy the multiclass weak learning condition with edge γ and sample complexity S if for all outcomes $(y_t)_{t=1}^T$ and cost matrices $(C_t)_{t=1}^T$ from the set \mathcal{C} adaptively chosen by the environment, we have*

$$\sum_{t=1}^T C_t(y_t, l_t) \leq \sum_{t=1}^T \mathbb{E}_{k \sim u_{\gamma, y_t}} [C_t(y_t, k)] + S,$$

where $u_{\gamma, y}$ is the distribution that samples uniform over $[K]$ with probability $(1 - \gamma)$ and otherwise picks the label y .

Given weak-learners as described above, the aim of multiclass boosting (using multiple instantiation of the weak learner) is to achieve arbitrarily low error/regret (measured by the number of mistakes made by the algorithm), keeping the number of samples received and the number of weak learners instantiated by the algorithm as low as possible. Foster et al. (2018) propose AdaBoost.OLM++ (Algorithm 4 in the Appendix), an extension of AdaBoost.OLM (Jung et al., 2017) and AdaBoost.OL (Beygelzimer et al., 2015) using their proposed algorithm for multiclass logistic regression as a subroutine. Overall they exhibit an algorithm that achieves a target error rate ϵ , after seeing $T = \tilde{\Omega}\left(\frac{1}{\epsilon\gamma^2} + \frac{KS}{\epsilon\gamma}\right)$ samples, leveraging at most $N = \tilde{O}\left(\frac{1}{\epsilon\gamma^2}\right)$ -many weak learners with edge γ . However since Foster et al. (2018) instantiate a logistic regression instance for every weak learner, their algorithm suffers a very high computational overhead per-step per weak-learner due to the use of the Aggregating Algorithm. In this section we demonstrate a more efficient reduction for the online multiclass boosting problem which reduces the problem a *binary* logistic regression problem. As a result we can leverage our proposal FOLKLORE or AIOLI achieving the same (up to $\log(T)$ factors) bounds as Foster et al. (2018) but significantly faster running time.

We now briefly revisit AdaBoost.OLM++ (detailed description and proofs found in Appendix C). The high level idea is to let each weak learner $i \in [N]$ iteratively improve the *aggregated* logits predicted. In time step t , let s_t^{i-1} denote the aggregation up to weak learner $i - 1$, then the cost matrix of the i -th weak learner is determined by the gradient of the logistic loss of s_t^{i-1} and its prediction l_t shifts the aggregation by $s_t^i = s_t^{i-1} + \alpha_t^i \mathbf{e}_{l_t}$. Foster et al. (2018) show that one obtains small overall error, as long as the adaptive aggregations $s_t^i = s_t^{i-1} + \alpha_t^i \mathbf{e}_{l_t}$ have small regret compared to the best fixed α^i in hindsight.

We formalize this intuition in the following.

Definition 11 (Boosting regression problem) *The problem proceeds as a T -round problem where at every step t , the environment (potentially adversarially and adaptively) picks and reveals to the algorithm, a logit $s_t \in \mathbb{R}^K$ and a label $l_t \in [K]$. The algorithm predicts a vector $\hat{s}_t \in \mathbb{R}^K$. Then a true label picked by the environment $y_t \in [K]$ is revealed to the algorithm. The aim of the algorithm is to ensure the following type of regret bound:*

$$\sum_{t=1}^T -\log([\sigma(\hat{s}_t)]_{y_t}) \leq \min_{\alpha \in [-2, 2]} \sum_{t=1}^T -\log([\sigma(s_t + \alpha \mathbf{e}_{l_t})]_{y_t}) + \mathcal{R}(T),$$

where $\mathcal{R}(T)$ is a regret function.

With the above definition, we can generalize the guarantee for the AdaBoost.OLM++ algorithm (Foster et al., 2018) (stated as Algorithm 4 in the Appendix (Section C)) as in the following proposition (proof provided in the Appendix (Section C)).

Proposition 12 (Variant of Theorem 8 and Proposition 9 (Foster et al., 2018)) *Given an algorithm solving the boosting regression problem with regret $\mathcal{R}(T)$, the predictions $(\hat{y}_t)_{t=1}^T$ generated by Algorithm 4 satisfy with probability at least $1 - \delta$,*

$$\sum_{t=1}^T \mathbb{I}\{\hat{y}_t \neq y_t\} = \mathcal{O}\left(\frac{T \log(K) + N\mathcal{R}(T)}{\sum_{i=1}^N \gamma_i^2} + \log(N/\delta)\right).$$

Suppose further that all weak learners satisfy the weak learning condition with edge γ and sample complexity S , then with probability at least $1 - \delta$, we have

$$\sum_{t=1}^T \mathbb{I}\{\hat{y}_t \neq y_t\} = \mathcal{O}\left(\frac{T \log(K)}{N\gamma^2} + \frac{\mathcal{R}(T)}{\gamma^2} + \frac{KS}{\gamma} + \log(N/\delta)\right).$$

Foster et al. (2018) provide a solution to the boosting regression problem with regret $\mathcal{R}(T) = \tilde{\mathcal{O}}(1)$ by recasting it as a regular regression problem, mapping α to $W_\alpha \in \mathbb{R}^{K \times 2K}$ and expanding the dimension to $d = 2K$. Unlike FOLKLORE, the regret bound of multiclass logistic regression provided by Foster et al. (2018) scales with the algebraic dimension of the comparator class \mathcal{W} , which is only 1 instead of $2K^2$. We show in the next paragraph a significantly more efficient reduction based on FOLKLORE.

Solving the boosting regression. We present a suitable reduction of the boosting regression to *binary* regression, which therefore does not scale with the number of classes K . Algorithm 3 first collapses the dimension of the prediction from K classes to a binary decision representing a prediction of whether the predicted label l_t is correct. Then it clips the magnitude of the logits to $\log(T)$ to ensure bounded R and runs FOLKLORE³ over the resulting binary regression problem. Finally it expands the binary result to K classes by predicting all classes $y \neq l_t$ proportional to their input logits $s_{t,y}$.

Algorithm 3 Boosting regression with FOLKLORE subroutine

Input: FOLKLORE with $K = 2, d = 2, R = 1 + \ln(T), B = 2$.

for $t = 1, \dots, T$ **do**

Receive s_t, l_t
 $\tilde{s}_t \leftarrow (s_{t,l_t}, \log \sum_{k \neq l_t} \exp(s_{t,k}))$
 $\tilde{x}_t = (\min\{\log(T), \max\{-\log(T), \frac{1}{2}(\tilde{s}_{t,1} - \tilde{s}_{t,2})\}, -1)$.
 $\zeta_t \leftarrow \text{FOLKLORE.Predict}(\tilde{x}_t)$
 $\hat{s}_{t,l_t} \leftarrow \zeta_{t,1}$
 $\forall k \in [K] \setminus \{l_t\} : \hat{s}_{t,k} \leftarrow s_{t,k} + \zeta_{t,2} - \log \sum_{k' \neq l_t} \exp(s_{t,k'})$.
Play s_t and receive true class y_t .
 $\text{FOLKLORE.Update}(\tilde{x}_t, 1 + \mathbb{I}\{y_t \neq l_t\})$

The following theorem provides a bound on the regret of Algorithm 3. We provide the proof in the Appendix (Section C).

3. AIOLI would work equally well, since we only require binary regression.

Theorem 13 *Algorithm 3* solves the boosting regression problem with $\mathcal{R}(T) = \mathcal{O}(\log^2(T))$.

Combining Theorem 13 with Proposition 12 directly leads to an efficient algorithm for online multiclass boosting. The running time complexity during each time-step of algorithm 3 is only $\mathcal{O}(K)$. Hence, in contrast to the work of Foster et al. (2018), our regression routine does not dominate the overall running time complexity of AdaBoost.OLM++, where computing the cost matrices C_i^t alone is at least linear in K .

4.3. Online-to-Batch Conversion

The FOLKLORE algorithm can be used to obtain a predictor for the batch statistical setting via the standard online-to-batch conversion. Given a sample set of size T drawn i.i.d. from the underlying distribution, this predictor runs FOLKLORE on the sample (with an arbitrary order) and stops at a random time $\tau \in \{1, 2, \dots, T\}$. Then, given a new input feature vector $x \in \mathcal{X}$, the predicted class probabilities are computed by solving the optimization problem (3.3) in FOLKLORE at time step τ using the procedure from Theorem 4. Standard online-to-batch conversion analysis (Cesa-Bianchi et al., 2004) then implies that the expected excess risk of this predictor over the optimal linear prediction W with $\|W\|_{2,\infty} \leq B$ is bounded by the average regret, i.e. by $\mathcal{O}\left(\frac{(BR + \ln(K))dK \ln(T)}{T}\right)$.

5. Conclusion and Future work

In this paper we gave an efficient online multiclass regression algorithm which enjoys logarithmic regret with running time scaling only quadratically in the dimension of the feature vectors, thus answering an open question of Foster et al. (2018). We also showed how to apply this algorithm to the online bandit multiclass prediction and online multiclass boosting problems via new reductions.

One open question remaining in this line of work is the following. The regret bound in (Foster et al., 2018) scales with the algebraic dimension of \mathcal{W} , while ours scales with the worst-case dimension Kd . Our main technique, specifically Lemma 7 breaks when one tries to reduce the dependency to the algebraic dimension. Designing an algorithm with similar performance guarantees to ours but with regret scaling with the algebraic dimension of \mathcal{W} would be quite interesting.

References

Jacob D. Abernethy and Alexander Rakhlin. An Efficient Bandit Algorithm for \sqrt{T} Regret in Online Multiclass Prediction? In *COLT*, 2009.

Francis Bach. Self-concordant analysis for logistic regression. *Electron. J. Statist.*, 4:384–414, 2010. doi: 10.1214/09-EJS521. URL <https://doi.org/10.1214/09-EJS521>.

Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39:357—365, 1944.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

Alina Beygelzimer, Satyen Kale, and Haipeng Luo. Optimal and adaptive algorithms for online boosting. In *International Conference on Machine Learning*, pages 2323–2331. PMLR, 2015.

Alina Beygelzimer, Francesco Orabona, and Chicheng Zhang. Efficient Online Bandit Multiclass Learning with $\tilde{O}(\sqrt{T})$ Regret. In *ICML*, pages 488–497, 2017.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inf. Theory*, 50(9):2050–2057, 2004. doi: 10.1109/TIT.2004.833339. URL <https://doi.org/10.1109/TIT.2004.833339>.

Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference On Learning Theory*, pages 167–208. PMLR, 2018.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

E. Hazan and S. Kale. Newtron: an efficient bandit algorithm for online multiclass prediction. In *Advances in Neural Information Processing Systems*, pages 891–899, 2011.

Elad Hazan. Introduction to online convex optimization. *CoRR*, abs/1909.05207, 2019. URL <http://arxiv.org/abs/1909.05207>.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 197–209, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL <https://proceedings.mlr.press/v35/hazan14a.html>.

Young Hun Jung, Jack Goetz, and Ambuj Tewari. Online multiclass boosting. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 920–929, 2017.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2085–2108. PMLR, 09–12 Jul 2020. URL <http://proceedings.mlr.press/v125/jezequel20a.html>.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Mixability made efficient: Fast online multi-class logistic regression. *HAL preprint: hal-03370530*, 2021.

Sham M Kakade and Andrew Y Ng. Online bounds for bayesian algorithms. In *Advances in neural information processing systems*, pages 641–648, 2005.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pages 440–447. ACM, 2008.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis R. Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In Alina Beygelzimer and Daniel Hsu, editors, *COLT*, volume 99 of *Proceedings of Machine Learning Research*, pages 2294–2340. PMLR, 2019.

H Brendan McMahan and Matthew Streeter. Open problem: Better bounds for online logistic regression. In *Conference on Learning Theory*, pages 44–1. JMLR Workshop and Conference Proceedings, 2012.

Jaouad Mourtada and Stéphane Gaiffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *CoRR*, abs/1912.10784, 2019. URL <http://arxiv.org/abs/1912.10784>.

Quoc Tran-Dinh, Yen-Huan Li, and Volkan Cevher. Composite convex minimization involving self-concordant-like cost functions. In *Modelling, Computation and Optimization in Information Systems and Management Sciences*, pages 155–168. Springer, 2015.

Dirk van der Hoeven. Exploiting the surrogate gap in online multiclass classification. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *NeurIPS*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/6ce8d8f3b038f737cefcdafcf3752452-Abstract.html>.

Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.

Appendix A. Preliminaries

We begin by providing some simple calculations regarding the gradient and Hessian of the multinomial logistic loss functions. For any x, y we have that

$$\begin{aligned}\nabla_W \ell(Wx, y) &= (\sigma(Wx) - y) \otimes x \\ \nabla_W^2 \ell(Wx, y) &= (\text{diag}(\sigma(Wx)) - \sigma(Wx)\sigma(Wx)^\top) \otimes (xx^\top).\end{aligned}$$

Note that the Hessian is independent of y . This fact will be useful in the analysis. The following lemma provides a simple bound on the Hessian.

Lemma 14 *For any x, y , we have that the hessian satisfies*

$$\nabla_W^2 \ell(Wx, y) \preceq \|x\|_2^2 \cdot \mathbf{I}_{Kd}.$$

Proof We have

$$\nabla_W^2 \ell(Wx, y) = (\text{diag}(\sigma(Wx)) - \sigma(Wx)\sigma(Wx)^\top) \otimes (xx^\top).$$

We have that $xx^\top \preceq \|x\|_2^2 \mathbf{I}_d$. Further, $\sigma(Wx)$ is a probability distribution and hence $(\text{diag}(\sigma(Wx)) - \sigma(Wx)\sigma(Wx)^\top) \preceq \mathbf{I}_K$. Combining these two inequalities completes the proof. \blacksquare

Appendix B. Omitted proofs

Proof of Lemma 6 For any W , define $\hat{L}_t(W) \triangleq \sum_{s=1}^t \hat{\ell}_s(W) + \lambda \|W\|_F^2$ with $L_0(W) \triangleq \lambda \|W\|_F^2$. Further define $\hat{W}_t = \operatorname{argmin}_{W \in \mathbb{R}^{k \times d}} \hat{L}_t(W)$. We now have the following consequence for any W ,

$$\begin{aligned} \text{Regret}(W) &= \sum_{t=1}^T [\ell_t(W_t) - \ell_t(W)] \leq \sum_{t=1}^T [\hat{\ell}_t(W_t) - \hat{\ell}_t(W)] \\ &= \sum_{t=1}^T [\hat{\ell}_t(W_t) + \hat{L}_{t-1}(\hat{W}_{t-1}) - \hat{L}_t(\hat{W}_t)] + \underbrace{\hat{L}_T(\hat{W}_T) - \hat{L}_T(W)}_{\leq 0 \text{ (since } \hat{W}_T \text{ minimizes } \hat{L}_T\text{)}} - \lambda(\|\hat{W}_1\|_F^2 - \|W\|_F^2) \\ &\leq \lambda \|W\|_F^2 + \sum_{t=1}^T [\hat{L}_t(W_t) - \hat{L}_t(\hat{W}_t) + \hat{L}_{t-1}(\hat{W}_{t-1}) - \hat{L}_{t-1}(W_t)]. \end{aligned}$$

The first inequality above follows from the fact that $\hat{\ell}_t(W_t) = \ell_t(W_t)$ and for all W , $\hat{\ell}_t(W_t) \leq \ell_t(W_t)$. The last inequality follows from the definition of \hat{L}_t .

Note that \hat{L}_t is a quadratic function with hessian $2A_t$. For any quadratic function f with invertible hessian A , it holds that

$$f(x) - \min_{x'} f(x') = \frac{1}{2} \|\nabla f(x)\|_{A^{-1}}^2.$$

Hence

$$\hat{L}_t(W_t) - \hat{L}_t(\hat{W}_t) + \hat{L}_{t-1}(\hat{W}_{t-1}) - \hat{L}_{t-1}(W_t) = \frac{1}{4} \left(\|\nabla \hat{L}_t(W_t)\|_{A_t^{-1}}^2 - \|\nabla \hat{L}_{t-1}(W_t)\|_{A_{t-1}^{-1}}^2 \right).$$

By definition of our strategy of selecting W_t in Algorithm 1, (since the optimization is unconstrained and is for a function which is strongly convex and continuously differentiable), we have that

$$\nabla \hat{L}_{t-1}(W_t) + \nabla \phi_t(W_t) = 0,$$

which directly gives $\nabla \hat{L}_t(W_t) = \nabla \ell_t(W_t) - \nabla \phi_t(W_t)$ and $\nabla \hat{L}_{t-1}(W_t) = -\nabla \phi_t(W_t)$. Plugging these into the equation above and combining everything completes the proof. \blacksquare

Proof of Lemma 1 Given W_t and a target W with the guarantee that $\|W x_t\|_\infty \leq BR$, define the following vectors

$$\omega = \sigma(W x_t) \text{ and } \nu = \sigma(W_t x_t).$$

Further denote $\ln(w) = [\ln(w_i)]_{i=1}^K$ as the element-wise logarithm of a vector w and define $\gamma \triangleq \frac{1}{RB + \ln(K)/2}$. The lower bound is defined as

$$\hat{\ell}_t(W) = \ell_t(W_t) + \langle \vec{W} - \vec{W}_t, \nabla \ell_t(W_t) \rangle + \gamma \|\vec{W} - \vec{W}_t\|_{\nabla^2 \ell_t(W_t)}^2,$$

which implies the following,

$$\begin{aligned} \hat{\ell}_t(W) - \ell_t(W) &= -\langle \ln(\nu), y_t \rangle + \langle \ln(\omega), y_t \rangle + \langle (W - W_t)x_t, \nu - y_t \rangle + \gamma \|(W - W_t)x_t\|_{\operatorname{diag}(\nu) - \nu \nu^\top}^2 \\ &= -\langle \ln(\nu), y_t \rangle + \langle \ln(\omega), y_t \rangle + \langle \ln(\omega) - \ln(\nu), \nu - y_t \rangle + \gamma \|\ln(\omega) - \ln(\nu)\|_{\operatorname{diag}(\nu) - \nu \nu^\top}^2 \\ &= \langle \ln(\omega) - \ln(\nu), \nu \rangle + \gamma \|\ln(\omega) - \ln(\nu)\|_{\operatorname{diag}(\nu) - \nu \nu^\top}^2, \end{aligned}$$

The second equality above follows using the facts that $\langle \nu - y_t, \mathbf{1}_K \rangle = 0$ and $(\text{diag}(\nu) - \nu\nu^\top)\mathbf{1}_K = 0$. Denote $\Omega = \{\sigma(Wx) \mid W \in \mathcal{W}, x \in \mathcal{X}\} \subset \Delta_K$ be the set of all possible distributions induced by $W \in \mathcal{W}$, then we have

$$\hat{\ell}_t(W) - \ell_t(W) \leq \max_{\omega \in \Omega} \max_{\nu \in \Delta_K} \underbrace{\langle \ln(\omega) - \ln(\nu), \nu \rangle + \gamma \|\ln(\omega) - \ln(\nu)\|_{\text{diag}(\nu) - \nu\nu^\top}^2}_{\triangleq F(\nu, \omega)}.$$

To analyse the above we first fix a $\omega \in \Omega$ and analyse the expression. Let ν^* be the optimum fixing ω . Note that since for all $W \in \mathcal{W}, X \in \mathcal{X}$ we have $\|Wx\|_\infty \leq BR$, we conclude that for any $i \in [K]$, we have $\omega_i \in (0, 1)$. Further, by the KKT conditions, at the optimal point $\nu^* \in \Delta_K$, there must exist $\lambda \in \mathbb{R}$ such that for all $i \in [K]$, one of the following must hold,

$$\begin{aligned} \frac{\partial}{\partial \nu_i} F(\nu^*, \omega) &= \lambda \\ \text{or } \frac{\partial}{\partial \nu_i} F(\nu^*, \omega) &< \lambda \text{ and } \nu_i^* = 0 \\ \text{or } \frac{\partial}{\partial \nu_i} F(\nu^*, \omega) &> \lambda \text{ and } \nu_i^* = 1. \end{aligned}$$

We can derive a closed form expression for the derivative given by

$$\frac{\partial}{\partial \nu_i} F(\nu^*, \omega) = \left(\ln\left(\frac{\omega_i}{\nu_i^*}\right) - 1 \right) (1 - 2\gamma \langle \ln(\omega) - \ln(\nu^*), \nu^* \rangle) + \gamma \ln^2\left(\frac{\omega_i}{\nu_i^*}\right) + 2\gamma \ln\left(\frac{\omega_i}{\nu_i^*}\right)$$

We observe that for any $i \in [k]$,

$$\lim_{\nu_i^* \rightarrow 0} \frac{\partial}{\partial \nu_i} F(\nu^*, \omega) = +\infty.$$

Therefore by the KKT conditions, there cannot be an $i \in [K]$, such $\nu_i^* = 0$. Since $\nu^* \in \Delta_K$, this also implies that there cannot be an $i \in [K]$, such $\nu_i^* = 1$. Therefore by the KKT conditions we must have that there exists a $\lambda \in \mathbb{R}$ such that for all i ,

$$\frac{\partial}{\partial \nu_i} F(\nu^*, \omega) = \lambda$$

The above condition can be re-written as for all $i \in [K]$,

$$\left(\ln\left(\frac{\omega_i}{\nu_i^*}\right) - 1 \right) (1 - 2\gamma \text{KL}(\omega, \nu^*)) + \gamma \ln^2\left(\frac{\omega_i}{\nu_i^*}\right) + 2\gamma \ln\left(\frac{\omega_i}{\nu_i^*}\right) = \lambda$$

Further observe that the LHS above across all i , is a fixed quadratic function in $\ln(\frac{\omega_i}{\nu_i^*})$ (Note that the KL term is fixed across constants). Therefore to satisfy the above equation, it must be the case that across all i , $\ln(\frac{\omega_i}{\nu_i^*})$ can acquire at most two distinct values. Formally, there must exist two values $\alpha_1, \alpha_2 \in \mathbb{R}$ and a subset $J \subseteq [K]$ such that the following holds,

$$\begin{aligned} \forall i \in J : \frac{\omega_i}{\nu_i^*} &= \alpha_1 \\ \forall i \in [K] \setminus J : \frac{\omega_i}{\nu_i^*} &= \alpha_2. \end{aligned}$$

Using the above define the following quantities, let

$$w \triangleq \sum_{i \in J} \omega_i, \quad v \triangleq \sum_{i \in J} \nu_i^*,$$

then we have that

$$F(\nu^*, \omega) = f(v, w) \triangleq v \ln\left(\frac{w}{v}\right) + (1-v) \ln\left(\frac{1-w}{1-v}\right) - \gamma \left(\left(v \ln\left(\frac{w}{v}\right) + (1-v) \ln\left(\frac{1-w}{1-v}\right) \right)^2 + v \ln^2\left(\frac{w}{v}\right) + (1-v) \ln^2\left(\frac{1-w}{1-v}\right) \right).$$

At this point we have effectively reduced the computation to the one-dimensional case. Note that $v \in [0, 1]$. Further since $\omega \in \Omega$, we have that

$$\ln(w) \geq \ln\left(\frac{\exp(-BR)}{(K-1)\exp(BR) + \exp(-BR)}\right) > \ln\left(\frac{\exp(-2BR)}{K}\right) = -2BR - \ln(K).$$

Therefore we can reduce the problem as follows

$$\max_{\omega \in \Omega} \max_{\nu \in \Delta_K} F(\nu, \omega) \leq \sup_{w \in (\frac{e^{-2BR}}{K}, 1)} \sup_{v \in [0, 1]} f(v, w) \quad (\text{B.1})$$

Next we show that

$$\sup_{w \in (\frac{e^{-2BR}}{K}, 1)} \sup_{v \in [0, 1]} f(v, w) \leq 0 \quad (\text{B.2})$$

which completes the proof. To characterize the sup of $f(v, w)$ over v for a given w , we consider two cases.

Case 1: Firstly, let $v \in \{0, 1\}$. In this case, $f(v, w)$ is $\ln(w) \leq 0$ and $\ln(1-w) \leq 0$ respectively.

Case 2: The next case is when $\frac{\partial}{\partial v} f(v, w) = 0$ for some $v \in (0, 1)$. The derivative is given by

$$\frac{\partial}{\partial v} f(v, w) = \left(\ln\left(\frac{w}{1-w}\right) - \ln\left(\frac{v}{1-v}\right) \right) \left(1 + \gamma(2 - (v - \frac{1}{2})(\ln\left(\frac{w}{1-w}\right) - \ln\left(\frac{v}{1-v}\right))) \right).$$

We will show that $\frac{\partial}{\partial v} f(v, w) = 0$ if and only if $v = w$. The forward direction is immediate. To see the backward direction, firstly, observe that for all $v \in (0, 1)$

$$(v - \frac{1}{2}) \ln\left(\frac{v}{1-v}\right) \geq 0, .$$

Therefore we have that for any $w \in (\frac{e^{-2BR}}{K}, 1)$ and any $v \in (0, 1)$,

$$\begin{aligned} \left(1 + \gamma(2 - (v - \frac{1}{2})(\ln\left(\frac{w}{1-w}\right) - \ln\left(\frac{v}{1-v}\right))) \right) &> 1 - \gamma(v - \frac{1}{2}) \ln\left(\frac{w}{1-w}\right) \\ &> 1 - \frac{1}{2RB + \ln(K)} \left| \ln\left(\frac{w}{1-w}\right) \right| \\ &> 1 + \frac{\min\{\ln(w), \ln(1-w)\}}{2RB + \ln(K)} \geq 0. \end{aligned}$$

Therefore we have that

$$\frac{\partial}{\partial v} f(v, w) = 0 \iff \left(\ln\left(\frac{w}{1-w}\right) - \ln\left(\frac{v}{1-v}\right) \right) = 0,$$

which happens iff and only if $v = w$. It can now be seen that when $v = w$, $f(v, w) = 0$, which establishes (B.2). Using (B.1), we have also established

$$\max_{\omega \in \Omega} \max_{\nu \in \Delta_K} F(\nu, \omega) \leq 0.$$

This completes the proof. ■

Proof of Lemma 7 To prove Lemma 7, as alluded to before, we first prove a restriction of the lemma for the case when $\ell_t \in \{\ell(Wx_t, \mathbf{e}_k) \mid k \in [K]\}$.

Lemma 15 *For any W , any p.d. matrix $A \succ 0$, any regularization function ϕ_t , and any $\ell_t(W) \in \{\ell(Wx_t, y) \mid y \in \Delta_K\}$, the instantaneous regret decomposes as*

$$\begin{aligned} \|\nabla \phi_t(W) - \nabla \ell_t(W)\|_{A^{-1}}^2 - \|\nabla \phi_t(W)\|_{A^{-1}}^2 \\ = -2\langle A^{-1} \nabla \ell_t(W), \nabla \phi_t(W) - b_t(W, A) \rangle + \text{Tr}(A^{-1} \nabla^2 \ell_t(W)), \end{aligned}$$

where

$$b_t(W; A) = \sigma(Wx_t) \otimes x_t - \frac{1}{2} A \text{diag}_{\otimes}(A^{-1})(\mathbf{1}_K \otimes x_t),$$

and diag_{\otimes} denotes the operator that sets all matrix entries besides those corresponding to the K -many $d \times d$ blocks on the diagonal to 0.

To use the above lemma first note that for any W

$$\nabla \ell_t(W) = \sum_{k \in [K]} y_k \nabla \ell(Wx_t, \mathbf{e}_k).$$

Therefore an application of Jensen's inequality implies that

$$\begin{aligned} \|\nabla \phi_t(W) - \nabla \ell_t(W)\|_{A^{-1}}^2 - \|\nabla \phi_t(W)\|_{A^{-1}}^2 \\ \leq \left(\sum_{k=1}^K ([y_t]_k \cdot \|\nabla \phi_t(W) - \nabla \ell(Wx_t, \mathbf{e}_k)\|_{A^{-1}}^2) \right) - \|\nabla \phi_t(W)\|_{A^{-1}}^2 \\ = \sum_{k=1}^K [y_t]_k \cdot (\|\nabla \phi_t(W) - \nabla \ell(Wx_t, \mathbf{e}_k)\|_{A^{-1}}^2 - \|\nabla \phi_t(W)\|_{A^{-1}}^2). \end{aligned}$$

Further using Lemma 15 we get that

$$\begin{aligned} \|\nabla \phi_t(W) - \nabla \ell_t(W)\|_{A^{-1}}^2 - \|\nabla \phi_t(W)\|_{A^{-1}}^2 \\ \leq \sum_{k=1}^K -2\langle A^{-1} \nabla \ell(Wx_t, \mathbf{e}_k), \nabla \phi_t(W) - b_t(W, A) \rangle + \text{Tr}(A^{-1} \nabla^2 \ell_t(W)) \\ \leq -2\langle A^{-1} \nabla \ell_t(W), \nabla \phi_t(W) - b_t(W, A) \rangle + \text{Tr}(A^{-1} \nabla^2 \ell_t(W)). \end{aligned}$$

Therefore all that is left to prove is Lemma 15 which we prove next

Proof of Lemma 15 Recall that for $\sigma_t = \sigma(Wx_t)$ and for a label y , the gradient and Hessian of the logistic loss are respectively

$$\begin{aligned}\nabla \ell(Wx_t, y) &= (\sigma_t - \mathbf{e}_y) \otimes x_t \\ \nabla^2 \ell(Wx_t, y) &= (\text{diag}(\sigma_t) - \sigma_t \sigma_t^\top) \otimes (x_t x_t^\top).\end{aligned}$$

Therefore we begin by analyzing the quadratic part of the term which is given by

$$\begin{aligned}\|\nabla \ell_t(W)\|_{A^{-1}}^2 &= \|(\sigma_t - \mathbf{e}_y) \otimes x_t\|_{A^{-1}}^2 \\ &= \text{Tr}((\text{diag}(\mathbf{e}_y) \otimes x_t x_t^\top) A^{-1}) + \langle A^{-1}(\sigma_t - 2\mathbf{e}_y) \otimes x_t, \sigma_t \otimes x_t \rangle \\ &= \text{Tr}((\text{diag}(\mathbf{e}_y - \sigma_t) \otimes x_t x_t^\top) A^{-1}) + 2\langle A^{-1}((\sigma_t - \mathbf{e}_y) \otimes x_t), \sigma_t \otimes x_t \rangle + \\ &\quad \underbrace{\text{Tr}(((\text{diag}(\sigma_t) - \sigma_t \sigma_t^\top) \otimes x_t x_t^\top) A^{-1})}_{=\text{Tr}(\nabla^2 \ell_t(W) A^{-1})}.\end{aligned}$$

Finally, note that

$$\begin{aligned}\text{Tr}((\text{diag}(\mathbf{e}_y - \sigma_t) \otimes x_t x_t^\top) A^{-1}) &= \text{Tr}((\text{diag}(\mathbf{e}_y - \sigma_t) \otimes x_t x_t^\top) \text{diag}_\otimes(A^{-1})) \\ &= \text{Tr}((\mathbf{1}_K \otimes x_t)((\mathbf{e}_y - \sigma_t) \otimes x_t)^\top \text{diag}_\otimes(A^{-1})) \\ &= \langle (\mathbf{e}_y - \sigma_t) \otimes x_t, \text{diag}_\otimes(A^{-1})(\mathbf{1}_K \otimes x_t) \rangle.\end{aligned}$$

Combining everything and replacing $(\sigma_t - \mathbf{e}_y) \otimes x_t = \nabla \ell_t(W)$ we get that,

$$\|\nabla \ell_t(W)\|_{A^{-1}}^2 = \langle A^{-1} \nabla \ell_t(W), -A \text{diag}_\otimes(A^{-1})(\mathbf{1}_K \otimes x_t) + 2\sigma_t \otimes x_t \rangle + \text{Tr}(\nabla^2 \ell_t(W) A^{-1})$$

Now noting the following expansion

$$\|\nabla \phi_t(W) - \nabla \ell_t(W)\|_{A^{-1}}^2 - \|\nabla \phi_t(W)\|_{A^{-1}}^2 = \|\nabla \ell_t(W)\|_{A^{-1}}^2 - 2\langle A^{-1} \nabla \ell_t(W), \nabla \phi_t(W) \rangle$$

and replacing the above finishes the proof. ■

Proof of Lemma 8 Recall that $\nabla \ell(Wx_t, \mathbf{e}_k) = (\sigma_t - \mathbf{e}_k) \otimes x_t$. Hence

$$\begin{aligned}\nabla \phi_t(W) &= (\sigma(Wx_t) - \frac{1}{K} \mathbf{1}_K) \otimes x_t + B_t = \sigma(Wx_t) \otimes x_t - \frac{1}{2} A_{t-1} \text{diag}_\otimes(A_{t-1}^{-1})(\mathbf{1}_K \otimes x_t) \\ &= b_t(W; A_{t-1}).\end{aligned}$$

Proof of Theorem 9 At any time t , with probability γ we sample a class uniformly at random. Otherwise we sample a class from σ_t of our algorithm. We update our algorithm only if the class was sampled by the extra exploration and if we selected the correct class. We have

$$\begin{aligned}
\mathbb{E} \left[\sum_{t=1}^T \mathbb{I}(\hat{y}_t \neq y_t) \right] &\leq \mathbb{E} \left[\sum_{t=1}^T (1 - \sigma_{t,y_t}) \right] + \gamma T \\
&\leq \mathbb{E} \left[\sum_{t=1}^T -\ln(\sigma_{t,y_t}) \right] + \gamma T \\
&\leq \frac{K}{\gamma} \mathbb{E} \left[\sum_{t=1}^T -\ln(\sigma_{t,y_t}) \mathbb{I}\{\hat{y}_t = y_t \wedge \text{explore}\} \right] + \gamma T \\
&\leq \frac{K}{\gamma} \mathbb{E} \left[\sum_{t=1}^T -\ln(\sigma_t(W)_{y_t}) \mathbb{I}\{\hat{y}_t = y_t \wedge \text{explore}\} + \text{Reg}_T \right] + \gamma T \\
&= -\sum_{t=1}^T \ln(\sigma_t(W)_{y_t}) + \frac{K}{\gamma} \text{Reg}_T + \gamma T
\end{aligned}$$

Plugging in the optimal choice of γ completes the proof. ■

Appendix C. Online Multiclass Boosting

AdaBoost.OLM++ (Algorithm 4) takes N weak learners $(\text{WL}^i)_{i=1}^N$, which are stateful objects that support the operations $\text{Predict}(x, C)$ and $\text{Update}(x, C, y)$ based on the weak learner protocol described in section 4.2. We use the index t to denote the number of updates of any stateful object. Additionally, we have a copy of an algorithm solving the boosting regression problem BoostReg for each weak learner with operations $\text{Predict}(s, l)$ and $\text{Update}(s, l, y)$. The algorithm maintains s_t^i , which are the weighted aggregated scores of the first i weak learners. They are iterately updated from s_t^{i-1} given the prediction l_t^i such that the regret $\sum_{t=1}^T (\log([\sigma(s_t^i)]_{y_t}) - \log([\sigma(s_t^{i-1} + \alpha \mathbf{e}_{l_t^i})]_{y_t}))$ is small with respect to some $\alpha \in [-2, 2]$. The predictions s_t^i induce a label $\hat{y}_t^i = \arg \max_{k \in [K]} (s_t^i)_k$, which are treated as expert recommendations. Over these expert predictions, AdaBoost.OLM++ runs the hedge algorithm (Freund and Schapire, 1997) to make its final decision. Finally, the cost matrices C_t^i are computed in the following way. Let

$$\widehat{C}_t^i(k, y) \triangleq \frac{\partial}{\partial z_k} \log([\sigma(z)]_y) \Big|_{z=s_t^{i-1}} = \sigma(s_t^{i-1})_k - \mathbb{I}\{k = y\},$$

then C_t^i is the translated and rescaled transformation of \widehat{C}_t^i that lies in \mathcal{C} :

$$C_t^i(k, y) = \frac{1}{K} \left(\widehat{C}_t^i(k, y) - \widehat{C}_t^i(y, y) \right). \quad (\text{C.1})$$

The following proof is included for completeness and follows up to minor modifications the work of Foster et al. (2018).

Algorithm 4 AdaBoost.OLM++ (Foster et al., 2018)

Input: weak learners WL^1, \dots, WL^N , boosting regression algorithm $BoostReg$
 For all $i \in [N]$, set $v_1^i \leftarrow 1$, initialize weak learner WL_1^i , and initialize copy of boosting regression algorithm $BoostReg_1^i$.

for $t = 1, \dots, n$ **do**

- Receive instance x_t .
- $s_t^0 \leftarrow 0 \in \mathbb{R}^K$.
- for** $i = 1, \dots, N$ **do**
- Compute cost matrix C_t^i from s_t^{i-1} using (C.1).
- $l_t^i \leftarrow WL_t^i.\text{Predict}(x_t, C_t^i)$.
- $s_t^i \leftarrow BoostReg_t^i.\text{Predict}(s_t^{i-1}, l_t^i)$.
- $\hat{y}_t^i \leftarrow \arg \max_k s_t^i(k)$.

 Sample i_t with $\Pr(i_t = i) \propto v_t^i$.

 Predict $\hat{y}_t = \hat{y}_t^{i_t}$ and receive true class $y_t \in [K]$.

for $i = 1, \dots, N$ **do**

- $WL_{t+1}^i \leftarrow WL_t^i.\text{Update}(x_t, C_t^i, y_t)$.
- $BoostReg_{t+1}^i \leftarrow BoostReg_t^i.\text{Update}(s_t^{i-1}, l_t^i, y_t)$.
- $v_{t+1}^i \leftarrow v_t^i \cdot \exp(-\mathbb{I}\{\hat{y}_t^i \neq y_t\})$.

Proof of Proposition 12 We begin with the first part of the proposition. Denote the number of mistakes of the i -th expert (which is the combination of the first i weak learners) by

$$M_i = \sum_{t=1}^T \mathbb{I}\{\hat{y}_t^i \neq y_t\} = \sum_{t=1}^T \mathbb{I}\{\arg \max_k s_t^i(k) \neq y_t\},$$

with the convention that $M_0 = T$. The weights v_t^i simply implement the multiplicative weights strategy, and so Lemma 17, which gives a concentration bound based on Freedman's inequality implies that with probability at least $1 - \delta$,⁴

$$\sum_{t=1}^T \mathbb{I}\{\hat{y}_t \neq y_t\} \leq 4 \min_i M_i + 2 \log(N/\delta). \quad (\text{C.2})$$

Note that if $k^* \triangleq \arg \max_k s_t^{i-1}(k) \neq y_t$, then $[\sigma(s_t^{i-1})]_{k^*} \geq [\sigma(s_t^{i-1})]_{y_t}$ and $\sigma(s_t^{i-1}) \in \Delta_K$ imply $[\sigma(s_t^{i-1})]_{y_t} \leq 1/2$, which then implies $\sum_{k \neq y_t} [\sigma(s_t^{i-1})]_k \geq 1/2$ and finally

$$-\sum_{t=1}^T \hat{C}_t^i(y_t, y_t) = \sum_{t=1}^T \sum_{k \neq y_t} [\sigma(s_t^{i-1})]_k \geq \frac{M_{i-1}}{2}. \quad (\text{C.3})$$

This also holds for $i = 1$ because $s_t^0 = 0$ and $-\hat{C}_t^1(y_t, y_t) = (K-1)/K \geq 1/2$.

4. Note that previous online boosting works (Beygelzimer et al., 2015; Jung et al., 2017) use a simpler Hoeffding bound at this stage, which picks up an extra \sqrt{T} term. For their results this is not a dominant term, but in our case it can spoil the improvement given by improper logistic regression, and so we use Freedman's inequality to remove it.

Define the difference between the total loss of the i -th and $(i-1)$ -th expert to be

$$\Delta_i = \sum_{t=1}^T \ell(s_t^i, y_t) - \ell(s_t^{i-1}, y_t).$$

By the assumption on the employed regression algorithm, we have

$$\Delta_i \leq \inf_{\alpha \in [-2, 2]} \left[\sum_{t=1}^T \ell(\alpha \mathbf{e}_{l_t^i} + s_t^{i-1}, y_t) - \ell(s_t^{i-1}, y_t) \right] + \mathcal{R}(T).$$

By [Lemma 18](#) each term in the sum above satisfies

$$\ell(\alpha \mathbf{e}_{l_t^i} + s_t^{i-1}, y_t) - \ell(s_t^{i-1}, y_t) \leq \begin{cases} (e^\alpha - 1)[\sigma(s_t^{i-1})]_{l_t^i} = (e^\alpha - 1)\widehat{C}_t^i(y_t, l_t^i), & l_t^i \neq y_t, \\ (e^{-\alpha} - 1)(1 - [\sigma(s_t^{i-1})]_{y_t}) = -(e^{-\alpha} - 1)\widehat{C}_t^i(y_t, y_t), & l_t^i = y_t. \end{cases}$$

With notation $w^i = -\sum_{t=1}^n \widehat{C}_t^i(y_t, y_t)$, $c_+^i = -\frac{1}{w^i} \sum_{t: l_t^i = y_t} \widehat{C}_t^i(y_t, y_t)$, and $c_-^i = \frac{1}{w^i} \sum_{t: l_t^i \neq y_t} \widehat{C}_t^i(y_t, l_t^i)$, we rewrite

$$\inf_{\alpha \in [-2, 2]} \left[\sum_{t=1}^n \ell(\alpha \mathbf{e}_{l_t^i} + s_t^{i-1}, y_t) - \ell(s_t^{i-1}, y_t) \right] = w^i \cdot \inf_{\alpha \in [-2, 2]} [(e^\alpha - 1)c_-^i + (e^{-\alpha} - 1)c_+^i].$$

One can verify that $w^i > 0$, $c_-^i, c_+^i \geq 0$, $c_+^i - c_-^i = \gamma_i \in [-1, 1]$ and $c_+^i + c_-^i \leq 1$. By [Lemma 19](#), it follows that

$$w^i \cdot \inf_{\alpha \in [-2, 2]} [(e^{-\alpha} - 1)c_-^i + (e^\alpha - 1)c_+^i] \leq -\frac{w^i \gamma_i^2}{2}.$$

Summing Δ_i over $i \in [N]$, we have

$$\sum_{t=1}^T \ell(s_t^N, y_t) - \sum_{t=1}^T \ell(s_t^0, y_t) = \sum_{i=1}^T \Delta_i \leq -\frac{1}{2} \sum_{i=1}^N w^i \gamma_i^2 + N\mathcal{R}(T). \quad (\text{C.4})$$

We lower bound the left hand side as

$$\sum_{t=1}^T \ell(s_t^N, y_t) - \sum_{t=1}^T \ell(s_t^0, y_t) \geq -\sum_{t=1}^T \ell(s_t^0, y_t) = -T \log(K),$$

where the inequality uses non-negativity of the logistic loss and the equality is a direct calculation from $s_t^0 = 0$. Next we upper bound the right-hand side of [\(C.4\)](#). Since $w^i = -\sum_{t=1}^T \widehat{C}_t^i(y_t, y_t)$, [Eq. \(C.3\)](#) implies

$$-\frac{1}{2} \sum_{i=1}^N w^i \gamma_i^2 \leq -\frac{1}{4} \sum_{i=1}^N M_{i-1} \gamma_i^2 \leq -\min_{i \in [N]} M_{i-1} \cdot \frac{1}{4} \sum_{i=1}^N \gamma_i^2 \leq -\min_{i \in [N]} M_i \cdot \frac{1}{4} \sum_{i=1}^N \gamma_i^2.$$

Combining our upper and lower bounds on $\sum_{i=1}^N \Delta_i$ now gives

$$-T \log(K) \leq -\frac{1}{2} \sum_{i=1}^N w^i \gamma_i^2 + N\mathcal{R}(T) \leq -\min_{i \in [N]} M_i \cdot \frac{1}{4} \sum_{i=1}^N \gamma_i^2 + N\mathcal{R}(T). \quad (\text{C.5})$$

Rearranging, we have

$$\min_{i \in [N]} M_i \leq O\left(\frac{T \log(K)}{\sum_{i=1}^N \gamma_i^2}\right) + O\left(\frac{N\mathcal{R}(T)}{\sum_{i=1}^N \gamma_i^2}\right).$$

Returning to (C.2), this implies that with probability at least $1 - \delta$,

$$\sum_{t=1}^T \mathbb{I}\{\hat{y}_t \neq y_t\} \leq O\left(\frac{T \log(K)}{\sum_{i=1}^N \gamma_i^2}\right) + O\left(\frac{N\mathcal{R}(T)}{\sum_{i=1}^N \gamma_i^2}\right) + 2 \log(N/\delta),$$

which finishes the first part of the proof.

By the definition of the cost matrices, the weak learning condition

$$\sum_{t=1}^T C_t^i(y_t, l_t^i) \leq \sum_{t=1}^T \mathbb{E}_{k \sim u_{\gamma, y_t}} [C_t^i(y_t, k)] + S$$

implies

$$\sum_{t=1}^T \widehat{C}_t^i(y_t, l_t^i) \leq \sum_{t=1}^T \mathbb{E}_{k \sim u_{\gamma, y_t}} [\widehat{C}_t^i(y_t, k)] + KS$$

Expanding the definitions of u_{γ, y_t} and \widehat{C}_t^i , we have

$$\mathbb{E}_{k \sim u_{\gamma, y_t}} [\widehat{C}_t^i(y_t, k)] = \left(\frac{1-\gamma}{K}\right) \left([\sigma(s_t^{i-1})]_{y_t} - 1 + \sum_{k \neq y_t} [\sigma(s_t^{i-1})]_k \right) + \gamma([\sigma(s_t^{i-1})]_{y_t} - 1) = \gamma \widehat{C}_t^i(y_t, y_t).$$

So we have

$$\sum_{t=1}^T \widehat{C}_t^i(y_t, l_t^i) \leq \gamma \sum_{t=1}^T \widehat{C}_t^i(y_t, y_t) + KS,$$

or, since $\widehat{C}_t^i(y_t, y_t) < 0$,

$$\gamma_i \geq \gamma - \frac{KS}{w^i},$$

where $w^i = -\sum_{t=1}^n \widehat{C}_t^i(y_t, y_t)$ as in the first part. Since $a \geq b - c$ implies $a^2 \geq b^2 - 2bc$ for non-negative a, b and c , we further have $\gamma_i^2 \geq \gamma^2 - 2\frac{\gamma KS}{w^i}$.

Returning to the first inequality in (C.5), the bound we just proved implies

$$\begin{aligned} -T \log(K) &\leq -\frac{1}{2} \sum_{i=1}^N w^i \gamma^2 + \gamma K S N + N \mathcal{R}(T) \\ &\leq -\frac{\gamma^2}{4} \sum_{i=1}^N M_{i-1} + \gamma K S N + N \mathcal{R}(T) \quad (\text{by (C.3)}) \\ &\leq \min_{i \in [N]} M_i \cdot \frac{\gamma^2 N}{4} + \gamma K S N + N \mathcal{R}(T). \end{aligned}$$

From here we proceed as in the first part of the proof to get the result. ■

Lemma 16 (Freedman's Inequality (Beygelzimer et al., 2011)) *Let $(Z_t)_{t \leq n}$ be a real-valued martingale difference sequence adapted to a filtration $(\mathcal{J}_t)_{t \leq n}$ with $|Z_t| \leq R$ almost surely. For any $\eta \in [0, 1/R]$, with probability at least $1 - \delta$,*

$$\sum_{t=1}^n Z_t \leq \eta(e-2) \sum_{t=1}^n \mathbb{E}[Z_t^2 | \mathcal{J}_t] + \frac{\log(1/\delta)}{\eta} \quad (\text{C.6})$$

for all $\eta \in [0, 1/R]$.

Lemma 17 (Lemma 23 (Foster et al., 2018)) *With probability at least $1 - \delta$, the predictions $(\hat{y}_t)_{t \leq n}$ generated by [Algorithm 4](#) satisfy*

$$\sum_{t=1}^T \mathbb{I}\{\hat{y}_t \neq y_t\} \leq 4 \min_i \sum_{t=1}^T \mathbb{I}\{\hat{y}_t^i \neq y_t\} + 2 \log(N/\delta).$$

Lemma 18 (Lemma 24 (Foster et al., 2018)) *The multiclass logistic loss satisfies for any $z \in \mathbb{R}^K$ and $y \in [K]$,*

$$\ell(z + \alpha \mathbf{e}_l, y) - \ell(z, y) \leq \begin{cases} (e^\alpha - 1)[\sigma(z)]_l, & l \neq y, \\ (e^{-\alpha} - 1)(1 - [\sigma(z)]_y), & l = y. \end{cases}$$

Lemma 19 (Jung et al. (2017)) *For any $A, B \geq 0$ with $A - B \in [-1, +1]$ and $A + B \leq 1$,*

$$\inf_{\alpha \in [-2, 2]} [A(e^\alpha - 1) + B(e^{-\alpha} - 1)] \leq -\frac{(A - B)^2}{2}.$$

Proof of Theorem 13 We reduce the problem successively to a regular regression problem for which we can induce Theorem 3. Observe that for all $k \neq l_t$, there exists a constant $c_t \in \mathbb{R}$ such that $\hat{s}_{t,k} = s_{t,k} + c$. Hence $[\sigma(\hat{s}_t)]_{k \in [K] \setminus y_t} \propto [\sigma(s_t)]_{k \in [K] \setminus y_t}$. For the loss, this implies

$$-\log([\sigma(\hat{s}_t)]_k) + \log([\sigma(s_t + \alpha \mathbf{e}_{l_t})]_k) = -\log(1 - [\sigma(\hat{s}_t)]_{l_t}) + \log(1 - [\sigma(s_t + \alpha \mathbf{e}_{l_t})]_{l_t}).$$

By construction $[\sigma(\hat{s}_t)]_{l_t} = [\sigma(\zeta_t)]_1$ and $[\sigma(s_t + \alpha \mathbf{e}_{l_t})]_{l_t} = [\sigma(\tilde{s}_t + \alpha \mathbf{e}_1)]_1$, this implies

$$\ell(\hat{s}_t, y_t) - \ell(s_t + \alpha \mathbf{e}_{l_t}, y_t) = \ell(\zeta_t, 1 + \mathbb{I}\{y_t \neq l_t\}) - \ell(\tilde{s}_t + \alpha \mathbf{e}_1, 1 + \mathbb{I}\{y_t \neq l_t\}).$$

Shifting the logits by a constant does not change the distribution or the loss, hence setting

$$W_\alpha = \begin{pmatrix} 1 & \frac{1}{2}\alpha \\ -1 & -\frac{1}{2}\alpha \end{pmatrix}, \quad x_t = \begin{pmatrix} \frac{1}{2}(\tilde{s}_{t,1} - \tilde{s}_{t,2}) \\ 1 \end{pmatrix},$$

we have

$$\ell(\tilde{s}_t + \alpha \mathbf{e}_1, 1 + \mathbb{I}\{y_t \neq l_t\}) = \ell(W_\alpha x_t, 1 + \mathbb{I}\{y_t \neq l_t\}).$$

The algorithm clips the logits to the range $[-\log(T), \log(T)]$ by $\tilde{x}_{t,1} = \min\{\log(T), \max\{-\log(T), x_{t,1}\}\}$. We bound the induced error by the clipping

$$\begin{aligned}
 & \max_{x_{t,1} \in \mathbb{R}, \alpha \in [-2,2], y \in \{1,2\}} \ell(W_\alpha \tilde{x}_t, y) - \ell(W_\alpha x_t, y) \\
 &= \max_{x_{t,1} \in \mathbb{R}, \alpha \in [-2,2]} \ell(W_\alpha \tilde{x}_t, 1) - \ell(W_\alpha x_t, 1) \quad (\text{due to symmetry over the labels}) \\
 &= \max_{x_{t,1} \in \mathbb{R}, \alpha \in [-2,2]} -\log\left(\frac{\exp(\tilde{x}_{t,1} + \frac{1}{2}\alpha)}{\exp(\tilde{x}_{t,1} + \frac{1}{2}\alpha) + \exp(-\tilde{x}_{t,1} - \frac{1}{2}\alpha)}\right) + \log\left(\frac{\exp(x_{t,1} + \frac{1}{2}\alpha)}{\exp(x_{t,1} + \frac{1}{2}\alpha) + \exp(-x_{t,1} - \frac{1}{2}\alpha)}\right) \\
 &= \max_{x_{t,1} \in \mathbb{R}, \alpha \in [-2,2]} \log\left(\frac{1 + \exp(-2\tilde{x}_{t,1} - \alpha)}{1 + \exp(-2x_{t,1} - \alpha)}\right) \\
 &= \max_{x_{t,1} \in \mathbb{R}, \alpha \in [-2,2]} \log\left(\frac{1 + \exp(-2\log(T) - \alpha)}{1 + \exp(-2x_{t,1} - \alpha)}\right) \quad (\tilde{x}_{t,1} < x_{t,1} \text{ implies } \tilde{x}_{t,1} = \log(T)) \\
 &= \log(1 + \exp(-2\log(T) + 2)) \leq \frac{e^2}{T^2}.
 \end{aligned}$$

Combining everything up till now yields

$$\begin{aligned}
 \sum_{t=1}^T \ell(\hat{s}_t, y_t) - \ell(s_t, y_t) &\leq T \frac{e^2}{T^2} + \sum_{t=1}^T \ell(\zeta_t, 1 + \mathbb{I}\{y_t \neq l_t\}) - \ell(W_\alpha \tilde{x}_t, 1 + \mathbb{I}\{y_t \neq l_t\}) \\
 &\leq \mathcal{O}(1) + \max_{W: \|W\|_{2,\infty} \leq 2} \sum_{t=1}^T \ell(\zeta_t, 1 + \mathbb{I}\{y_t \neq l_t\}) - \ell(W \tilde{x}_t, 1 + \mathbb{I}\{y_t \neq l_t\}).
 \end{aligned}$$

Finally invoking Theorem 3 with $d = 2, B = 2, R = \log(T) + 1, K = 2$ completes the proof. ■

Appendix D. Reduction to AIOLI for binary logistic regression

The algorithm AIOLI proposed in [Jézéquel et al. \(2020\)](#) for the case of $K = 2$ proposes to use a regularizer $\phi_t(W) \propto \sum_{y=1}^K \ell(W x_t, y)$ without the additional bias term B_t introduced by our algorithm. We note that the proposal in [Jézéquel et al. \(2020\)](#) is for an alternative formulation of binary case and uses a different multiplicative constant (i.e. BR as opposed to $BR + \ln(K)$), however the core of the algorithm is in the choice of the regularizer $\phi_t(W)$. In this section we show that for the binary case (due to the inherent symmetries of the problem) instantiating our algorithm with $B_t = 0$ (as opposed to our proposal) leads to the exact same predictions. We want to emphasize that this emerges from the special structure of the binary case, but does not hold for $K > 2$.

Let us begin by understanding certain symmetries of the logistic regression problem. Consider splitting the (vectorized) parameter space \mathbb{R}^{Kd} into two orthogonal spaces $\mathcal{V} \triangleq \{\mathbf{1}_K \otimes z \mid z \in \mathbb{R}^d\}$ and its orthogonal space \mathcal{V}^\perp . Given any $W \in \mathbb{R}^{K \times d}$ define the projections $W^\mathcal{V}$ and $W^{\mathcal{V}^\perp}$ obtained by projecting \vec{W} onto \mathcal{V} and \mathcal{V}^\perp , and reshaping into $K \times d$ matrices, so that

$$W = W^\mathcal{V} + W^{\mathcal{V}^\perp}.$$

Now note that, since $\sigma(z + \gamma \mathbf{1}_K) = \sigma(z)$ for any γ , and for any $x \in \mathbb{R}^d$ and $V \in \mathcal{V}$ we have $Vx \propto \mathbf{1}_K$, our predictions only depend on $W_t^{\mathcal{V}^\perp}$, i.e.

$$\sigma(W_t x_t) = \sigma(W_t^{\mathcal{V}^\perp} x_t).$$

Furthermore consider any $V \in \mathcal{V}$, and let $z \in \mathbb{R}^d$ be such that $V = \mathbf{1}_K \otimes z$. For any $x_t \in \mathbb{R}^d$ and for any W the hessian $\nabla^2 \ell_t(W)$ and gradient satisfy

$$\begin{aligned}\nabla^2 \ell_t(W)V &= ((\text{diag}(\sigma_t(W)) - \sigma_t(W)\sigma_t(W)^\top)\mathbf{1}_K) \otimes (x_t x_t^\top z_t) = 0 \\ \langle \nabla \ell_t(W), V \rangle &= \langle \sigma_t(W) - y_t, \mathbf{1}_K \rangle \langle x_t, z \rangle = 0.\end{aligned}$$

It can now be seen that the optimization problem solved at every step can therefore be split by considering $W = U + V$, as

$$\begin{aligned}\min_{W \in \mathbb{R}^{d \times K}} \lambda \|W\|_F^2 + \sum_{s=1}^{t-1} \hat{\ell}_s(W) + \phi_t(W) \\ = \min_{U \in \mathcal{V}^\perp} \left(\lambda \|U\|_F^2 + \sum_{s=1}^{t-1} \hat{\ell}_s(U) + \phi_t(U) \right) + \min_{V \in \mathcal{V}} (\lambda \|V\|_F^2 + \langle V, B_t \rangle).\end{aligned}$$

Note that the optimization over V is irrelevant for the eventual predictions. We now show that in the binary case, $B_t \in \mathcal{V}$ which means that for any $U \in \mathcal{V}^\perp$, $\langle B_t, U \rangle = 0$, thereby implying that setting it to 0 does not affect the eventual predictions.

Binary case. Let $K = 2$ and through the run of the algorithm denote $\sigma(W_t x_t)_1 = p_t$ and therefore we have that $\sigma(W_t x_t)_2 = 1 - p_t$. Denoting $\sigma_t = \sigma(W_t x_t)$, we have

$$\nabla^2 \ell_t(W_t) = (\text{diag}(\sigma_t) - \sigma_t \sigma_t^\top) \otimes (x_t x_t^\top) = p_t(1 - p_t) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes x_t x_t^\top.$$

Hence setting $M_t = \sum_{s=1}^t p_s(1 - p_s)x_s x_s^\top$, we have

$$A_t = \begin{pmatrix} \lambda \mathbf{I}_d + M_t & -M_t \\ -M_t & \lambda \mathbf{I}_d + M_t \end{pmatrix}.$$

With some algebra, we can show that

$$\begin{aligned}A_t^{-1} &= \begin{pmatrix} \lambda^{-1} \mathbf{I}_d - \tilde{M}_t & \tilde{M}_t \\ \tilde{M}_t & \lambda^{-1} \mathbf{I}_d - \tilde{M}_t \end{pmatrix} \\ \tilde{M}_t &= \lambda^{-1} (2M_t + \lambda \mathbf{I}_d)^{-1} M_t.\end{aligned}$$

Hence the bias term for $K = 2$ is given by

$$\begin{aligned}B_t &= \frac{1}{2} \mathbf{1}_2 \otimes x_t - \frac{1}{2} A_t \text{diag}_{\otimes}(A_t^{-1}) \mathbf{1}_2 \otimes x_t \\ &= \frac{1}{2} \mathbf{1}_K \otimes x_t - \frac{1}{2} \mathbf{1}_2 \otimes ((\mathbf{I}_d - \lambda \tilde{M}_t) x_t) \\ &= \frac{\lambda}{2} \mathbf{1}_K \otimes (\tilde{M}_t x_t) \in \mathcal{V}.\end{aligned}$$

Appendix E. Efficient Implementation of Algorithm 1

In this section we provide a proof of Theorem 4. We begin by noting that the algorithm can be implemented by computing the vector $z_t = W_t x_t$ and not the full matrix W_t . We show that this can be done efficiently. Note that the algorithm computes W_t as the following

$$W_t = \underset{W \in \mathbb{R}^{K \times d}}{\operatorname{argmin}} \|\vec{W}\|_{A_{t-1}}^2 + \langle \vec{W}, G_{t-1} \rangle + \phi_t(W).$$

Since $A_{t-1} \succeq \lambda I_{Kd}$, the above minimization has a unique solution which can be obtained via the following first order optimality condition,

$$2A_{t-1}\vec{W}_t + G_{t-1} - \frac{1}{2}A_{t-1} \operatorname{diag}_{\otimes}(A_{t-1}^{-1})(\mathbf{1}_K \otimes x_t) + \sigma(W_t x_t) \otimes x_t = 0.$$

Rearranging leads to

$$\vec{W}_t = -\frac{1}{2}A_{t-1}^{-1}(G_{t-1} + \sigma(W_t x_t) \otimes x_t) + \frac{1}{4} \operatorname{diag}_{\otimes}(A_{t-1}^{-1})(\mathbf{1}_K \otimes x_t).$$

Remember that $A_{t-1} \in \mathbb{R}^{Kd \times Kd}$. For a matrix $M \in \mathbb{R}^{Kd \times Kd}$, and for $i, j \in [K]$, we denote by $[[M]]_{i,j}$ the $d \times d$ matrix obtained by segmenting M into K^2 contiguous submatrix blocks of size $d \times d$ in the natural manner, and taking the $(i, j)^{th}$ block. Now define the matrix $\tilde{A} \in \mathbb{R}^{K \times K}$, whose $(i, j)^{th}$ entry is given by

$$[\tilde{A}]_{i,j} = \frac{1}{2}x_t^\top [[A_{t-1}^{-1}]]_{i,j} x_t.$$

Further define $\tilde{g} \in \mathbb{R}^K$ as a vector whose k^{th} entry is given by

$$[\tilde{g}]_k = -\frac{1}{2}\langle x_t, (A_{t-1}^{-1}G_{t-1})_k \rangle + \frac{1}{4}x_t^\top [[A_{t-1}^{-1}]]_{k,k} x_t.$$

We first note the following implications,

$$A_{t-1} \succ \lambda \mathbf{I}_{Kd} \Rightarrow A_{t-1}^{-1} \prec \lambda^{-1} \mathbf{I}_{Kd},$$

$$A_{t-1} \preceq (TR^2 + \lambda) \mathbf{I}_{Kd} \Rightarrow A_{t-1}^{-1} \succeq (TR^2 + \lambda)^{-1} \mathbf{I}_{Kd}$$

Furthermore note that for any $v \in \mathbb{R}^K$, we have that

$$v^\top \tilde{A} v = (v \otimes x_t)^\top A_{t-1}^{-1} (v \otimes x_t).$$

If $\lambda_{\max}(M)$ denotes the largest eigenvalue of a matrix M , then the above equation implies that $\lambda_{\max}(\tilde{A}) \leq \lambda_{\max}(A_{t-1}^{-1})\|x_t\|_2^2 = \frac{R^2}{\lambda}$.

In terms of the computation of z_t , it can be seen that z_t is the solution of the following equation

$$z_t = \tilde{g} - \tilde{A}\sigma(z_t),$$

which in turn is the first order optimality condition of

$$z_t = \underset{z \in \mathbb{R}^K}{\operatorname{argmin}} \left[\psi(z) \triangleq \frac{1}{2}\|z\|_{\tilde{A}^{-1}}^2 - \langle z, \tilde{A}^{-1}\tilde{g} \rangle + \log \left(\sum_{k=1}^K \exp(z_k) \right) \right].$$

We can precondition the above optimization problem as follows: set $\tilde{z} = \tilde{A}^{-1/2}z$ and instead solve the following optimization problem

$$\tilde{z}^* = \underset{\tilde{z} \in \mathbb{R}^K}{\operatorname{argmin}} \tilde{\psi}(\tilde{z}) \triangleq \underset{\tilde{z} \in \mathbb{R}^K}{\operatorname{argmin}} \frac{1}{2} \|\tilde{z}\|^2 - \langle \tilde{z}, \tilde{A}^{-1/2}\tilde{g} \rangle + \log \left(\sum_{k=1}^K \exp([\tilde{A}^{1/2}\tilde{z}]_k) \right). \quad (\text{E.1})$$

It can be observed that the Hessian of the log term above is bounded by $\lambda_{\max}(\tilde{A})\mathbf{I} \preceq \frac{R^2}{\lambda}\mathbf{I}_K$ and therefore the smoothness of the above optimization problem is bounded by $\left(1 + \frac{R^2}{\lambda}\right)$. In particular this implies that τ steps of gradient descent, with step size $= \left(\frac{1}{1 + \frac{R^2}{\lambda}}\right)$, generating the sequence $\{\tilde{z}_0 \dots \tilde{z}_\tau\}$ (\tilde{z}_0 will be specified momentarily), on the above problem satisfies the following bound

$$\|\tilde{z}_\tau - \tilde{z}^*\|^2 \leq \left(1 + \frac{R^2}{\lambda}\right) \exp \left(-\frac{\tau}{\left(1 + \frac{R^2}{\lambda}\right)} \right) (\tilde{\psi}(\tilde{z}_0) - \tilde{\psi}(\tilde{z}^*)).$$

This implies that

$$\|\tilde{A}^{1/2}\tilde{z}_\tau - z_t\|^2 \leq \left(1 + \frac{R^2}{\lambda}\right)^2 \exp \left(-\frac{\tau}{\left(1 + \frac{R^2}{\lambda}\right)} \right) (\tilde{\psi}(\tilde{z}_0) - \tilde{\psi}(\tilde{z}^*)).$$

$\tilde{\psi}$ is 1-strongly convex, and hence by the Polyak-Lojasiewicz (PL) inequality,

$$\tilde{\psi}(\tilde{z}_0) - \tilde{\psi}(\tilde{z}^*) \leq \frac{1}{2} \|\nabla \tilde{\psi}(\tilde{z}_0)\|_2^2.$$

Setting $\tilde{z}_0 = \tilde{A}^{-1/2}\tilde{g}$ yields

$$\|\nabla \tilde{\psi}(\tilde{z}_0)\|_2^2 = \|\tilde{A}^{1/2}\sigma(A^{1/2}\tilde{z}_0)\|_2^2 \leq \left(1 + \frac{R^2}{\lambda}\right).$$

We see that within τ steps of gradient descent on the problem (E.1) we can obtain a vector $\hat{z}_t = \tilde{A}^{1/2}\tilde{z}_\tau$ such that

$$\|\hat{z}_t - z_t\|^2 \leq \left(1 + \frac{R^2}{\lambda}\right)^3 \exp \left(-\frac{\tau}{\left(1 + \frac{R^2}{\lambda}\right)} \right).$$

Therefore setting $\tau = \left(1 + \frac{R^2}{\lambda}\right) \log \left(\epsilon^{-1} \left(1 + \frac{R^2}{\lambda}\right)^3 \right)$, we see that the error $\|\hat{z}_t - z_t\|^2 \leq \epsilon$.

In terms of computation, note that maintaining the inverse of A_{t-1} requires $O(d^2K^3)$ time. This can be seen by the fact that the update to A_{t-1} is of rank at most K at every step and the update can be performed via the Sherman-Morrison-Woodbury formula. Having computed A_{t-1}^{-1} , it can be seen that the quantities $\tilde{A}, \tilde{g}, \tilde{A}^{1/2}, \tilde{A}^{-1/2}$ can all be computed in time $O(d^2K^2 + K^3)$. Having computed these quantities it is easy to see that every step of gradient descent on (E.1) takes time at most $O(K^2)$. Therefore total running time for computing a vector \hat{z}_t such that $\|\hat{z}_t - z_t\|^2 \leq \epsilon$ in total time

$$O \left(d^2K^3 + K^2 \left(1 + \frac{R^2}{\lambda}\right) \log \left(\epsilon^{-1} \left(1 + \frac{R^2}{\lambda}\right)^3 \right) \right).$$

Appendix F. Challenges for Hessian-dominance based analysis

In this section we describe the difficulty in extending the analysis technique of [Jézéquel et al. \(2020\)](#) to online logistic regression with more than 2 classes. This analysis is based on a Hessian-dominance technique which works as follows. We start from the instantaneous regret defined from Lemma 6:

$$\|\nabla\phi_t(W_t) - \nabla\ell_t(W_t)\|_{A_t^{-1}}^2 - \|\nabla\phi_t(W_t)\|_{A_{t-1}^{-1}}^2.$$

Since $A_t \succeq A_{t-1}$, this can be upper bounded by

$$\begin{aligned} & \|\nabla\phi_t(W_t) - \nabla\ell_t(W_t)\|_{A_t^{-1}}^2 - \|\nabla\phi_t(W_t)\|_{A_t^{-1}}^2 \\ &= \text{Tr}((\nabla\phi_t(W_t) - \nabla\ell_t(W_t))(\nabla\phi_t(W_t) - \nabla\ell_t(W_t))^\top - \nabla\phi_t(W_t)\nabla\phi_t(W_t)^\top]A_t^{-1}). \end{aligned}$$

Then, if we choose a regularizer ϕ_t such that the following *Hessian-dominance* condition

$$(\nabla\phi_t(W_t) - \nabla\ell_t(W_t))(\nabla\phi_t(W_t) - \nabla\ell_t(W_t))^\top - \nabla\phi_t(W_t)\nabla\phi_t(W_t)^\top \preceq c\nabla^2\ell_t(W_t) \quad (\text{F.1})$$

holds for some constant c , then the instantaneous regret can be upper bounded by

$$\text{Tr}(c\nabla^2\ell_t(W_t)A_t^{-1}).$$

Summing this up from $t = 1$ to T yields a harmonic sum that can be bounded by standard techniques by $O(c \log T)$. Note that in this analysis we need $c = \text{poly}(B, R)$ to get polynomial dependence on B and R in the regret bound.

For the binary case $K = 2$, it is easy to check that both of the following choices of ϕ_t

$$\phi_t(W) = \ell(Wx_t, 1) + \ell(Wx_t, 2) \quad \text{or} \quad \phi_t(W) = \frac{1}{2}\ell(Wx_t, 1) + \frac{1}{2}\ell(Wx_t, 2)$$

can be used to satisfy (F.1). [Jézéquel et al. \(2020\)](#) use the first choice of ϕ_t to design AIOLI.

Unfortunately, this elegant Hessian dominance technique breaks down when we have more than 2 classes, at least for regularizers of a certain form which satisfy a certain *symmetry* condition (satisfied by the two choices for the binary case above). To describe the issue, it will be convenient to drop the t subscript since it is irrelevant to the analysis. We will assume that the regularizer is of the form $\phi(W) = \psi(Wx)$ for some function $\psi : \mathbb{R}^K \rightarrow \mathbb{R}$, and the symmetry condition we need is the following: for any two classes $y \neq y'$ and any $z \in \mathbb{R}^K$, if $z_y = z_{y'}$, then $\nabla\psi(z)_y = \nabla\psi(z)_{y'}$.

Let $K = 3$. Consider a regularizer ϕ satisfying the form and symmetry conditions above. Assume w.l.o.g. that the true label is 1. Then the gradients of the loss and regularizer are

$$\nabla\ell(W) = (\sigma(Wx) - \mathbf{e}_1) \otimes x \quad \text{and} \quad \nabla\phi(W) = \nabla\psi(Wx) \otimes x.$$

With some abuse of notation, for the rest of this section we will use σ to denote $\sigma(Wx)$ and ψ' to denote $\nabla\psi(Wx)$. Then the Hessian equals

$$(\text{diag}(\sigma) - \sigma\sigma^\top) \otimes (xx^\top).$$

The Hessian dominance condition (F.1) reduces to the following:

$$[(\sigma - \mathbf{e}_1 - \psi')(\sigma - \mathbf{e}_1 - \psi')^\top - \psi'\psi'^\top] \otimes (xx^\top) \preceq c(\text{diag}(\sigma) - \sigma\sigma^\top) \otimes (xx^\top). \quad (\text{F.2})$$

We can now show that the above condition cannot hold uniformly for all W and x such that $\|W\|_{2,\infty} \leq B$ and $\|x\| \leq R$ unless $c = \Omega(\exp(BR))$. In particular, choose

$$W = \text{diag}(0, 0, B) \quad \text{and} \quad x = (0, 0, R)^\top,$$

so that $Wx = (0, 0, BR)^\top$. Now let $v = (\mathbf{e}_1 - \mathbf{e}_2) \otimes x$. Then we have $v^\top \sigma = 0$, since $\sigma_1 = \sigma_2$, and $v^\top \psi' = 0$, since $\psi'_1 = \psi'_2$ by the symmetry condition. Using these facts, we have

$$v^\top \left([(\sigma - \mathbf{e}_1 - \psi')(\sigma - \mathbf{e}_1 - \psi')^\top - \psi' \psi'^\top] \otimes (xx^\top) \right) v = v^\top [(\mathbf{e}_1 \mathbf{e}_1^\top) \otimes (xx^\top)] v = \|x\|_2^4 = R^4.$$

Whereas, we have

$$v^\top \left((\text{diag}(\sigma) - \sigma \sigma^\top) \otimes (xx^\top) \right) v = v^\top \left(\text{diag}(\sigma) \otimes (xx^\top) \right) v = (\sigma_1 + \sigma_2) \|x\|_2^4 = (\sigma_1 + \sigma_2) R^4.$$

Since $\sigma_1 = \sigma_2 = \frac{1}{\exp(BR)+2}$, for the Hessian dominance condition (F.2) to hold, we must have $c \geq \frac{1}{2} \exp(BR) + 1$, as claimed.