Geometric Active Exploration in Markov Decision Processes: the Benefit of Abstraction

Riccardo De Santi [1 2] **Federico Arangath Joseph** [* 1] **Noah Liniger** [* 1] **Mirco Mutti** [3] **Andreas Krause** [1]

Abstract

How can a scientist use a Reinforcement Learning (RL) algorithm to design experiments over a dynamical system's state space? In the case of finite and Markovian systems, an area called *Active Exploration* (AE) relaxes the optimization problem of experiments design into Convex RL, a generalization of RL admitting a wider notion of reward. Unfortunately, this framework is currently not scalable and the potential of AE is hindered by the vastness of experiment spaces typical of scientific discovery applications. However, these spaces are often endowed with natural geometries, e.g., permutation invariance in molecular design, that an agent could leverage to improve the statistical and computational efficiency of AE. To achieve this, we bridge AE and MDP homomorphisms, which offer a way to exploit known geometric structures via abstraction. Towards this goal, we make two fundamental contributions: we extend MDP homomorphisms formalism to Convex RL, and we present, to the best of our knowledge, the first analysis that formally captures the benefit of abstraction via homomorphisms on sample efficiency. Ultimately, we propose the Geometric Active Exploration (GAE) algorithm, which we analyse theoretically and experimentally in environments motivated by problems in scientific discovery.

1. Introduction

The problem of optimal experimental design (OED) (Chaloner & Verdinelli, 1995) refers to the task of optimally selecting experiments to minimize a measure of uncertainty of an unknown *quantity of interest* $f : \mathcal{S} \to \mathbb{R}$, where $\mathcal{S}$ denotes a space of experiments. Typically, the problem considers a limited budget of resources, e.g., number of experiments, and assumes the possibility to directly sample f at arbitrary inputs $s \in \mathcal{S}$. Conceptually, an optimal design can be interpreted as a distribution over experiments determining the probability with which these should be carried out in order to minimize the uncertainty of f (Pukelsheim, 2006).

Interestingly, in a wide variety of applications the input space $\mathcal{S}$ corresponds to the state space of a dynamical system (Mutny et al., 2023). Therefore, the agent carrying out the experiments must respect the underlying dynamics and cannot teleport from any experiment, now interpreted as a state $s \in \mathcal{S}$, to any other experiment. For instance, consider the environmental sensing problem illustrated in Figure 1, where an agent aims to actively estimate the amount of air pollution caused by the diffusion of a chemical substance released from a point source. To address this problem, the agent chooses sampling policies to minimize an estimation error of the amount of pollutant f.

In the case of time-discrete and Markovian dynamical systems, this problem is known as Active Exploration (AE) (Mutny et al., 2023; Tarbouriech & Lazaric, 2019; Tarbouriech et al., 2020). AE frames the experiments design task as an instance of Convex Reinforcement Learning (Convex RL) (Hazan et al., 2019; Zahavy et al., 2021), a recent generalization of RL where the agent aims to minimize a convex functional of the state-action distribution induced by a policy interacting with the environment.

The AE formulation of the OED problem on dynamical systems is promising as it allows to learn (from data) optimal sampling policies that respect the system dynamics while minimizing a measure of uncertainty of f. Nonetheless, solving an instance of Convex RL typically entails solving a sequence of Markov Decision Processes (MDPs) and estimating the visitation density at each iteration (Hazan et al., 2019). As a consequence, current Active Exploration methods are not scalable, hindering their use in real-world scientific discovery problems where experiment spaces are generally immense (Wang et al., 2023; Thiede et al., 2022).

Luckily, these spaces are often endowed with natural geometries, as in the case of permutation invariances in molecular

[*]Equal contribution [1]Department of Computer Science, ETH Zurich, Zurich, Switzerland [2]ETH AI Center, Zurich, Switzerland [3]Technion, Haifa, Israel. Correspondence to: Riccardo De Santi <rdesanti@ethz.ch>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

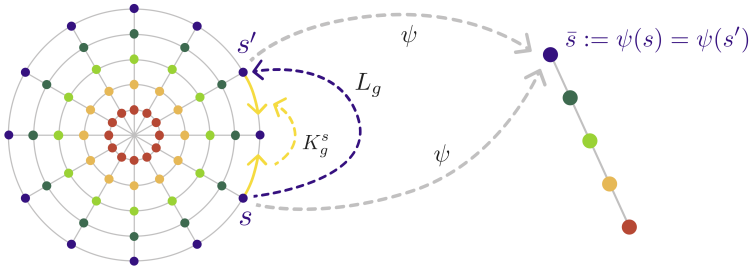


Figure 1. Radial diffusion process of a pollutant from a central point source. On the left, original MDP where each circle is an f-equivalence class, L_g denotes a state symmetry acting on f, K_g^s denotes a state-dependent action symmetry acting on P. On the right, the abstract MDP obtained via the MDP homomorphism $h = (\psi, \{\phi_s \mid s \in \mathcal{S}\})$, where ψ maps f-equivalence classes to abstract states.

design (Cheng et al., 2021; Elton et al., 2019), or roto-translation and reflection invariances in environmental sensing (Krause, 2008; van der Pol et al., 2020b). Let us take the example reported in Figure 1. One can expect locations that are equally distant from the point source (center) to have almost identical amounts of pollutant i.e., the quantity of interest f follows radial symmetries.[1] Therefore, by sampling f in one location, the agent gains information on f in all the other symmetric locations as well. As a consequence, in this work we first aim to answer the following question:

How can a RL agent exploit geometric structure to increase the statistical and computational efficiency of AE?

First, we introduce a novel geometric estimation error and corresponding AE objective (Sec. 3 and 4). Then, we find optimal sampling strategies for the introduced AE objective by bridging Active Exploration with the area of MDP homomorphisms (Ravindran & Barto, 2001; van der Pol et al., 2020b), which offers an algorithmic scheme to leverage known geometric structure in RL via abstraction. Unfortunately, MDP homomorphisms are not directly usable in AE as it is not a classic RL problem. Thus, we extend MDP homomorphisms to Convex RL and introduce **G**eometric **A**ctive **E**xploration (GAE), an algorithm that solves the AE problem by exploiting known geometric structure via abstraction (Sec. 5). To the best of our knowledge, we provide the first analysis that formally captures the benefit of abstraction on sample efficiency via MDP homomorphisms (Sec. 6). Finally, we showcase experimentally the statistical and computational advantages of GAE in illustrative environments inspired by scientific discovery problems (Sec. 7).

To sum up, we make the following contributions:

- An Active Exploration objective that leverages known invariances of the quantity of interest f (Sec. 3 and 4).
- **G**eometric **A**ctive **E**xploration (GAE), an algorithm that extends MDP homomorphism to Convex RL and solves AE via abstraction (Sec. 5).
- The first analysis that sheds light on the benefits of abstraction on sample efficiency via MDP homomorphisms, here specialized for the AE problem (Sec. 6).
- An experimental evaluation of the performance of GAE against a classic AE algorithm (Sec. 7).

Our analysis capturing the benefit of geometric structure on sample efficiency may be of independent interest for RL and Convex RL.

[1] Note that in the paper we conceptualize symmetries to be *exact* while they may be *approximate* in practice. Extending this work to deal with approximate symmetries is a nice direction for future works.

2. Background and Notation

Let X be a set, we denote with $\Delta(X)$ the probability simplex over X. We define $[N] \coloneqq \{1, \dots, N\}$, and given a number x, we denote $x^+ \coloneqq \max\{1, x\}$.

2.1. Markovian Processes and Active Exploration

In the following, we briefly introduce basic RL notions and the Active Exploration (AE) problem.

Discrete Markovian Processes. A (discrete) Controlled Markov Process (CMP) is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, \mu)$, where $\mathcal{S}$ is a finite state space ($|\mathcal{S}| = S$), $\mathcal{A}$ is a finite action space ($|\mathcal{A}| = A$), $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition model, such that $P(s'|s, a)$ denotes the conditional probability of reaching $s' \in \mathcal{S}$ when selecting $a \in \mathcal{A}$ in $s \in \mathcal{S}$, and $\mu : \Delta(\mathcal{S})$ is the initial state distribution. A CMP $\mathcal{M}$ paired with a function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, i.e., $\mathcal{M}_r \coloneqq \mathcal{M} \cup r$, is a Markov Decision Process (MDP) (Puterman, 2014).

An agent interacting with a CMP starts from an initial state $s_0 \sim \mu$. Then, at every time-step t, the agent takes an action a_t, collects a reward $r(s_t, a_t)$ (when defined) and transitions to $s_{t+1} \sim P(\cdot|s_t, a_t)$. The agent's actions are sampled from a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ such that $\pi(a|s)$ denotes the conditional probability of a in s.

Active Exploration. In the Active Exploration problem, an agent interacts with a MDP $\mathcal{M}_f = (\mathcal{S}, \mathcal{A}, P, \mu, f)$ where f is an unknown and deterministic *quantity of interest*

providing a noisy observation $x \sim y(s) = f(s) + \nu(s)$ at state s. Here, ν is a distribution with zero mean and unknown heteroscedastic variance $\sigma^2(s) \in [0, \sigma^2_{\max}]$ and $x \in [0, F_{\max}]$. In AE, the agent aims to learn a policy to minimize a measure of uncertainty over f through interactions with $\mathcal{M}_f$. Notice that, as a sub-case of AE, f can be interpreted as a reward function that an agent wishes to estimate rather than maximize (Lindner et al., 2022).

2.2. Invariances and MDP Homomorphisms

In the following, we introduce basic concepts from abstract algebra and the notion of MDP homomorphism.

Equivalence, Invariance, and Symmetries. When a function $f : \mathcal{X} \to \mathcal{Y}$ maps two inputs x, x' to the same value $f(x) = f(x')$, we say that x and x' are f-equivalent. The set $[x]$ of all points f-equivalent to x is called *equivalence class* of x. We say that f is invariant across $[x]$. Consider a transformation operator $L_g : \mathcal{X} \to \mathcal{X}$, where $\mathbb{G} = (\{L_g\}_{g \in G}, \cdot)$ is a group, G is an index set, and $\cdot$ denotes composition. Given a function $f : \mathcal{X} \to \mathcal{Y}$, if L_g satisfies:

$$f(x) = f(L_g[x]) \qquad \forall g \in G, \forall x \in \mathcal{X} \tag{1}$$

then we say that f is invariant to L_g, and we call $\{L_g\}_{g \in G}$ a set of *symmetries* of f.

MDP Homomorphisms. An *MDP homomorphism* h is a mapping from an *original* MDP $\mathcal{M}_r = (\mathcal{S}, \mathcal{A}, P, \mu, r)$ to an *abstract* MDP $\overline{\mathcal{M}}_r = (\overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{P}, \overline{\mu}, \overline{r})$ defined by a surjective map $h : \mathcal{S} \times \mathcal{A} \to \overline{\mathcal{S}} \times \overline{\mathcal{A}}$. In particular, h is composed of a tuple $(\psi, \{\phi_s \mid s \in \mathcal{S}\})$, where $\psi : \mathcal{S} \to \overline{\mathcal{S}}$ is the state map and $\phi_s : \mathcal{A} \to \overline{\mathcal{A}}$ is the state-dependent action map. These maps are built to satisfy the conditions:

$$\overline{r}(\psi(s), \phi_s(a)) = r(s, a) \tag{2}$$

$$\overline{P}(\psi(s') \mid \psi(s), \phi_s(a)) = \sum_{s'' \in [s']} P(s'' \mid s, a) \tag{3}$$

for all $s, s' \in \mathcal{S}, a \in \mathcal{A}$. Moreover, given a state s such that $\overline{s} \coloneqq \psi(s)$, we denote the equivalence class of s (and $\overline{s}$) induced by ψ as $[s] = [\overline{s}] \coloneqq \{s' \in \mathcal{S} : \psi(s') = \psi(s)\}$ and indicate with $E_s \coloneqq E_{\overline{s}} = |[s]|$ its cardinality.

Policy Lifting. Given a MDP homomorphism h, the Optimal Value Equivalence Theorem by Ravindran & Barto (2001) states that an optimal policy $\overline{\pi}$ for the abstract MDP can be transformed to an optimal policy π for the original MDP via the *lifting* operation:

$$\pi(a|s) \coloneqq \frac{\overline{\pi}(\overline{a}|\psi(s))}{|\{a \in \phi_s^{-1}(\overline{a})\}|} \qquad \forall s \in \mathcal{S}, a \in \phi_s^{-1}(\overline{a}) \tag{4}$$

3. Problem Setting

Consider the AE problem in a MDP $\mathcal{M}_f = (\mathcal{S}, \mathcal{A}, P, \mu, f)$ with known dynamics P, and unknown *quantity of interest* $f : \mathcal{S} \to \mathcal{B} \subset \mathbb{R}$. Typically, an agent aims to minimize an estimation error of f of the form:

(Classic) Estimation Error

$$\xi_n = \frac{1}{S} \sum_{s \in \mathcal{S}} \left| \hat{f}_n(s) - f(s) \right| \tag{5}$$

where $\hat{f}_n(s)$ denotes the empirical estimate of $f(s)$ after n steps in the environment (Tarbouriech & Lazaric, 2019).

In this work, we consider the case where f and the dynamics P have convenient geometric structures. In a vast variety of applications, the quantity of interest f is known to have certain group-structured symmetries $L_g : \mathcal{S} \to \mathcal{S}$ and state-dependent action symmetries $K^s_g : \mathcal{A} \to \mathcal{A}$. For all $g \in G, s \in \mathcal{S}, a \in \mathcal{A}$, f and P are invariant according to[2]

$$f(s, a) = f(L_g[s], K^s_g[a]) \tag{6}$$

$$P(s' \mid s, a) = P(L_g[s'] \mid L_g[s], K^s_g[a]) \tag{7}$$

For instance, in the diffusion process in Fig. 1, f follows radial symmetries, while P has roto-translation symmetries as in most physical systems (van der Pol et al., 2020b, Table 1).

An MDP with this structure, often denoted as *MDP with symmetries* (van der Pol et al., 2020b), naturally defines an MDP homomorphism $h = (\psi, \{\phi_s \mid s \in \mathcal{S}\})$ that can be efficiently built, as illustrated in Figure 1, by mapping state-action pairs across which f and P are invariant to a unique abstract state-action pair (van der Pol et al., 2020b; Ravindran & Barto, 2001). The main intuition with respect to our estimation process, is that all sets of states $[s]$ across which f is invariant, will map along ψ to an abstract state $\overline{s} \coloneqq \psi(s) = \psi(s') \in \overline{\mathcal{S}}\ \forall s, s' \in [s]$.

We consider the case where such an MDP homomorphism h encoding the underlying geometric structure is known. This is a fair assumption for a large class of applications, where geometric priors can easily be represented via a MDP homomorphism (van der Pol et al., 2020b). Nonetheless, in Section 9 we briefly discuss how the contributions presented in this work can be leveraged in the case of unknown MDP homomorphism. In the following, we introduce a novel geometric estimation error that makes it possible to leverage geometric priors both while learning an optimal sampling strategy, and in inference, when the gathered data is used to compute estimates of the unknown quantity of interest f.

3.1. Geometric Function Estimation

First, we introduce some quantities updated by the agent when obtaining a noisy realization of f at every time-step i, namely $x_i \sim y(s_i)$, where s_i indicates the current state at

[2]Here we extend f such that $f(s, a) \coloneqq f(s)\ \forall a \in \mathcal{A}$ and consider y to satisfy the same set of invariances as f.

time-step $i \in [t]$. After t interaction steps we have:

$$T_t(s) \coloneqq \sum_{i=1}^{t} \mathbb{I}\{s_i = s\} \tag{8}$$

$$\hat{f}_t(s) \coloneqq \frac{1}{T_t^+(s)} \sum_{i=1}^{t} x_i \mathbb{I}\{s_i = s\} \tag{9}$$

$$\hat{\sigma}_t^2(s) \coloneqq \frac{1}{T_t^+(s)} \sum_{i=1}^{t} x_i^2 \mathbb{I}\{s_i = s\} - \hat{f}_t(s)^2 \tag{10}$$

which are respectively the *visitation counts*, the *empirical mean* and *empirical variance*. Now, we define the geometric estimation error given n samples from f as follows.

Geometric Estimation Error

$$\bar{\xi}_n = \frac{1}{S} \sum_{s \in \mathcal{S}} \left| \hat{f}_n^A(s) - f(s) \right| \tag{11}$$

where $\hat{f}_n^A(s)$ is an empirical mean obtained by weighted averaging across all states within the same f-equivalence class $[s]$. Formally, given $T_n^+([s]) \coloneqq \sum_{s' \in [s]} T_n^+(s')$, we have:

$$\hat{f}_n^A(s) \coloneqq \frac{1}{T_n^+([s])} \sum_{s' \in [s]} T_n(s') \hat{f}_n(s') \tag{12}$$

Interestingly, the geometric estimation error (Eq. 11) generalizes the classic AE estimation error (Tarbouriech & Lazaric, 2019), which corresponds to the limit case of ours where there are no f-invariances and therefore every equivalence class is composed of only one state.

Given the geometric estimation error $\bar{\xi}_n$, any $\epsilon > 0$ and $\delta \in (0, 1)$, we say that an estimate of f is (ϵ, δ)-accurate if:

$$\mathbb{P}(\bar{\xi}_n \leq \epsilon) \geq 1 - \delta \tag{13}$$

Notice that, in this case, the RL agent is used as an algorithmic tool to estimate an external property of the environment and can be interpreted as an active sampler of the underlying Markov chain. In particular, we aim to design an algorithm that minimizes the sample complexity needed to estimate f nearly-optimally in high probability.

Definition 1 (Sample Complexity Geometric Estimation)**.** *Given an error $\epsilon > 0$ and a confidence level $\delta \in (0, 1)$, the sample complexity to solve the geometric function estimation problem is:*

$$n^{\bar{\xi}}(\epsilon, \delta) \coloneqq \min\{n \geq 1 : \mathbb{P}(\bar{\xi}_n \leq \epsilon) \geq 1 - \delta\} \tag{14}$$

4. From Experimental Design to Convex RL

In this section, we derive a principled objective to minimize the sample complexity of geometric estimation (Def. 1). We first show that $\bar{\xi}_n$ (Eq. 11) can be rewritten as a function of abstract states $\overline{s} \in \overline{\mathcal{S}}$, making it well-defined over the abstract MDP.

Proposition 1. *The geometry-aware estimation error $\bar{\xi}_n$ can be rewritten as a function of abstract states as:*

$$\bar{\xi}_n = \frac{1}{S} \sum_{\overline{s} \in \overline{\mathcal{S}}} E_{\overline{s}} \mid \hat{f}_n(\overline{s}) - f(\overline{s}) \mid \tag{15}$$

While Proposition 1 is proved in Appendix B, here we briefly mention the main intuition. Since the empirical estimator $\hat{f}_n^A(s)$ aggregates over experiments $s \in [s]$ across which f is invariant, the estimation error can be rewritten by considering only a representative of the f-equivalence class $[s]$, namely an abstract state $\overline{s} = \psi(s) \in \overline{\mathcal{S}}$. Then, equality is obtained by reweighting with the cardinality E_s of $[s]$.

4.1. Tractable formulation via Convex RL

Proposition 1 gives $\bar{\xi}_n$ as a function of abstract states, for which we derive the upper bound below (proof in Apx. B).

Proposition 2 (Convex Upper Bound of $\bar{\xi}_n$)**.** *With probability at least $1 - \delta$ and n interactions with f we have:*

$$\bar{\xi}_n \leq \frac{C(n, \overline{S}, \delta)}{S} \sum_{\overline{s} \in \overline{\mathcal{S}}} \overline{\mathcal{F}}(\overline{s};\, T_n^+) \tag{16}$$

with $C(n, \overline{S}, \delta) \coloneqq \max\left\{ \log(n\overline{S}/\delta), \sqrt{\log(n\overline{S}/\delta)} \right\}$ *and*

$$\overline{\mathcal{F}}(\overline{s};\, T_n^+) \coloneqq E_{\overline{s}} \left(\sqrt{\frac{2\sigma^2(\overline{s})}{T_n^+(\overline{s})}} + \frac{F_{\max}}{T_n^+(\overline{s})} \right)$$

where $T_n^+(\overline{s}) \coloneqq T_n^+([s])$ are the visitation counts of $\overline{s}$.

As an abstract state represents an f-equivalence class of original states, i.e., $\overline{s} = \psi(s) = \psi(s')\ \forall s, s' \in [\overline{s}]$, one can notice that Equation 16 captures two interesting facts. First, equivalence classes $[\overline{s}]$ that are under-visited (small $T_n^+(\overline{s})$) with high variance (large $\sigma^2(\overline{s})$) lead to higher estimation error. Second, the cardinality of an equivalence class $E_{\overline{s}}$ is proportional to how much its estimation quality impacts the overall estimation error.

While the upper bound in Equation 16 is convex, the constraint set of admissible visitation counts T_n^+ is non-convex (Tarbouriech et al., 2020), rendering this formalization a NP-hard problem (Welch, 1982; Tarbouriech & Lazaric, 2019). Nonetheless, problems of this form present a hidden convexity in the asymptotic relaxation ($n \to \infty$) of the dual problem. Given the set Λ of admissible asymptotic state-action distributions

$$\Lambda \coloneqq \{\lambda \in \Delta(\mathcal{S} \times \mathcal{A}) : \forall s \in \mathcal{S},$$

$$\sum_{b\in\mathcal{A}}\lambda(s,b) = \sum_{(s',a)\in\mathcal{S}\times\mathcal{A}} P(s|s',a)\lambda(s',a)\}$$

we introduce the following η-smoothened objective.

Geometric Estimation Objective

$$\overline{\mathcal{L}}_{\infty,\eta}(\lambda) \coloneqq \frac{1}{S}\sum_{\overline{s}\in\overline{\mathcal{S}}} E_{\overline{s}}\sqrt{\frac{2\sigma^2(\overline{s})}{\sum_{s\in[\overline{s}]}(\lambda(s)+\eta)}} \tag{17}$$

Where $\lambda(s) \coloneqq \sum_{a\in\mathcal{A}}\lambda(s,a)$. Crucially, in the following statement, we show that $\overline{\mathcal{L}}_{\infty,\eta}(\lambda)$ is an upper bound of the estimation error $\bar{\xi}_n$ and therefore that minimizing $\overline{\mathcal{L}}_{\infty,\eta}(\lambda)$ is a principled objective to minimize the sample complexity in Definition 1.

Proposition 3 (Tractable Convex Upper Bound of $\bar{\xi}_n$)**.** *Let an empirical state-action frequency at time t be defined as $\lambda_t(s,a) = T_t(s,a)/t$, then for $E_{\overline{s}}\eta \leq \frac{1}{n}$ we have:*

$$\bar{\xi}_n \leq \frac{2S}{\sqrt{n}}C(n,\overline{S},\delta)\left[\overline{\mathcal{L}}_{\infty,\eta}(\lambda_n) + \frac{\overline{S}F_{\max}}{S\sqrt{n}\eta}\right] \tag{18}$$

Interestingly, in the next section we show that this problem can be solved by computing optimal sampling policies for abstract MDPs only.

5. Geometric Active Exploration (GAE)

In this section, we introduce **G**eometric **A**ctive **E**xploration (GAE), an algorithm for AE that leverages the power of abstraction to improve statistical and computational efficiency. Alike classic AE algorithms (Tarbouriech & Lazaric, 2019; Hazan et al., 2019), GAE is based on a Frank-Wolfe (FW) scheme (Jaggi, 2013) that reduces the problem

$$\min_{\lambda\in\Lambda}\overline{\mathcal{L}}_{\infty,\eta}(\lambda) \tag{19}$$

to a sequence of K linear programs, each corresponding to a classic MDP $\mathcal{M}_f^k$ with reward r_λ^k defined as

$$\nabla\overline{\mathcal{L}}^+_{t_k-1}(\lambda)_{[s,a]} = \frac{-E_s\left[\sqrt{2\hat{\sigma}^2_{t_k-1}(\overline{s})} + \alpha(t_k-1,\overline{s},\delta)\right]}{2S\left(\sum_{s\in[s]}(\sum_{b\in\mathcal{A}}\lambda(s,b)+\eta)\right)^{\frac{3}{2}}}$$

where the unknown variances are optimistically bounded via a quantity $\alpha(t,\overline{s},\delta)$ according to the following result.

Lemma 5.1 (Variance Concentration (Panaganti & Kalathil, 2022))**.** *For all $\overline{s}\in\overline{\mathcal{S}}$, with probability at least $1-\delta$ we have*

$$\left|\sqrt{\sigma^2(\overline{s})} - \sqrt{\hat{\sigma}^2_t(\overline{s})}\right| \leq F_{\max}\sqrt{2\frac{\log(2\overline{S}t^2/\delta)}{T^+_t(\overline{s})}} \coloneqq \alpha(t,\overline{s},\delta)$$

We show that optimistic gradients (r_λ^k) of Equation 19 satisfy state-action invariances induced by f (proof in Apx. C).

Algorithm 1 **G**eometric **A**ctive **E**xploration (GAE)

Input: η, h, $\mathcal{M}$, δ, $\{\tau_k\}_{k\in[K-1]}$
Compute abstract CMP $\overline{\mathcal{M}}$ induced by $h,\mathcal{M}$
Initialize $\overline{\lambda}_1 = 1/\overline{S}\overline{A}$
for $k = 1,2,...,K-1$ **do**
  Compute abstract reward $\bar{r}^k_{\overline{\lambda}_k}\ \forall\overline{s}\in\overline{\mathcal{A}},\overline{a}\in\overline{\mathcal{A}}$

$$\bar{r}^k_{\overline{\lambda}_k}(\overline{s},\overline{a}) \coloneqq \frac{-E_s\left[\sqrt{2\hat{\sigma}^2_{t_k-1}(\overline{s})} + \alpha(t_k-1,\overline{s},\delta)\right]}{2S\left(\overline{\lambda}_k(\overline{s}) + E_{\overline{s}}\eta\right)^{\frac{3}{2}}}$$

  $\overline{\pi}^+_{k+1} \leftarrow$ MDP-SOLVER $\left[\overline{\mathcal{M}}^k_{\bar{r}} = \left(\overline{\mathcal{S}},\overline{\mathcal{A}},\overline{P},\overline{\mu},\bar{r}^k_{\overline{\lambda}_k}\right)\right]$
  Lift abstract policy

$$\pi^+_{k+1}(a|s) = \frac{\overline{\pi}^+_{k+1}(\overline{a}|\psi(s))}{|\{a\in\phi_s^{-1}(\overline{a})\}|}, \forall s\in\mathcal{S}, a\in\phi_s^{-1}(\overline{a})$$

  Deploy policy π^+_{k+1} in $\mathcal{M}_f$ for τ_k steps
  Compute $\hat{f}_{t_{k+1}-1}$ and $\tilde{v}_{k+1}$
  Aggregate estimates according to $\overline{\mathcal{M}}^k_f$

$$\hat{f}_{t_{k+1}-1}(\overline{s}) = \frac{1}{T^+_{t_{k+1}-1}(\overline{s})}\sum_{s\in[\overline{s}]} T^+_{t_{k+1}-1}(s)\hat{f}_{t_{k+1}-1}(s)$$

$$\tilde{\overline{v}}_{k+1}(\overline{s},\overline{a}) = \sum_{s\in[\overline{s}]}\tilde{v}_{k+1}(s,a)$$

  Update the abstract state-action frequency $\overline{\lambda}_{k+1}$

$$\overline{\lambda}_{k+1} = \frac{\tau_k}{t_{k+1}-1}\tilde{\overline{v}}_{k+1} + \frac{t_k-1}{t_{k+1}-1}\overline{\lambda}_k$$

end for
Return: $\hat{f}_{t_K-1}$

Proposition 4 (Gradient-Reward Invariances)**.** *If f is invariant over states s and s', then $\forall s,s'\in[s],\forall a,a'\in\mathcal{A}$*

$$\nabla_\lambda\overline{\mathcal{L}}^+_{t_k-1}(\lambda)[s,a] = \nabla_\lambda\overline{\mathcal{L}}^+_{t_k-1}(\lambda)[s',a']$$

Intuitively, Proposition 4 is due to the fact that the invariances of f propagate to the gradient via $\hat{\sigma}$, because of its definition in Equation 10. As a consequence, we can define the optimistic abstract reward as:

$$\bar{r}^k_{\overline{\lambda}}(\overline{s},\overline{a}) \coloneqq \frac{-E_s\left[\sqrt{2\hat{\sigma}^2_{t_k-1}(\overline{s})} + \alpha(t_k-1,\overline{s},\delta)\right]}{2S\left(\overline{\lambda}(\overline{s}) + E_{\overline{s}}\eta\right)^{\frac{3}{2}}} \tag{20}$$

where $\overline{\lambda}\in\overline{\Lambda}\subseteq\Delta(\overline{\mathcal{S}}\times\overline{\mathcal{A}})$ is an admissible abstract state-action distribution.[3] Given the reward $\bar{r}^k_{\overline{\lambda}}$ and the invariances on the dynamics P in Equation 3, the MDP $\mathcal{M}_f^k$ at each step $k\in[K]$ of the FW scheme can be solved by computing the optimal policy for the abstract MDP $\overline{\mathcal{M}}^k_f$ and then lifting it back along the MDP homomorphism h

[3] Notice that the set of admissible abstract state-action distributions $\overline{\Lambda}$ can be defined analogously to Λ.

(Eq. 4). This observation unlocks the power of abstraction for AE in MDPs, and leads to the GAE algorithm, for which we report the pseudocode in Algorithm 1.

First, GAE computes the abstract CMP $\overline{\mathcal{M}}$ given the homomorphism h and the original CMP $\mathcal{M}$ (line 2). This operation is computationally efficient as we can perform one sweep over $\mathcal{SA}$ to compute $\overline{\mathcal{SA}}$ by applying ψ and ϕ_s to original states and actions respectively. The empirical visitation frequency $\overline{\lambda}_1$ is initialized in line 3. At each iteration, GAE computes the optimal abstract policy $\overline{\pi}^+_{k+1}$ (line 6), e.g., via value iteration, for the abstract MDP $\overline{\mathcal{M}}^k_{\bar{r}} = \overline{\mathcal{M}} \cup \{\bar{r}^k_{\overline{\lambda}_k}\}$, where $\bar{r}^k_{\overline{\lambda}_k}$ is an estimate of the optimistic gradient in (20) based on the samples gathered via policy π^+_k during the previous iteration. Then, it computes an optimal policy for the original MDP, namely π^+_{k+1} by lifting the optimal abstract policy (line 7), which is deployed for τ_k steps (line 8). The gathered samples of f and the state-action visitation counts are used to update the abstract empirical mean of f, namely $\hat{f}_{t_{k+1}-1}(\overline{s})$, and compute the abstract empirical state-action distribution $\tilde{\overline{v}}_{k+1}$ by aggregating $\tilde{v}_{k+1}$ across states within the same equivalence class (lines 8-9). Then, the empirical visitation frequency $\overline{\lambda}_{k+1}$ is updated to serve for the gradient estimation at the next iteration. GAE outputs the aggregated estimates of f.

Benefits of Abstraction. Since AE typically entails solving a sequence of MDPs, encoding each instance via a (smaller) abstract MDP (line 6) gives significant computational benefits, as shown in Section 7. From a statistical perspective, the fundamental advantage of GAE is leveraging known invariances during the density estimation process (lines 8-10) common in previous works (Hazan et al., 2019; Tarbouriech & Lazaric, 2019), leading to faster convergence. However, how do different degrees of geometric structure benefit the statistical efficiency of the problem? In the next section, we formally answer this question by presenting a sample complexity result showcasing a geometric compression term. In the following, we denote by sampling strategy the empirical distribution λ_n over $\mathcal{S}\times\mathcal{A}$ induced by the policies $\{\pi^+_k\}_{k\in[K]}$.

6. Theoretical Analysis

In this section, we present an upper bound on the regret and the sample complexity achieved by GAE against an optimal sampling strategy. The latter result captures the impact of abstraction on the complexity of the problem via the following notion of geometric compression.

Definition 2 (Geometric Compression Term)**.** *We denote as geometric compression term Φ the ratio between the cardinalities of the abstract and original state spaces, formally:*

$$\Phi := \overline{S}/S \in (0, 1] \tag{21}$$

Before presenting these results, we state two assumptions we employed for deriving them.

Assumption 6.1 (Homogeneous Equivalence Classes)**.** *The equivalence classes induced by f over $\mathcal{S}$ are homogeneous i.e., they have same cardinality, $E_s = E_{s'}\ \forall s, s' \in \mathcal{S}$.*

Moreover, due to the non-episodic nature of our setting we need to assume the following.

Assumption 6.2 (Ergodicity)**.** *The Markov chain induced by any Markovian stationary policy is ergordic.*

6.1. Regret Analysis

Given an empirical state-action distribution $\lambda_n \in \Delta(\mathcal{S}\times\mathcal{A})$ induced by a sequence of sampling policies interacting with the environment, we define its regret against the optimal sampling strategy as follows

$$\mathcal{R}_n(\lambda_n) := \overline{\mathcal{L}}_{\infty,\eta}(\lambda_n) - \overline{\mathcal{L}}_{\infty,\eta}(\lambda^*) \tag{22}$$

where $\lambda^* := \arg\min_{\lambda\in\Lambda} \overline{\mathcal{L}}_{\infty,\eta}(\lambda)$ is an optimal state-action distribution. Notice that, while common in AE (Tarbouriech & Lazaric, 2019), this notion of regret is not standard in RL (Szepesvári, 2009).

Theorem 6.1 (Regret Guarantee)**.** *If algorithm* GAE *is run with a budget of n samples and $\tau_k = 3k^2 - 3k + 1$ then w.p. at least $1-\delta$, it holds that:*

$$\mathcal{R}_n = \widetilde{\mathcal{O}}\left(\left(\frac{\Phi^{\frac{1}{2}} S^{\frac{1}{2}} A F_{\max}\sqrt{\sigma^2_{\max}}}{\eta^{\frac{5}{2}}}\right)\frac{1}{n^{1/3}}\right)$$

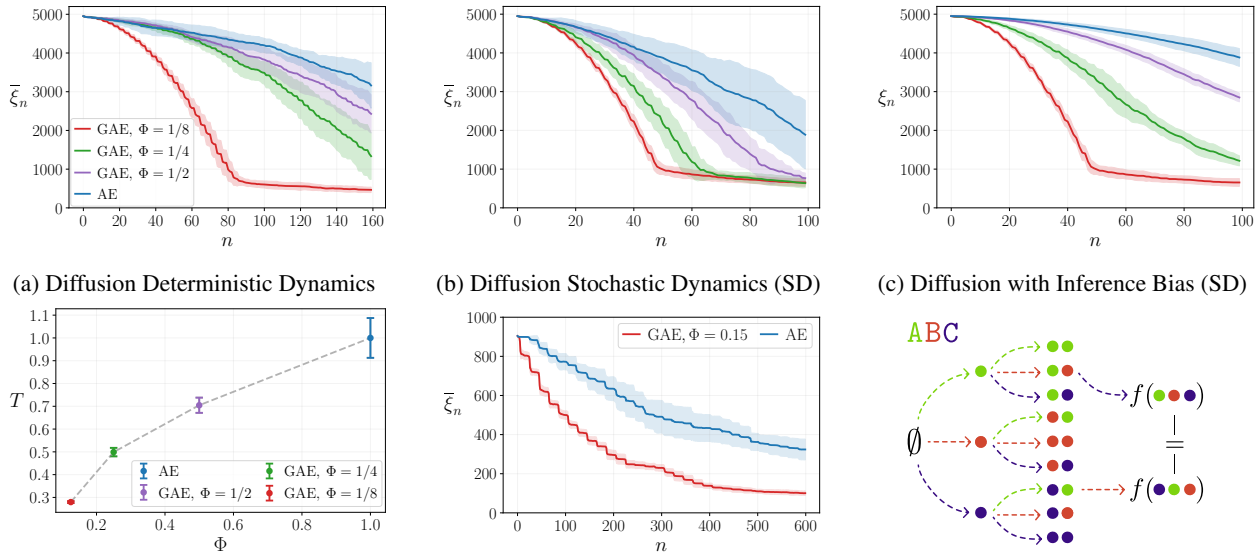
In the following, we present a brief sketch of the proof, while complete derivations are deferred to Apx. D.

Step 1. We derive the result w.r.t. the abstract variables via a Frank-Wolfe analysis, taking into account (i) the effect of the optimistic gradient and (ii) the error due to the gap between the empirical and stationary distribution induced by the policy at each iteration (Tarbouriech & Lazaric, 2019).

Step 2. Since the density estimation step of GAE is carried out w.r.t. a distribution defined over $\overline{\mathcal{S}}\times\overline{\mathcal{A}}$, we notice that it can be analysed with respect to the abstract variables.

Step 3. Finally, in order to state the result w.r.t. the original MDP variables, we leverage the *geometric compression* term Φ (Def. 2).

In the following, we report the sample complexity bound capturing the effect of abstraction on learning with high probability a nearly-optimal sampling strategy w.r.t. the geometric estimation objective (17).



(a) Diffusion Deterministic Dynamics (b) Diffusion Stochastic Dynamics (SD) (c) Diffusion with Inference Bias (SD)

(d) Diffusion Runtime (SD) (e) Strings Statistical Performance (f) Strings Drawing

Figure 2. Comparison of GAE with AE. GAE shows better statistical and computational efficiency. Experiments were carried out over 15 seeds and confidence intervals shown are ± one standard deviation. 2a the statistical advantage of GAE with compression Φ against AE for deterministic dynamics in the diffusion environment. 2b same setting as 2a, but with stochastic dynamics. 2c the (classic) estimation error taken over the abstract state space. 2d computational advantage of GAE over AE for different degrees of compression (standardized). 2e the statistical advantage of GAE over AE in the strings environment. 2f the strings environment and the invariance of f under permutation.

Theorem 6.2 (Sample Complexity of Geometric Estimation Objective)**.** *If algorithm* GAE *is run with* $\tau_k = 3k^2 - 3k + 1$*, for:*

$$n = \widetilde{\mathcal{O}}\left(\frac{\Phi^{\frac{3}{2}} S^{\frac{3}{2}} A^3 F_{\max}^3 (\sigma_{\max}^2)^{\frac{3}{2}}}{\eta^{\frac{15}{2}} \epsilon^3}\right)$$

samples, then we have that with probability at least $1-\delta$*:*

$$\mathbb{P}\left(\left|\overline{\mathcal{L}}_{\infty,\eta}(\lambda_n) - \overline{\mathcal{L}}_{\infty,\eta}(\lambda^*)\right| \leq \epsilon\right) \geq 1 - \delta$$

Crucially, setting $\Phi = 1$ recovers the case where abstraction is not leveraged by the algorithm.

6.2. Geometric Compression in MDP with Symmetries

If the MDP $\mathcal{M}_f$ has symmetries (see Eq. 6 and 7), it is possible to make explicit the dependency of Φ on the cardinality of the state symmetries group $\mathbb{G} = (\{L_g\}_{g \in G}, \cdot)$.

Proposition 5 (Compression via Group Cardinality)**.** *Given a set of group-structured state symmetries* $\mathbb{G} = (\{L_g\}_{g \in G}, \cdot)$ *and* $\mathrm{Stab}(s) = \mathrm{Stab}(s')\ \forall s, s' \in \mathcal{S}$ *then:*

$$\Phi = \frac{|\mathrm{Stab}(s)|}{|G|}$$

where $\mathrm{Stab}(s) \coloneqq \{g \in G : L_g[s] = s\}$.

Proposition 5, which we proved via the Orbit-Stabilizer Theorem (Rotman, 2010, Theorem C-1.16), sheds light on the intuition that a higher number of symmetries leads to a higher degree of compression.

7. Experiments

In this section, we perform a thorough experimental evaluation of GAE analysing its statistical and computational efficiency on two tasks where the unknown quantity f represents: (1) the amount of pollutant emerging from a point source (see Fig. 1), and (2) the toxicity of chemical compounds generated from a set of base elements (see Fig. 2f). In all experiments, unless otherwise specified, we compare the data gathering performances of GAE with classic AE, an implementation based on Convex RL and analogous to GAE, but not exploiting symmetries (Tarbouriech & Lazaric, 2019). More explicitly, AE is the same as GAE (see Alg. 1) in the case when the homomorphism h is an identity map and hence $\Phi = 1$ and $\overline{\mathcal{M}} = \mathcal{M}$.

(1) Pollutant Diffusion Process. We consider the problem of actively estimating the amount of pollution released in the environment from a point source and following a diffusion process with radial symmetries, as illustrated in Figure 1 and introduced in Section 2. The agent can measure the pollution at 30 different radii and at 8 different angles, resulting in $S = 240$ states, and can select actions

$\mathcal{A} = \{\text{in}, \text{out}, \text{clockwise}, \text{anticlockwise}, \text{stay}\}$. In Figure 2a, we show the sample efficiency of GAE compared with AE for several values of Φ in the case of deterministic dynamics. We observe a significant effect of different degrees of compression to the data efficiency of the algorithm. Similarly, Figure 2b shows the same comparison for stochastic dynamics. In Figure 2c, we show that omitting the inductive bias in the inference step, specifically the absence of weighted averaging across equivalence classes, worsens the performance of AE, thus showing the role of exploiting geometric structure also in inference. Ultimately, in Figure 2d, we compare the normalized runtime of GAE and AE for several degrees of compression, showing the effect of leveraging geometric structure on computational efficiency and hence practical scalability.

(2) Toxicity of Chemical Compounds. In this experiment, we consider the problem of actively estimating the toxicity of chemical compounds that can be generated using some base chemical elements. Similar to prior work (Thiede et al., 2022; Dong et al., 2022), we associate with states chemical compounds represented as strings (see Fig. 2f). Each character of the string thereby stands for a base chemical element. In our simplified setting, we consider three base elements `A`, `B`, `C` and every compound may consist of at most 5 base elements, resulting in a total of $S = 363$ states. The actions correspond to the three base elements and a stay action. If the agent picks an action that corresponds to a base element, this is appended to the current string, resulting in a new compound to which the agent transitions and gets a noisy observation of its toxicity. We consider toxicity to be invariant w.r.t. string permutations, resulting in $\overline{S} = 55$ states and $\Phi \approx 0.15$. Figure 2e shows the statistical performances of GAE compared with AE. We observe that even a relatively small compression leads to a significant statistical advantage.

An extensive discussion of the environments, homomorphisms, and implementation details is deferred to Apx. F.

8. Related Works

In the following, we present relevant works in MDP Homomorphisms, Active Exploration, and Convex RL.

MDP Homomorphisms. Ravindran & Barto (2001) were among the first to identify the benefits of solving MDPs via MDP homomorphisms. Recently, these have been extended to Deep RL (van der Pol et al., 2020b), approximate invariances (Ravindran & Barto, 2004; Jiang et al., 2014; Ravindran & Barto, 2002), and continuous domains (Rezaei-Shoshtari et al., 2022; Biza & Platt, 2018; Zhao, 2022). While in this work we have built a fundamental connection between Active Exploration and abstraction via MDP homomorphisms, extending the contributions presented to the mentioned settings, namely Deep RL, approximate abstraction, and continuous domains, is an interesting and relevant direction of future work.

Active Exploration. The AE problem has been introduced in (Tarbouriech & Lazaric, 2019) with a non-episodic setting assuming ergodicity and reversibility of the induced Markov chain. Afterwards, the framework has been extended to perform transition dynamics estimation in high probability (Tarbouriech et al., 2020). Compared with these works, by smoothening the objective (Sec. 4), we remove the need to solve an LP program at every iteration paving the way for more efficient dynamic programming methods, e.g., value iteration (Puterman, 2014, Chapter 6). Recently, Mutny et al. (2023) introduced an episodic version of the problem, where the agent can reset its state, but ergodicity is not required, while the unknown quantity f is assumed to be an element of a reproducing kernel Hilbert space with known kernel (Mutny et al., 2023). While this setting can capture the correlation structure of f, it does not leverage known geometric structure of the dynamics and therefore cannot *compress* the original MDP into an abstract one to render the algorithm more scalable from a computational perspective.

Convex RL. The algorithmic scheme presented in this work is an instance of a general framework that has received significant attention recently, generally under the name of *Convex RL* (Hazan et al., 2019; Zhang et al., 2020; Zahavy et al., 2021; Geist et al., 2022; Mutti et al., 2022a; 2023; 2022b). In this framework, a learning agent interacts with a CMP to optimize an objective formulated through a convex function of the state-action distribution induced by the agent policy. Different choices of this convex function allow to cover several domains of practical interest beyond active exploration, such as pure exploration (e.g., Hazan et al., 2019), imitation learning (e.g., Abbeel & Ng, 2004), and risk-averse RL (Garcıa & Fernández, 2015) among others.

9. Conclusions

In this paper, we presented how abstraction can be leveraged to solve the Active Exploration problem with better statistical complexity and computational efficiency. Before presenting some concluding remarks, we briefly mention a few important discussion points.

Beyond Known Geometric Structure. In a wide set of applications e.g., molecular design or environmental sensing, geometric priors on f and P are known and can be easily encoded in an abstract MDP as considered in this work. Nonetheless, for a arguably wider set, geometric structure is not known or it is not human-interpretable. In these cases, one can leverage algorithms that automatically discover symmetries in the environment (Angelotti et al., 2021; Narayanamurthy & Ravindran, 2008) or directly

learn a MDP homomorphism (Mavor-Parker et al., 2022; Biza & Platt, 2018; Wolfe & Barto, 2006; Mondal et al., 2022). Interestingly, in both these cases, the GAE algorithm can be run with the machine-learned homomorphism.

Abstraction in Convex RL. The presented algorithmic scheme (Alg. 1) and theoretical analysis (Sec. 6) are not tied to the AE problem treated in this paper and can be straightforwardly extended to leverage abstraction in a variety of Convex RL application areas, including those mentioned within Section 8.

Benefit of Abstraction on Statistical Efficiency. Abstraction, via MDP homomorphisms or close variants, has been leveraged in a large body of works (Rezaei-Shoshtari et al., 2022; van der Pol et al., 2020b;a; 2022; Soni & Singh, 2006; Ravindran, 2003; Zhu et al., 2022; Ravindran & Barto, 2002; Mahajan & Tulabandhula, 2017; Ravindran, 2003) showcasing experimental advantages on the sample complexity in the context of RL. Nonetheless, to the best of our knowledge, this work is the first that formally captures the effect of abstraction via MDP homomorphisms on sample complexity (Sec. 6). Moreover, we believe the main ideas within our analysis can be leveraged to treat a large extent of RL settings.

To summarize, in this work we have presented a principled Active Exploration objective that makes it possible to leverage geometric priors on an unknown quantity f and the system dynamics to solve the estimation problem via an abstraction process, thus increasing statistical and computational efficiency. We introduced an algorithm, Geometric Active Exploration, which we believe could render Active Exploration more practical for a wide variety of real-world settings. Then, we have presented, to the best of our knowledge, the first analysis capturing the effect of abstraction on the sample complexity of the Active Exploration problem, and more in general in RL. Ultimately, we have performed a thorough experimental evaluation of the proposed method on tasks resembling real-world scientific discovery problems while showing promising performances.

Acknowledgement

This publication was made possible by the ETH AI Center doctoral fellowship to Riccardo De Santi.

We would like to thank Emanuele Rossi, Ossama El Oukili, Massimiliano Viola, Marcello Restelli, and Michael Bronstein for collaborating on a previous version of this project. Moreover, we wish to thank Mojmír Mutný and Manish Prajapat for the insightful discussions.

Impact Statement

This paper presents work whose goal is to advance the field of Reinforcement Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, 2004.

Angelotti, G., Drougard, N., and Chanel, C. P. Expert-guided symmetry detection in Markov decision processes. *arXiv preprint arXiv:2111.10297*, 2021.

Biza, O. and Platt, R. Online abstraction with mdp homomorphisms for deep learning. In *International Conference on Autonomous Agents and Multiagent Systems*, 2018.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, pp. 273–304, 1995.

Cheng, Y., Gong, Y., Liu, Y., Song, B., and Zou, Q. Molecular design in drug discovery: a comprehensive review of deep generative models. *Briefings in Bioinformatics*, 22 (6), 2021.

Dong, J., Zhao, M., Liu, Y., Su, Y., and Zeng, X. Deep learning in retrosynthesis planning: datasets, models and tools. *Briefings in Bioinformatics*, 23(1), 2022.

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4 (4):828–849, 2019.

Garcıa, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R., and Pietquin, O. Concave utility reinforcement learning: The mean-field game viewpoint. In *International Conference on Autonomous Agents and Multiagent Systems*, 2022.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019.

Jaggi, M. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, 2013.

Jiang, N., Singh, S., and Lewis, R. Improving UCT planning via approximate homomorphisms. In *International Conference on Autonomous Agents and Multiagent Systems*, 2014.

Krause, A. *Optimizing sensing: Theory and applications*. PhD thesis, Carnegie Mellon University, 2008.

Lafferty, J., Liu, H., and Wasserman, L. Concentration of measure. *Online at* https://www.stat.cmu.edu/~larry/=sml/Concentration.pdf, 2008.

Lindner, D., Krause, A., and Ramponi, G. Active exploration for inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Mahajan, A. and Tulabandhula, T. Symmetry learning for function approximation in reinforcement learning. *arXiv preprint arXiv:1706.02999*, 2017.

Mavor-Parker, A. N., Sargent, M. J., Banino, A., Griffin, L. D., and Barry, C. A simple approach for state-action abstraction using a learned mdp homomorphism. *arXiv preprint arXiv:2209.06356*, 2022.

Mondal, A. K., Jain, V., Siddiqi, K., and Ravanbakhsh, S. EqR: Equivariant representations for data-efficient reinforcement learning. In *International Conference on Machine Learning*, 2022.

Mutny, M., Janik, T., and Krause, A. Active exploration via experiment design in Markov chains. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Mutti, M., De Santi, R., De Bartolomeis, P., and Restelli, M. Challenging common assumptions in convex reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022a.

Mutti, M., De Santi, R., and Restelli, M. The importance of non-Markovianity in maximum state entropy exploration. In *International Conference on Machine Learning*, 2022b.

Mutti, M., De Santi, R., De Bartolomeis, P., and Restelli, M. Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250):1–42, 2023.

Narayanamurthy, S. M. and Ravindran, B. On the hardness of finding symmetries in Markov decision processes. In *International Conference on Machine Learning*, 2008.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, 2014.

Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Pukelsheim, F. *Optimal design of experiments*. SIAM, 2006.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Ravindran, B. Smdp homomorphisms: An algebraic approach to abstraction in semi Markov decision processes. In *International Joint Conference on Artificial Intelligence*, 2003.

Ravindran, B. and Barto, A. G. Symmetries and model minimization in Markov decision processes. *Technical report*, 2001.

Ravindran, B. and Barto, A. G. Model minimization in hierarchical reinforcement learning. In *Abstraction, Reformulation, and Approximation: 5th International Symposium*, 2002.

Ravindran, B. and Barto, A. G. Approximate homomorphisms: A framework for non-exact minimization in Markov decision processes. *Technical report*, 2004.

Rezaei-Shoshtari, S., Zhao, R., Panangaden, P., Meger, D., and Precup, D. Continuous mdp homomorphisms and homomorphic policy gradient. In *Advances in Neural Information Processing Systems*, 2022.

Rotman, J. J. *Advanced Modern Algebra*, volume 114. American Mathematical Society, 2010.

Schreck, J. S., Coley, C. W., and Bishop, K. J. Learning retrosynthetic planning through simulated experience. *ACS Central Science*, 5(6):970–981, 2019.

Soni, V. and Singh, S. Using homomorphisms to transfer options across continuous reinforcement learning domains. In *AAAI Conference on Artificial Intelligence*, volume 6, pp. 494–499, 2006.

Szepesvári, C. Reinforcement learning algorithms for mdps. 2009.

Tarbouriech, J. and Lazaric, A. Active exploration in Markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Tarbouriech, J., Shekhar, S., Pirotta, M., Ghavamzadeh, M., and Lazaric, A. Active model estimation in markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Thiede, L. A., Krenn, M., Nigam, A., and Aspuru-Guzik, A. Curiosity in exploring chemical spaces: intrinsic rewards for molecular reinforcement learning. *Machine Learning: Science and Technology*, 3(3), 2022.

van der Pol, E., Kipf, T., Oliehoek, F. A., and Welling, M. Plannable approximations to mdp homomorphisms: Equivariance under actions. In *International Conference on Autonomous Agents and Multiagent Systems*, 2020a.

van der Pol, E., Worrall, D., van Hoof, H., Oliehoek, F., and Welling, M. Mdp homomorphic networks: Group symmetries in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020b.

van der Pol, E., van Hoof, H., Oliehoek, F. A., and Welling, M. Multi-agent mdp homomorphic networks. In *International Conference on Learning Representations*, 2022.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Welch, W. J. Algorithmic complexity: three np-hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, 15(1):17–25, 1982.

Wolfe, A. P. and Barto, A. G. Decision tree methods for finding reusable mdp homomorphisms. In *AAAI Conference on Artificial Intelligence*, 2006.

Zahavy, T., O'Donoghue, B., Desjardins, G., and Singh, S. Reward is enough for convex mdps. In *Advances in Neural Information Processing Systems*, 2021.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, 2020.

Zhao, R. Y. *Continuous Homomorphisms and Leveraging Symmetries in Policy Gradient Algorithms for Markov Decision Processes*. PhD thesis, 2022.

Zhu, Z.-M., Jiang, S., Liu, Y.-R., Yu, Y., and Zhang, K. Invariant action effect model for reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2022.

A. List of symbols

General Mathematical Objects

$\Delta(X)$	$\triangleq$	Probability simplex over X
x^+	$\triangleq$	$\max\{1, x\}$
$\overline{X}$	$\triangleq$	Abstract counterpart of variable X

MDP

$\mathcal{M}_r$	$\triangleq$	Markov decision process $\mathcal{M}_r = (\mathcal{S}, \mathcal{A}, P, \mu, r)$
$\mathcal{S}$	$\triangleq$	State space
$\mathcal{A}$	$\triangleq$	Action space
P	$\triangleq$	Transition model $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
r	$\triangleq$	Scalar state function e.g., reward $r : \mathcal{S} \times \mathcal{A} \to \Delta([0, R])$
μ	$\triangleq$	Initial state distribution $\mu \in \Delta(\mathcal{S})$
S	$\triangleq$	Size of the state space $S = \lvert\mathcal{S}\rvert$
A	$\triangleq$	Size of the action space $A = \lvert\mathcal{A}\rvert$
s	$\triangleq$	State $s \in \mathcal{S}$
a	$\triangleq$	Action $a \in \mathcal{A}$
π	$\triangleq$	Stationary Markovian policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
λ	$\triangleq$	Stationary state-action distribution $\lambda \in \Lambda$
Λ	$\triangleq$	Set of admissible state-action distribution, expression within Section 3

Symmetries, MDP Homomorphisms

L_g	$\triangleq$	state transformation or symmetry $L_g : \mathcal{S} \to \mathcal{S}$
K_g^s	$\triangleq$	state-dependent action transformation or symmetry $L_g : \mathcal{A} \to \mathcal{A}$
$\mathbb{G}$	$\triangleq$	Algebraic group structure e.g., $\mathbb{G} = (\{L_g\}_{g \in G}, \cdot)$[4]
g	$\triangleq$	Index of group element $g \in G$
$[s]$	$\triangleq$	Equivalence class of f-invariant states $[s] \subseteq \mathcal{S}$
h	$\triangleq$	MDP homomorphism between $\mathcal{M}_f$ and $\overline{\mathcal{M}}_f$, $h : \mathcal{S} \times \mathcal{A} \to \overline{\mathcal{S}} \times \overline{\mathcal{A}}$, $h = (\psi, \{\phi_s \mid s \in \mathcal{S}\})$
ψ	$\triangleq$	Homomorphism state map $\psi : \mathcal{S} \to \overline{\mathcal{S}}$
ϕ	$\triangleq$	Homomorphism state-dependent action map $\phi_s : \mathcal{A} \to \overline{\mathcal{A}}$
$[\overline{s}]$	$\triangleq$	Equivalence class of f-invariant states mapping to $\overline{s}$ along ψ, $[\overline{s}] = \{s' \in \mathcal{S} : \psi(s') = \psi(s)\}$
E_s	$\triangleq$	Cardinality of equivalence class $[s]$, $E_s = \lvert[s]\rvert = \lvert[\overline{s}]\rvert = E_{\overline{s}}$
Φ	$\triangleq$	Geometric compression coefficient, $\Phi = \overline{S}/S$

GAE Algorithm

f	$\triangleq$	Unknown state quantity $f : \mathcal{S} \to \mathcal{B} \subset \mathbb{R}$
ν	$\triangleq$	Noise random variable with zero mean and unknown variance σ^2
y	$\triangleq$	Noisy random variable $y(s) = f(s) + \nu(s)$
$T_t(s)$	$\triangleq$	Visitation counts for state s after t steps, see Equation (8)
$\hat{f}_t(s)$	$\triangleq$	Empirical mean for state s after t steps, see Equation (9)
$\hat{\sigma}_t^2(s)$	$\triangleq$	Empirical variance for state s after t steps, see Equation (10)
$\hat{f}_t^A(s)$	$\triangleq$	Average empirical mean for equivalence class [s], see Equation (12)
$\hat{f}_t(\overline{s})$	$\triangleq$	Empirical mean for abstract state $\overline{s}$ after t steps
$\bar{\xi}_n$	$\triangleq$	Geometric estimation error, see Equation (11)
ϵ	$\triangleq$	Controllable approximation error for PAC guarantees
δ	$\triangleq$	Controllable probability of error for PAC guarantees
$n^{\bar{\xi}}(\epsilon, \delta)$	$\triangleq$	Sample complexity for PAC estimation of geometric estimation error $\bar{\xi}_n$, see def. 1
n	$\triangleq$	Sample complexity for PAC optimality of geometric estimation objective, see def. 6.2

[4] we do not explicitly specify the identity and inverse elements of the group as we do not use them in the following

$\overline{\mathcal{L}}_n$	$\triangleq$	Finite-samples Convex RL Objective (31)
$\overline{\mathcal{L}}_{\infty,\eta}$	$\triangleq$	Asymptotic and smoothened Convex RL Objective (19)
$\bar{r}^k_{\overline{\lambda}_k}$	$\triangleq$	Abstract reward estimated at iteration k via empirical density $\overline{\lambda}_k$
$\overline{\pi}^+_{k+1}$	$\triangleq$	Optimal abstract policy w.r.t. MDP $\overline{\mathcal{M}}^k_{\bar{r}}$
π^+_{k+1}	$\triangleq$	Optimal original policy at iteration k obtained by lifting $\overline{\pi}_{k+1}$
τ_k	$\triangleq$	Length of trajectory of policy deployed at iteration k
$\tilde{v}_{k+1}$	$\triangleq$	Empirical state-action distribution induced during iteration k
$\bar{\tilde{v}}_{k+1}$	$\triangleq$	Abstract state action distribution obtained at iteration k by aggregating $\tilde{v}_{k+1}$
$\overline{\lambda}_{k+1}$	$\triangleq$	Updated state-action frequency at end of k-th iteration

Regret and Sample Complexity Analysis

$\overline{v}^+_{k+1}$	$\triangleq$	state-action distribution induced by $\overline{\pi}^+_{k+1}$
$\overline{\lambda}^*$	$\triangleq$	optimal abstract state-action distribution of the learning problem
λ^*	$\triangleq$	optimal state-action distribution of the learning problem
$\overline{v}^*_k$	$\triangleq$	optimal state-action distribution of the MDP $\overline{\mathcal{M}}^k_r$
$\nabla\widehat{\mathcal{L}}_{t_k-1}$	$\triangleq$	empirical gradient of objective $\overline{\mathcal{L}}_{\infty,\eta}$
$\nabla\widehat{\mathcal{L}}^+_{t_k-1}$	$\triangleq$	empirical optimistic gradient of objective $\overline{\mathcal{L}}_{\infty,\eta}$

B. Proofs Section 3

Proposition 1. *The geometry-aware estimation error $\bar{\xi}_n$ can be rewritten as a function of abstract states as:*

$$\bar{\xi}_n = \frac{1}{S} \sum_{\overline{s} \in \overline{\mathcal{S}}} E_{\overline{s}} \mid \hat{f}_n(\overline{s}) - f(\overline{s}) \mid \tag{15}$$

Proof. By the definition of $\bar{\xi}_n$ we have

$$\bar{\xi}_n \coloneqq \frac{1}{S} \sum_{s \in \mathcal{S}} |\hat{f}_n^A(s) - f(s)| \tag{23}$$

$$\overset{(1)}{=} \frac{1}{S} \sum_{s \in \mathcal{S}} \left| \frac{1}{T_n^+(\overline{s})} \sum_{s' \in [s]} T_n(s') \hat{f}_n(s') - f(s) \right| \tag{24}$$

$$\overset{(2)}{=} \frac{1}{S} \sum_{\overline{s} \in \overline{\mathcal{S}}} |[\overline{s}]| \left| \frac{1}{T_n^+(\overline{s})} \sum_{s' \in [\overline{s}]} T_n(s') \hat{f}_n(s') - f(\overline{s}) \right| \tag{25}$$

$$\overset{(3)}{=} \frac{1}{S} \sum_{\overline{s} \in \overline{\mathcal{S}}} E_{\overline{s}} \left| \hat{f}_n(\overline{s}) - f(\overline{s}) \right| \tag{26}$$

where in step (1) we used that $\hat{f}_n^A(s) \coloneqq \frac{1}{T_n^+([s])} \sum_{s' \in [\overline{s}]} T_n(s') \hat{f}_n(s')$, in step (2) we used f invariances and in step (3) we used that $E_{\overline{s}} \coloneqq |[\overline{s}]|$ and $\hat{f}_n(\overline{s}) \coloneqq \frac{1}{T_n^+(\overline{s})} \sum_{s \in [\overline{s}]} T_n(s) \hat{f}_n(s)$. □

Proposition 2 (Convex Upper Bound of $\bar{\xi}_n$)**.** *With probability at least $1 - \delta$ and n interactions with f we have:*

$$\bar{\xi}_n \leq \frac{C(n, \overline{S}, \delta)}{S} \sum_{\overline{s} \in \overline{\mathcal{S}}} \overline{\mathcal{F}}(\overline{s};\ T_n^+) \tag{16}$$

with $C(n, \overline{S}, \delta) \coloneqq \max \left\{ \log(n\overline{S}/\delta), \sqrt{\log(n\overline{S}/\delta)} \right\}$ *and*

$$\overline{\mathcal{F}}(\overline{s};\ T_n^+) \coloneqq E_{\overline{s}} \left(\sqrt{\frac{2\sigma^2(\overline{s})}{T_n^+(\overline{s})}} + \frac{F_{\max}}{T_n^+(\overline{s})} \right)$$

where $T_n^+(\overline{s}) \coloneqq T_n^+([s])$ are the visitation counts of $\overline{s}$.

Proof. To prove the statement, we employ a Bernstein type inequality from Lemma 7.37 in (Lafferty et al., 2008), where we upper bound $2/3$ by 1 in the second summand. Then, given a δ', we have that for all $t \in [n]$:

$$\mathbb{P}\left(\left| \hat{f}_t(\overline{s}) - f(\overline{s}) \right| \leq \sqrt{\frac{2\sigma^2(\overline{s}) \log(1/\delta')}{T_n^+(\overline{s})}} + \frac{F_{\max} \log(1/\delta')}{T_n^+(\overline{s})} \right) \geq 1 - \delta' \tag{27}$$

where $T_n^+(\overline{s}) = \sum_{s \in [\overline{s}]} T_n^+(s)$ and $\hat{f}_t(\overline{s}) = \frac{1}{T_t^+(\overline{s})} \sum_{s \in [\overline{s}]} T_t(s) \hat{f}_t(s)$. For notational simplicity, we will define:

$$B_{\delta'}(\overline{s}) = \sqrt{\frac{2\sigma^2(\overline{s}) \log(1/\delta')}{T_n^+(\overline{s})}} + \frac{F_{\max} \log(1/\delta')}{T_n^+(\overline{s})}$$

Then, by using a standard union bound, over $\overline{s} \in \overline{\mathcal{S}}$ and $t \in [n]$:

$$\mathbb{P}\left(\bigcap_{t \in [n]} \bigcap_{\overline{s} \in \overline{\mathcal{S}}} \left\{ \left| \hat{f}_t(\overline{s}) - f(\overline{s}) \right| \leq B_{\delta'}(\overline{s}) \right\} \right) = 1 - \mathbb{P}\left(\bigcup_{n \in [t]} \bigcup_{\overline{s} \in \overline{\mathcal{S}}} \left\{ \left| \hat{f}_t(\overline{s}) - f(\overline{s}) \right| > B_{\delta'}(\overline{s}) \right\} \right)$$

$$\geq 1 - \sum_{t\in[n]} \sum_{\overline{s}\in\overline{\mathcal{S}}} \mathbb{P}\left(\left|\hat{f}_t(\overline{s}) - f(\overline{s})\right| > B_{\delta'}(\overline{s})\right)$$
$$\overset{(4)}{\geq} 1 - n\overline{S}\delta' \tag{28}$$

where in step (4) we have used Equation (27). Then:

$$\begin{aligned}
1-\delta &\overset{(5)}{\leq} \mathbb{P}\left(\bigcap_{t\in[n]}\bigcap_{\overline{s}\in\overline{\mathcal{S}}}\left\{\left|\hat{f}_t(\overline{s}) - f(\overline{s})\right| \leq B_{\delta'}(\overline{s})\right\}\right) \\
&= \mathbb{P}\left(\bigcap_{t\in[n]}\bigcap_{\overline{s}\in\overline{\mathcal{S}}}\left\{E_{\overline{s}}\left|\hat{f}_t(\overline{s}) - f(\overline{s})\right| \leq E_{\overline{s}}B_{\delta'}(\overline{s})\right\}\right) \\
&\leq \mathbb{P}\left(\bigcap_{\overline{s}\in\overline{\mathcal{S}}}\left\{E_{\overline{s}}\left|\hat{f}_n(\overline{s}) - f(\overline{s})\right| \leq E_{\overline{s}}B_{\delta'}(\overline{s})\right\}\right) \\
&\overset{(6)}{\leq} \mathbb{P}\left(\sum_{\overline{s}\in\overline{\mathcal{S}}} E_{\overline{s}}\left|\hat{f}_n(\overline{s}) - f(\overline{s})\right| \leq \sum_{\overline{s}\in\overline{S}} E_{\overline{s}}B_{\delta'}(\overline{s})\right) \\
&= \mathbb{P}\left(S\bar{\xi}_n \leq \sum_{\overline{s}\in\overline{\mathcal{S}}} E_{\overline{s}}B_{\delta'}(\overline{s})\right) \\
&= \mathbb{P}\left(\bar{\xi}_n \leq \frac{1}{S}\sum_{\overline{s}\in\overline{\mathcal{S}}} E_{\overline{s}}B_{\delta'}(\overline{s})\right)
\end{aligned} \tag{29}$$

where in (5) we used Equation 28 and in step (6) we used the trivial fact that since the inequality in Equation 29 holds for every $\overline{s} \in \overline{\mathcal{S}}$ therefore it holds also for the respective sums over $\overline{\mathcal{S}}$. Then, we get that with probability at least $1-\delta$:

$$\begin{aligned}
\bar{\xi}_n &\leq \frac{1}{S}\sum_{\overline{s}\in\overline{\mathcal{S}}} E_{\overline{s}}\left(\sqrt{\frac{2\sigma^2(\overline{s})\log(n\overline{S}/\delta)}{T_n^+(\overline{s})}} + \frac{F_{\max}\log(n\overline{S}/\delta)}{T_n^+(\overline{s})}\right) \\
&\leq \frac{1}{S}\max\left\{\log(n\overline{S}/\delta), \sqrt{\log(n\overline{S}/\delta)}\right\}\sum_{\overline{s}\in\overline{\mathcal{S}}} E_{\overline{s}}\left(\sqrt{\frac{2\sigma^2(\overline{s})}{T_n^+(\overline{s})}} + \frac{F_{\max}}{T_n^+(\overline{s})}\right) \\
&= \frac{1}{S}\max\left\{\log(n\overline{S}/\delta), \sqrt{\log(n\overline{S}/\delta)}\right\}\sum_{\overline{s}\in\overline{\mathcal{S}}} \overline{\mathcal{F}}(\overline{s};\, T_n^+) \\
&= \frac{C(n,\overline{S},\delta)}{S}\sum_{\overline{s}\in\overline{\mathcal{S}}} \overline{\mathcal{F}}(\overline{s};\, T_n^+)
\end{aligned}$$

□

Proposition 3 (Tractable Convex Upper Bound of $\bar{\xi}_n$)**.** *Let an empirical state-action frequency at time t be defined as $\lambda_t(s,a) = T_t(s,a)/t$, then for $E_{\overline{s}}\eta \leq \frac{1}{n}$ we have:*

$$\bar{\xi}_n \leq \frac{2S}{\sqrt{n}} C(n,\overline{S},\delta)\left[\overline{\mathcal{L}}_{\infty,\eta}(\lambda_n) + \frac{\overline{S}F_{\max}}{S\sqrt{n\eta}}\right] \tag{18}$$

Proof. First, we define the following:

$$\overline{\mathcal{F}}(\overline{s}\,;\lambda) \coloneqq E_{\overline{s}}\left(\overline{\mathcal{F}}_1(\overline{s}\,;\lambda) + \overline{\mathcal{F}}_2(\overline{s}\,;\lambda)\right) \tag{30}$$

with

$$\overline{\mathcal{F}}_1(\overline{s}\,;\lambda) = \sqrt{\frac{2\sigma^2(\overline{s})}{\sum_{a\in\mathcal{A}}\sum_{s\in[\overline{s}]}\lambda(s,a)+\frac{1}{n}}}$$
$$\overline{\mathcal{F}}_2(\overline{s}\,;\lambda) = \frac{1}{\sqrt{n}}\frac{F_{\max}}{\sum_{a\in\mathcal{A}}\sum_{s\in[\overline{s}]}\lambda(s,a)+\frac{1}{n}}$$

and the following auxiliary objective:

$$\overline{\mathcal{L}}_n(\lambda) \coloneqq \frac{1}{S}\sum_{\overline{s}\in\overline{\mathcal{S}}}\overline{\mathcal{F}}(\overline{s}\,;\lambda) \tag{31}$$

Then, given an empirical state-action frequency at time t, defined as $\lambda_t(s,a) = T_t(s,a)/t$, we have that:

$$\begin{aligned}
\bar{\xi}_n &\overset{(7)}{\leq} \frac{2}{\sqrt{n}}C(n,\overline{S},\delta)\sum_{\overline{s}\in\overline{\mathcal{S}}}\overline{\mathcal{F}}(\overline{s}\,;\lambda_n)\\
&\overset{(8)}{=} \frac{2S}{\sqrt{n}}C(n,\overline{S},\delta)\overline{\mathcal{L}}_n(\lambda_n)\\
&\leq \frac{2S}{\sqrt{n}}C(n,\overline{S},\delta)\left[\overline{\mathcal{L}}_{\infty,\eta}(\lambda_n) + \frac{\overline{S}F_{\max}}{S\sqrt{n}\eta}\right]
\end{aligned}$$

where in step (7) we employed Equations 30 and 16 and the fact that $T_n^+(s) = \max(T_n(s),1) \geq (T_n(s)+1)/2$. In step (8) we used the definition of the objective in Equation 31 and in the last inequality we used Lemma D.4. □

C. Proofs Section 5

Lemma 5.1 (Variance Concentration (Panaganti & Kalathil, 2022)). *For all $\overline{s} \in \overline{\mathcal{S}}$, with probability at least $1-\delta$ we have*

$$\left|\sqrt{\sigma^2(\overline{s})} - \sqrt{\hat{\sigma}_t^2(\overline{s})}\right| \leq F_{\max}\sqrt{2\frac{\log(2\overline{S}t^2/\delta)}{T_t^+(\overline{s})}} := \alpha(t,\overline{s},\delta)$$

Proof. From (Panaganti & Kalathil, 2022), we have that for a fixed time $k \leq t$, it holds that with probability at least $1-\tilde{\delta}$:

$$\left|\sqrt{\sigma^2(\overline{s}_k)} - \sqrt{\hat{\sigma}_t^2(\overline{s}_k)}\right| \leq F_{\max}\sqrt{2\frac{\log(2\overline{S}/\delta)}{T_t^+(\overline{s})}}$$

Since we want the result to hold $\forall k \leq t$, we simply union bound over time which leads to the desired result. □

Proposition 4 (Gradient-Reward Invariances). *If f is invariant over states s and s', then $\forall s, s' \in [s], \forall a, a' \in \mathcal{A}$*

$$\nabla_\lambda \overline{\mathcal{L}}^+_{t_k-1}(\lambda)[s,a] = \nabla_\lambda \overline{\mathcal{L}}^+_{t_k-1}(\lambda)[s',a']$$

Proof. The proof simply follows by computing the gradient explicitly. In particular, we have that

$$\nabla_\lambda \widehat{\mathcal{L}}^+_{t_k-1}(\lambda)_{[s,a]} = \frac{-E_s\left[\sqrt{2\hat{\sigma}^2_{t_k-1}(\overline{s})} + \alpha(t_k-1,\overline{s},\delta)\right]}{2S\left(\sum_{s\in[s]}\sum_{b\in\mathcal{A}}\lambda(s,b) + E_{\overline{s}}\,\eta\right)^{\frac{3}{2}}} = \nabla_\lambda \widehat{\mathcal{L}}^+_{t_k-1}(\lambda)_{[s',a']}$$

Indeed, we notice that the gradient does not directly depend on the actions a and a' since in the denominator we are summing over all the possible actions. Furthermore, since both $s, s' \in [s]$, the gradient remains unchanged since it depends exclusively on the abstract state $\overline{s}$ and not directly on s and s'. In particular, since s and s' are in the same equivalence class, then the corresponding abstract state $\overline{s}$ is the same. □

D. Proofs Section 6

Theorem 6.1 (Regret Guarantee)**.** *If algorithm* GAE *is run with a budget of n samples and $\tau_k = 3k^2 - 3k + 1$ then w.p. at least $1-\delta$, it holds that:*

$$\mathcal{R}_n = \widetilde{\mathcal{O}}\left(\left(\frac{\Phi^{\frac{1}{2}} S^{\frac{1}{2}} A F_{\max}\sqrt{\sigma^2_{\max}}}{\eta^{\frac{5}{2}}}\right)\frac{1}{n^{1/3}}\right)$$

Proof. The analysis follows a classic Frank-Wolfe (FW) scheme analysis while taking into account the approximation error due to the optimistic estimate of the gradient, and the estimation error due to the gap between the asymptotic distribution associated with $\overline{\pi}^+_{k+1}$ and the induced empirical frequences $\overline{\lambda}_{k+1}$ at each iteration k. It diverges from previous analysis for non-episodic AE settings (Tarbouriech & Lazaric, 2019; Tarbouriech et al., 2020) in two main ways: (i) by leveraging ergodicity and hence uniqueness of the stationary distribution induced by Markovian stationary policies, we study the density estimation process via (Mutny et al., 2023, Lemma 5), (ii) we introduce a dependency on the geometric compression term Φ (Definition 2) in order to show the effect of compression on the final sample complexity result in Theorem 6.2.

As the regret in Convex RL is interpreted as a suboptimality gap (Mutti et al., 2023; 2022a; Tarbouriech & Lazaric, 2019), we first derive an upper bound on the approximation error, as defined in Equation 22, achieved at the end of iteration $k \in [K]$ of Algorithm 1. In the following we denote with t_{k+1} the number of samples gathered until the end of iteration k, formally $t_{k+1} := \sum_{j=1}^{k} \tau_j$, where τ_j is the number of steps policy π^+_{k+1} has been released in iteration k. Then by defining $\mathcal{L} := \overline{\mathcal{L}}_{\infty,\eta}$, we derive the following.

$$\begin{aligned}
\rho_{k+1} &:= \mathcal{L}(\overline{\lambda}_{k+1}) - \mathcal{L}(\overline{\lambda}^*) && (32)\\
&= \mathcal{L}((1-\beta_k)\overline{\lambda}_k + \beta_k\tilde{\overline{v}}_{k+1}) - \mathcal{L}(\overline{\lambda}^*) && (\text{where } \beta_k := \tau_k/(t_{k+1}-1))\\
&\overset{(9)}{\leq} \mathcal{L}(\overline{\lambda}_k) - \mathcal{L}(\overline{\lambda}^*) + \beta_k\langle\nabla\mathcal{L}(\overline{\lambda}_k), \tilde{\overline{v}}_{k+1} - \overline{\lambda}_k\rangle + C_\eta\beta_k^2\\
&= \mathcal{L}(\overline{\lambda}_k) - \mathcal{L}(\overline{\lambda}^*) + \beta_k\langle\nabla\mathcal{L}(\overline{\lambda}_k), \overline{v}^*_{k+1} - \overline{\lambda}_k\rangle + \beta_k\langle\nabla\mathcal{L}(\overline{\lambda}_k), \tilde{\overline{v}}_{k+1} - \overline{v}^*_{k+1}\rangle + C_\eta\beta_k^2\\
&\overset{(10)}{\leq} \mathcal{L}(\overline{\lambda}_k) - \mathcal{L}(\overline{\lambda}^*) + \beta_k\langle\nabla\mathcal{L}(\overline{\lambda}_k), \overline{\lambda}^* - \overline{\lambda}_k\rangle + \beta_k\langle\nabla\mathcal{L}(\overline{\lambda}_k), \tilde{\overline{v}}_{k+1} - \overline{v}^*_{k+1}\rangle + C_\eta\beta_k^2\\
&\overset{(11)}{\leq} (1-\beta_k)\rho_k + C_\eta\beta_k^2 + \beta_k\underbrace{\langle\nabla\mathcal{L}(\overline{\lambda}_k), \overline{v}^+_{k+1} - \overline{v}^*_{k+1}\rangle}_{\epsilon_{k+1}} + \beta_k\underbrace{\langle\nabla\mathcal{L}(\overline{\lambda}_k), \tilde{\overline{v}}_{k+1} - \overline{v}^+_{k+1}\rangle}_{\Delta_{k+1}} && (33)
\end{aligned}$$

where in step (9) we use C_η-smoothness of $\mathcal{L}$, in step (10) we use the definition of the update step of FW and the fact that $\overline{v}^*_{k+1}$ is optimal and in (11) we use again C_η-smoothness to bound $\langle\nabla\mathcal{L}(\tilde{\lambda}_k), \lambda^* - \tilde{\lambda}_k\rangle$. Note that in the first term ϵ_{k+1}, we take into account the discrepancy between the state-action distribution $\overline{v}^*_{k+1}$ induced by the optimal policy w.r.t. the MDP with the exact gradient as reward, and our exact solution $\overline{v}^+_{k+1}$ of the MDP with the optimistic gradient as reward. In the second term Δ_{k+1}, we take into account the error due to the gap between the state-action distribution $\overline{v}^+_{k+1}$ and the empirical distribution $\tilde{\overline{v}}_{k+1}$ induced by deploying policy π^+_{k+1} for τ_k steps. In the following, we upper bound independently the terms Δ_{k+1} and ϵ_{k+1}.

D.1. Upper Bound Δ_{k+1}

We derive a PAC guarantee on the density estimation error by using Lemma E.1 with $\eta_t(\overline{s}) = \eta(\overline{s}) = \overline{\lambda}(\overline{s})$, as in our case $\overline{\lambda}$ is fixed within one FW iteration and set $f^{\overline{s}}(\cdot) = f^{\overline{s}}_t(\cdot) := \frac{1}{T}\langle\cdot, \delta_{\overline{s}}\rangle$, where $\delta_{\overline{s}}$ is the vector of all zeros except in position s where it has a one and where $\langle\cdot,\cdot\rangle$ denotes the inner product. In particular, in this case, $f^{\overline{s}}$ corresponds to the evaluation functional of a probability distribution in state $\overline{s}$. Note that this functional is clearly linear as requested by the proposition as the inner product is linear. Furthermore, we can notice that $||f_{\overline{s}}||_\infty = \frac{1}{T}$ and hence we can restate the bound presented in the Lemma as follows:

$$\left|\frac{\sum_{t=1}^{\tau}\mathbb{I}\{\overline{s}_t = \overline{s}\}}{\tau} - \overline{\lambda}(\overline{s})\right| \leq \sqrt{\frac{2}{\tau}\log\left(\frac{2}{\delta'}\right)} \quad \text{with probability at least } 1-\delta'$$

Since we want the statement above to hold uniformly for every state $\overline{s} \in \overline{\mathcal{S}}$ and for every possible abstract policy, we set $\delta = \frac{\delta'}{\overline{SA}^{\overline{S}}}$ and apply a union bound, obtaining:

$$\left| \frac{\sum_{t=1}^{\tau} \mathbb{I}\{\overline{s}_t = \overline{s}\}}{\tau} - \overline{\lambda}(\overline{s}) \right| \leq \sqrt{\frac{2}{\tau} \log\left(\frac{2\overline{SA}^{\overline{S}}}{\delta}\right)} \quad \text{with probability at least } 1-\delta \tag{34}$$

Then, we can bound Δ_{k+1} as follows.

$$\begin{aligned}
\Delta_{k+1} &= \langle \nabla\mathcal{L}(\overline{\lambda}_k), \tilde{\overline{v}}_{k+1} - \overline{v}^{+}_{k+1} \rangle \\
&= -\frac{1}{2S} \sum_{\overline{s}} \frac{E_{\overline{s}} \sqrt{2\sigma^2(\overline{s})}}{\left(\overline{\lambda}_k(\overline{s}) + E_{\overline{s}}\eta\right)^{\frac{3}{2}}} \sum_{\overline{a}} \left(\tilde{\overline{v}}_{k+1}(\overline{s},\overline{a}) - \overline{v}^{+}_{k+1}(\overline{s},\overline{a})\right) \\
&\overset{(12)}{=} -\frac{1}{2S} \sum_{\overline{s}} \frac{E_{\overline{s}} \sqrt{2\sigma^2(\overline{s})}}{\left(\overline{\lambda}_k(\overline{s}) + E_{\overline{s}}\eta\right)^{\frac{3}{2}}} \left(\tilde{\overline{v}}_{k+1}(\overline{s}) - \overline{v}^{+}_{k+1}(\overline{s})\right) \\
&\overset{(13)}{\leq} -\frac{1}{2S} \sum_{\overline{s}} \frac{E_{\overline{s}}^{-\frac{1}{2}} \sqrt{2\sigma^2(\overline{s})}}{\eta^{\frac{3}{2}}} \left(\tilde{\overline{v}}_{k+1}(\overline{s}) - \overline{v}^{+}_{k+1}(\overline{s})\right) \\
&\overset{(14)}{\leq} \frac{\Phi^{1/2}\overline{S}\sqrt{\sigma^2_{\max}}}{S\eta^{3/2}} \left|\left|\tilde{\overline{v}}_{k+1} - \overline{v}^{+}_{k+1}\right|\right|_{\infty} \\
&\overset{(15)}{\leq} \frac{\Phi^{3/2}\sqrt{\sigma^2_{\max}}}{\eta^{3/2}} \sqrt{\frac{2}{\tau_k} \log\left(\frac{2\overline{SA}^{\overline{S}}}{\delta}\right)} \quad \text{with probability at least } 1-\delta
\end{aligned} \tag{35}$$

where in (12) we consider the state densities by summing over the actions, in (13) we lower bound $\overline{\lambda}(\overline{s}) \geq 0$, in (14) we use that $E_{\overline{s}} = \frac{1}{\Phi}$ in addition to taking the infinity norm and in (15) we use the PAC bound in Equation 34.

D.2. Upper Bound ϵ_{k+1}

Next, we define as $\nabla\widehat{\mathcal{L}}^{+}_{t_k-1}(\overline{\lambda})$ the empirical optimistic gradient it iteration k, defined as in Equation 20, but replacing the true variance with its empirical counterpart, and by using Lemma 5.1 we upper bound the true gradient $\nabla\mathcal{L}(\overline{\lambda})$ with a term containing the gradient estimate $\nabla\widehat{\mathcal{L}}_{t_k-1}(\overline{\lambda})$ (same as $\nabla\widehat{\mathcal{L}}^{+}_{t_k-1}(\overline{\lambda})$ but without the α). More explicitly, we have that:

$$\nabla\widehat{\mathcal{L}}^{+}_{t_k-1}(\overline{\lambda})(\overline{s},\overline{a}) = \frac{-E_{\overline{s}}\left[\sqrt{2\hat{\sigma}^2_{t_k-1}(\overline{s})} + \alpha(t_k-1,\overline{s},\delta)\right]}{2S\left(\overline{\lambda}(\overline{s}) + E_{\overline{s}}\eta\right)^{\frac{3}{2}}}$$

$$\nabla\widehat{\mathcal{L}}_{t_k-1}(\overline{\lambda})(\overline{s},\overline{a}) = \frac{-E_{\overline{s}}\sqrt{2\hat{\sigma}^2_{t_k-1}(\overline{s})}}{2S\left(\overline{\lambda}(\overline{s}) + E_{\overline{s}}\eta\right)^{\frac{3}{2}}}$$

$$\nabla\mathcal{L}(\overline{\lambda})(\overline{s},\overline{a}) = \frac{-E_{\overline{s}}\sqrt{2\sigma^2(\overline{s})}}{2S\left(\overline{\lambda}(\overline{s}) + E_{\overline{s}}\eta\right)^{\frac{3}{2}}}$$

In particular, we obtain that with probability at least $1-\delta$:

$$\nabla\widehat{\mathcal{L}}^{+}_{t_k-1}(\overline{\lambda})(\overline{s},\overline{a}) = \nabla\widehat{\mathcal{L}}_{t_k-1}(\overline{\lambda})(\overline{s},\overline{a}) - \frac{1}{2\Phi S} \frac{\alpha(t_k-1,\overline{s},\delta)}{\left(\overline{\lambda}(\overline{s}) + E_{\overline{s}}\eta\right)^{3/2}} \leq \nabla\mathcal{L}(\overline{\lambda})(\overline{s},\overline{a}) \quad \text{and} \tag{36}$$

$$\nabla\mathcal{L}(\overline{\lambda})(\overline{s},\overline{a}) \leq \nabla\widehat{\mathcal{L}}_{t_k-1}(\overline{\lambda})(\overline{s},\overline{a}) + \frac{1}{2\Phi S} \frac{\alpha(t_k-1,\overline{s},\delta)}{\left(\overline{\lambda}(\overline{s}) + E_{\overline{s}}\eta\right)^{3/2}} \tag{37}$$

Hence, we get that:

$$\begin{aligned}
\left\langle \nabla\mathcal{L}\left(\overline{\lambda}\right), \overline{v}^{+}_{k+1}\right\rangle &= \sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \nabla\mathcal{L}\left(\overline{\lambda}\right)(\overline{s},\overline{a}) \\
&\overset{(16)}{\leq} \sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \nabla\widehat{\mathcal{L}}_{t_k-1}\left(\overline{\lambda}\right)(\overline{s},\overline{a}) + \frac{1}{2\Phi S}\sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{\alpha\left(t_k-1,\overline{s},\delta\right)}{\left(\overline{\lambda}(\overline{s})+E_{\overline{s}}\eta\right)^{3/2}} \\
&\overset{(17)}{\leq} \sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \nabla\widehat{\mathcal{L}}^{+}_{t_k-1}\left(\overline{\lambda}\right)(\overline{s},\overline{a}) + \frac{1}{\Phi S}\sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{\alpha\left(t_k-1,\overline{s},\delta\right)}{\left(\overline{\lambda}(\overline{s})+E_{\overline{s}}\eta\right)^{3/2}} \\
&= \left\langle \nabla\widehat{\mathcal{L}}^{+}_{t_k-1}\left(\overline{\lambda}\right), \overline{v}^{+}_{k+1}\right\rangle + \frac{1}{\Phi S}\sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{\alpha\left(t_k-1,\overline{s},\delta\right)}{\left(\overline{\lambda}(\overline{s})+E_{\overline{s}}\eta\right)^{3/2}} \\
&\overset{(18)}{\leq} \left\langle \nabla\widehat{\mathcal{L}}^{+}_{t_k-1}\left(\overline{\lambda}\right), \overline{v}^{\star}_{k+1}\right\rangle + \frac{1}{\Phi S}\sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{\alpha\left(t_k-1,\overline{s},\delta\right)}{\left(\overline{\lambda}(\overline{s})+E_{\overline{s}}\eta\right)^{3/2}} \\
&\overset{(19)}{\leq} \left\langle \nabla\mathcal{L}\left(\overline{\lambda}\right), \overline{v}^{\star}_{k+1}\right\rangle + \frac{3}{2\Phi S}\sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{\alpha\left(t_k-1,\overline{s},\delta\right)}{\left(\overline{\lambda}(\overline{s})+E_{\overline{s}}\eta\right)^{3/2}}
\end{aligned} \tag{38}$$

where in (16) we used Equation 37, in (18) we used the optimality of $\overline{v}^{+}_{k+1}$ and in (17) and (19) we used Equation 36.

Using the definition of ϵ_{k+1} from Equation (33), by rearranging the terms in Equation (38) we get:

$$\begin{aligned}
\epsilon_{k+1} &= \langle \nabla\mathcal{L}(\overline{\lambda}_k), \overline{v}^{+}_{k+1} - \overline{v}^{*}_{k+1}\rangle \\
&\leq \frac{3}{2\Phi S}\sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{\alpha\left(t_k-1,\overline{s},\delta\right)}{\left(\overline{\lambda}_k(\overline{s})+E_{\overline{s}}\eta\right)^{3/2}} \\
&\leq \frac{3}{2\Phi S}\sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{\alpha\left(t_k-1,\overline{s},\delta\right)}{\left(E_{\overline{s}}\eta\right)^{3/2}}
\end{aligned} \tag{39}$$

where in the last inequality we simply lower-bound $\overline{\lambda}_k(\overline{s}) \geq 0$.

In the following we denote with $T_k(\overline{s}) \coloneqq T_{t_k-1}(\overline{s})$ the number of visits of state $\overline{s}$ from the start until iteration $k-1$ of the FW scheme (i.e., at time t_{k-1}). We now plug in the definition of $\alpha\left(t,\overline{s},\delta\right) = F_{\max}\sqrt{\frac{2\log(2\overline{S}t^2/\delta)}{T_t(\overline{s})}}$, coming from Lemma 5.1, in Equation 39, leading to:

$$\begin{aligned}
\epsilon_{k+1} &\leq \sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{3F_{\max}}{2\Phi S(\frac{1}{\Phi}\eta)^{3/2}} \sqrt{2\log\left(\frac{2\overline{S}\left(t_k-1\right)^2}{\delta}\right)} \frac{1}{\sqrt{T_k(\overline{s})}} \\
&= \sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{3\Phi^{1/2}F_{\max}}{2S\eta^{3/2}} \sqrt{2\log\left(\frac{2\overline{S}\left(t_k-1\right)^2}{\delta}\right)} \frac{1}{\sqrt{T_k(\overline{s})}} \\
&\leq c_0 \sum_{\overline{s},\overline{a}} \overline{v}^{+}_{k+1}(\overline{s},\overline{a}) \frac{1}{\sqrt{T_k(\overline{s})}} \\
&\overset{(20)}{=} \underbrace{c_0 \sum_{\overline{s},\overline{a}} \widetilde{\overline{v}}_{k+1}(\overline{s},\overline{a}) \frac{1}{\sqrt{T_k(\overline{s})}}}_{\gamma_k} + \underbrace{c_0 \sum_{\overline{s},a} \left(\overline{v}^{+}_{k+1}(\overline{s},\overline{a}) - \widetilde{\overline{v}}_{k+1}(\overline{s},\overline{a})\right) \frac{1}{\sqrt{T_k(\overline{s})}}}_{\Gamma_{k+1}}
\end{aligned} \tag{40}$$

with $c_0 = \frac{3\Phi^{1/2}F_{\max}}{2S\eta^{3/2}}\sqrt{2\log\left(2\frac{\overline{S}T^2}{\delta}\right)}$, where $T \coloneqq t_K - 1$ and K is the total number of FW iterations and where in (20) we simply added and substracted a term $c_0\sum_{\overline{s},\overline{a}} \widetilde{\overline{v}}^{+}_{k+1}(\overline{s},\overline{a})\frac{1}{\sqrt{T_k(\overline{s})}}$. We can now bound Γ_{k+1} similarly to Δ_{k+1} using Lemma

E.1. In particular, by upper-bounding $\frac{1}{\sqrt{T_k(\overline{s})}} \leq 1$, we get that with probability at least $1-\delta$, we have that:

$$\Gamma_{k+1} \leq c_0 \overline{S} \sqrt{\frac{2}{\tau_k} \log\left(\frac{2\overline{SA}^{\overline{S}}}{\delta}\right)}$$

Finally, by plugging in the bound of Γ_{k+1} we just derived into Equation 40, we have that with probability at least $1-\delta$:

$$\epsilon_{k+1} \leq c_0 \gamma_k + c_0 \overline{S} \sqrt{\frac{2}{\tau_k} \log\left(\frac{2\overline{SA}^{\overline{S}}}{\delta}\right)} \quad \text{where} \quad c_0 = \frac{3\Phi^{1/2} F_{\max}}{2S\eta^{3/2}} \sqrt{2\log\left(\frac{2\overline{S}T^2}{\delta}\right)} \quad \text{and} \quad \gamma_k = \sum_{\overline{s},\overline{a}} \frac{\tilde{\overline{v}}_{k+1}(\overline{s},\overline{a})}{\sqrt{T_k(\overline{s})}} \tag{41}$$

In the next steps, we will introduce an explicit time dependency on τ_k, which we will use to simplify the expression of the approximation error and finally perform a recursion leading to the final expression of the regret.
First, we recall from Equation 33 that with probability at least $1-\delta$, we have:

$$\rho_{k+1} \leq (1-\beta_k)\rho_k + C_\eta \beta_k^2 + \beta_k \epsilon_{k+1} + \beta_k \Delta_{k+1}$$

By plugging in the upper bounds of ϵ_{k+1} and Δ_{k+1} that we derived in (35) and (41), we get:

$$\begin{aligned}
\rho_{k+1} &\leq (1-\beta_k)\rho_k + C_\eta \beta_k^2 + \beta_k \left(c_0 \gamma_k + c_0 \overline{S} \sqrt{\frac{2}{\tau_k} \log\left(\frac{2\overline{SA}^{\overline{S}}}{\delta}\right)} \right) + \beta_k \frac{\Phi^{3/2}\sqrt{\sigma_{\max}^2}}{\eta^{3/2}} \sqrt{\frac{2}{\tau_k} \log\left(\frac{2\overline{SA}^{\overline{S}}}{\delta}\right)} \\
&= (1-\beta_k)\rho_k + C_\eta \beta_k^2 + \beta_k \underbrace{\left(c_0 \overline{S} + \frac{\Phi^{3/2}\sqrt{\sigma_{\max}^2}}{\eta^{3/2}} \right) \sqrt{\frac{2}{\tau_k} \log\left(\frac{2\overline{SA}^{\overline{S}}}{\delta}\right)}}_{c_1/\sqrt{\tau_k}} + \beta_k c_0 \gamma_k
\end{aligned} \tag{42}$$

Choosing $t_k = \tau_1 (k-1)^3 + 1$, we get:

$$\tau_k = t_{k+1} - t_k = \tau_1 (3k^2 - 3k + 1) \geq 3\tau_1 k^2 \quad \text{and} \quad \beta_k = \frac{\tau_k}{t_k - 1} \leq \frac{3}{k} \tag{43}$$

Hence, using (43) and the fact that $\tau_1 \geq 1$, we can further upper-bound together the second and third terms in Equation (42) as follows:

$$C_\eta \beta_k^2 + \beta_k \frac{c_1}{\sqrt{\tau_k}} \leq 9\frac{C_\eta}{k^2} + \frac{\sqrt{3}c_1}{k^2} =: \frac{b_\delta}{k^2} \tag{44}$$

Plugging this in (42) gives:

$$\rho_{k+1} \leq (1-\beta_k)\rho_k + \frac{b_\delta}{k^2} + \beta_k c_0 \gamma_k \tag{45}$$

We now want to compute the recursion on ρ in order to find the approximation error after K iterations. In order to do so, we will closely follow the steps from (Tarbouriech & Lazaric, 2019), which we will report for completeness. We choose $q \geq \left(\overline{S}/\tau_1\right)^{\frac{1}{3}} + 1$ for later use, and we introduce the sequence $(u_k)_{k \geq q}$ with $u_q = \rho_q$ and

$$u_{k+1} = \left(1 - \frac{1}{k}\right) u_k + \frac{b_\delta}{k^2} + \beta_k c_0 \gamma_k \tag{46}$$

From Inequality 45, we have $\rho_k \leq u_k$ and by induction we can see that $(u_k) \geq 0$. By rearranging the terms in (46), we get:

$$(k+1)u_{k+1} - ku_k = \frac{-u_k}{k} + \frac{b_\delta(k+1)}{k^2} + (k+1)\beta_k c_0 \gamma_k \leq \frac{b_\delta(k+1)}{k^2} + (k+1)\beta_k c_0 \gamma_k$$

Let $K \geq q$. We can now apply a telescoping sum starting at q and ending at K and exploit the fact that $\beta_k \leq 3/k \leq 6/(k+1)$ which leads to:

$$Ku_K - qu_q \leq 2b_\delta \sum_{k=q}^{K-1} \frac{1}{k} + 6c_0 \sum_{k=q}^{K-1} \gamma_k \leq 2b_\delta \log\left(\frac{K-1}{q-1}\right) + 6c_0 \sum_{k=q}^{K-1} \gamma_k \tag{47}$$

where the qu_q term appears as it corresponds to the first term of the telescoping sum.

By rearranging the terms in (47) and using that $\rho_K \leq u_K$ as previously observed, we have with probability at least $1-\delta$:

$$\rho_K \leq u_K \leq \frac{q\rho_q + 2b_\delta \log K}{K} + \frac{6c_0}{K} \sum_{k=q}^{K-1} \gamma_k \tag{48}$$

$$= \frac{\tau_1^{1/3}}{(t_K-1)^{1/3} + \tau_1^{1/3}} \left(q\rho_q + 2b_\delta \log K + 6c_0 \sum_{k=q}^{K-1} \gamma_k \right) \tag{49}$$

By employing (Tarbouriech & Lazaric, 2019, Lemma 6) with $q \geq \left(\overline{S}/\tau_1\right)^{1/3} + 1$ as previously set, we get the following upper bound:

$$\rho_K \leq \frac{\tau_1^{1/3}}{(t_K-1)^{1/3} + \tau_1^{1/3}} \left(q\rho_q + 2b_\delta \log K + 6c_0 \sqrt{\frac{\Sigma}{\tau_1}} \right)$$

where C_η is the smoothness constant of the objective function, which can be bounded as in Lemma E.2 and $\Sigma := \overline{S} \log(\frac{\tau_1 (K-1)^3}{\overline{S}})$.

To showcase better interpretability of the final result and in particular to show the advantage of exploiting symmetries as in our method, we now upper-bound the error ρ_K in such a way to make its dependence on the compression coefficient explicit. We choose the tightest possible q and by using the bounds for $q\rho_q$ and b_δ from Lemma D.3 and Lemma D.2 we get:

$$\begin{aligned}
\rho_K &= \widetilde{\mathcal{O}}\Bigg(\frac{\tau_1^{1/3}}{(t_K-1)^{1/3} + \tau_1^{1/3}} \Bigg[\frac{A\Phi^{\frac{1}{2}} F_{\max} \sqrt{\sigma^2_{\max}} \sqrt{S}}{\eta^{\frac{5}{2}}} + \frac{A\Phi^2 F_{\max} \sqrt{\sigma^2_{\max}} \sqrt{S}}{\eta^{\frac{5}{2}}} \\
&\qquad + \frac{3\Phi^{1/2} F_{\max}}{2S\eta^{\frac{3}{2}}} \sqrt{2\log\left(\frac{2\overline{S}T^2}{\delta}\right)} \sqrt{\frac{\overline{S}\log(\frac{\tau_1(K-1)^3}{\overline{S}})}{\tau_1}} \Bigg] \Bigg) \\
&= \widetilde{\mathcal{O}}\left(\left(\frac{A\Phi^{\frac{1}{2}} S^{\frac{1}{2}} F_{\max} \sqrt{\sigma^2_{\max}}}{\eta^{\frac{5}{2}}} \right) \frac{1}{t_K{}^{\frac{1}{3}}} \right) \\
&= \widetilde{\mathcal{O}}\left(\left(\frac{A\Phi^{\frac{1}{2}} S^{\frac{1}{2}} F_{\max} \sqrt{\sigma^2_{\max}}}{\eta^{\frac{5}{2}}} \right) \frac{1}{n^{\frac{1}{3}}} \right)
\end{aligned}$$

where in the last step we used that $n = \sum_{k=1}^K \tau_k = \sum_{k=1}^K \Theta(k^2) = \Theta(K^3) = \Theta(t_K)$ and where $T = t_K - 1$.

□

Theorem 6.2 (Sample Complexity of Geometric Estimation Objective)**.** *If algorithm* GAE *is run with* $\tau_k = 3k^2 - 3k + 1$*, for:*

$$n = \widetilde{\mathcal{O}}\left(\frac{\Phi^{\frac{3}{2}} S^{\frac{3}{2}} A^3 F_{\max}^3 (\sigma_{\max}^2)^{\frac{3}{2}}}{\eta^{\frac{15}{2}} \epsilon^3}\right)$$

samples, then we have that with probability at least $1 - \delta$*:*

$$\mathbb{P}\left(\left|\overline{\mathcal{L}}_{\infty,\eta}(\lambda_n) - \overline{\mathcal{L}}_{\infty,\eta}(\lambda^*)\right| \leq \epsilon\right) \geq 1 - \delta$$

Proof. The result simply follows by inverting the regret guarantee in Theorem 6.1. □

Lemma D.1. *We have the following bound for the approximation error at time* q*:*

$$\rho_q \leq \frac{\sqrt{\Phi}\sqrt{2\sigma_{\max}^2}}{\sqrt{\eta} \cdot q} + \frac{2b_\delta \log q}{q} + \frac{9 \cdot \Phi^{1/2} F_{\max}}{S\eta^{3/2}} \sqrt{2 \log\left(\frac{2\bar{S}q^6}{\delta}\right)}$$

Proof. We begin by noting that we can derive an equivalent bound of the error at time q in the same way as in Equation 48 by deploying the telescoping series from 1 to $q-1$ hence getting:

$$\begin{aligned}
\rho_q &\leq \frac{\rho_1 + 2b_\delta \log q}{q} + \frac{6c_0}{q} \sum_{k=1}^{q-1} \gamma_k \\
&= \frac{\rho_1 + 2b_\delta \log q}{q} + \frac{6c_0}{q} \sum_{k=1}^{q-1} \sum_{\overline{s},\overline{a}} \frac{\tilde{\overline{v}}_{k+1}(\overline{s},\overline{a})}{\sqrt{T_k(\overline{s})}} \\
&\leq \frac{\rho_1 + 2b_\delta \log q}{q} + \frac{6c_0}{q} \sum_{k=1}^{q-1} \underbrace{\sum_{\overline{s},\overline{a}} \tilde{\overline{v}}_{k+1}(\overline{s},\overline{a})}_{=1} \\
&\leq \frac{\rho_1 + 2b_\delta \log q}{q} + 6c_0
\end{aligned} \tag{50}$$

We now proceed in bounding the error ρ_1, by first recalling the definition we gave in Equation 22:

$$\begin{aligned}
\rho_1 &\coloneqq L\left(\bar{\lambda}_2\right) - \underbrace{L\left(\lambda^*\right)}_{\geq 0} \\
&\leq L\left(\bar{\lambda}_2\right) \\
&\leq \frac{\sqrt{\Phi}\sqrt{2\sigma_{\max}^2}}{\sqrt{\eta}}
\end{aligned}$$

where in the last inequality we use Lemma E.3.
Hence, by plugging this bound in Equation 50 and expliciting c_0 as given in Equation 41, we get:

$$\begin{aligned}
\rho_q &\leq \frac{\sqrt{\Phi}\sqrt{2\sigma_{\max}^2}}{\sqrt{\eta} \cdot q} + \frac{2b_\delta \log q}{q} + 6c_0 \\
&= \frac{\sqrt{\Phi}\sqrt{2\sigma_{\max}^2}}{\sqrt{\eta} \cdot q} + \frac{2b_\delta \log q}{q} + \frac{9 \cdot \Phi^{1/2} F_{\max}}{S\eta^{3/2}} \sqrt{2 \log\left(\frac{2\bar{S}q^6}{\delta}\right)}
\end{aligned}$$

□

Lemma D.2. *It holds that:*

$$b_\delta = \widetilde{\mathcal{O}}\left(\frac{A\Phi^2 F_{\max}\sqrt{\sigma_{\max}^2}\sqrt{S}}{\eta^{5/2}}\right)$$

Proof. Using the definition of b_δ from Equation 44, we have:

$$b_\delta = \sqrt{3}c_1 + 9C_\eta \tag{51}$$

$$\overset{(21)}{\leq} \sqrt{3}\left[\left(c_0\bar{S} + \frac{\Phi^{3/2}\sqrt{\sigma^2_{\max}}}{\eta^{3/2}}\right)\sqrt{2\log\left(\frac{2\bar{S}\overline{A}^{\bar{S}}}{\delta}\right)}\right] + 9\frac{A\sqrt{2\sigma^2_{\max}}\cdot\Phi^{5/2}}{\eta^{5/2}} \tag{52}$$

$$\overset{(22)}{=} \sqrt{3}\left(\left(\frac{3\Phi^{3/2}F_{\max}}{2\eta^{3/2}}\sqrt{2\log\left(\frac{\bar{S}T}{\delta}\right)} + \frac{\Phi^{3/2}\sqrt{\sigma^2_{\max}}}{\eta^{3/2}}\right)\sqrt{2\log\left(2\frac{2\bar{S}\overline{A}^{\bar{S}}}{\delta}\right)}\right) + 9\frac{A\sqrt{2\sigma^2_{\max}}\cdot\Phi^{5/2}}{\eta^{5/2}} \tag{53}$$

where in 21 we plug in the value of c_1 from Equation 42 and we use the upper bound of the smoothness constant from Lemma E.2 and in 22 we plug in the definition of c_0 from Equation 50. Hence, we have:

$$b_\delta = \widetilde{\mathcal{O}}\left(\left(\frac{3\Phi^{3/2}F_{\max}}{2\eta^{3/2}}\sqrt{2\log\left(\frac{2\bar{S}T}{\delta}\right)} + \frac{\Phi^{3/2}\sqrt{\sigma^2_{\max}}}{\eta^{3/2}}\right)\sqrt{\log\left(\frac{2\bar{S}}{\delta}\right) + \Phi S\log(\overline{A})} + \frac{A\sqrt{2\sigma^2_{\max}}\cdot\Phi^{5/2}}{\eta^{5/2}}\right) \tag{54}$$

$$= \widetilde{\mathcal{O}}\left(\frac{A\Phi^2F_{\max}\sqrt{\sigma^2_{\max}}\sqrt{S}}{\eta^{5/2}}\right) \tag{55}$$

□

Lemma D.3. *It holds that:*

$$q\rho_q = \widetilde{\mathcal{O}}\left(\frac{A\Phi^{\frac{1}{2}}\sqrt{\sigma^2_{\max}}F_{\max}\sqrt{S}}{\eta^{\frac{5}{2}}}\right)$$

Proof. From Lemma D.1 we get:

$$q\rho_q = \widetilde{\mathcal{O}}\left(\frac{\sqrt{\Phi}\sqrt{2\sigma^2_{\max}}}{\sqrt{\eta}} + 2b_\delta\log q + q\frac{9\cdot\Phi^{1/2}F_{\max}}{S\eta^{3/2}}\sqrt{2\log\left(\frac{2\bar{S}q^6}{\delta}\right)}\right)$$

$$= \widetilde{\mathcal{O}}\left(\frac{\sqrt{\Phi}\sqrt{2\sigma^2_{\max}}}{\sqrt{\eta}} + \frac{1}{3}2b_\delta\log\bar{S} + \bar{S}^{1/3}\frac{9\cdot\Phi^{1/2}F_{\max}}{S\eta^{3/2}}\right)$$

where in the second equality we used the fact that $q \geq (\overline{S}/\tau_1)^{1/3} + 1$ and $\frac{1}{\tau_1} \leq 1$. Next, we plug in the result of Lemma D.2 into b_δ in the last equation to get:

$$q\rho_q = \widetilde{\mathcal{O}}\left(\frac{\sqrt{\Phi}\sqrt{2\sigma^2_{\max}}}{\sqrt{\eta}} + \frac{A\Phi^2F_{\max}\sqrt{\sigma^2_{\max}}\sqrt{S}}{\eta^{5/2}}\log\bar{S} + \Phi^{\frac{1}{3}}S^{\frac{1}{3}}\frac{9\cdot\Phi^{1/2}F_{\max}}{S\eta^{3/2}}\right)$$

$$= \widetilde{\mathcal{O}}\left(\frac{A\Phi^{\frac{1}{2}}\sqrt{\sigma^2_{\max}}F_{\max}\sqrt{S}}{\eta^{\frac{5}{2}}}\right)$$

□

Lemma D.4. *For $E_{\overline{s}}\eta \leq \frac{1}{n}$ and $\forall\lambda \in \Lambda$, we can bound the non asymptotic objective function as:*

$$\overline{\mathcal{L}}_n(\lambda) \leqslant \overline{\mathcal{L}}_{\infty,\eta}(\lambda) + \frac{\overline{S}F_{\max}}{S\sqrt{n}\eta}$$

Proof.

$$\begin{aligned}
\overline{\mathcal{L}}_n(\lambda) &= \frac{1}{S}\sum_{\overline{s}\in\overline{S}} \overline{\mathcal{F}}(\overline{s},\lambda) \\
&= \frac{1}{S}\sum_{\overline{s}\in\overline{\mathcal{S}}} E_{\overline{s}}\left(\sqrt{\frac{2\sigma^2(\bar{s})}{\sum_{b\in A}\sum_{s\in[\bar{s}]}\lambda(s,b)+\frac{1}{n}}} + \frac{1}{\sqrt{n}}\frac{F_{\max}}{\sum_{b\in A}\sum_{s\in[\bar{s}]}\lambda(s,b)+\frac{1}{n}}\right) \\
&\overset{(23)}{\leq} \frac{1}{S}\sum_{\overline{s}\in\overline{\mathcal{S}}} E_{\overline{s}}\left(\sqrt{\frac{2\sigma^2(\bar{s})}{\sum_{b\in A}\sum_{s\in[\bar{s}]}\lambda(s,b)+E_{\overline{s}}\eta}} + \frac{1}{\sqrt{n}}\frac{F_{\max}}{\sum_{b\in A}\sum_{s\in[\bar{s}]}\lambda(s,b)+E_{\overline{s}}\eta}\right) \\
&= \overline{\mathcal{L}}_{\infty,\eta}(\lambda) + \frac{1}{S}\sum_{\overline{s}\in\overline{\mathcal{S}}} E_{\overline{s}}\frac{1}{\sqrt{n}}\frac{F_{\max}}{\sum_{b\in A}\sum_{s\in[\bar{s}]}\lambda(s,b)+E_{\overline{s}}\eta} \\
&\leq \overline{\mathcal{L}}_{\infty,\eta}(\lambda) + \frac{1}{S}\sum_{s\in[\bar{s}]}\frac{E_{\overline{s}}F_{\max}}{\sqrt{n}E_{\overline{s}}\eta} = \overline{\mathcal{L}}_{\infty,\eta}(\lambda) + \frac{\overline{S}F_{\max}}{S\sqrt{n}\eta}
\end{aligned}$$

where in (23) we used that $E_{\overline{s}}\eta \leq \frac{1}{n}$ and in the last inequality we lower bounded $\lambda(s,b) \geq 0$. □

E. Auxiliary Lemmas

Lemma E.1 ((Mutny et al., 2023), Lemma 5)**.** *Let $\{\eta_t\}_{t=1}$ be an adapted sequence of probability distributions on $\mathcal{X}, \mathcal{P}(\mathcal{X})$ with respect to filtration $\mathcal{F}_{t-1}$. Likewise let $\{f_t\}_{t=1}$ be an adapted sequence of linear functionals $f_t : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ s.t. $||f_t||_\infty \leq B_t$. Also, let $x_t \sim \eta_t$, and $\delta_t(x) = \mathbb{I}_{\{x_t = x\}}$, then:*

$$\mathbb{P}\left(\left|\sum_{t=1}^{T} f_t(\delta_t - \eta_t)\right| \geq \sqrt{2\sum_{t=1}^{T} B_t^2 \log\left(\frac{2}{\delta}\right)}\right) \leq \delta$$

Lemma E.2 (Bound of Smoothness constant)**.** *The smoothness constant C_η can be bounded by:*

$$C_\eta \leq \frac{A\sqrt{2\Phi^5}\sqrt{\sigma^2_{\max}}}{\eta^{5/2}}$$

Proof. Given the objective $\overline{\mathcal{L}}(\overline{\lambda})$, we have that its Hessian is made up of second order partial derivatives of the form

$$\frac{\partial^2\overline{\mathcal{L}}(\overline{\lambda})}{\partial\overline{\lambda}(\overline{s}',\overline{a}')^2} = \frac{3}{4}\frac{1}{\Phi S}\sqrt{\frac{2\sigma^2(\overline{s}')}{\left(\overline{\lambda}(\overline{s}')+E_{\overline{s}}\eta\right)^5}} \tag{56}$$

when both partial derivatives are taken w.r.t. the same coordinate, while the mixed second order partial derivatives are given by:

$$\frac{\partial\overline{\mathcal{L}}(\overline{\lambda})}{\partial\overline{\lambda}(s',a')\partial\overline{\lambda}(s'',a'')} = 0 \tag{57}$$

Hence the the Hessian $H(\overline{\mathcal{L}})$ is a diagonal matrix, thus containing only its eigenvalues. In particular, we can upper bound the value of the biggest eigenvalue as:

$$\max_{v\in\sigma(H(\overline{\mathcal{L}}))} v \leq \sum_{(\overline{s},\overline{a})\in\overline{\mathcal{S}}\times\overline{\mathcal{A}}} H(\overline{\mathcal{L}})(\overline{\lambda})_{((\overline{s},\overline{a}),(\overline{s},\overline{a}))} \tag{58}$$

which corresponds to summing all entries on the diagonal of the Hessian, and where $\sigma(A)$ stands for the spectrum of A. Hence, noting that $E_{\overline{s}} = \frac{1}{\Phi}$ we have that $\forall\ \overline{\lambda} \in \overline{\Lambda}$:

$$\max_{v\in\sigma(H(\overline{\mathcal{L}}))} v \leq \sum_{(\overline{s},\overline{a})\in\overline{\mathcal{S}}\times\overline{\mathcal{A}}} H(\overline{\mathcal{L}})(\overline{\lambda})_{((\overline{s},\overline{a}),(\overline{s},\overline{a}))} \tag{59}$$

$$= \sum_{(\overline{s},\overline{a})\in\overline{\mathcal{S}}\times\overline{\mathcal{A}}} \frac{\partial^2\overline{\mathcal{L}}(\overline{\lambda})}{\partial\overline{\lambda}(\overline{s}\,\overline{a})^2} \tag{60}$$

$$= \sum_{(\overline{s},\overline{a})\in\overline{\mathcal{S}}\times\overline{\mathcal{A}}} \frac{3}{4}\frac{1}{\Phi S}\sqrt{\frac{2\sigma^2(\overline{s})}{\left(\frac{1}{\Phi}\eta+\overline{\lambda}(\overline{s})\right)^5}} \tag{61}$$

$$\leq \sum_{(\overline{s},\overline{a})\in\overline{\mathcal{S}}\times\overline{\mathcal{A}}} \frac{1}{S}\frac{\sqrt{2\Phi^3\sigma^2(\overline{s})}}{\eta^{5/2}} \tag{62}$$

where in the last step we have used the fact that $\overline{\lambda}\in\overline{\Lambda}$.
Since $\overline{\mathcal{L}}$ is twice continuously differentiable over $\overline{\Lambda}$, by (Nesterov, 2014, Theorem 2.1.6) , we have that:

$$H(\overline{\mathcal{L}})\left(\overline{\lambda}\right) \preceq \sum_{(\overline{s},\overline{a})\in\overline{\mathcal{S}}\times\overline{\mathcal{A}}} \frac{1}{S}\frac{\sqrt{2\Phi^3\sigma^2(\overline{s})}}{\eta^{5/2}}\mathbb{I} \tag{63}$$

where $\mathbb{I}$ is the identity matrix, implying that for $C_\eta \leq \sum_{(\overline{s},\overline{a})\in\overline{\mathcal{S}}\times\overline{\mathcal{A}}} \frac{1}{S}\frac{\sqrt{2\Phi^3\sigma^2(\overline{s})}}{\eta^{5/2}}$, $\overline{\mathcal{L}}$ is C_η-smooth on $\overline{\Lambda}$.

From this, we have that:

$$C_\eta \leq \frac{\overline{A}\sqrt{2\Phi^5}\sqrt{\sigma^2_{\max}}}{\eta^{5/2}} \leq \frac{A\sqrt{2\Phi^5}\sqrt{\sigma^2_{\max}}}{\eta^{5/2}}$$

where in the last step we used the fact that $\overline{A}\leq A$, since we want to give a bound depending on the quantities of the original MDP. □

Lemma E.3. *The following bound holds for the asymptotic objective:*

$$\overline{\mathcal{L}}_{\infty,\eta}(\overline{\lambda}) \leq \frac{\sqrt{\Phi}\sqrt{2\sigma^2_{\max}}}{\sqrt{\eta}}$$

Proof. Using that $\overline{\lambda}(\overline{s},a) := \sum_{s\in[\overline{s}]}\lambda(s,a)$, we get:

$$\begin{aligned}
\overline{\mathcal{L}}_{\infty,\eta}\left(\overline{\lambda}\right) &= \frac{1}{\Phi S}\sum_{\overline{s}\in\overline{S}}\frac{\sqrt{2\sigma^2(\overline{s})}}{\sqrt{\overline{\lambda}(\overline{s})+E_{\overline{s}}\eta}} \\
&\leq \frac{1}{\Phi S}\sqrt{2\sigma^2_{\max}}\sum_{\overline{s}\in\overline{S}}\frac{1}{\sqrt{\overline{\lambda}(\overline{s})+E_{\overline{s}}\eta}} \\
&= \frac{1}{\Phi S}\sqrt{2\sigma^2_{\max}}\sum_{\overline{s}\in\overline{S}}\frac{1}{\sqrt{\overline{\lambda}(\overline{s})+E_{\overline{s}}\eta}} \\
&\overset{(24)}{\leq} \frac{\sqrt{2\sigma^2_{\max}}}{\Phi S}\sum_{\overline{s}\in\bar{S}}\frac{1}{\sqrt{E_{\overline{s}}\eta}} \\
&\overset{(25)}{=} \frac{\sqrt{2\sigma^2_{\max}}}{\Phi S}\frac{\bar{S}}{\sqrt{\frac{\eta}{\Phi}}} \\
&\overset{(26)}{=} \frac{\sqrt{2\sigma^2_{\max}}}{\sqrt{\Phi}S}\frac{\Phi S}{\sqrt{\eta}} \\
&= \frac{\sqrt{\Phi}\sqrt{2\sigma^2_{\max}}}{\sqrt{\eta}}
\end{aligned}$$

where in (24) we used that $\overline{\lambda}(\overline{s})\geq 0$, in (26) we used that $E_s=\frac{1}{\Phi}$ and in (26) we used that $\overline{S}=\Phi S$. □

Proposition 5 (Compression via Group Cardinality)**.** *Given a set of group-structured state symmetries* $\mathbb{G} = (\{L_g\}_{g\in G}, \cdot)$ *and* $\mathrm{Stab}(s) = \mathrm{Stab}(s')\ \forall s, s' \in \mathcal{S}$ *then:*

$$\Phi = \frac{|\mathrm{Stab}(s)|}{|G|}$$

where $\mathrm{Stab}(s) \coloneqq \{g \in G : L_g[s] = s\}$.

Proof. We consider the set $\mathcal{S}$, the group $\mathbb{G} = (\{L_g\}_{g\in G}, \cdot) = (G, \cdot)$ of transformations acting on $\mathcal{S}$ via the group action $* : G \times \mathcal{S} \to \mathcal{S}$. Due to Assumption 6.1 we have the following.

$$\frac{1}{\Phi} \overset{(27)}{=} E_s \tag{64}$$

$$= |[s]| \tag{65}$$

$$\overset{(28)}{=} |\mathrm{Orbit}(s)| \tag{66}$$

$$\overset{(29)}{=} \frac{|G|}{|\mathrm{Stab}(s)|} \tag{67}$$

where step (27) is due to Assumption 6.1, in step (28) we employ the definition of Orbit (Rotman, 2010, Chapter C-1, page 6), and in step (29) we leverage the Orbit-Stabilizer Theorem (Rotman, 2010, Theorem C-1.16). □

F. Experimental Details

In the following, we provide further details about the experiments carried out in this work. We first present the environments, their invariances and resulting abstract environments. Subsequently, we provide additional information regarding the implementation of GAE.

F.1. Pollutant Diffusion Process

MDP: We consider the problem of actively measuring the amount of pollutant released to the environment. The pollutant is released from a point source and spreads radially outwards through a diffusion process. The measurement setup is displayed in Figure 1. As can be seen, the measurement stations are aligned in two ways. **1)** Radially outwards from the point source. This allows to measure the variation of the pollutant amount in the radial direction. We will refer to a set of states that are aligned in this way as a ray. **2)** Along circles of different radii to measure the variation in the azimuthal direction. We will refer to a set of states aligned in this way as a circle. Each measurement station corresponds to a state in the MDP where the agent obtains noisy measurements of the pollutant amount. In our setup, there are a total of 30 circles and 8 rays, leading to a total of 240 states. The action space consists of five actions, $\{\mathrm{in}, \mathrm{out}, \mathrm{clockwise}, \mathrm{anticlockwise}, \mathrm{stay}\}$. We consider both deterministic and stochastic dynamics. In the deterministic case, if the agent chooses action in it moves one state closer to the point source along the ray (one further away for out). The actions leading to transitions along the circle are $\mathrm{clockwise}$ and $\mathrm{anticlockwise}$. The action stay, makes the agent to remain in the same state and repeat the experiment. In the case of stochastic transitions, the agent moves to it's intended state with probability q and with probability $1-q$ to another reachable state randomly chosen. In the experiments conducted we used $q = 0.98$.

Invariances of f: We consider the case where the diffusion process is radially symmetric, meaning that for all the states on a circle, the pollution is the same. Therefore, f is invariant over different rays, as illustrated already in 1.

Abstract MDP: The aforementioned invariances on f make it possible to define MDP homomorphisms h mapping the original MDP to abstract MDPs. In the experiments conducted, we considered three different homomorphisms, resulting in three different geometric compression terms Φ. The homomorphisms simply differ by how the rays are compressed. The first homomorphism h_1, maps two consecutive rays into one, resulting in a compression of the 8 rays into 4. The second homomorphism h_2, compresses the 8 rays into 2 and the third h_3 maps all 8 rays into 1 resulting in the compression illustrated in Figure 1. The state map ψ therefore maps states on consecutive rays together. The state-dependent action map ϕ_s corresponds to the identity map.

Implementation Details: The function $f(s)$, is modeled as increasing the closer the state is to the source. From the first equivalence class on the innermost circle to the 30th equivalence class consisting of the outermost circle, f decreases in steps of 300, starting at 9300. The noise $\nu(\overline{s})$ is modeled as increasing with function value as often large measurements are associated with larger variances. The corresponding standard deviations are hence also decreasing from the states closest to the source to the outermost states. The standard deviations decrease by steps of 100, starting at 3100. The distribution of $\nu(\overline{s})$ was taken to be a 0 mean Gaussian with the standard deviations given above. As an example, for a state s on the 21st circle we have $f(s) = 3000$ and $\nu(s) \sim \mathcal{N}(0, 1000)$. The smoothness parameter was chosen to be $\eta = 0.001$, and $\delta = 0.01$ for both, deterministic and stochastic dynamics. Furthermore, we found that in practice, a constant number of interactions $\tau_k = \tau$ for all the K iterations of GAE works well, especially for remarkably low τ. In this setting, we chose $\tau = 3$, which makes the algorithm more adaptive, resulting in the rapid exploration of different equivalence classes.To update the abstract state-action frequency $\overline{\lambda}_{k+1}$, we also use a constant update step of $0.005/\overline{S}$. The initial state of the agent was chosen on the outermost circle. n as 210, resulting in $K = 70$ iterations of GAE. All the experiments were repeated over 15 random seeds. The computational time was measured using a standard time library in Python. The main part of the computational time can be attributed to solving the MDP using value iteration. We applied the Bellman optimality operator until there was no change in the value function up to the 5th digit.

F.2. Toxicity of Chemical Compounds

MDP: As a second experiment we consider the experimental design problem of estimating the toxicity of chemical compounds. Similarly as in (Schreck et al., 2019; Dong et al., 2022; Thiede et al., 2022), we consider an MDP where a chemical compound is represented as a string where every character in the string stands for a base chemical element. The goal of the agent is to estimate the toxicity associated to all possible compounds that can be generated using the base chemical elements. In our setting we consider three base chemical elements `A`, `B`, and `C`. We limit the maximum length of

the string to 5. The state space therefore consists of all possible chemical compounds that have at most 5 base elements and the cardinality is $S = 363$. The action space consists of the three base elements, and an action that makes the agent stay in the same state $\mathcal{A} = \{\texttt{A}, \texttt{B}, \texttt{C}, \texttt{stay}\}$. By taking an action corresponding to a base element, the agent appends this element to the current compound. Once the agent reaches a compound with 5 base elements it can either choose to measure the toxicity of that compound again by picking the action $\texttt{stay}$ or pick a new base element to start another compound. The agent may therefore transition from one compound s of length l to another compound s' of length $l + 1$ if the first l base elements are the same.

Invariances of f: We assume that the toxicity of a chemical compound is invariant under permutations of the compound, such that $f(s) = f(s')$ if s is a permutation of s'. The abstract state-space therefore has a cardinality of $\overline{S} = 55$

Abstract MDP: These invariances on f again allow us to define an MDP homomorphism that maps the original MDP to an abstract one. In this case the state map ψ maps all the states which are equivalent up to permutation to one abstract state. The state-dependent action map ϕ_s is simply the identity. This results in an abstract MDP where the agent can transition from one abstract state $\overline{s}$ to another one $\overline{s}'$ if all the base compounds making up $\overline{s}$ are also contained in $\overline{s}'$ and the agent chooses the action corresponding to the chemical compound that is not yet in $\overline{s}'$. As an example consider $\overline{s} = \texttt{CABA}$ and $\overline{s}' = \texttt{AABCC}$, then the agent may transition from $\overline{s}$ to $\overline{s}'$ by choosing action $\texttt{C}$

Implementation Details: We model the toxicity of the chemical compounds and the noise as a piecewise constant functions of it's base chemical elements. For every $\texttt{A}$ in the compound, f is increased by 200, for every $\texttt{B}$, $\texttt{C}$ there is an increase of 400 and 600 respectively. Similarly, to the diffusion environment, we assume that higher measurements of toxicity are associated with higher standard deviations and that the noise has a Gaussian distribution with 0 mean. For every $\texttt{A}$ the standard deviation increases by 100, for every $\texttt{B}$, $\texttt{C}$, the standard deviations increase by 200 and 300 respectively. As an example consider the compound $\texttt{AABC}$, then $f(\texttt{AABC}) = 1400$ and $\nu(\texttt{AABC}) \sim \mathcal{N}(0, 700)$. We let GAE run for $n = 2400$ with a constant step size of $\tau_k = 20 \ \forall k \in [K]$, resulting in a total of $K = 80$ iterations of GAE. The smoothness constant chosen was $\eta = 0.0007$ and $\delta = 0.01$. To update the abstract state-action frequency $\overline{\lambda}_{k+1}$, we also use a constant update step of $0.005/\overline{S}$. The experiments were repeated over 15 different random seeds.