RECOMMENDER REGRET FOR IMPROVING REGULARISED RL ESTIMATION WITH BANDITS AND MCTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Regularised reinforcement learning (RL) algorithms rely on estimators, like Monte Carlo tree search or one-step look-ahead, to provide policy, value, or Q-value estimates in training and inference. However, existing bandit-based estimation frameworks evaluate sampling efficiency, not the quality of estimates. We address this by formalising estimate generation as a recommendation process, treating bandit-based sampling as an improvement operator. To evaluate the quality of a recommendation (estimate), we define regulariser-aware recommender regrets, analysing their lower bounds, and proving that simple algorithms like uniform sampling can achieve these bounds. Finally, we unify bandit- and search-based target improvement under recommender MCTS. Across synthetic trees, arcade games, and scientific discovery tasks, recommenders often yield faster target convergence and superior downstream performance. Our experiments characterise for the first time the potential of regularised MCTS in scientific applications.

1 INTRODUCTION

Reinforcement learning (RL) algorithms often require policy, value, or Q-value estimators during training and inference. For example, deep Q-networks (Mnih et al., 2013, DQN) and soft actor-critic (Haarnoja et al., 2018, SAC) generate Q-value targets by computing the one-sample Monte Carlo estimates of the next state's value, while AlphaZero uses Monte Carlo tree search (MCTS) to generate improved value and policy estimates (Silver et al., 2017). Despite the success of these estimator-based methods, the literature lacks a framework for optimally allocating sampling budgets to maximise the quality of target estimates.

Using the theoretical framework of multi-armed sampling (Pedramfar & Ravanbakhsh, 2026), which is based on entropy-regularised RL, we study how to efficiently estimate these targets for a more general class of regularisers. In multi-armed sampling, an algorithm sequentially chooses a sampling policy over an action space, samples an action from the sampling policy, and observes a reward. Using the information contained in the rewards, the goal is to efficiently match some target distribution, usually modelled as a softmax or Boltzmann distribution. Efficiency here refers to minimising the number of samples required to match the target distribution, thus motivating why prior formulations focused on evaluating the sampling policy against the target (Wang et al., 2021; Pedramfar & Ravanbakhsh, 2026). However, in RL estimation, we care about the quality of the *recommendation* of various target objects (*e.g.*, policies, values, and Q-values) produced at the end of the sampling allocation, rather than the cost of the sampling policy itself.

To bridge this gap, we extend the simple sampling regret of Pedramfar & Ravanbakhsh (2026) with the simple regret minimisation of Bubeck et al. (2009), to formalise target generation as an explicit **recommendation process**. Specifically, we 1) extend the simple sampling regret to allow for alternative regularisers, 2) expand the class of target objects to include values and Q-values, and 3) define divergences based on the choice regulariser and target object. This yields a regulariser-aware notion of regret we call the **recommender regret**. Our core contributions are as follows:

- We introduce regulariser-aware notions of **recommender regret** for policies, values, and Q-values across general regularised RL settings, deriving theoretical lower bounds for each
- We prove upper bounds on recommender regret, demonstrating that uniform sampling achieves the tightest bounds of the algorithms analysed. This is a surprising result, given that multi-armed sampling requires almost no exploration (Pedramfar & Ravanbakhsh, 2026)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- We extend recommender regret to tree-search settings by proposing **recommender MCTS**, a generalisation of prior regularised tree-search work for planning-based target estimation
- We conduct extensive empirical evaluations on synthetic trees, arcade games, and scientific discovery domains, demonstrating that recommender-based algorithms often converge faster and achieve superior performance compared to prior work.

2 Background and related work

2.1 Multi-armed sampling bandit problems

Multi-armed sampling bandits (Pedramfar & Ravanbakhsh, 2026) are instances of stochastic multi-armed bandit problems with different notions of regret, and thus different goals for the policy. Let $\mathcal{A}$ denote a set of available actions, also called *arms*. Stochastic multi-armed bandit instances can be represented as a collection of reward distributions $\phi = (\phi_a : a \in \mathcal{A})$, where ϕ_a is the reward distribution of action a, with mean $\mu_a = \mathbb{E}_{r\sim\phi_a}[r] \in \mathbb{R}$, and $K = |\mathcal{A}|$ (*i.e.*, there are K actions) (Lattimore & Szepesvári, 2020). A sampler interacts with the environment over T rounds sequentially, receiving feedback in the form of rewards which inform future interactions. At each time step $t \in [T]$ the sampler chooses a sampling policy $d_t \in \Delta_{\mathcal{A}}$, and samples an action $a_t \sim d_t$, receiving a randomly sampled reward $r_{t,a_t} \sim \phi_{a_t}$ from the environment. The history of interactions, $\mathcal{H}_t = (d_1, a_1, r_{1,a_1}, \ldots, d_{t-1}, a_{t-1}, r_{t-1,a_{t-1}})$, is maintained by the sampler and used to choose the next sampling policy: $d_{t+1}(\cdot|\mathcal{H}_{t+1})$.[1]

In multi-armed sampling, the sampler maintains a performance distribution, $f_t = f(\mathcal{H}_t)$, which is a deterministic function of $\mathcal{H}_t$. Given a divergence D, the sampling regret at time t is defined as:

$$\mathcal{R}_t^{f,D} = D(f_t, \pi^\star) \quad \text{with } \pi^\star = \text{softmax}(\mu) \quad \text{and } \mu = (\mu_a : a \in \mathcal{A}), \tag{1}$$

and can be extended to expected or cumulative regret (Pedramfar & Ravanbakhsh, 2026). By setting the performance distribution-generating function $f(\mathcal{H}_t)$, we can alter the properties of the sampling policy, d_t, that optimises the regret. This problem is similar to FairX bandits (Wang et al., 2021), where the considered policies are $\text{softmax}(\ln(g(\mu)))$, with $g : \mathbb{R}^K \to \mathbb{R}^K_{>0}$ being an arbitrary function, and the regret is defined specifically with D as the total variation distance (L_1 norm). Pedramfar & Ravanbakhsh (2026) found that exploration was barely required for sampling.

2.2 Regularised Markov decision processes

Regularised Markov decision processes (MDPs) were introduced by Geist et al. (2019), extending the usual MDP tuple with a strongly convex regularisation function. The regularised MDP tuple is $(\mathcal{S}, \mathcal{A}, r, P, \gamma, H, \Omega)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the environment's dynamics function, $\gamma \in [0, 1)$ is the discount factor, H is the horizon, and $\Omega : \Delta_{\mathcal{A}} \to \mathbb{R}$ is the regulariser. $\Delta_{\mathcal{S}}$ and $\Delta_{\mathcal{A}}$ denote the probability simplices over the state and action spaces, respectively.

Regularisation with Ω penalises the policy for deviating from some desirable behaviour. This can be seen in the regularised policy evaluation functions V_Ω^π and Q_Ω^π over states and state-action pairs respectively, where the regulariser is subtracted from the value function:

$$V_\Omega^\pi(s) = \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}[Q_\Omega^\pi(s,a)] - \frac{1}{\omega}\Omega(\pi) \qquad Q_\Omega^\pi(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim p(\cdot|s,a)}[V_\Omega^\pi(s')]\,. \tag{2}$$

For instance, when Ω is negative entropy, the optimal policy will be "more uniform," in that probability mass will be allocated throughout the action space rather than to a single optimal action (Neu et al., 2017; Haarnoja et al., 2018; Geist et al., 2019; Zhang et al., 2025). Solving for the optimal policy by maximising V_Ω^π with respect to the policy π yields the Legendre-Fenchel dual. The optimal regularised value and policy functions evaluate to the dual of the regulariser and its gradient, respectively: $V_\Omega^\star(s) = \Omega^*(Q_\Omega^\star(s,\cdot))$ and $\pi_\Omega^\star(\cdot|s) = \nabla\Omega^*(Q_\Omega^\star(s,\cdot))$ (Rockafellar, 1997; Geist et al., 2019). Table 1 lists common regularisers and their optimality equations where ω and α are inverse temperatures, π_{ref} is an arbitrary reference policy, $\pi_Q^\alpha = \text{softmax}(\alpha Q_\Omega^\star(s))$, and Tsallis entropy[2]

[1] With the history containing all the randomness of the environment (due to random sampling of the rewards from their distributions) and the policy (due to random sampling of actions from the policy).

[2] We always use $q = 2$ as the entropic index for Tsallis entropy, unless otherwise stated.

yields sparse policies via the sparsemax functions (Martins & Astudillo, 2016; Vieillard et al., 2020; Asadi & Littman, 2017; Gan et al., 2021; Lee et al., 2018; 2020).

2.3 Prior work in recommender regret

Xiao et al. (2019) introduced the stochastic softmax bandit problem which sought to minimise the exponentiated value regret in the entropy-regularised RL setting: $\mathcal{E}_t = \mathbb{E}\big[(\exp(\hat{V}_t) - \exp(V^\star))^2\big]$. Importantly, Xiao et al. (2019) did not study the Q-value or policy regrets of the stochastic softmax bandit problem. In contrast, our work provides lower and upper bounds on norm-based Q-value regrets as well as policy regrets with Bregman divergence and norm-based divergences. Thus, our setting is novel both in the notions of regret and the bounds on those regrets.

Generalising the algorithm of Xiao et al. (2019), Dam et al. (2021) proposed extended empirical exponential weight (E3W), which selects actions according to the convex combination of the regularised policy and a uniform exploration term ($\epsilon > 0$):

$$d_t^{\text{E3W}}(a|s) = (1 - \lambda_t)\nabla\Omega^*\left(\hat{Q}_t(s,\cdot)\right)(a) + \frac{\lambda_t}{|\mathcal{A}|} \quad \text{with } \lambda_t = \frac{\epsilon}{\log(t+1)}. \tag{3}$$

In Section 3.3 we show that E3W is not optimal in the recommender regret, in that uniform action selection can achieve tighter upper bounds.

Xiao et al. (2019) also introduced the maximum entropy tree search (MENTS) algorithm, later extended by Dam et al. (2021) to strongly convex regularisers like relative and Tsallis entropy (RENTS and TENTS, respectively) and to the soft mellowmax (SMM) operator by Vucetic & Gidel (2026).

3 Recommender regret of sampling bandits in regularised RL

We define recommenders in order to generate improved policy, value, and Q-value estimates for use in regularised RL training and inference. Recommender regret appends estimation problems to the sampling bandit framework. While sampling bandits update a general performance distribution $f_t = f(\mathcal{H}_t)$, recommenders specifically update estimates of the regularised RL optimality equations:

$$\hat{\pi}_{\Omega,t} = \nabla\Omega^*(\hat{Q}_{\Omega,t}); \quad \hat{V}_{\Omega,t} = \Omega^*(\hat{Q}_{\Omega,t}); \quad \hat{Q}_{\Omega,t} = \hat{r}_t, \tag{4}$$

where Ω is a strongly convex regulariser, and $\hat{r}_t = (\sum_{i=1}^t r_{i,a_i}\mathbb{I}(a_i = a)/N_t(a) : a \in \mathcal{A})$. From sampling bandits, we naturally arrive at the recommender policy regret by setting $f_t = \text{softmax}(\hat{r}_t)$. However, this only covers one policy function (the softmax) out of many possible through regularised RL, and does not extend to other objects such as values and Q-values. Lastly, the divergences $D \in \{\text{KL}, \|\cdot\|_1\}$ of the sampling bandit policy regret are not aligned with arbitrary regularisers, whose primal norm and Bregman divergence may be different (let alone divergences for objects which are not policies) (Bregman, 1967; Shalev-Shwartz & Singer, 2007). This highlights the need for a new, **regulariser-aware notion of regret** such that the geometry of D matches the regularised RL setting, along with a suitable analysis of lower and upper bounds for this regret.

3.1 Notions of recommender regret

The goal of a recommender is to match a target object $f^\star \in \{\pi_\Omega^\star, V_\Omega^\star, Q_\Omega^\star\}$. We formalise this idea using the notation of Pedramfar & Ravanbakhsh (2026), choosing performance functions

Table 1: Optimality equations for various regularisers.

Name	**Regulariser** $\Omega(\pi)$	**Value** $\Omega^*(Q_\Omega^\star(s))$	**Policy** $\nabla\Omega^*(Q_\Omega^\star(s))$
Negative Entropy	$\sum_a \pi(a\|s)\log\pi(a\|s)$	$\frac{1}{\omega}\text{LSE}\left(\omega Q_\Omega^\star(s)\right)$	$\text{softmax}(\omega Q_\Omega^\star(s))$
Relative Entropy	$\text{KL}(\pi(s)\|\|\pi_{\text{ref}}(s))$	$\frac{1}{\omega}\log\sum_a \pi_{\text{ref}}(a\|s)e^{\omega Q_\Omega^\star(s,a)}$	$\frac{\pi_{\text{ref}}(a\|s)e^{\omega Q_\Omega^\star(s,a)}}{\sum_b \pi_{\text{ref}}(b\|s)e^{\omega Q_\Omega^\star(s,b)}}$
Soft Mellowmax	$\text{KL}(\pi(s)\|\|\pi_Q^\alpha(s))$	$\frac{1}{\omega}\log\sum_a \pi_Q^\alpha(a\|s)e^{\omega Q_\Omega^\star(s,a)}$	$\text{softmax}((\alpha+\omega)Q_\Omega^\star(s))$
Tsallis Entropy	$\frac{1}{2}\sum_a \pi(a\|s)^2$	$\frac{1}{\omega}\text{sparsemax-val}\left(\omega Q_\Omega^\star(s)\right)$	$\text{sparsemax}\left(\omega Q_\Omega^\star(s)\right)$

$f_t \in \{\hat{\pi}_{\Omega,t}, \hat{V}_{\Omega,t}, \hat{Q}_{\Omega,t}\}$, which are deterministic functions of the history $\mathcal{H}_t$. A regulariser-aware divergence function D is also required, with its set of admissible functions changing depending on the object of interest. The simple recommender regret on an object of interest f with respect to a divergence D at time step$t \in [1, T]$, is:

$$\mathcal{SRR}_t^{f,D}(\mathcal{H}_t) = D(f_t, f^\star), \tag{5}$$

where $f^\star$ is the target of the object $f \in \{\pi, V, Q\}$. Algorithms that attempt to minimise the simple recommender regret must make f_t close to $f^\star$ with respect to the divergence D. We use the notation $\mathcal{SRR}_t^{f,D}$ when there is no ambiguity. The simple recommender regret is a random variable of the policy d_t and the environment $\phi = (\phi_a : a \in \mathcal{A})$ since f_t is updated according to the sampled rewards. The expected simple recommender regret is defined as:

$$\mathbb{E}\left[\mathcal{SRR}_t^{f,D}\right] = \mathop{\mathbb{E}}_{\substack{a_i \sim d_i \\ r_{i,a_i} \sim \phi_{a_i} \\ \forall i \in [1,t]}} \left[D(f_t, f^\star)\right]. \tag{6}$$

Policies When the object of interest is $f = \pi$, then $D(\pi_1, \pi_2) \in \{\|\pi_1 - \pi_2\|, B_\Omega(\pi_1, \pi_2)\}$, where $\|\cdot\|$ is the primal norm for which Ω is strongly convex, and B_Ω is the Bregman divergence under Ω: $B_\Omega(\pi_1, \pi_2) = \Omega(\pi_1) - \Omega(\pi_2) - \langle \nabla\Omega(\pi_2), \pi_1 - \pi_2 \rangle$. For example, negative entropy is strongly convex with respect to the L_1 norm, while the Bregman divergence is equal to the KL divergence.

Values Value functions output scalars, so any non-negative scalar error function would be appropriate. Let $f = V$ and choose $D(V_1, V_2) = |V_1 - V_2|$, the absolute value function, as the distance for value-based simple recommender regret.

Q-values For Q-value-based recommender regret, we define $D(Q_1, Q_2) = \|Q_1 - Q_2\|_*$, which is the dual norm of the policy's primal norm.

3.2 Lower bounds on the recommender regret

We first establish minimax lower bounds for the expected recommender regret under norm-based statistical distances when the objects of study are values and Q-values. The results are presented in Theorem 1 and hold for any arbitrary regulariser.

Theorem 1 (Lower bound for state and action values recommender regret)**.** *Let T and K be the horizon and the number of actions respectively, Ω be a regulariser, and $\mathcal{SG}(\sigma^2)$ denote the class of σ-sub-Gaussian bandits. Then, for any algorithm, we have*

$$\sup_{\phi \in \mathcal{SG}(\sigma^2)} \mathbb{E}\left[\left|\hat{V}_{\Omega,T} - V_\Omega^*\right|\right] \geq \sqrt{\frac{2}{T}} \frac{\sigma}{8e}$$

$$\max_{\phi \in \mathcal{SG}(\sigma^2)} \mathop{\mathbb{E}}_{\phi} \left[\left\|\hat{Q}_{\Omega,T} - Q_{\Omega,\phi}^*\right\|_*\right] \geq \sqrt{\frac{K}{2T}} \frac{\sigma \min_{a,a' \in [K], a \neq a'} \|\mathbf{e}_a - \mathbf{e}_{a'}\|_*}{8e},$$

where $\mathbf{e}_a$ is a K-dimensional a-th indexed canonical basis vector.

The result is obtained by creating a base instance and a perturbed instance, averaging their respective expected recommender regret as a lower bound, and applying Bretagnolle-Huber. We leverage the triangle inequality to get a quantity that depends on the perturbation, which highlights the $\sqrt{T^{-1}}$ rate. Details are deferred to Appendix A. Similarly to simple regret minimisation (Bubeck et al., 2009, Remark 2), the Q-value recommender regret scales with $\sqrt{K/T}$, i.e. larger horizons make the problem easier, but more actions make it harder. Interestingly, the value lower bound does not depend on K, i.e., regardless of how many actions are in the problem, the difficulty is constant.

For policy recommender regret, we note that the convexity of Ω^* by itself does not provide us with meaningful lower bounds because there may exist multiple perturbations that provide the same policy, thus trivially setting the regret to zero. We provide a sufficient condition in Proposition 2. Before stating the propostion, let $E := \{v \in \mathbb{R}^A \mid \sum_{a \in \mathcal{A}} v_a = 0\}$ be the vector space parallel to the simplex and let us define the dual norm restricted to E as $\|\cdot\|_{\star,E} := \sup_{v \in E \|v\| \leq 1} \langle \cdot, v \rangle$. We will also need to make the following assumption

Assumption 1. *Given $\mathcal{Q} \subset \mathbb{R}^{\mathcal{A}}$ the set of targets $Q^\star_\Omega$, we assume that $\Pi := \nabla\Omega^*(\mathcal{Q})$ is a subset of ri($\Delta_{\mathcal{A}}$),and that for all $\pi \in$ hull(Π) and $v \in E$, we have $\langle \nabla^2_E \Omega(\pi)v, v\rangle \leq L\|v\|^2$.*[3]

This assumption holds for instance for the entropy function for which, given that $\mathcal{Q}$ is compact we have that $\nabla\Omega^*(\mathcal{Q})$ is in the relative interior of the simplex and Ω^* has the smoothness property described above.

Proposition 2. *Under Ass. 1, $\forall Q, Q' \in \mathcal{Q}$ we have $\langle \nabla\Omega^*(Q) - \nabla\Omega^*(Q'), Q - Q'\rangle \geq \frac{1}{L}\|Q - Q'\|^2_{*,E}$.*

The proof is in Appendix B as well as a discussion on why strict convexity of Ω^* is not enough. It relies on a simplex-contrained version of the fact that this property of the regulariser Ω implies the dual Ω^* is $\frac{1}{L}$-strongly convex, which we use to lower-bound the expected recommender regret. As a concrete example, it follows from Proposition 2 that negative entropy is L-smooth on the policy space $\Pi = \{\pi : \pi(a) \geq \epsilon > 0, \forall a \in \mathcal{A}\}$. Note that Assumption 1 is made solely to obtain an interesting analysis of a broader class of regularisers on a subset of their domain. Indeed, compact sets of policies may be overly restrictive for some regularisers. We obtain the following lower bound on the policy recommender regret.

Theorem 3 (Lower bound for policy recommender regret)**.** *Let T and K be the horizon and the number of actions respectively and let us consider the class of σ-sub-Gaussian bandits, $\mathcal{SG}(\sigma^2)$ with target $Q^\star_\Omega \in \mathcal{Q}$, we call this set $\mathcal{SG}(\sigma^2, \mathcal{G})$. Then if, Ass 1 holds we have*

$$\max_{\phi \in \mathcal{SG}(\sigma^2,\mathcal{G})} \mathop{\mathbb{E}}_{\phi}\left[\left\|\hat{\pi}_{\Omega,T} - \pi^\star_\phi\right\|\right] \geq \sqrt{\frac{K}{2T}} \frac{\sigma \min_{a,a' \in [K], a \neq a'} \|\mathbf{e}_a - \mathbf{e}_{a'}\|_{*,E}}{8L}$$

$$\max_{\phi \in \mathcal{SG}(\sigma^2,\mathcal{G})} \mathop{\mathbb{E}}_{\phi}\left[B_\Omega(\hat{\pi}_{\Omega,T}, \pi^\star_\phi)\right] \geq \frac{K}{T} \frac{\beta\sigma^2 \min_{a,a' \in [K], a \neq a'} \|\mathbf{e}_a - \mathbf{e}_{a'}\|^2_{*,E}}{64L^2 e},$$

where $\mathbf{e}_a$ is a K-dimensional a-th indexed canonical basis vector.

The proof can be found in Appendix A, and reuses the same proof structure from Theorem 1, where the Bregman divergence is lower bounded using the β-strong convexity of Ω. Once again, these lower bounds indicate that larger horizons make the problem easier, whereas a larger number of actions and smaller smoothness constants L make the problem harder. Indeed, when L is small, vastly different policies can have quasi-identical Q-values. However, this result provides evidence that the multi-armed sampling lower bound of Pedramfar & Ravanbakhsh (2026) on the simple sampling regret using the total variation distance may be loose, since our lower bound is greater by a factor of K.

3.3 Algorithms for the recommender regret

We begin with an analysis of the notions of regret, assuming no knowledge of the form of the regulariser, other than its being strongly convex. Theorem 4 provides general upper bounds on each notion of simple recommender regret for any sampling policy d_t with non-zero probability mass on each action. Importantly, the upper bounds always achieve near-optimal dependence of on T, but only achieve optimal dependence on K when the dual norm is $\|\cdot\|_\infty$, or when $f = V$. This theorem allows us to conclude that uniform sampling is an effective algorithm for regularised RL estimation.

Theorem 4 (Sampling policies with non-zero probability mass achieve optimal regret bounds)**.** *Suppose Ω is β-strongly convex with dual $V_\Omega(s) = \Omega^*(Q(s,\cdot))$, with σ-sub-Gaussian reward distributions $\phi_a, \forall a \in \mathcal{A}$. For any $T \geq \frac{2\ln(2K/\delta)}{\zeta^2}$, with $\delta \in O(1/T^2)$, and any algorithm with sampling policy d_t satisfying $d_t(a) \geq \zeta > 0\ \forall a \in \mathcal{A}, \forall t \in [1, T]$, the following expected regret bounds hold:*

1. *When $\|\cdot\|_\star = \|\cdot\|_\infty$ or when $f = V$:*

$$\mathbb{E}\left[\mathcal{SRR}^{f,D}_T\right] \in \tilde{O}\left(\sqrt{\frac{1}{\zeta T}}\right), \qquad \mathbb{E}\left[\mathcal{SRR}^{\pi,B_\Omega}_T\right] \in \tilde{O}\left(\frac{1}{\zeta T}\right). \tag{7}$$

2. *When $\|\cdot\|_\star$ is a general dual norm with $\|\mathbf{e}'\|_\star = \max_a \|\mathbf{e}_a\|_\star$:*

$$\mathbb{E}\left[\mathcal{SRR}^{f,D}_T\right] \in \tilde{O}\left(K\|\mathbf{e}'\|_\star \sqrt{\frac{1}{\zeta T}}\right), \qquad \mathbb{E}\left[\mathcal{SRR}^{\pi,B_\Omega}_T\right] \in \tilde{O}\left(K^2\|\mathbf{e}'\|^2_\star \frac{1}{\zeta T}\right), \tag{8}$$

[3]The notation ri(C) and hull(C) respectively stand for the relative interior and the convex hull of C, for and ∇_E correspond to the gradient restricted to the vector space E.

where the left and right terms bound norm-based and Bregman divergence-based regrets, respectively. $\tilde{O}$ leaves out slow-growing logarithmic terms, following Pedramfar & Ravanbakhsh (2026).

Theorem 4, proof in Appendix D, makes it clear that an optimal sampling policy should maximise ζ, to minimise the upper bound. Intuitively, **uniform sampling** should achieve better rates than other sampling policies, like E3W (Equation 3). Corollary 13 makes this intuition concrete, demonstrating that E3W achieves upper bounds that are looser by a $\log(T)$ term. Our results match those of Pedramfar & Ravanbakhsh (2026) in terms of dependence on T (they don't report dependence on K) and highlight a theoretical tension between sampling and recommendation: while sampling requires minimal exploration, recommendation requires pure exploration, as in Bubeck et al. (2009).

Limitations of the analysis We sought to produce general upper bounds of the simple recommender regret that would hold regardless of the regulariser, and consequently, the form of the policy or value functions. As such, though our upper bounds fairly tight in T, they are still (sometimes) loose in K. Future work will propose regulariser-specific algorithms (*e.g.*, like Pedramfar & Ravanbakhsh (2026) for entropy-regularised RL in sampling regret) to achieve tighter upper bounds.

4 Monte Carlo planning for recommendations

Monte Carlo tree search is a planning algorithm that naturally extends recommender bandit problems to generating recommendations in multi-state RL environments. Following the algorithmic setup of Xiao et al. (2019) and Dam et al. (2021), we consider a search tree $\mathcal{T}$ rooted at s_0 of depth H with deterministic transition function $\tau : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ and deterministic non-leaf rewards $r(s, a)$. Each non-leaf node $s \in \mathcal{T}$ acts as a bandit instance operating over actions $a \in \mathcal{A}$.[4] Each node s of the tree maintains value estimates $\hat{V}_t(s)$, $\hat{Q}_t(s, a)$, and visit counts $N_t(s, a)$ across total simulation steps t.

At each simulation step t, a sampling policy $d_t(\cdot|s)$ navigates from s_0 through $\mathcal{T}$ along trajectory $(s_0, a_0) \to (s_1, a_1) \to \cdots \to s_L$ of length $L \leq H$, with $a_i \sim d_t(\cdot|s_i)$ and $s_{i+1} = \tau(s_i, a_i)$, adding s_L to the tree. Visit counts are updated recursively via $N_t(s_i, a_i) \leftarrow N_t(s_i, a_i) + 1$ along the trajectory. While standard MCTS algorithms like UCT (Kocsis & Szepesvári, 2006) and AlphaZero (Silver et al., 2017) use empirical average backups, regularised tree search algorithms use the Legendre-Fenchel dual $\Omega^*(\hat{Q})$ directly, without averaging:

$$\hat{V}_t(s_i) = \Omega^*\left(\hat{Q}_t(s_i, \cdot)\right); \quad \hat{Q}_t(s_i, a_i) = \begin{cases} r(s_i, a_i) + \gamma R(s_L) & i = L - 1 \\ r(s_i, a_i) + \gamma \hat{V}_t(\tau(s_i, a_i)) & i < L - 1 \end{cases}, \tag{9}$$

where $R(s_L)$ is a value estimate generated via a Monte Carlo rollout, or a learned value estimator (Silver et al., 2017; Xiao et al., 2019; Dam et al., 2021; Painter et al., 2023).

Our main contribution is the theoretical analysis of regularised tree search (Xiao et al., 2019; Dam et al., 2021) for generating recommendations. Theorem 5 shows that the simple recommender regret concentrates at the root of the tree, indicating that recommender MCTS can effectively generate regularised policy, value, and Q-value recommendations (proof in Appendix E).

Theorem 5. *For any notion of simple recommender regret $\mathcal{SRR}_T^{f,D}$ at the root node s_0 of a search tree of depth H with deterministic transitions and deterministic non-leaf rewards, a sampling policy $d_t(a|s) \geq \zeta > 0, \forall a \in \mathcal{A}, \forall s \in \mathcal{S}, \forall t \in [1, T]$, and $\epsilon \leq C_3\zeta$, then $\exists C_1, C_2, C_3 \in \mathbb{R}_{>0}$, such that*

$$\mathbb{P}\left(\mathcal{SRR}_T^{f,D} \geq \epsilon\right) \leq C_1(KT)^H \exp\left(-\frac{\zeta^H T\epsilon^2}{K^{2H}C_2}\right). \tag{10}$$

Corollary 6 integrates these concentration inequalities to achieve the first expected regret bounds under regularised MCTS (proof in Section F.3).[5]

Corollary 6 (Recommender MCTS expected regret bound)**.** *Let the Q-value error be bounded: $\left\|\hat{Q}_T(s_0, \cdot) - Q^\star(s_0, \cdot)\right\|_\infty \leq 2R_{\max} < \infty$. Under the hypotheses of Theorem 5, the uniform sampling policy $d_t(a|s) = 1/K$ maximizes $\zeta^\star = 1/K$. For $\delta \in (0, 1]$ this choice minimizes the search*

[4] MCTS can be intuitively viewed as a sequence of bandits stacked one on top of another.

[5] Prior work could not generate such bounds since they specified restrictive errors ϵ.

horizon $T \geq \frac{C_2 K^{2+2H}}{\zeta^H C_3^2} \ln\left(\frac{C_1 (TK)^H}{\delta}\right)$*, for which the following hold in expectation for MCTS:*

$$\mathbb{E}\left[\mathcal{SR}\mathcal{R}_T^{f,D}\right] \in \tilde{O}\left(\sqrt{\frac{K^{3H}}{T}}\right); \qquad \mathbb{E}\left[\mathcal{SR}\mathcal{R}_T^{\pi,B_\Omega}\right] \in \tilde{O}\left(\frac{K^{3H}}{T}\right), \tag{11}$$

where the terms bound norm-based and Bregman divergence-based regrets, respectively.

While algorithms like E3W also achieve exponential convergence in probability, they introduce logarithmic penalties in the exponent, as noted in Corollary 13 (Xiao et al., 2019; Dam et al., 2021). Uniform sampling thus achieves tighter MCTS rates in both concentration (Theorem 5) and expected regret (Corollary 6), again following the results of Bubeck et al. (2009) for simple regret minimisation. Remark 1 comments on the different rates achieved in prior work and our bound. These results make it clear that uniform sampling is optimal for recommender MCTS, with sample complexity scaling exponentially with tree depth, and polynomially with action space size.

5 Experiments

We evaluate recommender bandits and MCTS algorithms in three domains to showcase their convergence efficiency, and utility in diverse decision-making contexts. We first study E3W and uniform sampling in synthetic trees to empirically verify our theoretical claims (Section 5.1). We then demonstrate the potential of recommenders for finding high-value candidate solutions for arcade games in OCTAX (Section 5.2) and for biological sequence-design tasks (Section 5.3). Note that we are the first to employ regularised MCTS in either OCTAX or scientific discovery tasks, demonstrating for the first time the efficacy of regularised planning with recommenders in these domains. In all domains, the tree search algorithms described in Section 2.3 are employed, including the max-entropy (MENTS), relative-entropy/KL (RENTS), Tsallis-entropy (TENTS), and soft-mellowmax (SMM) backup operators, alongside AlphaZero's variant of UCT.

5.1 Synthetic tree experiments

Xiao et al. (2019) proposed a synthetic tree environment as a testbed for regularised MCTS convergence. We follow their experimental setup, analysing the Q-value error since by Lemma 11, Q-value error upper bounds all forms of simple recommender regret. Each tree is parameterised by a branching factor k and depth d. At all nodes of the tree, the reward is drawn from a normal distribution upon each visit i: $r_i(s,a) \sim \mathcal{N}(\mu_{s,a}, 1)$. For each tree, we randomly select $\mu_{s,a} \sim \mathcal{N}(0,1)$ for each action at each node, which is held constant for the whole simulation budget. Regret is averaged over five such μ seeds. Q-value error is measured in accordance with each regulariser's dual norm: MENTS, RENTS, and SMM use the L_∞ norm, while TENTS and UCT use the L_2 norm.

Figure 1 illustrates the differences in convergence speed between the MCTS methods. **Uniform sampling outperforms E3W across the presented trees**. This trend continues to hold for Bregman divergence-based regrets (Figure 9) across all trees, but not for value regret (Figure 5), nor norm-based policy regrets (Figure 7), where uniform sampling does not always achieve the lowest error curve for each MCTS algorithm, see Appendix G. This discrepancy is interesting, and reveals a gap in our theoretical analysis: the algorithms' dependence on K is not perfectly modelled, and the bounds on value and policy errors are somewhat loose. The higher error of TENTS an UCT in the figure is due to the difference in norms (L_2 vs. L_∞) between the regularised methods (*c.f.*, Figure 4). Finally, we note that our results differ from Bubeck et al. (2009), where UCB outperforms uniform sampling, despite the latter being theoretically optimal. In our experiments, uniform sampling often outperforms UCT and greedier sampling (E3W) in many tree and algorithmic settings.

5.2 Arcade games in OCTAX

OCTAX, (Radji et al., 2026), is a massively parallelized arcade-game environment that emulates a suite of classic arcade-style games (like Atari). For each of 19 OCTAX games, we train a separate Q-network with a parallelised Q-network (PQN)-style algorithm (Gallici et al., 2025). The frozen Q-network is used to estimate values inside regularised MCTS (*c.f.*, Equation 9, $R \leftarrow \Omega^*(Q_\theta(s', \cdot))$). We compare the learned policy and the MCTS-improved policies, greedily selecting the highest

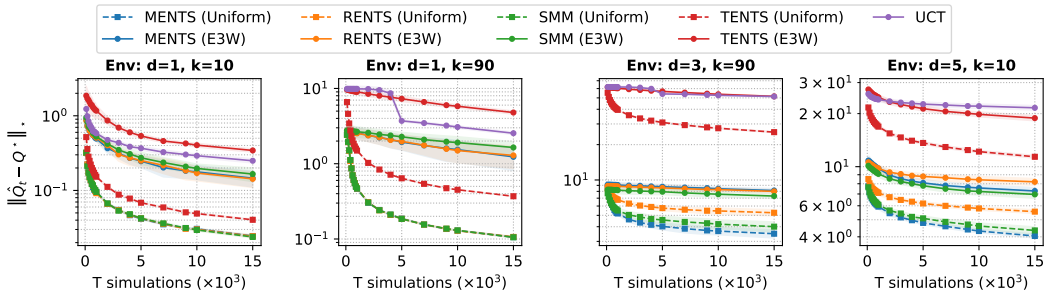

Figure 1: Q-value estimation error in recommender MCTS over 15k simulations and trees of varying depth d, and action space size k, with 100 random tree search seeds, for each of the five μ seeds.

Q-value action for each. The inverse temperature parameter $\omega = 5$ for all regularised tree search methods, and $\alpha = 1$ for SMM. The E3W exploration constant is fixed at $\epsilon = 0.1$. We sweep over simulation budgets to evaluate performance as simulations increase. Each configuration is evaluated over 100 random seeds.

In Figure 2, we report mean episodic return for a representative subset of the games. This subset was chosen by filtering out the games that saturated quickly (to study performance with T), games that didn't improve with search (*e.g.*, search is worse in Tank in Figure 16), and choosing a wide range of episode lengths (100 for Brix, 4000 for Filter, *c.f.*, Figure 17). Regularised MCTS clearly improves the performance of the PQN agent (baseline) in these games, especially as the number of simulations increases. Interestingly, uniform sampling and E3W perform similarly at high simulation counts, but E3W often outperforms uniform sampling at lower simulation counts. We speculate that this is because E3W preferentially samples greedily (with $\hat{\pi}_t$), so higher-reward actions are likely better-explored earlier than with uniform sampling. This would explain why uniform sampling achieves similar performance with more simulations. These results support the results of prior work, namely Xiao et al. (2019) and Dam et al. (2021), which showed that regularised MCTS algorithms could achieve superior performance to UCT in Atari games.See Appendix H for more experiment details and Appendix I for plots of execution times, episode lengths, and mean returns for all 19 games.

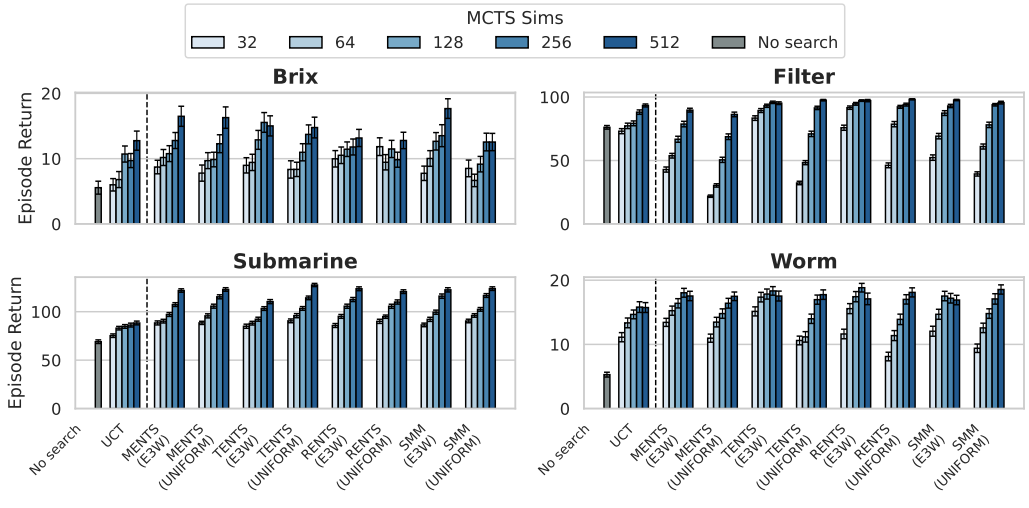

Figure 2: Episodic returns in select OCTAX tasks showing improved performance with increased MCTS simulations. The error bars represent the standard error, with 100 random seeds per setting.

5.3 SCIENTIFIC DISCOVERY

Following the experimental setup of Jiralerspong et al. (2025), we now consider a number of biological sequence design tasks popular in the generative flow network (GFlowNet) literature (Whitammer et al., 2022; Jain et al., 2023). Using the protein- and DNA-sequence datasets of Trabucco et al. (2022), a transformer model is trained as an autoregressive proxy reward model. The proxy reward model is then fixed, and used to train a trajectory general mellowmax (Jiralerspong et al., 2025, TGM) model, which we use as the baseline, and to estimate values inside regularised MCTS. TGM is chosen over other baselines like GFlowNets, because it was shown to achieve state-of-the-art performance in the three tasks we evaluate on: anti-microbial peptide (AMP), green fluorescent protein (GFP), and 5' untranslated region sequence (UTR), see Appendix J for details.

Search evaluation. We compare the TGM baseline model against MCTS-improved policies, sampling actions stochastically according to the policy estimator. This is because in biological sequence design, diversity is prioritised over return maximisation. Simulation budgets are swept from 32 to 1024 (the 1024-simulation tier is evaluated only for AMP and UTR). We sweep the inverse temperature parameter ω over from 0.025 to 40, for each configuration in each task, including the baseline model (whose logits are scaled by ω), and UCT (whose policy output temperature is set to $1/\omega$). We again set $\alpha = 1$ for SMM. For each omega, we generate 100 sequences, over five random seeds.

Metrics. Using the metrics in Jiralerspong et al. (2025) and Whitammer et al. (2022), we evaluate the diversity of high-reward sequences generated by each configuration, reporting a diversity-aware top-k reward: sequences are ranked by reward and greedily added to a selected set only if they are at least a task-specific Levenshtein distance δ from every sequence already selected. This measures each configuration's coverage of distinct modes the reward landscape rather than its best sample.

Results. Figure 3 illustrates the performance distribution of each configuration in the scientific discovery experiments. Notably, tree search-improved policies, particularly Tsallis entropy-based tree search (TENTS), were able to match or exceed the top-100 mean of the baseline in all tasks, especially with uniform sampling. In fact, uniform sampling outperforms E3W for almost all search algorithms, in almost all tasks. This claim is further supported by Table 3, which lists the top-100 mean reward for all methods. While prior work (Morozov et al., 2024) improved GFlowNet policies with MCTS, they only reported similarity metrics between reward functions and MCTS-improved models on synthetic GFlowNet tasks. Thus, our results indicate for the first time that recommender-based regularised MCTS can be useful for real-world scientific discovery applications.

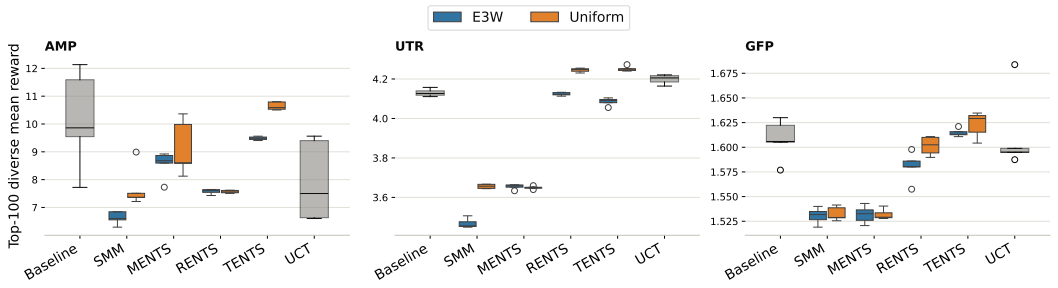

Figure 3: Top-100 diverse reward plots for the three scientific discovery tasks, with the TGM model as baseline, and as the value predictor in the MCTS tree. Tree search is competitive with, or beats, the mean performance of the baseline. Uniform sampling outperforms E3W for almost all algorithms.

6 CONCLUSION

We present recommender bandits, and new class of bandit-based estimators for regularised RL based on the multi-armed sampling framework. Theoretically, we show that our notions of recommender regret are well-posed, in that they produce non-trivial lower bounds, and that these bounds can be achieved with uniform policies, unlike the greedier policies required by multi-armed sampling. Experimentally, we demonstrate that recommender-motivated uniform policies for bandits and MCTS match or outperform prior work in synthetic trees, arcade games, and most notably, scientific discovery. Future work will tighten the upper bounds, using regulariser-specific algorithms.

AI USE STATEMENT

In this work, we used generative AI tools for a number of tasks including

- proof checking
- proof ideation, especially proposing certain tools to use, or techniques
- to help with learning certain mathematical tools
- to transcribe hand-written notes to Latex
- to review the theorem and lemma statements for correctness
- to check code for correctness
- to write plotting and latex table generation code
- to help write some experiment code, particularly boilerplate evaluation scripts and bash scripts
- reviewing the writing of the paper, especially to focus narratives and reduce fluff

We have not used generative AI tools for writing the paper, and took great care to ensure that all the written work presented in this manuscript was human-generated, but perhaps AI-assisted. We have also not blindly followed the written code, nor blindly used any theoretical tools that were presented in the early phases of this work. We again, took great care to ensure that our work was sensible, and that it was honestly reported and conducted. The other AI disclosures are not relevant to this work.

We have reviewed all AI-assisted work. In particular, the core codebase has been written entirely by humans, and has been thoroughly checked, tested, and debugged. The plotting code was already written, but AI was used to modify it for different plots and objectives. These were always thoroughly vetted for correctness. All latex transcriptions were treated as a starting point, and modified to fit the math we wrote down.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kavosh Asadi and Michael L Littman. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, pp. 243–252. PMLR, 2017.

Heinz H Bauschke, Jonathan M Borwein, et al. Legendre functions and the method of random bregman projections. *J. Convex Anal*, 4(1):27–67, 1997.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90040-7. URL https://www.sciencedirect.com/science/article/pii/0041555367900407.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and trends in Machine Learning*, 8(3-4):266–267, 2015.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pp. 23–37. Springer, 2009.

Tuan Q Dam, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. Convex regularization in monte-carlo tree search. In *International Conference on Machine Learning*, pp. 2365–2375. PMLR, 2021.

Ivo Danihelka, Arthur Guez, Julian Schrittwieser, and David Silver. Policy improvement by planning with gumbel. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=bERaNdoegnO.

Matteo Gallici, Mattie Fellows, Benjamin Ellis, Bartomeu Pou, Ivan Masmitja, Jakob Nicolaus Foerster, and Mario Martin. Simplifying deep temporal difference learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7IzeL0kflu.

Yaozhong Gan, Zhe Zhang, and Xiaoyang Tan. Stabilizing q learning via soft mellowmax operator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7501–7509, 2021.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Moksh Jain, Tristan Deleu, Jason Hartford, Cheng-Hao Liu, Alex Hernandez-Garcia, and Yoshua Bengio. Gflownets for ai-driven scientific discovery. *Digital Discovery*, 2(3):557–577, 2023.

Marco Jiralerspong, Esther Derman, Danilo Vucetic, Nikolay Malkin, Bilun Sun, Tianyu Zhang, Pierre-Luc Bacon, and Gauthier Gidel. Robust reinforcement learning for discrete compositional generation via general soft operators. *arXiv preprint arXiv:2506.17007*, 2025.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13(1):1865–1890, 2012.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. ISBN 9781108486828. URL https://books.google.ca/books?id=bydXzAEACAAJ.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018. doi: 10.1109/LRA.2018.2800085.

Kyungjae Lee, Sungyub Kim, Sungbin Lim, Sungjoon Choi, Mineui Hong, Jae In Kim, Yong-Lae Park, and Songhwai Oh. Generalized tsallis entropy reinforcement learning and its application to soft mobile robots. In *Robotics: science and systems*, volume 16, pp. 1–10, 2020.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013. URL https://arxiv.org/abs/1312.5602.

Nikita Morozov, Daniil Tiapkin, Sergey Samsonov, Alexey Naumov, and Dmitry Vetrov. Improving gflownets with monte carlo tree search. In *ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2024.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Michael Painter, Mohamed Baioumy, Nick Hawes, and Bruno Lacerda. Monte carlo tree search with boltzmann exploration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=NG4DaApavi.

Mohammad Pedramfar and Siamak Ravanbakhsh. Multi-armed sampling problem and the end of exploration. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=wVIdRkkbGe.

Waris Radji, Thomas Michel, and Hector Piteau. Octax: Accelerated CHIP-8 arcade environments for reinforcement learning in JAX. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=BdeUsYlyIf.

Philippe Rigollet. 18.s997: High dimensional statistics; lecture notes, 2015. URL https://ocw.mit.edu/courses/18-s997-high-dimensional-statistics-spring-2015/619e4ae252f1b26cbe0f7a29d5932978_MIT18_S997S15_CourseNotes.pdf.

R.T. Rockafellar. *Convex Analysis*, chapter 12. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1997. ISBN 9780691015866. URL https://books.google.ca/books?id=1TiOka9bx3sC.

S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications.* PhD thesis, The Hebrew University, 2007.

Shai Shalev-Shwartz and Yoram Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2):115–142, 2007.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning, Second Edition: An Introduction*. Bradford Books, Cambridge, Massachusetts London, England, 2020.

Brandon Trabucco, Xinyang Geng, Aviral Kumar, and Sergey Levine. Design-bench: Benchmarks for data-driven offline model-based optimization. In *International Conference on Machine Learning*, pp. 21658–21676. PMLR, 2022.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Remi Munos, and Matthieu Geist. Leverage the average: an analysis of kl regularization in reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12163–12174. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8e2c381d4dd04f1c55093f22c59c3a08-Paper.pdf.

Danilo Vucetic and Gauthier Gidel. Soft mellowmax monte carlo planning. In *2026 International Conference on Learning Representations Workshop on Recursive Self Improvement*, 2026. URL https://openreview.net/forum?id=SeZCSEGZTq.

M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019a. ISBN 9781108498029. URL https://books.google.ca/books?id=IluHDwAAQBAJ.

M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, chapter 2 and 3, pp. 21–97. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019b. ISBN 9781108498029. URL https://books.google.ca/books?id=IluHDwAAQBAJ.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, pp. 10686–10696. PMLR, 2021.

Esmeralda Whitammer, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022.

Chenjun Xiao, Ruitong Huang, Jincheng Mei, Dale Schuurmans, and Martin Müller. Maximum entropy monte-carlo planning. *Advances in Neural Information Processing Systems*, 32, 2019.

Ruipeng Zhang, Ya-Chien Chang, and Sicun Gao. When maximum entropy misleads policy optimization. In *Proceedings of the 42nd International Conference on Machine Learning, ICML 2025, Vancouver, Canada*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

A PROOF OF THE LOWER BOUNDS

Let $\mathcal{SG}(\sigma^2)$ be the class of σ^2-sub-Gaussian bandits, let $\mu := (\mu_1, \mu_2, \dots, \mu_K)$ denote a vector of means for a Gaussian K-Armed Bandit with variance σ^2 for all arms (i.e. the reward $r_{t,a} \sim \mathcal{N}(\mu_i, \sigma^2)$ for each arm $a \in [K]$) and $N_a(t)$ denote the number of times arm a has been played at time $t \in [1, \dots, T]$. We also introduce a perturbation vector $\epsilon \in \mathbb{R}^K \neq \mathbf{0}$, and a *perturbed* Gaussian instance whose means are perturbed by ϵ, with means $\mu + \epsilon$. We refer to these two instances as μ and $\mu + \epsilon$ directly. Note that the mean vectors of each instance should be interpreted as the Q-values for these instances. As we are looking for a minimax lower bound on the class of σ-sub-Gaussian instances $\mathcal{SG}(\sigma^2)$, investigating Gaussian instances yields a valid lower bound. The instances μ and $\mu+\epsilon$ are in $\mathcal{SG}(\sigma^2)$. We now introduce a lemma on the number of times a pair of arm can be played.

Lemma 7. *Let $K \geq 2$ be the number of arms and $T \geq 1$ be the time horizon. For any sequence of arm selections generated by an algorithm $\mathcal{F}$, let $N_a(T)$ denote the total number of pulls of arm $a \in [K]$ up to time T. Then, there exists a distinct pair of arms (i, j) with $i \neq j$ such that*

$$N_i(T) + N_j(T) \leq \frac{2T}{K}.$$

Proof. Let $N_1(T) \leq N_2(T) \leq \cdots \leq N_K(T)$ denote the number of pulls of each arm in a K-armed bandit. Then, $\sum_{a=1}^{K} N_a(T) = T$.

Now, suppose $N_1(T) + N_2(T) > \frac{2T}{K}$, then since $N_1(T) \leq N_2(T)$ it follows that

$$2N_2(T) \geq N_1(T) + N_2(T) > \frac{2T}{K} \implies N_2(T) > \frac{T}{K}.$$

For every arm $i \geq 3$, we have

$$N_i(T) \geq N_2(T) > \frac{T}{K}.$$

Then,

$$\begin{aligned}\sum_{a=1}^{K} N_a(T) &= N_1(T) + N_2(T) + \sum_{a=3}^{K} N_a(T) \\ &> \frac{2T}{K} + (K-2)\frac{T}{K} \\ &= T.\end{aligned}$$

This is a contradiction, as $\sum_{a=1}^{K} N_a(T) = T > T$. □

Note that although the indices rely on the sampling policies d_t generated by $\mathcal{F}$, such a pair always exists regardless of $\mathcal{F}$ and policies d_t.

For clarity, we denote $\mathbb{E}_\mu$ and $\mathbb{P}_\mu$ as the expected value and probability under the probability measure induced by the T round interaction between instance ϕ, algorithm $\mathcal{F}$ and its associated policies d_t. Similarly, f^*_ϕ, where $f \in \{V_\Omega, Q_\Omega, \pi_\Omega\}$, refers to the target object when the instance is ϕ. Lastly, we denote $\Delta_{\|\cdot\|,f} = \left\|f^*_\mu - f^*_{\mu+\epsilon}\right\|$ as a shorthand.

We provide two lemmas that will be used throughout the proofs. The first concerns recommender regret when the divergence D is a norm $\|\cdot\|$. The second is specifically when D is the Bregman divergence B_Ω. Throughout the proofs, $\mathbf{e}_a$ denotes a K-dimensional a-th indexed canonical basis vector and $\mathbb{1}$ denotes a K-dimensional all-ones vector.

Lemma 8. *Let f be the object of interest, T be the horizon, and denote $\Delta_{\|\cdot\|,f} = \left\|f^*_\mu - f^*_{\mu+\epsilon}\right\|$ for some arbitrary norm $\|\cdot\|$. Given perturbation vectors $\epsilon \in \left\{\sqrt{\frac{K}{2T}}\sigma\left(\mathbf{e}_i - \mathbf{e}_j\right), \sqrt{\frac{2}{T}}\sigma\mathbb{1}\right\}$, where (i, j) are the indices of the least played pair of arms under some arbitrary algorithm $\mathcal{F}$, the following holds for any subgaussian instance $\mathcal{SG}(\sigma^2)$:*

$$\min_{\mathcal{F}} \max_{\phi \in \mathcal{SG}(\sigma^2)} \mathop{\mathbb{E}}_{\phi}\left[\left\|f_T - f^*_\phi\right\|\right] \geq \frac{\Delta_{\|\cdot\|,f}}{8e}.$$

Proof. Consider the event $A := \left\{ \left\| f_\mu^* - f_T \right\| \geq g_T \right\}$. Then, choosing any algorithm $\mathcal{F}$, we have

$$\max_{\phi \in \mathcal{SG}(\sigma^2)} \mathop{\mathbb{E}}_{\phi} \left[\left\| f_T - f_\phi^* \right\| \right] \geq \frac{\mathbb{E}_\mu \left[\left\| f_T - f_\mu^* \right\| \right] + \mathbb{E}_{\mu+\epsilon} \left[\left\| f_T - f_{\mu+\epsilon}^* \right\| \right]}{2}. \tag{12}$$

Focusing on the numerator, we obtain

$$\begin{aligned} \mathop{\mathbb{E}}_{\mu} \left[\left\| f_T - f_\mu^* \right\| \right] + \mathop{\mathbb{E}}_{\mu+\epsilon} \left[\left\| f_T - f_{\mu+\epsilon}^* \right\| \right] &\geq \mathop{\mathbb{E}}_{\mu} \left[\left\| f_T - f_\mu^* \right\| \middle| A \right] \mathbb{P}_\mu(A) \\ &\quad + \mathop{\mathbb{E}}_{\mu+\epsilon} \left[\left\| f_T - f_{\mu+\epsilon}^* \right\| \middle| A^c \right] \mathbb{P}_{\mu+\epsilon}(A^c) \\ &\geq g_T \mathbb{P}_\mu(A) + (\Delta_{\|\cdot\|,f} - g_T) \mathbb{P}_{\mu+\epsilon}(A^c), \end{aligned} \tag{13}$$

where the last inequality comes from the fact that A, A^c occur to replace the expected values, combined with the triangle inequality. By choosing $g_T = \frac{\Delta_{\|\cdot\|,f}}{2}$, we obtain

$$\begin{aligned} \mathop{\mathbb{E}}_{\mu} \left[\left\| f_T - f_\mu^* \right\| \right] + \mathop{\mathbb{E}}_{\mu+\epsilon} \left[\left\| f_T - f_{\mu+\epsilon}^* \right\| \right] &\geq \frac{\Delta_{\|\cdot\|,f}}{2} \left(\mathbb{P}_\mu(A) + \mathbb{P}_{\mu+\epsilon}(A^c) \right) \\ &\geq \frac{\Delta_{\|\cdot\|,f}}{4} \left(\exp\left(-D_{KL}\left(\mathop{\mathbb{P}}_{\mu} \middle\| \mathop{\mathbb{P}}_{\mu+\epsilon} \right) \right) \right) \\ &= \frac{\Delta_{\|\cdot\|,f}}{4} \left(\exp\left(-\sum_{a=1}^{K} \mathop{\mathbb{E}}_{\mu} \left[N_a(T) \right] \frac{\epsilon_a^2}{2\sigma^2} \right) \right). \end{aligned} \tag{14}$$

The second line is obtained from the Bretagnolle-Huber inequality. The third line applies the KL-divergence decomposition lemma for bandits (Lemma 15.1, Sutton & Barto (2020)), then take the KL-divergence for Gaussian distributions. Now, if we choose $\epsilon = \sqrt{\frac{K}{2T}} \sigma (\mathbf{e}_i - \mathbf{e}_j)$, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\mu} \left[\left\| f_T - f_\mu^* \right\| \right] + \mathop{\mathbb{E}}_{\mu+\epsilon} \left[\left\| f_T - f_{\mu+\epsilon}^* \right\| \right] &\geq \frac{\Delta_{\|\cdot\|,f}}{4} \left(\exp\left(-\sum_{a=1}^{K} \mathop{\mathbb{E}}_{\mu} \left[N_a(T) \right] \frac{\epsilon_a^2}{2\sigma^2} \right) \right) \\ &= \frac{\Delta_{\|\cdot\|,f}}{4} \left(\exp\left(-\mathop{\mathbb{E}}_{\mu} \left[N_i(T) \right] \frac{\epsilon_i^2}{2\sigma^2} - \mathop{\mathbb{E}}_{\mu} \left[N_j(T) \right] \frac{\epsilon_j^2}{2\sigma^2} \right) \right) \\ &\geq \frac{\Delta_{\|\cdot\|,f}}{4} \left(\exp\left(-\frac{2T}{K} \left(\frac{\epsilon_i^2}{2\sigma^2} + \frac{\epsilon_j^2}{2\sigma^2} \right) \right) \right) \\ &= \frac{\Delta_{\|\cdot\|,f}}{4} \left(\exp\left(-\frac{T}{K} \left(\frac{K}{2T} + \frac{K}{2T} \right) \right) \right) \\ &\geq \frac{\Delta_{\|\cdot\|,f}}{4e}. \end{aligned} \tag{15}$$

Similarly, using $\epsilon = \sqrt{\frac{2}{T}} \sigma \mathbb{1}$ we obtain

$$\begin{aligned} \mathop{\mathbb{E}}_{\mu} \left[\left\| f_T - f_\mu^* \right\| \right] + \mathop{\mathbb{E}}_{\mu+\epsilon} \left[\left\| f_T - f_{\mu+\epsilon}^* \right\| \right] &\geq \frac{\Delta_{\|\cdot\|,f}}{4} \left(\exp\left(-\sum_{a=1}^{K} \mathop{\mathbb{E}}_{\mu} \left[N_a(T) \right] \frac{\epsilon_a^2}{2\sigma^2} \right) \right) \\ &= \frac{\Delta_{\|\cdot\|,f}}{4} \left(\exp\left(-\frac{1}{T} \sum_{a=1}^{K} \mathop{\mathbb{E}}_{\mu} \left[N_a(T) \right] \right) \right) \\ &= \frac{\Delta_{\|\cdot\|,f}}{4e}. \end{aligned} \tag{16}$$

Therefore, when $\epsilon \in \left\{\sqrt{\frac{K}{2T}}\sigma\left(\mathbf{e}_i - \mathbf{e}_j\right), \sqrt{\frac{2}{T}}\sigma\mathbb{1}\right\}$, since all of the above hold for any arbitrary algorithm $\mathcal{F}$, we have that

$$\min_{\mathcal{F}} \max_{\phi \in \mathcal{SG}(\sigma^2)} \mathop{\mathbb{E}}_{\phi}\left[\left\|f_T - f_\phi^*\right\|\right] \geq \frac{\Delta_{\|\cdot\|,f}}{8e}.$$

□

Lemma 9. *Let f be the object of interest, T be the horizon, and denote $\Delta_{B,f} = \left\|f_\mu^* - f_{\mu+\epsilon}^*\right\|^2$ for some arbitrary norm $\|\cdot\|$. Given $\epsilon = \sqrt{\frac{K}{2T}}\sigma\left(\mathbf{e}_i - \mathbf{e}_j\right)$, where (i,j) are the indices of the least played pair of arms under some arbitrary algorithm $\mathcal{F}$, the following holds for any subgaussian instance $\mathcal{SG}(\sigma^2)$:*

$$\min_{\mathcal{F}} \max_{\phi \in \mathcal{SG}(\sigma^2)} \mathop{\mathbb{E}}_{\phi}\left[B_\Omega(f_T, f_\phi^*)\right] \geq \frac{\beta\Delta_{B,f}}{32e},$$

where β is the strong convexity parameter of regularizer Ω.

Proof. By β-strong convexity of Ω on Π, $B_\Omega(p,q) \geq \frac{\beta}{2}\|p-q\|^2$. Thus,

$$\begin{aligned}\max_{\phi \in \mathcal{SG}(\sigma^2)} \mathop{\mathbb{E}}_{\mu}\left[B_\Omega(f_T, f_\phi^*)\right] &\geq \frac{\mathbb{E}_\mu\left[B_\Omega(f_T, f_\mu^*)\right] + \mathbb{E}_{\mu+\epsilon}\left[B_\Omega(f_T, f_{\mu+\epsilon}^*)\right]}{2} \\ &\geq \frac{\beta}{4}\left(\mathop{\mathbb{E}}_{\mu}\left[\left\|f_T - f_\mu^*\right\|^2\right] + \mathop{\mathbb{E}}_{\mu+\epsilon}\left[\left\|f_T - f_{\mu+\epsilon}^*\right\|^2\right]\right).\end{aligned} \tag{17}$$

Let $\Delta_{B,f} = \left\|f_\mu^* - f_{\mu+\epsilon}^*\right\|^2 = \Delta_{\|\cdot\|,f}^2$. Starting from $\Delta_{B,f}$ and using the triangle inequality, followed by $(a+b)^2 \leq 2a^2 + 2b^2$, we find $\frac{\Delta_{B,\pi}}{2} \leq \left\|f_\mu^* - f_T\right\|^2 + \left\|f_{\mu+\epsilon}^* - f_T\right\|^2$. Consider the event $A := \left\{\left\|f_\mu^* - f_T\right\|^2 \leq g_T\right\}$, then we get

$$\begin{aligned}\mathop{\mathbb{E}}_{\mu}\left[\|f_\mu^* - f_T\|^2\right] + \mathop{\mathbb{E}}_{\mu+\epsilon}\left[\|f_{\mu+\epsilon}^* - f_T\|^2\right] &\geq g_T \mathop{\mathbb{P}}_{\mu}(A^c) + \left(\frac{\Delta_{B,f}}{2} - g_T\right)\mathop{\mathbb{P}}_{\mu+\epsilon}(A) \\ &\geq \frac{\Delta_{B,f}}{4}\left(\mathop{\mathbb{P}}_{\mu}(A^c) + \mathop{\mathbb{P}}_{\mu+\epsilon}(A)\right) \\ &\geq \frac{\Delta_{B,f}}{8}\exp\left(-D_{KL}\left(\mathop{\mathbb{P}}_{\mu} \| \mathop{\mathbb{P}}_{\mu+\epsilon}\right)\right) \\ &\geq \frac{\Delta_{B,f}}{8e}.\end{aligned} \tag{18}$$

The second line results from choosing $g_T = \frac{\Delta_{B,\pi}}{4}$, the third applies Bretagnolle-Huber, and the last line uses the KL decomposition lemma then the fact that $\epsilon = \sqrt{\frac{K}{2T}}\sigma\left(\mathbf{e}_i - \mathbf{e}_j\right)$. As a result,

$$\min_{\mathcal{F}} \max_{\phi \in \mathcal{SG}(\sigma^2)} \mathop{\mathbb{E}}_{\phi}\left[B_\Omega(f_T, f^*)\right] \geq \frac{\beta\Delta_{B,f}}{32e}. \tag{19}$$

□

A.1 Values Recommender Regret (First half of Theorem 1)

We now establish the proof for the first half of Theorem 1: Value recommender regret. For this regret, the divergence $D(\hat{V}_{\Omega,T}, V_\Omega^*) = |\hat{V}_{\Omega,T} - V_\Omega^*|$ is the absolute value of the difference between the recommendation and the target value.

Proof. For values, $f^* = V_\Omega^* = \Omega^*(Q_\Omega^*)$, $f_T = \hat{V}_{\Omega,T}$, and $\|\cdot\| = |\cdot|$. Therefore, $\Delta_{|\cdot|,V} := |V_{\Omega,\mu}^* - V_{\Omega,\mu+\epsilon}^*|$. Note that the values are dependent on the instance Q-values, as $V_{\Omega,\mu}^* = \Omega^*(Q_{\Omega,\mu}^\star) = \Omega^*(\mu)$

and similarly for the perturbed instance. By choosing $\epsilon = \sqrt{\frac{2}{T}}\sigma\mathbb{1}$, we have that

$$\begin{aligned} V^*_{\Omega,\mu+\epsilon} - V^*_{\Omega,\mu} &= \max_{\pi} \left[\langle \pi, \mu+\epsilon\rangle - \Omega(\pi)\right] - \max_{\pi} \left[\langle \pi, \mu\rangle - \Omega(\pi)\right] \\ &\geq \langle \pi^\star_\mu, \mu+\epsilon\rangle - \Omega(\pi^\star_\mu) - \langle \pi^\star_\mu, \mu\rangle + \Omega(\pi^\star_\mu) \\ &= \langle \pi^\star_\mu, \mu+\epsilon\rangle - \langle \pi^\star_\mu, \mu\rangle \\ &= \sum_{a=1}^{K} \pi^\star_\mu(a)\epsilon_a \\ &= \sqrt{\frac{2}{T}}\sigma. \end{aligned} \tag{20}$$

Leveraging Lemma 8, we obtain

$$\begin{aligned} \frac{\Delta_{|\cdot|,f}}{8e} &= \frac{|V^*_{\Omega,\mu+\epsilon} - V^*_{\Omega,\mu}|}{8e} \\ &\geq \frac{\sqrt{2}\sigma}{\sqrt{T}8e}. \end{aligned} \tag{21}$$

Therefore, we have the following lower bound on the values Recommender Regret:

$$\min_{\mathcal{F}} \max_{\phi\in\mathcal{SG}(\sigma^2)} \mathop{\mathbb{E}}_{\phi} \left[\left\|\hat{V}_{\Omega,T} - V^*_{\Omega,\phi}\right\|\right] \geq \frac{\sqrt{2}\sigma}{\sqrt{T}8e}. \tag{22}$$

□

A.2 Q-values Recommender Regret (Second half of Theorem 1)

For Q-values recommender regret we have $f^* = Q^*_\Omega$ and $f_T = \hat{Q}_{\Omega,T}$, $f^*_\mu = Q^\star_{\Omega,\mu} = \mu$, and $f^*_{\mu+\epsilon} = Q^\star_{\Omega,\mu+\epsilon} = \mu+\epsilon$. Then, $\Delta_{\|\cdot\|_*,Q} = \left\|Q^*_\mu - Q^*_{\mu+\epsilon}\right\|_* = \|\epsilon\|_*$.

Proof. Recall that (i,j) are the indices of least played pair of arms under some algorithm $\mathcal{F}$ and that such a pair is guaranteed to exist regardless of the algorithm. By choosing $\epsilon = \sqrt{\frac{K}{2T}}\sigma\left(\mathbf{e}_i - \mathbf{e}_j\right)$, and applying Lemma 8, we obtain

$$\begin{aligned} \frac{\Delta_{\|\cdot\|_*,Q}}{8e} &= \frac{\|\epsilon\|_*}{8e} \\ &= \frac{\sqrt{K}\sigma\left\|\mathbf{e}_i - \mathbf{e}_j\right\|_*}{\sqrt{2T}8e}. \end{aligned} \tag{23}$$

Thus, we have the following lower bound on the Q-values Recommender Regret:

$$\min_{\mathcal{F}} \max_{\phi\in\mathcal{SG}(\sigma^2)} \mathop{\mathbb{E}}_{\phi} \left[\left\|\hat{Q}_{\Omega,T} - Q^*_{\Omega,\phi}\right\|_*\right] \geq \sqrt{\frac{K}{2T}}\frac{\sigma \min_{a,a'\in[K],a\neq a'}\left\|\mathbf{e}_a - \mathbf{e}_{a'}\right\|_*}{8e}. \tag{24}$$

This is a consequence of the fact that any algorithm $\mathcal{F}$ has a least played pair of arms arm (i,j), and so the lower bound for the $\max$ over instances must hold for the algorithm minimizing the right hand side.

□

A.3 Norm-based policy recommender regret (First half of Theorem 3)

We now demonstate the lower bound for the norm-based policy recommender regret when Assumption 1 is respected. For policy recommender regret we have $f^* = \pi^\star_\Omega$ and $f_T = \hat{\pi}_{\Omega,T}$. We now provide a lemma lower bounding $\Delta_{\|\cdot\|,\pi}$.

Lemma 10. *Let (i, j) denote the indices of the least played pair of arms under some algorithm $\mathcal{F}$. When the object of interest is π, and the divergence for recommender regret is $\|\cdot\|$, then the following lower bound for $\Delta_{\|\cdot\|,\pi}$ with perturbation $\epsilon = \sqrt{\frac{K}{2T}}\sigma(\mathbf{e}_i - \mathbf{e}_j)$ holds for L-smooth, β-strongly convex regulariser Ω*

$$\Delta_{\|\cdot\|,\pi} \geq \sqrt{\frac{K}{2T}} \frac{\sigma \min_{a,a' \in [K], a \neq a'} \|\mathbf{e}_a - \mathbf{e}_{a'}\|_*}{L}.$$

Proof. We first observe that

$$\Delta_{\|\cdot\|,\pi} = \left\|\pi^\star_{\mu+\epsilon} - \pi^\star_\mu\right\| = \left\|\nabla\Omega^*(Q^*_{\Omega,\mu}) - \nabla\Omega^*(Q^*_{\Omega,\mu+\epsilon})\right\| = \left\|\nabla\Omega^*(\mu+\epsilon) - \nabla\Omega^*(\mu)\right\|.$$

From Proposition 2, Ω^* is $\frac{1}{L}$-strongly convex on the space of policies Π. Using $\epsilon = \sqrt{\frac{K}{2T}}\sigma(\mathbf{e}_i - \mathbf{e}_j)$, and recalling that (i, j) are the indices of the least played pair of arms under some algorithm $\mathcal{F}$, we obtain

$$\begin{aligned}
\langle \nabla\Omega^*(\mu) - \nabla\Omega^*(\mu+\epsilon), \epsilon\rangle &\geq \frac{1}{L}\|\epsilon\|^2_{*,E} && \text{(Proposition 2)}\\
\|\nabla\Omega^*(\mu) - \nabla\Omega^*(\mu+\epsilon)\|\,\|\epsilon\|_{*,E} &\geq \frac{1}{L}\|\epsilon\|^2_{*,E} && \text{(Dual norm inequality)}\\
\Delta_{\|\cdot\|,\pi} &\geq \frac{1}{L}\|\epsilon\|_{*,E}\\
&\geq \sqrt{\frac{K}{2T}}\frac{\sigma\|\mathbf{e}_i - \mathbf{e}_j\|_{*,E}}{L}\\
&\geq \sqrt{\frac{K}{2T}}\frac{\sigma \min_{a,a' \in [K]}\|\mathbf{e}_a - \mathbf{e}_{a'}\|_{*,E}}{L}.
\end{aligned} \tag{25}$$

□

Then, for the first half of Theorem 3 we have this simple proof.

Proof. Combining Lemmas 8 and 10, we get

$$\min_{\mathcal{F}} \max_{\phi \in \mathcal{SG}(\sigma^2)} \mathbb{E}_\phi\left[\left\|\hat{\pi}_{\Omega,T} - \pi^\star_\phi\right\|\right] \geq \sqrt{\frac{K}{2T}}\frac{\sigma \min_{a,a' \in [K]}\|\mathbf{e}_a - \mathbf{e}_{a'}\|_{*,E}}{8L}. \tag{26}$$

□

A.4 Bregman Divergence policy recommender regret (Second half of Theorem 3)

We now demonstate the lower bound for the Bregman divergence policy recommender regret when Assumption 1 is respected. For policy recommender regret we have $f^* = \pi^\star_\Omega$ and $f_T = \hat{\pi}_{\Omega,T}$.

Proof. Starting from Lemma 9, we must lower bound $\Delta_{B,\pi}$. However, notice that $\Delta_{B,\pi} = \Delta^2_{\|\cdot\|,\pi}$. Using Lemma 10 with $\epsilon = \sqrt{\frac{K}{2T}}\sigma(\mathbf{e}_i - \mathbf{e}_j)$, we obtain

$$\Delta_{B,\pi} = \Delta^2_{\|\cdot\|,\pi} \geq \frac{K}{2T}\frac{\sigma^2 \min_{a,a' \in [K], a \neq a'}\|\mathbf{e}_a - \mathbf{e}_{a'}\|^2_{*,E}}{L^2} \tag{27}$$

directly, as taking the square is monotonically increasing on the range $[0, \infty]$.

Combining with Lemma 9, we get our result

$$\min_{\mathcal{F}} \max_{\phi \in \mathcal{SG}(\sigma^2)} \mathbb{E}_\phi\left[B_\Omega(\hat{\pi}_{\Omega,T}, \pi^\star_\phi)\right] \geq \frac{K}{T}\frac{\beta\sigma^2 \min_{a,a' \in [K], a \neq a'}\|\mathbf{e}_a - \mathbf{e}_{a'}\|^2_*}{64L^2 e}. \tag{28}$$

□

B Proof of Proposition 2

Let us first recall the statement of the proposition:

Given a set $\mathcal{Q} \subset \mathbb{R}^{\mathcal{A}}$, such that $\Pi := \nabla\Omega^*(\mathcal{Q})$ is a subset of $\mathrm{ri}(\Delta_{\mathcal{A}})$,[6] if, for all $\pi \in \mathrm{Hull}(\Pi)$ and $v \in E$, we have $\langle \nabla^2_E \Omega(\pi)v, v\rangle \leq L\|v\|^2$, then for any $Q, Q' \in \mathcal{Q}$, we have $\langle \nabla\Omega^*(Q) - \nabla\Omega^*(Q'), Q - Q'\rangle \geq \frac{1}{L}\|Q - Q'\|^2_{*,E}$.

Proof. Let us consider $\pi = \nabla\Omega^*(Q)$ and $\pi' = \nabla\Omega^*(Q')$. By (Geist et al., 2019, Proposition 1) we have that $\pi = \arg\max_{\pi' \in \Delta_A} \langle \pi', Q\rangle - \Omega(\pi')$ and since π is in the relative interior of the simplex we have that the first order optimality condition becomes

$$\nabla_E \Omega(\pi) = PQ \tag{29}$$

where $P = I - \frac{1}{|\mathcal{A}|}\mathbf{1}\mathbf{1}^\top$ is the orthogonal projection onto E. Similarly, we have $\nabla_E\Omega(\pi') = PQ'$. Now let us consider

$$P(Q - Q') = \nabla_E\Omega(\pi) - \nabla_E\Omega(\pi') = \int_0^1 \nabla^2_E\Omega(\pi' + t(\pi - \pi'))(\pi - \pi')dt = Hd \tag{30}$$

where we called $H = \nabla_E\Omega(\pi) - \nabla_E\Omega(\pi') = \int_0^1 \nabla^2_E\Omega(\pi' + t(\pi - \pi'))dt$ and $d = \pi - \pi'$. Now, using Cauchy-Schwartz inequality for the inner product $\langle H\cdot, \cdot\rangle$ we get

$$\langle Hd, v\rangle^2 \leq \langle Hd, d\rangle\langle Hv, v\rangle \leq \langle Hd, d\rangle L\|v\|^2 \tag{31}$$

where for the last inequality we used for all $\pi \in \mathrm{Hull}(\Pi)$ and $v \in E$, we have $\langle \nabla^2_E\Omega(\pi)v, v\rangle \leq L\|v\|^2$. Taking the supremum over $v \in E, \|v\| = 1$ we get

$$\|Hd\|^2_{\star,E} \leq \langle Hd, d\rangle L \tag{32}$$

Now we can look at,

$$\begin{aligned}
\langle \nabla\Omega^*(Q) - \nabla\Omega^*(Q'), Q - Q'\rangle &= \langle \pi - \pi', Q - Q'\rangle && (33)\\
&= \langle P(\pi - \pi'), Q - Q'\rangle && (34)\\
&= \langle \pi - \pi', P(Q - Q')\rangle && (35)\\
&= \langle d, Hd\rangle && (36)\\
&\geq \frac{1}{L}\|Hd\|^2_{\star,E} && (37)\\
&= \frac{1}{L}\|Q - Q'\|^2_{\star,E} && (38)
\end{aligned}$$

which concludes the proof. □

C Supporting theoretical results

Lemma 11 (Q-value error bounds value and policy error)**.** *Given a regulariser $\Omega : \Delta_{\mathcal{A}} \to \mathbb{R}$ that is β-strongly convex with respect to a norm $\|\cdot\|$, and whose dual is the regularised RL value function,*

$$V_\Omega(s) = \Omega^*(Q(s, \cdot)) = \max_\pi \left\{ \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)} [Q(s, a)] - \frac{1}{\omega}\Omega(\pi(\cdot|s)) \right\}.$$

The policy and value errors can be bounded as follows:

$$\begin{aligned}
\|\hat\pi - \pi^\star\| &= \left\|\nabla\Omega^*(\hat Q) - \nabla\Omega^*(Q^\star)\right\| \leq \frac{\omega}{\beta}\left\|\hat Q - Q^\star\right\|_\star \\
D_\Omega(\hat\pi, \pi^\star) &\leq \frac{\omega^2}{2\beta}\left\|\hat Q - Q^\star\right\|^2_\star \\
\left|\hat V - V^\star\right| &= \left|\Omega^*(\hat Q) - \Omega^*(Q^\star)\right| \leq \left\|\hat Q - Q^\star\right\|_\infty,
\end{aligned} \tag{39}$$

where $(\hat\pi, \hat Q)$ and $(\pi^\star, Q^\star)$ are primal-dual pairs, with the Qs not necessarily being unique.

[6] For a convex set C, the notation $\mathrm{ri}(C)$ stands for its relative interior.

Proof of Lemma 11. The policy error bounds are both the result of dual smoothness. In short, dual smoothness guarantees that when the primal function Ω is closed and convex, the dual function, Ω^*, is smooth. From (Shalev-Shwartz, 2007, Lemma 15(3)) and (Kakade et al., 2012, Theorem 3), we have that $\frac{1}{\omega}\Omega$ is β/ω-convex, thus admitting a dual smoothness constant ω/β:

$$\left\|\nabla\Omega^*(\hat{Q}) - \nabla\Omega^*(Q^\star)\right\| \leq \frac{\omega}{\beta}\left\|\hat{Q} - Q^\star\right\|_\star.$$

Similarly, the Bregman divergence can be upper-bounded using the equivalency between the dual and primal Bregman divergence, and the $\frac{\omega}{\beta}$-smoothness property of Ω^* via (Bauschke et al., 1997, Theorem 3.7(5)) and (Bubeck, 2015, Lemma 3.4), respectively. Therefore, we have that

$$\begin{aligned}\frac{1}{\omega}D_\Omega(\hat{\pi}, \pi^\star) &= D_{\Omega^*}(Q^\star, \hat{Q}) \\ &= \Omega^*(Q^\star) - \Omega^*(\hat{Q}) - \langle\nabla\Omega^*(\hat{Q}), Q^\star - \hat{Q}\rangle \\ &\leq \frac{\omega}{2\beta}\left\|\hat{Q} - Q^\star\right\|_\star^2,\end{aligned}$$

where the first line states the primal-dual Bregman divergence equivalency, and the third line uses the smoothness upper bound. The bound is finished by multiplying through by ω. Note that we have to divide by ω in the first place to state the divergence in terms of the function $\frac{1}{\omega}\Omega$.

Finally, we know that the regularised value functions $V = \Omega^*(Q)$ are non-expansive mappings via (Nachum et al., 2017, Lemma 8, Corollary 9), whose results hold for any strongly convex regulariser, despite only being proven for negative entropy. This gives:

$$\left|\hat{V} - V^\star\right| = \left|\Omega^*(\hat{Q}) - \Omega^*(Q^\star)\right| \leq \left\|\hat{Q} - Q^\star\right\|_\infty.$$

□

Lemma 12 (Policies with non-zero probability mass achieve optimal rates on the Q-value regret)**.** *Given some sampling policy d_t such that $d_t(a) \geq \zeta > 0, \forall a \in \mathcal{A}, \forall t \in [1, T]$, where $\hat{Q}_t = \hat{r}_t$ (Equation 4), and where the reward distributions ϕ_a are σ-sub-Gaussian, and $\|\mathbf{e}'\|_\star = \max_a \|\mathbf{e}_a\|_\star$ where $\mathbf{e}_a$ is the canonical basis vector indexed by a, the following Q-value error bound holds for any $t \geq \frac{2\ln(2K/\delta)}{\zeta^2}$ with probability $1-\delta$ when $\|\cdot\|_\star = \|\cdot\|_\infty$*

$$\left\|\hat{Q}_T - Q^\star\right\|_\infty \leq \sqrt{\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T}}, \tag{40}$$

and otherwise for any $\|\cdot\|_\star$ with probability $1-\delta$

$$\left\|\hat{Q}_T - Q^\star\right\|_\star \leq K\|\mathbf{e}'\|_\star\sqrt{\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T}}. \tag{41}$$

Proof of Lemma 12. We begin by the bounding the action selection count using a high-probability proof technique via the Azuma-Hoeffding concentration inequality (Wainwright, 2019b). For arm a up to time t, let $N_t(a) = \sum_{i=1}^t \mathbb{I}(a_i = a)$ be its empirical tally of samples, and let $\hat{N}_t(a) = \sum_{i=1}^t d_i(a) \geq \zeta t$ be the sum of its empirical sampling probabilities. To bound the difference between these two terms, we have two options for modelling action selection: 1) as a Bernoulli distribution at each time step using the Hoeffding inequality, and 2) as a martingale difference sequence using the Azuma-Hoeffding inequality. Both reach the same result, but we opt for the latter.

To use Azuma-Hoeffding we must prove that $X_t = d_t(a|\mathcal{H}_t) - \mathbb{I}(a = a_t)$ is a martingale difference sequence. This is immediate from $|X_t| \leq 1$ and

$$\mathbb{E}\left[X_t|\mathcal{H}_t\right] = \mathbb{E}\left[d_t(a|\mathcal{H}_t) - \mathbb{I}(a = a_t)|\mathcal{H}_t\right] = d_t(a|\mathcal{H}_t) - d_t(a|\mathcal{H}_t) = 0.$$

Applying the one-sided Azuma-Hoeffding inequality (with $X_0 = 0$) implies that for a single arm a:

$$\mathbb{P}\left(\hat{N}_t(a) - N_t(a) > \epsilon_1\right) \leq \exp\left(-\frac{2\epsilon_1^2}{t}\right).$$

Setting $\epsilon_1 = \sqrt{\frac{t\ln(1/\delta_1)}{2}}$, this inequality implies that with probability $1-\delta_1$ (taking the complement event),

$$N_t(a) \geq \hat{N}_t(a) - \sqrt{\frac{t\ln(1/\delta_1)}{2}} \geq \zeta t - \sqrt{\frac{t\ln(1/\delta_1)}{2}}.$$

This implies that for any $t \geq T_0 = \frac{2\ln(1/\delta_1)}{\zeta^2}$, the action count will be bound by $N_t(a) \geq \frac{\zeta t}{2}$.

Defining the event $E_a = \left\{\hat{N}_t(a) - N_t(a) \leq \sqrt{\frac{t\ln(1/\delta_1)}{2}}\right\}$ to represent these action count probabilities, and the more general event over all actions $E = \bigcap_{a\in\mathcal{A}} E_a$, we have that $\mathbb{P}(E) \geq 1 - K\delta_1$.

We now begin deriving the Q-value error bound. Recall that $\hat{Q}_t(a) = \sum_{i=1}^t r_{i,a_i}\mathbb{I}(a_i = a)/N_t(a)$, and that $Q^\star(a) = \mu_a$. This leads to the following relation:

$$\begin{aligned}\left|\hat{Q}_t(a) - Q^\star(a)\right| &= \left|\sum_{i=1}^t \frac{r_{i,a_i}\mathbb{I}(a_i = a)}{N_t(a)} - \mu_a\right| \\ &= \frac{1}{N_t(a)}\left|\sum_{i=1}^t (r_{i,a_i} - \mu_a)\mathbb{I}(a_i = a)\right|\end{aligned}$$

which can be used in a Hoeffding inequality, by rearranging the equation. By setting the error $\epsilon_2 = \sqrt{\frac{2\sigma^2\ln(1/\delta_2)}{N_t(a)}}$, we have the following:

$$\begin{aligned}\mathbb{P}\left(|\hat{Q}_t(a) - Q^\star(a)| > \epsilon_2\right) &= \mathbb{P}\left(\left|\sum_{i=1}^t (r_{i,a_i} - \mu_a)\mathbb{I}(a_i = a)\right| > N_t(a)\epsilon_2\right) \\ &\leq 2\exp\left(-\frac{\epsilon_2^2 N_t(a)}{2\sigma^2}\right) = 2\delta_2.\end{aligned}$$

To bound the absolute error across all actions and time steps (which we must do in order to take into account all the entire history $\mathcal{H}_t$), we can define an event on the reward error $G_{a,i} = \left\{\left|\hat{Q}_i(a) - Q^\star(a)\right| \leq \sqrt{\frac{2\sigma^2\ln(1/\delta_2)}{N_i(a)}}\right\}$, and a general event $G = \bigcap_{a\in\mathcal{A}}\bigcap_{i\in[1,t]} G_{a,i}$. It is now clear that $\mathbb{P}(G) \geq 1 - 2Kt\delta_2$.

Finally, we can solve for δ_1 and δ_2, so that we can get a $1-\delta$ statement.

$$\mathbb{P}(E \cap G) \geq 1 - \mathbb{P}(E^c) - \mathbb{P}(G^c) \geq 1 - K\delta_1 - 2Kt\delta_2 = 1 - \delta, \tag{42}$$

setting $\delta_1 = \frac{\delta}{2K}$ and $\delta_2 = \frac{\delta}{4Kt}$. Thus under the intersection event, we have that for each $a \in \mathcal{A}$, for $t \geq T_0$:

$$|\hat{Q}_t(a) - Q^\star(a)| \leq \sqrt{\frac{2\sigma^2\ln(4Kt/\delta)}{N_t(a)}} \leq \sqrt{\frac{4\sigma^2\ln(4Kt/\delta)}{\zeta t}}. \tag{43}$$

This statement is enough to prove the special case where $\|\cdot\|_\star = \|\cdot\|_\infty$, since we don't need to convert the dual norm to a suitable form for bounding under the concentration argument, above.

Finally, we can map this back to our dual-norm regret, let $\mathbf{e}_a$ represent the standard basis vector for arm a. By the triangle inequality, any error vector can be decomposed via its coordinate-wise ℓ_∞ maximum:

$$\left\|\hat{Q}_t - Q^\star\right\|_\star = \left\|\sum_{a\in\mathcal{A}}(\hat{Q}_t(a) - Q^\star(a))\mathbf{e}_a\right\|_\star \leq \left\|\hat{Q}_t - Q^\star\right\|_\infty \sum_{a\in\mathcal{A}}\|\mathbf{e}_a\|_\star \leq K\|\mathbf{e}'\|_\star\left\|\hat{Q}_t - Q^\star\right\|_\infty$$

which proves the lemma after applying the Q-error inequality from Equation 43, and with $\|\mathbf{e}'\|_\star = \max_a \|\mathbf{e}_a\|_\star$. □

D Proof of Theorem 4

Theorem 5 (Sampling policies with non-zero probability mass achieve optimal regret bounds)**.** *Suppose Ω is β-strongly convex with dual $V_\Omega(s) = \Omega^*(Q(s,\cdot))$, with σ-sub-Gaussian reward distributions $\phi_a, \forall a \in \mathcal{A}$. For any $T \geq \frac{2\ln(2K/\delta)}{\zeta^2}$, with $\delta \in O(1/T^2)$, and any algorithm with sampling policy d_t satisfying $d_t(a) \geq \zeta > 0\ \forall a \in \mathcal{A}, \forall t \in [1,T]$, the following expected regret bounds hold:*

1. When $\|\cdot\|_\star = \|\cdot\|_\infty$ or when $f = V$, then

$$\mathbb{E}\left[\mathcal{SRR}_T^{f,D}\right] \in \tilde{O}\left(\sqrt{\frac{1}{\zeta T}}\right), \qquad \mathbb{E}\left[\mathcal{SRR}_T^{\pi,B_\Omega}\right] \in \tilde{O}\left(\frac{1}{\zeta T}\right).$$

2. When $\|\cdot\|_\star$ is a general dual norm with $\|\mathbf{e}'\|_\star = \max_a \|\mathbf{e}_a\|_\star$:

$$\mathbb{E}\left[\mathcal{SRR}_T^{f,D}\right] \in \tilde{O}\left(K\|\mathbf{e}'\|_\star\sqrt{\frac{1}{\zeta T}}\right), \qquad \mathbb{E}\left[\mathcal{SRR}_T^{\pi,B_\Omega}\right] \in \tilde{O}\left(K^2\|\mathbf{e}'\|_\star^2\frac{1}{\zeta T}\right),$$

where the terms on the left bound norm-based policy, value, and Q-value regrets, and the terms on the right bound Bregman divergence-based regrets. Where $\tilde{O}$ leaves out slow-growing logarithmic terms, as was done in Pedramfar & Ravanbakhsh (2026).

Proof. From Lemma 11 and Lemma 12, we have high probability upper bounds on instances of the simple recommender regret, as reflected in Equation 44 for general dual norms $\|\cdot\|_\star$.

$$\begin{aligned}
\mathcal{SRR}_T^{Q,\|\cdot\|_\star} &= \left\|\hat{Q}_{\Omega,T} - Q^\star\right\|_\star \leq K\|\mathbf{e}'\|_\star\sqrt{\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T}} \\
\mathcal{SRR}_T^{\pi,\|\cdot\|} &= \|\hat{\pi}_{\Omega,T} - \pi^\star\| \leq \frac{\omega}{\beta}K\|\mathbf{e}'\|_\star\sqrt{\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T}} \\
\mathcal{SRR}_T^{\pi,B_\Omega} &= B_\Omega(\hat{\pi}_{\Omega,T},\pi^\star) \leq \frac{\omega^2}{2\beta}K^2\|\mathbf{e}'\|_\star^2\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T} \\
\mathcal{SRR}_T^{V,|\cdot|} &= \left|\hat{V}_{\Omega,T} - V^\star\right| \leq \sqrt{\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T}}.
\end{aligned} \tag{44}$$

Note that when $\|\cdot\|_\star = \|\cdot\|_\infty$, each of the above bounds become tighter, as is evident from Lemma 12. Equation 45 lists the bounds for the L_∞ dual norm.

$$\begin{aligned}
\mathcal{SRR}_T^{Q,\|\cdot\|_\infty} &= \left\|\hat{Q}_{\Omega,T} - Q^\star\right\|_\infty \leq \sqrt{\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T}} \\
\mathcal{SRR}_T^{\pi,\|\cdot\|} &= \|\hat{\pi}_{\Omega,T} - \pi^\star\| \leq \frac{\omega}{\beta}\sqrt{\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T}} \\
\mathcal{SRR}_T^{\pi,B_\Omega} &= B_\Omega(\hat{\pi}_{\Omega,T},\pi^\star) \leq \frac{\omega^2}{2\beta}\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T} \\
\mathcal{SRR}_T^{V,|\cdot|} &= \left|\hat{V}_{\Omega,T} - V^\star\right| \leq \sqrt{\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T}}.
\end{aligned} \tag{45}$$

The principal difference between the regret bounds of Equation 44 and Equation 45 are the constants in front of the Q-error concentration bound. We will henceforth use the constant C, to denote the constant, which can depend on K, as is evident, above.

These bounds are conditioned on the fact that, with probability $1-\delta$, the event $E \cap G$ holds, see Equation 42. We re-write the events here, for clarity:

$$G = \bigcap_{a\in\mathcal{A}}\bigcap_{i\in[1,T]}\left\{\left|\hat{Q}_i(a) - Q^\star(a)\right| \leq \sqrt{\frac{2\sigma^2\ln(4KT/\delta)}{N_i(a)}}\right\}$$

$$E = \bigcap_{a\in\mathcal{A}}\left\{\hat{N}_T(a) - N_T(a) \leq \sqrt{\frac{T\ln(2K/\delta)}{2}}\right\}.$$

Then, for the non-Bregman divergences, we have that for some $T \geq T_0$, there exists a constant C that depends on the type of recommender regret and the conversion between the dual norm to the L_∞ norm, such that

$$\begin{aligned}\mathbb{E}\left[\mathcal{SRR}_T^{f,D}\right] &= \mathbb{E}\left[\mathcal{SRR}_T^{f,D}|E\cap G\right]\mathbb{P}\left(E\cap G\right) + \mathbb{E}\left[\mathcal{SRR}_T^{f,D}|E^c\cup G^c\right]\mathbb{P}\left(E^c\cup G^c\right)\\ &\leq C\sqrt{\frac{4\sigma^2\ln(4KT/\delta)}{\zeta T}} + \delta C\,\mathbb{E}\left[\left\|\hat{Q}_T - Q^\star\right\|_\infty \middle| E^c\cup G^c\right].\end{aligned}$$

The term on the right is irreducible under the complement event, since we have no relations on $N_T(a)$, nor $\left\|\hat{Q}_T - Q^\star\right\|_\infty$ that provide bounds on the term in the expectation. Thus, we use the following properties of independent sub-Gaussian random variables:

- We know that random variables $Y_{i,a} \sim \text{subG}(\sigma^2), i \in [1, N_T(a)]$ are sub-Guassian by definition if $\mathbb{E}\left[Y_{i,a}\right] = 0$

  $$\mathbb{E}\left[\exp\left(\lambda Y_{i,a}\right)\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right); \forall\lambda\in\mathbb{R}.$$

  We may then trivially achieve zero-mean sub-Gaussian random variables on our reward distributions by setting $Y_{i,a} = r_{i,a} - \mu_a$. (Rigollet, 2015, Definition 1.2)
- We also know that sums of sub-Gaussian random variables are sub-Gaussian. The following is a specific case of (Rigollet, 2015, Theorem 1.6). For a random vector of sub-Gaussian random variables $Y_a = (Y_{i,a_i} : i \in [1,T], a_i = a)$ and a vector of constant factors $z_a = \frac{1}{N_T(a)}\mathbb{1}$, such that $z_a^\intercal Y_a = \frac{1}{N_T(a)}\sum_{i=1}^{N_T(a)} Y_{i,a}$:

  $$\mathbb{E}\left[\exp\left(\lambda z_a^\intercal Y_a\right)\right] = \prod_{i=1}^{N_T(a)}\mathbb{E}\left[\exp\left(\frac{\lambda Y_{i,a}}{N_T(a)}\right)\right] \leq \prod_{i=1}^{N_T(a)}\exp\left(\frac{\lambda^2\sigma^2}{2N_T^2(a)}\right) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right),$$

  where the last inequality follows by $1/N_T(a) \leq 1$.
- Finally, following (Rigollet, 2015, Theorem 1.14), define $\hat{Y}_T = (z_a^\intercal Y_a : a \in \mathcal{A})$, which, by the above, is a sub-Gaussian random vector with variance proxy σ^2. We have that

  $$\mathbb{E}\left[\max_{a\in\mathcal{A}}|\hat{Y}_{T,a}|\right] \leq \sigma\sqrt{2\log(2K)}; \quad \mathbb{P}\left(\max_{a\in\mathcal{A}}|\hat{Y}_{T,a}| > j\right) \leq 2K\exp\left(-\frac{j^2}{2\sigma^2}\right)$$

We can now continue by defining $\hat{Y}_{T,a} = \hat{Q}_T(a) - Q^\star(a) = \frac{1}{N_T(a)}\sum_{i=1}^{T}\mathbb{I}(a_i = a)\left(r_{i,a_i} - \mu_a\right)$. Note that we will have to evaluate moments of $\max_{a\in\mathcal{A}}|\hat{Y}_{T,a}|$, so the following will prove useful.

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{Q}_T - Q^\star\right\|_\infty^{2p}\right] &= \mathbb{E}\left[\max_{a\in\mathcal{A}}\left|\hat{Y}_{T,a}\right|^{2p}\right]\\ &= \int_0^\infty \mathbb{P}\left(\max_{a\in\mathcal{A}}\left|\hat{Y}_{T,a}\right|^{2p} > j\right)dj\\ &= \int_0^\infty \mathbb{P}\left(\max_{a\in\mathcal{A}}\left|\hat{Y}_{T,a}\right| > j^{1/2p}\right)dj\\ &= \int_0^u \mathbb{P}\left(\max_{a\in\mathcal{A}}\left|\hat{Y}_{T,a}\right| > j^{1/2p}\right)dj + \int_u^\infty \mathbb{P}\left(\max_{a\in\mathcal{A}}\left|\hat{Y}_{T,a}\right| > j^{1/2p}\right)dj\\ &\leq \int_0^u 1dt + 2K\int_u^\infty \exp\left(-\frac{j^{1/p}}{2\sigma^2}\right)dj\\ &= u + 2K\int_u^\infty \exp\left(-\frac{j^{1/p}}{2\sigma^2}\right)dj.\end{aligned}$$

We choose u to be a convenient constant: $u = (2\sigma^2\ln(2K))^p$. Then via a change of variables, we will evaluate the remaining integral term: $h = j^{1/p}$ such that $dj = ph^{p-1}dh$, with the lower integration bound becoming $u^{1/p} = 2\sigma^2\ln(2K)$.

$$\int_u^\infty \exp\left(-\frac{j^{1/p}}{2\sigma^2}\right)dj = \int_{2\sigma^2\ln(2K)}^\infty \exp\left(-\frac{h}{2\sigma^2}\right)ph^{p-1}dh.$$

Once again changing the integration bounds to recover a gamma-like function, let $x = \frac{h}{2\sigma^2}$, and $dh = 2\sigma^2 dx$, with the lower integration bound now being $\ln(2K)$:

$$\begin{aligned}\int_{2\sigma^2 \ln(2K)}^{\infty} \exp\left(-\frac{h}{2\sigma^2}\right) p h^{p-1} dh &= \int_{\ln(2K)}^{\infty} \exp(-x)\, p(2\sigma^2 x)^{p-1} 2\sigma^2 dx \\ &= p(2\sigma^2)^p \int_{\ln(2K)}^{\infty} x^{p-1} \exp(-x)\, dx \\ &\leq p(2\sigma^2)^p e^{-\ln(2K)} \sum_{k=0}^{p-1} \frac{(p-1)!}{k!} \ln(2K)^k \\ &= (2\sigma^2)^p \frac{1}{2K} \sum_{k=0}^{p-1} \frac{p!}{k!} \ln(2K)^k,\end{aligned}$$

which is a similar result to what is shown in (Wainwright, 2019a, Theorem 2.6).

We only care about powers of 2 and 4, since these will be used later, so taking $p = 1, 2$, we find the following bounds:

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{Q}_T - Q^\star\right\|_\infty^2\right] &\leq 2\sigma^2 \left(\ln(2K) + 1\right) \\ \mathbb{E}\left[\left\|\hat{Q}_T - Q^\star\right\|_\infty^4\right] &\leq 4\sigma^4 \left(\ln(2K)^2 + 2\ln(2K) + 2\right).\end{aligned}$$

Back to our difficult-to-reduce expectation, the following holds, using Cauchy-Schwartz:

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{Q}_T - Q^\star\right\|_\infty \middle| E^c \cup G^c\right] \mathbb{P}\left(E^c \cup G^c\right) &= \mathbb{E}\left[\left\|\hat{Q}_T - Q^\star\right\|_\infty \mathbb{I}(E^c \cup G^c)\right] \\ &\leq \sqrt{\mathbb{E}\left[\left\|\hat{Q}_T - Q^\star\right\|_\infty^2\right]} \sqrt{\mathbb{E}\left[\mathbb{I}(E^c \cup G^c)\right]} \\ &\leq \sigma\sqrt{2\delta(\log(2K) + 1)}.\end{aligned}$$

Thus, we have that

$$\begin{aligned}\mathbb{E}\left[\mathcal{SRR}_T^{f,D}\right] &= \mathbb{E}\left[\mathcal{SRR}_T^{f,D} | E \cap G\right] \mathbb{P}\left(E \cap G\right) + \mathbb{E}\left[\mathcal{SRR}_T^{f,D} | E^c \cup G^c\right] \mathbb{P}\left(E^c \cup G^c\right) \\ &\leq C\sqrt{\frac{4\sigma^2 \ln(4KT/\delta)}{\zeta T}} + \sigma C\sqrt{2\delta(\log(2K) + 1)}.\end{aligned}$$

For Bregman divergences, we follow the same proof technique, just with squared Q-value error, leading to a similar result.

$$\begin{aligned}\mathbb{E}\left[\mathcal{SRR}_T^{\pi,B_\Omega}\right] &= \mathbb{E}\left[\mathcal{SRR}_T^{\pi,B_\Omega} | E \cap G\right] \mathbb{P}\left(E \cap G\right) + \mathbb{E}\left[\mathcal{SRR}_T^{\pi,B_\Omega} | E^c \cup G^c\right] \mathbb{P}\left(E^c \cup G^c\right) \\ &\leq C\frac{4\sigma^2 \ln(4KT/\delta)}{\zeta T} + 2\sigma^2 C\sqrt{\delta\left(\ln(2K)^2 + 2\ln(2K) + 2\right)}.\end{aligned}$$

Allowing us to conclude that, so long as $\delta \in O(1/T^2)$, $\mathbb{E}\left[\mathcal{SRR}_T^{f,D}\right] \in \tilde{O}\left(C\sqrt{\frac{1}{\zeta T}}\right)$ for the non-Bregman divergences, and $\mathbb{E}\left[\mathcal{SRR}_T^{\pi,B_\Omega}\right] \in \tilde{O}\left(\frac{C}{\zeta T}\right)$ for the Bregman divergence-based regret. By choosing the dual norm, we can finally achieve the statements in the theorem, by cross-referencing Equation 44 and Equation 45.

Note that we chose the above proof technique to avoid bounding the reward distributions. If instead we bounded each $|r_{i,a_i}| \leq R_{\max} < \infty$, we would quite simply have that

$$\begin{aligned}\mathbb{E}\left[\mathcal{SRR}_T^{f,D} | E^c \cup G^c\right] \mathbb{P}\left(E^c \cup G^c\right) &\leq \delta C\, \mathbb{E}\left[\left\|\hat{Q}_T - Q^\star\right\|_\infty \middle| E^c \cup G^c\right] \\ &\leq 2CR_{\max}\delta.\end{aligned}$$

Then with $\delta \in O(1/T)$ we have both bounds presented above. □

E Proof of Theorem 5

Proof of Theorem 5. Using Lemma 11 with the fact that $\left\|\hat{Q}_{\Omega,t} - Q^\star\right\|_\star \leq C_\mathcal{A} \left\|\hat{Q}_{\Omega,t} - Q^\star\right\|_\infty$ for some constant $C_\mathcal{A}$ that depends on the dual norm and action space, we know that there exists a constant C_3 such that

$$\mathcal{SRR}_t^{f,D} \leq \begin{cases} C_3^2 \left\|\hat{Q}_{\Omega,t}(s_0,\cdot) - Q^\star(s_0,\cdot)\right\|_\infty^2 & \text{if } D \in \{B_\Omega\} \\ C_3 \left\|\hat{Q}_{\Omega,t}(s_0,\cdot) - Q^\star(s_0,\cdot)\right\|_\infty & \text{else.} \end{cases}$$

Without loss of generality, we proceed with the linear bound, noting that for Bregman divergences, we can afford a higher $\mathcal{SRR}$ lower bound in the probability statement: ϵ^2, which yields the same exponential concentration rate, after square rooting both sides of the inequality in the probability.

Our goal is now to analyse the following term at various tree depths H, for an inductive proof:

$$\mathbb{P}\left(\mathcal{SRR}_t^{f,D} \geq \epsilon\right) \leq \mathbb{P}\left(\left\|\hat{Q}_{\Omega,t}(s_0,\cdot) - Q^\star(s_0,\cdot)\right\|_\infty \geq \frac{\epsilon}{C_3}\right).$$

Namely, we seek to prove that $\exists C_1, C_2 \in \mathbb{R}_{>0}$ at the root s_0 of a search tree such that

$$\mathbb{P}\left(\left\|\hat{Q}_{\Omega,t}(s_0,\cdot) - Q^\star(s_0,\cdot)\right\|_\infty \geq \frac{\epsilon}{C_3}\right) \leq C_1 \exp\left(-\frac{N_t(s_0)\epsilon^2}{C_2}\right).$$

Let $N_t(s,a) = \sum_{i=1}^t \mathbb{I}(a_i = a)\mathbb{I}(s_i = s)$ be the empirical tally of samples, $N_t(s_0) := \sum_{a\in\mathcal{A}} N_t(s_0,a) = t$, and let $\hat{N}_t(s,a) = \sum_{i=1}^t d_i(a|s) \geq \zeta t$ be the sum of sampling policy d_i for a given state-action pair.

Base case, with $H = 1$ In the bandit setting, we can simply use Hoeffding's inequality to achieve the desired concentration bound. We will reuse elements of the proof technique from Lemma 12, specifically the concentration of action selection counts via the Azuma-Hoeffding inequality on martingale difference sequences (Wainwright, 2019b). Define the event $E_a = \left\{\hat{N}_t(s_0,a) - N_t(s_0,a) \leq \epsilon_1\right\}$ and the more general event over all actions $E = \bigcap_{a\in\mathcal{A}} E_a$. When conditioning on E_a, it will be useful to set $\epsilon_1 = \zeta N_t(s_0)/2$, so that we have $N_t(s_0,a) \geq \zeta N_t(s_0)/2$ (since $\hat{N}_t(s_0,a) \geq \zeta N_t(s_0)$). It is clear that

$$\begin{aligned} \mathbb{P}(E_a^c) &\leq \exp\left(-\frac{2\epsilon_1^2}{N_t(s_0)}\right) \\ &= \exp\left(-\frac{2}{N_t(s_0)}\frac{\zeta^2(N_t(s_0))^2}{4}\right) \\ &= \exp\left(-\frac{\zeta^2 N_t(s_0)}{2}\right). \end{aligned}$$

Define another event on the reward error $G_{a,t} = \left\{\left|\hat{Q}_{\Omega,t}(s_0,a) - Q^\star_\Omega(s_0,a)\right| \geq \epsilon_2\right\}$, and a general event $G_t = \bigcup_{a\in\mathcal{A}} G_{a,t}$. Notice how $\mathbb{P}\left(\left\|\hat{Q}_{\Omega,t}(s_0,\cdot) - Q^\star(s_0,\cdot)\right\|_\infty \geq \epsilon_2\right) \leq \mathbb{P}(G_t)$. Using Hoeffding's inequality (under the assumption that $N_t(s_0,a)$ is not a random variable), it is clear that

$$\begin{aligned} \mathbb{P}(G_{a,t}) &= \mathbb{P}\left(\left|\sum_{i=1}^t (r_{i,a_i} - Q^\star(s_0,a))\mathbb{I}(a_i = a)\right| > N_t(s_0,a)\epsilon_2\right) \\ &\leq 2\exp\left(-\frac{\epsilon_2^2 N_t(s_0,a)}{2\sigma^2}\right). \end{aligned}$$

To finally bound $\mathbb{P}(G_t)$, we have to use the event E_a to convert between child and parent action counts, and we have to convert those action counts from random variables to deterministic sample counts. Since we know that $N_t(s_0, a) \in [\zeta N_t(s_0)/2, N_t(s_0)]$ under the event E_a, we can sum over those action counts (through the union bound) to make the count deterministic:

$$\begin{aligned}
\mathbb{P}(G_{a,t}|E_a) &= \mathbb{P}\left(\left|\hat{Q}_{\Omega,t}(s_0,a) - Q^\star_\Omega(s_0,a)\right| \geq \epsilon_2 \middle| E_a\right) \\
&= \mathbb{P}\left(\exists n = N_t(s_0,a) \in \left[\frac{\zeta N_t(s_0)}{2}, N_t(s_0)\right] : \left|\hat{Q}_{\Omega,t}(s_0,a) - Q^\star_\Omega(s_0,a)\right| \geq \epsilon_2 \middle| E_a\right) \\
&\leq \mathbb{P}\left(\bigcup_{n=\zeta N_t(s_0)/2}^{N_t(s_0)} \left\{n = N_t(s_0,a) : \left|\hat{Q}_{\Omega,t}(s_0,a) - Q^\star_\Omega(s_0,a)\right| \geq \epsilon_2\right\} \middle| E_a\right) \\
&\leq \sum_{n=\zeta N_t(s_0)/2}^{N_t(s_0)} \mathbb{P}\left(\left\{n = N_t(s_0,a) : \left|\hat{Q}_{\Omega,t}(s_0,a) - Q^\star_\Omega(s_0,a)\right| \geq \epsilon_2\right\}\right) \\
&\leq 2 \sum_{n=\zeta N_t(s_0)/2}^{N_t(s_0)} \exp\left(-\frac{\epsilon_2^2 n}{2\sigma^2}\right) \\
&\leq 2N_t(s_0) \exp\left(-\frac{\epsilon_2^2 \zeta N_t(s_0)}{4\sigma^2}\right).
\end{aligned}$$

Completing the bound on $\mathbb{P}(G_t)$:

$$\begin{aligned}
\mathbb{P}(G_t) &= \mathbb{P}(G_t|E)\,\mathbb{P}(E) + \mathbb{P}(G_t|E^c)\,\mathbb{P}(E^c) \\
&\leq \mathbb{P}(G_t|E) + \mathbb{P}(E^c) \\
&= \sum_{a\in\mathcal{A}} \mathbb{P}(G_{a,t}|E_a) + \mathbb{P}(E_a^c) \\
&\leq 2KN_t(s_0) \exp\left(-\frac{\epsilon_2^2 \zeta N_t(s_0)}{4\sigma^2}\right) + K \exp\left(-\frac{\zeta^2 N_t(s_0)}{2}\right) \\
&\leq 2KN_t(s_0) \exp\left(-\frac{\epsilon_2^2 \zeta N_t(s_0)}{4\sigma^2}\right) + KN_t(s_0) \exp\left(-\frac{\zeta^2 N_t(s_0)}{2}\right).
\end{aligned}$$

From here, we need to draw some relations between ζ and ϵ_2 in order to get the bound into the correct form. We set $\epsilon_2 = \epsilon/C_3 \leq \zeta$ thus setting the assumption in the theorem statement. This invariably affects the size that T must take in order to satisfy the upper bound (i.e. T must be very large since ϵ_2 is smaller than ζ). Thus, using the facts that $-1 \leq -\zeta$ and that $-\zeta \leq -\epsilon_2$, we have the following bound

$$\begin{aligned}
\mathbb{P}(G_t) &\leq 2KN_t(s_0) \exp\left(-\frac{\epsilon_2^2 \zeta N_t(s_0)}{4\sigma^2}\right) + KN_t(s_0) \exp\left(-\frac{\zeta^2 N_t(s_0)}{2}\right) \\
&\leq 2KN_t(s_0) \exp\left(-\frac{\epsilon_2^2 \zeta N_t(s_0)}{4\sigma^2}\right) + KN_t(s_0) \exp\left(-\frac{\epsilon_2^2 \zeta N_t(s_0)}{2}\right) \\
&\leq 3KN_t(s_0) \exp\left(-\frac{\epsilon_2^2 \zeta N_t(s_0)}{\max(4\sigma^2, 2)}\right) \\
&\leq 3KN_t(s_0) \exp\left(-\frac{\epsilon_2^2 \zeta N_t(s_0)}{2(2\sigma^2+1)}\right),
\end{aligned}$$

where the last inequality come from the fact that $a + b \geq \max(a, b), \forall a, b \in \mathbb{R}_{\geq 0}$.

Finally, we use the relation between G_t and the L_∞ Q-value error to show the base case:

$$\begin{aligned}\mathbb{P}\left(\left\|\hat{Q}_{\Omega,t}(s_0,\cdot)-Q^\star(s_0,\cdot)\right\|_\infty \geq \frac{\epsilon}{C_3}\right) &\leq \mathbb{P}\left(G_t\right), \quad (\epsilon_2 = \epsilon/C_3)\\ &\leq 3KN_t(s_0)\exp\left(-\frac{\epsilon^2\zeta N_t(s_0)}{2C_3^2(2\sigma^2+1)}\right)\\ &= Z_1^{(1)}\exp\left(-\frac{\epsilon^2 N_t(s_0)}{Z_2^{(1)}C_3^2}\right),\end{aligned}$$

with $Z_1^{(1)} = 3KN_t(s_0)$ and $Z_2^{(1)} = \frac{\zeta}{2(2\sigma^2+1)}$.

Inductive step Assuming that the inequality holds for each depth $H-1$ tree, we will prove that it also holds for a search tree of depth H. We will reuse the events E_a and E, since they were applied at the root of our bandit, but equally apply to the root of a search tree. Let both Q-value estimators use the same deterministic reward, where s_a is the next state from state s_0 after a is taken: $\hat{Q}_{\Omega,t}(s_0,a) = r(s_0,a) + \gamma\Omega^*(\hat{Q}_{\Omega,t}(s_a,\cdot))$ and $Q^\star_{\Omega,t}(s_0,a) = r(s_0,a) + \gamma\Omega^*(Q^\star_{\Omega,t}(s_a,\cdot))$. Therefore, we have that

$$\begin{aligned}\mathbb{P}\left(\left\|\hat{Q}_{\Omega,t}(s_0,\cdot)-Q^\star(s_0,\cdot)\right\|_\infty \geq \frac{\epsilon}{C_3}\right) &\leq \mathbb{P}\left(\sum_{a\in\mathcal{A}}\left|\hat{Q}_{\Omega,t}(s_0,a)-Q^\star(s_0,a)\right| \geq \frac{\epsilon}{C_3}\right)\\ &= \mathbb{P}\left(\sum_{a\in\mathcal{A}}\gamma\left|\Omega^*(\hat{Q}_{\Omega,t}(s_a,\cdot)))-\Omega^*(Q^\star(s_a,\cdot))\right| \geq \frac{\epsilon}{C_3}\right)\\ &\leq \mathbb{P}\left(\sum_{a\in\mathcal{A}}\left\|\hat{Q}_{\Omega,t}(s_a,\cdot)-Q^\star(s_a,\cdot)\right\|_\infty \geq \frac{\epsilon}{C_3\gamma}\right)\\ &\leq \mathbb{P}\left(\bigcup_{a\in\mathcal{A}}\left\{\left\|\hat{Q}_{\Omega,t}(s_a,\cdot)-Q^\star(s_a,\cdot)\right\|_\infty \geq \frac{\epsilon}{KC_3\gamma}\right\}\right)\\ &\leq \sum_{a\in\mathcal{A}}\mathbb{P}\left(\left\|\hat{Q}_{\Omega,t}(s_a,\cdot)-Q^\star(s_a,\cdot)\right\|_\infty \geq \frac{\epsilon}{C_3K\gamma}\right),\end{aligned}$$

where the second line follows from the Q-value definitions, the third follows from Lemma 11, the fourth follows from the union bound (setting the resulting constant on the bound to $1/K$ so that the constants sum to 1), and the final inequality follows from the union bound.

Now, let $G'_{s_a} = \left\{\left\|\hat{Q}_{\Omega,t}(s_a,\cdot)-Q^\star(s_a,\cdot)\right\|_\infty \geq \frac{\epsilon}{C_3K\gamma}\right\}$. We follow the same proof technique from the base case, conditioning on E and solving for the upper bound:

$$\begin{aligned}\mathbb{P}\left(G'_{s_a}\right) &= \mathbb{P}\left(G'_{s_a}|E\right)\mathbb{P}\left(E\right) + \mathbb{P}\left(G'_{s_a}|E^c\right)\mathbb{P}\left(E^c\right)\\ &\leq \mathbb{P}\left(G'_{s_a}|E\right) + \mathbb{P}\left(E^c\right)\\ &\leq Z_1^{(H-1)}N_t(s_0)\exp\left(-\frac{\zeta N_t(s_0)\epsilon^2}{2Z_2^{(H-1)}C_3^2K^2\gamma^2}\right) + K\exp\left(-\frac{\zeta^2 N_t(s_0)}{2}\right)\\ &\leq Z_1^{(H-1)}N_t(s_0)\exp\left(-\frac{\zeta N_t(s_0)\epsilon^2}{2Z_2^{(H-1)}C_3^2K^2\gamma^2}\right) + KN_t(s_0)\exp\left(-\frac{\zeta N_t(s_0)\epsilon^2}{2}\right) \quad \text{(using } \zeta^2 \geq \epsilon^2\text{)}\\ &\leq (Z_1^{(H-1)}+K)N_t(s_0)\exp\left(-\frac{\zeta N_t(s_0)\epsilon^2}{2(Z_2^{(H-1)}C_3^2K^2\gamma^2+1)}\right).\end{aligned}$$

Finally, we have that the root Q-value error concentrates given that the sub-trees concentrate:

$$\begin{aligned}\mathbb{P}\left(\left\|\hat{Q}_{\Omega,t}(s_0,\cdot)-Q^\star(s_0,\cdot)\right\|_\infty \geq \frac{\epsilon}{C_3}\right) &\leq \sum_{a\in\mathcal{A}}\mathbb{P}\left(G'_{s_a}\right)\\ &\leq K(Z_1^{(H-1)}+K)N_t(s_0)\exp\left(-\frac{\zeta N_t(s_0)\epsilon^2}{2(Z_2^{(H-1)}C_3^2K^2\gamma^2+1)}\right)\\ &\leq Z_1^{(H)}\exp\left(-\frac{N_t(s_0)\epsilon^2}{Z_2^{(H)}}\right),\end{aligned}$$

for $Z_1^{(H)} = K(Z_1^{(H-1)}+K)N_t(s_0)$ and $Z_2^{(H)} = \frac{2(Z_2^{(H-1)}C_3^2K^2\gamma^2+1)}{\zeta}$. Analysing these relations for arbitrary tree depths H, we find the following relation, which will be used to construct our final bounds, later:

- **For $\mathbf{Z_1^{(H)}}$**:

$$\begin{aligned}Z_1^{(H)} &= K(Z_1^{(H-1)}+K)t \propto KtZ_1^{(H-1)}\\ &= K(K(Z_1^{(H-2)}+K)t+K)t \propto K^2t^2Z_1^{(H-2)}\\ &\vdots\\ &\propto (Kt)^H\end{aligned}$$

- **For $\mathbf{Z_2^{(H)}}$**:

$$\begin{aligned}Z_2^{(H)} &= \frac{2(Z_2^{(H-1)}C_3^2K^2\gamma^2+1)}{\zeta} \propto \frac{K^2}{\zeta}Z_2^{(H-1)}\\ &= \frac{2(\frac{2(Z_2^{(H-2)}C_3^2K^2\gamma^2+1)}{\zeta}C_3^2K^2\gamma^2+1)}{\zeta} \propto \frac{K^4}{\zeta^2}Z_2^{(H-2)}\\ &\vdots\\ &\propto \frac{K^{2H}}{\zeta^H}\end{aligned}$$

Conclusion The L_∞ Q-value error concentrates at the root of a search tree. Since this is the case, we also have that the recommender regret concentrates at the root of a search tree:

$$\begin{aligned}\mathbb{P}\left(\mathcal{SR}_t^{f,D}\geq\epsilon\right) &\leq \mathbb{P}\left(\left\|\hat{Q}_{\Omega,t}(s_0,\cdot)-Q^\star(s_0,\cdot)\right\|_\infty \geq \frac{\epsilon}{C_3}\right)\\ &\leq C_1(Kt)^H\exp\left(-\frac{\zeta^H N_t(s_0)\epsilon^2}{C_2K^{2H}}\right).\end{aligned}$$

For some $C_1, C_2 \in \mathbb{R}_{>0}$ that depend on $K, H, \gamma, \zeta, \sigma,$ and C_3. Specifically, $\exists C_1 : Z_1^{(H)} \leq C_1(Kt)^H$ and $\exists C_2 : Z_2^{(H)} \leq C_2\frac{K^{2H}}{\zeta^H}$. These exponential relations indicate that the tree must be well-explored prior to concentration occurring, and that ζ should be maximised to minimise the upper bound. □

Remark 1 (Comparison with prior regularised MCTS bounds)**.** *Prior regularised MCTS proofs produced exponential rates of order $O(T\exp(-T/\log^3 T))$, where depth dependence of $O(K^H)$ was hidden in constants. However, union bounding over execution histories $\sum_{i=1}^{T}\mathbb{P}(E_{r_i}^c)$ (c.f., (Dam et al., 2021, Lemma 3, 4)) incorporates non-decaying early-stage estimation errors ($\mathbb{P}\left(E_{r_1}^c\right) = \Omega(1)$, making history-based concentration bounds vacuous. By explicitly conditioning on valid visit allocations at each depth layer $h \in [1, H]$, our proof guarantees a rigorous concentration bound of $O(T^H\exp(-CT))$.*

An explanation of Remark 1: the issue with prior proofs. (Dam et al., 2021, Lemma 3) states that

$$\mathbb{P}\left(|\hat{N}_t(s,a) - N_t^*(s,a)| > \frac{Ct}{\log t}\right) \leq \hat{C}|A|t\exp\left\{-\frac{t}{(\log t)^3}\right\}.$$

To prove this statement, the following event is defined which concentrates according to the bound that follows.

$$E_{r_k} = \left\{\|r(s',\cdot) - \hat{r}_k(s',\cdot)\|_\infty < \frac{2\sigma}{\log(2+k)}\right\}$$

$$\mathbb{P}\left(\|r - \hat{r}_t\|_\infty \geq \frac{2\sigma}{\log(2+t)}\right) \leq 4|A|\exp\left(-\frac{t}{(\log(2+t))^3}\right).$$

After some analysis, a relation is arrived at which serves as the basis for the lemma statement:

$$\mathbb{P}\left(|\hat{N}_t(s,a) - N_t^*(s,a)| \geq \frac{Ct}{\log t}\right) \leq \sum_{k=1}^{t}\mathbb{P}(E_{r_k}^c).$$

It is clear that at the t-th step, we have concentration with regards to t. However, at some arbitrary k-th step, $k \in [1,t]$, all we have is an irreducible constant term:

$$\mathbb{P}\left(E_{r_k}^c\right) \leq 4|A|\exp\left(-\frac{k}{(\log(2+k))^3}\right).$$

Since there is no relation between k and t that allows us to convert the bound to one on t (nor should there be, since at some time step k we had a fixed error probability), we must conclude that the probability bound achieved is actually $\Omega(1)$, and does not prove concentration. This affects every theorem down the line, by making it impossible to prove tree convergence at any step, since the lemma is used frequently to convert child action counts to parent action counts. Note that this issue is inherited from (Xiao et al., 2019, Theorem 3), which uses the same logic to draw relations between child and parent action counts. □

F Deferred proofs

F.1 Proof of Corollary 13

Corollary 13 (Uniform sampling minimises simple recommender regret bounds)**.** *Among all sampling policies maintaining full support, uniform sampling ($d_t(a) = 1/K$) minimises the upper bounds in Theorem 4 across all notions of recommender regret. Table 2 lists the bounds for each notion of simple regret for an instance (i.e., not in expectation, but these results translate directly to the expectation).*

Table 2: Regret bounds comparing Uniform and E3W sampling policies.

Bounded Term	**Uniform**	**E3W**
$\left\|\hat{Q}_t - Q^\star\right\|_\star$	$2K^{3/2}\left\|\mathbf{e}_1\right\|_\star\sqrt{\frac{\sigma^2\ln(4Kt/\delta)}{t}}$	$4K^{3/2}\left\|\mathbf{e}_1\right\|_\star\sqrt{\frac{\sigma^2\ln(2+t)\ln(4Kt/\delta)}{t}}$
$\|\hat{\pi} - \pi^\star\|$	$\frac{1}{\beta}2K^{3/2}\left\|\mathbf{e}_1\right\|_\star\sqrt{\frac{\sigma^2\ln(4Kt/\delta)}{t}}$	$\frac{1}{\beta}4K^{3/2}\left\|\mathbf{e}_1\right\|_\star\sqrt{\frac{\sigma^2\ln(2+t)\ln(4Kt/\delta)}{t}}$
$D_\Omega(\hat{\pi},\pi^\star)$	$\frac{1}{\beta}K^3\left\|\mathbf{e}_1\right\|_\star^2\frac{\sigma^2\ln(4Kt/\delta)}{t}$	$\frac{2}{\beta}K^3\left\|\mathbf{e}_1\right\|_\star^2\frac{\sigma^2\ln(2+t)\ln(4Kt/\delta)}{t}$
$\left\|\hat{V} - V^\star\right\|$	$2\sigma\sqrt{\frac{K\ln(4Kt/\delta)}{t}}$	$4\sigma\sqrt{\frac{K\ln(2+t)\ln(4Kt/\delta)}{t}}$

Proof of Corollary 13. It is evident in Equation 41 that ζ is the principle quantity which controls the constants of the bound. If we want to minimise the bound, we should take ζ to its maximum possible value, across all actions. This implies that uniform sampling should produce the tightest Q-value error bound.

We start by proving the Q-error bound under E3W-based action sampling, and then prove a similar result for uniform sampling. Both results work off of the proof in Lemma 12.

E3W-based sampling Dam et al. (2021, Lemma 1) and Xiao et al. (2019, Lemma 4) show that $\hat{N}_t(a) \geq \frac{t}{2\ln(t+2)}$ for E3W-based action selection. Thus, with probability $1-\delta_1$ we have that

$$N_t(a) \geq \hat{N}_t(a) - \sqrt{\frac{t\ln(1/\delta_1)}{2}} \geq \frac{t}{2\ln(t+2)} - \sqrt{\frac{t\ln(1/\delta_1)}{2}}$$

Then, with $\frac{t}{(\ln(2+t))^2} \geq 2\ln(1/\delta_1)$, we see that $N_t(a) \geq \frac{t}{4\ln(t+2)}$. Following the derivation in Lemma 12, we can now achieve the following bound, under the intersection event $E \cap G$, which holds for each $a \in \mathcal{A}$ for $t \geq T_0$.

$$|\hat{Q}_t(a) - Q^\star(a)| \leq \sqrt{\frac{2\sigma^2\ln(4Kt/\delta)}{N_t(a)}} \leq \sqrt{\frac{16K\ln(t+2)\sigma^2\ln(4Kt/\delta)}{t}}. \tag{46}$$

Uniform sampling We may proceed trivially by setting $N_t(a) = t/K$, which is achieved on the K-th time step, $t = nK$, with a deterministic uniform sampling policy that samples each action sequentially over K time steps. The following bound holds, under the event G, for each $a \in \mathcal{A}$.

$$|\hat{Q}_t(a) - Q^\star(a)| \leq \sqrt{\frac{2\sigma^2\ln(4Kt/\delta)}{N_t(a)}} = \sqrt{\frac{2\sigma^2 K\ln(4Kt/\delta)}{t}}. \tag{47}$$

□

F.2 Proof of Corollary 14

This is pushed to the appendix because it does not directly give the expected regret bound, but is close.

Corollary 14 (Recommender MCTS optimal regret bound)**.** *Under the hypotheses of Theorem 5, the uniform sampling policy $d_t(a|s) = 1/K$ maximizes $\zeta^* = 1/K$. This choice minimizes both C_2/ζ^H and the search horizon T. In particular, for error probability $\delta \in (0, \min(C_1(TK)^H, 1)]$ and search horizon $T \geq \frac{C_2K^{2+3H}}{C_3^2}\ln\left(\frac{C_1(TK)^H}{\delta}\right)$, uniform sampling guarantees that with probability at least $1-\delta$:*

$$\mathcal{SRR}_T^{f,D} \leq \sqrt{\frac{C_2K^{3H}}{T}\ln\left(\frac{C_1(TK)^H}{\delta}\right)}. \tag{48}$$

Proof of Corollary 14. Summing the allocation constraint $d_t(a|s) \geq \zeta$ over all $K = |\mathcal{A}|$ actions yields $1 \geq K\zeta$, establishing that $\zeta \leq 1/K$. The maximum threshold $\zeta^* = 1/K$ is uniquely attained by the uniform sampling policy $d_t(a|s) = 1/K$. Choosing $\zeta^* = 1/K$ minimises C_2/ζ^H, which simultaneously minimises the regret upper bound and the required search horizon.

From Theorem 5, we have the failure probability

$$\begin{aligned}\delta &= C_1(TK)^H\exp\left(-\frac{\zeta^H T\epsilon^2}{C_2K^{2H}}\right)\\ &= C_1(TK)^H\exp\left(-\frac{T\epsilon^2}{C_2K^{3H}}\right). \qquad (\zeta^* = 1/K)\end{aligned}$$

Solving for ϵ gives

$$\epsilon = \sqrt{\frac{C_2K^{3H}}{T}\ln\left(\frac{C_1(TK)^H}{\delta}\right)}.$$

By Theorem 5, this concentration bound holds with probability at least $1-\delta$. Notice that for the bound to be non-vacuous, then T must be relatively large. Provided $\epsilon \leq C_3\zeta^* = C_3/K$, we can

substitute our expression for ϵ into this threshold and obtain

$$\sqrt{\frac{C_2K^{3H}}{T}\ln\left(\frac{C_1(TK)^H}{\delta}\right)} \le \frac{C_3}{K}$$
$$\frac{C_2K^{2+3H}}{C_3^2}\ln\left(\frac{C_1(TK)^H}{\delta}\right) \le T.$$

□

F.3 Proof of Corollary 6

Expected regret bound for MCTS.

Proof of Corollary 6. Because $\mathcal{SRR}_T^{f,D}$ is a non-negative random variable strictly bounded by

$$\mathcal{SRR}_T^{f,D} \le C_3\left\|\hat{Q}_t - Q^\star\right\|_\infty \le 2C_3R_{\max},$$

according to the assumptions of the Theorem, its expectation can be computed by integrating its cumulative distribution function:

$$\mathbb{E}[\mathcal{SRR}_T^{f,D}] = \int_0^{2C_3R_{\max}} \mathbb{P}\left(\mathcal{SRR}_T^{f,D} \ge \epsilon\right) d\epsilon.$$

Following Corollary 14, which bounds the instance-dependent $\mathcal{SRR}$, we choose $d_t(a|s) = 1/K$ to maximise $1/\zeta$, which consequently minimises the required search horizon for the bounds to hold:

$$T \ge \frac{C_2K^{2+3H}}{C_3^2}\ln\left(\frac{C_1(TK)^H}{\delta}\right),$$

which guarantees that

$$u = \sqrt{\frac{C_2K^{3H}}{T}\ln(C_1(TK)^H)} \le C_3\zeta^* = \frac{C_3}{K}.$$

We can then set $\delta = 1$ in the high-probability horizon bound from Corollary 14 to arrive at our choice for u: this means that when $\epsilon < u$, the error probability of this event is $\le \delta = 1$. Thus, we partition the integration into three intervals: $[0, u]$, $[u, C_3/K]$, and $[C_3/K, 2C_3R_{\max}]$.

Interval 1: $[0, u]$
We trivially bound the probability by 1:

$$\int_0^u \mathbb{P}\left(\mathcal{SRR}_T^{f,D} \ge \epsilon\right) d\epsilon \le \int_0^u 1\, d\epsilon = u = \sqrt{\frac{C_2K^{3H}}{T}\ln(C_1(TK)^H)}.$$

Interval 2: $[u, C_3/K]$
In this region, $\epsilon \le C_3/K$, so Theorem 5 applies. Extending the upper limit to infinity yields an upper bound. Applying the standard Gaussian tail inequality $\int_u^\infty e^{-ax^2}dx \le \frac{1}{2au}e^{-au^2}$ with $a = T/C_2$:

$$\int_u^{\frac{C_3}{K}} \mathbb{P}\left(\mathcal{SRR}_T^{f,D} \ge \epsilon\right) d\epsilon \le \int_u^\infty C_1(TK)^H \exp\left(-\frac{T\epsilon^2}{C_2K^{3H}}\right) d\epsilon$$
$$\le C_1(TK)^H\left(\frac{C_2K^{3H}}{2Tu}\right)\exp\left(-\frac{Tu^2}{C_2K^{3H}}\right).$$

By our definition of u, the term $\exp(-Tu^2/C_2K^{3H}) = \exp(-\ln(C_1(TK)^H)) = \frac{1}{C_1(TK)^H}$. Substituting this back cancels the leading coefficient:

$$\int_u^{\frac{C_3}{K}} \mathbb{P}\left(\mathcal{SRR}_T^{f,D} \ge \epsilon\right) d\epsilon \le \frac{C_2K^{3H}}{2Tu} = \frac{1}{2}\sqrt{\frac{C_2K^{3H}}{T\ln\left(C_1(TK)^H\right)}}.$$

Interval 3: $[C_3/K, 2C_3R_{\max}]$
For $\epsilon > C_3/K$, notice that the probability is monotonically decreasing, so it is bounded by its value at the threshold $\epsilon = C_3/K$, allowing us to leverage Theorem 5. The length of this interval is strictly less than $2C_3R_{\max}$, providing

$$\int_{\frac{C_3}{K}}^{2C_3R_{\max}} \mathbb{P}\left(\mathcal{SRR}_T^{f,D} \geq \epsilon\right) d\epsilon \leq \int_0^{2C_3R_{\max}} \mathbb{P}\left(\mathcal{SRR}_T^{f,D} \geq \frac{C_3}{K}\right) d\epsilon$$
$$\leq 2C_3R_{\max} \cdot C_1(TK)^H \exp\left(-\frac{TC_3^2}{C_2K^{2+3H}}\right).$$

Summing the bounds from the three intervals yields the final result in Equation 49. As $T \to \infty$, the first term scales as $\tilde{O}(\sqrt{\frac{K^{3H}}{T}})$ and dominates, while the second term scales as $\tilde{O}((T\ln T)^{-1/2})$, and the final term decays exponentially.

$$\begin{aligned}\mathbb{E}[\mathcal{SRR}_T^{f,D}] \leq & \sqrt{\frac{C_2K^{3H}}{T}\ln\left(C_1(TK)^H\right)} \\ & + \frac{1}{2}\sqrt{\frac{C_2K^{3H}}{T\ln\left(C_1(TK)^H\right)}} \\ & + 2C_3R_{\max}C_1(TK)^H \exp\left(-\frac{TC_3^2}{C_2K^{2+3H}}\right).\end{aligned} \tag{49}$$

It is also of interest to bound $\mathcal{SRR}_T^{\pi,B}$ for the Bregman divergence B_Ω. Noting the result in Theorem 5, we know that the Bregman divergence-based regret is bound by $C_3^2\left\|\hat{Q}_t - Q^\star\right\|_\infty^2$, which means that we can set the error lower bound in the probability statement to ϵ^2. Doing so yields a slightly modified regret bound and proof, but one of the expected rate: $\frac{1}{T}$.

$$\begin{aligned}\mathbb{E}[\mathcal{SRR}_T^{\pi,B_\Omega}] &= \int_0^{4C_3^2R_{\max}^2} \mathbb{P}\left(\mathcal{SRR}_T^{\pi,B_\Omega} \geq \epsilon\right) d\epsilon \\ &\leq \int_0^{4C_3^2R_{\max}^2} \mathbb{P}\left(C_3^2\left\|\hat{Q}_t - Q^\star\right\|_\infty^2 \geq \epsilon\right) d\epsilon \\ &= \int_0^{4C_3^2R_{\max}^2} \mathbb{P}\left(C_3\left\|\hat{Q}_t - Q^\star\right\|_\infty \geq \sqrt{\epsilon}\right) d\epsilon\end{aligned}$$

We can now choose

$$u = \frac{C_2K^{3H}}{T}\ln(C_1(TK)^H) \leq \frac{C_3^2}{K^2}.$$

Our high-probability bound from Corollary 14 still holds under $\sqrt{\epsilon}$, and we can still set $\delta = 1$. This means that the only difference in the above proof is the limits of integration, and what they resolve to. The following completes the proof.

Interval 1: $[0, u]$
We trivially bound the probability by 1:

$$\int_0^u \mathbb{P}\left(\mathcal{SRR}_T^{\pi,B_\Omega} \geq \epsilon\right) d\epsilon \leq \int_0^u 1\, d\epsilon = u = \frac{C_2K^{3H}}{T}\ln(C_1(TK)^H).$$

Interval 2: $[u, C_3^2/K^2]$
In this region, $\sqrt{\epsilon} \leq C_3/K$, so Theorem 5 applies. Extending the upper limit to infinity yields an

upper bound.

$$\begin{aligned}\int_u^{\frac{C_3^2}{K^2}} \mathbb{P}\left(\mathcal{SRR}_T^{\pi,B_\Omega} \geq \epsilon\right) d\epsilon &\leq \int_u^{\infty} C_1(TK)^H \exp\left(-\frac{T\epsilon}{C_2K^{3H}}\right) d\epsilon \\ &\leq -C_1(TK)^H \left(\frac{C_2K^{3H}}{T}\right) \exp\left(-\frac{T\epsilon}{C_2K^{3H}}\right)\Bigg|_u^{\infty} \\ &= C_1(TK)^H \left(\frac{C_2K^{3H}}{T}\right) \exp\left(-\frac{Tu}{C_2K^{3H}}\right) \\ &= \frac{C_2K^{3H}}{T}\end{aligned}$$

Interval 3: $[C_3^2/K^2, 4C_3^2R_{\max}^2]$
For $\sqrt{\epsilon} > C_3/K$, the probability is monotonically decreasing, so it is bounded by its value at the threshold $\sqrt{\epsilon} = C_3/K$. The length of this interval is strictly less than $4C_3^2R_{\max}^2$, providing:

$$\begin{aligned}\int_{\frac{C_3^2}{K^2}}^{4C_3^2R_{\max}^2} \mathbb{P}\left(\mathcal{SRR}_T^{\pi,B_\Omega} \geq \epsilon\right) d\epsilon &\leq \int_0^{4C_3^2R_{\max}^2} \mathbb{P}\left(\mathcal{SRR}_T^{\pi,B_\Omega} \geq \frac{C_3^2}{K^2}\right) d\epsilon \\ &\leq 4C_3^2R_{\max}^2 \cdot C_1(TK)^H \exp\left(-\frac{TC_3^2}{C_2K^{2+3H}}\right).\end{aligned}$$

Summing these terms, we get a bound of $\tilde{O}\left(\frac{K^H}{T}\right)$:

$$\begin{aligned}\mathbb{E}[\mathcal{SRR}_T^{\pi,B_\Omega}] \leq &\frac{C_2K^{3H}}{T} \ln(C_1(TK)^H) \\ &+ \frac{C_2K^{3H}}{T} \\ &+ 4C_3^2R_{\max}^2 \cdot C_1(TK)^H \exp\left(-\frac{TC_3^2}{C_2K^{2+3H}}\right).\end{aligned} \tag{50}$$

□

G ADDITIONAL DETAILS AND PLOTS FOR SYNTHETIC TREE EXPERIMENTS

The purpose of this experiment is to highlight the differences between E3W and uniform selection. To adequately compare E3W action selection to uniform action selection, we set the exploration parameter $\epsilon = 0.1$. In our testing, there was little difference in performance for different selections of ϵ, excluding of course when ϵ is large, which means E3W performs similarly to uniform sampling, for some time. We also change $\lambda_{s_t} = \frac{\epsilon}{\log(N_t(s_t)+2)}$ since this forces $\lambda \in [0, 1]$ (when $\lambda_{s_t} = \frac{\epsilon K}{\log(N_t(s_t)+2)} > 1$ like in Dam et al. (2021) we no longer have a convex combination in d_t^{E3W}).

We present complete plots for the synthetic tree experiments. These plots include all the depth and action space size settings that we tested. Note that any plots labelled "theoretical" use the primal or dual norm that is associated with the regulariser. For instance, entropy is strongly convex with respect to the L1 norm, and Tsallis entropy with respect to the L2 norm. Thus, the dual norm for entropy-regularised Q-values is the L-infinity norm, while for Tsallis it is L2.

Note how in the "theoretical" plots, Tsallis entropy and UCT often have significantly different error values to the other regularised RL methods. In the Bregman divergence plots, this is due to the KL divergence being unbounded, while the L2 policy error is bounded. In the Q-value error plots, this is because the L2 error is significantly higher than the L-infinity error.

In the main body, we elected to present plots that were similarly scaled through the use of the same error function. This meant that we sometimes couldn't accurately represent the regret curve of the bandit / tree search, but a proxy of that curve through another error function. This choice was made to avoid unnecessary confusion regarding which method was best, since the use of different norms would skew the positioning of the curves.

"Why does UCT drop suddenly?" and other related questions answered When we retrieve Q-value estimates from the tree, we always zero out the unvisited action's Q-values, because otherwise the default value is negative infinity. This means that we suffer constant regret on unvisited arms for those Q-value predictions. While papers like Danihelka et al. (2022) would use a sort of mixing operator to replace unvisited Q-values with the value of the parent, we do not use such a device. This may be useful for future investigation, but would likely not yield terribly different plots. Note that "constant regret on unvisited arms" means that UCT will especially have difficulty minimising Q-value error since it doesn't visit all arms until some high number of simulations, in our results around 4500 or 5000 simulations. Even the greedier sampling policies like E3W suffer this fate, since they tend to sample high performing arms more often (as is expected by the greedy policy $\hat{\pi}_t$).

The following lists the plots in this appendix.

- **Figure 4** illustrates the Q-value mean squared error plots in synthetic trees.
- **Figure 5** shows the value mean squared error plots in synthetic trees.
- **Figure 6** displays the L2 norm-based policy error plots in synthetic trees.
- **Figure 7** depicts the theoretical norm-based policy error plots in synthetic trees.
- **Figure 8** demonstrates the KL divergence-based value error plots in synthetic trees.
- **Figure 9** highlights the theoretical Bregman divergence-based value error plots in synthetic trees.
- **Figure 10** shows the Q-value error for each seed (random environment over which 100 random tree searches were run)
- **Figure 11** shows the value error for each seed (random environment over which 100 random tree searches were run)
- **Figure 12** shows the norm-based policy error for each seed (random environment over which 100 random tree searches were run)
- **Figure 13** shows the Bregman-based policy error for each seed (random environment over which 100 random tree searches were run)

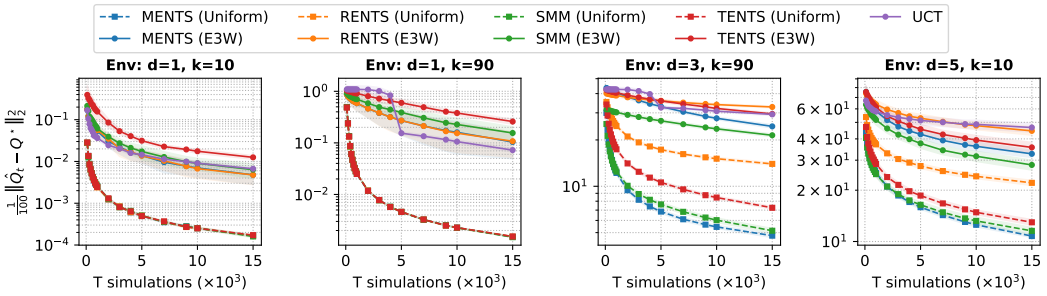

Figure 4: Q-value mean squared error plots in synthetic trees.

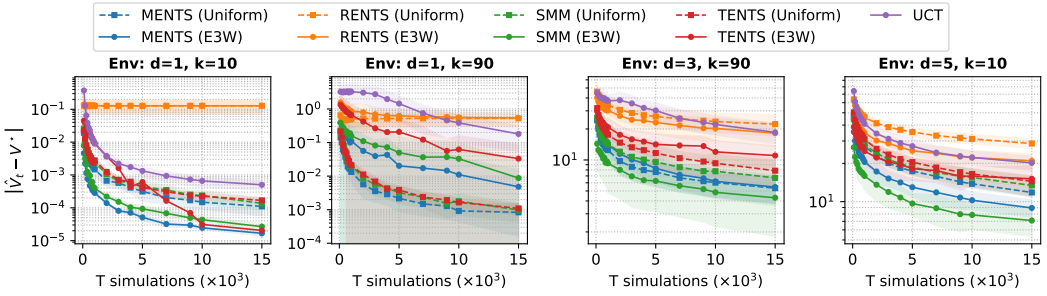

Figure 5: Value mean squared error plots in synthetic trees.

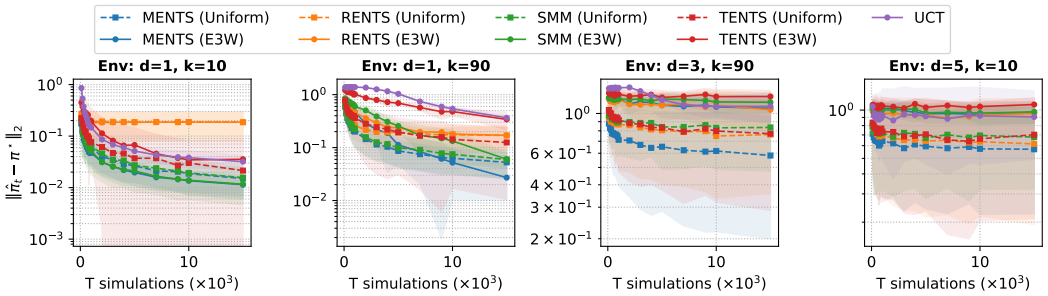

Figure 6: L2 norm-based policy error plots in synthetic trees.

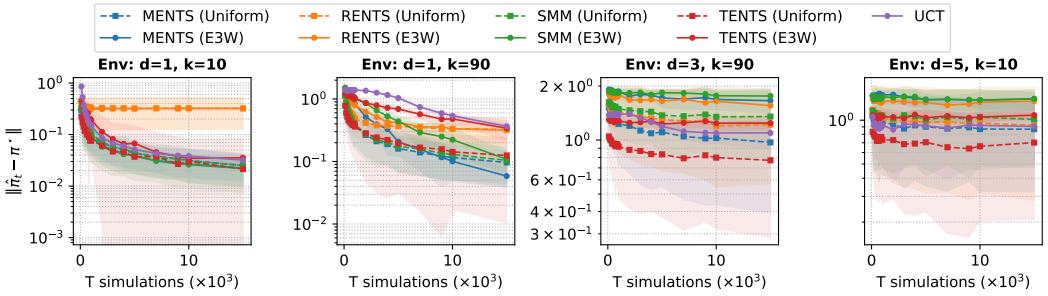

Figure 7: Theoretical norm-based policy error plots in synthetic trees.

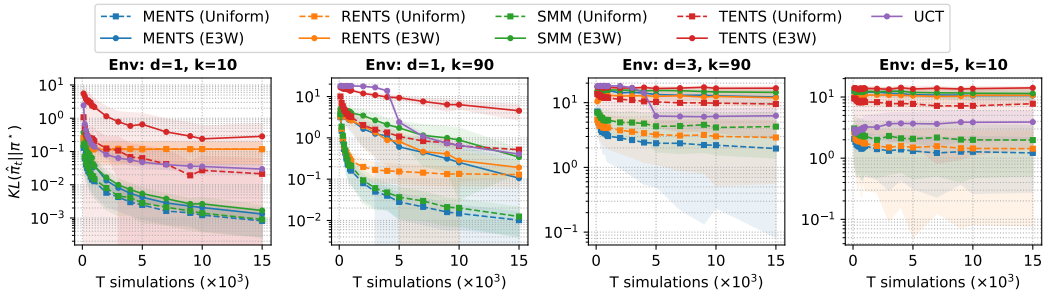

Figure 8: KL divergence-based error plots in synthetic trees.

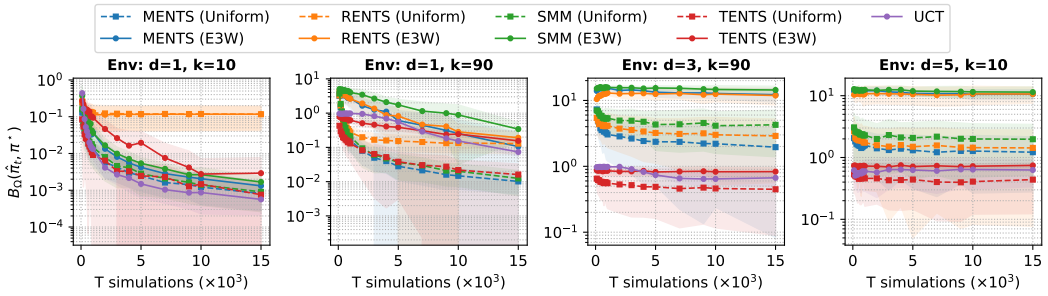

Figure 9: Theoretical Bregman divergence-based value error plots in synthetic trees.

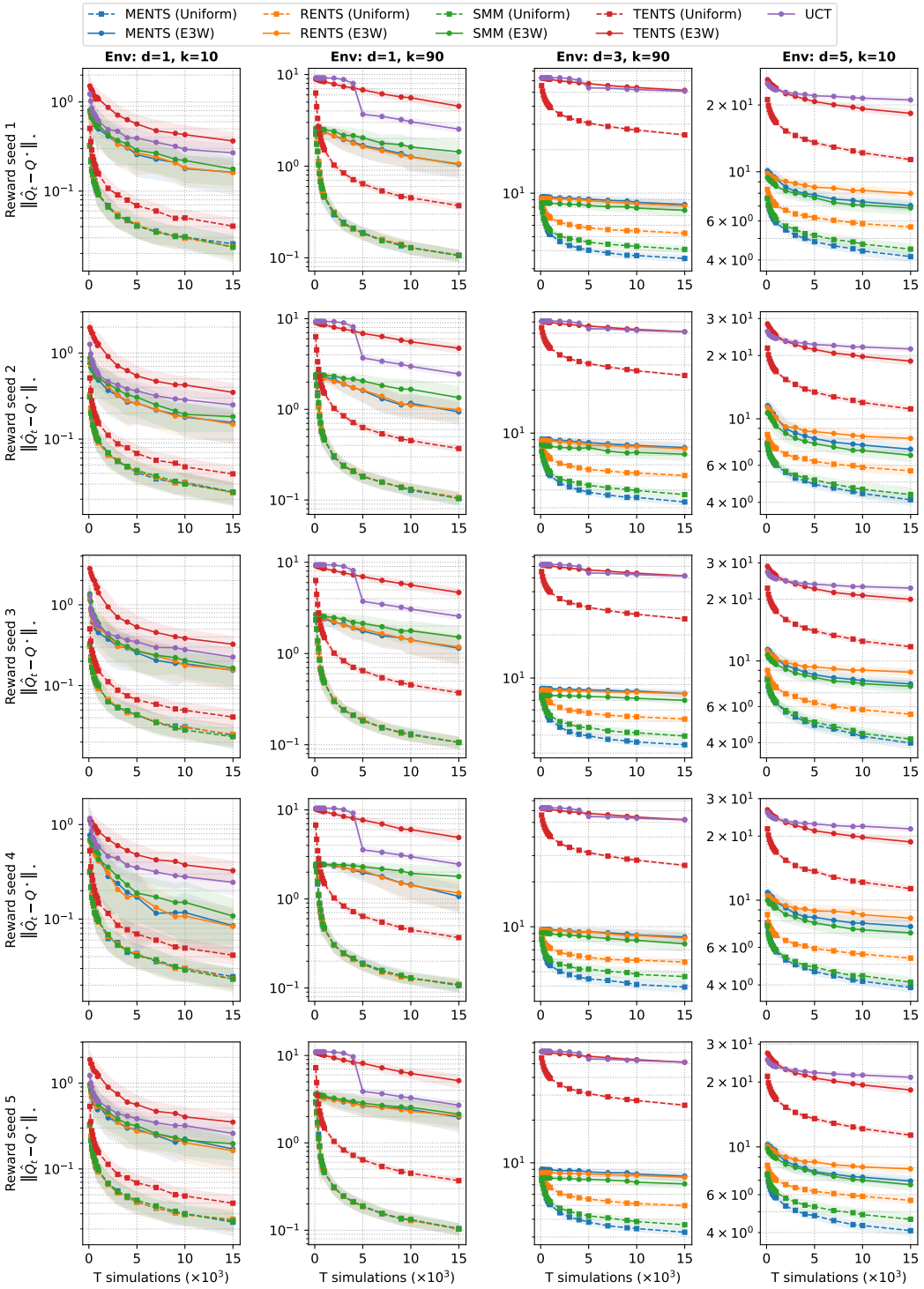

Figure 10: Q-error graphs per seed (5 seeds one per row)

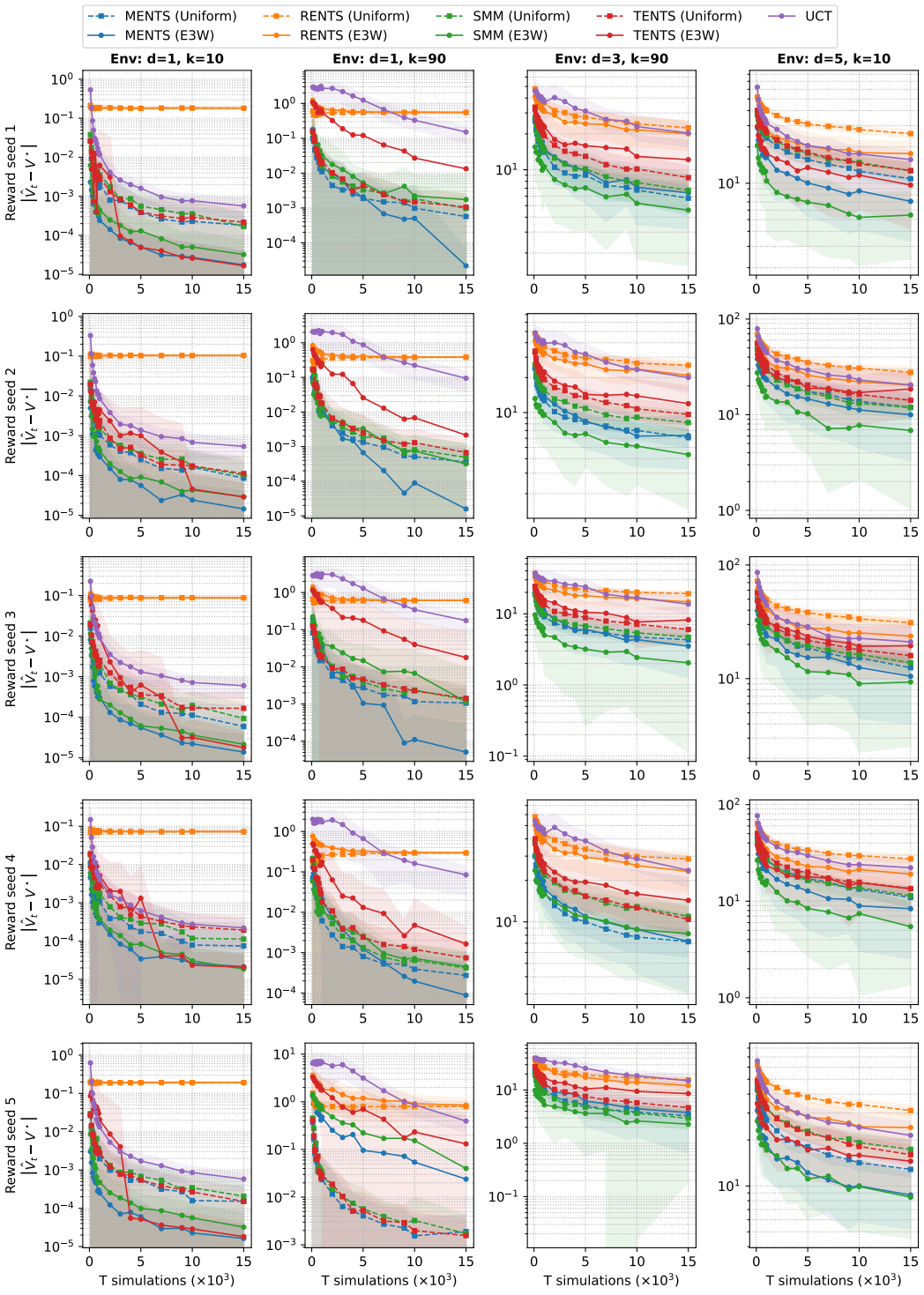

Figure 11: Value error graphs per seed (5 seeds one per row)

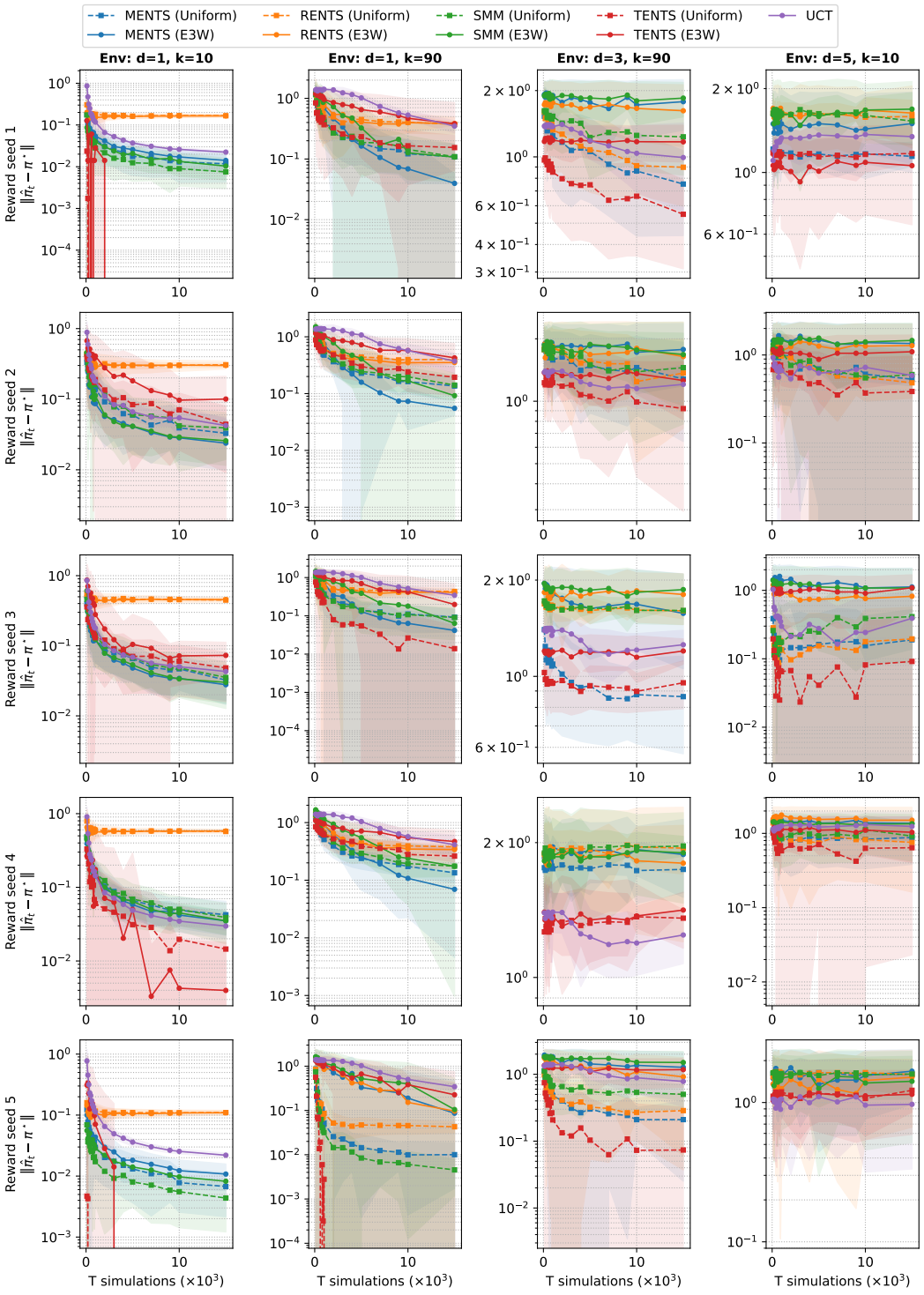

Figure 12: Norm-based policy error graphs per seed (5 seeds one per row)

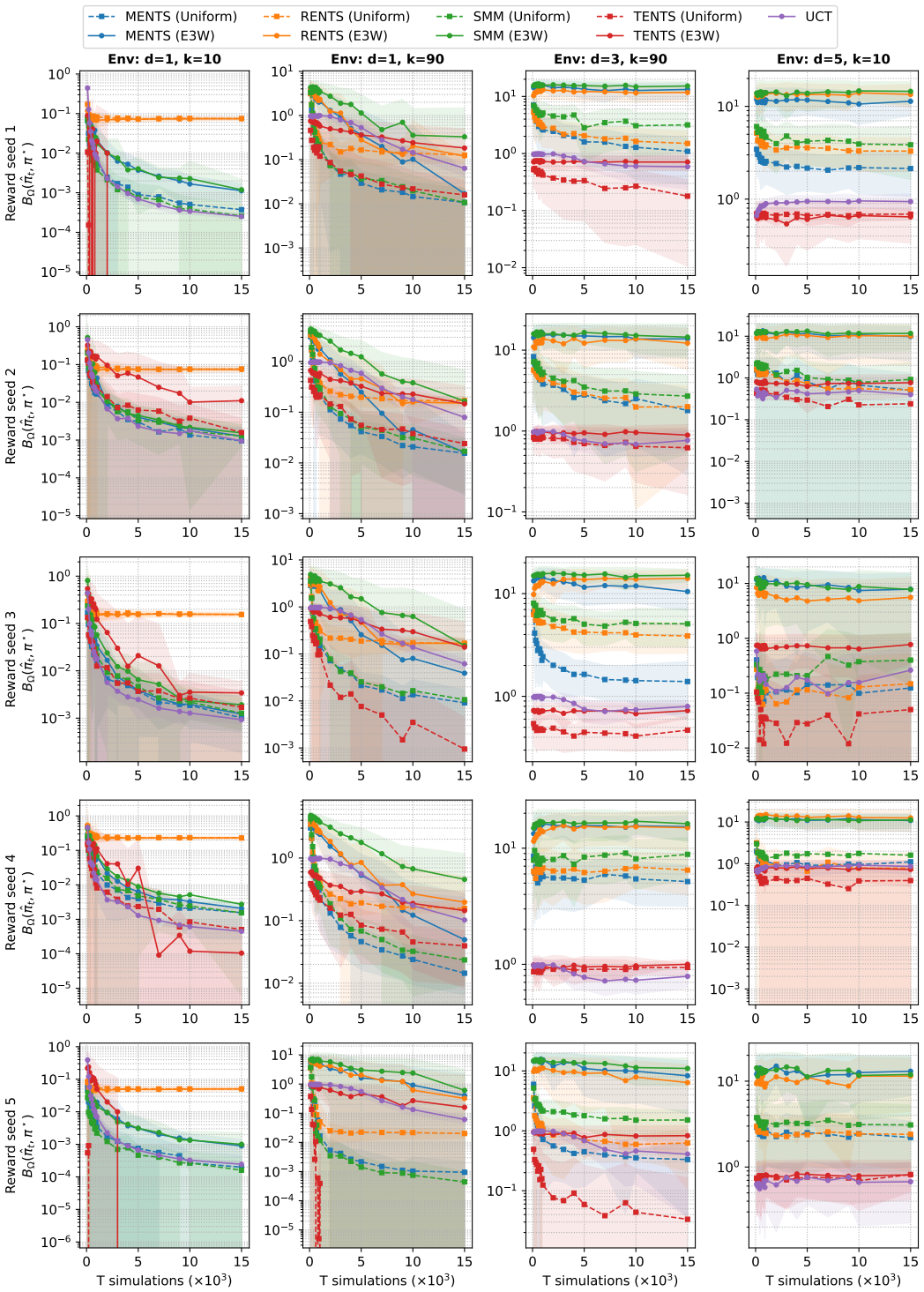

Figure 13: Bregman-based policy error graphs per seed (5 seeds one per row)

H Details of the OCTAX experiments

We follow a PQN-style (Gallici et al., 2025) implementation: 1024 vectorized environments, 16-step rollouts, 3 gradient updates over 16 minibatches per rollout, for a total of 10^7 environment steps. The network is a small convolutional encoder (two 3×3, stride-2 convolutions with 16 and 32 channels, each followed by LayerNorm and ReLU, into a 64-dimensional dense encoding) with a linear Q-head, trained with an ϵ-greedy behaviour policy over a 4-frame stack. We generate Q-value estimates from the Q-head by running $\hat{V}(s) = \Omega^*(Q_\theta(s, \cdot))$.

We swept the inverse temperature $\omega \in 5, 10, 20, 50, 100$ on 15 of the 19 games for the regularised tree searches and UCT. We discovered that performance didn't change considerably over these wider ω settings: an algorithm that outperforms the baseline with $\omega = 5$ still outperforms at $\omega = 20$, for instance. Since we aren't looking to demonstrate state-of-the-art performance in these tasks, we found it counterproductive, and computationally wasteful, to run full sweeps over five ω values. Tree search depth is capped at 10, and $\gamma = 1$ to match full-episode returns, with the number steps per episode fixed based on that game's typical episode length. For UCT specifically, we also swept over the Dirichlet hyperparameters to see how performance might improve with more or less exploration, but again did not observe significant changes in performance. In particular, all settings of ω and the exploration hyperparameters that we swept had returns within one standard deviation of one another, indicating that there was no significant differences in performance. For the OCTAX experiments, we followed the hyperparameters of Xiao et al. (2019) and Dam et al. (2021) for their arcade game experiments: $\epsilon = 0.1$, with $\lambda_{s_t} = \frac{\epsilon K}{\log(N_t(s_t)+2)}$.

The tasks that we filtered out were done so to choose which ones to run for the 512 simulation budget. These experiments took many days (*c.f.*, Figure 14), so we had to be sure which ones to run.

I Additional plots for OCTAX experiments

We first present the plots for environments where we ran 32 to 512 MCTS simulations per environment step. Then, we present the plots for all the environments together, where we show the results for 32 to 128 MCTS simulations per step.

- **Figure 14** illustrates the execution times for the large-sim task set.
- **Figure 15** illustrates the episode lengths for the large-sim task set.
- **Figure 16** illustrates the episodic return for all tasks, with 32 to 128 MCTS simulations per step.
- **Figure 17** illustrates the episode lengths for all tasks, with 32 to 128 MCTS simulations per step.
- **Figure 18** illustrates the execution times for all tasks, with 32 to 128 MCTS simulations per step.

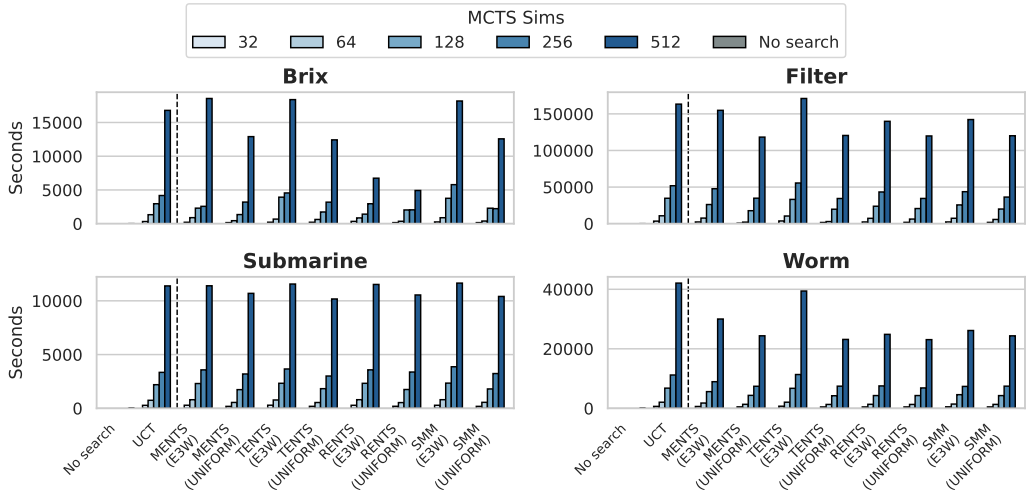

Figure 14: Execution times increase quickly with sims per step.

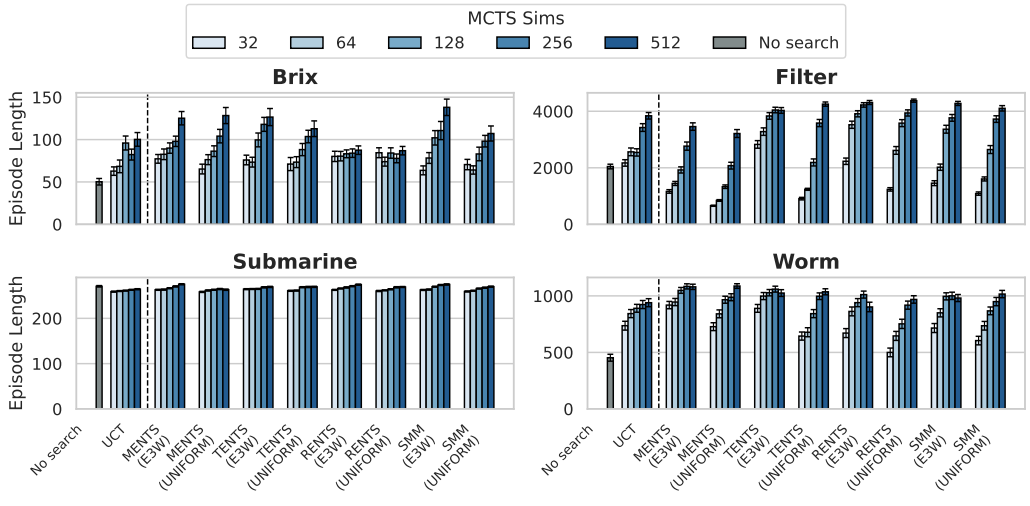

Figure 15: Episode lengths stay relatively similar with increased sims per step.



Figure 16: Episodic returns for all OCTAX tasks.

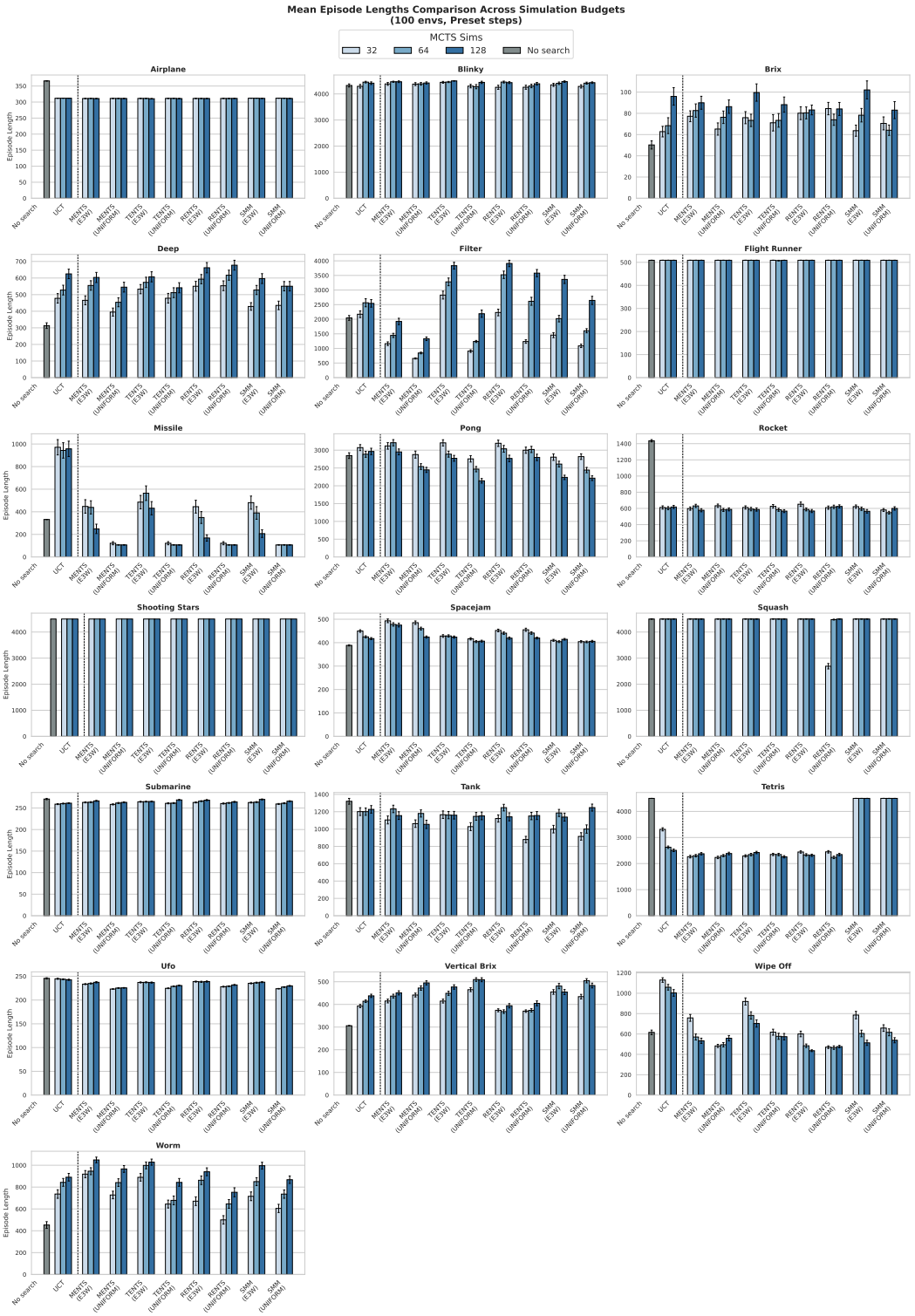

Figure 17: Episode lengths for all OCTAX tasks.

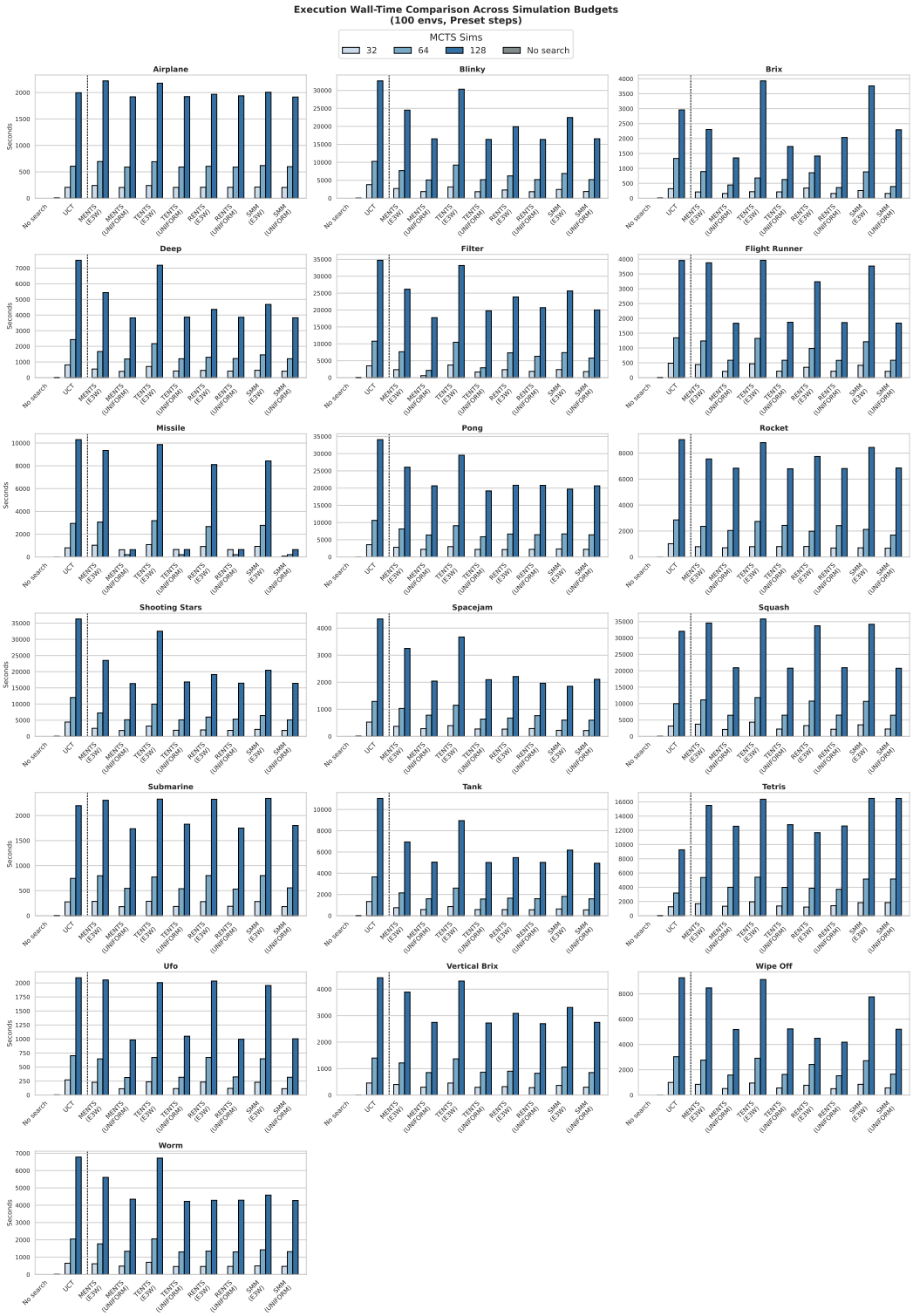

Figure 18: Execution times for all OCTAX tasks.

J Details of the scientific discovery experiments

On the top-k metric Because a policy that repeatedly regenerates one high-reward sequence is far less useful than one that discovers many distinct high-reward sequences, we report a diversity-aware top-k reward: sequences are ranked by reward and greedily added to a selected set only if they are at least a task-specific Levenshtein distance δ from every sequence already selected ($\delta = 10$ for AMP, $\delta = 60$ for GFP, $\delta = 13$ for UTR); we report the mean reward of the selected set at $k \in \{1, 10, 20, 50, 100\}$. This measures each method's ability to cover distinct modes of the reward landscape rather than only its best-case sample.

Details about the models. We re-use the code from Jiralerspong et al. (2025) to train the proxy reward models and to train the TGM models. For each task we train a causal sequence-Transformer policy (3 layers, 8 attention heads, 64-dimensional embeddings/hidden size, dropout 0.1, sinusoidal positional encoding) with a soft, regularization-consistent trajectory-balance-style objective; the network has no separate value head; its next-token logits are treated as prior Q-values for search, and the state value is reconstructed analytically from the same causal forward pass (summing the trained model's per-token credit along the prefix generated so far). Note that this last part is required for tree search since we need value estimates, not just Q-values estimates. According to (Jiralerspong et al., 2025, Table 2), we select the top-performing hyperparameters for each task, and use those to train our baselines: set $q = 0.75$ for UTR and AMP, and $q = 0.25$ for GFP.

We did notice that our proxy reward model and TGM models did not give as high of numbers as in Jiralerspong et al. (2025). This likely just boils down to implementation differences, different seeds being employed, or just differences between the trained proxy reward models, which are used in evaluation.

Details about the tree search. Rather than sweeping a single ω, we run a 17-point sweep over $\omega \in \{0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 2, 4, 8, 10, 15, 20, 25, 30, 35, 40\}$ for every algorithm/simulation-budget configuration, so that we can separately report both peak achievable reward and sensitivity to this hyperparameter. Each configuration generates 100 sequences in parallel, repeated over 5 random seeds for the regularized algorithms and their no-search and UCT baselines. This sweep is done to generate a larger set of diverse sequences, so that we can evaluate the diversity of high-reward objects generated by each configuration. Note that this is standard procedure for these tasks in the GFlowNet-adjacent literature, see specifically Whitammer et al. (2022) and Jiralerspong et al. (2025). It is also standard to compute the mean reward of the top-100 diverse sequences according to the thresholds on the Levenshtein distance. Finally, we acknowledge that the sweep over UCT's temperature hyperparameter is not necessarily the best way to search a reward landscape, but we also note that UCT is not designed for such tasks, so adapting it to fit scientific discovery better is not within the scope of this investigation.

Like in the OCTAX experiments, we set $\epsilon = 0.1$ with $\lambda_{s_t} = \frac{\epsilon K}{\log(N_t(s_t)+2)}$ for E3W, like prior work. Tree depth is capped at 9 for all tasks. For GFP, whose much longer episodes make search at every generation step prohibitively expensive, we instead run full tree search only every 10th token and sample directly from the trained policy otherwise, restricting that task's simulation budget to 32. For sake of transparency, running MCTS for each of the 237 token of the GFP task took roughly 20-30 minutes for a single ω at just 32 simulations per token. Scaling this up to every ω we swept over, and each simulation count, the task would take far too much compute, a quick estimate states roughly a GPU-month of compute. As such, our GFP results are limited by time and compute considerations. Note that it is standard practice in the research literature to run MCTS as reduced intervals. Xiao et al. (2019) did so in the beam rider task in Atari since the sequences were too long to run MCTS on each step.

K Additional results for the scientific discovery experiments

Table 3 lists the performance of all tree search methods and the no-search baseline for all 3 scientific discovery tasks. It is clear that uniform sampling performs better than E3W for almost all search methods, for almost all tasks. Uniform sampling with Tsallis entropy-based tree search (TENTS) is particularly performant in the scientific discovery tasks.

Table 3: Top-100 diverse mean reward (mean ± std across seeds) for each search algorithm and action-selection rule, at each task's highest-performance simulation budget: AMP/UTR: 1024 simulations for most methods, other than UCT; GFP: 32 simulations for all methods. Bold marks a configuration whose mean exceeds the no-search baseline for that task. †: fewer than 5 seeds had completed; mean and std are over the seeds available.

Method	AMP	UTR	GFP
Baseline (no search)	10.17 ± 1.76	4.13 ± 0.02	$\mathbf{1.61 \pm 0.02}$
SMM (E3W)	6.63 ± 0.23	3.47 ± 0.02	1.53 ± 0.01
SMM (Uniform)	7.69 ± 0.74	3.66 ± 0.01	1.53 ± 0.01
MENTS (E3W)	8.56 ± 0.48	3.65 ± 0.01	1.53 ± 0.01
MENTS (Uniform)	9.13 ± 0.98	3.65 ± 0.01	1.53 ± 0.01
RENTS (E3W)	$7.57 \pm 0.10^{\dagger}$	4.12 ± 0.01	1.58 ± 0.01
RENTS (Uniform)	7.57 ± 0.05	$\mathbf{4.24 \pm 0.01}^{\dagger}$	1.60 ± 0.01
TENTS (E3W)	$9.49 \pm 0.07^{\dagger}$	$4.09 \pm 0.02^{\dagger}$	$\mathbf{1.61 \pm 0.00}$
TENTS (Uniform)	$\mathbf{10.64 \pm 0.14}$	$\mathbf{4.25 \pm 0.01}$	$\mathbf{1.62 \pm 0.01}$
UCT	7.94 ± 1.45	$\mathbf{4.20 \pm 0.02}$	$\mathbf{1.61 \pm 0.04}$

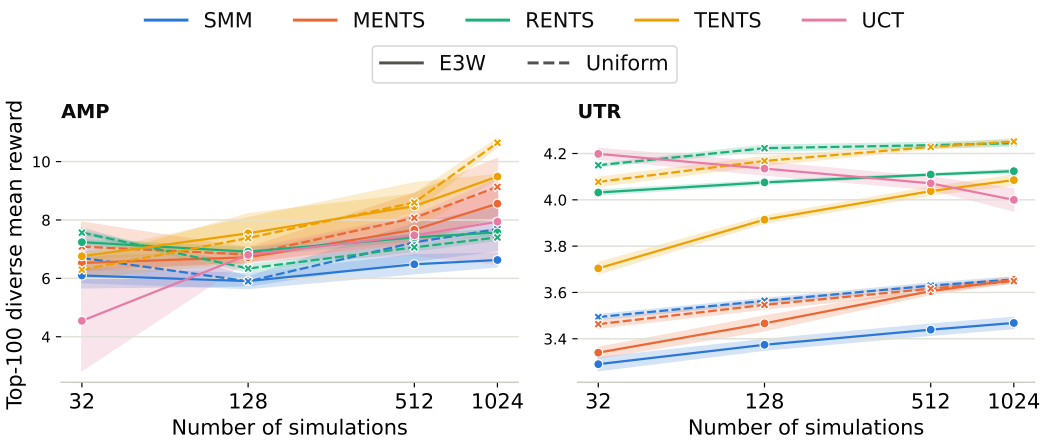

Figure 19: Performance of tree search in the AMP and UTR tasks, with increasing MCTS simulations per environment step. The top-100 diverse mean reward consistently increases for all tasks, with uniform sampling (dotted lines) often outperforming E3W. UCT did not finish at the 1024 simulation budget. The mean and standard deviation are computed over 5 random seeds of the generation process, sweeping inverse temperatures between 0.025 to 40. GFP is not presented on this figure because it takes too long to run with higher simulation budgets.

Figure 19 shows the performance of the algorithms on AMP and UTR, as the number of simulations increases from 32 to 1024. Performance clearly improves with the number of simulations. We also clearly see that UCT doesn't finish at the 1024 mark for either of the tasks.

Figure 20 shows plots across top-K settings **for a single seed**. It is clear that the baseline model is very strong at lower values of K: 1, 10, and 20 are virtually unbeaten. For top-50 and top-100 mean rewards, it is clear that regularised MCTS algorithms are able to produce higher average rewards, especially as the number of simulations increases. AMP is the only task where no tree search method seems to clear the maximum average reward, but as is visible in the box-and-whisker plots of Figure 3, TENTS and RENTS are able to achieve competitive, or higher, rewards at 1024 simulations per step. Like in Figure 19, we see that performance improves as sims increase. Finally, it is interesting to note that uniform sampling often outperforms E3W in UTR, but not in AMP, except at 1024 simulations. We do not have an explanation of this phenomenon.

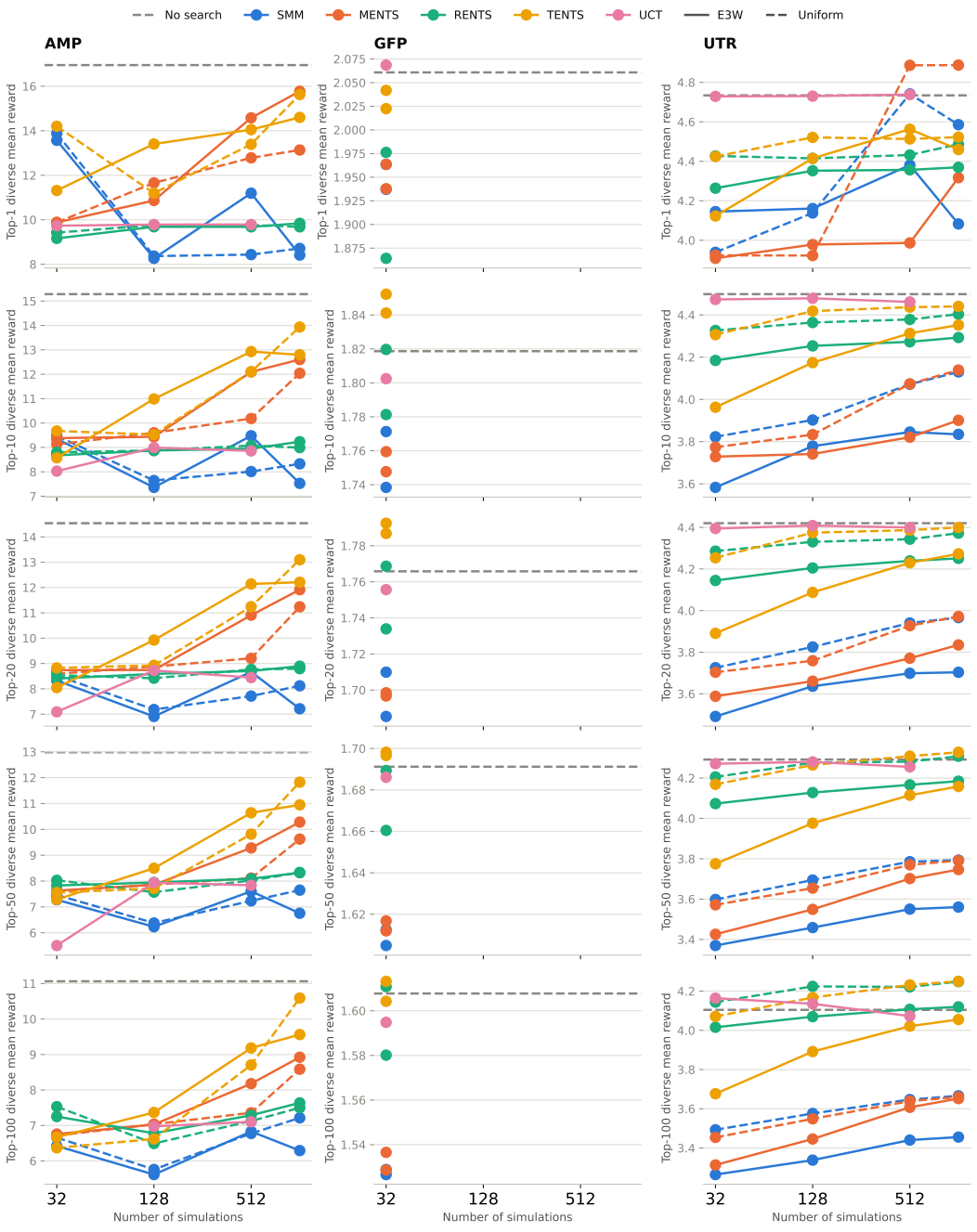

Figure 20: Performance of tree search across all scientific discovery tasks at different top$-k$ levels, with increasing simulation budgets.