Refined Risk Bounds for Unbounded Losses via Transductive Priors

Jian Qian   JIANQIAN@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Alexander Rakhlin   RAKHLIN@MIT.EDU
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Nikita Zhivotovskiy   ZHIVOTOVSKIY@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA

Abstract

We revisit the sequential variants of linear regression with the squared loss, classification problems with hinge loss, and logistic regression, all characterized by unbounded losses in the setup where no assumptions are made on the magnitude of design vectors and the norm of the optimal vector of parameters. The key distinction from existing results lies in our assumption that the set of design vectors is known in advance (though their order is not), a setup sometimes referred to as *transductive* online learning. While this assumption might seem similar to fixed design regression or denoising, we demonstrate that the sequential nature of our algorithms allows us to convert our bounds into statistical ones with random design without making any additional assumptions about the distribution of the design vectors-an impossibility for standard denoising results. Our key tools are based on the exponential weights algorithm with carefully chosen transductive (design-dependent) priors, which exploit the full horizon of the design vectors, as well as additional aggregation tools that address the possibly unbounded norm of the vector of the optimal solution.

Our classification regret bounds have a feature that is only attributed to bounded losses in the literature: they depend solely on the dimension of the parameter space and on the number of rounds, independent of the design vectors or the norm of the optimal solution. For linear regression with squared loss, we further extend our analysis to the sparse case, providing sparsity regret bounds that depend additionally only on the magnitude of the response variables. We argue that these improved bounds are specific to the transductive setting and unattainable in the worst-case sequential setup. Our algorithms, in several cases, have polynomial time approximations and reduce to sampling with respect to log-concave measures instead of aggregating over hard-to-construct ε-covers of classes.

Keywords: Transductive online learning, unbounded losses, logistic regression, hinge loss.

1. Introduction

Sequential learning algorithms are widely used in practice, even when the entire training sample is known in advance (*batch* setup), with stochastic gradient descent being one of the most well-known examples. The model we study in this paper represents an intermediate approach, combining sequential procedures with their *batch* counterparts. A standard sequential regression problem can be described as follows: at round t, one observes the design vector x_t and makes a prediction $\widehat{y}_t$ based on previous observations, aiming to approximate the true value y_t, which is revealed only after the prediction $\widehat{y}_t$ is made. Our setup, often referred to as *transductive online learning*, provides the learner with additional information, as the entire set of design vectors $x_1, \ldots, x_T$ (though not their order) is revealed to the learner in advance and can be utilized. In the terminology of Vapnik (2013), transduction refers to situations where the aim is not to infer a general prediction rule for arbitrary future points, but to predict the labels of a specific, fixed set of points known beforehand. This exactly matches our setting, where the goal is to assign labels to a predetermined sequence rather than to predict each point in an arbitrary sequence as it arrives.

One of our primary motivations for studying the transductive online model stems from its connection to the batch setting, where the entire training sample is known in advance. It is well established, due to the so-called *online-to-batch* conversion arguments, that regret bounds in sequential settings correspond to risk bounds in statistical settings. We make the following simple observation: instead of proving regret bounds in the more challenging *worst-case sequential setting*, one can instead prove regret bounds in the *easier transductive sequential setting* and obtain statistical risk bounds with provably better guarantees. Our analysis leverages certain versions of the classical exponential weights algorithm, where the prior distributions depend on the entire set of design vectors $x_1, \ldots, x_T$, thus motivating the term *transductive priors*.

Our key focus is on the separation between the regret bounds achievable in the transductive setting compared to the standard online setting, particularly in the case of linear regression with squared loss, including sparse linear regression, as well as classification problems like logistic regression and classification with hinge loss. Some of our key findings are simple enough to be formulated in the introduction. Let us start with classification with logarithmic loss: in the standard online setup (Hazan et al., 2014; Foster et al., 2018), we observe the vectors $x_t \in \mathbb{R}^d$ round by round, where $t = 1, \ldots, T$, and assign the probabilities $\widehat{p}(x_t, 1)$ to label 1 and $\widehat{p}(x_t, -1)$ to label -1, after which we observe the true label $y_t \in \{1, -1\}$. For some $b > 0$, our aim is to minimize the regret with respect to the logarithmic loss, that is,

$$\text{Reg}_T = \sum_{t=1}^{T} -\log(\widehat{p}(x_t, y_t)) - \inf_{\theta \in \mathbb{R}^d, \|\theta\|_2 \leq b} \sum_{t=1}^{T} -\log(\sigma(y_t \langle x_t, \theta \rangle)), \tag{1}$$

where $\sigma(z) = 1/(1 + \exp(-z))$ is the sigmoid function. Here, for the sake of example, we assumed $\|x_t\|_2 \leq 1$, though importantly, the results in this paper will completely bypass *any boundedness assumptions* on the design vectors. The approaches based on choosing the *proper* predictor, that is, a predictor corresponding to the logistic regression model, $\widehat{p}(x_t, 1) = \sigma(\langle x_t, \widehat{\theta}_t \rangle)$, where $\widehat{\theta}_t$ is independent of x_t, are quite limited. As shown by Hazan, Koren, and Levy (2014), and rather informally skipping some assumptions, the following

lower bound holds:

$$\text{Reg}_T = \Omega\left(\min\left\{d\exp(b), b\sqrt{T}\right\}\right).$$

Motivated by the question of McMahan and Streeter (2012), a number of works have focused on improving the exponential dependence on b in the above bound. A significant improvement on the regret bound can be achieved by using *improper* predictions, namely mixtures of sigmoid functions, as shown by the results of Kakade and Ng (2004); Foster, Kale, Luo, Mohri, and Sridharan (2018) (see also (Cesa-Bianchi and Lugosi, 2006, Section 12.2)). Their results generate a sequence of probability assignments $\widehat{p}(x_t, \cdot)$ whose regret, as defined by (1), satisfies

$$\text{Reg}_T = O(d\log(b \cdot T)), \tag{2}$$

demonstrating only logarithmic dependence on b. Moreover, it appears that this result is generally unimprovable, as the lower bound shows that $b < \infty$ is indeed required for regret that scales better than $O(T)$. Remarkably, the constructions by Foster et al. (2018) demonstrating the necessity of the dependence on b are related to the classical example in sequential binary classification, due to Littlestone (1988), where the thresholds are impossible to learn with a finite number of mistakes. At the same time, thresholds are easily learnable in the batch setup, or alternatively in the sequential setup when the sequence $\{x_1, \ldots, x_T\}$ is known in advance (Ben-David et al., 1997). As we show in what follows, this favorable phenomenon extends well beyond the example of realizable binary classification and can be applied to classification and regression problems with several loss functions in the agnostic case.

It is important to note that the assumption $b < \infty$ in (2) is not natural. For instance, in the favorable scenario where all data points belong to the same class $+1$, the optimal parameter vector $\theta^\star$ could have an arbitrarily large norm. As discussed earlier, our focus in this paper is on the transductive setup, where the set of vectors $\{x_1, \ldots, x_T\}$ is known to the learner beforehand. By denoting the regret in this setup as $\text{Reg}_T^{\text{trd}}$, we demonstrate that

$$\text{Reg}_T^{\text{trd}} = O(d\log(T)), \tag{3}$$

where there is *no dependency* on the norm of the optimizer $\theta^\star$ or the norms of the x_t's. Notably, the probability assignments generated by our algorithm do not require a computationally prohibitive discretization of the function space, which is often seen in density estimation and regression approaches (Yang and Barron, 1999; Vovk, 2006; Bilodeau et al., 2023). Instead, we leverage variants of continuous exponential weights, which, while not always practical, can be implemented in polynomial time in certain scenarios.

While the distinction between the regret bounds we prove in this paper-comparing the existing $\text{Reg}_T = O(d\log(b \cdot T))$ with $\text{Reg}_T^{\text{trd}} = O(d\log(T))$-might appear quantitatively minor, it is, in fact, qualitatively significant. Our findings show an essential separation: they reveal what is learnable, with sublinear regret achievable in one framework, namely the transductive setup (which also extends to the classical batch setting, as discussed below), versus what remains unlearnable in the traditional online learning setup.

Additionally, we introduce a variant of the online-to-batch conversion argument to illustrate the application of our results in the statistical setup. In the statistical learning context, one observes i.i.d. sample $(X_t, Y_t)_{t=1}^T$ of copies of (X, Y), where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. Our analysis shows that there exists a probability assignment $\widetilde{p}(\cdot, \cdot)$, constructed based on the training sample $(X_t, Y_t)_{t=1}^T$, such that

$$\mathbb{E}[-\log(\widetilde{p}(X, Y))] - \inf_{\theta \in \mathbb{R}^d} \mathbb{E}[-\log(\sigma(Y \langle X, \theta \rangle))] = O\left(\frac{d \log T}{T}\right), \tag{4}$$

with expectations taken over both the training sample $(X_t, Y_t)_{t=1}^T$ and an independent test observation (X, Y). A key feature of our excess risk bound is that the right-hand side of (4) is *completely independent* of the distributions of X and Y. Moreover, the curvature of the logarithmic loss function enables the $d \log(T)/T$ convergence rate without requiring additional realizability/low noise assumptions. Despite the extensive literature on logistic regression in the statistical framework (for comparison, see, e.g., (Bach, 2010; Hazan et al., 2014; Plan et al., 2017; Ostrovskii and Bach, 2021; Puchkin and Zhivotovskiy, 2023; Kuchelmeister and van de Geer, 2024; Hsu and Mazumdar, 2024) for a non-exhaustive list of existing risk upper bounds in this context), this fully assumption-free convergence rate has not been previously established even for mixtures of sigmoids proposed in (Kakade and Ng, 2004; Foster et al., 2018) and other batch-specific results as in (Mourtada and Gaïffas, 2022; Vijaykumar, 2021; van der Hoeven et al., 2023). We emphasize again that our approach relies on using variants of the continuous exponential weights algorithm.

Our findings go well beyond the logarithmic loss. For example, we consider the case of sparse linear regression with *squared loss*. Specifically, in the transductive setting with squared loss, assuming $|y_t| \leq m$ for all $t = 1, \ldots, T$ and comparing against any $\theta^\star$ with $\|\theta^\star\|_0 \leq s$, we prove a regret bound of

$$\mathrm{Reg}_T^{\mathrm{trd}} = O\left(sm^2 \log\left(\frac{dT}{s\kappa_s^2}\right)\right),$$

where the dependence on the compatibility constant κ_s is logarithmic, contrasting with the polynomial dependence found in most known results for predictors that satisfy sparsity regret or risk bounds.

Furthermore, in the setup of classification with *hinge loss*, we prove the so-called *first-order* regret bounds. In more detail, we observe the vectors $x_t \in \mathbb{R}^d$ round by round, where $t = 1, \ldots, T$, and output a real-valued score $\widehat{f}(x_t)$, trying to minimize, for some $\gamma > 0$, the total hinge loss:

$$\sum_{t=1}^T \frac{(\gamma - y_t \widehat{f}(x_t))_+}{\gamma}.$$

In the transductive setup, where we know the set $\{x_1, \ldots, x_T\}$, if additionally the dataset is linearly separable with margin γ, the result of our Theorem 21 implies, in particular, that

$$\sum_{t=1}^T \frac{(\gamma - y_t \widehat{f}(x_t))_+}{\gamma} = O\left(d \log(T)\right).$$

Prediction is again achieved using a version of the continuous exponential weights algorithm, and the right-hand side of the above cumulative error bound does not depend on γ, which need not hold in the inductive (unknown-set) setup.

1.1 Key contributions and structure of the paper

Our key contributions can be summarized through the regret bounds we derive for the transductive online learning framework, considering the squared loss, logistic loss, and hinge loss. These bounds explicitly reveal distinctions between the transductive and inductive setups across different loss functions. Moreover, our algorithm exhibits computational advantages through sampling from log-concave densities. Concretely, we summarize them as follows:

- In Section 2, we revisit the classical problem of online squared loss regression. We recover the optimal rate of $O(dm^2 \log T)$ for the transductive setup originally established by Gaillard et al. (2019) with a transductive prior without making any assumptions about the boundedness of covariates or the norm of the optimal solution (comparator). In the sparse linear regression setting, we introduce a sparse prior that achieves regret bounds independent of the norm of $\theta^\star$ via the evaluation of Gaussian-type integrals. For completeness, we also include results by combining our transductive prior and the discrete sparse prior proposed by Rigollet and Tsybakov (2011, 2012) in Appendix A, where a better regret bound for sparse linear regression is achieved, albeit through a method involving sampling from discrete measures, which may require a prohibitively large number of samples from these distributions in the worst case.

  Then, we present a hard case where the celebrated predictor from Vovk (2001); Azoury and Warmuth (2001) suffers $\Omega(T)$ regret, which shows a separation between the transductive setup and the inductive setup. Finally, we show that such a separation exists for any prediction strategy when comparing with the clipped version of the squared loss.

- In Section 3, we cover transductive online logistic regression. In this section, we present a construction of slab-experts that combine predictions of exponential weights algorithms with a transductive Gaussian prior and linear separators. We show that the optimal slab-expert always obtains the optimal *assumption-free* bound of $O(d \log T)$, and by mixability, such a bound is achievable in general with the exponential weights algorithm over the slab-experts. We also mention that such assumption-free bounds are not obtainable in the inductive setup, establishing a desired separation.

- In Section 4, we investigate the hinge loss in both the inductive and transductive setup. The hinge loss is unbounded above and not mixable. Nevertheless, we present the mistake bound in the realizable case scaling as $O(d \log(bT/\gamma))$ in the realizable case, where γ is the margin of separation in the inductive case and b is the norm of the optimal weight vector. This regret bound is the parametric counterpart of the celebrated $O(1/\gamma^2)$ mistake bound for perceptron. We then combine the slab-expert idea with the aforementioned clipping trick to obtain an assumption-free mistake bound in the realizable case of $O(d \log T)$ in the transductive setup. The separation between the inductive and transductive setups also exists for the hinge loss.

- In Section 5, we highlight our computational benefit through sampling from log-concave densities. More concretely, since the squared loss and the logistic loss are convex, the exponential weights algorithm, therefore, predicts log-concave densities. We showcase how to compute the exponential weights algorithm for sparse linear regression and logistic regression. For the latter, we demonstrate how to sample the slab-experts, though our

constructions lead to computational complexities exponential in dimension in the worst case.

- In Section 6, the online-to-batch conversion is restated with the reduction from the transductive online learning setup to the batch setup. Our results Sections 2 to 4, together with this reduction, bring forth new batch results for corresponding losses. Moreover, for the special case of logistic regression and under mild additional assumptions on the design, we obtain a polynomial time algorithm that gives a $O((d + \log T) \log T / T)$ excess risk upper bound.

Alongside, in Section 7, we discuss several directions of future work. In Appendix C, we revisit sequential linear regression and discuss short proofs of several existing bounds.

1.2 Contextual review of related literature

In this section, we provide an overview focused on the most relevant papers. Given the vast volume of existing literature, we concentrate only on the most pertinent directions, with additional references typically found in the respective papers. Further discussion of related literature is deferred to the corresponding sections.

Our results cover three unbounded loss functions, namely squared loss, logarithmic loss, and hinge loss, using a unified approach. It is worth mentioning that in the case of the squared loss, the results similar to ours are relatively developed. In the context of squared loss, the closest papers to our work are (Bartlett et al., 2015; Gaillard et al., 2019), which were the first to emphasize the improved rates for the celebrated sequential online regression predictor of Vovk (2001); Azoury and Warmuth (2001). In fact, the paper (Gaillard et al., 2019) achieves the same regret bounds we achieve with our methods. However, in the case of the squared loss, we make some further extensions and work with the sparse case. When working with the squared loss, a result that is somewhat close in spirit to our result is the sequential bound of Gerchinovitz (2013), which is a version of the earlier non-sequential fixed design sparse regression papers of Dalalyan and Tsybakov (2008, 2012a,b). We make a more detailed comparison in Section 2.1.

The case of logistic regression is more challenging and requires further work. Analyzing this problem requires some form of boundedness for the vector of optimal solutions $\theta^\star$. Because this limitation is mainly recognized in the case of proper probability assignments, where the assigned probability corresponds to the sigmoid function (Hazan et al., 2014), our approach builds on (Kakade and Ng, 2004; Foster et al., 2018) (see also (Cesa-Bianchi and Lugosi, 2006, Section 11.10)) and uses the mixture approach or improper predictors, where assigned probabilities are not necessarily given by sigmoid functions. This technique allows for improving the dependence on the norm of the optimal solution from being *exponential* with respect to the norm of $\theta^\star$ to a *logarithmic* dependence on this norm. However, without additional assumptions, the lower bounds in (Foster et al., 2018) require that the norm of $\theta^\star$ is bounded. Our results show that if the design vectors are known in advance, an exponential weighting algorithm-based predictor can achieve a regret bound dependent only on the dimension and the number of rounds. The only studies we know that address unbounded optimal solutions are the recent works by (Drmota et al., 2024, 2025). These works consider a similar transductive online setup with x_t generated uniformly at random on the sphere and

6

compute the precise asymptotics of the first term of the expected regret and leave the additive asymptotic term in their bounds. While asymptotic results are standard in information theory, our results provide deterministic regret bounds without structural assumptions on design vectors, are finite samples, but do not recover the asymptotically optimal constant in front of the leading term $d \log(T)$. The methods differ, as we focus on adapting exponential weights algorithms with transductive priors, which also usually leads to polynomial time implementation, whereas Drmota et al. (2024, 2025) analyze Shtarkov's sum using analytic combinatorics tools. We note that the analysis of Shtarkov's sum is not known to lead to polynomial time predictors. We also refer to the paper (Wu et al., 2022), where connections between online transductive setup and the result of Shtarkov (1987) are made.

Choosing the right prior for exponential weights is a classical question studied extensively in the information theory literature, primarily focusing on logarithmic loss. When computing the regret with respect to realizable i.i.d. data, the works of Clarke and Barron (1990, 1994) show that the *Jeffreys* prior, the prior induced by the Fisher information, provides asymptotically optimal expected regret bounds. The classical work of Rissanen (1996) demonstrates that the i.i.d. data assumption is not necessary to achieve optimal asymptotic regret with the Jeffreys prior. This is done under the assumption that the maximum likelihood estimates satisfy the Central Limit Theorem, a weaker assumption than i.i.d. data. These important results are asymptotic in nature and do not fully align with our goals. In some special cases, such as probability assignment, the analysis of the renowned estimator by Krichevsky and Trofimov (1981) can be seen as follows. First, one computes the *asymptotically* optimal Jeffreys prior, but then uses it to obtain sharp non-asymptotic regret bounds for the exponential weights algorithm (see the derivations in (Cesa-Bianchi and Lugosi, 2006, Section 9.7)). Another approach to non-asymptotic bounds with the Jeffreys prior in exponential weights algorithms with i.i.d. data is discussed in (Barron, 1999). This approach derives non-asymptotic expected regret bounds under a local quadratic behavior of the Fisher information. Unfortunately, the method as is does not work in the context of logistic regression with unbounded target parameters due to possible very imprecise quadratic approximations when the target vector has coordinates that are too large. However, a modification of this idea, involving control over large coordinates of the target vector of parameters, is used in our analysis.

In the machine learning literature, transductive online learning has been explored in several papers (Ben-David et al., 1997; Kakade and Kalai, 2005; Cesa-Bianchi and Shamir, 2013). These studies primarily demonstrate an equivalence between batch learning in classification tasks using binary loss—assuming random and i.i.d. samples—and transductive settings, focusing on *expected regret* rather than worst-case scenarios. Simplifying these results, they suggest that a computationally efficient i.i.d. learner can be converted to an efficient transductive online learner. Our work, in contrast, focuses on establishing deterministic regret bounds for several unbounded losses, which in some cases can lead to computationally efficient learners in the batch setup.

Results closely related to our regret bounds are known in the batch setup, where one observes the full sample of i.i.d. points and focuses on bounding the *excess risk*. The result of Forster and Warmuth (2002) complements the sequential regression bound of Vovk (2001) (see also (Vaškevičius and Zhivotovskiy, 2023) for the relation of this result to the standard least squares estimator, and (Mourtada et al., 2022) for the high probability version of their

excess risk bound) by focusing on the batch setup. The excess risk bound of Forster and Warmuth (2002) shares similar properties with our regret bound: it depends neither on the properties of the design vectors nor on the norm of the optimal parameter $\theta^\star$. In the context of removing dependencies on the norm of $\theta^\star$, the recent work on *localized priors* (originally introduced in (Catoni, 2007)) in the PAC-Bayesian scenario Mourtada et al. (2023) is particularly relevant. The advantage of that work is that, specifically in the batch setting, it not only removes the dependence on the norm of the optimal parameter $\theta^\star$ but also eliminates superfluous logarithmic factors related to the sample size, although it still requires additional boundedness assumptions that are easier to circumvent in our sequential setup.

Another closely related area of literature is the PAC-Bayesian analysis and its specific realization with *exchangeable priors*. Originally introduced in (Catoni, 2007; Audibert, 2004), this approach was primarily developed for the analysis of exponential-weights type prediction rules in classification problems, particularly when learning VC classes of classifiers. While selecting a prior distribution for exponential weights can be problematic for discrete but infinite VC classes, the method becomes feasible if one introduces a ghost sample and projects the VC class onto it, rendering the class finite and allowing a uniform distribution over the projection to serve as a prior. We refer to the recent works (Grünwald et al., 2021; Mourtada et al., 2023) that utilize exchangeable priors, as well as the survey (Alquier, 2024) for an extensive discussion of PAC-Bayesian results. Our online learning-based approach is more flexible: it operates in the sequential setup instead of relying on i.i.d. data, and it allows us to easily incorporate improvements due to transductive priors as well as the mixability of the loss. This flexibility is reflected in the difference between $O(\log T)$ and $O(\sqrt{T})$ regret bounds, without the need to assume the boundedness of the loss, which is typically required in PAC-Bayesian analysis.

Finally, we consider the hinge loss. To the best of our knowledge, except for the recent work of Mai (2024) (see also (Alquier et al., 2016)) focusing on sparsity and PAC-Bayesian tools, studies exploring the connections between exponential weights and the hinge loss are limited. In this context, even without access to $x_1, \ldots, x_T$, i.e., in the standard online learning setup, we derive a mistake bound complementary to the classical perceptron error bound. While the classic perceptron algorithm incurs $O(1/\gamma^2)$ mistakes, as demonstrated by the result of Novikoff (1962), our bound of $O(d\log(T/\gamma))$ can be tighter in low-dimensional settings when the margin is particularly small. The closest in spirit result in the literature is the mistake bound of Gilad-Bachrach, Navot, and Tishby (2004), which applies specifically to the realizable case, using a method based on predictions from the center of gravity of the current version space. Furthermore, in the transductive online setup, where all $x_1, \ldots, x_T$ are known, we show that, in the realizable case, the total normalized hinge error will scale as $d\log(T)$, independent of the margin γ. This phenomenon is similar to what we obtain for logistic regression, where the regret depends only on the dimension and the number of rounds.

Notation. In what follows, I_d stands for the $d \times d$ identity matrix. We use $\|\cdot\|_p$ for ℓ_p norm of the vector, $p \geq 1$. The symbol $\|v\|_0$ denotes the number of non-zero coordinates of $v \in \mathbb{R}^d$. In particular, a vector $v \in \mathbb{R}^d$ is s-sparse if $\|v\|_0 = s$. For a pair of measures φ, μ the symbol $\mathcal{KL}(\varphi\|\mu)$ denotes the Kullback-Leibler divergence between φ and μ. The symbol $\mathcal{N}(\nu, \sigma^2)$ stands for the Gaussian distribution with mean ν and variance σ^2. For any real number x, the symbol $\lfloor x \rfloor$ denotes the largest integer smaller or equal to x. For

any matrix A, $\det(A)$ denotes the determinant of the matrix, and $A^\top$ denotes its transpose. For any set $\mathcal{X}$, the symbol $|\mathcal{X}|$ denotes its cardinality. We define $O(\cdot), \Omega(\cdot)$, and $\mathrm{poly}(\cdot)$ following standard non-asymptotic big-O notation. For any positive real number m, we define the clipping operator $\mathrm{clip}_m(x) = \min\{m, \max\{-m, x\}\}$. We use $\mathbb{1}(\mathcal{E})$ to denote the indicator function for event $\mathcal{E}$. We use $\mathcal{S}^{d-1}$ to denote the unit sphere in $\mathbb{R}^d$, i.e., $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$. For any positive number $a > 0$, we use $a\mathcal{S}^{d-1}$ to denote the unit sphere scaled by a, that is, $a\mathcal{S}^{d-1} = \{ax \mid x \in \mathcal{S}^{d-1}\}$. For any bounded measurable set $\mathcal{X}$ in $\mathbb{R}^d$, we use $\mathrm{Unif}(\mathcal{X})$ to denote the uniform distribution on $\mathcal{X}$.

2. Linear Regression with Transductive Priors

We begin this section by recalling the regret bound in (Gaillard et al., 2019, Theorem 7), which itself is an improvement over the earlier result in (Bartlett et al., 2015). This result is the closest in spirit to the bounds presented in this paper and serves as an existing benchmark.

Theorem 1 (Gaillard, Gerchinovitz, Huard, and Stoltz (2019)) *Assume that the sequence $\{(x_t, y_t)\}_{t=1}^T$, where $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$ for all $1 \le t \le T$, is such that $\max_t |y_t| \le m$ and that the matrix $\sum_{t=1}^T x_t x_t^\top$ is invertible[1]. For $x \in \mathbb{R}^d$, consider the sequence of estimators*

$$\widehat{\theta}_{t,x} = \arg \min_{\theta \in \mathbb{R}^d} \left\{ \sum_{i=1}^{t-1} (y_i - \langle x_i, \theta \rangle)^2 + (\langle x, \theta \rangle)^2 + \lambda \sum_{i=1}^T (\langle x_i, \theta \rangle)^2 \right\}.$$

Then, the following regret bound holds:

$$\sum_{t=1}^T (y_t - \langle x_t, \widehat{\theta}_{t,x_t} \rangle)^2 \le \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \langle x_t, \theta \rangle)^2 \right\} + \lambda m^2 T + dm^2 \log \left(1 + \frac{1}{\lambda} \right).$$

In particular, choosing $\lambda = 1/T$ yields

$$\sum_{t=1}^T (y_t - \langle x_t, \widehat{\theta}_{t,x_t} \rangle)^2 \le \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \langle x_t, \theta \rangle)^2 \right\} + m^2(1 + d \log(1 + T)). \tag{5}$$

The key aspect of this regret bound is that the term $m^2(1 + d \log(1 + T))$ depends neither on the x_t-s nor on the norm of $\theta^\star$, where $\theta^\star = \arg \min_{\theta \in \mathbb{R}^d} \sum_{t=1}^T (y_t - \langle x_t, \theta \rangle)^2$. Throughout the paper, we are aiming for results of this type.

The proof of Theorem 1 in (Gaillard et al., 2019) employs a linear transformation of the feature vectors, thereby reducing the problem to the classical regret bounds of Vovk (2001) and Azoury and Warmuth (2001) using linear algebraic arguments. Our approach, instead, follows the original path of Vovk (2001) by exploiting the mixability of the loss (details to follow), with the main difference being the choice of a prior distribution that depends on the design vectors $x_1, \ldots, x_T$. Theorem 1 can be immediately recovered from the general bounds

---

1. This assumption can be easily bypassed. One way around this is by restricting the problem to the linear span of $x_1, \ldots, x_T$.

for the exponential weights algorithm and the mixability of the squared loss discussed in (Vovk, 2001) by choosing the following multivariate Gaussian prior μ in $\mathbb{R}^d$:

$$\mu(\theta) = \left(\frac{\lambda}{\pi}\right)^{d/2} \sqrt{\det\left(\sum_{t=1}^{T} x_t x_t^\top\right)} \exp\left(-\lambda\theta^\top \left(\sum_{t=1}^{T} x_t x_t^\top\right)\theta\right). \tag{6}$$

We do not provide the derivation here but instead present several versions of the derivation throughout the paper. This direct approach is a natural first step when analyzing the logarithmic loss, hinge loss, and absolute loss. In Appendix C, we discuss some general deterministic identities for online linear regression that are capable of recovering existing sequential linear regression bounds, including the regret bound of Theorem 1.

The prior (6) has a close connection with the *Jeffreys prior*, where $\sqrt{\det\left(\sum_{t=1}^{T} x_t x_t^\top\right)}$ plays the role of this prior and the term $\exp\left(-\lambda\theta^\top \left(\sum_{t=1}^{T} x_t x_t^\top\right)\theta\right)$ plays the role of the exponential tilting that makes this prior integrable. We refer to (Barron, 1999, Section 3) as well as (Polyanskiy and Wu, 2024, Chapter 13.4). Connections with Jeffreys priors in the context of existing information theory results will be made throughout the paper.

2.1 Sparse linear regression

In this section, as an initial demonstration of our approach utilizing transductive priors for exponential weights, we aim to extend Theorem 1. Specifically, we will focus on linear regression, with an emphasis on deriving *sparsity* regret bounds within the transductive setup. We remark that the techniques used in this section can be readily applied in other contexts, including logistic regression and classification with hinge loss.

The regret bounds of this section compare favorably to the existing bounds with a fixed design, such as those presented in (Dalalyan and Tsybakov, 2008, 2012a,b), as well as their online counterparts presented in (Gerchinovitz, 2013). Our regret bounds do not depend on the magnitude of the optimal solution, which is the case in the aforementioned papers in the *batch* setup, where the whole data sample is revealed to us in advance. This aspect has already been noticed in (Rigollet and Tsybakov, 2012), where discrete sparse priors are used to achieve sharp oracle inequalities without any dependence on the norm of the optimal solution but in the context of fixed design batch linear regression with Gaussian noise, a denoising result less general than our sequential result.

First, we introduce the following *sparsity inducing* data-dependent prior. Fix some $\tau > 0$ and assume without loss of generality that for all j, it holds that $\sum_{t=1}^{T}(x_t^{(j)})^2 \neq 0$, where $x_t^{(j)}$ denotes the j-th coordinate of x_t. Indeed, otherwise, we may ignore this coordinate. The prior μ over $\mathbb{R}^d$ is defined as follows:

$$\mu(\theta) = \prod_{j=1}^{d} \frac{3 \cdot \sqrt{\sum_{t=1}^{T}(x_t^{(j)})^2}/\tau}{2\left(1 + |\theta^{(j)}| \cdot \sqrt{\sum_{t=1}^{T}(x_t^{(j)})^2}/\tau\right)^4}. \tag{7}$$

The prior is similar to the sparse priors used in (Dalalyan and Tsybakov, 2008, 2012a,b) and (Gerchinovitz, 2013), but, as will always be the case in this paper, it explicitly depends on

$x_1, \ldots, x_T$. The estimator we use has an explicit integral form following Vovk's algorithm (Theorem 6). Given m, which is the upper bound on $\max_t |y_t|$, define for $x \in \mathbb{R}^d$,

$$\widehat{f_t}(x) = \frac{m}{2} \cdot \log \left(\frac{\int_{\mathbb{R}^d} \exp \left(-\frac{1}{2m^2}(m - \langle x, \theta \rangle)^2 - \frac{1}{2m^2} \sum_{i=1}^{t-1}(y_i - \langle x_i, \theta \rangle)^2 \right) \mu(\theta) d\theta}{\int_{\mathbb{R}^d} \exp \left(-\frac{1}{2m^2}(-m - \langle x, \theta \rangle)^2 - \frac{1}{2m^2} \sum_{i=1}^{t-1}(y_i - \langle x_i, \theta \rangle)^2 \right) \mu(\theta) d\theta} \right). \quad (8)$$

We discuss the computational aspects behind this estimator in Section 5. We emphasize that the analysis of the related predictor in (Gerchinovitz, 2013) is based on exp-concavity and requires an additional clipping step, which prevents computational efficiency. As we mentioned, the computational aspects of sparse sequential linear regression were raised as an open question in (Lattimore and Szepesvári, 2020, Chapter 23).

We need an additional scale-invariant assumption on the design vectors $x_1, \ldots, x_T$. This assumption is quite standard in the sparse recovery literature and is discussed and used in (Castillo et al., 2015; Ray and Szabó, 2022) among other papers.

Definition 2 (Smallest scaled singular value) *Assume that there is a constant $\kappa_s > 0$ such that for any non-zero s-sparse vector $v \in \mathbb{R}^d$, it holds that*

$$\frac{\sqrt{\sum_{t=1}^{T}(\langle x_t, v \rangle)^2}}{\max_j \sqrt{\sum_{t=1}^{T}(x_t^{(j)})^2} \cdot \|v\|_2} \geq \kappa_s.$$

The following result is the main bound of this section.

Theorem 3 *Assume that the sequence $\{(x_t, y_t)\}_{t=1}^{T}$, where $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$ for all $1 \leq t \leq T$, is such that $\max_t |y_t| \leq m$ and that the smallest scaled singular value condition is satisfied with constant κ_s. Then, for any s-sparse $\theta^\star \in \mathbb{R}^d$, the following regret bound holds for the estimator (8) with $\tau = m\sqrt{\frac{s}{d}}$,*

$$\sum_{t=1}^{T}(y_t - \widehat{f_t}(x_t))^2 \leq \sum_{t=1}^{T}(y_t - \langle x_t, \theta^\star \rangle)^2 + sm^2 \left(1 + 8 \log \left(1 + \frac{2}{\kappa_s}\sqrt{\frac{dT}{s}} \right) \right). \quad (9)$$

There are multiple interesting properties of the above result. First, the main term in (9) is $O\left(sm^2 \log \left(\frac{dT}{\kappa_s^2 s} \right) \right)$ and does not depend on the magnitude of design vectors and the magnitude of $\theta^\star$. Second, the dimensionality of $\theta^\star$ enters the regret bound only through $s \log d$. Finally, we note that the dependence on $\frac{1}{\kappa_s}$ is logarithmic, which is not typical for sparse recovery problems, where dependencies of this type are usually polynomial.

It is time to introduce the exponential weights algorithm (Littlestone and Warmuth, 1994; Vovk, 1990; Freund and Schapire, 1997). Fix $\eta > 0$ and let $\ell_\theta(\cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a set of loss functions parameterized by some $\Theta \subseteq \mathbb{R}^d$. For notational simplicity, we usually use the same symbols for distributions and their densities. Fix some prior μ over Θ. Given the sequence $(x_1, y_1), \ldots, (x_T, y_T)$, define $\rho_1 = \mu$. For $t \geq 2$, we set

$$\rho_t(\theta) = \frac{\exp \left(-\eta \sum_{i=1}^{t-1} \ell_\theta(x_i, y_i) \right) \mu(\theta)}{\mathbb{E}_{\theta' \sim \mu} \exp \left(-\eta \sum_{i=1}^{t-1} \ell_{\theta'}(x_i, y_i) \right)}. \quad (10)$$

11

The following identity is standard and appears in, e.g., (Vovk, 2001, Lemma 1).

Lemma 4 *The exponential weights algorithm satisfies the following identity:*

$$\sum_{t=1}^{T} -\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \rho_t} \exp(-\eta \ell_\theta(x_t, y_t)) \right) = -\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \mu} \exp \left(-\sum_{t=1}^{T} \eta \ell_\theta(x_t, y_t) \right) \right). \quad (11)$$

Each summand on the left-hand side of (11) is called the *mix-loss*. Our next lemma is also standard in the online learning literature and is an immediate corollary of the Donsker-Varadhan variational formula.

Lemma 5 *In the setup of Theorem 4, the following identity holds:*

$$-\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \mu} \exp \left(-\sum_{t=1}^{T} \eta \ell_\theta(x_t, y_t) \right) \right) = \inf_{\varphi} \left(\mathbb{E}_{\theta \sim \varphi} \sum_{t=1}^{T} \ell_\theta(x_t, y_t) + \frac{\mathcal{KL}(\varphi \| \mu)}{\eta} \right).$$

Finally, we introduce another key lemma. This form is due to (Vovk, 2001, Lemma 2 and computations after it). The important aspect of it is that it does not require boundedness of the class of functions (in our case, it is the unbounded linear class), which appears to be important for our analysis. We additionally refer to (Van Erven et al., 2015) for a detailed account of mixability. For completeness, we provide a proof also in Appendix B.

Lemma 6 (Mixability of the squared loss) *Consider a class $\mathcal{F}$ of functions $f_\theta : \mathcal{X} \to \mathbb{R}$ (possibly unbounded) parameterized by $\Theta \subseteq \mathbb{R}^d$. Assume that y is such that $|y| \leq m$ and choose $\eta = \frac{1}{2m^2}$. Given any distribution ρ over Θ, define the predictor*

$$\widehat{f}(x) = \frac{m}{2} \log \left(\frac{\mathbb{E}_{\theta \sim \rho} \exp(-\eta(m - f_\theta(x))^2)}{\mathbb{E}_{\theta \sim \rho} \exp(-\eta(-m - f_\theta(x))^2)} \right). \quad (12)$$

Then, the following holds:

$$(y - \widehat{f}(x))^2 \leq -\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \rho} \exp(-\eta(y - f_\theta(x))^2) \right).$$

Proof [Proof of Theorem 3] In what follows, $\eta = \frac{1}{2m^2}$. Our first observation is that if we choose $\rho = \rho_t$ in Theorem 6, where ρ_t is the exponential weights algorithm with prior μ given by (7) and $\ell_\theta(x, y) = (y - \langle x, \theta \rangle)^2$, then the predictor (12) coincides exactly with the predictor $\widehat{f}_t$ given by (8). This implies by Theorem 6, Theorem 4 and Theorem 5:

$$\sum_{t=1}^{T} (y_t - \widehat{f}_t(x_t))^2 \leq \sum_{t=1}^{T} -\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \rho_t} \exp \left(-\eta(y_t - \langle x_t, \theta \rangle)^2 \right) \right)$$

$$= -\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \mu} \exp \left(-\sum_{t=1}^{T} \eta(y_t - \langle x_t, \theta \rangle)^2 \right) \right)$$

$$\leq \mathbb{E}_{\theta \sim \mu^\star} \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 + \frac{\mathcal{KL}(\mu^\star \| \mu)}{\eta}, \quad (13)$$

where $\mu^\star$ is any distribution over Θ. It remains to bound both terms in the last line. First, given an s-sparse target vector $\theta^\star$, we choose

$$\mu^\star(\theta) = \prod_{j=1}^{d} \frac{3 \cdot \sqrt{\sum_{t=1}^{T} (x_t^{(j)})^2 / \tau}}{2 \left(1 + |\theta^{(j)} - \theta^{\star(j)}| \cdot \sqrt{\sum_{t=1}^{T} (x_t^{(j)})^2 / \tau} \right)^4}.$$

By symmetry $\mathbb{E}_{\theta \sim \mu^\star} \theta = \theta^\star$. Moreover, a simple computation shows that $\mathbb{E}_{\theta \sim \mu}(\theta^{(j)})^2 = \tau^2 / (\sum_{t=1}^{T} (x_t^{(j)})^2)$ (recall that μ is defined by (7)). Therefore, we have

$$\mathbb{E}_{\theta \sim \mu^\star} \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 = \sum_{t=1}^{T} (y_t - \langle x_t, \theta^\star \rangle)^2 + \sum_{t=1}^{T} \mathbb{E}_{\theta \sim \mu^\star}(\langle x_t, \theta - \theta^\star \rangle)^2$$

$$= \sum_{t=1}^{T} (y_t - \langle x_t, \theta^\star \rangle)^2 + \sum_{j=1}^{d} \sum_{t=1}^{T} \mathbb{E}_{\theta \sim \mu}(x_t^{(j)} \theta^{(j)})^2$$

$$= \sum_{t=1}^{T} (y_t - \langle x_t, \theta^\star \rangle)^2 + \sum_{j=1}^{d} \sum_{t=1}^{T} \frac{(\tau x_t^{(j)})^2}{\sum_{t'=1}^{T} (x_{t'}^{(j)})^2}$$

$$= \sum_{t=1}^{T} (y_t - \langle x_t, \theta^\star \rangle)^2 + \tau^2 d.$$

Now we proceed with the second term in (13) exactly the same way as in (Dalalyan and Tsybakov, 2008, Theorem 5) or, more specifically, as in (Gerchinovitz, 2013, Lemma 23). We have

$$\mathcal{KL}(\mu^\star \| \mu) \leq 4 \sum_{j=1}^{d} \log \left(1 + \frac{|\theta^{\star(j)}| \sqrt{\sum_{t=1}^{T} (x_t^{(j)})^2}}{\tau} \right)$$

$$\leq 4 \|\theta^\star\|_0 \max_j \left\{ \log \left(1 + \frac{|\theta^{\star(j)}| \sqrt{\sum_{t=1}^{T} (x_t^{(j)})^2}}{\tau} \right) \right\}. \tag{14}$$

Let us now compare any s-sparse $\theta^\star$ with the solution $\theta_0 = 0$. We may assume without loss of generality that

$$\sum_{t=1}^{T} (y_t - \langle x_t, \theta^\star \rangle)^2 \leq \sum_{t=1}^{T} (y_t - \langle x_t, \theta_0 \rangle)^2 = \sum_{t=1}^{T} y_t^2. \tag{15}$$

Otherwise, θ_0, satisfying $\|\theta_0\|_0 = 0$, will make both terms on the right-hand side of (9) smaller than for a given $\theta^\star$. Thus, we restrict our attention only to those solutions satisfying (15). Now, we use the smallest singular value condition. This assumption implies

$$|\theta^{\star(j)}| \sqrt{\sum_{t=1}^{T} (x_t^{(j)})^2} \leq \|\theta^\star\|_2 \sqrt{\sum_{t=1}^{T} (x_t^{(j)})^2} \leq \frac{1}{\kappa_s} \sqrt{\sum_{t=1}^{T} (\langle x_t, \theta^\star \rangle)^2}.$$

Finally, using (15) and $|y_t| \leq m$, we have

$$\frac{1}{\kappa_s}\sqrt{\sum_{t=1}^{T}(\langle x_t, \theta^\star \rangle)^2} \leq \frac{1}{\kappa_s}\sqrt{2\sum_{t=1}^{T}\left((y_t - \langle x_t, \theta^\star \rangle)^2 + y_t^2\right)} \leq \frac{2m\sqrt{T}}{\kappa_s}.$$

Plugging this into (14) and choosing $\tau = m\sqrt{\frac{s}{d}}$ in (13), we conclude the proof. ∎

We make several remarks regarding the result. The bound of Theorem 3 depends logarithmically on $\frac{1}{\kappa_s}$. However, even if κ_s is close to zero, our bound remains meaningful as it depends at most logarithmically on $\|\theta^\star\|$ and $\max_t \|x_t\|_2$, provided both are bounded. This follows directly from upper bounding (14) when both $\theta^\star$ and all x_t-s are bounded.

We can also derive a regret upper bound independent of the parameter κ_s by combining the transductive prior in (6) with the discrete sparse prior proposed by Rigollet and Tsybakov (2012). We include this result for completeness in Appendix A. Note that for predictors corresponding to such priors no polynomial time implementation is known.

2.2 Separation between inductive and transductive settings

In this section, we present a challenging case for Vovk's online linear regression predictor in the standard online setting and prove a separation result between transductive and standard online learning.

Lower bounds for Vovk's aggregating algorithm in the standard online setup. It is open whether the transductive setup that we are considering here shares the same minimax regret bound (up to constants) with the classical inductive setup considered by Vovk (2001) and Azoury and Warmuth (2001) (see the discussion in (Gaillard et al., 2019, Section 4.1)). Namely, the minimax regret bound for the transductive setup scales with $\Theta(dm^2 \log T)$ (Gaillard et al., 2019) whereas the best known upper bound for the inductive setup scales with $O(dm^2 \log(1+rbT))$, where $\|x_t\| \leq r$ and $\|\theta^\star\| \leq b$ are given by assumption. Moreover, the upper bound for the inductive setup is obtained through the classical Vovk-Azoury-Warmuth algorithm with appropriate regularization. In the following, we show that this algorithm will incur linear regret in the inductive case when the comparator $\theta^\star$ is not assumed to be bounded. Concretely, we have the following simple result.

Proposition 7 *Let $d = 1$. Consider the Vovk-Azoury-Warmuth algorithm, which computes, upon observing x_t at time t,*

$$\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}} \left\{ \sum_{i=1}^{t-1}(y_i - x_i\theta)^2 + (x_t\theta)^2 + \lambda\theta^2 \right\},$$

where λ is the regularization fixed in advance. There exists a sequence $\{(x_t, y_t)\}_{t=1}^{T}$, where $y_t \in \{\pm 1\}$, such that

$$\sum_{t=1}^{T}(y_t - x_t\hat{\theta}_t)^2 - \inf_{\theta \in \mathbb{R}}\left\{\sum_{t=1}^{T}(y_t - x_t\theta)^2\right\} \geq \Omega(T).$$

Proof [Proof of Theorem 7] We consider two cases. In the first case, the regularization is given at a large scale $\lambda \geq T$. For this case, consider $x_t = 1$ and $y_t = 1$ for all t. Then we have

$$\inf_{\theta \in \mathbb{R}} \left\{ \sum_{t=1}^{T} (y_t - x_t\theta)^2 \right\} = T \inf_{\theta \in \mathbb{R}} (1 - \theta)^2 = 0.$$

However, at each time step t, we have

$$\hat{\theta}_t = \frac{\sum\limits_{i=1}^{t-1} x_i y_i}{\lambda + \sum\limits_{i=1}^{t} x_i^2} \leq \frac{t-1}{t+T} \leq \frac{1}{2}.$$

Then, the total loss is lower bounded by

$$\sum_{t=1}^{T} (y_t - x_t\hat{\theta}_t)^2 \geq T(1 - 1/2)^2 = T/4.$$

In the second case, the regularization is given at a relatively small scale with $\lambda < T$. We let $x_t = 2^t T$ and $y_t = (-1)^t$ for $t \geq 1$. Then, for any $t \geq 2$,

$$x_t \hat{\theta}_t = 2^t T \cdot \frac{\sum\limits_{i=1}^{t-1} x_i y_i}{\lambda + \sum\limits_{i=1}^{t} x_i^2} = 2^t T \cdot \frac{T \cdot \sum\limits_{i=1}^{t-1} (-2)^i}{\lambda + T^2 \cdot \sum\limits_{i=1}^{t} 4^i} = 2^t T^2 \cdot \frac{-2\left(1 - (-2)^{t-1}\right)/3}{\lambda + 4T^2(4^t - 1)/3}.$$

Then, we have for any $t \geq 2$,

$$\begin{aligned}
|y_t - x_t\hat{\theta}_t| &= \left| (-1)^t - 2^t T^2 \cdot \frac{-2\left(1 - (-2)^{t-1}\right)/3}{\lambda + 4T^2(4^t - 1)/3} \right| \\
&= \left| 1 + \frac{T^2 2^{t+1}\left((-1)^t + 2^{t-1}\right)/3}{\lambda + 4T^2(4^t - 1)/3} \right| \\
&\geq \left| 1 + \frac{T^2 2^{2t-1}/3}{5T^2 4^t/3} \right| \geq 11/10,
\end{aligned}$$

where the second last inequality is due to the fact that for $t \geq 2$, we have $4^t - 1 \geq \frac{1}{2}4^t$ and also $\lambda < T \leq \frac{T^2 4^t}{12}$, so $\lambda + \frac{4T^2(4^t - 1)}{3} \leq \frac{5T^2 4^t}{3}$. This establishes the lower bound of

$$\sum_{t=1}^{T} (y_t - x_t\hat{\theta}_t)^2 - \inf_{\theta \in \mathbb{R}} \left\{ \sum_{t=1}^{T} (y_t - x_t\theta)^2 \right\} \geq \sum_{t=2}^{T} (y_t - x_t\hat{\theta}_t)^2 - \sum_{t=1}^{T} y_t^2$$
$$\geq (11/10)^2(T-1) - T \geq \Omega(T),$$

where the first inequality is by $\inf_{\theta \in \mathbb{R}} \left\{ \sum_{t=1}^{T} (y_t - x_t\theta)^2 \right\} \leq \sum_{t=1}^{T} y_t^2$. The claim follows. ∎

15

Although our result does not resolve the question in (Gaillard et al., 2019, Section 4.1), it hints, in view of Theorem 1, at a separation between the transductive setup and the inductive setup. In the following section, we show that the separation indeed exists for the clipped linear function class.

Separation for clipped linear class. We now investigate the clipped linear class and demonstrate a separation between transductive online learning and inductive setups when dealing with unbounded covariates and comparators. Recall that for the linear class, the learner can compare with the optimal loss of

$$\inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 \right\}$$

with at most a regret of $O(dm^2 \log T)$ as indicated in (5). In fact, in the transductive setup even more is possible, as the learner can compare with an even smaller loss obtained by the clipped linear class, that is,

$$\inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} (y_t - \mathrm{clip_m} \left(\langle x_t, \theta \rangle \right))^2 \right\} \tag{16}$$

with a regret upper bound of $O(dm^2 \log T)$ following a covering argument, which we detail in Appendix D. To show a separation between the transductive setup and the inductive setup, we prove that any algorithm must suffer linear regret of $\Omega(T)$ in the worst case in the inductive setting for the clipped linear class, i.e., comparing to the optimal loss achieved by the clipped linear class as in (16). Concretely, we have the following result.

Proposition 8 *Let $d = 1$ and $m = 1$. For any (inductive) algorithm outputting $\hat{y}_1, \ldots, \hat{y}_T$ as predictions, there exists a sequence $\{(x_t, y_t)\}_{t=1}^{T}$, where $y_t \in \{\frac{1}{8}, \frac{7}{8}\}$ such that*

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 - \inf_{\theta \in \mathbb{R}} \left(\sum_{t=1}^{T} (y_t - \mathrm{clip_m} \left(x_t \theta \right))^2 \right) \geq \Omega(T).$$

Proof [Proof of Theorem 8] With a slight abuse of the notation, denote $x^{(i)} = 10^i$ and $\theta^{(i)} = \frac{15}{16x^{(i)}}$ for $i \in \{0, 1, \ldots, 2^T - 2\}$. We have

$$\begin{cases} \theta^{(i)} x^{(j)} > 7/8 & \text{if } j \geq i, \\ \theta^{(i)} x^{(j)} < 1/8 & \text{if } j < i. \end{cases}$$

This thresholding structure enables the construction of a sequence $(x_1, y_1), \ldots, (x_T, y_T)$ and a parameter $\widetilde{\theta}$ such that for any predictions $\hat{y}_1, \ldots, \hat{y}_T$ and for all $t \in [T]$,

$$\begin{cases} x_t \widetilde{\theta} > y_t = 7/8 & \text{if } \hat{y}_t \leq 1/2, \\ x_t \widetilde{\theta} < y_t = 1/8 & \text{if } \hat{y}_t > 1/2 \end{cases} \tag{17}$$

16

with the following binary searching construction. To start with, we set the left and the right end to be respectively $a_1 = 0$ and $b_1 = 2^T - 2$. At each round t, the adversary presents the midpoint index between a_t and b_t, that is, $i_t = \lfloor \frac{a_t + b_t}{2} \rfloor$ and $x_t = x^{(i_t)}$. Then if $\hat{y}_t \leq 1/2$, then we set $y_t = 7/8$, $a_{t+1} = i_t + 1$, and $b_{t+1} = b_t$, else we set $y_t = 1/8$, $a_{t+1} = a_t$, and $b_{t+1} = i_t - 1$. At the beginning, we have $n_1 = 2^T - 1$ covariates in between a_1 and b_1 and each time we are left with $(n_t - 1)/2$ covariates in between a_{t+1} and b_{t+1}. Then by induction, $n_t = 2^{T-t+1} - 1$. Thus, $n_T = 1$. Finally, it is clear that $\widetilde{\theta} = \theta^{(i_T)}$ satisfies (17). Then, with the sequence of $x_t, y_t, \hat{y}_t$ and $\widetilde{\theta}$ satisfying (17), since $|y_t - \hat{y}_t| > 3/8$ and $\left| y_t - \text{clip}_{\text{m}} \left(x_t \widetilde{\theta} \right) \right| < 1/8$, this implies the following lower bound on the regret

$$\sum_{t=1}^{T} (y_t - \hat{y}_t)^2 - \inf_{\theta \in \mathbb{R}} \left(\sum_{t=1}^{T} (y_t - \text{clip}_{\text{m}} (x_t \theta))^2 \right) \geq 9T/64 - \sum_{t=1}^{T} \left(y_t - \text{clip}_{\text{m}} \left(x_t \widetilde{\theta} \right) \right)^2$$
$$\geq 9T/64 - T/64 = T/8 = \Omega(T).$$

Thus, the claim follows. ∎

3. Logistic Regression

In this section, we introduce our main result for logistic regression with exponential weights algorithm via transductive priors. Concretely, we establish an assumption-free tight guarantee for logistic regression in the transductive online setting. The prior is carefully tailored to the logistic case as a mixture of the multivariate Gaussian prior defined in (6) for a subset of the covariates together with the Vapnik-Chervonenkis-style argument to select these covariates. We will discuss more the intuition behind the choice of the prior as we develop our result.

For the transductive setting of log loss, it is well-known that the exact minimax risk is characterized by the Shtarkov sum (Shtarkov, 1987; Wu et al., 2022), and this exact minimax risk is achieved by the Normalized Maximum Likelihood (NML) algorithm. However, the Shtarkov sum is difficult to interpret, and the NML algorithm is computationally prohibitive, with the computational complexity scaling as $\Omega(2^T)$ in the case of logistic regression. In terms of estimating the Shtarkov sum, Jacquet et al. (2022) notably proved that for the logistic loss, the Shtarkov sum scales as $\frac{d}{2} \log \frac{T}{2\pi} + \log \int \sqrt{\det(J_F(\theta; x_1, \ldots, x_T))} \, d\theta + o(1)$ when the covariates $x_1, \ldots, x_T$ are selected from a finite set of size $N = o(\sqrt{T})$ and dimension $d \leq N$, where $J_F(\theta; x_1, \ldots, x_T)$ represents the Jeffreys prior given $x_1, \ldots, x_T$. This result leaves the interpretation of the Jeffreys prior term open for general covariates $x_1, \ldots, x_T$. For the special case where the covariates are uniformly sampled from the unit sphere, an asymptotic upper bound of $O(d \log T)$ for the Shtarkov sum can be inferred. Recent work by Drmota, Jacquet, Wu, and Szpankowski (2024) claims to achieve an upper bound of $O(d \log(T/d))$ through a covering argument, which we formally detail in Appendix D but with worse constants in the regret bound. However, the covering-based argument lacks implementable algorithms for constructing the covering.

Meanwhile, the sequential learning of logistic loss in the inductive setting has been extensively studied, both statistically and computationally, with various boundedness assump-

tions on the covariates $x_1, \ldots, x_T$ and the optimal comparator $\theta^\star$ (see Kakade and Ng (2004); Foster et al. (2018); Shamir (2020); Jézéquel et al. (2021) and references therein). However, it is crucial to obtain a regret bound without such assumptions, as, for instance, in the simplest case of linearly separable data, the norm of $\theta^\star$ can be infinite.

We provide the first non-asymptotic, tight upper bound on the regret for transductive logistic regression without any assumptions on the covariates $x_1, \ldots, x_T$ or the optimal comparator $\theta^\star$. Importantly, in Section 5, we discuss practical implementable versions of our algorithms that are guaranteed to converge to the optimal solution, with the only caveat that the number of iterations required may grow exponentially with the dimension in the worst-case scenario. In the specific case of random design logistic regression, we achieve in Section 6.2 a truly polynomial time implementation.

3.1 Background for exponential weights algorithm for logistic loss

Similar to the mixability of the squared loss (Theorem 6), the mixability of logistic loss facilitates the analysis of exponential weights algorithms (Vovk, 1990), or equivalently, Bayesian learning algorithms (Kakade and Ng, 2004).

For logistic loss defined as $\ell_\theta(x, y) = -\log(\sigma(y\langle x, \theta \rangle))$, taking the (improper) probabilistic prediction $\hat{p}(y \mid x) = \mathbb{E}_{\theta \sim \rho} \sigma(y\langle x, \theta \rangle)$, we have

$$-\log \hat{p}(y \mid x) = -\log \mathbb{E}_{\theta \sim \rho} \exp(-\ell_\theta(x, y)), \tag{18}$$

so the incurred loss equals the mix-loss (with $\eta = 1$). The Bayesian learning algorithm combined with the Jeffreys prior has been used to establish asymptotic optimality for parametric cases under the assumption that maximum likelihood estimates satisfy the Central Limit Theorem (Rissanen, 1996). A simpler presentation can be found in Polyanskiy and Wu (2024, Section 13.4). These rest on the celebrated asymptotic Laplace approximation which states the right-hand-side of (11) for any prior μ is asymptotically

$$
\begin{aligned}
&-\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \mu} \exp \left(-\sum\nolimits_{t=1}^{T} \eta \ell_\theta(x_t, y_t) \right) \right) \\
&\approx -\frac{1}{\eta} \log \int \mu(\theta^\star) \exp \left(-\sum_{t=1}^{T} \eta \ell_{\theta^\star}(x_t, y_t) - \frac{1}{2}(\theta - \theta^\star)^\top \mathsf{H}_t(\theta^\star)(\theta - \theta^\star) \right) d\theta,
\end{aligned}
\tag{19}
$$

where $\theta^\star := \arg\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} \ell_\theta(x_t, y_t)$ and $\mathsf{H}_t(\theta^\star) = \frac{x_t x_t^\top}{(1+\exp((\theta^\star)^\top x_t))(1+\exp(-(\theta^\star)^\top x_t))}$ is the Hessian matrix of $\ell_\theta(x_t, y_t)$ at $\theta^\star$. Applying an exact Laplace formula for quadratic forms (Theorem 16 below) to (19) together with Theorem 4 and Eq. (18), the asymptotic logistic loss of an exponential weights algorithm with prior μ is approximately

$$-\sum_{t=1}^{T} \log \left(\mathbb{E}_{\theta \sim \rho_t}[\sigma(y_t \langle x_t, \theta \rangle)] \right) \approx \sum_{t=1}^{T} \ell_{\theta^\star}(x_t, y_t) + \frac{d}{2} \log \frac{T}{2\pi} + \log \left(\frac{\sqrt{\det\left(\frac{1}{T} \sum_{t=1}^{T} \mathsf{H}_t(\theta^\star) \right)}}{\mu(\theta^\star)} \right). \tag{20}$$

The above approximation motivates the choice of a prior density $\mu(\theta) \propto \sqrt{\det\left(\frac{1}{T}\sum_{t=1}^{T}\mathsf{H}_t(\theta)\right)}$ which is the Jeffreys prior. The right-hand side of (20) is also obtained as an upper bound for the Shtarkov sum (thus for the normalized maximum likelihood algorithm) by Jacquet et al. (2022). However, for the purpose of non-asymptotic analysis, which is the focus of this paper, the Jeffreys prior for the logistic loss can be computed but is hard to work with, as it is proportional to

$$\mu(\theta) \propto \sqrt{\det\left(\frac{1}{T}\sum_{t=1}^{T}\frac{x_t x_t^\top}{(1+\exp(\theta^\top x_t))(1+\exp(-\theta^\top x_t))}\right)}.$$

Instead, Kakade and Ng (2004) employs a Gaussian prior and obtains an upper bound through quadratic approximation of the logistic loss. Concretely, by choosing the prior distribution $\mu = \mathcal{N}(0, \nu^2 I)$ to be gaussian for some $\nu > 0$ to be specified, by Theorems 4 and 5 and Eq. (18), the loss of the exponential weights algorithm satisfies

$$-\sum_{t=1}^{T}\log\left(\mathbb{E}_{\theta\sim\rho_t}[\sigma(y_t\langle x_t,\theta\rangle)]\right) = \inf_{\varphi}\left(-\sum_{t=1}^{T}\mathbb{E}_{\theta\sim\varphi}[\log(\sigma(y_t\langle x_t,\theta\rangle))] + \mathcal{KL}(\varphi\|\mu)\right).$$

By choosing $\varphi = \mathcal{N}(\theta^\star, \varepsilon^2 I)$ for some $\varepsilon > 0$, together with $|\partial^2(-\log\sigma(y_t z))/\partial z^2| \leq 1/4$, the quadratic approximation of $-\mathbb{E}_{\theta\sim\varphi}\sum_{t=1}^{T}\log\sigma(y_t\langle x_t,\theta\rangle)$ near $\theta^\star$ is upper bounded by

$$-\mathbb{E}_{\theta\sim\varphi}\sum_{t=1}^{T}\log\sigma(y_t\langle x_t,\theta\rangle) \leq -\sum_{t=1}^{T}\log\sigma(y_t\langle x_t,\theta^\star\rangle) + \varepsilon^2\sum_{t=1}^{T}\|x_t\|^2/8.$$

This approximation is accurate up to an additive constant with the assumption that $\|x_t\| \leq 1$ and ε chosen to be $O(1/T)$. The difficulty of the choice of a Gaussian prior arises in the KL-divergence term, which is

$$\mathcal{KL}(\varphi\|\mu) = d\log\nu + \frac{1}{2\nu^2}\left(\|\theta^\star\|^2 + d\varepsilon^2\right) - \frac{d}{2} - d\log\varepsilon.$$

In fact, $\theta^\star$ can be unbounded in the logistic regression case. Moreover, even $\min_t|\langle\theta^\star, x_t\rangle|$ can be unbounded, as can be seen from the realizable case where all x_t's have the same label. Indeed, since a lower bound for the inductive case is established by Foster et al. (2018, Theorem 5) in the following theorem, it is not possible to obtain our result with any prior that does not depend on $x_1, \ldots, x_T$.

Theorem 9 (Foster et al. (2018)) *Consider the binary logistic regression problem over the class of linear predictors with parameter set $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq b\}$ with $b = \Omega(\sqrt{d}\log(T))$. Then for any algorithm for prediction with the binary logistic loss, there is a sequence of examples $(x_t, y_t) \in \mathbb{R}^d \times \{-1, 1\}$ for $t \in [T]$ with $\|x_t\|_2 \leq 1$ such that the regret of the algorithm is*

$$\Omega\left(d\log\left(\frac{b}{\sqrt{d}\log(T)}\right) \wedge T\right).$$

This theorem, in conjunction with our upper bound Theorem 13, establishes a learnability separation in terms of assumptions between the transductive and inductive settings for logistic loss. Specifically, it demonstrates that the transductive setting is crucial for achieving assumption-free regret bounds. We note that Dzhamtyrova and Kalnishkan (2021) obtained an asymptotically tight lower bound for a related multiclass softmax problem under KL loss with the norm in the bound; while their setting differs, it would be interesting to see if similar ideas improve Theorem 9.

3.2 Slab-experts

As discussed, in logistic regression, both $\theta^\star$ and $\min_t |\langle \theta^\star, x_t \rangle|$ can be unbounded. To address this difficulty, we introduce the concept of *slab-experts*, which operate by modifying the Gaussian prior to avoid dependence on $\|\theta^\star\|$ or $\max_t \langle x_t, \theta^\star \rangle$. Specifically, a slab-expert, based on its own belief regarding a slab, partitions the set of covariates into two regions: the large margin region and the small margin region. On the small margin covariates, the slab-expert applies exponential weights using the transductive Gaussian prior restricted to those covariates. On the large margin covariates, the slab-expert deterministically predicts the label according to the slab. Before detailing the slab-expert methodology, we first present a simple comparison lemma that controls the margin when the prediction is incorrect. The lemma is stated as follows.

Lemma 10 *For any sequence $\{(x_t, y_t)\}_{1 \leq t \leq T}$ where $(x_t, y_t) \in \mathbb{R}^d \times \{\pm 1\}$ for all $1 \leq t \leq T$, suppose $\theta^\star$ is the optimal comparator in hindsight, i.e.,*

$$\theta^\star := \arg\min_{\theta \in \mathbb{R}^d} -\sum_{t=1}^T \log(\sigma(y_t \langle x_t, \theta \rangle)). \tag{21}$$

For this comparator[2] and any $1 \leq t \leq T$, if $y_t \langle x_t, \theta^\star \rangle \leq 0$, then $|\langle \theta^\star, x_t \rangle| \leq T \log(2)$.

Proof [Proof of Theorem 10] For any t,

$$-\log(\sigma(y_t \langle x_t, \theta^\star \rangle)) \leq \sum_{t=1}^T -\log(\sigma(y_t \langle x_t, \theta^\star \rangle)) \leq \sum_{t=1}^T -\log(\sigma(0)) = T \log 2,$$

where the second inequality is by comparing $\theta^\star$ to $0 \in \mathbb{R}^d$. Thus,

$$|\langle x_t, \theta^\star \rangle| \leq \log(1 + \exp(|\langle x_t, \theta^\star \rangle|)) = -\log(\sigma(y_t \langle x_t, \theta^\star \rangle)) \leq T \log 2.$$

Thus, the claim follows. ∎

Theorem 10 states that for any x_t wrongly predicted covariate vector by $\theta^\star$, the margin $|\langle \theta^\star, x_t \rangle|$ will not exceed T. This motivates the following separation of the covariates: either $|\langle \theta^\star, x_t \rangle| \leq T$ or $y_t \langle \theta^\star, x_t \rangle > T$.

---

2. The parameter $\theta^\star$ is not always well-defined in $\mathbb{R}^d$. For instance, when the sequence is realizable, $\theta^\star$ will tend to be infinite in the separating direction. Since this is a minor technical issue, we avoid the complication in this paper. For a more rigorous treatment, we refer to Mourtada and Gaïffas (2022).

In the following, we first introduce a partial exponential weights algorithm that depends on any chosen subsets of the covariates that only predicts on the chosen subsets. Then, we use these partial exponential weights algorithms to build slab-experts.

Definition 11 (Partial exponential weights) *For any subset (potentially a multiset) $K \subset \{x_t\}_{t \in [T]}$, regularization parameter $\alpha > 0$, loss function $\ell_\theta(x, y)$ for $\theta, x \in \mathbb{R}^d$ and $y \in \{\pm 1\}$, and learning rate η, we use $\text{EWA}_{K,\alpha,\ell,\eta}$ to denote the partial exponential weights algorithm with loss function ℓ restricted to predictions on covariates in K with prior*

$$\mu_{K,\alpha}(\theta) := \left(\frac{\alpha}{\pi}\right)^{d/2} \sqrt{\det\left(\sum_{x \in K} xx^\top\right)} \exp\left(-\alpha \theta^\top \left(\sum_{x \in K} xx^\top\right) \theta\right).$$

We omit the dependence on the loss function and the learning rate when they are clear from the context.

Now, we introduce the construction of a slab-expert and present our main theorem.

Definition 12 (Slab-expert) *A slab-expert $B_{\theta,\alpha}$ is an expert that is specified by a vector $\theta \in \mathbb{R}^d$ and a regularization parameter $\alpha > 0$. Given $x_1, \ldots, x_T$, the expert constructs a slab $K_\theta = \{x_t : |\langle x_t, \theta \rangle| \leq T\}$ and predicts labels in the following way: At time step t, if $x_t \in K_\theta$, the expert predicts on x_t according to $\text{EWA}_{K_\theta,\alpha}$ with logistic loss and learning rate 1. If $x_t \notin K_\theta$, then the expert predicts $\hat{y}_t = \text{sign}(\langle x_t, \theta \rangle)$ with probability 1. We denote by $B_{\theta,\alpha}(x_t, y_t)$ the probability of predicting y_t upon observing x_t.*

Algorithm 1 Aggregation over slab-experts

Require: The set of (unordered) covariates $\{x_t\}_{t \in [T]}$ and let $\alpha = 1/T^3$.

  Construct the set of all unique slab-experts $\mathcal{A} = \{B_{\theta,\alpha} \mid \theta \in \mathbb{R}^d\}$ (removing copies) with logistic loss and learning rate 1.

  for $t = 1, 2, \ldots, T$ **do**

    The learner receives x_t.

    The learner predicts with aggregation on the slab-experts

$$\tilde{\rho}_t(x_t, \cdot) = \sum_{B \in \mathcal{A}} p_t(B) B(x_t, \cdot) \quad \text{where} \quad p_t(B) \propto \exp\left(\sum_{i=1}^{t-1} \log B(x_i, y_i)\right) \tag{22}$$

    and $B(x_i, y_i)$ is the probability of B predicting y_i on x_i.

    Nature reveals y_t.

    Update all B in $\mathcal{A}$ with (x_t, y_t).

  end for

Theorem 13 *The estimators $\tilde{\rho}_1, \ldots, \tilde{\rho}_T$ produced by Algorithm 1 with logistic loss and learning rate 1 satisfy the total loss upper bound of*

$$-\sum_{t=1}^{T} \log \tilde{\rho}_t(x_t, y_t) \leq -\sum_{t=1}^{T} \log(\sigma(y_t \langle x_t, \theta^\star \rangle)) + 6d \log(eT),$$

where $\theta^\star \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} -\log \sigma(y_t \langle x_t, \theta \rangle)$.

There are several noteworthy aspects of this result. First and foremost, it provides the first assumption-free guarantee, meaning that no assumptions are required on the covariates $x_1, \ldots, x_T$ or the comparator $\theta^\star$. In Appendix D, we will present an alternative covering approach that achieves the same upper bound, where the covering is constructed in a manner analogous to the slab division on the covariates. However, our focus remains on the slab-experts approach, as the covering approach lacks implementable algorithms for constructing the covering. Indeed, our algorithm can be implemented in time polynomial in T, though exponential in d in the worst case. This efficiency stems from the slab-experts themselves, each of which consists of an exponential weight algorithm with a Gaussian prior and a deterministic linear separator. The exponential weight algorithm with a Gaussian prior can be computed in polynomial time through log-concave sampling, and the aggregation of the slab-experts can be performed in $\text{poly}(T^d)$ time. Importantly, under mild additional assumptions on the distribution in the i.i.d. batch setup, we can reduce the number of experts to 2 with pretraining, providing a polynomial time algorithm. For further details, see Sections 5 and 6. Additionally, the slab-experts approach has two notable aspects:

1. No clipping is needed for unbounded losses when computing exponential weights, unlike in Gerchinovitz (2013) or the covering argument in Appendix D, where clipping would disrupt the convexity and computational feasibility of sampling the slab-experts.

2. Despite using two layers of exponential weights-one for slab-experts and another for covariates-the algorithm remains a Bayesian learning algorithm with a single prior, demonstrating that such an algorithm can achieve the tight assumption-free regret upper bound.

The analysis of the algorithm relies on three key facts. First, the number of slab-experts is at most $\text{poly}(T^d)$. Second, the optimal slab-expert $B_{\theta^\star,\alpha}$ with $\alpha = 1/T^3$ incurs a regret of no more than $O(d \log T)$. Finally, the aggregation over the slab-experts results in an additional regret compared to the optimal slab-expert of at most $O(\log(1/\mu(B_{\theta^\star,\alpha})))$, where μ is a prior over any set of experts that includes the optimal slab-expert $B_{\theta^\star,\alpha}$. Consequently, using a uniform prior over the slab-experts, as in Algorithm 1, guarantees a regret upper bound of $O(d \log T)$. We first derive an upper bound on the number of slab-experts.

Lemma 14 *If $T \geq d$, then there are at most $(eT/d)^{4d}$ distinct slab-experts.*

Proof [Proof of Theorem 14] By the Vapnik-Chervonenkis-Sauer-Shelah lemma (Shalev-Shwartz and Ben-David (2014, Lemma 6.10)), the number of all possible linear separations on the set $\{x_t\}_{t \in [T]}$ is upper bounded by $(eT/(d+1))^{(d+1)} \leq (eT/d)^{2d}$, whenever $T \geq d$. Therefore, since slab-experts can be formulated as intersections of two linear separations on the set $\{x_t\}_{t \in [T]}$, the number of slab-experts is upper bounded by $(eT/d)^{4d}$. ∎

We will upper-bound the loss for the slab-expert $B_{\theta^\star,\alpha}$ and compare the performance of Algorithm 1 to that of the slab-expert $B_{\theta^\star,\alpha}$. To do this, we first prove the following performance guarantee for the partial exponential weights algorithm $\text{EWA}_{K,\alpha}$, where K is a subset of the covariate set and α is a regularization parameter. For simplicity, let y_x be the labeling for $x \in \{x_t\}_{t \in [T]}$, that is $y_{x_t} = y_t$.

Proposition 15 *Let $K \subset \{x_t\}_{t \in [T]}$ and $\theta^\star \in \mathbb{R}^d$. Then the partial exponential weights algorithm $\mathrm{EWA}_{K,\alpha}$ as defined in Theorem 11 with prediction restricted on covariates in K satisfies a loss upper bound of*

$$- \sum_{x \in K} \log \left(\mathbb{E}_{\theta \sim \rho_x}[\sigma(y_x \langle x, \theta \rangle)] \right)$$

$$\leq - \sum_{x \in K} \log(\sigma(y_x \langle x, \theta^\star \rangle)) + \alpha \, (\theta^\star)^\top \left(\sum_{x \in K} xx^\top \right) \theta^\star + \frac{d}{2} \log \left(1 + \frac{1}{8\alpha} \right)$$

on the set K where ρ_x is the posterior distribution (over θ) used by $\mathrm{EWA}_{K,\alpha}$ on x. Specifically, suppose that $\theta^\star$ as defined in Eq. (21) satisfies $|\langle x, \theta^\star \rangle| \leq T$ for all $x \in K$ and let $\alpha = 1/T^3$. We have

$$- \sum_{x \in K} \log \left(\mathbb{E}_{\theta \sim \rho_x}[\sigma(y_x \langle x, \theta \rangle)] \right) \leq - \sum_{x \in K} \log(\sigma(y_x \langle x, \theta^\star \rangle)) + 2d \log T.$$

Before providing the proof, we present the following standard result that computes exponential quadratic integrals based on the multivariate Gaussian integral.

Lemma 16 (Exact Laplace formula for quadratic forms) *Let $Q(\theta) = \theta^\top A \theta + b^\top \theta + c$, where $A \in \mathbb{R}^{d \times d}$ is a positive definite matrix, $b \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Then*

$$\int_{\mathbb{R}^d} \exp(-Q(\theta)) d\theta = \exp \left(- \inf_{\theta \in \mathbb{R}^d} Q(\theta) \right) \frac{\pi^{d/2}}{\sqrt{\det(A)}}.$$

Proof [Proof of Theorem 15] By Theorem 4, we have

$$- \sum_{x \in K} \log \left(\mathbb{E}_{\theta \sim \rho_x}[\sigma(y_x \langle x, \theta \rangle)] \right) = - \log \left(\int_{\mathbb{R}^d} \exp \left(\sum_{x \in K} \log(\sigma(y_x \langle x, \theta \rangle)) \right) \mu_{K,\alpha}(d\theta) \right).$$

Let $f(z) = - \log(\sigma(z))$. Since $|f''(z)| \leq 1/4$, using the Taylor expansion for f we have for any $x \in K$,

$$- \log(\sigma(y_x \langle x, \theta \rangle)) \leq - \log(\sigma(y_x \langle x, \theta^\star \rangle)) - \frac{1}{1 + \exp(y_x \langle x, \theta^\star \rangle)} \cdot (y_x \langle x, \theta - \theta^\star \rangle) + \frac{1}{8}(\langle x, \theta - \theta^\star \rangle)^2.$$

Summing these inequalities together, we have

$$- \sum_{x \in K} \log(\sigma(y_x \langle x, \theta \rangle)) \leq - \sum_{x \in K} \log(\sigma(y_x \langle x, \theta^\star \rangle)) - \sum_{x \in K} \frac{1}{1 + \exp(y_x \langle x, \theta^\star \rangle)} \cdot (y_x \langle x, \theta - \theta^\star \rangle)$$

$$+ \frac{1}{8}(\theta - \theta^\star)^\top \left(\sum_{x \in K} xx^\top \right) (\theta - \theta^\star). \tag{23}$$

Thus, we have

$$
-\log\left(\int_{\mathbb{R}^d} \exp\left(\sum_{x\in K} \log(\sigma(y_x\langle x, \theta\rangle))\right) \mu_{K,\alpha}(d\theta)\right)
$$

$$
\leq -\sum_{x\in K} \log(\sigma(y_x\langle x, \theta^\star\rangle)) - \log\left(\int_{\mathbb{R}^d} \exp\left(\sum_{x\in K} \frac{1}{1+\exp(y_x\langle x, \theta^\star\rangle)} \cdot (y_x\langle x, \theta-\theta^\star\rangle)\right.\right.
$$

$$
\left.\left. -\frac{1}{8}\cdot(\theta-\theta^\star)^\top\left(\sum_{x\in K} xx^\top\right)(\theta-\theta^\star)\right)\mu_{K,\alpha}(d\theta)\right).
$$

Using Theorem 16 to compute the integral of the quadratic form we have

$$
-\log\left(\int_{\mathbb{R}^d} \exp\left(\sum_{x\in K} \frac{1}{1+\exp(y_x\langle x, \theta^\star\rangle)} \cdot (y_x\langle x, \theta-\theta^\star\rangle)\right.\right.
$$

$$
\left.\left. -\frac{1}{8}\cdot(\theta-\theta^\star)^\top\left(\sum_{x\in K} xx^\top\right)(\theta-\theta^\star)\right)\mu_{K,\alpha}(d\theta)\right)
$$

$$
= -\log\left(\left(\frac{\alpha}{\pi}\right)^{\frac{d}{2}} \cdot \left(\frac{\pi}{\alpha+\frac{1}{8}}\right)^{\frac{d}{2}} \exp\left(-\inf_{\theta\in\mathbb{R}^d}\left(\alpha\theta^\top\left(\sum_{x\in K} xx^\top\right)\theta\right.\right.\right.
$$

$$
+\frac{1}{8}(\theta-\theta^\star)^\top\left(\sum_{x\in K} xx^\top\right)(\theta-\theta^\star)
$$

$$
\left.\left.\left.-\sum_{x\in K} \frac{1}{1+\exp(y_x\langle x, \theta^\star\rangle)} \cdot (y_x\langle x, \theta-\theta^\star\rangle)\right)\right)\right)
$$

$$
\leq \alpha\,(\theta^\star)^\top\left(\sum_{x\in K} xx^\top\right)\theta^\star + \frac{d}{2}\log\left(1+\frac{1}{8\alpha}\right),
$$

where the last step is by taking $\theta=\theta^\star$. This concludes our proof. ∎

Proof [Proof of Theorem 13] We first upper bound the total loss by the loss of slab-expert $B_{\theta^\star,\alpha}$ by the definition of the exponential weights algorithm,

$$
\begin{aligned}
-\sum_{t=1}^{T} \log \tilde{\rho}_t(x_t, y_t) &= -\sum_{t=1}^{T} \log \left(\frac{\sum\limits_{B \in \mathcal{A}} \exp\left(\sum\limits_{i=1}^{t} \log B(x_i, y_i) \right)}{\sum\limits_{B \in \mathcal{A}} \exp\left(\sum\limits_{i=1}^{t-1} \log B(x_i, y_i) \right)} \right) \\
&= -\log \left(\frac{1}{|\mathcal{A}|} \sum_{B \in \mathcal{A}} \exp\left(\sum_{i=1}^{T} \log B(x_i, y_i) \right) \right) \\
&\leq -\sum_{i=1}^{T} \log B_{\theta^\star,\alpha}(x_i, y_i) + \log |\mathcal{A}| \\
&\leq -\sum_{i=1}^{T} \log B_{\theta^\star,\alpha}(x_i, y_i) + 4d \log(eT/d),
\end{aligned}
\tag{24}
$$

where the last inequality is due to Theorem 14. For the slab-expert $B_{\theta^\star,\alpha}$ with $\alpha = 1/T^3$, it is clear that inside the slab $K_{\theta^\star}$ as defined in Theorem 12, the assumption of Theorem 15 is satisfied, thus

$$
-\sum_{x \in K_{\theta^\star}} \log B_{\theta^\star,\alpha}(x, y_x) \leq -\sum_{x \in K_{\theta^\star}} \log(\sigma(y_x \langle x, \theta^\star \rangle)) + 2d \log T.
\tag{25}
$$

Meanwhile, for all the covariates outside the slab, the prediction of $B_{\theta^\star,\alpha}$ coincides with the linear separator with vector $\theta^\star$. By Lemma 10, if $y_x \langle x, \theta^\star \rangle \leq 0$ then $|\langle x, \theta^\star \rangle| \leq T \log 2 < T$, hence such x must belong to $K_{\theta^\star}$. Therefore for $x \notin K_{\theta^\star}$ we have $y_x \langle x, \theta^\star \rangle > 0$ and the expert predicts correctly with probability 1. Thus, we have as a result

$$
-\sum_{x \in \{x_t\}_{t \in [T]} \setminus K_{\theta^\star}} \log B_{\theta^\star,\alpha}(x, y_x) = 0.
\tag{26}
$$

Plug (25) and (26) back into (24), we obtain

$$
-\sum_{t=1}^{T} \log \tilde{\rho}_t(x_t, y_t) \leq -\sum_{t=1}^{T} \log(\sigma(y_t \langle x_t, \theta^\star \rangle)) + 6d \log(eT).
$$

The claim follows. ∎

4. Non-curved losses

In the previous sections, we focused on mixable losses, such as the logistic loss and the squared loss. These losses are generally known to yield $O(\log T)$-type regret bounds. We now turn our attention to another unbounded loss function, namely the hinge loss, which lacks mixability properties and thus compels us to focus on the so-called *first-order* regret bounds. These bounds scale as $O(\sqrt{T})$ in the worst case but can improve if the total loss of the optimal solution is close to zero.

4.1 Regret bounds for hinge loss

In this section, we first obtain new margin-dependent bounds for the hinge loss in the sequential learning setup using an exponential weight algorithm; then, we proceed to use a transductive prior to achieve an assumption-free regret bound. In the classification setup with classes $1, -1$, given $\gamma > 0$, we define the normalized hinge loss of a predictor $\widehat{f}$ as follows

$$\frac{(\gamma - y\widehat{f}(x))_+}{\gamma}.$$

There is a standard connection with the binary loss. Indeed,

$$\mathbb{1}(\text{sign}(\widehat{f}(x)) \neq y) \leq \frac{(\gamma - y\widehat{f}(x))_+}{\gamma}. \tag{27}$$

In the agnostic case where the design vectors are not known in advance, the state-of-the-art regret bound is given by the *second order perceptron* algorithm of Cesa-Bianchi, Conconi, and Gentile (2005). Denoting the output of this predictor with regularization parameter $\lambda > 0$ at round t by $\widehat{y}_t \in \{1, -1\}$ (see (Cesa-Bianchi and Lugosi, 2006, Section 12.2) for the exact definition of this predictor), we have for any $\theta^\star \in \mathbb{R}^d$,

$$\sum_{t=1}^{T} \mathbb{1}(\widehat{y}_t \neq y_t) \leq \sum_{t=1}^{T} \frac{(\gamma - y_t\langle x_t, \theta^\star\rangle)_+}{\gamma} + \frac{1}{\gamma}\sqrt{\left(\lambda\|\theta^\star\|^2 + \|A_T^{1/2}\theta^\star\|^2\right)\sum_{i=1}^{d}\log\left(1 + \frac{\lambda_i}{\lambda}\right)}, \tag{28}$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of the matrix $A_T = \sum_{t=1}^{T} x_t x_t^\top \mathbb{1}(\widehat{y}_t \neq y_t)$.

This regret bound (28), although corresponding to a computationally efficient algorithm, has several features we are aiming to avoid. One of the key problems is that the bound depends on the term $\|A_T^{1/2}\theta^\star\|^2$ which scales badly with the magnitude of $\max_t |\langle x_t, \theta^\star\rangle|$. Our first result in this section corresponds to the analysis of the exponential weights algorithm, whose prior is given by the general covariance matrix. Given a positive-definite matrix $A \in \mathbb{R}^{d \times d}$ and $\lambda > 0$, set

$$\mu(\theta) = \sqrt{\det(\lambda A)} \cdot \exp\left(-\pi\lambda\theta^\top A\theta\right). \tag{29}$$

Fix $\eta > 0$. Using the general formula for the exponential weights algorithm (10), we set $\rho_1 = \mu$ and for $t \geq 2$,

$$\rho_t(\theta) = \frac{\exp\left(-\eta\sum_{i=1}^{t-1}(\gamma - y_i\langle x_i, \theta\rangle)_+\right)\mu(\theta)}{\mathbb{E}_{\theta'\sim\mu}\exp\left(-\eta\sum_{i=1}^{t-1}(\gamma - y_i\langle x_i, \theta'\rangle)_+\right)}.$$

Given the positive-definite matrix A defining the Gaussian prior (29), the prediction of our algorithm at round t is given by

$$\widehat{f}_t(x_t) = \mathbb{E}_{\theta\sim\rho_t} \text{clip}_\gamma\left(\langle x_t, \theta\rangle\right). \tag{30}$$

The computation of such prediction depends on the ability of sampling from ρ_t, which is a log-concave measure. Indeed, once we can approximately sample from ρ_t, we can use a simple estimate based on Hoeffding's inequality to identify the number of samples required from ρ_t in order to approximate $\widehat{f}_t(x_t)$ with the desired accuracy.

Proposition 17 *For any $\eta \in [0, 3/(10\gamma)]$, $\beta > 0$, and any $\theta^\star \in \mathbb{R}^d$ the following bound holds for the predictor (30):*

$$\sum_{t=1}^{T} \frac{(\gamma - y_t \widehat{f}_t(x_t))_+}{\gamma} \le (1 + 2\eta\gamma) \cdot \left(\sum_{t=1}^{T} \frac{(\gamma - y_t \langle x_t, \theta^\star \rangle)_+}{\gamma} + \frac{1}{\pi\gamma\sqrt{\beta}} \sum_{t=1}^{T} \|A^{-1/2} x_t\|_2 \right.$$
$$\left. + \frac{\pi\lambda(\theta^\star)^\top A \theta^\star}{\eta\gamma} + \frac{d}{2\eta\gamma} \left(\log\left(\frac{\beta}{\lambda}\right) - \frac{1}{2} + \frac{\lambda}{2\beta} \right) \right).$$

Proof The proof follows the standard lines. First, observe that Jensen's inequality implies

$$\sum_{t=1}^{T} (\gamma - y_t \widehat{f}_t(x_t))_+ \le \sum_{t=1}^{T} \mathbb{E}_{\theta \sim \rho_t} (\gamma - y_t \operatorname{clip}_\gamma (\langle x_t, \theta \rangle))_+. \tag{31}$$

Using Bernstein's inequality for the moment generating function and

$$\mathbb{E}_{\theta \sim \rho} (\gamma - \operatorname{clip}_\gamma (y \langle x, \theta \rangle))_+^2 \le 2\gamma \mathbb{E}_{\theta \sim \rho} (\gamma - \operatorname{clip}_\gamma (y \langle x, \theta \rangle))_+,$$

we have for any distribution ρ over $\mathbb{R}^d$,

$$\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \rho} \exp \left(-\eta(\gamma - \operatorname{clip}_\gamma (y \langle x, \theta \rangle))_+ \right) \right)$$
$$\le -\mathbb{E}_{\theta \sim \rho}(\gamma - \operatorname{clip}_\gamma (y \langle x, \theta \rangle))_+ + \frac{\eta\gamma \mathbb{E}_{\theta \sim \rho}(\gamma - \operatorname{clip}_\gamma (y \langle x, \theta \rangle))_+}{1 - 2\eta\gamma/3}.$$

Combining this inequality with (31) and using Theorem 4, we obtain

$$\sum_{t=1}^{T} \left(1 - \frac{\eta\gamma}{1 - 2\eta\gamma/3} \right) (\gamma - y_t \widehat{f}_t(x_t))_+ \le -\sum_{t=1}^{T} \frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \rho_t} \exp \left(-\eta(\gamma - \operatorname{clip}_\gamma (y_t \langle x_t, \theta \rangle))_+ \right) \right)$$
$$\le -\sum_{t=1}^{T} \frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \rho_t} \exp \left(-\eta(\gamma - y_t \langle x_t, \theta \rangle)_+ \right) \right)$$
$$= -\frac{1}{\eta} \log \left(\mathbb{E}_{\theta \sim \mu} \exp \left(-\eta \sum_{t=1}^{T} (\gamma - y_t \langle x_t, \theta \rangle)_+ \right) \right),$$

where the second inequality uses $(\gamma - \operatorname{clip}_\gamma (z))_+ \le (\gamma - z)_+$ for all z. We upper bound the last quantity using the Donsker-Varadhan formula. Fix $\beta > 0$ and let $\mu^\star$ be the multivariate Gaussian distribution with density

$$\mu^\star(\theta) = \sqrt{\det(\beta A)} \cdot \exp \left(-\pi(\theta - \theta^\star)^\top \beta A (\theta - \theta^\star) \right).$$

27

and By Theorem 5, we have

$$-\frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\mu}\exp\left(-\eta\sum_{t=1}^{T}(\gamma-y_t\langle x_t,\theta\rangle)_+\right)\right)\le\mathbb{E}_{\theta\sim\mu^\star}\sum_{t=1}^{T}(\gamma-y_t\langle x_t,\theta\rangle)_++\frac{\mathcal{KL}(\mu^\star\|\mu)}{\eta}.$$

We analyze the two terms of the right-hand side of the last inequality separately. Define the density

$$\mu_0(\theta)=\sqrt{\frac{\det(\beta A)}{\pi^d}}\exp(-\pi\theta^\top\beta A\theta).$$

Using the standard properties of the Gaussian distribution we have

$$\mathbb{E}_{\theta\sim\mu^\star}\sum_{t=1}^{T}(1-y_t\langle x_t,\theta\rangle)_+\le\sum_{t=1}^{T}(1-y_t\langle x_t,\theta^\star\rangle)_++\sum_{t=1}^{T}\mathbb{E}_{\theta\sim\mu_0}|\langle\theta,x_t\rangle|$$

$$=\sum_{t=1}^{T}(1-y_t\langle x_t,\theta^\star\rangle)_++\frac{1}{\pi\sqrt{\beta}}\sum_{t=1}^{T}\|A^{-1/2}x_t\|_2.$$

Another standard formula implies

$$\mathcal{KL}(\mu^\star\|\mu)=\pi\lambda(\theta^\star)^\top A\theta^\star+\frac{d}{2}\left(\log\left(\frac{\beta}{\lambda}\right)-\frac{1}{2}+\frac{\lambda}{2\beta}\right).$$

Combining all the bounds together, we get

$$\sum_{t=1}^{T}\left(1-\frac{\eta\gamma}{1-2\eta\gamma/3}\right)\cdot\frac{(\gamma-y_t\widehat{f}_t(x_t))_+}{\gamma}$$

$$\le\sum_{t=1}^{T}\frac{(\gamma-y_t\langle x_t,\theta^\star\rangle)_+}{\gamma}+\frac{1}{\pi\gamma\sqrt{\beta}}\sum_{t=1}^{T}\|A^{-1/2}x_t\|_2+\frac{\pi\lambda(\theta^\star)^\top A\theta^\star}{\eta\gamma}$$

$$+\frac{d}{2\eta\gamma}\left(\log\left(\frac{\beta}{\lambda}\right)-\frac{1}{2}+\frac{\lambda}{2\beta}\right).\tag{32}$$

Finally, we observe that for any $\eta\le\frac{3}{10\gamma}$, we have

$$\left(1-\frac{\eta\gamma}{1-2\eta\gamma/3}\right)^{-1}\le1+2\eta\gamma.$$

Plugging this into (32), we prove the claim. ∎

There are some immediate corollaries of the result of Theorem 17. First, we consider the standard bounded online setup where the vectors $x_1,\dots,x_T$ are not known in advance.

Corollary 18 *Assume that $\|x_t\|\le r$ and let $\theta^\star\in\mathbb{R}^d$ be such that $\|\theta^\star\|\le b$. Then, for any $\eta\in[0,3/(10\gamma)]$, the predictor (30) with $A=I_d$ and $\lambda=\frac{1}{b^2}$ satisfies*

$$\sum_{t=1}^{T}\frac{(\gamma-y_t\widehat{f}_t(x_t))_+}{\gamma}\le(1+2\eta\gamma)\left(\sum_{t=1}^{T}\frac{(\gamma-y_t\langle x_t,\theta^\star\rangle)_+}{\gamma}+\frac{\frac{1}{\pi}+\pi+\frac{d}{2}\log\left(1+\eta^2T^2r^2b^2\right)}{\eta\gamma}\right).$$

In particular, when the data is linearly separable with margin γ, that is, for the same $\theta^\star$ we have $y_t\langle x_t, \theta^\star\rangle \geq \gamma$ for $t = 1, \ldots, T$, by fixing $\eta = 3/(10\gamma)$ we obtain

$$\sum_{t=1}^{T} \frac{(\gamma - y_t\widehat{f}_t(x_t))_+}{\gamma} \leq \frac{7}{3}\left(\frac{1}{\pi} + \pi + \frac{d}{2}\log\left(1 + \frac{9T^2r^2b^2}{100\gamma^2}\right)\right). \tag{33}$$

Proof The first claim follows from Theorem 17 if we plug in $A = I_d$, $\lambda = \frac{1}{b^2}$, $\beta = \frac{1}{b^2} + \eta^2T^2r^2$. The second part of the claim follows from the fact that in the linearly separable case with margin γ we have $\sum_{t=1}^{T}(\gamma - y_t\langle x_t, \theta^\star\rangle)_+ = 0$. The claim follows. ∎

Finally, we present our main result for the transductive setup, where we employ the transductive prior by selecting a data-dependent matrix A in (29). Additionally, we note that a similar sparsity regret bound can be provided, analogous to the one in Theorem 3.

4.2 Slab-experts for hinge loss

In this section, we show that slab experts can also be leveraged to obtain assumption-free bounds for hinge loss. To start, we present the following comparison lemma for hinge loss.

Lemma 19 *For any sequence $\{(x_t, y_t)\}_{1 \leq t \leq T}$ where $(x_t, y_t) \in \mathbb{R}^d \times \{\pm 1\}$ for all $1 \leq t \leq T$, suppose $\theta^\star$ is the optimal comparator in hindsight, i.e.,*

$$\theta^\star := \arg\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} \frac{(\gamma - y_t\langle x_t, \theta\rangle)_+}{\gamma} \tag{34}$$

For this comparator and any $1 \leq t \leq T$, if $y_t\langle x_t, \theta^\star\rangle \leq 0$, then $|\langle x_t, \theta^\star\rangle| \leq T\gamma$.

Proof [Proof of Theorem 19] For any $t \in [T]$, suppose $y_t\langle x_t, \theta^\star\rangle \leq 0$. Since $\theta^\star$ minimizes the loss, we have

$$\frac{\gamma + |\langle x_t, \theta^\star\rangle|}{\gamma} = \frac{(\gamma - y_t\langle x_t, \theta^\star\rangle)_+}{\gamma} \leq \sum_{t=1}^{T} \frac{(\gamma - y_t\langle x_t, \theta^\star\rangle)_+}{\gamma} \leq T.$$

Reorganizing terms, we obtain the desired result. ∎

Definition 20 (Slab-expert for hinge loss) *A slab-expert for the hinge loss $B_{\theta,\alpha,\gamma}$ is an expert that is specified by a vector $\theta \in \mathbb{R}^d$, a regularization parameter α, and a margin $\gamma > 0$. The expert constructs a subset of covariates that belong to a single slab, namely $K_{\theta,\gamma} = \{x_t : |\langle x_t, \theta\rangle| \leq \gamma T\}$ and predicts labels in the following way: At time step t, if $x_t \in K_{\theta,\gamma}$, the expert predicts $\widehat{f}_t(x_t)$ given by the partial EWA on $K_{\theta,\gamma}$ (regularization parameter α, hinge loss, rate η), i.e., $\widehat{f}_t(x_t) = \mathbb{E}_{\vartheta \sim \rho_t} \mathrm{clip}_\gamma(\langle x_t, \vartheta\rangle)$. If $x_t \notin K_{\theta,\gamma}$, then the expert predicts $\hat{y}_t = \gamma\,\mathrm{sign}(\langle x_t, \theta\rangle)$. We denote by $B_{\theta,\alpha,\gamma}(x_t)$ the value predicted on receiving the covariate x_t.*

Algorithm 2 Aggregation over slab-experts for the hinge loss

Require: The set of (unordered) covariates $\{x_t\}_{t \in [T]}$, margin γ, learning rate η, regularization parameter $\alpha = 1/(\gamma^2 T^3)$.

Construct the set of unique slab-experts for the hinge loss $\mathcal{A} = \{B_{\theta,\alpha,\gamma} \mid \theta \in \mathbb{R}^d\}$ with the hinge loss function $\ell_\theta(x, y) = (\gamma - y\langle x, \theta \rangle)_+$ and learning rate η.

for $t = 1, 2, \ldots, T$ **do**

The learner receives x_t.

The learner predicts with aggregation on the slab-experts

$$\widehat{f}_t(x_t) = \sum_{B \in \mathcal{A}} p_t(B) B(x_t) \text{ where } p_t(B) \propto \exp\left(\sum_{i=1}^{t-1} -\eta(\gamma - y_i B(x_i))_+\right).$$

Nature reveals y_t.

Update all B in $\mathcal{A}$ with (x_t, y_t).

end for

Theorem 21 *For any margin $\gamma > 0$, $\eta \in [0, 3/(10\gamma)]$, and $\theta^\star$ as defined in (34), the estimators $\widehat{f}_1(x_1), .., \widehat{f}_T(x_T)$ produced by Algorithm 2 satisfy the total loss upper bound of*

$$\sum_{t=1}^{T} \frac{(\gamma - y_t \widehat{f}_t(x_t))_+}{\gamma} \leq (1 + 2\eta\gamma)^2 \left(\sum_{t=1}^{T} \frac{(\gamma - y_t \langle x_t, \theta^\star \rangle)_+}{\gamma}\right) + \frac{7d}{\eta\gamma}(1 + 2\eta\gamma)^2 \log(\pi T)$$
$$+ \frac{1}{\eta\gamma}(1 + 2\eta\gamma)^2.$$

In particular, when the data is linearly separable with margin γ, that is, for the same $\theta^\star$ we have $y_t \langle x_t, \theta^\star \rangle \geq \gamma$ for $t = 1, \ldots, T$, by fixing $\eta = 3/(10\gamma)$ we obtain

$$\sum_{t=1}^{T} \frac{(\gamma - y_t \widehat{f}_t(x_t))_+}{\gamma} \leq 9 + 60d \log(\pi T). \tag{35}$$

Observe that due to (27), the bound of (35) upper-bounds the number of mistakes made by our predictor.

Proof [Proof of Theorem 21] Similar to the logistic case, we first upper bound the total loss by the loss of slab-expert $B_{\theta^\star, \alpha, \gamma}$ as defined in Theorem 20. First, observe that Jensen's inequality implies

$$\sum_{t=1}^{T} (\gamma - y_t \widehat{f}_t(x_t))_+ \leq \sum_{t=1}^{T} \mathbb{E}_{B \sim p_t}(\gamma - y_t B(x_t))_+. \tag{36}$$

Using Bernstein's inequality for the moment generating function together with $\mathbb{E}_{B \sim p_t}(\gamma - y_t B(x_t))_+^2 \leq 2\gamma \mathbb{E}_{B \sim p_t}(\gamma - y_t B(x_t))_+$, we have for any distribution p over $\mathbb{R}$,

$$\frac{1}{\eta} \log \left(\mathbb{E}_{y' \sim p} \exp\left(-\eta(\gamma - yy')_+\right)\right) \leq -\mathbb{E}_{y' \sim p}(\gamma - yy')_+ + \frac{\eta\gamma \mathbb{E}_{y' \sim p}(\gamma - yy')_+}{1 - 2\eta\gamma/3}.$$

Combining this inequality with (36) and using Lemma Theorem 4, we obtain

$$\sum_{t=1}^{T} \left(1 - \frac{\eta\gamma}{1 - 2\eta\gamma/3}\right) \cdot (\gamma - y_t \widehat{f}_t(x_t))_+$$

$$\leq -\sum_{t=1}^{T} \frac{1}{\eta} \log \left(\mathbb{E}_{B \sim p_t} \exp\left(-\eta(\gamma - y_t B(x_t))_+\right)\right)$$

$$= -\sum_{t=1}^{T} \frac{1}{\eta} \log \left(\frac{\sum_{B \in \mathcal{A}} \exp\left(-\eta \sum_{i=1}^{t} (\gamma - y_i B(x_i))_+\right)}{\sum_{B \in \mathcal{A}} \exp\left(-\eta \sum_{i=1}^{t-1} (\gamma - y_i B(x_i))_+\right)}\right)$$

$$= -\frac{1}{\eta} \log \left(\frac{1}{|\mathcal{A}|} \sum_{B \in \mathcal{A}} \exp\left(-\eta \sum_{i=1}^{T} (\gamma - y_i B(x_i))_+\right)\right)$$

$$\leq -\frac{1}{\eta} \log \left(\exp\left(-\eta \sum_{i=1}^{T} (\gamma - y_i B_{\theta^\star,\alpha,\gamma}(x_i))_+\right)\right) + \log|\mathcal{A}|$$

$$\leq \sum_{i=1}^{T} (\gamma - y_i B_{\theta^\star,\alpha,\gamma}(x_i))_+ + \frac{4d}{\eta} \log(eT/d),$$

where the last inequality is due to Theorem 14. We apply Theorem 17 to the slab-expert $B_{\theta^\star,\alpha,\gamma}$ with $\alpha = \pi\lambda = 1/(\gamma^2 T^3)$ and $A = \sum_{x \in K_{\theta^\star,\gamma}} xx^\top$ to obtain

$$\sum_{x \in K_{\theta^\star,\gamma}} \frac{(\gamma - y_x B_{\theta^\star,\alpha,\gamma}(x))_+}{\gamma}$$

$$\leq (1 + 2\eta\gamma) \cdot \left(\sum_{x \in K_{\theta^\star,\gamma}} \frac{(\gamma - y_x \langle x, \theta^\star\rangle)_+}{\gamma} + \frac{1}{\pi\gamma\sqrt{\beta}} \sum_{x \in K_{\theta^\star,\gamma}} \|A^{-1/2}x\|_2 \quad (37)\right.$$

$$\left. + \frac{\pi\lambda(\theta^\star)^\top A\theta^\star}{\eta\gamma} + \frac{d}{2\eta\gamma} \left(\log\left(\frac{\beta}{\lambda}\right) - \frac{1}{2} + \frac{\lambda}{2\beta}\right)\right).$$

Set $\beta = T/(\pi^2\gamma^2)$. We can bound the additive terms separately by

$$\frac{1}{\pi\gamma\sqrt{\beta}} \sum_{x \in K_{\theta^\star,\gamma}} \|A^{-1/2}x\|_2 = \frac{1}{\sqrt{T}} \sum_{x \in K_{\theta^\star,\gamma}} \sqrt{x^\top A^{-1}x} \leq \frac{1}{\sqrt{T}} \sqrt{T \sum_{x \in K_{\theta^\star,\gamma}} x^\top A^{-1}x} \leq \sqrt{d}$$

and

$$\frac{\pi\lambda(\theta^\star)^\top A\theta^\star}{\eta\gamma} \leq \frac{1}{\eta\gamma^3 T^3} \sum_{x \in K_{\theta^\star,\gamma}} \langle x, \theta^\star\rangle^2 \leq \frac{1}{\eta\gamma^3 T^3}\gamma^2 T^3 = \frac{1}{\eta\gamma}.$$

Plug the above inequalities to (37), we can obtain

$$\sum_{x \in K_{\theta^\star,\gamma}} \frac{(\gamma - y_x B_{\theta^\star,\alpha,\gamma}(x))_+}{\gamma}$$

$$\leq (1 + 2\eta\gamma) \cdot \left(\sum_{x \in K_{\theta^\star,\gamma}} \frac{(\gamma - y_x\langle x, \theta^\star \rangle)_+}{\gamma} + \sqrt{d} + \frac{1}{\eta\gamma} + \frac{2d}{\eta\gamma} \log(\pi T) \right)$$

$$\leq (1 + 2\eta\gamma) \cdot \left(\sum_{x \in K_{\theta^\star,\gamma}} \frac{(\gamma - y_x\langle x, \theta^\star \rangle)_+}{\gamma} + \frac{1}{\eta\gamma} + \frac{3d}{\eta\gamma} \log(\pi T) \right).$$

Meanwhile, for all the covariates outside the slab, the prediction of $B_{\theta^\star,\alpha}$ coincides with the linear separator with vector $\theta^\star$. By Lemma 19, if $y_x\langle x, \theta^\star \rangle \leq 0$ then $|\langle x, \theta^\star \rangle| \leq \gamma T$, hence such x must lie in $K_{\theta^\star,\gamma}$. Therefore outside the slab we have $y_x\langle x, \theta^\star \rangle > 0$ and the hinge loss is 0. Thus, we have as a result

$$\sum_{x \in \{x_t\}_{t \in [T]} \setminus K_{\theta^\star,\gamma}} \frac{(\gamma - y_x B_{\theta^\star,\alpha,\gamma}(x))_+}{\gamma} = 0. \tag{38}$$

Altogether, we obtain

$$\sum_{t=1}^{T} \frac{(\gamma - y_t \widehat{f}_t(x_t))_+}{\gamma}$$

$$\leq (1 + 2\eta\gamma)\left(\sum_{i=1}^{T} \frac{(\gamma - y_i B_{\theta^\star,\alpha,\gamma}(x_i))_+}{\gamma} + \frac{4d}{\eta\gamma} \log(eT/d) \right)$$

$$\leq (1 + 2\eta\gamma)\left((1 + 2\eta\gamma) \cdot \left(\sum_{x \in K_{\theta^\star,\gamma}} \frac{(\gamma - y_x\langle x, \theta^\star \rangle)_+}{\gamma} + \frac{1}{\eta\gamma} + \frac{3d}{\eta\gamma} \log(\pi T) \right) + \frac{4d}{\eta\gamma} \log(eT) \right)$$

$$\leq (1 + 2\eta\gamma)^2 \left(\sum_{t=1}^{T} \frac{(\gamma - y_t\langle x_t, \theta^\star \rangle)_+}{\gamma} \right) + \frac{7d}{\eta\gamma}(1 + 2\eta\gamma)^2 \log(\pi T) + \frac{1}{\eta\gamma}(1 + 2\eta\gamma)^2.$$

The claim follows. ∎

5. Computational Aspects

In this section, we briefly discuss the computational properties of our algorithms in the problems considered so far, namely, sparse linear regression, logistic regression, and hinge loss regression.

5.1 Computation for sparse linear regression

In this section, we discuss the computational properties of our methods in Theorem 3 and Theorem 34. Concretely, we would like to compute the quantity $\widehat{f}_t(x_t)$ as defined in (8).

We provide three possible approaches for this computation. First, set

$$L_m(\theta) = \frac{1}{2m^2}(m - \langle x_t, \theta \rangle)^2 + \frac{1}{2m^2}\sum_{i=1}^{t-1}(y_i - \langle x_i, \theta \rangle)^2,$$

$$L_{-m}(\theta) = \frac{1}{2m^2}(-m - \langle x_t, \theta \rangle)^2 + \frac{1}{2m^2}\sum_{i=1}^{t-1}(y_i - \langle x_i, \theta \rangle)^2.$$

The first approach is to notice that computing $\widehat{f}_t(x_t)$ is equivalent to computing

$$\frac{\int\limits_{\mathbb{R}^d} \exp\left(-L_m(\theta)\right)\mu(\theta)d\theta}{\int\limits_{\mathbb{R}^d} \exp\left(-L_{-m}(\theta)\right)\mu(\theta)d\theta} = \int_{\mathbb{R}^d} \frac{\exp(-V(\theta))}{\int_{\mathbb{R}^d} \exp(-V(\theta'))d\theta'} \cdot \exp\left(\frac{2}{m}\langle x_t, \theta \rangle\right)d\theta,$$

where

$$V(\theta) = L_{-m}(\theta) + 4\sum_{j=1}^{d} \log\left(1 + |\theta^{(j)}| \cdot \sqrt{\sum_{t=1}^{T}(x_t^{(j)})^2/\tau}\right).$$

Thus, we can sample from the distribution with density proportional to $\exp(-V(\theta))$, following almost verbatim the approach in Dalalyan and Tsybakov (2012b, Section 4), who utilized the Langevin Monte Carlo method[3]. Using these samples, we can evaluate the expectation of $\exp\left(\frac{2}{m}\langle x_t, \theta \rangle\right)$ under that distribution.

The second approach is to notice that

$$\frac{\int\limits_{\mathbb{R}^d} \exp\left(-L_m(\theta)\right)\mu(\theta)d\theta}{\int\limits_{\mathbb{R}^d} \exp\left(-L_{-m}(\theta)\right)\mu(\theta)d\theta} = \frac{\mathbb{E}_{\theta \sim \xi_1}\left[\prod_{j=1}^{d}\left(1 + |\theta^{(j)}| \cdot \sqrt{\sum_{t=1}^{T}(x_t^{(j)})^2/\tau}\right)^{-4}\right]}{\mathbb{E}_{\theta \sim \xi_2}\left[\prod_{j=1}^{d}\left(1 + |\theta^{(j)}| \cdot \sqrt{\sum_{t=1}^{T}(x_t^{(j)})^2/\tau}\right)^{-4}\right]},$$

where

$$\xi_1(\theta) \propto \exp\left(-\frac{1}{2m^2}(m - \langle x_t, \theta \rangle)^2 - \frac{1}{2m^2}\sum_{i=1}^{t-1}(y_i - \langle x_i, \theta \rangle)^2\right),$$

$$\xi_2(\theta) \propto \exp\left(-\frac{1}{2m^2}(-m - \langle x_t, \theta \rangle)^2 - \frac{1}{2m^2}\sum_{i=1}^{t-1}(y_i - \langle x_i, \theta \rangle)^2\right)$$

are both Gaussian densities. If sufficiently many Gaussian samples from ξ_1 and ξ_2 are obtained to approximate both expectations up to multiplicative constants close to one, the value of $\widehat{f}_t(x_t)$ can then be approximated within the desired precision level.

---

3. Note that, in our setting, it remains unclear whether the Langevin Monte Carlo method admits a worst-case polynomial-time implementation. This limitation also applies to the results of Dalalyan and Tsybakov (2012b).

Finally, we can analyze Laplace's approximation, which is computationally efficient. Specifically, we can compute two least squares solutions, $\widehat{\theta}_m = \arg\min_{\theta\in\mathbb{R}^d} L_m(\theta)$ and $\widehat{\theta}_{-m} = \arg\min_{\theta\in\mathbb{R}^d} L_{-m}(\theta)$. When the asymptotic error of the Laplace's method is small, i.e.,

$$\begin{cases} \int\limits_{\mathbb{R}^d} \exp\left(-L_m(\theta)\right)\mu(\theta)\,d\theta \approx \exp\left(-L_m(\widehat{\theta}_m)\right)\mu(\widehat{\theta}_m), \\ \int\limits_{\mathbb{R}^d} \exp\left(-L_{-m}(\theta)\right)\mu(\theta)\,d\theta \approx \exp\left(-L_{-m}(\widehat{\theta}_{-m})\right)\mu(\widehat{\theta}_{-m}). \end{cases}$$

Using these solutions, we can express the following approximate identity:

$$\frac{\int\limits_{\mathbb{R}^d} \exp\left(-L_m(\theta)\right)\mu(\theta)\,d\theta}{\int\limits_{\mathbb{R}^d} \exp\left(-L_{-m}(\theta)\right)\mu(\theta)\,d\theta} \approx \frac{\exp\left(-L_m(\widehat{\theta}_m)\right)\mu(\widehat{\theta}_m)}{\exp\left(-L_{-m}(\widehat{\theta}_{-m})\right)\mu(\widehat{\theta}_{-m})}.$$

We leave a rigorous analysis of the last two approaches as an interesting direction for future research.

5.2 Computation for logistic regression

Motivated by the high computational complexity of the algorithm proposed by Foster et al. (2018), the computational problem of online logistic regression has been further studied in (Jézéquel et al., 2020, 2021). However, since their algorithms are designed for the online setup, the lower bound in Theorem 9 shows that they cannot achieve the regret upper bound achieved by our Algorithm 1. On the other hand, the Normalized Maximum Likelihood (NML) algorithm is known to achieve exact minimax regret (Wu et al., 2022), but its computation requires exhausting all possible label sequences $y_1, \ldots, y_T \in \{\pm 1\}^T$, which scales as 2^T. Consequently, without additional algorithmic insights, the NML algorithm cannot be computed in polynomial time. Methods based on ε-nets typically work for the problems we consider but suffer from the lack of known implementable algorithms constructing the ε-nets (see Appendix D for detailed discussion).

In this section, we discuss a computable variant of Algorithm 1, where, however, the computational cost can scale in the worst case as $O(\text{poly}(T^d))$ The computation of Algorithm 1 presents two main challenges:

1. How do we compute a prediction from a slab-expert within the slab? This is equivalent to determining how to compute the probability assigned when the covariate has a small margin with respect to a slab-expert, i.e., how to compute the Gaussian integral in this case.

2. How do we aggregate the predictions from the slab-experts?

We address both in what follows.

5.2.1 COMPUTING SLAB-EXPERT PREDICTION

For any $\theta \in \mathbb{R}^d$, to produce our prediction for the slab-expert $B_{\theta,\alpha}$ as defined in Theorem 11 at time step t on a point x_t inside a slab $K_\theta \subset \{x_i\}_{i\in[T]}$, we need to estimate probability

assignments defined as $B_{\theta,\alpha}(x_t, 1) = \int \sigma(\langle x_t, \vartheta \rangle)\, \pi_t(d\vartheta)$ where the density is

$$\pi_t(\vartheta) \propto \exp\left(\sum_{x \in K_{t-1}} \log(\sigma(y_x \langle x, \vartheta \rangle)) - \alpha \vartheta^\top \left(\sum_{x \in K_\theta} xx^\top \right) \vartheta \right),$$

where $K_{t-1} = \{x_i\}_{i \in [t-1]} \cap K_\theta$. We can estimate $B_{\theta,\alpha}(x_t, 1)$ by sampling ϑ according to π_t and averaging over $\sigma(\langle x_t, \vartheta \rangle)$. Sampling from the distribution π_t has been studied in the literature (Dalalyan, 2017, Example 2). The LMC algorithm (Dalalyan, 2017) can sample from a distribution ε-close to the distribution π_t, in TV distance with a number of steps scaling with

$$O\left(\left(\frac{\alpha + 1/4}{\alpha}\right)^2 d\varepsilon^{-2} \left(2\log(1/\varepsilon) + (d/2)\log\left(\frac{\alpha + 1/4}{\alpha}\right) \right)^2 \right) = O(\mathrm{poly}(T, d))$$

when $\varepsilon = 1/T$ and recall $\alpha = 1/T^3$ in Algorithm 1. Thus, since the function $\sigma(\langle x_t, \vartheta \rangle)$ is bounded in $[0, 1]$, by sampling T^2 times, and using Hoeffding's inequality for concentration, we can estimate the assignment $B_{\theta,\alpha}(x_t, 1)$ with accuracy to the level of $O(1/T)$.

5.2.2 AGGREGATING THE SLAB-EXPERTS

Suppose μ is a prior on any set of experts that contains the optimal slab-expert $B_{\theta^\star,\alpha}$. Recall from Section 3.2, the mixability of the logistic loss implies that the regret of the exponential weights algorithm over this set will suffer at most a regret scaling with $O(\log(1/\mu(B_{\theta^\star,\alpha})) + d\log T)$. While it is challenging to identify all the slab-experts due to computational constraints, we can consider a broader class of generalized slab-experts. These generalized slab-experts extend the concept of slab-experts by allowing the slab to be defined as the intersection of two halfspaces.

Consider all the half-space separations $\mathcal{S} = \{(K_+, K_-)\}$ where $K_+ \cup K_- = \{x_t\}_{t \in [T]}$, and there exists a vector θ and an intercept b such that either for all $x \in K_+$, $\langle x, \theta \rangle - b \geq 0$ and for all $x \in K_-$, $\langle x, \theta \rangle - b < 0$.

Definition 22 (Generalized slab-expert) *Let $S_1 = (K_{1,+}, K_{1,-})$ and $S_2 = (K_{2,+}, K_{2,-})$ be any two elements in the set of the half-space separations $\mathcal{S}$. If $K_{1,-} \cap K_{2,+} = \emptyset$, then denote by A_{S_1,S_2} the generalized slab-expert that runs exponential weights on set $K = K_{1,+} \cap K_{2,-} = \{x_t\}_{t \in [T]} \setminus (K_{1,-} \cup K_{2,+})$, predicts -1 on $K_{1,-}$, and predicts 1 on $K_{2,+}$. Denote the set of all such generalized slab-experts by $\mathcal{A}^+$.*

It is clear that the slab experts are such experts as well. Thus, $\mathcal{A} \subseteq \mathcal{A}^+$. Since the number of half-spaces is upper bounded by $|\mathcal{S}| \leq (eT/d)^{2d}$, we have $|\mathcal{A}^+| \leq |\mathcal{S}|^2 \leq (eT/d)^{4d}$. The generalized slab-experts are constructed through all possible halfspaces. This reduces the task of aggregating slab-experts to the simpler problem of enumerating/sampling halfspaces. Indeed, in the following, we show that it is possible to enumerate all the halfspaces with the help of the SVM algorithm. Moreover, it is possible to use rejection sampling where each halfspace will be chosen with probability at least $1/(2T)^{d+1}$ and with acceptance at least $1/2$. This, together with the Metropolis-Hastings (MH) algorithm for evaluating exponential weighted averaging, constitutes an implementable algorithm.

Enumerating the generalized slab-experts. In order to enumerate the generalized slab-experts, we only need to enumerate all the possible halfspaces. Fortunately, the hard-margin SVM for halfspaces in $\mathbb{R}^d$ with n input labeled data points can output at most $d+1$ labeled data points that capture all the labeling information (specified in the proof of Theorem 23) with computational cost $O(\text{poly}(d, n))$ (Boser et al., 1992). Utilizing this fact, we can enumerate all the halfspaces for any given covariate set in polynomial time, as shown by the following lemma.

Lemma 23 *For any set K of $T > 0$ covariates, the halfspaces restricted to K defined as $\text{HalfSpace}_K = \{h_{\theta,b}|_K : h_{\theta,b} = \text{sign}(\langle x, \theta \rangle - b), \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$ can be enumerated using at most $O((2T)^d)$ calls to the SVM algorithm.*

Proof [Proof of Theorem 23] The SVM algorithm for $\mathbb{R}^d$ takes as input labeled data points $\mathcal{D} = \{(x_i, y_i)\}_{i \in [n]}$ with $n > 0$ and outputs $(\widehat{\mathcal{D}}, \widehat{h}) = \text{SVM}(\mathcal{D})$, where $\widehat{\mathcal{D}} \subseteq \mathcal{D}$ is a subset with at most $d+1$ labeled data points and $\widehat{h}$ is a linear separator such that $\widehat{h}(x_i) = y_i$ for all $i \in [n]$. Moreover, the composition of the SVM algorithm is stationary in the sense that if $(\overline{\mathcal{D}}, \overline{h}) = \text{SVM}(\widehat{\mathcal{D}})$, then $\overline{\mathcal{D}} = \widehat{\mathcal{D}}$ and $\overline{h} = \widehat{h}$ (For a proof of this fact, we refer to Long and Long (2020), originally established by Vapnik and Chervonenkis (1974)).

We first enumerate all the subsets of K with at most $d+1$ elements, i.e., $\mathcal{K}_{d+1} = \{K' \subseteq K : |K'| \le d+1\}$. For each of the subset K' in $\mathcal{K}_{d+1}$, we consider the set of all possible labelings $\mathcal{D}(K') = \{\{\pm 1\}^{K'} : K' \in \mathcal{K}_{d+1}\}$. For any of the labeling $\mathcal{D} = \{(x, y_x)\}_{x \in K'} \in \mathcal{D}(K')$, we can apply the SVM algorithm to obtain $(\widehat{\mathcal{D}}, \widehat{h}) = \text{SVM}(\mathcal{D})$. With this process, we can define the following set of linear separators

$$\mathcal{H}_{K,\text{SVM}} = \left\{ \widehat{h}|_K \;\middle|\; \begin{array}{c} (\widehat{\mathcal{D}}, \widehat{h}) = \text{SVM}(\mathcal{D}), \;\; \mathcal{D} \in \mathcal{D}(K'), \\ \mathcal{D}(K') = \{(x, y_x)_{x \in K'} : (y_x)_{x \in K'} \in \{\pm 1\}^{K'}\}, \\ \mathcal{K}_{d+1} = \{K' \subseteq K : |K'| \le d+1\} \end{array} \right\}.$$

We claim that $\text{HalfSpace}_K = \mathcal{H}_{K,\text{SVM}}$. For any hypothesis $h|_K \in \text{HalfSpace}_K$, let $(\widehat{\mathcal{D}}, \widehat{h}) = \text{SVM}(\{x, h(x)\}_{x \in K})$. By the property of SVM, we have $|\widehat{\mathcal{D}}| \le d+1$, $h|_K = \widehat{h}|_K$ and $(\widehat{\mathcal{D}}, \widehat{h}) = \text{SVM}(\widehat{\mathcal{D}})$.

Let $\widehat{\mathcal{D}} = \{(x_i, y_i)\}_{i \in [k]}$, then since $\widehat{\mathcal{D}} \in \mathcal{D}(\{x_i\}_{i \in [k]})$ and $\{x_i\}_{i \in [k]} \in \mathcal{K}_{d+1}$, thus $\widehat{h}|_K \in \mathcal{H}_{K,\text{SVM}}$ as desired. Moreover, to construct the hypothesis class $\mathcal{H}_{K,\text{SVM}}$, we need to enumerate all possible labelings of covariates subset of size no larger than $d+1$, which results in at most $O((2T)^{d+1})$ labelings and calls to the SVM algorithm. Thus, the claim follows. ∎

With the enumeration of halfspaces using the SVM algorithm, we state the following result.

Proposition 24 *Algorithm 1, when using generalized slab-experts $\mathcal{A}^+$ instead of the slab-experts $\mathcal{A}$, can be implemented with a computational cost $O(\text{poly}(T^d))$ and achieves a regret upper bound of $O(d \log T)$.*

Proof [Proof of Theorem 24] Applying Theorem 23, we know that we can enumerate the halfspaces on given covariate set with computational cost $O(\text{poly}(T^d))$. Once the halfspaces are enumerated, the generalized slab-experts can be constructed by the combinations of all

halfspaces. Then Algorithm 1 with generalized slab-experts $\mathcal{A}^+$ aggregates each round with sampling cost at most $O(\mathrm{poly}(T,d))$ as shown in Section 5.2.1 and computational cost at most $O(\mathrm{poly}(T^d))$ (since there are $\mathrm{poly}(T^d)$ generalized slab-experts). In all, Algorithm 1, when using generalized slab-experts $\mathcal{A}^+$ instead of the slab-experts $\mathcal{A}$, can be implemented with a computational cost that scales as at most $O(\mathrm{poly}(T^d))$. Moreover, recall that the number of generalized slab-experts is upper bounded by $(eT/d)^{4d}$. Thus, the uniform prior on the generalized slab-experts puts at least $\Omega(\mathrm{poly}(1/T^d))$ probability on the optimal slab-expert $B_{\theta^\star,\alpha}$. Therefore, the regret can be upper bounded by $O(d\log T)$. The claim follows. $\blacksquare$

Sample the generalized slab-expert. To further improve the implementability of Algorithm 1 with generalized slab-experts $\mathcal{A}^+$ instead of the slab-experts $\mathcal{A}$, we can replace the enumeration of the generalized slab-experts with sampling. Concretely, we first sample a random subset of at most $d+1$ points of the covariate set K. Next, we sample the labels randomly. Finally, we apply SVM to obtain a halfspace. Since any halfspace is sampled with probability at least $\Omega(\mathrm{poly}(1/T^d))$, the optimal slab-expert $B_{\theta^\star,\alpha}$ will also be sampled with probability at least $\Omega(\mathrm{poly}(1/T^d))$. Thus, the regret can be upper bounded by $O(d\log T)$. The advantage of such a sampling scheme is that, for a set of covariates in general position (no d points are linearly dependent), random labeling of any $d+1$ subset will typically lead to a halfspace. This is validated by Cover (1965, Theorem 1). Specifically, for any $d+1$ points in general position, the number of distinct halfspaces is at least $2\sum_{k=0}^{d-1}\binom{d}{k} = 2^{d+1}-2 \geq 2^d$ according to Cover (1965, Theorem 1). However, there are at most 2^{d+1} different labeling available. Thus, random labeling allows for a halfspace separation with probability at least $1/2$. Moreover, since the aggregation follows the exponentially weighted averaging

$$\tilde{\rho}_t(x_t,\cdot) = \sum_{B\in\mathcal{A}^+} p_t(B)B(x_t,\cdot), \quad \text{where} \quad p_t(B) \propto \exp\left(\sum_{i=1}^{t-1}\log B(x_i,y_i)\right).$$

The distribution p_t on the set of generalized slab-experts $\mathcal{A}^+$ is a Gibbs-type distribution that can be represented as the stationary distribution of a Markov chain generated by the Metropolis-Hastings (MH) algorithm (Robert and Casella, 2004, Section 7.3), in a manner similar to that in (Rigollet and Tsybakov, 2012, Section 7). Note that this approach provides an implementable sampling procedure rather than offering any computational advantages.

6. Online-to-Batch Conversion: Fixed Design vs Random Design

In the problem of statistical learning, the learner is given a sample $S_T = \{(X_1,Y_1),\ldots, (X_T,Y_T)\}$ consisting of i.i.d. pairs sampled according to some unknown distribution $\mathcal{D}$ on the space $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}$ is the covariate space, and $\mathcal{Y}$ is the observation space. The learner is further given a function class $\mathcal{F}:\mathcal{X}\to\mathcal{Y}$ and a loss function $\ell(\cdot,\cdot):\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$. The learner's goal is to output a (possibly randomized) predictor $\widehat{f}_T$ to minimize the (expected)

excess risk defined as

$$\mathbb{E}_{S_T}\left[\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(\widehat{f}_T(X),Y)] - \inf_{f\in\mathcal{F}}\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(f(X),Y)]\right].$$

We use transductive learning with the online-to-batch technique to obtain the following random design excess risk bound.

Proposition 25 (Transductive online-to-batch conversion) *For any function class $\mathcal{F}$, number of rounds T, and a loss function $\ell(\cdot,\cdot)$, convex with respect to the first argument, consider a deterministic sequence $x_1,\ldots,x_T$. Suppose there exists an algorithm that produces a sequence of predictors $\widehat{f}_1,\ldots,\widehat{f}_T$ in the sequential transductive setting, satisfying*

$$\sum_{t=1}^{T}\ell(\widehat{f}_t(x_t),y_t) - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t),y_t) \leq \mathrm{Reg}_T^{\mathrm{trd}}, \tag{39}$$

where $\mathrm{Reg}_T^{\mathrm{trd}}$ is an upper bound on the regret, independent of x_t and y_t.
Now, given a sample S_T of size T and an additional test point (X,Y), both drawn i.i.d. from an unknown distribution $\mathcal{D}$, construct the sequence $\widehat{f}_1,\ldots,\widehat{f}_{T+1}$, where $\widehat{f}_{t+1}$ is based on the labeled sample $(X_1,Y_1),\ldots,(X_t,Y_t)$ and the unlabeled set $\{X_1,\ldots,X_T,X\}$[4] and satisfies the deterministic regret bound (39) with an upper bound on the regret $\mathrm{Reg}_{T+1}^{\mathrm{trd}}$. The final predictor, defined as $\widetilde{f} = \frac{1}{T+1}\sum_{t=1}^{T+1}\widehat{f}_t$, satisfies the following excess risk bound:

$$\mathbb{E}_{S_T}\left[\mathbb{E}_{(X,Y)}[\ell(\widetilde{f}(X),Y)] - \inf_{f\in\mathcal{F}}\mathbb{E}_{(X,Y)}[\ell(f(X),Y)]\right] \leq \frac{\mathrm{Reg}_{T+1}^{\mathrm{trd}}}{T+1}.$$

Proof [Proof of Theorem 25] Denote $(X_{T+1},Y_{T+1}) = (X,Y)$. To highlight the dependence of any predictor f_t on both the unlabeled set $\{X_s\}_{s=1}^{T+1}$ and the labeled sequence of observations $X_{1:t-1},Y_{1:t-1}$, we use the following notation for the loss $\ell(\widehat{f}_t(X_t \mid X_{1:t-1},Y_{1:t-1},\{X_s\}_{s=1}^{T+1}),Y_t)$ at the observation X_t, where $X_{1:t-1}$ denotes the sequence $X_1,\ldots$,X_{t-1} and $Y_{1:t-1}$ stands for the sequence $Y_1,\ldots,Y_{t-1}$, respectively. We first break the loss into summands of conditional expectations with Jensen's inequality:

$$(T+1)\mathbb{E}_{S_T}\left[\mathbb{E}_{(X,Y)}[\ell(\widetilde{f}(X),Y)]\right]$$

$$\leq \mathbb{E}_{S_T}\mathbb{E}_{(X_{T+1},Y_{T+1})}\left[\sum_{t=1}^{T+1}[\ell(\widehat{f}_t(X_{T+1}|X_{1:t-1},Y_{1:t-1},\{X_s\}_{s=1}^{T+1}),Y_{T+1})]\right]$$

$$= \mathbb{E}_{X_1,\ldots,X_{T+1}}\left[\sum_{t=1}^{T+1}\mathbb{E}_{(Y_s|X_s)_{s=1}^{t-1},Y_{T+1}|X_{T+1}}\left[\ell(\widehat{f}_t(X_{T+1}|X_{1:t-1},Y_{1:t-1},\{X_s\}_{s=1}^{T+1}),Y_{T+1})\right]\right]$$

$$= \sum_{t=1}^{T+1}\mathbb{E}_{X_1,\ldots,X_{T+1}}\left[\mathbb{E}_{(Y_s|X_s)_{s=1}^{t-1},Y_{T+1}|X_{T+1}}\left[\ell(\widehat{f}_t(X_{T+1}|X_{1:t-1},Y_{1:t-1},\{X_s\}_{s=1}^{T+1}),Y_{T+1})\right]\right]$$

---

4. Accordingly, we assume that at the level of the entire unlabeled dataset, the order of the observations does not influence the output of the algorithm, a condition consistently met in our applications. Whether this assumption is indeed required for transductive online-to-batch conversions in general remains open.

Now we notice that for each summand, there is exchangeability between (X_t, Y_t) and (X_{T+1}, Y_{T+1}). Concretely, let $(X'_t, Y'_t) = (X_{T+1}, Y_{T+1})$ and $(X'_{T+1}, Y'_{T+1}) = (X_t, Y_t)$, together with the fact that the transductive algorithm is indifferent to the ordering of $X_1, \ldots, X_{T+1}$, we have

$$\mathbb{E}_{X_1,\ldots,X_{T+1}} \left[\mathbb{E}_{(Y_s|X_s)_{s=1}^{t-1}, Y_{T+1}|X_{T+1}} \left[\ell(\widehat{f}_t(X_{T+1}|X_{1:t-1}, Y_{1:t-1}, \{X_s\}_{s=1}^{T+1}), Y_{T+1}) \right] \right]$$

$$= \mathbb{E}_{X_1,\ldots,X'_t,\ldots,X_T,X'_{T+1}} \left[\mathbb{E}_{(Y_s|X_s)_{s=1}^{t-1}, Y'_t|X'_t} \left[\ell(\widehat{f}_t(X'_t|X_{1:t-1}, Y_{1:t-1}, \{X_1,\ldots,X'_t,\ldots,X_T,X'_{T+1}\}), Y'_t) \right] \right]$$

$$\overset{(*)}{=} \mathbb{E}_{X_1,\ldots,X_T,X_{T+1}} \left[\mathbb{E}_{(Y_s|X_s)_{s=1}^{t}} \left[\ell(\widehat{f}_t(X_t|X_{1:t-1}, Y_{1:t-1}, \{X_s\}_{s=1}^{T+1}), Y_t) \right] \right]$$

$$= \mathbb{E}_{S_T} \mathbb{E}_{(X_{T+1}, Y_{T+1})}[\ell(\widehat{f}_t(X_t), Y_t)].$$

where for $(*)$, we use the fact that the distribution is not changed with the exchange. Then,

$$(T+1)\mathbb{E}_{S_T} \left[\mathbb{E}_{(X_{T+1},Y_{T+1})}[\ell(\widetilde{f}(X_{T+1}), Y_{T+1})] - \inf_{f \in \mathcal{F}} \mathbb{E}_{(X_{T+1},Y_{T+1})}[\ell(f(X_{T+1}), Y_{T+1})] \right]$$

$$\leq \sum_{t=1}^{T+1} \mathbb{E}_{S_T} \mathbb{E}_{(X_{T+1},Y_{T+1})}[\ell(\widehat{f}_t(X_t), Y_t)] - \inf_{f \in \mathcal{F}} \mathbb{E}_{S_T} \mathbb{E}_{(X_{T+1},Y_{T+1})} \sum_{t=1}^{T+1} \ell(f(X_t), Y_t)$$

$$\leq \mathbb{E}_{S_T} \mathbb{E}_{(X_{T+1},Y_{T+1})} \left[\sum_{t=1}^{T+1} \ell(\widehat{f}_t(X_t), Y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T+1} \ell(f(X_t), Y_t) \right]$$

$$\leq \mathrm{Reg}_{T+1}^{\mathrm{trd}}.$$

The claim follows. ∎

6.1 Applications of online-to-batch conversion

In this section, we demonstrate two applications of the online-to-batch conversion result presented in Theorem 25.

Regression with squared loss. Consider the regression problem with squared loss. Let the covariates $X_1, \ldots, X_T$ and response variables $Y_1, \ldots, Y_T$ be i.i.d. copies of the random pair (X, Y) drawn from an unknown distribution. Denote the dataset as $S_T = \{(X_1, Y_1), \ldots, (X_T, Y_T)\}$. We make no assumptions on the distribution of $X \in \mathbb{R}^d$, but assume that $|Y| \leq m$ almost surely. By combining the regret bound from Theorem 1 (with $\lambda = 1/T$) and the online-to-batch conversion argument from Theorem 25, we immediately construct a computationally efficient batch predictor $\widetilde{f}$, based on S_T, that satisfies the following excess risk bound:

$$\mathbb{E}_{S_T} \left[\mathbb{E}(Y - \widetilde{f}(X))^2 - \inf_{\theta \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, \theta \rangle)^2 \right] \leq \frac{m^2(1 + d\log(T+2))}{T+1}.$$

This bound, up to a logarithmic factor, matches the batch-specific bounds for the estimator proposed by Forster and Warmuth (2002) and the clipped linear least squares estimator analyzed by Mourtada et al. (2022).

Excess risk bounds for logistic regression. Now consider the case of logistic regression, where the covariates $X_1, \ldots, X_T$ and the labels $Y_1, \ldots, Y_T$ are i.i.d. copies of (X, Y). Again, we make no assumptions about the distribution of $X \in \mathbb{R}^d$, and assume that $Y \in \{-1, 1\}$. Define the dataset $S_T = \{(X_1, Y_1), \ldots, (X_T, Y_T)\}$. The following result holds.

Corollary 26 *In the logistic regression setting described above, without assumptions on the distribution of (X, Y), there exists a probability assignment $\widetilde{p}$, based on S_T, that satisfies*

$$\mathbb{E}_{S_T}\left[-\log(\widetilde{p}(X, Y)) - \inf_{\theta \in \mathbb{R}^d} \mathbb{E}\left(-\log(\sigma(Y\langle X, \theta\rangle))\right)\right] \leq \frac{6d\log(e(T+1))}{T+1}.$$

In particular, if we assume that $\mathbb{P}(Y = 1|X)$ follows a Bernoulli distribution with parameter $\sigma(\langle X, \theta^\star\rangle)$ (denoted as $\mathrm{Ber}(\sigma(\langle X, \theta^\star\rangle))$), then

$$\mathbb{E}_{S_T}\mathbb{E}_X\mathcal{KL}\left(\mathrm{Ber}(\sigma(\langle X, \theta^\star\rangle)) \| \mathrm{Ber}(\widetilde{p}(X, 1))\right) \leq \frac{6d\log(e(T+1))}{T+1},$$

where $\widetilde{p}(X, 1)$ denotes the predicted conditional probability of label $+1$ given X.

This result differs from previous excess risk bounds for logistic regression found in the literature (see, e.g., (Bach, 2010; Ostrovskii and Bach, 2021; Foster et al., 2018; Mourtada and Gaïffas, 2022; Vijaykumar, 2021; Bilodeau et al., 2023; van der Hoeven et al., 2023)) in that it imposes no assumptions on the marginal distribution of the covariates, and the regret bound does not depend on the norm of the minimizer $\theta^\star$.

Proof [Proof of Theorem 26] The proof follows directly from combining Theorem 13 with Theorem 25. The second statement follows from the standard computation, showing that in the well-specified case, the logarithmic loss excess risk corresponds to the Kullback-Leibler divergence between the corresponding distributions. ■

6.2 Polynomial time algorithm for batch logistic regression

In this section, we focus on minimizing the logistic loss in the batch setting for logistic distributions on the scaled sphere $\sqrt{d}\mathcal{S}^{d-1}$ using polynomial-time algorithms. Specifically, we assume $(X, Y) \sim \mathcal{D}_{\theta^\star}$, where X is sampled uniformly from the scaled sphere $\sqrt{d}\mathcal{S}^{d-1}$. Conditioned on X, we draw $Y \in \{\pm 1\}$ with $\mathbb{P}(Y = 1 \mid X) = \sigma(\langle X, \theta^\star\rangle)$ (and $\mathbb{P}(Y = -1 \mid X) = 1 - \sigma(\langle X, \theta^\star\rangle) = \sigma(-\langle X, \theta^\star\rangle))$, for some unknown parameter $\theta^\star$.

We will present the proof for the distribution $\mathcal{D}_{\theta^\star}$ where the covariate X follows a uniform distribution on the scaled sphere $\sqrt{d}\mathcal{S}^{d-1}$. Nevertheless, our proof only requires mild assumptions (Theorem 29 and Theorem 30) on the marginal distribution of the covariates X. In effect, our proof extends to and well beyond the standard Gaussian distribution.

Our algorithm attains an excess risk upper bound of $O\left(\frac{(d+\log T)\log T}{T}\right)$ using the transductive online-to-batch method with two experts, both of which can be implemented in polynomial time. Besides being implementable in polynomial time, our excess risk upper bound, combined with the lower bound established by Hsu and Mazumdar (2024) for proper predictors, reveals a separation between proper and improper approaches in terms of the dependence of the minimax excess risk on the norm of $\theta^\star$.

As a starting point, to address potential issues with unbounded loss, we introduce the probability assignment induced by the clipped linear separator.

Definition 27 (Clipped linear separator) *Let $T > 1$ be an integer. For any $\theta \in \mathbb{R}^d$, the clipped linear separator $k_\theta : \mathbb{R}^d \times \{\pm 1\} \to [0,1]$ is defined as*

$$k_\theta(x, y) = \begin{cases} 1 - \frac{1}{2T}, & y\langle x, \theta \rangle \geq 0, \\ \frac{1}{2T}, & otherwise. \end{cases}$$

Namely, the expert assigns probability $1 - 1/(2T)$ to the label predicted by the linear separator and $1/(2T)$ to the alternative one.

We state our main theorem of this section as follows.

Algorithm 3 Batch algorithm for logistic regression with pretraining

---

Require: The set of the data $\{(X_t, Y_t)\}_{t \in [T]} \cup \{(X'_t, Y'_t)\}_{t \in [T]}$ and the point of prediction
$\quad$ X. Let $\alpha = 1/(T+1)^9$ and $K = \{X_t\}_{t \in [T]} \cup \{X\}$.
$\quad$ Solve the hard margin SVM for the pretraining dataset $\{(X'_t, Y'_t)\}_{t \in [T]}$.
$\quad$ **if** $\{(X'_t, Y'_t)\}_{t \in [T]}$ is separable **then**
$\quad\quad$ Obtain $\hat{\theta}$ that separates $\{(X'_t, Y'_t)\}_{t \in [T]}$.
$\quad$ **else**
$\quad\quad$ Let $\hat{\theta} = e_1$ the unit vector in the first dimension.
$\quad$ **end if**
$\quad$ Let $\mathcal{A} = \{\text{EWA}_{K,\alpha}, k_{\hat{\theta}}\}$ be the set of the two predictors, where $\text{EWA}_{K,\alpha}$ is the slab-
$\quad$ expert defined as in Theorem 11.
$\quad$ **for** $t = 1, 2, \ldots, T+1$ **do**
$\quad\quad$ Let the prediction be the aggregation on the experts

$$\tilde{\rho}_t(\cdot, \cdot) = \sum_{B \in \mathcal{A}} p_t(B) B(\cdot, \cdot) \quad \text{where} \quad p_t(B) \propto \exp\left(\sum_{i=1}^{t-1} \log B(X_i, Y_i)\right) \quad (40)$$

$\quad\quad$ and $B(X_i, Y_i)$ is probability of B predicting Y_i on X_i.
$\quad$ **end for**
$\quad$ Return the prediction $\tilde{p}(X, \cdot) = \frac{1}{T+1} \sum_{t=1}^{T+1} \tilde{\rho}_t(X, \cdot)$.

Theorem 28 *Let $T > 1$ be an integer. Suppose $S_T = \{(X_t, Y_t)\}_{t \in [T]}$, $S'_T = \{(X'_t, Y'_t)\}_{t \in [T]}$, and (X, Y) are i.i.d. drawn from $\mathcal{D}_{\theta^*}$. Then the $\tilde{p}$ returned by Algorithm 3 satisfies*

$$\mathbb{E}_{S_T, S'_T, (X,Y)}[-\log \tilde{p}(X, Y)] \leq \mathbb{E}_{(X,Y)}[-\log \sigma(Y\langle X, \theta^* \rangle)] + \frac{C(d + \log T)\log T}{T},$$

where $C > 0$ is an absolute constant.

The proof proceeds as follows: We use the transductive online-to-batch method as described in Section 6. There are two experts, one based on the transductive prior, which incorporates

41

all covariates, and the other obtained by solving the hard-margin SVM on a set of pretraining data. If the hard-margin SVM admits a solution $\widehat{\theta}$, we include the clipped linear separator with this solution $\widehat{\theta}$; otherwise, we use the unit vector along the first coordinate.

There are two cases to analyze. The first case assumes that the norm of θ^* is bounded by T^4. In this scenario, by setting $\alpha < 1/T^9$, the expert using the transductive prior achieves a rate of $O(d \log T / T)$ (Theorem 15).

In the second case, where the norm of θ^* exceeds T^4, we further divide the analysis into three subcases: 1. The first subcase is when the data is not separable by θ^*, which occurs with small probability due to the large norm of θ^* (Theorem 31). 2. The second subcase is when $\widehat{\theta}$ deviates from θ^* by more than $(d + \log T)/T$, which also happens with small probability (Theorem 32). 3. The final subcase occurs when $\widehat{\theta}$ is close to θ^*, in which case the clipped linear separator with $\widehat{\theta}$ achieves a rate of $O((d + \log T) \log T / T)$ (Theorem 33). This completes the proof.

We provide a proof for covariates distributed uniformly on the scaled sphere $\sqrt{d}\mathcal{S}^{d-1}$, relying on two key properties of the covariate distribution, as described in Theorem 29 and Theorem 30. These properties also hold when the covariates follow a standard Gaussian distribution, thus extending our proof to that case as well.

Assumption 29 (Bounded eigenvalue) *A random vector X in $\mathbb{R}^d$ has a bounded eigenvalue if the largest eigenvalue of the covariance matrix $\mathbb{E}[XX^\top]$ is upper bounded by 1, i.e., $\lambda_{\max}\left(\mathbb{E}[XX^\top]\right) \leq 1$. Here, for any positive semi-definite matrix $A \in \mathbb{R}^{d \times d}$, we denote by $\lambda_{\max}(A)$ the largest eigenvalue of the matrix A.*

The choice of the constant one in the upper bound $\lambda_{\max}\left(\mathbb{E}[XX^\top]\right) \leq 1$ can be relaxed to any constant and will appear as a log factor in the final risk bound. Both the uniform distribution on the scaled sphere sphere and the standard Gaussian distribution satisfy the bounded eigenvalue assumption (Theorem 29). For the former, we have for any θ on the sphere $\mathcal{S}^{d-1}$:

$$\theta^\top \mathbb{E}_{X \sim \mathrm{Unif}(\sqrt{d}\mathcal{S}^{d-1})}[XX^\top]\theta = d\mathbb{E}_{X \sim \mathrm{Unif}(\mathcal{S}^{d-1})}[\langle X, e_1 \rangle^2] = d\mathbb{E}_{X \sim \mathrm{Unif}(\mathcal{S}^{d-1})}\left[\frac{1}{d}\sum_{i=1}^{d} X_i^2\right] = 1.$$

For the standard Gaussian distribution, the covariance matrix is the identity, thus it also satisfies the assumption.

Assumption 30 (Small probability in thin slabs) *A random vector X in $\mathbb{R}^d$ has small probability in thin slabs if for any $\varepsilon \in (0, 1)$ and the $\theta^\star$ used to define the conditional distribution of Y, we have*

$$\mathbb{P}_X\left(|\langle X, \theta^\star/\|\theta^\star\|\rangle| \leq \varepsilon\right) \leq \frac{\sqrt{d}\varepsilon}{\sqrt{\pi}}.$$

The choice of the constant $\sqrt{d}/\sqrt{\pi}$ can be relaxed to any polynomial dependence on d and will appear as a log factor in the final risk bound. Both the uniform distribution on the scaled sphere $\sqrt{d}\mathcal{S}^{d-1}$ and the standard Gaussian distribution satisfy Theorem 30. For the former, if $d = 1$, $\langle X, \theta \rangle = \pm 1$, thus

$$\mathbb{P}_X\left(|\langle X, \theta \rangle| \leq \varepsilon\right) = 0.$$

Moreover, for $d \geq 2$,

$$\mathbb{P}_X\left(|\langle X, \theta \rangle| \leq \varepsilon\right) = \int_{-\varepsilon/\sqrt{d}}^{\varepsilon/\sqrt{d}} \frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}(1-x^2)^{\frac{d-3}{2}} \, dx \leq \frac{\sqrt{d}\varepsilon}{\sqrt{\pi}},$$

where the first equality is due to the formula for the projection of a uniform distribution on the interval $[-1, 1]$, and the inequality follows from the fact that the density is upper bounded by $\frac{d}{2\sqrt{\pi}}$. For covariates X distributed according to the standard Gaussian, Theorem 30 also holds up to a constant factor since the projection in any direction is standard normal with density upper bounded by $1/\sqrt{2\pi}$.

With Theorem 30, when $\|\theta^\star\|$ is large, then with high probability, the data generated will be separable as shown by the following Theorem 31.

Lemma 31 *Let $T > 1$ be an integer. For any $\theta^\star \in \mathbb{R}^d$, with $\|\theta^\star\| > T^4$, with probability at least $1 - d/T$, the set $\{(X_t', Y_t')\}_{t=1}^T$ drawn i.i.d. from $\mathcal{D}_{\theta^\star}$ satisfies $Y_t' = \mathrm{sign}(\langle X_t', \theta^\star \rangle)$ for all t.*

Proof [Proof of Theorem 31] If $d \geq T^2$, we have nothing to prove. Thus, we assume $d < T^2$ for the following.

By the independence of the data points, we have

$$\mathbb{P}\left(\forall t, Y_t'\langle X_t', \theta^\star \rangle > 0\right) = \left(\mathbb{P}\left(Y_1'\langle X_1', \theta^\star \rangle > 0\right)\right)^T.$$

Furthermore, by the conditional dependence of Y_1' on X_1', we have

$$\mathbb{P}\left(Y_1'\langle X_1', \theta^\star \rangle > 0\right) \geq \mathbb{P}\left(Y_1'\langle X_1', \theta^\star \rangle > 0, |\langle X_1', \theta^\star \rangle| > T\right)$$

$$\geq \mathbb{P}\left(|\langle X_1', \theta^\star \rangle| > T\right) \cdot \frac{1}{1 + e^{-T}}.$$

Then, by Theorem 30, we have

$$\mathbb{P}\left(|\langle X_1', \theta^\star \rangle| > T\right) = 1 - \mathbb{P}\left(|\langle X_1', \theta^\star \rangle| \leq T\right)$$

$$\geq 1 - \mathbb{P}\left(|\langle X_1', \theta^\star/\|\theta^\star\| \rangle| \leq 1/T^3\right)$$

$$\geq 1 - \frac{\sqrt{d}}{\sqrt{\pi}T^3}.$$

Altogether, we have shown that

$$\mathbb{P}\left(\forall t, Y_t'\langle X_t', \theta^\star \rangle > 0\right) \geq \left(\left(1 - \frac{\sqrt{d}}{\sqrt{\pi}T^3}\right) \cdot \frac{1}{1 + e^{-T}}\right)^T$$

$$\geq 1 - \frac{d}{T}.$$

This concludes our proof. ■

Whenever the data $\{(X_t', Y_t')\}_{t \in [T]}$ is separable, the hard margin SVM will, with high probability, output a classifier $\hat{\theta}$ where the disagreement region for the linear separators induced by $\hat{\theta}$ and $\theta^\star$ is only $O\left(\frac{d + \log T}{T}\right)$, according to the recent result of Bousquet et al. (2020). This is concretized by the following Theorem 32.

Lemma 32 *Let $T > 1$ be an integer. For any $\theta^\star \in \mathbb{R}^d$, with $\|\theta^\star\| > T^2$, suppose the $\{X_t', Y_t'\}_{t=1}^T$ are drawn i.i.d. from the distribution $\mathcal{D}_{\theta^\star}$. Then, if $\{\forall t, Y_t' = \text{sign}(\langle X_t', \theta^\star \rangle)\}$, then with probability at least $1 - d/T$, the SVM solution $\hat{\theta}$ on $\{(X_t', Y_t')\}_{t \in [T]}$ achieves*

$$\hat{\theta} \in \arg\min_\theta \sum_{t=1}^T \mathbb{1}(Y_t' \langle X_t', \theta \rangle < 0)$$

satisfies

$$\mathbb{P}_{X, Y \sim \mathcal{D}_{\theta^\star}}\left(\langle X, \theta^\star \rangle \langle X, \hat{\theta} \rangle < 0 \right) \le \frac{C_1(d + \log T)}{T},$$

for some constant $C_1 > 0$.

Proof [Proof of Theorem 32] If $d > T$, then the argument automatically holds. Thus, we assume hereafter that $d \le T$. This proof is a revised version of Hsu and Mazumdar (2024, Lemma 14). For any θ, the error of the linear separator defined by θ in $0/1$ loss is defined as

$$
\begin{aligned}
\text{err}_{\theta^\star}(\theta) &:= \mathbb{E}[\mathbb{1}(Y \langle X, \theta \rangle < 0)] \\
&= \mathbb{E}[\sigma(-\text{sign}(\langle X, \theta \rangle) \langle X, \theta^\star \rangle)] \\
&= \mathbb{E}[(1 - \sigma(-|\langle X, \theta \rangle|)) \mathbb{1}(\langle X, \theta^\star \rangle \langle X, \theta \rangle < 0)] + \mathbb{E}[\sigma(-|\langle X, \theta \rangle|) \mathbb{1}(\langle X, \theta^\star \rangle \langle X, \theta \rangle \ge 0)] \\
&= \mathbb{P}(\langle X, \theta^\star \rangle \langle X, \theta \rangle < 0) - 2\mathbb{E}[\sigma(-|\langle X, \theta \rangle|) \mathbb{1}(\langle X, \theta^\star \rangle \langle X, \theta \rangle < 0)] + \mathbb{E}[\sigma(-|\langle X, \theta \rangle|)]
\end{aligned}
$$

This implies

$$
\begin{aligned}
\mathbb{P}(\langle X, \theta^\star \rangle & \langle X, \theta \rangle < 0) \\
&\le \text{err}_{\theta^\star}(\theta) - \mathbb{E}[\sigma(-|\langle X, \theta \rangle|)] + 2\mathbb{E}[\sigma(-|\langle X, \theta \rangle|) \mathbb{1}(\langle X, \theta^\star \rangle \langle X, \theta \rangle < 0)] \qquad (41) \\
&\le \text{err}_{\theta^\star}(\theta) + \text{err}_{\theta^\star}(\theta^\star).
\end{aligned}
$$

Now we note that by Theorem 30,

$$
\begin{aligned}
\text{err}_{\theta^\star}(\theta^\star) &= \mathbb{E}[\sigma(-|\langle X, \theta^\star \rangle|)] \\
&\le \mathbb{P}(|\langle X, \theta^\star \rangle| \le T) + \sigma(-T) \\
&\le \mathbb{P}(|\langle X, \theta^\star / \|\theta^\star\| \rangle| \le 1/T) + \frac{1}{1 + e^T} \qquad (42) \\
&\le \frac{d}{T}.
\end{aligned}
$$

Now, for any θ, we let the error rate under conditioning be

$$\widetilde{\text{err}}_{\theta^\star}(\theta) := \mathbb{E}[\mathbb{1}(Y \langle X, \theta \rangle < 0) \mid Y = \text{sign}(\langle X, \theta^\star \rangle)].$$

By the risk bound for SVM (Bousquet et al., 2020, Theorem 12), we have, with probability at least $1 - d/T$, that

$$\widetilde{\text{err}}_{\theta^\star}(\hat{\theta}) \le \frac{C'(d + \log T)}{T} \qquad (43)$$

for some constant $C' > 0$. Finally, we have

$$
\begin{aligned}
\mathrm{err}_{\theta^\star}(\hat{\theta}) &= \mathbb{E}\Big[\mathbb{1}(Y\langle X,\hat{\theta}\rangle < 0)\Big] \\
&= \mathbb{P}(Y = \mathrm{sign}(\langle X,\theta^\star\rangle))\mathbb{E}\Big[\mathbb{1}(Y\langle X,\hat{\theta}\rangle < 0) \mid Y = \mathrm{sign}(\langle X,\theta^\star\rangle)\Big] \\
&\quad + \mathbb{P}(Y \neq \mathrm{sign}(\langle X,\theta^\star\rangle))\mathbb{E}\Big[\mathbb{1}(Y\langle X,\hat{\theta}\rangle < 0) \mid Y \neq \mathrm{sign}(\langle X,\theta^\star\rangle)\Big] \\
&\leq \widetilde{\mathrm{err}}_{\theta^\star}(\hat{\theta}) + \mathrm{err}_{\theta^\star}(\theta^\star).
\end{aligned}
\tag{44}
$$

Combining Equations (41) to (44) concludes the proof. ∎

Finally, whenever $\|\theta^\star\|$ is large enough, and the $\hat{\theta}$ is close enough to $\theta^\star$, the clipped linear predictor induced by $\hat{\theta}$ admits a good performance as shown by the following Theorem 33.

Lemma 33 *Let $T > 1$ be an integer. For any $\theta^\star \in \mathbb{R}^d$, with $\|\theta^\star\| > T^2$, and $\hat{\theta} \in \mathcal{S}^{d-1}$ such that $\mathbb{P}\Big(\langle X,\theta^\star\rangle\langle X,\hat{\theta}\rangle \leq 0\Big) \leq \varepsilon$, then we have the logistic loss of the clipped linear separator $k_{\hat{\theta}}$ under distribution $\mathcal{D}_{\theta^\star}$ is upper bounded by*

$$
\mathbb{E}_{(X,Y)\sim\mathcal{D}_{\theta^\star}}\Big[-\log k_{\hat{\theta}}(X,Y)\Big] \leq \frac{2d\log(2T)}{T} + \log(2T)\varepsilon.
$$

Proof [Proof of Theorem 33] We have

$$
\begin{aligned}
\mathbb{E}_{(X,Y)\sim\mathcal{D}_{\theta^\star}}&\Big[-\log k_{\hat{\theta}}(X,Y)\Big] \\
&\leq \mathbb{E}_{(X,Y)\sim\mathcal{D}_{\theta^\star}}\Big[-\log k_{\hat{\theta}}(X,Y)\mathbb{1}(|\langle X,\theta^\star\rangle| \leq T)\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim\mathcal{D}_{\theta^\star}}\Big[-\log k_{\hat{\theta}}(X,Y)\mathbb{1}\Big(\langle X,\theta^\star\rangle\langle X,\hat{\theta}\rangle \leq 0\Big)\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim\mathcal{D}_{\theta^\star}}\Big[-\log k_{\hat{\theta}}(X,Y)\mathbb{1}\Big(|\langle X,\theta^\star\rangle| > T, \langle X,\theta^\star\rangle\langle X,\hat{\theta}\rangle > 0\Big)\Big] \\
&\leq \log(2T)\Big(\mathbb{P}(|\langle X,\theta^\star\rangle| \leq T) + \mathbb{P}\Big(\langle X,\theta^\star\rangle\langle X,\hat{\theta}\rangle \leq 0\Big)\Big) \\
&\quad + \mathbb{E}_{(X,Y)\sim\mathcal{D}_{\theta^\star}}\Big[-\log k_{\hat{\theta}}(X,Y)\mathbb{1}\Big(|\langle X,\theta^\star\rangle| > T, \langle X,\theta^\star\rangle\langle X,\hat{\theta}\rangle > 0\Big)\Big],
\end{aligned}
$$

where the second inequality is by $-\log k_{\hat{\theta}}(X,Y) \leq \log(2T)$. We first estimate the probability in the slab by Theorem 30, since $\|\theta^\star\| \geq T^2$, we have $T/\|\theta^\star\| \leq 1/T$ and

$$
\begin{aligned}
\mathbb{P}(|\langle X,\theta^\star\rangle| \leq T) &= \mathbb{P}(|\langle X,\theta^\star/\|\theta^\star\|\rangle| \leq T/\|\theta^\star\|) \\
&\leq \mathbb{P}(|\langle X,\theta^\star/\|\theta^\star\|\rangle| \leq 1/T) \\
&\leq \frac{\sqrt{d}}{\sqrt{\pi}T}.
\end{aligned}
$$

Finally, we estimate the risk in the correctly labeled region as

$$
\begin{aligned}
\mathbb{E}_{(X,Y)\sim\mathcal{D}_{\theta^\star}} & \left[-\log k_{\hat{\theta}}(X,Y)\mathbb{1}\Big(|\langle X,\theta^\star\rangle| > T, \langle X,\theta^\star\rangle\langle X,\hat{\theta}\rangle > 0\Big)\right] \\
& \leq \sup_{x:|\langle x,\theta^\star\rangle|>T} \mathbb{E}[-\log k_{\hat{\theta}}(x,Y) \mid X = x] \\
& \leq -\frac{1}{1+e^{-T}}\log\left(1 - \frac{1}{2T}\right) + \frac{e^{-T}}{1+e^{-T}}\log(2T) \\
& \leq \frac{1}{T}.
\end{aligned}
$$

Altogether, we conclude our proof. ∎

Proof [Proof of Theorem 28] Denote $(X_{T+1}, Y_{T+1}) = (X, Y)$. We consider two cases. For the first case, suppose $\|\theta^\star\| \leq T^4$. We have by the exponential weights

$$
\begin{aligned}
-\sum_{t=1}^{T+1}\log\tilde{\rho}_t(X_t, Y_t) &= -\sum_{t=1}^{T+1}\log\left(\frac{\sum_{B\in\mathcal{A}}\exp\left(\sum_{i=1}^{t}\log B(X_i,Y_i)\right)}{\sum_{B\in\mathcal{A}}\exp\left(\sum_{i=1}^{t-1}\log B(X_i,Y_i)\right)}\right) \\
&= -\log\left(\frac{1}{|\mathcal{A}|}\sum_{B\in\mathcal{A}}\exp\left(\sum_{i=1}^{T+1}\log B(X_i,Y_i)\right)\right) \\
&\leq \min\left\{-\sum_{i=1}^{T+1}\log(\text{EWA}_{K,\alpha}(X_i,Y_i)), -\sum_{i=1}^{T+1}\log\big(k_{\hat{\theta}}(X_i,Y_i)\big)\right\} + \log 2.
\end{aligned}
$$

Furthermore, we have by Theorem 15

$$
\begin{aligned}
-\sum_{i=1}^{T+1}&\log(\text{EWA}_{K,\alpha}(X_i,Y_i)) \\
&\leq -\sum_{t=1}^{T+1}\log(\sigma(Y_t\langle X_t,\theta^\star\rangle)) + \alpha(\theta^\star)^\top\left(\sum_{t=1}^{T+1}X_tX_t^\top\right)\theta^\star + \frac{d}{2}\log\left(1 + \frac{1}{8\alpha}\right) \\
&\leq -\sum_{t=1}^{T+1}\log(\sigma(Y_t\langle X_t,\theta^\star\rangle)) + \alpha\sum_{t=1}^{T+1}\langle X_t,\theta^\star\rangle^2 + 3d\log(2T).
\end{aligned}
$$

Take expectation, we have, by the choice of $\alpha = 1/(T+1)^9$ and the case assumption $\|\theta^\star\| \leq T^4$, that

$$
\begin{aligned}
\mathbb{E}_{S_T,S_T',(X,Y)}[-\log\widetilde{p}(X,Y)] &\leq \frac{1}{T+1}\mathbb{E}\left[-\sum_{t=1}^{T+1}\log\tilde{\rho}_t(X_t,Y_t)\right] \\
&\leq \mathbb{E}_{(X,Y)}[-\log\sigma(Y\langle X,\theta^\star\rangle)] + \alpha\lambda_{\max}\Big(\mathbb{E}[XX^\top]\Big)\|\theta^\star\|^2 + \frac{4d\log(2T)}{T+1} \\
&\leq \mathbb{E}_{(X,Y)}[-\log\sigma(Y\langle X,\theta^\star\rangle)] + \frac{5d\log(2T)}{T},
\end{aligned}
$$

where the last inequality uses Theorem 29. For the second case, suppose $\|\theta^\star\| > T^4$. For this, we have

$$\mathbb{E}_{S_T, S_T', (X,Y)}[-\log \widetilde{p}(X, Y)] = \frac{1}{T+1}\mathbb{E}\left[-\sum_{t=1}^{T+1}\log \tilde{\rho}_t(X_t, Y_t)\right]$$
$$\leq \mathbb{E}\left[-\log\bigl(k_{\hat{\theta}}(X, Y)\bigr)\right] + \frac{\log 2}{T+1}.$$

Then we divide into cases as

$$\mathbb{E}\left[-\log\bigl(k_{\hat{\theta}}(X, Y)\bigr)\right]$$
$$\leq \mathbb{E}\left[-\log\bigl(k_{\hat{\theta}}(X, Y)\bigr)\left(\mathbb{1}\bigl(\exists t, Y_t' \neq \operatorname{sign}(\langle X_t', \theta^\star\rangle)\bigr)\right.\right.$$
$$+ \mathbb{1}\left(\forall t, Y_t' = \operatorname{sign}(\langle X_t', \theta^\star\rangle), \mathbb{P}_{X,Y\sim\mathcal{D}_{\theta^\star}}\Bigl(\langle X, \theta^\star\rangle\langle X, \hat{\theta}\rangle < 0\Bigr) > \frac{C_1(d + \log T)}{T}\right)$$
$$\left.\left.+ \mathbb{1}\left(\mathbb{P}_{X,Y\sim\mathcal{D}_{\theta^\star}}\Bigl(\langle X, \theta^\star\rangle\langle X, \hat{\theta}\rangle < 0\Bigr) \leq \frac{C_1(d + \log T)}{T}\right)\right)\right]$$
$$\leq \log(2T)\mathbb{P}\bigl(\exists t, Y_t' \neq \operatorname{sign}(\langle X_t', \theta^\star\rangle)\bigr)$$
$$+ \log(2T)\mathbb{P}\left(\forall t, Y_t' = \operatorname{sign}(\langle X_t', \theta^\star\rangle), \mathbb{P}_{X,Y\sim\mathcal{D}_{\theta^\star}}\Bigl(\langle X, \theta^\star\rangle\langle X, \hat{\theta}\rangle < 0\Bigr) > \frac{C_1(d + \log T)}{T}\right)$$
$$+ \mathbb{E}\left[-\log\bigl(k_{\hat{\theta}}(X, Y)\bigr)\mathbb{1}\left(\mathbb{P}_{X,Y\sim\mathcal{D}_{\theta^\star}}\Bigl(\langle X, \theta^\star\rangle\langle X, \hat{\theta}\rangle < 0\Bigr) \leq \frac{C_1(d + \log T)}{T}\right)\right].$$

Then, by Theorems 31 to 33, we have

$$\mathbb{E}\left[-\log\bigl(k_{\hat{\theta}}(X, Y)\bigr)\right] \leq \frac{C_2(d + \log T)\log T}{T}.$$

The claim follows. ∎

7. Directions of Future Research

Finally, we outline several promising directions for future research.

- Our online-to-batch conversion argument in Section 6 only provides in-expectation results in important applications, such as logistic regression. While recent martingale-based arguments allow for high-probability excess risk bounds (see, e.g., (van der Hoeven, Zhivotovskiy, and Cesa-Bianchi, 2023)), these existing arguments do not extend straightforwardly to the transductive setup, where the natural martingale structure is not present. An even more challenging question would be to provide computationally efficient methods for logistic regression with no assumptions on the design, for which the excess risk bound of Theorem 26 holds with high probability.

- For computational considerations, it is also worthwhile to explore algorithms such as Follow-the-Regularized-Leader (FTRL), where regularization is incorporated. To

align with the context of our current work, we may examine transductive regularization where the regularizer depends on the set of covariates, as studied by Gaillard et al. (2019) in the case of squared loss. For logistic loss and hinge loss, however, it remains unclear whether FTRL with transductive regularization can achieve similar assumption-free excess risk bounds.

- Our results currently apply only to linear classes; a natural next step is to extend them to more general classes. At present, however, we lack the technical tools to construct natural, nontrivial priors for such classes, beyond, for example, the uniform distribution on ε-covers.

- While some of our techniques, such as exponential weights and mixability, extend to broader loss families, our analysis also exploits specific structural properties of the considered models, including the form of the sigmoid link and the resulting Hessian in the parameter space. For instance, our logistic regression results rely on the slab construction and need not hold for the same logarithmic loss in more general GLM settings. As in the case of general classes, it would be interesting to study other losses and different link functions in the context of our setup.

Acknowledgments

References

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016.

Jean-Yves Audibert. *PAC-Bayesian statistical learning theory*. PhD thesis, Université Paris VI, 2004.

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

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

Andrew R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting, June 6-10, 1998*, pages 27–52. Oxford University Press, 1999.

Peter L Bartlett, Wouter M Koolen, Alan Malek, Eiji Takimoto, and Manfred K Warmuth. Minimax fixed-design linear regression. In *Conference on Learning Theory*, pages 226–239. PMLR, 2015.

Shai Ben-David, Eyal Kushilevitz, and Yishay Mansour. Online learning versus offline learning. *Machine Learning*, 29:45–63, 1997.

Blair Bilodeau, Dylan J Foster, and Daniel M Roy. Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 51(2):762–790, 2023.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal svm bound. In *Conference on Learning Theory*, pages 582–609. PMLR, 2020.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi and Ohad Shamir. Efficient transductive online learning via randomized rounding. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 177–194, 2013.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. A second-order perceptron algorithm. *SIAM Journal on Computing*, 34(3):640–668, 2005.

Bertrand S Clarke and Andrew R Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

Bertrand S Clarke and Andrew R Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965. doi: 10.1109/PGEC.1965.264137.

Arnak Dalalyan and Alexandre B Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1):39–61, 2008.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

Arnak S Dalalyan and Alexandre B Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012a.

Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012b.

Michael Drmota, Philippe Jacquet, Changlong Wu, and Wojciech Szpankowski. Minimax regret with unbounded weights. In *ISIT 2024-International Symposium on Information Theory*, 2024.

Michael Drmota, Philippe Jacquet, Changlong Wu, and Wojciech Szpankowski. Precise regularized minimax regret with unbounded weights. *IEEE Transactions on Information Theory*, 2025.

Raisa Dzhamtyrova and Yuri Kalnishkan. A lower bound for a prediction algorithm under the kullback-leibler game. In *Conformal and Probabilistic Prediction and Applications*, pages 39–51. PMLR, 2021.

Jürgen Forster. On relative loss bounds in generalized linear regression. In *Fundamentals of Computation Theory: 12th International Symposium*, pages 269–280, 1999.

Jürgen Forster and Manfred K Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.

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, 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.

Pierre Gaillard, Sébastien Gerchinovitz, Malo Huard, and Gilles Stoltz. Uniform regret bounds over $\mathbb{R}^d$ for the sequential linear regression problem with the square loss. In *Algorithmic Learning Theory*, pages 404–432. PMLR, 2019.

Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14:729–769, 2013.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. Bayes and Tukey meet at the center point. In *17th Annual Conference on Learning Theory, COLT 2004*, pages 549–563, 2004.

Peter Grünwald, Thomas Steinke, and Lydia Zakynthinou. PAC-Bayes, MAC-Bayes and conditional mutual information: Fast rate bounds that handle general VC classes. In *Conference on Learning Theory*, pages 2217–2247. PMLR, 2021.

László Győrfi, Michael Kohler, Adam Krzyzak, Harro Walk, et al. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Elad Hazan, Tomer Koren, and Kfir Y Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209. PMLR, 2014.

Daniel Hsu and Arya Mazumdar. On the sample complexity of parameter estimation in logistic regression with normal design. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2418–2437, 2024.

Philippe Jacquet, Gil I Shamir, and Wojciech Szpankowski. Precise minimax regret for logistic regression. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 444–449. IEEE, 2022.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In *Conference on Learning Theory*, pages 2085–2108. PMLR, 2020.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Mixability made efficient: Fast online multiclass logistic regression. *Advances in Neural Information Processing Systems*, 34: 23692–23702, 2021.

Sham Kakade and Adam T Kalai. From batch to transductive online learning. *Advances in Neural Information Processing Systems*, 18, 2005.

Sham M Kakade and Andrew Ng. Online bounds for Bayesian algorithms. *Advances in Neural Information Processing Systems*, 17, 2004.

Jyrki Kivinen and Manfred K Warmuth. Averaging expert predictions. In *European Conference on Computational Learning Theory*, pages 153–167. Springer, 1999.

Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

Felix Kuchelmeister and Sara van de Geer. Finite sample rates for logistic regression with small noise or few samples. *Sankhya A*, pages 1–70, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2:285–318, 1988.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Philip M Long and Raphael J Long. On the complexity of proper distribution-free learning of linear classifiers. In *Algorithmic Learning Theory*, pages 583–591. PMLR, 2020.

The Tien Mai. High-dimensional sparse classification using exponential weighting with empirical hinge loss. *Statistica Neerlandica*, 2024.

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 Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *Journal of Machine Learning Research*, 23(31):1–49, 2022.

Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Distribution-free robust linear regression. *Mathematical Statistics and Learning*, 4(3):253–292, 2022.

Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Local risk bounds for statistical aggregation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5697–5698. PMLR, 2023.

Albert B. Novikoff. On convergence proofs on perceptrons. *Proceedings of the Symposium on the Mathematical Theory of Automata*, 12:615–622, 1962.

Dmitrii M Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1):326–391, 2021.

Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.

Nikita Puchkin and Nikita Zhivotovskiy. Exploring local norms in exp-concave statistical learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1993–2013, 2023.

Kolyan Ray and Botond Szabó. Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281, 2022.

Philippe Rigollet and Alexandre B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011. doi: 10.1214/10-AOS854.

Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

Jorma J Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, New York, 2nd edition, 2004.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Gil I Shamir. Logistic regression regret: What's the catch? In *Conference on Learning Theory*, pages 3296–3319. PMLR, 2020.

Yurii Shtarkov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.

Dirk van der Hoeven, Nikita Zhivotovskiy, and Nicolò Cesa-Bianchi. High-probability risk bounds via sequential predictors. *arXiv preprint arXiv:2308.07588*, 2023.

Tim Van Erven, Peter D Grünwald, Nishant A Mehta, Mark D Reid, and Robert C Williamson. Fast rates in statistical and online learning. *The Journal of Machine Learning Research*, 16(1):1793–1861, 2015.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 2013.

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

Tomas Vaškevičius and Nikita Zhivotovskiy. Suboptimality of constrained least squares and improvements via non-linear predictors. *Bernoulli*, 29(1):473–495, 2023.

Suhas Vijaykumar. Localization, convexity, and star aggregation. *Advances in Neural Information Processing Systems*, 34:4570–4581, 2021.

Vladimir Vovk. Metric entropy in competitive on-line prediction. *arXiv preprint cs/0609045*, 2006.

Volodimir Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–386. ACM, 1990.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Sequential vs. fixed design regrets in online learning. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 438–443. IEEE, 2022.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Fedor Zhdanov and Yuri Kalnishkan. An identity for kernel ridge regression. *Theoretical Computer Science*, 473:157–178, 2013.

Fedor Zhdanov and Vladimir Vovk. Competing with gaussian linear experts. *arXiv preprint arXiv:0910.4683*, 2009.

Appendix A. Additional Results for Sparse Regression

It is worth making the connection with sparsity priors used in (Rigollet and Tsybakov, 2011, 2012). These authors discuss that certain *discrete priors* in the batch fixed design linear regression problem with Gaussian noise allow bypassing any dependence on compatibility constants such as the constant κ_s used above. Rigollet and Tsybakov (2011, 2012) also make comparisons with classical Lasso and BIC-type oracle inequalities. While Lasso has computational advantages, its bounds involve compatibility constants such as the restricted eigenvalue assumption. In contrast, discrete sparsity priors allow bypassing any compatibility constants akin to what is observed for $\| \cdot \|_0$ penalization.

Before discussing these discrete priors, we note that despite the minor penalty of the logarithm of the inverse smallest scaled singular value constant κ_s, there is a significant advantage: the predictor in Theorem 3 has an explicit Gaussian-type integral form. On the other hand, the predictor derived from sparsity priors can be quite challenging to exploit, as it might require too many samples from this prior (or equivalently, sampling from the posterior distribution ρ_t is hard). However, from a practical implementation standpoint, it can still be effective, as claimed by simulations in (Rigollet and Tsybakov, 2011, 2012).

We use the sparsity priors as in (Rigollet and Tsybakov, 2011). Consider the set $\mathcal{P} = \{1, \ldots, d\}$. For each $p \subseteq \mathcal{P}$ assign the probability mass

$$\mu(p) = \left(\binom{d}{|p|} \exp(|p|) H_d \right)^{-1}, \tag{45}$$

where $H_d = \sum_{i=1}^{d} \exp(-i)$. Although this prior is not data-dependent, it can be naturally combined with the data-dependent prior (6) by introducing a product prior that first samples the sparsity pattern vector according to (45), then uses (6) restricted to the coordinates corresponding to this sparsity pattern. Finally, the exponential weights algorithm is run with respect to this data-dependent prior. To slightly simplify the analysis, we will use a

version of Theorem 1, where the additional clipping step is performed, and then we use aggregation using the exp-concavity of the squared loss in the bounded case.

For any $p \subseteq \mathcal{P}$ let $\Theta_p = \{\theta \in \mathbb{R}^d : \theta^{(j)} = 0 \text{ if } j \notin p\}$. That is, Θ_p contains all vectors in $\mathbb{R}^d$, which are equal to zero for all coordinates outside of the set p. For each $p \in \mathcal{P}$, define the sequence of weights

$$\widetilde{\theta}_{t,p} = \arg\min_{\theta \in \Theta_p} \left\{ \sum_{i=1}^{t-1} (y_i - \langle x_i, \theta \rangle)^2 + \lambda \sum_{s=1}^{T} (\langle x_s, \theta \rangle)^2 \right\},$$

where in the case of a non-unique minimizer, we choose the one with the smallest Euclidean norm, which corresponds to the pseudo-inverse matrix in the least squares solution formula. Define the sequence ρ_t of distributions over subsets of $\mathcal{P}$, by $\rho_1 = \mu$ and for any $t \geq 2$, set

$$\rho_t(p) = \frac{\exp\left(-\frac{1}{8m^2} \sum\limits_{i=1}^{t-1} \left(y_i - \text{clip}_m\left(\langle \widetilde{\theta}_{t,p}, x_i \rangle\right)\right)^2\right) \mu(p)}{\sum\limits_{p' \in \mathcal{P}} \exp\left(-\frac{1}{8m^2} \sum\limits_{i=1}^{t-1} \left(y_i - \text{clip}_m\left(\langle \widetilde{\theta}_{t,p'}, x_i \rangle\right)\right)^2\right) \mu(p')}.$$

Finally, we introduce the sequence of predictors

$$\widetilde{f}_t(x) = \sum_{p \in \mathcal{P}} \rho_t(p) \, \text{clip}_m\left(\langle x, \widetilde{\theta}_{t,p} \rangle\right). \tag{46}$$

The work of Rigollet and Tsybakov (2012) suggests a Metropolis-Hastings type algorithm for such sampling, though in the worst case its convergence rate can be quite slow.

The following result provides the desired regret bound for this predictor. Given the established regret bounds in the non-sparse case, the proof follows a standard approach and is included for reference.

Proposition 34 *Assume that the sequence $\{(x_t, y_t)\}_{t=1}^{T}$, where $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$ for all $1 \leq t \leq T$, satisfies $\max_t |y_t| \leq m$. Then, for any s-sparse $\theta^\star \in \mathbb{R}^d$, the following regret bound holds for the predictor (46) with $\lambda = \frac{s}{T}$,*

$$\sum_{t=1}^{T} (y_t - \widetilde{f}_t(x_t))^2 \leq \sum_{t=1}^{T} (y_t - \langle x_t, \theta^\star \rangle)^2 + sm^2 \left(\log\left(1 + \frac{T}{s}\right) + 16 \log\left(\frac{ed}{s}\right) + 5 \right).$$

Proof Fix any $p \subseteq \mathcal{P}$ such that $|p| = s$. When restricted to the coordinates corresponding to the set p, the bound of Theorem 41 in Appendix C gives the regret bound

$$\sum_{t=1}^{T} \left(y_t - \text{clip}_m\left(\langle x_t, \widetilde{\theta}_{t,p} \rangle\right)\right)^2 \leq \inf_{\theta \in \Theta_p} \left\{ \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 \right\} + \lambda T m^2 + 4sm^2 \log\left(1 + \frac{1}{\lambda}\right).$$

It remains to notice that for any $p \in \mathcal{P}$ such that $|p| = s$, the prediction $\text{clip}_m\left(\langle x_t, \widetilde{\theta}_{t,p} \rangle\right)$ is absolutely bounded by m. Finally, combining the exp-concavity of the squared loss in the bounded case (Vovk, 2001, Remark 3) (thus, we choose $\eta = \frac{1}{8m^2}$ in the exponential weights algorithm) and the fact that the predictor (46) is a posterior average with respect to the

exponential weights algorithm with prior μ over $\mathcal{P}$, and by Theorem 5 with $\log\left(\frac{1}{\mu(p)}\right) = 2s\log\left(\frac{ed}{s}\right) + \frac{1}{2}$ by (Rigollet and Tsybakov, 2012, Inequality (5.4)), we obtain that for any s-sparse $\theta^\star \in \mathbb{R}^d$,

$$\sum_{t=1}^{T}(y_t - \widetilde{f}_t(x_t))^2 \leq \sum_{t=1}^{T}(y_t - \langle x_t, \theta^\star\rangle)^2 + sm^2 + 4sm^2\log\left(1 + \frac{T}{s}\right) + 8m^2\left(2s\log\left(\frac{ed}{s}\right) + \frac{1}{2}\right).$$

The claim follows. ∎

Appendix B. Mixability of the Squared Loss for Unbounded Sets of Predictors

For completeness, we present the following mixability result, which is implicitly used in the analysis of sequential linear regression by Vovk (2001). Our proof closely follows the arguments in (Vovk, 2001), with the key clarification that the boundedness of the class of predictors is not required, a point that may not be clear from the original paper.

Lemma 35 (Restatement of Theorem 6) *Consider a class $\mathcal{F}$ of functions $f_\theta : \mathcal{X} \to \mathbb{R}$ (possibly unbounded) parameterized by $\Theta \subseteq \mathbb{R}^d$. Assume that y is such that $|y| \leq m$ and choose $\eta = \frac{1}{2m^2}$. Given any distribution ρ over Θ, define the predictor*

$$\widehat{f}(x) = \frac{m}{2}\log\left(\frac{\mathbb{E}_{\theta\sim\rho}\exp(-\eta(m - f_\theta(x))^2)}{\mathbb{E}_{\theta\sim\rho}\exp(-\eta(-m - f_\theta(x))^2)}\right).$$

Then, the following holds:

$$(y - \widehat{f}(x))^2 \leq -\frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\rho}\exp(-\eta(y - f_\theta(x))^2)\right).$$

Proof Without loss of generality, assume $m = 1$. The classical mixability result from (Vovk, 2001, Lemmas 2 and 3) shows that for the set $\mathcal{F}$ of 1-absolutely bounded prediction functions, there exists a predictor $\widetilde{f}$ such that for all $y \in [-1, 1]$:

$$(y - \widetilde{f}(x))^2 \leq -2\log\left(\mathbb{E}_{\theta\sim\rho}\exp\left(-(y - \text{clip}_1(f_\theta(x)))^2/2\right)\right).$$

By the definition of the clipping function, this implies that for the same predictor and all $y \in [-1, 1]$,

$$(y - \widetilde{f}(x))^2 \leq -2\log\left(\mathbb{E}_{\theta\sim\rho}\exp\left(-(y - f_\theta(x))^2/2\right)\right). \tag{47}$$

Now, define a predictor $\widehat{f}$ as

$$\widehat{f}(x) = \arg\inf_{\widehat{y}}\sup_{y\in[-1,1]}\left((y - \widehat{y})^2 + 2\log\left(\mathbb{E}_{\theta\sim\rho}\exp\left(-(y - f_\theta(x))^2/2\right)\right)\right),$$

which satisfies the same inequality (47) as $\widetilde{f}$ by construction. The proof of Lemma 3 in (Vovk, 2001) shows that the function h, defined as

$$h(y) = (y - \widehat{f}(x))^2 + 2\log\left(\mathbb{E}_{\theta\sim\rho}\exp\left(-(y - f_\theta(x))^2/2\right)\right),$$

satisfies $\frac{\partial^2 h(y)}{\partial y^2} \geq 0$, making h convex. Importantly, the derivation of $\frac{\partial^2 h(y)}{\partial y^2}$ does not impose any boundedness condition on $f_\theta(x)$. Since h is convex, its maximum is attained at either -1 or 1. Therefore, by symmetry, $\widehat{f}$ satisfies

$$(1 - \widehat{f}(x))^2 + 2\log\left(\mathbb{E}_{\theta \sim \rho} \exp\left(-(1 - f_\theta(x))^2/2\right)\right)$$
$$= (-1 - \widehat{f}(x))^2 + 2\log\left(\mathbb{E}_{\theta \sim \rho} \exp\left(-(-1 - f_\theta(x))^2/2\right)\right).$$

Solving this equation for $\widehat{f}(x)$ completes the proof. ∎

Appendix C. Revisiting Existing Bounds for Sequential Linear Regression

This section contains several results related to linear regression that might be of independent interest. The proof strategy goes beyond the transductive setting and follows the derivations of Vovk (2001) with the following twist: instead of exploiting the *mixability* of the squared loss, we consider a similar problem of conditional density estimation for the class of conditional densities

$$\left\{ p(y|x, \theta) = \exp\left(-\pi(y - x^\top \theta)^2\right) : \ \theta \in \mathbb{R}^d \right\}. \tag{48}$$

Our loss function is the logarithmic function, and we simply compute the total error for the *mix-loss* as suggested by Theorem 4. This straightforward computation is sufficient to recover all existing sequential online bounds, including the one in (Vovk, 2001), the bound for clipped ridge regression in (Kivinen and Warmuth, 1999), and the bound of Gaillard et al. (2019), which has already been discussed in Section 2. In the following, we discuss how essentially the same strategy, specific to sequential ridge regression, was previously employed by Zhdanov and Vovk (2009). Here, we mainly provide a detailed exposition of this approach and extend it to our transductive setup.

Let $A \in \mathbb{R}^{d \times d}$ be any positive definite matrix and given the sequence $x_1, \ldots, x_T \in \mathbb{R}^d$ define the *sequential leverage scores* of x_i as follows:

$$h_{x_i} = x_i^\top \left(\sum_{t=1}^{i} x_t x_t^\top + A\right)^{-1} x_i.$$

We remark that the results hold if the matrix A is positive semi-definite and thus not necessarily invertible. In this case, we should replace the inverse with the pseudo-inverse. The central result of this section is the following identity.

Proposition 36 (Generalization of (Azoury and Warmuth, 2001; Zhdanov and Vovk, 2009))
Let $A \in \mathbb{R}^{d \times d}$ be a positive definite matrix. Given any sequence $\{(x_t, y_t)\}_{t=1}^{T} \in (\mathbb{R}^d \times \mathbb{R})^T$, define the ridge regression-type predictor

$$\widehat{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{t-1} (y_i - \langle x_i, \theta \rangle)^2 + \theta^\top A \theta.$$

The following identity holds:

$$\sum_{t=1}^{T}(1 - h_{x_t})(y_t - \langle x_t, \widehat{\theta}_t \rangle)^2 = \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T}(y_t - \langle x_t, \theta \rangle)^2 + \theta^\top A \theta \right\}.$$

Although the bound of Theorem 36 can be derived directly using linear algebraic arguments (which is essentially done in the proof of Theorem 4.6 in (Azoury and Warmuth, 2001)), we believe that the following proof strategy is more insightful. It provides this identity without prior knowledge of its possible existence. We note that the proof strategy and explicit identity in Theorem 36 closely align with previous work, specifically with Zhdanov and Vovk (2009, Theorem 1), which establishes this identity for $A = I_d$ and acknowledges that a similar identity appears in Azoury and Warmuth (2001). Additionally, the work by Zhdanov and Kalnishkan (2013) explores sequential formulations of kernel ridge regression with the focus on similar identities. Although this formulation is not novel, it is included here for completeness and to support the transductive prior framework used in this paper. The extension to a general positive definite matrix A as the regularizer corresponds directly to the transductive priors considered, simplifying derivations of bounds such as that in Gaillard, Gerchinovitz, Huard, and Stoltz (2019), as well as additional technical results used throughout the paper.

We will first introduce some auxiliary lemmas. The first result is standard.

Lemma 37 *Let $B \in \mathbb{R}^{n \times n}$ be a full-rank matrix, let x be a vector in $\mathbb{R}^n$, and let $A = B + xx^\top$. Assuming that both A and B are invertible, we have*

$$\frac{\det(B)}{\det(A)} = 1 - x^\top A^{-1} x, \quad and \quad \frac{\det(A)}{\det(B)} = 1 + x^\top B^{-1} x.$$

We also need the following derivation.

Lemma 38 *Assume that $Z \sim \mathcal{N}(\nu, \sigma^2)$. We have for any $y \in \mathbb{R}$,*

$$-\log \left(\mathbb{E} \exp(-(y - Z)^2)) \right) = \frac{(\nu - y)^2}{2\sigma^2 + 1} + \frac{1}{2} \log(2\sigma^2 + 1).$$

Proof Using Theorem 16, we have

$$
\begin{aligned}
-\log \left(\mathbb{E} \exp(-(y - Z)^2)) \right) &= -\log \left(\frac{1}{\sqrt{2\pi}\sigma} \int \exp(-(y - z)^2 - \frac{(z - \nu)^2}{2\sigma^2}) dz \right) \\
&= -\log \left(\frac{\sqrt{\pi}}{\sqrt{2\pi}\sigma \sqrt{1 + 1/(2\sigma^2)}} \exp \left(-\frac{(\nu^2 - 2\nu y + y^2)}{2\sigma^2 + 1} \right) \right) \\
&= -\log \left(\frac{1}{\sqrt{2\sigma^2 + 1}} \exp \left(-\frac{(\nu - y)^2}{2\sigma^2 + 1} \right) \right) \\
&= \frac{(\nu - y)^2}{2\sigma^2 + 1} + \frac{1}{2} \log(2\sigma^2 + 1).
\end{aligned}
$$

∎

We are ready to provide a proof of Theorem 36.

Proof Our prior distribution μ is multivariate Gaussian in $\mathbb{R}^d$ with its density given by

$$\mu(\theta) = \sqrt{\det(A)} \cdot \exp\left(-\pi\theta^\top A\theta\right).$$

We want to compute the right-hand side of (11) for the class (48). The first computation uses Theorem 16. We have

$$-\log\left(\mathbb{E}_{\theta\sim\mu}\exp\left(-\sum_{t=1}^{T}\pi(y_t - \langle x_t, \theta\rangle)^2\right)\right) \tag{49}$$

$$= -\log\left(\sqrt{\det(A)}\int_{\theta\in\mathbb{R}^d}\exp\left(-\sum_{t=1}^{T}\pi(y_t - \langle x_t, \theta\rangle)^2 - \pi\theta^\top A\theta\right)\right)$$

$$= \inf_{\theta\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\pi(y_t - \langle x_t, \theta\rangle)^2 + \pi\theta^\top A\theta\right\} - \frac{1}{2}\log(\det(A)) + \frac{1}{2}\log\left(\det\left(\sum_{t=1}^{T}x_t x_t^\top + A\right)\right). \tag{50}$$

Another perspective on (49) is the following. We assume that at step t we predict with the distribution ρ_t over $\mathbb{R}^d$ using the exponential weights over the class (48) with respect to the logarithmic loss. By Theorem 16, we have

$$-\log\left(\mathbb{E}_{\theta\sim\mu}\exp\left(-\sum_{t=1}^{T}\pi(y_t - \langle x_t, \theta\rangle)^2\right)\right) = -\sum_{t=1}^{T}\log\left(\mathbb{E}_{\theta\sim\rho_t}\exp\left(-\pi(y_t - \langle x_t, \theta\rangle)^2\right)\right). \tag{51}$$

It is only left to compute individual terms (mix-losses) in the right-hand side of the above identity. By the definition of the exponential weights algorithm, we have

$$\rho_t \propto \exp\left(-\sum_{i=1}^{t-1}\pi(y_i - \langle x_i, \theta\rangle)^2 - \pi\theta^\top A\theta\right).$$

Clearly, the distribution ρ_t is multivariate Gaussian. Using Theorem 16, we can immediately obtain

$$\rho_t \sim \mathcal{N}\left(\widehat{\theta}_t, \frac{1}{2\pi}\left(\sum_{i=1}^{t-1}x_i x_i^\top + A\right)^{-1}\right).$$

Observe that if $\theta \sim \rho_t$, then $\langle x_t, \theta\rangle \sim \mathcal{N}\left(\langle\widehat{\theta}_t, x_t\rangle, \frac{1}{2\pi}x_t^\top\left(\sum_{i=1}^{t-1}x_i x_i^\top + A\right)^{-1}x_t\right)$, which implies by Theorem 37 that the following holds

$$\langle x_t, \theta\rangle \sim \mathcal{N}\left(\langle\widehat{\theta}_t, x_t\rangle, \frac{1}{2\pi}\frac{h_{x_t}}{1 - h_{x_t}}\right).$$

Finally, we have by Theorem 38

$$-\log\left(\mathbb{E}_{\theta\sim\rho_t}\exp\left(-\pi(y_t-\langle x_t,\theta\rangle)^2\right)\right) = -\log\left(\mathbb{E}_{\theta\sim\rho_t}\exp\left(-(\sqrt{\pi}y_t-\sqrt{\pi}\langle x_t,\theta\rangle)^2\right)\right)$$

$$= \frac{(\sqrt{\pi}\langle\widehat{\theta}_t,x_t\rangle-\sqrt{\pi}y_t)^2}{\frac{h_{x_t}}{1-h_{x_t}}+1} + \frac{1}{2}\log\left(\frac{h_{x_t}}{1-h_{x_t}}+1\right)$$

$$= \pi(1-h_{x_t})(\langle\widehat{\theta}_t,x_t\rangle-y_t)^2 + \frac{1}{2}\log\left(\frac{1}{1-h_{x_t}}\right).$$

Summing the above identity with respect t and using (11) we obtain

$$-\log\left(\mathbb{E}_{\theta\sim\mu}\exp\left(-\sum_{t=1}^T\pi(y_t-\langle x_t,\theta\rangle)^2\right)\right)$$

$$= \sum_{t=1}^T\left(\pi(1-h_{x_t})(\langle\widehat{\theta}_t,x_t\rangle-y_t)^2 + \frac{1}{2}\log\left(\frac{1}{1-h_{x_t}}\right)\right). \tag{52}$$

Denote $M_t = x_t^\top\left(\sum_{i=1}^{t-1}x_ix_i^\top+A\right)^{-1}x_t$. Finally, by Theorem 38 we have

$$\sum_{t=1}^T\frac{1}{2}\log\left(\frac{1}{1-h_{x_t}}\right) = \frac{1}{2}\sum_{t=1}^T\log\left(\frac{\det(M_{t+1})}{\det(M_t)}\right)$$

$$= -\frac{1}{2}\log(\det(A)) + \frac{1}{2}\log\left(\det\left(\sum_{t=1}^Tx_tx_t^\top+A\right)\right). \tag{53}$$

where in the last computation, we used the telescopic sum. Combining (50), (52), and (53), we prove the claim. ∎

The result of Theorem 36 immediately implies the following classical result.

Corollary 39 (The regret of the non-linear predictor of Vovk (2001)) *Fix $\lambda > 0$ and for $x \in \mathbb{R}^d$ consider a sequence of estimators*

$$\widehat{\theta}_{t,x} = \arg\min_{\theta\in\mathbb{R}^d}\sum_{i=1}^{t-1}(y_i-\langle x_i,\theta\rangle)^2 + \lambda\|\theta\|^2 + (\langle x,\theta\rangle)^2.$$

Assume that $|y_t| \le m$ for $t = 1,\dots,T$. Then, the following regret bound holds

$$\sum_{t=1}^T(y_t-\langle x_t,\widehat{\theta}_{t,x_t}\rangle)^2 \le \inf_{\theta\in\mathbb{R}^d}\left\{\sum_{t=1}^T(y_t-\langle x_t,\theta\rangle)^2 + \lambda\|\theta\|^2\right\} + m^2\log\left(\det\left(I_d+\frac{1}{\lambda}\sum_{t=1}^Tx_tx_t^\top\right)\right).$$

Proof To see how the implication, one should first verify, using Theorem 37, that

$$\langle x_t,\widehat{\theta}_{t,x_t}\rangle = (1-h_{x_t})\langle x_t,\widehat{\theta}_t\rangle,$$

where

$$\widehat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{t-1} (y_i - \langle x_i, \theta \rangle)^2 + \lambda \|\theta\|^2, \quad \text{and} \quad h_{x_i} = x_i^\top \left(\sum_{s=1}^{i} x_s x_s^\top + \lambda I_d \right)^{-1} x_i. \quad (54)$$

Combining this with Theorem 36, we have

$$\sum_{t=1}^{T} (y_t - \langle x_t, \widehat{\theta}_{t,x_t} \rangle)^2 = \sum_{t=1}^{T} (1 - h_{x_t})(y_t - \langle x_t, \widehat{\theta}_t \rangle)^2 + \sum_{t=1}^{T} (h_{x_t}^2 - h_{x_t})(\langle x_t, \widehat{\theta}_t \rangle)^2 + \sum_{t=1}^{T} h_{x_t} y_t^2$$

$$= \inf_{\theta \in \mathbb{R}^d} \left(\sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 + \lambda \|\theta\|^2 \right) + \sum_{t=1}^{T} (h_{x_t}^2 - h_{x_t})(\langle x_t, \widehat{\theta}_t \rangle)^2$$

$$+ \sum_{t=1}^{T} h_{x_t} y_t^2.$$

This is the exact expression of regret. We can further simplify it. First, we have

$$\sum_{t=1}^{T} (h_{x_t}^2 - h_{x_t})(\langle x_t, \widehat{\theta}_t \rangle)^2 \leq 0,$$

since $h_{x_t} \in [0, 1]$. It is only left to bound $\sum_{t=1}^{T} h_{x_t} y_t^2$, which, using (53) and $x \leq \log(1/(1 - x))$ for $x \in (0, 1)$, can be done as follows:

$$\sum_{t=1}^{T} h_{x_t} y_t^2 \leq m^2 \sum_{t=1}^{T} \log \left(\frac{\det \left(\sum_{i=1}^{t} x_i x_i^\top + \lambda I_d \right)}{\det \left(\sum_{i=1}^{t-1} x_i x_i^\top + \lambda I_d \right)} \right) \leq m^2 \log \left(\det \left(I_d + \frac{1}{\lambda} \sum_{t=1}^{T} x_t x_t^\top \right) \right).$$

$$(55)$$

The claim follows. ∎

In the following example, we demonstrate that Theorem 36 implies the bound of Kivinen and Warmuth (1999) (see also (Forster, 1999)) for clipped ridge regression. This bound is also revisited in (Vovk, 2001, Theorem 4) using exp-concavity arguments applied to the squared loss. However, there is an issue with that proof. The exp-concavity argument in (Vovk, 2001, Theorem 4) necessitates that both $|y_t| \leq m$ and $|\langle x_t, \theta \rangle| \leq m$ for θ-s within the considered range. The second assumption is not mentioned in the statement of Theorem 4 in (Vovk, 2001), and indeed, it can be bypassed. Unlike the previous analysis, our approach does not rely on exp-concavity and, therefore, successfully recovers the claimed result without additional boundedness assumptions. The proof below was also presented in (Zhdanov and Vovk, 2009, Corollary 1), and we include it here primarily for the reader's convenience.

Corollary 40 (The regret of the clipped ridge predictor of Kivinen and Warmuth (1999)) *In the setup of Theorem 39, we have*

$$\sum_{t=1}^{T} \left(y_t - \text{clip}_{\mathrm{m}} \left(\langle x_t, \widehat{\theta}_t \rangle \right) \right)^2$$

$$\leq \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 + \lambda \|\theta\|^2 \right\} + 4m^2 \log \left(\det \left(I_d + \frac{1}{\lambda} \sum_{t=1}^{T} x_t x_t^\top \right) \right).$$

where $\widehat{\theta}_t$ is the ridge regression estimator given by (54).

Proof Using Theorem 36, the definition of the clipping operator and (55), we have

$$
\sum_{t=1}^{T} \left(y_t - \text{clip}_m \left(\langle x_t, \widehat{\theta}_t \rangle \right) \right)^2 - \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 + \lambda \|\theta\|^2 \right\}
$$

$$
= \sum_{t=1}^{T} \left(y_t - \text{clip}_m \left(\langle x_t, \widehat{\theta}_t \rangle \right) \right)^2 - \sum_{t=1}^{T} (1 - h_{x_t}) \left(y_t - \langle x_t, \widehat{\theta}_t \rangle \right)^2
$$

$$
\leq \sum_{t=1}^{T} \left(y_t - \text{clip}_m \left(\langle x_t, \widehat{\theta}_t \rangle \right) \right)^2 - \sum_{t=1}^{T} (1 - h_{x_t}) \left(y_t - \text{clip}_m \left(\langle x_t, \widehat{\theta}_t \rangle \right) \right)^2
$$

$$
= \sum_{t=1}^{T} h_{x_t} \left(y_t - \text{clip}_m \left(\langle x_t, \widehat{\theta}_t \rangle \right) \right)^2 \leq 4m^2 \sum_{t=1}^{T} h_{x_t} \leq 4m^2 \log \left(\det \left(I_d + \frac{1}{\lambda} \sum_{t=1}^{T} x_t x_t^\top \right) \right).
$$

The claim follows. ∎

Finally, to recover Theorem 1, originally due to Gaillard, Gerchinovitz, Huard, and Stoltz (2019), using Theorem 36. We exactly repeat the lines of the proof of Theorem 39 with the only change. Now, the matrix A in Theorem 36 is not equal to λI_d, but is equal to $\lambda \sum_{t=1}^{T} x_t x_t^\top$, which is assumed to be invertible. Finally, using an analog of (55), the total loss of the predictor is bounded by

$$
\inf_{\theta \in \mathbb{R}^d} \left(\sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 + \lambda \sum_{t=1}^{T} (\langle x_t, \theta \rangle)^2 \right) + m^2 \log \left(\frac{\det \left(\lambda \sum_{t=1}^{T} x_t x_t^\top + \sum_{t=1}^{T} x_t x_t^\top \right)}{\det \left(\lambda \sum_{t=1}^{T} x_t x_t^\top \right)} \right)
$$

$$
= \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 + \lambda \sum_{t=1}^{T} (\langle x_t, \theta \rangle)^2 \right\} + dm^2 \log \left(1 + \frac{1}{\lambda} \right)
$$

$$
\leq \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 \right\} + \lambda T m^2 + dm^2 \log \left(1 + \frac{1}{\lambda} \right),
$$

where the last line follows from exactly the same computation as in the proof of Theorem 7 in (Gaillard et al., 2019).

Similarly, repeating these lines but instead using the matrix $A = \lambda \sum_{t=1}^{T} x_t x_t^\top$ in the proof of Theorem 40, we obtain the following result.

Corollary 41 *Fix $\lambda > 0$ and consider a sequence of estimators*

$$
\widetilde{\theta}_t = \arg \min_{\theta \in \mathbb{R}^d} \left\{ \sum_{i=1}^{t-1} (y_i - \langle x_i, \theta \rangle)^2 + \lambda \sum_{i=1}^{T} (\langle x_i, \theta \rangle)^2 \right\}.
$$

Assume that $|y_t| \leq m$ for $t = 1, \ldots, T$. Then, the following regret bound holds

$$
\sum_{t=1}^{T} \left(y_t - \text{clip}_m \left(\langle x_t, \widetilde{\theta}_t \rangle \right) \right)^2 \leq \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} (y_t - \langle x_t, \theta \rangle)^2 \right\} + \lambda T m^2 + 4dm^2 \log \left(1 + \frac{1}{\lambda} \right).
$$

Appendix D. Achieving Similar Regret Bounds Using the ε-nets

We note that an approach based on ε-nets typically works for the problems we consider. The idea behind this approach is to build a covering for the corresponding classes using the information on $x_1, \ldots, x_T$, and then run the exponential weights algorithm over the cover. This can be used to obtain the realizable bounds for the binary loss, as implied by the inequality (33) in Theorem 18 when reduced to the binary loss. Indeed, one can discretize the space of linear predictors at the level γ and then obtain an $O(d\log(1/\gamma))$-type regret bound using the Halving algorithm. However, such a procedure is impractical even in the realizable case, as the discretization would require us to aggregate over $O\left(1/\gamma^d\right)$ experts. Similarly, for linear regression, one can construct an ε-net for the clipped class of linear functions as in (Mourtada et al., 2022, Proof of Theorem 3) and aggregate over this net using the exponential weights algorithm with the uniform prior on this ε-net. Similar ε-net based constructions are also popular in the analysis of density estimators and usually lead to the information-theoretically optimal regret bound, but do not result in implementable algorithms; see (Yang and Barron, 1999; Bilodeau, Foster, and Roy, 2023) for related results. We quote Vovk (2006) for a general comment on the ε-net based methods: "Another disadvantage of the metric entropy method is that it is not clear how to implement it efficiently, whereas many other methods are computationally very efficient. Therefore, the results obtained by this method are only a first step, and we should be looking for other prediction strategies, both computationally more efficient and having better performance guarantees." Despite that, for completeness, in what follows, we choose to present the ε-nets construction for the logistic case for assumption-free regret bounds. In our construction of ε-nets, we first clip the logistic function to obtain bounded functions. That is, for any $\theta, x \in \mathbb{R}^d$ and $y \in \{\pm 1\}$, we define the clipped logistic function as

$$f_\theta(x, y) = -\log\left(\frac{1}{2T} + \left(1 - \frac{1}{T}\right)\sigma(y\langle x, \theta\rangle)\right) = -\log\left(\frac{1}{2T} + \left(1 - \frac{1}{T}\right)\frac{1}{1 + e^{-y\langle x, \theta\rangle}}\right). \quad (56)$$

Our goal is to cover the function class $\mathcal{F}_X = \{f_\theta(\cdot, \cdot) : X \times \{\pm 1\} \to [0, 1] \mid \theta \in \mathbb{R}^d\}$ in the L^∞ norm, which is defined as $\|f\|_\infty = \sup_{x,y} |f(x, y)|$. Note that working with a clipped class for the log loss causes only a small overhead (Cesa-Bianchi and Lugosi, 2006, Lemma 9.5) (see also Bilodeau et al. (2023)). Specifically, running the exponential weights algorithm with a uniform prior over the ε-net for a properly chosen scale ε provides an $O(d\log(T))$ regret bound.

Lemma 42 *For any $u_1, u_2 \in [0, 1]$, we have*

$$\left|\log\left(\frac{1}{2T} + \left(1 - \frac{1}{T}\right)u_1\right) - \log\left(\frac{1}{2T} + \left(1 - \frac{1}{T}\right)u_2\right)\right| \le 2T|u_1 - u_2|.$$

Proof [Proof of Theorem 42] Let $h(u) = -\log(1/(2T) + (1 - 1/T)u)$ for $u \in [0, 1]$. Then

$$|h'(u)| = \frac{(1 - 1/T)}{1/(2T) + (1 - 1/T)u} \le 2T.$$

The claim follows. ∎

Lemma 43 (Lemma 6 of Drmota et al. (2024)) *For any covariate set $X = \{x_t\}_{t \in [T]}$, there exists an ε-covering $\mathcal{G}_{X,\varepsilon}$ of the function class $\mathcal{G}_X = \{\sigma(\langle \cdot, \theta \rangle) : X \to [0,1] \mid \theta \in \mathbb{R}^d\}$ in L^∞ norm with at most $O((T/\varepsilon)^d)$ elements.*

One may alternatively prove Lemma 43 using the bound on the covering numbers of VC-subgraph classes (see Györfi et al. (2002, Theorem 9.4)).

Lemma 44 *For any covariate set $X = \{x_t\}_{t \in [T]}$, there exists an ε-covering $\mathcal{F}_{X,\varepsilon}$ of the function class $\mathcal{F}_X = \{f_\theta(\cdot, \cdot) : X \times \{\pm 1\} \to [0,1] \mid \theta \in \mathbb{R}^d\}$ (see (56)) in L^∞ norm with at most $O((2T/\varepsilon)^{2d})$ elements.*

Proof [Proof of Theorem 44] For any $\varepsilon > 0$, let $\varepsilon' = \varepsilon/(2T)$. Let $\mathcal{G}_{X,\varepsilon'}$ be an ε'-covering of the function class $\mathcal{G}_X$ in L^∞ norm with at most $O((2T/\varepsilon)^{2d})$ elements. Then consider

$$\mathcal{F}_{X,\varepsilon} = \left\{ f_{\widetilde{\theta}}(\cdot, \cdot) \mid \sigma(\langle \cdot, \widetilde{\theta} \rangle) \in \mathcal{G}_{X,\varepsilon} \right\}.$$

Then, for any $f_\theta \in \mathcal{F}_X$, there exists $\sigma(\langle \cdot, \widetilde{\theta} \rangle) \in \mathcal{G}_{X,\varepsilon}$ such that $\|\sigma(\langle \cdot, \theta \rangle) - \sigma(\langle \cdot, \widetilde{\theta} \rangle)\|_\infty \leq \varepsilon$. By Theorem 42, we note that for any x,

$$\left| f_\theta(x, 1) - f_{\widetilde{\theta}}(x, 1) \right| \leq 2T \left| \sigma(\langle x, \theta \rangle) - \sigma(\langle x, \widetilde{\theta} \rangle) \right| \leq 2T\varepsilon' = \varepsilon$$

and

$$\left| f_\theta(x, -1) - f_{\widetilde{\theta}}(x, -1) \right| \leq 2T \left| \sigma(-\langle x, \theta \rangle) - \sigma(-\langle x, \widetilde{\theta} \rangle) \right| = 2T \left| \sigma(\langle x, \theta \rangle) - \sigma(\langle x, \widetilde{\theta} \rangle) \right| \leq \varepsilon,$$

where the first equality is due to $\sigma(-u) = 1 - \sigma(u)$ for any $u \in \mathbb{R}$. This implies that

$$\left\| f_\theta(\cdot, \cdot) - f_{\widetilde{\theta}}(\cdot, \cdot) \right\|_\infty = \sup_{x,y} \left| f_\theta(x, y) - f_{\widetilde{\theta}}(x, y) \right| \leq \varepsilon.$$

Thus, $\mathcal{F}_{X,\varepsilon}$ is an ε-covering of $\mathcal{F}_X$. The claim follows. ■