Eluder dimension: localise it!

Alireza Bakhtiari
University of Alberta
sbakhtia@ualberta.ca

Alex Ayoub
University of Alberta
aayoub@ualberta.ca

Samuel Robertson
University of Alberta
smrobert@ualberta.ca

David Janz
University of Oxford
david.janz@stats.ox.ac.uk

Csaba Szepesvári
University of Alberta
szepesva@ualberta.ca

Abstract

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

1 Introduction

We study decision-making problems where the regret admits a first-order (small-cost) bound of the form

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

with $\eta(a_\star)$ the optimal mean per-round cost and Γ_n, Γ'_n instance and model dependent complexities.

The challenge in obtaining first-order bounds is in how we measure the complexity of the task. The (global) eluder dimension (Russo and Van Roy, 2013) is a standard complexity measure used to provide worst-case guarantees, but, as we shall argue, it ignores the local structure of the problem: analyses based on the global eluder dimension often introduce a factor in Γ_n that scales like a per-step worst-case information/curvature parameter (denoted κ), which cancels out $\eta(a_\star)$ and destroys first-order gains. This subtle issue is present in much of the previous work on first-order bounds.

We show that by localising the eluder dimension—restricting it to a small neighbourhood of the optimal model—these κ terms can be moved into the lower-order Γ'_n term. This tightens the link between exploration and the actual difficulty of the instance and yields genuinely first-order bounds.

Our contributions may be summarised as follows:

1. *Localised ℓ_1-eluder dimension.* We define a localised ℓ_1-eluder dimension (in the sense of Liu et al., 2022) over a small-excess-loss neighbourhood, which avoids the κ dependency in the classic generalised linear bandit setting (Filippi et al., 2010; Faury et al., 2020).
2. *Necessity of localisation.* We prove lower bounds showing that the global ℓ_1- and ℓ_2-eluder dimensions (Russo and Van Roy, 2013; Liu et al., 2022) must scale with κ in the generalised linear setting, and that this negates small-cost or variance-dependent improvements.
3. *Stochastic bandits.* We propose a version-space optimistic algorithm, ℓ-UCB, which takes a loss ℓ and builds confidence sets for the cost function. Under (a) bounded loss, (b) a Bernstein variance condition, and (c) a triangle condition (Foster and Krishnamurthy, 2021), ℓ-UCB achieves a small-cost bound using analysis based on our localised eluder dimension.
4. *Reinforcement learning.* We extend our method to online RL, giving ℓ-GOLF, and obtain the first κ-free first-order regret bound for finite-horizon RL with bounded rewards/costs.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1.1 Related work

Small-cost in bandits. Adversarial small-cost bounds are known (Neu, 2015; Allen-Zhu et al., 2018; Foster and Krishnamurthy, 2021; Ito et al., 2020; Olkhovskaya et al., 2023). However, adversarial algorithms are often conservative in stochastic regimes (Lattimore and Szepesvári, 2020, Ch. 18) and do not transfer cleanly to reinforcement learning, where optimism (Lai and Robbins, 1985) remains central for low regret (Ayoub et al., 2020; Weisz et al., 2023; Wu et al., 2025; Moulin et al., 2025).

In stochastic bandits, first-order bounds typically assume distributional knowledge (e.g., noise/cost models) (Abeille et al., 2021; Faury et al., 2022; Janz et al., 2024; Liu et al., 2024; Lee et al., 2024). We consider stochastic bandits with function approximation and *unknown* bounded cost distributions, aligning with adversarial-style uncertainty but in a stochastic environment.

Small-cost in reinforcement learning. Online first-order bounds have been shown by Wang et al. (2023) with the (strong) distributional Bellman completeness assumption. This assumption was then removed by Ayoub et al. (2024), but in the offline setting; Wang et al. (2024) extended this result back to the online setting. However, by relying on the global eluder dimension, Wang et al. (2023, 2024) suffer a (hidden) κ-dependence in the leading term that undermines first-order gains.

Reward-first-order vs cost-first-order. Reward-first-order bounds help when the optimal reward is small (Jin et al., 2020; Wagenmaker et al., 2022); this is very different from the cost-first-order guarantees we target (our results actually hold for both small-cost and small-reward settings). Small-cost results have been previously shown in structured settings such as tabular Markov decision processes (Lee et al., 2020) and linear-quadratic regulators (Kakade et al., 2020).

Instance-optimal exploration. Pure-exploration studies instance-dependent sample complexity, with notable works including policy-difference estimation for tabular reinforcement learning (Narang et al., 2024) and PEDEL for linear function approximation (Wagenmaker and Jamieson, 2022). The algorithm-agnostic lower bounds of Al-Marjani et al. (2022) show that PEDEL is near instance-optimal for tabular MDPs. These works are complementary to our regret-focused results.

2 Background on generalised linear models & loss functions

We collect the notation and standing assumptions for generalised linear models (GLMs) and losses used throughout. Fix a dimension $d \in \mathbf{N}_+$, and let $\mathcal{A}, \Theta \subset \mathbf{R}^d$ be closed sets. Let $U \subset \mathbf{R}$ be a closed interval and let $\mu : U \to [0,1]$ be increasing. The GLM class with link μ and parameter set Θ is

$$\mathrm{GLM}(\mu, \Theta) = \{a \mapsto \mu(\langle a, \theta \rangle) \colon a \in \mathbf{R}^d, \theta \in \Theta, \langle a, \theta \rangle \in U\}\,.$$

We also consider losses $\ell \colon [0,1] \times [0,1] \to \mathbf{R}$, where $\ell(y, \hat{y})$ evaluates a prediction $\hat{y}$ against an outcome y. The next assumption records the structural conditions on $(\mathcal{A}, \Theta, \mu, \ell)$ used in our analysis; the final condition is satisfied when ℓ is the negative log-likelihood of the GLM associated with μ.

Assumption 1. *We make the following assumptions:*

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

Remark 1. *The requirement that $M, \kappa \ge 1$ in Assumption 1 is there solely to simplify our bounds.*

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

Example 1. *The log-loss function $\ell_{\mathrm{X}}(y, p) = -y \log p - (1-y)\log(1-p)$ together with the sigmoid link function $\mu_{\mathrm{X}} \colon [-S, S] \to [0,1]$ given by $u \mapsto 1/(1 + e^{-u})$ satisfies Assumption 1 with $L = 1/4$, $M = 1$ and $\kappa = 3e^S$.*

Algorithm 1 the ℓ-UCB bandit algorithm

input loss function ℓ, model $\mathcal{F}$, nonnegative, nondecreasing confidence widths $(\beta_t)_{t\geq 1}$
for time-step $t \in \mathbf{N}_+$ **do**
let $\mathcal{F}_t$ be the subset of the model given by

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

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

Example 2. *The Poisson loss function $\ell_{\mathrm{P}}(y,p) = p - y\log p$ together with the exponential link function $\mu_{\mathrm{P}} \colon [-S, 0] \to [0,1]$ given by $u \mapsto e^u$ satisfies Assumption 1 with $L = M = 1$ and $\kappa = e^S$.*

3 Bandits with bounded costs and the ℓ-UCB algorithm

Our bandit setting comprises a set of actions $\mathcal{A}$ and a corresponding set of action-dependent cost distributions $\mathcal{P} = \{P_a \colon a \in \mathcal{A}\}$ supported on the interval $[0,1]$ (we will write $\operatorname{supp} P$ 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

$$R_n = \sum_{t=1}^{n} \eta(A_t) - \eta(a_\star) \quad \text{where} \quad \eta \colon a \mapsto \int y P_a(dy) \quad \text{and} \quad a_\star \in \arg\min_{a\in\mathcal{A}} \eta(a).$$

The learner may base its choice of A_t on the past observations $A_1, Y_1, \ldots, A_{t-1}, Y_{t-1}$, any extra randomness independent of the observations (say, for tie-breaking), and prior knowledge in the form of a model class: a set $\mathcal{F}$ of functions $\mathcal{A} \to [0,1]$ known to contain η. The key assumptions here are:

Assumption 2 (Bounded costs)**.** *We have $\cup_{a\in\mathcal{A}} \operatorname{supp} P_a \subset [0,1]$.*

Assumption 3 (Realisability)**.** *We have that $\eta \in \mathcal{F}$.*

Algorithm Our algorithm, ℓ-UCB (Algorithm 1) is an implementation of optimism with empirical risk minimisation-based confidence intervals.

At each time-step $t \in \mathbf{N}_+$, the algorithm constructs a confidence set $\mathcal{F}_t$ for η composed of functions in $\mathcal{F}$ for which the empirical risk under the loss function $\ell \colon [0,1] \times [0,1] \to \mathbf{R}$ does not exceed that of the empirical risk minimiser by more than β_t, where $(\beta_t)_{t\geq 1}$ is a problem-dependent nonnegative, nondecreasing sequence of confidence widths. The algorithm then computes an optimistic function-action pair $(f_t, A_t) \in \mathcal{F}_t \times \mathcal{A}$ such that

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

and plays A_t. Optimising over $\mathcal{F}_t \times \mathcal{A}$ is difficult without further assumptions. In Appendix A we detail a standard convex relaxation of this optimisation problem applicable to self-concordant models.

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

Definition 1. Let $\mathcal{F}$ be a model class and $\ell \colon [0,1] \times [0,1] \to \mathbf{R}$ a loss function. For each $f \in \mathcal{F}$, we define the excess loss $\varphi_f \colon [0,1] \times \mathcal{A} \to \mathbf{R}$ and expected excess loss $\bar{\varphi}_f \colon \mathcal{A} \to \mathbf{R}_+$ as

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

We will write $\Phi(\mathcal{F}) = \{\varphi_f \colon f \in \mathcal{F}\}$ and $\bar{\Phi}(\mathcal{F}) = \{\bar{\varphi}_f \colon f \in \mathcal{F}\}$ for the respective loss classes.

Let $\Delta\colon [0,1]\times[0,1]\to \mathbf{R}_+$ be the triangular discrimination function given by

$$\Delta(0,0)=0 \quad \text{and} \quad \Delta(p,q)=\frac{(p-q)^2}{p+q} \quad \text{otherwise.}$$

Assumption 4 (Loss function assumptions)**.** *There exist constants $b, c, \gamma > 0$ such that for all $(f,a)\in\mathcal{F}\times\mathcal{A}$, letting $Y\sim P_a$, the following three bounds hold:*

$$|\varphi_f(Y,a)| \le b \text{ a.s.}\,, \qquad \text{(bounded loss)}$$
$$\operatorname{Var}\varphi_f(Y,a) \le c\bar{\varphi}_f(a)\,, \qquad \text{(variance condition)}$$
$$\Delta(f(a),\eta(a)) \le \gamma\bar{\varphi}_f(a)\,. \qquad \text{(triangle condition)}$$

The first two conditions in Assumption 4, boundedness and the variance condition, allow for a Bernstein-type concentration on the excess loss class. The triangle condition is used in the regret decomposition to move from fast concentration to small-cost bounds. The conditions in Assumption 4 implicitly depend on η, and thus ought to hold uniformly for all $\eta\in\mathcal{F}$. Recall our two losses:

- Log-loss satisfies the triangle condition with $\gamma=2$ (Proposition 20); for any $f\in\mathcal{F}$ such that $\|\varphi_f\|_\infty\le b$, φ_f satisfies the variance condition with $c=b+4$ (Proposition 16).
- Poisson loss satisfies the triangle condition with $\gamma=5$ (Proposition 21); for any $f\in\mathcal{F}$ such that $\|\varphi_f\|_\infty\le b$, φ_f satisfies the variance condition with $c=b+2$ (Proposition 17).

The squared loss function fails to satisfy the triangle condition; Theorem 2 of Foster and Krishnamurthy (2021) shows that squared loss cannot lead to the small-cost bounds we seek.

4 The localised eluder dimension & first-order regret bounds for bandits

We define the (global) ℓ_1-eluder dimension of Liu et al. (2022) as follows; we will henceforth refer to this quantity as *the* eluder dimension, forgoing the ℓ_1 quantifier.

Definition 2 (Eluder dimension)**.** 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 length n 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 $\sum_{t=1}^n|\psi(z_t)|\le\varepsilon$ and $|\psi(x)|>\varepsilon$.
2. We say that z is an ε-eluder sequence with respect to Ψ if for all $t\le n$, z_t is ε-independent of $z_1,\dots,z_{t-1}$ with respect to Ψ.
3. The ε-eluder dimension $\dim_{\mathrm{elud}}(\varepsilon;\Psi)$ of Ψ is the length n of the longest ω-eluder sequence with respect to Ψ for any $\omega\ge\varepsilon$.

Remark 2. *The ℓ_2-eluder dimension at scale ε of a function class Ψ is equivalent to the ℓ_1-eluder dimension of the function class $\{z\mapsto\psi(z)^2\colon \psi\in\Psi\}$ at scale ε^2. We think of the ℓ_1-eluder dimension as loss-agnostic, whereas of the ℓ_2-eluder dimension as baking in the squared loss.*

Our upcoming regret bound will require the choice of a localised model class $\mathcal{F}'\subset\mathcal{F}$, and will depend on it through the following two quantities:

1. The eluder dimension of $\bar{\Phi}(\mathcal{F}')$, the expected excess loss class induced by the localised model class.
2. The number of time-steps $t\in\mathbf{N}_+$ for which the optimistic function f_t is not within the localised set $\mathcal{F}'$.

The localised eluder dimension will feature in the leading term; by taking $\mathcal{F}'$ small, we can make it independent of κ. Taking $\mathcal{F}'$ small may increase the second term, the number of optimistic functions falling outside of $\mathcal{F}'$, but this contributes to the regret as an additive term only.

Theorem 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 2 to 4. Let N_n denote the $1/n$-covering number of $\Phi(\mathcal{F})$ with respect to the uniform metric, and for each $t\in\mathbf{N}_+$ let*

$$\beta_t = 5/2 + 15(b+c)\log(N_n h_t/\delta) \quad \textit{where} \quad h_t = e+\log(1+t)\,.$$

Let $\mathcal{F}' \subset \mathcal{F}$, and denote by d_n the $1/n$-eluder dimension of $\bar{\Phi}(\mathcal{F}')$. Define

$$\Gamma_n = \gamma(1 + (d_n + 1)b + 4d_n\beta_n \log(1 + nb)) .$$

Suppose a learner uses Algorithm 1, ℓ-UCB, over the course of n-many interactions with $\mathcal{P}$, with model class $\mathcal{F}$, loss function ℓ and confidence widths $(\beta_t)_{t\geq 1}$. Then, with probability at least $1-\delta$,

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

The proof of Theorem 1 is located in Appendix D.

Remark 3. *The localised model class $\mathcal{F}'$ does not need to be chosen to run the algorithm. The regret of the algorithm scales automatically with the best possible choice of $\mathcal{F}'$. The localised eluder dimension does not need to be computed to run the algorithm; it is an analysis-only quantity.*

Remark 4. *The covering number N_n featuring in the confidence widths $(\beta_t)_{t\geq 1}$ does not depend on κ; the confidence widths themselves thus do not bring in any κ dependence.*

4.1 Why localisation matters: eluder dimension lower bound for generalised linear models

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 $a_\star$.

Theorem 2 (GLM ℓ_1-eluder dimension lower bound)**.** *Let (μ, ℓ) satisfy the last four properties of Assumption 1 (link L-Lipschitz, M-self-concordant, link-derivative lower bound, and link–loss compatibility). Fix $S \geq 4/M$ and assume that $[-S, 0] \subset U$. Write*

$$\tilde{\kappa} = \frac{\dot{\mu}(0)}{2\dot{\mu}(-S/2)} \in (0, \infty) , \qquad b = \min\{\lfloor S \rfloor, d-1\} .$$

Then, there exist $\mathcal{A} \subset \mathbf{B}_2^d$ and $\Theta \subset S\mathbf{B}_2^d$ such that $(\mathcal{A}, \Theta, \mu, \ell)$ satisfy Assumption 1 and for every $\varepsilon \leq \dot{\mu}(0)/(2M^2)$ the eluder dimension of the expected excess-loss class $\bar{\Phi}(\mathcal{F})$ with $\mathcal{F} = \mathrm{GLM}(\mu, \Theta)$ satisfies

$$\dim_{\mathrm{elud}}(\varepsilon; \bar{\Phi}(\mathcal{F})) \geq \frac{d-1}{4b} \exp\left\{\min\left(\frac{b}{16}, \frac{\log(\tilde{\kappa})^2}{8SM^2 + 4\log(\tilde{\kappa})}\right)\right\},$$

for a sequence of actions taking values in $\mathcal{A}$.

The proof of Theorem 2 is given in Appendix G.

The quantity $\tilde{\kappa} > 0$ can be thought of as the ratio of the information gain in the middle of the parameter set and that at a large step in the negative direction; in common GLMs, $\tilde{\kappa} \approx \kappa$.

Corollary 3. *Consider the setting of Theorem 2 with the log-loss ℓ_{X} and the sigmoid link function $\mu(u) = 1/(1 + e^{-u})$. Then, $M = 1$ and $\dot{\mu}(0) = 1/4$. Therefore, for $S \geq 4$, $d \geq 2$ and any $\varepsilon \leq 1/8$, the ε-eluder dimension of $\bar{\Phi}(\mathrm{GLM}(\mu, \Theta))$ exceeds $\frac{d-1}{4b}\exp\{b/4300\}$, where $b = \min\{\lfloor S \rfloor, d-1\}$.*

The corollary follows by substituting the relevant quantities into Theorem 2; the proof is omitted.

To understand the implications of Theorem 2, consider the setting of logistic bandits in the usual low-information regime, where $\langle a_\star, \theta_\star\rangle \approx -S$; think clickthrough rates in online advertising, where even the best adverts rarely get clicked on. Then $\eta(a_\star) \approx \dot{\mu}(\langle a^\star, \theta_\star\rangle) \approx \exp(-S)$, 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 $\exp(S)$, completely cancelling out the benefit of the $\eta(a_\star)$ small-cost term. This results in a bound that fails to truly adapt to the problem instance. We now show how localisation helps.

4.2 Regret upper bound with localisation for the generalised linear model setting

We now instantiate Theorem 1 for generalised linear models; see Appendix F.3 for the relevant proofs.

Consider the GLM setting. For any f_t let $\theta_t \in \Theta$ be such that $f_t(\cdot) = \mu(\langle \cdot, \theta_t\rangle)$. For any $r > 0$, let

$$\Theta'(r) = \{\theta \in \Theta \colon \forall a \in \mathcal{A}\,, |\langle a, \theta - \theta_\star\rangle| \leq r\}$$

be the r-localised set of parameters and define the corresponding r-localised model class

$$\mathcal{F}'(r) = \{\mu(\langle\cdot,\theta\rangle) \colon \theta \in \Theta'(r)\}\,.$$

The eluder dimension of $\bar{\Phi}(\mathcal{F}'(r))$ can be upper-bounded as a function of r as follows:

Proposition 4. *Let Assumption 1 hold. Then, there exists a universal constant $C > 0$ such that for any $r, \varepsilon > 0$, the ε-eluder dimension of $\bar{\Phi}(\mathcal{F}'(r))$ is bounded as*

$$\dim_{\text{elud}}(\varepsilon; \bar{\Phi}(\mathcal{F}'(r))) \leq Cd\exp(rM)\log(1 + S^2 L\exp(rM)/\varepsilon)\,.$$

In the above bound, taking $r = S$ we end up with an e^S dependence, which is an upper bound on κ (up to constant factors; the upper bound is tight in the logistic setting). However, localising to a $1/M$ neighbourhood, we instead obtain a bound for the ε-eluder dimension of $\bar{\Phi}(\mathcal{F}'(1/M))$ of

$$\dim_{\text{elud}}(\varepsilon; \bar{\Phi}(\mathcal{F}'(1/M))) \leq Cd\log(1 + S^2 L/\varepsilon)\,.$$

Localisation thus allows for first-order bounds where the effect of $\dot{\mu}(a_\star)$ is not overshadowed by κ.

The cost of this $1/M$-localisation in terms of the additive term in the regret is bounded as follows:

Proposition 5. *Under Assumption 1, on the high-probability event of Theorem 1, for any $n \in \mathbf{N}_+$,*

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

That this additive term depends on κ is to be expected; all algorithms for generalised linear models without κ in the leading term have such an additive dependence (Abeille et al., 2021). (Note, our bound depends on the sequence $A_1, \dots, A_n$ rather than holding for all actions—this is fine.)

Using the bounds of Propositions 4 and 5, we obtain the following specialisation of Theorem 1:

Proposition 6 (Regret for ℓ-UCB with the logistic model)**.** *Let $\delta \in (0,1)$, $S > 0$ and $n \in \mathbf{N}_+$. Consider the setting of Theorem 1, with the model class $\mathcal{F} = \mathrm{GLM}(\mu, \Theta)$ where $\mu(u) = 1/(1+e^{-u})$ and the logistic loss function ℓ_{X}. Consider running ℓ-UCB with confidence widths $(\beta_t)_{t\in\mathbf{N}_+}$ given by*

$$\beta_t = 5/2 + 60(2S+1)\left[d\log(1 + 8Sn) + \log(h_t/\delta)\right], \qquad h_t = e + \log(1+t)\,.$$

Then, for a constant $C > 0$, with probability at least $1 - \delta$, the resulting regret satisfies the bound

$$R_n \leq C\sqrt{n\eta(a_\star)d\beta_n}\log(1 + Sn) + Cd\beta_n\left[(\log(1+Sn))^2 + e^{2S} d\log(1+\beta_n)\right]\,.$$

The same result of Proposition 6 also holds, up to constant factors, for the Poisson model; each log-loss specific result used in the proof of Proposition 6 has a Poisson equivalent in Appendix C.

Observe that the regret bound of Proposition 6 holds as soon as Assumptions 1 to 3 are met. *Importantly, we do not assume that the rewards are generated by a generalised linear model.* However, much of the literature does make that assumption, so we make comparisons in that setting.

4.3 Discussion in the logistic bandit & maximum likelihood estimation settings

The maximum likelihood estimation (MLE) setting is the well-studied setting where the costs are sampled from a known generalised linear model (one might also call this a 'well-specified' setting).

A special case of the MLE setting with bounded rewards is the logistic bandit setting. Here, η is given by a generalised linear model with the sigmoid link function and the responses are given by

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

In this setting, the leading term in the regret bound of Proposition 6 nearly matches the lower bound given by Abeille et al. (2021), which states that there exists a $C > 0$ such that

$$R_n \geq Cd\sqrt{nv(a_\star)} \quad \text{where} \quad v(a_\star) = \eta(a_\star)(1 - \eta(a_\star))\,.$$

Likewise, Proposition 6 almost matches the upper bounds of Faury et al. (2022) for their logistic-bandit-specific algorithm, which guarantee that for some $C > 0$, with probability at least $1 - \delta$,

$$R_n \leq CSd\sqrt{nv(a_\star)}\log(n/\delta) + CS^6 d\kappa(\log(n/\delta))^2\,.$$

The suboptimality of Proposition 6 here is in that it depends on $\eta(a_\star)$, providing only a small-cost bound, rather than on $v^\star(a_\star)$; the latter allows for a simultaneous small-cost and small-reward bound. This is because Proposition 6 only assumes that the triangle condition is met on one side of the reward interval, and so only allows for small-cost bounds; strengthening the assumption to be two-sided (which is satisfied by the logistic model) would allow us to recover the $v(a_\star)$. (We do not do this, as it would rule out, for example, the Poisson GLM, which only gives small-cost bounds.)

Interestingly, while Faury et al. (2022) only consider the logistic bandit setting, their analysis actually shows a regret bound for the wider bounded reward setting. The distinction between our work and that of Faury et al. (2022) is that where we use an analysis-only localisation technique to move κ to an additive term, Faury et al. (2022) use an explicit algorithmic warm-up procedure to do this. That is, they run an approximation of an optimal design at the start of interaction, until their confidence sets have shrunk to a neighbourhood of the true parameter (on the good event where the confidence sets do indeed contain the true parameter).[1] *The change from algorithmic localisation to analysis-only localisation is vital for the upcoming reinforcement learning setting where, because we do not have random access to state-action pairs, the solving of an optimal design is not feasible.*

The works of Lee et al. (2024) and Emmenegger et al. (2024) also do away with the warm-up employed in Faury et al. (2022), using techniques based on likelihood ratios. However, they rely on their likelihood ratios forming a martingale, which restricts the results to the MLE setting.

Remark 5 (Bernoullisation)**.** *Any algorithm* `A` *that yields first-order regret for the logistic setting can be used to obtain first-order regret for the bounded reward setting using Bernoullisation. The trick is thus: for each time-step $t \in \mathbf{N}_+$, upon observing $Y_t \in [0,1]$, we sample*

$$Y'_t \sim \mathrm{Bernoulli}(Y_t)$$

and feed Y'_t to the algorithm `A`*. Since the conditional means of Y_t and Y'_t are the same, first-order properties are preserved. Bernoullisation, however, destroys any second-order adaptivity of the algorithm. 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* $\mathrm{Bernoulli}(1/2)$ *random variables leads to an $\Omega(1/\sqrt{n})$ absolute error in the estimate.*

5 First-order regret bounds for online reinforcement learning

We consider the episodic reinforcement learning setting with horizon $H \in \mathbf{N}_+$. Let $M = (\mathcal{S}, \mathcal{A}, c, P, s_1)$ be a Markov decision process (MDP) with states $\mathcal{S}$, actions $\mathcal{A}$, a cost function $c = (c_1, \dots, c_H)$ with $c_h \colon \mathcal{S} \times \mathcal{A} \to [0,1]$, a deterministic starting state $s_1 \in \mathcal{S}$, and a transition kernel $P = (P_1, \dots, P_H)$ with P_h mapping from $\mathcal{S} \times \mathcal{A}$ to probability measures over $\mathcal{S}$.

The learner interacts with the MDP M for $n \in \mathbf{N}_+$ episodes. At the start of each episode $t \in [n]$, the learner specifies a deterministic policy $\pi^t = (\pi^t_1, \dots, \pi^t_H)$, where $\pi^t_h \colon \mathcal{S} \to \mathcal{A}$ for each $h \in [H]$. We allow the policy π^t to depend on the states, actions and costs observed prior to the start of the tth episode, but not 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^π_h be the value function of policy π in M, given by

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

for each $s \in \mathcal{S}$, where $\mathbf{E}_\pi[\,\cdot \mid S_h = s]$ denotes the expectation with respect to the states $S_h, \dots, S_H$ induced by following the policy π in the MDP M starting at $S_h = s$. Then, letting $v^t_h := v^{\pi^t}_h$ ($h \in [H]$), the n-episode regret is given by

$$R_n = \sum_{t=1}^{n} v^t_1(s_1) - v^\star_1(s_1)$$

[1] Faury et al. (2022) also propose an online data-rejection procedure that can be used instead of a warm-up. This is, however, again, an algorithmic tool, in contrast to our analysis-only approach.

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

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

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

$$f^t \in \operatorname*{arg\,min}_{f \in \mathcal{F}^t} f_1(s_1, \pi_f(s_1))$$

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

where $v_h^\star$ ($h \in [H]$) is the optimal value function, defined formally just after Eq. (1).

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

Assumption 5. *Costs are nonnegative and sum to at most one over each episode.*

5.1 Preliminaries on Q-functions, Bellman optimality operators and greedy policies

Let $\mathcal{Q}$ be the set of all maps $\mathcal{S} \times \mathcal{A} \to [0,1]^H$. For $q \in \mathcal{Q}$, we write q_h for the map $(s,a) \mapsto (q(s,a))_h$ (entry h of $q(s,a)$), and we write $q^{\wedge}$ for the function $\mathcal{S} \to [0,1]^H$ defined by

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

For convenience, we may augment each q with $q_{H+1} = 0$, to reflect the standard boundary condition. We let $\mathcal{T} \colon \mathcal{Q} \to \mathcal{Q}$ denote the Bellman optimality operator for the MDP M, given by

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

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

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

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

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

$$\pi_h^q(s) \in \operatorname*{arg\,min}_{a \in \mathcal{A}} q_h(s,a) \quad \text{for all } s \in S \text{ and } h \in [H] \,.$$

5.2 The ℓ-GOLF algorithm, model and loss assumptions & regret bound

Our ℓ-GOLF algorithm (Algorithm 2) is an extension of ℓ-UCB to the episodic online reinforcement learning setting, generalising the GOLF algorithm of Jin et al. (2021) to arbitrary loss functions (GOLF is recovered by taking ℓ to be the squared loss). The algorithm requires the specification of a loss function $\ell \colon [0,1]^2 \to \mathbf{R}$, 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 $q^\star$, and the model $\mathcal{G}$ contains candidates for estimating $\mathcal{T} f$ for $f \in \mathcal{F}$. We will use the following two assumptions of Antos et al. (2008):

Assumption 6 (Realisability)**.** *We assume that $q^\star \in \mathcal{F}$.*

Assumption 7 (Generalised completeness)**.** *We assume that $\mathcal{T}\mathcal{F} \subset \mathcal{G}$.*

The algorithm proceeds to, in each episode $t \in \mathbf{N}_+$, construct a confidence set $\mathcal{F}^t \subset \mathcal{F}$ containing action-value functions that are close to satisfying the Bellman optimality condition $f = \mathcal{T} f$ on the data observed thus far, with errors penalised according to ℓ. It then selects an optimistic function $f^t \in \mathcal{F}$, and plays the policy $\pi^t := \pi^{f_t}$ greedy with respect to f^t. Our analysis will use the following:

Definition 3. For any $f \in \mathcal{Q}$, $h \in [H]$, $x \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}$, we let

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

be the response under the model f. For $(f, g) \in \mathcal{F} \times \mathcal{G}$, we define the excess Bellman loss function

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

and the expected excess Bellman loss function

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

We write $\Phi(\mathcal{F}, \mathcal{G})$ and $\bar{\Phi}(\mathcal{F}, \mathcal{G})$ for the classes of excess and expected excess Bellman losses.

Our assumptions on the loss function here mirror those of the bandit setting:

Assumption 8 (RL loss function assumptions)**.** *There exist constants $b, c, \gamma > 0$ such that for all $(f, g) \in \mathcal{F} \times \mathcal{G}$, $h \in [H]$, $x \in \mathcal{S} \times \mathcal{A}$, $S' \sim P_h(x)$, the following hold:*

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

$$\operatorname{Var} \varphi_h^{f,g}(x, S') \le c \bar{\varphi}_h^{f,g}(x) \,, \qquad \text{(RL variance condition)}$$

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

Our main result is captured by the following theorem. Note, the eluder complexity defined therein is equivalent to an ℓ_1 version of the Bellman eluder dimension of Jin et al. (2021).

Theorem 7. *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 6 to 8. 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 $\Phi(\mathcal{F}, \mathcal{G})$ with respect to the uniform metric, and let*

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

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

$$x \mapsto \sum_{h=1}^H \varphi_h^{f,f}(x_h) \quad \text{for some } f \in \mathcal{F}' \,.$$

Let P_f denote the state-action occupancy measure on $(\mathcal{S} \times \mathcal{A})^H$ induced by the interconnection of M and the policy greedy with 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

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

Suppose a learner uses Algorithm 2, ℓ-GOLF, over the course of n-many episodes with M, with 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

$$R_n \le 3\sqrt{H n v_1^\star(s_1) \Gamma_n} + 6 H \Gamma_n + \operatorname{card}\{t \le n \colon f_t \notin \mathcal{F}'\} \,.$$

Theorem 7 is established in Appendix E. For context, the closest results to ours are those of Wang et al. (2023, 2024) for online RL. Both provide a small-cost regret bound scaling with the Bellman eluder dimension; however, without our notion of a *localised* dimension, their regret bound scales with κ in the leading term for logistic linear models. This entirely offsets any benefit of their small-cost analysis; the bound is not truly instance-adaptive. Moreover, Wang et al. (2023) assumes that the distributional Bellman operator (Bellemare et al., 2017) lies in their model class, an assumption that is significantly stronger than our Assumption 7 (as discussed in Ayoub et al., 2024). An argument for extending the results from costs to rewards was given in Ayoub et al. (2025).

6 Conclusion

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

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

Acknowledgements

Alex Ayoub gratefully acknowledges funding from Netflix. Csaba Szepesvári gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii, NSERC and Netflix.

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.

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.

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.

A. Ayoub, Z. Jia, C. Szepesvári, M. Wang and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, 2020.

A. Ayoub, D. Szepesvári, A. Bakhtiari, C. Szepesvári and D. Schuurmans. Rectifying Regression in Reinforcement Learning. In *Reinforcement Learning Conference*, 2025.

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.

M. G. Bellemare, W. Dabney and R. Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, 2017.

S. Boucheron, G. Lugosi and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

S. Dong, T. Ma and B. Van Roy. On the performance of Thompson sampling on logistic bandits. In *Conference on Learning Theory*, 2019.

S. S. Du, S. M. Kakade, R. Wang and L. F. Yang. Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning? In *International Conference on Learning Representations*, 2020.

N. Emmenegger, M. Mutny and A. Krause. Likelihood ratio confidence sets for sequential decision-making. *Advances in Neural Information Processing Systems*, 2024.

T. Erven, P. Grünwald, M. D. Reid and R. C. Williamson. Mixability in Statistical Learning. In *Advances in Neural Information Processing Systems*, 2012.

L. Faury, M. Abeille, C. Calauzènes and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, 2020.

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.

S. Filippi, O. Cappe, A. Garivier and C. Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 2010.

D. J. Foster and A. Krishnamurthy. Efficient first-order contextual bandits: prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 2021.

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.

D. Janz, S. Liu, A. Ayoub and C. Szepesvári. Exploration via linearly perturbed loss minimisation. In *International Conference on Artificial Intelligence and Statistics*, 2024.

C. Jin, A. Krishnamurthy, M. Simchowitz and T. Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 2020.

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

S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi and W. Sun. Information theoretic regret bounds for online nonlinear control. In *Advances in Neural Information Processing Systems*, 2020.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 1985.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

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

J. Lee, S.-Y. Yun and K.-S. Jun. A Unified Confidence Sequence for Generalized Linear Models, with Applications to Bandits. In *Conference on Neural Information Processing Systems*, 2024.

Q. Liu, A. Chung, C. Szepesvári and C. Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, 2022.

S. Liu, A. Ayoub, F. Sentenac, X. Tan and C. Szepesvári. Almost Free: Self-concordance in Natural Exponential Families and an Application to Bandits. In *Advances in Neural Information Processing Systems*, 2024.

A. Al-Marjani, A. Tirinzoni and E. Kaufmann. Towards instance-optimality in online PAC reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

A. Moulin, G. Neu and L. Viano. Optimistically Optimistic Exploration for Provably Efficient Infinite-Horizon Reinforcement and Imitation Learning. In *Conference on Learning Theory*, 2025.

A. Narang, A. Wagenmaker, L. Ratliff and K. G. Jamieson. Sample complexity reduction via policy difference estimation in tabular reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

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

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

Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, 2013.

T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: a recipe for Newton-type methods. *Mathematical Programming*, 2019.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

A. Wagenmaker and K. G. Jamieson. Instance-dependent near-optimal policy identification in linear MDPs via online experiment design. In *Advances in Neural Information Processing Systems*, 2022.

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

K. Wang, N. Kallus and W. Sun. The central role of the loss function in reinforcement learning. *arXiv preprint arXiv:2409.12799*, 2024.

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

G. Weisz, A. György and C. Szepesvári. Online RL in Linearly q^π-Realizable MDPs Is as Easy as in Linear MDPs If You Learn What to Ignore. In *Advances in Neural Information Processing Systems*, 2023.

J. Whitehouse, Z. S. Wu and A. Ramdas. Time-uniform self-normalized concentration for vector-valued processes. *arXiv preprint arXiv:2310.09100*, 2023.

R. Wu, A. Sekhari, A. Krishnamurthy and W. Sun. Computationally efficient RL under linear Bellman completeness for deterministic dynamics. In *International Conference on Learning Representations*, 2025.

Appendices

A Self-concordance & convex relaxation

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

Assumption 9 (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 the convex set Θ) and $u, v \in \mathbf{R}^d$,*

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

where $D_u^3 \ell_z(\theta) \in \mathbf{R}^{d\times d}$ denotes the third directional derivative of ℓ_z at in the direction u evaluated 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.

In particular, the generalised linear models introduced in Example 1 and Example 2 are $M = 1$ self-concordant (Faury et al., 2020; Lee et al., 2024). As shown in Janz et al. (2024), Assumption 9 is equivalent to requiring that $|\ddot{\mu}(x)| \le M\,\dot{\mu}(x)$ for all x in the domain of μ, which holds for these GLMs. Moreover, a recent result by Liu et al. (2024) shows that many GLMs satisfy Assumption 9.

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

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

These can be enclosed within an ellipsoid as follows.

Theorem 8. *Under Assumption 9, for all $t \in \mathbf{N}_+$,*

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

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

Lemma 9 (Proposition 10 of Sun and Tran-Dinh (2019))**.** *Let $g(x) = \frac{\exp(x)-x-1}{x^2}$. For any $\theta, \theta' \in \Theta$, under Assumption 9,*

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

Proof of Theorem 8. From Lemma 9, and observing that since $\hat{\theta}_t$ 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 $\theta \in \Theta$,

$$g(-M\|\theta - \hat{\theta}_t\|_2)\, \|\theta - \hat{\theta}_t\|^2_{\nabla^2 \mathcal{L}_t(\hat{\theta}_t)} \le \mathcal{L}_t(\theta) - \mathcal{L}_t(\hat{\theta}_t)\,.$$

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

B A uniform Bernstein concentration inequality

The following uniform Bernstein inequality, proven over the course of this section, will be needed to prove both the bandit and the reinforcement learning regret bounds. It is the use of this inequality that necessitates the variance condition and boundedness condition for the excess loss classes.

Theorem 10 (Uniform Bernstein inequality)**.** *Let $\mathcal{Z}$ be a set, $(Z_t)_t$ be a $\mathcal{Z}$-valued process adapted to a filtration $(\mathbf{F}_t)_t$, and Φ a set of real-valued functions on $\mathcal{Z}$. Assume that:*

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

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

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

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

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

To prove Theorem 10, we will need the following definitions and results:

Definition 4 (CGF-like)**.** We say a twice differentiable function $\psi : [0, \lambda_{\max}) \to \mathbf{R}_+$ is *CGF-like* if ψ is strictly convex, $\psi(0) = \psi'(0) = 0$ and $\psi''(0)$ exists.

Definition 5 (sub-ψ process)**.** Let $\mathbf{F}$ be a filtration, $\psi : [0, \lambda_{\max}) \to \mathbf{R}_+$ be a CGF-like function and let $(S_t)_t$ and $(V_t)_t$ be respectively $\mathbf{R}$-valued and $\mathbf{R}_+$-valued $\mathbf{F}$-adapted processes. We say that $(S_t, V_t)_t$ is a sub-ψ process if, for every $\lambda \in [0, \lambda_{\max})$, there exists an $\mathbf{F}$-adapted supermartingale $L(\lambda)$ such that

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

Definition 6 (Sub-gamma process)**.** We say that a random process $(S_t, V_t)_t$ is sub-gamma with parameter $\vartheta > 0$ if it is sub-ψ for the CGF-like function $\psi\colon [0, 1/\vartheta) \to \mathbf{R}_+$ mapping $\lambda \mapsto \frac{\lambda^2}{2(1-\vartheta\lambda)}$.

Theorem 11 (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 $1-\delta$, for all $t \ge 1$,*

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

Theorem 11 is a consequence of Theorem 3.1 of Whitehouse et al. (2023); we will prove it shortly.

Proposition 12. *Let $\mathbf{F}$ be a filtration and let X be a martingale with respect to $\mathbf{F}$, satisfying $X_t \le \mathbf{E}[X_t \mid \mathbf{F}_{t-1}] + b$ for all $t \in \mathbf{N}_+$. Then, for*

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

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

Proof of Proposition 12. If the random variables $X = (X_t)_{t\in\mathbf{N}_+}$ are independent, the result follows directly from Theorem 2.10 (Bernstein's inequality) in Boucheron et al. (2013) combined with the discussion immediately after Corollary 2.11 therein. For the martingale result, use the tower property of the conditional expectation to extend the independent case. ■

Proof of Theorem 10. 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 $\Phi(\varepsilon) \subset \Phi$ be a uniform ε-cover of Φ with cardinality N. Then for any $\varphi \in \Phi$ there exists some $\hat{\varphi} \in \Phi(\varepsilon)$, such that for any $n \in \mathbf{N}_+$,

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

Now observe that for any $\hat{\varphi} \in \Phi(\varepsilon)$, from Proposition 12 and our assumptions on Φ, we have that

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

is a sub-gamma process with parameter $b/3$. Applying Theorem 11 with $\rho = b$ and a confidence parameter δ/N, and taking a union bound over the N functions in $\Phi(\varepsilon)$, we conclude that with probability at least $1 - \delta$,

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

where we have upper bounded the H_n therein, defined as in Theorem 11, by $h_n = e + \log(1+n)$. Next, by the variance condition and Young's inequality,

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

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

We now prove Theorem 11, which is an application of the following result:

Theorem 13 (Theorem 3.1, Whitehouse et al. (2023))**.** *Let $(S_t, V_t)_{t \ge 0}$ be a sub-ψ process for a CGF-like function $\psi : [0, c) \to \mathbf{R}_+$ satisfying $\lim_{\lambda \uparrow c} \psi'(\lambda) = \infty$. Let $\alpha > 1$, $\beta > 0$, $\delta \in (0,1)$ and let $h : \mathbf{R}_+ \to \mathbf{R}_+$ be an increasing function such that $\sum_{k \in \mathbf{N}} 1/h(k) \le 1$. Define the function $\ell_\beta : \mathbf{R}_+ \to \mathbf{R}_+$ 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, δ, h). Then

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

where ψ^* is the convex conjugate of ψ.

Proof of Theorem 11. The result follows from applying Theorem 13 to our sub-gamma process with $\alpha = e$, $\beta = \rho^2$ and $h(k) = (k+e)^2$, and bounding the result crudely. In particular, for our choices of α and h, we have the bound

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

Now, since for our choice of ψ, $\psi^{*-1}(t) = \sqrt{2t} + tc$, the bound from Theorem 13 can be further bounded as

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

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

C Analysis of the log-loss and Poisson loss functions

For convenience, we restate our loss function conditions:

Assumption 4 (Loss function assumptions)**.** *There exist constants $b, c, \gamma > 0$ such that for all $(f, a) \in \mathcal{F} \times \mathcal{A}$, letting $Y \sim P_a$, the following three bounds hold:*

$$|\varphi_f(Y, a)| \le b \text{ a.s.}\,, \qquad \text{(bounded loss)}$$
$$\operatorname{Var} \varphi_f(Y, a) \le c\bar{\varphi}_f(a)\,, \qquad \text{(variance condition)}$$
$$\Delta(f(a), \eta(a)) \le \gamma\bar{\varphi}_f(a)\,. \qquad \text{(triangle condition)}$$

We now establish that the variance condition and triangle condition hold for the log-loss excess loss class Φ_{X} induced by the loss function ℓ_{X} and the Poisson loss excess loss class Φ_{P} induced by ℓ_{P}.

C.1 Establishing the variance condition

For our proof of the variance condition, we will assume that all $\varphi \in \Phi_{\mathrm{X}} \cup \Phi_{\mathrm{P}}$ satisfy the pointwise bound

$$\|\varphi\|_\infty \le b\,.$$

This being satisfied relies on the choice of the model class $\mathcal{F}$. In Appendix F.1, we will verify that for a compatible GLM with parameter norm $S > 0$, the boundedness condition holds with $b = 4S$.

To establish the variance condition, we will use the following result of Erven et al. (2012), and in particular a special case stated and proven immediately afterwards.

Lemma 14 (Lemma 10, Erven et al. (2012))**.** *Let $g(x) = (e^x - x - 1)/x^2$ for $x \ne 0$ and $g(0) = 1/2$, and let X be a random variable satisfying $|X| \le b$. Then, for all $t > 0$, there exists a $C_t \ge g(-tB)$ such that*

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

Lemma 15. *Suppose that X is a random variable satisfying*

1. *Boundedness: $|X| \le b < \infty$; and*
2. *Stochastic mixability: $\mathbf{E}[\exp\{-X/2\}] \le 1$.*

Then, $\operatorname{Var} X \le (b + 4)\mathbf{E}X$.

Proof of Lemma 15. Applying Lemma 14, with $t = 1/2$, we obtain that there exists a $C \ge g(-B/2)$ such that

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

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

We are now ready to prove the variance condition for the log-loss and Poisson loss functions.

Proposition 16 (Log-loss variance condition)**.** *Every $\varphi \in \Phi_{\mathrm{X}}$ uniformly bounded by $b > 0$ satisfies the variance condition with constant $c = b + 4$.*

Proof. The result follows from Lemma 15 combined with that every $\varphi \in \Phi_{\mathrm{X}}$ is stochastically mixable, which we establish now. Observe that every $\varphi \in \Phi_{\mathrm{X}}$ is of the form

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

for some $f \in \mathcal{F}$. Therefore, for a fixed a, letting $Y \sim P_a$ with $\mathbf{E}[Y] = \eta(a)$, we have

$$\begin{aligned}\mathbf{E}\exp\{-\varphi(Y, a)/2\} &= \mathbf{E}\left[\left(\frac{f(a)}{\eta(a)}\right)^{\frac{Y}{2}}\left(\frac{1 - f(a)}{1 - \eta(a)}\right)^{\frac{1-Y}{2}}\right] \\ &\le \frac{\mathbf{E}[Y]}{\eta(a)}\frac{f(a)}{2} + \frac{1 - \mathbf{E}[Y]}{1 - \eta(a)}\frac{1 - f(a)}{2} = \frac{1}{2}\,,\end{aligned}$$

where the inequality follows by the application of AM-GM. ■

Proposition 17 (Poisson loss variance condition)**.** *Every $\varphi \in \Phi_{\mathrm{P}}$ uniformly bounded by $b > 0$ satisfies the variance condition with $c = b + 2$.*

Proof. We establish that for every $\varphi \in \Phi_{\mathrm{P}}$, 2φ is stochastically mixable. The result then follows from Lemma 15, after looking at how each side of the variance condition scales with the change $\varphi \mapsto 2\varphi$. Observe that every $\varphi \in \Phi_{\mathrm{P}}$ is of the form

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

for some $f \in \mathcal{F}$. Thus, for a fixed a, letting $Y \sim P_a$ with $\mathbf{E}[Y] = \eta(a)$, we have

$$\mathbf{E} \exp\{-\varphi(Y, a)\} = \exp\{\eta(a) - f(a)\} \mathbf{E}\left[\left(\frac{f(a)}{\eta(a)}\right)^Y\right].$$

Now, noting that for any $x > 0$ and $y \in [0, 1]$, by convexity, $x^y \le xy + 1 - y$,

$$\mathbf{E}\left[\left(\frac{f(a)}{\eta(a)}\right)^Y\right] \le f(a)\frac{\mathbf{E}[Y]}{\eta(a)} + 1 - \mathbf{E}[Y] = 1 + f(a) - \eta(a) \le \exp\{f(a) - \eta(a)\}\,.$$

$(1 + x \le e^x$ for all $x \in \mathbf{R})$

Combining with the previous display, we have that $\mathbf{E} \exp\{-(2\varphi(Y, a))/2\} \le 1$. ■

C.2 Establishing the triangle condition

To establish the triangle condition, we first sandwich Δ with an easier-to-work-with quantity:

Lemma 18. *For any $p, q \in [0, 1]$,*

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

Proof. If $p, q = 0$ the statement is trivial. Assume that one of p and q is nonzero. Using the algebraic identity $(a - b)(a + b) = a^2 - b^2$, we have

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

Rearranging the above display gives the lower bound:

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

For the upper bound, note that

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

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

Proposition 19. *For any two measures P, Q on the same measurable space with densities p and q with respect to some common dominating measure μ,*

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

We are now ready to prove that the log-loss and Poisson loss functions satisfy the triangle condition.

Proposition 20 (Log-loss triangle condition)**.** *The expected excess log-loss class $\bar{\Phi}_{\mathrm{X}}$ satisfies the triangle condition with constant $\gamma = 2/\log_2(e)$.*

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

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

and that

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

Using Lemma 18, the definition of H^2 and Proposition 19, we have that

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

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

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

Proposition 21 (Poisson loss triangle condition)**.** *The expected excess Poisson loss class $\bar{\Phi}_{\mathrm{P}}$ satisfies the triangle condition with constant $\gamma = 4\sqrt{e}/\log_2(e)$.*

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

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

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

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

Hence,

$$\begin{aligned}
\Delta(p,q) &\leq 2(\sqrt{p} - \sqrt{q})^2 && \text{(Lemma 18)}\\
&\leq 4\sqrt{e}(1 - \exp\{-(\sqrt{p} - \sqrt{q})^2/2\}) && \text{(Eq. (2))}\\
&= 4\sqrt{e}H^2(Q,P) && \text{(definition of } H^2\text{)}\\
&\leq \frac{4\sqrt{e}}{\log_2(e)}\,\mathrm{KL}(Q\|P)\,. && \text{(Proposition 19)}
\end{aligned}$$

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

$$\mathbf{E}[\ell_{\mathrm{P}}(Y,p) - \ell_{\mathrm{P}}(Y,q)] = \mathbf{E}\left[p - q + Y \log \frac{q}{p}\right] = \mathrm{KL}(Q\|P)\,. \qquad \blacksquare$$

D Proof of the cost-sensitive regret bound in the bandit setting, Theorem 1

Recall the following assumptions, and the statement of Theorem 1, which we shall now prove.

Assumption 2 (Bounded costs). *We have* $\cup_{a\in\mathcal{A}} \operatorname{supp} P_a \subset [0,1]$.

Assumption 3 (Realisability). *We have that* $\eta \in \mathcal{F}$.

Definition 1. Let $\mathcal{F}$ be a model class and $\ell\colon [0,1]\times[0,1] \to \mathbf{R}$ a loss function. For each $f \in \mathcal{F}$, we define the excess loss $\varphi_f\colon [0,1]\times\mathcal{A} \to \mathbf{R}$ and expected excess loss $\bar\varphi_f\colon \mathcal{A} \to \mathbf{R}_+$ as

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

We will write $\Phi(\mathcal{F}) = \{\varphi_f\colon f \in \mathcal{F}\}$ and $\bar\Phi(\mathcal{F}) = \{\bar\varphi_f\colon f \in \mathcal{F}\}$ for the respective loss classes.

Assumption 4 (Loss function assumptions). *There exist constants* $b, c, \gamma > 0$ *such that for all* $(f,a) \in \mathcal{F}\times\mathcal{A}$*, letting* $Y \sim P_a$*, the following three bounds hold:*

$$\begin{aligned} |\varphi_f(Y,a)| &\le b \text{ a.s.}\,, && \text{(bounded loss)}\\ \operatorname{Var}\varphi_f(Y,a) &\le c\bar\varphi_f(a)\,, && \text{(variance condition)}\\ \Delta(f(a), \eta(a)) &\le \gamma\bar\varphi_f(a)\,. && \text{(triangle condition)}\end{aligned}$$

Theorem 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 2 to 4. Let* N_n *denote the* $1/n$*-covering number of* $\Phi(\mathcal{F})$ *with respect to the uniform metric, and for each* $t \in \mathbf{N}_+$ *let*

$$\beta_t = 5/2 + 15(b+c)\log(N_n h_t/\delta) \quad \textit{where} \quad h_t = e + \log(1+t)\,.$$

Let $\mathcal{F}' \subset \mathcal{F}$*, and denote by* d_n *the* $1/n$*-eluder dimension of* $\bar\Phi(\mathcal{F}')$*. Define*

$$\Gamma_n = \gamma(1 + (d_n+1)b + 4d_n\beta_n\log(1+nb))\,.$$

Suppose a learner uses Algorithm 1, ℓ*-UCB, over the course of* n*-many interactions with* $\mathcal{P}$*, with model class* $\mathcal{F}$*, loss function* ℓ *and confidence widths* $(\beta_t)_{t\ge1}$*. Then, with probability at least* $1-\delta$*,*

$$R_n \le 3\sqrt{n\eta(a_\star)\Gamma_n} + 6\Gamma_n + \operatorname{card}\{t \le n\colon f_t \notin \mathcal{F}'\}\,.$$

Proof of Theorem 1. Our proof will rely on Theorem 10, the uniform Bernstein inequality established in Appendix B.

Validity of confidence sequence Let $\mathbf{F}$ be the filtration given by $\mathbf{F}_t = \sigma(A_1, Y_1, \dots, A_t, Y_t, A_{t+1})$ for each $t \in \mathbf{N}$. We apply our uniform Bernstein inequality (Theorem 10) to the $\mathbf{F}$-adapted process $(Y_t, A_t)_{t\in\mathbf{N}_+}$ with the function class $\Phi = \{\varphi_f\colon f \in \mathcal{F}\}$ and the choice $\varepsilon = 1/n$ (the two requirements in Theorem 10 are satisfied due to the boundedness and variance condition parts of Assumption 4). From this, we conclude the first part of the following proposition:

Proposition 22. *There exists an event* $\mathcal{E}_\delta$ *satisfying* $\mathbf{P}(\mathcal{E}_\delta) \ge 1-\delta$*, whereon, for all* $f \in \mathcal{F}$ *and* $t \in \mathbf{N}_+$*,*

$$\sum_{i=1}^t \bar\varphi_f(A_i) \le 2\sum_{i=1}^t \varphi_f(Y_i, A_i) + 2\beta_t\,. \tag{3}$$

Moreover, on $\mathcal{E}_\delta$*,*

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

Proof. The second conclusion of Proposition 22, that on $\mathcal{E}_\delta$, $\eta \in \cap_{t\in\mathbf{N}_+}\mathcal{F}_t$, is not immediate. For this, observe that the left-hand side of Eq. (3) is nonnegative (as ensured by the triangle condition of Assumption 4), we conclude that on $\mathcal{E}_\delta$, for all $t \in \mathbf{N}_+$,

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

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

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

Lemma 23. *For $x, y, z > 0$ with $y \leq z$, we have that $x - z \leq 3\sqrt{z\Delta(x,y)} + 6\Delta(x,y)$.*

Lemma 24 (Ayoub et al. (2024))**.** *For $x, z \geq 0$, $z \leq 3x + \Delta(x,z)$.*

Proof of Lemma 23. Observe that

$$
\begin{aligned}
x - z &\leq x - y && \text{(by assumption)} \\
&\leq \sqrt{x+y}\sqrt{\Delta(x,y)} && \text{(defn. } \Delta(x,y)\text{)} \\
&\leq \sqrt{4x+\Delta(x,y)}\sqrt{\Delta(x,y)} && \text{(Lemma 24)} \\
&\leq 2\sqrt{x\Delta(x,y)} + \Delta(x,y)\,. && (4)
\end{aligned}
$$

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

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

which yields that $x \leq 6\Delta(x,y) + 2z$; using this and Eq. (4) gives

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

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

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

Hence, we may apply Lemma 23 with $x = \eta(A_t)$, $y = f_t(A_t)$ and $z = \eta(a_\star)$ to obtain the bound

$$r_t := \eta(A_t) - \eta(a_\star) \leq 3\sqrt{\eta(a_\star)\Delta(f_t(A_t), \eta(A_t))} + 6\Delta(f_t(A_t), \eta(A_t))\,. \tag{5}$$

Regret decomposition Let $I_n = \{t \leq n \colon f_t \in \mathcal{F}'\}$ and observe that the maximal per-step regret is bounded by 1, by Assumption 2. Thus, for any $n \in \mathbf{N}_+$,

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

Using Eq. (5), Cauchy-Schwarz, and that $\operatorname{card} I_n \leq n$, we have that on $\mathcal{E}_\delta$,

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

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 4 to obtain

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

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

Proposition 25 (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 $\psi_1, \dots, \psi_n$ 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 $t \leq n$,*

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

where d is the ω-Eluder dimension of Ψ.

Proposition 26. *On the event $\mathcal{E}_\delta$ of Proposition 22, we have that for all $t \in \mathbf{N}_+$,*

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

With Proposition 26, for any $n \in \mathbf{N}_+$, we may apply Proposition 25 with $\beta := 4\beta_n$, $\omega = 1/n$, and with the upper bound b from Assumption 4, to conclude that on $\mathcal{E}_\delta$,

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

This concludes the proof of Theorem 1. ∎

Proof of Proposition 26. On $\mathcal{E}_\delta$, for any $t \in \mathbf{N}_+$,

$$\begin{aligned}
\sum_{i \in I_{t-1}} \bar{\varphi}_{f_t}(A_i) &\le \sum_{i=1}^{t-1} \bar{\varphi}_{f_t}(A_i) && \text{(by the triangle condition, } \bar{\varphi} \text{ is nonnegative)} \\
&\le 2\left[\sum_{i=1}^{t-1} \varphi_{f_t}(Y_i, A_i) + \beta_t\right] && \text{(on } \mathcal{E}_\delta \text{ Eq. (3) holds)} \\
&= 2\left[\sum_{i=1}^{t-1} \ell(Y_i, f_t(A_i)) - \sum_{i=1}^{t-1} \ell(Y_i, \eta(A_i)) + \beta_t\right] \\
&\le 2\left[\sum_{i=1}^{t-1} \ell(Y_i, f_t(A_i)) - \inf_{\hat{f} \in \mathcal{F}} \sum_{i=1}^{t-1} \ell(Y_i, \hat{f}(A_i)) + \beta_t\right] && \text{(on } \mathcal{E}_\delta, \eta \in \mathcal{F}_t) \\
&\le 4\beta_t \,. && (f_t \in \mathcal{F}_t \text{ and the definition of } \mathcal{F}_t)
\end{aligned}$$

∎

E Proof of RL cost-sensitive regret bound, Theorem 7

Recall the following assumptions, and the statement of Theorem 7, which we shall now prove.

Assumption 5. *Costs are nonnegative and sum to at most one over each episode.*

Assumption 6 (Realisability)**.** *We assume that* $q^\star \in \mathcal{F}$*.*

Assumption 7 (Generalised completeness)**.** *We assume that* $\mathcal{T}\mathcal{F} \subset \mathcal{G}$*.*

Definition 3. For any $f \in \mathcal{Q}$, $h \in [H]$, $x \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}$, we let

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

be the response under the model f. For $(f, g) \in \mathcal{F} \times \mathcal{G}$, we define the excess Bellman loss function

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

and the expected excess Bellman loss function

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

We write $\Phi(\mathcal{F}, \mathcal{G})$ and $\bar{\Phi}(\mathcal{F}, \mathcal{G})$ for the classes of excess and expected excess Bellman losses.

Assumption 8 (RL loss function assumptions)**.** *There exist constants* $b, c, \gamma > 0$ *such that for all* $(f, g) \in \mathcal{F} \times \mathcal{G}$*,* $h \in [H]$*,* $x \in \mathcal{S} \times \mathcal{A}$*,* $S' \sim P_h(x)$*, the following hold:*

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

$$\operatorname{Var} \varphi_h^{f,g}(x, S') \le c\bar{\varphi}_h^{f,g}(x)\,, \qquad \text{(RL variance condition)}$$

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

Theorem 7. *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 6 to 8. 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* $\Phi(\mathcal{F}, \mathcal{G})$ *with respect to the uniform metric, and let*

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

Let $\mathcal{F}' \subset \mathcal{F}$*, and define* $\mathcal{Z}$ *to be the set of functions* $(\mathcal{S} \times \mathcal{A})^H \to \mathbf{R}$ *mapping*

$$x \mapsto \sum_{h=1}^{H} \varphi_h^{f,f}(x_h) \quad \textit{for some } f \in \mathcal{F}'\,.$$

Let P_f *denote the state-action occupancy measure on* $(\mathcal{S} \times \mathcal{A})^H$ *induced by the interconnection of* M *and the policy greedy with respect to* $f \in \mathcal{F}$*, and let* Ψ *be the family of functionals on* $\mathcal{Z}$ *mapping*

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

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

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

Suppose a learner uses Algorithm 2, ℓ*-GOLF, over the course of* n*-many episodes with* M*, with 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*

$$R_n \le 3\sqrt{Hnv_1^\star(s_1)\Gamma_n} + 6H\Gamma_n + \operatorname{card}\{t \le n \colon f_t \notin \mathcal{F}'\}\,.$$

Within the upcoming proofs, we will use the shorthand

$$X_h^t = (S_h^t, A_h^t)\,.$$

Proof of Theorem 7. Our proof will rely on Theorem 10, the uniform Bernstein inequality of Appendix B. The structure of the proof is broadly the same as that of our bandit result, Theorem 1.

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 by

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

Now for each $h \in [H]$, we apply our uniform Bernstein inequality (Theorem 10) to the $\mathbf{F}_h$-adapted process $((X_h^t, S_{h+1}^t))_{t \in \mathbf{N}}$ with the function class $\Phi_h = \{\varphi_h^{f,g} \colon (f,g) \in (\mathcal{F} \times \mathcal{G})\}$ and with $\varepsilon = 1/n$. From this, we conclude the first part of the following proposition:

Proposition 27. *There exists an event $\mathcal{E}_\delta$ satisfying $\mathbf{P}(\mathcal{E}_\delta) \geq 1-\delta$, whereon, for all $(f,g) \in (\mathcal{F} \cup \mathcal{G})^2$, $t \in \mathbf{N}_+$ and $h \in [H]$,*

$$\sum_{i=1}^t \bar{\varphi}_h^{f,g}(X_h^i) \leq 2 \sum_{i=1}^t \varphi_h^{f,g}(X_h^i, S_{h+1}^i) + 2\beta_t\,, \tag{6}$$

Moreover, on $\mathcal{E}_\delta$,

$$q^\star \in \bigcap_{t \leq n} \mathcal{F}^t\,.$$

Proof. The second conclusion of Proposition 27, that on $\mathcal{E}_\delta$, $q^\star \in \cap_{t \in \mathbf{N}_+} \mathcal{F}^t$, is not immediate. For that, fix some $h \in [H]$. Now, observe that the left-hand side of Eq. (6) is nonnegative (as ensured by the RL triangle condition of Assumption 8). Hence, on $\mathcal{E}_\delta$, for all $t \in \mathbf{N}_+$,

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

which implies that

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

Comparing the above inequality with the form of our confidence set $\mathcal{F}_t$ yields that on $\mathcal{E}_\delta$, we have that $q^\star \in \bigcap_{t \leq n} \mathcal{F}^t$, as desired. ■

Let $I_n = \{t \leq n \colon f_t \in \mathcal{F}'\}$ and observe that the maximal per-step regret is bounded by 1, by Assumption 5. Then,

$$\sum_{t=1}^n v_1^t(S_1) - v_1^\star(S_1) \leq \sum_{t \in I_n} \Big(v_1^t(S_1) - v_1^\star(S_1) \Big) + \mathrm{card}([n] \setminus I_n)\,.$$

We bound only the regret on episodes $t \in I_n$. For this, we observe that on $\mathcal{E}^\delta$, $q^\star \in \mathcal{F}^t$ (Proposition 27). Thus, we can apply our inequality on the triangular discrimination, Lemma 23, with $x = v_1^t(S_1)$, $y = f_1^t(S_1)$ and $z = v_1^\star(S_1)$, to obtain

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

Lemma 28 (Contraction Lemma)**.** *Let $f \in \mathcal{F}$, and let $\pi := \pi^f$ and $v = v^\pi$. Then,*

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

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

We prove Lemma 28 presently. Now, let $\mathbf{E}_{\pi^t}$ denote the expectation over trajectories generated by π^t in M, with dummy integration variables $(X_h)_{h=1}^H$, and let $\Delta_{t,h} := \Delta(f_h^t(X_h), (\mathcal{T} f^t)_h(X_h))$.

Observe that

$$
\begin{aligned}
&3\sum_{t\in I_n}\sqrt{v_1^\star(S_1)\Delta(v_1^t(S_1),f_1^t(S_1))}+6\sum_{t\in I_n}\Delta(v_1^t(S_1),f_1^t(S_1))\\
&\le 3\sum_{t\in I_n}\sum_{h=1}^H\sqrt{v_1^\star(S_1)\mathbf{E}_{\pi^t}\Delta_{t,h}}+6\sum_{t\in I_n}\Big\{\sum_{h=1}^H\sqrt{\mathbf{E}_{\pi^t}\Delta_{t,h}}\Big\}^2 && \text{(Lemma 28)}\\
&\le 3\sqrt{Hnv_1^\star(S_1)\sum_{t\in I_n}\sum_{h=1}^H\mathbf{E}_{\pi^t}\Delta_{t,h}}+6H\sum_{t\in I_n}\sum_{h=1}^H\mathbf{E}_{\pi^t}\Delta_{t,h}\,. && \text{(Cauchy-Schwarz, } \operatorname{card} I_n\le n)
\end{aligned}
$$

Now, from the triangle condition, we have that

$$
\sum_{t\in I_n}\sum_{h=1}^H\mathbf{E}_{\pi^t}\Delta_{t,h}\le\gamma\sum_{t\in I_n}\sum_{h=1}^H\mathbf{E}_{\pi^t}\bar\varphi_h^{f^t,f^t}(X_h^t)\,, \tag{7}
$$

and it remains to upper bound the right-hand side by our complexity measure Γ_n. For this, consider the following result that bounds the cumulative expected excess risk on past observations.

Lemma 29. *On $\mathcal{E}_\delta$, for all $t\in\mathbf{N}_+$, $h\in[H]$,*

$$
\sum_{i=1}^{t-1}\bar\varphi_h^{f^t,f^t}(X_h^i)\le 4\beta_t\,.
$$

Proof. By Proposition 27, on $\mathcal{E}_\delta$,

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

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

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

This completes the proof of Lemma 29. ∎

Combining the result in the thus established Lemma 29 with the usual eluder dimension argument of Proposition 25, with $\omega=1/n$ and the upper bound b from Assumption 8, we obtain that Γ_n is indeed an upper bound on the right-hand side of (7). ∎

We now move to prove the contraction lemma, Lemma 28. We will need the following simple result, which follows from the joint convexity of $(x,y)\mapsto-\sqrt{xy}$ for $x,y\ge 0$.

Lemma 30. *For $x,y\ge0$, the map $(x,y)\mapsto(\sqrt{x}-\sqrt{y})^2$ is jointly convex in its arguments.*

Proof of Lemma 28. Let for each $h \in [H]$, let μ_h denote the joint 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 18, we have that

$$\sqrt{\Delta(f_1(S_1, \pi(S_1)), v_1(S_1))} \leq 2\|\sqrt{g_1} - \sqrt{v_1}\|_{\mu_1} .$$

We shall shortly establish the inequality

$$(\forall h \in [H]) \quad \|\sqrt{f_h} - \sqrt{v_h}\|_{\mu_h} \leq \|\sqrt{f_h} - \sqrt{\mathcal{T} f_h}\|_{\mu_h} + \|\sqrt{f_{h+1}} - \sqrt{v_{h+1}}\|_{\mu_{h+1}}, \tag{8}$$

where we interpret $\|\sqrt{f_{H+1}} - \sqrt{v_{H+1}}\|\mu_{H+1} = 0$. Unrolling this over $h = 1, \ldots, H$, we obtain

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

The result follows from applying the other side of Lemma 18 to the terms above.

We now establish Eq. (8). 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

$$\begin{aligned}
\|\sqrt{\mathcal{T} f_h} - \sqrt{v_h}\|_{\mu_h} &= \left\| \sqrt{c(\cdot) + \int f_{h+1}^{\wedge} dP(\cdot)} - \sqrt{c(\cdot) + \int v_{h+1} dP(\cdot)} \right\|_{\mu_h} \\
&\leq \left\| \sqrt{\int f_{h+1}^{\wedge} dP(\cdot)} - \sqrt{\int v_{h+1} dP(\cdot)} \right\|_{\mu_h} \\
&\qquad\qquad (\forall c, a, b \geq 0, |\sqrt{c+a} - \sqrt{c+b}| \leq |\sqrt{a} - \sqrt{b}|) \\
&\leq \|\sqrt{f_{h+1}} - \sqrt{v_{h+1}}\|_{\mu_{h+1}} . \qquad \text{(Jensen's, justified by Lemma 30)}
\end{aligned}$$

This establishes Eq. (8), and with it the lemma. ■

F On self-concordant GLMs with compatible losses

We restate our compatible GLM assumption for convenience.

Assumption 1. *We make the following assumptions:*

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

In this section, we prove that the following hold under Assumption 1:

1. the excess loss class $\Phi(\mathcal{F})$ is uniformly bounded and admits a rather small uniform cover
2. that for a suitable localised model class $\mathcal{F}' \subset \mathcal{F}$, the number of rogue steps under our ℓ-UCB algorithm is bounded
3. the localised expected excess loss class $\bar{\Phi}(\mathcal{F}')$ has a small eluder dimension

These results combined with Theorem 1 yield Proposition 6.

We will use the following lemma repeatedly.

Lemma 31. *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

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

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 $u \in U^\circ$. ■

F.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 32. *Let $(\Theta, \mathcal{A}, \mu, \ell)$ be compatible according to Assumption 1. Then, for all $\theta, \theta' \in \Theta$ and $(y, a) \in [0,1] \times \mathcal{A}$,*

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

Proof. Let $\theta(t) = t\theta + (1-t)\theta'$ and note that for any $(y, a) \in [0,1] \times \mathcal{A}$, by Lemma 31,

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

■

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

Proposition 33. *Under Assumption 1, the ε-covering number of $\Phi(\mathrm{GLM}(\mu,\Theta))$ with respect to the uniform norm is upper bounded by $(1+8S/\varepsilon)^d$.*

Proof. Write $\mathcal{F}=\mathrm{GLM}(\mu,\Theta)$. Let $\theta_\star\in\Theta$ be such that $\eta(a)=\mu(\langle a,\theta_\star\rangle)$ (such a parameter exists by Assumption 3, realisability) and define the map $\rho\colon \Theta\to\Phi(\mathcal{F})$ 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 $\theta_0,\theta_1\in\Theta$,

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

we have by Lemma 32 that ρ is 2-Lipschitz as a map from $(\Theta,\|\cdot\|_2)\to(\mathcal{L}(\mathcal{F}),\|\cdot\|_\infty)$. Now, if $\mathcal{C}_{\varepsilon/2}$ is an $\varepsilon/2$-cover of $(S\mathbf{B}_2^d,\|\cdot\|_2)$, then by said 2-Lipschitzness, $\rho(\mathcal{C}_{\varepsilon/2})$ is an ε-external-cover of $(\Phi(\mathcal{F}),\|\cdot\|_\infty)$. Finally, the 2-norm $\varepsilon/4$-covering number of $S\mathbf{B}_2^d$ is an upper bound on the $\varepsilon/2$-covering number of Θ (Vershynin, 2018, Exercise 4.2.9), and the former quantity is at most $(1+8S/\varepsilon)^d$ (Vershynin, 2018, Corollary 4.2.13). ■

F.2 Rogue steps bound and the localised eluder dimension

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

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

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

$$\bar\varphi(a)=\int \ell(y,\mu(\langle x,\theta\rangle))-\ell(y,\mu(\langle x,\theta_\star\rangle))P_a(dy)=:\bar\varphi(a,\theta)\quad\text{for some }\theta,\theta_\star\in S\mathbf{B}_2^d\,.$$

Lemma 34. *For any $\theta\in\mathbf{R}^d$, letting $\theta(t)=t\theta+(1-t)\theta_\star$ for $t\in[0,1]$, we have that*

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

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

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

Proof. By Lemma 31, for any $(y,a)\in[0,1]\times\mathbf{B}_2^d$,

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

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

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

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

F.2.1 Proof of bound on the number of rogue steps, Proposition 5

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

Lemma 35 (Lemma 2 of Janz et al. (2024))**.** *Fix $\lambda,\gamma>0$. Let $a_1,a_2,a_3\in\mathbf{R}^d$ be a sequence of vectors with $\|a_t\|_2\le 1$ for all $t\ge 1$. Define $V_t(\lambda)=\lambda I+\sum_{s=1}^t a_ta_t^\mathsf{T}$. Then the number of times that $\|a_t\|_{V_{t-1}^{-1}(\lambda)}\ge\gamma$ is upper bounded by*

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

Proposition 5. *Under Assumption 1, on the high-probability event of Theorem 1, for any $n \in \mathbf{N}_+$,*

$$\operatorname{card}\{t \le n \colon |\langle A_t, \theta_t - \theta_\star \rangle| > 1/M\} \le 64 d\kappa M^2 \beta_n \log\left(1 + (64/3)\kappa^2 M^2 S^2 \beta_n\right) .$$

Proof of Proposition 5. For any $\lambda \ge 0$ and $\theta \in \mathbf{R}^d$, we define the positive semidefinite matrices

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

where α is defined as in Lemma 34. Then, using that by Lemma 34, for any $\theta \in \Theta$, there exists a $\zeta \in \mathbf{R}$ satisfying $|\zeta| \le S$ such that $\alpha(A_i, \theta) = \dot{\mu}(\zeta)$, we have that for all $\theta \in \Theta$, from the definition of κ in Assumption 1,

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

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

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

By the triangle inequality and then using Proposition 26, we have that on the good event $\mathcal{E}_\delta$ (as defined in Proposition 22),

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

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

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

We lower bound x^2 by $y = 1/(8M^2(\kappa\beta_t + 1))$ and apply Lemma 35 together with the bound $\log(1+y) \ge 3/(4y)$, twice (which holds because $y \le 1/16$), to obtain that the count in question is no greater than

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

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

F.2.2 Proof of the upper bound on the eluder dimension bound, Proposition 4

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

Proposition 36. *Let $\mu \colon U \to [0,1]$ be an M-self-concordant link function. Then, for any $u, u' \in U^\circ$ satisfying $|u - u'| \le c$, $\dot{\mu}(u) \le \exp(cM)\dot{\mu}(u')$.*

Lemma 37. *Suppose that $a > 1$, $x \ge 1$, $b \ge 0$ and $a^x \le bx + 1$. Then,*

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

Proof of Lemma 37. Let $f(x) = a^x$ and $g(x) = bx + 1$. Since f is convex and g is affine, they intersect at no more than two points. Since they intersect at 0, we have that the set of x satisfying $f(x) \le g(x)$ is of the form $[0, y]$ for some $y \ge 0$. Now, let $y' = \log(1 + b/\log(a))/\log(a)$. Then, a quick calculation shows that $f(y') > g(y')$. Thus, $y' > y$. ■

Proposition 4. *Let Assumption 1 hold. Then, there exists a universal constant $C > 0$ such that for any $r, \varepsilon > 0$, the ε-eluder dimension of $\bar{\Phi}(\mathcal{F}'(r))$ is bounded as*

$$\dim_{\mathrm{elud}}(\varepsilon; \bar{\Phi}(\mathcal{F}'(r))) \le C d \exp(rM) \log(1 + S^2 L \exp(rM)/\varepsilon) .$$

Proof of Proposition 4. Let $a_1, \ldots, a_k$ and $\theta_1, \ldots, \theta_k$ be witness to the eluder dimension in question, in that they satisfy

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

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 positive semidefinite matrix

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

For each $i \leq t \leq k$, use Lemma 34 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, by Section 4.2 for all $i \leq t \leq k$,

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

we have by Proposition 36 that, for all $i \leq t \leq k$,

$$\exp(-rM)\dot{\mu}(\langle a_i, \theta_\star \rangle) \leq \dot{\mu}(\zeta_{i,t}) \leq \exp(rM)\dot{\mu}(\langle a_i, \theta_\star \rangle) \,. \tag{9}$$

Hence, using Lemma 34 and Eq. (9), we have the bound

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

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

$$\begin{aligned} \|\theta_t - \theta_\star\|^2_{H_{t-1}(\lambda)} &\leq \|\theta_t - \theta_\star\|^2_{H_{t-1}(0)} + \lambda\|\theta_t - \theta_\star\|^2 \\ &\leq 2\exp(rM)\omega + 4\lambda S^2 \\ &\leq 2\omega(\exp(rM) + 1) \,. \end{aligned} \tag{10}$$

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

$$\begin{aligned} \omega &\leq \bar{\varphi}(a_t, \theta_t) && \text{(definition of } \omega, a_t, \theta_t) \\ &= \frac{\dot{\mu}(\zeta_{t,t})}{2} \langle a_t, \theta_t - \theta_\star \rangle^2 && \text{(definition of } \zeta_{t,t}) \\ &\leq \frac{\exp(rM)}{2} \dot{\mu}(\langle a_t, \theta_\star \rangle) \langle a_t, \theta_t - \theta_\star \rangle^2 && \text{(Eq. (9))} \\ &\leq \frac{\exp(rM)}{2} \dot{\mu}(\langle a_t, \theta_\star \rangle) \|a_t\|^2_{H^{-1}_{t-1}(\lambda)} \|\theta_t - \theta_\star\|^2_{H_{t-1}(\lambda)} && \text{(Cauchy-Schwarz)} \\ &\leq \omega \exp(rM)(\exp(rM) + 1) \|x_t\|^2_{H^{-1}_{t-1}(\lambda)} \,. && \text{(Eq. (10))} \end{aligned}$$

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

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

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

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

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

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

Putting the two inequalities together yields the inequality

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

Now, applying Lemma 37 with $a = 1 + c$, $x = k/d$ and $b = L/\lambda = 2S^2 L/\omega$, we obtain

$$k \leq d \log\left(1 + \frac{2S^2 L}{\omega \log(1 + c)} \right) / \log(1 + c) \leq d e^{2rM} \log(1 + 2S^2 L e^{2rM}/\omega) \,,$$

where the second inequality follows by substituting in the definition of c and using that $e^x(1 + e^x) \geq e^{2x}$ for $x \geq 0$. Since the above bound is decreasing with $\omega \geq \varepsilon$, it is maximised at $\omega = \varepsilon$. ■

F.3 Regret bound for ℓ-UCB with the logistic model

Proposition 6 (Regret for ℓ-UCB with the logistic model)**.** *Let $\delta \in (0,1)$, $S > 0$ and $n \in \mathbf{N}_+$. Consider the setting of Theorem 1, with the model class $\mathcal{F} = \mathrm{GLM}(\mu, \Theta)$ where $\mu(u) = 1/(1+e^{-u})$ and the logistic loss function ℓ_{X}. Consider running ℓ-UCB with confidence widths $(\beta_t)_{t \in \mathbf{N}_+}$ given by*

$$\beta_t = 5/2 + 60(2S+1)\big[d \log(1+8Sn) + \log(h_t/\delta)\big], \qquad h_t = e + \log(1+t)\,.$$

Then, for a constant $C > 0$, with probability at least $1-\delta$, the resulting regret satisfies the bound

$$R_n \leq C\sqrt{n\eta(a_\star)d\beta_n}\log(1+Sn) + Cd\beta_n\big[(\log(1+Sn))^2 + e^{2S} d \log(1+\beta_n)\big]\,.$$

Proof. By Proposition 16, $c = b+4$; by Lemma 32, $b = 4S$; by Proposition 33, $N_n \leq (1+8Sn)^d$. Combined with Theorem 1, these results yield the confidence widths

$$\beta_t = \frac{5}{2} + 60(2S+1)\big[d\log(1+8Sn) + \log(h_t/\delta)\big]\,.$$

Recall that for the logistic model, $L = 1/4$, $M = 1$, $\kappa = 3e^S$, and, by Proposition 20, $\gamma = 2/\log_2(e)$. We consider the localised class $\mathcal{F}'(1)$, for which, by Proposition 4 and the discussion immediately thereafter, for some $C > 0$, the $\frac{1}{n}$-eluder dimension satisfies

$$d_n \leq Cd\log(1+Sn)\,.$$

This yields that for some $C' > 0$,

$$\Gamma_n \leq C' d\beta_n (\log(1+Sn))^2\,.$$

Moreover, by Proposition 5, for some $C'' > 0$,

$$\mathrm{card}\{t \leq n \colon f_t \notin \mathcal{F}'(1)\} \leq C' e^{2S} d\beta_n \log(1+\beta_n)\,.$$

Combining these estimates with the upper bound for regret given by Theorem 1, we obtain the claimed result. ∎

G Proof of lower bound on the eluder dimension in GLMs, Theorem 2

This section establishes our lower bound on the eluder dimension for generalised linear models. The construction is based on the technique of Dong et al. (2019).

Theorem 2 (GLM ℓ_1-eluder dimension lower bound)**.** *Let (μ, ℓ) satisfy the last four properties of Assumption 1 (link L-Lipschitz, M-self-concordant, link-derivative lower bound, and link–loss compatibility). Fix $S \geq 4/M$ and assume that $[-S, 0] \subset U$. Write*

$$\tilde{\kappa} = \frac{\dot{\mu}(0)}{2\dot{\mu}(-S/2)} \in (0, \infty)\,, \qquad b = \min\{\lfloor S \rfloor, d-1\}\,.$$

Then, there exist $\mathcal{A} \subset \mathbf{B}_2^d$ and $\Theta \subset S\mathbf{B}_2^d$ such that $(\mathcal{A}, \Theta, \mu, \ell)$ satisfy Assumption 1 and for every $\varepsilon \leq \dot{\mu}(0)/(2M^2)$ the eluder dimension of the expected excess-loss class $\bar{\Phi}(\mathcal{F})$ with $\mathcal{F} = \mathrm{GLM}(\mu, \Theta)$ satisfies

$$\dim_{\mathrm{elud}}(\varepsilon; \bar{\Phi}(\mathcal{F})) \geq \frac{d-1}{4b} \exp\left\{\min\left(\frac{b}{16}, \frac{\log(\tilde{\kappa})^2}{8SM^2 + 4\log(\tilde{\kappa})}\right)\right\},$$

for a sequence of actions taking values in $\mathcal{A}$.

Our proof will use the following lemma, given as Lemma A.1 in Du et al. (2020).

Lemma 38 (Johnson-Lindenstrauss packing lemma)**.** *For any integer $D \geq 2$ and any parameter $\zeta \in (0, 1)$, there exists a finite set $\Phi \subset \mathbf{S}^{D-1} := \{x \in \mathbf{R}^D : \|x\|_2 = 1\}$ with $|\Phi| \geq \lfloor \exp(D\zeta^2/8) \rfloor$ and*

$$|\langle x, y\rangle| \leq \zeta \qquad \textit{for all distinct } x, y \in \Phi\,.$$

Proof of Theorem 2. Let $\zeta \in (0, 1)$. For $N = \lfloor \exp(b\zeta^2/8) \rfloor$, let $x_1, \ldots, x_N \in \mathbf{R}^b$ satisfy $\|x_i\| = 1$ for all $i \leq N$ and $|\langle x_i, x_j\rangle| \leq \zeta$ for $i, j \leq N$ with $i \neq j$ (such vectors exist by Lemma 38).

Let $e_1, \ldots, e_d$ denote the basis vectors of $\mathbf{R}^d$. Let $m = \lfloor (d-1)/b \rfloor \geq 1$ be the number of length b blocks that fit into the $d-1$ dimensions spanned by $e_2, \ldots, e_d$. Let $E_i \colon \mathbf{R}^b \to \mathbf{R}^d$ insert $v \in \mathbf{R}^b$ into the coordinates of the ith such block; that is, for $i \in [m]$,

$$E_i(v) = \sum_{\ell=1}^{b} v_\ell e_{1+(i-1)b+\ell}\,.$$

Define the optimal parameter vector $\theta_\star = -2^{-1/2} S e_1$ such that $\eta(a) = \mu(\langle a, \theta_\star\rangle)$. We take Θ consisting of $\theta_\star$ and the vectors

$$\theta_{ij} = \theta_\star + 2^{-1/2} S E_i(x_j)\,, \qquad (i, j) \in [m] \times [N]\,.$$

We take the arm-set $\mathcal{A}$ to consist of the vectors

$$a_{ij} = -\frac{\theta_\star}{S} + 2^{-1/2} E_i(x_j)\,, \qquad (i, j) \in [m] \times [N]\,.$$

With this construction, we have the following properties:

$$\begin{aligned}
&\langle a_{ij}, \theta_\star\rangle = -S/2 && \forall (i,j) \in [m] \times [N] \\
&\langle a_{ij}, \theta_{ij}\rangle = 0 && \forall (i,j) \in [m] \times [N] \\
&\langle a_{ij}, \theta_{i'j}\rangle = -S/2 && \forall i, i' \in [m] \text{ with } i \neq i' \\
&\langle a_{ij}, \theta_{ij'}\rangle \in [-(S/2)(1+\zeta), -(S/2)(1-\zeta)] && \forall (i,j) \in [m] \times [N] \text{ and all } j' \in [N]\,.
\end{aligned}$$

For each $(i, j) \in [m] \times [N]$, let $\bar{\varphi}_{ij} \in \bar{\Phi}(\mathcal{F})$ denote the expected excess loss comparing θ_{ij} against $\theta_\star$:

$$\bar{\varphi}_{ij}(a) = \ell(\mu(\langle a, \theta_\star\rangle), \mu(\langle a, \theta_{ij}\rangle)) - \ell(\mu(\langle a, \theta_\star\rangle), \mu(\langle a, \theta_\star\rangle))\,.$$

Consider the sequence of actions (a_{ij}) given by

$$a_{1,1}, \ldots, a_{1,N}, a_{2,1} \ldots, a_{2,N}, \ldots, a_{m,1}, \ldots, a_{m,N}\,, \tag{11}$$

where we index by enumerating $[m] \times [N]$ in lexicographic order. We will show that for a choice of $\zeta \in (0, 1)$, (a_{ij}) is an ω-eluder sequence for $\omega = \dot{\mu}(0)/(2M^2)$, and lower bound the resulting sequence length nM (which depends on ζ). We will take $\{\bar\varphi_{ij}\colon (i,j)\in[m]\times[N]\}$ as the witnessing functions: at step (i,j) function $\bar\varphi_{ij}$ certifies that the sequence (a_{ij}) is ω-eluder.

Large deviation at new action. Let $f(u)=\ell(\mu(-S/2),\mu(u))$. Observe that

$$\dot f(-S/2)=\mu(-S/2)-\mu(-S/2)=0\,,\qquad \ddot f(-S/2)=\dot\mu(-S/2)\,,$$

by the link and loss compatibility properties in Assumption 1. By Taylor's expansion with integral remainder, around $-S/2$, noting that $\dot f(-S/2)=0$, we have

$$\begin{aligned}\bar\varphi_{ij}(a_{ij})=f(0)-f(-S/2)&=\int_0^1(1-t)(S/2)^2\ddot f\left(-\tfrac{S}{2}+\tfrac{tS}{2}\right)\,dt\\&=f(-S/2)+\int_0^1(1-t)(S/2)^2\dot\mu\left(-\tfrac{S}{2}+\tfrac{tS}{2}\right)dt\\&=(S/2)^2\int_0^1(1-t)\dot\mu\left(-\tfrac{S}{2}+\tfrac{tS}{2}\right)\,dt\,.\end{aligned}$$

By Proposition 36, we have the bound

$$\dot\mu\left(-\tfrac{S}{2}+\tfrac{tS}{2}\right)\ge\dot\mu(0)\exp\left(-\tfrac{MS}{2}(1-t)\right)\,.$$

Combined with the previous display, writing $\alpha=\frac{MS}{2}$, this bound gives that

$$\bar\varphi_{ij}(a_{ij})\ge(S/2)^2\dot\mu(0)\int_0^1(1-t)\exp\left(-\alpha(1-t)\right)dt=\frac{\dot\mu(0)}{M^2}\left(1-\exp(-\alpha)(1+\alpha)\right)\ge\frac{\dot\mu(0)}{2M^2}\,,$$

where the final inequality uses that $\alpha\ge2$ (since we assumed $S\ge4/M$), and that $x\mapsto1-e^{-x}(1+x)$ is increasing on $(0,\infty)$.

Small cumulative deviation. At index (i,j), the cumulative deviation is

$$\sum_{t=1}^{i-1}\sum_{\ell=1}^{N}\bar\varphi_{ij}(a_{t\ell})+\sum_{\ell=1}^{j-1}\bar\varphi_{ij}(a_{i\ell})=\sum_{\ell=1}^{j-1}\bar\varphi_{ij}(a_{i\ell})\,,$$

where we used that $i'\neq i$, $\bar\varphi_{ij}(a_{i'j})=0$. Now, by the self-concordance lower bound of Lemma 9 applied to the one-dimensional function $u\mapsto\ell(\mu(-S/2),\mu(u))$ we obtain that for any $\ell<j$,

$$\bar\varphi_{ij}(a_{i\ell})\le\frac{\dot\mu(-S/2)}{M^2}\exp\{MS\zeta/2\}\,,$$

and there are at most $N\le\exp\{b\zeta^2/8\}$ such terms in the sum. Therefore,

$$\sum_{\ell=1}^{j-1}\bar\varphi_{ij}(a_{i\ell})\le\frac{\dot\mu(-S/2)}{M^2}\exp\{b\zeta^2/8+M\zeta S/2\}\,.$$

This sum is upper bounded by $\dot\mu(0)/(2M^2)=\omega$ whenever

$$b\zeta^2/8+M\zeta S/2\le\log\tilde\kappa\,.$$

Using $b\le S$, it suffices that $S\zeta^2/8+M\zeta S/2\le\log\tilde\kappa$, or equivalently that

$$\zeta^2+4M\zeta-\frac{8\log\tilde\kappa}{S}\le0\,.$$

This implies that we require ζ to be in the interval

$$0\le\zeta\le2\Big(\sqrt{M^2+(2/S)\log\tilde\kappa}-M\Big)\,.$$

Length of eluder sequence. With ζ chosen to be the largest feasible and using $\lfloor x\rfloor\ge x/2$ for $x\ge1$, we get

$$N\ge\frac12\exp\{b\zeta^2/8\}\ge\frac12\min\left\{\exp\{b/8\}\,,\,\exp\left\{\frac{\log(\tilde\kappa)^2}{8SM^2+4\log(\tilde\kappa)}\right\}\right\},$$

and $m\ge(d-1)/(2b)$. Hence, $\dim_{\text{elud}}(\varepsilon;\bar\Phi(\mathcal F))\ge mN$, which is lower bounded by the stated quantity. ■

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

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

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

11. **Safeguards**

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

    Answer: [NA]

    Justification:

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification:

    Guidelines:

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

13. **New assets**

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

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.