WHEN ARE GFLOWNETS SCALABLE? A REGRET ANALYSIS OF SAMPLE EFFICIENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Flow Networks (GFlowNets) are designed to learn distributions over combinatorial objects whose underlying state spaces can be prohibitively large. While substantial progress has been made in improving GFlowNet training, through better trajectory sampling, exploration, and gradient estimation, a basic theoretical question remains unsolved: *when is a GFlowNet statistically scalable, and how does the structure of the task determine its sample efficiency?* We study this question through regret, where the performance gap is the trajectory KL divergence from the reward-proportional target distribution. Our central finding is that the relevant statistical scale is the representation dimension d, rather than the state space size. We first establish an oracle-sampling lower bound of $\Omega(\frac{dH}{1+\log d}\log(1+K/d))$. This bound reveals a data-starved regime: when $K \ll d$, the worst-case average performance gap does not vanish, so the trajectory budget must reach the scale $K = \Omega(d)$ to leave this regime. We then investigate whether the statistical scaling of the lower bound can be approached constructively. We develop an optimistic policy-learning algorithm that leverages a shared d-dimensional representation of policy values. A first-order analysis yields regret $O(dH\sqrt{K\log(1+K/d)})$, and a second-order analysis further improves the dependence on the trajectory budget K to be logarithmic under the stated assumptions. Together, these results identify representation dimension as a fundamental measure of GFlowNet learnability: enormous state spaces of GFlowNet can remain statistically scalable when their relevant value structure admits a compact shared representation, whereas without such compression the effective dimension d can approach the state space size. This provides a formal step toward the theory of the statistical scalability of GFlowNets.

1 INTRODUCTION

Generative Flow Networks (GFlowNets) learn distributions over discrete and combinatorial objects. Given a nonnegative reward function $R(x)$ over $\mathcal{X}$, the goal is to construct a generative policy whose terminal distribution satisfies $P^\star(x) = \frac{R(x)}{Z^\star}$ with $Z^\star = \sum_{x\in\mathcal{X}} R(x)$. Unlike optimization methods that seek only a maximizer of R, GFlowNets aim to sample diverse objects in proportion to their rewards. They achieve this through sequential construction on a directed acyclic graph (DAG), where states $s \in \mathcal{S}$ represent partially constructed objects and trajectories terminate at elements of $\mathcal{X}$ (Bengio et al., 2021a; 2023).

A central difficulty is that the resulting construction space can be prohibitively large. Existing GFlowNet methods improve how information from sampled trajectories is used. Flow Matching imposes local flow-consistency constraints, while Subtrajectory Balance propagates training signals over longer portions of trajectories (Bengio et al., 2023; Madan et al., 2023; He et al., 2025). Policy-based approaches instead optimize the generative policy directly, using gradient estimators motivated by variational formulations (Malkin et al., 2022b; Deleu et al., 2024; Niu et al., 2026). Other work improves trajectory collection and reuse through exploration, policy ensembles, replay, and local search (Rector-Brooks et al., 2023; Shen et al., 2023; Kim et al., 2023). Nevertheless, when the state space can not be enumerated, learning must rely on finitely many sampled trajectories. This raises a basic theoretical question: *when is a GFlowNet statistically scalable, and how does the structure of the task determine its sample efficiency?*

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 study this question through regret analysis. Under the soft Bellman formulation considered in this paper, $V_0^{\pi}(s_0) = \log Z^{\star} - D_{\mathrm{KL}}(P_{\pi} \| P^{\star})$. So the performance gap of a policy is exactly its trajectory KL divergence from the reward-proportional target distribution. Consequently,

$$\mathrm{Reg}(K) = \sum_{k=1}^{K} D_{\mathrm{KL}}(P_{\pi^k} \| P^{\star})$$

directly measures cumulative generative error. Regret analysis has a long history in bandits and reinforcement learning, where optimism in the face of uncertainty, as exemplified by UCB methods, balances estimated values with uncertainty bonuses to obtain sublinear performance regret Lattimore & Szepesvári (2020). This principle has been extended to finite-horizon MDPs through methods such as UCB-RL and UCB-VI (Auer et al., 2008; Azar et al., 2017; Jin et al., 2018; Ménard et al., 2021). However, tabular regret guaranties retain explicit dependence on the number of states, which becomes problematic for combinatorial state spaces in GlowNets. Our focus is therefore different: we use regret not only as a performance measure but also as a tool to characterize the statistical learnability of a target GFlowNet and to identify the structural quantity that governs its sample efficiency.

Our central perspective is that the statistical difficulty of GFlowNet learning need not scale with the raw number of states. When the relevant policy values admit a shared d-dimensional representation, observations collected in one part of the DAG can provide information about many others. We therefore use d to measure the compressibility of the policy values relevant to learning. Without such compression, the representation can degenerate to the tabular scale, with d approaching $|\mathcal{S}|$. We first isolate the intrinsic statistical difficulty through an oracle-sampling benchmark, where the learner observes independent trajectories directly from the unknown target distribution and therefore does not need to explore the DAG. Even under this stronger information access and with at most two actions per state, we establish the lower bound $\Omega\left(\frac{dH}{1+\log d}\log(1+\frac{K}{d})\right)$ for a given DAG horizon H. Thus, the lower bound is nearly linear in d and logarithmic in K, up to the factor $\log d$. More importantly, it reveals a data-starved regime: when $K \ll d$, increasing the number of trajectories does not reduce the worst-case average performance gap. Hence, the trajectory budget must reach at least the scale $K = \Omega(d)$ to leave this regime.

Motivated by this lower bound, we next ask whether its dimension-dependent statistical scaling can be approached constructively. For deterministic GFlowNet DAGs, we propose an optimistic policy-learning algorithm that uses exact Bellman relabeling together with a shared d-dimensional representation of continuation values. A first-order analysis gives $\mathbb{E}[\mathrm{Reg}(K)] = O\Big(dH\sqrt{K\log\big(1+\frac{K}{d}\big)}\Big)$, while a second-order analysis improves the dependence on K to $\mathbb{E}[\mathrm{Reg}(K)] = O\big(dH^2\log(1+K)\big)$. under the stated assumptions. These results show that the statistical scaling suggested by the lower bound can be approached by a computationally implementable algorithm.

Taken together, the lower and upper bounds identify d, rather than the potentially enormous $|\mathcal{S}|$, as a fundamental statistical scale for GFlowNet learning. The lower bound identifies the trajectory scale $K = \Omega(d)$ required to escape the data-starved regime, while the upper bounds show that compact Bellman representations can be exploited algorithmically, with logarithmic dependence on K achievable under the second-order analysis. These results provide a regret-theoretic characterization of GFlowNet learnability: enormous construction spaces can remain statistically scalable when their relevant value structure admits a compact shared representation, whereas without such compression, d can approach the complexity of $|\mathcal{S}|$. In this sense, representation dimension provides a theoretical measure of when GFlowNets can scale statistically beyond the size of their underlying state spaces.

2 PRELIMINARIES

For a finite set $\mathcal{A}(s)$, $\Delta(\mathcal{A}(s))$ denotes the probability simplex. For distributions p, q on the same finite support, $D_{\mathrm{KL}}(p\|q) = \sum_a p(a)\log\frac{p(a)}{q(a)}$. For a positive-definite matrix M, $\|x\|_M = (x^{\top}Mx)^{1/2}$. We use $[H]_0 = \{0, 1, \ldots, H-1\}$. The total number of sampled states is $T = HK$.

2.1 GFlowNet DAG, trajectories, and target distribution

A GFlowNet generates a structured object by sequentially constructing it along a directed acyclic graph (DAG). We focus on a graded DAG, $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H$ with $\mathcal{S}_0 = \{s_0\}$ and $\mathcal{S}_H = \mathcal{X}$. This notation produces no loss of generality since any DAG can be converted into an equivalent graded DAG by inserting dummy states (Malkin et al., 2022b). Every action (edge) $s \to s'$ is a directed edge from $s \in \mathcal{S}_h$ to $s' \subseteq \mathcal{S}_{h+1}$ that is in the child state set $\mathcal{A}(s)$. Thus, the environment dynamics are deterministic once an edge is chosen. A complete trajectory is $\tau = (s_0 \to s_1 \to \ldots \to s_{H-1}, s_H = x)$, and $\mathcal{T}$ denotes the set of all complete trajectories. A Markov forward policy $\pi = \{\pi_h\}_{h=0}^{H-1}$ induces $P_\pi(\tau) = \prod_{h=0}^{H-1} \pi_h(s_{h+1} \mid s_h)$. Let $R : \mathcal{X} \to (0, \infty)$ be the terminal reward, and $Z^\star = \sum_{x \in \mathcal{X}} R(x)$ and $\overleftarrow{\pi}^\star = \{\overleftarrow{\pi}^\star_h\}_{h=0}^{H-1}$ be some given backward policy. The target (oracle) terminal distribution is $P^\star(x) = R(x)/Z^\star$ and the target trajectory distribution is $P_{\pi^\star}(\tau) = P^\star(\tau) = \frac{R(x)}{Z^\star} \prod_{h=0}^{H-1} \overleftarrow{\pi}^\star_h(s_h \mid s_{h+1})$.

The objective of GFlowNet is to ensure that $P_\pi(x)$ matches $P_{\pi^\star}(x)$. Since summing $P^\star(\tau)$ across all backward trajectories that end at a particular x yields $R(x)/Z^\star$, it is enough to align $P_\pi(\tau)$ with $P^\star(\tau)$ to recover the terminal distribution (Malkin et al., 2022a). This trajectory-wise matching over entire trajectories can be reduced to state-wise matching. Define the desired state and edge flows by $F^\star_h(s) := P^*(s)Z^\star$ and $F^\star_h(s, s') := P^*(s, s')Z^\star$, respectively. The flow matching requirement (Bengio et al., 2021b) is that $F^\star(x) = R(x)$ and, recursively,

$$F^\star_h(s, s') = \pi^\star_h(s'|s)F^*_h(s) = F^*_{h+1}(s')\overleftarrow{\pi}^\star_h(s|s') \tag{1}$$

Therefore, for any forward flow $F(s, s')$, the derived terms $F(s) = \sum_{s'} F(s, s')$ and $\pi = F(s, s')/F(s)$ that meet the flow-matching constraint yields $P_\pi(x) = P_{\pi^\star}(x)$.

2.2 Flow matching via Soft Bellman equations

The Soft Bellman equation for a general policy π is presented as follows (Niu et al., 2024). Define $V^\pi_H(x) = \log R(x)$ and recursively

$$Q^\pi_h(s, s') = \rho_h(s, s') + V^\pi_{h+1}(s'), \quad V^\pi_h(s) = \mathbb{E}_{s' \sim \pi_h(\cdot|s)} \left[Q^\pi_h(s, s') - \log \pi_h(s' \mid s)\right] \tag{2}$$

where $\rho_h(s, s') = \log \overleftarrow{\pi}^\star_h(s|s')$ By recursively replacing Q^π_h in the expression of V^π_h with $\rho_h + V^\pi_{h+1}$ for $h = 0, \ldots, H-1$, we have

$$V^\pi(s_0) = \mathbb{E}_\pi \left[\sum_{h=0}^{H-1} \log \frac{\overleftarrow{\pi}^\star_h(s_h|s_{h+1})}{\pi_h(s_{h+1}|s_h)} + \log R(x)\right] = \log Z^* - D_{\mathrm{KL}}(P_\pi \| P^\star). \tag{3}$$

Consequently, the *optimality* of π is determined by the maximization of $\arg\max_\pi V^\pi_0(s_0)$. The flow balance condition 1 admits a logarithmic transform, related to the maximization of the Soft Bellman equation. To be specific, define $Q^\star_h(s, s') = \log F^\star_h(s, s')$, $V^\star_h(s) = \log F^\star_h(s)$, $\rho_h(s, s') = \log \overleftarrow{\pi}^\star_h(s \mid s')$. The flow-matching condition (1) becomes (Tiapkin et al., 2024; Niu et al., 2026)

$$Q^\star_h(s, s') = \rho_h(s, s') + V^\star_{h+1}(s'), \quad V^\star_h(s) = \log \sum_{s' \in \mathcal{A}(s)} \exp Q^\star_h(s, s'), \tag{4}$$

$$\pi^\star_h(\cdot \mid s) = \frac{\exp Q^\star_h(s, \cdot)}{\exp V^\star_h(s)} = \arg\max_{\pi_h} \mathbb{E}_{s' \sim \pi_h(\cdot|s)} \left[Q^\star_h(s, s') - \log \pi_h(s' \mid s)\right]. \tag{5}$$

Here, the second equality in (5) holds, as the expectation can be rewritten as $V^\star_h(s) - D_{\mathrm{KL}}(\pi_h(\cdot \mid s) \| \pi^\star_h(\cdot \mid s))$. Therefore, backward induction yields $V^\star_h \geq V^\pi_h$ for every policy π. Since $Q^\star_h(s, s') - Q^\pi_h(s, s') = V^\star_{h+1}(s') - V^\pi_{h+1}(s')$, we obtain $Q^\star_h \geq Q^\pi_h$ point-wise, justifying the optimality of $\pi^\star$.

Assumption 2.1 (Bounded rewards, backward probabilities, and action sets)**.** *There exist constants $-\infty < \underline{\iota} \leq \overline{\iota} < \infty$, $\underline{\rho} > -\infty$, and $a_{\max} < \infty$, such that, for every terminal state x, every admissible edge $s \to s' = s'$, and every intermediate state s,*

$$\underline{\iota} \leq \log R(x) \leq \overline{\iota}, \qquad \underline{\rho} \leq \log \overleftarrow{\pi}^\star_h(s \mid s') \leq \overline{\rho} \leq 0, \qquad 1 \leq |\mathcal{A}(s)| \leq a_{\max}. \tag{6}$$

Consequently, $\underline{\rho} \leq \rho_h(s, s') \leq \overline{\rho} \leq 0$ and $\underline{\iota} \leq \log R(x) \leq \overline{\iota}$.

3 ORACLE-SAMPLING LOWER BOUND

We study the statistical difficulty of identifying the target GFlowNet policy from trajectory samples when the underlying DAG may be large but every state has only a constant number of outgoing edges (actions). To separate this estimation difficulty from the additional challenge of exploration, we consider an oracle-sampling experiment in which the learner is given independent trajectories drawn directly from the unknown target trajectory distribution. Since the data are generated by the target policy itself, the learner does not need to explore the DAG. Therefore, the resulting lower bound characterizes the intrinsic statistical difficulty of reconstructing a high-dimensional target GFlowNet distribution from trajectory observations. We emphasize that the resulting lower bound is fully information-theoretic: the estimator may be randomized, may yield an arbitrary Markov policy, and need not belong to the parametric family used in the construction.

Formally, consider a GFlowNet parameterized by an unknown parameter θ over a known DAG $\mathcal{G}$. The instance induces a target (oracle) forward policy $\pi_\theta^\star$ and the corresponding target trajectory distribution $P_\theta^\star (:= P_{\pi_\theta^\star})$. Before trajectory episode k, the learner only observes $\{\tau^1, \ldots, \tau^{k-1}\}$ independently sampled from $P_\theta^\star$, while the explicit expression of the oracle defined by $\overleftarrow{\pi}^\star$ and R is not revealed to the learner. Based on the observed oracle trajectories, an arbitrary, possibly random, learning rule $\mathbb{A}_k$ yields a Markov forward policy $\widehat{\pi}^k = \mathbb{A}_k\big(\tau^1, \ldots, \tau^{k-1}\big)$, whose induced trajectory distribution is denoted by $P_{\widehat{\pi}^k}$. The estimator is not required to belong to the parametric family used in the lower-bound construction. We measure the cumulative statistical error of these sequential estimators through the same trajectory-KL performance gap used to evaluate GFlowNet policies (He et al., 2022; Zhou et al., 2021a). That is,

$$\mathrm{Reg}^\star(K) := \sum_{k=1}^{K} \left[V_\theta^\star(s_0) - V^{\widehat{\pi}^k}(s_0) \right] = \sum_{k=1}^{K} D_{\mathrm{KL}} \left(P_{\widehat{\pi}^k} \| P_\theta^\star \right). \tag{7}$$

Therefore, $\mathrm{Reg}^\star(K)$ measures the cumulative error incurred when reconstructing the unknown target GFlowNet policy from an increasing number of oracle trajectory samples.

Theorem 3.1 (Oracle Regret lower bound)**.** *Let $d \geq 2$ and $H \geq 1$. There exist GFlowNets for some graded DAG $\mathcal{G}$, with parameter θ, Horizon H, and action cardinality $|\mathcal{A}(s)| \leq 2$ for every state s, such that the following statements hold. For every sequential estimator $\{\mathbb{A}_k\}_{k=1}^K$ with τ^k sampled independently from $P_\theta^\star$, Consequently, for some universal constant c,*

$$\inf_{\{\mathbb{A}_k\}_{k=1}^K} \sup_{\theta} \mathbb{E}_{\tau^{1:K-1} \sim P_\theta^\star} [\mathrm{Reg}^\star(K)] \geq c \left(d \left\lfloor \frac{H}{1 + \log_2 d} \right\rfloor \log \left(1 + \frac{K}{d} \right) \right). \tag{8}$$

The details are provided in Appendix A. Theorem 3.1 shows that a low-dimensional representation alone does not make the target policy statistically trivial to identify. Even when every state has only two actions and trajectories are sampled directly from the target policy, the cumulative estimation error is at least $\Omega\Big(\frac{dH}{\log d} \log\big(1 + \frac{K}{d}\big) \Big)$, where the $\log d$ factor is the cost of encoding a d-way coordinate choice using binary local actions. Therefore, up to the algorithmic factor, the oracle benchmark exhibits nearly linear dependence on the representation dimension d and logarithmic dependence on the number of trajectories K.

The dependence on K/d also reveal a basic sample-efficiency threshold. In the data-poor regime, $K/d \to 0$, $\log\big(1 + \frac{K}{d}\big) = \frac{K}{d}(1 + o(1))$, and therefore, the average regret satisfies $\frac{1}{K} \inf_{\{\mathbb{A}_k\}_{k=1}^K} \sup_\theta \mathbb{E}[\mathrm{Reg}^\star(K)] = \Omega\Big(\frac{H}{1+\log d} \Big)$. This implies that increasing K while remaining in the regime $K \ll d$ does not reduce the worst-case average estimation error. Thus, the oracle benchmark predicts a data-starved regime in which the target policy cannot be learned to the shrinking average error. The representation dimension thus determines a natural sample-efficiency scale: one needs at least $K = \Omega(d)$ trajectories to leave this data-starved regime.

4 UPPER BOUND UNDER OPTIMISTIC POLICY OPTIMIZATION

Having established in Section 3 an information-theoretic lower bound on the regret of any sequential policy learner, we now turn to the constructive side of the problem: whether one can design a

computationally implementable GFlowNet learning algorithm whose regret approaches the statistical scaling suggested by this lower bound. The oracle result identifies the representation dimension d, the horizon H, and the number of sampled trajectories K as the fundamental quantities governing the difficulty of learning the target policy. This naturally raises the following question: *does there exist an GFlowNet learning algorithm whose regret approaches the established lower bound as closely as possible?* To address this question, we introduce an optimistic policy-optimization algorithm and establish provable regret upper bounds for it. Before going into it, we first make the following assumption.

Assumption 4.1 (Linear Bellman representation)**.** *For every stage $h \in [H]_0$ there is a known feature map*

$$\phi_h : \mathcal{S}_h \times \mathcal{A} \to \mathbb{R}^d, \qquad \|\phi_h(s, s')\|_2 \le 1, \tag{9}$$

and a vector $\theta_h(V) \in \mathbb{R}^d$ for every value function V that can be generated by Algorithm 1, such that

$$V(s') = \phi_h(s, s')^\top \theta_h(V), \qquad \|\theta_h(V)\|_2 \le L. \tag{10}$$

For every stored edge (s_h^t, s_{h+1}^t) and every pointwise-evaluable current V, the scalar $V(s_{h+1}^t)$ can be evaluated exactly.

Remark 4.1 (Representation dimension as compressibility)**.** *Assumption 4.1 is a representation assumption rather than a claim that the underlying GFlowNet is intrinsically linear. The dimension d quantifies how compactly the relevant policy values can be embedded in a shared linear subspace. In the absence of any compressible structure, one may take $\phi(s, \cdot)$ to be the canonical basis over all possible successor states $s' \in \mathcal{S}$, in which case the dimension d reduces to the tabular scale $\approx |\mathcal{S}|$ (Jin et al., 2020; Zhou et al., 2021b;a).*

As V can be represented by the elements of θ_h in the subspace with dimension d, it is reasonable in the second condition to assumed that the absolute value of the i-th element $|\theta_{h,i}|$ is no greater than some constant c_L. Then, $\|\theta_h\|_2 = \sqrt{\sum_{i=1}^d [\theta_h]_i^2} \le \sqrt{d(c_L)^2} = c_L\sqrt{d}$. The third condition is immediate for deterministic dynamics once the current value can be evaluated at the stored child.

The full algorithm is demonstrated in Algorithm 1. In episode k, it constructs its policy from the previous trajectories and then samples a complete trajectory $\tau^k = (s_0^k \to s_1^k \to \ldots \to s_H^k = x^k) \sim P_{\pi^k}$. For a collection of random variables and/or sigma-algebras, we use $\sigma(\cdot)$ to denote the sigma-algebra generated jointly by the listed objects. Let $\mathcal{F}_k := \sigma(\mathcal{F}_0, \tau^1, \ldots, \tau^k)$ for $k = 1, \ldots, K$, with $\mathcal{F}_0$ denoting the information available before the first trajectory is sampled. Thus, $\mathcal{F}_{k-1}$ contains all trajectory information available before episode k. In particular, the episode-k quantities π^k, Q^k, V^k, and δ^k are $\mathcal{F}_{k-1}$-measurable. Accordingly, we further introduce the stage-wise filtration $\mathcal{F}_{k,h} := \sigma\big(\mathcal{F}_{k-1}, s_0^k, s_1^k, \ldots, s_h^k\big)$ for $h = 0, \ldots, H-1$ and $\mathcal{F}_{k,H} = \mathcal{F}_k$. Thus, $\mathcal{F}_{k,h}$ contains all information available before s_{h+1}^k is sampled.

We now describe the backward relabeling and regression step. At the state h of episode k, historical trajectories provide the stage-wise design points $\{(s_h^t, s_{h+1}^t) : t < k\}$. The backward recursion starts from the exact terminal value $V_H^k = \log R$. Once the current value V_{h+1}^k have been constructed, every historical stage-h edge is assigned the refreshed label $y_{t,h}^k$ for $t < k$, and by assumption 4.1 $y_{t,h}^k = \phi_h(s_h^t, s_{h+1}^t)\theta_h^k$. Correspondingly, we define the stage-h ridge estimator by

$$w_h^k = \arg\min_{w \in \mathbb{R}^d} \sum_{t<k} \big(y_{t,h}^k - \phi_h\left(s_h^t, s_{h+1}^t\right)^\top w\big)^2 + \lambda \|w\|_2^2 \tag{11}$$

It is straight-forward to obtain the closed-form solution of w_h^k as follows:

$$\Lambda_h^k = \lambda I_d + \sum_{t<k} \phi_h(s_h^t, s_{h+1}^t)\phi_h(s_h^t, s_{h+1}^t)^\top, \quad b_h^k = \sum_{t=1}^{k-1} \phi_h(s_h^t, s_{h+1}^t) y_{t,h}^k. \tag{12}$$

$$w_h^k = (\Lambda_h^k)^{-1} b_h^k, \quad \widehat{V}_{h+1}^k(s, s') = \phi_h(s, s')^\top w_h^k. \tag{13}$$

The matrix Λ_h^k depends only on the historical design points and therefore admits the rank-one update $\Lambda_h^{k+1} = \Lambda_h^k + \phi_h(s_h^k, s_{h+1}^k)\phi_h(s_h^k, s_{h+1}^k)^\top$. In contrast, b_h^k must generally be recomputed at every

Algorithm 1 Optimistic policy optimization

Require: $(\overleftarrow{\pi}^\star, R)$, $\{\phi_h\}_{h=0}^{H-1}$, $\lambda > 0$, $\beta = \sqrt{\lambda}L$ and $\{[\ell_h, u_h]\}_{h=0}^{H-1}$

Initialize $\mathcal{D}_h^1 \leftarrow \emptyset$ and $\Lambda_h^1 \leftarrow \lambda I_d$ for $h \in [H]_0$
for $k = 1, 2, \ldots, K$ **do**
  Define $V_H^k := \log R$ ▷ Backward pass: fit the episode-k value
  for $h = H-1, H-2, \ldots, 0$ **do**
    $y_{t,h}^k \leftarrow V_{h+1}^k(s_{h+1}^t)$ for $s_{h+1}^t \in \mathcal{D}_h^k$; $b_h^k \leftarrow \sum_{t=1}^{k-1} \phi_h(s_h^t, s_{h+1}^t) y_{t,h}^k$; $w_h^k \leftarrow (\Lambda_h^k)^{-1} b_h^k$
    Compute $V_h^k(s_h^t)$ for $s_h^t \in \mathcal{D}_h^k$
  end for
  $s_0^k \leftarrow s_0$ ▷ Forward pass: evaluate the frozen value lazily
  for $h = 0, 1, \ldots, H-1$ **do**
    Compute $Q_h^k(s_h^k, \cdot)$ for s_h^k; $\pi_h^k(\cdot \mid s_h^k) \leftarrow \frac{\exp\{Q_h^k(s_h^k, \cdot)\}}{\sum_{s' \in \mathcal{A}(s_h^k)} \exp\{Q_h^k(s_h^k, s')\}}$; $s_{h+1}^k \sim \pi_h^k(\cdot \mid s_h^k)$
  end for
  for $h = 0, 1, \ldots, H-1$ **do** ▷ Update the design only after the trajectory ends
    $\mathcal{D}_h^{k+1} \leftarrow \mathcal{D}_h^k \cup \{(s_h^k, s_{h+1}^k)\}$; $\Lambda_h^{k+1} \leftarrow \Lambda_h^k + \phi_h(s_h^k, s_{h+1}^k)\phi_h(s_h^k, s_{h+1}^k)^\top$
  end for
end for

episode because all historical edges are relabeled using the current value function V_{h+1}^k. For a confidence level $\delta \in (0,1)$, define the confidence width

$$\Gamma_h^k(s, s') = \sqrt{\lambda} L \, \|\phi_h(s, s')\|_{(\Lambda_h^k)^{-1}} , \tag{14}$$

Define the optimistic versions of the edge-wise and state-wise value functions as

$$Q_h^k(s,\cdot) := \mathrm{clip}_{[\ell_h, u_h]} \left(\rho_h(s,\cdot) + \widehat{V}_{h+1}^k(s,\cdot) + \Gamma_h^k(s,\cdot) \right), \; V_h^k(s) := \log \sum_{s' \in \mathcal{A}(s)} \exp Q_h^k(s, s') \tag{15}$$

In analogy to the π^*, the definition of π^k is

$$\pi_h^k(\cdot \mid s) := \frac{\exp\{Q_h^k(s,\cdot)\}}{\exp V_h^k(s)} = \arg\max_{\pi_h} \mathbb{E}_{s' \sim \pi_h(\cdot|s)} \left[Q_h^k(s, s') - \log \pi_h(s' \mid s) \right]. \tag{16}$$

Clipping uses the known structural range in Assumption 2.1, ensures that $\rho_h(s, s') + V_{h+1}^k(s') \in [\ell_h, u_h]$, and $V_h^k(s) \in [\ell_h', u_h']$ as demonstrated in Lemma B.1 in Appendix B.1.

4.1 Regret Analysis

Let $P_{\mathbb{A}}$ denote the joint distribution of the trajectories generated by Algorithm 1, $\mathbb{P}_{\mathbb{A}}(\tau^{1:K}) := \prod_{k=1}^K P_{\pi^k}(\tau^k \mid \mathcal{F}_{k-1})$, where, by definition, π^k is $\mathcal{F}_{k-1}$-measurable by $\tau^k \mid \mathcal{F}_{k-1} \sim P_{\pi^k}$. We measure cumulative regret (performance gap) by

$$\mathrm{Reg}(K) = \sum_{k=1}^K \left[V^\star(s_0) - V^{\pi^k}(s_0) \right] = \sum_{k=1}^K D_{\mathrm{KL}}(P_{\pi^k} \| P^\star). \tag{17}$$

Theorem 4.2 (Regert upper bound)**.** *Suppose Assumptions 4.1–2.1 hold for Algorithm 1. Let* $\lambda = 1$, $L = c_L\sqrt{d}$ *and* $g = \max_{h \in [H]_0}(u_h - \ell_h)$. *Then,*

$$\mathbb{E}_{\tau^{1:K} \sim P_{\mathbb{A}}}[\mathrm{Reg}(K)] \le c_L H d \sqrt{8K \log\left(1 + \frac{K}{d}\right)}. \tag{18}$$

and with probability at least $1 - \zeta$,

$$\mathrm{Reg}(K) \le c_L H d \sqrt{8K \log\left(1 + \frac{K}{d}\right)} + g\sqrt{\frac{1}{2} K H \log\frac{1}{\zeta}}. \tag{19}$$

The proof is given in Appendix C.1. Consequently, $\mathbb{E}_{\tau^{1:K}\sim P_{\mathbb{A}}}[\mathrm{Reg}(K)] = O(dH\sqrt{K\log(1+\frac{K}{d})})$. Thus, the regret scales linearly with the feature dimension d, while growing as $\sqrt{K}$ with the number of trajectories. The sublinear dependency on K also means that $\mathbb{E}[\mathrm{Reg}(K)]/K = O(dH\sqrt{\frac{1}{K}\log(1+K/d)}) \to 0$ as $K \to \infty$. Thus, the average performance gap of Algorithm 1 vanishes asymptotically. In the following, we present Theorem 4.3 that complements this result with a purely logarithmic dependence on K.

Theorem 4.3 (Logarithmic rate regret bound)**.** *Suppose Assumptions 4.1 and 2.1 hold, and Algorithm 1 uses* $\lambda = \frac{g^2}{4L^2}$*, and* $g = \max_{h\in[H]_0}(u_h - \ell_h)$*. Then,*

$$\mathbb{E}_{\tau^{1:K}\sim P_{\mathbb{A}}}[\mathrm{Reg}(K)] \le g^2H^2 d\log\left(1+\frac{4L^2K}{g^2d}\right), \tag{20}$$

and with a probability of at least $1-\zeta$*,*

$$\mathrm{Reg}(K) \le \frac{g^2H^2}{1-e^{-1}}\left[d\log\left(1+\frac{4L^2K}{g^2d}\right)+\frac{1}{2}\log\frac{1}{\zeta}\right]. \tag{21}$$

The proof is given in D.2. Consequently, when g and $L = c_L\sqrt{d}$ are treated as problem-dependent constants, $\mathbb{E}_{\tau^{1:K}\sim P_{\mathbb{A}}}[\mathrm{Reg}(K)] = O(dH^2\log(1+K))$.

Taken together, the upper and lower bounds highlight the representation dimension d as a central source of statistical difficulty in GFlowNet learning. The first-order bound $O(dH\sqrt{K\log(1+K/d)})$ has the same explicit dH dependence as the oracle lower bound $\Omega(\frac{dH}{\log d}\log(1+K/d))$, up to the logarithmic factor $\log d$. Its main remaining weakness is the dependence on the number of trajectories with a gap ratio of $\sqrt{K/\log(1+K/d)}$. In particular, in the data-starved regime $K \ll d$, where Theorem 3.1 shows that the average oracle regret cannot vanish, the first-order guarantee is also too loose to certify decreasing average regret. This is consistent with the lower bound's implication that the representation dimension determines a natural sample-efficiency scale and that the number of trajectories must at least reach the order of d before vanishing average error becomes possible.

The second-order bound serves a complementary role. Its $O(dH^2\log K)$ rate shows that the $\sqrt{K}$ dependence of the first-order analysis is not fundamental: logarithmic dependence on K, as suggested by the oracle benchmark for fixed d in the large-sample regime, is achievable. Once the number of trajectories reaches the representation scale, $K = \Omega(d)$, the second-order analysis is substantially more consistent with the sample-efficiency threshold identified by the oracle lower bound. The remaining discrepancy arises mainly from the additional horizon dependence, the binary-routing $\log d$ factor in the lower bound, and the problem-dependent factors involving L and g.

4.2 Proof Sketch

We next summarize the mechanisms behind the two upper bounds. The first-order proof discards the negative comparator terms and accumulates residuals linearly. The second-order proof keeps the comparator KL term and uses it to cancel the linear residual, so that only squared residuals remain. First of all, the following quantities are defined to aid in the regret analysis. We define the Bellman residual of the optimistic surrogate by

$$\delta_h^k(s,s') := Q_h^k(s,s') - \rho_h(s,s') - V_{h+1}^k(s'). \tag{22}$$

Then, we define the state-wise difference between the optimistic value and the value of π^k, as well as the difference between the optimistic value and the value of π^* as

$$\widetilde{\Delta}_h^k(s) := V_h^k(s) - V_h^{\pi^k}(s), \quad \Delta_h^k(s) := V_h^\star(s) - V_h^k(s). \tag{23}$$

Moreover, we define the edge-wise value difference as

$$\widetilde{\nabla}_h^k(s,s') := Q_h^k(s,s') - Q_h^{\pi^k}(s,s'), \tag{24}$$

and it is straightforward to obtain $\widetilde{\Delta}_h^k(s) = \mathbb{E}_{s'\sim\pi_h^k(\cdot|s)}\left[\widetilde{\nabla}_h^k(s,s')\right]$. For the k-th sampled episode τ^k, we define the Martingale difference as

$$M_h^k := \widetilde{\Delta}_h^k(s_h^k) - \widetilde{\nabla}_h^k(s_h^k, s_{h+1}^k). \tag{25}$$

Since π_h^k, Q_h^k, and hence $\widetilde{\nabla}_h^k$ are $\mathcal{F}_{k-1}$-measurable, $\mathbb{E}\left[M_h^k \,\middle|\, \mathcal{F}_{k,h}\right] = \widetilde{\Delta}_h^k(s_h^k) - \mathbb{E}_{s'\sim\pi_h^k(\cdot|s)}\left[\widetilde{\nabla}_h^k(s,s')\right] = 0$. Therefore, $\{M_h^k\}_{k,h}$ forms a martingale-difference sequence when the episode-stage pairs (k,h) are ordered chronologically.

Both upper bounds rely on the following optimistic mechanisms. First, the quantity Γ_h^k serves as an optimism bonus (confidence width) added to Q_h^k to compensate for the Bellman residual δ_h^k. Specifically, it bounds the absolute difference between the estimated value $\widehat{V}^k$ and the learned value V^k in episode k, and therefore also yields an upper bound on the Bellman residual δ_h^k. The formal statement is provided in the following lemma.

Lemma 4.4 (Optimism bonus)**.** *Under Assumptions 2.1 and 4.1, the following inequality $\mathcal{E}$ holds for every $k \le K$, $h \in [H]_0$, and $(s,s') \in \mathcal{S}\times\mathcal{S}$:*

$$\left|\widehat{V}_{h+1}^k(s,s') - V_{h+1}^k(s')\right| \le \Gamma_h^k(s,s'), \quad 0 \le \delta_h^k(s,s') \le 2\Gamma_h^k(s,s'). \tag{26}$$

Consequently , on Γ^k the learned Q^k and V^k act as optimistic value functions, in the sense that-meaning they upper-bound their optimal counterparts, as shown in the following lemma.

Lemma 4.5 (Optimistic propagation)**.** *Suppose that $V_H^k = V_H^\star = \log R$ and for every $k \le K$, $h \in [H]_0$, and $(s,s') \in \mathcal{S}\times\mathcal{S}$, the Bellman residual satisfies $\delta_h^k(s,s') \ge 0$. Then, the algorithmic action and state values are optimistic:*

$$Q_h^k(s,s') \ge Q_h^\star(s,s'), \qquad V_h^k(s) \ge V_h^\star(s). \tag{27}$$

The proofs of the above two lemmas are given in Appendices B.2 and B.3 respectively.

4.2.1 First-order proof: Linear residual accumulation

The first-order proof starts with the following decomposition of the gap between the optimal value and value of the current policy.

Proposition 4.6 (Regret decomposition)**.** *For every episode k, denote $\epsilon_h^k = -D_{\mathrm{KL}}(\pi_h^\star\|\pi_h^k) - \delta_h^k$. Then, $V^\star(s_0) - V^{\pi^k}(s_0)$ is equal to*

$$\sum_{h=0}^{H-1} \mathbb{E}_{P_{\pi^\star}}[\epsilon_h^k] + \mathbb{E}_{P_{\pi^k}}[\delta_h^k(s_h,s_{h+1})] = \sum_{h=0}^{H-1} \mathbb{E}_{P_{\pi^\star}}[\epsilon_h^k] + \delta_h^k(s_h^k,s_{h+1}^k) + M_h^k. \tag{28}$$

Proof. Insert the surrogate value $V_0^k(s_0)$ between the optimal and executed-policy values:

$$V_0^\star(s_0) - V_0^{\pi^k}(s_0) = \left(V_0^\star(s_0) - V_0^k(s_0)\right) + \left(V_0^k(s_0) - V_0^{\pi^k}(s_0)\right). \tag{29}$$

Then, applying Lemmas C.1 and C.3 gives the first term in (28). Applying Lemmas C.2 and C.3 gives the second term. □

Recall that under assumptions 2.1 and 4.1, we have $0 \le \delta_h^k < 2\Gamma_h^k$. In addition, since $D_{\mathrm{KL}}(\pi_h^\star\|\pi_h^k) \ge 0$, it follows that $\epsilon_h^k \le 0$. Therefore, the value gap $V_0^\star(s_0) - V_0^{\pi^k}(s_0)$ can be simplified further and upper bounded by the inequality stated below.

Proposition 4.7 (Regret inequality)**.** *Suppose that for every $k \le K$, $h \in [H]_0$, and $(s,s') \in \mathcal{S}\times\mathcal{S}$: $0 \le \delta_h^k(s,s') \le 2\Gamma_h^k(s,s')$. Then, $V^\star(s_0) - V^{\pi^k}(s_0)$ is no greater than*

$$2\mathbb{E}_{\tau\sim P_{\pi^k}}\left[\sum_{h=0}^{H-1} \Gamma_h^k(s_h,s_{h+1})\right], \quad \sum_{h=0}^{H-1} 2\Gamma_h^k(s_h^k,s_{h+1}^k) + M_h^k. \tag{30}$$

Proof. Substituting (26) into the two terms (28) respectively, we have (30). □

The remaining tasks are to derive upper bounds for Γ_h^k and M_h^k. By applying the self-normalized concentration bound shown in Lemma B.2 in together with the log-determinant bounds B.3 in Appendix B.1, we obtain an upper bound, $\sum_{k=1}^K \Gamma_h^K \leq L\sqrt{K(\lambda+1)\, d \log\left(1+\frac{K}{\lambda d}\right)}$. Moreover, using the Azuma-Hoeffding inequality (Wainwright, 2019) as stated in Lemma B.4 yields a high-probability upper bound, $\sum_{k=1}^K \sum_{h=0}^{H-1} M_h^k \leq g\sqrt{\frac{KH}{2}\log\frac{1}{\zeta}}$.

4.2.2 Second-order Proof: comparator KL cancellation

The second-order proof starts from the same Bellman decomposition but does not discard the comparator KL terms. First of all, we define the total Bellman residual along an episode k by $D_k(\tau) := \sum_{h=0}^{H-1} \delta_h^k(s_h, s_{h+1})$. Then, the regret decomposition (28) can be written as

$$\mathbb{E}_{P_{\pi^k}}[D_k] - \mathbb{E}_{P_{\pi^\star}}[D_k] - D_{\mathrm{KL}}(P_{\pi^\star} \| P_{\pi^k}). \tag{31}$$

Using variational inequality (Hiriart-Urruty & Lemaréchal, 2004), the last two terms are bounded by some second-order quantities, while they produce a first-order cancellation to $E_{\pi^{\pi^k}}[D_k]$. Consequently, we can simplify the regret further by upper-bounding it in the following way.

Lemma 4.8 (Second-order Regret decomposition)**.** *Suppose that $V_H^k = V_H^\star = \log R$ and for every $k \leq K$, $h \in [H]_0$, and $(s, s') \in \mathcal{S} \times \mathcal{S}$, the Bellman residual satisfies $\delta_h^k(s, s') \geq 0$. Then, the regret can be*

$$V^\star(s_0) - V^{\pi^k}(s_0) \leq \frac{1}{2}\mathbb{E}_{\tau\sim P_{\pi^k}}\left[D_k(\tau)^2\right] \leq \frac{H}{2}\sum_{h=0}^{H-1}\mathbb{E}_{\tau\sim P_{\pi^k}}\left[\delta_h^k(s_h, s_{h+1})^2\right]. \tag{32}$$

The proof is given in Appendix D.1. The remaining task is to bound $\delta_h^k(s, s')^2$. Under $\lambda = \frac{g^2}{4L^2}$, Lemma B.5 gives $\sum_{k=1}^K \delta_h^k(s_h^k, s_{h+1}^k)^2 \leq 2g^2 d \log\left(1 + \frac{4L^2K}{g^2 d}\right)$. Taking expectations again over the conditional expectations given $\mathcal{F}_{k-1}$ in Lemma 4.8 results in the desired expected upper bound stated in Theorem 4.3.

For the high-probability upper bound, it is noted that D_k^2 is bounded by $[0, H^2g^2]$. Then, the exponential random sequence $\exp\{(1-e^{-1})\mathbb{E}[\bar{X}_k|\mathcal{F}_{k-1}] - \bar{X}_k\}$ for $\bar{X}_k := D_k^2/H^2g^2$ can be used to construct a nonnegative super-martingale. Therefore, Ville's inequality, gives, with a probability of at least $1-\zeta$, $\sum_{k=1}^K \mathbb{E}_{P_{\pi^k}}[D_k^2] \leq \frac{1}{1-e^{-1}}\left[\sum_{k=1}^K D_k^2 + H^2g^2 \log\frac{1}{\zeta}\right]$ (Wang & Ramdas, 2025). This predictable-to-realized concentration step plays a role analogous to the Azuma–Hoeffding inequality used for the martingale terms M_h^k in the first-order analysis, but it exploits the non-negativity of the squared trajectory residual and therefore avoids introducing a $\sqrt{K}$ deviation term.

5 Conclusion

We study the statistical scalability of GFlowNets, asking when a target reward-proportional policy can be learned efficiently from finitely many trajectories despite an enormous underlying state space. Our upper and lower bounds both identify the representation dimension d, rather than the raw state-space size, as a fundamental bottleneck for GFlowNet learning. In particular, the lower bound reveals a data-starved regime when $K \ll d$: the average performance gap cannot decrease with additional samples when fall into this regime. Thus, the number of trajectories must become comparable to the representation dimension, $K = \Omega(d)$, to leave this regime. Together with our dimension-dependent regret upper bounds, these results characterize d as the natural statistical scale of GFlowNet learning. A remaining limitation is that the current upper and lower bounds are not fully matching; closing this gap under a common model and obtaining matching dependence on d, H, and K is an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools as assistive tools for writing, literature retrieval and discovery, and mathematical derivations.

For writing, the scientific content, arguments, structure, and technical presentation were developed by the authors. Generative AI was used primarily to improve clarity, grammar, organization, and wording of author-written text, without determining the scientific content or conclusions.

For literature retrieval and discovery, the research questions, relevant areas of literature, and judgments about which prior works were pertinent to this paper were determined by the authors. Generative AI was used to accelerate the search for potentially relevant papers and related terminology. All cited works, bibliographic information, and claims about prior literature were checked by the authors against the original sources.

For mathematical derivations and proofs, the main mathematical ideas, theorem formulations, assumptions, proof strategies, and key technical arguments were developed by the authors. Generative AI was used primarily to accelerate derivations under the authors' direction, work through intermediate calculations, test possible extensions or alternative derivations, check algebraic and logical steps, and help refine preliminary proof drafts. The authors independently reviewed and verified all mathematical arguments and proof steps included in the final paper.

In all of these uses, generative AI served as an assistive tool rather than as the source of the paper's scientific contributions. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including its text, claims, proofs, citations, and other artifacts produced with the aid of generative AI.

REFERENCES

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pp. 263–272. PMLR, 2017.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021a.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. GFlowNet foundations. *arXiv e-prints*, pp. arXiv–2111, 2021b.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Tristan Deleu, Padideh Nouri, Nikolay Malkin, Doina Precup, and Yoshua Bengio. Discrete probabilistic inference as control in multi-path environments. *arXiv preprint arXiv:2402.10309*, 2024.

Haoran He, Emmanuel Bengio, Qingpeng Cai, and Ling Pan. Random policy evaluation uncovers policies of generative flow networks. In *Forty-second International Conference on Machine Learning*, pp. 22351–22367. PMLR, 2025.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 4259–4280. PMLR, 2022.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 2020.

Minsu Kim, Taeyoung Yun, Emmanuel Bengio, Dinghuai Zhang, Yoshua Bengio, Sungsoo Ahn, and Jinkyoo Park. Local search gflownets. In *The Twelfth International Conference on Learning Representations*, 2023.

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

Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jain, Andrei Cristian Nica, Tom Bosc, Yoshua Bengio, and Nikolay Malkin. Learning GFlowNets from partial episodes for improved convergence and stability. In *International Conference on Machine Learning*, pp. 23467–23483. PMLR, 2023.

Nikolay Malkin, 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, 2022a.

Nikolay Malkin, Salem Lahlou, Tristan Deleu, Xu Ji, Edward J Hu, Katie E Everett, Dinghuai Zhang, and Yoshua Bengio. Gflownets and variational inference. In *The Eleventh International Conference on Learning Representations*, 2022b.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pp. 7599–7608. PMLR, 2021.

Puhua Niu, Shili Wu, Mingzhou Fan, and Xiaoning Qian. Gflownet training by policy gradients. In *Forty-first International Conference on Machine Learning*, 2024.

Puhua Niu, Shili Wu, and Xiaoning Qian. Evaluating GFlownet from partial episodes for stable and flexible policy-based training. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jarrid Rector-Brooks, Kanika Madan, Moksh Jain, Maksym Korablyov, Cheng-Hao Liu, Sarath Chandar, Nikolay Malkin, and Yoshua Bengio. Thompson sampling for improved exploration in gflownets. In *ICML 2023 Workshop on Structured Probabilistic Inference $\{\backslash\&\}$ Generative Modeling*, 2023.

Max W Shen, Emmanuel Bengio, Ehsan Hajiramezanali, Andreas Loukas, Kyunghyun Cho, and Tommaso Biancalani. Towards understanding and improving gflownet training. In *International Conference on Machine Learning*, pp. 30956–30975. PMLR, 2023.

Daniil Tiapkin, Morozov Nikita, Naumov Alexey, and Vetrov P Dmitry. Generative flow networks as entropy-regularized rl. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Hongjian Wang and Aaditya Ramdas. The extended Ville's inequality for nonintegrable nonnegative supermartingales. *Bernoulli*, 31(4):2723–2746, 2025. doi: 10.3150/24-BEJ1823.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021b.

A ORACLE-SAMPLING LOWER BOUNDS

A.1 BINARY ROUTING CONSTRUCTION

Without loss of generality, we assume that $d = 2^r$ for an integer $r = \log_2 d$ for analysis simplicity. At every intermediate state, the local action set is $\mathcal{A} = \{-1, +1\}$ and $|\mathcal{A}| = 2$. Partition the horizon into $B := \left\lfloor \frac{H}{r+1} \right\rfloor$ complete blocks. Let $m := H - B(r+1) \in \{0, \dots, r\}$ be the number of remaining stages.

- Each block b contains $r + 1$ binary actions. The first r binary routing actions are $A_{b,1}, \dots, A_{b,r} \in \{-1, +1\}$. One additional binary action is $Y_b \in \{-1, 1\}$. Because $d = 2^r$, there is a bijection $\mathbb{J} : \{-1, +1\}^r \to [d] (:= \{1, \dots, d\})$ between the 2^r routing-bit strings and the indices in $[d]$. So $J_b = \mathbb{J}(A_{b,1}, \dots, A_{b,r})$ tell us which coordinate is sampled in $[d]$, and Y_b carries information about that coordinate.
- The final m stages are nuisance binary actions $\bar{A}_1, \dots, \bar{A}_m \in \{-1, +1\}$ that carry no unknown parameters. They are introduced only so that every complete trajectory has exactly H actions.

The deterministic DAG is the complete binary tree of depth H. More precisely, let $\mathcal{S}_0 = s_0$, where s_0 denotes the root, and for every $h \in [H]$ let $\mathcal{S}_h = \{-1, +1\}^h$. Hence, a state $s_h = (a_1, \dots, a_h) \in \mathcal{S}_h$ is identified with the complete sequence of binary actions taken along the unique path from the root to depth h. In particular, $|\mathcal{S}_h| = 2^h$. From every intermediate state $s_h = (a_1, \dots, a_h), h < H$, taking action $a \in \{-1, +1\}$ leads deterministically to $s_{h+1} = (a_1, \dots, a_h, a)$. Thus, every intermediate state has exactly two children, while every non-root state has exactly one parent. Under the block notation, every terminal state $x \in \mathcal{S}_H$ has the unique representation

$$x = \left(A_{1,1}, \dots, A_{1,r}, Y_1, ; \dots, ; A_{B,1}, \dots, A_{B,r}, Y_B, ; \bar{A}_1, \dots, \bar{A}_m\right). \tag{33}$$

Since $B(r+1)+m = H$, this representation contains exactly H binary decisions, and consequently $|\mathcal{S}_H| = 2^H$. Because the construction is a tree, a special DAG without merging states, every non-root state has a unique parent. We therefore set $\overleftarrow{\pi}^\star_h(s_h \mid s_{h+1}) = 1$ on every valid backward edge. Associate one vector $\theta_b = (\theta_{b,1}, \dots, \theta_{b,d}) \in [-1/2, 1/2]^d$ with every complete block $b \in [B]$. Thus, the full unknown parameter is $(\theta_1, \dots, \theta_B) \in [-1/2, 1/2]^{dB}$ containing exactly dB unknown scalar coordinates. For the terminal object corresponding to a complete action sequence,

$$R_\theta(x) := \prod_{b=1}^{B} (1 + Y_b \cdot \theta_{b,J_b}). \tag{34}$$

The reward does not depend on the nuisance actions $\bar{A}_1, \dots, \bar{A}_m$. Since $|\theta_{b,j}| \le \frac{1}{2}$, all terminal rewards satisfy $R_\theta(x) > 0$.

Partition function By definitions, summing over the r routing bits of a block b, is equivalent to summing over $J_b \in [d]$, meaning

$$\sum_{A_{b,1:r}} \sum_{Y_b} (1 + Y_b \cdot \theta_{b,J_b}) = \sum_{j=1}^{d} \sum_{y=\pm 1} (1 + y \cdot \theta_{b,j}) = \sum_{j=1}^{d} (1 + \theta_{b,j} + 1 - \theta_{b,j}) = 2d. \tag{35}$$

Then,

$$Z^\star_\theta := \sum_x R_\theta(x) = \sum_{A_{1:B,1:r},\, Y_{1:B},\, \bar{A}_{1:m}} \prod_{b=1}^{B} (1 + Y_b \cdot \theta_{b,J_b}) \tag{36}$$

Because the B blocks are independent in this product, the sum factorizes. Besides, each of the remaining m nuisance actions has two possible values and does not affect the reward. Therefore,

$$Z^\star_\theta := \prod_{b=1}^{B} \sum_{A_{b,1:r}} \sum_{Y_b} (1 + Y_b \cdot \theta_{b,J_b}) \sum_{\bar{A}_{1:m}} 1 = (2d)^B \sum_{\bar{A}_{1:m}} 1 = (2d)^B 2^m. \tag{37}$$

Since $d = 2^r$, we have $2d = 2^{r+1}$. Thus,

$$Z^\star_\theta = 2^{B(r+1)} 2^m = 2^{B(r+1)+m} = 2^H. \tag{38}$$

Trajectory distribution As the backward policy is identically one, substituting the definitions of $R_\theta(x)$ and $Z_\theta^\star$ gives

$$P_\theta^\star(\tau) = \frac{R_\theta(x)}{Z_\theta^\star} = \frac{\prod_{b=1}^{B}(1 + Y_b \cdot \theta_{b,J_b})}{(2d)^B 2^m} = \prod_{b=1}^{B} \frac{1 + Y_b \cdot \theta_{b,J_b}}{2d} \prod_{\ell=1}^{m} \frac{1}{2}. \tag{39}$$

Because the target trajectory distribution factorizes across blocks, the block variables $(J_1, Y_1), \ldots, (J_B, Y_B)$ are mutually independent under $P_\theta^\star$. The nuisance variables $\bar{A}_1, \ldots, \bar{A}_m$ are also independent of the informative blocks and are uniformly distributed by $\mathrm{Unif}(\{-1, 1\}^m)$. Thus, we can conclude that

$$\Pr_{\theta_b}(J_b, Y_b) = \frac{1 + Y_b \cdot \theta_{b,J_b}}{2d} \tag{40}$$

as the summation over $J_b \in [d]$ and $Y_b \in \{-1, 1\}$ is equal to 1 according to (35). Consequently,

$$\Pr_\theta(J_b = j) = \sum_{y \in \{-1,+1\}} \Pr_\theta(J_b = j, Y_b = y) = \frac{1 + \theta_{b,j}}{2d} + \frac{1 - \theta_{b,j}}{2d} = \frac{1}{d}. \tag{41}$$

$$\Pr_\theta(Y_b = y \mid J_b = j) = \frac{\Pr_\theta(J_b = j, Y_b = y)}{\Pr_\theta(J_b = j)} = \frac{(1 + y \cdot \theta_{b,j})/(2d)}{1/d} = \frac{1 + y \cdot \theta_{b,j}}{2}. \tag{42}$$

$$\mathbb{E}_\theta[Y_b \mid J_b = j] = \frac{1 + \theta_{b,j}}{2} - \frac{1 - \theta_{b,j}}{2} = \theta_{b,j}. \tag{43}$$

Hence, in every complete block b, an oracle trajectory first selects one coordinate $J_b \sim \mathrm{Unif}([d])$ and then provides one binary observation $Y_b \in \{-1, +1\}$ whose conditional mean is the selected parameter coordinate $= \theta_{b,j}$. Since the block pairs (J_b, Y_b) are independent across $b \in [B]$, one oracle trajectory supplies B independent sparse-coordinate observations, one for each parameter vector θ_b.

A.2 From trajectory KL to parameter estimation

The learner is only required to output a policy $\widehat{\pi}^k = \mathbb{A}_n(\tau^1, \ldots, \tau^k)$, with induced trajectory distribution $P_{\widehat{\pi}^k}$. In other word, the learner is not required to output an estimate of the latent parameter θ, nor is $P_{\widehat{\pi}^k}$ required to belong to the hard parametric family. We therefore need to extract a parameter estimate from an arbitrary learned trajectory distribution. Let $\widehat{P}$ be any distribution over complete binary trajectories. For block b, let $\hat{p}_b$ denote the marginal distribution of (J_b, Y_b) under $\widehat{P}$. Define a Hellinger projection

$$\widehat{\theta}_b \in \operatorname*{arg\,min}_{\omega \in [-1/2, 1/2]^d} D_{\mathrm{HL}}(\hat{p}_b, p_\omega). \tag{44}$$

A minimizer exists because the parameter set is compact, and the Hellinger distance is continuous.

For distributions p, q on a finite set, we can write the Total Variation distance and the square of the Hellinger distance as (Tsybakov & Zaiats, 2009)

$$D_{\mathrm{TV}}(p, q) := \frac{1}{2} \sum_x |p(x) - q(x)|, \quad D_{\mathrm{HL}}^2(p, q) := 1 - \sum_x \sqrt{p(x)q(x)} \tag{45}$$

Lemma A.1 (Blockwise Hellinger Distance)**.** *For any $\theta, \omega \in [-1/2, 1/2]^d$, suppose the distributions $p_\omega(j)$ and $p_\theta(j)$ are uniform for $j \in [d]$. Then, the joint distributions $p_\omega(j, y) := \frac{1+y\omega_j}{2d}$ and $p_\theta(j, y) := \frac{1+y\theta_j}{2d}$ for $j \in [d]$ and $y = \pm 1$ satisfy*

$$D_{\mathrm{HL}}^2(p_\theta, p_\omega) \geq \frac{1}{8d} \|\theta - \omega\|_2^2. \tag{46}$$

Proof. By assumption, $p_\theta(y|j) := p_\theta(j,y)/p_\theta(j) = \frac{1+y\theta_j}{2}$. Therefore,

$$D^2_{\mathrm{HL}}(p_\theta, p_\omega) := 1 - \sum_{j=1}^{d} \sum_{y=\pm 1} \sqrt{p_\theta(j,y) p_\omega(j,y)} \tag{47}$$

$$= \frac{1}{d} \sum_{j=1}^{d} 1 - \frac{1}{d} \sum_{j=1}^{d} \sum_{y=\pm 1} \sqrt{p_\theta(y|j) p_\omega(y|j)} \tag{48}$$

$$= \frac{1}{d} \sum_{j=1}^{d} D^2_{\mathrm{HL}} \left(\mathrm{Ber}_{\frac{1+\theta_j}{2}}, \mathrm{Ber}_{\frac{1+\omega_j}{2}} \right), \tag{49}$$

where Ber represents the Bernoulli distributions. It is known that for any two distributions p, q,

$$D_{\mathrm{TV}}(p,q) \le \sqrt{2}\, D_{\mathrm{HL}}(p,q), \tag{50}$$

and

$$D_{\mathrm{TV}} \left(\mathrm{Ber}_{\frac{1+\theta_j}{2}}, \mathrm{Ber}_{\frac{1+\omega_j}{2}} \right) = \frac{1}{2} \sum_{y=\pm 1} \left| \frac{1+y\theta_j}{2} - \frac{1+y\omega_j}{2} \right| = \frac{|\theta_j - \omega_j|}{2}. \tag{51}$$

Consequently,

$$D^2_{\mathrm{HL}} \left(\mathrm{Ber}_{\frac{1+\theta_j}{2}}, \mathrm{Ber}_{\frac{1+\omega_j}{2}} \right) \ge \frac{1}{2} \left(\frac{|\theta_j - \omega_j|}{2} \right)^2 = \frac{(\theta_j - \omega_j)^2}{8}. \tag{52}$$

Substitution into (49) gives (46). □

Lemma A.2 (Trajectory KL controls blockwise parameter error)**.** *For every trajectory distribution $\widehat{P}$ and every $(\theta_1, \ldots, \theta_B) \in [-1/2, 1/2]^{dB}$, we have*

$$D_{\mathrm{KL}} \left(\widehat{P} \| P^\star_\theta \right) \ge \frac{1}{16d} \sum_{b=1}^{B} \|\widehat{\theta}_b - \theta_b\|_2^2. \tag{53}$$

Proof. Let $\hat{p}_{\mathrm{rem}}$ and p_{rem} denote the joint marginal distributions of the remaining nuisance variables, induced by $\widehat{P}$ and $P^\star_\theta$ respectively. Recall that the target trajectory distribution factorizes as:

$$P^\star_\theta(\tau) = \left[\prod_{b=1}^{B} p_{\theta_b}(J_b, Y_b) \right] p_{\mathrm{rem}}, \tag{54}$$

Then, with $\overline{P}(\tau) := \prod_{b=1}^{B} \hat{p}_b(J_b, Y_b) \hat{p}_{\mathrm{rem}}(\bar{A}_{1:m})$

$$D_{\mathrm{KL}}(\widehat{P} \| P^\star_\theta) = D_{\mathrm{KL}} \left(\widehat{P} \| \overline{P} \right) + \sum_{b=1}^{B} D_{\mathrm{KL}}(\hat{p}_b \| p_{\theta_b}) + D_{\mathrm{KL}}(\hat{p}_{\mathrm{rem}} \| p_{\mathrm{rem}}). \tag{55}$$

Consequently, by non-negativity of KL divergence,

$$D_{\mathrm{KL}}(\widehat{P} \| P^\star_\theta) \ge \sum_{b=1}^{B} D_{\mathrm{KL}}(\hat{p}_b \| p_{\theta_b}). \tag{56}$$

By the definition of a minimizer 44,

$$D_{\mathrm{HL}} \left(\hat{p}_b, p_{\widehat{\theta}_b} \right) \le D_{\mathrm{HL}} \left(\hat{p}_b, p_{\theta_b} \right). \tag{57}$$

By the triangle inequality for Hellinger distance,

$$D_{\mathrm{HL}} \left(p_{\widehat{\theta}_b}, p_{\theta_b} \right) \le D_{\mathrm{HL}} \left(p_{\widehat{\theta}_b}, \hat{p}_b \right) + D_{\mathrm{HL}} \left(\hat{p}_b, p_{\theta_b} \right) \le 2 D_{\mathrm{HL}} \left(\hat{p}_b, p_{\theta_b} \right). \tag{58}$$

Combining with Lemma A.1 gives

$$D^2_{\mathrm{HL}} \left(\hat{p}_b, p_{\theta_b} \right) \ge \frac{1}{4} D^2_{\mathrm{HL}} \left(p_{\widehat{\theta}_b}, p_{\theta_b} \right) \ge \frac{1}{32d} \|\widehat{\theta}_b - \theta_b\|_2^2. \tag{59}$$

By monotonicity of the Rényi divergence in its order,

$$D_{\mathrm{KL}}(p\|q) \geq D_{1/2}(p\|q) := -2\log \sum_x \sqrt{p(x)q(x)}. \tag{60}$$

for arbitrary distribution p and q. Using $-\log x \geq 1-x$ and $D^2_{\mathrm{HL}}(p,q) = 1 - \sum_x \sqrt{p(x)q(x)}$, we further obtain

$$D_{\mathrm{KL}}(p\|q) \geq 2D^2_{\mathrm{HL}}(p,q). \tag{61}$$

Hence, by Lemma A.1,

$$D_{\mathrm{KL}}\left(\hat{p}_b \| p_{\theta_b}\right) \geq \frac{1}{16d}\|\widehat{\theta}_b - \theta_b\|_2^2. \tag{62}$$

Summing (62) for all $b \in [B]$ and combing it with (56) proves (53). □

A.3 Coordinatewise van Trees bound

We now lower bound the mean-squared estimation error of each unknown coordinate $\theta_{b,j}$ by applying the scalar van Trees inequality (Tsybakov & Zaiats, 2009).

Choice of prior For each scalar coordinate $\theta_{b,j}$, define the prior density

$$\eta(x) := 2\cos^2(\pi x), \qquad x \in [-1/2, 1/2]. \tag{63}$$

Since $2\int_{-1/2}^{1/2} \cos^2(\pi x)\,\partial x = 1$, η is a probability density on $[-1/2, 1/2]$. Moreover, $\eta(-1/2) = \eta(1/2) = 0$. so the standard boundary regularity condition for the scalar van Trees inequality is satisfied. Its Fisher information (Tsybakov & Zaiats, 2009) is

$$\begin{aligned}\mathbb{I}(\eta) &:= \mathbb{E}_\eta\left[\left(\frac{\partial \log \eta(x)}{\partial x}\right)^2\right] = \int_{-1/2}^{1/2} \frac{[\eta'(x)]^2}{\eta(x)}\,\partial x \\ &= \int_{-1/2}^{1/2} \frac{4\pi^2 \sin^2(2\pi x)}{2\cos^2(\pi x)}\,\partial x = \int_{-1/2}^{1/2} 8\pi^2 \sin^2(\pi x)\,\partial x = 4\pi^2.\end{aligned} \tag{64}$$

Finally, we place an independent product prior on the unknown parameters by assuming $\theta_{b,j} \sim \eta$ for $b \in [B]$ and $j \in [d]$.

Fisher information of one oracle trajectory As block variables are mutually independent according to (39), the log-likelihood of block b is

$$\log p_{\theta_b}(J_b, Y_b) = -\log(2d) + \log\left(1 + Y_b \theta_{b,J_b}\right). \tag{65}$$

Therefore, the Fisher information in one oracle trajectory for $\theta_{b,j}$ is

$$\begin{aligned}\mathbb{I}_{b,j}(\theta) &:= \mathbb{E}_{\tau\sim P^\star_\theta}\left[\left(\frac{\partial}{\partial \theta_{b,j}} \log P^\star_\theta(\tau)\right)^2\right] \\ &= \mathbb{E}_{\tau\sim P^\star_\theta}\left[\mathbf{1}\{J_b = j\}\frac{1}{(1+Y_b\theta_{b,j})^2}\right] \\ &= \sum_{y=\pm 1} \Pr_\theta(J_b = j, Y_b = y)\frac{1}{(1+y\theta_{b,j})^2} \\ &= \frac{1}{d}\sum_{y=\pm 1}\frac{1+y\theta_{b,j}}{2}\frac{1}{(1+y\theta_{b,j})^2} \\ &= \frac{1}{2d}\left[\frac{1}{1+\theta_{b,j}} + \frac{1}{1-\theta_{b,j}}\right] = \frac{1}{d(1-\theta^2_{b,j})}.\end{aligned} \tag{66}$$

Since $|\theta_{b,j}| \leq \frac{1}{2}$, we have $1 - \theta_{b,j}^2 \geq \frac{3}{4}$. Consequently, for every θ, the one-trajectory Fisher information satisfies $\mathbb{I}_{b,j}(\theta) \leq \frac{4}{3d}$.

$$\mathbb{I}_{b,j}(\theta) \leq \frac{4}{3d}. \tag{67}$$

For k independent trajectory samples $\{\tau^1, \ldots, \tau^k\} \sim P_\theta^\star$, $\mathbb{I}_{b,j}$ is additive. Therefore,

$$k\mathbb{I}_{b,j}(\theta) := \mathbb{I}_{b,j}^{(k)}(\theta) \leq \frac{4k}{3d}. \tag{68}$$

Application of the scalar van Trees inequality. Let $\widehat{\theta}_b$ be the Hellinger projection extracted from the arbitrary output trajectory distribution $P_{\widehat{\pi}^k}$ as defined in (44). For every coordinate (b, j), the scalar van Trees inequality gives

$$\mathbb{E}_{\theta\sim\eta,\tau^{1:k}\sim P_\theta^\star}\left[(\widehat{\theta}_{b,j} - \theta_{b,j})^2\right] \geq \frac{1}{\mathbb{E}_{\theta\sim\eta}[\mathbb{I}_{b,j}^{(k)}(\theta)] + \mathbb{I}(\eta)}. \tag{69}$$

Using (64) and (68),

$$\mathbb{E}_{\theta\sim\eta,\tau^{1:k}\sim P_\theta^\star}\left[(\widehat{\theta}_{b,j} - \theta_{b,j})^2\right] \geq \frac{1}{\frac{4k}{3d} + 4\pi^2} = \frac{d}{4(\frac{k}{3} + \pi^2 d)}. \tag{70}$$

Since $\frac{k}{3} + \pi^2 d \leq \pi^2(k + d)$, we further obtain

$$\mathbb{E}_{\theta\sim\eta,\tau^{1:k}\sim P_\theta^\star}\left[(\widehat{\theta}_{b,j} - \theta_{b,j})^2\right] \geq \frac{d}{4\pi^2(k + d)}. \tag{71}$$

Summing (71) over all $b \in [B]$ and $j \in [d]$ yields

$$\mathbb{E}_{\theta\sim\eta,\tau^{1:k}\sim P_\theta^\star}\left[\sum_{b=1}^{B}\sum_{j=1}^{d}(\widehat{\theta}_{b,j} - \theta_{b,j})^2\right] = \mathbb{E}_{\theta\sim\eta,\tau^{1:k}\sim P_\theta^\star}\left[\sum_{b=1}^{B}\|\widehat{\theta}_b - \theta_b\|_2^2\right] \geq \frac{d^2 B}{4\pi^2(k + d)}. \tag{72}$$

A.4 Proof of Theorem 3.1

Proof. Lemma A.2 gives, pointwise in θ and in the observed data,

$$D_{\mathrm{KL}}\left(P_{\widehat{\pi}^k} \| P_\theta^\star\right) \geq \frac{1}{16d}\sum_{b=1}^{B}\|\widehat{\theta}_b - \theta_b\|_2^2. \tag{73}$$

Before episode k, the estimator has observed exactly $k - 1$ oracle trajectory samples. Take the expectation jointly over the fixed product prior and the oracle samples. By (72),

$$\mathbb{E}_{\theta\sim\eta}\mathbb{E}_{\tau^{1:k-1}\sim P_\theta^\star}\left[D_{\mathrm{KL}}\left(P_{\widehat{\pi}^k} \| P_\theta^\star\right)\right] \geq \frac{1}{16d}\frac{d^2 B}{4\pi^2(k - 1 + d)} = \frac{dB}{64\pi^2(k - 1 + d)}. \tag{74}$$

Summing over $k = 1, \ldots, K$,

$$\mathbb{E}_{\theta\sim\eta}\mathbb{E}_{\tau^{1:K-1}\sim P_\theta^\star}[\mathrm{Reg}^\star(K)] \geq \frac{dB}{64\pi^2}\sum_{k=1}^{K}\frac{1}{d + k - 1} = \frac{dB}{64\pi^2}\sum_{n=0}^{K-1}\frac{1}{d + n}. \tag{75}$$

Since $1/(d + x)$ is decreasing for $x \geq 0$,

$$\begin{aligned}\sum_{n=0}^{K-1}\frac{1}{d + n} &\geq \sum_{n=0}^{K-1}\int_n^{n+1}\frac{1}{d + x}\,\partial x \geq \int_0^K \frac{1}{d + x}\partial x \\ &= \log(d + K) - \log d = \log\left(1 + \frac{K}{d}\right).\end{aligned} \tag{76}$$

Therefore,

$$\mathbb{E}_{\theta\sim\eta}\mathbb{E}_{\tau^{1:K-1}\sim P_\theta^\star}[\mathrm{Reg}^\star(K)] \geq \frac{dB}{64\pi^2}\log\left(1 + \frac{K}{d}\right). \tag{77}$$

Finally, the supremum over θ is at least the average under the fixed prior and holds for any arbitrary sequential estimator. This proves (8) when $d = 2^r$.

The preceding derivations and proofs assumed that$d = 2^r$. For arbitrary $d \geq 2$, define

$$d_0 := 2^{\lfloor \log_2 d \rfloor}. \tag{78}$$

Then

$$\frac{d}{2} \leq d_0 \leq d. \tag{79}$$

Apply the preceding construction using only d_0 coordinates. Its routing depth is

$$r_0 = \log_2 d_0 \leq \log_2 d. \tag{80}$$

. Therefore,

$$B_0 := \left\lfloor \frac{H}{1 + \log_2 d_0} \right\rfloor \geq \left\lfloor \frac{H}{1 + \log_2 d} \right\rfloor. \tag{81}$$

The lower bound is

$$\Omega\left(d_0 B_0 \log\left(1 + \frac{K}{d_0}\right)\right). \tag{82}$$

Using $d_0 \geq \frac{d}{2}$ and $\log\left(1 + \frac{K}{d_0}\right) \geq \log\left(1 + \frac{K}{d}\right)$, we obtain (8) for arbitrary $d > 2$. □

B Confidence Analysis

Lemma B.1 (Validity of the clipping intervals)**.** *Initialize the terminal-value bounds as $\ell'_H = \underline{\iota}$ and $u'_H = \overline{\iota}$. For $h = H-1, H-2, \ldots, 0$, recursively define*

$$\ell_h = \underline{\rho} + \ell'_{h+1}, \qquad u_h = \overline{\rho} + u'_{h+1}, \tag{83}$$

$$\ell'_h = \ell_h, \qquad u'_h = u_h + \log a_{\max}. \tag{84}$$

Suppose Assumption 2.1 holds and set $V_H^k = \log R$. For every episode k, stage $h \in [H]_0$, state $s \in \mathcal{S}_h$, and $s' \in \mathcal{A}(s)$,

$$\rho_h(s, s') + V_{h+1}^k(s') \in [\ell_h, u_h], \quad V_h^k(s) \in [\ell'_h, u'_h], \tag{85}$$

Moreover, for any Markov policy $\pi = \{\pi_h\}_{h=0}^{H-1}$, its value functions satisfy

$$Q_h^\pi(s, s') \in [\ell_h, u_h],\ V_h^\pi(s) \in [\ell'_h, u'_h]. \tag{86}$$

Proof. We prove the claims by backward induction. At the terminal stage, both the algorithmic and policy values satisfy $V_H^k(x) = V_H^\pi(x) = \log R(x)$. By Assumption 2.1, $\underline{\iota} \leq \log R(x) \leq \overline{\iota}$, and therefore

$$V_H^k(x),\ V_H^\pi(x) \in [\ell'_H, u'_H].$$

Now suppose that, for some $h \in [H]_0$, $V_{h+1}^k(s'),\ V_{h+1}^\pi(s') \in [\ell'_{h+1}, u'_{h+1}]$ for every $s' \in \mathcal{S}_{h+1}$.

First, consider the algorithmic value. By Assumption 2.1, $\underline{\rho} \leq \rho_h(s, s') \leq \overline{\rho}$. Therefore,

$$\ell_h = \underline{\rho} + \ell'_{h+1} \leq \rho_h(s, s') + V_{h+1}^k(s') \leq \overline{\rho} + u'_{h+1} = u_h. \tag{87}$$

Thus, the exact deterministic Bellman backup lies inside the $[\ell_h, u_h]$. Since $Q_h^k \in [\ell_h, u_h]$ by clipping,

$$\ell_h \leq V_h^k(s) = \log \sum_{s' \in \mathcal{A}(s)} \exp\{Q_h^k(s, s')\} \leq u_h + \log |\mathcal{A}(s)|. \tag{88}$$

Because $|\mathcal{A}(s)| \leq a_{\max}$,

$$\ell'_h \leq V_h^k(s) \leq u'_h.$$

We next consider an arbitrary Markov policy π. Its Bellman equation gives $Q_h^\pi(s,s') = \rho_h(s,s') + V_{h+1}^\pi(s')$. By Assumption 2.1,

$$\ell_h = \underline{\rho} + \ell'_{h+1} \le Q_h^\pi(s,s') \le \overline{\rho} + u'_{h+1} = u_h.$$

Furthermore,

$$V_h^\pi(s) = \mathbb{E}_{s'\sim\pi_h(\cdot|s)}\left[Q_h^\pi(s,s') - \log \pi_h(s' \mid s)\right] \tag{89}$$

$$= \mathbb{E}_{s'\sim\pi_h(\cdot|s)}\left[Q_h^\pi(s,s')\right] + \mathcal{H}(\pi_h(\cdot \mid s)), \tag{90}$$

where $\mathcal{H}$ denotes the policy entropy. Since

$$\ell_h \le Q_h^\pi(s,s') \le u_h$$

and

$$0 \le \mathcal{H}(\pi_h(\cdot \mid s)) \le \log|\mathcal{A}(s)| \le \log a_{\max},$$

it follows that

$$\ell'_h = \ell_h \le V_h^\pi(s) \le u_h + \log a_{\max} = u'_h.$$

Thus, all the stated bounds hold at stage h, completing the backward induction. □

B.1 First-order Confidence Analysis

Lemma B.2 (self-normalized norm bound)**.** *Under the assumption* $\|\phi_h(s,s')\|_2 \le 1$ *for* $(s,s') \in \mathcal{S}\times\mathcal{S}$, $\lambda > 0$ *and* $\Lambda_h^1 = \lambda I_d$,

$$\sum_{k=1}^K \left\|\phi_h(s_h^k, s_{h+1}^k)\right\|_{(\Lambda_h^k)^{-1}} \le \sqrt{K\left(1+\frac{1}{\lambda}\right)\log\frac{\det\Lambda_h^{K+1}}{\det\Lambda_h^1}}. \tag{91}$$

Proof. By definition, $\Lambda_h^{k+1} = \Lambda_h^k + \phi_h(s_h^k, s_{h+1}^k)\phi_h(s_h^k, s_{h+1}^k)^\top$. The matrix determinant lemma for invertible matrices gives

$$\det\Lambda_h^{k+1} = \det\Lambda_h^k\left(1 + \left\|\phi_h(s_h^k, s_{h+1}^k)\right\|^2_{(\Lambda_h^k)^{-1}}\right). \tag{92}$$

Taking logs and summing,

$$\log\frac{\det\Lambda_h^{K+1}}{\det\Lambda_h^1} = \sum_{k=1}^K \log\left(1 + \left\|\phi_h(s_h^k, s_{h+1}^k)\right\|^2_{(\Lambda_h^k)^{-1}}\right). \tag{93}$$

Since $(\Lambda_h^k)^{-1} \preceq \frac{I}{\lambda}$ and $\|\phi_h(s_h^k, s_{h+1}^k)\|_2 \le 1$, we have $\|\phi_h\|^2_{(\Lambda_h^k)^{-1}} \le \frac{\|\phi_h\|_2^2}{\lambda} \le \frac{1}{\lambda}$. For $x \ge 0$, $\frac{x}{1+x} \le \log(1+x)$, so

$$\begin{aligned}
\sum_{k=1}^K \left\|\phi_h(s_h^k, s_{h+1}^k)\right\|^2_{(\Lambda_h^k)^{-1}} &\le \sum_{k=1}^K \left(1 + \left\|\phi_h(s_h^k, s_{h+1}^k)\right\|^2_{(\Lambda_h^k)^{-1}}\right)\log\left(1 + \left\|\phi_h(s_h^k, s_{h+1}^k)\right\|^2_{(\Lambda_h^k)^{-1}}\right) \\
&\le \sum_{k=1}^K \left(1+\frac{1}{\lambda}\right)\log\left(1 + \left\|\phi_h(s_h^k, s_{h+1}^k)\right\|^2_{(\Lambda_h^k)^{-1}}\right) \\
&= \left(1+\frac{1}{\lambda}\right)\log\frac{\det\Lambda_h^{K+1}}{\det\Lambda_h^1}.
\end{aligned} \tag{94}$$

Further applying the Cauchy–Schwarz inequality, we arrive at (91). □

Lemma B.3 (Log-determinant bound)**.** *Under the assumption* $\|\phi(s,s')\|_2 \le 1$ *for* $(s,s') \in \mathcal{S}\times\mathcal{S}$, $\lambda > 0$ *and* $\Lambda^1 = \lambda I_d$,

$$\frac{\det(\Lambda^{K+1})}{\det(\Lambda^1)} \le \left(1 + \frac{K}{\lambda d}\right)^d. \tag{95}$$

Proof. Let eig^k denote the eigenvalues of Λ^k. Since $\Lambda^{K+1} \succ 0$, we have $\mathrm{eig}^k \succ 0$. By the arithmetic–geometric mean inequality,

$$\left(\prod_{i=1}^{d}[\mathrm{eig}^k]_i\right)^{1/d} \le \frac{1}{d}\sum_{i=1}^{d}[\mathrm{eig}^k]_i. \tag{96}$$

Therefore,

$$\det(\Lambda^{K+1}) = \prod_{i=1}^{d}[\mathrm{eig}^k]_i \le \left(\frac{1}{d}\sum_{i=1}^{d}[\mathrm{eig}^k]_i\right)^d = \left(\frac{\mathrm{tr}(\Lambda_{K+1})}{d}\right)^d, \tag{97}$$

where tr denotes the operator of matrix trace. Next, by the linearity of the trace,

$$\mathrm{tr}(\Lambda^{K+1}) = \mathrm{tr}(\lambda I_d) + \sum_{k=1}^{K}\mathrm{tr}(\phi(s^k, s'^k)\phi_h(s^k, s'^k)^\top) = \lambda d + \sum_{k=1}^{K}\|\phi_h(s^k, s'^k)\|_2^2. \tag{98}$$

Since $\|\phi(s, s')\|_2 \le 1$,

$$\mathrm{tr}(\Lambda^{K+1}) \le \lambda d + K. \tag{99}$$

Substituting (99) into (97),

$$\det(\Lambda^{K+1}) \le \left(\frac{\lambda d + K}{d}\right)^d = \left(\lambda + \frac{K}{d}\right)^d. \tag{100}$$

Since $\Lambda_1 = \lambda I_d$, we have $\det(\Lambda_1) = \lambda^d$, and

$$\frac{\det(\Lambda^{K+1})}{\det(\Lambda^1)} \le \frac{\left(\lambda + \frac{K}{d}\right)^d}{\lambda^d} = \left(1 + \frac{K}{\lambda d}\right)^d. \tag{101}$$

□

Lemma B.4 (Azuma-Hoeffding inequality (Wainwright, 2019)). *Let $\{X_t, \mathcal{F}_t\}_{t=1}^n$ be a martingale-difference sequence satisfying $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0$, and X_t lie in the interval of $[\ell_t, u_t]$ almost surely. Then, for every $x > 0$,*

$$\mathbb{P}\left(\sum_{t=1}^{n} X_t \ge x\right) \le \exp\left(-\frac{2x^2}{\sum_{t=1}^{n}(u_t - \ell_t)^2}\right). \tag{102}$$

Equivalently, with probability at least $1 - \zeta$,

$$\sum_{t=1}^{n} X_t \le \sqrt{\frac{1}{2}\left(\sum_{t=1}^{n}(u_t - \ell_t)^2\right)\log\frac{1}{\zeta}}. \tag{103}$$

B.2 Proof of Lemma 4.4

Proof. Write $\theta_h^k := \theta_h(V_{h+1}^k)$. By Assumption 4.1, for every $t < k$,

$$y_{t,h}^k = \phi_h(s_h^t, s_{h+1}^t)^\top \theta_h^k. \tag{104}$$

Therefore,

$$\Lambda_h^k w_h^k = \sum_{t<k}\phi_h(s_h^t, s_{h+1}^t)\phi_h(s_h^t, s_{h+1}^t)^\top\theta_h^k = (\Lambda_h^k - \lambda I)\theta_h^k. \tag{105}$$

Consequently,

$$w_h^k - \theta_h^k = -\lambda(\Lambda_h^k)^{-1}\theta_h^k. \tag{106}$$

Then, by Cauchy–Schwarz in the $(\Lambda_h^k)^{-1}$ norm,

$$\begin{aligned}\left|\phi_h^\top(s, s')(w_h^k - \theta_h^k)\right| &= \lambda\left|\phi_h^\top(s, s')(\Lambda_h^k)^{-1}\theta_h^k\right| \\ &\le \lambda\|\phi(s, s')\|_{(\Lambda_h^k)^{-1}}\|\theta_h^k\|_{(\Lambda_h^k)^{-1}}.\end{aligned} \tag{107}$$

Since $\Lambda_h^k \succeq \lambda I$,

$$\|\theta_h^k\|_{(\Lambda_h^k)^{-1}} \leq \frac{\|\theta_h^k\|_2}{\sqrt{\lambda}} \leq \frac{L}{\sqrt{\lambda}}. \tag{108}$$

Therefore, (107) becomes

$$\left|\widehat{V}_{h+1}^k(s,s') - V_{h+1}^k(s')\right| = \left|\phi_h^\top(s,s')(w_h^k - \theta_h^k)\right| \tag{109}$$

$$\leq L\sqrt{\lambda}\|\phi_h(s,s')\|_{(\Lambda_h^k)^{-1}} = \Gamma_h^k(s,s'). \tag{110}$$

Denote the unclipped function and the corresponding Bellman residuals as

$$\widehat{Q}_h^k(s,s') = \rho_h(s,s') + \widehat{V}_{h+1}^k(s,s') + \Gamma_h^k(s,s') \tag{111}$$

$$\hat{\delta}_h^k(s,s') = \widehat{Q}_h^k(s,s') - \rho_h(s,s') - V_{h+1}^k(s') \tag{112}$$

According to (110),

$$0 \leq \widehat{V}_{h+1}^k(s,s') + \Gamma_h^k(s,s') - V_{h+1}^k(s') \leq 2\Gamma_h^k(s,s'). \tag{113}$$

Therefore,

$$0 \leq \hat{\delta}_h^k(s,s') \leq 2\Gamma_h^k(s,s'). \tag{114}$$

Because $\rho_h(s,s') + V_{h+1}^k(s') \in [\ell_h, u_h]$, projecting $\widetilde{Q}_h^k(s,s')$ onto $[\ell_h, u_h]$ cannot move it below $\rho_h(s,s') + V_{h+1}^k(s') \in [\ell_h, u_h]$. Projection also cannot increase its distance above $\rho_h(s,s') + V_{h+1}^k(s') \in [\ell_h, u_h]$. Hence,

$$0 \leq \delta_h^k(s,s') \leq 2\Gamma_h^k(s,s'). \tag{115}$$

□

B.3 Proof of Lemma 4.5

Proof. We prove (27) by backward induction. At the terminal stage,

$$V_H^k(x) = \log R(x) = V_H^\star(x).$$

Suppose that $V_{h+1}^k(s') \geq V_{h+1}^\star(s')$ for every $s' \in \mathcal{S}_{h+1}$. By Lemma B.1, both Q_h^k and $\rho_h + V_{h+1}^k$ lie in $[\ell_h, u_h]$. Therefore, by the assumption $\delta_h^k \geq 0$,

$$Q_h^k(s,s') \geq \rho_h(s,s') + V_{h+1}^k(s') \geq \rho_h(s,s') + V_{h+1}^\star(s') = Q_h^\star(s,s').$$

Because log-sum-exp is element-wise monotone,

$$V_h^k(s) = \log \sum_a \exp\{Q_h^k(s,s')\} \geq \log \sum_a \exp\{Q_h^\star(s,s')\} = V_h^\star(s). \tag{116}$$

This completes the backward induction. □

B.4 Second-order Confidence Analysis

Lemma B.5 (Squared-residual bound)**.** *Suppose $\|\phi_h(s,s')\|_2 \leq 1$ and choose $\lambda = \frac{g^2}{4L^2}$. Then, for every episode $k \leq K$ and $h \in [H]_0$,*

$$\sum_{k=1}^K \delta_h^k(s_h^k, s_{h+1}^k)^2 \leq 2g^2 d \log\left(1 + \frac{4L^2K}{g^2 d}\right). \tag{117}$$

Proof. Since both the clipped optimistic backup $Q_h^k(s,s')$ and the exact backup $\rho_h(s,s') + V_{h+1}^k(s')$ belong to $[\ell_h, u_h]$ at each stage h, we have:

$$0 \leq \delta_h^k(s,s') \leq u_h - \ell_h \leq g. \tag{118}$$

Furthermore, using Lemma 4.4 ,

$$\delta_h^k(s_h^k, s_{h+1}^k)^2 \le \min\left\{4\lambda L^2 \|\phi_h(s_h^k, s_{h+1}^k)\|^2_{(\Lambda_h^k)^{-1}}, g^2\right\}. \tag{119}$$

Under the choice $\lambda = \frac{g^2}{4L^2}$, $4\lambda L^2 = g^2$ and

$$\delta_h^k(s_h^k, s_{h+1}^k)^2 \le g^2 \min\{1, \|\phi_h(s_h^k, s_{h+1}^k)\|^2_{(\Lambda_h^k)^{-1}}.\}. \tag{120}$$

For every $x \ge 0$, $\min\{1, x\} \le 2\log(1+x)$.

$$\sum_{k=1}^K \min\{1, \|\phi_h(s_h^k, s_{h+1}^k)\|^2_{(\Lambda_h^k)^{-1}}\} \le 2\sum_{k=1}^K \log(1 + \|\phi_h(s_h^k, s_{h+1}^k)\|^2_{(\Lambda_h^k)^{-1}}) \tag{121}$$

$$= 2\log\frac{\det(\Lambda_h^{K+1})}{\det(\lambda I)} \tag{122}$$

$$\le 2d\log\left(1 + \frac{K}{\lambda d}\right), \tag{123}$$

where the equality holds by (93) in Lemma B.2 and the last inequality holds by Lemma B.3. Combining (123), and (119),

$$\sum_{k=1}^K \delta_h^k(s_h^k, s_{h+1}^k)^2 \le 2g^2 d\log\left(1 + \frac{K}{\lambda d}\right). \tag{124}$$

Finally, substituting $\lambda = \frac{g^2}{4L^2}$ gives (117). □

We use the following multiplicative concentration lemma that plays the role that Azuma–Hoeffding played in the first-order proof, but it exploits the fact that the random quantity being concentrated is nonnegative.

Lemma B.6. *Let $\{X_t, \mathcal{F}_t\}_{t=1}^n$ be an adapted sequence satisfying $0 \le X_t \le \beta$ almost surely. Define $\mu_t := \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$. Then, for every $\eta > 0$, with probability at least $1-\zeta$, simultaneously for all $n \ge 1$,*

$$\sum_{t=1}^n \mu_t \le \frac{\eta}{1-e^{-\eta}} \sum_{t=1}^n X_t + \frac{\beta}{1-e^{-\eta}} \log\frac{1}{\zeta}. \tag{125}$$

Proof. Set $\bar{X}_t := \frac{X_t}{\beta} \in [0,1]$, and $\bar{\mu}_t := \mathbb{E}[\bar{X}_t \mid \mathcal{F}_{t-1}] = \frac{\mu_t}{\beta}$. Since $e^{-\eta x}$ is convex, we have

$$e^{-\eta x} \le (1-x)e^0 + xe^{-\eta} = 1 - (1-e^{-\eta})x. \tag{126}$$

Accordingly, let $c_\eta := 1 - e^{-\eta}$ and use $1 - x \le e^{-x}$

$$\mathbb{E}[e^{-\eta \bar{X}_t} \mid \mathcal{F}_{t-1}] \le 1 - c_\eta \bar{\mu}_t \le e^{-c_\eta \bar{\mu}_t}. \tag{127}$$

Define $M_n := \exp\left\{\sum_{t=1}^n \left(c_\eta \bar{\mu}_t - \eta \bar{X}_t\right)\right\}$ with $M_0 = \exp(0) = 1$. Then,

$$M_n = M_{n-1} e^{c_\eta \bar{\mu}_n - \eta \bar{X}_n} \tag{128}$$

$$\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] = M_{n-1}\mathbb{E}[e^{c_\eta \bar{\mu}_n - \eta \bar{X}_n} \mid \mathcal{F}_{n-1}] \le M_{n-1}, \tag{129}$$

means that the adapted sequence $\{M_n : n \ge 0\}$ is a supermartingale. For some $\zeta \in (0,1)$, Ville's inequality (Wang & Ramdas, 2025) gives:

$$\Pr\left(\sup_{n\ge 0} M_n \ge \frac{1}{\zeta}\right) = \Pr\left(\sup_{n\ge 1} M_n \ge \frac{1}{\zeta}\right) \le \zeta. \tag{130}$$

Thus, with probability at least $1-\zeta$,

$$c_\eta \sum_{t=1}^n \bar{\mu}_t - \eta \sum_{t=1}^n \bar{X}_t \le \log\frac{1}{\zeta}.$$

Multiplication by β/c_η proves (125). □

C Regret decompositions and Proofs

Lemma C.1 (Current-policy Bellman telescope)**.** *For every episode k,*

$$V_0^k(s_0) - V_0^{\pi^k}(s_0) = \mathbb{E}_{\tau\sim P_{\pi^k}}\left[\sum_{h=0}^{H-1} \delta_h^k(s_h, s_{h+1})\right]. \tag{131}$$

Proof. First of all,

$$\widetilde{\Delta}_h^k(s) = \mathbb{E}_{s'\sim\pi^k(\cdot|s)}[\widetilde{\nabla}_h^k(s,s')]. \tag{132}$$

By the Bellman equations of π^k,

$$\widetilde{\nabla}_h^k(s,s') = Q_h^k(s,s') - \rho_h(s,s') - V_{h+1}^{\pi^k}(s').$$

Using the definition of Bellman residuals,

$$\begin{aligned}\widetilde{\nabla}_h^k(s) &= Q_h^k(s,s') - \rho_h(s,s') - V_{h+1}^k(s') + V_{h+1}^k(s') - V_{h+1}^{\pi^k}(s') \\ &= \delta_h^k(s,s') + \widetilde{\Delta}_{h+1}^k(s').\end{aligned} \tag{133}$$

Substituting (133) into (132), we have

$$\widetilde{\Delta}_h^k(s) = \mathbb{E}_{s'\sim\pi_h^k(\cdot|s)}[\delta_h^k(s,s') + \widetilde{\Delta}_{h+1}^k(s')].$$

Recursively expanding the above equation from $h = 0, \ldots, H-1$ yields

$$\widetilde{\Delta}_0^k(s_0) = \mathbb{E}_{P_{\pi^k}}\left[\sum_{h=0}^{H-1} \delta_h^k(s_h^k, s_{h+1}^k) + \widetilde{\Delta}_H^k(s_H^k)\right]. \tag{134}$$

Since $\widetilde{\Delta}_H^k = \log R - \log R = 0$, we arrive at (131). □

Corollary C.2 (Realized current-policy Bellman telescope)**.** *Fix an episode k,*

$$V_0^k(s_0) - V_0^{\pi^k}(s_0) = \sum_{h=0}^{H-1} \delta_h^k(s_h^k, s_{h+1}^k) + \sum_{h=0}^{H-1} M_h^k. \tag{135}$$

Proof. We first relate the state-value error $\widetilde{\Delta}_h^k$ to the action-value error E_h^k. By the definitions of V_h^k and $V_h^{\pi^k}$,

$$\widetilde{\Delta}_h^k(s) = \mathbb{E}_{s'\sim\pi_h^k(\cdot|s)}\left[Q_h^k(s,s') - Q_h^{\pi^k}(s,s')\right] = \mathbb{E}_{s'\sim\pi_h^k(\cdot|s)}\left[\widetilde{\nabla}_h^k(s,s')\right]. \tag{136}$$

Using the definition of M_h^k,

$$\widetilde{\Delta}_h^k(s_h^k) = \widetilde{\nabla}_h^k(s_h^k, s_{h+1}^k) + M_h^k. \tag{137}$$

Using (133) in the proof of Lemma C.1, at the realized state s_h^k,

$$\widetilde{\Delta}_h^k(s_h^k) = \delta_h^k(s_h^k, s_{h+1}^k) + \widetilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) + M_h^k. \tag{138}$$

Summing (138) over $h = 0, \ldots, H-1$ yields

$$\widetilde{\Delta}_0^k(s_0) = \sum_{h=0}^{H-1} \delta_h^k(s_h^k, s_{h+1}^k) + \sum_{h=0}^{H-1} M_h^k + \widetilde{\Delta}_H^k(s_H^k). \tag{139}$$

Since $\widetilde{\Delta}_H^k(s_H^k) = V_H^k(s_H^k) - V_H^{\pi^k}(s_H^k) = 0$, we arrive at (135). □

Lemma C.3 (Comparator Bellman telescope)**.** *For every episode k,*

$$V_0^\star(s_0) - V_0^k(s_0) = \mathbb{E}_{\tau\sim P_{\pi^\star}}\left[\sum_{h=0}^{H-1} -D_{\mathrm{KL}}\left(\pi_h^\star(\cdot \mid s_h) \| \pi_h^k(\cdot \mid s_h)\right) - \delta_h^k(s_h, s_{h+1})\right]. \tag{140}$$

Proof. By the definitions δ_h^k and Δ_h^k , the optimal value can be written as

$$V_h^\star(s) = \mathbb{E}_{s'\sim\pi_h^\star(\cdot|s)}\left[\rho_h(s,s') + V_{h+1}^\star(s') - \log\pi_h^\star(s' \mid s)\right] \tag{141}$$

$$= \mathbb{E}_{s'\sim\pi_h^\star(\cdot|s)}\Big[Q_h^k(s,s') - \delta_h^k(s,s') + \Delta_{h+1}^k(s') - \log\pi_h^\star(s' \mid s)\Big]. \tag{142}$$

On the other hand, by the definition of the optimistic value,

$$V_h^k(s) = \mathbb{E}_{s'\sim\pi_h^k(\cdot|s)}\left[Q_h^k(s,s') - \log\pi_h^k(s' \mid s)\right]. \tag{143}$$

Therefore,

$$\begin{aligned}\Delta_h^k(s) =&\mathbb{E}_{s'\sim\pi_h^\star(\cdot|s)}\left[Q_h^k(s,s') - \log\pi_h^\star(s' \mid s)\right] - \mathbb{E}_{s'\sim\pi_h^k(\cdot|s)}\left[Q_h^k(s,s') - \log\pi_h^k(s' \mid s)\right]\\ &+ \mathbb{E}_{s'\sim\pi_h^\star(\cdot|s)}\left[-\delta_h^k(s,s') + \Delta_{h+1}^k(s')\right],\end{aligned} \tag{144}$$

where the conditioning on s is suppressed to lighten notation. Adding and subtracting $\mathbb{E}_{\pi_h^\star}\left[\log\pi_h^k(s' \mid s)\right]$, we have

$$\begin{aligned}\Delta_h^k(s) =&\mathbb{E}_{\pi_h^\star}\left[Q_h^k - \log\pi_h^k\right] - \mathbb{E}_{\pi_h^k}\left[Q_h^k - \log\pi_h^k\right] - D_{\mathrm{KL}}\left(\pi_h^\star(\cdot \mid s)\|\pi_h^k(\cdot \mid s)\right)\\ &+ \mathbb{E}_{\pi_h^\star}\left[-\delta_h^k(s,s') + \Delta_{h+1}^k(s')\right] &(145)\\ =&- D_{\mathrm{KL}}\left(\pi_h^\star(\cdot \mid s)\|\pi_h^k(\cdot \mid s)\right) + \mathbb{E}_{\pi_h^\star}\left[-\delta_h^k(s,s') + \Delta_{h+1}^k(s')\right] &(146)\end{aligned}$$

where the second equation holds in that the first two terms on the right-hand side of the first equation equals

$$\mathbb{E}_{s'\sim\pi_h^\star(\cdot|s)}\left[V_h^k(s)\right] - \mathbb{E}_{s'\sim\pi_h^\star(\cdot|s)}\left[V_h^k(s)\right] = 0.$$

Iterating (146) from $h = 0$ to $H - 1$ and using the fact that $\Delta_H^k(s_H) = V_H^\star(s_H) - V_H^k(s_H) = 0$, we arrive at (140). □

C.1 Proof of Theorem 4.2

Proof. Using Proposition 4.7, we have

$$\mathbb{E}_{\tau^{1:K}\sim P_{\mathbb{A}}}[\mathrm{Reg}(K)] \le 2\sum_{k=1}^K \mathbb{E}_{\tau^{1:k-1}}\mathbb{E}_{\tau\sim P_{\pi^k}|\mathcal{F}_{k-1}}\left[\sum_{h=0}^{H-1}\Gamma_h^k(s_h,s_{h+1})\right]. \tag{147}$$

Using Lemma B.2 and B.3 for $\sum_{k=1}^K \Gamma_h^k(s_h^k, s_{h+1}^k)$, we have

$$\sum_{k=1}^K \Gamma_h^k(s_h^k,s_{h+1}^k) \le L\sqrt{K\,(1+\lambda)\log\frac{\det\Lambda_h^{K+1}}{\det\Lambda_h^1}} \le L\sqrt{K\,(1+\lambda)\,d\log\left(1+\frac{K}{\lambda d}\right)}. \tag{148}$$

Inserting (148) into (147) gives

$$\mathbb{E}_{\tau^{1:K}\sim P_{\mathbb{A}}}[\mathrm{Reg}(K)] \le 2HL\sqrt{K\,(1+\lambda)\,d\log\left(1+\frac{K}{\lambda d}\right)} \tag{149}$$

We now prove the high-probability upper bound. Using Proposition 4.7,

$$\mathrm{Reg}(K) \le \sum_{k=1}^K\sum_{h=0}^{H-1} 2\Gamma_h^k(s_h^k,s_{h+1}^k) + M_h^k. \tag{150}$$

For each martingale difference, by Lemma 4.5,

$$\widetilde{\nabla}_h^k(s,s') := Q_h^k(s,s') - Q_h^{\pi^k}(s,s') \ge Q_h^\star(s,s') - Q_h^{\pi^k}(s,s') \ge 0. \tag{151}$$

Moreover, both $Q_h^k(s,s')$ and $Q_h^{\pi^k}(s,s')$ lie in the valid stage-h interval $[\ell_h, u_h]$. Hence

$$0 \le \widetilde{\nabla}_h^k(s,s') \le u_h - \ell_h \le g. \tag{152}$$

Conditional on $\mathcal{F}_{k,h}$, $\widetilde{\Delta}_h^k(s_h^k)$ is the conditional mean of $\widetilde{\nabla}_h^k(s_h^k, s_{h+1}^k)$. Thus, M_h^k lies almost surely in an interval of length at most $u_h - \ell_h \le g$. Applying Lemma B.4 to the chronologically ordered martingale-difference sequence $\{M_h^k\}_{k\in[K],\, h\in[H]_0}$ gives, with probability at least $1-\zeta$,

$$\sum_{k=1}^{K}\sum_{h=0}^{H-1} M_h^k \le \sqrt{\frac{1}{2}\left(\sum_{k=1}^{K}\sum_{h=0}^{H-1}(u_h-\ell_h)^2\right)\log\frac{1}{\zeta}} \le g\sqrt{\frac{KH}{2}\log\frac{1}{\zeta}}. \tag{153}$$

Inserting (148) and (153) into (150), we have a probability of at least $1-\zeta$,

$$\mathrm{Reg}(K) \le 2LH\sqrt{K\,(1+\lambda)\,d\log\left(1+\frac{K}{\lambda d}\right)} + g\sqrt{\frac{1}{2}KH\log\frac{1}{\zeta}} \tag{154}$$

□

D LOGARITHMIC REGRET DECOMPOSITION AND PROOF

D.1 PROOF OF LEMMA 4.8

Proof. By Lemma C.1, C.3 and the definition of D_k, we have

$$V^\star(s_0) - V^{\pi^k}(s_0) = \mathbb{E}_{P_{\pi^k}}[D_k] - \mathbb{E}_{P_{\pi^\star}}[D_k] - \sum_{h=0}^{H-1}\mathbb{E}_{\tau\sim P_{\pi^\star}}\left[D_{\mathrm{KL}}\left(\pi_h^\star(\cdot\mid s_h)\,\|\pi_h^k(\cdot\mid s_h)\right)\right] \tag{155}$$

$$= \mathbb{E}_{P_{\pi^k}}[D_k] - \mathbb{E}_{P_{\pi^\star}}[D_k] - D_{\mathrm{KL}}\left(P_{\pi^\star}\,\|P_{\pi^k}\right). \tag{156}$$

We now use the variational inequality (Hiriart-Urruty & Lemaréchal, 2004)

$$\mathbb{E}_P[f] - D_{\mathrm{KL}}(P\|Q) \le \log\mathbb{E}_Q[e^f],, \tag{157}$$

valid for any distributions (P, Q) with P absolutely continuous with respect to Q. Take

$$P = P_{\pi^\star}, \qquad Q = P_{\pi^k}, \qquad f = -D_k. \tag{158}$$

Then, (157) implies

$$-\mathbb{E}_{P_{\pi^\star}}[D_k] - D_{\mathrm{KL}}\left(P_{\pi^\star}\|P_{\pi^k}\right) \le \log\mathbb{E}_{P_{\pi^k}}\left[e^{-D_k}\right].$$

Substituting this into (156),

$$V^\star(s_0) - V^{\pi^k}(s_0) \le \mathbb{E}_{P_{\pi^k}}[D_k] + \log\mathbb{E}_{P_{\pi^k}}\left[e^{-D_k}\right]. \tag{159}$$

By Lemma 4.4, every Bellman residual is nonnegative, so $D_k \ge 0$. For every $x \ge 0$, $e^{-x} \le 1 - x + \frac{x^2}{2}$. Therefore,

$$\mathbb{E}_{P_{\pi^k}}\left[e^{-D_k}\right] \le 1 - \mathbb{E}_{P_{\pi^k}}[D_k] + \frac{1}{2}\mathbb{E}_{P_{\pi^k}}[D_k^2]. \tag{160}$$

Using $\log x \le x - 1$,

$$\log\mathbb{E}_{P_{\pi^k}}\left[e^{-D_k}\right] \le -\mathbb{E}_{P_{\pi^k}}[D_k] + \frac{\mathbb{E}_{P_{\pi^k}}[D_k^2]}{2}. \tag{161}$$

Substitution into (159) cancels the first-order term and proves the first inequality in (32). Finally,

$$\left(\sum_{h=0}^{H-1}\delta_h^k\right)^2 \le H\sum_{h=0}^{H-1}(\delta_h^k)^2 \tag{162}$$

by Cauchy-Schwarz inequality. Taking expectation proves the second inequality in (32). □

D.2 Proof of Theorem 4.3

Proof. Using Lemma 4.8, summing $V^\star(s_0) - V^{\pi^k}(s_0)$ over $k = 1, \ldots, K$ and taking expectation,

$$\mathbb{E}_{\tau^{1:K}\sim P_{\mathbb{A}}}[\mathrm{Reg}(K)] \leq \frac{H}{2}\sum_{k=1}^{K}\mathbb{E}_{\tau^{1:k-1}}\mathbb{E}_{\tau\sim P_{\pi^k}|\mathcal{F}_{k-1}}\left[\sum_{h=0}^{H-1}\delta_h^k(s_h^k, s_{h+1}^k)^2\right]. \tag{163}$$

Further using Lemma B.5, we have

$$\mathbb{E}_{\tau^{1:K}\sim P_{\mathbb{A}}}[\mathrm{Reg}(K)] \leq g^2H^2d\log\left(1+\frac{4L^2K}{g^2d}\right). \tag{164}$$

Now, we prove the high-probability bound. Using Lemma 4.8, and summing $V^\star(s_0) - V^{\pi^k}(s_0)$ over $k = 1, \ldots, K$,

$$\mathrm{Reg}(K) \leq \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}_{\tau^k\sim P_{\pi^k}}\left[D_k(\tau^k)^2\right]. \tag{165}$$

By Cauchy-Schwarz inequality,

$$(D_k)^2 = \left(\sum_{h=0}^{H-1}\delta_h^k\right)^2 \leq H\sum_{h=0}^{H-1}(\delta_h^k)^2. \tag{166}$$

It is shown in 118 that $\delta_h^k \leq g$, thereby

$$(D_k)^2 \leq H\sum_{h=0}^{H-1}g^2 = H^2g^2. \tag{167}$$

Apply Lemma B.6 and for $(D_k)^2 \leq H^2g^2$, and $\eta = 1$. With probability at least $1-\zeta$,

$$\sum_{k=1}^{K}\mathbb{E}_{\tau^k\sim P_{\pi^k}}\left[D_k(\tau^k)^2\right] \leq \frac{1}{1-e^{-1}}\left[\sum_{k=1}^{K}D_k(\tau^k)^2 + H^2g^2\log\frac{1}{\zeta}\right]. \tag{168}$$

Using Lemma B.5 and (166), we have with probability at least $1-\zeta$,

$$\sum_{k=1}^{K}\mu_k \leq \frac{H^2g^2}{1-e^{-1}}\left[2d\log\left(1+\frac{4L^2K}{g^2d}\right)+\log\frac{1}{\zeta}\right].$$

Substitution into (165) yields (21). □