Journal of Machine Learning Research 24 (2023) 1-42 Submitted 12/22; Published 6/23

Convex Reinforcement Learning in Finite Trials

Mirco Mutti MIRCO.MUTTI@POLIMI.IT
Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20133 Milan, Italy

Riccardo De Santi* RDESANTI@ETHZ.CH
ETH Zürich
Rämistrasse 101, 8092 Zürich, Switzerland

Piersilvio De Bartolomeis* PDEBARTOL@ETHZ.CH
ETH Zürich
Rämistrasse 101, 8092 Zürich, Switzerland

Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20133 Milan, Italy

Editor: Kilian Weinberger

Abstract

Convex Reinforcement Learning (RL) is a recently introduced framework that generalizes the standard RL objective to any convex (or concave) function of the state distribution induced by the agent's policy. This framework subsumes several applications of practical interest, such as pure exploration, imitation learning, and risk-averse RL, among others. However, the previous convex RL literature implicitly evaluates the agent's performance over infinite realizations (or trials), while most of the applications require excellent performance over a handful, or even just one, trials. To meet this practical demand, we formulate convex RL in finite trials, where the objective is any convex function of the empirical state distribution computed over a finite number of realizations. In this paper, we provide a comprehensive theoretical study of the setting, which includes an analysis of the importance of non-Markovian policies to achieve optimality, as well as a characterization of the computational and statistical complexity of the problem in various configurations.

Keywords: Reinforcement Learning, Convex Reinforcement Learning, General Utilities, Finite Trials, Non-Markovian Policies

1. Introduction

Although Reinforcement Learning (RL, Sutton and Barto, 2018) provides a powerful and flexible framework to model sequential decision-making problems, many relevant applications do not fit naturally into the standard RL framework (Abel et al., 2021). Especially, the objective function of RL can be seen as a linear combination between a reward vector and the state distribution induced by the agent's policy. However, some applications cannot be cast into a linear objective function. Several works have thus extended the standard RL formulation to address non-linear objectives of practical interest.

* Riccardo and Piersilvio contributed to the works (Mutti et al., 2022b,a) that are extended by this paper.

This family of objectives includes *imitation learning* (Hussein et al., 2017; Osa et al., 2018), or the problem of finding a policy that minimizes the distance between the induced state distribution and the state distribution provided by experts' interactions (Abbeel and Ng, 2004; Ho and Ermon, 2016; Kostrikov et al., 2019; Lee et al., 2019; Ghasemipour et al., 2020; Dadashi et al., 2020; Kim et al., 2021; Freund et al., 2023), *risk-averse RL* (Garcıa and Fernández, 2015), in which the objective is sensitive to the tail behavior of the agent's policy (Tamar and Mannor, 2013; Prashanth and Ghavamzadeh, 2013; Tamar et al., 2015; Chow et al., 2015, 2017; Bisi et al., 2020; Zhang et al., 2021b; Greenberg et al., 2022; Eldowa et al., 2022; Bonetti et al., 2023; Hau et al., 2023), *pure exploration* (Hazan et al., 2019), where the goal is to find a policy that maximizes the entropy of the induced state distribution (Lee et al., 2019; Mutti and Restelli, 2020; Mutti et al., 2021; Zhang et al., 2021a; Guo et al., 2021; Liu and Abbeel, 2021b; Seo et al., 2021; Yarats et al., 2021; Mutti et al., 2022d,b; Nedergaard and Cook, 2022; Yang and Spaan, 2023; Tiapkin et al., 2023; Mutti, 2023), *diverse skills discovery* (Gregor et al., 2017; Eysenbach et al., 2018; Hansen et al., 2019; Sharma et al., 2020; Campos et al., 2020; Liu and Abbeel, 2021a; He et al., 2022; Zahavy et al., 2023; Mutti et al., 2022c), *constrained RL* (Altman, 1999; Achiam et al., 2017; Brantley et al., 2020; Miryoosefi et al., 2019; Qin et al., 2021; Yu et al., 2021; Bai et al., 2022; Germano et al., 2023), *active learning* in Markov decision processes (Tarbouriech and Lazaric, 2019; Tarbouriech et al., 2020; Wagenmaker and Jamieson, 2022; Mutny et al., 2023), and others.

All this large body of work has been recently unified into a unique broad framework, called *convex* RL (Hazan et al., 2019; Zhang et al., 2020; Zahavy et al., 2021; Geist et al., 2022), which generalizes the RL objective to any convex (or concave) function of the state distribution induced by the agent's policy. The convex RL problem has been shown to be largely tractable either computationally, as it admits a dual formulation akin to standard RL (Puterman, 2014), or statistically, as principled algorithms achieving sub-linear regret rates that are slightly worse than standard RL have been developed (Zhang et al., 2020; Zahavy et al., 2021).

However, the convex RL formulation presented in the previous literature implicitly evaluates the agent's performance over infinite realizations (or trials), as the objective function is computed on the (expected) state distribution. In practice, we can only draw a finite number of realizations instead, inducing an empirical state distribution that can be significantly far from its expectation (Weissman et al., 2003). In several applications, it is crucial to achieve a good performance in those finite number of realizations rather than in expectation over infinite trials. A typical example is the problem of learning a risk-sensitive policy for financial markets. Even if we have access to a simulator to train our policy on many realizations, we get just one realization when deploying the policy to the market. It is crucial that the deployed policy achieves a good performance in this single trial. Analogously, we might train a robot to imitate human behavior over tons of simulated realizations. However, we want our robot to effectively imitate the demonstrated behavior once it is deployed to the physical world, where we can typically get a few realizations. All of the previous considerations are meaningless in standard RL, as a linear objective function implies that the policy optimized over infinite realizations is also optimal when deployed over any finite number of realizations. Instead, we argue that accounting for the number of available trials is paramount in convex RL.

In this paper, we formulate the convex RL problem in *finite trials* to close the gap between the theoretical formulation of convex RL that is considered in the literature and the objective that should be optimized in practice.

In Section 3, we formalize the finite-trials convex RL objective as any convex (or concave) function of the empirical state distribution induced by the agent's policy over n realizations. Then, we compare the finite-trials formulation with its infinite-trials counterpart, demonstrating a crucial mismatch between their objective functions (Section 3.1). Especially, we show that a policy optimized for the infinite-trials objective can be significantly sub-optimal when evaluated over n trials, where the sub-optimality scales with a factor of $O(1/\sqrt{n})$ (Section 3.2). Supported by these results, we advocate for directly optimizing the finite-trials objective.

In Section 4, we provide a comprehensive study of the latter problem in its most extreme formulation, which is convex RL in a *single trial*. First, we demonstrate the importance of non-Markovian policies when optimizing the single-trial objective (Section 4.1). Especially, we show that the problem always admits a deterministic non-Markovian optimal policy, whereas the best policy within the space of Markovian policies has to be randomized. We prove that this randomization degrades the single-trial performance w.r.t. the optimal non-Markovian policy. Then, we provide a negative result on the tractability of computing the optimal non-Markovian policy when the environment is known, showing that the problem is NP-hard in general (Section 4.2). Finally, we provide an analysis of the statistical complexity of the corresponding learning problem, which demonstrates that $O(\sqrt{K})$ regret can be achieved while interacting with an unknown environment over K rounds (Section 4.3). The latter result gives some hope to the design of provably efficient algorithms that rely on approximate solvers to overcome the computational intractability of the problem.

In Section 5, we complement the previous results with the study of the convex RL problem in a *handful of trials*, where the objective is computed over $1 < n \ll \infty$ realizations. For the latter problem, we provide separate analyses for the settings in which the realizations are sampled in a sequence (Section 5.1) or in parallel, where we further differentiate between the perfect communication scenario (Section 5.2) and the scenario without communication (Section 5.3). Our results show that the sequential and the parallel communicating settings can be translated into an equivalent single-trial convex RL problem, thus inheriting analogous optimality of deterministic non-Markovian policies, as well as computational and statistical properties. Instead, the parallel non-communicating setting is crucially different from the others, as it generally admits an optimal stochastic non-Markovian policy.

Finally, we report a brief numerical validation of the presented claims (Section 6), as well as a discussion of the most relevant related work (Section 7) and interesting future directions (Section 8). Some of the proofs have been (partially) omitted for the sake of readability, and they are reported in Appendix A.

This paper unifies and extends the previous works (Mutti et al., 2022b,a). The former demonstrates the importance of non-Markovian policies to optimize a specific convex RL application, which is the state entropy maximization for pure exploration. The latter instead formulates convex RL in finite trials, highlighting the crucial mismatch between the finite-trials objective and the infinite-trials formulation that was previously considered in the literature. Specifically, here we extend the contributions of (Mutti et al., 2022b,a) as follows:

- We generalize the results in (Mutti et al., 2022b) from pure exploration to the broader convex RL framework;
- We improve the result in Lemma 4.6 of (Mutti et al., 2022b) by deriving a more informative version of the bounds (Lemma 5), which are now provided in a single cumulative expression rather than in per-step contributions;
- We sharpen the preliminary regret analysis of (Mutti et al., 2022a, Section 5.3) to derive the statistical complexity of convex RL in a single trial (Section 4.3);
- We report a novel study of convex RL in a handful of trials (Section 5), which was not analyzed in (Mutti et al., 2022b,a).

With this paper, we aim to provide a useful guide to convex RL in finite trials, and we hope to spark a research area that will bring convex RL closer to practical applications.

2. Background

In this section, we introduce the essential background notions for the remainder of the paper. We will denote with $[N]$ a set of integers $\{0, \dots, N-1\}$, and with $\mathbb{N}, \mathbb{R}$ natural and real numbers respectively. For two vectors $v = (v_1, \dots, v_n), u = (u_1, \dots, u_d)$ of any dimension, we denote with $v \oplus u = (v_1, \dots, v_n, u_1, \dots, u_d)$ their concatenation. When v, u have the same length $n = d$, we define the inner product $v \cdot u = \sum_{i=1}^{n} v_i u_i$.

For a measurable space $\mathcal{X}$, we will denote with $\Delta_{\mathcal{X}}$ the probability simplex over $\mathcal{X}$, and with $p \in \Delta_{\mathcal{X}}$ a probability measure over $\mathcal{X}$. For two probability measures $p, q \in \Delta_{\mathcal{X}}$, we define their ℓ^p-distance as

$$\|p - q\|_p := \Big(\sum_{x \in \mathcal{X}} \big|p(x) - q(x)\big|^p \Big)^{1/p},$$

where $\|p - q\|_\infty = \sup_{x \in \mathcal{X}} \big|p(x) - q(x)\big|$. We further define the Kullback-Leibler (KL) divergence between p and q as

$$\mathrm{KL}(p||q) := \sum_{x \in \mathcal{X}} p(x) \log \big(p(x)/q(x)\big).$$

Let X be a random variable having a cumulative density function $F_X(x) = Pr(X \leq x)$. We denote with $\mathbb{E}[X]$ its expected value, and its α-percentile is denoted as $\mathrm{VaR}_\alpha(X) = \inf \big\{x \mid F_X(x) \geq \alpha\big\} = F_X^{-1}(\alpha)$, where $\alpha \in (0, 1)$ is a confidence level, and VaR_α stands for Value at Risk (VaR) at level α. We denote the expected value of X within its α-percentile as $\mathrm{CVaR}_\alpha(X) = \mathbb{E}\big[X \mid X \leq \mathrm{VaR}_\alpha(X)\big]$, where CVaR_α stands for Conditional Value at Risk (CVaR) at level α.

2.1 Markov Decision Processes

A Markov Decision Process (MDP, Puterman, 2014) is a powerful framework to model sequential decision problems. A finite-horizon MDP is defined through the tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, T, \mu, r)$, in which $\mathcal{S}$ is a discrete space of $|\mathcal{S}| = S$ states, $\mathcal{A}$ is a discrete space of

$|\mathcal{A}| = A$ actions, $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition model, such that $P(s'|s,a)$ denotes the probability of transitioning to state s' by taking action a in state s, T is the horizon of an episode, $\mu \in \Delta_{\mathcal{S}}$ is the initial state distribution, and $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is a reward function that assigns a numeric reward $r(s,a)$ for taking action a in state s.[1]

A decision-maker, which is usually called the *agent*, interacts with the MDP over several episodes. In each episode, an initial state s_0 is drawn from the initial state distribution $s_0 \sim \mu$. The agent observes the state s_0 and picks an action a_0, therefore collecting the reward $r(s_0, a_0)$, while the MDP transitions to the next state s_1 drawn from $P(\cdot|s_0, a_0)$. Then, the agent observes s_1 and takes action a_1, collecting $r(s_1, a_1)$ while the MDP transitions to $s_2 \sim P(\cdot|s_1, a_1)$. This interaction process is repeated for each step $t \in [T]$ until the last reward $r(s_{T-1}, a_{T-1})$ is collected, and the episode ends. We call the sequence $h_T = (s_t, a_t)_{t=0}^{T-1}$ of states and actions encountered during the episode the *history* of interactions,[2] and we denote as $\mathcal{H}_T$ the set of all the histories of length T. We further denote as h_t a sub-history of length t and as $\mathcal{H}_t$ the set of all such sub-histories. Finally, we denote as $\mathcal{H} = \bigcup_{t=1}^{T} \mathcal{H}_t$ the set of all the histories up to length T.

2.2 Policies and Policy Spaces

The decision strategy of the agent, i.e., how the agent selects the action to take at each step, is defined through a *policy* π. A policy consists of a sequence of decision rules $\pi = (\pi_t)_{t\in[T]}$, one for each interaction step. In its most general formulation, a decision rule maps an history of interactions at step t with a probability distribution over actions $\pi_t : \mathcal{H}_t \to \Delta_{\mathcal{A}}$. The latter is called a *non-Markovian* decision rule. Instead, a *Markovian* decision rule maps the state at step t with a probability distribution over actions $\pi_t : \mathcal{S} \to \Delta_{\mathcal{A}}$, neglecting the previous history. A *deterministic* decision rule maps either the history or the state to a unique action, $\pi_t : \mathcal{H}_t \to \mathcal{A}$ or $\pi_t : \mathcal{S} \to \mathcal{A}$ respectively. We then define some relevant policy spaces as follows:

- A policy $\pi = (\pi_t)_{t\in[T]}$ composed of non-Markovian decision rules π_t is called a non-Markovian policy. Π_{NM} denotes the space of all the non-Markovian policies.
- A non-Markovian policy $\pi \in \Pi_{\text{NM}}$ composed of deterministic decision rules is called a deterministic non-Markovian policy. $\Pi_{\text{NM}}^{\text{D}} \subset \Pi_{\text{NM}}$ denotes the space of deterministic non-Markovian policies;
- A policy $\pi = (\pi_t)_{t\in[T]}$ composed of Markovian decision rules π_t is called a Markovian policy. Π_{M} denotes the space of all the Markovian policies;
- A Markovian policy $\pi \in \Pi_{\text{M}}$ composed of deterministic decision rules is called a deterministic Markovian policy. $\Pi_{\text{M}}^{\text{D}} \subset \Pi_{\text{M}}$ denotes the space of deterministic Markovian policies.

Finally, we will denote as Π a general policy space, such that it holds $\Pi_{\text{M}} \subset \Pi_{\text{NM}} \equiv \Pi$.

[1] In the following, we will sometimes define the reward as a *per-state* function $r : \mathcal{S} \to [0,1]$. Note that this is coherent with the previous definition by taking $r(s,a) = r(s), \forall a \in \mathcal{A}$.

[2] The sequence of rewards $(r_t)_{t=0}^{T-1}$ is omitted from the history definition, as it can be recovered from h_T through a deterministic mapping.

2.3 State Distributions

An agent interacting with an MDP over n episodes induces a sequence of n histories $(h_{T,i})_{i=1}^n$.[3] From those histories, we can compute the *empirical state distribution* $d_n \in \Delta_{\mathcal{S}}$ as

$$d_n(s) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \mathbb{1}(s_{t,i} = s),$$

such that $s_{t,i}$ is the state at the step t in the history $h_{T,i}$. A policy $\pi \in \Pi$ induces a particular probability measure p_n^π over the sequence of n histories $(h_{T,i})_{i\in[n]}$, and thus over the empirical state distribution d_n. Especially, we have

$$p_n^\pi\big((h_{T,i})_{i\in[n]}\big) = \prod_{i=1}^{n} \mu(s_{0,i}) \prod_{t=0}^{T-1} \pi(a_{t,i}|s_{t,i}) P(s_{t+1,i}|s_{t,i}, a_{t,i}).$$

With a slight overload of notation, we denote as $d_n \sim p_n^\pi$ an empirical state distribution obtained from a sequence of histories $(h_{T,i})_{i\in[n]} \sim p_n^\pi$. We further denote with $d_1 \sim p_1^\pi$ an empirical state distribution obtained from a single history $h_T \sim p_1^\pi$, and with $h_t \sim p_{1,t}^\pi$ a history of $t < T$ steps drawn from p_1^π.

Finally, we denote the expectation of the empirical state distribution under the policy π as $d^\pi(s) = \mathbb{E}_{d_n \sim p_n^\pi}[d_n(s)]$, such that $d^\pi \in \Delta_{\mathcal{S}}$ is called the *state distribution* induced by π.

2.4 Planning and Reinforcement Learning

The goal of an agent interacting with an MDP $\mathcal{M}$ is to find a decision policy that maximizes the expected sum of the rewards collected during an episode. Thus, the objective function of the agent can be written as[4]

$$\max_{\pi\in\Pi} \mathop{\mathbb{E}}_{\pi} \left[\sum_{t=1}^{T-1} r(s_t)\right] = \max_{\pi\in\Pi} \left(r \cdot d^\pi\right) =: \mathcal{J}(\pi), \tag{1}$$

and a policy $\pi^* \in \arg\max_{\pi\in\Pi} \mathcal{J}(\pi)$ is called an *optimal* policy. It is well-known (Puterman, 2014) that an MDP admits a deterministic Markovian optimal policy $\pi^* \in \Pi_{\mathrm{M}}^{\mathrm{D}}$. Moreover, solving (1) when the MDP is fully known, which is called the *planning* problem, is computationally efficient, as the optimal policy can be recovered from a linear program (Schweitzer and Seidmann, 1985; De Farias and Van Roy, 2003).

Reinforcement Learning (RL, Sutton and Barto, 2018) deals with the problem of *learning* a near-optimal policy from sampled interactions with an unknown MDP. Without having access to the transition model P, the RL agent optimizes a sampled-based version of (1) through running statistics computed on the collected episodes. This learning process is statistically efficient in tabular MDPs, as we can learn a policy $\widehat{\pi}$ such that $Pr(\mathcal{J}(\pi^*) - \mathcal{J}(\widehat{\pi}) \geq \epsilon) \leq \delta$ for any $\epsilon > 0, \delta \in (0, 1)$ by taking a polynomial number of episodes (Kearns and Singh, 2002; Kakade, 2003; Strehl and Littman, 2008).

[3] For each $h_{T,i}$, the first subscript denote the history length, the second subscript is the episode index. We will omit the first subscript when it is clear from the context.

[4] With a slight abuse of notation, we can equivalently represent a reward function $r : \mathcal{S} \to [0, 1]$ with a S-dimensional vector $r = (r(s))_{s\in\mathcal{S}} \in [0, 1]^S$, such that $r \cdot d^\pi = \sum_{s\in\mathcal{S}} r(s) d^\pi(s)$ is a well-defined inner product between the vectors $r \in [0, 1]^S$ and $d^\pi \in \Delta_{\mathcal{S}}$ which both lie in a subspace of $\mathbb{R}^S$.

2.5 Partially Observable MDPs

A Partially Observable Markov Decision Process (POMDP, Astrom, 1965; Kaelbling et al., 1998) generalizes the MDP model described in Section 2.1 to partially observable decision problems. A POMDP is described by $\mathcal{M}_\Omega := (\mathcal{S}, \mathcal{A}, P, T, \mu, r, \Omega, O)$, where $\mathcal{S}, \mathcal{A}, P, T, \mu, r$ are defined as in an MDP, Ω is a finite observation space, and $O : \mathcal{S} \to \Delta_\Omega$ is the observation function, such that $O(o|s)$ denotes the conditional probability of the observation $o \in \Omega$ when the POMDP is in state $s \in \mathcal{S}$. Crucially, while interacting with a POMDP the agent cannot observe the state $s \in \mathcal{S}$, but just the observation $o \in \Omega$. The performance of a policy π is defined as in an MDP (see (1)).

3. Convex Reinforcement Learning

Even though the RL formulation covers a wide range of sequential decision-making problems, several relevant applications cannot be expressed, as in (1), through the inner product between a reward vector r and a state distribution d^π (Abel et al., 2021; Silver et al., 2021). These include imitation learning, pure exploration, constrained problems, and risk-sensitive objectives, among others. Recently, a *convex* RL formulation (Hazan et al., 2019; Zhang et al., 2020; Zahavy et al., 2021; Geist et al., 2022) has been proposed to unify these applications in a unique general framework. In the latter framework, the agent interacts with an unknown *convex* MDP $\mathcal{M}_\mathcal{F} := (\mathcal{S}, \mathcal{A}, P, T, \mu, \mathcal{F})$, where $\mathcal{S}, \mathcal{A}, P, T, \mu$ are defined as in the MDP model described in Section 2.1, and the utility function $\mathcal{F}$ replaces the reward function r. For any $F < \infty$, the utility function $\mathcal{F} : \Delta_\mathcal{S} \to (-\infty, F]$ is a F-bounded concave function[5] of the state distribution[6] d^π that allows for a generalization of the learning objective, which becomes[7]

$$\max_{\pi \in \Pi} \Big(\mathcal{F}(d^\pi) \Big) =: \zeta_\infty(\pi). \tag{2}$$

To give a few examples, the utility $\mathcal{F}$ can be the entropy function in pure exploration setting (Hazan et al., 2019), a KL divergence in imitation learning (Ghasemipour et al., 2020), or some risk functional in risk-sensitive RL (Tamar et al., 2015). In Table 1, we recap some of the most relevant problems that fall under the convex RL formulation, along with their specific utility function $\mathcal{F}$. Note that the convex RL objective $\zeta_\infty(\pi)$ (2) reduces to the traditional RL objective (1) when $\mathcal{F}$ is a linear function.

Although convex RL is a generalization of the standard RL problem, previous works have demonstrated that convex RL enjoys similar computational and statistical properties. Hazan et al. (2019) note that the objective $\zeta_\infty(\pi)$ (2), being concave (convex) w.r.t. the state distribution d^π, can still be non-concave (non-convex) w.r.t. the policy parameters. However, they show that $\zeta_\infty(\pi)$ (2) admits a concave (convex) formulation that is instead convenient for optimization. While there exists an optimal Markovian policy $\pi^* \in \arg\max_{\pi \in \Pi} \mathcal{F}(d^\pi)$ for any $\mathcal{M}_\mathcal{F}$, the policy π^* can be stochastic (Hazan et al., 2019),

[5] In this context, we use the term *convex* RL to distinguish it from the standard *linear* RL objective (1). However, in practice, the function $\mathcal{F}$ can be either convex, concave, or even non-convex. In the following, we will assume $\mathcal{F}$ is concave if not mentioned otherwise.

[6] The utility function can be alternatively defined over state-action distributions.

[7] In general, problem (2) takes the form of a max problem for concave utilities $\mathcal{F} : \Delta_\mathcal{S} \to (-\infty, F]$, or a min problem for convex utilities $\mathcal{F} : \Delta_\mathcal{S} \to [F, +\infty)$. The meaning of the infinity subscript of ζ_∞ will be made clear in the next section.

Utility $\mathcal{F}$		Application	Infinite $\equiv$ Finite
$r \cdot d$	$r \in \mathbb{R}^S, d \in \Delta_{\mathcal{S}}$	RL	✓
$\|d - d_E\|_p^p$ $\mathrm{KL}(d \| d_E)$	$d, d_E \in \Delta_{\mathcal{S}}$	Imitation Learning	✗
$-d \cdot \log(d)$	$d \in \Delta_{\mathcal{S}}$	Pure Exploration	✗
$\mathrm{CVaR}_\alpha[r \cdot d]$ $r \cdot d - \mathbb{V}\mathrm{ar}[r \cdot d]$	$r \in \mathbb{R}^S, d \in \Delta_{\mathcal{S}}$	Risk-Averse RL	✗
$r \cdot d$, s.t. $\lambda \cdot d \leq c$	$r, \lambda \in \mathbb{R}^S, c \in \mathbb{R}, d \in \Delta_{\mathcal{S}}$	Linearly Constrained RL	✓
$-\mathbb{E}_z \mathrm{KL}(d_z \| \mathbb{E}_k d_k)$	$z \in \mathbb{R}^d, d_z, d_k \in \Delta_{\mathcal{S}}$	Diverse Skill Discovery	✗

Table 1: Relevant convex RL objectives and applications. The last column states the equivalence between infinite-trials and finite-trials settings (more details below) as derived in Proposition 6 (Appendix A).

differently from standard MDPs which always admit a deterministic optimal policy. Learning an optimal policy π^* from sampled interactions with $\mathcal{M}_{\mathcal{F}}$ has been demonstrated to be provably efficient, both in terms of sample complexity (Hazan et al., 2019) and regret (Zahavy et al., 2021).

3.1 Convex Reinforcement Learning in Finite Trials

In the previous section, we have denoted the convex RL objective as $\zeta_\infty(\pi)$ (2), with an infinity subscript, to underline that the state distribution d^π used to compute the objective can be only obtained asymptotically over the number of episodes (trials). Instead, in any practical simulated or real-world scenario, we can only draw a finite number of episodes $n \in \mathbb{N}$ with a policy π. From these episodes, we obtain an empirical state distribution $d_n \sim p_n^\pi$ rather than the actual state distribution d^π. This can cause a mismatch from the objective that is typically considered in convex RL (e.g., see Hazan et al., 2019; Zhang et al., 2020; Zahavy et al., 2021) and what can be optimized in practice. To overcome this mismatch, we generalize the convex RL problem to its finite-trials formulation.

Definition 1 (Finite-Trials Objective) *Let $\mathcal{M}_{\mathcal{F}}$ be a convex MDP and let $n \in \mathbb{N}$ a number of evaluation episodes. The corresponding n-trials convex RL objective is given by*

$$\max_{\pi \in \Pi} \left(\mathop{\mathbb{E}}_{d_n \sim p_n^\pi} \left[\mathcal{F}(d_n) \right] \right) =: \zeta_n(\pi). \tag{3}$$

In $\zeta_n(\pi)$ (3) the objective function is expressed in terms of the utility $\mathcal{F}$ associated to the empirical state distribution d_n obtained within n episodes, for which we then take the expectation by considering the probability of drawing d_n with the policy π.

In the following theorem, we show that the finite-trials convex RL objective is not equivalent to the usual formulation in general.

Theorem 1 (Finite-Trials Mismatch) *Let $\mathcal{M}_{\mathcal{F}}$ be a convex MDP and let $n \in \mathbb{N}$ the number of evaluation episodes. The corresponding convex RL $\zeta_\infty(\pi)$ (2) and finite-trials convex RL $\zeta_n(\pi)$ (3) objectives are not equivalent, i.e., $\zeta_\infty(\pi) \neq \zeta_n(\pi)$ in general.*

Proof Let us recall that $d^\pi = \mathbb{E}_{d_n \sim p_n^\pi}[d_n]$. Through Jensen's inequality, we can write

$$\zeta_\infty(\pi) = \mathcal{F}(d^\pi) = \mathcal{F}(\mathop{\mathbb{E}}_{d_n \sim p_n^\pi}[d_n]) \geq \mathop{\mathbb{E}}_{d_n \sim p_n^\pi}[\mathcal{F}(d_n)] = \zeta_n(\pi).$$

When $n < \infty$ and the utility function $\mathcal{F}$ is strictly concave (convex), the inequality is strict, meaning there is a mismatch between the two objectives. Instead, when $n \to \infty$, we have that

$$\lim_{n\to\infty} \zeta_n(\pi) = \lim_{n\to\infty} \mathop{\mathbb{E}}_{d_n \sim p_n^\pi}[\mathcal{F}(d_n)] = \mathop{\mathbb{E}}_{d^\pi \sim p_\infty^\pi}[\mathcal{F}(d^\pi)] = \zeta_\infty(\pi).$$

For this reason, we alternatively call $\zeta_\infty(\pi)$ (2) the *infinite-trials* convex RL objective. Finally, when $\mathcal{F}$ is a linear function, e.g., a reward function r, we can write

$$\zeta_\infty(\pi) = r \cdot d^\pi = r \cdot \mathop{\mathbb{E}}_{d_n \sim p_n^\pi}[d_n] = \mathop{\mathbb{E}}_{d_n \sim p_n^\pi}[r \cdot d_n] = \zeta_n(\pi),$$

which means that the mismatch between infinite and finite trials vanishes for the standard RL objective (1). ■

As a consequence of Theorem 1, optimizing the infinite-trials objective $\zeta_\infty(\pi)$ does not necessarily guarantee an optimal policy for the finite-trials objective $\zeta_n(\pi)$. This is a crucial difference between the standard RL problem, which does not suffer from this mismatch, and its convex generalization. Whereas in standard RL we can always design our learning algorithms in the well-founded, infinite-trials realm, in convex RL an algorithm designed for the infinite-trials formulation can output a policy that results sub-optimal when it is evaluated on a finite number of episodes. Specifically, in the next section, we characterize the approximation error of using the infinite-trials convex RL objective $\zeta_\infty(\pi)$ as a proxy of the finite-trials objective $\zeta_n(\pi)$.

3.2 Approximating the Finite-Trials Objective with Infinite Trials

Despite the evident mismatch between the finite trials and the infinite trials formulation of the convex RL problem, most existing works consider $\zeta_\infty(\pi)$ (2) as the standard objective, even if only a finite number of episodes can be drawn in practice. Thus, it is worth investigating how much we can lose by approximating a finite-trials objective with an infinite-trials one. First, we report a useful assumption on the structure of the function $\mathcal{F}$.

Assumption 1 (Lipschitz) *A function $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ is Lipschitz-continuous for some constant $L < \infty$, or L-Lipschitz for short, if it holds*

$$\big|\mathcal{F}(x) - \mathcal{F}(y)\big| \leq L\big\|x - y\big\|_1, \qquad \forall (x, y) \in \mathcal{X}^2.$$

Then, we provide an upper bound on the approximation error.

Theorem 2 (Approximation Error) *Let $\mathcal{M}_\mathcal{F}$ be a convex MDP with L-Lipschitz utility function $\mathcal{F}$, let $n \in \mathbb{N}$ be a number of evaluation episodes, let $\delta \in (0, 1]$ be a confidence level, let $\pi^\dagger \in \arg\max_{\pi\in\Pi} \zeta_n(\pi)$ and $\pi^\star \in \arg\max_{\pi\in\Pi} \zeta_\infty(\pi)$. Then, it holds with probability at least $1 - \delta$*

$$err := \big|\zeta_n(\pi^\dagger) - \zeta_n(\pi^\star)\big| \leq 4LT\sqrt{\frac{2S\log(4T/\delta)}{n}}.$$

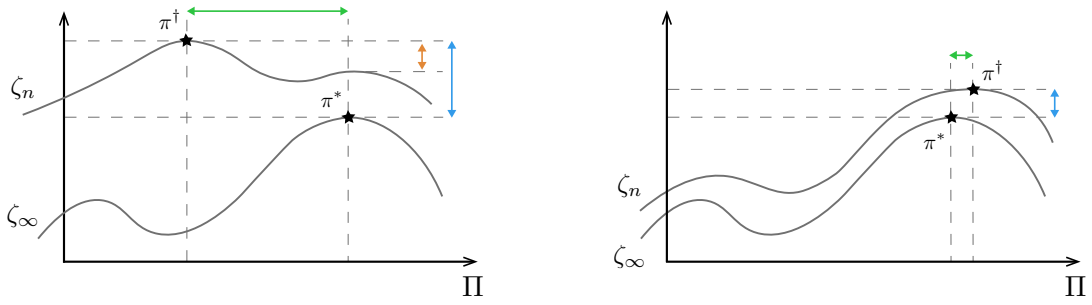


Figure 1: The two illustrations report an abstract visualization of ζ_n and ζ_∞ for small values of n (left) and large values of n (right) respectively. The green bar visualize the distance $\left\|d_n - d^{\pi^\star}\right\|_1$, in which $d_n \sim p_n^{\pi^\dagger}$. The blue bar visualize the distance $\left|\zeta_n(\pi^\dagger) - \zeta_\infty(\pi^\star)\right|$. The orange bar visualize the approximation error, i.e., the distance $\left|\zeta_n(\pi^\dagger) - \zeta_n(\pi^\star)\right|$.

Proof Sketch Starting from the definition of the approximation error, we can write

$$
\begin{aligned}
err := \left|\zeta_n(\pi^\dagger) - \zeta_n(\pi^\star)\right| &\leq \left|\zeta_n(\pi^\dagger) - \zeta_\infty(\pi^\dagger)\right| + \left|\zeta_\infty(\pi^\star) - \zeta_n(\pi^\star)\right| && (4)\\
&\leq \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\dagger}}\left[\left|\mathcal{F}(d_n) - \mathcal{F}(d^{\pi^\dagger})\right|\right] + \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\star}}\left[\left|\mathcal{F}(d_n) - \mathcal{F}(d^{\pi^\star})\right|\right] && (5)\\
&\leq \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\dagger}}\left[L\left\|d_n - d^{\pi^\dagger}\right\|_1\right] + \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\star}}\left[L\left\|d_n - d^{\pi^\star}\right\|_1\right] && (6)\\
&\leq 2L \max_{\pi \in \{\pi^\dagger, \pi^\star\}} \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi}}\left[\|d_n - d^{\pi}\|_1\right] && (7)
\end{aligned}
$$

where we obtain (5) from (4) through algebraic manipulations, we apply Assumption 1 to write (6), and we take the maximum over the policies in (7). Then, we apply an Höeffding-type concentration inequality for empirical distributions (Weissman et al., 2003, Theorem 2.1) to bound (7) with high probability. See Appendix A for complete derivations. ■

The previous result establishes an approximation error rate $err = O(LT\sqrt{S/n})$ that is polynomial in the number of evaluation episodes n. Unsurprisingly, the guarantees over the approximation error scale with $O(1/\sqrt{n})$, as one can expect the empirical distribution d_n to concentrate around its expected value for large n (Weissman et al., 2003). This implies that approximating the finite-trials objective $\zeta_n(\pi)$ with the infinite-trials $\zeta_\infty(\pi)$ can be particularly harmful in those settings in which n is necessarily small.

For instance, consider training a robot through a simulator and deploying the obtained policy in the real world, where the performance measure is often based on a single episode ($n = 1$). The performance we experience from the deployment can be much lower than the expected $\zeta_\infty(\pi)$, which might result in undesirable or unsafe behaviors.

However, Theorem 2 only reports an instance-agnostic upper bound, and it does not necessarily imply that there would be a significant approximation error in a specific instance, i.e., a specific convex MDP $\mathcal{M}_\mathcal{F}$. Nevertheless, in this paper, we argue that the upper bound of the approximation error is not vacuous in several relevant applications. We provide an illustrative numerical corroboration of this claim in Section 6.

Finally, in Figure 1, we report a visual representation[8] of the approximation error defined in Theorem 2. Notice that the finite-trials objective ζ_n converges uniformly to the infinite-trials objective ζ_∞ as a trivial consequence of Theorem 2. This is particularly interesting as it results in $\pi^\dagger$ converging to $\pi^\star$ in the limit of large n as shown Figure 1.

Having established a significant approximation error in optimizing the infinite-trials $\zeta_\infty(\pi)$ in place of the finite-trials $\zeta_n(\pi)$, in the following sections we will instead focus on the optimization of the finite-trials objective. In Section 4, we study the most extreme version of $\zeta_n(\pi)$ in which we have a single evaluation episode ($n = 1$). In Section 5, we study the optimization of $\zeta_n(\pi)$ for $n > 1$.

4. Convex RL in a Single Trial

In most real-world applications, the autonomous agent, which has often been trained in simulation, is deployed in the test environment over a single evaluation episode (or trial). For instance, in a financial application, we evaluate our agent in a single realization, as we cannot reset the market to understand the agent's performance over a handful, or infinite, trials. Similarly, an autonomous-driving vehicle has to maximize the utility in every trial, to ensure the safety of the passengers. Those examples motivate the study of the *single-trial* convex RL formulation, in which the convex utility is evaluated in expectation of a single realization. The corresponding objective function is given by

$$\max_{\pi \in \Pi} \left(\mathop{\mathbb{E}}_{d_1 \sim p_1^\pi} \left[\mathcal{F}(d_1) \right] \right) =: \zeta_1(\pi). \tag{8}$$

Whereas the infinite-trials convex RL problem enjoys favorable computational and statistical properties, we proved (see Theorem 2) that the resulting policy can be significantly sub-optimal w.r.t. the objective $\zeta_1(\pi)$ (8). Instead, it is worth investigating whether directly optimizing the single-trial objective $\zeta_1(\pi)$, thus avoiding the approximation error, is also suitable for optimization and statistically efficient.

First, in Section 4.1, we investigate the optimality of the common policy spaces, and we show that non-Markovian policies Π_{NM} are in general necessary to optimize $\zeta_1(\pi)$ (8). Then, in Section 4.2, we show that the corresponding optimization problem is, unfortunately, computationally intractable. Finally, in Section 4.3, we prove that the problem is at least statistically tractable, giving hope to design provably efficient methodologies that rely on approximate solvers to overcome the computational hardness of (8).

4.1 Optimality and The Importance of Non-Markovianity

First, we introduce a refined tool to evaluate the performance of a policy π beyond the value of the objective function $\zeta_1(\pi)$ (8), which will be convenient for our analysis.

Definition 2 (Value Gap) *Consider a policy $\pi \in \Pi$ interacting with a convex MDP $\mathcal{M}_\mathcal{F}$ over an episode of T steps. We define the* value gap $\mathcal{V}_T(\pi)$ *of the policy π as*

$$\mathcal{V}_T(\pi) = \mathcal{F}^* - \mathop{\mathbb{E}}_{d_1 \sim p_1^\pi} \left[\mathcal{F}\big(d_1\big) \right],$$

[8] Note that it is not possible to represent the objective functions in two dimensions in general. Nevertheless, we provide an abstract one-dimensional representation of the policy space to convey the intuition.

where $\mathcal{F}^* = \max_{\pi^* \in \Pi} \mathbb{E}_{d_1 \sim p_1^{\pi^*}} \left[\mathcal{F}\big(d_1\big)\right]$ *is the value achieved by an optimal policy* $\pi^* \in \Pi$ *over* T *steps. We further denote with* $\mathcal{V}_t(\pi, s)$ *the value gap induced by* π *over* t *steps starting from the state* s*, such that* $\mathcal{V}_T(\pi) = \mathbb{E}_{s \sim \mu}[\mathcal{V}_T(\pi, s)]$ *and* $\mathcal{V}_0(\pi, s) = 0, \forall s \in \mathcal{S}$*.*

The value gap essentially evaluates the policy π in relation to the optimal value $\mathcal{F}^*$ of $\zeta_1(\pi)$ (8) that can be achieved by any policy $\pi \in \Pi$ in $\mathcal{M}_\mathcal{F}$. It is interesting to assess whether a zero value gap can be achieved within the space of Markovian policies Π_M or non-Markovian policies Π_NM, and what is the corresponding minimal value gap otherwise.

Before formally stating the results, we introduce the following assumption to ease the notation without losing generality.[9]

Assumption 2 (Unique Optimal Action) *For every convex MDP* $\mathcal{M}_\mathcal{F}$ *and trajectory* $h_t \in \mathcal{H}$*, there exists a unique optimal action* $a^* \in \mathcal{A}$ *w.r.t.* $\zeta_1(\pi)$ (8)*.*

First, we show that the class of deterministic non-Markovian policies is sufficient for the minimization of the value gap, and thus for the maximization of $\zeta_1(\pi)$ (8).

Lemma 1 *For every convex MDP* $\mathcal{M}_\mathcal{F}$*, there exists a deterministic non-Markovian policy* $\pi_\text{NM} \in \Pi_\text{NM}^\text{D}$ *such that* $\pi_\text{NM} \in \arg\max_{\pi \in \Pi_\text{NM}} \mathbb{E}_{d_1 \sim p_1^\pi} \left[\mathcal{F}(d_1)\right]$*, which suffers zero value gap* $\mathcal{V}_T(\pi_\text{NM}) = 0$*.*

Proof It is straightforward to note the existence of a non-Markovian policy $\pi \in \Pi_\text{NM}$ such that $\mathcal{V}_T(\pi) = 0$, as the set Π_NM is the most general policy space. We need to prove that there exists one such policy that is deterministic. To this purpose, we reduce the convex MDP $\mathcal{M}_\mathcal{F}$ to an equivalent $\mathcal{M}_\ell = (\mathcal{S}_\ell, \mathcal{A}_\ell, P_\ell, T_\ell, \mu_\ell, r_\ell)$ that we call the *temporally-extended* MDP. We construct $\mathcal{M}_\ell$ from $\mathcal{M}_\mathcal{F}$ as follows:

- We build $\mathcal{S}_\ell$ by defining a state s_ℓ for each history h_t that can be induced in $\mathcal{M}_\mathcal{F}$, i.e., $s_\ell \in \mathcal{S}_\ell \iff h_t \in \mathcal{H}$;
- We keep $\mathcal{A}_\ell, P_\ell, T_\ell, \mu_\ell$ equivalent to $\mathcal{A}, P, T, \mu$ of $\mathcal{M}_\mathcal{F}$, where for the extended transition model $P_\ell(s'_\ell | s_\ell, a)$ we solely consider the last state in the history (corresponding to) s_ℓ to define the conditional probability to the next history (corresponding to) s'_ℓ;
- We define the reward function $r_\ell : \mathcal{S}_\ell \to \mathbb{R}$ such that $r_\ell(s_\ell) = \mathcal{F}(d_{s_\ell})$ for all the histories s_ℓ of length T and $r_\ell(s_\ell) = 0$ otherwise, where we denoted with d_{s_ℓ} the empirical state distribution induced by the history (corresponding to) s_ℓ.

From (Puterman, 2014), we know that there exists an optimal deterministic Markovian policy $\pi_\ell = (\pi_t : \mathcal{S}_\ell \to \mathcal{A}_\ell)_{t=0}^{T-1}$ for $\mathcal{M}_\ell$. Since $\mathcal{S}_\ell$ corresponds to the set of histories of the original MDP $\mathcal{M}_\mathcal{F}$, π_ℓ maps to a non-Markovian policy $\pi_\text{NM} \in \Pi_\text{NM}$ in $\mathcal{M}_\mathcal{F}$. Finally, it is straightforward to note that the optimality of π_ℓ for $\mathcal{M}_\ell$ implies the optimality of π_NM for $\zeta_1(\pi)$ (8), which concludes the proof.[10] ■

[9] Note that this assumption could be easily removed by partitioning the action space in h_t as $\mathcal{A}(h_t) = \mathcal{A}_{opt}(h_t) \cup \mathcal{A}_{sub-opt}(h_t)$, such that $\mathcal{A}_{opt}(h_t)$ are optimal actions and $\mathcal{A}_{sub-opt}(h_t)$ are sub-optimal, and substituting any term $\pi(a^*|h_t)$ with $\sum_{a \in \mathcal{A}_{opt}(h_t)} \pi(a|h_t)$ in the results.

[10] Note that the construction of the extended MDP cannot be computed in polynomial time, as it requires to enumerate all of the histories in $\mathcal{M}_\mathcal{F}$, and it only serves as a theoretical tool.

Then, in order to prove that the class of non-Markovian policies is also necessary for value gap minimization, it is useful to prove, as an intermediate step, that Markovian policies rely on randomization to optimize $\zeta_1(\pi)$ (8) in general.

Lemma 2 *Let $\pi_{\text{NM}} \in \Pi^{\text{D}}_{\text{NM}}$ be an optimal deterministic non-Markovian policy for $\zeta_1(\pi)$ (8) in the convex MDP $\mathcal{M}_{\mathcal{F}}$. For a fixed history $h_t \in \mathcal{H}_t$ ending in state s, the variance of the event of an optimal Markovian policy $\pi_{\text{M}} \in \arg\max_{\pi \in \Pi_{\text{M}}} \mathbb{E}_{d_1 \sim p_1^{\pi}} \big[\mathcal{F}(d_1)\big]$ taking $a^* = \pi_{\text{NM}}(h_t)$ in s is given by*

$$\mathbb{V}\text{ar}\left[\mathcal{B}(\pi_{\text{M}}(a^*|s_t))\right] = \mathop{\mathbb{V}\text{ar}}_{hs \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right]\right],$$

where $hs \in \mathcal{H}_t$ is any history of length t such that the final state is s, i.e., $hs := (h_{t-1} \in \mathcal{H}_{t-1}) \oplus s$, and $\mathcal{B}(x)$ is a Bernoulli with parameter x.

Proof Sketch We can prove the result through the Law of Total Variance (LoTV) (see Bertsekas and Tsitsiklis, 2002) on the variance of the event in which the optimal Markovian policy π_{M} takes the optimal action a^*. The latter gives

$$\mathbb{V}\text{ar}\left[\mathcal{B}(\pi_{\text{M}}(a^*|s_t))\right] = \mathop{\mathbb{E}}_{hs \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\mathbb{V}\text{ar}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right]\right] + \mathop{\mathbb{V}\text{ar}}_{hs \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right]\right].$$

Then, exploiting the determinism of π_{NM} through Lemma 1, it is straightforward to see that $\mathbb{E}_{hs \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\mathbb{V}\text{ar}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right]\right] = 0$, which concludes the proof.[11] ■

Unsurprisingly, Lemma 2 shows that whenever the optimal strategy for $\zeta_1(\pi)$ (8) in $\mathcal{M}_{\mathcal{F}}$ (i.e., the deterministic non-Markovian π_{NM}) requires to adapt its decision in a state s according to the history that led to it (hs), an optimal Markovian policy for the same objective (i.e., π_{M}) must necessarily be a stochastic policy. We can show that this randomization is harmful to the performance of the optimal Markovian policy, which incurs a positive value gap in general, meaning that it cannot match the performance of the optimal non-Markovian policy for $\zeta_1(\pi)$ (8). In the following result, we make use of the Lemma 2 to characterize lower and upper bounds to value gap of any Markovian policy that optimizes $\zeta_1(\pi)$ (8).

Lemma 3 *Let π_{M} be an optimal Markovian policy for $\zeta_1(\pi)$ (8) in the convex MDP $\mathcal{M}_{\mathcal{F}}$. It holds $\underline{\mathcal{V}}_T(\pi_{\text{M}}) \leq \mathcal{V}_T(\pi_{\text{M}}) \leq \overline{\mathcal{V}}_T(\pi_{\text{M}})$ such that*

$$\underline{\mathcal{V}}_T(\pi_{\text{M}}) = (\mathcal{F}^* - \mathcal{F}_2^*) \sum_{t=0}^{T-1} \mathop{\mathbb{E}}_{h_t \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\frac{\prod_{j=0}^{t-1} \pi_{\text{M}}(a_j^*|s_j)}{\pi_{\text{M}}(a_t^*|s_t)} \mathop{\mathbb{V}\text{ar}}_{hs_t \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\text{NM}}(a_t^*|hs_t))\right]\right]\right],$$

$$\overline{\mathcal{V}}_T(\pi_{\text{M}}) = (\mathcal{F}^* - \mathcal{F}_*) \sum_{t=0}^{T-1} \mathop{\mathbb{E}}_{h_t \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\frac{\prod_{j=0}^{t-1} \pi_{\text{M}}(a_j^*|s_j)}{\pi_{\text{M}}(a_t^*|s_t)} \mathop{\mathbb{V}\text{ar}}_{hs_t \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\text{NM}}(a_t^*|hs_t))\right]\right]\right],$$

where $\pi_{\text{NM}} \in \arg\max_{\pi \in \Pi^{\text{D}}_{\text{NM}}} \mathbb{E}_{d_1 \sim p_1^{\pi}}\big[\mathcal{F}(d_1)\big]$, and $\mathcal{F}_2^*, \mathcal{F}_*$ are given by

$$\mathcal{F}_2^* = \max_{\pi \in \{\Pi \setminus \pi_{\text{NM}}\}} \mathop{\mathbb{E}}_{d_1 \sim p_1^{\pi}}\big[\mathcal{F}(d_1)\big], \qquad \mathcal{F}_* = \min_{\pi \in \Pi} \mathop{\mathbb{E}}_{d_1 \sim p_1^{\pi}}\big[\mathcal{F}(d_1)\big].$$

[11] Note that the determinism of π_{NM} does not also imply $\mathbb{V}\text{ar}_{hs \sim p_t^{\pi_{\text{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right]\right] = 0$, as the optimal action $\overline{a} = \pi_{\text{NM}}(hs)$ may vary for different histories, which results in the inner expectations $\mathbb{E}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right]$ being either 1 (when $\overline{a} = a^*$) or 0 (when $\overline{a} \neq a^*$).

Proof Sketch The derivation of $\underline{\mathcal{V}}_T(\pi_{\mathrm{M}}), \overline{\mathcal{V}}_T(\pi_{\mathrm{M}})$ is based on computing, for each step t, the probability of the event in which π_{M} takes the optimal action $a^* = \pi_{\mathrm{NM}}(h_t)$, such that the value gap does not increase, and to bound the cost of taking a sub-optimal action optimistically and pessimistically for the lower and upper bounds respectively. Especially, starting from the definition of the value gap (Definition 2), we can write

$$\begin{aligned}\mathcal{V}_T(\pi_{\mathrm{M}}) &= \mathcal{F}^* - \mathop{\mathbb{E}}_{h_T \sim p_1^{\pi_{\mathrm{M}}}} \left[\mathcal{F}(d_{h_T})\right] \\ &\leq \mathcal{F}^* - \mathop{\mathbb{E}}_{s_0 \sim \mu} \Big[\pi_{\mathrm{M}}(a_0^*|s_0) \mathop{\mathbb{E}}_{s_1 \sim P(\cdot|s_0, a_0^*)} \left[\mathcal{V}_{T-1}(\pi_{\mathrm{M}}, s_1)\right] + \big(1 - \pi_{\mathrm{M}}(a_0^*|s_0)\big)\mathcal{F}_*\Big],\end{aligned}$$

where the value gap associated with the optimal action, which is taken with probability $\pi_{\mathrm{M}}(a_0^*|s_0)$, only depends on the expected value gap of the next step $\mathcal{V}_{T-1}(\pi_{\mathrm{M}}, s_1)$, whereas a sub-optimal action, which is taken with probability $1 - \pi_{\mathrm{M}}(a_0^*|s_0)$, incurs in the pessimistic value $\mathcal{F}_*$. By repeatedly applying this decomposition for all the remaining $T-1$ steps, we get

$$\mathcal{V}_T(\pi_{\mathrm{M}}) \leq (\mathcal{F}^* - \mathcal{F}_*) \sum_{t=0}^{T-1} \mathop{\mathbb{E}}_{h_t \sim p_{1,t}^{\pi_{\mathrm{NM}}}} \left[\Big(\prod_{j=0}^{t-1} \pi_{\mathrm{M}}(a_j^*|s_j)\Big)\big(1 - \pi_{\mathrm{M}}(a_t^*|s_t)\big)\right].$$

Finally, we note that $\pi_{\mathrm{M}}(a_t^*|s_t)(1 - \pi_{\mathrm{M}}(a_t^*|s_t)) = \mathbb{V}\mathrm{ar}\left[\mathcal{B}(\pi_{\mathrm{M}}(a_t^*|s_t))\right]$ from the definition of the Bernoulli distribution, and that $\mathbb{V}\mathrm{ar}\left[\mathcal{B}(\pi_{\mathrm{M}}(a_t^*|s_t))\right] = \mathbb{V}\mathrm{ar}_{hs_t \sim p_{1,t}^{\pi_{\mathrm{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\mathrm{NM}}(a_t^*|hs_t))\right]\right]$ through Lemma 2 to derive the upper bound $\overline{\mathcal{V}}_T(\pi_{\mathrm{M}})$. The lower bound $\underline{\mathcal{V}}_T(\pi_{\mathrm{M}})$ can be derived following similar steps but considering the optimistic value $\mathcal{F}_2^*$ whenever π_{M} takes a sub-optimal action. Complete derivations can be found in Appendix A. ■

The lower and upper bounds on the value gap of an optimal Markovian policy provided by Lemma 3 have a very similar structure. They are composed of an instance-dependent constant factor, i.e., $(\mathcal{F}^* - \mathcal{F}_2^*)$ and $(\mathcal{F}^* - \mathcal{F}_*)$ respectively, which accounts for the cost of taking a sub-optimal action, and a second factor that measures the randomization of the optimal Markovian policy across the time steps, and it relates this randomization to how much the optimal non-Markovian policy adapts its strategy according to the history, which is given by $\mathbb{V}\mathrm{ar}_{hs_t \sim p_{1,t}^{\pi_{\mathrm{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\mathrm{NM}}(a^*|hs_t))\right]\right]$.

Finally, through the combination of Lemma 1 and Lemma 3, we can state the following optimality result in single-trial convex RL.

Theorem 3 (Single-Trial Optimality) *For every convex MDP $\mathcal{M}_{\mathcal{F}}$, the space of deterministic non-Markovian policies $\Pi_{\mathrm{NM}}^{\mathrm{D}}$ is sufficient to optimize $\zeta_1(\pi)$ (8), while the space of Markovian policies Π_{M} incurs in a positive value gap $\mathcal{V}_T(\pi) \geq 0$ in general.*

The result of Theorem 3 highlights the importance of non-Markovianity for single-trial convex RL, as the class of Markovian policies is dominated by the class of non-Markovian policies. Most importantly, Lemma 3 shows that non-Markovian policies are strictly better than Markovian policies in several convex MDPs of practical interest, i.e., those in which an optimal Markovian policy has to be randomized to maximize $\zeta_1(\pi)$ (8). The intuition behind this result is that a Markovian policy would randomize to make up for the uncertainty over the history, whereas a non-Markovian policy does not suffer from this partial observability, and it can deterministically select an optimal action instead.

4.2 Computational Complexity

Having established the importance of non-Markovianity in dealing with convex RL in a single-trial regime, it is worth considering how hard it is to optimize the objective $\zeta_1(\pi)$ (8) within the space of non-Markovian policies. Especially, for a given convex MDP $\mathcal{M}_{\mathcal{F}}$, we aim at characterizing the complexity of the problem

$$\Psi_0 := \max_{\pi \in \Pi_{\text{NM}}} \zeta_1(\pi).$$

First, we provide a couple of useful definitions, whereas we leave to (Arora and Barak, 2009) an extended review of complexity theory.

Definition 3 (Many-to-one Reductions) *We denote as $A \leq_m B$ a many-to-one reduction from A to B.*

Definition 4 (Polynomial Reductions) *We denote as $A \leq_p B$ a polynomial-time (Turing) reduction from A to B.*

Then, we recall that Ψ_0 can be rewritten as the problem of finding an optimal Markovian policy for a convenient extended MDP $\mathcal{M}_\ell$ obtained from $\mathcal{M}_{\mathcal{F}}$ (see the proof of Lemma 1 for further details on how to build $\mathcal{M}_\ell$). We call this problem $\Psi_{0\ell}$ and we note that $\Psi_{0,\ell} \in \text{P}$, since a reward-maximizing policy can be computed in polynomial time for any MDP (Papadimitriou and Tsitsiklis, 1987). However, the following lemma shows that it does not exist a many-to-one reduction from Ψ_0 to $\Psi_{0\ell}$.

Lemma 4 *A reduction $\Psi_0 \leq_m \Psi_{0\ell}$ does not exist.*

Proof We can prove the result by showing that coding an instance of Ψ_0 in the representation required by $\Psi_{0\ell}$, which is an extended MDP $\mathcal{M}_\ell$, holds exponential complexity w.r.t. the input of Ψ_0, i.e., a convex MDP $\mathcal{M}_{\mathcal{F}}$. Indeed, to build the extended MDP $\mathcal{M}_\ell$ from $\mathcal{M}_{\mathcal{F}}$, we need to define the transition probabilities $P_\ell(s'_\ell | s_\ell, a_\ell)$ for every $s'_\ell \in \mathcal{S}_\ell, a_\ell \in \mathcal{A}_\ell, s_\ell \in \mathcal{S}_\ell$. Whereas the extended action space is $\mathcal{A}_\ell = \mathcal{A}$, we recall that the extended state space $\mathcal{S}_\ell$ is the set of all the histories $h_t \in \mathcal{H}$ of the convex MDP $\mathcal{M}_{\mathcal{F}}$. Thus, $\mathcal{S}_\ell$ has cardinality $|\mathcal{S}_\ell| = (SA)^T$ in general, which grows exponentially in T. ■

The latter result informally suggests that $\Psi_0 \notin \text{P}$. Indeed, we can now prove that Ψ_0 is NP-hard under the common assumption that $\text{P} \neq \text{NP}$.

Theorem 4 (Complexity of Single-Trial Convex MDPs) Ψ_0 *is* NP-hard.

Proof Sketch To prove the result, it is sufficient to show that there exists a problem $\Psi_c \in$ NP-hard that is at least as hard as Ψ_0. We obtain the latter through the chain of reductions

$$\Psi_0 \geq_m \text{specific class of POMDPs} \geq_p \text{3SAT}$$

starting with the original problem Ψ_0 of solving a single-trial convex MDP $\mathcal{M}_{\mathcal{F}}$, which is reduced to the problem of solving a particular class of POMDPs, which is then reduced to 3SAT, a notoriously NP-complete problem (Arora and Barak, 2009).

The first reduction $\Psi_0 \geq_m$ POMDP is obtained from a construction similar to the one of the extended MDP $\mathcal{M}_\ell$ described in the proof of Lemma 1. Specifically, we define

$\mathcal{S}_\ell, \mathcal{A}_\ell, P_\ell, T_\ell, \mu_\ell, r_\ell$ in the same way as in $\mathcal{M}_\ell$, and we further include an observation space Ω and an observation function O to obtain the POMDP $\mathcal{M}_{\ell,\Omega,O}$. The observation space is defined as the original state space $\Omega = \mathcal{S}$, whereas the observation function $O : \mathcal{S}_\ell \to \Omega$ takes as input an extended state $s_\ell \in \mathcal{S}_\ell$ (a history of the original $\mathcal{M}_\mathcal{F}$) and returns the observation $o \in \Omega$ that corresponds to the last state in the history s_ℓ. With this construction, we can map a reward-maximizing policy for $\mathcal{M}_{\ell,\Omega,O}$ to an optimal policy for $\mathcal{M}_\mathcal{F}$, which means we can solve the POMDP to solve Ψ_0.

Then, we carry out the reduction POMDP $\geq_p$ 3SAT by first reducing the problem of solving this specific class of POMDPs to the policy existence problem in the same class (Lusena et al., 2001, Section 3), and we rework the proof from (Mundhenk et al., 2000, Theorem 4.13) to reduce the latter policy existence problem to 3SAT. Finally, since 3SAT $\in$ NP-complete and $\Psi_0 \geq_p$ 3SAT, we can conclude that $\Psi_0 \in$ NP-hard. ■

Having established the computational hardness of solving convex MDPs in a single trial, i.e., maximizing the objective $\zeta_1(\pi)$ (8) within the set of non-Markovian policies, it is worth considering whether the problem admits at least a favorable statistical complexity.

4.3 Statistical Complexity

Although we provided a negative result on the computational complexity of solving a single-trial convex MDP exactly, reliable approximate solvers might be developed nonetheless. Thus, it is interesting to assess whether the corresponding learning problem is at least statistically efficient. For the purpose of this analysis, we assume to have access to a planning oracle that can solve any given convex MDP efficiently, while we speculate some potential directions for implementing approximate solutions in Section 8.

Assumption 3 (Planning Oracle) *Given a convex MDP $\mathcal{M}_\mathcal{F}$, the planning oracle returns a policy $\pi^* \leftarrow \texttt{Plan}(\mathcal{M}_\mathcal{F})$ such that $\pi^* \in \arg\max_{\pi \in \Pi} \zeta_1(\pi)$.*

With this assumption, we consider a learning setting in which the agent interacts with an unknown convex MDP $\mathcal{M}_\mathcal{F}$ over K episodes. In each of them, the agent deploys a policy $\widehat{\pi}_k$ to draw a history $h_{(k)}$ from $\mathcal{M}_\mathcal{F}$, receiving a single feedback $\mathcal{F}(d_{(k)})$ at the end of the episode, where $d_{(k)}$ is the empirical state distribution induced by $h_{(k)}$. Then, the agent makes use of the collected information to compute the policy $\widehat{\pi}_{k+1}$ to be deployed in the subsequent episode. In this online learning setting, the goal of the agent is typically to minimize the cumulative *regret* caused by deploying sub-optimal policies instead of an optimal decision strategy. The regret is defined as follows.

Definition 5 (Regret) *Let $\mathcal{M}_\mathcal{F}$ be an unknown convex MDP, and let* `Alg` *be a learning algorithm interacting with $\mathcal{M}_\mathcal{F}$. The K-episodes regret $\mathcal{R}(K)$ of* `Alg` *is given by*

$$\mathcal{R}(K) := \sum_{k=1}^{K} \Big(\zeta_1(\pi^*) - \zeta_1(\widehat{\pi}_k) \Big) = \sum_{k=1}^{K} \Big(\mathcal{F}^* - \mathop{\mathbb{E}}_{d_1 \sim p_1^{\widehat{\pi}_k}} \big[\mathcal{F}(d_1) \big] \Big),$$

where $\pi^* \in \arg\max_{\pi \in \Pi} \zeta_1(\pi)$, and $\widehat{\pi}_k$ is the policy deployed by `Alg` *in the episode k.*

Having defined the performance measure, we look for a learning algorithm that achieves a regret rate that is sub-linear in K, such as the $O(\sqrt{K})$ that can be achieved by online RL

Algorithm 1 UCBVI with history labels (Chatterji et al., 2021)

Input: convex MDP components $\mathcal{S}$, $\mathcal{A}$, T, μ, basis functions ϕ
initialize visitation counts $N_0(\cdot,\cdot) = 0$ and $N_0(\cdot,\cdot,\cdot) = 0$
randomly initialize $\widehat{\pi}_0$
for $k = 0, \dots$ **do**
  draw history $h_{(k)} \sim p_1^{\widehat{\pi}_k}$, collect $\mathcal{F}(d_{(k)})$, and update $N_k(\cdot,\cdot), N_k(\cdot,\cdot,\cdot)$
  compute the transition model $\widehat{P}_k(s'|s,a) = N_k(s,a,s')/N_k(s,a)$
  solve a regression problem $\widehat{\boldsymbol{w}}_k = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{d_{\boldsymbol{w}}}} \mathcal{L}_k(\boldsymbol{w})$ with a cross-entropy loss $\mathcal{L}_k$
  compute $\widehat{\mathcal{F}}_k(\cdot) = \widehat{\boldsymbol{w}}_k^\top \phi(\cdot)$ and build the optimistic convex MDP $\widehat{\mathcal{M}}_{\widehat{\mathcal{F}}}$
  call the planning oracle $\widehat{\pi}_{k+1} \leftarrow \texttt{Plan}(\widehat{\mathcal{M}}_{\widehat{\mathcal{F}}})$
end for

algorithms. However, the learning problem is inherently harder than standard RL. On the one hand, the feedback is sparse, as it only comes at the end of an episode. Previous works considered episode feedback in RL (e.g., Efroni et al., 2021), but they usually assume that the feedback is computed from an unknown reward function nonetheless. Instead, here we consider the feedback that comes from an unknown convex function of the empirical state distribution, which is akin to a non-Markovian reward, and indeed requires non-Markovian policies to be maximized.

A viable strategy (see Chatterji et al., 2021) is to estimate the utility function $\mathcal{F}$ with the feedback from the collected data, i.e., instantiating a regression problem to find the best approximation of $\mathcal{F}$ within a pre-specified function class, and then computing the policy that maximizes the approximated utility function. Here we assume that the true function $\mathcal{F}$ lies in a particular class of linear models, specified as follows.

Assumption 4 (Linear Realizability) *The function $\mathcal{F}$ is linearly-realizable if it holds*

$$\mathcal{F}(d_1) = \boldsymbol{w}_*^\top \phi(h),$$

where $h \in \mathcal{H}_T$ is an history that induces the empirical state distribution d_1, $\boldsymbol{w}_* \in \mathbb{R}^{d_{\boldsymbol{w}}}$ is a vector of parameters such that $\|\boldsymbol{w}_*\|_2 \le B$ for some known $B > 0$, and $\phi(h) = (\phi_j(h))_{j=1}^{d_{\boldsymbol{w}}}$ is a known vector of basis functions such that $\|\phi(h)\|_2 \le 1, \forall h \in \mathcal{H}_T$.

Note that the latter assumption does not reduce the problem to standard RL, as the features $\phi_j(h)$ are (possibly non-linear) functions of the whole history h, and we cannot decompose the utility in per-state rewards in general. Moreover, we do not lose generality by assuming linear realizability, since we can perfectly encode any history h through a sufficiently large features vector $\phi(h)$, while $\boldsymbol{w}_*$ induces an ordering over histories. However, as we shall see, the size $d_{\boldsymbol{w}}$ of the features vector negatively impacts the regret rate. Finally, as in several convex RL settings, the utility $\mathcal{F}$ is known, even assuming to have access to the feature vector is arguably reasonable. We leave as future work the problem of learning the features from data as well.

Now we have all of the ingredients to provide a result on the regret rate that can be achieved in single-trial convex RL. To this purpose, we reduce our problem setting to the once-per-episode RL framework discussed in (Chatterji et al., 2021). Then, we apply

their modified version of UCBVI (Azar et al., 2017) to work with history feedback. The procedure, for which we report an abstract pseudocode in Algorithm 1,[12] is a model-based algorithm that repeatedly solves a regression problem to approximate $\mathcal{F}$ from data and applies optimism to ensure the sufficient exploration. In the next theorem, we report its regret rate.

Theorem 5 (Regret Upper Bound) *Let $\mathcal{M}_{\mathcal{F}}$ be an unknown convex MDP with linearly-realizable utility $\mathcal{F}$. For any $\delta \in (0, 1]$, the K-episodes regret of* UCBVI with history labels *is upper bounded as*

$$\mathcal{R}(K) \leq O\Big(\Big[d_{\boldsymbol{w}}^{7/2} B^{3/2} T^2 S A^{1/2}\Big]\sqrt{K}\Big)$$

with probability $1 - \delta$.

Proof Sketch To prove the result, we show that the described online learning setting can be translated into the once-per-episode framework (Chatterji et al., 2021). The main difference between the setting in (Chatterji et al., 2021) and ours is that they assume a binary feedback $y_k \in \{0, 1\}$ coming from a logistic model

$$y_k | h_{(k)} = \begin{cases} 1 & \text{with prob.} \quad \sigma(\boldsymbol{w}_*^\top \phi(h_{(k)})) \\ 0 & \text{with prob.} \quad 1 - \sigma(\boldsymbol{w}_*^\top \phi(h_{(k)})), \end{cases} \qquad \sigma(x) = \frac{1}{1 + \exp(-x)}, \forall x \in \mathbb{R},$$

instead of our richer $\mathcal{F}(d_{(k)})$. To transform the latter in the binary reward y_k, we note that $\mathcal{F}(d_{(k)}) = \boldsymbol{w}_*^\top \phi(h_{(k)})$ through linear realizability (Assumption 4), then we filter $\mathcal{F}(d_{(k)})$ through a logistic model to obtain $y_k = \sigma\big(\mathcal{F}(d_{(k)})\big)$, which is then used as feedback for UCBVI (Algorithm 1). In this way, we can call Theorem 3.2 of (Chatterji et al., 2021) to obtain the same regret rate up to a constant factor[13] $C = \mathcal{F}^* - \mathcal{F}_*$, which is caused by the different range of per-episode contributions in the regret (see Definition 5). For detailed derivations and the complete regret upper bound see (Chatterji et al., 2021). ■

Theorem 5 demonstrates the existence of a principled algorithm achieving a $O(\sqrt{K})$ regret rate for convex RL in a single trial. We can conclude that single-trial convex RL is statistically efficient under the given assumptions. Since we are only providing an upper bound on the regret, it is fair to wonder what is the statistical barrier in this problem setting. Comparing our regret rate $O(d_{\boldsymbol{w}}^{7/2} B^{3/2} T^2 S\sqrt{AK})$ with the minimax regret of standard RL $O(\sqrt{TSAK})$, we notice that we are paying additional factors of T and S, while the rate is tight in A, K. Moreover, the linear-realizability assumption impacts the regret with additional $d_{\boldsymbol{w}}, B$ factors. Finally, it is worth noticing that when $\mathcal{F}$ is known, we have $d_{\boldsymbol{w}} = 1, B = 1$, and the regret rate reduces to $O(T^2 S\sqrt{AK})$.

Future works might study a lower bound on the regret for single-trial convex RL with linear realizability, to assess whether the additional factors w.r.t. standard RL are unavoidable. Other interesting directions include improving the procedure to exploit the richer feedback of our setting w.r.t. the one in (Chatterji et al., 2021), as well as incorporating in the analysis the error induced by approximate solvers in place of the planning oracle.

[12] See (Azar et al., 2017) and (Chatterji et al., 2021) for more detailed descriptions of the algorithm.
[13] Recall that $\mathcal{F}^* = \max_{\pi \in \Pi} \zeta_1(\pi)$ and $\mathcal{F}_* = \min_{\pi \in \Pi} \zeta_1(\pi)$.

5. Convex RL in a Handful of Trials

Whereas the real world is essentially single-trial, as we cannot truly reset a system to a previous state, most of the empirical work in RL optimizes the decision policy by drawing a batch of episodes from the environment, which is usually modeled through a simulator with reset. This practice is theoretically grounded in the standard RL setting since the policy that optimizes a linear utility over a batch of episodes maximizes the expected utility in a single trial as well (see the proof of Theorem 1). Instead, we demonstrated that this useful property does not hold when the utility is concave (or convex). Thus, it is worth providing a separate analysis for this setting, which we call convex RL in a *handful of trials*, to differentiate it from the single-trial formulation and to highlight that the number of evaluation episodes is $1 < n \ll \infty$ (typically dozens). We recall that the corresponding objective function for this setting is $\zeta_n(\pi)$ (3).

On the one hand, convex RL in a handful of trials is closer to the infinite-trials setting, as the empirical state distribution $d_n \sim p_n^\pi$ computed over n histories concentrates around its expected value d^π as n increases. However, the gap between the value $\zeta_n(\pi^\dagger)$ of an optimal infinite-trials policy $\pi^\dagger \in \arg\max_{\pi\in\Pi} \zeta_\infty(\pi)$ and the optimal value $\max_{\pi\in\Pi} \zeta_n(\pi)$ can still be significant, as it scales with $O(\sqrt{S/n})$ (see Theorem 2).

Even if in most of the convex RL applications the feedback may be available at the end of each episode, there might be good reasons to prefer a formulation with a handful of trials. For example, averaging the feedback over a handful of trials reduces its variability in general, to the benefit of the stability of the learning process. Moreover, optimizing the policy with each new piece of information, i.e., at the end of any episode, might cause a significant computational cost, which is usually called the *switching cost*. This is even more true when a single sweep of optimization might require an exponential cost (see Theorem 4). These considerations warrant the study of convex RL in a handful of trials.

In this section, we analyze three relevant modes to collect the batch of episodes in this setting, for which we provide specific results in terms of optimality, computational and statistical complexity. In Section 5.1, we consider the setting in which the histories h_i in the batch $(h_i)_{i=1}^n$ are collected sequentially, such that the agent can possibly exploit the information gathered in previous histories to adapt decisions. In Section 5.2, we consider the setting in which the batch $(h_i)_{i=1}^n$ is obtained through parallel sampling processes, but the workers can still communicate their respective state to others. Finally, in Section 5.3, we consider parallel sampling without communication between the workers, which means the histories h_i are sampled independently.

5.1 Sequential Sampling

In the setting with sequential sampling, the interaction process proceeds as follows. The state $s_{0,1}$ of the first history h_1 is drawn from μ, the agent takes an action $a_{0,1} \sim \pi(\cdot|s_{0,1})$ and the environment transitions to $s_{1,1} \sim P(\cdot|s_{0,1}, a_{0,1})$. This sequence is repeated for T steps until the episode 1 ends, and the initial state of the next history is sampled $s_{0,2} \sim \mu$. This process goes on until the last state $s_{T-1,n}$ of the last history h_n is reached, and the

agent receives a feedback $\mathcal{F}(d_n)$, where d_n is the empirical state distribution computed on the sampled batch of histories $(h_i)_{i\in[n]}$.[14]

Now we aim to characterize the computational and statistical complexity of the described setting, as well as whether non-Markovian policies are necessary to optimize the utility in a handful of trials. The following theorem shows that this setting can be actually translated to single-trial convex RL.

Proposition 1 *Let $\mathcal{M}_\mathcal{F}$ be a convex MDP, and let n be a number of episodes sampled sequentially. Optimizing the problem $\max_{\pi\in\Pi}\zeta_n(\pi)$ in $\mathcal{M}_\mathcal{F}$ is equivalent to solving a single-trial convex RL problem $\max_{\pi\in\Pi}\zeta_1(\pi)$ in a conveniently constructed convex MDP $\widetilde{\mathcal{M}}_\mathcal{F}$.*

Proof To prove the result, we start from the convex MDP $\mathcal{M}_\mathcal{F} = (\mathcal{S}, \mathcal{A}, P, T, \mu, \mathcal{F})$ to construct a convenient convex MDP $\widetilde{\mathcal{M}}_\mathcal{F}$, in which we see a sequence of histories in $\mathcal{M}_\mathcal{F}$ as a single long history. To construct the convex MDP $\widetilde{\mathcal{M}}_\mathcal{F} = (\mathcal{S}, \mathcal{A}, \widetilde{P}, \widetilde{T}, \mu, \mathcal{F})$ we proceed as follows:

- We keep the same $\mathcal{S}, \mathcal{A}, \mu, \mathcal{F}$ of the original convex MDP $\mathcal{M}_\mathcal{F}$;
- We set the horizon $\widetilde{T} = nT$;
- We construct a (time-inhomogeneous) transition model $\widetilde{P} = (\widetilde{P}_t)_{t=0}^{\widetilde{T}-1}$ such that each component is given by

$$\widetilde{P}_t(s'|s,a) = \begin{cases} \mu(s') & \text{if } t \bmod T = 0 \\ P(s'|s,a) & \text{if } t \bmod T \neq 0 \end{cases} \qquad \forall (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}.$$

With the latter construction, any sequence of histories $(h_i)_{i\in[n]}$ such that $h_i \in \mathcal{H}$ in $\mathcal{M}_\mathcal{F}$ can be mapped to an history $\widetilde{h} \in \widetilde{\mathcal{H}}$, where $\widetilde{\mathcal{H}}$ is the space of all the $\widetilde{T}$ steps histories in $\widetilde{\mathcal{M}}_\mathcal{F}$. Thus, an optimal policy $\widetilde{\pi}^* \in \arg\max_{\pi\in\Pi}\zeta_1(\pi)$ in $\widetilde{\mathcal{M}}_\mathcal{F}$ corresponds to an optimal policy $\pi^* \in \arg\max_{\pi\in\Pi}\zeta_n(\pi)$ in $\mathcal{M}_\mathcal{F}$. The last missing piece to prove the equivalence between the two settings is that we need a time-inhomogeneous transition model to construct $\widetilde{P}$. However, we can easily translate the latter in a time-homogeneous transition model defined over an extended state space $\widetilde{\mathcal{S}}$, in which each state is replicated for the different stages, such that $|\widetilde{\mathcal{S}}| = nS$. ■

The equivalence result of Theorem 1 implies that convex RL in a handful of trials with sequential sampling also admits a non-Markovian deterministic optimal policy. Hence, this setting also inherits the computational intractability of single-trial convex RL, as well as its favorable statistical complexity. Specifically, the upper bound to the regret (Theorem 5) deteriorates of a factor $O(n^{5/2})$ by replacing T with nT, S with nS, and K with K/n.

[14] Note that the histories $(h_i)_{i\in[n]}$ are not sampled independently, as the action taken at step $t \in [T]$ of the history $i \in [n]$ by the policy π depends on an all the previous steps of the previous histories.

5.2 Parallel Sampling with Communication

When we have access to parallel sampling with communication, the interaction process proceeds as follows. We deploy n parallel workers, each of them interacting with a copy of the environment, and we take the actions with a centralized policy. First, n initial states $(s_{0,1}, \dots, s_{0,n})$ are sampled independently from μ. Then, a vector of n actions is drawn from the centralized policy $(a_{0,1}, \dots, a_{0,n}) \sim \pi(\cdot|(s_{0,1}, \dots, s_{0,n}))$, and the copies of the environment transitions to their respective next states $(s_{1,1}, \dots, s_{1,n})$ independently, i.e., $s_{1,i} \sim P(\cdot|s_{0,i}, a_{0,i})$. This sequence is repeated until the last vector of states $(s_{T-1,1}, \dots, s_{T-1,n})$ is reached, and the agent collects a feedback $\mathcal{F}(d_n)$, where d_n is the empirical distribution induced by the parallel histories $(h_i)_{i\in[n]}$.[15]

As we shall see in the next theorem, the learning problem associated to this interaction process also translates to single-trial convex RL.

Proposition 2 *Let $\mathcal{M}_\mathcal{F}$ be a convex MDP, and let n be a number of episodes sampled in parallel with perfect communication. Optimizing the problem $\max_{\pi\in\Pi} \zeta_n(\pi)$ in $\mathcal{M}_\mathcal{F}$ is equivalent to solving a single-trial convex RL problem $\max_{\pi\in\Pi} \zeta_1(\pi)$ in a conveniently constructed convex MDP $\widetilde{\mathcal{M}}_\mathcal{F}$.*

Proof We prove the result as in the previous section. We start from the convex MDP $\mathcal{M}_\mathcal{F} = (\mathcal{S}, \mathcal{A}, P, T, \mu, \mathcal{F})$ to construct a convenient convex MDP $\widetilde{\mathcal{M}}_\mathcal{F}$, in which a vector of state in $\mathcal{M}_\mathcal{F}$ corresponds to a single state in $\widetilde{\mathcal{M}}_\mathcal{F}$, and a vector of actions in $\mathcal{M}_\mathcal{F}$ to a single action in $\widetilde{\mathcal{M}}_\mathcal{F}$. Specifically, to construct the convex MDP $\widetilde{\mathcal{M}}_\mathcal{F} = (\widetilde{\mathcal{S}}, \widetilde{\mathcal{A}}, \widetilde{P}, T, \mu, \mathcal{F})$ we proceed as follows:

- We keep the same $T, \mu, \mathcal{F}$ of the original convex MDP $\mathcal{M}_\mathcal{F}$;
- We construct the state space $\widetilde{\mathcal{S}}$ such that each $\widetilde{s} \in \widetilde{\mathcal{S}}$ corresponds to $(s_i)_{i=1}^n \in \mathcal{S}^n$;
- We construct the action space $\widetilde{\mathcal{A}}$ such that each $\widetilde{a} \in \widetilde{\mathcal{A}}$ corresponds to $(a_i)_{i=1}^n \in \mathcal{A}^n$;
- We construct the transition model $\widetilde{P}$ as

$$\widetilde{P}(\widetilde{s}\,'|\widetilde{s}, a) = \prod_{i=1}^{n} P(s_i'|s_i, a_i), \quad \forall(\widetilde{s}, \widetilde{a}, \widetilde{s}\,') \in \widetilde{\mathcal{S}} \times \widetilde{\mathcal{A}} \times \widetilde{\mathcal{S}}.$$

With this construction, any sequence of histories $(h_i)_{i=1}^n$ such that $h_i \in \mathcal{H}$ in $\mathcal{M}_\mathcal{F}$ can be mapped to an history $\widetilde{h} \in \widetilde{\mathcal{H}}$, where $\widetilde{\mathcal{H}}$ is the space of all the $\widetilde{T}$ steps histories in $\widetilde{\mathcal{M}}_\mathcal{F}$. Thus, an optimal policy $\widetilde{\pi}^* \in \arg\max_{\pi\in\Pi} \zeta_1(\pi)$ in $\widetilde{\mathcal{M}}_\mathcal{F}$ corresponds to an optimal policy $\pi^* \in \arg\max_{\pi\in\Pi} \zeta_n(\pi)$ in $\mathcal{M}_\mathcal{F}$, which proves the equivalence. ■

Exactly as for the sequential sampling, Theorem 2 demonstrates that convex RL in a handful of trials with parallel sampling with communication is not crucially different than convex RL in a single trial, hence displaying similar computational and statistical properties, while it admits an optimal deterministic non-Markovian policy. However, it is worth considering that the exponential growth of the state and action spaces $\widetilde{\mathcal{S}}, \widetilde{\mathcal{A}}$ means that the upper bound of the regret also scales with $O((SA)^n)$.

[15] Note that the histories $(h_i)_{i\in[n]}$ are not sampled independently, as the vector of actions taken at step $t \in [T]$ by the policy π depends on the previous steps of all the histories.

5.3 Parallel Sampling without Communication

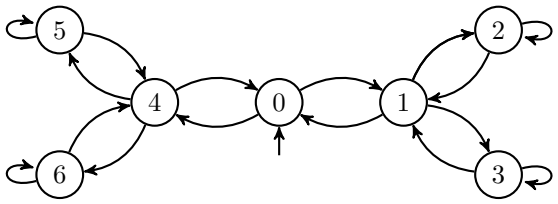
In the setting in which we have access to parallel sampling, but the workers cannot communicate their state to the others, the interaction process is as follows. As in the previous section, we deploy n parallel workers, each of them equipped with their own copies of the environment and the policy, which is thus decentralized. First, an initial state $s_{0,i}$ is sampled independently for each episode $i \in [n]$. Then, each worker draws an action $a_{0,i} \sim \pi(\cdot|s_{0,i})$ with their copy of the policy, so that the sampled actions $(a_{0,i})_{i\in[n]}$ only depends on the history of their respective episode. Finally, each worker updates their state by drawing $s_{1,i} \sim P(\cdot|s_{0,i}, a_{0,i})$ with their copy of the environment. This sequence is repeated by each worker until the episode ends, and the agent receives centralized feedback $\mathcal{F}(d_n)$ where d_n is the empirical state distribution computed on the independent histories $(h_i)_{i\in[n]}$.

It is worth wondering whether convex RL in a handful of trials with the described sampling process can also be translated into an equivalent single-trial convex RL problem. The following proposition demonstrates that this setting is crucially different from convex RL in a single trial.

Proposition 3 *Let $\mathcal{M}_\mathcal{F}$ be a convex MDP, and let n be a number of episodes sampled independently. The optimization problem $\max_{\pi\in\Pi} \zeta_n(\pi)$ in $\mathcal{M}_\mathcal{F}$ cannot be translated to an equivalent single-trial convex RL problem in general.*

Proof We prove the result by providing an instance of convex RL in a handful of trials with parallel non-communicating sampling that can only be optimized within the space of *stochastic* non-Markovian policies. Since we know that a convex MDP $\mathcal{M}_\mathcal{F}$ in a single trial always admits an optimal *deterministic* non-Markovian policy (see Lemma 1), the two problem settings cannot be equivalent.

Let us consider the following instance



with $S = 7$ states, $A \leq 3$ actions (one to go right and one to go left in $0, 2, 3, 5, 6$, one to go right/left, up, down in $1, 4$), a deterministic transition model, horizon $T = 7$, initial state distribution $\mu(0) = 1$, utility function $\mathcal{F}(d) = -d \cdot \log d$ given by the entropy of the empirical state distribution d. It is easy to see that, for every $n > 1$, $\max_{\pi\in\Pi} \zeta_n(\pi) = \mathbb{E}_{d_n\sim p_n^\pi}[-d_n \cdot \log d_n]$ is attained by a policy $\pi \in \Pi_{\text{NM}}$ that randomizes between actions *up* and *down* when reaching states $1, 4$ from 0. ■

The latter result shows that, when the histories $(h_i)_{i\in[n]}$ are sampled independently, then the policy that optimizes the utility in a handful of trials is stochastic in general. This is in stark contrast with the single-trial formulation, as well as the handful of trials with sequential sampling or parallel sampling with perfect communication, which all admit a

deterministic non-Markovian optimal policy. In the following, we provide a better characterization of the *importance of randomization* in convex RL in a handful of (independent) trials, first considering the simpler setting with deterministic transitions and then the more general setting with stochastic transitions.

5.3.1 Deterministic Transitions

Let us consider convex MDPs $\mathcal{M}_\mathcal{F}$ with a deterministic transition model $P : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$. For this class of instances, we can show that the optimal deterministic policy in a handful of (independent) trials is the same deterministic policy that optimizes the single-trial utility.

Lemma 5 *Let $\mathcal{M}_\mathcal{F}$ be a convex MDP with deterministic transitions. Then, the policy $\pi^\ddagger \in \arg\max_{\pi \in \Pi_{\text{NM}}} \zeta_1(\pi)$ in $\mathcal{M}_\mathcal{F}$ is also $\pi^\ddagger \in \arg\max_{\pi \in \Pi^{\text{D}}_{\text{NM}}} \zeta_n(\pi)$ in $\mathcal{M}_\mathcal{F}$.*

Proof To prove the result, we note that deterministic transitions imply that the history h induced by a deterministic policy π is also deterministic, and thus the corresponding utility $\mathcal{F}(d)$ is deterministic as well. Hence, we have

$$\zeta_1(\pi) = \mathop{\mathbb{E}}_{d_1 \sim p_1^\pi}[\mathcal{F}(d_1)] = \mathcal{F}(d_1) = \mathop{\mathbb{E}}_{d_n \sim p_n^\pi}[\mathcal{F}(d_n)] = \zeta_n(\pi), \quad \forall \pi \in \Pi^{\text{D}}_{\text{NM}}.$$

Since equality holds for any deterministic non-Markovian policy, it also holds for the policy $\pi^\ddagger$, which proves the result. ■

Having demonstrated that the optimal single-trial policy $\pi^\ddagger$ is also the optimal deterministic policy in a handful of (independent) trials, we now provide a characterization of the value gap between $\pi^\ddagger$ and the optimal (stochastic) policy in a handful of (independent) trials.

Proposition 4 *Let $\mathcal{M}_\mathcal{F}$ be a convex MDP with deterministic transitions and L-Lipschitz utility $\mathcal{F}$, let n be a number of independent trials, let $\delta \in (0, 1]$ be a confidence level, let $\pi^\dagger \in \arg\max_{\pi \in \Pi} \zeta_n(\pi)$ and $\pi^\ddagger \in \arg\max_{\pi \in \Pi^{\text{D}}_{\text{NM}}} \zeta_n(\pi)$. Then it holds with probability at least $1 - \delta$*

$$|\zeta_n(\pi^\dagger) - \zeta_n(\pi^\ddagger)| \leq O\big(LT\sqrt{S}\big).$$

Proof To prove the result, we write

$$\begin{aligned}
|\zeta_n(\pi^\dagger) - \zeta_n(\pi^\ddagger)| &= |\zeta_n(\pi^\dagger) - \zeta_1(\pi^\ddagger)| && (9)\\
&\leq |\zeta_n(\pi^\dagger) - \zeta_1(\pi^\dagger)| && (10)\\
&\leq |\zeta_n(\pi^\dagger) - \zeta_\infty(\pi^\dagger)| + |\zeta_\infty(\pi^\dagger) - \zeta_1(\pi^\dagger)| && (11)
\end{aligned}$$

where we obtained (9) from Lemma 5, (10) from the definition of $\pi^\ddagger$ that implies $\zeta_1(\pi^\ddagger) \geq \zeta_1(\pi), \forall \pi \in \Pi$, and we got (11) by adding $\pm\zeta_\infty(\pi^\dagger)$ then applying the triangle inequality. Finally, we can bound the two terms on the right-hand side of (11) with high probability as in the proof of Theorem 2, such that it holds

$$|\zeta_n(\pi^\dagger) - \zeta_n(\pi^\ddagger)| \leq 4LT\sqrt{\frac{2S\log(4T/\delta)}{n}} + 4LT\sqrt{2S\log(4T/\delta)}$$

with probability $1 - \delta$. The result follows by noting that the second term is dominating. ■

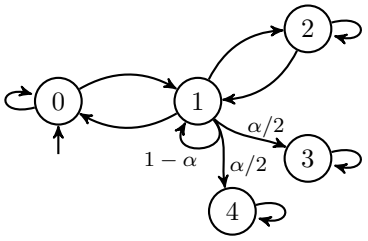
5.3.2 Stochastic Transitions

Let us consider convex MDPs $\mathcal{M}_\mathcal{F}$ with a stochastic transition model $P : \mathcal{S} \times \mathcal{A} \to \Delta_\mathcal{S}$. We can show that for this class of instances, which generalizes the one with deterministic transitions of the previous section, the result in Lemma 5 does not hold anymore. In the following result, we show that the optimal deterministic policy for a handful of trials is not necessarily the optimal single-trial policy.

Proposition 5 *Let $\mathcal{M}_\mathcal{F}$ be a convex MDP with stochastic transitions. Then, the policy $\pi^\ddagger \in \arg\max_{\pi \in \Pi_{\text{NM}}} \zeta_1(\pi)$ in $\mathcal{M}_\mathcal{F}$ does not maximize $\zeta_n(\pi)$ for $\pi \in \Pi^{\text{D}}_{\text{NM}}$ in general.*

Proof We prove the result by providing an instance in which the optimal deterministic policy $\pi^\dagger \in \arg\max_{\pi \in \Pi^{\text{D}}_{\text{NM}}} \zeta_n(\pi)$ for n independent trials is different than the optimal single-trial policy $\pi^\ddagger \in \arg\max_{\pi \in \Pi_{\text{NM}}} \zeta_1(\pi)$ (that is also deterministic, as stated in Lemma 1).

Let us consider the following instance



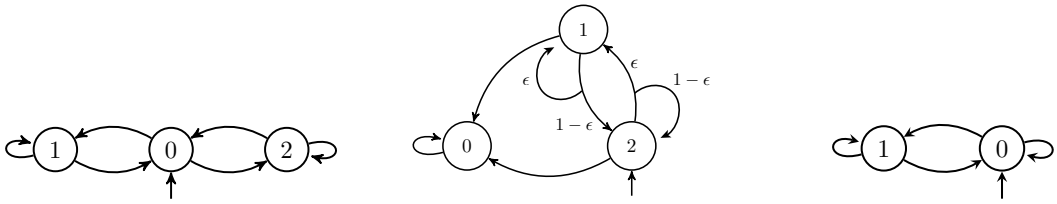
with $S = 5$ states, $A \leq 2$ actions, a stochastic transition model for action *down* in state 1, horizon $T = 2$, initial state distribution $\mu(0) = 1$, utility function $\mathcal{F}(d) = -d \cdot \log d$ given by the entropy of the empirical state distribution d. It is easy to see that the single-trial policy $\pi^\ddagger$ takes action *up* in 1 to generate the history of states $(0, 1, 2)$ with probability one. Instead, for every $n > 1$ and $\alpha \to 1$, the deterministic policy $\pi^\dagger$ takes action *down* to produce even visits at the states $3, 4$. ■

The latter result is a further testament of the essential difference between convex RL in a handful of (independent) trials and the settings that we have analyzed in previous sections, which we can all trace back to a single-trial problem. Future works might focus on a better understanding of this setting, including computational and statistical complexity, as well as extending Proposition 4 to stochastic transitions.

6. Numerical Validation

In this section, we provide a numerical validation on the single-trial convex RL problem.[16] We compare the performance (computed with the single-trial objective $\zeta_1(\pi)$) achieved by a policy $\pi^\dagger \in \arg\max_{\pi \in \Pi} \zeta_1(\pi)$ that maximizes the single-trial utility $\zeta_1(\pi)$ with the

[16] For the sake of clarity, here we restrict our empirical validation to the single-trial setting (i.e., $n = 1$), but similar results can be easily extended to finite-trials settings ($n > 1$) with sequential sampling, as described in Section 5.1, or parallel sampling with perfect communication, as described in Section 5.2. We leave as future work an empirical study of the n-trials setting with parallel sampling without communication (Section 5.3).



(a) Pure exploration (b) Risk-averse RL (c) Imitation learning

Figure 2: Visualization of the convex MDP instances $\mathcal{M}_{\mathcal{F}}$. In **(b)**, state 0 is a low-reward ($r = 1$) low-risk state, state 2 is a high-reward ($R = 10$) high-risk state, and state 1 is a penalty state with zero reward.

performance of a policy $\pi^\star \in \arg\max_{\pi \in \Pi} \zeta_\infty(\pi)$ that maximizes the infinite-trials utility $\zeta_\infty(\pi)$ instead.

The latter infinite-trials $\pi^\star$ is obtained by first solving a dual optimization of the convex MDP $\mathcal{M}_{\mathcal{F}}$ (see Sec. 6.2 in (Mutti and Restelli, 2020)),

$$\max_{\omega \in \Delta_{\mathcal{S}\times\mathcal{A}}} \mathcal{F}(\omega), \qquad \text{subject to } \sum_{a\in\mathcal{A}} \omega(s,a) = \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} P(s|s',a')\omega(s',a'),\ \forall s \in \mathcal{S},$$

and then constructing $\pi^\star$ as $\pi^\star(a|s) = \omega^*(s,a)/\sum_{a\in\mathcal{A}} \omega^*(s,a), \forall (s,a) \in \mathcal{S}\times\mathcal{A}$, where ω^* are the optimal dual variables. To get the finite-trials $\pi^\dagger$, we first recover the extended MDP $\mathcal{M}_\ell$ as explained in the proof of Theorem 1, and then we apply standard dynamic programming (Bellman, 1957) on $\mathcal{M}_\ell$ to get $\pi^\dagger$. Note that $\pi^\dagger$ is a deterministic non-Markovian policy $\pi^\dagger \in \Pi^{\mathrm{D}}_{\mathrm{NM}}$, while $\pi^\star$ is a stochastic Markovian policy $\pi^\star \in \Pi_{\mathrm{M}}$.

In the experiments, we show that optimizing the infinite-trials objective can lead to sub-optimal policies across a wide range of applications. In particular, we cover examples from imitation learning, risk-averse RL, and pure exploration. We carefully selected convex MDPs that are as simple as possible in order to stress the generality of our results (see Figure 2 for the instances).

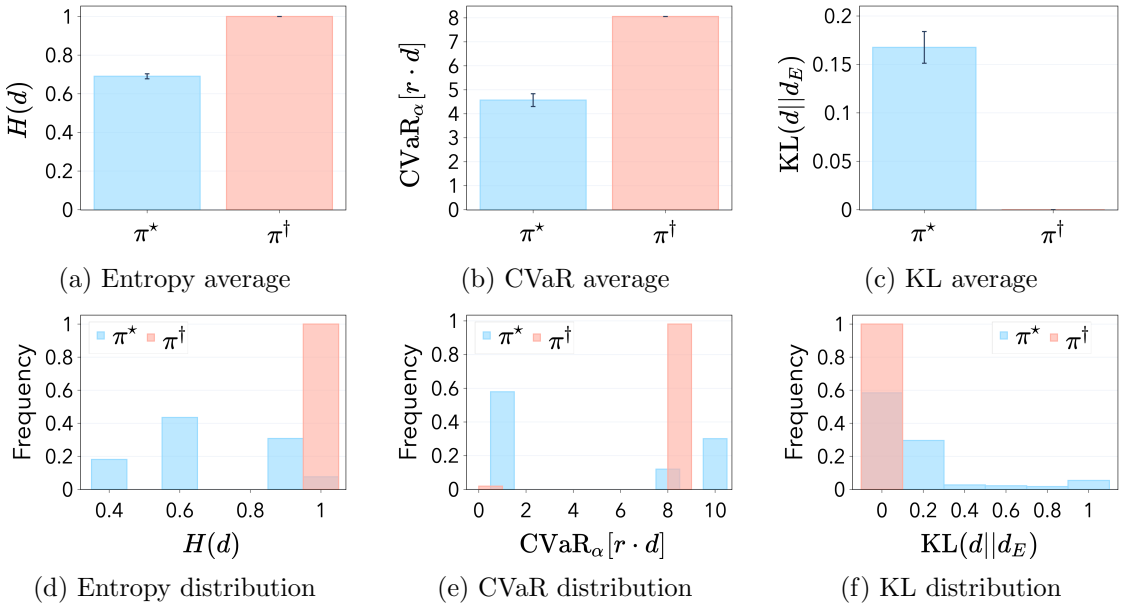
6.1 Pure Exploration

For the pure exploration setting, we consider the state entropy utility (Hazan et al., 2019), i.e.,

$$\mathcal{F}(d) = H(d) = -d \cdot \log d,$$

and the convex MDP in Figure 2a. In this example, the agent aims to maximize the state entropy over a finite-length episode of T steps. Notice that this happens when a policy induces an empirical state distribution that is close to a uniform distribution.

In Figure 3a, we compare the utility $H(d)$ induced by the optimal single-trial policy $\pi^\dagger$ and the optimal infinite-trials policy $\pi^\star$. An agent following the policy $\pi^\dagger$ always achieves a uniform empirical state distribution, which leads to the maximum utility with probability 1, as $\pi^\dagger$ is a deterministic policy. In contrast, the policy π^* is randomized in all three states. As a result, this policy induces sub-optimal empirical state distributions with *strictly positive* probability, as shown in Figure 3d.



(a) Entropy average (b) CVaR average (c) KL average

(d) Entropy distribution (e) CVaR distribution (f) KL distribution

Figure 3: $\pi^\dagger$ denotes an optimal single-trial policy, $\pi^\star$ denotes an optimal infinite-trials policy. In **(a, d)** we report the average and the empirical distribution of the single-trial utility $H(d)$ achieved in the pure exploration convex MDP ($T = 6$) of Figure 2a. In **(b, e)** we report the average and the empirical distribution of the single-trial utility $\mathrm{CVaR}_\alpha[r \cdot d]$ (with $\alpha = 0.4$) achieved in the risk-averse convex MDP ($T = 5$) of Figure 2b. In **(c, f)** we report the average and the empirical distribution of the single-trial utility $\mathrm{KL}(d||d_E)$ (with expert distribution $d_E = (1/3, 2/3)$) achieved in the imitation learning convex MDP ($T = 12$) of Figure 2c. For all the results, we provide 95 % c.i. over 1000 runs.

6.2 Risk-Averse RL

For the risk-averse RL setting, we consider a Conditional Value-at-Risk (CVaR) utility (Rockafellar and Uryasev, 2000) given by

$$\mathcal{F}(d) = \mathrm{CVaR}_\alpha[r \cdot d],$$

where $r \in [0, 1]^S$ is a reward vector, and the convex MDP in Figure 2b, in which the agent aims to maximize the CVaR over a finite-length episode of T steps.

First, notice that financial semantics can be attributed to the given MDP. An agent, starting in state 2, can decide whether to invest in risky assets, e.g., crypto-currencies, or in safe assets, e.g., treasury bills. Because the transitions are stochastic, a policy needs to be reactive to the realization in order to maximize the single-trial utility. This kind of behavior is achieved by an optimal single-trial policy $\pi^\dagger$. Indeed, $\pi^\dagger$ is a non-Markovian deterministic policy, which can take decisions as a function of history, and thus takes into account the current realization. On the other hand, an optimal infinite-trials policy π^* is a Markovian policy, and it cannot take into account the current history. As a result, the policy π^* induces sub-optimal trajectories with *strictly positive* probability (see Figure 3e).

Finally, in Figure 3b we compare the single-trial utility induced by the optimal single-trial policy $\pi^\dagger$ and the optimal infinite-trials policy $\pi^\star$. Overall, $\pi^\dagger$ performs significantly better than $\pi^\star$.

6.3 Imitation Learning

For the imitation learning setting, we consider the distribution matching utility (Kostrikov et al., 2019), i.e.,

$$\mathcal{F}(d) = \mathrm{KL}\left(d||d_E\right),$$

and the convex MDP in Figure 2c. The agent aims to learn a policy π inducing an empirical state distribution d close to the empirical state distribution d_E demonstrated by an expert.

In Figure 3c, we compare the single-trial utility induced by the optimal single-trial policy $\pi^\dagger$ and the optimal infinite-trials policy $\pi^\star$. An agent following $\pi^\dagger$ induces an empirical state distribution that perfectly matches the expert. In contrast, an agent following π^* induces sub-optimal realizations with *strictly positive* probability (see Figure 3f).

7. Related Work

In this section, we revise the relevant literature and how it relates with our findings.

To the best of our knowledge, Hazan et al. (2019) were the first to introduce the convex RL problem, as a generalization of the standard RL formulation to non-linear utilities, especially the entropy of the state distribution. They show that the convex RL objective, while being concave (convex) in the state distribution, can be non-concave (non-convex) in the policy parameters. Anyway, they provide a provably efficient algorithm that overcomes the non-convexity through a Frank-Wolfe approach. Zhang et al. (2020) study the convex RL problem under the name of RL with general utilities. Especially, they investigated a hidden convexity of the convex RL objective that allows for statistically efficient policy optimization in the infinite-trials setting. Recently, the infinite-trials convex RL formulation has been reinterpreted from game-theoretic perspectives (Zahavy et al., 2021; Geist et al., 2022). The former (Zahavy et al., 2021) notes that the convex RL problem can be seen as a min-max game between the policy player and a cost player. The latter (Geist et al., 2022) shows that the convex RL problem is a subclass of mean-field games.

Another relevant branch of literature is the one investigating the expressivity of (Markovian) rewards (Abel et al., 2021; Silver et al., 2021; Abel et al., 2022; Bowling et al., 2022). Especially, Abel et al. (2021) show that not all the notions of tasks, such as inducing a set of admissible policies, a (partial) policy ordering, or a trajectory ordering, can be naturally encoded with a scalar reward function. Whereas the convex RL formulation extends the expressivity of traditional RL w.r.t. all these three notions of tasks, it is still not sufficient to cover every instance. Convex RL is powerful in terms of the policy order it can induce, but it is inherently limited on the trajectory ordering, as it only accounts for the infinite-trials state distribution. Instead, the finite-trials convex RL setting that we presented in this paper is naturally expressive in terms of trajectory orderings, at the expense of a diminished expressivity on the policy orderings w.r.t. infinite-trials convex RL.

Previous works concerning RL in the presence of history feedback are also related to this work. Most of this literature assumes an underlying scalar reward model (e.g., Efroni et al.,

2021) which only delays the feedback at the end of the episode. One notable exception is the once-per-episode formulation in (Chatterji et al., 2021). In their setting, the agent receives binary feedback at the end of an episode, where the feedback is obtained from a logistic model whose input is a function of the history. This problem formulation is close to ours, and we relied on their regret analysis to give our statistical complexity results (Theorem 5). Our paper generalizes the once-per-episode framework beyond the single-trial setting and the binary feedback, and it provides complementing results in terms of optimality and computational complexity. Another interesting form of history feedback is considered in RL with preference feedback, where the agent draws two independent histories and receives a binary preference between them. The work by (Novoseller et al., 2020; Xu et al., 2020; Pacchiano et al., 2021) study the sample complexity of preference-based RL.

Finally, the work in (Cheung, 2019a,b) considers infinite-horizon MDPs with vectorial rewards as a mean to encode convex objectives in RL with a multi-objective flavor. They show that stationary policies are in general sub-optimal for the introduced online learning setting, where non-stationarity becomes essential. In this setting, they provide principled procedures to learn an optimal policy with sub-linear regret. Their work essentially complements our analysis in the infinite-horizon problem formulation, where the difference between finite trials and infinite trials fades away.

8. Conclusion and Future Directions

In this paper, we provided a comprehensive study of convex RL in finite trials.

First, we formally defined the finite-trials convex RL objective. We demonstrated a crucial mismatch between the latter and the infinite-trials formulation that is usually considered in the literature but seldom contemplated in practice. In addition, we characterized the approximation error when optimizing the infinite-trials objective in place of the finite-trials one, showing that the error can be significant when the number of trials is small.

Especially, we reported an in-depth analysis of the extreme single-trial setting, which demonstrates the importance of non-Markovianity when optimizing the single-trial objective, but provides a negative result over the computational tractability of the problem. Nonetheless, we showed that the problem is at least statistically tractable, giving some hope to develop provably efficient algorithms that rely on approximate solvers.

Then, we complemented our analysis with the study of convex RL in a handful of trials, which is the standard in the empirical RL literature. We identified three relevant settings, in which the trials are drawn sequentially or in parallel, with or without communication between the processes in the latter case. We demonstrated that the sequential setting and the parallel setting with communication reduce to the single-trial setting, inheriting analogous computational and statistical properties. We showed that the parallel setting without communication is instead essentially different, as it requires randomized policies to achieve optimal performance.

Improving the analysis Whereas we believe to have answered some of the main questions over convex RL in finite trials, our analysis can be improved in many directions.

On the one hand, our results make little use of the properties of the specific instance $\mathcal{M}_{\mathcal{F}}$ and the utility function $\mathcal{F}$. An instance-dependent analysis could provide additional insights, especially as it is known that some of the utility functions $\mathcal{F}$ allow for efficient

computation even in a single-trial formulation. Characterizing a minimal set of assumptions over $\mathcal{F}$ and/or the transition model P for which finite-trials convex RL is computationally tractable would be extremely valuable.

On the statistical side, our analysis solely guarantees the existence of a provably efficient algorithm for the single-trial setting (and the n-trials setting with sequential sampling or parallel sampling with communication). However, we still do not know whether we can further improve over the provided rate. Proper statistical barriers, such as a minimax lower bound on the regret and a matching algorithm, are yet to be established. An instance-dependent statistical characterization of the problem is also uncharted.

Finally, our understanding of convex RL in a handful of trials with parallel sampling and without communication is still fairly limited. Our results showed that the optimal policy is stochastic in general, which hints that the problem is crucially different from the other finite-trials settings we considered. This warrants further studies to see whether this setting enjoys better computational or statistical properties.

Developing practical methodologies While our analysis provides a generally negative result over the computational tractability of convex RL in finite trials, we believe it is not hopeless to learn near-optimal finite-trials policies in practice.

In the paper, we considered non-Markovian policies that condition their decisions on histories of arbitrary length, which causes an exponential blowup in the number of policy parameters. One can instead condition the decisions on a finite-length history, obtained from a sliding window over past interactions. This restricted policy space can still provide significant benefits over the space of Markovian policies while keeping the computational tractability of the latter. Similarly, one can consider compact representations of the full history, such as implementing the non-Markovian policies through deep recurrent architectures (e.g., Hochreiter and Schmidhuber, 1997) or transformers (Chen et al., 2021).

Another option to sidestep the exponential blowup on the policy parameters is to draw actions from the optimal non-Markovian policy without ever computing it, e.g., by employing a Monte-Carlo Tree Search (MCTS) approach (e.g., Kocsis and Szepesvári, 2006) to select the next action to take. Given the current state as a root, we can build the tree of future histories from the root through repeated simulations of potential action sequences. With a sufficient number of simulations and a sufficiently deep tree, we are guaranteed to select the optimal action at the root. If the episode horizon is too long, we can still cut the tree at any depth and approximately evaluate a leaf node with the utility induced by the partial history, i.e., the path from the root to the leaf. The drawback of this procedure is that we require to access a simulator with reset (or a reliable estimate of the transition model) to actually build the tree.

To conclude, we hope to have shed some light on the convex RL problem in finite trials, which was previously neglected by the literature but is paramount for properly implementing convex RL in both simulated and real-world domains. This work aims to inspire future theoretical and empirical contributions toward fully mastering convex RL.

Appendix A. Missing Proofs

In this section, we report the proofs and derivations that were previously omitted.

A.1 Proofs of Section 3

Theorem 2 (Approximation Error) *Let $\mathcal{M}_{\mathcal{F}}$ be a convex MDP with L-Lipschitz utility function $\mathcal{F}$, let $n \in \mathbb{N}$ be a number of evaluation episodes, let $\delta \in (0, 1]$ be a confidence level, let $\pi^\dagger \in \arg\max_{\pi \in \Pi} \zeta_n(\pi)$ and $\pi^\star \in \arg\max_{\pi \in \Pi} \zeta_\infty(\pi)$. Then, it holds with probability at least $1 - \delta$*

$$err := \left|\zeta_n(\pi^\dagger) - \zeta_n(\pi^\star)\right| \leq 4LT\sqrt{\frac{2S\log(4T/\delta)}{n}}.$$

Proof Let us first upper bound the approximation error as

$$\begin{aligned}
err := \left|\zeta_n(\pi^\dagger) - \zeta_n(\pi^\star)\right| &\leq \left|\zeta_n(\pi^\dagger) - \zeta_\infty(\pi^\dagger)\right| + \left|\zeta_\infty(\pi^\dagger) - \zeta_n(\pi^\star)\right| && (12)\\
&\leq \left|\zeta_n(\pi^\dagger) - \zeta_\infty(\pi^\dagger)\right| + \left|\zeta_\infty(\pi^\star) - \zeta_n(\pi^\star)\right| && (13)\\
&\leq \left| \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\dagger}} [\mathcal{F}(d_n)] - \mathcal{F}(d^{\pi^\dagger})\right| + \left| \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\star}} [\mathcal{F}(d_n)] - \mathcal{F}(d^{\pi^\star})\right| && (14)\\
&\leq \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\dagger}} \left[\left|\mathcal{F}(d_n) - \mathcal{F}(d^{\pi^\dagger})\right|\right] + \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\star}} \left[\left|\mathcal{F}(d_n) - \mathcal{F}(d^{\pi^\star})\right|\right] && (15)\\
&\leq \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\dagger}} \left[L \left\|d_n - d^{\pi^\dagger}\right\|_1\right] + \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi^\star}} \left[L \left\|d_n - d^{\pi^\star}\right\|_1\right] && (16)\\
&\leq 2L \max_{\pi \in \{\pi^\dagger, \pi^\star\}} \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi}} \left[\|d_n - d^{\pi}\|_1\right] && (17)\\
&\leq 2L \max_{\pi \in \{\pi^\dagger, \pi^\star\}} \mathop{\mathbb{E}}_{d_n \sim p_n^{\pi}} \left[\max_{t \in [T]} \|d_{n,t} - d_t^{\pi}\|_1\right], && (18)
\end{aligned}$$

where (12) is obtained by adding $\pm\zeta_\infty(\pi^\dagger)$ and then applying the triangle inequality, (13) follows by noting that $\zeta_\infty(\pi^\star) \geq \zeta_\infty(\pi^\dagger)$, we derive (14) by plugging the definitions of ζ_n, ζ_∞ in (13), then we obtain (15) from $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$, we apply the Lipschitz assumption on $\mathcal{F}$ to write (16) from (15), we maximize over the policies to write (17), and we finally obtain (18) through a maximization over the episode's step by noting that $d_n = \frac{1}{T}\sum_{t \in [T]} d_{n,t}$ and $d^\pi = \frac{1}{T}\sum_{t \in [T]} d_t^\pi$, where $d_{n,t}$ and d_t^π are the empirical distribution and the expected distribution over s_t respectively. Then, we seek to bound with high probability

$$\begin{aligned}
Pr\Big(\max_{\pi \in \{\pi^\dagger, \pi^\star\}} \max_{t \in [T]} \|d_{n,t} - d_t^\pi\|_1 \geq \epsilon\Big) &\leq Pr\Big(\bigcup_{\pi,t} \|d_{n,t} - d_t^\pi\|_1 \geq \epsilon\Big) && (19)\\
&\leq \sum_{\pi,t} Pr\Big(\|d_{n,t} - d_t^\pi\|_1 \geq \epsilon\Big) && (20)\\
&\leq 2T\ Pr\Big(\|d_{n,t} - d_t^\pi\|_1 \geq \epsilon\Big), && (21)
\end{aligned}$$

where $\epsilon > 0$ is a positive constant, and we applied a union bound to get (20) from (19). From concentration inequalities for empirical distributions (see Theorem 2.1 in (Weissman

et al., 2003) and Lemma 16 in (Efroni et al., 2021)) we have

$$Pr\left(\|d_{n,t} - d_t^\pi\|_1 \geq \sqrt{\frac{2S\log(2/\delta')}{n}} \right) \leq \delta'. \tag{22}$$

By setting $\delta' = \delta/2T$ in (22), and then plugging (22) in (21), and again (21) in (18), we have that with probability at least $1-\delta$

$$\left|\zeta_n(\pi^\dagger) - \zeta_n(\pi^\star)\right| \leq 4LT\sqrt{\frac{2S\log(4T/\delta)}{n}},$$

which concludes the proof. ■

Proposition 6 (Finite Trials vs Infinite Trials) *Here we provide equivalence results between the finite-trials and the infinite-trials formulations of the objectives reported in Table 1.*

(i) *Let* $\mathcal{F}(d) = r \cdot d$ *then* $\min_{\pi\in\Pi} \zeta_\infty(\pi) = \min_{\pi\in\Pi} \zeta_n(\pi),\ \forall n \in \mathbb{N}$

(ii) *Let* $\mathcal{F}(d) = r \cdot d$ *s.t.* $\lambda \cdot d \leq c$ *then* $\min_{\pi\in\Pi} \zeta_\infty(\pi) = \min_{\pi\in\Pi} \zeta_n(\pi),\ \forall n \in \mathbb{N}$

(iii) *Let* $\mathcal{F}(d) = \|d - d_E\|_2^2$ *then* $\min_{\pi\in\Pi} \zeta_\infty(\pi) < \min_{\pi\in\Pi} \zeta_n(\pi),\ \forall n \in \mathbb{N}$

(iv) *Let* $\mathcal{F}(d) = -d \cdot \log(d) = H(d)$ *then* $\min_{\pi\in\Pi} \zeta_\infty(\pi) < \min_{\pi\in\Pi} \zeta_n(\pi),\ \forall n \in \mathbb{N}$

(v) *Let* $\mathcal{F}(d) = \mathrm{KL}(d||d_E)$ *then* $\min_{\pi\in\Pi} \zeta_\infty(\pi) < \min_{\pi\in\Pi} \zeta_n(\pi),\ \forall n \in \mathbb{N}$

Proof We report below the corresponding derivations.

(i) $\min_{\pi\in\Pi} \zeta_\infty(\pi) = \min_{\pi\in\Pi} r \cdot d^\pi = \min_{\pi\in\Pi} r \cdot \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[d_n] = \min_{\pi\in\Pi} \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[r \cdot d_n] = \min_{\pi\in\Pi} \zeta_n(\pi)$

(ii) $\min_{\pi\in\Pi} \zeta_\infty(\pi) = \min_{\pi\in\Pi,\lambda\cdot d^\pi\leq c} r \cdot d^\pi = \min_{\pi\in\Pi,\lambda\cdot d^\pi\leq c} r \cdot \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[d_n] = \min_{\pi\in\Pi, r\cdot d^\pi\leq c} \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[r \cdot d_n] = \min_{\pi\in\Pi} \zeta_n(\pi)$

(iii) $\min_{\pi\in\Pi} \zeta_\infty(\pi) = \min_{\pi\in\Pi} \| \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[d_n] - d_E\|_2^2 < \min_{\pi\in\Pi} \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[\|d_n - d_E\|_2^2] = \min_{\pi\in\Pi} \zeta_n(\pi)$

(iv) $\min_{\pi\in\Pi} \zeta_\infty(\pi) = \min_{\pi\in\Pi} \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[d_n] \cdot \log \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[d_n] < \min_{\pi\in\Pi} \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[d_n \cdot \log d_n] = \min_{\pi\in\Pi} \zeta_n(\pi)$

(v) $\min_{\pi\in\Pi} \zeta_\infty(\pi) = \min_{\pi\in\Pi} \mathrm{KL}(\mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[d_n]||d_E) < \min_{\pi\in\Pi} \mathop{\mathbb{E}}_{d_n\sim p_n^\pi}[\mathrm{KL}(d_n \,||\, d_E)] = \min_{\pi\in\Pi} \zeta_n(\pi)$

■

A.2 Proofs of Section 4

Lemma 2 *Let $\pi_{\text{NM}} \in \Pi^{\text{D}}_{\text{NM}}$ be an optimal deterministic non-Markovian policy for $\zeta_1(\pi)$ (8) in the convex MDP $\mathcal{M}_{\mathcal{F}}$. For a fixed history $h_t \in \mathcal{H}_t$ ending in state s, the variance of the event of an optimal Markovian policy $\pi_{\text{M}} \in \arg\max_{\pi \in \Pi_{\text{M}}} \mathbb{E}_{d_1 \sim p_1^{\pi}} \left[\mathcal{F}(d_1)\right]$ taking $a^* = \pi_{\text{NM}}(h_t)$ in s is given by*

$$\mathbb{V}\text{ar}\left[\mathcal{B}(\pi_{\text{M}}(a^*|s_t))\right] = \mathop{\mathbb{V}\text{ar}}_{hs \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right]\right],$$

where $hs \in \mathcal{H}_t$ is any history of length t such that the final state is s, i.e., $hs := (h_{t-1} \in \mathcal{H}_{t-1}) \oplus s$, and $\mathcal{B}(x)$ is a Bernoulli with parameter x.

Proof Let us consider the random variable $A \sim \mathcal{P}$ denoting the event "the agent takes action $a^* \in \mathcal{A}$". Through the law of total variance (Bertsekas and Tsitsiklis, 2002), we can write the variance of A given $s \in \mathcal{S}$ and $t \geq 0$ as

$$\begin{aligned}
\mathbb{V}\text{ar}\left[A|s,t\right] &= \mathbb{E}\left[A^2|s,t\right] - \mathbb{E}\left[A|s,t\right]^2 \\
&= \mathop{\mathbb{E}}_{h}\left[\mathbb{E}\left[A^2|s,t,h\right]\right] - \mathop{\mathbb{E}}_{h}\left[\mathbb{E}\left[A|s,t,h\right]\right]^2 \\
&= \mathop{\mathbb{E}}_{h}\left[\mathbb{V}\text{ar}\left[A|s,t,h\right] + \mathbb{E}\left[A|s,t,h\right]^2\right] - \mathop{\mathbb{E}}_{h}\left[\mathop{\mathbb{E}}_{\pi}\left[A|s,t,h\right]\right]^2 \\
&= \mathop{\mathbb{E}}_{h}\left[\mathbb{V}\text{ar}\left[A|s,t,h\right]\right] + \mathop{\mathbb{E}}_{h}\left[\mathbb{E}\left[A|s,t,h\right]^2\right] - \mathop{\mathbb{E}}_{h}\left[\mathbb{E}\left[A|s,t,h\right]\right]^2 \\
&= \mathop{\mathbb{E}}_{h}\left[\mathbb{V}\text{ar}\left[A|s,t,h\right]\right] + \mathop{\mathbb{V}\text{ar}}_{h}\left[\mathbb{E}\left[A|s,t,h\right]\right]. \qquad (23)
\end{aligned}$$

Now let the conditioning event h be distributed as $h \sim p_{t-1}^{\pi_{\text{NM}}}$, so that the condition s,t,h becomes hs where $hs = (s_0, a_0, s_1, \ldots, s_t = s) \in \mathcal{H}_t$, and let the variable A be distributed according to the distribution $\mathcal{P}$ maximizing the objective $\zeta_1(\pi)$ (8) given the conditioning. Hence, we have that the variable A on the left hand side of (23) is distributed as a Bernoulli $\mathcal{B}(\pi_{\text{M}}(a^*|s,t))$, where $\pi_{\text{M}} \in \arg\max_{\pi \in \Pi_{\text{M}}} \mathbb{E}_{d_1 \sim p_1^{\pi}}[\mathcal{F}(d_1)]$, and the variable A on the right hand side of (24) is distributed as a Bernoulli $\mathcal{B}(\pi_{\text{NM}}(a^*|hs))$, where $\pi_{\text{NM}} \in \arg\max_{\pi \in \Pi_{\text{NM}}} \mathbb{E}_{d_1 \sim p_1^{\pi}}[\mathcal{F}(d_1)]$.[17] Thus, we obtain

$$\mathbb{V}\text{ar}\left[\mathcal{B}(\pi_{\text{M}}(a^*|s,t))\right] = \mathop{\mathbb{E}}_{hs \sim p_t^{\pi_{\text{NM}}}}\left[\mathbb{V}\text{ar}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right]\right] + \mathop{\mathbb{V}\text{ar}}_{hs \sim p_t^{\pi_{\text{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right]\right]. \quad (24)$$

Under Assumption 2, we know from Lemma 1 that the policy π_{NM} is deterministic, i.e., $\pi_{\text{NM}} \in \Pi^{\text{D}}_{\text{NM}}$, so that $\mathbb{V}\text{ar}\left[\mathcal{B}(\pi_{\text{NM}}(a^*|hs))\right] = 0$ for every hs, which concludes the proof. ■

Lemma 3 *Let π_{M} be an optimal Markovian policy for $\zeta_1(\pi)$ (8) in the convex MDP $\mathcal{M}_{\mathcal{F}}$. It holds $\underline{\mathcal{V}}_T(\pi_{\text{M}}) \leq \mathcal{V}_T(\pi_{\text{M}}) \leq \overline{\mathcal{V}}_T(\pi_{\text{M}})$ such that*

$$\underline{\mathcal{V}}_T(\pi_{\text{M}}) = (\mathcal{F}^* - \mathcal{F}_2^*)\sum_{t=0}^{T-1} \mathop{\mathbb{E}}_{h_t \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\frac{\prod_{j=0}^{t-1}\pi_{\text{M}}(a_j^*|s_j)}{\pi_{\text{M}}(a_t^*|s_t)} \mathop{\mathbb{V}\text{ar}}_{hs_t \sim p_{1,t}^{\pi_{\text{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\text{NM}}(a_t^*|hs_t))\right]\right]\right],$$

[17] Note that the random variable A has the same distribution on both sides of (23) but different conditioning, which makes them result in two distinct Bernoulli.

$$\overline{\mathcal{V}}_T(\pi_{\mathrm{M}}) = (\mathcal{F}^* - \mathcal{F}_*) \sum_{t=0}^{T-1} \mathop{\mathbb{E}}_{h_t \sim p_{1,t}^{\pi_{\mathrm{NM}}}} \left[\frac{\prod_{j=0}^{t-1} \pi_{\mathrm{M}}(a_j^*|s_j)}{\pi_{\mathrm{M}}(a_t^*|s_t)} \mathop{\mathbb{V}\mathrm{ar}}_{hs_t \sim p_{1,t}^{\pi_{\mathrm{NM}}}} \left[\mathbb{E}\left[\mathcal{B}(\pi_{\mathrm{NM}}(a_t^*|hs_t))\right] \right] \right],$$

where $\pi_{\mathrm{NM}} \in \arg\max_{\pi \in \Pi_{\mathrm{NM}}^{\mathrm{D}}} \mathbb{E}_{d_1 \sim p_1^{\pi}}\left[\mathcal{F}(d_1)\right]$*, and* $\mathcal{F}_2^*, \mathcal{F}_*$ *are given by*

$$\mathcal{F}_2^* = \max_{\pi \in \{\Pi \setminus \pi_{\mathrm{NM}}\}} \mathop{\mathbb{E}}_{d_1 \sim p_1^{\pi}} \left[\mathcal{F}(d_1)\right], \qquad \mathcal{F}_* = \min_{\pi \in \Pi} \mathop{\mathbb{E}}_{d_1 \sim p_1^{\pi}} \left[\mathcal{F}(d_1)\right].$$

Proof We first derive the upper bound $\overline{\mathcal{V}}_T(\pi_{\mathrm{M}})$. From the definition of the value gap (Definition 2), we can write

$$\begin{aligned}
\mathcal{V}_T(\pi_{\mathrm{M}}) &= \mathcal{F}^* - \mathop{\mathbb{E}}_{h \sim p_{1,T}^{\pi_{\mathrm{M}}}} \left[\mathcal{F}(d_h)\right] && (25)\\
&\leq \mathcal{F}^* - \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[\pi_{\mathrm{M}}(a_0^*|s_0) \mathop{\mathbb{E}}_{s_1 \sim P(\cdot|s_0, a_0^*)} \left[\mathcal{V}_{T-1}(\pi_{\mathrm{M}}, s_1)\right] + \big(1 - \pi_{\mathrm{M}}(a_0^*|s_0)\big)\mathcal{F}_* \right] && (26)\\
&= \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[\mathcal{F}^* - \pi_{\mathrm{M}}(a_0^*|s_0)\mathcal{F}^* - \big(1 - \pi_{\mathrm{M}}(a_0^*|s_0)\big)\mathcal{F}_* \right] \\
&\qquad + \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[\mathop{\mathbb{E}}_{s_1 \sim P(\cdot|s_0, a_0^*)} \left[\pi_{\mathrm{M}}(a_0^*|s_0) \mathop{\mathbb{E}}_{h \sim p_{1,T-2}^{\pi_{\mathrm{M}}}} \left[\mathcal{F}(d_{(s_0, s_1) \oplus h})\right] \right] \right] && (27)\\
&= (\mathcal{F}^* - \mathcal{F}_*) \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[\big(1 - \pi_{\mathrm{M}}(a_0^*|s_0)\big) \right] \\
&\qquad + \mathop{\mathbb{E}}_{h_t \sim p_{1,1}^{\pi_{\mathrm{NM}}}} \left[\pi_{\mathrm{M}}(a_0^*|s_0) \mathop{\mathbb{E}}_{h \sim p_{1,T-2}^{\pi_{\mathrm{M}}}} \left[\mathcal{F}(d_{h_t \oplus h})\right] \right] && (28)\\
&\leq (\mathcal{F}^* - \mathcal{F}_*) \mathop{\mathbb{E}}_{h_t \sim p_{1,1}^{\pi_{\mathrm{NM}}}} \left[\big(1 - \pi_{\mathrm{M}}(a_0^*|s_0)\big) + \pi_{\mathrm{M}}(a_0^*|s_0)\big(1 - \pi_{\mathrm{M}}(a_1^*|s_1)\big) \right] \\
&\qquad + \mathop{\mathbb{E}}_{h_t \sim p_{1,2}^{\pi_{\mathrm{NM}}}} \left[\pi_{\mathrm{M}}(a_0^*|s_0)\pi_{\mathrm{M}}(a_1^*|s_1) \mathop{\mathbb{E}}_{h \sim p_{1,T-3}^{\pi_{\mathrm{M}}}} \left[\mathcal{F}(d_{h_t \oplus h})\right] \right] && (29)\\
&\leq (\mathcal{F}^* - \mathcal{F}_*) \sum_{t=0}^{T-1} \mathop{\mathbb{E}}_{h_t \sim p_{1,t}^{\pi_{\mathrm{NM}}}} \left[\left(\prod_{j=0}^{t-1} \pi_{\mathrm{M}}(a_j^*|s_j) \right) \big(1 - \pi_{\mathrm{M}}(a_t^*|s_t)\big) \right] && (30)
\end{aligned}$$

where we obtain (26) from (25) by separating the events in which the policy π_{M} takes the optimal action a^* or a sub-optimal action, and weighting the probabilities for the value gap at the next step $\mathcal{V}_{T-1}(\pi_{\mathrm{M}}, s_1)$ and the pessimistic value gap $\mathcal{F}_*$ respectively, we apply Definition 2 to write (27), we note that $\mu(s_0)P(s_1|s_0, a_0^*) = \mu(s_0)P(s_1|s_0, a_0^*)\pi_{\mathrm{NM}}(a^*|s_0) = p_{1,1}^{\pi_{\mathrm{NM}}}(s_0, a_0^*, s_1)$ to derive (28), and we repeatedly apply the previous steps to get (29) and then (30). Finally, we note that $\pi_{\mathrm{M}}(a_t^*|s_t)(1 - \pi_{\mathrm{M}}(a_t^*|s_t)) = \mathbb{V}\mathrm{ar}\left[\mathcal{B}(\pi_{\mathrm{M}}(a_t^*|s_t))\right]$ from the definition of the Bernoulli distribution, and we apply Lemma 2 on the right-hand side $\mathbb{V}\mathrm{ar}\left[\mathcal{B}(\pi_{\mathrm{M}}(a_t^*|s_t))\right] = \mathbb{V}\mathrm{ar}_{hs_t \sim p_{1,t}^{\pi_{\mathrm{NM}}}}\left[\mathbb{E}\left[\mathcal{B}(\pi_{\mathrm{NM}}(a_t^*|hs_t))\right]\right]$ to derive the upper bound $\overline{\mathcal{V}}_T(\pi_{\mathrm{M}})$.

Following similar steps, we can derive the lower bound $\underline{\mathcal{V}}_T(\pi_{\mathrm{M}})$. We write

$$\begin{aligned}
\mathcal{V}_T(\pi_{\mathrm{M}}) &= \mathcal{F}^* - \mathop{\mathbb{E}}_{h \sim p_{1,T}^{\pi_{\mathrm{M}}}} \left[\mathcal{F}(d_h)\right] && (31)\\
&\geq \mathcal{F}^* - \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[\pi_{\mathrm{M}}(a_0^*|s_0) \mathop{\mathbb{E}}_{s_1 \sim P(\cdot|s_0, a_0^*)} \left[\mathcal{V}_{T-1}(\pi_{\mathrm{M}}, s_1)\right] + \big(1 - \pi_{\mathrm{M}}(a_0^*|s_0)\big)\mathcal{F}_2^* \right] && (32)\\
&= \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[\mathcal{F}^* - \pi_{\mathrm{M}}(a_0^*|s_0)\mathcal{F}^* - \big(1 - \pi_{\mathrm{M}}(a_0^*|s_0)\big)\mathcal{F}_2^* \right]
\end{aligned}$$

$$+ \mathop{\mathbb{E}}_{s_0\sim\mu}\Big[\mathop{\mathbb{E}}_{s_1\sim P(\cdot|s_0,a_0^*)}\Big[\pi_{\mathrm{M}}(a_0^*|s_0)\mathop{\mathbb{E}}_{h\sim p_{1,T-2}^{\pi_{\mathrm{M}}}}\big[\mathcal{F}(d_{(s_0,s_1)\oplus h})\big]\Big]\Big] \tag{33}$$

$$= (\mathcal{F}^* - \mathcal{F}_2^*)\mathop{\mathbb{E}}_{s_0\sim\mu}\Big[\big(1-\pi_{\mathrm{M}}(a_0^*|s_0)\big)\Big]$$
$$+ \mathop{\mathbb{E}}_{h_t\sim p_{1,1}^{\pi_{\mathrm{NM}}}}\Big[\pi_{\mathrm{M}}(a_0^*|s_0)\mathop{\mathbb{E}}_{h\sim p_{1,T-2}^{\pi_{\mathrm{M}}}}\big[\mathcal{F}(d_{h_t\oplus h})\big]\Big] \tag{34}$$

$$\geq (\mathcal{F}^* - \mathcal{F}_2^*)\mathop{\mathbb{E}}_{h_t\sim p_{1,1}^{\pi_{\mathrm{NM}}}}\Big[\big(1-\pi_{\mathrm{M}}(a_0^*|s_0)\big) + \pi_{\mathrm{M}}(a_0^*|s_0)\big(1-\pi_{\mathrm{M}}(a_1^*|s_1)\big)\Big]$$
$$+ \mathop{\mathbb{E}}_{h_t\sim p_{1,2}^{\pi_{\mathrm{NM}}}}\Big[\pi_{\mathrm{M}}(a_0^*|s_0)\pi_{\mathrm{M}}(a_1^*|s_1)\mathop{\mathbb{E}}_{h\sim p_{1,T-3}^{\pi_{\mathrm{M}}}}\big[\mathcal{F}(d_{h_t\oplus h})\big]\Big] \tag{35}$$

$$\geq (\mathcal{F}^* - \mathcal{F}_2^*)\sum_{t=0}^{T-1}\mathop{\mathbb{E}}_{h_t\sim p_{1,t}^{\pi_{\mathrm{NM}}}}\Big[\Big(\prod_{j=0}^{t-1}\pi_{\mathrm{M}}(a_j^*|s_j)\Big)\big(1-\pi_{\mathrm{M}}(a_t^*|s_t)\big)\Big] \tag{36}$$

and then we apply the definition of the variance of a Bernoulli distribution and the Lemma 2 as before to obtain $\underline{\mathcal{V}}_T(\pi_{\mathrm{M}})$. ■

Theorem 4 (Complexity of Single-Trial Convex MDPs) Ψ_0 *is* NP-hard*.*

Proof To prove the result, it is sufficient to show that there exists a problem $\Psi_c \in$ NP-hard such that $\Psi_c \leq_p \Psi_0$. We show this by reducing 3SAT, a well-known NP-complete problem, to Ψ_0. To derive the reduction, we consider two intermediate problems, namely Ψ_1 and Ψ_2. Especially, we aim to show that the following chain of reductions hold

$$\Psi_0 \geq_m \Psi_1 \geq_p \Psi_2 \geq_p \text{3SAT}.$$

First, we define Ψ_1 as the problem of solving a conveniently constructed POMDP $\mathcal{M}_{\ell,\Omega,O} = (\mathcal{S}_\ell, \mathcal{A}_\ell, P_\ell, T_\ell, \mu_\ell, r_\ell, \Omega, O)$ within the space of Markovian policies Π_{M}. The latter is obtained as follows:

- We construct $\mathcal{S}_\ell, \mathcal{A}_\ell, P_\ell, T_\ell, \mu_\ell, r_\ell$ in the same way as in the extended MDP $\mathcal{M}_\ell$ construction described in the proof of Lemma 1;
- We define the observation space $\Omega = \mathcal{S}$, which means that each observation $o \in \Omega$ corresponds to a state $s \in \mathcal{S}$ of the original convex MDP $\mathcal{M}_{\mathcal{F}}$;
- We define a *deterministic* observation function $O : \mathcal{S}_\ell \to \Omega$, such that the observation $o = O(s_\ell)$ corresponds to the last state of the history $s_\ell \in \mathcal{S}_\ell$.

Then, the reduction $\Psi_0 \geq_m \Psi_1$ works as follows. We denote as $\mathcal{I}_{\Psi_i}$ the set of possible instances of problem Ψ_i. We show that Ψ_0 is harder than Ψ_1 by defining the polynomial-time functions ψ and ϕ such that any instance of Ψ_1 can be converted through ψ as an instance of Ψ_0, and a solution $\pi_0^* \in \Pi_{\mathrm{NM}}$ for Ψ_0 can be converted through ϕ into a solution $\pi_1^* \in \Pi_{\mathrm{M}}$ for Ψ_1. The chain of conversions can be visualized as

$$\begin{array}{ccc} \mathcal{I}_{\Psi_1} & \xrightarrow{\ \psi\ } & \mathcal{I}_{\Psi_0} \\ & & \downarrow \\ \pi_{\mathrm{M}}^* & \xleftarrow[\ \phi\ ]{} & \pi_{\mathrm{NM}}^* \end{array}$$

The function ψ constructs $\mathcal{M}_{\mathcal{F}}$ from $\mathcal{M}_{\ell,\Omega,O}$ by setting $\mathcal{S} = \Omega, \mathcal{A} = \mathcal{A}_\ell, T = T_\ell, \mu = \mu_\ell$ and recovering $\mathcal{F}, P$ from r_ℓ, P_ℓ. The function ϕ converts a solution π_0^* of Ψ_0 by computing

$$\pi_1^*(a|o) = \sum_{ho \in \mathcal{H}_o} p_1^{\pi_0^*}(ho)\pi_0^*(a|ho)$$

where $\mathcal{H}_o$ stands for the set of histories $h \in \mathcal{H}$ ending in the observation $o \in \Omega$. Since π_1^* is a solution for Ψ_1, we have that $\Psi_0 \geq_m \Psi_1$.

We now define Ψ_2 as the policy existence problem (see Lusena et al., 2001) in the same class of POMDPs of Ψ_1. The policy existence is the problem of determining whether there exists a policy $\pi \in \Pi_{\text{M}}$ having a value greater than 0 in $\mathcal{M}_{\ell,\Omega,O}$. Since computing an optimal policy in POMDPs is in general harder than the relative policy existence problem (Lusena et al., 2001, Section 3), we have that $\Psi_1 \geq_p \Psi_2$.[18]

For the last reduction, i.e., $\Psi_2 \geq_p$ 3SAT, we extend the proof of Theorem 4.13 in (Mundhenk et al., 2000), which states that the policy existence problem for POMDPs is NP-complete. In particular, we show that this holds for the restricted class of POMDPs that we defined earlier. The restrictions on the POMDPs class are the following:

1. The reward function can be different than zero only in the subset of states $\mathcal{C} \subset \mathcal{S}_\ell$ that correspond to histories of T steps;

2. It holds the relation $|\mathcal{S}_\ell| = |\Omega|^T$ between the cardinality of state and observation spaces.

The latter restrictions can be overcome as follows:

1. It suffices to add states with deterministic transitions so that $T = m \cdot n$ can be defined a priori, where T is the number of steps needed to reach a state with positive reward through every possible path. Here m is the number of clauses, and n is the number of variables in the 3SAT instance, as defined in (Mundhenk et al., 2000);

2. Noticing that the set of observations corresponds with the set of variables and that $T = m \cdot n$ from the previous point, we have that $|\Omega|^T = n^{m \cdot n}$, while the class of POMDPs defined earlier has $|\mathcal{S}_\ell| = m \cdot n^2$. Notice that $n \geq 2$ and $m \geq 1$ implies that $n^{m \cdot n} \geq m \cdot n^2$. Moreover, notice that every instance of 3SAT has $m \geq 1$ and $n \geq 3$. Hence, to extend the proof to the class of POMDPs of interest, it is sufficient to add a set of states $\mathcal{D}$ such that $r_\ell(s_\ell) = 0, \ \forall s_\ell \in \mathcal{D}$.

Since the chain $\Psi_0 \geq_m \Psi_1 \geq_p \Psi_2 \geq_p$ 3SAT holds, we have that $\Psi_0 \geq_p$ 3SAT. Moreover, since 3SAT $\in$ NP-complete, we can conclude that Ψ_0 is NP-hard. ■

[18] The latter statement can be trivially verified as follows: If we solve Ψ_1 to obtain the policy π_1^*, then we can easily solve the policy existence problem by testing whether the value of π_1^* is greater than zero. The latter is a necessary and sufficient condition for the policy existence since π_1^* is the policy attaining the maximum value in the corresponding POMDP.

References

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

David Abel, Will Dabney, Anna Harutyunyan, Mark K Ho, Michael Littman, Doina Precup, and Satinder Singh. On the expressivity of Markov reward. In *Advances in Neural Information Processing Systems*, 2021.

David Abel, André Barreto, Michael Bowling, Will Dabney, Steven Hansen, Anna Harutyunyan, Mark K Ho, Ramana Kumar, Michael L Littman, Doina Precup, et al. Expressing non-Markov reward to a Markov agent. In *Multidisciplinary Conference on Reinforcement Learning and Decision Making*, 2022.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, 2017.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Sanjeev Arora and Boaz Barak. *Computational complexity: A modern approach*. Cambridge University Press, 2009.

Karl J Astrom. Optimal control of Markov decision processes with incomplete state estimation. *Journal Mathematical Analysis and Applications*, 1965.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *AAAI Conference on Artificial Intelligence*, 2022.

Richard Bellman. Dynamic programming. *Princeton University Press*, 1957.

Dimitri P Bertsekas and John N Tsitsiklis. *Introduction to probability*. Athena Scientific Belmont, MA, 2002.

L Bisi, L Sabbioni, E Vittori, M Papini, and M Restelli. Risk-averse trust region optimization for reward-volatility reduction. In *International Joint Conference on Artificial Intelligence*, 2020.

Massimiliano Bonetti, Lorenzo Bisi, and Marcello Restelli. Risk-averse optimization of reward-based coherent risk measures. *Artificial Intelligence*, 316:103845, 2023.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. *arXiv preprint arXiv:2212.10420*, 2022.

Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances in Neural Information Processing Systems*, 2020.

Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giró-i Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. In *International Conference on Machine Learning*, 2020.

Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. In *Advances in Neural Information Processing Systems*, 2021.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, 2021.

Wang Chi Cheung. Exploration-exploitation trade-off in reinforcement learning on online Markov decision processes with global concave rewards. *arXiv preprint arXiv:1905.06466*, 2019a.

Wang Chi Cheung. Regret minimization for reinforcement learning with vectorial feedback and complex objectives. In *Advances in Neural Information Processing Systems*, 2019b.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: A cvar optimization approach. In *Advances in Neural Information Processing Systems*, 2015.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Robert Dadashi, Leonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal Wasserstein imitation learning. In *International Conference on Learning Representations*, 2020.

Daniela Pucci De Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *AAAI Conference on Artificial Intelligence*, 2021.

Khaled Eldowa, Lorenzo Bisi, and Marcello Restelli. Finite sample analysis of mean-volatility actor-critic for risk-averse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2018.

Gideon Freund, Elad Sarafian, and Sarit Kraus. A coupled flow approach to imitation learning. In *International Conference on Machine Learning*, 2023.

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

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Oliver Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. In *International Conference on Autonomous Agents and Multiagent Systems*, 2022.

Jacopo Germano, Francesco Emanuele Stradi, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. A best-of-both-worlds algorithm for constrained mdps with long-term constraints. *arXiv preprint arXiv:2304.14326*, 2023.

Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, 2020.

Ido Greenberg, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Efficient risk-averse reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Karol Gregor, Danilo Rezende, and Daan Wierstra. Variational intrinsic control. *International Conference on Learning Representations, Workshop Track*, 2017.

Zhaohan Daniel Guo, Mohammad Gheshlagi Azar, Alaa Saade, Shantanu Thakoor, Bilal Piot, Bernardo Avila Pires, Michal Valko, Thomas Mesnard, Tor Lattimore, and Rémi Munos. Geometric entropic exploration. *arXiv preprint arXiv:2101.02055*, 2021.

Steven Hansen, Will Dabney, Andre Barreto, David Warde-Farley, Tom Van de Wiele, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. In *International Conference on Learning Representations*, 2019.

Jia Lin Hau, Marek Petrik, and Mohammad Ghavamzadeh. Entropic risk optimization in discounted mdps. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019.

Shuncheng He, Yuhang Jiang, Hongchang Zhang, Jianzhun Shao, and Xiangyang Ji. Wasserstein unsupervised reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2022.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys*, 50(2):1–35, 2017.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning.* PhD Thesis, University College London, 2003.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49:209–232, 2002.

Kuno Kim, Akshat Jindal, Yang Song, Jiaming Song, Yanan Sui, and Stefano Ermon. Imitation with neural density models. In *Advances in Neural Information Processing Systems*, 2021.

Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In *European Conference on Machine Learning*, 2006.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations*, 2019.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Hao Liu and Pieter Abbeel. APS: Active pretraining with successor features. In *International Conference on Machine Learning*, 2021a.

Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. In *Advances in Neural Information Processing Systems*, 2021b.

Christopher Lusena, Judy Goldsmith, and Martin Mundhenk. Nonapproximability results for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 14(1):83–103, 2001.

Sobhan Miryoosefi, Kianté Brantley, Hal Daume III, Miro Dudik, and Robert E Schapire. Reinforcement learning with convex constraints. In *Advances in Neural Information Processing Systems*, 2019.

Martin Mundhenk, Judy Goldsmith, Christopher Lusena, and Eric Allender. Complexity of finite-horizon Markov decision process problems. *Journal of the ACM*, 47(4):681–720, 2000.

Mojmir Mutny, Tadeusz Janik, and Andreas Krause. Active exploration via experiment design in markov chains. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Mirco Mutti. *Unsupervised reinforcement learning via state entropy maximization.* PhD Thesis, Università di Bologna, 2023.

Mirco Mutti and Marcello Restelli. An intrinsically-motivated approach for learning highly exploring and fast mixing policies. In *AAAI Conference on Artificial Intelligence*, 2020.

Mirco Mutti, Lorenzo Pratissoli, and Marcello Restelli. Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate. In *AAAI Conference on Artificial Intelligence*, 2021.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022a.

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

Mirco Mutti, Stefano Del Col, and Marcello Restelli. Reward-free policy space compression for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2022c.

Mirco Mutti, Mattia Mancassola, and Marcello Restelli. Unsupervised reinforcement learning in multiple environments. In *AAAI Conference on Artificial Intelligence*, 2022d.

Alexander Nedergaard and Matthew Cook. k-means maximum entropy exploration. *arXiv preprint arXiv:2205.15623*, 2022.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2):1–179, 2018.

Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

LA Prashanth and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive mdps. In *Advances in Neural Information Processing Systems*, 2013.

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

Zengyi Qin, Yuxiao Chen, and Chuchu Fan. Density constrained reinforcement learning. In *International Conference on Machine Learning*, 2021.

R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Paul J Schweitzer and Abraham Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110 (2):568–582, 1985.

Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*, 2021.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Aviv Tamar and Shie Mannor. Variance adjusted actor critic algorithms. *arXiv preprint arXiv:1310.3697*, 2013.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, 2015.

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

Jean Tarbouriech, Shubhanshu Shekhar, Matteo Pirotta, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active model estimation in Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Pierre Perrault, Yunhao Tang, Michal Valko, and Pierre Menard. Fast rates for maximum entropy exploration. *arXiv preprint arXiv:2303.08059*, 2023.

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

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Technical Report*, 2003.

Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees. In *Advances in Neural Information Processing Systems*, 2020.

Qisong Yang and Matthijs TJ Spaan. CEM: Constrained entropy maximization for task-agnostic safe exploration. In *AAAI Conference on Artificial Intelligence*, 2023.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, 2021.

Tiancheng Yu, Yi Tian, Jingzhao Zhang, and Suvrit Sra. Provably efficient algorithms for multi-objective competitive rl. In *International Conference on Machine Learning*, 2021.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. In *Advances in Neural Information Processing Systems*, 2021.

Tom Zahavy, Yannick Schroecker, Feryal Behbahani, Kate Baumli, Sebastian Flennerhag, Shaobo Hou, and Satinder Singh. Discovering policies with domino: Diversity optimization maintaining near optimality. In *International Conference on Learning Representations*, 2023.

Chuheng Zhang, Yuanying Cai, Longbo Huang, and Jian Li. Exploration by maximizing Rényi entropy for reward-free RL framework. In *AAAI Conference on Artificial Intelligence*, 2021a.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, 2020.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Mean-variance policy iteration for risk-averse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2021b.