A Unified Confidence Sequence for Generalized Linear Models, with Applications to Bandits

Jungyhun Lee, Se-Young Yun
Kim Jaechul Graduate School of AI
KAIST
Seoul, Republic of Korea
{jh_lee00, yunseyoung}@kaist.ac.kr

Kwang-Sung Jun
Department of Computer Science
University of Arizona
Tucson, AZ, USA
kjun@cs.arizona.edu

Abstract

We present a unified likelihood ratio-based confidence sequence (CS) for *any* (self-concordant) generalized linear models (GLMs) that is guaranteed to be convex and numerically tight. We show that this is on par or improves upon known CSs for various GLMs, including Gaussian, Bernoulli, and Poisson. In particular, for the first time, our CS for Bernoulli has a $\mathrm{poly}(S)$-free radius where S is the norm of the unknown parameter. Our first technical novelty is its derivation, which utilizes a time-uniform PAC-Bayesian bound with a uniform prior/posterior, despite the latter being a rather unpopular choice for deriving CSs. As a direct application of our new CS, we propose a simple and natural optimistic algorithm called `OFUGLB` applicable to *any* generalized linear bandits (**GLB**; Filippi et al. (2010)). Our analysis shows that the celebrated optimistic approach simultaneously attains state-of-the-art regrets for various self-concordant (not necessarily bounded) **GLB**s, and even $\mathrm{poly}(S)$-free for bounded **GLB**s, including logistic bandits. The regret analysis, our second technical novelty, follows from combining our new CS with a new proof technique that completely avoids the previously widely used self-concordant control lemma (Faury et al., 2020, Lemma 9), which may be of independent interest. Finally, we verify numerically that `OFUGLB` significantly outperforms the prior state-of-the-art (Lee et al., 2024) for logistic bandits.

1 Introduction

One paramount task in statistics and machine learning is to estimate the uncertainty of the underlying model from (possibly noisy) observations. For example, in interactive machine learning scenarios such as bandits (Lattimore and Szepesvári, 2020; Robbins, 1952; Thompson, 1933) and recently reinforcement learning with human feedback (RLHF; Christiano et al. (2017); Ouyang et al. (2022)), at each time step t, the learner chooses an action $\boldsymbol{x}_t$ from an available set of actions $\mathcal{X}_t$ and observes reward or outcome r_t that is modeled as a distribution whose mean is an unknown function f^* of $\boldsymbol{x}_t$; i.e., $r_t \sim p(\cdot|\boldsymbol{x}_t; f^*)$. One popular choice of such a model is generalized linear model (GLM; McCullagh and Nelder (1989)) that extends exponential family distributions to have linear structure in its natural parameter, i.e., $\langle \boldsymbol{x}, \boldsymbol{\theta}_\star \rangle$ where $\boldsymbol{\theta}_\star$ is an unknown parameter, which means that the mean function is $f^*(\boldsymbol{x}) = \mu(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star \rangle)$ for some inverse link function μ. This encompasses a wide range of distributions, which in turn makes it ubiquitous in various real-world applications, such as news recommendations (Bernoulli; Li et al. (2010, 2012)), social network influence maximization (Poisson; Gisselbrecht et al. (2015); Lage et al. (2013)), and more. In such tasks, the learner must estimate the uncertainty about $\boldsymbol{\theta}_\star$ *at each time step* $t \geq 1$, given observations $\{(\boldsymbol{x}_s, r_s)\}_{s=1}^{t-1}$, to plan his/her course of actions and make wise decisions. One popular and useful way to capture the uncertainty is via a *time-uniform confidence sequence (CS)* $\{\mathcal{C}_t(\delta)\}_{t=1}^{\infty}$, which takes the form of $\mathbb{P}[\exists t \geq 1 : \boldsymbol{\theta}_\star \notin \mathcal{C}_t(\delta)] \leq \delta$. Recently, CS has been described as one of the key components for *safe*

Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

anytime-valid inference (SAVI) that can ensure the validity/safeness of sequentially adaptive statistical inference (Ramdas et al., 2023).

There has been much work on deriving CS for specific families of distributions. Many common distributions are in a smaller family, called *generalized linear models (GLMs)*. Existing CSs for GLM, however, are far from ideal. Much of the prior works focus on obtaining CS for specific instantiations of GLMs, such as Gaussian (Abbasi-Yadkori et al., 2011; Flynn et al., 2023) and Bernoulli (Abeille et al., 2021; Faury et al., 2020, 2022; Lee et al., 2024). Especially for Bernoulli, all the existing CSs suffer from $\text{poly}(S)$ factor in the radius, where S is the norm of the unknown parameter $\boldsymbol{\theta}_\star$. Emmenegger et al. (2023); Jun et al. (2017); Li et al. (2017) proposed generic CSs that work for any convex GLMs, but their radii all suffer from a globally worst-case curvature of μ, which is detrimental in many cases (e.g., for Bernoulli, it scales as e^S).

Contributions. First, we propose a *unified* construction of likelihood ratio-based CS for any convex GLMs (Theorem 3.1) and then instantiate it as an ellipsoidal CS for *self-concordant GLMs*, including Bernoulli, Gaussian, and Poisson distributions (Theorem 3.2). *Notably, we keep track of all the constants so that any practitioner can directly implement it without trouble.* The proof uses ingredients from time-uniform PAC-Bayesian bounds (Chugg et al., 2023) – martingale + Donsker-Varadhan representation of KL + Ville's inequality. The main technical novelty lies in using *uniform* prior/posterior for the analysis, inspired by various literature on portfolios (Blum and Kalai, 1999) and fast rates in statistical/online learning (Foster et al., 2018; Grünwald and Mehta, 2020; Hazan et al., 2007; van Erven et al., 2015).

Secondly, we apply our novel CSs to contextual generalized linear bandits (**GLB**; Filippi et al. (2010)) with changing (and adversarial) arm sets, and propose a new algorithm called **Optimism in the Face of Uncertainty for Generalized Linear Bandits** (`OFUGLB`). `OFUGLB` employs the simple and standard optimistic approach, choosing an arm that maximizes the upper confidence bound (UCB) computed by our CS (Abbasi-Yadkori et al., 2011; Auer, 2002). We show that `OFUGLB` achieves the state-of-the-art regret bounds for self-concordant (possibly *unbounded*) **GLB** (Theorem 4.1). This is the first time a *purely* optimistic strategy attains such $\text{poly}(S)$-free regret for logistic bandits in the sense that `OFUGLB` does not involve an explicit warmup phase. Our other significant main technical contribution is the analysis of `OFUGLB` since naïvely applying existing analysis techniques for optimistic algorithms (Abeille et al., 2021; Lee et al., 2024) yields a regret bound whose leading term scales with $\text{poly}(S)$. We identify the key reason for such additional dependency as the use of self-concordance control lemma (Faury et al., 2020, Lemma 9), and provide an alternate analysis that completely bypasses it, which may be of independent interest in the bandits community and beyond.

2 Problem Setting

We consider the realizable (online) regression with the **generalized linear model** (GLM; McCullagh and Nelder (1989)) whose conditional density of r is given as

$$dp(r|\boldsymbol{x};\boldsymbol{\theta}_\star) = \exp\left(\frac{r\langle \boldsymbol{x}, \boldsymbol{\theta}_\star\rangle - m(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star\rangle)}{g(\tau)} + h(r,\tau)\right) d\nu, \tag{1}$$

where τ is some known scaling (temperature) parameter and ν is some known base measure (e.g., Lebesgue, counting). We assume the following:

Assumption 1. *The domain X for arm (context) $\boldsymbol{x}$ satisfies $X \subseteq \mathcal{B}^d(1)$.*

Assumption 2. *$\boldsymbol{\theta}_\star \in \Theta \subseteq \mathcal{B}^d(S) := \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \leq S\}$ for some known $S > 0$. Also, Θ is nonempty, compact, and convex with intrinsic dimension[1] d.*

Assumption 3. *m is three times differentiable and convex, i.e., m''' exists and $\dot{\mu} := m'' \geq 0$.*

In **GLB** problem, at each time $t \in [T]$, the learner observes a time-varying, arbitrary (often called adversarial) arm set $\mathcal{X}_t \subseteq X$, chooses a $\boldsymbol{x}_t \in \mathcal{X}_t$, and receives a reward $r_t \sim p(\cdot|\boldsymbol{x}_t, \boldsymbol{\theta}_\star)$. Let $\mathcal{X}_{[T]} := \cup_{t=1}^T \mathcal{X}_t$ and $\Sigma_{t+1} := \sigma(\Sigma_t, r_t, \boldsymbol{x}_{t+1})$ with $\Sigma_0 = \sigma(\boldsymbol{x}_1)$ be the filtration in the canonical bandit model (Lattimore and Szepesvári, 2020, Chapter 4.6). From well-known properties of GLMs, we have that $\mathbb{E}[r_t|\Sigma_t] = m'(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star\rangle) \triangleq \mu(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star\rangle)$ and $\text{Var}[r_t|\Sigma_t] = g(\tau)\dot{\mu}(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star\rangle)$, where μ is

[1] the linear-algebraic dimension (minimum number of basis vectors spanning it) of the affine span of Θ in $\mathbb{R}^d$.

the *inverse link function*. We also define the following quantities:

$$R_{\mu,\star} := \max_{\boldsymbol{x}\in\mathcal{X}_{[T]}} |\mu(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star\rangle)|, \quad R_{\dot{\mu}} := \max_{\boldsymbol{x}\in\mathcal{X}_{[T]}, \boldsymbol{\theta}\in\Theta} \dot{\mu}(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle). \tag{2}$$

Note that many common distributions, such as Gaussian ($\mu(z) = z$), Poisson ($\mu(z) = e^z$), and Bernoulli ($\mu(z) = (1 + e^{-z})^{-1}$), fall under the umbrella of GLM.

3 Unified Likelihood Ratio-based Confidence Sequence for GLMs

The learner's goal is to output a time-uniform confidence sequence (CS) for $\boldsymbol{\theta}_\star$, $\mathbb{P}[\exists t \geq 1 : \boldsymbol{\theta}_\star \notin \mathcal{C}_t(\delta)] \leq \delta$, where $\mathbb{P}$ is w.r.t. the randomness of the confidence sets $\mathcal{C}_t(\delta)$. In this work, we are particularly interested in the log-likelihood-based confidence set "centered" at the *norm-constrained*, batch maximum likelihood estimator (MLE):

$$\mathcal{C}_t(\delta) := \left\{ \boldsymbol{\theta} \in \Theta : \mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \leq \beta_t(\delta)^2 \right\}, \tag{3}$$

where $\beta_t(\delta)^2$ is the "radius" of the CS that we will define later, and $\mathcal{L}_t(\boldsymbol{\theta})$ is the negative log-likelihood of $\boldsymbol{\theta}$ w.r.t. data collected up to $t - 1$, and

$$\mathcal{L}_t(\boldsymbol{\theta}) := \sum_{s=1}^{t-1} \left\{ \ell_s(\boldsymbol{\theta}) \triangleq \frac{-r_s\langle \boldsymbol{x}_s, \boldsymbol{\theta}\rangle + m(\langle \boldsymbol{x}_s, \boldsymbol{\theta}\rangle)}{g(\tau)} \right\}, \quad \widehat{\boldsymbol{\theta}}_t := \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_t(\boldsymbol{\theta}). \tag{4}$$

Note that $h(r_s, \tau)$ is omitted as it plays no role in the confidence set nor the MLE.

The form of the confidence set is the same as Lee et al. (2024) in that it leverages the batched constrained MLE as opposed to the batch regularized MLE (Abbasi-Yadkori et al., 2011), sequential (regularized) MLE (Abbasi-Yadkori et al., 2012; Emmenegger et al., 2023; Faury et al., 2022; Jun et al., 2017; Wasserman et al., 2020), or expected loss over some distribution (e.g., Gaussian) without committing to an estimator (Flynn et al., 2023). As one can see later, our derivation of the CS also starts from an expectation of loss over a prior distribution of $\boldsymbol{\theta}$ without committing to an estimator, yet we introduce the estimator to avoid the computational difficulty of evaluating the expectation.

Our first main contribution is the following unified confidence sequence for *any* GLMs, regardless of whether it is bounded or not, as long as the corresponding log-likelihood loss is Lipschitz:

Theorem 3.1 (Unified CS for GLMs)**.** *Let $L_t := \max_{\boldsymbol{\theta}\in\Theta} \left\|\nabla\mathcal{L}_t(\boldsymbol{\theta})\right\|_2$ be the Lipschitz constant*[a] *of $\mathcal{L}_t(\cdot)$ that may depend on $\{(\boldsymbol{x}_s, r_s)\}_{s=1}^{t-1}$. Then, we have $\mathbb{P}[\exists t \geq 1 : \boldsymbol{\theta}_\star \notin \mathcal{C}_t(\delta)] \leq \delta$, where*

$$\beta_t(\delta)^2 = \log\frac{1}{\delta} + \inf_{c\in(0,1]} \left\{ d\log\frac{1}{c} + 2SL_t c \right\} \leq \log\frac{1}{\delta} + d\log\left(e \vee \frac{2eSL_t}{d} \right), \tag{5}$$

where the last inequality follows from the choice $c = 1 \wedge \frac{d}{2SL_t}$.

[a] **Rademacher's theorem** (Federer, 1996, Theorem 3.1.6): for differentiable function $\mathcal{L} : \Theta \to \mathbb{R}$, $\inf\left\{ L \geq 0 : |\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}')| \leq L\left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2, \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta \right\} = \max_{\boldsymbol{\theta}\in\Theta} \left\|\nabla\mathcal{L}(\boldsymbol{\theta})\right\|_2$.

Practically, the computation of L_t involves a potentially non-concave maximization over a convex set, which is NP-hard in general (Murty and Kabadi, 1987). In Table 1, we provide *closed-form* (up to absolute constants), high-probability upper bounds for L_t's for various GLMs:

Comparisons to Prior Works. There have been some works on providing CSs for either generic GLMs (Emmenegger et al., 2023; Jun et al., 2017; Li et al., 2017) or specific GLMs (linear: Flynn et al. (2023), logistic: Abeille et al. (2021); Faury et al. (2020); Lee et al. (2024)). The generic CSs are generally not tight as the "radius" often scales with $\kappa := \left(\min_{\boldsymbol{x}\in X, \boldsymbol{\theta}\in\Theta} \dot{\mu}(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle)\right)^{-1}$, which scales exponentially in S for Bernoulli (Faury et al., 2020). For instance, Theorem 1 of Jun et al. (2017) and Theorem 1 of Li et al. (2017) proved CSs of the form $\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{V}_t}^2 \leq \zeta_1(t, \delta)$, with ζ_1 always

Table 1: Instantiations of L_t's for various GLMs.

GLM	Upper bounds for L_t	Proof
Bounded[b] by M	$(M + 2SR_{\dot\mu})(t-1)/g(\tau)$	Trivial from triangle inequality
Bernoulli	$(1 + S/2)(t-1)$	Trivial from above
Gaussian[a]	$\sigma^{-2}\left(St + \sigma\sqrt{t \log \frac{d}{\delta}}\right)$	Appendix C.1
Poisson[a]	$e^S t + \log \frac{d}{\delta}$	Appendix C.2

[a] Here we omit the absolute constants; these are made explicit in the proofs.
[b] as in $\max_{\boldsymbol{x} \in \mathcal{X}_{[T]}, \boldsymbol{\theta} \in \Theta} |r - \mu(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star \rangle)| \leq M < \infty$.

scaling with κ. Emmenegger et al. (2023) proposed a CS using weighted, sequential likelihood testing that is empirically shown to be superior to other approaches. However, their Theorem 3, which rewrites the likelihood-based CS as the form $D(\boldsymbol{\theta}, \boldsymbol{\theta}_\star) \leq \zeta_2(t, \delta)$ for some well-defined Bregman divergence $D(\cdot, \cdot)$ and ζ_2, always scales with κ as well and thus a direct comparison with our CS is not possible. Refer to Appendix A for further discussions on CSs for exponential family. On the other hand, the CSs for specific GLMs are inapplicable to GLM models beyond what they are designed for and may not be tight enough. For the Bernoulli distribution, the prior state-of-the-art (likelihood ratio-based) CS radius is $\mathcal{O}\left(S \log \frac{1}{\delta} + d \log \frac{St}{d}\right)$ of Lee et al. (2024), while our theorem gives us $\mathcal{O}\left(\log \frac{1}{\delta} + d \log \frac{St}{d}\right)$. We *completely* remove the $\mathrm{poly}(S)$-dependency from the radius, resolving one of the open problems posited by Lee et al. (2024). Later in Section 4, we show that this improvement is significant, both theoretically *and* numerically, for logistic bandits.

3.1 Ellipsoidal Confidence Sequence for Self-Concordant GLMs

Having an ellipsoidal version of CS is often beneficial, as this is easier to implement in practice. In particular, in the context of bandits, this allows one to equivalently rewrite the optimistic optimization in the UCB algorithm as a *closed-form bonus-based UCB algorithm*, even if the MLE requires an iterative algorithm. This section provides the ellipsoidal version of Theorem 3.1 for the following class of GLMs whose inverse link function μ satisfies the following:

Assumption 4 (Russac et al. (2021)). *μ is* **(generalized) self-concordant***, i.e., the following quantity is well-defined (finite):* $R_s := \inf\left\{R \geq 0 : |\ddot\mu(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)| \leq R\dot\mu(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle), \ \forall \boldsymbol{x} \in X, \boldsymbol{\theta} \in \Theta\right\}$.

For instance, Bernoulli satisfies this with $R_s = 1$, and more generally, GLM bounded by R a.s. satisfy this assumption with $R_s = R$ (Sawarni et al., 2024, Lemma 2.1). Many unbounded GLMs also satisfy this assumption, such as Gaussian ($\mu(z) = z \Rightarrow R_s = 0$), Poisson ($\mu(z) = e^z \Rightarrow R_s = 1$), and Exponential ($\mu(z) = 1 \Rightarrow R_s = 0$).

For this class of GLMs, we have the following slightly relaxed *ellipsoidal* CS, whose proof is deferred to Appendix C.3:

Theorem 3.2 (Ellipsoidal CS for Self-Concordant GLMs)**.** *With the same notations as Theorem 3.1, we have* $\mathbb{P}[\exists t \geq : \boldsymbol{\theta}_\star \notin \mathcal{E}_t(\delta)] \leq \delta$, *where*

$$\mathcal{E}_t(\delta) := \left\{ \boldsymbol{\theta} \in \Theta : \left\| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t \right\|^2_{\nabla^2 \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) + \frac{1+SR_s}{2S^2} \boldsymbol{I}_d} \leq \gamma_t(\delta)^2 \triangleq 2(1 + SR_s)(1 + \beta_t(\delta)^2) \right\}. \tag{6}$$

Let us denote $A \lesssim B$ if $A \leq cB$ for some absolute constant $c > 0$. Note that the relaxation is "strict" (i.e., Theorem 3.2 is strictly looser than Theorem 3.1) when $R_s > 0$. For Gaussian distribution, we have that $R_s = 0$; thus, the ellipsoidal relaxation is *exact*! We then have that $\nabla^2 \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) = \frac{1}{\sigma^2} \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\top =: \frac{1}{\sigma^2} \boldsymbol{V}_t$, and $L_t \lesssim St$ with high probability (Proposition C.1). Combining everything, we have $\left\| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t \right\|^2_{\boldsymbol{V}_t} \lesssim \sigma^2 \left(\log \frac{t}{\delta} + d \log \frac{St}{d} \right)$, which *completely* matches the prior state-of-the-art radius as in Lemma D.10 of Flynn et al. (2023) with $c = \sigma^2 S^2$.

3.2 Proof of Theorem 3.1 – PAC-Bayes Approach with Uniform Prior

We consider $M_t(\boldsymbol{\theta}) := \exp\left(\mathcal{L}_t(\boldsymbol{\theta}_\star) - \mathcal{L}_t(\boldsymbol{\theta})\right)$, the log-likelihood ratio between the (estimated) distribution corresponding to $\boldsymbol{\theta}$ and the true distribution corresponding to $\boldsymbol{\theta}_\star$. This has been the subject of study for over 50 years (Darling and Robbins, 1967a,b; Lai, 1976; Robbins and Siegmund, 1972) and recently revisited by statistics and machine learning communities (Emmenegger et al., 2023; Flynn et al., 2023; Ramdas et al., 2023; Wasserman et al., 2020).

We follow the usual recipes for deriving time-uniform PAC-Bayesian bound (Alquier, 2024; Chugg et al., 2023). We start with the following time-uniform property:

Lemma 3.1. *Let $\delta \in (0,1)$. For any data-independent probability measure $\mathbb{Q}$ on Θ, we have:*

$$\mathbb{P}\left(\exists t \geq 1 : \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}}[M_t(\boldsymbol{\theta})] \geq \frac{1}{\delta}\right) \leq \delta, \tag{7}$$

where $\mathbb{P}$ is over the randomness of the data (and thus randomness of $\mathcal{L}_t$'s).

Proof. First, it is easy to see that $M_t(\boldsymbol{\theta}) = \prod_{s=1}^t \frac{dp(r_s|x_s;\boldsymbol{\theta})}{dp(r_s|x_s;\boldsymbol{\theta}_\star)}$ is a nonnegative martingale w.r.t. Σ_t, as

$$\mathbb{E}[M_t(\boldsymbol{\theta})|\Sigma_{t-1}] = M_{t-1}(\boldsymbol{\theta})\mathbb{E}\left[\frac{dp(r_t|x_t;\boldsymbol{\theta})}{dp(r_t|x_t;\boldsymbol{\theta}_\star)}\middle|\Sigma_{t-1}\right] = M_{t-1}(\boldsymbol{\theta})\underbrace{\int_{\mathcal{R}} \frac{dp(r|x_t;\boldsymbol{\theta})}{dp(r|x_t;\boldsymbol{\theta}_\star)} dp(r|x_t;\boldsymbol{\theta}_\star)}_{=1},$$

where $\mathcal{R}$ is the support of the GLM. (Note that this property is not specific to GLMs and holds for any distributions over measurable spaces.)

Now consider the random variable $\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}}[M_t(\boldsymbol{\theta})]$, which is adapted to Σ_t. This is a martingale, as

$$\mathbb{E}[\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}}[M_t(\boldsymbol{\theta})]|\Sigma_{t-1}] \overset{(*)}{=} \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}}[\mathbb{E}[M_t(\boldsymbol{\theta})|\Sigma_{t-1}]] = \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}}[M_{t-1}(\boldsymbol{\theta})]$$

where $(*)$ follows from the Tonelli's theorem. The desired statement then follows from Ville's inequality (Ville, 1939). □

We recall the variational representation of the KL divergence:

Lemma 3.2 (Theorem 2.1 of Donsker and Varadhan (1983))**.** *For two probability measures $\mathbb{P}, \mathbb{Q}$ over Θ, we have the following: $D_{\mathrm{KL}}(\mathbb{P}||\mathbb{Q}) = \sup_{g:\Theta\to\mathbb{R}} \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[g(\boldsymbol{\theta})] - \log \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}}[e^{g(\boldsymbol{\theta})}]$.*

We then have the following:

Lemma 3.3. *For any data-independent prior $\mathbb{Q}$ and any sequence of adapted posterior distributions (possibly learned from the data) $\{\mathbb{P}_t\}$, the following holds: for any $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists t \geq 1 : \mathcal{L}_t(\boldsymbol{\theta}_\star) - \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}_t}[\mathcal{L}_t(\boldsymbol{\theta})] \geq \log\frac{1}{\delta} + D_{\mathrm{KL}}(\mathbb{P}_t||\mathbb{Q})\right) \leq \delta. \tag{8}$$

Proof. Note that

$$\log \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}}[M_t(\boldsymbol{\theta})] - \mathcal{L}_t(\boldsymbol{\theta}_\star) = \log \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}}[\exp\left(-\mathcal{L}_t(\boldsymbol{\theta})\right)] \overset{(*)}{\geq} \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}_t}[-\mathcal{L}_t(\boldsymbol{\theta})] - D_{\mathrm{KL}}(\mathbb{P}_t||\mathbb{Q}),$$

where $(*)$ follows from Lemma 3.2 with $g(\cdot) = -\mathcal{L}_t(\cdot)$. By Lemma 3.1, we have that $\mathbb{P}\left(\exists t \geq 1 : \log\frac{1}{\delta} \leq \log \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}}[M_t(\boldsymbol{\theta})]\right) \leq \delta$. Rearranging gives the desired statement. □

Remark 1 (Choice of KL)**.** *One can replace KL with other divergences with similar variational formulations (Ohnishi and Honorio, 2021). As we will show later, KL suffices for our purpose.*

Up to now is well-known in the PAC-Bayes literature. Our main technical novelty lies in how to choose $\mathbb{Q}$ and $\mathbb{P}_t$, which is as follows: for $c \in (0,1]$ to be determined later,

$$\mathbb{Q} = \mathrm{Unif}(\Theta), \quad \mathbb{P}_t = \mathrm{Unif}(\widetilde{\Theta}_t \triangleq (1-c)\widehat{\boldsymbol{\theta}}_t + c\Theta), \tag{9}$$

where $\mathrm{Unif}(\cdot)$ is the (continuous) uniform distribution in the Lebesgue measure and $\boldsymbol{a}+\Theta = \{\boldsymbol{a}+\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta\}$ for a vector $\boldsymbol{a} \in \mathbb{R}^d$.

Algorithm 1: `OFUGLB`

Initialize $\mathcal{C}_1 = \Theta$;
Pull a random arm $\boldsymbol{x}_1 \in \mathcal{X}_1$ and receive a reward r_1;
for $t = 2, 3, \cdots$ **do**
  Compute the norm-constrained MLE: $\widehat{\boldsymbol{\theta}}_t \leftarrow \arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_t(\boldsymbol{\theta})$;
  Update the confidence set: $\mathcal{C}_t \leftarrow \left\{ \boldsymbol{\theta} \in \Theta : \mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) \leq \beta_t(\delta)^2 \right\}$, where $\beta_t(\delta)^2$ is defined as in Theorem 3.1;
  UCB step: $(\boldsymbol{x}_t, \boldsymbol{\theta}_t) \leftarrow \arg\max_{\boldsymbol{x} \in \mathcal{X}_t, \boldsymbol{\theta} \in \mathcal{C}_t} \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$;
  Pull the arm $\boldsymbol{x}_t$ and receive a reward r_t;

Then, note that

$$D_{\mathrm{KL}}(\mathbb{P}_t || \mathbb{Q}) = \log \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(\widetilde{\Theta})} = \log \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}\left((1-c)\widehat{\boldsymbol{\theta}}_t + c\Theta \right)} = \log \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(c\Theta)} = \log \frac{\mathrm{vol}(\Theta)}{c^d \mathrm{vol}(\Theta)} = d \log \frac{1}{c},$$

where $\mathrm{vol}(\cdot)$ is the volume measured by the Lebesgue measure in $\mathbb{R}^d$.

Remark 2 (Our choice of posterior)**.** *The main intuition behind the translated/shrunken posterior is to show that a sufficiently large volume of Θ is sufficiently near $\widehat{\boldsymbol{\theta}}_t$. Indeed, in the literature, such choice has been considered for the first time in proof of Theorem 1 of Blum and Kalai (1999), and later in fast rates in online learning (Foster et al., 2018; Hazan et al., 2007). To our knowledge, this is the first time such a translated/shrunken posterior has been used in the PAC-Bayes context.*

We also have that

$$\mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}_t}[\mathcal{L}_t(\boldsymbol{\theta})] = \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) + \mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}_t}[\mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t)] \leq \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) + 2SL_t c,$$

where the last inequality follows from the Lipschitzness of $\mathcal{L}_t(\cdot)$ and the observation that for $\boldsymbol{\theta} = (1-c)\widehat{\boldsymbol{\theta}}_t + c\tilde{\boldsymbol{\theta}} \in \widetilde{\Theta}_t$, $\left\| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t \right\|_2 = c \left\| \tilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_t \right\|_2 \leq 2Sc$. Combining everything and minimizing over $c \in (0, 1]$, the first part of the statement is done. □

4 `OFUGLB`: A Generic, State-of-the-Art UCB Algorithm for Self-Concordant Generalized Linear Bandits

As a direct application of our CS, we consider self-concordant **GLB** (Filippi et al., 2010; Janz et al., 2024), where at each time t, the learner chooses a $\boldsymbol{x}_t \in \mathcal{X}_t$ dependent on the history $\{(x_s, r_s)\}_{s=1}^{t-1}$ and receives $r_t \sim p(\cdot | \boldsymbol{x}_t, \boldsymbol{\theta}_\star)$. The learner's goal is to minimize the (pseudo-)regret:

$$\mathsf{Reg}(T) := \sum_{t=1}^{T} \left\{ \mu(\langle \boldsymbol{x}_{t,\star}, \boldsymbol{\theta}_\star \rangle) - \mu(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star \rangle) \right\}, \tag{10}$$

where $\boldsymbol{x}_{t,\star} := \arg\max_{\boldsymbol{x} \in \mathcal{X}_t} \mu(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star \rangle)$ is the optimal action at time t.

Inspired by the optimism principle (Abbasi-Yadkori et al., 2011; Auer, 2002), based on our new, improved confidence sequence (Theorem 3.1), we propose `OFUGLB` (Algorithm 1), a generic UCB-type algorithm that applies to *any* instantiations of **GLB**. Through a new proof technique that allows us to circumvent κ- and $\mathrm{poly}(S)$-dependencies in the leading term, our unified algorithm attains or improves the known state-of-the-art regret bound for the class of *self-concordant* **GLB**, which encompasses a zoo of well-studied stochastic bandits such as linear (Abbasi-Yadkori et al., 2011; Auer, 2002), Poisson (Gisselbrecht et al., 2015), logistic (Abeille et al., 2021; Faury et al., 2020), etc.

We define the following problem difficulty quantities: denoting $\overline{\mathcal{X}}(T) := \bigcup_{t \in [T]} \mathcal{X}_t$,

$$\kappa_\star(T) := \frac{1}{\frac{1}{T}\sum_{t \in [T]} \dot{\mu}(\langle \boldsymbol{x}_{t,\star}, \boldsymbol{\theta}_\star \rangle)}, \quad \kappa(T) := \max_{\boldsymbol{x} \in \overline{\mathcal{X}}(T), \boldsymbol{\theta} \in \Theta} \frac{1}{\dot{\mu}(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)}. \tag{11}$$

These may scale exponentially in S, e.g., for logistic bandits (Faury et al., 2020; Filippi et al., 2010), but we will later show that through our new analysis, the leading term of the regret scales *inversely* with $\kappa_\star(T)$, and the transient term scales linearly with $\kappa(T)$.

We now present the *unified & state-of-the-art* regret guarantee for self-concordant **GLB**s:

Theorem 4.1 (`OFUGLB` for Self-Concordant **GLB**). *`OFUGLB` attains the following regret bound for self-concordant* **GLB** *with probability at least* $1-\delta$:

$$\mathsf{Reg}(T) \lesssim d\sqrt{\frac{g(\tau)T}{\kappa_\star(T)} \log \frac{SL_T}{d} \log \frac{R_{\dot{\mu}}ST}{d}} + d^2 R_s R_{\dot{\mu}} \sqrt{g(\tau)}\kappa(T) \log\left(1 + \frac{ST}{dg(\tau)\kappa(T)}\right), \tag{12}$$

where L_T *is as defined in Theorem 3.1 and we assume that* $\log\frac{1}{\delta} = \mathcal{O}\left(d \log \frac{SL_T}{d}\right)$.

Proof Sketch. We first emphasize that even though we have a tight CS (Theorem 3.1), naïvely combining it with existing regret analyses of logistic bandits (Abeille et al., 2021; Lee et al., 2024) *still* results in the leading term of $dS\sqrt{T/\kappa_\star(T)}$. This is because their proofs rely on a bound on $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\nabla^2 \mathcal{L}_t(\boldsymbol{\theta}_\star)}$ (Lee et al., 2024, Lemma 6) which in turn relies on the self-concordant control (Abeille et al., 2021, Lemma 8) that incurs extra factor of S.

Three key technical novelties/ingredients allow us to bypass the issues. *First,* we derive a novel self-concordance lemma that bounds the difference of $\dot{\mu}$'s with the difference of μ's times R_s; see Lemma D.3. This later leads to an implicit inequality of the form $X \leq A\sqrt{B + R_s X} + C$ (similar to Abeille et al. (2021)) , which can be solved for X to obtain the final regret bound. This does not incur any dependency on S as it avoids the self-concordance control lemma. *Second,* we introduce a novel regret decomposition to show that the UCB implicitly performs warm-up by dividing the regret into two terms: one corresponding to the timesteps in which the "warmup conditions" are satisfied and the remaining term. The second term is at most constant w.r.t. T due to the elliptical potential *count* lemma (Gales et al., 2022, Lemma 7), which was also the main argument used to avoid S dependency in Lee et al. (2024). *The final* ingredient is how we deal with the first term. We further decompose it by introducing intermediate points $\bar{\boldsymbol{\theta}}_t, \boldsymbol{\nu}_t \in \bigcup_{b \in [t,T]} \mathcal{C}_b$, where $[t,T] := \{t, t+1, \cdots, T\}$. These points lying in the union of future confidence sets, combined with the fact that the current summands satisfy the "warmup conditions" allow for the elliptical potential lemma (Abbasi-Yadkori et al., 2011, Lemma 11) to be directly applicable; see Lemma D.5. This leads to a $\mathrm{poly}(S)$-free leading term of the form $\sqrt{T/\kappa_\star(T)}$. See Appendix D for the full detailed proof. □

In Table 2, we instantiate Theorem 4.1 for various self-concordant **GLB**s. It can be seen that our `OFUGLB` attains state-of-the-art regret guarantees in all considered scenarios, either by achieving (linear) or improving upon (bounded, logistic) the known rates! Note that the instantiation for (sub-)Gaussian linear bandits is meant to be a sanity check because tighter confidence sets are available in Flynn et al. (2023) and Chowdhury et al. (2023, Appendix F).

The only works dealing with generic self-concordant **GLB**s (possibly unbounded) are Jun et al. (2017) and Janz et al. (2024). The former work incurs a regret bound scaling with $\kappa_\star(T)$ in the leading term, and the latter is interestingly a scalable, randomized exploration-based approach:

Remark 3 (Randomized exploration for **GLB**s). *Janz et al. (2024) proposed `EVILL`, a randomized exploration algorithm by linearly perturbing the regularized log-likelihood loss. It attains a regret bound of* $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T/\kappa_\star(T)})$ *omitting factors of* S*, for fixed arm-set. Regret-wise, it suffers an extra factor of* $\sqrt{d}$*, similar to other Thompson sampling-based approaches to* **GLB***s (Abeille and Lazaric, 2017; Dong et al., 2019; Kim et al., 2023; Kveton et al., 2020). An interesting question is whether the intuitions from our new CS can be used to improve Thompson sampling for* **GLB***s.*

Below, we provide more discussions on bounded **GLB** and logistic bandits; see Appendix A for some discussions on Poisson bandits as well.

Bounded GLB. The only work that applies to general bounded **GLB** is Sawarni et al. (2024), where the authors propose `RS-GLinCB` with the regret as in Table 2. Compared to our regret, they are

Table 2: Regret bounds of `OFUGLB` for various self-concordant **GLB**s. Logarithmic factors are omitted to avoid a cognitive overload. Here, we denote $\kappa_{\mathcal{X}}(T) := \max_{\boldsymbol{x} \in \cup_{t=1}^T \mathcal{X}_t} \frac{1}{\dot{\mu}(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star \rangle)}$.

GLB	Our regret bound	Prior state-of-the-art
Bounded[a]	$d\sqrt{\frac{T}{\kappa_\star(T)}} + d^2 R R_{\dot{\mu}} \kappa(T)$	$d\sqrt{\frac{T}{\kappa_\star(T)}} + d^2 R^5 S^2 \kappa_{\mathcal{X}}(T)$ (Sawarni et al., 2024, Theorem 4.2)
Logistic	$d\sqrt{\frac{T}{\kappa_\star(T)}} + d^2 \kappa(T)$	$d\sqrt{\frac{T}{\kappa_\star(T)}} + d^2 S^2 \kappa_{\mathcal{X}}(T)$ (Sawarni et al., 2024, Theorem 4.2)
Linear[b]	$\sigma d \sqrt{T}$	$\sigma d \sqrt{T}$ (Flynn et al., 2023, Lemma D.10)
Poisson	$dS\sqrt{\frac{T}{\kappa_\star(T)}} + d^2 e^{2S} \kappa(T)$	None

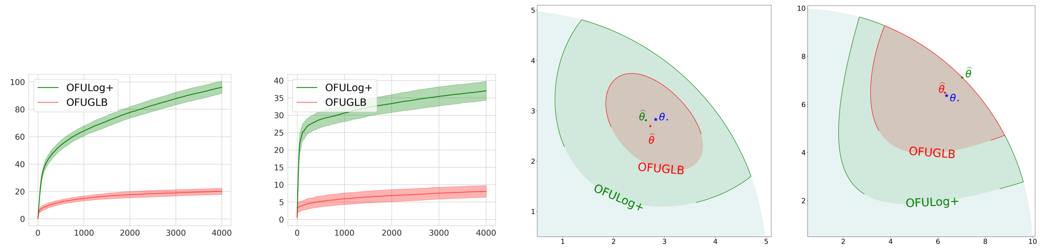
[a] $|r_t| \leq R \; a.s.$, and $g(\tau) = \mathcal{O}(1)$.
[b] We choose $c = \sigma^2 S^2$ in Lemma D.10 of Flynn et al. (2023).

slightly better as their transient term scales as $\kappa_{\mathcal{X}}(T)$ while ours scales as $\kappa(T)$. Despite this seeming gap, as `RS-GLinCB` relies on an explicit warm-up scheme, our `OFUGLB` is expected to have superior numerical performance as it avoids excessive exploration in the early phase. We will elaborate more on this issue in the later paragraph on logistic bandits. Still, `RS-GLinCB` has its own advantages in that it only requires $\Omega(\log^2 T)$ switches while we require $\Omega(T)$ switches; it is an interesting open problem whether a lazy variant of `OFUGLB` with same (or better) regret guarantee is possible.

Logistic Bandits. Although the logistic bandit is a special case of the bounded **GLB**, the number of prior works and its practical applicability to recommender systems (Li et al., 2010, 2012) deserve separate discussions. We first review the prior works on (contextual) logistic bandits. Faury et al. (2020) was the first to obtain a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T} + d^2\kappa(T))$ (up to some dependencies on S) that is κ-free in the leading term. Subsequently, a local minimax regret lower bound of $\Omega\left((d/S)\sqrt{T/\kappa_\star(T)}\right)$ was proven (Abeille et al., 2021, Theorem 2)[2], suggesting that more nonlinearity helps, and several works have focused on proposing and analyzing algorithms with matching upper bounds. One line of works (Abeille et al., 2021; Lee et al., 2024), including this work, focuses on getting a tight *convex* CS for logistic losses, which then directly gives an OFUL-type algorithm. Abeille et al. (2021) proposed a somewhat loose (in S) likelihood ratio-based CS, and their algorithm, `OFULog-r`, attain a regret bound of $\widetilde{\mathcal{O}}(dS^{5/2}\sqrt{T/\kappa_\star(T)} + R_{\mathcal{X}}(T))$. Lee et al. (2024) propose a new framework for converting an achievable online learning algorithm to a CS and use the resulting tighter CS with UCB to obtain $\widetilde{\mathcal{O}}(dS\sqrt{T/\kappa_\star(T)} + R_{\mathcal{X}}(T))$. From a computational perspective, Faury et al. (2022) proposed an online Newton step-based algorithms that attain the regret bound of $\widetilde{\mathcal{O}}(dS\sqrt{T/\kappa_\star(T)} + d^2S^6\kappa(T))$ using only $\mathcal{O}(\log t)$ computational cost *and* $\mathcal{O}(1)$ storage per iteration; the computational cost was later improved to $\mathcal{O}(1)$ in Zhang and Sugiyama (2023). Another line of works (Mason et al., 2022; Sawarni et al., 2024) proposed optimal design-based algorithms that perform an *explicit* warm-up in the early stages of the algorithms. Thanks to the explicit warmup, both attain regret with $\mathrm{poly}(S)$-free leading term, e.g., $\widetilde{\mathcal{O}}(d\sqrt{T/\kappa_\star(T)} + d^2S^2\kappa_{\mathcal{X}}(T))$ by Sawarni et al. (2024). However, the explicit warmup typically lasts for $\widetilde{\Omega}(\kappa(T))$ or $\widetilde{\Omega}(\kappa_{\mathcal{X}}(T))$ steps, and given how both scales as e^S (Faury et al., 2020), it is practically problematic.

It is known in some cases that such κ-scaling transient term can be avoided (Abeille et al., 2021, Section 4). Indeed, this discrepancy follows from the algorithm design and is shown in the *transient term* of the regret bounds. For the prior OFUL-type algorithms (Abeille et al., 2021; Lee et al., 2024), the transient term $R_{\mathcal{X}}(T)$ is defined as $R_{\mathcal{X}}(T) := \sum_{t=1}^T \mu(\langle \boldsymbol{x}_{t,\star}, \boldsymbol{\theta}_\star \rangle) \mathbb{1}[\boldsymbol{x}_t \in \mathcal{X}_-(t)]$, where $\mathcal{X}_-(t)$ is the set of *detrimental arms* with a large reward gap and little information (small conditional variance). $R_{\mathcal{X}}(T)$ is *adaptive to the arm-set geometry* and can be *completely independent* of κ for certain arm geometries (Abeille et al., 2021, Proposition 2). For the warmup-based algorithms (Faury et al., 2022; Mason et al., 2022; Sawarni et al., 2024), the transient term *always* scale with κ, which is not adaptive to the arm-set geometry.

[2] In their statement, dependency on S is ignored. By tracking their lower bound proof, one can see that it leads to an extra factor of $1/S$.



(a) $S = 5$ (b) $S = 10$ (c) $S = 5$ (d) $S = 10$

Figure 1: (a,b) Regret plots for considered algorithms. (c,d) Confidence sets at $t = 4000$ from a single run: red is from `OFUGLB` and green is from `OFULog+`.

In this context, our `OFUGLB` is the first purely optimism-based UCB algorithm (no explicit warmup) that attains a $\mathrm{poly}(S)$-free leading term in the regret for the first time. However, as our regret analysis utilizes "implicit warmup", our transient term scales with $\kappa(T)$, which *is not* adaptive to the arm-set geometry. Thus, the natural question is whether a similar, arm-set geometry adaptive transient term is attainable for logistic bandits, *while* keeping the optimal $\mathrm{poly}(S)$-free leading term. Currently, it seems that the regret decomposition used in our analysis is incompatible with the arm-set geometry-dependent analysis, and we leave to future work on obtaining both characteristics ($\mathrm{poly}(S)$-free leading term, arm-set geometry-dependent transient term) for logistic bandits and **GLB**s in general.

Remark 4 (Detrimental arms for **GLB**s.)**.** *In Abeille et al. (2021), one other key component for allowing such transient term that is adaptive to arm-set geometry is that there exists a $\mathcal{Z} \subseteq \mathbb{R}$ such that $\sup_{z \in \mathcal{Z}} \ddot{\mu}(z) \leq 0$; for logistic case ($\mu(z) = (1 + e^{-z})^{-1}$), $\mathcal{Z} = (-\infty, 0]$. For general $\mathcal{Z}$, we can define the set of detrimental arms as $\mathcal{X}_-(t) := \{\boldsymbol{x} \in \mathcal{X}_t : \langle \boldsymbol{x}, \boldsymbol{\theta}_\star \rangle \in \mathcal{Z}\}$. Of course, the scaling of $R_{\mathcal{X}}(T)$ depends on various factors, whose precise characterization for μ's beyond the logistic function is left for future work.*

To complement the improvement in our regret bounds and CS, we perform experiments on logistic bandits by comparing our `OFUGLB` to `OFULog+` (Lee et al., 2024). Following the setting of Lee et al. (2024), for `OFUGLB` and `OFULog+`, we utilize Sequential Least SQuares Programming (SLSQP) implemented in SciPy (Virtanen et al., 2020) for precise computation of the norm-constrained MLE at each time step for a fairer comparison. For the parameters, we set $T = 4000$, $d = 2$, $|\mathcal{A}| = 20$, and $\delta = 0.05$, and we average over 10 independent random trials for the regret comparison. We use $\boldsymbol{\theta}_\star = \frac{S-1}{\sqrt{d}} \mathbf{1}$ for $S \in \{5, 10\}$, and time-varying arm-set by sampling in the unit ball at random at each t. The regret curves shown in Figure 1(a) and 1(b) clearly show that `OFUGLB` numerically outperforms `OFULog+`. The confidence sets at $t = 4000$ shown in Figure 1(c) and 1(d) indicates that, indeed, our confidence set from Theorem 3.1 is much smaller than that of Lee et al. (2024), which shows the practical benefit of our novel CS.

5 Conclusion

This paper introduces a novel and *unified* likelihood ratio-based CS for generic (convex) GLMs, encompassing widely-used models such as Gaussian, Bernoulli, and Poisson. Especially for Bernoulli, this leads to the first $\mathrm{poly}(S)$-free CS, resolving an open problem posed in Lee et al. (2024). Our CS is equipped with exact constants for various scenarios, making it suitable for any practitioner to use. The proof involves leveraging key techniques from PAC-Bayes bounds along with a uniform prior/posterior, which may be of independent interest. We then propose `OFUGLB`, a generic UCB algorithm applicable to *any* **GLB**s, achieving state-of-the-art regret bounds across various instantiations (linear, logistic, GLM). The proof involves novel regret decomposition and maximally avoiding the self-concordance control lemma (Faury et al., 2020, Lemma 9), which may also be of independent interest. Notably, for logistic bandits, `OFUGLB` is the first pure-optimism-based algorithm that achieves $\mathrm{poly}(S)$-free leading term in the theoretical regret and is numerically verified to be the best-performing. This work opens up various fruitful future directions, which we relegate to Appendix B.

Acknowledgments and Disclosure of Funding

J. Lee thanks G. Neu for hosting him at a mini-workshop after AISTATS 2024 at UPF, during which many insightful discussions led to the current PAC-Bayesian proof. J. Lee also thanks A. Ramdas and T. van Erven for insightful discussions and comments during AISTATS 2024. J. Lee and S.-Y. Yun were supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) (No. RS-2022-II220311, Development of Goal-Oriented Reinforcement Learning Techniques for Contact-Rich Robotic Manipulation of Everyday Objects and No.RS-2019-II190075 Artificial Intelligence Graduate School Program (KAIST)). K.-S. Jun was supported in part by the National Science Foundation under grant CCF-2327013.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1–9. PMLR, 21–23 Apr 2012. URL https://proceedings.mlr.press/v22/abbasi-yadkori12.html.

Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165 – 5197, 2017. doi: 10.1214/17-EJS1341SI. URL https://doi.org/10.1214/17-EJS1341SI.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3691–3699. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/abeille21a.html.

Pierre Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024. ISSN 1935-8237. doi: 10.1561/2200000100. URL http://dx.doi.org/10.1561/2200000100.

Peter Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3:397–422, Nov 2002. URL https://jmlr.csail.mit.edu/papers/v3/auer02a.html.

Avrim Blum and Adam Kalai. Universal Portfolios With and Without Transaction Costs. *Machine Learning*, 35(3):193–205, Jun 1999. ISSN 1573-0565. doi: 10.1023/A:1007530728748. URL https://doi.org/10.1023/A:1007530728748.

Gavin C. Cawley, Gareth J. Janacek, and Nicola L. C. Talbot. Generalised Kernel Machines. In *2007 International Joint Conference on Neural Networks*, pages 1720–1725, 2007. doi: 10.1109/IJCNN.2007.4371217. URL https://ieeexplore.ieee.org/document/4371217.

Sayak Ray Chowdhury and Aditya Gopalan. On Kernelized Multi-armed Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/chowdhury17a.html.

Sayak Ray Chowdhury, Patrick Saux, Odalric Maillard, and Aditya Gopalan. Bregman Deviations of Generic Exponential Families. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 394–449. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/chowdhury23a.html.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30, pages 4302–4310. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds. *Journal of Machine Learning Research*, 24(372):1–61, 2023. URL http://jmlr.org/papers/v24/23-0401.html.

D. A. Darling and Herbert Robbins. Confidence Sequences for Mean, Variance, and Median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967a. doi: 10.1073/pnas.58.1.66. URL https://www.pnas.org/doi/abs/10.1073/pnas.58.1.66.

D. A. Darling and Herbert Robbins. Iterated Logarithm Inequalities. *Proceedings of the National Academy of Sciences*, 57(5):1188–1192, 1967b. doi: 10.1073/pnas.57.5.1188. URL https://www.pnas.org/doi/abs/10.1073/pnas.57.5.1188.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Provably Sample Efficient RLHF via Active Preference Optimization. *arXiv preprint arXiv:2402.10500*, 2024. URL https://arxiv.org/abs/2402.10500.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3): 1902 – 1933, 2004. doi: 10.1214/009117904000000397. URL https://doi.org/10.1214/009117904000000397.

Shi Dong, Tengyu Ma, and Benjamin Van Roy. On the Performance of Thompson Sampling on Logistic Bandits. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1158–1160. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/dong19a.html.

M. D. Donsker and S. R. S. Varadhan. Asymptotic Evaluation of Certain Markov Process Expectations for Large Time. IV. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983. doi: https://doi.org/10.1002/cpa.3160360204. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160360204.

Nicolas Emmenegger, Mojmír Mutný, and Andreas Krause. Likelihood Ratio Confidence Sets for Sequential Decision Making. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=4anryczeED.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved Optimistic Algorithms for Logistic Bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3052–3060. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/faury20a.html.

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly Efficient and Optimal Algorithms for Logistic Bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 546–580. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/faury22a.html.

Herbert Federer. *Geometric Measure Theory*. Classics in Mathematics. Springer Berlin, Heidelberg, 1996.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric Bandits: The Generalized Linear Case. In *Advances in Neural Information Processing Systems*, volume 23, pages 586–594. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Hamish Flynn, David Reeb, Melih Kandemir, and Jan Peters. Improved Algorithms for Stochastic Linear Bandits Using Tail Bounds for Martingale Mixtures. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=TXoZiUZywf.

Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic Regression: The Importance of Being Improper. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 167–208. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/foster18a.html.

Spencer B. Gales, Sunder Sethuraman, and Kwang-Sung Jun. Norm-Agnostic Linear Bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 73–91. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/gales22a.html.

Thibault Gisselbrecht, Sylvain Lamprier, and Patrick Gallinari. Policies for Contextual Bandit Problems with Count Payoffs. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 542–549, 2015. doi: 10.1109/ICTAI.2015.85. URL https://ieeexplore.ieee.org/document/7372181.

Peter D. Grünwald and Nishant A. Mehta. Fast Rates for General Unbounded Loss Functions: From ERM to Generalized Bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020. URL http://jmlr.org/papers/v21/18-488.html.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, Dec 2007. ISSN 1573-0565. doi: 10.1007/s10994-007-5016-8. URL https://doi.org/10.1007/s10994-007-5016-8.

David Janz, Shuai Liu, Alex Ayoub, and Csaba Szepesvári. Exploration via linearly perturbed loss minimisation. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 721–729. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/janz24a.html.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. A Short Note on Concentration Inequalities for Random Vectors with SubGaussian Norm. *arXiv preprint arXiv:1902.03736*, 2019. URL https://arxiv.org/abs/1902.03736.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems*, volume 30, pages 98–108. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/28dd2c7955ce926456240b2ff0100bde-Paper.pdf.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved Confidence Bounds for the Linear Logistic Model and Applications to Bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5148–5157. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jun21a.html.

Emilie Kaufmann and Wouter M. Koolen. Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021. URL http://jmlr.org/papers/v22/18-798.html.

Abbas Kazerouni and Lawrence M. Wein. Best arm identification in generalized linear bandits. *Operations Research Letters*, 49(3):365–371, 2021. ISSN 0167-6377. doi: https://doi.org/10.1016/j.orl.2021.03.011. URL https://www.sciencedirect.com/science/article/pii/S0167637721000523.

Wonyoung Kim, Kyungbok Lee, and Myunghee Cho Paik. Double Doubly Robust Thompson Sampling for Generalized Linear Contextual Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8300–8307, Jun. 2023. doi: 10.1609/aaai.v37i7.26001. URL https://ojs.aaai.org/index.php/AAAI/article/view/26001.

Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved Regret Analysis for Variance-Adaptive Linear Bandits and Horizon-Free Linear Mixture MDPs. In *Advances in Neural Information Processing Systems*, volume 35, pages 1060–1072. Curran Associates, Inc., 2022. URL https://openreview.net/forum?id=U_YPSEyN2ls.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvári, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized Exploration in Generalized Linear Bandits. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2066–2076. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/kveton20a.html.

Ricardo Lage, Ludovic Denoyer, Patrick Gallinari, and Peter Dolog. Choosing which message to publish on social networks: A contextual bandit approach. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 620–627, 2013. doi: 10.1145/2492517.2492541. URL https://ieeexplore.ieee.org/document/6785767.

Tze Leung Lai. On Confidence Sequences. *The Annals of Statistics*, 4(2):265 – 280, 1976. doi: 10.1214/aos/1176343406. URL https://doi.org/10.1214/aos/1176343406.

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

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. Improved Regret Bounds of (Multinomial) Logistic Bandits via Regret-to-Confidence-Set Conversion. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4474–4482. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/lee24c.html.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL https://doi.org/10.1145/1772690.1772758.

Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. An Unbiased Offline Evaluation of Contextual Bandit Algorithms with Generalized Linear Models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, volume 26 of *Proceedings of Machine Learning Research*, pages 19–36, Bellevue, Washington, USA, 02 Jul 2012. PMLR. URL https://proceedings.mlr.press/v26/li12a.html.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably Optimal Algorithms for Generalized Linear Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2071–2080. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/li17c.html.

Elliott H Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973. ISSN 0001-8708. doi: https://doi.org/10.1016/0001-8708(73)90011-X. URL https://www.sciencedirect.com/science/article/pii/000187087390011X.

Blake Mason, Kwang-Sung Jun, and Lalit Jain. An Experimental Design Approach for Regret Minimization in Logistic Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7736–7743, Jun. 2022. doi: 10.1609/aaai.v36i7.20741. URL https://ojs.aaai.org/index.php/AAAI/article/view/20741.

Benedict C. May, Nathan Korda, Anthony Lee, and David S. Leslie. Optimistic Bayesian Sampling in Contextual-Bandit Problems. *Journal of Machine Learning Research*, 13(67):2069–2106, 2012. URL http://jmlr.org/papers/v13/may12a.html.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 2 edition, 1989.

Hans-Georg Müller and Ulrich Stadtmüller. Generalized functional linear models. *The Annals of Statistics*, 33(2):774 – 805, 2005. doi: 10.1214/009053604000001156. URL https://doi.org/10.1214/009053604000001156.

Katta G. Murty and Santosh N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, Jun 1987. ISSN 1436-4646. doi: 10.1007/BF02592948. URL https://doi.org/10.1007/BF02592948.

Mojmír Mutný and Andreas Krause. No-regret Algorithms for Capturing Events in Poisson Point Processes. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7894–7904. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mutny21a.html.

Yuki Ohnishi and Jean Honorio. Novel Change of Measure Inequalities with Applications to PAC-Bayesian Bounds and Monte Carlo Estimation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1711–1719. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/ohnishi21a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576 – 601, 2023. doi: 10.1214/23-STS894. URL https://doi.org/10.1214/23-STS894.

Herbert Robbins. Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952. URL https://projecteuclid.org/journals/bulletin-of-the-american-mathematical-society/volume-58/issue-5/Some-aspects-of-the-sequential-design-of-experiments/bams/1183517370.full.

Herbert Robbins and David Siegmund. A Class of Stopping Rules for Testing Parametric Hypotheses. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Health*, 6(4):37–41, 1972. URL https://projecteuclid.org/proceedings/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Sixth-Berkeley-Symposium-on-Mathematical-Statistics-and/Chapter/A-class-of-stopping-rules-for-testing-parametric-hypotheses/bsmsp/1200514454.

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. Self-Concordant Analysis of Generalized Linear Bandits with Forgetting. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 658–666. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/russac21a.html.

Ayush Sawarni, Nirjhar Das, Siddharth Barman, and Gaurav Sinha. Optimal Regret with Limited Adaptivity for Generalized Linear Contextual Bandits. *arXiv preprint arXiv:2404.06831*, 2024. URL https://arxiv.org/abs/2404.06831.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of The 27th International Conference on Machine Learning*, pages 1015–1022, 2010. URL https://arxiv.org/abs/0912.3995.

William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444. URL http://www.jstor.org/stable/2332286.

Joel A. Tropp. An Introduction to Matrix Concentration Inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015. ISSN 1935-8237. doi: 10.1561/2200000048. URL http://dx.doi.org/10.1561/2200000048.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast Rates in Statistical and Online Learning. *Journal of Machine Learning Research*, 16(54): 1793–1861, 2015. URL http://jmlr.org/papers/v16/vanerven15a.html.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010. URL https://arxiv.org/abs/1011.3027.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Jean Ville. *Étude critique de la notion de collectif*. Monographies des Probabilités. Paris: Gauthier-Villars, 1939. URL http://eudml.org/doc/192893.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2. URL https://www.nature.com/articles/s41592-019-0686-2.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. doi: 10.1073/pnas.1922664117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1922664117.

Yu-Jie Zhang and Masashi Sugiyama. Online (Multinomial) Logistic Bandit: Improved Regret and Constant Computation Cost. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=ofa1U5BJVJ.

Contents

A Relations to Prior Works

CSs for Exponential Family. Lai (1976) derived the first generic CS for the exponential family based on a generalized likelihood ratio. Their CS, however, only applies to scalar-valued unknown parameters, and instantiating it often requires solving an equation with no closed-form solution (e.g., f_n and g_n in Lai (1976)). Recently, Chowdhury et al. (2023) proposed a generic CS for exponential family expressed in the local Bregman geometry induced by the log-partition function. The proof relies on the method of mixtures (de la Peña et al., 2004; Kaufmann and Koolen, 2021), which resembles our PAC-Bayesian approach that utilizes a mixture of log-likelihood functions. One drawback is that their main result (Chowdhury et al., 2023, Theorem 3) is instantiated for scalar parameters (e.g., $\mu \in [0, 1]$ for Bernoulli without observed feature vectors), and not for GLMs. While one can attempt to instantiate it to GLMs, we speculate that the resulting confidence set may not be convex since the prior itself is centered at the true parameter, unlike our choice of the prior. While we believe their second method (Chowdhury et al., 2023, Theorem 7) results in a convex set when instantiated to GLMs, the authors do not provide any computationally efficient way to evaluate the integral over the unknown parameter except for the Gaussian GLM. We mention in passing that their CS for Gaussian (Chowdhury et al., 2023, Appendix F) improves upon Abbasi-Yadkori et al. (2011) in the same manner ($\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$) that Flynn et al. (2023) and ours do.

Fast Rates in Statistical Learning. Our goal is to obtain a tight CS for $\boldsymbol{\theta}_\star$, which is quite different from that of statistical learning, which is to obtain the optimal decay rate of the ERM. Although it is not immediately clear, we believe they have a connection. To illustrate our suspicion, we recall Example 10 of Grünwald and Mehta (2020). By taking a uniform prior over a function space $\mathcal{F}$[3] and taking the posterior to be randomly sampling from ε-ball centered at $\hat{f}$, the KL term becomes the metric entropy of $\mathcal{F}$, $\log \mathcal{N}(\mathcal{F}, \varepsilon)$. Combining this with the Bernstein condition with exponent β, the ERM obtains the minimax rate of $\widetilde{\mathcal{O}}(n^{-1/(2-\beta)})$, which interpolates between the slow rate $\widetilde{\mathcal{O}}(1/\sqrt{n})$ and the fast rate $\widetilde{\mathcal{O}}(1/n)$, where n is the number of samples. This is similar to what we obtain by considering discrete uniform prior in our proof; see Appendix E for more details. We also remark that our proof of taking a prior over $\mathcal{L}_t$ resembles improper learning and the v-central condition (Foster et al., 2018; van Erven et al., 2015), which also outputs a mixture of predictors to obtain fast rates.

Poisson Bandits. Despite its potential to model various real-world problems involving count feedback, Poisson bandits have not been studied often in the literature. Gisselbrecht et al. (2015) was the first to consider contextual Poisson bandits and proposed UCB and optimistic Bayesian-based algorithms (May et al., 2012), but without any regret guarantees. To our knowledge, this is the first regret bound for the (finite-dimensional) contextual Poisson bandits without reward boundedness assumption. On a slightly related note, Mutný and Krause (2021) consider Poisson bandits with the intensity function in an RKHS. Their formulation is, however, incomparable to ours, as they consider Poisson to be a linear model in the RKHS; see their Appendix A.1 for further discussions on why this is incompatible with the log-linear formulation as in our GLM.

B Further Future Works

Here, we propose some more interesting directions. One is to extend the techniques used here to kernelized or functional GLM (Cawley et al., 2007; Müller and Stadtmüller, 2005), which would be an interesting nonlinear generalization of the linear kernel bandits (Chowdhury and Gopalan, 2017; Srinivas et al., 2010). The optimality of our obtained CS radius as well as the leading term in the regret of **GLB**, especially with respect to S, is an important question. In the era of LLMs and RLHFs, it would be interesting to see if there are any improvements in the pure exploration (best arm identification) of **GLB**s from our new CS (Jun et al., 2021; Kazerouni and Wein, 2021), which would have direct implications in sample efficient RLHF (Das et al., 2024).

[3] satisfying some regularity conditions including Lipschitzness and boundedness

C Missing Results and Proofs

C.1 Bounding L_t for Gaussian Distribution

We first recall some definitions:

Definition C.1. *A random variable $X \in \mathbb{R}$ is σ-subGaussian, if* $\mathbb{P}(\left|X - \mathbb{E}[X]\right| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$, $\forall t \in \mathbb{R}$.

Definition C.2 (Definition 3 of Jin et al. (2019))**.** *A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is σ-norm-subGaussian, if* $\mathbb{P}(\left\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\right\|_2 \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$, $\forall t \in \mathbb{R}$.

Here is the full statement:

Proposition C.1. *Suppose the GLM is σ-subGaussian. Then, for any $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists t \geq 1 : L_t > \frac{2}{g(\tau)}\left(R_{\dot{\mu}} S(t-1) + 2\pi\sigma\sqrt{(t-1)\log\frac{\pi^2 d t^2}{3\delta}}\right)\right) \leq \delta. \tag{13}$$

Proof. Here, as $\max_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\theta}\in\Theta} |\dot{\mu}(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle)| \leq R_{\dot{\mu}}$, we have that

$$\begin{aligned}
L_t &= \frac{1}{g(\tau)} \max_{\boldsymbol{\theta}\in\Theta} \left\| \sum_{s=1}^{t-1} (r_s - \mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta}\rangle))\boldsymbol{x}_s \right\|_2 \\
&\leq \frac{1}{g(\tau)} \max_{\boldsymbol{\theta}\in\Theta} \left\| \sum_{s=1}^{t-1} (\mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta}\rangle) - \mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star\rangle))\boldsymbol{x}_s \right\|_2 + \frac{1}{g(\tau)} \left\| \sum_{s=1}^{t-1} \underbrace{(r_s - \mu(\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star\rangle))\boldsymbol{x}_s}_{\triangleq \boldsymbol{y}_s} \right\|_2 \\
&\leq \frac{2R_{\dot{\mu}} S(t-1)}{g(\tau)} + \frac{1}{g(\tau)} \left\| \sum_{s=1}^{t-1} \boldsymbol{y}_s \right\|_2 .
\end{aligned}$$

We now utilize subGaussian concentrations from Jin et al. (2019). First note that $\boldsymbol{y}_s$ is a martingale difference sequence adapted to Σ_s and is norm-subGaussian with (conditional) variance σ^2 be given. Then, by Corollary 7 of Jin et al. (2019), we have that

$$\mathbb{P}\left(\left\| \sum_{s=1}^{t-1} \boldsymbol{y}_s \right\|_2 \leq 4\pi\sigma\sqrt{(t-1)\log\frac{2d}{\delta}} \right) \geq 1 - \delta, \quad \forall t \geq 1. \tag{14}$$

The exact constant 4π is not available in Jin et al. (2019), as all the constants are hidden under c. This is not useful, especially for practitioners wanting to use the concentration directly. Thus, we tracked the constant from their Corollary 7, the details of which we provide in Lemma C.1.

We then conclude by parametrizing δ as δ/t^2, applying union bound over $t \geq 1$, and using the Basel sum. □

Lemma C.1 (Lemma 2 of Jin et al. (2019); originally Lemma 5.5 of Vershynin (2010))**.** *For any σ-norm-subGaussian random vector $\boldsymbol{X}$, we have that* $\sup_{p\in\mathbb{N}} p^{-1/2}\left(\mathbb{E}[\|\boldsymbol{X}\|^p]\right)^{1/p} \leq \sqrt{\pi}\sigma$.

Proof. This follows from brute-force computation. First, we have that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{X}\|^p] = \int_0^\infty \mathbb{P}[\|\boldsymbol{X}\|^p \geq t]dt = p\int_0^\infty \mathbb{P}[\|\boldsymbol{X}\| \geq t] t^{p-1} dt \leq 2p\int_0^\infty t^{p-1}\exp\left(-\frac{t^2}{2\sigma^2}\right)dt \\
= 2^{\frac{p-1}{2}}\sigma^p p\Gamma\left(\frac{p}{2}\right).
\end{aligned}$$

Then, for any $p \in \mathbb{N}$,

$$p^{-1/2}(\mathbb{E}[\|\boldsymbol{X}\|^p])^{1/p} = \sigma 2^{\frac{p-1}{2p}} p^{\frac{1}{p}-\frac{1}{2}} \left(\sqrt{\pi}\frac{(p-2)!!}{2^{\frac{p-1}{2}}}\right)^{1/p} = \sigma \underbrace{p^{\frac{1}{p}-\frac{1}{2}} \left(\sqrt{\pi}(p-2)!!\right)^{1/p}}_{\triangleq f(p)}.$$

Using WolframAlpha, we can conclude that $f(p)$ is decreasing, and we conclude by noting that $f(1) = \sqrt{\pi}$. ☐

C.2 Bounding L_t for Poisson Distribution

We have the following result for Poisson, which may be of independent interest (to our knowledge, this is the first explicit martingale concentration for Poisson):

Proposition C.2. *For the Poisson distribution, we have that for any $\delta \in (0, 1)$: when $S > 1$,*

$$\mathbb{P}\left(L_t \leq C(S)(t-1) + \frac{2}{1-2e^{-S}} \log \frac{\pi^2(d+1)t^2}{3\delta}\right) \geq 1 - \delta, \quad \forall t \geq 1, \tag{15}$$

where $C(S) := \frac{1}{4}(1-2e^{-S})(e^S + 2S + 2\log\frac{2(1-2e^{-S})}{e}) + 2Se^S$. When $S \leq 1$,

$$\mathbb{P}\left(L_t \leq \tilde{C}(S)(t-1) + 4\log \frac{\pi^2(d+1)t^2}{3\delta}\right) \geq 1 - \delta, \quad \forall t \geq 1, \tag{16}$$

where $\tilde{C}(S) := \frac{1}{16}\left(e^S + 4S + 4\log(8 + 2e^S)\right) + 2Se^S$.

Proof. Proceeding similarly as in the previous subsection, we first have that

$$L_t \leq 2Se^S(t-1) + \left\|\sum_{s=1}^{t-1} \boldsymbol{y}_s\right\|_2, \tag{17}$$

where $\boldsymbol{y}_s = (r_s - e^{\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star\rangle})\boldsymbol{x}_s$ is the martingale difference sequence satisfying $\mathbb{E}[\boldsymbol{y}_s|\Sigma_s] = \boldsymbol{0}$ as $r_s|\Sigma_s \sim \mathrm{Poi}(\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star\rangle)$.

We now modify the proof of Corollary 7 of Jin et al. (2019) (which is based upon the celebrated Chernoff-Cramér method) for the Poisson martingale vectors, details of which we provide here for completeness.

First, we consider the following MGF bound of the Poisson distribution whose proof is deferred to the end of this subsection:

Lemma C.2. *Suppose that the random vector $\boldsymbol{y}$ is of the form $\boldsymbol{y} = (r - \lambda)\boldsymbol{x}$ for some fixed $\boldsymbol{x} \in \mathcal{B}^d(1)$, $r \sim \mathrm{Poi}(\lambda)$, and $\lambda > 0$. Then, for the Hermitian dilation (Tropp, 2015, Definition 2.1.5) of $\boldsymbol{y}$, $\boldsymbol{Y} := \begin{bmatrix} 0 & \boldsymbol{y}^\top \\ \boldsymbol{y} & \boldsymbol{0} \end{bmatrix}$, we have that $\mathbb{E}e^{\theta \boldsymbol{Y}} \preceq \exp\left(F(\theta, \lambda)\right) \boldsymbol{I}_{d+1}$ for $|\theta| < \frac{1}{2}$, where $F(\theta, \lambda) := \lambda|\theta| + \log(2|\theta|) + \log\left(\frac{e^{-\frac{\lambda}{2}}}{\frac{1}{2}-|\theta|} + \lambda\right)$.*

We also recall the Lieb's trace inequality:

Theorem C.3 (Theorem 6 of Lieb (1973))**.** *Let $\boldsymbol{A}$ be a fixed symmetric matrix, and let $\boldsymbol{Y}$ be a random symmetric matrix. Then,*

$$\mathbb{E}\,\mathrm{tr}(\exp(\boldsymbol{A} + \boldsymbol{Y})) \leq \mathrm{tr}\exp(\boldsymbol{A} + \log \mathbb{E}e^{\boldsymbol{Y}}) \tag{18}$$

Now let $0 < \theta < \frac{1}{2}$ be fixed, and let us denote $\lambda_s := e^{\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star\rangle}$ and $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot|\Sigma_s]$ for $s \leq t-1$. We start by noting that

$$\mathbb{E}\,\mathrm{tr}\exp\left(-\theta^2 \boldsymbol{I}_{d+1}\sum_{s=1}^{t-1} F(\theta, \lambda_s) + \theta\sum_{s=1}^{t-1}\boldsymbol{Y}_s\right)$$

$$
\begin{aligned}
&= \mathbb{E}\left[\mathbb{E}_{t-1}\left[\operatorname{tr}\exp\left(-\theta^2 \boldsymbol{I}_{d+1}\sum_{s=1}^{t-1} F(\theta,\lambda_s) + \theta\sum_{s=1}^{t-1}\boldsymbol{Y}_s\right)\right]\right] \\
&\leq \mathbb{E}\left[\operatorname{tr}\exp\left(-\theta^2 \boldsymbol{I}_{d+1}\sum_{s=1}^{t-1} F(\theta,\lambda_s) + \theta\sum_{s=1}^{t-2}\boldsymbol{Y}_s + \log\mathbb{E}_{t-1}\left[e^{\theta\boldsymbol{Y}_{t-1}}\right]\right)\right] && \text{(Theorem C.3)} \\
&\leq \mathbb{E}\left[\operatorname{tr}\exp\left(-\theta^2 \boldsymbol{I}_{d+1}\sum_{s=1}^{t-1} F(\theta,\lambda_s) + \theta\sum_{s=1}^{t-2}\boldsymbol{Y}_s + F(\theta,\lambda_{t-1})\boldsymbol{I}_{d+1}\right)\right] \\
& && \text{(Lemma C.2, } \boldsymbol{A}\preceq\boldsymbol{B}\Rightarrow e^{\boldsymbol{C}+\boldsymbol{A}}\preceq e^{\boldsymbol{C}+\boldsymbol{B}}\text{)} \\
&\leq \mathbb{E}\left[\operatorname{tr}\exp\left(-\theta^2 \boldsymbol{I}_{d+1}\sum_{s=1}^{t-2} F(\theta,\lambda_s) + \theta\sum_{s=1}^{t-2}\boldsymbol{Y}_s\right)\right] \\
&\leq \cdots \leq \operatorname{tr}\exp(0\boldsymbol{I}_{d+1}) = d+1.
\end{aligned}
$$

Thus, for any $\rho \geq 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{s=1}^{t-1}\boldsymbol{y}_s\right\| \geq \theta\sum_{s=1}^{t-1}F(\theta,\lambda_s) + \frac{\rho}{\theta}\right) \\
&= \mathbb{P}\left(\left\|\sum_{s=1}^{t-1}\boldsymbol{Y}_s\right\| \geq \theta\sum_{s=1}^{t-1}F(\theta,\lambda_s) + \frac{\rho}{\theta}\right) \\
& && \left(\textstyle\sum_s \boldsymbol{Y}_s \text{ is a rank-2 matrix with eigenvalues } \pm\left\|\sum_s \boldsymbol{y}_s\right\|_2\right) \\
&= 2\mathbb{P}\left(\lambda_{\max}\left(\sum_{s=1}^{t-1}\boldsymbol{Y}_s\right) \geq \theta\sum_{s=1}^{t-1}F(\theta,\lambda_s) + \frac{\rho}{\theta}\right) && (\boldsymbol{Y}_s\text{'s are symmetric}) \\
&= 2\mathbb{P}\left(\lambda_{\max}\left(\exp\left(\theta\sum_{s=1}^{t-1}\boldsymbol{Y}_s\right)\right) \geq \exp\left(\theta^2\sum_{s=1}^{t-1}F(\theta,\lambda_s) + \rho\right)\right) \\
&\leq 2\mathbb{P}\left(\operatorname{tr}\exp\left(\theta\sum_{s=1}^{t-1}\boldsymbol{Y}_s\right) \geq \exp\left(\theta^2\sum_{s=1}^{t-1}F(\theta,\lambda_s) + \rho\right)\right) \\
&\leq 2e^{-\rho}\mathbb{E}\operatorname{tr}\exp\left(-\theta^2\sum_{s=1}^{t-1}F(\theta,\lambda_s) + \theta\sum_{s=1}^{t-1}\boldsymbol{Y}_s\right) && \text{(Markov's inequality)} \\
&\leq 2(d+1)e^{-\rho}. && \text{(Lemma C.2)}
\end{aligned}
$$

Finally, by reparametrizing, we have that for any $\delta \in (0,1)$,

$$
\mathbb{P}\left(\left\|\sum_{s=1}^{t-1}\boldsymbol{y}_s\right\| \geq \inf_{\theta\in(0,1/2)}\left\{\theta\sum_{s=1}^{t-1}F(\theta,\lambda_s) + \frac{1}{\theta}\log\frac{2d}{\delta}\right\}\right) \leq \delta, \tag{19}
$$

where we recall that $F(\theta,\lambda) = \lambda\theta + \log(2\theta) + \log\left(\frac{e^{-\frac{\lambda}{2}}}{\frac{1}{2}-\theta} + \lambda\right)$ for $\theta > 0$.

First, when $S > 1$, let us choose $\theta = \frac{1}{2} - e^{-S}$, which is guaranteed to be positive. Noting that $\lambda_s = e^{\langle \boldsymbol{x}_s, \boldsymbol{\theta}_\star \rangle} \leq e^S$, we have

$$F\left(\frac{1}{2} - e^{-S}, \lambda_s\right) \leq e^S \left(\frac{1}{2} - e^{-S}\right) + \log(1 - 2e^{-S}) + \log(2e^S) = \frac{1}{2}e^S + S + \log\frac{2(1-2e^{-S})}{e}.$$

Thus, the RHS of Eqn. (19)

$$\frac{(1-2e^{-S})(e^S + 2S + 2\log\frac{2(1-2e^{-S})}{e})}{4}(t-1) + \frac{2}{1-2e^{-S}}\log\frac{2(d+1)}{\delta}. \tag{20}$$

For the case $S \leq 1$, choosing $\theta = \frac{1}{4}$, the RHS becomes

$$\frac{e^S + 4S + 4\log(8+2e^S)}{16}(t-1) + 4\log\frac{2(d+1)}{\delta}. \tag{21}$$

Finally, we conclude by parametrizing δ as δ/t^2, applying union bound over $t \geq 1$, and using the Basel sum. □

Proof of Lemma C.2. We first have that

$$\mathbb{E}e^{\theta \boldsymbol{Y}} \overset{(*)}{=} \boldsymbol{I}_{d+1} + \sum_{p=1}^{\infty} \frac{\theta^p \mathbb{E}\boldsymbol{Y}^{2p}}{(2p)!} \preceq \boldsymbol{I}_{d+1} + \sum_{p=1}^{\infty} \frac{\theta^{2p}\mathbb{E}\left\|\boldsymbol{y}\right\|^{2p}}{(2p)!}\boldsymbol{I}_{d+1} = \mathbb{E}\left[\frac{e^{\theta\|\boldsymbol{y}\|} + e^{-\theta\|\boldsymbol{y}\|}}{2}\right]\boldsymbol{I}_{d+1} \preceq \mathbb{E}\left[e^{|\theta||r-\lambda|}\right]\boldsymbol{I}_{d+1},$$

where $(*)$ follows from the observation that $\mathbb{E}\boldsymbol{Y}^{2p+1} = \boldsymbol{0}$. We now recall a well-known concentration for Poisson distribution (taken from a note by C. Canonne):

Lemma C.3. $\mathbb{P}(|r-y| \geq x) \leq 2e^{-\frac{x^2}{2(\lambda+x)}}$.

Then, we have that

$$\begin{aligned}
\mathbb{E}[e^{|\theta||r-\lambda|}] &= \int_0^\infty \mathbb{P}(e^{|\theta||r-\lambda|} \geq k)dk && (dk \text{ is the Lebesgue measure}) \\
&\leq 1 + \int_1^\infty \mathbb{P}(e^{|\theta||r-\lambda|} \geq k)dk \\
&\leq 2\int_1^\infty e^{-\frac{(\log k/|\theta|)^2}{2(\lambda+\log k/|\theta|)}}dk && (\text{Lemma C.3}) \\
&= 2|\theta|\int_0^\infty e^{-\frac{u^2}{2(\lambda+u)}+|\theta|u}du \\
&= 2|\theta|\left\{\int_\lambda^\infty e^{-\frac{u^2}{2(\lambda+u)}+|\theta|u}du + \int_0^\lambda e^{-\frac{u^2}{2(\lambda+u)}+|\theta|u}du\right\} \\
&\leq 2|\theta|\left\{\int_\lambda^\infty e^{-(\frac{1}{2}-|\theta|)u}du + \lambda e^{|\theta|\lambda}\right\} && \left(\tfrac{u^2}{2(\lambda+u)} \geq \tfrac{1}{2}u \text{ for } u \geq \lambda\right) \\
&\leq 2|\theta|\left(\frac{1}{\frac{1}{2}-|\theta|}e^{-(\frac{1}{2}-|\theta|)\lambda} + \lambda e^{|\theta|\lambda}\right) \\
&= \exp\left(F(\theta,\lambda) \triangleq \lambda|\theta| + \log(2|\theta|) + \log\left(\frac{e^{-\frac{\lambda}{2}}}{\frac{1}{2}-|\theta|} + \lambda\right)\right).
\end{aligned}$$

□

C.3 Proof of Theorem 3.2 – Ellipsoidal Confidence Sequence

First, similarly to prior works on logistic bandits (Abeille et al., 2021; Lee et al., 2024), let us define the following quantities:

$$\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}, \boldsymbol{\nu}) := \frac{1}{g(\tau)}\sum_{s=1}^{t-1}\tilde{\alpha}_s(\boldsymbol{\theta}, \boldsymbol{\nu})\boldsymbol{x}_s\boldsymbol{x}_s^\top, \ \tilde{\alpha}_s(\boldsymbol{\theta}, \boldsymbol{\nu}) := \int_0^1 (1-v)\dot{\mu}\left(\langle \boldsymbol{x}_s, \boldsymbol{\theta} + v(\boldsymbol{\nu} - \boldsymbol{\theta})\rangle\right)dv.$$

(We will later come back to these quantities in the regret analysis.)

Then, by Taylor's theorem with integral remainder, we have that for any $\lambda \geq 0$ to be chosen later,

$$\begin{aligned}
\beta_t(\delta)^2 \geq \mathcal{L}_t(\boldsymbol{\theta}) - \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t) &= \underbrace{\langle \nabla \mathcal{L}_t(\widehat{\boldsymbol{\theta}}_t), \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t \rangle}_{=0} + \left\| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t \right\|^2_{\widetilde{\boldsymbol{G}}_t(\widehat{\boldsymbol{\theta}}_t, \boldsymbol{\theta})} \\
&= \left\| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t \right\|^2_{\widetilde{\boldsymbol{G}}_t(\widehat{\boldsymbol{\theta}}_t, \boldsymbol{\theta}) + \lambda \boldsymbol{I}_d} - \lambda \left\| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t \right\|_2^2 \\
&\geq \left\| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t \right\|^2_{\widetilde{\boldsymbol{G}}_t(\widehat{\boldsymbol{\theta}}_t, \boldsymbol{\theta}) + \lambda \boldsymbol{I}_d} - 4S^2 \lambda.
\end{aligned}$$

We conclude by choosing $\lambda = \frac{1}{4S^2}$ and the self-concordance control for $\widetilde{\boldsymbol{G}}$ (Abeille et al., 2021, Lemma 8), which we recall here:

Lemma C.4 (A slight extension of Lemma 8 of Abeille et al. (2021))**.** *Let μ be increasing ($\dot{\mu} \geq 0$, which is basically Assumption 3) and self-concordant with constant R_s (as in Assumption 4). Let $\mathcal{Z} \subset \mathbb{R}$ be bounded. Then, the following holds for any $z_1, z_2 \in \mathcal{Z}$:*

$$\int_0^1 (1 - v) \dot{\mu}(z_1 + v(z_2 - z_1)) dv \geq \frac{\dot{\mu}(z_1)}{2 + R_s |z_1 - z_2|}.$$

This then implies that $\widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}, \boldsymbol{\nu}) \succeq \frac{1}{2 + 2SR_s} \nabla^2 \mathcal{L}_t(\boldsymbol{\theta})$.

D Proof of Theorem 4.1 – Regret Bound of `OFUGLB`

D.1 Supporting Lemmas

Before diving into the proof, we recall or prove some important supporting lemmas that we will be using throughout the proof.

First, we recall the elliptical potential arguments:

Lemma D.1 (Elliptical Potential Count Lemma; EPCL[4]). *For $X, L > 0$, let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T \in \mathcal{B}^d(X)$ be a sequence of vectors, $\boldsymbol{V}_t := \lambda \boldsymbol{I} + \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal$, and let us define the following: $\mathcal{H}_T := \left\{ t \in [T] : \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}^2 > L \right\}$. Then, we have that*

$$|\mathcal{H}_T| \leq \frac{2d}{\log(1+L^2)} \log\left(1 + \frac{X^2}{\lambda \log(1+L^2)}\right). \tag{22}$$

Lemma D.2 (Elliptical Potential Lemma; EPL[5]). *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T \in \mathcal{B}^d(X)$ be a sequence of vectors and $\boldsymbol{V}_t := \lambda \boldsymbol{I} + \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal$. Then, we have that*

$$\sum_{t=1}^T \min\left\{1, \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}^2\right\} \leq 2d \log\left(1 + \frac{X^2 T}{d\lambda}\right). \tag{23}$$

We have the following self-concordance lemma that will be frequently used throughout the proof:

Lemma D.3. *For $\boldsymbol{\theta}, \boldsymbol{\nu} \in \mathbb{R}^d$, $|\dot{\mu}_t(\boldsymbol{\theta}) - \dot{\mu}_t(\boldsymbol{\nu})| \leq R_s |\mu_t(\boldsymbol{\theta}) - \mu_t(\boldsymbol{\nu})|$*

Proof. This follows from direct computation:

$$\begin{aligned}
|\dot{\mu}_t(\boldsymbol{\theta}) - \dot{\mu}_t(\boldsymbol{\nu})| &= \left| \langle \boldsymbol{x}_t, \boldsymbol{\theta} - \boldsymbol{\nu} \rangle \int_0^1 \ddot{\mu}_t(\boldsymbol{\nu} + v(\boldsymbol{\theta} - \boldsymbol{\nu}) dv \right| \\
&\leq \left| \langle \boldsymbol{x}_t, \boldsymbol{\theta} - \boldsymbol{\nu} \rangle \right| \int_0^1 \left| \ddot{\mu}_t(\boldsymbol{\nu} + v(\boldsymbol{\theta} - \boldsymbol{\nu}) \right| dv \\
&\leq R_s \left| \langle \boldsymbol{x}_t, \boldsymbol{\theta} - \boldsymbol{\nu} \rangle \right| \int_0^1 \left| \dot{\mu}_t(\boldsymbol{\nu} + v(\boldsymbol{\theta} - \boldsymbol{\nu}) \right| dv & \text{(Assumption 4)} \\
&= R_s \left| \langle \boldsymbol{x}_t, \boldsymbol{\theta} - \boldsymbol{\nu} \rangle \int_0^1 \dot{\mu}_t(\boldsymbol{\nu} + v(\boldsymbol{\theta} - \boldsymbol{\nu}) dv \right| \\
& & (m \text{ is convex, and thus } \dot{\mu} = m'' \geq 0) \\
&= R_s \left| \mu_t(\boldsymbol{\theta}) - \mu_t(\boldsymbol{\nu}) \right|.
\end{aligned}$$

□

This later leads to an implicit inequality of the form $X \leq A\sqrt{B + R_s X} + C$, leading to the final regret bound. We also remark that this self-concordant result is distinct from the original self-concordance control lemma (Faury et al., 2020, Lemma 9) and does not incur any dependency on S.

Throughout the proof, we denote $\mu_t(\cdot) := \mu(\langle \boldsymbol{x}_t, \cdot \rangle)$ and $[a, b] := \{a, a+1, \cdots, b\}$ for two integers $a \leq b$. We recall the following quantities:

$$R_{\mu,\star} := \max_{\boldsymbol{x} \in X} |\mu(\langle \boldsymbol{x}, \boldsymbol{\theta}_\star \rangle)|, \quad R_{\dot{\mu}} := \max_{\boldsymbol{x} \in X, \boldsymbol{\theta} \in \Theta} \dot{\mu}(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle). \tag{24}$$

We now define the following crucial quantities: for $\lambda > 0$ to be chosen later,

$$\bar{\boldsymbol{\theta}}_t := \mathop{\arg\min}_{\boldsymbol{\theta} \in \bigcup_{b \in [t,T]} \mathcal{C}_b} \dot{\mu}_t(\boldsymbol{\theta}), \quad (b(t), \boldsymbol{\nu}_t) := \mathop{\arg\max}_{b \in [t,T], \boldsymbol{\theta} \in \mathcal{C}_b} \left| \mu_t(\boldsymbol{\theta}) - \mu_t(\widehat{\boldsymbol{\theta}}_b) \right|, \tag{25}$$

[4]This is a generalization of Exercise 19.3 of Lattimore and Szepesvári (2020), presented (in parallel) at Lemma 7 of Gales et al. (2022) and Lemma 4 of Kim et al. (2022).

[5]Lemma 11 of Abbasi-Yadkori et al. (2011).

$$\bar{\boldsymbol{H}}_t := 2g(\tau)\lambda \boldsymbol{I} + \sum_{s=1}^{t-1} \dot{\mu}_s(\bar{\boldsymbol{\theta}}_s)\boldsymbol{x}_s\boldsymbol{x}_s^\top, \quad \boldsymbol{V}_t := 2g(\tau)\kappa(T)\lambda \boldsymbol{I} + \sum_{s=1}^{t-1} \boldsymbol{x}_s\boldsymbol{x}_s^\top, \tag{26}$$

and

$$\tilde{\alpha}_t(\boldsymbol{\theta}, \boldsymbol{\nu}) := \int_0^1 (1-v)\dot{\mu}_t\left(\boldsymbol{\theta} + v(\boldsymbol{\nu} - \boldsymbol{\theta})\right) dv, \quad \widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}, \boldsymbol{\nu}) := \lambda \boldsymbol{I} + \frac{1}{g(\tau)} \sum_{s=1}^{t-1} \tilde{\alpha}_s(\boldsymbol{\theta}, \boldsymbol{\nu})\boldsymbol{x}_s\boldsymbol{x}_s^\top. \tag{27}$$

These points in the union of future confidence sets, combined with the "warmup conditions" allow for the elliptical potential lemma (Lemma D.2) to be directly applicable, avoiding dependencies on $\mathrm{poly}(S)$ and κ in the leading term. Also, note that $\bar{\boldsymbol{\theta}}_s$ bears some resemblance to additional linear constraints introduced in `Logistic-UCB-2` of Faury et al. (2020).

This is formalized in the following set of properties:

Lemma D.4. *For any $\boldsymbol{\theta}, \boldsymbol{\nu} \in \mathbb{R}^d$, $\frac{1}{2g(\tau)\kappa(T)} \leq \tilde{\alpha}_t(\boldsymbol{\theta}, \boldsymbol{\nu}) \leq \frac{R_{\dot{\mu}}}{2}$, and thus, $\frac{1}{2g(\tau)\kappa(T)} \boldsymbol{V}_t \preceq \widetilde{\boldsymbol{G}}_t(\boldsymbol{\theta}, \boldsymbol{\nu})$.*

Proof. Follows from straightforward computation. □

In the following two lemmas, $b(t)$ is as defined in Eqn. (25).

Lemma D.5. $\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t) \succeq \frac{1}{2g(\tau)} \bar{\boldsymbol{H}}_t$.

Proof. For each $s \leq b(t)$,

$$\tilde{\alpha}_s(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t) = \int_0^1 (1-v)\dot{\mu}_s\left(\widehat{\boldsymbol{\theta}}_{b(t)} + v(\boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)})\right) dv \overset{(*)}{\geq} \dot{\mu}_s(\bar{\boldsymbol{\theta}}_s) \int_0^1 (1-v)dv = \frac{1}{2}\dot{\mu}_s(\bar{\boldsymbol{\theta}}_s),$$

where $(*)$ follows from the observations that $\boldsymbol{\nu}_t, \widehat{\boldsymbol{\theta}}_{b(t)} \in \mathcal{C}_{b(t)}$ and $\mathcal{C}_{b(t)}$ is convex. We then conclude by noting that $b(t) \geq t$, and thus $\bar{\boldsymbol{H}}_{b(t)} \succeq \bar{\boldsymbol{H}}_t$. □

Lemma D.6. *For any $t \geq 1$ and $\boldsymbol{\theta}, \boldsymbol{\nu} \in \mathcal{C}_{b(t)}$, we have the following:*

(i) $\left\|\boldsymbol{\nu} - \widehat{\boldsymbol{\theta}}_{b(t)}\right\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu})} \leq \sqrt{4\lambda S^2 + \beta_T(\delta)^2}$,

(ii) $\left|\mu_t(\boldsymbol{\nu}) - \mu_t(\boldsymbol{\theta})\right| \leq 2R_{\dot{\mu}}\sqrt{2\left(4\lambda S^2 + \beta_T(\delta)^2\right)\kappa(T)}\, \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}$.

Proof. *(i)* follows from Taylor's theorem with integral remainder and our definition of $\mathcal{C}_{b(t)}$:

$$\begin{aligned}
\beta_T(\delta)^2 \geq \mathcal{L}_{b(t)}(\boldsymbol{\nu}) - \mathcal{L}_{b(t)}(\widehat{\boldsymbol{\theta}}_t) &= \underbrace{\langle \nabla \mathcal{L}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}), \boldsymbol{\nu} - \widehat{\boldsymbol{\theta}}_{b(t)} \rangle}_{=0} + \left\|\boldsymbol{\nu} - \widehat{\boldsymbol{\theta}}_{b(t)}\right\|^2_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}) - \lambda \boldsymbol{I}} \\
&\geq \left\|\boldsymbol{\nu} - \widehat{\boldsymbol{\theta}}_{b(t)}\right\|^2_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu})} - 4\lambda S^2.
\end{aligned}$$

(ii) follows from *(i)* and similar arguments:

$$\begin{aligned}
\left|\mu_t(\boldsymbol{\nu}) - \mu_t(\boldsymbol{\theta})\right| &= \left|\langle \boldsymbol{x}_t, \boldsymbol{\nu} - \boldsymbol{\theta}\rangle \int_0^1 \dot{\mu}_t(\boldsymbol{\theta} + v(\boldsymbol{\nu} - \boldsymbol{\theta}))dv\right| \\
&\leq R_{\dot{\mu}} \left\{ \left\|\boldsymbol{\nu} - \widehat{\boldsymbol{\theta}}_{b(t)}\right\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\theta})} + \left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{b(t)}\right\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\theta})} \right\} \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\theta})^{-1}} \\
&\qquad \text{(Cauchy-Schwartz \& triangle inequalities)} \\
&\leq 2R_{\dot{\mu}}\sqrt{2\left(4\lambda S^2 + \beta_T(\delta)^2\right)\kappa(T)}\, \|\boldsymbol{x}_t\|_{\boldsymbol{V}_{b(t)}^{-1}} \qquad \text{((i), Lemma D.4)} \\
&\leq 2R_{\dot{\mu}}\sqrt{2\left(4\lambda S^2 + \beta_T(\delta)^2\right)\kappa(T)}\, \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}. \qquad (b(t) \geq t)
\end{aligned}$$

□

D.2 Proof of Theorem 4.1

Throughout, let us assume that the event $\{\forall t \geq 1,\ \boldsymbol{\theta}_\star \in \mathcal{C}_t\}$ holds, which is with probability at least $1-\delta$ by Theorem 3.1.

Define the set of timesteps satisfying the "warmup conditions":

$$\mathcal{I}_T := \left\{ t \in [T] : \left(\left\| \sqrt{\dot{\mu}_t(\bar{\boldsymbol{\theta}}_t)} \boldsymbol{x}_t \right\|_{\bar{\boldsymbol{H}}_t^{-1}} \geq 1 \right) \vee \left(\|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}} \geq 1 \right) \right\} \subseteq [T]. \tag{28}$$

First, we have

$$\begin{aligned}
&\mathsf{Reg}(T) \\
&= \sum_{t \in \mathcal{I}_T} \{\mu(\langle \boldsymbol{x}_{t,\star}, \boldsymbol{\theta}_\star \rangle) - \mu(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star \rangle)\} + \underbrace{\sum_{t \notin \mathcal{I}_T} \{\mu(\langle \boldsymbol{x}_{t,\star}, \boldsymbol{\theta}_\star \rangle) - \mu(\langle \boldsymbol{x}_t, \boldsymbol{\theta}_\star \rangle)\}}_{\triangleq \mathsf{Reg}_{\mathcal{I}}(T)} \\
&\leq 2R_{\mu,\star}|\mathcal{I}_T| + \mathsf{Reg}_{\mathcal{I}}(T) \\
&\leq 2R_{\mu,\star} \sum_{t \in [T]} \mathbb{1}\left[\left\| \sqrt{\dot{\mu}_t(\bar{\boldsymbol{\theta}}_t)} \boldsymbol{x}_t \right\|_{\bar{\boldsymbol{H}}_t^{-1}} \geq 1 \right] + 2R_{\mu,\star} \sum_{t \in [T]} \mathbb{1}\left[\|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}} \geq 1 \right] + \mathsf{Reg}_{\mathcal{I}}(T) \\
&\hspace{25em} \text{(definition of } \mathcal{I}_T\text{)} \\
&\leq \frac{4dR_{\mu,\star}}{\log 2} \left\{ \log\left(1 + \frac{R_{\dot{\mu}}}{2\lambda g(\tau) \log 2}\right) + \log\left(1 + \frac{1}{2\kappa(T)\lambda g(\tau) \log 2}\right) \right\} + \mathsf{Reg}_{\mathcal{I}}(T). \\
&\hspace{25em} \text{(EPCL and Lemma D.4)}
\end{aligned}$$

We now focus on bounding the last term:

$$\begin{aligned}
\mathsf{Reg}_{\mathcal{I}}(T) &= \sum_{t \notin \mathcal{I}_T} \left\{ \mu_{t,\star}(\boldsymbol{\theta}_\star) - \mu_t(\widehat{\boldsymbol{\theta}}_t) \right\} + \sum_{t \notin \mathcal{I}_T} \left\{ \mu_t(\widehat{\boldsymbol{\theta}}_t) - \mu_t(\boldsymbol{\theta}_\star) \right\} \\
&\hspace{8em} (\mu_t(\cdot) := \mu(\langle \boldsymbol{x}_t, \cdot \rangle),\ \mu_{t,\star}(\cdot) := \mu(\langle \boldsymbol{x}_{t,\star}, \cdot \rangle)) \\
&\leq \sum_{t \notin \mathcal{I}_T} \left\{ \mu_t(\boldsymbol{\theta}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_t) \right\} + \sum_{t \notin \mathcal{I}_T} \left\{ \mu_t(\widehat{\boldsymbol{\theta}}_t) - \mu_t(\boldsymbol{\theta}_\star) \right\} \\
&\hspace{8em} \text{(optimism -- line 7 of Algorithm 1)} \\
&\leq 2 \sum_{t \notin \mathcal{I}_T} \max_{b \in [t,T]} \max_{\boldsymbol{\theta} \in \mathcal{C}_b} \left| \mu_t(\boldsymbol{\theta}) - \mu_t(\widehat{\boldsymbol{\theta}}_b) \right| \\
&= 2 \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\boldsymbol{\nu}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right|. \hspace{8em} \text{(Eqn. (25))}
\end{aligned}$$

Using Taylor's theorem with integral remainder form, we have that for $t \notin \mathcal{I}_T$,

$$\begin{aligned}
&\left| \mu_t(\boldsymbol{\nu}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right| \\
&= \left| \dot{\mu}_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \langle \boldsymbol{x}_t, \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \rangle + \int_{\mu_t(\widehat{\boldsymbol{\theta}}_{b(t)})}^{\mu_t(\boldsymbol{\nu}_t)} (\mu_t(\boldsymbol{\nu}_t) - z) \ddot{\mu}_t(z) dz \right| \\
&\leq \dot{\mu}_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \left| \langle \boldsymbol{x}_t, \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \rangle \right| + \langle \boldsymbol{x}_t, \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \rangle^2 \int_0^1 (1-v) \left| \ddot{\mu}_t\left(\widehat{\boldsymbol{\theta}}_{b(t)} + v(\boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)}) \right) \right| dv \\
&\hspace{20em} \text{(triangle inequality, reparametrization)} \\
&\leq \dot{\mu}_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \left| \langle \boldsymbol{x}_t, \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \rangle \right| + R_s \langle \boldsymbol{x}_t, \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \rangle^2 \underbrace{\int_0^1 (1-v) \dot{\mu}_t\left(\widehat{\boldsymbol{\theta}}_{b(t)} + v(\boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)}) \right) dv}_{= \tilde{\alpha}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)} \\
&\hspace{25em} \text{(Assumption 4)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) \left| \langle \boldsymbol{x}_t, \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \rangle \right| + \left| \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) - \dot{\mu}_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right| \left| \langle \boldsymbol{x}_t, \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \rangle \right| \\
&\quad + R_s \langle \boldsymbol{x}_t, \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \rangle^2 \tilde{\alpha}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t) \\
&\leq \underbrace{\dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) \left\| \boldsymbol{x}_t \right\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)^{-1}} \left\| \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \right\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)}}_{\triangleq A_t} \\
&\quad + \underbrace{\left| \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) - \dot{\mu}_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right| \left\| \boldsymbol{x}_t \right\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)^{-1}} \left\| \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \right\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)}}_{\triangleq B_t} \\
&\quad + \underbrace{R_s \left\| \boldsymbol{\nu}_t - \widehat{\boldsymbol{\theta}}_{b(t)} \right\|^2_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)} \tilde{\alpha}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t) \left\| \boldsymbol{x}_t \right\|^2_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)^{-1}},}_{\triangleq C_t} \\
&\hspace{25em} \text{(Cauchy-Schwartz inequality)}
\end{aligned}
$$

where $\widetilde{\boldsymbol{G}}$ is as defined in Eqn. (27).

We bound each sum separately:

Bounding $\sum_t A_t$

$$
\begin{aligned}
\sum_{t \notin \mathcal{I}_T} A_t &\leq \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sum_{t \notin \mathcal{I}_T} \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) \left\| \boldsymbol{x}_t \right\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)^{-1}} && (\boldsymbol{\nu}_t \in \mathcal{C}_{b(t)}, \text{Lemma D.6 } (i)) \\
&\leq \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sqrt{\sum_{t \notin \mathcal{I}_T} \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t)} \sqrt{\sum_{t \notin \mathcal{I}_T} \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) \left\| \boldsymbol{x}_t \right\|^2_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)^{-1}}} \\
& && \text{(Cauchy-Schwartz inequality)} \\
&\leq \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sqrt{\sum_{t \notin \mathcal{I}_T} \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t)} \sqrt{2g(\tau) \sum_{t \notin \mathcal{I}_T} \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) \left\| \boldsymbol{x}_t \right\|^2_{\bar{\boldsymbol{H}}_t^{-1}}} && \text{(Lemma D.5)} \\
&\leq \sqrt{2g(\tau) \left(4\lambda S^2 + \beta_T(\delta)^2\right)} \sqrt{\sum_{t \notin \mathcal{I}_T} \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t)} \sqrt{\sum_{t \in [T]} \min\left\{1, \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) \left\| \boldsymbol{x}_t \right\|^2_{\bar{\boldsymbol{H}}_t^{-1}}\right\}} \\
& && \text{(Definition of } \mathcal{I}_T) \\
&\leq 2\sqrt{dg(\tau)(4\lambda S^2 + \beta_T(\delta)^2) \log\left(1 + \frac{R_{\dot{\mu}} T}{d\lambda}\right)} \sqrt{\sum_{t \notin \mathcal{I}_T} \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t)}. && \text{(EPL (Lemma D.2))} \\
&\leq 2\sqrt{dg(\tau)(4\lambda S^2 + \beta_T(\delta)^2) \log\left(1 + \frac{R_{\dot{\mu}} T}{d\lambda}\right)} \sqrt{\sum_{t \in [T]} \dot{\mu}_{t,\star}(\boldsymbol{\theta}_\star) + \sum_{t \notin \mathcal{I}_T} \left\{\dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) - \dot{\mu}_{t,\star}(\boldsymbol{\theta}_\star)\right\}} \\
& && (\mu_{t,\star}(\cdot) := \mu(\langle \boldsymbol{x}_{t,\star}, \cdot \rangle)) \\
&= 2\sqrt{dg(\tau)(4\lambda S^2 + \beta_T(\delta)^2) \log\left(1 + \frac{R_{\dot{\mu}} T}{d\lambda}\right)} \sqrt{\frac{T}{\kappa_\star(T)} + \sum_{t \notin \mathcal{I}_T} \left\{\dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) - \dot{\mu}_{t,\star}(\boldsymbol{\theta}_\star)\right\}}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\sum_{t \notin \mathcal{I}_T} \left\{\dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) - \dot{\mu}_{t,\star}(\boldsymbol{\theta}_\star)\right\} &= \sum_{t \notin \mathcal{I}_T} \left\{\dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) - \dot{\mu}_t(\boldsymbol{\theta}_\star)\right\} + \sum_{t \notin \mathcal{I}_T} \left\{\dot{\mu}_t(\boldsymbol{\theta}_\star) - \dot{\mu}_{t,\star}(\boldsymbol{\theta}_\star)\right\} \\
&\leq R_s \left\{ \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\bar{\boldsymbol{\theta}}_t) - \mu_t(\boldsymbol{\theta}_\star) \right| + \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\boldsymbol{\theta}_\star) - \mu_{t,\star}(\boldsymbol{\theta}_\star) \right| \right\} \\
& \hspace{15em} \text{(Lemma D.3)}
\end{aligned}
$$

$$\leq R_s \left\{ \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\boldsymbol{\nu}_t) - \mu_t(\boldsymbol{\theta}_\star) \right| + \underbrace{\sum_{t \notin \mathcal{I}_T} \{ \mu_{t,\star}(\boldsymbol{\theta}_\star) - \mu_t(\boldsymbol{\theta}_\star) \}}_{= \mathsf{Reg}_{\mathcal{I}}(T)} \right\}$$

$$\leq R_s \left\{ \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\boldsymbol{\nu}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_t) \right| + \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\boldsymbol{\theta}_\star) - \mu_t(\widehat{\boldsymbol{\theta}}_t) \right| + \mathsf{Reg}_{\mathcal{I}}(T) \right\}$$

$$\leq 4R_s \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\boldsymbol{\nu}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right|. \qquad \text{(Definition of } (\boldsymbol{\nu}_t, b(t)))$$

Bounding $\sum_t B_t$

$$\sum_{t \notin \mathcal{I}_T} B_t \leq \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sum_{t \notin \mathcal{I}_T} \left| \dot{\mu}_t(\bar{\boldsymbol{\theta}}_t) - \dot{\mu}_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right| \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{G}}_{b(t)}^{-1}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)}$$

$$(\boldsymbol{\nu}_t \in \mathcal{C}_{b(t)}, \text{ Lemma D.6 } (i))$$

$$\leq R_s \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\bar{\boldsymbol{\theta}}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right| \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{G}}_{b(t)}^{-1}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)} \qquad \text{(Lemma D.3)}$$

$$\leq R_s \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sqrt{2g(\tau)\kappa(T)} \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\bar{\boldsymbol{\theta}}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right| \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}$$

$$\text{(Lemma D.4, } b(t) \geq t)$$

$$\leq R_s \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sqrt{2g(\tau)\kappa(T)} \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\boldsymbol{\nu}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right| \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}$$

$$\text{(Definition of } \boldsymbol{\nu}_t \text{ (Eqn. (25)))}$$

$$\leq 4R_s R_{\dot{\mu}} \kappa(T)(4\lambda S^2 + \beta_T(\delta)^2) \sqrt{g(\tau)} \sum_{t \notin \mathcal{I}_T} \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}^2$$

$$(\boldsymbol{\nu}_t, \widehat{\boldsymbol{\theta}}_{b(t)} \in \mathcal{C}_{b(t)}, \text{ Lemma D.6 } (ii))$$

$$\leq 4R_s R_{\dot{\mu}} \kappa(T)(4\lambda S^2 + \beta_T(\delta)^2) \sqrt{g(\tau)} \sum_{t \in [T]} \min\left\{1, \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}^2\right\} \qquad \text{(Definition of } \mathcal{I}_T)$$

$$\leq 8dR_s R_{\dot{\mu}} \kappa(T)(4\lambda S^2 + \beta_T(\delta)^2) \sqrt{g(\tau)} \log\left(1 + \frac{T}{2dg(\tau)\kappa(T)\lambda}\right).$$

$$\text{(EPL (Lemma D.2))}$$

Bounding $\sum_t C_t$

$$\sum_{t \notin \mathcal{I}_T} C_t \leq R_s \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sum_{t \notin \mathcal{I}_T} \tilde{\alpha}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t) \|\boldsymbol{x}_t\|_{\widetilde{\boldsymbol{G}}_{b(t)}(\widehat{\boldsymbol{\theta}}_{b(t)}, \boldsymbol{\nu}_t)^{-1}}^2$$

$$(\boldsymbol{\nu}_t \in \mathcal{C}_{b(t)}, \text{ Lemma D.6 } (i))$$

$$\leq R_s R_{\dot{\mu}} g(\tau)\kappa(T) \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sum_{t \notin \mathcal{I}_T} \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}^2 \qquad \text{(Lemma D.4, } b(t) \geq t)$$

$$\leq R_s R_{\dot{\mu}} g(\tau)\kappa(T) \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \sum_{t \in [T]} \min\left\{1, \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}^2\right\} \qquad \text{(Definition of } \mathcal{I}_T)$$

$$\leq 2dR_s R_{\dot{\mu}} g(\tau)\kappa(T) \sqrt{4\lambda S^2 + \beta_T(\delta)^2} \log\left(1 + \frac{T}{2dg(\tau)\kappa(T)\lambda}\right) \qquad \text{(EPL (Lemma D.2))}$$

Let us choose $\lambda = \frac{1}{4S^2}$. Then, combining everything, we have:

$$\sum_{t \notin \mathcal{I}_T} \left| \mu_t(\boldsymbol{\nu}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right|$$

$$\leq \sum_{t \notin \mathcal{I}_T} A_t + \sum_{t \notin \mathcal{I}_T} B_t + \sum_{t \notin \mathcal{I}_T} C_t$$

$$\lesssim \beta_T(\delta) \sqrt{d g(\tau) \log\left(1 + \frac{R_{\dot{\mu}} S T}{d}\right)} \sqrt{\frac{T}{\kappa_\star(T)} + R_s \sum_{t \notin \mathcal{I}_T} \left| \mu_t(\boldsymbol{\nu}_t) - \mu_t(\widehat{\boldsymbol{\theta}}_{b(t)}) \right|}$$

$$+ d R_s R_{\dot{\mu}} \kappa(T) \beta_T(\delta)^2 \sqrt{g(\tau)} \log\left(1 + \frac{ST}{d g(\tau) \kappa(T)}\right).$$

where we denote $A \lesssim B$ if $A \leq cB$ for some absolute constant $c > 0$, and we note that the upper bound for $\sum_t C_t$ is asymptotically negligible compared to $\sum_t B_t$.

This is of the form $X \lesssim A\sqrt{B + R_s X} + C$, which implies the bound of $X \lesssim A\sqrt{B} + A\sqrt{R_s} + C$ up to absolute constants via an elementary polynomial inequality (Abeille et al., 2021, Proposition 7). Combining everything gives us the desired statement. □

E Alternate CS via Discrete Uniform Prior and ε-net Argument

In this Appendix, instead of the PAC-Bayes with a continuous uniform prior/posterior as in the main text, we explore an alternate derivation of CS using a discrete uniform prior. This is a supplementary discussion for the "**Fast Rates in Statistical Learning**" paragraph in Section **??** in the main text.

We present the alternate CS, which is strictly looser than our Theorem 3.1:

Theorem E.1 (Slightly Looser, Unified CS for GLMs)**.** *Let* $L_t := \max_{\boldsymbol{\theta}\in\Theta} \left\|\nabla\mathcal{L}_t(\boldsymbol{\theta})\right\|_2$ *be the Lipschitz constant of* $\mathcal{L}_t(\cdot)$ *that may depend on* $\{(\boldsymbol{x}_s, r_s)\}_{s=1}^{t-1}$*. Then, we have* $\mathbb{P}[\exists t \geq 1 : \boldsymbol{\theta}_\star \notin \mathcal{C}_t(\delta)] \leq \delta$*, where*

$$\beta_t(\delta)^2 = \log\frac{\pi^2 t^2}{6\delta} + \inf_{c\in(0,5S]}\left\{d\log\frac{5S}{c} + cL_t\right\} \leq 1 + \log\frac{\pi^2t^2}{6\delta} + d\log(5SL_t), \tag{29}$$

where the last inequality follows from the choice $c = 1 \vee \frac{1}{L_t}$*.*

Proof. Consider $p \sim \mathcal{U}(\{\boldsymbol{\theta}_i\}_{i\in[N]})$, where the $\boldsymbol{\theta}_i$'s will be determined later. In that case, we have:

$$\begin{aligned}
\log\mathbb{E}_{\boldsymbol{\theta}}[M_t(\boldsymbol{\theta})] &= \mathcal{L}_t(\boldsymbol{\theta}_\star) + \log\mathbb{E}_{\boldsymbol{\theta}}[\exp\left(-\mathcal{L}_t(\boldsymbol{\theta})\right)] \\
&= \mathcal{L}_t(\boldsymbol{\theta}) + \log\left\{\frac{1}{N}\sum_{i=1}^N \exp\left(-\mathcal{L}_t(\boldsymbol{\theta}_i)\right)\right\} \\
&\geq \mathcal{L}_t(\boldsymbol{\theta}_\star) + \log\left\{\frac{1}{N}\max_{i\in[N]}\exp\left(-\mathcal{L}_t(\boldsymbol{\theta}_i)\right)\right\} \\
&= \mathcal{L}_t(\boldsymbol{\theta}_\star) - \max_{i\in[N]}\mathcal{L}_t(\boldsymbol{\theta}_i) + \log\frac{1}{N}.
\end{aligned}$$

By the Markov's inequality, we have

$$\mathbb{P}\left(\mathcal{L}_t(\boldsymbol{\theta}_\star) - \max_{i\in[N]}\mathcal{L}_t(\boldsymbol{\theta}_i) \leq \log\frac{N}{\delta}\right) \geq 1-\delta, \quad \forall t \geq 1.$$

By taking the union bound over $t \geq 1$ and $i \in [N]$, we have that

$$\mathbb{P}\left[\exists t \geq 1 : \max_{i\in[N]} M_t(\boldsymbol{\theta}_i) \geq N\frac{\pi^2t^2}{6\delta}\right] \leq \delta.$$

Here, we reparametrize δ as $\frac{\delta}{t^2}$ and use the Basel sum.

Taking the log and recalling that $M_t(\boldsymbol{\theta}) = \exp(\mathcal{L}_t(\boldsymbol{\theta}_\star) - \mathcal{L}_t(\boldsymbol{\theta}))$, above is equivalent to

$$\mathbb{P}\left[\exists t \geq 1 : \mathcal{L}_t(\boldsymbol{\theta}_\star) - \min_{i\in[N]}\mathcal{L}_t(\boldsymbol{\theta}_i) \geq \log N + \log\frac{\pi^2t^2}{6\delta}\right] \leq \delta.$$

With the above, we have that with probability at least $1-\delta$: for all $t \geq 1$,

$$\begin{aligned}
\mathcal{L}_t(\boldsymbol{\theta}_\star) - \min_{\boldsymbol{\theta}\in\Theta}\mathcal{L}_t(\boldsymbol{\theta}) &\leq \log\frac{\pi^2t^2}{6\delta} + \log N + \min_{i\in[N]}\mathcal{L}_t(\boldsymbol{\theta}_i) - \min_{\boldsymbol{\theta}\in\Theta}\mathcal{L}_t(\boldsymbol{\theta}) \\
&\leq \log\frac{\pi^2t^2}{6\delta} + \log N + L_t\min_{i\in[N]}\left\|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_t\right\|_2,
\end{aligned}$$

where we recall that L_t is the Lipschitz constant of $\mathcal{L}_t(\cdot)$.

We now choose $\{\boldsymbol{\theta}_i\}$ to be the c-net (as in the ε-net) of Θ for $c \in (0, 5S]$. As $\Theta \subseteq \mathcal{B}^d(S)$, we have that $N \leq \left(\frac{5S}{c}\right)^d$ (Vershynin, 2018, Corollary 4.2.13).

Then, with probability at least $1-\delta$, for all $t \geq 1$,

$$\mathcal{L}_t(\boldsymbol{\theta}_\star) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_t(\boldsymbol{\theta}) \leq \log \frac{\pi^2 t^2}{6\delta} + d \log \frac{5S}{c} + cL_t,$$

We then conclude by optimizing over c. □