Proceedings of Machine Learning Research vol 125:1–24, 2020 33rd Annual Conference on Learning Theory

Logistic Regression Regret: What's the Catch?

Gil I. Shamir GSHAMIR@GOOGLE.COM
Google

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We address the problem of the achievable regret rates with online logistic regression. We derive lower bounds with logarithmic regret under L_1, L_2, and L_∞ constraints on the parameter values. The bounds are dominated by $d/2 \log T$, where T is the horizon and d is the dimensionality of the parameter space. We show their achievability for $d = o(T^{1/3})$ in all these cases with Bayesian methods, that achieve them up to a $d/2 \log d$ term. Interesting different behaviors are shown for larger dimensionality. Specifically, on the negative side, if $d = \Omega(\sqrt{T})$, any algorithm is guaranteed regret of $\Omega(d \log T)$ (greater than $\Theta(\sqrt{T})$) under L_∞ constraints on the parameters (and the example features). On the positive side, under L_1 constraints on the parameters, there exist Bayesian algorithms that can achieve regret that is sub-linear in d for the asymptotically larger values of d. For L_2 constraints, it is shown that for large enough d, the regret remains linear in d but no longer logarithmic in T. Adapting the *redundancy-capacity* theorem from information theory, we demonstrate a principled methodology based on grids of parameters to derive lower bounds. Grids are also utilized to derive some upper bounds. Our results strengthen results by Kakade and Ng (2005) and Foster et al. (2018) for upper bounds for this problem, introduce novel lower bounds, and adapt a methodology that can be used to obtain such bounds for other related problems. They also give a novel characterization of the asymptotic behavior when the dimension of the parameter space is allowed to grow with T. They additionally strengthen connections to the information theory literature, demonstrating that the actual regret for logistic regression depends on the richness of the parameter class, where even within this problem, richer classes lead to greater regret.

Keywords: Logistic regression, online learning, Bayesian methods, convex optimization, regret, redundancy capacity theorem.

1. Introduction

Logistic regression plays a significant role in many learning applications, where a set of parameters representing the effects of different features on the outcome (label) is learned from a training data set with known *labels*. The learned parameters are then used to predict the true labels of, yet unseen, data examples. Examples include predicting the probability some person carries some disease based on features that are, e.g., hereditary or environmental; or predicting the click-through-rate of ads shown in online advertising. Many applications may require to operate in the online learning (or online convex optimization) setting. In this setting, an algorithm consumes the data in rounds. At round t, predictions can be based on all examples seen up to round $t-1$, including on their true labels (but not on data beyond round $t-1$), to predict the label of the example at round t.

The performance of an online algorithm is measured by its *regret*, which is defined as the extra loss it incurs beyond that of an algorithm that is playing, at all rounds, some *comparator* value $\theta^* \in \Theta$, where Θ is a predefined space of possible values. The values of the parameters θ^* can be those that minimize the cumulative loss over all rounds up to the horizon T. While regret is defined for the online setting, it is directly connected to the *convergence rate*, which measures an expected loss on an unseen example at round $T+1$, based on training on the first T examples.

Paper Outline: In Section 2, we outline our contributions. We present a summary of related work in Section 3. Section 4 formulates the problem. In Section 5, we frame and extend results from the literature, setting them to prove our results. Section 6 describes regret lower bounds for any algorithm. Section 7 shows upper bounds that can be achieved with *Bayesian mixture* algorithms and apply to logistic regression when the *feature vector* x_t is observed prior to predicting a label.

2. Summary of Contributions and Methods

We consider several settings with a d-dimensional parameter space limited by B (B can be function of T, including $o(1)$) , specifically, $\Theta \stackrel{\triangle}{=} \{\theta^* : \|\theta^*\|_\rho \leq B\}$, for $\rho = 1, 2, \infty$. Define $\gamma = B/(\log T)$ as the count of $\log T$ units constituting B. We focus on the case in which the norm of the *example* (or feature value vector) x_t at t is bounded in L_∞, i.e., $|x_{t,i}| \leq 1, \forall i = 1, 2, \ldots, d$, (or $\|x_t\|_\infty \leq 1$). (This setup generalizes the practical setup with binary features). However, with proper adjustments (which decrease the bounds), the results transform also to the more restrictive $\|x_t\|_2 \leq 1$.

Our contributions include:

- Comprehensive characterization of the regret for the logistic regression problem, including the asymptotic behavior in the dimensionality d, showing regret bounds logarithmic in T and linear in d for lower regions of d.
- Novel bounds that lead to this characterization, especially, lower bounds showing limitations on regret in the different settings.
- Specific negative results that demonstrate that in cases such as L_∞ constraints, for $d = \Omega(\sqrt{T})$, we are guaranteed regret rates of at least $\Omega(\sqrt{T}\log T)$.
- Specific positive results that demonstrate that for upper regions of d, there exist Bayesian algorithms with regret rates $o(d)$ (for L_1 constraints with $d = \omega(B\sqrt{T})$), as well as regret rates that are linear in d, and no longer logarithmic in T (for L_2 constraints with $d = \Omega(B^2T)$).
- Strengthened adaptation of a principled methodology from the information theory literature, that allows derivation of lower bounds for this and related problems.

The sub-linear regret in d for L_1 is very interesting for logistic regression because the dot product, used for prediction, is a linear combination of the parameters, making L_1 constraints very realistic, especially in sparse real-worlds problems that have binary feature vectors (see, e.g., McMahan et al. (2013)).

Our results characterize the behavior of the regret for the various regions of d. For smaller d, we show lower bounds of $(d/2)\log(T/d)$, $(d/2)\log(T/d^2)$, and $(d/2)\log(T/d^3)$ for the cases where $\|\theta^*\|_\rho \leq B$ and ρ is $\infty, 2, 1$, respectively, and upper bounds of $(d/2)\log(B^2T)$, $(d/2)\log(B^2T/d)$, and $(d/2)\log(B^2T/d^2)$ for the respective norm constraints. (An additional d term in the denominator of the logarithm in all bounds applies to the setting in which $\|x_t\|_2 \leq 1$.) The difference between

the constraints on different norms illustrates that regret is a function of the richness of Θ. The richer is Θ (e.g., L_∞ constraints are richer than L_2, which are richer than L_1) the greater is the regret. As the dimension d is allowed to grow, the bounds change when the denominator of the logarithmic term above equals the numerator. They lead to different regions of the different bounds, with different behavior in each region. Table 1 shows the different lower and upper bounds for different d and different norm constraints on the space Θ, and summarizes the results in Theorems 3-5. For simplicity of the table, we omitted the lower limit on d for each row, but it should be understood as $\Theta(n)$ where the upper limit of the previous row is $o(n)$ (for the first row in each block, the previous $n = 1$). The lower bounds that are $\Theta(T^\alpha)$ should be understood as $\Theta(T^{\alpha(1-\varepsilon)})$ for some small $\varepsilon > 0$ which can be as small as $O(\log\log T/(\log T))$. This is again omitted for simplicity. For the setting in which $\|x_t\|_2 \leq 1$, the additional d term in the denominators of the logarithm leads to earlier transitions between regions of d, for all cases (as well as adding a $d/2$ upper regret region for L_∞). Table 2 compares results in this paper to previously reported results (described in more detail in Section 3). We omit middle ranges of d that are in Table 1. Some results in this paper are extended from Kakade and Ng (2005) and adapted to the setup $\|x_t\|_\infty \leq 1$. A footnote marks these with a proper explanation. For multi labels, we use $\theta^{*(m)}$ to denote the d-dimensional projection of the parameter space for label m. Results for L_1 and L_∞ (that were not directly derived) are generalized from results that were derived for L_2 and are described in the "previous results" column.

Table 1: Summary of regret bounds

Norm Constraint	[1] Dimension d	[2] Lower Bound	Upper Bound
$L_1 : \|\theta^*\|_1 \leq B$	$o\left((\gamma T)^{1/3}\right)$	$\frac{d}{2}\log\frac{\gamma T}{d^3}$	$\frac{d}{2}\log\frac{B^2T}{d^2}$
	$o\left(B\sqrt{T}\right)$	$\Theta\left((\gamma T)^{1/3}\right)$	$\frac{d}{2}\log\frac{B^2T}{d^2}$
	$\omega\left(B\sqrt{T}\right)$	$\Theta\left((\gamma T)^{1/3}\right)$	$o(d)$
$L_2 : \|\theta^*\|_2 \leq B$	$o\left(\sqrt{\gamma T}\right)$	$\frac{d}{2}\log\frac{\gamma T}{d^2}$	$\frac{d}{2}\log\frac{B^2T}{d}$
	$o\left(B^2T\right)$	$\Theta\left(\sqrt{\gamma T}\right)$	$\frac{d}{2}\log\frac{B^2T}{d}$
	$\Omega\left(B^2T\right)$	$\Theta\left(\sqrt{\gamma T}\right)$	$\frac{d}{2}$
$L_\infty : \|\theta^*\|_\infty \leq B$	$o\left(\gamma T\right)$	$\frac{d}{2}\log\frac{\gamma T}{d}$	$\frac{d}{2}\log\left(B^2T\right)$
	$\Omega\left(\gamma T\right)$	$\Theta\left(\gamma T\right)$	$\frac{d}{2}\log\left(B^2T\right)$

To prove lower bounds, we adapt techniques based on the *redundancy-capacity theorem* (see, e.g., Davisson (1973); Merhav and Feder (1995); Shamir (2006a)) from the information theory literature. Specifically, we set a *grid* of points in the parameter space that are distinguishable by the observed label sequence for some example sequence. The logarithm of the cardinality of the grid is a lower bound on the regret. The concept of distinguishability was used somewhat differently by Hazan et al. (2014) to prove regret lower bounds. Upper bounds for L_2 and L_∞, but not for L_1, can be derived by manipulating the *Bayesian mixture* approach in Kakade and Ng (2005), (adjusted to

1. The dimension column shows an upper limit on the shown range. The lower limit should be understood as $\Theta(n)$ where the upper limit for the previous row is $o(n)$, ($n = 1$ for the row "previous" to the first one in a block).
2. Lower bounds that are $\Theta(T^\alpha)$ should be understood as $\Theta(T^{\alpha(1-\varepsilon)})$ for some small $\varepsilon > 0$.

Table 2: Comparison of regret bounds in this paper with previously reported bounds

Problem Setting		Previous Results	This Paper
Binary Labels, $d = 1$	:	$\mathcal{R} = \frac{1}{2}\log T$ [Davisson (1973)] [Krichevsky and Trofimov (1981)] [McMahan and Streeter (2012)] [1]	
Multi Labels [2] $\omega(1) = m = o(T)$ $d = 1$	:	$\mathcal{R} = \frac{m}{2}\log(T/m)$ [Krichevsky and Trofimov] [Orlitsky and Santhanam (2004)] [Shamir (2006a)]	
Binary Labels [3] Multi Dimensions d $L_1 : \|\theta^*\|_1 \leq B$	:	$O(B\sqrt{dT})$ [Xiao (2010)] $\mathcal{R} \leq \frac{d}{2}\log(1+T)$ [Kakade and Ng (2005)] $\mathcal{R} \leq \frac{d}{2}\log\left(\frac{B^2T}{d} + e\right)$ [Kakade and Ng] [4] $\mathcal{R} \leq 5d\log\left(\frac{BT}{d} + e\right)$ [Foster et al. (2018)]	$d = o\left((\gamma T)^{1/3}\right)$: $\frac{d}{2}\log\frac{\gamma T}{d^3} \leq \mathcal{R} \leq \frac{d}{2}\log\frac{B^2T}{d^2}$ $d = \Omega\left(B\sqrt{T}\right)$: $\Theta\left((\gamma T)^{1/3}\right) \leq \mathcal{R} = o(d)$
Binary Labels Multi Dimensions d $L_2 : \|\theta^*\|_2 \leq B$	:	$O(B\sqrt{dT})$ [Xiao (2010)] $\mathcal{R} \leq \frac{d}{2}\log(1+T)$ [Kakade and Ng (2005)] $\mathcal{R} \leq \frac{d}{2}\log\left(\frac{B^2T}{d} + e\right)$ [Kakade and Ng] [4] $\Omega(d) \leq \mathcal{R} \leq 5d\log\left(\frac{BT}{d} + e\right)$ [Foster et al.]	$d = o\left(\sqrt{\gamma T}\right)$: $\frac{d}{2}\log\frac{\gamma T}{d^2} \leq \mathcal{R} \leq \frac{d}{2}\log\frac{B^2T}{d}$ $d = \Omega\left(B^2T\right)$: $\Theta\left(\sqrt{\gamma T}\right) \leq \mathcal{R} \leq \frac{d}{2}$
Binary Labels [3] Multi Dimensions d $L_\infty : \|\theta^*\|_\infty \leq B$	:	$O(dB\sqrt{T})$ [McMahan (2017)] $\mathcal{R} \leq \frac{d}{2}\log\left(B^2T\right)$ [Kakade and Ng (2005)] [4] $\Omega(d)$ [Foster et al. (2018)]	$d = o\left(\gamma T\right)$: $\frac{d}{2}\log\frac{\gamma T}{d} \leq \mathcal{R} \leq \frac{d}{2}\log\left(B^2T\right)$ $d = \Omega\left(\gamma T\right)$: $\Theta\left(\gamma T\right) \leq \mathcal{R} \leq \frac{d}{2}\log\left(B^2T\right)$
Multi Labels m Multi Dimensions d	:	L_2 constraints: $\mathcal{R} \leq 5md\log\left(\frac{BT}{dm} + e\right)$ [Foster et al. (2018)]	$\|\theta^{*(m)}\|_\infty \leq B$: $\mathcal{R} \geq \frac{d(m-1)}{2}\log\left(\frac{T}{d\cdot m}\right)$
Binary Labels, Multi-d, proper	:	$\Omega(\sqrt{BT})$ [Hazan et al. (2014)]	

our setup). Using a normal prior with *large variance* can attain the proper rates, with the respective constants. However, for $\|\theta^*\|_1 \leq B$, we combine this approach with the method of grids, applying a discrete uniform prior on some $\Theta_m \subseteq \Theta$. Applying the method in Kakade and Ng (2005), $\log|\Theta_m|$ initially dominates an upper bound, with additional contribution from the effective *quantization* of the parameters by the mixture only on a discrete subset of the space. This method can also be used for L_2 and L_∞, and achieves a similar bound for L_∞, but slightly weaker one for L_2.

1. The single dimensional results were known in the information theory literature, and derived using Bayesian mixture methods. McMahan and Streeter (2012) demonstrated their achievability with a *Follow The Regularized Leader (FTRL)* gradient method.
2. Results for $m = O(1)$ were known in the information theory literature since Davisson (1973) and perhaps even before that. The KT estimator achieves the upper bound also for $m = \omega(1)$. Lower bounds were derived in the references cited.
3. The *previous* results in the table for L_1 and L_∞ are implied from results for L_2 in the literature.
4. These upper bounds are derived by extending the derivation from Kakade and Ng (2005) as in Theorem 4 in our paper. For L_2 this was also shown in Foster et al. (2018).

3. Related Work

Prior results in both the machine learning literature (see, e.g. Azoury and Warmuth (2001); Cesa-Bianchi et al. (2002); Littlestone (1989)) and the information theory literature (see, e.g., Krichevsky and Trofimov (1981); Merhav and Feder (1995); Rissanen (1984)) illustrate that the performance of the regret (*redundancy* in information theory) of the online setting normalized by T meets batch convergence rates at least to first order. Hence, studying online regret also implies to generalization ability. The setup of a logistic regression problem, whether online or batch is very similar to the setups of the universal compression problems in the information theory literature. In these problems, the redundancy in predicting multi label outcomes in a setup that is equivalent to single dimensional logistic regression with binary features was studied. It was shown (see, e.g., the seminal work in Rissanen (1984), subsequent work in Drmota and Szpankowski. (2004); Orlitsky and Santhanam (2004); Shamir (2006a); Szpankowski and Weinberger (2012), and references therein) that for these problems, regret of $\frac{m}{2}\log(T/m)$ is achievable to first order, where m is the number of labels. However, the concepts presented by Rissanen (1984) should apply also to more general d dimensional problems, where d is the number of parameters that affect the label outcome. Specifically, in Rissanen (1984), central limit arguments, that are also satisfied in the logistic regression setting, were used to prove $\frac{d}{2}\log T$ redundancy bounds, when $d = \Theta(1)$. The subsequent results in Drmota and Szpankowski. (2004); Orlitsky and Santhanam (2004); Shamir (2006a,b), however, extended the redundancy results to $\frac{m}{2}\log(T/m)$, even when $m = T^{1-\varepsilon} = o(T)$ (for some small fixed $\varepsilon > 0$) but were more specific to the equivalent of single dimensional logistic regression with multi m labels.

The machine learning literature considered general online convex optimization, and derived minimax-optimal algorithms for both the linear and strongly convex settings (see, e.g., Abernethy et al. (2008)), with logarithmic regret in the strongly convex setting. General minimax bounds for a wide variety of loss functions and references classes, including large and nonparametric classes, were provided in a series of papers by Rakhlin et al. (2010a,b); Rakhlin and Sridharan (2014, 2015); Rakhlin et al. (2017). For weakly convex settings (which generally includes logistic regression), regret rates of $O(dB\sqrt{T})$ have been shown to be achievable (see, e.g., Zinkevich (2003), and references therein), and later, rates of $O(B\sqrt{dT})$ (see, e.g., Xiao (2010)), where B is the radius of the L_2 ball defining the allowed values of the parameter θ_t, played at round t, and the space Θ of values of a possible comparator θ^*.

To the best of our knowledge, in Kakade and Ng (2005), a first result suggesting that regret of $O(d\log(T/d))$ is achievable for logistic regression, and in fact for other generalized linear models, was presented. Instead of using *gradient methods*, (typically used for this problem) in which the training algorithm updates the learned parameters taking a step against the gradient on the loss, the method took from the Bayesian literature to apply *Bayesian Model Averaging* (or *Bayesian mixture*) to show a regret upper bound (but not a lower bound) that achieves this rate. In addition, however, the algorithm pays an additional penalty that depends on the prior selection as well as on the squared L_2 norm of a comparator θ^* (which can be the loss minimizing parameter in hindsight). If $\|\theta^*\|_2^2$ is larger than the $O(d\log(T/d))$ term, this penalty term could dominate the bound (depending on the selected prior). The proof of the bound utilized variational techniques, and also, in part, resembled

some of the central limit arguments used in Rissanen (1984) to show upper bounds on redundancy. The use of Bayesian methods is also justified in the information theory literature (see, e.g., Davisson (1973); Krichevsky and Trofimov (1981); Rissanen (1984)). Specifically, Merhav and Feder (1995) showed that a mixture code is as good as the best code in terms of regret (and thus can be better but not worse than any other type of code).

McMahan and Streeter (2012) demonstrated that with binary feature values, using the *Follow-The-Regularized-Leader (FTRL)* methodology (see, e.g., Hazan (2012); McMahan (2011); Rakhlin et al. (2005); Shalev-Shwartz (2007) and references therein) with a Beta regularizer, $O(\log T)$ regret can be achieved for the single dimensional problem. In the special case of a Beta regularizer with $\alpha = \beta = 1/2$, their FRTL algorithm coincides with the well-known (add-$1/2$) Krichevsky and Trofimov (1981) (KT) estimator that, in fact, achieves the lower bound on the regret for this problem of $0.5\log(T)$. It is interesting to note, however, that the KT method is derived using a Bayesian mixture with the Dirichlet-$1/2$ prior. Thus for the single dimensional case, both the FTRL methodology and the Bayesian mixture one result in the same estimator. Unfortunately, this result does not generalize to larger dimensions.

While the lower bound can be achieved for the single dimensional case for binary features with an FTRL gradient method, McMahan and Streeter (2012) posed a problem of what happens in larger dimensions. The results in Kakade and Ng (2005) hint in the direction of Bayesian methods, but still fall short of achieving $d/2\log(T/d)$ regret due to the additional penalty on the prior. (Although, as we demonstrate, these results with a proper, perhaps unexpected, choice of prior could lead to the desired rates and constants in some cases, but, to the best of our knowledge, such a result was not reported in the literature.) A series of papers Bach (2010); Bach and Moulines (2013); Bach (2014) studied the convergence rate of gradient methods for logistic regression, and concluded, that while logistic loss is not globally strongly convex, it can, depending on the actual data, locally exhibit strong convexity (referred to as the *self-concordance* property). Then, gradient methods can achieve convergence rate of $O(1/\lambda T)$, where λ is the smallest eigenvalue of the Hessian at the global optimum. This implies that gradient methods can, in many case, achieve logarithmic regret, but there do exist situations where gradient methods fail to achieve $O(d\log(T/d))$ regret (when λ is small).

Hazan et al. (2014) studied the problem, in which Bayesian methods are not possible to apply directly, where the feature values are unknown when playing θ_t at round t, and are only revealed later, together with the label. Bayesian methods do condition the predicted label probability on the observed feature values, and if such are not available, they would require also mixing on the feature values. It was shown, that in this setting, which is more difficult to the algorithm, regret of $O(B^3T^{1/3})$ is achieved for the single dimensional problem where only $\Omega(B^{2/3}T^{1/3})$ is possible, and $\Omega(\sqrt{BT})$ is only possible for any larger dimensions even for $d = 2$.

Foster et al. (2018) separated the problem posed in McMahan and Streeter (2012) to the case considered by Hazan et al. (2014), where the algorithm plays θ_t with no knowledge of the feature values x_t, which is referred to as a *proper* setting, and to the *mixable* setting where the feature values are revealed to the algorithm prior to generating a prediction, referred to as the *improper* setting. Using Bayesian model averaging with a uniform prior with an approach that resembles

that in Kakade and Ng (2005), an upper bound of $O(md\log(T/md))$ was shown for the multi label d-dimensional (with d distinct features) logistic regression problem, where m is the number of distinct labels, under L_2 constraints on θ^*. A lower bound of $\Omega(d)$ was shown for the binary labels / binary features setting under the constraints that $B = \Omega(\sqrt{d}\log T)$. The upper bound matches the logarithmic order of the bound expected from the information theory problems, but not the constant, and the lower bound is lower in order.

The results summarized above suggest that there are, in fact, two different sets of online logistic problems considered. In the first, the features x_t are revealed prior to playing θ_t or to generating a prediction, and in the second, θ_t is played before the feature values are revealed. The first problem allows the use of Bayesian methods, while the second will require such methods to also mix over the unseen x_t. For the first problem, logarithmic regret is possible for low dimensionalities, whereas for the second extreme case, it is not in many settings, even in the single dimensional problem. In this paper, we give a comprehensive characterization of the regret behavior for the first problem, including the asymptotic regime, where d is allowed to grow with T. The lower bounds we derive apply to any case, including the second problem, but the upper bounds are specific to the first one.

4. Problem Formulation, Notation and Definitions

We consider online convex optimization over a series of rounds $t \in \{1, 2, \ldots, T\}$ as in (McMahan, 2014) (see also Boyd and Vandenberghe (2004); Rockafellar (1997); Shalev-Shwartz (2012)). Each round t, a d-dimensional *example* feature vector $x_t \stackrel{\triangle}{=} \{x_{t,1}, x_{t,2}, \ldots, x_{t,d}\} \in \mathcal{X}$ and a label $y_t \in \mathcal{Y}$ are observed. For the binary labels, we use $\mathcal{Y} = \{-1, 1\}$. We assume, without loss of generality, that $|x_{t,i}| \leq 1$, as features can be normalized. We denote a subsequence up to time t by $x^t \stackrel{\triangle}{=} \{x_1, x_2, \ldots, x_t\}$. For the example/label pair, we also use $S_t \stackrel{\triangle}{=} \{(x_s, y_s)\}_{s=1}^t$. Capital letters denote random variables. A learning algorithm $\mathcal{A}$ is a function that, given a sequence S_{t-1}, an example x_t, and an arbitrary label $y \in \mathcal{Y}$, returns at round t a probability for the label

$$\mathcal{A}(S_{t-1}, x_t, y) \stackrel{\triangle}{=} P[Y_t = y | X_t = x_t, S_{t-1}]. \tag{1}$$

To produce a prediction, an algorithm may play a weight vector $\theta_t \in \Theta$, or perform a Bayesian mixture over $\theta \in \Theta_m \subseteq \Lambda \subseteq \mathbb{R}^d$. For a given model θ, the probability of a label for example x is given by $p\left(y|x,\theta\right) \stackrel{\triangle}{=} \frac{1}{1+\exp(-y\cdot x^T\theta)}$. The loss at t for model θ is $\ell(\theta, x_t, y_t) \stackrel{\triangle}{=} -\log p(y_t|x_t,\theta) = \log\left(1+\exp(-y_t\cdot x_t^T\theta)\right)$, where it will sometimes be convenient to use the dot product $z \stackrel{\triangle}{=} x^T\theta$. Similarly, the loss of $\mathcal{A}$ at t is $\ell(\mathcal{A}, x_t, y_t) \stackrel{\triangle}{=} -\log[\mathcal{A}(S_{t-1}, x_t, y_t)]$. The total loss for model θ on sequence S_T is $L(\theta, S_T) \stackrel{\triangle}{=} \sum_{t=1}^T \ell(\theta, x_t, y_t)$. Similarly, $L(\mathcal{A}, S_T) \stackrel{\triangle}{=} \sum_{t=1}^T \ell(\mathcal{A}, x_t, y_t)$.

The *regret* of $\mathcal{A}$ for a given example/label pair sequence S_T relative to a comparator model $\theta^* \in \Theta \stackrel{\triangle}{=} \{\theta : \|\theta\|_\rho \leq B\}$, where B constrains the norm of θ^*, is defined as

$$\text{Regret}\left(\mathcal{A}, S_T, \theta^*\right) \stackrel{\triangle}{=} L\left(\mathcal{A}, S_T\right) - L\left(\theta^*, S_T\right). \tag{2}$$

We limit the comparator such that $\theta^* \in \Theta$, and consider the different cases where $\rho \in \{1, 2, \infty\}$. It is reasonable to assume that $B = \gamma \log T$ for some $\gamma > 0$. *Bayesian mixture* algorithms could have

support in $\Theta_m \subseteq \Lambda$ where $\Lambda \supseteq \Theta$. The regret of $\mathcal{A}$ relative to the best comparator is given by

$$\text{Regret}\left(\mathcal{A}, S_T\right) \triangleq \sup_{\theta^* \in \Theta} \text{Regret}\left(\mathcal{A}, S_T, \theta^*\right). \tag{3}$$

A mixture algorithm $\mathcal{A}$ that may rely on the values of x_t in its predictions of y_t, predicts

$$p_{\mathcal{A}}(y^t|x^t) \triangleq \int_{\theta \in \Theta_m} p\left(y^t|x^t,\theta\right) \cdot p_0(\theta) d\theta = \int_{\theta \in \Theta_m} \prod_{\tau=1}^{t} p\left(y_\tau|x_\tau,\theta\right) \cdot p_0(\theta) d\theta \tag{4}$$

where $p_0(\theta) \triangleq p_{0,\mathcal{A}}(\theta)$ is some initial *prior* for $\mathcal{A}$ on the distribution of the parameter vector θ, and $\Theta_m \subseteq \Lambda$ is the support of the mixture, which may be different form Θ. The probability (4) assigned to y^t can also be expressed as a set of equations that sequentially update a *posterior* distribution over θ at round t from the prior at t, which is the posterior at $t-1$, i.e.,

$$p_{\mathcal{A}}(\theta|S_t) = \frac{\prod_{\tau=1}^{t} p\left(y_\tau|x_\tau,\theta\right) \cdot p_0(\theta)}{\int_\theta \prod_{\tau=1}^{t} p\left(y_\tau|x_\tau,\theta\right) \cdot p_0(\theta) d\theta} \triangleq \frac{p_{\mathcal{A}}(\theta, y^t|x^t)}{p_{\mathcal{A}}(y^t|x^t)}. \tag{5}$$

The prediction of y_t is then given by

$$p_{\mathcal{A}}(y_t|x_t, S_{t-1}) = \int_\theta p(y_t|x_t,\theta) \cdot p_{\mathcal{A}}(\theta|S_{t-1}) d\theta. \tag{6}$$

As seen in (6), the prediction is also conditioned on the feature values (example) vector x_t. The prior distribution $p_0(\theta)$ of $\mathcal{A}$ is shown to be continuous in (4)-(6). However, Θ_m can be set to be a discrete set, and then (4) can be re-rewritten as

$$p_{\mathcal{A}}(y^t|x^t) \triangleq \sum_{\theta \in \Theta_m} p\left(y^t|x^t,\theta\right) \cdot p_0(\theta) = \sum_{\theta \in \Theta_m} \prod_{\tau=1}^{t} p\left(y_\tau|x_\tau,\theta\right) \cdot p_0(\theta). \tag{7}$$

5. Useful Methods

5.1. Lower Bounds on Regret - The Redundancy Capacity Theorem

A lower bound on regret is meaningful only when stated in terms of existence of a sequence S_T for every possible algorithm, for which the regret is at least the lower bound. Davisson (1973) formulated such a notion connecting universal compression redundancy with problems like hypothesis testing by the *redundancy-capacity theorem*, which shows that the redundancy (or regret) can be lower bounded by the mutual information $I(\Theta; \mathcal{S}_T)$ between the parameter and the observed data sequence, induced by the prior on Θ of a mixture model. A specific interesting case is when the prior is uniform on a discrete subset $\Theta_m \subseteq \Theta$ of the parameter space, and the elements in Θ_m are *distinguishable* by the observation $\mathcal{S}_T$, i.e., observing S_T is sufficient to determine which $\theta \in \Theta_m$ generated $\mathcal{S}_T$ with error probability $P_e \to 0$ as $T \to \infty$. This case leads to a weaker lower bound than the bounds described in Davisson (1973) and subsequent works, but is sufficient for showing redundancy bounds in many cases (see, e.g., Merhav and Feder (1995); Shamir (2006a)), and also for regret bounds for our problem. We frame this result to regret, and prove it by mirroring the part of the derivation in Davisson (1973) that is sufficient for the result we need, but described in terms that apply to the regret problem. We next state the theorem, which is proved in Appendix A.

Theorem 1 *Distinguishable Grid Regret (adapted from Davisson (1973)): Let $T \to \infty$. Let $\Theta_m \subseteq \Theta$ be a set of $M \to \infty$ distinct values of θ. Draw $\Phi \in \Theta_m$ with a uniform prior, and generate $\mathcal{S}_T$ from the distribution determined by Φ. Let $\hat{\Phi} \stackrel{\triangle}{=} f(\mathcal{S}_T)$ be some estimator of $\Phi \in \Theta_m$ from the observed $\mathcal{S}_T$. Then, if $P_e \stackrel{\triangle}{=} P(\hat{\Phi} \neq \Phi) \to 0$, the regret of any algorithm $\mathcal{A}$ for the worst sequence S_T is lower bounded by*

$$\sup_{S_T} \textit{Regret}\,(\mathcal{A}, S_T) \geq (1 - o(1)) \log M. \tag{8}$$

Similarly, for a fixed x^{*T}, if we draw Y^T instead of S_T and the conditions above hold, (8) also holds.

5.2. Variational Approach for Upper Bounds

Upper bounds can be obtained by showing Bayesian methods that can achieve low regret and bounding their regret. For simplicity, one can select priors $p_0(\cdot)$ with a diagonal covariance. Kakade and Ng (2005) selected a normal prior, whereas Foster et al. (2018) used a uniform one. We follow Kakade and Ng (2005) and manipulate their approach to obtain tighter bounds for L_2 and L_∞, and then use the method of grids with a uniform discrete prior combined with their method to derive an L_1 bound. We first describe their approach. Define a distribution $Q(\theta)$ on $\Theta_q \subseteq \Theta_m$ where $E_q(\theta) = \theta^*$, and $E_q[(\theta_i - \theta_i^*)(\theta_j - \theta_j^*)] \leq \eta_q^2 \cdot \delta(i-j)$, where $\delta(n) = 1$ if $n = 0$ and is 0, otherwise, i.e, diagonal covariance matrix, where η_q^2 is an upper bound on elements of the diagonal. (Note that Θ_q can be a subset of Θ, but does not have to, and in fact, is not for the normal prior). Let $D(Q||p_0)$ be the KL-divergence between Q and p_0. Then the following theorem holds.

Theorem 2 *Kakade and Ng (2005): The regret of a Bayesian algorithm $\mathcal{A}^*$ with prior p_0 for sequence S_T and comparator θ^* is upper bounded by*

$$\textit{Regret}\,(\mathcal{A}^*, S_T, \theta^*) \leq D(Q||p_0) + \frac{dT}{8}\eta_q^2 \tag{9}$$

The proof of Theorem 2 is in Kakade and Ng (2005), but needs to be modified a bit because they restricted $\|x_t\|_2 \leq 1$, while we assume $|x_{t,i}| \leq 1$ ($\|x_t\|_\infty \leq 1$). We rely on their proof, except where it needs to be modified. The proof is in Appendix B.

6. Regret Lower Bounds

We now use Theorem 1 to derive lower bounds for the binary label case. A lower bound for the multi label case is given in Appendix E. We first define

$$\gamma \stackrel{\triangle}{=} \min\left\{\frac{B}{\log T}, \{\alpha : \alpha \cdot \min[T^{1-\varepsilon}, d] = T^{1-\varepsilon}\}\right\} \tag{10}$$

as the effective count of $\log T$ units in B (where if d is very large, we will only consider a clipped portion of Θ for Θ_m to ensure distinguishability. This will guarantee that $\gamma \min[T^{1-\varepsilon}, d] = o(T)$). The lower bounds are stated in the following theorem.

Theorem 3 *Fix an arbitrary $\varepsilon > 0$, let $T \to \infty$. Then, for every algorithm $\mathcal{A}$ there exists a sequence S_T, for which the regret is lower bounded by*

$$\textit{Regret}(\mathcal{A}, S_T) \geq \begin{cases} (1-o(1))\frac{d}{2}\log\frac{T}{d}; & \textit{for } d = O(1), \\ (1-o(1))\frac{d}{2}\log\frac{4\gamma T}{d}; & \textit{for } \|\theta^*\|_\infty \leq B \textit{ and } d < \frac{4}{e}\gamma T^{1-\varepsilon}, \\ (1-o(1))\frac{2}{e}\gamma T^{1-\varepsilon}; & \textit{for } \|\theta^*\|_\infty \leq B \textit{ and } d \geq \frac{4}{e}\gamma T^{1-\varepsilon}, \\ (1-o(1))\frac{d}{2}\log\frac{2\pi e\gamma T}{d^2}; & \textit{for } \|\theta^*\|_2 \leq B \textit{ and } d < \sqrt{\frac{2\pi}{e}\gamma T^{1-\varepsilon}}, \\ (1-o(1))\sqrt{\frac{2\pi}{e}\gamma T^{1-\varepsilon}}; & \textit{for } \|\theta^*\|_2 \leq B \textit{ and } d \geq \sqrt{\frac{2\pi}{e}\gamma T^{1-\varepsilon}}, \\ (1-o(1))\frac{d}{2}\log\frac{4e^2\gamma T}{d^3}; & \textit{for } \|\theta^*\|_1 \leq B \textit{ and } d < \left(\frac{4\gamma T^{1-\varepsilon}}{e}\right)^{1/3}, \\ (1-o(1))\frac{3}{2}\left(\frac{4\gamma T^{1-\varepsilon}}{e}\right)^{1/3}; & \textit{for } \|\theta^*\|_1 \leq B \textit{ and } d \geq \left(\frac{4\gamma T^{1-\varepsilon}}{e}\right)^{1/3}. \end{cases} \tag{11}$$

Theorem 3 shows that for small d each feature/dimension contributes $0.5\log(T/d)$ to the worst case regret. Generally, for L_∞ each parameter costs $0.5\log(\gamma T/d)$. For L_2 there is a factor d reduction inside the logarithm. An additional similar reduction is observed between L_2 and L_1. These relations are expected, because they reflect the logarithm of the ratio between the respective volumes of the parameter spaces, dictated by the constraints. The greater the volume, the harder the algorithm has to work to match the best comparator, and the larger the regret penalty it pays. This is similar to observations in the information theory literature and in results as Rakhlin and Sridharan (2015), which tie the regret to the richness of the class. The dependence on B is through γ. Each interval of $\log T$ consists of roughly $\sqrt{T/d}$ distinguishable parameters. Hence, the ratio between B and $\log T$ dictates how many parameter regions are in an interval of diameter $2B$. Thus this ratio, represented by γ, dominates the effect of B, which is normally, in practice, negligible relative to the effects of T and d. (In practice, we would normally limit θ to some reasonable range, which is usually $O(1)$. However, theoretically, γ can be larger, in which case it does influence the bound.) If B is too large, the effective γ in (10) guarantees that $\gamma d = o(T)$, and if $d = \Omega(T)$, it guarantees this with respect to the largest value $T^{1-\varepsilon}$ used for the bound.

An interesting behavior is observed for all cases L_∞, L_2 and L_1. When we reach $d = O(T)$, $d = O(\sqrt{T})$ and $d = O(T^{1/3})$, respectively, a threshold phenomenon happens, and the bound becomes constant for every greater d. It is not clear how much of this is a result of the bounding techniques and how much is real. However, as we see in the upper bounds for L_1 in the following section, there exist Bayesian algorithms that achieve $o(d)$ regret for L_1 constraints. We also observe a decrease in rate in the upper bounds for L_2 at $d = O(T)$. Together, these results imply, that there are, in fact, cases in which the regret does not grow linearly with d for large enough d. The L_∞ bounds demonstrate that there are situations in which we are guaranteed regret rates of $\Omega(\sqrt{T}\log T)$. In fact, the regret could even be linear with d up to $d = T^{1-\varepsilon} = o(T)$.

To prove the first region of Theorem 3, we partition x^{*T} into d separate equal length segments, where in each segment only one component $x^*_{t,i}; i = 1, 2, \ldots, d$; of x^*_t is 1 and the rest are 0. This transforms the problem to a standard universal compression problem in d different segments, in each a single parameter is to be estimated. In each segment, we now have a grid of $\sqrt{T/d}^{1-\varepsilon}$ points, which are spaced (in the space of label probability they induce) at a distance $\sqrt{d/T}^{1-\varepsilon}$ from one

another. The total grid Ψ is the power set of the individual grids over the segments. Large deviation typical sets analysis (see Cover and Thomas (2006)) with the union bound over the segments is used to show that each of these points is distinguishable from the others. Finally, applying Theorem 1 with a fixed x^{*T} gives the lower bound. For diminishing large deviation exponent to dominate over the union bound, we need to use $d = o(T)$. For larger values of d, we use a grid that varies only in the first $T^{1-\varepsilon}$ components of θ, and apply the resulting bound.

For the remaining regions, we fix the first component of the parameter at the maximal point $\theta_1 = B$. Then, $x_{t,1}$ would scale it by factors of s/γ, where s is an integer, taking all values from $-\gamma$ to γ. This will induce $2\gamma + 1$ priors, that make $(2\gamma + 1)$ distinct distinguishable regions of a second nonzero component θ_i of θ that occurs in the same examples as θ_1. We partition each of now $d - 1$ segments, where in each of these segments a different component $x^*_{t,i}$ is 1 for $i > 1$, while the remaining ones are 0, to $(2\gamma + 1)$ subsegments corresponding to the different values of $x_{t,1}$. We show that the points on the grid, now constructed as a power set of $d - 1$ grids of $(2\gamma + 1)\sqrt{T/d}\gamma^{1-\varepsilon}$ points, are, again, distinguishable with the fixed x^{*T}. Using Theorem 1, the logarithm of the cardinality of the power grid lower bounds the regret. However, for L_2 and L_1, only the components of $\theta \in \Psi$ which satisfy the constraints are included in the grid. This reduces the bounds, and leads to a threshold point, in which the lower bounds become useless for the value of d, since the remaining space no longer contains parameters for which all components of θ are nonzero. We can thus use the value of d, which is lower, but achieves the largest bound. This leads to regions 3, 5, and 7 in the bound. The proof of Theorem 3 is presented in Appendix C.

7. Regret Upper Bounds for Bayesian Methods

Theorem 2 allows us to prove the following two theorems:

Theorem 4 *There exist Bayesian algorithms $\mathcal{A}^*$ that for every sequence S_T and comparator $\theta^* \in \Theta$, achieve regret*

$$\textit{Regret}(\mathcal{A}^*, S_T, \theta^*) \le \begin{cases} (1+o(1)) \cdot \frac{d}{2} \log\left(\frac{B^2 T e}{4} + e\right); & \textit{for } \|\theta^*\|_\infty \le B, \\ (1+o(1)) \cdot \frac{d}{2} \log\left(\frac{B^2 T e}{4d} + e\right); & \textit{for } \|\theta^*\|_2 \le B, \\ (1+o(1)) \cdot \frac{d}{2} \log\left(\frac{B^2 T e^3}{4d^2}\right); & \textit{for } \|\theta^*\|_1 \le B \textit{ and } d = o(B\sqrt{T}), \\ (1+o(1)) \cdot \left(\frac{d}{2} \log(4e) + \frac{B\sqrt{T}}{2}\right); & \textit{for } \|\theta^*\|_1 \le B \textit{ and } d = \Theta(B\sqrt{T}), \\ (1+o(1)) \cdot \left(\frac{d}{2} + \sqrt{2dB\sqrt{T}}\right); & \textit{for } \|\theta^*\|_1 \le B \textit{ and } d = \Omega(B\sqrt{T}) \end{cases} \tag{12}$$

Theorem 5 *Let $d = \omega(B\sqrt{T})$. Then, there exists a Bayesian algorithm $\mathcal{A}^*$ that for every sequence S_T and comparator $\theta^* \in \Theta$ with $\|\theta^*\|_1 \le B$, achieves regret*

$$\textit{Regret}(\mathcal{A}^*, S_T, \theta^*) = O\left(T^{1/5} d^{3/5} B^{2/5}\right) = o(d). \tag{13}$$

Theorem 4 shows that regret logarithmic with T and linear with d is achievable in all three cases. The bounds asymptotically differ only by a factor of d inside the logarithm. The wider

allowable range of θ^* gives an upper bound where T in the logarithmic term is not normalized by d. The smaller comparator region, where θ^* norms are restricted by L_2, reduces the logarithmic cost from $\log T$ to $\log(T/d)$. A similar reduction is achieved from L_2 to L_1. Both L_2 and L_1 have interesting threshold behavior, which matches the behavior with the lower bounds. For L_1, as long as $d = o(B\sqrt{T})$, we observe regret linear in d and logarithmic in T. For larger dimensions, though, we observe only linear behavior in d, without the logarithmic terms. Furthermore, if we tighten the bounds further, Theorem 5 shows that even sub-linear behavior in d is possible ($o(d)$, which as long as $d = o(B^6T^3)$ is $o(B\sqrt{dT})$). For L_2, transition from $O(d \log T)$ to $O(d)$ occurs at $d = O(B^2T)$.

The bounds in Theorem 4 are derived for our setting of $\|x_t\|_\infty \leq 1$. Kakade and Ng (2005) and others considered the setting where $\|x_t\|_2 \leq 1$. In their setting, all the bounds in (12) will have an additional factor of d in the denominator of the logarithmic term. This means that the transition from $O(d \log T)$ to $O(d)$ rate (and for L_1 to $o(d)$) will now occur at $d = O(T^{1/3})$, $d = O(\sqrt{T})$ and $d = O(T)$ for L_1, L_2, and L_∞, respectively. (Such a transition will now also happen for L_∞.)

Unlike Theorem 3, B is present in the upper bounds instead of γ. For distinguishability on an individual feature, each region of $\log T$ in a single dimension consists of only $\sqrt{T}$ distinguishable points, and not $\sqrt{T} \log T$. However, when dimensions are mixed through the dot product, it is hard to disentangle the dimensions. This leads to the difference between the lower and the upper bounds.

The proof of Theorems 4 and 5 is based on Theorem 2. For L_∞ and L_2, we use a Gaussian prior with a Gaussian $Q(\cdot)$. We derive the bounds on the terms of (9), find the value of the parameter that gives the smallest bound and apply it. We then find the variance of the prior that gives the tightest bound, and apply it. For L_1, we construct a grid whose points are assigned a uniform probability. We construct $Q(\cdot)$ as a Bernoulli distribution in each dimension, giving nonzero probability only to the surrounding neighbors of the ith component θ^*_i of θ^*, but ensuring that θ^* is the expectation of $Q(\cdot)$. Then, in the same manner, we upper bound the terms of (9), and optimize the free parameter for the tightest bound. Finally, since a threshold occurs, where the bound becomes useless, we use a union bound on lower dimensions of the parameter space. We find the dimension that gives the maximal element of the sum over all dimensions, and use it to upper bound all dimensions. Applying this method more tightly gives some tedious algebra, but yields a bound of $o(d)$ for the upper region of the L_1 constraint problem. The proof of both theorems is presented in Appendix D.

8. Conclusions

We studied logistic regression regret, and derived lower and upper bounds for settings constraining the norm of a comparator. We presented a comprehensive characterization of the regret for the different settings, including the asymptotic behavior in the dimensionality. Adapting a methodology from the universal compression literature, we derived lower bounds on the regret, showing initial logarithmic in T, linear in d regret, with rates whose growth slows with larger dimensions of the feature space. Matching upper bounds confirm the general behavior of the lower bounds. Specifically, we demonstrated that under L_1 constraints, for large enough d, regret becomes sub-linear in d, and for L_2 constraints, it drops from linear in d and logarithmic in T to just linear in d. On the negative side, under L_∞ constraints, regrets of $\Omega(\sqrt{T} \log T)$ are guaranteed for $d = \Omega(\sqrt{T})$.

Acknowledgments

The author acknowledges Matt Streeter for numerous invaluable discussions and insights. The author also acknowledges Wojciech Szpankowski and Yury Polyanskiy for helpful discussions, and the reviewers for their insightful and helpful comments that helped improve the paper.

References

Jacob Abernethy, Peter L Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. 2008.

K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Advances in neural information processing systems*, pages 773–781, 2013.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Nicoló Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. In *Advances in neural information processing systems*, pages 359–366, 2002.

T. M. Cover and J. A. Thomas. *Elements of Information Theory—*. John Wiley & Sons, second edition, 2006.

L. D. Davisson. Universal noiseless coding. *IEEE Trans. Inf. Theory*, IT-19(6):783–795, Nov. 1973.

M. Drmota and W. Szpankowski. Precise minimax redundancy and regrets. *IEEE Trans. Inf. Theory*, IT-50:2686–2707, 2004.

Dylan J Foster, Satyen Kale, Haipeng Luo andMehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *COLT - Conference on Learning Theory*, 2018.

E. Hazan. The convex optimization approach to regret minimization. In *S. Sra, S. Nowozin, and S. Wright, editors, Optimization for Machine Learning*, pages 287–303. MIT press, 2012.

E. Hazan, T. Koren, and K. Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *The 27th Conference on Learning Theory, COLT 2014*, page 197209. MIT press, 2014.

Sham M Kakade and Andrew Y. Ng. Online bounds for bayesian algorithms. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 641–648. MIT Press, 2005. URL http://papers.nips.cc/paper/2637-online-bounds-for-bayesian-algorithms.pdf.

R. E. Krichevsky and V. K. Trofimov. he performance of universal encoding. *IEEE Trans. Inform. Theory*, IT-27(2):199?207, Mar. 1981.

Nick Littlestone. From on-line to batch learning. In *Proceedings of the Second Annual Workshop on Computational Learning Theory*, COLT ?89, page 269?284, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc. ISBN 1558600868.

H. B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *AISTATS*, 2011.

H. B. McMahan. Analysis techniques for adaptive online learning. *CoRR*, abs/1403.3465, 2014. URL http://arxiv.org/abs/1403.3465.

H. B. McMahan and M. J. Streeter. Open problem: Better bounds for online logistic regression. In *Journal of Machine Learning Research-Proceedings Track, 23*, 2012.

H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. Mar Hrafnkelsson, T. Boulos, and J. Kubica. Ad click prediction: A view from the trenches. In *KDD*, 2013.

H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017.

N. Merhav and M. Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Trans. Inf. Theory*, 41(3):714–722, May. 1995.

A. Orlitsky and N. P. Santhanam. Speaking of infinity. *IEEE Trans. Inf. Theory*, 50(10):2215–2230, Oct. 2004.

A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 2005.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression with general loss functions. *arXiv preprint arXiv:1501.06598*, 2015.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1984–1992. Curran Associates, Inc., 2010a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *arXiv preprint arXiv:1006.1138*, 2010b.

Alexander Rakhlin, Karthik Sridharan, Alexandre B Tsybakov, et al. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, IT-30(4):629–636, Jul. 1984.

R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1997.

S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 2012.

G. I. Shamir. On the MDL principle for i.i.d. sources with large alphabets. *IEEE Trans. Inform. Theory*, 52(5):1939–1955, May 2006a.

G. I. Shamir. Universal lossless compression with unknown alphabets - the average case. *IEEE Trans. Inform. Theory*, 52(11):4915–4944, Nov. 2006b.

Szpankowski and M. J. Weinberger. Minimax pointwise redundancy for memoryless models over large alphabets. *IEEE Trans. Inform. Theory*, 58(7):4094–4104, Jul. 2012.

V. Vovk. Aggregating strategies. In *Third Annual Workshop on Computational Learning Theory*, pages 371–383, 1990.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

Appendix A. Proof of Theorem 1

The following lemma is needed to prove Theorem 1.

Lemma 6 *Let $\theta^* \in \Theta_m$ be a parameter in the support of a Bayesian algorithm $\mathcal{A} = \mathcal{A}^*$ that predicts as described in (4)-(6) or in (7). Then,*

$$\textit{Regret}(\mathcal{A}^*, S_T, \theta^*) = \log p_{\mathcal{A}^*}(\theta^*|S_T) - \log p_0(\theta^*) \tag{14}$$

where $p_{\mathcal{A}^*}(\theta^*|S_T)$ is the posterior that $\mathcal{A}^*$ assigns to θ^* in (5) at $t = T$.

Proof By the definition in (2), the regret of $\mathcal{A}^*$ over label sequence y^T is the difference between the log-likelihood assigned to the complete label sequence y^T by the comparator θ^* and that assigned to y^T by $\mathcal{A}^*$

$$\begin{aligned}\text{Regret}(\mathcal{A}^*, S_T, \theta^*) &= \log p(y^T|x^T, \theta^*) - \log p_{\mathcal{A}^*}(y^T|x^T) \\ &= \log \frac{p(y^T|x^T, \theta^*)p_0(\theta^*)}{p_{\mathcal{A}^*}(y^T|x^T)p_0(\theta^*)} = \log p_{\mathcal{A}^*}(\theta^*|S_T) - \log p_0(\theta^*) \qquad (15)\end{aligned}$$

where $p_{\mathcal{A}^*}(y^T|x^T)$ is defined in (4) for $\mathcal{A} = \mathcal{A}^*$ and $t = T$. The equalities are obtained by multiplication and division by $p_0(\theta^*)$, which by the conditions of the lemma ($\theta^* \in \Theta_m$) is greater than 0, and by identifying the posterior of $\mathcal{A}^*$ defined in (5) for $\mathcal{A} = \mathcal{A}^*$, $\theta = \theta^*$, and $t = T$. ■

Proof of Theorem 1: Let $\mathcal{A}^*$ be a Bayesian algorithm as defined in (7) on a discrete Θ_m. Let $p_0(\theta)$ now be uniform with $p_0(\theta) = 1/M, \forall\theta \in \Theta_m$, and $\|\Theta_m\| = M$ (where M can be a function of T). Now,

$$\begin{aligned}\sup_{S_T} \text{Regret}\left(\mathcal{A}, S_T\right) &= \sup_{S_T}\sup_{\theta^*\in\Theta} \text{Regret}\left(\mathcal{A}, S_T, \theta^*\right) \overset{(a)}{\geq} \sup_{\theta^*\in\Theta_m}\sup_{S_T}\left[L(\mathcal{A}, S_T) - L(\theta^*, S_T)\right] \\ &\overset{(b)}{\geq} \sup_{\theta^*\in\Theta_m} E_{\mathcal{S}_T|\theta^*}\left[\log p(Y^T|X^T, \theta^*) - \log \mathcal{A}(S_T, X^T, Y^T)\right] \\ &\overset{(c)}{\geq} E\log p(Y^T|X^T, \Phi) - E_{\mathcal{S}_T}E_{\Theta_m|\mathcal{S}_T}\log p_{\mathcal{A}^*}(Y^T|X^T, \mathcal{S}_T) \\ &= E[\text{Regret}(\mathcal{A}^*, \mathcal{S}_T, \Phi)] \overset{(d)}{=} E\log p_{\mathcal{A}^*}(\Phi|\mathcal{S}_t) - E\log p_0(\Phi) \qquad (16)\end{aligned}$$

Step (a) follows for exchanging order of supremums, and shrinking Θ to Θ_m. Substituting loss definitions, and lower bounding the supremum on S_T by an expectation over S_T conditioned on θ^* leads to (b). Step (c) follows from lower bounding the supremum on Θ_m by expectation of Θ_m w.r.t. $p_0(\theta)$. This yields expectation w.r.t. θ^* and S_T for the left term. Performing this expectation on the right term implies expectation on Y^T with a distribution that is the one assigned to Y^T by $\mathcal{A}^*$. This negative logarithm is minimized with predictions given from $\mathcal{A}^*$, leading to the right term in step (c), which, similarly to the left term, performs the expectation on both $\mathcal{S}_T$ and Θ_m. The resulting expression is the expectation of the regret of $\mathcal{A}^*$ on Θ_m. Applying Lemma 6, gives (d). By the uniform construction of $p_0(\cdot)$, the right term is $\log M$. The left term can be bounded using P_e (see, e.g., Fano's inequality, Cover and Thomas (2006))

$$-E\log p_{\mathcal{A}^*}(\Phi|\mathcal{S}_t) \leq h_2(P_e) + P_e\log(M-1) \leq 1 + P_e\cdot\log M, \qquad (17)$$

where $h_2(p) \overset{\triangle}{=} -p\log p - (1-p)\log(1-p)$ is the binary entropy function. This is proved by expectation over Φ, and then, breaking the events of the value of $\hat{\Phi}$ into the event $\hat{\Phi} = \Phi$ and $\hat{\Phi} \neq \Phi$, and then hierarchically separating the latter into the $M-1$ different possible values of $\hat{\Phi}$, upper bounding the conditional entropy on $\hat{\Phi} \neq \Phi$ by that of a uniform distribution. Combining both terms of (16), given $M \to \infty$ and $P_e \to 0$, concludes the proof of the first statement of Theroem 1.

The second statement follows the exact same derivation conditioned on a fixed x^{*T}, after lower bounding the supremum over $\mathcal{S}_T$ by that for this fixed x^{*T}. ■

Appendix B. Proof of Theorem 2

Proof of Theorem 2: Let $L(Q, S_T) \triangleq \int_{\theta \in \Theta_q} L(\theta, S_T) Q(\theta) d\theta$. (Similarly, if Θ_q is discrete, the integral is replaced by a sum). Then, for a Bayesian algorithm $\mathcal{A}^*$ with prior distribution $p_0(\cdot)$ on a logistic regression model,

$$\text{Regret}\left(\mathcal{A}^*, S_T, \theta^*\right) = \underbrace{L\left(\mathcal{A}^*, S_T\right) - L\left(Q, S_T\right)}_{\leq D(Q||p_0)} + \underbrace{L\left(Q, S_T\right) - L\left(\theta^*, S_T\right)}_{\leq d \cdot T \cdot \eta_q^2/8} \tag{18}$$

The first term is bounded in Lemma 2.1 in Kakade and Ng (2005). For the second term, recall the dot product $z_t \triangleq x_t^T \theta$. Then, for some θ, round t, and example/label pair $\{x_t, y_t\}$, define $f(z) \triangleq \ell(\theta, x_t, y_t)$ as the per example/label loss, which can be expressed as just a function of z. Then, using Taylor expansion,

$$E_q[f(z)] - f(z^*) = f'(z^*) \cdot 0 + E_q\left[f''(\xi(z)) \frac{(z - z^*)^2}{2}\right] \leq \frac{d \cdot \eta_q^2}{8} \tag{19}$$

where the first term is 0 by definition of the expectation w.r.t. $Q(\cdot)$. Then $\xi(z)$ is some point between z^* and z for which equality is satisfied for the second order Taylor approximation. The second derivative of logarithmic loss w.r.t. the dot product for a diagonal term is $x_{t,i}^2 p(1-p)$ for some probability p, and thus upper bounded by $1/4$. By construction of a diagonal covariance matrix, $\sigma_z^2 \leq \eta_q^2 x^T x \leq d\eta_q^2$. Since this bound is added on T examples, this term is multiplied by T. This concludes the proof. ■

Appendix C. Proof of Theorem 3:

Proof of Theorem 3: We construct grids $\Psi = \Theta_m$ of points ψ_i for all dimensions i in the parameter space, fix x^{*T}, show that the points are distinguishable given x^{*T}, and use the version of Theorem 1 for a fixed x^{*T} to lower bound the regret by the logarithm of the cardinality of the grid Ψ. For the first region, partition x^{*T} into d separate length T/d segments. For segment $i, i = 1, 2, \ldots, d$; $x_{t,i}^* = 1$ and all $x_{t,j} = 0$ for $j \neq i$. The grid Ψ is a power set of grids for dimension i, for all i. Define

$$p_{i,j} \triangleq \frac{1}{1 + \exp(-\psi_{i,j})}; i = 1, 2, \ldots, d; j = 0, 1, \ldots, \lfloor \sqrt{T/d}^{1-\varepsilon} \rfloor \triangleq k \tag{20}$$

for some fixed $\varepsilon > 0$. Then, the elements of $\psi_{i,j}$ are defined such that

$$p_{i,j+1} - p_{i,j} = \sqrt{\frac{d}{T}}^{1-\varepsilon} \triangleq \delta; \forall i, j \tag{21}$$

with $p_{i,0} = 0$. In this step we can use all $k+1$ values of $\psi_{i,j}$. For the subsequent regions of the bound, we omit the first and the last i ($i = 0$ and $i = k$). Note that we can similarly define the grid to be uniformly spaced w.r.t. ψ_j, where spacing is $\delta \cdot (\log T)$. This makes the distinguishability a little more tedious to prove, but does preserve a uniform grid (which will be necessary for L_2 and L_1 results). For simplicity, and without loss of generality, we will show distinguishability with the current definition.

The setup above transformed the problem to a standard well-known universal compression problem Rissanen (1984), but for d different segments, where in each a single parameter is to be estimated. The cardinality of the grid is $M = k^d$. Distinguishability is proved using the union bound on the d segments applying large deviations typical sets analysis (see Cover and Thomas (2006)). We skip this step, and perform it for the next regions. The method we use for the next regions also applies here. Applying Theorem 1 with a fixed x^{*T} gives the lower bound for the first region, taking the logarithm of M.

We now consider larger $B = \gamma \log T$, and the behavior with asymptotically larger d. For simplicity, we assume that γ satisfies (10). (If it does not, we lower the value of B for which the analysis is done.) If $d > T^{1-\varepsilon}$, the analysis will assume $d = T^{1-\varepsilon}$, and the resulting bound will be also applied for large d. In single dimension, for a sequence of length T, we can only distinguish between parameters in $[-0.5 \log T, 0.5 \log T]$. Any parameter greater than $0.5 \log T$ will appear with non-diminishing probability the same as the $0.5 \log T$ parameter. Similarly, parameters smaller than $-0.5 \log T$ will not be distinguishable from $-0.5 \log T$. Therefore, we cannot use the method described for smaller B to enhance the bound for the larger B. We will thus have to manipulate the wider region to achieve the desired results. The idea is to "sacrifice" one dimension from the parameters to serve as a prior, which for the other parameters maps different regions in $[-B, B]$ to $[-0.5 \log T, 0.5 \log T]$. Each such distinct region will be considered in a separate segment, in which distinguishability will be shown for parameter values in that region. We note that because we actually consider intervals $[-0.5 \log(T/d), 0.5 \log(T/d)]$, instead of a factor γ, we would actually have a factor $\gamma \cdot \kappa$ in the lower bounds, where $\kappa \stackrel{\triangle}{=} (\log T)/\log(T/d) \geq 1$. For simplicity, however, we chose to omit this term from the bounds. We proceed without this sidestep.

Recall that $\psi_{i,j}$ is the jth grid point for the value of parameter $i; 1 \leq i \leq d$. We now omit the extreme points $j = 0$ and $j = k$ to prevent duplications between partitions of the region $[-B, B]$. For $i = 1$, we now only include one grid point $\psi_{1,1} = B$. Instead of d segments as before, we now have $d-1$ segments, one for each of the remaining $d-1$ components of ψ. We further segment each of the $d-1$ segments of x^{*T} defined earlier, each into $2\gamma + 1$ subsegments. (For simplicity, we ignore negligible integer length constraints on γ.) Let $s \in \{-\gamma, -\gamma+1, \ldots, 0, 1, \ldots, \gamma\}$ be a subsegment index. Then, for subsegment s, $x^*_{t,1} = s/\gamma$. We still have for segment i representing dimension i, $x^*_{t,i} = 1$ for one i, and for the remaining components $j \neq 1, j \neq i$, $x^*_{t,j} = 0$. The grids $\psi_{i,j}$ for dimensions $i = 2, 3, \ldots, d$ now consist of a grid which is a union of $2\gamma + 1$ sub-grids. The sth sub-grid of dimension i is on the range $[-0.5 \log T, 0.5 \log T] - s \log T$. By adding the contribution of component 1, which is $B \cdot s/\gamma = s \log T$, the region of sub-grid s in subsegment s is mapped back into $[-0.5 \log T, 0.5 \log T]$. With this mapping in mind, we place the points in the sub-grid s such that they either map into probabilities as in (20), with the contribution of component

$i = 1$, or they are just uniformly spaced in $[-0.5 \log T, 0.5 \log T] - s \log T$. The spacing in each sub-grid must be adjusted (increased) from the case of the first region of the bound to account for the reduction in subsegment length, and is set to

$$\delta = \sqrt{\frac{d\gamma}{T}}^{1-\varepsilon}. \tag{22}$$

The construction described yields a total $(d-1)(2\gamma+1)$ subsegments, each of length $T/[(d-1)(2\gamma+1)]$, with a total of $(2\gamma+1)/\delta$ grid points in each of the $d-1$ original segments. Thus

$$\begin{aligned} \log M &= (d-1)\left[\log(2\gamma+1) + (1-\varepsilon)\frac{1}{2}\log\frac{T}{d\gamma}\right] \\ &\geq (1-o(1))\frac{d}{2}\log\frac{4\gamma T}{d}. \end{aligned} \tag{23}$$

Let S_T be generated by $\theta \in \Psi$, and let $\hat{\theta}$ be estimated from S_T. Define $\delta_e = \hat{\theta} - \theta$. For making an error between two grid points we have to have $\delta_e \geq 0.5\sqrt{d\gamma/T}^{1-\varepsilon}$. Using large deviation typical sets analysis, there are at most $T/[(d-1)(2\gamma+1)]$ different types for an error event per subsegment. Then using the union bound on $2\gamma+1$ subsegments, and then again on $d-1$ segments, we get a multiplier of T. Thus, absorbing lower order terms in $o(1)$, we have

$$\begin{aligned} P_e &\leq (1+o(1)) \cdot T \cdot \exp\left\{-\frac{T}{2d\gamma}\min_{\hat{\theta}\neq\theta} D(P_{\hat{\theta}}||P_\theta)\right\} \\ &\leq (1+o(1)) \cdot \exp\left\{\log T - \frac{1}{4}\left(\frac{T}{\gamma d}\right)^{\varepsilon}\right\} \end{aligned} \tag{24}$$

where the second inequality follows from the relation between the KL divergence and L_1 norm $D(P_{\hat{\theta}}||P_\theta) \geq 2\delta_e^2$ taking the minimum δ_e for an error event. From the definition of γ in (10) and from the assumption made that $d \leq T^{1-\varepsilon}$, we have $\gamma d = o(T)$, which with a fixed $\varepsilon > 0$ yields $P_e = o(1)$. We note that this can also be achieved with diminishing $\varepsilon = O(\log\log T/(\log T))$, with large enough constant. This, together with the bound of (23) and Theorem 1, concludes the proof for the second region. The derivation of the error for the first region is very similar to the one above. The third region is proved by taking d that maximizes the bound in the second region and applying its bound to every greater value of d. (This is the justification for replacing $d > T^{1-\varepsilon}$ earlier by $T^{1-\varepsilon}$.)

It remains to derive the lower bounds for L_2 and L_1. Utilizing a uniform version of the grid for L_∞, and since we already proved distinguishability, this remains as a counting problem, of which portions of Ψ satisfy the L_2 and L_1 constraints. For L_2, using the volume of a d-dimensional ball, we have

$$\begin{aligned} \log M &= (1-o(1)) \cdot \left(\frac{d}{2}\log\pi + d\log\gamma - \frac{d}{2}\log\frac{d}{2e} + \frac{1-\varepsilon}{2}d\log\frac{T}{d\gamma}\right) \\ &\geq (1-o(1))\frac{d}{2}\log\frac{2\pi e\gamma T^{1-\varepsilon}}{d^2}. \end{aligned} \tag{25}$$

This gives the fourth region of the bound.

We observe that as $d = \Theta(\sqrt{T})$ the bound becomes negative. This is because we can no longer fit d-dimensional cubes in the d-dimensional balls. Instead, cubes with lower dimensions can be fit in a lower dimension ball. This implies that we can have grid points that are sparse in the sense of having many 0 components. The set of possible points for large dimensions can include all the combinations of points in lower dimensions. We can thus lower bound the regret by taking the dimension that maximizes the lower bound in (25). The value $d = \sqrt{2\pi\gamma T^{1-\varepsilon}/e}$ maximizes the bound. Plugging it into (25) gives the fifth region of the bound.

Similarly, normalizing the size of the grid by $d!$ gives a lower bound on M for L_1

$$\log M \geq (1 - o(1))\frac{d}{2}\log\frac{4e^2\gamma T^{1-\varepsilon}}{d^{3-\varepsilon}}. \tag{26}$$

This gives the sixth region of the bound. Again, a similar issue as for L_2 occurs, but now at $d = O(T^{1/3})$. Optimizing, again, for a lower value of d, gives an optimizer at $d = \left(4\gamma T^{1-\varepsilon}/e\right)^{1/3}$. Plugging this value to (26) gives the last region of the bound, thus concluding the proof. ■

Appendix D. Proof of Theorems 4 and 5

Proof of Theorems 4 and 5: We apply Theorem 2. For L_∞ and L_2, we extend Kakade and Ng (2005), using Gaussian distributions with diagonal covariance matrices for both the prior p_0 and Q. For L_1, the Gaussian distributions cannot work, and we use a uniform prior p_0 on a grid with Q with diagonal covariance. (This also works for the other two cases, and gives identical bound for L_∞ but a weaker bound for L_2.) The first term in (9) dominates for L_∞, and the first regions of L_1 and L_2, but the second term for the second regions of L_1 and L_2.

Let $p_0 \triangleq \mathcal{N}(\theta; 0, \nu^2 I_d)$, 0-mean normal with diagonal covariance with ν^2 variance. Let $Q(\theta) \triangleq \mathcal{N}(\theta; \theta^*, \varepsilon^2 I_d)$ be a normal distribution with θ^* mean and diagonal covariance with variances ε^2. Then, as Kakade and Ng (2005) showed

$$D(Q||p_0) = d\log\nu + \frac{1}{2\nu^2}\left(\|\theta^*\|_2^2 + d\varepsilon^2\right) - \frac{d}{2} - d\log\varepsilon. \tag{27}$$

By definition of Q, the second term in (9) is $Td\varepsilon^2/8$ (where $\eta_q^2 = \varepsilon^2$). Combining the terms, minimizing for ε, we have $\varepsilon^2 = 4\nu^2/(4 + T\nu^2)$. (This is slightly different from Kakade and Ng (2005), because the d term is omitted due to the different constraints on the norm of x_t.) Plugging ε,

$$\text{Regret}(\mathcal{A}^*, S_T, \theta^*) \leq \frac{1}{2\nu^2}\|\theta^2\|_2^2 + \frac{d}{2}\log\left(1 + \frac{T\nu^2}{4}\right). \tag{28}$$

Extending Kakade and Ng (2005), we now consider what variance of p_0 would give the smallest bound. This is achieved with $\nu^2 = \|\theta^*\|_2^2/d$, which implies that we must have large variance on p_0 so that the cost of the prior does not dominate the bound. This may be unexpected, but in order to encapsulate the whole allowed range Θ, it is reasonable that the prior has variance, large enough,

to include the far ends. For the worst case with L_∞, $\|\theta^*\|_2^2 = dB^2$, giving $\nu^2 = B^2$. For L_2, $\|\theta^*\|_2^2 = B^2$, giving $\nu^2 = B^2/d$. In both cases, the first term of the bound becomes $d/2$. Plugging ν^2 to the second term as well, combining both terms into the logarithm, gives the first two regions of the bound.

For L_1, we use a uniformly distributed grid for the support of p_0. This method resembles approaches in Vovk (1990), but requires careful design of the prior and the variational distribution $Q(\cdot)$. The grid is defined by

$$\Psi = \Theta_m = \left\{ \psi : \begin{array}{l} \|\psi\|_1 \le B; \\ \psi_{j,i} = i \cdot \varepsilon; \\ \qquad i \in \{-\lfloor B/\varepsilon \rfloor - 1, -\lfloor B/\varepsilon \rfloor, -\lfloor B/\varepsilon \rfloor + 1, \dots \lfloor B/\varepsilon \rfloor, \lfloor B/\varepsilon \rfloor + 1\}; \\ \qquad \forall j \in \{1, 2, \dots, d\} \end{array} \right\}$$

where j denotes the dimension. The grid consists of ε spaced points in each dimension, including 0, in $[-B-\varepsilon, B+\varepsilon]$, that satisfy the L_1 constraints. The spacing parameter ε will be optimized later. Allowing for integer length constraints and accounting for 0, in each dimension, we upper bound the number of grid points by $2(B+\varepsilon)/\varepsilon + 1$, giving an additional ε margin (that would not actually matter for the asymptotic results). (Note that the points in Q must be in Θ_m, so that $D(Q||p_0)$ is finite. Hence, the margin outside $[-B, B]$ is needed.) The volume of a subspace of a cube in $\mathbb{R}^d$ which satisfies an L_1 constraint is the $d!$ fraction of the space. Hence, to bound the actual number of grid points we divide the number of points in the cube by $d!$. For sufficiently small $d = o(B\sqrt{T})$, this will suffice with the extra margin. Hence, we have

$$M \le \frac{\left(\frac{2(B+\varepsilon)}{\varepsilon} + 1\right)^d}{d!}. \tag{29}$$

Assume that the ith dimension θ_i^* of θ^* falls between adjacent grid points ζ_1 and ζ_2, for which $\zeta_2 - \zeta_1 = \varepsilon$ in the projection of Ψ to dimension i. We then define the distribution Q as a product of independent components

$$Q(\theta) \stackrel{\triangle}{=} \prod_{i=1}^{d} q_i(\theta_i) \tag{30}$$

where

$$q_i(\theta_i) \stackrel{\triangle}{=} \left\{ \begin{array}{ll} \alpha; & \theta_i = \zeta_1, \\ 1-\alpha; & \theta_i = \zeta_2, \\ 0; & \text{otherwise} \end{array} \right. \tag{31}$$

where $0 \le \alpha \le 1$ is determined such that

$$E_{q_i}(\theta_i) = \alpha\zeta_1 + (1-\alpha)\zeta_2 = \theta_i^*. \tag{32}$$

By definition of ζ_1, ζ_2 and the expectation of $q_i(\theta_i)$, we have $\alpha = (\zeta_2 - \theta_i^*)/\varepsilon$, and we can show that the variance of $q_i(\cdot)$ is

$$E_{q_i}(\theta_i - \theta_i^*)^2 = (1-\alpha)\alpha\varepsilon^2 \le \frac{\varepsilon^2}{4} \stackrel{\triangle}{=} \eta_q^2. \tag{33}$$

We can now apply the bound of Theorem 2. Since $p_0(\cdot)$ is uniform over the grid with M points, and the support of $Q(\cdot)$ is a subset of the support of $p_0(\cdot)$,

$$D(Q||p_0) \leq \log M \tag{34}$$

where the negative entropy of $Q(\cdot)$ is bounded by 0 (which can be the case if θ^* falls on a grid point $\psi \in \Psi$). The second term of (9) is bounded from (33). Substituting (29) for M (absorbing low order terms in the $o(1)$ term) and using Stirling approximation for the factorial, we thus have

$$\text{Regret}\left(\mathcal{A}^*, S_T, \theta^*\right) \leq (1+o(1))\left[d\log(2B) - d\log\frac{d}{e} - \frac{1}{2}\log(2\pi d) - d\log\varepsilon + \frac{Td\varepsilon^2}{32}\right]. \tag{35}$$

Differentiating w.r.t. ε gives $\varepsilon^2 = 16/T$, which gives the minimal bounds. Substituting this value, gives the third region of the bound in Theorem 4.

As long as $d = o(B\sqrt{T})$, the bound on M includes sparse points with many components which are 0 (this is guaranteed by the additional margins and 1 term in (29), which are basically negligible in this region). However, for larger d, we no longer have full d-dimensional cubes with side ε that can fit in the allowable volume of the L_1 constrained space. We still, however, have points that are included in this space, that have a large fraction of 0 coordinates, but for which the bound in (29) is no longer sufficient. We note that this problem is not only an artifact of the grid approach. The first term of (9) becomes negligible for $d = O(T)$ with L_2, and if we used $\|x_t\|_2 \leq 1$ in our setting, this would also happen at $d = O(T)$ for L_∞, and at $d = O(\sqrt{T})$ for L_2. This implies that because the constraints shrink the parameter space, the variance of p_0 will be small enough, such that Q and p_0 almost match.

When the bound in (29) starts diminishing, the normalization in $d!$ eliminates many grid points that include 0 coordinates from the count. To account for these points, we can upper bound M with a union bound over all subsets of d with n nonzero coordinates.

$$M \leq (1+o(1)) \cdot \sum_{n=1}^{d} \binom{d}{n} \frac{B^n 2^n}{n!\varepsilon^n}. \tag{36}$$

The term to the right of the combination is maximized for $n_o = 2B/\varepsilon = B\sqrt{T}/2$. For $d = \Theta(B\sqrt{T})$, $n_o = \Theta(d)$. We can use a (loose) upper bound of 2^d on the combination number in (36), to obtain a union bound

$$\log M \leq (1+o(1)) \cdot \left[d\log 2 + \frac{B\sqrt{T}}{2} + \log d\right], \tag{37}$$

Combining this bound with the bound of $d/2$ on the right term of (9) we obtain the fourth region of the bound of Theorem 4.

To prove the last region, we can use a tighter bound on the combination term in (36), for $n_o \leq d/2$. For the last region this is satisfied. We thus bound

$$M \leq (1+o(1)) \cdot \sum_{n=1}^{d} \left(\frac{de}{n}\right)^n \frac{B^n 2^n}{n!\varepsilon^n}. \tag{38}$$

The largest element for the sum is obtained with $n_0^2 = dB\sqrt{T}/2 = o(d)$ in this region. Plugging n_o, still using $\varepsilon = 4/\sqrt{T}$, using a union bound on all d elements of the sum, taking the logarithm of M, and adding the $d/2$ bound for the right term of (9) gives the last region of the bound of Theorem 4, thus concluding its proof.

The only change from the last region to prove Theorem 5 is that now we first bound M using the optimizing $n_o^2 = 2Bd/\varepsilon$, and then we find ε that minimizes the joint bound. Bounding all terms with a parameter ε gives

$$\text{Regret}\left(\mathcal{A}^*, S_T, \theta^*\right) \leq (1 + o(1)) \cdot \left[\log d + \sqrt{\frac{8Bd}{\varepsilon}} + \frac{Td\varepsilon^2}{32}\right]. \tag{39}$$

The minimizing ε is then

$$\varepsilon = \frac{2^{9/5} B^{1/5}}{T^{2/5} d^{1/5}}. \tag{40}$$

Plugging this value in the bound gives

$$\text{Regret}\left(\mathcal{A}^*, S_T, \theta^*\right) \leq (1 + o(1)) \cdot \frac{5}{4} \cdot 2^{3/5} B^{2/5} d^{3/5} T^{1/5} = o(d). \tag{41}$$

This concludes the proof of Theorem 5.

We note that the grid approach used for the L_1 bounds requires an algorithm to know the horizon T in advance, to perform the mixture. However, in a *strongly sequential* setting, when the horizon is not known *a-priori*, we can start with some hypothesized horizon. Once it is reached, the next horizon can be squared (or exponentiated with some exponent $1 + \varepsilon$, for some small ε), and the current posterior (or prior of the next example), can be split from each grid point to the new T^{ε} nearest grid points. This will incur some additional negligible regret relative to the logarithmic regret in T, but can still achieve the current bounds. ■

Appendix E. Multi Label Lower Bound

Similarly to Theorem 3, we can state a regret lower bound for multi label logistic regression. We only state the bound for L_∞. Let $\theta^{*(m)}$ be the projection of the parameter space on label m, i.e., a d-dimensional parameter space for label m.

Theorem 7 *Let $\|\theta^{*(m)}\|_\infty \leq B$. Fix an arbitrary $\varepsilon > 0$, let $T \to \infty$, and let the number of labels $m = o((T/d)^{1-\varepsilon})$. Then, for every algorithm $\mathcal{A}$ there exists a sequence S_T, for which the regret for multi m label logistic regression is lower bounded by*

$$\textit{Regret}(\mathcal{A}, S_T) \geq (1 - o(1)) \cdot \frac{d(m-1)}{2} \log \frac{T}{m \cdot d}. \tag{42}$$

The $m - 1$ factor is used to indicate that there are only $m - 1$ free parameters per feature. It does not affect the asymptotic behavior. The proof of Theorem 7 is similar to that of the first region of Theorem 3 in segmenting x^{*T} into d segments, where in each only a single dimension exists.

However, handling each segment, especially if $m = \omega(1)$, is substantially more difficult. The behavior for a single segment, however, appears as Theorem 1 in Shamir (2006a). For tight behavior it requires a nonuniform grid, which is described in the proof. Because of the nonuniform grid, distinguishability (proven in Appendix A in Shamir (2006a)) is more difficult to prove. Borrowing the proof from Shamir (2006a), all is left to do is sum up the size of the grid over the d segments, and apply the union bound on the d segments for proving distinguishability. In each segment, there are $\log M_d = (1 - o(1))0.5m \log T/(md)$ distinguishable parameters. The normalization in d is because the segment is of length T/d by the partitioning of x^{*T}. The m parameter appears due to effectively L_1 constraints imposed by the fact that the probabilities on all labels must sum to 1. As in the proof of Theorem 3, for $d > T^{1-\varepsilon}$, we clip the analysis at $d_m = T^{1-\varepsilon}$. As in Theorem 3, we will have a threshold that depends on both d and m over which the bound becomes negative and useless. We can derive bounds, as those in the other regions of Theorem 3 for these cases, by taking the values of d and m that produce a maximal value of the bound, and lower bounding the regret for the larger d and m by the regret for the maximizing d and m.