Confidence Sequences for Generalized Linear Models via Regret Analysis

Eugenio Clerico [1], **Hamish Flynn**[1],
Wojciech Kotłowski[2], **and Gergely Neu**[1]

[1] *Universitat Pompeu Fabra, Barcelona, Spain.*
e-mail: eugenio.clerico@gmail.com; hamishedward.flynn@upf.edu; gergely.neu@gmail.com

[2] *Poznań University of Technology, Poznań, Poland.*
e-mail: kotlow@gmail.com

Abstract: We develop a methodology for constructing confidence sets for parameters of statistical models via a reduction to sequential prediction. Our key observation is that for any generalized linear model (GLM), one can construct an associated game of sequential probability assignment such that achieving low regret in the game implies a high-probability upper bound on the excess likelihood of the true parameter of the GLM. This allows us to develop a scheme that we call *online-to-confidence-set conversions*, which effectively reduces the problem of proving the desired statistical claim to an algorithmic question. We study two varieties of this conversion scheme: 1) *analytical* conversions that only require proving the existence of algorithms with low regret and provide confidence sets centered at the maximum-likelihood estimator 2) *algorithmic* conversions that actively leverage the output of the online algorithm to construct confidence sets (and may be centered at other, adaptively constructed point estimators). The resulting methodology recovers all state-of-the-art confidence set constructions within a single framework, and also provides several new types of confidence sets that were previously unknown in the literature.

MSC2020 subject classifications: Primary 62F25, 62L12, 62J12; secondary 68Q87, 68Q32.
Keywords and phrases: confidence sequences, generalized linear models, sequential probability assignment, online learning, regret analysis.

1. Introduction

Building confidence sets for parameters of statistical models is one of the most fundamental questions of statistics. In this paper, we consider this problem in the context of generalized linear models (GLMs), where one has access to a data set of observations $(X^n, Y^n) = (X_t, Y_t)_{t=1}^n$. Here, $X_t \in \mathbb{R}^d$ is a vector of *covariates* (or *features*), $Y_t \in \mathbb{R}$ is a real-valued *label*, and the likelihood of the labels Y_t is given by the exponential-family model

$$p(y|X_t, \theta^\star) = \exp\big(\langle \theta^\star, X_t\rangle\, y - \psi(\langle \theta^\star, X_t\rangle)\big)h(y)\,,$$

with $\theta^\star \in \mathbb{R}^d$ the unknown parameter, $\psi : \mathbb{R} \to \mathbb{R}$ a convex function (often called the *log-partition function*), and $h : \mathbb{R} \to \mathbb{R}$ the *reference distribution* (independent of X_t or $\theta^\star$). The model can be alternatively written as $Y_t = \mu(\langle \theta^\star, X_t\rangle) + \varepsilon_t$, where $\mu = \psi'$ and ε_t is zero-mean noise whose precise distribution depends on $\langle \theta^\star, X_t \rangle$. We allow the covariates to be selected sequentially, which means that the distribution of X_{t+1} is determined by the sequence (X^t, Y^t). We will use $\mathcal{F}_t = \sigma(X_1, Y_1, X_2, Y_2, \ldots, X_t, Y_t, X_{t+1})$ to denote the σ-field generated by the sequence of observations up to index t and the covariate X_{t+1}. This model captures many problems of fundamental interest: the choice $\psi : z \mapsto \frac{z^2}{2}$ yields linear models with Gaussian noise, while logistic regression can be recovered by considering Bernoulli labels $Y_t = \text{Bernoulli}\big(\mu(\langle \theta^\star, X_t \rangle)\big)$, with $\mu : z \mapsto \frac{1}{1+e^{-z}}$ being the sigmoid link function. The corresponding log-partition function in this case is the logistic loss $\psi(z) = \log(1 + e^z)$.

In this work, we aim to develop tight *confidence sequences* for the true parameter $\theta^\star$, which are sequences of sets $\Theta_1, \ldots, \Theta_n \subseteq \mathbb{R}^d$ such that each Θ_t is determined by (X^t, Y^t) and $\theta^\star$ is included in *all* sets with high probability. Formally, we require $\mathbb{P}\left[\exists n : \theta^\star \notin \Theta_n\right] \leq \delta$ for some given $\delta > 0$. A set satisfying $\mathbb{P}\left[\theta^\star \notin \Theta_n\right] \leq \delta$ for a specific value of n will be referred to as a *confidence set.* We will largely focus on confidence sets and sequences in which each set is defined as the set of parameters whose likelihood is nearly maximal:

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \frac{\prod_{t=1}^n p(Y_t | X_t, \theta)}{\sup_{\theta' \in \mathbb{R}^d} \prod_{t=1}^n p(Y_t | X_t, \theta')} \geq e^{-\beta_n} \right\}. \tag{1}$$

The challenge then is to design suitable choices of the *confidence width* $\beta_n > 0$ such that $\Theta_1, \Theta_2, \ldots$ is guaranteed to be a valid confidence sequence. In this paper, we develop a methodology that reduces the *statistical* problem of designing an appropriate sequence of confidence widths to the *algorithmic* problem of showing the existence of a sequential prediction algorithm with guaranteed bounds on their performance (as measured by the standard notion of *regret*, to be defined shortly).

Our methodology can be most easily introduced via the following simple example that considers the linear-Gaussian case. For such models, the confidence set of Equation (1) can be equivalently written in terms of the squared loss function $\ell_t(\theta) = \frac{1}{2}\left(\langle \theta, X_t \rangle - Y_t\right)^2$, as

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \inf_{\theta' \in \mathbb{R}^d} \sum_{t=1}^n \left(\ell_t(\theta) - \ell_t(\theta')\right) \leq \beta_n \right\}. \tag{2}$$

For the sake of simplicity, let us now suppose that the maximum-likelihood estimator (MLE) $\widehat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^n \ell_t(\theta)$ exists. In this case, it can be easily shown that the set Θ_n is an ellipsoid centered at $\widehat{\theta}_n$. In order to provide a suitable choice of the width β_n of the ellipsoid, we will consider a game between an *online learner* and its *environment*, where in each round $t = 1, 2 \ldots, n$, the following steps are repeated:

1. the environment reveals X_t to the online learner,
2. the online learner picks a parameter $\theta_t \in \mathbb{R}^d$,
3. the environment reveals Y_t to the online learner, and

4. the online learner incurs the loss $\ell_t(\theta_t)$.

The performance of the online learner is measured in terms of its *regret* against a (potentially data-dependent) comparator $\theta' \in \mathbb{R}^d$, defined as

$$\mathrm{regret}_n(\theta') = \sum_{t=1}^{n} \big(\ell_t(\theta_t) - \ell_t(\theta')\big).$$

Our key observation is that the excess log-likelihood of the true parameter $\theta^\star$ relative to the MLE can be decomposed as follows:

$$\sum_{t=1}^{n} \left(\ell_t\left(\theta^\star\right) - \ell_t\left(\widehat{\theta}_n\right)\right) = \underbrace{\sum_{t=1}^{n} \left(\ell_t(\theta_t) - \ell_t\left(\widehat{\theta}_n\right)\right)}_{\mathrm{regret}_n\left(\widehat{\theta}_n\right)} + \underbrace{\sum_{t=1}^{n} \left(\ell_t(\theta^\star) - \ell_t(\theta_t)\right)}_{\leq \log\frac{1}{\delta} \text{ with high prob}}.$$

Here, we have noticed that the first term in the decomposition corresponds to the regret of the online learner against $\widehat{\theta}_n$, and that since the online learner has to make its prediction θ_t before seeing the label Y_t, it cannot have lower squared-loss prediction error than the true parameter $\theta^\star$, in expectation at least. As a result, the last term in the decomposition can be shown to be a light-tailed supermartingale. Notably, the argument holds for *any* online learning algorithm, and thus in order to construct a confidence set of the form given in Equation (2) with small β_n, one only needs to show the *existence* of an online learner with low regret against the data-dependent comparator point $\widehat{\theta}_n$.

For the online-learning game considered above, the best possible regret guarantees are of the order $\max_t Y_t^2 \cdot d \log n$, achieved by the so-called Vovk–Azoury–Warmuth forecaster (cf. Azoury and Warmuth, 2001, Vovk, 2001, and Section 11.8 in Cesa-Bianchi and Lugosi, 2006). Ultimately, this leads to a confidence width β_n of the order $d\left(\log n\right)^2 + \log\frac{1}{\delta}$, which is inferior to the optimal rate of $d \log n + \log\frac{1}{\delta}$, suggesting that the reduction we propose above is unlikely to yield tight confidence sets, at least in the simple form defined above. We address this issue by considering a closely related online learning game of *sequential probability assignment*, which allows us to derive confidence sequences of *optimal* width in a number of important cases of interest.

Besides making the above argument fully rigorous, in what follows we will extend this simple approach along multiple axes, such as going beyond linear models, providing reductions to more flexible and powerful online algorithms, or considering confidence sets of alternative shapes. We will refer to our general methodology as *Online-to-Confidence-Set Conversions* (following the terminology introduced by Abbasi-Yadkori, Pál and Szepesvári (2012), as discussed below). One aspect of the conversion scheme that we will particularly develop is the manner in which the online algorithm is used for constructing the confidence sequence. In the argument above, we have used the online algorithm only for the sake of analysis: there was no need to ever run the algorithm, as all that the argument needed was showing the existence of a method with low regret. We will call such conversions *analytic*. Alternatively, one can use the output of the online algorithm more actively in constructing the confidence sequence, by centering the sets at parameters other than the MLE $\widehat{\theta}_n$. We will refer to such conversion schemes as *algorithmic*.

The idea of using sequential predictions to construct confidence sets is far from being new: it goes back at least to the work of Robbins and Siegmund (1970). More recently, the same idea has resurfaced in the work of Abbasi-Yadkori, Pál and Szepesvári (2012) in a form very similar to our framework. We acknowledge this similarity by adopting their aptly chosen term "online-to-confidence-set conversion", whose meaning we expand significantly in our work. In particular, the results of Abbasi-Yadkori, Pál and Szepesvári (2012) fall into the category of *algorithmic* online-to-confidence-set conversions; our framework highlights the broader context of their work by placing it into a larger system of reduction schemes. More recently, online-to-confidence-set conversions were used in a variety of settings, but mostly in an ad-hoc way specialized to narrow subclasses of GLMs (Jun et al., 2017; Lee, Yun and Jun, 2024a; Emmenegger, Mutny and Krause, 2023). Further related works that make use of similar ideas without making the connection with regret analysis explicit include Gales, Sethuraman and Jun (2022), Flynn et al. (2023), and Lee, Yun and Jun (2024b). The results in the present work recover all of these as special cases, and in several cases improve over them in various aspects.

More broadly, this work fits into a recent wave of statistical literature sometimes referred to as *algorithmic statistics*. Algorithmic statistics is intended here[1] as the statistical counterpart of *algorithmic probability theory*, an alternative foundation to classic measure-theoretic probability theory based on game-theoretic and algorithmic constructions rooted in the later works of Kolmogorov—see the book of Shafer and Vovk (2001) for a detailed development of this theory, and Vovk and Shafer (2003) for a brief historical overview. The most notable representatives of the line of work we refer to as algorithmic statistics are the works Orabona and Jun (2024); Waudby-Smith and Ramdas (2023), which reduce the problem of mean estimation of bounded random variables to a sequential prediction problem, constructing a martingale whose concentration properties are exploited. In particular, Waudby-Smith and Ramdas (2023) takes an approach where by explicitly running a sequential online strategy one builds the martingale, while the take of Orabona and Jun (2024) is closer to our analytic method, where the sequential strategy does not need to be played in practice as the desired result directly follows from the existence of a suitable regret bound for the strategy considered. A similar analytic reduction was also proposed in the online-to-PAC framework of Lugosi and Neu (2024), where the existence of a regret bound for a specifically designed online game yields a generalization bound for a statistical learning problem. All of these works drew inspiration (at least implicitly) from Rakhlin and Sridharan (2017), who were the first to explicitly point out the tight connection between concentration inequalities and regret analysis, and have proved a general *equivalence* result between martingale

[1] While closely related, the work of Gács, Tromp and Vitányi (2001) titled "Algorithmic Statistics" studies a more narrow set of statistical approaches, which we regard as only one part algorithmic statistics in our terminology.

tail bounds and regret bounds. In a way, all works listed above are merely turning this general observation into concrete, quantitative results by repurposing regret minimization algorithms for statistical inference, often leading to major improvements over the best previously known results in this context. Our work fits directly into this theme as well.

An important additional note about related work is that the results presented in the present manuscript have a considerable overlap with the very recent preprint of Kirschner et al. (2025) that proposes a general framework for deriving confidence sequences. The basic principles underlying their framework are essentially the same as the ones we have used as starting point, and many general observations appear in both works. Most notably, our analytic online-to-confidence-set conversion appears in nearly identical form in their Section 3.3. Unlike our work that exhibits a broad range of confidence sets from this specific reduction, Kirschner et al. (2025) operate at a different level of generality and present online-to-confidence-set conversions as one of several methodologies for developing confidence sets via sequential likelihood ratios. We regard this parallel[2] work as complimentary to ours.

The rest of the paper is organized as follows. In Section 2, we fill the gaps in the definitions given above by introducing the full set of our assumptions and describing our online-to-confidence-set framework in full technical detail. In Section 3, we present a variety of regret bounds that we will combine with our reduction framework in Section 4 to derive concrete confidence sets for GLMs. We close with a final discussion of the framework, the results, and future research directions in Section 5.

Notation. For any measurable space $\mathcal{S}$, we use $\Delta_{\mathcal{S}}$ to denote the set of all probability distributions on $\mathcal{S}$, and we will use the notation $s^n = (s_1, s_2, \ldots, s_n)$ to denote sequences in $\mathcal{S}^n$. We will use $\mathbb{I}$ to denote the 0-1 indicator function, taking value $\mathbb{I}_{\{E\}} = 1$ if the logical expression E is true and $\mathbb{I}_{\{E\}} = 0$ otherwise. For any positive integer d, $[d] = \{1, \ldots, d\}$ is the set of integers from 1 to d.

2. Online-to-confidence-set conversions

We now introduce our general framework for constructing confidence sets via a reduction to regret analysis of online algorithms. We will consider generalized linear models (GLMs) as described at the beginning of Section 1, and will use the notation

$$\ell_t(\theta) = -\langle \theta, X_t \rangle Y_t + \psi(\langle \theta, X_t \rangle) - \log h(Y_t)\,,$$

corresponding to the negative log likelihood $-\log(p(Y_t|X_t, \theta))$ of Y_t given X_t and θ. In our framework, we will often think of ℓ_t as a *loss function* associated with the statistical model, and study algorithms that aim to "minimize" this loss in an appropriate sense. Notice that the last term in ℓ_t is independent of θ, and can be often omitted when analyzing such algorithms.

[2] We have verified via personal communication with the authors that they agree with this interpretation.

Instead of the online-learning game outlined in Section 1, we will consider the more flexible setup of *sequential probability assignment* (also often called *online prediction under the logarithmic loss*). A crucial component of this setup is the *logarithmic loss* (or simply the *log loss*) associated with a distribution $P \in \Delta_{\mathcal{Y}}$ with density p and example X_t, Y_t, defined as

$$\mathcal{L}_t(p) = -\log p(Y_t).$$

Concretely, we we will consider a sequential game between an online learner and its environment, with the following steps repeated in each round $t = 1, 2, \ldots, n$:

1. the environment reveals X_t to the online learner,
2. the online learner picks a distribution $P_t \in \Delta_{\mathcal{Y}}$ with density function p_t,
3. the environment reveals Y_t to the online learner, and
4. the online learner incurs the log loss $\mathcal{L}_t(p_t) = -\log p_t(Y_t)$.

The performance of an online learning algorithm producing outputs $p^n = (p_1, \ldots, p_n)$ is measured in terms of its *regret* against a comparator strategy that makes its predictions according to $p(\cdot|X_t, \bar{\theta})$ for some fixed $\bar{\theta} \in \Theta$:

$$\text{regret}_{p^n}(\bar{\theta}) = \sum_{t=1}^{n} \left(\mathcal{L}_t(p_t) - \mathcal{L}_t\left(p(\cdot|X_t, \bar{\theta})\right)\right) = \sum_{t=1}^{n} \left(\mathcal{L}_t(p_t) - \ell_t(\bar{\theta})\right) , \tag{3}$$

where we noticed in the last step that $\mathcal{L}_t\left(p(\cdot|X_t, \bar{\theta})\right) = \ell_t(\bar{\theta})$.

We will play special attention to *mixture forecasters* that pick a distribution $q_t \in \Delta_\Theta$ over the parameter space in each round t, and use the following mixture distribution for predicting Y_t:

$$p_t = \int_\Theta p(\cdot|X_t, \theta) \mathrm{d}q_t(\theta).$$

With a slight abuse of our earlier notation, we use $\mathcal{L}_t$ to denote loss associated with a mixture $q \in \Delta_\Theta$:

$$\mathcal{L}_t(q) = -\log \int_\Theta p(Y_t|X_t, \theta) \mathrm{d}q(\theta) = -\log \int_\Theta e^{-\ell_t(\theta)} \mathrm{d}q(\theta). \tag{4}$$

The regret of a mixture forecaster producing the sequence $q^n = (q_1, \ldots, q_n)$ is defined accordingly as

$$\text{regret}_{q^n}(\bar{\theta}) = \sum_{t=1}^{n} \left(\mathcal{L}_t(q_t) - \ell_t(\bar{\theta})\right) . \tag{5}$$

If $q^n = (q_1, \ldots, q_n)$ is a sequence of Dirac delta distributions supported on $\theta^n = (\theta_1, \ldots, \theta_n)$, the game corresponds to the setup described in Section 1, and we write

$$\text{regret}_{\theta^n}(\bar{\theta}) = \sum_{t=1}^{n} \left(\ell_t(\theta_t) - \ell_t(\bar{\theta})\right) . \tag{6}$$

The problem of designing algorithms with guaranteed bounds on their worst-case regret is well-studied within theoretical computer science, statistics and machine-learning theory. We review the most effective methods and state the guarantees most relevant to our setting in Section 3.

In what follows, we propose a comprehensive methodology for deriving confidence sets for GLMs using various reductions to the sequential prediction game defined above. We will collectively refer to these techniques as *online-to-confidence-set conversions*, and classify them into the following two types: *analytic* conversions that prove validity of confidence sets centred at arbitrary data-dependent parameters, and only use regret analysis within the proofs, and *algorithmic* conversions that make use of the sequence of predictions made by the online algorithm for constructing the confidence set. Both conversion schemes build on the concept of *sequential likelihood ratios*, which we present first below, followed by the detailed description of the conversions.

2.1. Sequential likelihood ratios

For any $\mathcal{F}_{t-1}$-measurable prediction $p_t \in \Delta_{\mathcal{Y}}$, the difference $\big(\ell_t(\theta^\star) - \mathcal{L}_t(p_t)\big)$ can be seen to be the logarithm of a likelihood ratio statistic:

$$\ell_t(\theta^\star) - \mathcal{L}_t(p_t) = \log p_t(Y_t) - \log(p(Y_t|X_t, \theta^\star)) = \log\left(\frac{p_t(Y_t)}{p(Y_t|X_t,\theta^\star)}\right).$$

Thus, the sum of these differences is a *sequential likelihood ratio*, to which the following classic result applies:

Proposition 2.1. *Let $p^n = (p_1, \dots, p_n)$ be a sequence of distributions over $\mathcal{Y}$ such that each p_t is $\mathcal{F}_{t-1}$-measurable. Then, for any $\delta > 0$,*

$$\mathbb{P}\left[\exists n \geq 1 : \sum_{t=1}^{n} \ell_t(\theta^\star) - \sum_{t=1}^{n} \mathcal{L}_t(p_t) \geq \log(1/\delta)\right] \leq \delta\,. \tag{7}$$

Proof. First, note that we can write

$$\begin{aligned}\sum_{t=1}^{n}\big(\ell_t(\theta^\star) - \mathcal{L}_t(p_t)\big) &= \sum_{t=1}^{n} \log p_t(Y_t) - \sum_{t=1}^{n} \log(p(Y_t|X_t,\theta^\star)) \\ &= \log\left(\frac{\prod_{t=1}^n p_t(Y_t)}{\prod_{t=1}^n p(Y_t|X_t,\theta^\star)}\right).\end{aligned}$$

Let $M_n = \prod_{t=1}^n \frac{p_t(Y_t)}{p(Y_t|X_t,\theta^\star)}$. Using that p_t and X_t are $\mathcal{F}_{t-1}$ measurable, we easily see that this defines a non-negative martingale. Indeed,

$$\mathbb{E}[M_t|\mathcal{F}_{t-1}] = M_{t-1}\int_{\mathcal{Y}} \frac{p_t(y)p(y|X_t,\theta^\star)}{p(y|X_t,\theta^\star)}\mathrm{d}y = M_{t-1}\int_{\mathcal{Y}} p_t(y)\mathrm{d}y = M_{t-1}\,.$$

Applying Ville's inequality yields

$$\mathbb{P}\left[\forall n \geq 1 : \log M_n \leq \log\frac{1}{\delta}\right] \geq 1 - \delta\,.$$

□

This result is of course classic: it can be traced back to Ville (1939), Wald (1945), and Robbins (1970), and is sometimes known as the "no-hypercompression inequality" (cf. Grünwald (2007), Chapter 3). In words, it states that no matter how each prediction p_t is selected, the associated total log loss will typically be larger than the loss incurred by the "oracle" predictor that knows the true parameter $\theta^\star$. This observation can already be used to construct valid confidence sequences for $\theta^\star$: indeed, one just needs to execute an online learning algorithm to produce a sequence of predictions $p_1, \dots, p_n$, and consider the set of all parameters θ with log-likelihood not much smaller than $\sum_{t=1}^n \mathcal{L}_t(p_t)$. This is in fact a classic recipe that has been used for designing confidence sets at least since the work of Robbins and Siegmund (1970). The tightness of the resulting confidence sets, however, depends on the quality of the sequence of predictions: the smaller the total total loss $\sum_{t=1}^n \mathcal{L}_t(p_t)$ is, the better the obtained confidence sets. The most straightforward way to achieve this goal is then to use one of the many known online algorithms guaranteeing low worst-case regret to generate the sequence $p_1, \dots, p_n$, but this approach has several downsides. Most notably, algorithms guaranteeing low regret are often computationally intractable, and thus obtaining tight bounds may be very expensive (or not affordable at all). This limitation is often significant, in that cheaper alternatives may very well result in much worse guarantees (as is the case in the linear regression example described in Section 1). This limitation is addressed by the conversion scheme to be described next.

2.2. Analytic online-to-confidence-set conversions

As an alternative to the naïve method suggested above, one may opt to define confidence sets of the following simpler form instead:

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^n \left(\ell_t(\theta) - \ell_t(\overline{\theta}_n) \right) \leq \beta_n \right\}, \tag{8}$$

where $\overline{\theta}_n \in \mathbb{R}^d$ is a data-dependent *reference parameter*, which may be much cheaper to compute than a sequence of well-performing predictions $p_1, \dots, p_n$. An easy choice can be the MLE $\widehat{\theta}_n = \min_{\theta \in \mathbb{R}^d} \sum_{t=1}^n \ell_t(\theta)$ (given that it exists). For such a confidence set, the main challenge is to choose the confidence-width parameter β_n in a way that Θ_n includes the true parameter $\theta^\star$ with high probability. Our first main result is the following theorem, which shows that a suitable β_n can be found via a reduction to regret analysis.

Theorem 2.2. *Let $p^n = (p_1, \dots, p_n)$ be a sequence of distributions over $\mathcal{Y}$ such that each p_t is $\mathcal{F}_{t-1}$-measurable. Then, for any $\delta \in (0,1)$, the set defined in Equation (8) satisfies $\mathbb{P}\left[\exists n : \theta^\star \notin \Theta_n\right] \leq \delta$ with the choice*

$$\beta_n = \mathrm{regret}_{p^n}(\overline{\theta}_n) + \log \frac{1}{\delta}.$$

Proof. Fix an online learning algorithm and, for any $\theta \in \mathbb{R}^d$, write

$$\sum_{t=1}^{n} \left(\ell_t\left(\theta\right) - \ell_t\left(\overline{\theta}_n\right)\right) = \underbrace{\sum_{t=1}^{n} \left(\mathcal{L}_t(p_t) - \ell_t\left(\overline{\theta}_n\right)\right)}_{\text{regret}_{p^n}\left(\overline{\theta}_n\right)} + \underbrace{\sum_{t=1}^{n} \left(\ell_t(\theta) - \mathcal{L}_t(p_t)\right)}_{M_n(\theta)},$$

where we defined $M_n(\theta)$ in the last line, and identified the first sum in the decomposition as the regret of the online algorithm against $\overline{\theta}_n$. The claim then follows by applying Proposition 2.1 to show that $\sup_{n \geq 1} M_n(\theta^\star) \leq \log \frac{1}{\delta}$ holds with probability at least $1 - \delta$. □

In words, Theorem 2.2 states that if we can upper bound the worst-case regret of an online algorithm, then the confidence sequence of Equation (8) will include the true parameter $\theta^\star$ with high probability. Note that the online algorithm does not need to be executed for constructing the confidence set, and one only needs to show the *existence* of an algorithm with low regret for the sake of analysis—which is the reason why we refer to this conversion technique as "analytic". This addresses the computational concerns with the naïve sequential likelihood-ratio method we discussed in Section 2.1: since there is no need to actually run the online learning algorithm, we can choose the method with the best regret bound regardless of how impractical it may be. Furthermore, it is important to note that typical regret bounds are guaranteed to hold with probability 1 for all sequences (X_t, Y_t), and thus the conversion scheme suggested by the above theorem cleanly splits the complex statistical problem of designing a suitable confidence set into a purely statistical question (bounding a supermartingale via Proposition 2.1) and a purely algorithmic question (bounding the regret).

Not having to run the online algorithm comes with other major conceptual advantages. In particular, the algorithm used in this analytic construction may make use of information that would not be available for a less sophisticated approach. Most remarkably, the online learner may even have access to $\theta^\star$ and use it in making its predictions. Other improvements can be achieved when the sequence of covariates $X_1, \dots, X_n$ is fixed in advance or is drawn i.i.d., by revealing this sequence to the online learner ahead of time (known as *transductive online learning*—see Section 4.1.2.) Finally, note that the tightness of the set is also influenced by the choice of the reference parameter $\overline{\theta}_n$, as can be readily observed after rewriting Θ_n as

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \ell_t(\theta) \leq \sum_{t=1}^{n} \ell_t(\overline{\theta}_n) + \text{regret}_{p^n}(\overline{\theta}_n) + \log \frac{1}{\delta} \right\}.$$

Thus, the confidence width is subject to a tradeoff between the total loss of the reference point and the best achievable regret against the chosen $\overline{\theta}_n$. One can thus restrict the online algorithm to only aim to compete with specific choices of $\overline{\theta}_n$ that are known to be "good enough" in terms of total loss, and enjoy the benefits of reduced regret against this restricted comparator. We demonstrate the power of this general approach in Section 4.1, where we use it to recover and tighten a number of state-of-the-art concentration inequalities.

2.3. Algorithmic online-to-confidence-set conversions

The analytic conversions we presented in the previous section provide a rigorous alternative to the naïve sequential likelihood ratio method outlined in Section 2.1. The tightness of the resulting bounds depended on the best achievable regret against the data-dependent comparator $\overline{\theta}_n$. It is natural to ask if it is possible to design confidence sets whose size depends on the best achievable regret against the *true parameter* $\theta^\star$ instead, to make use of potential additional structure that may be present in the problem at hand. In this section, we provide a methodology that achieves this. In what follows, we present a simplified recipe for deterministic forecasters that select each q_t as a point mass concentrated on θ_t, and provide a tighter (but technically more involved) result in Section 4.2.

Our algorithmic conversions are stated using the concept of *η-shifted losses*, defined for any $\eta \in (0, 1]$ as

$$\ell_t^{(\eta)}(\theta) = \ell_t\big(\eta\theta + (1-\eta)\theta^\star\big)\,.$$

Clearly, we have that for $\eta = 1$ this loss coincides with ℓ_t. We remark that, for any η, we have that $\ell_t^{(\eta)}(\theta^\star) = \ell_t(\theta^\star)$. The following proposition gives a result analogous to Proposition 2.1 for the total losses of mixture forecasters. Let us also define the associated *shifted log loss* of a mixture forecaster as

$$\mathcal{L}_t^{(\eta)}(q) = -\log \int e^{-\ell_t^{(\eta)}(\theta)} \mathrm{d}q(\theta)\,,$$

where $q \in \Delta_\Theta$. For $\eta = 1$, this corresponds to the loss $\mathcal{L}_t$ for the mixture forecaster defined in Equation (4).

Proposition 2.3. *Let $q^n = (q_1, \ldots, q_n)$ be a sequence of distributions on Θ such that each q_t is $\mathcal{F}_{t-1}$-measurable. Then, for any $\eta \in (0, 1]$, and any $\delta > 0$,*

$$\mathbb{P}\left[\exists n \geq 1 \,:\, \sum_{t=1}^{n} \ell_t(\theta^\star) - \sum_{t=1}^{n} \mathcal{L}_t^{(\eta)}(q_t) \geq \log(1/\delta)\right] \leq \delta\,.$$

Proof. First, note that

$$\begin{aligned}\mathbb{E}\left[e^{\ell_n(\theta^\star) - \mathcal{L}_t^{(\eta)}(q_n)}\middle|\,\mathcal{F}_{n-1}\right] &= \int \mathbb{E}\left[e^{\ell_n(\theta^\star) - \ell_t^{(\eta)}(\theta)}\middle|\,\mathcal{F}_{n-1}\right] \mathrm{d}q_n(\theta)\\ &= \int \frac{\mathbb{E}\left[e^{\eta\langle X_t, \theta - \theta^\star\rangle Y_t}\middle|\,\mathcal{F}_{n-1}\right]}{\exp\big(\psi\big(\langle\theta^\star, X_t\rangle + \eta\langle\theta - \theta^\star, X_t\rangle\big) - \psi\big(\langle\theta^\star, X_t\rangle\big)\big)} \mathrm{d}q_n(\theta) = 1\,,\end{aligned}$$

where we used that for an exponential family, the equality

$$\log \mathbb{E}[e^{\beta Y_t}|\mathcal{F}_{t-1}] = \psi(\beta + \langle\theta^\star, X_t\rangle) - \psi(\langle\theta^\star, X_t\rangle)$$

holds for any $\mathcal{F}_{t-1}$-measurable β. So, we have proved that for any t, we have

$$\mathbb{E}[M_t|\mathcal{F}_{t-1}] = M_{t-1}\mathbb{E}[\exp(\ell_t(\theta^\star) - \mathcal{L}_t^{(\eta)}(q_t))|\mathcal{F}_{t-1}] = M_{t-1}\,.$$

with $M_t = \exp\Big(\sum_{s=1}^t \ell_s(\theta^\star) - \sum_{s=1}^n \mathcal{L}_s^{(\eta)}(q_s)\Big)$. This yields that M_t is a non-negative martingale, and the desired result follows by Ville's inequality. □

Using the above result, we will derive confidence sets in terms of the function

$$d_\psi(z,z') = \frac{1}{2}\psi(z) + \frac{1}{2}\psi(z') - \psi(z/2 + z'/2).$$

For any ψ, d_ψ is symmetric and satisfies $d_\psi(z,z) = 0$ for every z. Thanks to the convexity of ψ, d_ψ is nonnegative (and also positive definite if ψ is strictly convex, i.e., $d_\psi(z,z') \neq 0$ when $z \neq z'$). However, in general, d_ψ is not a proper distance because it does not satisfy the triangle inequality. Loosely speaking, the rate at which $d_\psi(z,z')$ increases as z' moves further away from z is determined by the curvature of ψ. Our main result of this section provides a confidence set for $\theta^\star$ based on the sequence $q_1,\dots,q_n$ output by an online learning algorithm, in terms of the function d_ψ.

Theorem 2.4. *Let $\theta^n = (\theta_1,\dots,\theta_n)$ be a sequence of parameters in Θ such that each θ_t is $\mathcal{F}_{t-1}$-measurable. Then, for any $\delta \in (0,1)$,*

$$\mathbb{P}\left[\exists n \geq 1 \,:\, \sum_{t=1}^n d_\psi(\langle \theta_t, X_t\rangle, \langle \theta^\star, X_t\rangle) \leq \frac{1}{2}\mathrm{regret}_{\theta^n}(\theta^\star) + \log\frac{1}{\delta}\right] \leq \delta\,.$$

Proof. Let us introduce the shorthand $D_t(\theta) = \frac{1}{2}\ell_t(\theta) + \frac{1}{2}\ell_t(\theta^\star) - \ell_t^{(1/2)}(\theta)$, and observe that $D_t(\theta) = d_\psi(\langle\theta, X_t\rangle, \langle\theta^\star, X_t\rangle)$. Then, we have

$$\sum_{t=1}^n D_t(\theta_t) = \frac{1}{2}\sum_{t=1}^n (\ell_t(\theta_t) - \ell_t(\theta^*)) + \sum_{t=1}^n \left(\ell_t(\theta^*) - \ell_t^{(1/2)}(\theta_t)\right)$$

The claim then follows from applying Proposition 2.1 with $\eta = 1/2$. □

Whereas the previously presented constructions use *either* the predictions of an online learning algorithm (Section 2.1) *or* a regret bound (Section 2.2), this one uses a bound on the regret *and* the predictions of an online learning algorithm. In return, the size of the confidence sets is determined by the regret against $\theta^\star$ rather than the regret against $\hat\theta_n$, which can be advantageous when, for instance, $\theta^\star$ is known to enjoy some additional structure (e.g., sparsity).

Supposing that the online algorithm comes with a regret bound of the form $\mathrm{regret}_{\theta^n}(\theta) \leq B_n(\theta)$, the result above implies that the following is a valid confidence sequence:

$$\Theta_n = \left\{\theta \in \mathbb{R}^d \,:\, \sum_{t=1}^n d_\psi(\langle\theta_t, X_t\rangle, \langle\theta, X_t\rangle) \leq \frac{1}{2}B_n(\theta) + \log\frac{1}{\delta}\right\}. \tag{9}$$

If a tight upper bound on $\mathrm{regret}_{\theta^n}(\theta^\star)$ is known, the comparator-specific $B_n(\theta)$ term can be replaced by said bound while maintaining the validity of the confidence sequence. Unfortunately, d_ψ is not always a convex function, which means that the sets in Theorem 2.4 are not always convex sets. If ψ is strongly convex, this can be exploited to construct convex confidence ellipsoids that contain the sets in Equation (9), though they may be substantially larger. We provide concrete examples (as well as a tighter confidence set of a similar shape) in Section 2.3.

3. Regret bounds

In order to derive meaningful confidence sets from the framework described in the previous section, we need to exhibit sequential prediction algorithms that come with guaranteed bounds on their regret. Recall that our main result (Theorem 2.2) only requires demonstrating the *existence* of such methods, without concern for computational aspects. This allows us to consider algorithms that are generally hard (or impossible) to implement efficiently. For this reason, we will focus on variants of two specific algorithms that come with tight regret bounds, but are not necessarily easily implementable: the exponentially weighted average (EWA) and normalized maximum likelihood (NML) forecasters. We formally introduce these methods in the sections below, as well as provide bounds on their regret, stated in rather general terms. Some results will concern the special case of *transductive online learning*, meaning that the sequence of covariates $X_1, X_2, \dots, X_n$ is chosen obliviously of the predictions made by the online learner, and that this sequence is known ahead of time. All bounds we provide in this section are guaranteed to hold with probability 1. We will instantiate these bounds in Section 4 below to provide some concrete confidence sets for GLMs. For further reading on sequential prediction with the log loss, we refer to the excellent book of Cesa-Bianchi and Lugosi (2006) (and particularly Chapter 9 therein).

3.1. The Exponentially Weighted Average Forecaster

One of the most fundamental algorithms for sequential probability assignment is the *exponentially weighted average* (EWA) forecaster, first proposed and studied by Vovk (1990) (with later developments by Littlestone and Warmuth, 1994; Freund and Schapire, 1997 and numerous further applications throughout all of online learning, cf. Cesa-Bianchi and Lugosi, 2006). This strategy is a mixture forecaster, which takes as input a *prior* q_1 over the parameter space Θ, and then produces each subsequent mixture for $t = 2, \dots, n$ according to the update rule

$$\frac{\mathrm{d}q_t}{\mathrm{d}q_1}(\theta) = \frac{e^{-\lambda \sum_{s=1}^{t-1} \ell_s(\theta)}}{\int e^{-\lambda \sum_{s=1}^{t-1} \ell_s} \mathrm{d}q_1}.$$

Here, the parameter $\lambda > 0$ can be interpreted as a *learning rate* (or *stepsize*). The EWA forecaster comes with guarantees on its *scaled log loss* with scale parameter λ defined for each mixture $q \in \Delta_\Theta$ as

$$\mathcal{L}_{t,\lambda}(q) = -\frac{1}{\lambda} \log \left(\int e^{-\lambda \ell_t(\theta)} \mathrm{d}q(\theta) \right),$$

with the scaled regret defined as $\mathrm{regret}_{q^n,\lambda}(\bar{\theta}) = \sum_{t=1}^n \left(\mathcal{L}_{t,\lambda}(q_t) - \ell_t(\bar{\theta}) \right)$. Indeed, through a standard telescoping sum argument (cf. Lemma B.1), one can show that for any comparator $\bar{\theta}$ and any $\lambda > 0$, the regret of the EWA strategy satisfies

$$\mathrm{regret}_{q^n,\lambda}(\bar{\theta}) = -\frac{1}{\lambda} \log \left(\int e^{-\lambda \sum_{t=1}^n \left(\ell_t(\theta) - \ell_t(\bar{\theta}) \right)} \mathrm{d}q_1(\theta) \right). \tag{10}$$

There are several ways in which one can turn the above generic bound into concrete, quantitative guarantees. A result we will repeatedly use provides an upper bound in terms of a quantity we call the *Bregman information gain*, defined in terms of the so-called Bregman divergence associated with the log-likelihood of the GLM we study. For a precise definition, let $\rho : \mathbb{R}^d \to \mathbb{R}$ be any convex and differentiable function such that $\int \exp(-\rho(\theta))\mathrm{d}\theta < \infty$ and define the regularized cumulative loss

$$Z^{\rho}_{n,\lambda}(\theta) = \lambda \sum_{t=1}^{n} \ell_t(\theta) + \rho(\theta)\,.$$

For any convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, define the Bregman divergence as

$$\mathcal{B}_f(\theta, \theta') = f(\theta) - f(\theta') - \langle \theta - \theta', \nabla f(\theta') \rangle\,,$$

which is the difference between $f(\theta)$ and the first-order Taylor approximation of f around θ'. Then, we define the Bregman information gain as

$$\gamma^{\rho}_{n,\lambda} = -\log\left(\frac{\int \exp(-\mathcal{B}_{Z^{\rho}_{n,\lambda}}(\theta, \widehat{\theta}_{n,\lambda}))\mathrm{d}\theta}{\int \exp(-\rho(\theta))\mathrm{d}\theta}\right)\,, \tag{11}$$

where $\widehat{\theta}_{n,\lambda} \in \arg\min_{\theta} Z^{\rho}_{n,\lambda}(\theta)$, assuming this exists (which will be the case in all applications we consider). The term "Bregman information gain" is borrowed from Chowdhury et al. (2023), who justify this naming convention by the observation that in the case of linear models, this quantity is equal to the mutual information between the function values $\langle \theta^{\star}, X_1 \rangle, \ldots, \langle \theta^{\star}, X_n \rangle$ and the labels Y^n, which is often thought of as a measure of "information gain" about the true parameter $\theta^{\star}$ given the observed samples in a Bayesian model.

The following proposition provides an upper bound on the regret of the EWA forecaster in terms of the Bregman information gain, when using the prior $q_1 \propto e^{-\rho}$.

Proposition 3.1. *For all comparators $\bar{\theta}$ such that $\rho(\bar{\theta}) < \infty$, the regret of the EWA forecaster with prior density $q_1 \propto e^{-\rho}$ satisfies*

$$\mathrm{regret}_{q^n,\lambda}(\bar{\theta}) \le \frac{\rho(\bar{\theta}) + \gamma^{\rho}_{n,\lambda}}{\lambda}\,.$$

Proof. Recalling the expression (10) of the regret of the EWA forecaster, we write

$$\begin{aligned}
\mathrm{regret}_{q^n,\lambda}(\bar{\theta}) &= -\frac{1}{\lambda}\log\left(\int \exp\left(-\lambda\sum_{t=1}^{n}\ell_t(\theta) + \lambda\sum_{t=1}^{n}\ell_t(\bar{\theta})\right)\mathrm{d}q_1(\theta)\right) \\
&= -\frac{1}{\lambda}\log\left(\frac{\int \exp\left(-Z^{\rho}_{n,\lambda}(\theta) + Z^{\rho}_{n,\lambda}(\bar{\theta})\right)\mathrm{d}\theta}{\int \exp\left(-\rho(\theta)\right)\mathrm{d}\theta}\right) + \frac{\rho(\bar{\theta})}{\lambda} \\
&\le -\frac{1}{\lambda}\log\left(\frac{\int \exp\left(-Z^{\rho}_{n,\lambda}(\theta) + Z^{\rho}_{n,\lambda}(\widehat{\theta}_{n,\lambda})\right)\mathrm{d}\theta}{\int \exp\left(-\rho(\theta)\right)\mathrm{d}\theta}\right) + \frac{\rho(\bar{\theta})}{\lambda}
\end{aligned}$$

$$= -\frac{1}{\lambda}\log\left(\frac{\int \exp\left(-\mathcal{B}_{Z^{\rho}_{n,\lambda}}(\theta, \widehat{\theta}_{n,\lambda})\right)\mathrm{d}\theta}{\int \exp\left(-\rho(\theta)\right)\mathrm{d}\theta}\right) + \frac{\rho(\bar{\theta})}{\lambda} = \frac{\gamma^{\rho}_{n,\lambda} + \rho(\bar{\theta})}{\lambda}\,,$$

where we used that $Z^{\rho}_{n,\lambda}(\theta) - Z^{\rho}_{n,\lambda}(\widehat{\theta}_{n,\lambda}) = \mathcal{B}_{Z^{\rho}_{n,\lambda}}(\theta, \widehat{\theta}_{n,\lambda})$, which holds since $\widehat{\theta}_{n,\lambda}$ minimizes $Z^{\rho}_{n,\lambda}$. □

Under mild conditions, the Bregman information gain (and hence this regret bound) grows asymptotically as $d\log n$ in the worst case (Grünwald, 2007). As our focus is on finite-sample bounds, we will derive more precise bounds in the forthcoming sections, under concrete assumptions about the likelihood, the parameter $\theta^{\star}$ and the sequence of covariates. Naturally, the best results are achieved when picking the regularizer ρ (and thus the prior q_1) in a way that utilizes the problem structure at hand effectively. For instance, ρ may depend on the smoothness of the log-likelihood function, the sequence of covariates $X_1, \dots, X_n$ (if these are known in advance) or even $\theta^{\star}$. Our applications in Section 4 will make use of a number of such problem-dependent choices.

We provide below an additional result: a variation of EWA using a subset-selection prior inspired by Alquier and Lounici (2011), which enjoys regret guarantees that are adaptive to the *sparsity* of the comparator, improving the worst-case dependence of order $d\log n$ to $s\log n$ in the special case where the comparator $\bar{\theta}$ is s-sparse. We define a distribution π over subsets $S \subseteq [d]$ by

$$\pi(S) = \frac{2^{-|S|}}{\binom{d}{|S|}\sum_{s=0}^{d} 2^{-s}}\,.$$

For each subset $S \subseteq [d]$, we let $\Theta_S = \{\theta \in \mathbb{R}^d : \mathrm{supp}(\theta) \subseteq S\}$ and define the probability measure $q_S \in \Delta_{\Theta}$, which has support contained in Θ_S and whose density w.r.t. the Lebesgue measure on Θ_S is $q_S(\theta) = \frac{\exp(-\rho(\theta))}{\int_{\Theta_S}\exp(-\rho(\theta))\mathrm{d}\theta}$. Finally, we construct our prior as $q_1 = \sum_{S\subseteq[d]} \pi(S) q_S$. Our regret bound for the EWA forecaster with this prior depends on a quantity we call the *restricted Bregman information gain*. For any subset $S \subseteq [d]$, we let $\widehat{\theta}_{n,\lambda,S} = \arg\min_{\theta\in\Theta_S}\{Z^{\rho}_{n,\lambda}(\theta)\}$. We then define the restricted Bregman information gain as

$$\gamma^{\rho,S}_{n,\lambda} = -\log\left(\frac{\int_{\Theta_S}\exp(-Z^{\rho}_{n,\lambda}(\theta) + Z^{\rho}_{n,\lambda}(\widehat{\theta}_{n,\lambda,S}))\mathrm{d}\theta}{\int_{\Theta_S}\exp(-\rho(\theta)\mathrm{d}\theta}\right)\,. \tag{12}$$

We refer the curious reader to Appendix A, where we show that $\gamma^{\rho,S}_{n,\lambda}$ can be equivalently defined in terms of a Bregman divergence associated with the log-likelihood (which not only justifies the notation and choice of name, but will also come handy when putting this result to use in Section 4.2). For any comparator $\bar{\theta}$, with support $\bar{S}$, Proposition 3.2 provides an upper bound on the regret of the EWA forecaster that depends on the restricted Bregman information gain $\gamma^{\rho,\bar{S}}_{n,\lambda}$.

Proposition 3.2. *Let $\bar{S} = \operatorname{supp}(\bar{\theta})$. For all comparators $\bar{\theta}$ such that $\rho(\bar{\theta}) < \infty$, the regret of the EWA forecaster with prior $q_1 = \sum_{S \subseteq [d]} \pi(S) q_S$ satisfies*

$$\operatorname{regret}_{q^n,\lambda}(\bar{\theta}) \leq \frac{1}{\lambda}\left(\gamma_{n,\lambda}^{\rho,\bar{S}} + \rho(\bar{\theta}) + |\bar{S}| \log(2ed/|\bar{S}|) + \log 2\right) .$$

To our knowledge, this concrete regret bound has not appeared in previous literature, although it bears a strong similarity with a previous result by Gerchinovitz (2013). Their method is also based on an application of the EWA forecaster, although with a different sparsity-inducing prior inspired by Dalalyan and Tsybakov (2008, 2012). We provide further comparisons between the two methods in Section 4.2.2 where we instantiate our regret bound in the context of confidence sets for sparse models.

Proof. The proof is almost the same as the proof of Proposition 3.1. Starting again from (10), we have

$$\begin{aligned}
\operatorname{regret}_{q^n,\lambda}(\bar{\theta}) &= -\frac{1}{\lambda} \log\left(\int \exp\left(-\lambda \sum_{t=1}^{n} \ell_t(\theta) + \lambda \sum_{t=1}^{n} \ell_t(\bar{\theta})\right) \mathrm{d}q_1(\theta)\right) \\
&= -\frac{1}{\lambda} \log\left(\sum_{S \subseteq [d]} \pi(S) \frac{\int_{\Theta_S} \exp\left(-Z_{n,\lambda}^{\rho}(\theta) + Z_{n,\lambda}^{\rho}(\bar{\theta}) - \rho(\bar{\theta})\right) \mathrm{d}\theta}{\int_{\Theta_S} \exp\left(-\rho(\theta)\right) \mathrm{d}\theta}\right) \\
&\leq -\frac{1}{\lambda} \log\left(\frac{\int_{\Theta_{\bar{S}}} \exp\left(-Z_{n,\lambda}^{\rho}(\theta) + Z_{n,\lambda}^{\rho}(\bar{\theta})\right) \mathrm{d}\theta}{\int_{\Theta_{\bar{S}}} \exp\left(-\rho(\theta)\right) \mathrm{d}\theta}\right) + \frac{\rho(\bar{\theta}) + \log \frac{1}{\pi(\bar{S})}}{\lambda} \\
&\leq \frac{\gamma_{n,\lambda}^{\rho,\bar{S}} + \rho(\bar{\theta}) + \log \frac{1}{\pi(\bar{S})}}{\lambda} ,
\end{aligned}$$

where the first inequality follows from the fact that for any non-negative mapping f we have $\sum_S \pi(S) f(S) \geq \pi(\bar{S}) f(\bar{S})$. Now we are just left with controlling $\log \frac{1}{\pi(\bar{S})}$. By the construction of π, we have $\frac{1}{\pi(\bar{S})} \leq \binom{d}{\bar{S}} 2^{1+|\bar{S}|}$, from which the claim follows by using that $\binom{d}{s} \leq (ed/s)^s$. □

3.2. Normalized Maximum Likelihood

The second family of methods we will consider is that of *Normalized Maximum Likelihood* (NML) forecasters, first proposed by Shtarkov (1987); Rissanen (1996) and later studied extensively in a long sequence of works including Barron, Rissanen and Yu (1998); Takeuchi and Barron (1997); Xie and Barron (2000); Cesa-Bianchi and Lugosi (2001); Liang and Barron (2006); Bartlett et al. (2013); Grünwald and Harremoës (2009); Grünwald and Mehta (2019); Jacquet, Shamir and Szpankowski (2022), and in particular the excellent book of Grünwald (2007). Rather than playing distributions on Θ in each round, NML forecasters work directly with distributions on $\mathcal{Y}$ in each round. We recall that for a generic

$p \in \Delta_{\mathcal{Y}}$, the log loss is defined as $\mathcal{L}_t(p) = -\log(p(Y_t))$, and the regret associated with a sequence $p^n = p_1, \dots, p_n$ as

$$\mathrm{regret}_{p^n}(\bar{\theta}) = \sum_{t=1}^{n} \mathcal{L}_t(p_t) - \sum_{t=1}^{n} -\log(p(Y_t|X_t, \bar{\theta}))\,.$$

The standard Normalized Maximum Likelihood (NML) forecaster (Shtarkov, 1987) is defined in terms of the joint distribution over sequences in $\mathcal{Y}^n$ with density defined as

$$P_n(y^n) = \frac{\sup_{\theta \in \Theta}\{\prod_{t=1}^{n} p(y_t|X_t, \theta)\}}{\int_{\mathcal{Y}^n} \sup_{\theta \in \Theta}\{\prod_{t=1}^{n} p(\widetilde{y}_t|X_t, \theta)\}\mathrm{d}\widetilde{y}^n}\,, \tag{13}$$

whenever the normalization constant can be guaranteed to be finite. Since this is typically not the case in the setting we consider[3], the standard definition can be modified in a number of ways that lead to well-defined probability distributions. Our of the many possibilities[4], and in line with the spirit of the EWA forecaster described previously, we address this here by introducing a "prior" (or sometimes called "luckiness function" in the NML literature) $\rho : \Theta \to \mathbb{R}$, and considering the following distribution over sequences in $\mathcal{Y}^n$:

$$P_n(y^n) = \frac{\sup_{\theta \in \Theta}\{\prod_{t=1}^{n} p(y_t|X_t, \theta)e^{-\rho(\theta)}\}}{\int_{\mathcal{Y}^n} \sup_{\theta \in \Theta}\{\prod_{t=1}^{n} p(\widetilde{y}_t|X_t, \theta)e^{-\rho(\theta)}\}\mathrm{d}\widetilde{y}^n}\,. \tag{14}$$

Having defined this joint distribution, the NML strategy consists in playing the conditional distributions extracted from P_n via the formula

$$p_t(y_t) = \frac{\int_{\mathcal{Y}^{n-t}} P_n(y^n)\mathrm{d}y_{t+1}\cdots\mathrm{d}y_n}{\int_{\mathcal{Y}^{n-t+1}} P_n(y^n)\mathrm{d}y_t\cdots\mathrm{d}y_n}\,.$$

Notice that each p_t depends on the entire sequence of covariates X^n, which means the NML forecaster as stated is only suitable for the transductive setting where all covariates are known ahead of time.

The denominator in the expression of Equation (14) plays a special role in the regret analysis of NML. Following the conventions in the literature, we call this quantity the *Shtarkov sum* (Grünwald, 2007), and denote it by

$$\mathcal{S}(\Theta, \rho, X^n) = \int_{\mathcal{Y}^n} \sup_{\theta \in \Theta}\left\{\prod_{t=1}^{n} p(y_t|X_t, \theta)e^{-\rho(\theta)}\right\}\mathrm{d}y^n. \tag{15}$$

Within the transductive setting, the NML forecaster can be shown to be the unique minimax optimal forecaster for sequential probability assignment, and

[3] For example, in the case of linear regression with $n = 1$, $p(y_1|X_1, \theta)$ is maximized by any θ satisfying $\langle \theta, X_t \rangle = y_t$; as a result, $\sup_\theta p(y_1|X_1, \theta) = (2\pi)^{-1/2}$ does not depend on y_1, and thus cannot be normalized.

[4] For alternative techniques leading to well-defined distributions, we refer to Chapter 11 of Grünwald, 2007, where this method is called "LNML-2". For simplicity, we will refer to this method as NML below.

for any sequence X^n, the minimax regret is equal to the logarithm of the Shtarkov sum (up to the regularization due to the prior ρ). Indeed, the worst-case regularized regret of the NML forecaster can be written as

$$\begin{aligned}\sup_{\theta\in\Theta}\{\mathrm{regret}_{p^n}(\theta)-\rho(\theta)\} &= \sum_{t=1}^{n}\mathcal{L}_t(p_t)+\sup_{\theta\in\Theta}\left\{\sum_{t=1}^{n}\log(p(Y_t|X_t,\theta))-\rho(\theta)\right\}\\ &= \log\left(\frac{1}{p_n(Y^n)}\right)+\log\left(\sup_{\theta\in\Theta}\left\{\prod_{t=1}^{n}p(Y_t|X_t,\theta)e^{-\rho(\theta)}\right\}\right)\\ &= \log(\mathcal{S}(\Theta,\rho,X^n))\,.\end{aligned}$$

We note that $\mathcal{S}(\Theta,\rho,X^n)$ does not depend on Y^n, which means that the NML forecaster is an *equalizer*: it achieves the same regret on every realization of the labels Y^n. Together with the the fact that any forecaster induces a probability distribution p_n on $\mathcal{Y}^n$ (and vice versa), this can be used to show that the NML forecaster is the unique minimax optimal forecaster in terms of the regularized regret (see, e.g., Section 9.4 in Cesa-Bianchi and Lugosi, 2006 and Section 11.3 in Grünwald, 2007). There exist variations of the standard NML forecaster that are able to deal with adaptively chosen covariates as well, including the Sequential Normalized Maximum Likelihood (SNML) forecaster (Roos and Rissanen, 2008; Kotłowski and Grünwald, 2011), and the contextual NML (cNML) forecaster of Liu, Attias and Roy (2024) which enjoys a minimax-optimality property similar to the one satisfied by NML.

While it is generally hard to evaluate the Shtarkov sum in concrete settings of interest, there are some important special cases where this can be done and the minimax regret be evaluated. The most prominent example falling into the scope of the present paper is that of linear models (i.e., the setting we consider with the choice $\psi : z \mapsto z^2/2$). In this setting, the sequence of predictions produced by NML as defined above with $\rho : \theta \mapsto \frac{1}{2\gamma^2}\|\theta\|_2^2$ coincide with those of the EWA forecaster with prior $q_1 \propto e^{-\rho}$ (see Kakade, Seeger and Foster, 2005 and Section 11.3 in Grünwald, 2007). This shows that EWA is minimax optimal in the setting where the covariates X^n are known ahead of time. Curiously, EWA requires no prior knowledge of the sequence of covariates (and in fact is also equivalent to SNML in this case), implying that there is no gap in difficulty of the fixed- and adaptive-design models in this setting. Accordingly, the regret bounds of EWA we will make use of below are all minimax optimal for linear models with fixed and obliviously chosen covariates. For several other settings, the Shtarkov sum is known to grow asymptotically as $\Theta\big(\frac{d}{2}\log n\big)$ (Jacquet, Shamir and Szpankowski, 2022). Since our aim in this paper is to derive explicit finite-sample guarantees, we will not instantiate these asymptotic results below. Finally, several of the bounds we provide will show explicit dependence on the sequence covariates, leading to rates that are potentially better than minimax in benign cases (e.g., when the sequence of covariates does not span the full space).

4. Applications

In this section, we instantiate our online-to-confidence-set technique in a number of specific cases of interest. Throughout the section, we will assume that the log-partition function of the GLM is M*-smooth* in the sense that ψ is twice-differentiable with its second derivative satisfying $\psi''(z) \leq M$ for some positive M and all $z \in \mathbb{R}$. Many GLMs of practical interest satisfy this condition, including the classic linear model $\psi(z) = \frac{z^2}{2}$ with $M = 1$ and the logistic model $\psi(z) = \log(1 + e^z)$ with $M = \frac{1}{4}$ [5]. We note that this condition is equivalent to assuming that the random variable $Y_t - \mathbb{E}\left[Y_t \middle| \mathcal{F}_{t-1}\right]$ is $\sqrt{M}$-subgaussian[6].

Additionally, some of the results below will also assume that ψ is m*-strongly convex* on an interval $[-b, b]$, meaning that $\psi''(z) \geq m$ holds uniformly for all $z \in [-b, b]$ and some $m \geq 0$ (and some $b > 0$). Obviously, $m \leq M$ holds for all GLMs, and the two are equal if and only if ψ is quadratic (i.e., the GLM is linear). We will sometimes call the ratio of the two constants the *condition number*[7] of the GLM and denote it by $\kappa = M/m$. We highlight that assuming strong convexity of ψ is generally a very strong assumption, and the constant m might often scale poorly with problem parameters such as the dimension d. For instance, for logistic regression, the strong convexity assumption only holds whenever Θ is compact and the covariates are all bounded, and even then m can be exponentially small with d. Thus, the guarantees stated below without assuming strong convexity are to be regarded as much less restrictive than the ones requiring this condition.

Most results we show below are derived from the analytic conversion scheme presented in Section 2.2 and the confidence sets have the shape given in Equation (8). For each of these applications, our technique yields the best known upper bounds on the width parameter β_n. Later results also make use of the algorithmic conversion scheme of Section 2.3, which also leads to improved confidence sets in comparison with previous work. We provide a detailed discussion of the relevant literature after stating each result, and point out the concrete improvements explicitly.

4.1. Analytic conversions

To apply the analytic conversion scheme suggested by Theorem 2.2, one needs to find a suitable reference point $\overline{\theta}_n$ and demonstrate the existence of an algorithm with low regret. In what follows, we consider a number of concrete settings, where we demonstrate useful choices of these hyperparameters and present the resulting confidence sets.

[5] For $\psi(z) = \log(1 + e^z)$ and $\mu(z) = 1/(1 + e^{-z}) \in [0, 1]$, $\psi''(z) = \mu(z)(1 - \mu(z)) \in [0, \frac{1}{4}]$.

[6] This follows from observing that the centred moment generating function satisfies $\mathbb{E}\left[\exp(\lambda(Y_t - \mathbb{E}[Y_t|\mathcal{F}_{t-1}]))\middle| \mathcal{F}_{t-1}\right] = \mathcal{B}_\psi(\gamma + \lambda \| \gamma) \leq M\lambda^2/2$.

[7] This is not to be confused with the condition number of the covariance matrix associated with the data, which does not appear in any of our bounds.

4.1.1. Adaptively chosen covariates

We first consider the most general version of our setting, where each covariate X_t is allowed to depend on the previous sequence of outcomes $(X_k, Y_k)_{k<t}$ in an arbitrary fashion. Allowing such dependences is extremely important in applications of high practical interest, for instance in sequential decision-making problems such as online learning in (generalized) linear bandits or Markov decision processes (Lattimore and Szepesvári, 2020). Our main result for this setting is the following theorem.

Theorem 4.1. *Suppose that ψ is M-smooth, and fix $\gamma > 0$. Set $\rho(\theta) = \frac{\|\theta\|_2^2}{2\gamma^2}$ and let $\overline{\theta}_n = \widehat{\theta}_n =$ arg min$_\theta \sum_{t=1}^n \ell_t(\theta) + \rho(\theta)$. Then, for any $\delta \in (0,1)$, the set defined in Equation (8) satisfies $\mathbb{P}\left[\exists n : \theta^\star \notin \Theta_n\right] \le \delta$ with the choice*

$$\beta_n = \frac{\|\widehat{\theta}_n\|_2^2}{2\gamma^2} + \frac{1}{2}\log\det\left(\gamma^2 M\Lambda_n + \mathrm{Id}\right) + \log\frac{1}{\delta}\,. \tag{16}$$

In particular, if all the covariates are bounded as $\|X_t\|_2 \le L$, we have

$$\beta_n \le \frac{\|\widehat{\theta}_n\|_2^2}{2\gamma^2} + \frac{d}{2}\log\left(1 + \frac{\gamma^2 M L^2 n}{d}\right) + \log\frac{1}{\delta}\,. \tag{17}$$

The second claim is a simple consequence of the first one, by the inequality of arithmetic and geometric means applied in the form

$$\det(\gamma^2 M\Lambda_n + \mathrm{Id}) \le \left(1 + \frac{\gamma^2 M n L^2}{d}\right)^d.$$

The main claim follows from instantiating the reduction scheme of Theorem 2.2 with the online learning algorithm choosen as the EWA forecaster described in Section 3.1. Specifically, the proof follows immediately from applying Proposition 3.1 along with the following upper bound on the Bregman information gain for smooth GLMs.

Lemma 4.2. *Suppose that ψ is M-smooth, fix $\gamma, \lambda > 0$ and set $\rho(\theta) = \frac{\|\theta\|_2^2}{2\gamma^2}$. Letting $\Lambda_n = \sum_{t=1}^n X_t X_t^\intercal$, the Bregman information gain satisfies*

$$\gamma_{n,\lambda}^\rho \le \frac{1}{2}\log\det\left(\lambda M\gamma^2\Lambda_n + \mathrm{Id}\right)\,.$$

Proof. The smoothness of ψ implies that $\mathrm{Hess}[Z_{n,\lambda}^\rho] \preceq \lambda M\Lambda_n + \frac{\mathrm{Id}}{\gamma^2}$, and so

$$\mathcal{B}_{Z_{n,\lambda}^\rho}(\theta,\theta') \le \frac{1}{2\gamma^2}\|\theta-\theta'\|_{\lambda\gamma^2 M\Lambda_n+\mathrm{Id}}^2\,.$$

Thus, the Bregman information gain can be upper-bounded in terms of a Gaussian integral, which can be evaluated via straightforward calculations to give

$$\gamma_{n,\lambda}^\rho \le -\log\left(\frac{\int\exp\left(-\frac{1}{2\gamma^2}\|\theta-\widehat{\theta}_{\lambda,n}\|_{\lambda\gamma^2 M\Lambda_n+\mathrm{Id}}^2\right)\mathrm{d}\theta}{\int\exp(-\frac{1}{2\gamma^2}\|\theta\|_2^2)\mathrm{d}\theta}\right) = \frac{1}{2}\log\det(\lambda M\gamma^2\Lambda_n+\mathrm{Id})\,.$$

This concludes the proof. □

Comparison with state of the art. To ease discussion of the above result, it is more practical to rewrite the confidence set in terms of the *regularized losses* $\widetilde{\ell}_t = \ell_t + \rho$ as follows:

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d : \sum_{t=1}^{n} \left(\widetilde{\ell}_t(\theta) - \widetilde{\ell}_t(\widehat{\theta}_n) \right) \leq \widetilde{\beta}_n \right\} . \tag{18}$$

With this notation, the main claim of Theorem 4.1 can be rewritten as a bound on the confidence-width parameter $\widetilde{\beta}_n$ as

$$\begin{aligned} \widetilde{\beta}_n &= \frac{\left\|\theta^\star\right\|_2^2}{2\lambda\gamma^2} + \frac{1}{2\lambda} \log\det\left(\gamma^2\lambda M \Lambda_n + \mathrm{Id}\right) + \log\frac{1}{\delta} \\ &\leq \frac{\left\|\theta^\star\right\|_2^2}{2\lambda\gamma^2} + \frac{d}{2\lambda} \log\left(1 + \frac{\gamma^2\lambda M L^2 n}{d}\right) + \log\frac{1}{\delta} . \end{aligned}$$

If it is known in advance that $\left\|\theta^\star\right\|_2 \leq B$, then the choice $\gamma = B$ yields

$$\widetilde{\beta}_n \leq \frac{1 + \log(\det(B^2\lambda M \Lambda_n + \mathrm{Id}))}{2\lambda} \leq \frac{1}{2\lambda}\left(1 + d\log\left(1 + \frac{B^2\lambda M L^2 n}{d}\right)\right) .$$

In the special case of linear models, this recovers the classic results of Abbasi-Yadkori, Pál and Szepesvári (2011) (see also de la Peña, Lai and Shao, 2009 and Flynn et al., 2023). For GLMs, the most directly comparable result is due to Lee, Yun and Jun (2024b), who work under the slightly more restrictive assumption that the parameter set is compact, and provide confidence sets with width as given in our Equation (17) (which is a looser version of the width bound guaranteed by our result, given in Equation (16)). All known previous bounds from the literature suffer from additional constant factors such as a uniform lower bound on the second derivative ψ'' (Jun et al., 2017; Li, Lu and Zhou, 2017; Emmenegger, Mutny and Krause, 2023). For the special case of parameter estimation of exponential-family distributions (i.e., GLMs with fixed covariates), our bound recovers the confidence sequence proposed by Chowdhury et al. (2023), with width dependent on the Bregman information gain (which name was in fact coined in said work).

To the best of our knowledge, the data-dependent regret bound implied by Lemma 4.2 does not appear explicitly in any prior work on EWA. In the worst case, it recovers a classic regret bound of Kakade and Ng (2004) for the EWA forecaster under the same boundedness and smoothness assumptions and with the same prior. However, our data-dependent bound can be much smaller in practice, especially if L is a loose upper bound on the norm of the largest covariate. In addition, the data-dependent regret bound is adaptive to certain easy sequences of covariates. For instance, the data-dependent regret bound can be used to obtain

$$\mathrm{regret}_{q^n}(\bar{\theta}) \leq \frac{\|\bar{\theta}\|_2^2}{2\lambda\gamma^2} + \frac{\mathrm{rank}(\Lambda_n)}{2\lambda} \log\left(1 + \frac{\gamma^2 M L^2 n}{\mathrm{rank}(\Lambda_n)}\right) . \tag{19}$$

Evidently, this improvement is inherited by our confidence width parameter as well, demonstrating a clear improvement over the best previously known results of Lee, Yun and Jun (2024b). This fact follows from a small modification of the determinant-trace inequality in Lemma 10 of Abbasi-Yadkori, Pál and Szepesvári (2011), which accounts for the fact that Λ_n may not have full rank (cf. Lemma B.2).

4.1.2. Obliviously chosen covariates

The guarantees provided in the previous section can be tightened by making a stronger assumption about the sequence of covariates: that each X_t is chosen independently of the realized labels Y_k (for all $k \neq t$). Some well-studied special cases are the *fixed-design* setting where the sequence of covariates is arbitrary but fixed before the labels Y_t are drawn, and the *i.i.d.* setting where each X_t is independently from some fixed distribution, independently of all labels and the other covariates. More generally, the set of covariates can be drawn from any joint distribution as long as it is independent of the realized labels.

The lack of dependence between covariates and labels allows us to use online learning algorithms that have prior access to the sequence of covariates—a setting that is thoroughly studied in the literature under the name *transductive online learning* or *sequential prediction with transductive priors*. Since this setting only makes sense for sequences of fixed length, the results we prove below will naturally hold for a fixed sample size n. Without loss of generality[8], we will assume that the matrix $\Lambda_n = \sum_{t=1}^n X_t X_t^\intercal$ is full rank.

The result we will state below will assume that ψ is globally M-smooth on $\mathbb{R}$ and locally m-strongly convex on $[-b, b]$, and furthermore we will suppose that $|\langle \theta^\star, X_t \rangle| \leq b$ holds. For the given sequence of covariates $X_1, \dots, X_n$, we will define the associated *polar set* (at scale b) as $\mathcal{S}_{n,b} = \{\theta \in \mathbb{R}^d : \max_{t\in[n]} |\langle \theta, X_t \rangle| \leq b\}$, and note that this is a convex set that is guaranteed to include $\theta^\star$. For the comparator in the regret analysis, we will use the constrained MLE $\widehat{\theta}_{n,b} = \arg\min_{\theta \in \mathcal{S}_{n,b}} \sum_{t=1}^n \ell_t(\theta)$. The following theorem is our main result about smooth and strongly convex GLMs in this setting, which follows from instantiating Theorem 2.2 with a transductive online algorithm.

Theorem 4.3. *Suppose that $\theta^\star$ satisfies $|\langle \theta^\star, X_t \rangle| \leq b$ and that that ψ is M-smooth on $\mathbb{R}$ and m strongly convex on $[-b, b]$, and denote the condition number by $\kappa = \frac{M}{m}$. Let $\widehat{\theta}_{n,b} = \arg\min_{\theta \in \mathcal{S}_{n,b}} \sum_{t=1}^n \ell_t(\theta)$, define $\Psi : \Theta \to \mathbb{R}$ as $\Psi(\theta) = \sum_{t=1}^n \psi\big(\langle X_t, \theta \rangle\big)$ for all θ, and let $\mathcal{B}_\Psi$ denote the associated Bregman divergence. Then, for any $\delta > 0$, the set defined as*

$$\Theta_n = \left\{ \theta : \mathcal{B}_\Psi\big(\theta, \widehat{\theta}_{b,n}\big) \leq d \log\left(1 + 2\kappa\right) + 2 \log \frac{1}{\delta} \right\} \tag{20}$$

satisfies $\mathbb{P}\left[\theta^\star \notin \Theta_n\right] \leq \delta$.

[8] If this assumption does not hold without preprocessing, we can work in the subspace spanned by the covariates and aim to estimate the projection of $\theta^\star$ to said space. Obviously, estimating the orthogonal component is impossible in such a situation.

Notably, the width of the confidence set above does not show *any* dependence on the sample size n, and in particular it removes the $\log n$ factor that appeared in the previously stated guarantees for adaptively chosen covariates. Also, the bound is completely independent of the realization of the covariate sequence, and in particular is invariant to linear transformations of the coordinates of $\theta^\star$ and X^n. We contextualize this improvement in more detail below, after presenting the proof. At a high level, the construction for the proof involves executing EWA with a prior that takes into account the complete sequence of covariates, the true parameter $\theta^\star$, and the log-partition function ψ.

Proof. The proof is based on instantiating the regret bound of EWA executed with $\rho(\theta) = \frac{1}{\gamma^2}\left\|\theta - \theta^\star\right\|^2_{\Lambda_n}$. A crucial observation is that, thanks to the strong convexity of ψ on $[-b, b]$, the function Ψ is strongly convex on $\mathcal{S}_{n,b}$, with respect to the weighted norm $\left\|\cdot\right\|_{\Lambda_n}$, in the following sense:

$$\forall \theta, \theta' \in \mathcal{S}_{n,b},\ \frac{m}{2}\|\theta - \theta'\|^2_{\Lambda_n} \leq \mathcal{B}_\Psi(\theta, \theta')\,. \tag{21}$$

We define $Z_{n,\gamma}(\theta) = \sum_{t=1}^n \ell_t(\theta) + \frac{1}{2\gamma^2}\|\theta - \theta^\star\|^2_{\Lambda_n}$ and $\widehat{\theta}_{n,\gamma} = \arg\min_{\theta\in\mathbb{R}^d} Z_{n,\gamma}(\theta)$. Due to the smoothness of ψ on $\mathbb{R}$, the Bregman divergence induced by $Z_{n,\gamma}$ can be upper bounded by a quadratic function of θ. In particular, for any $\theta, \theta' \in \mathbb{R}^d$,

$$\mathcal{B}_{Z_{n,\gamma}}(\theta, \theta') = \mathcal{B}_\Psi(\theta, \theta') + \frac{1}{2\gamma^2}\|\theta - \theta'\|^2_{\Lambda_n} \leq \frac{1}{2}(M + 1/\gamma^2)\|\theta - \theta'\|^2_{\Lambda_n}\,.$$

Now, using the generic EWA regret bound in Proposition 3.1 (with $\lambda = 1$ and the prior described above) and applying the upper bound given above, we get

$$\begin{aligned}
\operatorname{regret}_{q^n}(\widehat{\theta}_n) &\leq -\log\left(\frac{\int \exp(-\mathcal{B}_{Z_{n,\gamma}}(\theta, \widehat{\theta}_{n,\gamma}))\mathrm{d}\theta}{\int \exp(-\frac{1}{2\gamma^2}\|\theta - \theta^\star\|^2_{\Lambda_n})\mathrm{d}\theta}\right) + \frac{1}{2\gamma^2}\|\widehat{\theta}_n - \theta^\star\|^2_{\Lambda_n} \\
&\leq -\log\left(\frac{\int \exp(-\frac{1}{2}(M + \frac{1}{\gamma^2})\|\theta - \widehat{\theta}_{n,\gamma}\|^2_{\Lambda_n})\mathrm{d}\theta}{\int \exp(-\frac{1}{2\gamma^2}\|\theta - \theta^\star\|^2_{\Lambda_n})\mathrm{d}\theta}\right) + \frac{1}{m\gamma^2}\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_n) \\
&= \frac{d}{2}\log(1 + \gamma^2 M) + \frac{1}{m\gamma^2}\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_n)\,,
\end{aligned}$$

where the last step follows from evaluating the Gaussian integral appearing in the second line. Along with Theorem 2.2, this implies that

$$\sum_{t=1}^n \Big(\ell_t(\theta^\star) - \ell_t(\widehat{\theta}_{n,b})\Big) \leq \frac{d}{2}\log(1 + \gamma^2 M) + \frac{1}{m\gamma^2}\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) + \log\frac{1}{\delta} \tag{22}$$

holds with probability at least $1 - \delta$. To proceed from here, notice that by the first-order optimality condition on the set $\Theta_{n,b}$, we have that $\widehat{\theta}_{n,b}$ satisfies $\langle \theta - \widehat{\theta}_{n,b}, \sum_{t=1}^n \nabla \ell_t(\widehat{\theta}_{n,b})\rangle \geq 0$ for all $\theta \in \mathcal{S}_{n,b}$, and in particular for $\theta^\star$ too. Therefore,

$$\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) \leq \sum_{t=1}^n \Big(\ell_t(\theta^\star) - \ell_t(\widehat{\theta}_{n,b})\Big)\,.$$

Combining this with the bound of Equation (22), we get

$$\left(1 - \frac{1}{m\gamma^2}\right) \mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) \leq \frac{d}{2}\log(1+\gamma^2 M) + \log\frac{1}{\delta}\,,$$

Picking $\gamma^2 = 2/m$ gives

$$\mathcal{B}_\Psi(\theta^\star, \widehat{\theta}_{n,b}) \leq d\log(1+2\kappa) + 2\log(1/\delta)\,,$$

which concludes the proof. □

Comparison with state of the art. We start by noting that the confidence set given in Theorem 4.3 is generally not equal to the set defined in Equation (8), due to the restriction of $\widehat{\theta}_{n,b}$ to the polar set $\mathcal{S}_{n,b}$. Note however that the two sets coincide if the MLE lies within $\mathcal{S}_{n,b}$, which can be verified empirically when evaluating the confidence set. For the special case of linear regression, b can be clearly set as $+\infty$ and the condition number becomes $\kappa = 1$. In this case, our result recovers a classic result of Cochran (1934) with a slightly worse constant. For other GLMs, we are not aware of any comparable result, and in particular we believe that our confidence set is the first to have a width independent of n. Furthermore, our construction property correctly accounts for the parametrization-invariance of GLMs: when transforming all covariates by the invertible linear map A and the true parameter $\theta^\star$ by A^{-1}, the distribution of labels remains unchanged. Since the polar set $\mathcal{S}_{n,b}$ used in our construction retains this invariance property, our confidence set also remains invariant to such reparametrizations. Accordingly, the scale of the covariates does not appear in the confidence width either. This phenomenon has been thoroughly studied in the context of transductive online learning, where several works have shown that knowing the covariates ahead of time can enable proving scale-invariant regret bounds (Gaillard et al., 2019; Qian, Rakhlin and Zhivotovskiy, 2024). Additionally, these factors have been shown to be impossible to remove without prior knowledge of the covariate sequence (Kotłowski, 2020; Foster et al., 2018; Qian, Rakhlin and Zhivotovskiy, 2024).

4.2. Algorithmic conversions

We now turn to deriving confidence sequences based on the algorithmic conversion scheme proposed in Theorem 2.4. We recall that this conversion scheme produces confidence sets whose size depends on the regret of an online algorithm against the true parameter $\theta^\star$. This allows us to derive tighter confidence sets when $\theta^\star$ has additional structure. To demonstrate this, the focus of this section is mainly on results for *sparse* GLMs, where $\theta^\star$ contains mostly zeros. In particular, we assume that $\theta^\star$ is s-sparse for some known s (i.e., that $\sum_{i=1}^d \mathbb{I}\{\theta_i \neq 0\} = s$). Throughout this section, we focus solely on the setting where each covariate X_t can depend on the previous observations $(X_k, Y_k)_{k=1}^{t-1}$ in an arbitrary fashion.

As mentioned in Section 2.3, the function d_ψ that appears in the algorithmic conversion of Theorem 2.4 is generally not convex, which means that the resulting confidence sets may not be convex either. To address this issue, we assume below that the log-partition function ψ is m-strongly convex on the interval $[-b, b]$, which implies that $d_\psi(z, z') \geq \frac{m(z-z')^2}{8}$ for all $z, z' \in [-b, b]$. In particular, this will allow us to derive convex confidence sets that can be stated conveniently in terms of a well-chosen set of *pseudo-labels*. Indeed, defining the truncation operator $[z]_b = \max\{\min\{z, b\} - b\}$ and the pseudo-label $\widehat{Y}_t = [\langle \theta_t, X_t \rangle]_b$, the confidence set of Theorem 2.4 can be shown to be included in the convex set

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d \ : \ \frac{1}{2} \sum_{t=1}^{n} \left(\langle \theta, X_t \rangle - \widehat{Y}_t \right)^2 \leq \beta_n \right\} \tag{23}$$

for some appropriately chosen β_n.

To emphasize the improvement of our new algorithmic conversions upon similar existing results, which use deterministic forecasters, we first focus on the generic algorithmic conversion scheme for deterministic forecasters in Theorem 2.4. This alone allows us to recover tighter (by constant factors) algorithmic conversions than those in Abbasi-Yadkori, Pál and Szepesvári (2012) and Jun et al. (2017). However, since all of these conversions use deterministic forecasters, the resulting confidence sets have sub-optimal confidence width by at least a factor of $\log(n)$. In Section 4.2.2 we address this issue with a tighter algorithmic conversion that is tailored to the EWA forecaster.

4.2.1. Algorithmic Conversions for Deterministic Forecasters

We start with a straightforward application of Theorem 2.4 for the case of strongly convex log-likelihoods.

Theorem 4.4. *Suppose that $\theta^\star$ satisfies $|\langle \theta^\star, X_t \rangle| \leq b$ and that ψ is m-strongly convex on $[-b, b]$. Then, for any $\delta \in (0, 1)$, the set defined in Equation (23) with $\widehat{Y}_t = [\langle \theta_t, X_t \rangle]_b$ satisfies $\mathbb{P}[\exists n : \theta^\star \notin \Theta_n] \leq \delta$ with the choice*

$$\beta_n = \frac{2}{m} \operatorname{regret}_{\theta^n}(\theta^\star) + \frac{4}{m} \log \frac{1}{\delta} .$$

Proof. Due to Lemma B.4, $d_\psi([\langle \theta_t, X_t \rangle]_b, \langle \theta^\star, X_t \rangle) \leq d_\psi(\langle \theta_t, X_t \rangle, \langle \theta^\star, X_t \rangle)$. The claim then follows by combining Theorem 2.4 and the discussion above. □

If we apply any of these confidence set to sparse linear models, where $m = 1$, and use the algorithm of Gerchinovitz (2013) to generate the sequence $\theta_1, \ldots, \theta_n$, then we obtain confidence sets of the form

$$\Theta_n = \left\{ \theta \in \mathbb{R}^d \ : \ \sum_{t=1}^{n} \widehat{\ell}_t(\theta) = \mathcal{O}\left(\max_t Y_t^2 s \log(\tfrac{dn}{s}) + \log \tfrac{1}{\delta} \right) \right\} . \tag{24}$$

Here, the big-O notation hides large numerical constants and logarithmic factors of problem parameters such as $\|\theta^\star\|$ and $\sup_t \|X_t\|$. The bound we will state in

Theorem 4.5 will spell all such dependencies out, and make strict improvements over the above guarantee. Since the above result uses a deterministic forecaster, the bound features a factor of $\max_t Y_t^2$, which may generally contribute an extra factor of $b + M \log n$ to the confidence width (for $\sqrt{M}$-sub-Gaussian noise). We will address this limitation in the next section.

Comparison with state of the art. In the case of linear models, our result is directly comparable with (and in fact inspired by) the result of Theorem 1 in Abbasi-Yadkori, Pál and Szepesvári (2012), which shows that the confidence set defined in Equation (23) is valid with the choice

$$\beta_n = \frac{1}{2} + 2\,\mathrm{regret}_{\theta^n}(\theta^\star) + 16\log\left(\frac{\sqrt{8} + \sqrt{1 + 2\,\mathrm{regret}_{\theta^n}(\theta^\star)}}{\delta}\right).$$

More generally, Theorem 1 by Jun et al. (2017) provides a comparable result for M-smooth and m-strongly convex GLMs, showing a confidence width of

$$\beta_n = \frac{1}{2} + \frac{2}{m}\,\mathrm{regret}_{\theta^n}(\theta^\star) + \frac{4M}{m^2}\log\left(\frac{1}{\delta}\sqrt{4 + \frac{8}{m}\,\mathrm{regret}_{\theta^n}(\theta^\star) + \frac{64M^2}{4m^4\delta^2}}\right).$$

In both cases, our confidence set has a much simpler expression, tighter numerical constants, and an improved dependence on m in the case of GLMs. In both cases, the improvement is due to our use of Proposition 2.3 instead of the self-normalized concentration inequality developed by Abbasi-Yadkori, Pál and Szepesvári (2012).

4.2.2. Algorithmic Conversions for the EWA Forecaster

We now provide an improved result that makes use of randomized forecasters, which will allow us to tighten the bounds above by removing the unnecessary $\max_t Y_t^2$ factor from the bound. Our confidence sets will take the same form as before, except that we will use pseudo-labels generated by the EWA forecaster with learning rate $\lambda = 1/2$, defined as follows:

$$\widehat{Y}_t = \int [\langle\theta, X_t\rangle]_b \, \mathrm{d}q_{t+1}(\theta)\,. \tag{25}$$

The following theorem is an improvement of Theorem 2.4, which is specialized to the EWA forecaster.

Theorem 4.5. *Let $q_1, q_2, \ldots$ be the sequence of distributions played by the EWA forecaster with learning rate $\lambda = 1/2$. Then,*

$$\mathbb{P}\left[\exists n \,:\, \sum_{t=1}^{n} \log \int e^{d_\psi(\langle\theta, X_t\rangle, \langle\theta^\star, X_t\rangle)} \mathrm{d}q_{t+1}(\theta) \geq \frac{1}{2}\,\mathrm{regret}_{q^n,1/2}(\theta^\star) + \log\frac{1}{\delta}\right] \leq \delta\,.$$

Proof. As in the proof of Theorem 2.4, we use the shorthand $D_t(\theta) = \frac{1}{2}\ell_t(\theta) + \frac{1}{2}\ell_t(\theta^\star) - \ell_t^{(1/2)}(\theta)$, and observe that $D_t(\theta) = d_\psi(\langle \theta, X_t\rangle, \langle \theta^\star, X_t\rangle)$. From the definition of the EWA forecaster, $\frac{\mathrm{d}q_{t+1}}{\mathrm{d}q_t}(\theta) \propto e^{-\frac{1}{2}\ell_t(\theta)}$. Therefore, we have

$$\int e^{D_t(\theta)}\mathrm{d}q_{t+1}(\theta) = \frac{\int e^{-\ell_t^{(1/2)}(\theta)}\mathrm{d}q_t(\theta)}{\int e^{-\frac{1}{2}\ell_t(\theta)}\mathrm{d}q_t(\theta)} e^{\frac{1}{2}\ell_t(\theta^\star)}\,.$$

Taking logarithms and rearranging terms, we obtain

$$\sum_{t=1}^n \log \int e^{D_t(\theta)}\mathrm{d}q_{t+1}(\theta) = \frac{1}{2}\,\mathrm{regret}_{q^n,1/2}(\theta^\star) + \sum_{t=1}^n \ell_t(\theta^\star) - \sum_{t=1}^n \mathcal{L}_t^{(1/2)}(q_t)\,.$$

The statement then follows from applying Proposition 2.3. □

Using this result, we can prove the following improved version of Theorem 4.4, which transforms the pseudo-labels generated by the EWA forecaster and the regret of the EWA forecaster against $\theta^\star$ into a confidence sequence.

Theorem 4.6. *Suppose that $\theta^\star$ satisfies $|\langle \theta^\star, X_t\rangle| \leq b$ and that ψ is m-strongly convex on $[-b, b]$. Let $q_1, q_2, \ldots$ are the distributions played by the EWA forecaster with learning rate $\lambda = 1/2$. Then, for any $\delta \in (0, 1)$, the set defined in Equation (23) with $\widehat{Y}_t$ defined in Equation (25) satisfies $\mathbb{P}[\exists n : \theta^\star \notin \Theta_n] \leq \delta$ with the choice*

$$\beta_n = \frac{2}{m}\,\mathrm{regret}_{q^n,1/2}(\theta^\star) + \frac{4}{m}\log\frac{1}{\delta}\,.$$

Proof. Using Jensen's inequality, the strong convexity of ψ, Lemma B.4 and then Jensen's inequality again, we obtain

$$\begin{aligned}
\frac{1}{2}\left(\langle \theta^\star, X_t\rangle - \widehat{Y}_t\right)^2 &\leq \frac{1}{2}\int ([\langle \theta, X_t\rangle]_b - \langle \theta^\star, X_t\rangle)^2 \mathrm{d}q_{t+1}(\theta) \\
&\leq \frac{4}{m}\int d_\psi([\langle \theta, X_t\rangle]_b, \langle \theta^\star, X_t\rangle)\mathrm{d}q_{t+1}(\theta) \\
&\leq \frac{4}{m}\int d_\psi(\langle \theta, X_t\rangle, \langle \theta^\star, X_t\rangle)\mathrm{d}q_{t+1}(\theta) \\
&\leq \frac{4}{m}\log\int e^{d_\psi(\langle \theta, X_t\rangle, \langle \theta^\star, X_t\rangle)}\mathrm{d}q_{t+1}(\theta)\,.
\end{aligned}$$

The claim then follows by summing both sides over t from 1 to n, and then applying Theorem 4.5. □

Notably, the bound now depends on the regret of the EWA forecaster on the sequential probability assignment game with the *logarithmic loss*, which allows us to remove the spurious $\max_t Y_t^2$ factor from the previous bound (that has resulted from using online algorithms with deterministic predictions).

Theorem 4.7. *Suppose that ψ is M-smooth on $\mathbb{R}$ and m-strongly convex on $[-b, b]$. In addition, suppose that $\theta^\star$ is s-sparse, $\|\theta^\star\|_2 \leq B$ and $\max_{t\geq 1} \|X_t\|_\infty \leq L_\infty$. Let $q_1, q_2, \ldots$ be the sequence of distributions played by the EWA forecaster with $\lambda = 1/2$ and the prior $q_1 = \sum_{S\subseteq[d]} \pi(S)q_S$, where $\rho(\theta) = \|\theta\|_2^2/(2B^2)$. Then the sequence of sets $(\Theta_t)_{t\geq 1}$, with Θ_n defined in Equation (23) and $\widehat{Y}_t$ defined in Equation (25), is a $1-\delta$ confidence sequence with the choice*

$$\beta_n = \frac{4s}{m} \log \frac{2ed\sqrt{1 + MB^2L_\infty^2 n/2}}{s} + \frac{4}{m} \log \frac{2\sqrt{e}}{\delta}\,.$$

Proof. Let $S^\star = \operatorname{supp}(\theta^\star)$ denote the support of $\theta^\star$. For any vector $x \in \mathbb{R}^d$, we define $x(S^\star) \in \mathbb{R}^s$ to be the s-dimensional subvector indexed by the elements of $S^\star$ (cf. Appendix A). Due to Theorem 4.6, we only need to bound $\operatorname{regret}_{q^n,1/2}(\theta^\star)$. We thus instantiate the regret bound of Proposition 3.2 with the choice $\rho(\theta) = \frac{\|\theta\|_2^2}{2B^2}$, which yields

$$\operatorname{regret}_{q^n,1/2}(\theta^\star) \leq 2\left(\gamma_{n,1/2}^{\rho,S^\star} + \frac{\|\theta^\star\|_2^2}{2B^2} + s\log\frac{2ed}{s} + \log 2\right)\,. \tag{26}$$

By assumption, $\|\theta^\star\|_2^2/B^2 \leq 1$. Using Lemma 4.2, the restricted Bregman information gain is bounded as

$$\gamma_{n,1/2}^{\rho,S^\star} \leq \frac{1}{2}\log\det(\tfrac{MB^2}{2}\Lambda_n(S^\star) + \mathrm{Id}_s)\,,$$

where $\Lambda_n(S^\star) = \sum_{t=1}^n X_t(S^\star)X_t(S^\star)^\top$ and Id_s is the $s \times s$ identity matrix. Since, $\|X_t(S^\star)\|_2 \leq \sqrt{s}L_\infty$, Lemma B.2 tells us that

$$\frac{1}{2}\log\det(\tfrac{MB^2}{2}\Lambda_n(S^\star) + \mathrm{Id}_s) \leq \frac{s}{2}\log(1 + MB^2L_\infty^2 n/2)\,.$$

Combining this upper bound on the restricted Bregman information gain with the inequality in Equation (26), we see that

$$\operatorname{regret}_{q^n,1/2}(\theta^\star) \leq 2s\log\frac{2ed\sqrt{1 + MB^2L_\infty^2 n/2}}{s} + 2\log(2\sqrt{e})\,.$$

The statement then follows by substituting this regret bound into the expression for β_n in Theorem 4.6. □

Comparison with state of the art. As promised, the bound in Theorem 4.7 removes the dependence on the maximum label that has appeared in previous bounds like Equation (24). This is a clear improvement over the best previous results by Abbasi-Yadkori, Pál and Szepesvári (2012) and Jun et al. (2017) that both suffer from this factor. Additionally, the bound given above comes with small, explicit constants and a much simpler overall expression. These latter improvements come from using the regret bound of Proposition 3.2 for EWA with a sparsity-inducing prior, instead of relying on the algorithm of Gerchinovitz (2013) whose regret bounds are much more complex to state. This complexity largely comes from their algorithm being completely parameter-free in the sense that it does not require any prior knowledge about the sparsity parameter s or the norm of $\theta^\star$. In our setting, knowledge of these parameters is necessary either way to evaluate the confidence width, which allows us to forgo this otherwise desirable parameter-free property.

5. Discussion

We have introduced a framework that establishes a link between statistical inference and sequential prediction by providing a reduction scheme that allows constructing confidence sets for a broad class of statistical models. This work fits into a line of work on algorithmic statistics initiated by Rakhlin and Sridharan (2017), which bridges statistics and the theory of algorithms by establishing similar reductions between the problems in either of the two areas. We have demonstrated the effectiveness of this framework in yet another context, and applied our framework to recover and improve state-of-the-art results for statistical inference in generalized linear models, as well as provide completely new confidence sets for these models. Besides these concrete results, our framework enables further progress in the area by streamlining the process of proving new concentration inequalities, effectively reducing the otherwise complex task of proving concentration inequalities to the relatively simpler task of regret analysis of online algorithms. Accordingly, any new result in this actively researched area can be immediately turned into new concentration bounds. In this section, we close by discussing some further aspects our framework and results, as well as discuss questions that we leave open for future research.

Extensions. For sake of concreteness and simplicity, we have focused on the class of generalized linear models with finite-dimensional parameters. Our framework can be extended in several straightforward ways. First, we mention that all of our results can be shown to also hold for “sub-exponential families” whose moment-generating functions satisfy the inequality $\mathbb{E}\left[\exp(\beta Y_t)|\mathcal{F}_{t-1}\right] \leq \exp(\psi(\beta + \langle\theta^\star, X_t\rangle) - \psi(\langle\theta^\star, X_t\rangle))$ for some ψ. Indeed, it is easy to see that the conclusion of Proposition 2.1 continues to hold under these relaxed conditions, and thus all subsequent results can be generalized in this way. In particular, all results proved for linear models can be also shown to hold under sub-Gaussian noise. Second, we mention that our results can be directly extended to infinite-dimensional models where the linear function $\langle\theta^\star, \cdot\rangle$ is replaced by some $f^\star$ that belongs to a reproducing kernel Hilbert space (see, e.g., Abbasi-Yadkori, 2012, Chapter 3, Emmenegger, Mutny and Krause, 2023 or Flynn and Reeb, 2024). We opted to not include results for this case to keep the paper less technical and easier to read, and leave working out the (potentially nontrivial) details for future work.

The tightness of our confidence sets. For all cases we have studied in Section 4, our results either recover the best known guarantees or make improvements over them. In some important special cases, these results also match the best achievable bounds: for linear models, the bounds under adaptively chosen covariates (Section 4.1.1) match the lower bound of Lattimore (2023), and the bounds for obliviously chosen covariates (Section 4.1.2) can only be improved in terms of constant factors (see, e.g., Example 15.14 in Wainwright, 2019). This suggests that our reduction technique does not introduce any gaps between the best achievable regret and the best achievable concentration inequalities (which is consistent with the general equivalence results between the two types of bounds proved by Rakhlin and Sridharan, 2017). As for the tightness of the confidence sets and regret bounds we provide in this paper, we believe that several improvements should be possible. In particular, under certain regularity conditions (such as boundedness of the covariates and the true parameter $\theta^\star$) the minimax regret for sequential probability assignment with GLMs is known to scale asymptotically as $\frac{d \log n}{2}$ (Grünwald, 2007). Finite-sample guarantees that match these rates are quite rare in the literature. One notable exception beyond linear models is the work of Jacquet, Shamir and Szpankowski (2022) that provides a finite-sample upper bound on the regret of the NML forecaster for logistic regression with bounded covariates, which matches the minimax rate up to some explicitly given additional terms that vanish as n grows large. In comparison to these bounds, the ones we provide are loose by a factor of the strong convexity constant $1/m$, which is often very large in GLMs of practical interest (e.g., as noted earlier, it can be exponentially large in d for the important case of logistic regression). It remains unclear if further improvements are possible at the level of generality that we have considered in this work, and we highlight this question as the most exciting one for future research.

Connections with mean estimation. It is insightful to instantiate our confidence sets in the special case of (one-dimensional) mean estimation, recovered by setting $X_t = 1$ for all t. In case the noise is sub-Gaussian, instantiating our results for obliviously chosen covariates (Theorem 4.3) correctly recovers Hoeffding's inequality without any additional $\log n$ factors (up to a small difference in constants). The case of observations that are almost surely bounded in $[0, 1]$ can be handled by noticing that such random variables are "sub-Bernoulli" in the sense defined in the paragraph above, and thus a reduction to logistic regression ($\psi : z \mapsto \log(1 + e^z)$) can be used. In this case, the shape of our confidence bounds matches the shape implied by Orabona and Jun (2024), which has been recently shown to be optimal in an appropriate sense (Clerico, 2024). Notably however, the width of our confidence interval in this case is only guaranteed to be meaningfully bounded if the range of the parameter $p = \mathbb{E}\left[Y_t\right]$ is constrained to lie strictly within the limits of the unit interval. Concretely, if it is known that $p \in [p_0, (1 - p_0)]$ holds for some known $p_0 > 0$, then the loss function ℓ_t can be shown to be strongly convex with parameter $m = \frac{1}{p_0(1-p_0)}$, and applying Theorem 4.3 gives a confidence width of order $\log\left(\frac{1}{p_0(1-p_0)}\right)$. This recovers the confidence interval of Orabona and Jun (2024) whenever $p_0 = \Omega\left(\frac{1}{n}\right)$, which we consider to be a reasonable regime. More generally, our bounds recover the results of Chowdhury et al. (2023) in the case of one-dimensional exponential families (as can be seen by directly combining our Theorem 2.2 with Proposition 3.1).

NML and minimax optimality. As we have noted in Section 3.2, the Normalized Maximum Likelihood forecaster is minimax optimal for the problem of sequential probability assignment (up to a regularization term in the version we have described). Unfortunately however, directly bounding the regret of NML (as expressed by the Shtarkov sum) is very challenging even for relatively simple models like GLMs. Beyond the simple linear case (where NML is essentially equivalent to EWA), the only bounds we are aware of only hold asymptotically or for mean estimation without covariates (Rissanen, 1996; Takeuchi and Barron, 1997; Szpankowski, 1998; Cesa-Bianchi and Lugosi, 2006; Grünwald, 2007; Suzuki and Yamanishi, 2018; Jacquet, Shamir and Szpankowski, 2022). We also remark that the equalizer property of NML-style forecasters is not always a desirable feature in the context of our study. Indeed, notice that several of our guarantees are stated in terms of the realized sequence of covariates, and the worst-case upper bounds we provide in terms of problem parameters like d or n are often loose (see, e.g., the bound of Equation (19) and the discussion around it). Thus, equalizer strategies that take into account the impact of predictions on the sequence of covariates (such as the cNML strategy of Liu, Attias and Roy, 2024) may result in unnecessarily large regret when the realized sequence turns out to be benign. Studying the best achievable regret for sequential probability assignment remains an actively researched area, and we remain hopeful that new developments (such as Qian, Rakhlin and Zhivotovskiy, 2024; Jia, Polyanskiy and Rakhlin, 2025) will enable proving tighter finite-sample bounds and a more fine-grained understanding of the best possible rates for special cases like GLMs and beyond.

Uncertainty quantification beyond GLMs. The reduction framework we propose in this paper is clearly applicable for more general models, and we chose to focus on GLMs for the sake of concreteness. Indeed, it is easy to see that our arguments continue to hold regardless of the form of the functional dependence between the parameters of the model and the likelihood, and that one can derive convex confidence sets using our method whenever the log-likelihood is convex (as is the case, for example, for additive models and generalized additive models, cf. Hastie and Tibshirani, 1986). It is much more interesting to study the possibility of extending our framework towards non-convex models such as deep neural networks, and develop a rigorous methodology for statistical inference for modern machine learning systems. We believe that our methodology could prove to be a solid basis for addressing “uncertainty quantification” tasks for machine learning, such as providing confidence sets for local minimizers of the population loss, or valid prediction intervals for individual data points. Obviously, our methodology is already applicable as it is for non-convex log-likelihoods, but its usefulness is limited because in such settings it leads to non-convex confidence sets and prediction intervals. A promising direction towards removing this limitation is to build on recently popular techniques for local linearization of nonlinear models (such as Laplace approximations, cf. MacKay, 1992; Khan et al., 2019; Daxberger et al., 2021; Immer, Korzepa and Bauer, 2021; Cinquin et al., 2024), and building confidence sets on the resulting linear model. We are optimistic that our method that directly addresses statistical questions via reductions to algorithmic questions can prove to be powerful in this highly challenging setting, especially given the abundance of algorithmic results and the relative scarcity of statistical guarantees in this context.

Appendix A: Restricted Bregman Information Gain

In this section, we justify why $\gamma_{n,\lambda}^{\rho,S}$ is called the restricted Bregman information gain. For any subset $S \subseteq [d]$ and any vector $x \in \mathbb{R}^d$, we define $x(S) \in \mathbb{R}^{|S|}$ to be the subvector indexed by elements of S. This can be written as $x(S) = \Pi_S x$, where $\Pi_S \in \mathbb{R}^{|S| \times d}$ is the matrix with rows $(e_i)_{i \in S}$, and $e_1, \dots, e_d$ are the standard basis vectors of $\mathbb{R}^d$. For each subset $S \subseteq [d]$, we will define a restricted version of $Z_{n,\lambda}^{\rho}$ which has domain $\mathbb{R}^{|S|}$. For any vector $v \in \mathbb{R}^{|S|}$, we define the restricted loss ℓ_t^S as

$$\ell_t^S(v) = -\langle v, X_t(S)\rangle Y_t + \psi(\langle v, X_t(S)\rangle) - \log h(Y_t)\,.$$

In addition, we define a collection of convex functions $\rho^S : \mathbb{R}^{|S|} \to \mathbb{R}$ such that for all $S \subseteq [d]$ and $\theta \in \Theta_S$, $\rho(\theta) = \rho^S(\theta(S))$. Note that if $\rho(\theta) = \|\theta\|_p^p$, then we can define ρ^S by $\rho^S(v) = \|v\|_p^p$. We can now define $Z_{n,\lambda}^{\rho,S} : \mathbb{R}^{|S|} \to \mathbb{R}$ to be the function

$$Z_{n,\lambda}^{\rho,S}(v) = \sum_{t=1}^{n} \lambda \ell_t^S(v) + \rho^S(v)\,.$$

We notice that for all $S \subseteq [d]$ and $\theta \in \Theta_S$, $Z_{n,\lambda}^{\rho,S}$ satisfies $Z_{n,\lambda}^{\rho}(\theta) = Z_{n,\lambda}^{\rho,S}(\theta(S))$. If we let $\widehat{\theta}_{n,\lambda,S} = \arg\min_{\theta \in \Theta_S}\{Z_{n,\lambda}^{\rho}(\theta)\}$, then $\widehat{\theta}_{n,\lambda,S}(S) = \arg\min_{v \in \mathbb{R}^{|S|}}\{Z_{n,\lambda}^{\rho,S}(v)\}$. This means that, for any $S \subseteq [d]$ and $\theta \in \Theta_S$,

$$Z_{n,\lambda}^{\rho}(\theta) - Z_{n,\lambda}^{\rho}(\widehat{\theta}_{n,\lambda,S}) = Z_{n,\lambda}^{\rho,S}(\theta(S)) - Z_{n,\lambda}^{\rho,S}(\widehat{\theta}_{n,\lambda,S}(S)) = \mathcal{B}_{Z_{n,\lambda}^{\rho,S}}(\theta(S), \widehat{\theta}_{n,\lambda,S}(S))\,.$$

Therefore, for any subset $S \subseteq [d]$, the definition of the restricted Bregman information gain in Equation (12) is equivalent to

$$\gamma_{n,\lambda}^{\rho,S} = -\log\left(\frac{\int_{\Theta_S} \exp(-\mathcal{B}_{Z_{n,\lambda}^{\rho,S}}(\theta(S), \widehat{\theta}_{n,\lambda,S}(S)))\mathrm{d}\theta}{\int_{\Theta_S} \exp(-\rho^S(\theta(S))\mathrm{d}\theta}\right)\,.$$

Appendix B: Technical Lemmas

Lemma B.1. *Let $q_1, \dots, q_n$ be the sequence of distributions played by the EWA forecaster with learning rate $\lambda > 0$. Then*

$$\mathrm{regret}_{q^n,\lambda}(\bar{\theta}) = -\frac{1}{\lambda}\log\int e^{-\lambda\sum_{t=1}^{n}\left(\ell_t(\theta) - \ell_t(\bar{\theta})\right)}\mathrm{d}q_1(\theta)\,.$$

Proof. By definition of the EWA forecaster,

$$\frac{\mathrm{d}q_t}{\mathrm{d}q_1}(\theta) = \frac{e^{-\lambda \sum_{k=1}^{t-1} \ell_k(\theta)}}{\int e^{-\lambda \sum_{k=1}^{t-1} \ell_k(\theta)} \mathrm{d}q_1(\theta)} \, .$$

Therefore, the total loss of the EWA forecaster is

$$\begin{aligned} \sum_{t=1}^{n} \mathcal{L}_{t,\lambda}(q_t) &= \sum_{t=1}^{n} -\frac{1}{\lambda} \log \int e^{-\lambda \ell_t(\theta)} \mathrm{d}q_t(\theta) \\ &= \sum_{t=1}^{n} -\frac{1}{\lambda} \log \left(\frac{\int e^{-\lambda \sum_{k=1}^{t} \ell_k(\theta)} \mathrm{d}q_1(\theta)}{\int e^{-\lambda \sum_{k=1}^{t-1} \ell_k(\theta)} \mathrm{d}q_1(\theta)} \right) \\ &= -\frac{1}{\lambda} \log \int e^{-\lambda \sum_{t=1}^{n} \ell_t(\theta)} \mathrm{d}q_1(\theta) \, . \end{aligned}$$

The statement follows by subtracting the total loss of $\bar{\theta}$ from both sides. □

Lemma B.2. *Let $x_1, \dots, x_n \in \mathbb{R}^d$ be any sequence satisfying $\max_{t\in[n]} \|x_t\|_2 \le L$ and set $\Lambda_n = \sum_{t=1}^n x_t x_t^\top$. Then for every $\alpha > 0$,*

$$\log\det(\alpha\Lambda_n + \mathrm{Id}) \le \mathrm{rank}(\Lambda_n) \log \left(1 + \frac{\alpha n L^2}{\mathrm{rank}(\Lambda_n)}\right) \, .$$

Proof. Let $k = \mathrm{rank}(\Lambda_n)$ and let $\lambda_1, \dots, \lambda_d$ be the eigenvalues of $\alpha\Lambda_n + \mathrm{Id}$, arranged in descending order. Since Λ_n has rank k, $\lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_d = 1$. This means that $\det(\alpha\Lambda_n + \mathrm{Id}) = \prod_{i=1}^k \lambda_k$. Similarly, $\mathrm{tr}(\alpha\Lambda_n + \mathrm{Id}) = \sum_{i=1}^k \lambda_k + (k - d)$. Using the inequality of arithmetic and geometric means, we have

$$\det(\alpha\Lambda_n + \mathrm{Id}) = \prod_{i=1}^{k} \lambda_k \le \left(\frac{\sum_{i=1}^k \lambda_i}{k} \right)^k = \left(\frac{\mathrm{tr}(\alpha\Lambda_n + \mathrm{Id}) + (k-d)}{k} \right)^k \, .$$

Finally, we notice that

$$\mathrm{tr}(\alpha\Lambda_n + \mathrm{Id}) = \alpha \sum_{t=1}^{n} \|x_t\|_2^2 + d \le \alpha n L^2 + d \, .$$

Combining this with the previous inequality and taking the logarithm of both sides gives the statement of the lemma. □

Lemma B.3. *For any $z' \in \mathbb{R}$, the map $z \mapsto d_\psi(z, z')$ is non-decreasing on the interval $[z', \infty)$ and non-increasing on the interval $(-\infty, z']$.*

Proof. Let $f(z) = d_\psi(z, z')$. From the definition of d_ψ, we have

$$f'(z) = \frac{1}{2}\psi'(z) - \frac{1}{2}\psi'(z/2 + z'/2) \, .$$

Since ψ is convex and differentiable, ψ' is non-decreasing. We notice that for any $z \geq z'$, $z \geq z/2 + z'/2$, which means that $f'(z) \geq 0$. The same argument shows that for any $z \leq z'$, $f'(z) \leq 0$. □

Lemma B.4. *For any $b > 0$, any $z' \in [-b, b]$ and all $z \in \mathbb{R}$, $d_\psi([z]_b, z') \leq d_\psi(z, z')$.*

Proof. Suppose $z \geq z' \geq -b$. If $z \leq b$, then $d_\psi([z]_b, z') = d_\psi(z, z')$. Otherwise, if $z > b$, then $z > [z]_b \geq z'$. Thus, $d_\psi([z]_b, z') \leq d_\psi(z, z')$ follows from Lemma B.3. The same argument works when $z \leq z'$. □

Acknowledgments

The authors would like to thank the following colleagues for insightful discussions throughout the preparation of this work: Yoav Freund, Kwang-Sung Jun, Johannes Kirschner, Junghyun Lee, Nishant Mehta, Francesco Orabona, Sasha Rakhlin, Aaditya Ramdas, Csaba Szepesvári.

Funding

This project has received funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180).

References

Abbasi-Yadkori, Y. (2012). Online learning for linearly parametrized control problems, PhD thesis, University of Alberta.

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, Cs. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24**.

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, Cs. (2012). Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics* 1–9.

Alquier, P. and Lounici, K. (2011). PAC-Bayesian Theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics* **5** 127–145.

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

Barron, A., Rissanen, J. and Yu, B. (1998). The Minimum Description Length Principle in Coding and Modeling. *IEEE Transactions on Information Theory* **44** 2743-2760.

Bartlett, P., Grünwald, P., Harremoës, P., Hedayati, F. and Kotłowski, W. (2013). Horizon-Independent Optimal Prediction with Log-Loss in

Exponential Families. In *Conference on Learning Theory (COLT)* **30** 639–661. PMLR.

Cesa-Bianchi, N. and Lugosi, G. (2001). Worst-Case Bounds for the Logarithmic Loss of Predictors. *Machine Learning* **43** 247–264.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games.* Cambridge university press.

Chowdhury, S. R., Saux, P., Maillard, O. and Gopalan, A. (2023). Bregman Deviations of Generic Exponential Families. *Conference on learning theory* **195**.

Cinquin, T., Pförtner, M., Fortuin, V., Hennig, P. and Bamler, R. (2024). FSP-Laplace: Function-Space Priors for the Laplace Approximation in Bayesian Deep Learning. *Advances in Neural Information Processing Systems* **36**.

Clerico, E. (2024). On the optimality of coin-betting for mean estimation. *arXiv preprint arXiv:2412.02640*.

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society* **30** 178–191.

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

Dalalyan, A. S. and Tsybakov, A. B. (2012). Mirror averaging with sparsity priors. *Bernoulli* **18** 914–944.

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M. and Hennig, P. (2021). Laplace redux – effortless Bayesian deep learning. *Advances in Neural Information Processing Systems* **34** 20089–20103.

de la Peña, V. H., Lai, T. L. and Shao, Q.-M. (2009). *Self-normalized processes: Limit theory and Statistical Applications.* Springer.

Emmenegger, N., Mutny, M. and Krause, A. (2023). Likelihood ratio confidence sets for sequential decision making. *Advances in Neural Information Processing Systems* **36**.

Flynn, H. and Reeb, D. (2024). Tighter Confidence Bounds for Sequential Kernel Regression. *arXiv preprint arXiv:2403.12732*.

Flynn, H., Reeb, D., Kandemir, M. and Peters, J. (2023). Improved Algorithms for Stochastic Linear Bandits Using Tail Bounds for Martingale Mixtures. *Advances in Neural Information Processing Systems* **36**.

Foster, D. J., Kale, S., Luo, H., Mohri, M. and Sridharan, K. (2018). Logistic regression: The importance of being improper. In *Conference on learning theory* 167–208.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55** 119–139.

Gács, P., Tromp, J. T. and Vitányi, P. M. (2001). Algorithmic statistics. *IEEE Transactions on Information Theory* **47** 2443–2463.

Gaillard, P., Gerchinovitz, S., Huard, M. and Stoltz, G. (2019). Uniform regret bounds over $\mathbb{R}^d$ for the sequential linear regression problem with the square loss. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory* **98** 404–432. PMLR.

Gales, S. B., Sethuraman, S. and Jun, K.-S. (2022). Norm-agnostic linear bandits. In *International Conference on Artificial Intelligence and Statistics* 73–91.

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

Grünwald, P. D. (2007). *The minimum description length principle.* MIT press.

Grünwald, P. and Harremoës, P. (2009). Finiteness of redundancy, regret, Shtarkov sums, and Jeffreys integrals in exponential families. In *2009 IEEE International Symposium on Information Theory* 714-718.

Grünwald, P. D. and Mehta, N. A. (2019). A tight excess risk bound via a unified PAC-Bayesian–Rademacher–Shtarkov–MDL complexity. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory* **98** 433–465. PMLR.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical science* **1** 297–310.

Immer, A., Korzepa, M. and Bauer, M. (2021). Improving predictions of Bayesian neural nets via local linearization. In *International conference on artificial intelligence and statistics* 703–711.

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

Jia, Z., Polyanskiy, Y. and Rakhlin, A. (2025). On the Minimax Regret of Sequential Probability Assignment via Square-Root Entropy. *arXiv preprint arXiv:2503.17823.*

Jun, K.-S., Bhargava, A., Nowak, R. and Willett, R. (2017). Scalable generalized linear bandits: Online computation and hashing. *Advances in Neural Information Processing Systems* **30**.

Kakade, S. M. and Ng, A. (2004). Online bounds for Bayesian algorithms. *Advances in neural information processing systems* **17**.

Kakade, S. M., Seeger, M. W. and Foster, D. P. (2005). Worst-case bounds for Gaussian process models. *Advances in Neural Information Processing Systems* **18**.

Khan, M. E. E., Immer, A., Abedi, E. and Korzepa, M. (2019). Approximate inference turns deep networks into Gaussian processes. *Advances in neural information processing systems* **32**.

Kirschner, J., Krause, A., Meziu, M. and Mutny, M. (2025). Confidence Estimation via Sequential Likelihood Mixing. *arXiv preprint arXiv:2502.14689.*

Kotłowski, W. (2020). Scale-invariant unconstrained online learning. *Theoretical Computer Science* **808** 139–158.

Kotłowski, W. and Grünwald, P. (2011). Maximum Likelihood vs. Sequential Normalized Maximum Likelihood in On-line Density Estimation. In *Conference on Learning Theory (COLT)* **19** 457–476. JMLR.

Lattimore, T. (2023). A lower bound for linear and kernel regression with adaptive covariates. In *The Thirty Sixth Annual Conference on Learning Theory* 2095–2113.

Lattimore, T. and Szepesvári, Cs. (2020). *Bandit algorithms.* Cambridge University Press.

Lee, J., Yun, S.-Y. and Jun, K.-S. (2024a). Improved Regret Bounds of (Multinomial) Logistic Bandits via Regret-to-Confidence-Set Conversion. In *International Conference on Artificial Intelligence and Statistics* 4474–4482.

Lee, J., Yun, S.-Y. and Jun, K.-S. (2024b). A Unified Confidence Sequence for Generalized Linear Models, with Applications to Bandits. *arXiv preprint arXiv:2407.13977.*

Li, L., Lu, Y. and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning* 2071–2080.

Liang, F. and Barron, A. (2006). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Trans. Inf. Theor.* **50** 2708–2726.

Littlestone, N. and Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation* **108** 212–261.

Liu, Z., Attias, I. and Roy, D. M. (2024). Sequential probability assignment with contexts: Minimax regret, contextual shtarkov sums, and contextual normalized maximum likelihood. *Advances in Neural Information Processing Systems* **37** 13982–14015.

Lugosi, G. and Neu, G. (2024). Online-to-PAC Conversions: Generalization Bounds via Regret Analysis.

MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural computation* **4** 448–472.

Orabona, F. and Jun, K.-S. (2024). Tight Concentrations and Confidence Sequences From the Regret of Universal Portfolio. *IEEE Transactions on Information Theory* **70** 436-455.

Qian, J., Rakhlin, A. and Zhivotovskiy, N. (2024). Refined risk bounds for unbounded losses via transductive priors. *arXiv preprint arXiv:2410.21621.*

Rakhlin, A. and Sridharan, K. (2017). On equivalence of martingale tail bounds and deterministic regret inequalities. In *Conference on Learning Theory* 1704–1722. PMLR.

Rissanen, J. (1996). Fisher Information and Stochastic Complexity. *IEEE Trans. Information Theory* **IT-42** 40-47.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* **41** 1397–1409.

Robbins, H. and Siegmund, D. (1970). Boundary Crossing Probabilities for the Wiener Process and Sample Sums. *The Annals of Mathematical Statistics* **41** 1410–1429.

Roos, T. and Rissanen, J. (2008). On sequentially normalized maximum likelihood models. In *Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08).*

Shafer, G. and Vovk, V. (2001). *Probability and finance: it's only a game!*

491. John Wiley & Sons.

Shtarkov, A. (1987). Universal sequential coding of single messages. *Problems of Information Transmission* **23** 175–186.

Suzuki, A. and Yamanishi, K. (2018). Exact Calculation of Normalized Maximum Likelihood Code Length Using Fourier Analysis. In *2018 IEEE International Symposium on Information Theory (ISIT)* 1211-1215.

Szpankowski, W. (1998). On the asymptotics of the minimax redundancy arising in a universal coding. *Problems of Information Transmission* **34** 142–146.

Takeuchi, J. and Barron, A. (1997). Asymptotically minimax regret for exponential families. In *Proceedings SITA '97* 665–668.

Ville, J. (1939). *Étude critique de la notion de collectif* **3**. Gauthier-Villars Paris.

Vovk, V. (1990). Aggregating strategies. In *Proceedings of the third annual workshop on Computational learning theory (COLT)* 371–386.

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

Vovk, V. G. and Shafer, G. R. (2003). Kolmogorov's Contributions to the Foundations of Probability. *Problems of Information Transmission* **39** 21–31.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge university press.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics* **16** 117–186.

Waudby-Smith, I. and Ramdas, A. (2023). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86** 1-27.

Xie, Q. and Barron, A. R. (2000). Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inf. Theory* **46** 431–445.