
Eluder dimension: localise it!

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 We establish a lower bound on the eluder dimension in generalised linear model
2 classes, showing that standard eluder dimension-based analysis cannot lead to
3 first-order regret bounds. To address this, we introduce a localisation method for
4 the eluder dimension; our analysis immediately recovers and improves on classic
5 results for Bernoulli bandits, and allows for the first genuine first-order bounds for
6 finite-horizon reinforcement learning tasks with bounded cumulative returns.

7 1 Introduction

8 A decision-maker (or learner) is said to be *small-cost adaptive* if its cumulative excess cost over n
9 rounds—the n -step regret R_n —can be bounded in terms of the average cost of the optimal policy:

$$R_n \lesssim \sqrt{n\eta(a_\star)\Gamma_n} + \Gamma'_n,$$

10 where \lesssim denotes inequality up to constant factors, $\eta(a_\star)$ is the cost of an optimal action a_\star , and
11 Γ_n, Γ'_n are measures of the complexity of the learning task. In particular, when $\eta(a_\star) \lesssim 1/n$ and
12 Γ_n, Γ'_n are only logarithmic in n , this regret bound becomes logarithmic in n . Such bounds provide
13 instance-adaptivity: the learner adapts to the difficulty of the instance as measured by $\eta(a_\star)$.

14 A central challenge in achieving such bounds lies in how we measure the complexity of the learning
15 task. The *eluder dimension* (Russo and Van Roy, 2013) is a well-known capacity measure that has
16 been successfully applied to derive worst-case regret bounds in decision-making problems (Wen and
17 Van Roy, 2013; Osband and Van Roy, 2014; Ayoub et al., 2020; Wang et al., 2020). However, we
18 show in this paper that the eluder dimension, as originally defined by Russo and Van Roy (2013),
19 is intrinsically ill-suited for first-order regret analysis: it is designed for worst-case guarantees and
20 ignores local structure near the optimal predictor. We show that even with an algorithm that is
21 small-cost adaptive, using analysis based on the standard eluder dimension yields a κ term in the
22 leading complexity Γ_n which more than cancels out $\eta(a_\star)$ (κ is approximately $\max_a 1/\eta(a)$).

23 To address this gap, we introduce a *localisation* technique for the eluder dimension. By focusing on
24 the expected excess loss of predictors in a neighbourhood of the truth our analysis connects the eluder
25 dimension directly to the difficulty of the instance, rather than the global worst-case complexity.
26 This refinement enables us to obtain genuinely first-order regret bounds in several important settings.
27 Specifically, in the case of bandits with bounded costs this localisation recovers and improves existing
28 results; we then extend our analysis to finite-horizon reinforcement learning with bounded returns.

29 1.1 Contributions

30 We introduce the *localised ℓ_1 -eluder dimension*, a refinement of the standard (global) eluder dimension
31 (Russo and Van Roy, 2013) which leads to tighter analysis in settings where small-cost bounds
32 are achievable. Here, the choice to localise the ℓ_1 -eluder dimension (Liu et al., 2022) is key to
33 capture the small-cost behaviour of the algorithm. This localisation technique allows us to eliminate

the dependence on a worst-case per-step information gain parameter, denoted κ in the setting of generalised linear bandits (Faury et al., 2020). We establish the need for this localisation argument by showing that both the global ℓ_1 - (Liu et al., 2022) and ℓ_2 -eluder dimensions (Russo and Van Roy, 2013) scale with κ , a term that cancels out the benefits of small-cost bounds and second-order bounds in certain cases such as logistic or Poisson bandits.

A number of works rely on the ℓ_1 - (Wang et al., 2023, 2024a,b; Wang et al., 2025) or ℓ_2 - (Jia et al., 2024; Pacchiano, 2024; Ye et al., 2025) eluder dimension to achieve their small-cost bounds. Our lower bounds on κ in the setting of generalised bandits imply that the benefits of the small-cost/variance dependent bounds from prior work are negated by the use of the (global) eluder-dimension—the regret bounds no longer match the best known upper bounds (Abeille et al., 2021; Janz et al., 2024; Lee et al., 2024; Liu et al., 2024) when accounting for κ . *Our lower bound almost resolves the open problem of Li et al. (2022)—we are a $\log n$ factor away from a full resolution of the stated problem.*¹

To demonstrate our results, we provide a simple instance-adaptive version-space upper confidence bound algorithm, ℓ -UCB. This algorithm takes a loss function ℓ as input, which it uses to construct confidence sets for the cost function of an unknown problem. We show that if the loss function (i) is bounded, (ii) satisfies a standard variance condition², which enables the confidence sets to shrink rapidly in small-cost settings, and (iii) satisfies a certain triangle condition (Foster and Krishnamurthy, 2021), then the resulting algorithm attains a small-cost regret bound.

We extend ℓ -UCB to the setting of online reinforcement learning, yielding the ℓ -GOLF algorithm. This extension achieves the first real (κ -free) first-order bound for reinforcement learning with bounded costs.

Proofs of all results presented in the main body of this manuscript may be deferred to the appendix without explicit reference. The appendix begins with a guide to where each result is to be found.

1.2 Related work

Small-cost bounds have been established in adversarial bandits (Neu, 2015; Allen-Zhu et al., 2018; Foster and Krishnamurthy, 2021; Ito et al., 2020; Olkhovskaya et al., 2023) and, under stronger assumptions, in the stochastic setting where the cost distribution is known to the learner (Abeille et al., 2021; Faury et al., 2022; Janz et al., 2024; Liu et al., 2024; Lee et al., 2024). We study stochastic bandits with function approximation and *unknown* bounded cost distributions, mirroring the adversarial setting but without assuming any additional structure on rewards, such as knowledge of the distribution of the cost for each action. Our result is the first small-cost bound in this setting.

This distinction is important because algorithms designed for adversarial bandits tend to be unnecessarily conservative in stochastic environments (see Lattimore and Szepesvári, 2020), and do not extend naturally to our target setting of online reinforcement learning. In reinforcement learning, where both transitions and costs are stochastic, optimism (Lai and Robbins, 1985) remains a crucial principle for achieving low regret (Weisz et al., 2023; Wu et al., 2024; Moulin et al., 2025).

Small-cost bounds in reinforcement learning have been studied under various structural assumptions. Wang et al. (2023) established such bounds in online RL under the distributional Bellman completeness assumption, which effectively requires the learner to know the distribution of returns for each state-action pair. This assumption was relaxed to standard Bellman completeness in the offline setting by Ayoub et al. (2024), and the result with the relaxed assumption was then extended back to the online setting by Wang et al. (2024a). However, their analysis relies on a global (unlocalised) ℓ_1 -eluder dimension, causing regret to scale with κ and negating any advantages of small-cost adaptivity.

Other works such as Jin et al. (2020a) and Wagenmaker et al. (2022) obtain first-order regret bounds that scale with the optimal policy’s reward, not cost. These bounds improve upon earlier results (e.g., Azar et al., 2017; Yang and Wang, 2019; Jin et al., 2020b) only when the optimal policy accumulates very little reward. Small-cost bounds have been shown in more structured RL settings, such as tabular MDPs (Lee et al., 2020) and linear quadratic regulators (Kakade et al., 2020).

¹See Appendix C.4.

²Also known as the Bernstein condition; closely related to the Tsybakov condition and stochastic mixability.

83 2 Background on generalised linear models & loss functions

Suppose that \mathcal{A} is a closed subset of \mathbf{B}_2^d , let Θ be a closed subset of $S\mathbf{B}_2^d$ for some $S > 0$. Let $U \subset \mathbf{R}$ be a closed interval and let $\mu: U \rightarrow [0, 1]$ be an increasing function. The generalised linear model GLM on \mathcal{A} with parameter set Θ and link function μ is the set of functions

$$\text{GLM}(\mu, S) = \{a \mapsto \mu(\langle a, \theta \rangle) : a \in \mathcal{A}, \theta \in \Theta\}.$$

84 We will make the following assumption on the link and loss functions and the model. These abstract
85 away the idea of a loss function ℓ being the negative log-likelihood function associated with the
86 given generalised linear model class, allowing us to avoid introducing standard definitions relating to
87 natural exponential families. The requirement that $M, \kappa \geq 1$ is there solely to simplify our bounds.
88 Here, \mathbf{B}_2^d denotes the unit ball in \mathbf{R}^d .

89 **Assumption 2.1.** We make the following assumptions:

	$\mathcal{A} \subset \mathbf{B}_2^d$	(action set bound)
($\exists S > 0$)	$\Theta \subset S\mathbf{B}_2^d$	(parameter set bound)
($\forall (a, \theta) \in \mathcal{A} \times \Theta$)	$\langle a, \theta \rangle \in U$	(valid domain)
($\exists L > 0, \forall u, u' \in U$)	$ \mu(u) - \mu(u') \leq L u - u' $	(L -Lipschitz link)
($\exists M \geq 1, \forall u \in U^\circ$)	$ \ddot{\mu}(u) \leq M\dot{\mu}(u)$	(M -self-concordant link)
($\exists 1 \leq \kappa < \infty$)	$\kappa \geq \sup_{u \in U^\circ} 1/\dot{\mu}(u)$	(link derivative lower-bound)
($\forall y \in [0, 1], \forall u \in U$)	$\partial_u \ell(y, \mu(u)) = \mu(u) - y$	(link and loss are compatible)

90 Examples of loss and link combinations that satisfy our assumptions include the log-loss with the
91 sigmoid link function and the Poisson loss with the exponential link function:

92 *Example 2.1.* The log-loss function ℓ_X together with the sigmoid link function $\mu_X: \mathbf{R} \mapsto [0, 1]$ given
93 by $u \mapsto 1/(1 + e^{-u})$ satisfies Assumption 2.1 with: $L = 1/2$, $M = 1$ and $\kappa = 3e^S$.

94 *Example 2.2.* The Poisson loss function ℓ_P together with the exponential link function $\mu_P: [-S, 0] \mapsto$
95 $[0, 1]$ given by $u \mapsto e^u$ satisfies Assumption 2.1 with: $L = 1$, $M = 1$ and $\kappa = e^S$.

96 3 Problem setting: Bandits with bounded costs & the ℓ -UCB algorithm

Our bandit setting comprises a set of actions \mathcal{A} and a corresponding set of arm-dependent cost distributions $\mathcal{P} = \{P_a: a \in \mathcal{A}\}$ supported on the interval $[0, 1]$ (we will write $\text{supp } P$ for the support of a measure P). At each round $t \in \mathbf{N}_+$, a learner selects an action $A_t \in \mathcal{A}$ and receives a cost $Y_t \sim P_{A_t}$. We measure the learner's performance over $n \in \mathbf{N}_+$ rounds by the n -step regret

$$R_n = \sum_{t=1}^n \eta(A_t) - \eta(a_\star),$$

97 where $\eta: a \mapsto \int y P_a(dy)$ is the regression function associated with the bandit problem and $a_\star \in$
98 $\arg \min_{a \in \mathcal{A}} \eta(a)$ is any optimal arm (assumed to exist). We allow the learner to base its decision at
99 time-step t on the past observations $A_1, Y_1, \dots, A_{t-1}, Y_{t-1}$, any extra randomness independent of
100 the observations (say, for tie-breaking), and prior knowledge, taking the form of a set \mathcal{F} of functions
101 $\mathcal{A} \rightarrow [0, 1]$ known to contain η . The key assumptions here are:

102 **Assumption 3.1** (Bounded costs). We have $\cup_{a \in \mathcal{A}} \text{supp } P_a \subset [0, 1]$.

103 **Assumption 3.2** (Realisability). We have that $\eta \in \mathcal{F}$.

Our algorithm for the bandit setting, ℓ -UCB (Algorithm 1), is just an implementation of optimism with empirical risk minimisation-based confidence intervals. At each time-step $t \in \mathbf{N}_+$, the algorithm then constructs a confidence set $\mathcal{F}_t \subset \mathcal{F}$ for η , composed of all functions in \mathcal{F} for which the empirical risk on the past observations, based on a specified loss function $\ell: [0, 1]^2 \rightarrow \mathbf{R}$, is no more than β_t -suboptimal relative to the empirical risk minimiser, where $(\beta_t)_{t \geq 1}$ is a problem dependent non-negative, nondecreasing sequence of confidence widths. The algorithm then computes an optimistic function-action pair $(f_t, A_t) \in \mathcal{F}_t \times \mathcal{A}$, and plays the action A_t . The optimistic pair is defined as that satisfying

$$f_t(A_t) \leq f(a), \quad \forall (f, a) \in \mathcal{F}_t \times \mathcal{A}.$$

Algorithm 1 the ℓ -UCB bandit algorithm

input loss function ℓ , model \mathcal{F} , nonnegative confidence widths $(\beta_t)_t$
for time-step $t \in \mathbf{N}_+$ **do**
 let \mathcal{F}_t be the subset of the model given by

$$\mathcal{F}_t = \left\{ f \in \mathcal{F} : \sum_{i=1}^{t-1} \ell(Y_i, f(A_i)) \leq \inf_{\hat{f} \in \mathcal{F}} \sum_{i=1}^{t-1} \ell(Y_i, \hat{f}(A_i)) + \beta_t \right\},$$

 compute an optimistic function $f_t \in \mathcal{F}_t$ and action $A_t \in \mathcal{A}$ that satisfy

$$f_t(A_t) \leq f(a), \quad \forall (f, a) \in \mathcal{F}_t \times \mathcal{A},$$

 and play action A_t
end for

104 The optimisation over $\mathcal{F}_t \times \mathcal{A}$ is difficult without further assumptions; see Theorem 3.2 in Lee
105 et al. (2024) for a standard convex relaxation of this optimisation problem that is applicable to
106 self-concordant models as defined in our Section 4.1.

107 The crucial component to the ℓ -UCB algorithm obtaining small-cost adaptivity is the right choice of
108 the loss function ℓ used to construct the confidence intervals (and well-chosen confidence widths,
109 based on the loss function and model class). Our requirements will be stated in the form of an
110 assumption on the offset versions of the loss functions, and their expectations, defined thus.

Definition 3.1. Fix a model class \mathcal{F} and let $\ell: [0, 1]^2 \rightarrow \mathbf{R}$ be a loss function. Then, we define the
excess loss function $\varphi_f: [0, 1] \times \mathcal{A} \rightarrow \mathbf{R}$ and the expected excess loss function $\bar{\varphi}_f: \mathcal{A} \rightarrow \mathbf{R}_+$ of
each model $f \in \mathcal{F}$ by

$$\varphi_f(y, a) = \ell(y, f(a)) - \ell(y, \eta(a)) \quad \text{and} \quad \bar{\varphi}_f(a) = \int \varphi_f(\cdot, a) dP_a.$$

111 Our final assumption, the triangle condition, is expressed in terms of the triangular discrimination
112 function, given by

$$\Delta(p, q) = \frac{(p - q)^2}{p + q}$$

113 for any $p, q \in [0, 1]$. With that, our assumptions on the loss function are thus:

114 **Assumption 3.3** (Loss function assumptions). There exist constants $b, c, \gamma > 0$ such that for all
115 $(f, a) \in \mathcal{F} \times \mathcal{A}$, letting $Y \sim P_a$, the following three bounds hold:

$$\begin{aligned} |\varphi_f(Y, a)| &\leq b \text{ a.s.}, && \text{(boundedness)} \\ \text{Var } \varphi_f(Y, a) &\leq c \bar{\varphi}_f(a), && \text{(variance condition)} \\ \Delta(f(a), \eta(a)) &\leq \gamma \bar{\varphi}_f(a). && \text{(triangle condition)} \end{aligned}$$

116 The first two conditions in Assumption 3.3, boundedness and the variance condition, allow for a
117 Bernstein-type concentration across the loss class. The triangle condition is used in the regret decom-
118 position to move from fast concentration to a small-cost bound. The conditions in Assumption 3.3
119 implicitly depend on η , and thus ought to hold uniformly for all $\eta \in \mathcal{F}$. Recall our two losses:

- 120 • The log-loss function satisfies the triangle condition with $\gamma \leq 2$ (Proposition A.16).
121 Moreover, for any $f \in \mathcal{F}$ such that $\|\varphi_f\|_\infty \leq b$, φ_f satisfies the variance condition
122 with $c = b + 4$ (Proposition A.12).
- 123 • The Poisson loss satisfies the triangle condition with $\gamma \leq 5$ (Proposition A.17). Moreover,
124 for any $f \in \mathcal{F}$ such that $\|\varphi_f\|_\infty \leq b$, φ_f satisfies the variance condition with $c = b + 2$
125 (Proposition A.13).

126 The often used squared loss function fails to satisfy the triangle condition of Assumption 3.3, and
127 cannot lead to the type of small-cost bounds we seek; see Theorem 2 of Foster and Krishnamurthy
128 (2021) for an illustrative lower bound and discussion.

4 Main result: localised eluder dimension & first-order bounds

We now define two measures of complexity that govern the regret of ℓ -UCB. A covering number of the whole class, and a localised notion of the ℓ_1 -eluder dimension (Liu et al., 2022), which we refer to as just *the* eluder dimension henceforth. The latter is defined as follows:

Definition 4.1. Let \mathcal{Z} be a set and Ψ be a class of real-valued functions on \mathcal{Z} , and let $z = (z_1, z_2, \dots, z_n)$ be a sequence in \mathcal{Z} . We define the following:

1. We say $x \in \mathcal{Z}$ is ε -independent of z with respect to Ψ if there exists a $\psi \in \Psi$ such that $\sum_{t=1}^n |\psi(z_t)| \leq \varepsilon$ and $|\psi(x)| > \varepsilon$.
2. We say that z is an ε -eluder sequence with respect to Ψ if for all $t \leq n$, z_t is ε -independent of z_1, \dots, z_{t-1} with respect to Ψ .
3. The ε -eluder dimension of Ψ is the length of the longest ω -eluder sequence with respect to Ψ for any $\omega \geq \varepsilon$.

Let $\mathcal{F}' \subset \mathcal{F}$ be a localised model class and define the excess loss and the localised expected excess loss function classes by

$$\Phi(\mathcal{F}) = \{\varphi_f : f \in \mathcal{F}\} \quad \text{and} \quad \bar{\Phi}(\mathcal{F}') = \{\bar{\varphi}_f : f \in \mathcal{F}'\}.$$

Our regret bound for Algorithm 1 is as follows.

Theorem 4.1 (Regret bound for ℓ -UCB in bandits). *Fix $\delta \in (0, 1)$, $n \in \mathbb{N}_+$, bandit instance \mathcal{P} , model class \mathcal{F} and a loss function ℓ . Suppose that $(\mathcal{P}, \mathcal{F}, \ell)$ satisfy Assumptions 3.1 to 3.3. Let $h_t = e + \log(1 + t)$ for each $t \in [n]$, let N_n denote the $1/n$ -covering number of the loss class $\Phi(\mathcal{F})$ with respect to the uniform metric, and let*

$$\beta_t = 5/2 + 15(b + c) \log(N_n h_t / \delta), \quad t \in \mathbb{N}_+.$$

Let $\mathcal{F}' \subset \mathcal{F}$, and denote by d_n the $1/n$ -eluder dimension of the localised expected excess loss class $\bar{\Phi}(\mathcal{F}')$. Define our complexity measure

$$\Gamma_n = \gamma((d_n + 1)b + d_n \beta_n \log(nb) + 1).$$

Suppose a learner uses Algorithm 1, ℓ -UCB, over the course of n -many interactions with \mathcal{P} , with model class \mathcal{F} and loss function ℓ and with confidence widths $(\beta_t)_{t \in \mathbb{N}_+}$.

Then, with probability at least $1 - \delta$, the learner's regret is bounded as

$$R_n \leq 3\sqrt{n\eta(a_*)\Gamma_n} + 6\Gamma_n + \text{card}\{t \leq n : f_t \notin \mathcal{F}'\}.$$

We prove Theorem 4.1 over the course of Appendix A.

Our regret bound will depend on two factors:

1. The confidence widths, $(\beta_t)_{t \in \mathbb{N}_+}$, which are themselves based $1/n$ -covering number of $\Phi(\mathcal{F})$ in the uniform norm, responsible for extending a pointwise Bernstein-type inequality to the whole class $\Phi(\mathcal{F})$ (the pointwise inequality is Theorem A.7; the resulting uniform inequality is Theorem A.1). This is completely standard.
2. The ℓ_1 -norm $1/n$ -eluder dimension of the localised expected excess loss class $\bar{\Phi}(\mathcal{F}')$.
3. The number of times the algorithm selects an optimistic function f_t outside the localised model class \mathcal{F} (we refer to these as the *rogue* steps).

The dependence on the ε -eluder dimension of $\bar{\Phi}(\mathcal{F}')$ will be our focus; this can be thought of as measuring the number of times we are ‘surprised’ at the scale $\varepsilon > 0$, in that there was a model $f \in \mathcal{F}'$ with low expected excess loss on past inputs that has high expected excess loss on some unseen action. The following lower bound shows that it is vital to only consider ‘surprises’ near a_* .

Theorem 4.2 (Eluder dimension lower bound). *Let ℓ and $\mathcal{F} = \text{GLM}(\mu, S)$, $S \geq 4$, $d \geq 2$, satisfy Assumption 2.1 with parameter M . Then there exist a universal constant $C > 0$, a parameter set Θ and an action set \mathcal{A} such that for $\varepsilon \leq \frac{\dot{\mu}(0)}{M^2}$, the ε -eluder dimension of $\bar{\Phi}(\mathcal{F})$ (defined in Definition 3.1), is lower bounded by*

$$\frac{Cd}{S} \min(\exp(S), \exp(\frac{\ln(\tilde{\kappa})^2}{8SM^2 + 4\ln(\tilde{\kappa})})) \quad \text{where} \quad \tilde{\kappa} = \frac{\dot{\mu}(0)}{\dot{\mu}(-S/2)}.$$

The quantity $\tilde{\kappa} > 0$ can be thought of as measuring the change of information gained between the middle of the parameter set and a large step in the negative direction; in all the commonly considered generalised linear models, we have $\tilde{\kappa} \approx \kappa$.

To understand the implications of Theorem 4.2, consider the setting of logistic bandits in the usual low-information regime, where $\langle a_*, \theta_* \rangle \approx -S$; think clickthrough rates in online advertising, where even the best adverts rarely get clicked on. Then, $\eta(a_*) \approx \dot{\mu}(\langle a^*, \theta_* \rangle) \approx \exp(-S)$, which suggests that our regret should be excellent; but at the same time $\tilde{\kappa} \approx \exp(S)$, and thus the eluder dimension scales as $\Omega(\exp(S))$, completely cancelling out the benefit of the $\eta(a_*)$ small cost term. This results in a bound that fails to truly adapt to the problem instance. We now show how localisation helps.

To reason about localised eluder in generalised linear models, first associate each f_t with a parameter $\theta_t \in \Theta$ such that $f_t(\cdot) = \mu(\langle \cdot, \theta_t \rangle)$, and define the localised parameter classes with radius $r > 0$ and the implied localised model classes as follows:

$$\Theta'(r) = \{\theta \in \Theta : \forall a \in \mathcal{A}, |\langle a, \theta - \theta_* \rangle| \leq r\}, \quad \mathcal{F}'(r) = \{\mu(\langle \cdot, \theta \rangle) : \theta \in \Theta'(r)\} \quad \text{for } r > 0.$$

We have the following upper bound on the eluder dimension of $\mathcal{F}'(r)$ as a function of r :

Proposition 4.3. *Let Assumption 2.1 hold. Then, there exists a universal constant $C > 0$ such that for any $r, \varepsilon > 0$, the ε -eluder dimension of $\Phi(\mathcal{F}'(r))$ does not exceed*

$$Cd \exp(rM) \log(1 + S^2 L \exp(rM) / \varepsilon).$$

Crucially, taking $r = 1/M$, we obtain an upper bound on the ε -eluder dimension of the form

$$Cd \log(1 + S^2 L / \varepsilon).$$

This depends only logarithmically on S , thus allowing for the main term to scale with e^{-S} , which is a huge improvement over the non-adaptive result obtained via the global eluder dimension.

Now, consider the following bound on the number of rogue steps taken by the ℓ -UCB algorithm.

Proposition 4.4 (Number of rogue steps bound). *Under Assumption 2.1, on the high-probability event of Theorem 4.1, for any $n \in \mathbb{N}_+$,*

$$\text{card}\{t \leq n : |\langle A_t, \theta_t - \theta_* \rangle| > 1/M\} \leq 64d\kappa M^2 \beta_n \log(1 + (64/3)\kappa^2 M^2 S^2 \beta_n).$$

We see that this number of rogue steps does depend on the large quantity κ ; but, examining the bound of Theorem 4.1, we see that this quantity features as an additive term, and not in the leading term. Our localisation trick has allowed us to recover a truly instance-adaptive bound using eluder dimension.

We have two final remarks on our results and the notion of eluder dimension that we consider:

Remark 4.1. Li et al. (2022) pose an open problem on providing tight bounds on the eluder dimension for generalised linear models. A lower bound we state in Appendix C almost does this: we are missing a $\log 1/\varepsilon$.

Remark 4.2. The eluder dimension introduced in Russo and Van Roy (2013) is defined with ℓ_2 -norm in place of the ℓ_1 -norm, and is applied to the model class (here \mathcal{F}) directly, rather than to (a localised) excess loss function class $\Phi(\mathcal{F}')$. In effect, this builds a squared-loss assumption into the eluder dimension, which per Foster and Krishnamurthy (2021, Theorem 2) is not compatible with fast rates. Furthermore, since the ℓ_1 -norm eluder dimension is never greater than the ℓ_2 -eluder (Liu et al., 2022), we believe the definition of Liu et al. (2022) used here should be preferred.

4.1 Comparison with existing results in the logistic bandit setting

The most directly comparable results to our work are those for the logistic bandits. Note, however, that while Assumption 2.1 assumes a generalised linear model for the responses, it does not assume that the responses are generated by such a model—merely that they are bounded in $[0, 1]$, and that their mean can be realised by some function in the generalised linear model class. The latter occurs as soon as the GLM can predict every mean value in $[0, 1]$. In the remarks and discussions that follow, we will refer to the setting where the responses *are* generated by the model as the maximum likelihood estimation (MLE) setting, and note that this is a strict subset of the setting we consider (for bounded costs).

Previous results in MLE settings subsumed by our work, like the logistic bandit setting of Faury et al. (2022), rely on a warm-up phase to ensure the accuracy of the confidence intervals. This would take

the form of either uniform exploration or the approximate solving of an optimal design problem. Our proof technique eliminates the need for this: the additional regret term R'_n we pay, which would usually be associated with such a warm-up, appears in the analysis only. *This is vital for the upcoming reinforcement learning setting, where because we do not have random access to state-action pairs, uniform sampling or the solving of an optimal design is infeasible.* Moreover, since the warm-up length in previous algorithms would be based on worst-case quantities, the additive cost R'_n in our algorithm is more adaptive, and thus lower.

Our result in Proposition C.1 subsumes the classic logistic bandit setting, where the responses are generated by

$$Y_t \sim \text{Bernoulli}(\eta(A_t)), \quad \text{for each } t \in \mathbf{N}_+,$$

where η is given by a generalised linear model with the sigmoid link function (Example 2.1). We now discuss how our approach relates to the results available for this common special case, and for other related generalised linear bandit settings.

Faury et al. (2022) obtain a bound for the logistic bandit setting that scales with $\eta(a_*)(1 - \eta(a_*))$, rather than just $\eta(a_*)$. Our analysis to the log-loss could recover that same dependency (this would not work for the Poisson loss). While the bound of Faury et al. (2022) also holds for the ‘misspecified’ setting of $[0, 1]$ rewards, their arguments require that the optimistic functions $(f_t)_{t \in \mathbf{N}_+}$ to be in a neighbourhood of η at every step (equivalent to our \mathcal{F}'), which they achieve through the use of an explicit warm-up procedure. Such algorithmic mechanisms are fine in the bandit setting, but cannot be translated to the online reinforcement learning, where we do not have random access to state-action pairs. In contrast, our localised eluder dimension and number of rogue steps arguments eliminate the need for an explicit warm-up.

The work of Lee et al. (2024) does away with the warm-up employed in Faury et al. (2022) using a PAC-Bayes technique, self-concordance and likelihood ratios. However, their results only hold for the MLE setting, since their proof relies on the observation that in the MLE setting, ratios of likelihoods give rise to a martingale (Lemma 3.1 in their paper). But when the probability model is misspecified—such as when the rewards are uniformly distributed while one assumes that they come from a Bernoulli distribution—it is easy to verify that the likelihood ratios no longer form a martingale. As such, the results of Lee et al. (2024) paper are not applicable to our setting where we make no specific parametric assumptions about the reward distributions beyond boundedness.

Emmenegger et al. (2024) also use a likelihood-ratio method for constructing confidence sets, together with an online-to-batch reduction. Their method has the advantage of not requiring the knowledge of the covering numbers at runtime, but adapting to the problem instance automatically. However, their work, too, is rooted in the MLE assumption and does not extend to our bounded-reward setting.

Finally, note that any small-cost result derived for the maximum likelihood logistic setting can be immediately converted to a result for bounded random variables using the Bernoullisation trick. In particular, if we have a small-cost adaptive algorithm A for $\{0, 1\}$ responses, and observe responses $(Y_t)_{t \in \mathbf{N}_+}$ in the interval $[0, 1]$, then we can simply sample

$$Y'_t \sim \text{Bernoulli}(Y_t) \quad \text{for each } t \in \mathbf{N}_+,$$

and feed $(Y'_t)_{t \in \mathbf{N}_+}$ to our algorithm A . This procedure retains the same first-order guarantees, but destroys any second-order adaptivity the algorithm may show. Indeed, consider the case where the $(Y_t)_{t \in \mathbf{N}_+}$ are equal to $1/2$ almost surely. Then, running empirical loss minimisation with the log-loss on $(Y_t)_{t \in \mathbf{N}_+}$ converges to $1/2$ after a single observation, but running the same procedure on the corresponding sequence $(Y'_t)_{t \in \mathbf{N}_+}$ of independent $\text{Bernoulli}(1/2)$ random variables leads to an $\Omega(1/\sqrt{n})$ absolute error in the estimate of the mean after n observations.

5 Application: online reinforcement learning

We now extend our insights from the bandit setting to online reinforcement learning, where the elimination of a warm-up is vital, and the bounded cost assumption is standard.

We consider an episodic reinforcement learning setting with a fixed horizon $H \in \mathbf{N}_+$. Here, we have an MDP $M = (\mathcal{S}, \mathcal{A}, c, P, s_1)$, where \mathcal{S} are the states, \mathcal{A} are the actions, $c = (c_1, \dots, c_H)$ with $c_h: \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$; a deterministic starting state $s_1 \in \mathcal{S}$; and a transition kernel $P = (P_1, \dots, P_h)$ with P_h mapping $\mathcal{S} \times \mathcal{A}$ to a measure over \mathcal{S} .

The learner interacts with the MDP M for $n \in \mathbf{N}_+$ episodes. At the start of each episode $t \in [n]$, it specifies a deterministic policy $\pi^t = (\pi_1^t, \dots, \pi_H^t)$, where $\pi_h^t: \mathcal{S} \rightarrow \mathcal{A}$ for each $h \in [H]$. For $(h, t) \in [H] \times [n]$, we let S_h^t, A_h^t and $C_h^t = c(S_h^t, A_h^t)$ denote the state, action and cost that occurs in step h of episode t . We allow the policy π^t to depend on the states, actions and costs up to the start of the t th episode. The policies may not depend directly on the cost function c or the dynamics P , as these are assumed to be unknown. The learner's aim will be to minimise the expected cumulative cost incurred over the n episodes. To formalise this, let v_h^t be the value function of policy π^t in M , given by

$$v_h^\pi(s) = \mathbf{E}_\pi \left[\sum_{i=h}^H c(S_i, \pi(S_i)) \mid S_h = s \right],$$

256 for each $s \in \mathcal{S}$, where \mathbf{E}_π denotes the expectation with respect to the transitions induced by following
 257 the policy π in the MDP M starting at stage h . Let the value of following policy π^t be denoted by
 258 $v^t = v^{\pi^t}$. Then the n -episode regret is given by

$$R_n = \sum_{t=1}^n v_1^t(s_1) - v_1^*(s_1),$$

259 where v^* is the optimal value function, defined formally just after Equation (1).

260 The key assumption that our learner will be allowed to exploit is the following:

261 **Assumption 5.1.** The costs are non-negative and sum to at most one over each episode.

Model structure For each $h \in [H]$, we let \mathcal{Q}_h be the set of all functions mapping $\mathcal{S}_h \times \mathcal{A}_h \rightarrow [0, 1]$ and define the set of all action-value functions to be

$$\mathcal{Q} = \mathcal{Q}_1 \times \dots \times \mathcal{Q}_H,$$

262 where each $q \in \mathcal{Q}$ is to be interpreted as a map from $\mathcal{S} \times \mathcal{A}$ to $[0, 1]^H$. For any $q \in \mathcal{Q}$ we write q^\wedge
 263 for the function $\mathcal{S} \rightarrow [0, 1]^H$ defined by

$$s \mapsto \min_{a \in \mathcal{A}} q_h(s, a) \quad \text{for all } s \in \mathcal{S} \text{ and } h \in [H].$$

264 We let $\mathcal{T}: \mathcal{Q} \rightarrow \mathcal{Q}$ denote the Bellman optimality operator for the MDP M , given by

$$\mathcal{T}: q \mapsto c + \int q^\wedge dP,$$

265 where, with slight abuse of notation, the integral is to be understood as with respect to the product
 266 $P = P_1 \times \dots \times P_H$. We define the optimal action-value function q^* for M to be the element of \mathcal{Q}
 267 satisfying

$$\mathcal{T}q^* = q^*, \tag{1}$$

268 and define the value function v^* for M to be $v^* = q^{*\wedge}$.

For any function $q \in \mathcal{Q}$, we write π^q for the policy greedy with respect to q , defined by

$$\pi_h^f(s) \in \arg \min_{a \in \mathcal{A}} q_h(s, a) \quad \text{for all } s \in \mathcal{S} \text{ and } h \in [H].$$

Algorithm Our algorithm, Algorithm 2, is an extension of our earlier bandit algorithm to the episodic online reinforcement learning setting. The algorithm requires the specification of a loss function $\ell: [0, 1]^2 \rightarrow \mathbf{R}$, confidence widths $(\beta_t)_{t \in [n]}$ and function classes $\mathcal{G}, \mathcal{F} \subset \mathcal{Q}$. The model \mathcal{F} contains candidate functions for estimating q^* , and the model \mathcal{G} contains candidates for estimating $\mathcal{T}f$, for any $f \in \mathcal{F}$. For convenience, we augment every $f \in \mathcal{F} \cup \mathcal{G}$ with

$$f_{H+1} = 0,$$

269 which serves to enforce the zero value boundary condition for the end of an episode.

270 We make the realisability and generalised completeness assumptions of Antos et al. (2008):

271 **Assumption 5.2** (Realisability). We assume that $q^* \in \mathcal{F}$.

272 **Assumption 5.3** (Generalised completeness). We assume that $\mathcal{T}\mathcal{F} \subseteq \mathcal{G}$.

Algorithm 2 The ℓ -GOLF algorithm

input loss function ℓ , models \mathcal{F} and \mathcal{G} , nonnegative confidence widths $(\beta_t)_t$
for episode $t \in \mathbf{N}_+$ **do**
 for each $h \in [H]$ **let**

$$\mathcal{L}_h^{t-1}(f, f') = \sum_{i=1}^{t-1} \ell(1 \wedge (C_h^i + f^\wedge(S_{h+1}^i)), f'(S_h^i, A_h^i))$$

 and let \mathcal{F}^t be the subset of \mathcal{F} given by

$$\mathcal{F}^t = \left\{ f \in \mathcal{F} : \mathcal{L}_h^{t-1}(f_{h+1}, f_h) \leq \inf_{g \in \mathcal{G}_h} \mathcal{L}_h^{t-1}(f_{h+1}, g) + \beta_t, \forall h \in [H] \right\}$$

 compute an optimistic function

$$f^t \in \arg \min_{f \in \mathcal{F}^t} f(s_1, \pi_f(s_1))$$

 and play the policy $\pi^t := \pi^{f^t}$ greedy with respect to f^t
end for

273 Now, in each episode $t \in \mathbf{N}_+$, the algorithm constructs confidence sets $\mathcal{F}^t \subset \mathcal{F}$ containing action-
274 value functions that are close to satisfying the bellman optimality condition $f = \mathcal{T}f$ on the data
275 observed thus far, with errors penalised according to ℓ . It then selects an optimistic function $f^t \in \mathcal{F}$,
276 defined as that for which the greedy policy maximises the value for the episode, and plays $\pi^t := \pi^{f^t}$,
277 the greedy policy with respect to f^t .

278 The ℓ -GOLF algorithm generalises the GOLF algorithm of Jin et al. (2021) to arbitrary loss functions,
279 with GOLF recovered by taking ℓ to be the squared loss. Again, the right choice of the loss will
280 be crucial in getting the ℓ -GOLF algorithm to adapt to small costs, with squared loss not achieving
281 small-cost behaviour (see Foster and Krishnamurthy (2021, Theorem 2)).

The conditions on the loss are much the same as in the bandit setting, but applying instead to Bellman residual errors, and it is to the spaces of these residuals, measured using our specified loss function, that we assign the localised eluder dimension. The overall result, which is stated in full in Appendix B, is again of the form: for any $\delta > 0$, with probability at least $1 - \delta$, the learner's regret is bounded as

$$R_n \leq 3\sqrt{Hn\eta(a_\star)\Gamma_n} + 6H\Gamma_n + \text{card}\{t \leq n : f_t \notin \mathcal{F}'\}.$$

282 Since the bandit setting is a special case of the reinforcement learning setting, our discussion regarding
283 lower bounds holds here too: localisation is still necessary if using the eluder dimension.

284 For context, the closest results to ours are those of Wang et al. (2023, 2024a) for online RL. Both
285 provide a small-cost regret bound scaling with the Bellman eluder dimension; however, without our
286 notion of a *localised* dimension, their regret bound scales with κ in the leading term for logistic
287 linear models. This entirely offsets any benefit of their small-cost analysis; the bound is not truly
288 instance-adaptive. Moreover, Wang et al. (2023) assumes that the distributional Bellman operator
289 (Bellemare et al., 2017) lies in their model class, an assumption that is significantly stronger than our
290 Assumption 5.3 (as shown in Ayoub et al., 2024).

291 6 Conclusion

292 We have shown that standard eluder dimension analysis inherently fails to achieve first-order regret
293 bounds in generalised linear model settings. By introducing the localised ℓ_1 -eluder dimension, we
294 overcome this limitation, removing problematic worst-case dependencies and achieving genuinely
295 adaptive, first-order regret bounds. Our refined analysis recovers and sharpens classical results in
296 Bernoulli bandit scenarios and demonstrates clear practical advantages through the ℓ -UCB algorithm.

297 Moreover, our localisation approach successfully extends to finite-horizon reinforcement learning
298 via the ℓ -GOLF algorithm, providing the first genuine first-order regret bounds in this setting. This
299 highlights the crucial role of localisation techniques in developing instance-adaptive algorithms,
300 opening promising avenues for further exploration in broader learning contexts.

References

- M. Abeille, L. Faury and C. Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, 2021. Cited on page 2.
- Z. Allen-Zhu, S. Bubeck and Y. Li. Make the minority great again: First-order regret bound for contextual bandits. In *International Conference on Machine Learning*, 2018. Cited on page 2.
- A. Antos, C. Szepesvári and R. Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2008. Cited on page 8.
- A. Ayoub, Z. Jia, C. Szepesvari, M. Wang and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, 2020. Cited on page 1.
- A. Ayoub, K. Wang, V. Liu, S. Robertson, J. McInerney, D. Liang, N. Kallus and C. Szepesvári. Switching the Loss Reduces the Cost in Batch Reinforcement Learning. In *International Conference on Machine Learning*, 2024. Cited on pages 2, 9, 14.
- M. G. Azar, I. Osband and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017. Cited on page 2.
- M. G. Bellemare, W. Dabney and R. Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, 2017. Cited on page 9.
- S. Boucheron, G. Lugosi and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. Cited on page 17.
- S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003. Cited on page 29.
- S. Dong, T. Ma and B. Van Roy. On the performance of thompson sampling on logistic bandits. In *Conference on Learning Theory*, pages 1158–1160. PMLR, 2019. Cited on page 29.
- N. Emmenegger, M. Mutny and A. Krause. Likelihood ratio confidence sets for sequential decision-making. *Advances in Neural Information Processing Systems*, 2024. Cited on page 7.
- T. Erven, P. Grünwald, M. D. Reid and R. C. Williamson. Mixability in Statistical Learning. In *Advances in Neural Information Processing Systems*, 2012. Cited on page 18.
- L. Faury, M. Abeille, C. Calauzènes and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, 2020. Cited on pages 2, 35.
- L. Faury, M. Abeille, K.-S. Jun and C. Calauzènes. Jointly Efficient and Optimal Algorithms for Logistic Bandits. In *International Conference on Artificial Intelligence and Statistics*, 2022. Cited on pages 2, 6, 7.
- D. J. Foster and A. Krishnamurthy. Efficient first-order contextual bandits: prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 2021. Cited on pages 2, 4, 6, 9.
- S. Ito, S. Hirahara, T. Soma and Y. Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 2020. Cited on page 2.
- D. Janz, S. Liu, A. Ayoub and C. Szepesvári. Exploration via linearly perturbed loss minimisation. In *International Conference on Artificial Intelligence and Statistics*, 2024. Cited on pages 2, 26, 35.
- Z. Jia, J. Qian, A. Rakhlin and C.-Y. Wei. How does variance shape the regret in contextual bandits? In *Advances in Neural Information Processing Systems*, 2024. Cited on page 2.
- C. Jin, A. Krishnamurthy, M. Simchowitz and T. Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 2020. Cited on page 2.

345 C. Jin, Q. Liu and S. Miryoosefi. Bellman eluder dimension: new rich classes of RL problems, and
346 sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 2021. Cited on
347 page 9.

348 C. Jin, Z. Yang, Z. Wang and M. I. Jordan. Provably efficient reinforcement learning with linear
349 function approximation. In *Conference on Learning Theory*, 2020. Cited on page 2.

350 S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi and W. Sun. Information theoretic regret
351 bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 2020.
352 Cited on page 2.

353 T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied
354 mathematics*, 1985. Cited on page 2.

355 T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. Cited on
356 pages 2, 26.

357 C.-W. Lee, H. Luo, C.-Y. Wei and M. Zhang. Bias no more: High-probability data-dependent regret
358 bounds for adversarial bandits and MDPs. *Advances in Neural Information Processing Systems*,
359 2020. Cited on page 2.

360 J. Lee, S.-Y. Yun and K.-S. Jun. A Unified Confidence Sequence for Generalized Linear Models, with
361 Applications to Bandits. In *Conference on Neural Information Processing Systems*, 2024. Cited on
362 pages 2, 4, 7, 35.

363 G. Li, P. Kamath, D. J. Foster and N. Srebro. Understanding the eluder dimension. *Advances in
364 Neural Information Processing Systems*, 2022. Cited on pages 2, 6, 29, 33.

365 Q. Liu, A. Chung, C. Szepesvári and C. Jin. When is partially observable reinforcement learning not
366 scary? In *Conference on Learning Theory*, 2022. Cited on pages 1, 2, 5, 6, 15.

367 S. Liu, A. Ayoub, F. Sentenac, X. Tan and C. Szepesvari. Almost Free: Self-concordance in Natural
368 Exponential Families and an Application to Bandits. In *Advances in Neural Information Processing
369 Systems*, 2024. Cited on pages 2, 35.

370 A. Moulin, G. Neu and L. Viano. Optimistically Optimistic Exploration for Provably Efficient Infinite-
371 Horizon Reinforcement and Imitation Learning. *arXiv preprint arXiv:2502.13900*, 2025. Cited on
372 page 2.

373 G. Neu. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*,
374 2015. Cited on page 2.

375 J. Olkhovskaya, J. Mayo, T. van Erven, G. Neu and C.-Y. Wei. First- and Second-Order Bounds for
376 Adversarial Linear Contextual Bandits. In *Advances in Neural Information Processing Systems*,
377 2023. Cited on page 2.

378 I. Osband and B. Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances
379 in Neural Information Processing Systems*, 2014. Cited on page 1.

380 A. Pacchiano. Second order bounds for contextual bandits with function approximation. *arXiv preprint
381 arXiv:2409.16197*, 2024. Cited on page 2.

382 Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University
383 Press, 2025. Cited on page 19.

384 D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In
385 *Advances in Neural Information Processing Systems*, 2013. Cited on pages 1, 2, 6, 33.

386 T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: a recipe for Newton-type methods.
387 *Mathematical Programming*, 2019. Cited on pages 27, 31, 35.

388 R. Vershynin. *High-dimensional probability: An introduction with applications in data science*,
389 volume 47. Cambridge University Press, 2018. Cited on page 26.

390 A. J. Wagenmaker, Y. Chen, M. Simchowitz, S. Du and K. Jamieson. First-order regret in reinforcement
391 learning with linear function approximation: A robust estimation approach. In *International*
392 *Conference on Machine Learning*, 2022. Cited on page 2.

393 K. Wang, N. Kallus and W. Sun. The central role of the loss function in reinforcement learning. *arXiv*
394 *preprint arXiv:2409.12799*, 2024. Cited on pages 2, 9.

395 K. Wang, O. Oertel, A. Agarwal, N. Kallus and W. Sun. More benefits of being distributional: Second-
396 order bounds for reinforcement learning. In *International Conference on Machine Learning*, 2024.
397 Cited on page 2.

398 K. Wang, K. Zhou, R. Wu, N. Kallus and W. Sun. The benefits of being distributional: Small-loss
399 bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.
400 Cited on pages 2, 9.

401 R. Wang, R. R. Salakhutdinov and L. Yang. Reinforcement learning with general value function
402 approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural*
403 *Information Processing Systems*, 2020. Cited on page 1.

404 Z. Wang, D. Zhou, J. C. Lui and W. Sun. Model-based RL as a minimalist approach to horizon-free
405 and second-order bounds. In *International Conference on Learning Representations*, 2025. Cited
406 on page 2.

407 G. Weisz, A. Györfy and C. Szepesvári. Online RL in Linearly q^π -Realizable MDPs Is as Easy as in
408 Linear MDPs If You Learn What to Ignore. *Advances in Neural Information Processing Systems*,
409 2023. Cited on page 2.

410 Z. Wen and B. Van Roy. Efficient exploration and value function generalization in deterministic
411 systems. *Advances in Neural Information Processing Systems*, 2013. Cited on page 1.

412 J. Whitehouse, Z. S. Wu and A. Ramdas. Time-uniform self-normalized concentration for vector-
413 valued processes. *arXiv preprint arXiv:2310.09100*, 2023. Cited on pages 16, 17.

414 R. Wu, A. Sekhari, A. Krishnamurthy and W. Sun. Computationally efficient RL under linear bellman
415 completeness for deterministic dynamics. *arXiv preprint arXiv:2406.11810*, 2024. Cited on page 2.

416 L. Yang and M. Wang. Sample-Optimal Parametric Q-Learning Using Linearly Additive Features. In
417 *International Conference on Machine Learning*, 2019. Cited on page 2.

418 C. Ye, Y. Jin, A. Agarwal and T. Zhang. Catoni contextual bandits are robust to heavy-tailed rewards.
419 *arXiv preprint arXiv:2502.02486*, 2025. Cited on page 2.

420 Appendix

421 The appendix begins with Appendix A, where we prove our main result in the bandit setting,
 422 Theorem 4.1. Along the way (in Appendix A.1) we establish the crucial Theorem A.1, which is
 423 also used to prove our result in RL. Then we show that our exemplary loss functions, the Poisson
 424 loss and the log-loss, satisfy the variance condition and the triangle condition in Appendices A.2
 425 and A.3 respectively. In Appendix B we prove the extension of our cost sensitive regret bound to
 426 RL, which is stated in Theorem B.1. All major sections begin with a high-level summary of their
 427 contents/intuitive proof sketch.

428 A Proof of Theorem 4.1, cost sensitive regret bound

429 This section outlines the key steps involved in proving the cost-sensitive regret bound for our bandit
 430 algorithm, as stated in Theorem 4.1. The proof unfolds as follows:

- 431 • After restating the relevant definitions and assumptions from the main text for convenience,
 432 we begin by introducing a uniform Bernstein inequality (Theorem A.1), which is the
 433 concentration result that gives us confidence in our confidence sets. Since it is somewhat
 434 technical and cumbersome, we defer the proof of the uniform Bernstein inequality to
 435 Appendix A.1.
- 436 • Then in Proposition A.2 we apply the uniform Bernstein inequality to establish that, with
 437 high probability, the regression function η is in the confidence set \mathcal{F}_t at every round $t \in \mathbf{N}_+$.
- 438 • Having handled the probabilistic aspects of the proof, we switch to the regret decomposition,
 439 and specifically separate the regret into two parts: the regret incurred when the algorithm
 440 selects a function from the confidence set \mathcal{F}_t (well-behaved steps) and the regret incurred
 441 when it selects a function from the set \mathcal{F}' (rogue steps).
- 442 • Finally, we bound the contribution of well-behaved steps to the regret on the good event \mathcal{E}_δ
 443 using the triangle condition of the loss function.

444 The following assumptions and definitions from the main text are necessary for the proof of The-
 445 orem 4.1—we also repeat the theorem statement for ease of reference.

446 **Assumption 3.1** (Bounded costs). We have $\cup_{a \in \mathcal{A}} \text{supp } P_a \subset [0, 1]$.

447 **Assumption 3.2** (Realisability). We have that $\eta \in \mathcal{F}$.

Definition 3.1. Fix a model class \mathcal{F} and let $\ell: [0, 1]^2 \rightarrow \mathbf{R}$ be a loss function. Then, we define the
 excess loss function $\varphi_f: [0, 1] \times \mathcal{A} \rightarrow \mathbf{R}$ and the expected excess loss function $\bar{\varphi}_f: \mathcal{A} \rightarrow \mathbf{R}_+$ of
 each model $f \in \mathcal{F}$ by

$$\varphi_f(y, a) = \ell(y, f(a)) - \ell(y, \eta(a)) \quad \text{and} \quad \bar{\varphi}_f(a) = \int \varphi_f(\cdot, a) dP_a.$$

448 **Assumption 3.3** (Loss function assumptions). There exist constants $b, c, \gamma > 0$ such that for all
 449 $(f, a) \in \mathcal{F} \times \mathcal{A}$, letting $Y \sim P_a$, the following three bounds hold:

$$\begin{aligned} |\varphi_f(Y, a)| &\leq b \text{ a.s.}, && \text{(boundedness)} \\ \text{Var } \varphi_f(Y, a) &\leq c \bar{\varphi}_f(a), && \text{(variance condition)} \\ \Delta(f(a), \eta(a)) &\leq \gamma \bar{\varphi}_f(a). && \text{(triangle condition)} \end{aligned}$$

Theorem 4.1 (Regret bound for ℓ -UCB in bandits). Fix $\delta \in (0, 1)$, $n \in \mathbf{N}_+$, bandit instance \mathcal{P} ,
 model class \mathcal{F} and a loss function ℓ . Suppose that $(\mathcal{P}, \mathcal{F}, \ell)$ satisfy Assumptions 3.1 to 3.3. Let
 $h_t = e + \log(1 + t)$ for each $t \in [n]$, let N_n denote the $1/n$ -covering number of the loss class $\Phi(\mathcal{F})$
 with respect to the uniform metric, and let

$$\beta_t = 5/2 + 15(b + c) \log(N_n h_t / \delta), \quad t \in \mathbf{N}_+.$$

Let $\mathcal{F}' \subset \mathcal{F}$, and denote by d_n the $1/n$ -eluder dimension of the localised expected excess loss class
 $\bar{\Phi}(\mathcal{F}')$. Define our complexity measure

$$\Gamma_n = \gamma((d_n + 1)b + d_n \beta_n \log(nb) + 1).$$

450 Suppose a learner uses Algorithm 1, ℓ -UCB, over the course of n -many interactions with \mathcal{P} , with
 451 model class \mathcal{F} and loss function ℓ and with confidence widths $(\beta_t)_{t \in \mathbf{N}_+}$.

Then, with probability at least $1 - \delta$, the learner's regret is bounded as

$$R_n \leq 3\sqrt{n\eta(a_*)\Gamma_n} + 6\Gamma_n + \text{card}\{t \leq n: f_t \notin \mathcal{F}'\}.$$

452 Our bandit result and our reinforcement learning result in the next section both hinge upon the
 453 following uniform Bernstein inequality, which we shall apply to the excess loss class:

454 **Theorem A.1** (Uniform Bernstein inequality). *Let \mathcal{Z} be a set, $(Z_t)_t$ be a \mathcal{Z} -valued process adapted
 455 to a filtration $(\mathbf{F}_t)_t$, and Φ a set of real-valued functions on \mathcal{Z} . Assume that:*

456 1. *For some $b > 0$, for all $\varphi \in \Phi$, $t \in \mathbf{N}_+$, $\mathbf{E}[\varphi(Z_t) \mid \mathbf{F}_{t-1}] \leq b + \varphi(Z_t)$ almost surely.*

457 2. *For some $c > 0$, for all $\varphi \in \Phi$ and $t \in \mathbf{N}_+$, $\text{Var}(\varphi(Z_t) \mid \mathbf{F}_{t-1}) \leq c\mathbf{E}[\varphi(Z_t) \mid \mathbf{F}_{t-1}]$.*

458 Let $\delta \in (0, 1)$, $\varepsilon > 0$ and let N be the ε -covering number of Φ in the uniform metric. For any
 459 $n \in \mathbf{N}_+$, define

$$\beta(n, \delta, \varepsilon, N) = \frac{5n\varepsilon}{2} + 15(b + c) \log(Nh_n/\delta),$$

460 where $h_n = e + \log(1 + n)$. Then, with probability at least $1 - \delta$, for all $\varphi \in \Phi$ and $n \in \mathbf{N}_+$,

$$\sum_{t=1}^n \mathbf{E}[\varphi(Z_t) \mid \mathbf{F}_{t-1}] \leq 2 \sum_{t=1}^n \varphi(Z_t) + 2\beta(n, \delta, \varepsilon, N).$$

461 The proof of the above concentration inequality is contained in Appendix A.1. The rest of this section
 462 contains the proof of Theorem 4.1.

463 **Validity of confidence sequence** Let \mathbf{F} be the filtration $\mathbf{F}_t = \sigma(A_1, Y_1, \dots, A_t, Y_t, A_{t+1})$ for
 464 each $t \in \mathbf{N}$. We apply our uniform Bernstein inequality (Theorem A.1) to the \mathbf{F} adapted pro-
 465 cess $(Y_t, A_t)_{t \in \mathbf{N}_+}$ with the function class $\Phi = \{\varphi_f: f \in \mathcal{F}\}$ and the choice $\varepsilon = 1/n$ (the two
 466 requirements in Theorem A.1 are satisfied due to the boundedness and variance condition parts of
 467 Assumption 3.3). From this, we conclude the first part of the following proposition:

468 **Proposition A.2.** *There exists an event \mathcal{E}_δ satisfying $\mathbf{P}(\mathcal{E}_\delta) \geq 1 - \delta$, whereon, for all $f \in \mathcal{F}$ and
 469 $t \in \mathbf{N}_+$,*

$$\sum_{i=1}^t \bar{\varphi}_f(A_i) \leq 2 \sum_{i=1}^t \varphi_f(Y_i, A_i) + 2\beta_t. \quad (2)$$

Moreover, on \mathcal{E}_δ ,

$$\eta \in \cap_{t \in \mathbf{N}_+} \mathcal{F}_t.$$

Proof. The second conclusion of Proposition A.2, that on \mathcal{E}_δ , $\eta \in \cap_{t \in \mathbf{N}_+} \mathcal{F}_t$, is not immediate.
 For this, observe that the left-hand side of Equation (2) is nonnegative (as ensured by the triangle
 condition of Assumption 3.3), we conclude that on \mathcal{E}_δ , for all $t \in \mathbf{N}_+$,

$$0 \leq \inf_{\hat{f} \in \mathcal{F}} \sum_{i=1}^t \varphi_{\hat{f}}(Y_i, A_i) + \beta_t \iff \sum_{i=1}^t \ell(Y_i, \eta(A_i)) \leq \inf_{\hat{f} \in \mathcal{F}} \sum_{i=1}^t \ell(Y_i, \hat{f}(A_i)) + \beta_t.$$

470 Comparing the right-hand side of the above implication with the form of our confidence set \mathcal{F}_t yields
 471 the second conclusion. \blacksquare

472 **Per-step regret bound** Bounding the per-step regret will use the following simple inequality for
 473 the triangle discrimination, based on an inequality of Ayoub et al. (2024).

474 **Lemma A.3.** *For $x, y, z > 0$ with $y \leq z$, we have that $x - z \leq 3\sqrt{z\Delta(x, y)} + 6\Delta(x, y)$.*

475 **Lemma A.4** (Lemma B.7 of Ayoub et al. (2024)). *For $x, z \geq 0$, $z \leq 3x + \Delta(x, z)$.*

476 **Proof of Lemma A.3.** Observe that

$$\begin{aligned}
x - z &\leq x - y && \text{(assumption)} \\
&= \sqrt{x + y} \sqrt{\Delta(x, y)} && \text{(defn. } \Delta(x, y)) \\
&\leq \sqrt{4x + \Delta(x, y)} \sqrt{\Delta(x, y)} && \text{(Lemma A.4)} \\
&= 2\sqrt{x\Delta(x, y)} + \Delta(x, y) && (3)
\end{aligned}$$

Hence, applying Young's inequality, we obtain the inequality

$$x \leq 2\sqrt{x\Delta(x, y)} + \Delta(x, y) + z \leq \frac{x}{2} + 3\Delta(x, y) + z,$$

477 which yields that $x \leq 6\Delta(x, y) + 2z$; using this and Equation (3) gives

$$x - z \leq 2\sqrt{(6\Delta(x, y) + 2z)\Delta(x, y)} + \Delta(x, y).$$

478 We finish by applying $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$ in the above, and bounding constants. ■

We now apply Lemma A.3 to bound per-step regret. For this, note that on \mathcal{E}_δ , for any $t \in \mathbf{N}_+$, by definition of the pair (f_t, A_t) and since $\eta \in \mathcal{F}_t$, we have

$$f_t(A_t) \leq \eta(A_t).$$

479 Hence, we may apply Lemma A.3 with $x = \eta(A_t)$, $y = f_t(A_t)$ and $z = \eta(a_\star)$ to obtain the bound

$$r_t := \eta(A_t) - \eta(a_\star) \leq 3\sqrt{\eta(a_\star)\Delta(\eta(f_t(A_t)), \eta(A_t))} + 6\Delta(f_t(A_t), \eta(A_t)). \quad (4)$$

480 **Regret decomposition** Let $I_n = \{t \leq n : f_t \in \mathcal{F}'\}$ and observe that the maximal per-step regret is
481 bounded by 1, by Assumption 3.1. Thus, for any $n \in \mathbf{N}_+$,

$$R_n = \sum_{t=1}^n r_t = \sum_{t \in I_n} r_t + \text{card}([n] \setminus I_n).$$

Using Equation (4), Cauchy-Schwarz, and that $\text{card } I_n \leq n$, we have that on \mathcal{E}_δ ,

$$\sum_{t \in I_n} r_t \leq 3\sqrt{n\eta(a_\star) \sum_{t \in I_n} \Delta(f_t(A_t), \eta(A_t))} + 6 \sum_{t \in I_n} \Delta(f_t(A_t), \eta(A_t)).$$

Bounding the triangles The result will be complete once we establish that for all $n \in \mathbf{N}_+$,

$$\sum_{t \in I_n} \Delta(f_t(A_t), \eta(A_t)) \leq \Gamma_n.$$

To this end, we first use the triangle condition of Assumption 3.3 to obtain

$$\sum_{t \in I_n} \Delta(f_t(A_t), \eta(A_t)) \leq \gamma \sum_{t \in I_n} \bar{\varphi}_{f_t}(A_t).$$

482 Now, consider carefully the following proposition from Liu et al. (2022), and the lemma thereafter,
483 which shall allow us to apply the proposition to bound the above sum:

Proposition A.5 (Proposition 21, Liu et al. (2022)). *Let \mathcal{X} be a set and Ψ a set of real-valued functions on \mathcal{X} . Suppose that the functions in Ψ are uniformly bounded by some $B > 0$. Let ψ_1, \dots, ψ_n be a sequence in Ψ and x_1, \dots, x_n a sequence in \mathcal{X} , such that for some $\beta > 0$, for all $t \leq n$, $\sum_{i=1}^{t-1} \psi_t(x_i) \leq \beta$. Then, for all $\omega > 0$ and $t \leq n$,*

$$\sum_{i=1}^t \psi_i(x_i) \leq (d+1)B + d\beta \log(B/\omega) + t\omega,$$

484 where d is the ω -Eluder dimension of Ψ .

485 **Proposition A.6.** *On the event \mathcal{E}_δ of Proposition A.2, we have that for all $t \in \mathbf{N}_+$,*

$$\sum_{i=1}^{t-1} \bar{\varphi}_{f_t}(A_i) \leq 4\beta_t.$$

486 **Proof of Proposition A.6.** On \mathcal{E}_δ , for any $t \in \mathbf{N}_+$,

$$\begin{aligned} \sum_{i \in I_{t-1}} \bar{\varphi}_{f_t}(A_i) &\leq \sum_{i=1}^{t-1} \bar{\varphi}_{f_t}(A_i) && \text{(by the triangle condition, } \bar{\varphi} \text{ is nonnegative)} \\ &\leq 2 \left[\sum_{i=1}^{t-1} \varphi_{f_t}(Y_i, A_i) + \beta_t \right] && \text{(on } \mathcal{E}_\delta \text{ Equation (2) holds)} \\ &= 2 \left[\sum_{i=1}^{t-1} \ell(Y_i, f_t(A_i)) - \sum_{i=1}^{t-1} \ell(Y_i, \eta(A_i)) + \beta_t \right] \\ &\leq 2 \left[\sum_{i=1}^{t-1} \ell(Y_i, f_t(A_i)) - \inf_{\hat{f} \in \mathcal{F}} \sum_{i=1}^{t-1} \ell(Y_i, \hat{f}(A_i)) + \beta_t \right] && \text{(on } \mathcal{E}_\delta, \eta \in \mathcal{F}_t) \\ &\leq 4\beta_t. && (f_t \in \mathcal{F}_t \text{ and the definition of } \mathcal{F}_t) \end{aligned}$$

487 ■

488 With Proposition A.6, for any $n \in \mathbf{N}_+$, we may apply Proposition A.5 with $\beta := 4\beta_n$, $\omega = 1/n$, and
489 with the upper bound b from Assumption 3.3, to conclude that on \mathcal{E}_δ ,

$$\gamma \sum_{t \in I_n} \bar{\varphi}_{f_t}(A_t) \leq \gamma((d_n + 1)b + 4d_n\beta_n \log(nb) + 1) = \Gamma_n. \quad \blacksquare$$

490 **A.1 Proof of the uniform Bernstein inequality, Theorem A.1**

491 To prove Theorem A.1, we will need the following definitions and results:

492 **Definition A.1** (CGF-like). We say a twice differentiable function $\psi : [0, c) \rightarrow \mathbf{R}_+$ is *CGF-like* if ψ
493 is strictly convex, $\psi(0) = \psi'(0) = 0$ and $\psi''(0)$ exists.

Definition A.2 (sub- ψ process). Let \mathbf{F} be a filtration, $\psi : [0, c) \rightarrow \mathbf{R}_+$ be a CGF-like function and let $(S_t)_{t \geq 0}$ and $(V_t)_{t \geq 0}$ be respectively \mathbf{R} -valued and \mathbf{R}_+ -valued \mathbf{F} -adapted processes. We say that $(S_t, V_t)_{t \geq 0}$ is a sub- ψ process if, for every $\lambda \in [0, c)$, there exists an \mathbf{F} -adapted supermartingale $L(\lambda)$ such that

$$M_t(\lambda) := \exp\{\lambda S_t - \psi(\lambda) V_t\} \leq L_t(\lambda) \quad \text{almost surely for all } t \geq 0.$$

494 **Definition A.3** (Sub-gamma process). We say that a random process $(S_t, V_t)_t$ is sub-gamma with
495 parameter $\vartheta > 0$ if it is sub- ψ for the CGF-like function $\psi : [0, 1/\vartheta) \rightarrow \mathbf{R}$ mapping $\lambda \mapsto \lambda^2/(2(1 - \vartheta\lambda))$.
496

Theorem A.7 (Sub-gamma concentration). *For a sub-gamma process $(S_t, V_t)_t$ with parameter $c > 0$, and any $\rho > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$S_t \leq 4\sqrt{V_t \log(H_t/\delta)} + 11(c + \rho) \log(H_t/\delta) \quad \text{where} \quad H_t = \log(1 + V_t/\rho^2) + 2.$$

497 Theorem A.7 is an application of Theorem 3.1 of Whitehouse et al. (2023) (Theorem A.9 here) to
498 subgamma processes. We provide a proof immediately after the proof of Theorem A.1.

499 **Proposition A.8.** *Let \mathbf{F} be a filtration and let X be an \mathbf{F} -adapted, centred random process uniformly
500 bounded by some $b > 0$. Then, for*

$$S_t = \sum_{i=1}^t \mathbf{E}[X_t \mid \mathbf{F}_{t-1}] \quad \text{and} \quad V_t = \sum_{i=1}^t \text{Var}(X_t \mid \mathbf{F}_{t-1})$$

501 *the process $(S_t, V_t)_{t \in \mathbf{N}_+}$ is sub-gamma with parameter $b/3$.*

Proof of Proposition A.8. This result is rather standard, so we only sketch it. If the random variables $X = (X_t)_{t \in \mathbb{N}_+}$ are independent, the result follows directly from Theorem 2.10 (Bernstein's inequality) in Boucheron et al. (2013) and the discussion immediately after Corollary 2.11 therein. For the adapted process structure assumed here, simply use the tower property of the conditional expectation to extend the result from the independent case. ■

Proof of Theorem A.1. We will write \mathbf{E}_{t-1} and Var_{t-1} to denote \mathbf{F}_{t-1} -conditional expectation and variance operators, respectively. Now let $\Phi(\varepsilon)$ be a minimal uniform ε -cover of Φ and let N denote its cardinality. Then for any $\varphi \in \Phi$ there exists some $\hat{\varphi} \in \Phi(\varepsilon)$, such that for any $n \in \mathbb{N}_+$,

$$\sum_{t=1}^n \mathbf{E}_{t-1} \varphi(Y_t) - \varphi(Y_t) \leq 2n\varepsilon + \sum_{t=1}^n \mathbf{E}_{t-1} \hat{\varphi}(Y_t) - \hat{\varphi}(Y_t).$$

Now observe that for any $\hat{\varphi} \in \Phi(\varepsilon)$, from Proposition A.8 and our assumptions on Φ , we have that

$$\left(\sum_{i=1}^t \mathbf{E}_{i-1} \hat{\varphi}(Y_i) - \hat{\varphi}(Y_i), \sum_{i=1}^t \text{Var}_{i-1} \hat{\varphi}(Y_i) \right)_{t \in \mathbb{N}_+}$$

is a sub-gamma process with parameter $b/3$. Applying Theorem A.7 with $\rho = b$ and a confidence parameter δ/N , and taking a union bound over the N functions in $\Phi(\varepsilon)$, we conclude that

$$\sum_{t=1}^n \mathbf{E}_{t-1} \hat{\varphi}(Y_t) - \hat{\varphi}(Y_t) \leq 4 \sqrt{\sum_{i=1}^n \text{Var}_{i-1} \hat{\varphi}(Y_i) \log(Nh_n/\delta)} + \frac{44b}{3} \log(Nh_n/\delta)$$

where we have upper bounded the H_n therein, defined as in Theorem A.7, by $h_n = e + \log(1 + n)$. Next, by the variance condition and Young's inequality,

$$4 \sqrt{\sum_{i=1}^n \text{Var}_{i-1} \hat{\varphi}(Y_i) \log(Nh_n/\delta)} \leq \frac{1}{2} \sum_{t=1}^n \mathbf{E}_{t-1} \hat{\varphi}(Y_t) + 8c \log(Nh_n/\delta).$$

We arrive at the desired result by bounding $\mathbf{E}_{t-1} \hat{\varphi}(Y_t) \leq \mathbf{E}_{t-1} \varphi(Y_t) + \varepsilon$, combining this with the previous inequalities and bounding the constants slightly for convenience. ■

We now prove Theorem A.7, which is an application of the following result:

Theorem A.9 (Theorem 3.1 of Whitehouse et al. (2023)). *Let $(S_t, V_t)_{t \geq 0}$ be a sub- ψ process for a CGF-like function $\psi : [0, c) \rightarrow \mathbf{R}_+$ satisfying $\lim_{\lambda \uparrow c} \psi'(\lambda) = \infty$. Let $\alpha > 1$, $\beta > 0$, $\delta \in (0, 1)$ and let $h : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ be an increasing function such that $\sum_{k \in \mathbf{N}} 1/h(k) \leq 1$. Define the function $\ell_\beta : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ by*

$$\ell_\beta(v) = \log h \left(\log_\alpha \left(\frac{v \vee \beta}{\beta} \right) \right) + \log \left(\frac{1}{\delta} \right),$$

where, for brevity, we have suppressed the dependence of ℓ_β on α and h . Then,

$$\mathbf{P} \left(\exists t \geq 0 : S_t \geq (V_t \vee \beta) \cdot (\psi^*)^{-1} \left(\frac{\alpha}{V_t \vee \beta} \ell_\beta(V_t) \right) \right) \leq \delta,$$

where ψ^* is the convex conjugate of ψ .

Proof of Theorem A.7. The result follows from applying Theorem A.9 to our sub-gamma process with $\alpha = e$, $\beta = \rho^2$ and $h(k) = (k + 2)^2$, and bounding the result crudely. In particular, for our choices of α and h , we have the bound

$$\ell_{\rho^2}(V_t) = \log(\log(\rho^{-2}V_t \vee 1) + 2)^2 + \log 1/\delta \leq 2 \log((\log(1 + V_t/\rho^2) + 2)/\delta) = 2 \log(H_t/\delta).$$

Now, since for our choice of ψ , $\psi^{*-1}(t) = \sqrt{2t} + tc$, the bound from Theorem A.9 can be further bounded as

$$\begin{aligned} (V_t \vee \beta) \cdot (\psi^*)^{-1} \left(\frac{\alpha}{V_t \vee \beta} \ell_\beta(V_t) \right) &= \sqrt{2e(V_t \vee \rho^2) \ell_{\rho^2}(V_t)} + e c \ell_{\rho^2}(V_t) \\ &\leq 2\sqrt{e(V_t \vee \rho^2) \log(H_t/\delta)} + 2ec \log(H_t/\delta) \\ &\leq 2\sqrt{e V_t \log(H_t/\delta)} + 2(\rho\sqrt{e} + ce) \log(H_t/\delta), \end{aligned}$$

where the final inequality uses that for $a, b > 0$, $\sqrt{a \vee b} \leq \sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ and that since $\log(H_t/\delta) \geq 1$, $\sqrt{\log(H_t/\delta)} \leq \log(H_t/\delta)$. ■

521 A.2 Establishing the variance condition

We now establish that the variance condition holds for excess losses classes associated with the log-loss and Poisson loss, denoted respectively Φ_X and Φ_P , under the boundedness assumption: for all $\varphi \in \Phi_X \cup \Phi_P$, and all $(y, x) \in [0, 1] \times \mathbf{B}_2^d$,

$$|\varphi(y, x)| \leq b.$$

522 The boundedness assumption relies on the choice of the model class \mathcal{F} , and not only on the properties
523 of the loss functions. In Appendix C.1, we will verify that for the associated GLMs with parameter
524 norm at most $S > 0$, the boundedness condition holds with $b = 4S$.

525 We establish the variance condition for both losses using following lemma, which is itself a simplification
526 of Lemma 10 in Erven et al. (2012) (Lemma A.11 here).

527 **Lemma A.10.** *Let Ψ be a set of real-valued functions on set \mathcal{Z} and suppose that Z is a \mathcal{Z} -valued*
528 *random variable. Suppose that all $\psi \in \Psi$ satisfy*

- 529 1. *Boundedness:* $|\psi(Z)| \leq b < \infty$; and
- 530 2. *Stochastic mixability:* $\mathbf{E}[\exp\{-\psi(Z)/2\}] \leq 1$.

531 *Then, for all $\psi \in \Psi$, $\text{Var } \psi(Z) \leq (b + 4)\mathbf{E}\psi(Z)$.*

Lemma A.11 (Lemma 10, Erven et al. (2012)). *Let $g(x) = (e^x - x - 1)/x^2$ for $x \neq 0$ and $g(0) = 1/2$, and let X be a random variable satisfying $|X| \leq b$. Then, for all $t > 0$, there exists a $C_t \geq g(-tB)$ such that*

$$\mathbf{E}X = \frac{1}{t}(1 - \mathbf{E}e^{-tX}) + C_t t \mathbf{E}X^2.$$

Proof of Lemma A.10. Let $X = \psi(Z)$. Applying Lemma A.11, with $t = 1/2$, we obtain that there exists a $C \geq g(-B/2)$ such that

$$\mathbf{E}X \geq 2(1 - \mathbf{E}e^{-X/2}) + \frac{C}{2}\mathbf{E}X^2 \geq \frac{C}{2}\mathbf{E}X^2 \geq \frac{g(-B/2)}{2}\mathbf{E}X^2.$$

532 The result follows by using the numerical inequality $g(x) \geq 1/(2 - x)$ that holds for all $x \leq 0$; that
533 $\mathbf{E}X^2 \geq \text{Var } X$ for every random variable with a finite variance; and rearranging. ■

534 **Proposition A.12** (Log-loss variance condition). *Every $\varphi \in \Phi_X$ uniformly bounded by $b > 0$ satisfies*
535 *the variance condition with constant $c = b + 4$.*

Proof. The result follows from Lemma A.10 combined with that every $\varphi \in \Phi_X$ is stochastically mixable, which we establish now. Observe that every $\varphi \in \Phi_X$ is of the form

$$\varphi(y, x) = -\log\left(\frac{f(x)}{\eta(x)}\right)^y - \log\left(\frac{1 - f(x)}{1 - \eta(x)}\right)^{1-y}$$

536 for some $f \in \mathcal{F}$. Therefore,

$$\begin{aligned} \mathbf{E}e^{-\frac{\varphi(y, x)}{2}} &= \mathbf{E}\left[\left(\frac{f(X)}{\eta(X)}\right)^{\frac{y}{2}} \left(\frac{1 - f(X)}{1 - \eta(X)}\right)^{\frac{1-y}{2}}\right] \\ &\leq \mathbf{E}\left[\frac{\mathbf{E}[Y | X] f(X)}{\eta(X)} + \frac{1 - \mathbf{E}[Y | X] 1 - f(X)}{1 - \eta(X)}\right] = \frac{1}{2}, \end{aligned}$$

537 where the inequality follows by the application of AM-GM, and then using the tower rule to condition
538 on X . The final equality follows by recalling that $\mathbf{E}[Y | X] = \eta(X)$. ■

539 **Proposition A.13** (Poisson loss variance condition). *Every $\varphi \in \Phi_P$ uniformly bounded by $b > 0$*
540 *satisfies the variance condition with $c = b + 2$.*

Proof. We establish that for every $\varphi \in \Phi_P$, 2φ is stochastically mixable. The result then follows from Lemma A.10, after looking at how each side of the variance condition scales with the change $\varphi \mapsto 2\varphi$. Observe that every $\varphi \in \Phi_P$ is of the form

$$\varphi(y, x) = -(\eta(x) - f(x)) - y \log\left(\frac{f(x)}{\eta(x)}\right)^y$$

541 for some $f \in \mathcal{F}$. Thus,

$$\begin{aligned} \mathbf{E} \exp\{-\varphi(Y, X)\} &= \mathbf{E} \left[\exp\{\eta(X) - f(X)\} \left(\frac{f(X)}{\eta(X)} \right)^Y \right] \\ &= \mathbf{E} \left[\exp\{\eta(X) - f(X)\} \mathbf{E} \left[\left(\frac{f(X)}{\eta(X)} \right)^Y \mid X \right] \right]. \end{aligned} \quad (\text{tower rule})$$

542 Now, noting that for any $a > 0$ and $y \in [0, 1]$, by convexity, $a^y \leq ay + 1 - y$, and recalling that
 543 $\mathbf{E}[Y \mid X] = \eta(X)$ (Assumption 3.2), we have the bound

$$\begin{aligned} \mathbf{E} \left[\left(\frac{f(X)}{\eta(X)} \right)^Y \mid X \right] &\leq f(X) \frac{\mathbf{E}[Y \mid X]}{\eta(X)} + 1 - \mathbf{E}[Y \mid X] \\ &= 1 + f(X) - \eta(X) \\ &\leq \exp\{f(X) - \eta(X)\}. \end{aligned} \quad (1 + x \leq e^x \text{ for all } x \in \mathbf{R})$$

544 Hence, for all $\varphi \in \Phi_P$, $\mathbf{E} \exp\{-(2\varphi(Y, X))/2\} \leq 1$, which is what we sought to establish. \blacksquare

545 A.3 Establishing the triangle condition

546 We now establish that the excess losses classes associated with the log-loss (Φ_X) and Poisson loss (Φ_P)
 547 satisfy the triangle condition with constants $\gamma_X = 2/\log_2(e)$ and $\gamma_P = 4\sqrt{e}/\log_2(e)$ respectively,
 548 for any random variables Y with support bounded on $[0, 1]$.

549 We first sandwich Δ within a constant multiple of an easier-to-work-with quantity:

550 **Lemma A.14.** *For any $p, q \in [0, 1]$,*

$$(\sqrt{p} - \sqrt{q})^2 \leq \Delta(p, q) \leq 2(\sqrt{p} - \sqrt{q})^2$$

551 **Proof.** Using the algebraic identity $(a - b)(a + b) = a^2 - b^2$, we have

$$(\sqrt{p} + \sqrt{q})^2 (\sqrt{p} - \sqrt{q})^2 = (p - q)^2.$$

552 Rearranging the above display gives the lower bound:

$$(\sqrt{p} - \sqrt{q})^2 = \frac{(p - q)^2}{(\sqrt{p} + \sqrt{q})^2} \leq \frac{(p - q)^2}{p + q} = \Delta(p, q).$$

553 For the upper bound, note that

$$\Delta(p, q) = \frac{(p - q)^2}{p + q} \leq 2 \frac{(p - q)^2}{(\sqrt{p} + \sqrt{q})^2} \leq 2(\sqrt{p} - \sqrt{q})^2. \quad \blacksquare$$

554 We will also need the following relation between the squared Hellinger distance and Kullback-Leibler
 555 divergence, which appears as Equation 7.33 in Polyanskiy and Wu (2025).

Proposition A.15. *For any two measures P, Q on the same measurable space with densities p and q with respect to some common dominating measure μ ,*

$$\text{KL}(P\|Q) \geq \log_2 e \cdot H^2(P, Q) \quad \text{where} \quad H^2(P, Q) := \int (\sqrt{p} - \sqrt{q})^2 d\mu.$$

556 **Proposition A.16** (Log-loss triangle condition). *The expected excess log-loss class $\bar{\Phi}_X$ satisfies the*
 557 *triangle condition with constant $\gamma = 2/\log_2(e)$.*

558 **Proof.** Let P, Q be Bernoulli distributions with parameters $p, q \in [0, 1]$ respectively, and recall that

$$H^2(P, Q) = (\sqrt{p} - \sqrt{q})^2 + (\sqrt{1 - p} + \sqrt{1 - q})^2$$

559 and

$$\text{KL}(P\|Q) = q \log \frac{q}{p} + (1 - q) \log \frac{1 - q}{1 - p}.$$

560 Using Lemma A.14 and Proposition A.15, we have that

$$\Delta(p, q) \leq 2(\sqrt{p} - \sqrt{q})^2 \leq 2H(P, Q) \leq (2/\log_2(e)) \text{KL}(P\|Q).$$

561 We conclude by observing that for any random variable $Y \in [0, 1]$ with mean q ,

$$\mathbf{E}[\ell_X(Y, p) - \ell_X(Y, q)] = \mathbf{E}\left[Y \log \frac{q}{p} + (1 - Y) \log \frac{1 - q}{1 - p}\right] = \text{KL}(P\|Q). \quad \blacksquare$$

562 **Proposition A.17** (Poisson loss triangle condition). *The expected excess log-loss class $\bar{\Phi}_P$ satisfies*
 563 *the triangle condition with constant $\gamma = 4\sqrt{e}/\log_2(e)$.*

564 **Proof.** Let P, Q be Poisson distributions with parameters $p, q \in [0, 1]$ respectively, and recall that

$$H(P, Q) = 1 - \exp\{-(\sqrt{p} - \sqrt{q})^2/2\} \quad \text{and} \quad \text{KL}(Q\|P) = p - q + q \log \frac{q}{p}.$$

565 Observe that for all $x \in [0, 1]$, we have the numerical inequality

$$1 - e^{-x/2} \geq x/(2\sqrt{e}).$$

566 Hence,

$$\begin{aligned} \Delta(p, q) &\leq 2(\sqrt{p} - \sqrt{q})^2 && \text{(Lemma A.14)} \\ &\leq 4\sqrt{e}(1 - e^{-(\sqrt{p} - \sqrt{q})^2/2}) && \text{(Equation (A.3))} \\ &= 4\sqrt{e}H^2(P, Q) && \text{(definition of } H^2) \\ &= 4\sqrt{e}H^2(Q, P) && \text{(symmetry of } H^2) \\ &\leq \frac{4\sqrt{e}}{\log_2(e)} \text{KL}(Q\|P). && \text{(Proposition A.15)} \end{aligned}$$

567 Now, observe that for any random variable $Y \in [0, 1]$ with $\mathbf{E}Y = q$,

$$\mathbf{E}[\ell_P(Y, p) - \ell_P(Y, q)] = \mathbf{E}\left[p - q + y \log \frac{q}{p}\right] = \text{KL}(Q\|P). \quad \blacksquare$$

568 B Proof of Theorem B.1, RL cost sensitive regret bound

569 The proof of our result in RL follows almost identically to the proof of the bandit setting, with an
 570 additional non-expansion argument that is standard when analyzing fitted-Q-iteration style algorithms.
 571 More specifically, the proof proceeds as follows:

- 572 • As in the bandit setting, we begin by leveraging a uniform Bernstein inequality (The-
 573 orem A.1) to construct confidence sets, this time for value functions at each step of the
 574 horizon.
- 575 • Then, in Proposition B.2, we use this inequality to establish a high-probability event \mathcal{E}_δ . On
 576 this event, our confidence sets \mathcal{F}^t are guaranteed to contain the true optimal Q-function q^* ,
 577 and a per-step concentration bound holds.
- 578 • Next, we decompose the total regret. We separate the regret into terms corresponding to
 579 episodes where the chosen model f_t is in a restricted class \mathcal{F}' (well-behaved episodes) and
 580 episodes where it is not (rogue episodes).
- 581 • For well-behaved episodes, we bound the regret by applying the properties of the good
 582 event \mathcal{E}_δ , the RL triangle condition from Assumption B.1, and a key contraction argument
 583 (Lemma B.3).
- 584 • The cumulative excess Bellman error is then bounded using an eluder dimension argument
 585 (Proposition A.5 and Lemma B.4).

586 Our result for reinforcement learning is thus:

Theorem B.1. Fix $\delta \in (0, 1)$, $n \in \mathbf{N}_+$, MDP M , model classes \mathcal{F} and \mathcal{G} and a loss function ℓ . Suppose that $(M, \mathcal{F}, \mathcal{G}, \ell)$ satisfy Assumptions 5.2, 5.3 and B.1. Let $h_t = e + \log(1 + t)$ for each $t \in [n]$, let N_n be the $1/n$ -covering number of the function class $\Phi(\mathcal{F} \cup \mathcal{G})$ with respect to the uniform metric, and let

$$\beta_t = 5/2 + 15(b + c) \log(N_n h_t / \delta), \quad t \in \mathbf{N}_+.$$

587 Let $\mathcal{F}' \subset \mathcal{F}$, and define \mathcal{Z} to be the set of functions $\mathcal{S} \times \mathcal{A} \mapsto \mathbf{R}$ mapping

$$x \mapsto \sum_{h=1}^H \varphi_h^{f, \mathcal{T}f}(x_h) \quad \text{for some } f \in \mathcal{F}'.$$

588 Let P_f denote the measure on $\mathcal{S} \times \mathcal{A}$ induced by the interconnection of M and the policy greedy with
589 respect to $f \in \mathcal{F}$, and let Ψ be the family of functionals on \mathcal{Z} mapping

$$z \mapsto \int z dP_f \quad \text{for each } f \in \mathcal{F}'.$$

Let d_n denote the $1/n$ -eluder dimension of Ψ . Define

$$\Gamma_n = \gamma((d_n + 1)b + d_n \beta_n \log(nb) + 1).$$

590 Suppose a learner uses Algorithm 2, ℓ -GOLF, over the course of n -many episodes with M , with
591 model classes \mathcal{F} and \mathcal{G} , loss function ℓ and confidence widths $(\beta_t)_{t \in \mathbf{N}_+}$.

Then, with probability at least $1 - \delta$, the learner's regret is bounded as

$$R_n \leq 3\sqrt{Hn\eta(a_*)\Gamma_n} + 6H\Gamma_n + \text{card}\{t \leq n : f_t \notin \mathcal{F}'\}.$$

592 **Definition B.1.** For any $f \in \mathcal{F} \cup \mathcal{G}$, $h \in [H]$, $x \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}$, we let

$$y_h^f : (x, s') \mapsto 1 \wedge (c_h(x) + f_{h+1}^\wedge(s'))$$

593 be the response under the model f . With the same symbols, we define the excess Bellman loss
594 function

$$\varphi_h^{f,g}(x, s') = \ell(y_h^f(x, s'), g_h(x)) - \ell(y_h^f(x, s'), (\mathcal{T}f)_h(x)),$$

595 and the expected excess Bellman loss function

$$\bar{\varphi}_h^{f,g}(x) = \int \varphi_h^f(x, \cdot) dP_h(x).$$

596 With that, our assumptions on the loss function for the reinforcement learning setting mirror the
597 bandit assumptions:

598 **Assumption B.1** (RL loss function assumptions). There exist constants $b, c, \gamma > 0$ such that for all
599 $f \in \mathcal{F} \cup \mathcal{G}$, $h \in [H]$, $x \in \mathcal{S} \times \mathcal{A}$, $S' \sim P_h(x)$, the following hold:

$$|\varphi_h^f(x, S')| \leq b \text{ a.s.}, \quad (\text{RL boundedness})$$

$$\text{Var } \varphi_h^f(x, S') \leq c \bar{\varphi}_h^f(x), \quad (\text{RL variance condition})$$

$$\Delta(f_h(x), (\mathcal{T}f)_h(x)) \leq \gamma \bar{\varphi}_h^f(x). \quad (\text{RL triangle condition})$$

600 Our reinforcement learning proof closely follows that for the bandit setting. Recall the following
601 definitions, assumptions, and the statement of the theorem to be established.

602 **Assumption 5.1.** The costs are non-negative and sum to at most one over each episode.

603 **Definition B.1.** For any $f \in \mathcal{F} \cup \mathcal{G}$, $h \in [H]$, $x \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}$, we let

$$y_h^f : (x, s') \mapsto 1 \wedge (c_h(x) + f_{h+1}^\wedge(s'))$$

604 be the response under the model f . With the same symbols, we define the excess Bellman loss
605 function

$$\varphi_h^{f,g}(x, s') = \ell(y_h^f(x, s'), g_h(x)) - \ell(y_h^f(x, s'), (\mathcal{T}f)_h(x)),$$

606 and the expected excess Bellman loss function

$$\bar{\varphi}_h^{f,g}(x) = \int \varphi_h^f(x, \cdot) dP_h(x).$$

607 **Assumption 5.2** (Realisability). We assume that $q^* \in \mathcal{F}$.

608 **Assumption 5.3** (Generalised completeness). We assume that $\mathcal{TF} \subseteq \mathcal{G}$.

609 **Assumption B.1** (RL loss function assumptions). There exist constants $b, c, \gamma > 0$ such that for all
 610 $f \in \mathcal{F} \cup \mathcal{G}$, $h \in [H]$, $x \in \mathcal{S} \times \mathcal{A}$, $S' \sim P_h(x)$, the following hold:

$$|\varphi_h^f(x, S')| \leq b \text{ a.s.}, \quad (\text{RL boundedness})$$

$$\text{Var } \varphi_h^f(x, S') \leq c\bar{\varphi}_h^f(x), \quad (\text{RL variance condition})$$

$$\Delta(f_h(x), (\mathcal{T}f)_h(x)) \leq \gamma\bar{\varphi}_h^f(x). \quad (\text{RL triangle condition})$$

611 **Proof.** We now prove Theorem B.1.

612 **Validity of the confidence sequence** For each $h \in [H]$, let $\mathbf{F}_h = (\mathbf{F}_h^t)_{t \in \mathbf{N}_+}$ be the filtration given
 613 by

$$\mathbf{F}_h^t = \sigma(X_h^1, S_{h+1}^1, \dots, X_h^t, S_{h+1}^t, X_h^{t+1}) \text{ for each } t \in \mathbf{N}.$$

614 Now for each $h \in [H]$, we apply our uniform Bernstein inequality (Theorem A.1) to the \mathbf{F}_h -adapted
 615 process $((X_h^t, S_{h+1}^t))_{t \in \mathbf{N}}$ with the function class $\Phi_h = \{\varphi_h^{f,g} : (f, g) \in (\mathcal{F} \cup \mathcal{G})^2\}$ and with $\varepsilon = 1/n$.
 616 From this, we conclude the first part of the following proposition:

617 **Proposition B.2.** *There exists an event \mathcal{E}_δ satisfying $\mathbf{P}(\mathcal{E}_\delta) \geq 1 - \delta$, whereon, for all $(f, g) \in (\mathcal{F} \cup \mathcal{G})^2$,
 618 $t \in \mathbf{N}_+$ and $h \in [H]$,*

$$\sum_{i=1}^t \bar{\varphi}_h^{f,g}(X_h^i) \leq 2 \sum_{i=1}^t \varphi_h^{f,g}(X_h^i, S_{h+1}^i) + 2\beta_t,$$

619 where, $\beta_t := \beta(t, \delta, 1/n, N_n)$ for β defined as in Theorem A.1. Moreover, on \mathcal{E}_δ ,

$$q^* \in \cap_{t \in \mathbf{N}_+} \mathcal{F}^t.$$

620 **Proof.** The second conclusion of Proposition A.2, that on \mathcal{E}_δ , $q^* \in \cap_{t \in \mathbf{N}_+} \mathcal{F}^t$, is not immediate.
 621 For that, fix some $h \in [H]$. Now, observe that the left-hand side of Equation (2) is nonnegative (as
 622 ensured by the RL triangle condition of Assumption B.1). Hence, on \mathcal{E}_δ , for all $t \in \mathbf{N}_+$,

$$0 \leq \inf_{g \in \mathcal{G}} \sum_{i=1}^t \varphi_h^{q^*, g}(X_h^i, S_{h+1}^i) + \beta_t,$$

623 which implies that

$$\sum_{i=1}^t \ell(y_h^{q^*}(X^i, S_{h+1}^i), (\mathcal{T}q^*)_h(X_h^i)) \leq \inf_{g \in \mathcal{G}} \sum_{i=1}^t \ell(y_h^{q^*}(X^i, S_{h+1}^i), g_h(X_h^i)) + \beta_t.$$

624 Comparing the above inequality with the form of our confidence set \mathcal{F}_t yields that on \mathcal{E}_δ , we have
 625 that $q^* \in \cap_{t \in \mathbf{N}_+} \mathcal{F}^t$, as desired. ■

626 Let $I_n = \{t \leq n : f_t \in \mathcal{F}'\}$ and observe that the maximal per-step regret is bounded by 1, by
 627 Assumption 3.1. Then,

$$\sum_{t=1}^n v_1^t(S_1) - v_1^*(S_1) \leq \sum_{t \in I_n} (v_1^t(S_1) - v_1^*(S_1)) + \text{card}([n] \setminus I_n)$$

628 We bound only the regret on episodes $t \in I_n$. For this, we observe that on \mathcal{E}_δ , $q^* \in \mathcal{F}^t$ (Propo-
 629 sition B.2). Thus, we can apply our inequality on the triangular discrimination, Lemma A.3, with
 630 $x = v_1^t(S_1)$, $y = f_1^t(S_1)$ and $z = v_1^*(S_1)$, to obtain

$$\sum_{t \in I_n} v_1^t(S_1) - v_1^*(S_1) \leq \sum_{t \in I_n} 3\sqrt{v_1^*(S_1)\Delta(v_1^t(S_1), f_1^t(S_1))} + 6 \sum_{t \in I_n} \Delta(v_1^t(S_1), f_1^t(S_1)).$$

631 **Lemma B.3** (Contraction Lemma). *Let $f \in \mathcal{F}$, and let $\pi := \pi^f$ and $v = v^\pi$. Then,*

$$\sqrt{\Delta(f_1(S_1, \pi(S_1)), v_1(S_1))} \leq \sum_{h=1}^H \sqrt{\mathbf{E}_\pi [\Delta(f_h(X_h), \mathcal{T}f_h(X_h))]},$$

632 *where \mathbf{E}_π denotes the expectation over the state-action pairs $(X_h)_{h=1}^H$ resulting from following the*
 633 *policy π in the MDP M .*

634 We prove Lemma B.3 presently. Now, let $\Delta_{t,h} := \Delta(f_h^t(X_h), (\mathcal{T}f_h^t)(X_h))$, let \mathbf{E}_{π^t} denote the
 635 expectation over trajectories generated by the policy π^t in the MDP M , and observe that

$$\begin{aligned} & 3 \sum_{t \in I_n} \sqrt{v_1^*(S_1) \Delta(v_1^t(S_1), f_1^t(S_1))} + 6 \sum_{t \in I_n} \Delta(v_1^t(S_1), f_1^t(S_1)) \\ & \leq 3 \sum_{t \in I_n} \sum_{h=1}^H \sqrt{v_1^*(S_1) \mathbf{E}_{\pi^t} \Delta_{t,h}} + 6 \sum_{t \in I_n} \left\{ \sum_{h=1}^H \sqrt{\mathbf{E}_{\pi^t} \Delta_{t,h}} \right\}^2 \quad (\text{Lemma B.3}) \\ & \leq 3 \sqrt{H n v_1^*(S_1) \sum_{t \in I_n} \sum_{h=1}^H \mathbf{E}_{\pi^t} \Delta_{t,h}} + 6H \sum_{t \in I_n} \sum_{h=1}^H \mathbf{E}_{\pi^t} \Delta_{t,h} \quad (\text{Cauchy-Schwarz, card } I_n \leq n) \end{aligned}$$

636 Now, from the triangle condition, we have that

$$\sum_{t \in I_n} \sum_{h=1}^H \mathbf{E}_{\pi^t} \Delta_{t,h} \leq \gamma \sum_{t \in I_n} \sum_{h=1}^H \mathbf{E}_{\pi^t} \bar{\varphi}_h^{f^t, \mathcal{T}f^t}(X_h^t), \quad (5)$$

637 and it remains to upper bound the right-hand side by our complexity measure Γ_n . For this, consider
 638 the following result that bounds the cumulative expected excess risk on past observations (this is
 639 broadly equivalent to Proposition A.6 of the bandit proof).

640 **Lemma B.4.** *On \mathcal{E}_δ , for all $t \in \mathbf{N}_+$, $h \in [H]$,*

$$\sum_{i=1}^{t-1} \bar{\varphi}_h^{f^t, \mathcal{T}f^t}(X_h^i) \leq 4\beta_t.$$

641 **Proof.** By Proposition B.2, on \mathcal{E}_δ ,

$$\sum_{i=1}^{t-1} \bar{\varphi}_h^{f^t, \mathcal{T}f^t}(X_h^i) \leq 2 \sum_{i=1}^{t-1} \varphi_h^{f^t, \mathcal{T}f^t}(X_h^i, S_{h+1}^i) + 2\beta_t.$$

642 Now, let g^t denote an element of \mathcal{G} attaining the infimum in the definition of \mathcal{F}^t (for convenience,
 643 suppose that this exists; otherwise, the argument goes through by introducing an approximate
 644 minimiser). Then,

$$\begin{aligned} \sum_{i=1}^{t-1} \varphi_h^{f^t, \mathcal{T}f^t}(X_h^i, S_{h+1}^i) &= \sum_{i=1}^{t-1} \ell(y^{f^t}(X_h^i, S_{h+1}^i), f_h^t(X_h^i)) - \ell(y^{f^t}(X_h^i, S_{h+1}^i), f_h^t(X_h^i)) \\ &= \underbrace{\sum_{i=1}^{t-1} \ell(y^{f^t}(X_h^i, S_{h+1}^i), f_h^t(X_h^i)) - \ell(y^{f^t}(X_h^i, S_{h+1}^i), g_h^t(X_h^i))}_{\leq \beta_t} \quad (\text{using that } f^t \in \mathcal{F}^t \text{ and the definition of } g^t \text{ and } \mathcal{F}^t) \\ &\quad + \underbrace{\sum_{i=1}^{t-1} \ell(y^{f^t}(X_h^i, S_{h+1}^i), g_h^t(X_h^i)) - \ell(y^{f^t}(X_h^i, S_{h+1}^i), (\mathcal{T}f^t)_h(X_h^i))}_{\leq 0} \\ &\quad (\text{using the definition of } g^t \text{ as the empirical risk minimiser over } \mathcal{G}, \text{ and that } \mathcal{T}\mathcal{F} \subset \mathcal{G}) \end{aligned}$$

645 This completes the proof of Lemma B.4. ■

Combining the result in the thus established Lemma B.4 with the usual eluder dimension argument of Proposition A.5, with $\omega = 1/n$ and the upper bound b from Assumption B.1, we obtain that Γ_n is indeed an upper bound on the right-hand side of (5). ■

B.1 Proof of Lemma B.3, Contraction

The following lemma follows from the joint convexity of $(x, y) \mapsto -\sqrt{xy}$ for $x, y \geq 0$:

Lemma B.5. *For $x, y \geq 0$, the map $(x, y) \mapsto (\sqrt{x} - \sqrt{y})^2$ is jointly convex in its arguments.*

Proof of Lemma B.3. Let for each $h \in [H]$, let μ_h denote the distribution of $(S_h, \pi(S_h))$ when following the policy π , and let $\|\cdot\|_{\mu_h}$ denote the $L_2(\mu_h)$ norm.

To start with, by Lemma A.14, we have that

$$\sqrt{\Delta(f_1(S_1, \pi(S_1)), v_1(S_1))} \leq 2\|\sqrt{g_1} - \sqrt{v_1}\|_{\nu_1}.$$

We shall shortly establish the inequality

$$(\forall h \in [H]) \quad \|\sqrt{f_h} - \sqrt{v_h}\|_{\mu_h} \leq \|\sqrt{f_h} - \sqrt{\mathcal{T}f_h}\|_{\mu_h} + \|\sqrt{f_{h+1}} - \sqrt{v_{h+1}}\|_{\mu_{h+1}}. \quad (6)$$

Unrolling this over $h = 1, \dots, H$ and using our boundary condition, we obtain that

$$\|\sqrt{f_1} - \sqrt{v_1}\|_{\nu_1} \leq \sum_{h=1}^H \|\sqrt{f_h} - \sqrt{\mathcal{T}f_h}\|_{\mu_h}.$$

The result follows from applying the other side of Lemma A.14 to the terms above.

We now establish Equation (6). Fix $h \in [H]$. By the triangle inequality,

$$\|\sqrt{f_h} - \sqrt{v_h}\|_{\mu_h} \leq \|\sqrt{f_h} - \sqrt{\mathcal{T}f_h}\|_{\mu_h} + \|\sqrt{\mathcal{T}f_h} - \sqrt{v_h}\|_{\mu_h}.$$

The first term is of the form we want. For the second term, we have

$$\begin{aligned} \|\sqrt{\mathcal{T}f_h} - \sqrt{v_h}\|_{\mu_h} &= \left\| \sqrt{c(\cdot) + \int f_{h+1}^\wedge dP(\cdot)} - \sqrt{c(\cdot) + \int v_{h+1} dP(\cdot)} \right\|_{\mu_h} \\ &\leq \left\| \sqrt{\int f_{h+1}^\wedge dP(\cdot)} - \sqrt{\int v_{h+1} dP(\cdot)} \right\|_{\mu_h} \\ &\quad (\forall c, a, b \geq 0, |\sqrt{c+a} - \sqrt{c+b}| \leq |\sqrt{a} - \sqrt{b}|) \\ &\leq \|\sqrt{f_{h+1}} - \sqrt{v_{h+1}}\|_{\mu_{h+1}}. \quad (\text{Jensen's, justified by Lemma B.5}) \end{aligned}$$

This establishes Equation (6), and with it the lemma. ■

C On self-concordant GLMs with compatible losses

We restate the GLM assumption for convenience:

Assumption 2.1. We make the following assumptions:

	$\mathcal{A} \subset \mathbf{B}_2^d$	(action set bound)
$(\exists S > 0)$	$\Theta \subset S\mathbf{B}_2^d$	(parameter set bound)
$(\forall (a, \theta) \in \mathcal{A} \times \Theta)$	$\langle a, \theta \rangle \in U$	(valid domain)
$(\exists L > 0, \forall u, u' \in U)$	$ \mu(u) - \mu(u') \leq L u - u' $	(L -Lipschitz link)
$(\exists M \geq 1, \forall u \in U^\circ)$	$ \ddot{\mu}(u) \leq M\dot{\mu}(u)$	(M -self-concordant link)
$(\exists 1 \leq \kappa < \infty)$	$\kappa \geq \sup_{u \in U^\circ} 1/\dot{\mu}(u)$	(link derivative lower-bound)
$(\forall y \in [0, 1], \forall u \in U)$	$\partial_u \ell(y, \mu(u)) = \mu(u) - y$	(link and loss are compatible)

In this section, we prove that under our GLM assumption (Assumption 2.1), the excess loss class $\Phi(\mathcal{F})$ is uniformly bounded and admits a rather small uniform cover; that for a suitable localised model class $\mathcal{F}' \subset \mathcal{F}$, the number of rogue steps under our ℓ -UCB algorithm is bounded, and the localised expected excess loss class $\bar{\Phi}(\mathcal{F}')$ has a small eluder dimension. These results combined with Theorem 4.1 yield Proposition C.1.

Proposition C.1. *Let $\delta \in (0, 1)$ and $n \in \mathbf{N}_+$. Consider the setting of Theorem 4.1, with the model structure described in Assumption 2.1. Let*

$$\beta_t = \theta(d(S+1) \log((1+St)/\delta)), \quad t \in \mathbf{N}_+.$$

Then, ℓ -UCB with the confidence widths $(\beta_t)_{t \in \mathbf{N}_+}$ leads to regret that is upper bounded with probability $1 - \delta$ as

$$R_n \leq 3\sqrt{n\eta(a_\star)\Gamma_n} + 6\Gamma_n + R'_n$$

where for some $C > 0$,

$$\Gamma_n \leq C\gamma d \log(1 + S^2 Ln) \cdot (S + \beta_n \log(1 + Sn)) = \tilde{O}(\gamma d^2 S),$$

and for some $C' > 0$,

$$R'_n \leq C'd\kappa M^2 \beta_n \cdot \log(1 + \kappa MS \beta_n) = \tilde{O}(d^2 M^2 S \kappa).$$

We will use the following lemma repeatedly.

Lemma C.2. *Fix some $(y, a) \in [0, 1] \times \mathbf{B}_2^d$ and let $h(\theta) = \ell(y, \mu(\langle a, \theta \rangle))$. Let $\theta, \theta' \in \mathbf{R}^d$ and write $\theta(t) = t\theta + (1-t)\theta'$. Then,*

$$h(\theta) - h(\theta') = \int_0^1 \partial_t h(\theta(t)) dt = \partial_t h(\theta') + \frac{1}{2} \int_0^1 (1-t) \partial_t^2 h(\theta(t)) dt,$$

where

$$\begin{aligned} \partial_t h(\theta(t)) &= (\mu(\langle a, \theta(t) \rangle) - y) \langle a, \theta - \theta' \rangle, \quad \text{and} \\ \partial_t^2 h(\theta(t)) &= \dot{\mu}(\langle a, \theta(t) \rangle) \langle a a^\top (\theta - \theta'), \theta - \theta' \rangle. \end{aligned}$$

Proof sketch. The proof follows from the fundamental theorem of calculus for the first equality, and then a Taylor expansion followed by another application of the fundamental theorem of calculus for the second equality. The absolute continuity requisite for the fundamental theorem of calculus is ensured by the L -Lipschitz continuity of the link function for the first application, and by the second derivative of the loss being bounded, which may be seen from its form, combined with L being an upper bound on $\dot{\mu}(u)$ for all $u \in U^\circ$. ■

C.1 Boundedness of excess losses & covering number bound

We first establish the boundedness of the excess risk class with $b = 4S$, which is implied from the following proposition together with our realisability assumption:

Lemma C.3. *Let $(\mu, \ell, \text{GLM}(\mu, S))$ be compatible according to Assumption 2.1. Then, for all $\theta, \theta' \in SB_2^d$ and $(y, a) \in [0, 1] \times \mathcal{A}$,*

$$|\ell(y, \mu(\langle x, \theta \rangle)) - \ell(y, \mu(\langle a, \theta' \rangle))| \leq 2\|\theta - \theta'\| \leq 4S.$$

Proof. Let $\theta(t) = t\theta + (1-t)\theta'$ and note that for any $(y, a) \in [0, 1] \times \mathcal{A}$, by Lemma C.2 and using the notation defined therein,

$$\begin{aligned} |\ell(y, \mu(\langle a, \theta \rangle)) - \ell(y, \mu(\langle a, \theta' \rangle))| &= \left| \int_0^1 (\mu(\langle a, \theta(t) \rangle) - y) \langle a, \theta - \theta' \rangle dt \right| \\ &\leq \left| \int_0^1 (\mu(\langle a, \theta(t) \rangle) - y) dt \right| \|\langle a, \theta - \theta' \rangle\| \\ &\leq 2\|\theta - \theta'\|. \end{aligned}$$

Now we establish a bound on the corresponding uniform covering number:

691 **Proposition C.4.** *Under Assumption 2.1, the ε -covering number of $\Phi(\text{GLM}(\mu, \Theta))$ with respect to*
692 *the uniform norm is upper bounded by $(1 + 8S/\varepsilon)^d$.*

Proof. Write $\mathcal{F} = \text{GLM}(\mu, \Theta)$. Let $\theta_\star \in \Theta$ be such that $\eta(a) = \mu(\langle a, \theta_\star \rangle)$ (such a parameter exists by Assumption 3.2, realisability) and define the map $\rho: \Theta \rightarrow \mathcal{L}(\mathcal{F})$ as that taking each $\theta \in \Theta$ to the map

$$(y, a) \mapsto \ell(y, \mu(\langle a, \theta \rangle)) - \ell(y, \eta(a)).$$

Observe that $\Phi(\mathcal{F}) = \rho(\Theta)$, and that since for any $\theta_0, \theta_1 \in \Theta$,

$$\|\rho(\theta_0) - \rho(\theta_1)\|_\infty = \sup_{(y, a) \in [0, 1] \times \mathcal{A}} |\ell(y, \mu(\langle a, \theta_0 \rangle)) - \ell(y, \mu(\langle a, \theta_1 \rangle))|,$$

693 we have by Lemma C.3 that ρ is 2-Lipschitz as a map from $(\Theta, \|\cdot\|_2) \rightarrow (\mathcal{L}(\mathcal{F}), \|\cdot\|_\infty)$. Now, if
694 $\mathcal{C}_{\varepsilon/2}$ is an $\varepsilon/2$ -cover of $(SB_2^d, \|\cdot\|_2)$, then by said 2-Lipschitzness, $\rho(\mathcal{C}_{\varepsilon/2})$ is an ε -external-cover
695 of $(\Phi(\mathcal{F}), \|\cdot\|_\infty)$. Finally, the 2-norm $\varepsilon/4$ -covering number of SB_2^d is an upper bound on the
696 $\varepsilon/2$ -covering number of Θ (Vershynin, 2018, Exercise 4.2.9), and the former quantity is at most
697 $(1 + 8S/\varepsilon)^d$ (Vershynin, 2018, Corollary 4.2.13). ■

698 C.2 Rogue steps bound and the localised eluder dimension

699 In the following, we associate with each function f_t selected by the ℓ -UCB algorithm a parameter
700 $\theta_t \in \Theta$ such that $f_t(\cdot) = \mu(\langle \cdot, \theta_t \rangle)$, and localise the GLM to the function class

$$\mathcal{F}' = \{\mu(\langle \cdot, \theta \rangle) : \theta \in \Theta'\} \quad \text{for} \quad \Theta' = \{\theta \in \Theta : \forall a \in \mathcal{A}, |\langle a, \theta - \theta_\star \rangle| \leq 1/M\}.$$

701 By realisability, we can write any $\bar{\varphi} \in \bar{\Phi}(\text{GLM}(\mu, S))$ in the form

$$\bar{\varphi}(a) = \int \ell(y, \mu(\langle x, \theta \rangle)) - \ell(y, \mu(\langle x, \theta_\star \rangle)) P_a(dy) := \bar{\varphi}(a, \theta) \quad \text{for some } \theta, \theta_\star \in SB_2^d.$$

702 **Lemma C.5.** *For any $\theta \in \mathbf{R}^d$, letting $\theta(t) = t\theta + (1-t)\theta_\star$ for $t \in [0, 1]$, we have that*

$$\bar{\varphi}(a, \theta) = \frac{1}{2} \|\theta - \theta_\star\|_{\alpha(a, \theta)aa^\top}^2 \quad \text{where} \quad \alpha(a, \theta) = \int_0^1 (1-t) \dot{\mu}(\langle a, \theta(t) \rangle) dt.$$

Moreover, there exists a real number $\zeta(a, \theta) \in \{\langle a, \theta(t) \rangle : t \in [0, 1]\}$ such that

$$\dot{\mu}(\zeta(a, \theta)) = \alpha(a, \theta).$$

703 **Proof.** By Lemma C.2, for any $(y, a) \in [0, 1] \times \mathbf{B}_2^d$,

$$\ell(y, \mu(\langle a, \theta \rangle)) - \ell(y, \mu(\langle a, \theta_\star \rangle)) = (\mu(\langle a, \theta_\star \rangle) - y) \langle a, \theta - \theta_\star \rangle + \frac{1}{2} \alpha(a, \theta) \langle aa^\top (\theta - \theta_\star), \theta - \theta_\star \rangle.$$

704 Integrating both sides with respect to $P_a(dy)$, and noting that, by our realisability assumption,
705 $\int y P_a(dy) = \mu(\langle a, \theta_\star \rangle)$, which leads to the first term dropping out, we obtain

$$\int \ell(y, \mu(\langle a, \theta \rangle)) - \ell(y, \mu(\langle a, \theta_\star \rangle)) P_a(dy) = \frac{1}{2} \|\theta - \theta_\star\|_{\alpha(a, \theta)aa^\top}^2.$$

706 For the second result, repeat the argument with the Lagrange form of the remainder. ■

707 C.2.1 Proof of bound on the number of rogue steps, Proposition 4.4

708 We will use the following extension of exercise 19.3 in Lattimore and Szepesvári (2020).

709 **Lemma C.6** (Lemma 2 of Janz et al. (2024)). *For any $\lambda, \gamma > 0$, the number of times $\|a_t\|_{V_{t-1}^{-1}(\lambda)} \geq \gamma$*
710 *is no greater than*

$$\frac{3d}{\log(1 + \gamma^2)} \log \left(1 + \frac{1}{\lambda \log(1 + \gamma^2)} \right).$$

711 **Proof of Proposition 4.4.** For any $\lambda \geq 0$ and $\theta \in \mathbf{R}^d$, we define the positive semidefinite matrices

$$V_t(\lambda) = \sum_{i=1}^t A_i A_i^\top + \lambda I \quad \text{and} \quad G_t(\theta, \lambda) = \sum_{i=1}^t \alpha(A_i, \theta) A_i A_i^\top + \lambda I,$$

712 where α is defined as in Lemma C.5. Then, using that by Lemma C.5, for any $\theta \in \Theta$, there exists a
 713 $\zeta \in \mathbf{R}$ satisfying $|\zeta| \leq S$ such that $\alpha(A_i, \theta) = \dot{\mu}(\zeta)$ we have that for all $\theta \in \Theta$, from the definition
 714 of κ in Assumption 2.1,

$$V_t(0) \preceq \kappa G_t(\theta, 0).$$

715 Now suppose $t \in \mathbf{N}_+$ is such that $1/M \leq |\langle A_t, \theta_t - \theta_\star \rangle|$. Then,

$$\begin{aligned} 1/M &\leq |\langle A_t, \theta_t - \theta_\star \rangle| \\ &\leq \|A_t\|_{V_{t-1}^{-1}(\lambda)} \|\theta_t - \theta_\star\|_{V_{t-1}(\lambda)} && \text{(Cauchy-Schwarz)} \\ &\leq \|A_t\|_{V_{t-1}^{-1}(\lambda)} \cdot \sqrt{\kappa} \|\theta_t - \theta_\star\|_{G_{t-1}(\theta_t, \lambda)}. && \text{(Equation (C.2.1))} \end{aligned}$$

716 By the triangle inequality and then using Proposition A.6, we have that on \mathcal{E}_δ ,

$$\|\theta_t - \theta_\star\|_{G_{t-1}(\theta_t, \lambda)} \leq \|\theta_t - \theta_\star\|_{G_{t-1}(\theta_t, 0)} + \sqrt{\lambda} \|\theta_t - \theta_\star\| \leq 2(\sqrt{\beta_t} + S\sqrt{\lambda}).$$

717 Hence, taking $\lambda = 1/(S^2 \kappa)$, the number of times that $1/M \leq |\langle A_t, \theta_t - \theta_\star \rangle|$ on \mathcal{E}_δ is no greater
 718 than the number of times that

$$x := \frac{1}{2M(\sqrt{\kappa\beta_t} + 1)} \leq \|A_t\|_{V_{t-1}^{-1}(\lambda)}.$$

719 We lower bound x^2 by $y = 1/(8M^2(\kappa\beta_t + 1))$ and apply Lemma C.6 together with the bound
 720 $\log(1 + y) \geq 3/(4y)$, twice (which holds because $y \leq 1/16$), to obtain that the count in question is
 721 no greater than

$$\frac{4d}{y} \log \left(1 + \frac{4}{\lambda y} \right) \leq 64d\kappa M^2 \beta_t \log \left(1 + \frac{64}{3} \kappa^2 M^2 S^2 \beta_t \right).$$

722 Finally, observe that $\beta_t \leq \beta_n$ for any $t \leq n$. ■

723 C.2.2 Proof of the upper bound on the eluder dimension bound, Proposition 4.3

724 The following proposition is a special case of proposition 8 of Sun and Tran-Dinh (2019). The lemma
 725 thereafter is a simple numerical inequality that will come in handy.

726 **Proposition C.7.** Let $\mu: U \rightarrow [0, 1]$ be an M -self-concordant link function. Then, for any $u, u' \in U^\circ$
 727 satisfying $|u - u'| \leq c$, $\dot{\mu}(u) \leq \exp(cM)\dot{\mu}(u')$.

Lemma C.8. Suppose that $a, x \geq 1$, $b \geq 0$ and $a^x \leq bx + 1$. Then,

$$x \leq \log(1 + b/\log(a))/\log(a).$$

728 **Proof of Lemma C.8.** Let $f(x) = a^x$ and $g(x) = bx + 1$. Since f is convex and g is affine, they
 729 intersect at no more than two points. Since they intersect at 0, we have that the set of x satisfying
 730 $f(x) \leq g(x)$ is of the form $[0, y]$ for some $y \geq 0$. Now, let $y' = \log(1 + b/\log(a))/\log(a)$. Then, a
 731 quick calculation shows that $f(y') > g(y')$. Thus, $y' > y$. ■

732 **Proof of Proposition 4.3.** Let a_1, \dots, a_k and $\theta_1, \dots, \theta_k$ be witness to the eluder dimension in ques-
 733 tion, in that they satisfy

$$\sum_{i=1}^{t-1} \bar{\varphi}(a_i, \theta_t) \leq \omega \quad \text{and} \quad \bar{\varphi}(a_t, \theta_t) \geq \omega$$

734 for some $\omega \geq \varepsilon$ and all $t \leq k$, where k is the ε -eluder dimension. Also, for any $\lambda \geq 0$, define the
 735 positive semidefinite matrix

$$H_{t-1}(\lambda) = \sum_{i=1}^{t-1} \dot{\mu}(\langle a_i, \theta_\star \rangle) a_i a_i^\top + \lambda I$$

For each $i \leq t \leq k$, use Lemma C.5 to construct a real number $\zeta_{i,t}$ on the interval connecting $\langle a_i, \theta_t \rangle$ and $\langle a_i, \theta_\star \rangle$ that satisfies $\dot{\mu}(\zeta_{i,t}) = \alpha(a_i, \theta_t)$. Now, since for all $i \leq t \leq k$,

$$|\zeta_{i,t} - \langle a_i, \theta_\star \rangle| \leq |\langle a_i, \theta_t - \theta_\star \rangle| \leq c,$$

736 we have by Proposition C.7 that, for all $i \leq t \leq k$,

$$\exp(-cM)\dot{\mu}(\langle a_i, \theta_\star \rangle) \leq \dot{\mu}(\zeta_{i,t}) \leq \exp(cM)\dot{\mu}(\langle a_i, \theta_\star \rangle).$$

Hence, using Lemma C.5 and Equation (C.2.2), we have the bound

$$\omega \geq \sum_{i=1}^{t-1} \bar{\varphi}(a_i, \theta_t) \geq \frac{1}{2\exp(cM)} \|\theta_t - \theta_\star\|_{H_{t-1}(0)}.$$

737 Taking $\lambda = \omega/(2S^2)$, this

$$\|\theta_t - \theta_\star\|_{H_{t-1}(\lambda)}^2 \leq \|\theta_t - \theta_\star\|_{H_{t-1}(0)}^2 + \lambda \|\theta_t - \theta_\star\|^2 \leq 2\exp(cM)\omega + 4\lambda S^2 \leq 2\omega(\exp(cM) + 1).$$

738 Now, letting $x_t = \dot{\mu}(\langle a_t, \theta_\star \rangle)^{1/2} a_t$, we have that

$$\begin{aligned} \omega &\leq \bar{\varphi}(a_t, \theta_t) && \text{(definition of } \omega, a_t, \theta_t) \\ &= \frac{\dot{\mu}(\zeta_{t,t})}{2} \langle a_t, \theta_t - \theta_\star \rangle^2 && \text{(definition of } \zeta_{t,t}) \\ &\leq \frac{\exp(cM)}{2} \dot{\mu}(\langle a_t, \theta_\star \rangle) \langle a_t, \theta_t - \theta_\star \rangle^2 && \text{(Equation (C.2.2))} \\ &\leq \frac{\exp(cM)}{2} \dot{\mu}(\langle a_t, \theta_\star \rangle) \|a_t\|_{H_{t-1}^{-1}(\lambda)}^2 \|\theta_t - \theta_\star\|_{H_{t-1}(\lambda)}^2 && \text{(Cauchy-Schwarz)} \\ &\leq \omega \exp(cM) (\exp(cM) + 1) \|x_t\|_{H_{t-1}^{-1}(\lambda)}^2. && \text{(Equation (C.2.2))} \end{aligned}$$

739 Whence, we conclude that for all $t \leq k$,

$$\|x_t\|_{H_{t-1}^{-1}(\lambda)}^2 \geq \exp(-cM) (\exp(cM) + 1)^{-1} =: c.$$

Using this lower bound and the matrix determinant lemma, we have

$$\det H_k(\lambda) = \lambda^d \prod_{t=1}^k (1 + \|x_t\|_{H_{t-1}^{-1}(\lambda)}^2) \geq \lambda^d (1 + c)^k.$$

On the other hand, using the AM-GM inequality and that $\|x_t\|^2 = \dot{\mu}(\langle a_t, x_t \rangle) \|a_t\|^2 \leq L$, we have the upper bound

$$\det H_k(\lambda) \leq \left(\frac{\text{tr}(H_k(\lambda))}{d} \right)^d \leq \left(\lambda + \frac{kL}{d} \right)^d.$$

Putting the two inequalities together yields the inequality

$$(1 + c)^{\frac{k}{d}} \leq \frac{kL}{d\lambda} + 1.$$

740 Now, applying Lemma C.8 with $a = 1 + c$, $x = k/d$ and $b = L/\lambda = 2S^2L/\omega$, we obtain

$$\begin{aligned} k &\leq d \log \left(1 + \frac{2S^2L}{\omega \log(1 + c)} \right) / \log(1 + c) \\ &= d \frac{\log \left(1 + \frac{2S^2L}{\omega \log(1 + \exp(-cM)(1 + \exp(cM))^{-1})} \right)}{\log(1 + \exp(-cM)(1 + \exp(cM))^{-1})} \\ &\leq d \frac{\log(1 + 8S^2L \exp(2cM)/\omega)}{\log(1 + \exp(-cM)(1 + \exp(cM))^{-1})} \\ &\leq 4d \exp(2cM) \log \left(1 + \frac{8S^2L \exp(2cM)}{\omega} \right) \end{aligned}$$

741 where the inequalities follow from Lemma C.9. Since the above bound is decreasing with $\omega \geq \varepsilon$, it is
742 maximised at $\omega = \varepsilon$. ■

743 **Lemma C.9.** For every real number $c > 0$ define

$$g(c) = \left[\ln \left(1 + \frac{e^{-c}}{1+e^c} \right) \right]^{-1}.$$

744 Then

$$g(c) \leq 4e^{2c}.$$

745 **Proof.** Let

$$u(c) = \frac{e^{-c}}{1+e^c} \quad (0 < u(c) < \frac{1}{2} \text{ for } c > 0).$$

746 For every $u \in (0, 1)$ the elementary inequality

$$\ln(1+u) \geq \frac{u}{1+u} \tag{7}$$

747 holds (easily verified from the concavity of \ln). Applying Equation (7) with $u = u(c)$ gives

$$\begin{aligned} \ln(1+u(c)) &\geq \frac{u(c)}{1+u(c)} \geq \frac{u(c)}{2} \quad (\text{since } u(c) < \frac{1}{2}) \\ &= \frac{e^{-c}}{2(1+e^c)}. \end{aligned}$$

748 Therefore

$$g(c) = \frac{1}{\ln(1+u(c))} \leq \frac{2(1+e^c)}{e^{-c}} = 2e^c(1+e^c) = 2e^c + 2e^{2c} \leq 4e^{2c},$$

749 because $2e^c \leq 2e^{2c}$ for $c > 0$. This completes the proof. ■

750 C.3 Lower bound on the eluder dimension

751 This section establishes lower bounds on various forms of the eluder dimension for generalized linear
752 models. The construction method is inspired by Dong et al. (2019). It is noteworthy that these lower
753 bounds also apply to the star-number of the model classes, which itself is a lower bound on the eluder
754 dimension, as detailed in Li et al. (2022).

755 We begin by stating the Johnson-Lindenstrauss lemma, which guarantees the existence of nearly
756 orthogonal vectors on the unit sphere. We first start with the following lemma, the proof of which can
757 be found in Dasgupta and Gupta (2003).

758 **Lemma C.10** (Johnson-Lindenstrauss Lemma). For any $\zeta \in (0, 1)$, there exists a subset $\Phi \subset \mathbf{S}_2^{d-1}$
759 such that:

760 1. For every distinct pair $x_i, x_j \in \Phi$, we have $|\langle x_i, x_j \rangle| \leq \zeta$.

761 2. The size of the set satisfies $|\Phi| \geq \lfloor \exp \left(\frac{d\zeta^2}{8} \right) \rfloor$.

762 **Corollary C.11.** For $\ell = \ell_X$, $\mu(x) = \frac{1}{1+e^{-x}}$, and $\mathcal{F} = \text{GLM}(\mu, S)$, the ε -eluder dimension of the
763 class $\Phi(\mathcal{F})$ is lower bounded by

$$\frac{d-1}{6S} \exp \left(\frac{S}{40} \right),$$

764 for $S \geq 4$, $d \geq 2$, and $\varepsilon \leq \frac{1}{4}$.

765 **Proof.** The self-concordance parameter of the specified class is $M = 1$. Moreover, $\dot{\mu}(x) \leq \mu(x) \leq$
766 $\exp(x)$ for all $x \in \mathbf{R}$, and $\dot{\mu}(0) = \frac{1}{4}$, so we have

$$\ln(\iota) = \ln \left(\frac{\dot{\mu}(0)}{\dot{\mu}(-S/2)} \right) \geq \ln \left(\frac{\exp(S/2)}{4} \right) = S/2 - 4,$$

767 and

$$\frac{\ln(\iota)^2}{8S + 4\ln(\iota)} \geq \frac{(S/2 - 4)^2}{8S + 4(S/2 - 4)} = \frac{S^2/4 + 16 - 4S}{10S - 16} \geq \frac{S^2/4 - 4S}{10S} \geq \frac{S}{40} - 0.4$$

768 Therefore, the bound in Theorem 4.2 bound evaluates to

$$\begin{aligned} \frac{d-1}{4S} \min \left(\exp \left(\frac{S}{16} \right), \exp \left(\frac{\ln(\iota)^2}{8S + 4\ln(\iota)} \right) \right) &\geq \frac{d-1}{4S} \min \left(\exp \left(\frac{S}{16} \right), \exp \left(\frac{S}{40} - 0.4 \right) \right) \\ &\geq \frac{d-1}{4Se^{0.4}} \min \left(\exp \left(\frac{S}{16} \right), \exp \left(\frac{S}{40} \right) \right) \\ &\geq \frac{d-1}{6S} \exp \left(\frac{S}{40} \right). \end{aligned}$$

769

770 **Theorem 4.2** (Eluder dimension lower bound). *Let ℓ and $\mathcal{F} = \text{GLM}(\mu, S)$, $S \geq 4$, $d \geq 2$, satisfy*
 771 *Assumption 2.1 with parameter M . Then there exist a universal constant $C > 0$, a parameter set Θ*
 772 *and an action set \mathcal{A} such that for $\varepsilon \leq \frac{\dot{\mu}(0)}{M^2}$, the ε -eluder dimension of $\bar{\Phi}(\mathcal{F})$ (defined in Definition 3.1),*
 773 *is lower bounded by*

$$\frac{Cd}{S} \min(\exp(S), \exp(\frac{\ln(\tilde{\kappa})^2}{8SM^2 + 4\ln(\tilde{\kappa})})) \quad \text{where} \quad \tilde{\kappa} = \frac{\dot{\mu}(0)}{\dot{\mu}(-S/2)}.$$

774 **Proof.** Let ζ be a free parameter, $d' = \lfloor S \rfloor$, $N = \lfloor \exp \left(\frac{d'\zeta^2}{8} \right) \rfloor$, and $M = \lfloor \frac{d-1}{d'} \rfloor$. By Lemma C.10,
 775 there exists $x_1, \dots, x_N \in \mathbf{S}_2^{d'-2} \subset \mathbf{R}^{d'-1}$ such that $|\langle x_i, x_j \rangle| \leq \zeta$ for all $i \neq j \in [N]$.

776 Next, for $(i, j) \in [M] \times [N]$ define $\theta_*, \theta_{i,j}, a_{i,j} \in \mathbf{R}^d$ as

$$\begin{aligned} \theta_* &= S \cdot \left(-\frac{1}{\sqrt{2}}, 0, \dots, 0 \right), \\ \theta_{i,j} &= S \cdot \left(-\frac{1}{\sqrt{2}}, \underbrace{0, \dots, 0}_{d'(i-1)}, \frac{1}{\sqrt{2}} \cdot x_j, 0, \dots, 0 \right), \\ a_{i,j} &= \left(\frac{1}{\sqrt{2}}, \underbrace{0, \dots, 0}_{d'(i-1)}, \frac{1}{\sqrt{2}} \cdot x_j, 0, \dots, 0 \right), \end{aligned}$$

777 and let $\eta(a) = \mu(\langle a, \theta_* \rangle)$ in Definition 3.1. We will show that the sequence

$$a_{1,1}, a_{1,2}, \dots, a_{1,N}, a_{2,1}, a_{2,2}, \dots, a_{2,N}, \dots, a_{M,1}, a_{M,2}, \dots, a_{M,N},$$

778 is an ω -eluder for $\omega = \frac{\dot{\mu}(0)}{M^2}$, which implies a lower bound on the ε -eluder dimension of the class
 779 $\bar{\Phi}(\mathcal{F})$ for $\varepsilon \leq \frac{\dot{\mu}(0)}{M^2}$. To this end, we will show that the two conditions

$$\sum_{\{(t,l):(t-1)N+l < (i-1)N+j\}} \bar{\varphi}_{i,j}(a_{t,l}) \leq \frac{\dot{\mu}(0)}{M^2} \quad (8)$$

$$\text{and} \quad \bar{\varphi}_{i,j}(a_{i,j}) \geq \frac{\dot{\mu}(0)}{M^2}, \quad (9)$$

780 hold for all $\bar{\varphi}_{i,j} \in \bar{\Phi}(\mathcal{F})$ where

$$\begin{aligned} \bar{\varphi}_{i,j}(a) &= \int \ell(\cdot, \mu(\langle a, \theta_{i,j} \rangle)) - \ell(\cdot, \mu(\langle a, \theta_* \rangle)) dP_a, \\ &= \ell(\mu(\langle a, \theta_* \rangle), \mu(\langle a, \theta_{i,j} \rangle)) - \ell(\mu(\langle a, \theta_* \rangle), \mu(\langle a, \theta_* \rangle)), \end{aligned}$$

781 where the second equality follows from the fact that ℓ is linear in the first argument and the fact that
 782 $\eta(a) = \mu(\langle a, \theta_* \rangle)$.

783 **Step 1: Show that Equation (9) holds.** Firstly, for all $(i, j) \in [M] \times [N]$, we have

$$\langle a_{i,j}, \theta_{i,j} \rangle = 0, \quad \langle a_{i,j}, \theta_\star \rangle = -\frac{S}{2},$$

784 which gives

$$\bar{\varphi}_{i,j}(a_{i,j}) = \ell(\mu(-S/2), \mu(0)) - \ell(\mu(-S/2), \mu(-S/2)).$$

785 Note that by Assumption 2.1, we have

$$\frac{\partial \ell(y, \mu(u))}{\partial u} = \mu(u) - y, \text{ and } \frac{\partial^2 \ell(y, \mu(u))}{\partial u^2} = \dot{\mu}(u).$$

786 Next, by Assumption 2.1 we know that $\ell(a, \mu(\cdot))$ is a self concordant function with parameter M ,
787 which gives

$$\begin{aligned} \bar{\varphi}_{i,j}(a_{i,j}) &= \ell(\mu(-S/2), \mu(0)) - \ell(\mu(-S/2), \mu(-S/2)) \\ &\geq (S/2)^2 \cdot \dot{\mu}(0) \cdot \left(\frac{\exp(-MS/2) + MS/2 - 1}{(-MS/2)^2} \right) \\ &= \dot{\mu}(0) \cdot \left(\frac{\exp(-MS/2) + MS/2 - 1}{M^2} \right) \\ &\geq \dot{\mu}(0) \cdot \left(\frac{MS/2 - 1}{M^2} \right) \\ &\geq \frac{\dot{\mu}(0)}{M^2} \quad (\text{since } S \geq 4/M), \end{aligned}$$

788 where the first inequality follows from Sun and Tran-Dinh (2019, Proposition 10) where we dropped
789 the linear term as the gradient is zero at $\ell(\mu(S/2), \mu(S/2))$.

790 **Step 2: Show that Equation (8) holds.** First, note that for all $i \neq i' \in [M]$ and $j \in [N]$ we have

$$\langle a_{i,j}, \theta_{i',j} \rangle = -S/2,$$

791 which gives

$$\bar{\varphi}_{i,j}(a_{i',j}) = \ell(\mu(-S/2), \mu(-S/2)) - \ell(\mu(-S/2), \mu(-S/2)) = 0. \quad (10)$$

792 Next, for all $i \in [M]$ and $j \neq j' \in [N]$ we have

$$-\frac{S(1+\zeta)}{2} \leq \langle a_{i,j}, \theta_{i,j'} \rangle \leq -\frac{S(1-\zeta)}{2},$$

793 which, again, by Sun and Tran-Dinh (2019, Proposition 10) gives

$$\begin{aligned} \bar{\varphi}_{i,j}(a_{i,j'}) &= \ell(\mu(-S/2), \mu(v)) - \ell(\mu(-S/2), \mu(-S/2)) \\ &\leq (\zeta S/2)^2 \cdot \dot{\mu}(-S/2) \cdot \left(\frac{\exp(M\zeta S/2) - M\zeta S/2 - 1}{(M\zeta S/2)^2} \right) \\ &\leq \frac{\dot{\mu}(-S/2)}{M^2} \cdot \exp(M\zeta S/2). \end{aligned} \quad (11)$$

794 Therefore, we have

$$\begin{aligned} \sum_{\{(t,l):(t-1)N+l < (i-1)N+j\}} \bar{\varphi}_{i,j}(a_{t,l}) &= \sum_{l=0}^{j-1} \bar{\varphi}_{i,j}(a_{i,l}) \\ &\leq \sum_{l=0}^{j-1} \frac{\dot{\mu}(-S/2)}{M^2} \cdot \exp(M\zeta S/2) \\ &\leq \exp\left(\frac{S\zeta^2}{8}\right) \cdot \frac{\dot{\mu}(-S/2)}{M^2} \cdot \exp(M\zeta S/2) \\ &= \frac{\dot{\mu}(-S/2)}{M^2} \exp\left(\frac{S\zeta^2}{8} + \frac{M\zeta S}{2}\right), \end{aligned}$$

795 where the first equality follows from Equation (10), the first inequality follows from Equation (11),
 796 and the last inequality follows from the upper bound on N from Lemma C.10. Hence, it suffices to
 797 set ζ such that

$$\frac{\dot{\mu}(-S/2)}{M^2} \exp\left(\frac{S\zeta^2}{8} + \frac{M\zeta S}{2}\right) \leq \frac{\dot{\mu}(0)}{M^2}$$

798 which is equivalent to

$$\zeta^2 + 4M\zeta - \frac{8 \ln\left(\frac{\dot{\mu}(0)}{\dot{\mu}(-S/2)}\right)}{S} \leq 0,$$

799 which is a quadratic function in ζ , and the valid range of (positive) ζ is

$$0 \leq \zeta \leq 2 \left(\sqrt{M^2 + \frac{2 \ln(\iota)}{S}} - M \right),$$

800 where $\iota = \frac{\dot{\mu}(0)}{\dot{\mu}(-S/2)}$. Next, from an application of Lemma C.12 and noting that ζ needs to be smaller
 801 than 1 due to Lemma C.10, we have

$$0 \leq \zeta \leq \min \left(\frac{4 \ln(\iota)}{S(2M + \sqrt{2 \ln(\iota)/S})}, 1 \right). \quad (12)$$

802 **Step 3: Compute the length of the eluder sequence.** Thus, we have shown that the sequence
 803 $a_{1,1}, a_{1,2}, \dots, a_{M,N}$, is a $\frac{\dot{\mu}(0)}{M^2}$ -eluder sequence with the length of $M \cdot N$. Firstly, note that

$$N = \left\lfloor \exp \left(\frac{d'\zeta^2}{8} \right) \right\rfloor \geq \frac{1}{2} \exp \left(\frac{S\zeta^2}{16} \right)$$

804 where the inequality follows from the fact that $\lfloor x \rfloor \geq x/2$ for $x \geq 1$. This leads to $\frac{1}{2} \exp(S/16)$ if
 805 the min in Equation (12) evaluates to 1, otherwise we have

$$\begin{aligned} N &\geq \frac{1}{2} \exp \left(\frac{S\zeta^2}{16} \right) \\ &= \frac{1}{2} \exp \left(\frac{S}{16} \left(\frac{4 \ln(\iota)}{S(2M + \sqrt{2 \ln(\iota)/S})} \right)^2 \right) \\ &= \frac{1}{2} \exp \left(\frac{\ln(\iota)^2}{S(2M + \sqrt{2 \ln(\iota)/S})^2} \right) \\ &\geq \frac{1}{2} \exp \left(\frac{\ln(\iota)^2}{8SM^2 + 4 \ln(\iota)} \right), \end{aligned}$$

806 where the last inequality follows from the fact that $(a+b)^2 \leq 2a^2 + 2b^2$. On the other hand, we have

$$M = \left\lfloor \frac{d-1}{\lfloor S \rfloor} \right\rfloor \geq \frac{d-1}{2\lfloor S \rfloor} \geq \frac{d-1}{2S},$$

807 which by putting everything together gives

$$M \cdot N \geq \frac{d-1}{4S} \min \left(\exp \left(\frac{S}{16} \right), \exp \left(\frac{\ln(\iota)^2}{8SM^2 + 4 \ln(\iota)} \right) \right).$$

808 ■

809 **Lemma C.12.** For every $A > 0$ and $c \geq 0$,

$$\sqrt{A^2 + c} \geq A + \frac{c}{2A + \sqrt{c}}.$$

810 **Proof.** Rewrite the square root as

$$\sqrt{A^2 + c} = A + \frac{c}{\sqrt{A^2 + c} + A}. \quad (13)$$

811 Because $(A + \sqrt{c})^2 = A^2 + 2A\sqrt{c} + c \geq A^2 + c$, we have $\sqrt{A^2 + c} \leq A + \sqrt{c}$. Substituting this
812 upper bound for the denominator in (13) gives

$$\sqrt{A^2 + c} \geq A + \frac{c}{(A + \sqrt{c}) + A} = A + \frac{c}{2A + \sqrt{c}},$$

813 which is the claimed inequality. ■

814 C.4 Lower bound on the ℓ_2 -eluder dimension

815 Li et al. (2022, Section 4) question the optimality of the eluder dimension bound for generalized
816 linear models, particularly the presence of the $(L_1/L_2)^2$ term, where L_1 and L_2 are the lower and
817 upper bounds on the derivative of the link function, respectively. The following theorem demonstrates
818 an instance where this $(L_1/L_2)^2$ term is indeed necessary. For the specified model class, $L_2 = \frac{1}{4}$
819 and $L_1 \geq c \exp(-S/2)$ for some constant c . We show that the eluder dimension is lower bounded by
820 $\Omega(d \exp(S))$. This result nearly matches the upper bound in Li et al. (2022, Proposition 4), differing
821 only by a $\ln(\frac{1}{\varepsilon})$ factor, thereby highlighting the necessity of the $(L_1/L_2)^2$ term.

822 **Definition C.1** (ℓ_2 -eluder dimension, Russo and Van Roy (2013, Definition 4)). The ε - ℓ_2 -eluder
823 dimension of a class of functions \mathcal{F} is the length of the longest sequence a_1, \dots, a_τ such that for
824 some $\varepsilon' \geq \varepsilon$

$$w_k := \sup \left\{ (f_{\theta_1} - f_{\theta_2})(a_k) : \sqrt{\sum_{i=1}^{k-1} (f_{\theta_1}(a_i) - f_{\theta_2}(a_i))^2} \leq \varepsilon', \text{ for } \theta_1, \theta_2 \in \Theta \right\} > \varepsilon', \quad (14)$$

825 for each $k \leq \tau$.

826 **Theorem C.13** (ℓ_2 -eluder dimension lower bound). *There exists a parameter set $\Theta \subseteq S \cdot \mathbf{S}_2^{d-1}$, and
827 an action set \mathcal{A} such that the ε - ℓ_2 -eluder dimension of $\mathcal{F} = \{\sigma(\langle \theta, \cdot \rangle) : \theta \in \Theta\}$ for $\sigma(x) = \frac{1}{1+e^{-x}}$
828 is lower bounded by*

$$\frac{d-1}{4S} \exp\left(\frac{S}{126}\right),$$

829 for $S \geq 4$, $d \geq 2$, and $\varepsilon \leq \sqrt{\frac{1}{8}}$.

830 **Proof.** Consider the same construction as in Theorem 4.2, namely the parameter set Θ and the action
831 set \mathcal{A} , and the true parameter θ_* . We will show that the sequence

$$a_{1,1}, a_{1,2}, \dots, a_{1,N}, a_{2,1}, a_{2,2}, \dots, a_{2,N}, a_{3,1}, \dots, a_{M-1,N}, a_{M,1}, a_{M,2}, \dots, a_{M,N},$$

832 satisfy the condition in Equation (14) for $\varepsilon' = \frac{1}{8}$, which implies a lower bound on the ε - ℓ_2 -eluder
833 dimension for $\varepsilon \leq \frac{1}{8}$.

834 To this end, it suffices to show that the two conditions

$$\sum_{\{(t,l):(t-1)N+l \leq (i-1)N+j\}} (\sigma(\theta_{i,j}^\top a_{t,l}) - \sigma(\theta_*^\top a_{t,l}))^2 \leq \frac{1}{8} \quad (15)$$

$$\text{and} \quad (\sigma(\theta_{i,j}^\top a_{i,j}) - \sigma(\theta_*^\top a_{i,j}))^2 \geq \frac{1}{8}, \quad (16)$$

835 hold for all $\theta_{i,j} \in \Theta$.

836 **Step 1: Show that Equation (16) holds.** We have

$$(\sigma(\theta_{i,j}^\top a_{i,j}) - \sigma(\theta_*^\top a_{i,j}))^2 = (\sigma(0) - \sigma(-S/2))^2 = \left(\frac{1}{2} - \sigma(-S/2)\right)^2 = \left(\frac{1}{2} - \frac{1}{1+e^{-S/2}}\right)^2 \geq \frac{1}{8},$$

837 where the last inequality holds for $S \geq 4$.

838 **Step 2: Show that Equation (15) holds.** First, note that for all $i \neq i' \in [M]$ and $j \in [N]$ we have

$$(\sigma(\theta_{i,j}^\top a_{i',j}) - \sigma(\theta_\star^\top a_{i',j}))^2 = (\sigma(-S/2) - \sigma(-S/2))^2 = 0, \quad (17)$$

839 which gives

$$\begin{aligned} \sum_{\{(t,l):(t-1)N+l < (i-1)N+j\}} (\sigma(\theta_{i,j}^\top a_{t,l}) - \sigma(\theta_\star^\top a_{t,l}))^2 &= \sum_{l=0}^{j-1} (\sigma(\theta_{i,j}^\top a_{i,l}) - \sigma(\theta_\star^\top a_{i,l}))^2 \\ &\leq \sum_{l=0}^{j-1} (\exp(\theta_{i,j}^\top a_{i,l}) - \exp(\theta_\star^\top a_{i,l}))^2 \\ &\leq \sum_{l=0}^{j-1} \exp\left(-\frac{S}{2}(1-\zeta)\right)^2 \left(\zeta \frac{S}{2}\right)^2 \\ &\leq N \exp(-S(1-\zeta)) \left(\zeta \frac{S}{2}\right)^2 \\ &\leq \exp\left(\frac{S\zeta^2}{8}\right) \exp(-S(1-\zeta)) \left(\zeta \frac{S}{2}\right)^2 \\ &\leq \exp\left(S\left(\frac{\zeta^2}{8} + \zeta - 1\right)\right) \left(\zeta \frac{S}{2}\right)^2, \end{aligned}$$

840 where the first equality follows from Equation (17), the first inequality follows from Lemma C.14, and
841 the second inequality follows by Taylor expansion and upper bounding the first derivative. Moreover,
842 by setting $\zeta = \frac{1}{3}$, we can further upper bound the last term as

$$\begin{aligned} \exp\left(S\left(\frac{\zeta^2}{8} + \zeta - 1\right)\right) \left(\zeta \frac{S}{2}\right)^2 &\leq \exp\left(S\left(\frac{9\zeta}{8} - 1\right)\right) \left(\zeta \frac{S}{2}\right)^2 \\ &= \exp\left(-\frac{5}{8}S\right) \left(\frac{S}{6}\right)^2 \\ &= \frac{1}{36} \left(\frac{5}{8}\right)^{-2} \exp\left(-\frac{5}{8}S\right) \left(\frac{5}{8}S\right)^2 \\ &\leq \frac{1}{36} \left(\frac{5}{8}\right)^{-2} \cdot \frac{4}{e^2} \leq \frac{1}{8}, \end{aligned}$$

843 where the last inequality follows from the fact that $e^{-t}(t^2) \leq \frac{4}{e^2}$ for $t \geq 0$.

844 **Step 3: Compute the length of the eluder sequence.** Thus, we have shown that the sequence
845 $a_{1,1}, a_{1,2}, \dots, a_{M,N}$, satisfies the condition in Equation (14) for $\varepsilon' = \frac{1}{8}$. As shown in the proof of
846 Theorem 4.2, $M \geq \frac{d-1}{2S}$, and we have

$$\begin{aligned} N \cdot M &\geq \left\lfloor \exp\left(\frac{d'\zeta^2}{8}\right) \right\rfloor \cdot \frac{d-1}{2S} \\ &= \left\lfloor \exp\left(\frac{d'}{63}\right) \right\rfloor \cdot \frac{d-1}{2S} \\ &= \frac{d-1}{4S} \exp\left(\frac{S}{126}\right), \end{aligned}$$

847 where the last inequality follows from the fact that $\lfloor x \rfloor \geq x/2$ for $x \geq 1$. ■

848 **Lemma C.14.** For all $x, y \in \mathbb{R}$,

$$(\sigma(x) - \sigma(y))^2 \leq (e^x - e^y)^2,$$

849 where $\sigma(t) = \frac{1}{1+e^{-t}}$.

850 **Proof.** We have

$$\begin{aligned}
(\sigma(x) - \sigma(y))^2 &= (\sigma(-x) - \sigma(-y))^2 \\
&= \left(\frac{1}{1+e^x} - \frac{1}{1+e^y} \right)^2 \\
&= \left(\frac{e^y - e^x}{(1+e^x)(1+e^y)} \right)^2 \\
&\leq (e^x - e^y)^2.
\end{aligned}$$

851

852 C.5 Self-concordance & convex relaxation

853 Take a parametric model class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ where $\Theta \subset \mathbf{R}^d$ is a convex parameter set satisfying
854 $\|\theta\|_2 \leq S$ for some $S > 0$, for all $\theta \in \Theta$. For any $(y, a) \in \mathcal{Y} \times \mathcal{A}$, let $\ell_{(y,a)} : \mathbf{R}^d \rightarrow \mathbf{R}$ be given by
855 $\theta \mapsto \ell(y, f_\theta(a))$. Consider the following self-concordance assumption.

Assumption C.1 (Self-concordance of losses). Assume that for all $z \in \mathcal{Y} \times \mathcal{A}$, ℓ_z is convex and thrice differentiable. Moreover, assume that there exists an $M > 0$ such that for all $z \in \mathcal{Y} \times \mathcal{A}$, $\theta \in \Theta^\circ$ (the interior of Θ) and $u, v \in \mathbf{R}^d$,

$$|\langle D_u^3 \ell_z(\theta) v, v \rangle| \leq M \|u\|_2 \langle \nabla^2 \ell_z(\theta) v, v \rangle,$$

856 where $D_u^3 \ell_z(\theta) \in \mathbf{R}^{d \times d}$ denotes the third directional derivative of ℓ_z at θ in the direction u evaluated
857 at θ , and $\nabla^2 \ell_z(\theta) \in \mathbf{R}^{d \times d}$ is a matrix of the second order partial derivatives of ℓ_z evaluated at θ .

858 In particular, the generalised linear models introduced in Example 2.1 and Example 2.2 are $M = 1$
859 self-concordant (Fauray et al., 2020; Lee et al., 2024). As shown in Janz et al. (2024), Assumption C.1
860 is equivalent to requiring that $|\ddot{\mu}(x)| \leq M \dot{\mu}(x)$ for all x in the domain of μ , which holds for these
861 GLMs. Moreover, a recent result by Liu et al. (2024) shows that many generalised linear models
862 satisfy Assumption C.1.

Let $\mathcal{L}_t(\theta) = \sum_{i=1}^{t-1} \ell(Y_i, f_\theta(A_i))$ be the empirical risk for a parameter $\theta \in \Theta$ on the first $t - 1$ observations, and $\hat{\theta}_t \in \Theta$ be an ERM. Consider the confidence sets of the form

$$\Theta_t = \{\theta \in \Theta : \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t) \leq \beta_t\}, \quad \beta_t > 0, \quad t \in \mathbf{N}_+,$$

863 These can be enclosed within an ellipsoid as follows.

Theorem C.15. Under Assumption C.1, for all $t \in \mathbf{N}_+$,

$$\Theta_t \subset \{\theta \in \Theta : \|\theta - \hat{\theta}_t\|_{\nabla^2 \mathcal{L}_t(\hat{\theta}_t)} \leq 2(1 + SM)\beta_t\}.$$

864 We provide a proof for completeness, but this result is well known (see, for example, Lee et al., 2024).

Lemma C.16 (Proposition 10 of Sun and Tran-Dinh (2019)). For any $\theta, \theta' \in \Theta$, under Assumption C.1,

$$g(-M\|\theta - \theta'\|_2) \|\theta - \theta'\|_{\nabla^2 \mathcal{L}_t(\theta')}^2 \leq \mathcal{L}_t(\theta) - \mathcal{L}_t(\theta') - \langle \nabla \mathcal{L}_t(\theta'), \theta - \theta' \rangle.$$

865 where $g(x) = \frac{\exp(x) - x - 1}{x^2}$.

Proof of Theorem C.15. From Lemma C.16, and observing that since $\hat{\theta}$ is an ERM and Θ is convex, $\langle \nabla \mathcal{L}_t(\hat{\theta}_t), \theta - \hat{\theta}_t \rangle$ is nonnegative for any $\theta \in \Theta$, we have that for any $\theta \in \Theta$,

$$g(-M\|\theta - \hat{\theta}_t\|) \|\theta - \hat{\theta}_t\|_{\nabla^2 \mathcal{L}_t(\hat{\theta}_t)}^2 \leq \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t).$$

866 Using that $\frac{\exp(x) - x - 1}{x^2} \geq \frac{1}{-x+2}$ whenever $x \leq 0$, bounding $\|\theta - \hat{\theta}_t\|_2 \leq 2S$ and staring at the result
867 a little ought to convince the reader of the veracity of our claim. ■

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

1076 generate deepfakes for disinformation. On the other hand, it is not needed to point out
1077 that a generic algorithm for optimizing neural networks could enable people to train
1078 models that generate Deepfakes faster.

- 1079 • The authors should consider possible harms that could arise when the technology is
1080 being used as intended and functioning correctly, harms that could arise when the
1081 technology is being used as intended but gives incorrect results, and harms following
1082 from (intentional or unintentional) misuse of the technology.
- 1083 • If there are negative societal impacts, the authors could also discuss possible mitigation
1084 strategies (e.g., gated release of models, providing defenses in addition to attacks,
1085 mechanisms for monitoring misuse, mechanisms to monitor how a system learns from
1086 feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

1088 Question: Does the paper describe safeguards that have been put in place for responsible
1089 release of data or models that have a high risk for misuse (e.g., pretrained language models,
1090 image generators, or scraped datasets)?

1091 Answer: [NA]

1092 Justification:

1093 Guidelines:

- 1094 • The answer NA means that the paper poses no such risks.
- 1095 • Released models that have a high risk for misuse or dual-use should be released with
1096 necessary safeguards to allow for controlled use of the model, for example by requiring
1097 that users adhere to usage guidelines or restrictions to access the model or implementing
1098 safety filters.
- 1099 • Datasets that have been scraped from the Internet could pose safety risks. The authors
1100 should describe how they avoided releasing unsafe images.
- 1101 • We recognize that providing effective safeguards is challenging, and many papers do
1102 not require this, but we encourage authors to take this into account and make a best
1103 faith effort.

12. Licenses for existing assets

1105 Question: Are the creators or original owners of assets (e.g., code, data, models), used in
1106 the paper, properly credited and are the license and terms of use explicitly mentioned and
1107 properly respected?

1108 Answer: [NA]

1109 Justification:

1110 Guidelines:

- 1111 • The answer NA means that the paper does not use existing assets.
- 1112 • The authors should cite the original paper that produced the code package or dataset.
- 1113 • The authors should state which version of the asset is used and, if possible, include a
1114 URL.
- 1115 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 1116 • For scraped data from a particular source (e.g., website), the copyright and terms of
1117 service of that source should be provided.
- 1118 • If assets are released, the license, copyright information, and terms of use in the package
1119 should be provided. For popular datasets, paperswithcode.com/datasets has
1120 curated licenses for some datasets. Their licensing guide can help determine the license
1121 of a dataset.
- 1122 • For existing datasets that are re-packaged, both the original license and the license of
1123 the derived asset (if it has changed) should be provided.
- 1124 • If this information is not available online, the authors are encouraged to reach out to
1125 the asset's creators.

13. New assets

1127 Question: Are new assets introduced in the paper well documented and is the documentation
1128 provided alongside the assets?

1129 Answer: [NA]
 1130 Justification:
 1131 Guidelines:
 1132 • The answer NA means that the paper does not release new assets.
 1133 • Researchers should communicate the details of the dataset/code/model as part of their
 1134 submissions via structured templates. This includes details about training, license,
 1135 limitations, etc.
 1136 • The paper should discuss whether and how consent was obtained from people whose
 1137 asset is used.
 1138 • At submission time, remember to anonymize your assets (if applicable). You can either
 1139 create an anonymized URL or include an anonymized zip file.

1140 **14. Crowdsourcing and research with human subjects**
 1141 Question: For crowdsourcing experiments and research with human subjects, does the paper
 1142 include the full text of instructions given to participants and screenshots, if applicable, as
 1143 well as details about compensation (if any)?
 1144 Answer: [NA]
 1145 Justification:
 1146 Guidelines:
 1147 • The answer NA means that the paper does not involve crowdsourcing nor research with
 1148 human subjects.
 1149 • Including this information in the supplemental material is fine, but if the main contribu-
 1150 tion of the paper involves human subjects, then as much detail as possible should be
 1151 included in the main paper.
 1152 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
 1153 or other labor should be paid at least the minimum wage in the country of the data
 1154 collector.

1155 **15. Institutional review board (IRB) approvals or equivalent for research with human**
 1156 **subjects**
 1157 Question: Does the paper describe potential risks incurred by study participants, whether
 1158 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
 1159 approvals (or an equivalent approval/review based on the requirements of your country or
 1160 institution) were obtained?
 1161 Answer: [NA]
 1162 Justification:
 1163 Guidelines:
 1164 • The answer NA means that the paper does not involve crowdsourcing nor research with
 1165 human subjects.
 1166 • Depending on the country in which research is conducted, IRB approval (or equivalent)
 1167 may be required for any human subjects research. If you obtained IRB approval, you
 1168 should clearly state this in the paper.
 1169 • We recognize that the procedures for this may vary significantly between institutions
 1170 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
 1171 guidelines for their institution.
 1172 • For initial submissions, do not include any information that would break anonymity (if
 1173 applicable), such as the institution conducting the review.

1174 **16. Declaration of LLM usage**
 1175 Question: Does the paper describe the usage of LLMs if it is an important, original, or
 1176 non-standard component of the core methods in this research? Note that if the LLM is used
 1177 only for writing, editing, or formatting purposes and does not impact the core methodology,
 1178 scientific rigorousness, or originality of the research, declaration is not required.
 1179 Answer: [NA]

1180

Justification:

1181

Guidelines:

1182

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

1183

1184

- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

1185