Efficient Model-Based Concave Utility Reinforcement Learning through Greedy Mirror Descent

Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard, Nadia Oudjane

Efficient Model-Based Concave Utility Reinforcement Learning through Greedy Mirror Descent

Bianca Marin Moreno
Inria THOTH
EDF R&D

Margaux Brégère
Sorbonne Université
EDF R&D

Pierre Gaillard
Inria THOTH

Nadia Oudjane
EDF R&D

Abstract

Many machine learning tasks can be solved by minimizing a convex function of an occupancy measure over the policies that generate them. These include reinforcement learning, imitation learning, among others. This more general paradigm is called the Concave Utility Reinforcement Learning problem (CURL). Since CURL invalidates classical Bellman equations, it requires new algorithms. We introduce MD-CURL, a new algorithm for CURL in a finite horizon Markov decision process. MD-CURL is inspired by mirror descent and uses a non-standard regularization to achieve convergence guarantees and a simple closed-form solution, eliminating the need for computationally expensive projection steps typically found in mirror descent approaches. We then extend CURL to an online learning scenario and present Greedy MD-CURL, a new method adapting MD-CURL to an online, episode-based setting with partially unknown dynamics. Like MD-CURL, the online version Greedy MD-CURL benefits from low computational complexity, while guaranteeing sub-linear or even logarithmic regret, depending on the level of information available on the underlying dynamics.

1 Introduction

We consider the concave utility reinforcement learning (CURL) problem, which consists on minimizing a convex function (or maximising a concave one) over state-action distributions induced by an agent's policy:

$$\min_{\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} \left\{ F(\mu^{\pi,p}) := \sum_{n=1}^{N} f_n(\mu_n^{\pi,p}) \right\}. \qquad (1)$$

Here, we consider an episodic Markov decision process (MDP) with finite state space $\mathcal{X}$, finite action space $\mathcal{A}$, episodes of length N, and probability transition kernel $p := (p_n)_{n\in[N]}$ such that $p_n : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0,1]$. For all $s \in \mathbb{N}$ we denote $[s] := \{1, \dots, s\}$. Letting $\Delta_{\mathcal{S}}$ be the simplex over a finite set $\mathcal{S}$, we denote $\mu^{\pi,p} := (\mu_n)_{0\le n\le N} \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$ the state-action distributions over an episode induced by the policy π in the MDP with dynamics p.

Many machine learning tasks are special cases of Problem (1). For instance, for the **reinforcement learning** (RL) task (Sutton and Barto, 2018), $F(\mu^{\pi,p}) := -\langle \mu^{\pi,p}, r\rangle$, i.e. the inner product between the state-action distribution induced by π and a reward r. For the **imitation learning** problem (Ghasemipour et al., 2020), $F(\mu^{\pi,p}) := D_f(\mu^{\pi,p}, \mu^*)$, where D_f represents a Bregman divergence induced by a function f. For some instances of the **mean field control** (MFC) problem (Bensoussan et al., 2013), $F(\mu^{\pi,p}) := -\langle \mu^{\pi,p}, r(\mu^{\pi,p})\rangle$, where the reward function also depends on the agents' state-action distribution. For **mean field game** (MFG) problems having the gradient of F as reward, finding a Nash Equilibrium amounts to solving Problem (1) (Geist et al., 2022; Lavigne and Pfeiffer, 2023).

Contribution 1**:** We present a new iterative algorithm focusing on solving Problem (1) called MD-CURL. It is inspired by the mirror descent algorithm (Beck and Teboulle, 2003), and we prove a convergence rate of order $1/\sqrt{K}$ where K is the number of iterations. Our main new ingredient is the use of a non-standard regularization, which enables us to find both a simple closed-form solution - meaning we avoid the generally costly projection step that mirror descent algorithms undergo - and a convergence proof.

Under review.

Until now, there have been few algorithms for solving the general framework of Problem (1). The first two approaches were proposed, by Hazan et al. (2019) on the basis of the Frank-Wolfe algorithm (Frank and Wolfe, 1956), and Zhang et al. (2020) on the basis of policy gradient methods, both of which have theoretical guarantees. In the mean field community, Geist et al. (2022) prove that all algorithms for solving MFGs in discrete-time RL can be applied for solving CURL. Laurière et al. (2022) survey existing algorithms and perform numerical experiments, showing that the adaptation of online mirror descent (OMD) for MFGs presented by Pérolat et al. (2022) has the best performance, outperforming the previously mentioned approaches. However, this method has no proof of convergence for discrete iterations. Our proposed algorithm, as we demonstrate with showcase experiments, has the same performance as OMD for MFGs while having theoretical guarantees of convergence.

Online extension of CURL: An interesting extension of Problem (1) is the online learning scenario, in which we consider computing a sequence of policies $(\pi^t)_{t\in[T]}$ for T episodes with the objective of minimizing a total loss

$$L_T := \sum_{t=1}^{T} F^t(\mu^{\pi^t,p}), \tag{2}$$

where we allow the objective function F^t to change arbitrarily over time (and only be revealed at the end of each episode t).

Here, we consider dynamics such that $(x_0, a_0) \sim \mu_0(\cdot)$, and for all steps $n \in [N]$,

$$x_{n+1} := g_n(x_n, a_n, \varepsilon_n), \tag{3}$$

where $(\varepsilon_n)_{n\in[N]}$ is an independent sequence of external noises with $\varepsilon_n \sim h_n(\cdot)$ for h_n a distribution.

Different variants of this problem can be considered, depending on the prior information available on the dynamics. Here, we consider the case where the agent has prior knowledge of the dynamics (g_n is known), but may be subject to unknown external interference (h_n is unknown). This includes scenarios such as: An energy central controlling the average consumption of electrical appliances. The temperature evolution equation is known, but consumer behavior is unknown and can interfere with the dynamics (Coffman et al., 2023). Controlling a fleet of drones in a known environment, subject to external influences due to weather conditions or human intervention. Controlling the state of charge of electric vehicles so that their average consumption follows an energy production target that changes every day and is not known in advance. The dynamics of loading are known, but the arrival and departure of users are not (Séguret et al., 2021).

Contribution 2: We propose Greedy MD-CURL, an online learning algorithm for CURL with dynamics as in Equation (3) when g_n is known but the noise distribution h_n is unknown. At each episode, we play a policy π^t, observe the agent's behavior, update an estimate of the external noise, and use the estimated dynamics to compute the next policy using MD-CURL. Greedy MD-CURL achieves state-of-the-art sub-linear regrets with low complexity and simple closed-form solutions. We further avoid the $\sqrt{|\mathcal{X}|}$ term paid in UCRL approaches (see Section 2) by showing a weaker control on the difference between the true and estimated probability kernels, being an advantage for models with large state spaces.

Balancing exploration and exploitation is challenging when both g_n and h_n are unknown, which can be computationally expensive. Greedy MD-CURL offers a low-complexity algorithm that achieves sub-linear regret, even without explicit exploration (see Remark 5.3). Although its regret bound is not the state of the art, Greedy MD-CURL is a good option for scenarios where exploration is already induced by the objective function or by noisy models.

2 Related Work

Online MDPs have mostly been studied in specific cases of CURL rather than in its general form, and draw inspiration from online learning problems (Cesa-Bianchi and Lugosi, 2006). In model-based RL Even-Dar et al. (2009) were the first to propose a method dealing with adversarial functions, supposing the transition kernel is fully known in advance. Neu et al. (2012) extended this work to the case of unknown transition kernels with adversarial rewards, using techniques inspired by UCRL-2 (Upper Confidence Reinforcement Learning) (Jaksch et al., 2008). Recently, UC-O-REPS was proposed by Rosenberg and Mansour (2019), which extends Zimin and Neu (2013) O-REPS algorithm to the case of unknown dynamics and improves upon the regret bound of Neu et al. (2012).

In the mean-field community, most approaches for unknown dynamics consider model-free scenarios, such as Angiuli et al. (2020, 2023); Carmona and Laurière (2022). M3-UCRL, proposed by Pasztor et al. (2021), is the only model-based algorithm for mean-field control problems with unknown dynamics. It uses the principle of optimism under uncertainty with UCRL-2 techniques, but only provides regret bounds for Gaussian process dynamics and does not consider online adversarial objective functions.

We introduce the first algorithm for the online CURL problem with theoretical guarantees. Unlike UCRL and PSRL approaches (Osband et al., 2013), which are

generally computationally expensive, our algorithm is nearly greedy and still achieves the same regret bounds with lower computational complexity, depending on the dynamics information available.

3 General Problem Formulation

Consider an episodic Markov decision process (MDP) with finite state space $\mathcal{X}$, finite action space $\mathcal{A}$, episodes of length N, and a sequence of transition probabilities $p := (p_n)_{n\in[N]}$ where $p_n : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0,1]$. At time step n, an agent in state x_n choosing action a_n transitions to state x_{n+1} with probability $p_{n+1}(x_{n+1}|x_n, a_n)$. At the start of an episode, the agent's first state-action couple follows a fixed distribution $\mu_0 \in \Delta_{\mathcal{X}\times\mathcal{A}}$. Actions are chosen by means of a policy $\pi_n : \mathcal{X} \to \Delta_{\mathcal{A}}$ at each time step. In an episode, when an agent follows a sequence of strategies $\pi := (\pi_n)_{n\in[N]}$, we define $\mu^{\pi,p} := (\mu_n^{\pi,p})_{0\le n\le N}$ the state-action distribution sequence induced by the policy π in the MDP with probability kernel p recursively for all $(x', a') \in \mathcal{X} \times \mathcal{A}$ and all $n \in [N]$:

$$\mu_0^{\pi,p}(x', a') := \mu_0(x', a') \tag{4}$$
$$\mu_n^{\pi,p}(x', a') := \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mu_{n-1}^{\pi,p}(x, a) p_n(x'|x, a)\pi_n(a'|x').$$

We let $\|\cdot\|_1$ be the L_1 norm, and for all $v := (v_n)_{n\in[N]}$, such that $v_n \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ we define $\|v\|_{\infty,1} := \sup_{0\le n\le N} \|v_n\|_1$. We define the objective function $F(\mu) := \sum_{n=1}^N f_n(\mu_n)$ where $f_n : \Delta_{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}$ are convex and ℓ-Lipschitz functions with respect to the norm $\|\cdot\|_1$.

Offline optimization setting (Section 4) To solve the CURL problem, we propose a learning protocol that consists in following an iterative method. At each iteration $k \in [K]$, the learner computes a new policy by solving an auxiliary optimization problem. This auxiliary optimization problem, that we denote by $\mathfrak{F}$, depends on the previous policy π^{k-1}, the model dynamics p, and the objective function F, i.e. $\pi^k := \mathfrak{F}(\pi^{k-1}, p, F)$. In Section 4, we show how to construct $\mathfrak{F}$ such that $\min_{k\in[K]} F(\mu^{\pi^k,p}) - F(\mu^{\pi^*,p})$ is bounded by a term of order $1/\sqrt{K}$, with K the number of iterations, where π^* is an optimal policy.

Online learning setting (Section 5) In the online extension of CURL, the objective function at episode $t \in [T]$ is denoted as $F^t := \sum_{n=1}^N f_n^t$, where T is the total number of episodes, and $f_n^t : \Delta_{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}$. We assume that f_n^t is convex and ℓ-Lipschitz with respect to the $\|\cdot\|_1$ norm. The functions F^t are only revealed to the learner at the end of episode t. The learner's objective is to compute a sequence of strategies $(\pi^t)_{t\in[T]}$ minimizing their total loss defined in Equation (2), and the learner's performance is measured by comparison to the best stationary policy, using the following regret:

$$R_T := \sum_{t=1}^T F^t(\mu^{\pi^t,p}) - \min_{\pi\in(\Delta_{\mathcal{A}})^{\mathcal{X}\times N}} \sum_{t=1}^T F^t(\mu^{\pi,p}). \tag{5}$$

We consider the dynamics of Equation (3) when g_n is known but h_n is unknown. In order to choose a sequence of policies that minimize their total loss, the learner must then both optimize the objective function and learn the noise distribution through observations. The learner's online protocol is in Algorithm 1.

At each episode t, the learner chooses a policy π^t, send it to M independent agents, observes the external noise $(\varepsilon_1^{j,t}, \dots, \varepsilon_N^{j,t})$ for each agent $j \in [M]$ over all N steps (to retrieve the noises, it is enough to observe the agent's trajectory and for g_n to be invertible), computes an estimate $\hat{p}^{t+1}$ of the probability kernel using the observations, observes the objective function F^t, and calculates the policy for the next episode by applying the auxiliary problem $\mathfrak{F}$ on π^t, F^t, and $\hat{p}^{t+1}$.

To compute a strategy sequence with sub-linear regret the learner faces two challenges: how to estimate $\hat{p}^t$ from the data and how to define the auxiliary optimization problem $\mathfrak{F}$. In Section 5, we show that by considering the same auxiliary optimization problem $\mathfrak{F}$ as in the optimization of the offline CURL problem, and by taking $\hat{p}^{t+1}$ as the empirical mean estimator, we can build an algorithm that achieves sub-linear regret. This result is studied in detail in Section 5. Observing M independent agents following the same policy is relevant to many applications, such as controlling the charging state of a set of electric vehicles. This allows us to explicitly express the dependence on M. Note that if we set $M = 1$, we recover the standard case.

4 CURL as an optimisation problem

4.1 Reformulation of learner's objective

The CURL problem as presented in Equation (1) is problematic in that it is not convex in π, and calculating the gradient of F with respect to the strategy π can be intractable. Therefore, we reformulate the learner's objective to obtain a convex problem. We define

$$\mathcal{M}_{\mu_0}^p := \Big\{ \mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N \Big| \sum_{a'\in\mathcal{A}} \mu_n(x', a') = \tag{6}$$
$$\sum_{x\in\mathcal{X}, a\in\mathcal{A}} p_n(x'|x, a)\mu_{n-1}(x, a)\ , \forall x' \in \mathcal{X}, \forall n \in [N] \Big\},$$

Algorithm 1 Learner's Online Protocol

Input: initial state-action distribution μ_0, initial strategy sequence π^1.
for $t = 1, \dots, T$ **do**
  for $j = 1, \dots, M$ **do**
    the j-th agent playing episode t starts at $(x_0^{j,t}, a_0^{j,t}) \sim \mu_0(\cdot)$
    for $n = 1, \dots, N$ **do**
      environment draws new state $x_n^{j,t} \sim p_n(\cdot|x_{n-1}^{j,t}, a_{n-1}^{j,t})$
      learner observes agent's j external noise $\varepsilon_n^{j,t}$
      agent j chooses an action $a_n^{j,t} \sim \pi_n^t(\cdot|x_n^{j,t})$
    end for
  end for
  learner computes, for all $n \in [N]$, new estimate $\hat{p}_n^{t+1}$ from data $(\varepsilon_n^{j,s})_{s\in[t],j\in[M]}$
  objective function F^t is exposed
  learner computes $\pi^{t+1} = \mathfrak{F}(\pi^t, \hat{p}^{t+1}, F^t)$ and send to all agents
end for
return π^T

as the set of state-action distribution sequences satisfying the Bellman-flow in the MDP with transition kernel p and initial state-action distribution μ_0. For now, we assume that the probability kernel p is known and, to minimize notations, we let $\mu^\pi := \mu^{\pi,p}$ and $\mathcal{M}_{\mu_0} := \mathcal{M}_{\mu_0}^p$. We also assume μ_0 is always known.

For any $\mu \in \mathcal{M}_{\mu_0}^p$, there exists a strategy π such that $\mu^\pi = \mu$. It suffices to take $\pi_n(a|x) \propto \mu_n(x,a)$ when the normalization factor is non-zero, and arbitrarily defined otherwise. This result is formally enunciated and proved in Proposition A.1 in Appendix A (see also Puterman (1994)). We therefore have the equivalence

$$\min_{\pi\in(\Delta_\mathcal{A})^{\mathcal{X}\times N}} F(\mu^\pi) \equiv \min_{\mu\in\mathcal{M}_{\mu_0}} F(\mu).$$

Note that the optimization problem over μ is convex.

4.2 The Algorithm

To build an algorithm solving the CURL problem, we need to build the auxiliary optimization problem $\mathfrak{F}$ discussed in Section 3. Let $\mathcal{M}_{\mu_0}^*$ denote the subset of $\mathcal{M}_{\mu_0}$ where the corresponding policies π are such that $\pi_n(a|x) \neq 0$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, $n \in [N]$. We define a regularization function $\Gamma : \mathcal{M}_{\mu_0} \times \mathcal{M}_{\mu_0}^* \to \mathbb{R}$ as

$$\Gamma(\mu^\pi, \mu^{\pi'}) := \sum_{n=1}^N \mathbb{E}_{(x,a)\sim\mu_n^\pi(\cdot)}\left[\log\left(\frac{\pi_n(a|x)}{\pi'_n(a|x)}\right)\right], \quad (7)$$

that is well defined thanks to the bijection between strategies and state-action distributions satisfying the Bellman flow (see Proposition A.1). We define the following iterative scheme with $\tau_k > 0$ and $\langle \nabla F(\mu^k), \mu^\pi\rangle := \sum_{n=1}^N \langle \nabla f_n(\mu_n^k), \mu_n^\pi\rangle$:

$$\mu^{k+1} \in \arg\min_{\mu^\pi\in\mathcal{M}_{\mu_0}} \left\{\langle \nabla F(\mu^k), \mu^\pi\rangle + \frac{1}{\tau_k}\Gamma(\mu^\pi, \mu^k)\right\}, \quad (8)$$

where the idea is, at iteration $k+1$, to choose μ^π minimizing a linearization of the objective function around μ^k, the distribution sequence found at the previous iteration, and at the same time penalizing the distance between policy π inducing μ^π and π^k inducing μ^k. Our first main result of this section is in Theorem 4.1. It shows that, due to the choice of penalizing strategies, the iterative scheme in Equation (8) can be solved through dynamic programming (Bertsekas, 2005) by building a Bellman recursion:

Theorem 4.1. *Let $k \geq 0$. The solution of Problem (8) is $\mu^{k+1} = \mu^{\pi^{k+1}}$, where for all $n \in [N]$, and $(x,a) \in \mathcal{X}\times\mathcal{A}$,*

$$\pi_n^{k+1}(a|x) := \frac{\pi_n^k(a|x)\exp\big(\tau_k \tilde{Q}_n^k(x,a)\big)}{\sum_{a'\in\mathcal{A}} \pi_n^k(a'|x)\exp\big(\tau_k \tilde{Q}_n^k(x,a')\big)}, \quad (9)$$

where $\tilde{Q}$ is a regularized Q-function satisfying the following recursion

$$\begin{aligned}
\tilde{Q}_N^k(x,a) &= -\nabla f_N(\mu_N^k)(x,a)\\
\tilde{Q}_n^k(x,a) &= \max_{\pi_{n+1}\in(\Delta_\mathcal{A})^\mathcal{X}} \Bigg\{ -\nabla f_n(\mu_n^k)(x,a) + \\
&\quad \sum_{x'} p_{n+1}(x'|x,a) \sum_{a'} \pi_{n+1}(a'|x') \quad (10)\\
&\quad \left[-\frac{1}{\tau_k}\log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right) + \tilde{Q}_{n+1}^k(x',a')\right]\Bigg\}.
\end{aligned}$$

Proof. See Appendix B.1. □

It is not obvious at first sight, but we can show that Γ is a Bregman divergence, making the iterative scheme an instance of mirror descent (Beck and Teboulle, 2003). Therefore, we can state the convergence result of Algorithm 2, MD-CURL, in Theorem 4.2, the second main result of this section. Solving a mirror descent (MD) instance usually includes a projection step that is generally computationally expensive. The low-complexity methods existing in the literature can only offer approximate solutions (Dick et al., 2014; Rosenberg and Mansour, 2019). We show that with the judicious choice of divergence as in Equation (8), MD can be solved accurately avoiding all costly projection steps.

Theorem 4.2. *Let π^* be a minimizer of Problem (1). Define $L := \ell N$ where ℓ is the Lipschitz constant of f_n with respect to $\|\cdot\|_1$ for all $n \in [N]$. Applying K iterations of MD-CURL to this problem, with, for each $1 \leq k \leq K$, $\tau_k := L^{-1}\sqrt{2\Gamma(\mu^{\pi^*}, \mu^0)/K}$, gives the following convergence rate*

$$\min_{0\leq k\leq K} F(\mu^{\pi^k}) - F(\mu^{\pi^*}) \leq L\frac{\sqrt{2\Gamma(\mu^{\pi^*}, \mu^0)}}{\sqrt{K}}.$$

Algorithm 2 MD-CURL

Input: number of iterations K, initial sequence of policies $\pi^0 \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ such that $\mu^0 := \mu^{\pi_0} \in \mathcal{M}^*_{\mu_0}$, objective function $F := \sum_{n=1}^N f_n$, probability kernel $p = (p_n)_{n\in[N]}$, initial state-action distribution μ_0, sequence of non-negative learning rates $(\tau_k)_{k\le K}$.
for $k = 0, \dots, K-1$ **do**
  $\mu^k = \mu^{\pi^k}$ as in Equation (4)
  $\tilde{Q}^k_N(x,a) = -\nabla f_N(\mu^k_N)(x,a)$, $\forall (x,a) \in \mathcal{X}\times\mathcal{A}$
  for $n = N, \dots, 1$ **do**
    $\forall (x,a) \in \mathcal{X}\times\mathcal{A}$:
    $\pi^{k+1}_n(a|x) = \frac{\pi^k_n(a|x)\exp\left(\tau_k \tilde{Q}^k_n(x,a)\right)}{\sum_{a'} \pi^k_n(a'|x)\exp\left(\tau_k \tilde{Q}^k_n(x,a')\right)}$
    $\tilde{Q}^k_{n-1}(x,a)$ using the recursion in Equation (10)
  end for
end for
return π^K

Proof. For ease of notation, for any probability measure $\eta \in \Delta_E$, whatever the (finite) space E, we introduce the neg-entropy function, with the convention that $0\log(0) = 0$, $\phi(\eta) := \sum_{x\in E} \eta(x)\log\eta(x)$.

Proposition 4.3. *Let $\mu \in \mathcal{M}_{\mu_0}$ with marginal given by $\rho \in (\Delta_{\mathcal{X}})^N$. The divergence Γ is a Bregman divergence induced by*

$$\psi(\mu) := \sum_{n=1}^N \phi(\mu_n) - \sum_{n=1}^N \phi(\rho_n).$$

Also, ψ is 1-strongly convex with respect to $\|\cdot\|_{\infty,1}$.

The proof is in Appendix B.2 and consists in showing that the Γ divergence taking values on the sequence of state-action distributions is in fact the KL divergence on the joint distribution. Next, if f_n is convex and ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$ for all $n \in [N]$, then F is also convex and Lipschitz with constant $L := \ell N$ with respect to the norm $\|\cdot\|_{\infty,1}$ (see Appendix B.2). Since the set $\mathcal{M}_{\mu_0}$ is convex, all convergence assumptions of MD (Beck and Teboulle, 2003) are satisfied, and the rate of convergence follows. □

5 Online learning extension of CURL

We consider here the online variant of Problem (1), where the learner must compute a sequence of strategies while facing unknown external noise and arbitrarily changing objective functions. We introduce Greedy MD-CURL, a new algorithm achieving sub-linear regret with a simple closed-form solution. At episode t, Greedy MD-CURL solves an optimization problem in the MDP induced by the estimated probability kernel $\hat{p}^t$ using one iteration of MD-CURL. We refer to p as the true probability kernel and $\hat{p}^t$ as the estimated one.

5.1 Learning the model

Since the learner does not know the noise dynamics, it has to estimate it from its experience. To obtain a sub-linear regret, the learner must learn $\hat{p}^t$ in such a way that its distance to the real probability kernel decreases with t with high probability. Let us denote M^t_n the number of times the learner observes step n until the start of episode t, and ε^s_n the s-th noise observed at step n. Recall that the dynamics follow Equation (3), and that the learner observes the noise values from the agent's trajectory. Let δ_x be the Dirac distribution centered in x. We define

$$\hat{p}^t_n(\cdot|x,a) := \frac{1}{M^t_n}\sum_{s=1}^{M^t_n} \delta_{g_n(x,a,\varepsilon^s_n)}(\cdot). \tag{11}$$

For any function $\Lambda : \mathcal{X} \to \mathbb{R}$, for all $n \in [N]$ and $(x,a) \in \mathcal{X}\times\mathcal{A}$ we introduce the notation

$$\left(p_n - \hat{p}^t_n\right)(\Lambda)(x,a) := \sum_{x'\in\mathcal{X}} \left(p_n(x'|x,a) - \hat{p}^t_n(x'|x,a)\right)\Lambda(x').$$

We have the following concentration result.

Lemma 5.1. *Let $\gamma > 0$. For any $0 < \delta < 1$ and any function $\Lambda : \mathcal{X} \to \mathbb{R}$ such that $|\Lambda(x')| \le \sqrt{\gamma}/2$ for all $x' \in \mathcal{X}$,*

$$\left(p_n - \hat{p}^t_n\right)(\Lambda)(x,a) \le \sqrt{\frac{\gamma}{2M^t_n}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)}$$

holds with probability $1-\delta$ simultaneously for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, steps $n \in [N]$, and episodes $t \in [T]$.

Proof. See Appendix C.1. □

In the literature (Jaksch et al., 2008; Rosenberg and Mansour, 2019), it is common to bound the L_1 deviation between p and $\hat{p}^t$ instead. However, this deteriorates the bound by an additional factor of $\sqrt{|\mathcal{X}|}$, which we can avoid here because of our dynamics hypothesis. This means that in the final regret analysis, we only pay the number of states in a term proportional to $\sqrt{\log(|\mathcal{X}|)}$, which is an advantage for problems with large state spaces or even to discretize continuous state space problems.

We further state Lemma 5.2, which is proven in Appendix C.2 and used later to prove the regret bound of Greedy MD-CURL.

Lemma 5.2. *For any vector $v \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|}$, for any strategy π and for all $0 \le n \le N$,*

$$\langle v, \mu^{\pi,p}_n - \mu^{\pi,\hat{p}^t}_n\rangle = \sum_{i=1}^n \sum_{y\in\mathcal{X}\times\mathcal{A}} \mu^{\pi,\hat{p}^t}_{i-1}(y)\left(p_i - \hat{p}^t_i\right)(\Lambda^{i,n,\pi}_v)(y),$$

where $\Lambda_v^{i,n,\pi} : \mathcal{X} \to \mathbb{R}$ *is a function depending on* v, i, n *and* π *defined in Equation (25). Also, if* $\|v\|_\infty := \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |v(x,a)| \leq V$, *then* $\|\Lambda_v^{i,n,\pi}\|_\infty \leq V$.

5.2 Optimization problem

Recall that the learner follows the online protocol in Algorithm 1. At each episode, the learner estimates $\hat{p}^t$ from the noise observations using Equation (11), We denote by $\mathcal{M}_{\mu_0}^t := \mathcal{M}_{\mu_0}^{\hat{p}^t}$ the set induced by this estimate (as in Equation (6)). At every episode the learner solves

$$\mu^{t+1} \in \arg\min_{\mu\in\mathcal{M}_{\mu_0}^{t+1}} \{\tau\langle \nabla F^t(\mu^t), \mu\rangle + \Gamma(\mu, \tilde{\mu}^t)\}, \qquad (12)$$

where, $\mu^t := \mu^{\pi^t,\hat{p}^t}$ and $\tilde{\mu}^t := \mu^{\tilde{\pi}^t,\hat{p}^t}$ with

$$\tilde{\pi}^t := (1-\alpha_t)\pi^t + \tfrac{\alpha_t}{|\mathcal{A}|}, \qquad (13)$$

and $\alpha_t \in (0, 1/2)$ is an exploration parameter.

In Theorem 4.1, we have already shown that the optimization problem of Equation (12) with Bregman divergence Γ has the format of an exponential twist as in Equation (9). Consequently, we can build Greedy MD-CURL in Algorithm 3. Note that to compute the policy for episode $t+1$, we perform one iteration of MD-CURL using π^t to compute μ^t as in line 3 of Algorithm 2, $\tilde{\pi}^t$ to compute the exponential twist in line 7 and to compute $\tilde{Q}$ recursively in line 8, the objective function F^t and the estimated probability kernel $\hat{p}^{t+1}$.

Remark 5.3. *We call our algorithm Greedy because it solves the optimization problem (12) at each episode using the empirically estimated dynamics (11) as if they were the true ones, without confidence intervals or exploration bonuses related to visit counts as usually is the case (Jaksch et al., 2008; Rosenberg and Mansour, 2019; Azar et al., 2017).*

5.3 Regret analysis

In this section, we prove the regret bound of Greedy MD-CURL. For that, we use the results from Subsection 5.1 and some results of OMD (Shalev-Shwartz, 2012), while also having to handle an online optimization problem with varying constraint sets. We decompose the regret (5) into three terms,

$$\begin{aligned} R_T &= \sum_{t=1}^T F^t(\mu^{\pi^t,p}) - F^t(\mu^{\pi^t,\hat{p}^t}) \\ &+ \sum_{t=1}^T F^t(\mu^{\pi^t,\hat{p}^t}) - F^t(\mu^{\pi^*,\hat{p}^{t+1}}) \\ &+ \sum_{t=1}^T F^t(\mu^{\pi^*,\hat{p}^{t+1}}) - F^t(\mu^{\pi^*,p}) \\ &:= R_T^{MDP}((\pi^t)_{t\in[T]}) + R_T^{policy} + R_T^{MDP}(\pi^*), \end{aligned}$$

Algorithm 3 Greedy MD-CURL

Input: number of episodes T, initial sequence of policies $\pi^1 \in (\Delta_\mathcal{A})^{\mathcal{X}\times N}$, number of observations per episode M, initial state-action distribution μ_0, learning rate $\tau > 0$, sequence of parameters $(\alpha_t)_{t\in[T]}$.
Initialization: $\forall (x,a)$, $p^1(\cdot|x,a) = \frac{1}{|\mathcal{X}|}$
for $t = 1, \dots, T$ **do**
  for $j = 1, \dots, M$ **do**
    j-th agent starts at $(x_0^{j,t}, a_0^{j,t}) \sim \mu_0(\cdot)$
    for $n = 1, \dots, N$ **do**
      environment draws new state $x_n^{j,t} \sim p_n(\cdot|x_{n-1}^{j,t}, a_{n-1}^{j,t})$
      learner observes agent j's external noise $\varepsilon_n^{j,t}$
      agent j chooses an action $a_n^{j,t} \sim \pi_n^t(\cdot|x_n^{j,t})$
    end for
  end for
  update probability kernel estimate for all (x,a):

$$\hat{p}_n^{t+1}(\cdot|x,a) := \frac{1}{Mt}\sum_{j=1}^M \delta_{g_n(x,a,\varepsilon_n^{j,t})} + \frac{t-1}{t}\hat{p}_n^t(\cdot|x,a)$$

  compute policy for the next episode:
    $\pi^{t+1} :=$ MD-CURL$(1, \pi^t \backslash \tilde{\pi}^t, F^t, \hat{p}^{t+1}, \mu_0, \tau)$
  compute $\tilde{\pi}^{t+1}$ as in Equation (13)
end for
return $(\pi^t)_{t\in[T]}$

where $\pi^* := \arg\min_{\pi\in(\Delta_\mathcal{A})^{\mathcal{X}\times N}} \sum_{t=1}^T F^t(\mu^{\pi,p})$. The terms $R_T^{MDP}((\pi^t)_{t\in[T]})$ and $R_T^{MDP}(\pi^*)$ pay for the error due to not knowing the true probability kernel, and the term R_T^{policy} pays for calculating sub-optimal policies using MD-CURL with constraint sets varying with each episode. Propositions 5.5 and 5.7 bound each of these terms, yielding our main result:

Theorem 5.4. *Consider an episodic MDP with finite state space* $\mathcal{X}$, *finite action space* $\mathcal{A}$, *episodes of length* N, *and probability kernel* $p := (p_n)_{n\in[N]}$. *Let* $F^t := \sum_{n=1}^N f_n^t$ *convex with* f_n^t ℓ*-Lipschitz with respect to the norm* $\|\cdot\|_1$ *for all* $n \in [N], t \in [T]$. *Let*

$$b := \Big(\sum_{t=1}^T 2\Big[N\alpha_t + \tfrac{N^2}{t}\log\big(\tfrac{|\mathcal{A}|}{\alpha_t}\big) + N^2\big(\tfrac{1}{t}+\alpha_t\big)^2\Big] + \big(N\log(|\mathcal{A}|)\big)\Big)^{\frac{1}{2}} \qquad (14)$$

Then, with probability $1-\delta$, *Greedy MD-CURL obtains, for* $\tau = \frac{b}{L\sqrt{T}}$,

$$R_T \leq 2\ell N b\sqrt{T} + 2\ell N^2\sqrt{\frac{2T}{M}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)}.$$

In particular, choosing $\alpha_t = T^{-1}$ for all $t \in [T]$, yields $R_T = O(\sqrt{T}\log(T))$.

5.3.1 Bounding R_T^{MDP}

Here we show the bounds on $R_T^{MDP}\big((\pi^t)_{t\in[T]}\big)$ and $R_T^{MDP}(\pi^*)$. Both indicate the difference between the loss of playing a sequence of policies over T episodes in the actual MDP and the loss of playing the same sequence of policies but in the estimated MDP. For the first term, the sequence is that produced by Greedy MD-CURL, i.e. $(\pi^t)_{t\in[T]}$, and for the second term, it is the best stationary policy over the horizon T, i.e. π^*. The results are presented in Proposition 5.5 and use the lemmas from Subsection 5.1.

Proposition 5.5. *Under the same hypothesis as in Theorem 5.4, with probability $1-\delta$, Greedy MD-CURL obtains,*

$$R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) \leq \ell N^2 \sqrt{\frac{2T}{M}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)}.$$

The exact same result being also valid for $R_T^{MDP}(\pi^*)$.

Proof. See Appendix C.3. □

5.3.2 Bounding R_T^{policy}

The term R_T^{policy} pays for the loss associated with the convergence of MD-CURL. Our main challenge is to deal with the terms concerning variable constraint sets $\mathcal{M}_{\mu_0}^t$. They depend on a bound on the difference between the state-action distributions induced by two consecutive probability kernel estimates, i.e. $\|\mu^{\pi,\hat{p}^t} - \mu^{\pi,\hat{p}^{t+1}}\|_{\infty,1}$ stated in Lemma 5.6. We also need a bound on $\|\nabla\psi(\mu^{\pi,p})\|_{\infty,1}$, the function inducing the Bregman divergence, justifying our construction in Equation (13). The result is stated in Proposition 5.7.

Lemma 5.6. *For any policy sequence $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$, the estimation of the probability kernel for two consecutive episodes done by Greedy MD-CURL satisfies, for all episodes $t \in [T-1]$, the following inequality*

$$\|\mu^{\pi,\hat{p}^{t+1}} - \mu^{\pi,\hat{p}^t}\|_{\infty,1} \leq \frac{2N}{t}.$$

Proposition 5.7. *Under the same hypothesis as in Theorem 5.4, let b be defined as in Equation (14). Then, Greedy MD-CURL obtains, for $\tau = \frac{b}{L\sqrt{T}}$,*

$$R_T^{policy} \leq 2\ell N b\sqrt{T}.$$

Proof. See Appendix D.5. □

Remark 5.8. *Appendix E shows that Greedy MD-CURL also has sub-linear regret in T when both g_n and h_n are unknown and the learner observes the trajectory of state-action pairs that each agent follows. Although its regret is not the best in the state of the art, Greedy MD-CURL is a good option when exploration is already induced by the environment.*

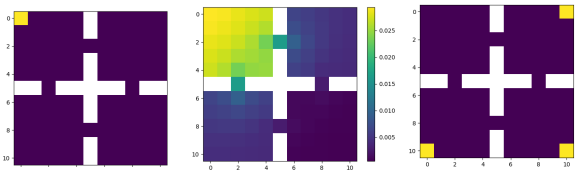


Figure 1: [left] Initial agent distribution; [middle] Distribution induced by the uniform policy; [right] The three targets.

6 Showcase experiments

In this section, we evaluate the performance of MD-CURL and Greedy MD-CURL on the *entropy maximisation* and *multi-objectives* problems, both introduced by Geist et al. (2022). To test Greedy MD-CURL's ability to learn the unknown dynamics, we consider a version with fixed, non-adversarial objective functions and the same probability kernel for all $n \in [N]$. Appendix F provides further experimental results.

6.1 Environments

We consider a model where the state space is a 11×11 four-room dimensional grid world with a single door connecting adjacent rooms. At each step, the agent can choose to stay still, go right, left, up or down, provided that there are no walls in the way:

$$x_{n+1} = x_n + a_n + \varepsilon_n, \tag{15}$$

with $a_n \in \{(0,0),(0,1),(1,0),(-1,0),(0,-1)\}$. The external noise ε_n represents a perturbation that pushes the agent to a neighbor state with a certain probability. We suppose the initial distribution is a Dirac at the upper left corner of the grid as in Figure 1 [left].

Entropy maximisation At each step, $f_n(\mu_n^{\pi,p}) := \langle \rho_n^{\pi,p}, \log(\rho_n^{\pi,p})\rangle$, where $\rho_n^{\pi,p}(x) := \sum_{a\in\mathcal{A}} \mu_n^{\pi,p}(x,a)$. Thus minimizing $F := \sum_{n=1}^N f_n$ means maximizing the entropy, so the optimal value is when the distribution is uniform over the state space (obs.: contrary to intuition, the uniform policy does not provide an optimal solution, as can be seen in Figure 1, [middle]).

Multi-objectives The goal is for the distribution to be concentrated on the three targets in Figure 1, [right], by the final step N. We let $f_n(\mu_n^{\pi,p}) := -\sum_{k=1}^3 (1 - \langle \rho_n^{\pi,p}, e^k\rangle)^2$, where $e^k \in \mathbb{R}^{|\mathcal{X}|}$ is a vector with zero everywhere and 1 in the element corresponding to a target state. Note that the target may not be reachable by any policy.

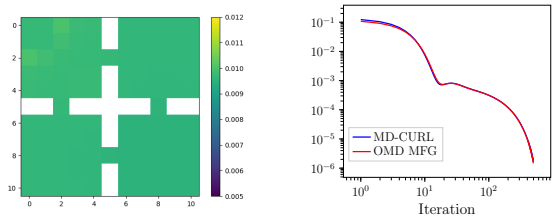


Figure 2: Entropy maximisation: [left] MD-CURL distribution at $N = 40$; [right] Log regret.

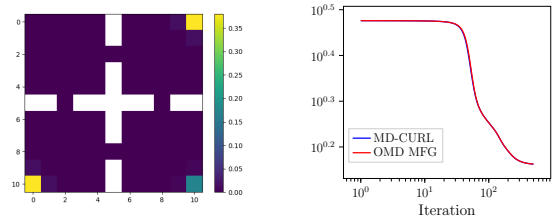


Figure 3: Multi-objectives: [left] MD-CURL distribution at $N = 40$; [right] Log regret.

6.2 Numerical experiments

For all experiments we consider $N = 40$. Figures 2 and 3 show at left the state distribution at $N = 40$ computed after 500 iterations of MD-CURL for each setting, and at right its log regret per iteration compared to that of OMD for MFG. The OMD algorithm for MFGs is the state-of-the-art method for the problems addressed in this paper, as shown by Laurière et al. (2022), but have no convergence results for discrete iterations. Therefore, MD-CURL is a good alternative for achieving state-of-the-art performance, with the advantage of having theoretical results. Note that both algorithms converge similarly, we leave the analysis of their differences for future work.

We now examine Greedy MD-CURL for online CURL. We add a noise ε_n that follows a categorical distribution h_n, with a 0.2 probability of going up and 0 for other directions. We suppose g_n is known but h_n is unknown, and we take $M = 10$. Figure 4 compares the log regret per iteration for Greedy MD-CURL (blue), MD-CURL with known noise dynamics (green), and MD-CURL with unknown noise dynamics, where the learner never learns the noise distribution, i.e. $\hat{p}_n^t(\cdot|x,a) = \delta_{g_n(x,a,0)}$ for all (x,a) (red). We see that Greedy MD-CURL quickly matches MD-CURL with known noise dynamics, and that never learning the noise is sub-optimal. We do not compare Greedy MD-CURL to other algorithms in the literature as none is well-suited to our scenario, and the ones that could be adapted use UCRL techniques making them computationally expensive or intractable.

Finally, Greedy MD-CURL achieves sub-linear regret

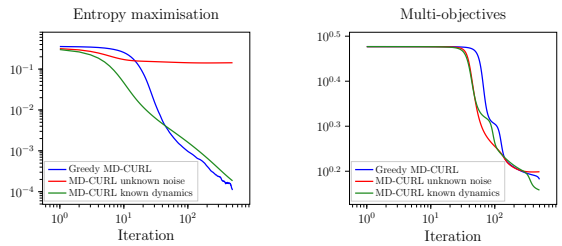


Figure 4: Log regret per iteration, $N = 40$: [left] Entropy maximisation; [right] Multi-objectives.

even with unknown dynamics (see Appendix E). Figure 5 shows how it learns the full dynamics for both the entropy maximization problem (right) and the multi-objective problem with 0.2 probability of being perturbed in any reachable neighboring state (left). It exploits the fact that maximizing entropy and perturbations with high probability favors exploration.

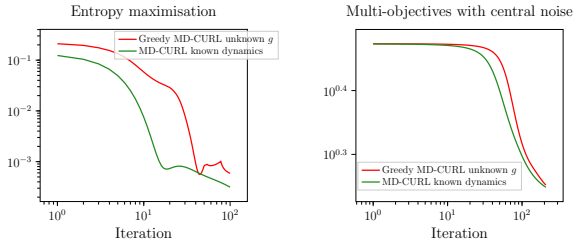


Figure 5: Log regret for Greedy MD-CURL with unknown g_n and h_n: [left] Entropy maximisation; [right] Multi-objectives.

7 Conclusion and future works

In this paper we analyzed two versions of the CURL problem in episodic MDPs with finite state and action spaces. For the offline optimization problem, where the dynamics g_n and h_n are known, we proposed an algorithm based on mirror descent converging with a rate of $O\big(1/\sqrt{K}\big)$ for K iterations. For the online learning extension with adversarial costs, we proposed an algorithm with a simple closed-form solution, and regret of $O\big(\sqrt{T}\log(T)\big)$ when g_n is known but h_n is unknown. Also, we showed that for this specific dynamic, we can improve the bounds of existing work and pay the number of states only in a term proportional to $\sqrt{\log(|\mathcal{X}|)}$.

A future direction is to investigate if we can achieve optimal regrets for variants of Greedy MD-CURL under more general assumptions about the available dynamics information. For example, by considering the case where g_n is a parametric function with unknown parameters, rather than being completely known.

References

Angiuli, A., Fouque, J., and Laurière, M. (2020). Unified reinforcement q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34:217 – 271.

Angiuli, A., Fouque, J.-P., and Lauriere, M. (2023). *Reinforcement Learning for Mean Field Games, with Applications to Economics*, page 393–425. Cambridge University Press.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, volume 70, pages 263–272.

Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175.

Bensoussan, A., Yam, P., and Frehse, J. (2013). *Mean Field Games and Mean Field Type Control Theory*. SpringerBriefs in Mathematics. Springer.

Bertsekas, D. P. (2005). *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, Belmont, MA, USA, 3rd edition.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3–4):231–357.

Carmona, R. and Laurière, M. (2022). Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: II—the finite horizon case. *The Annals of Applied Probability*, 32(6):4065 – 4105.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Coffman, A., Bušić, A., and Barooah, P. (2023). A unified framework for coordination of thermostatically controlled loads. *Automatica*, 152:111002.

Dick, T., György, A., and Szepesvári, C. (2014). Online learning in markov decision processes with changing cost sequences. In *International Conference on International Conference on Machine Learning*, volume 32, page I–512–I–520.

Even-Dar, E., Kakade, S. M., and Mansour, Y. (2009). Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110.

Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R., and Pietquin, O. (2022). Concave utility reinforcement learning: The mean-field game viewpoint. In *International Conference on Autonomous Agents and Multiagent Systems*, page 489–497, Richland, SC.

Ghasemipour, S. K. S., Zemel, R., and Gu, S. (2020). A divergence minimization perspective on imitation learning methods. In *Proceedings of the Conference on Robot Learning*, volume 100, pages 1259–1277.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019). Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, volume 97, pages 2681–2691.

Jaksch, T., Ortner, R., and Auer, P. (2008). Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600.

Laurière, M., Perrin, S., Geist, M., and Pietquin, O. (2022). Learning mean field games: A survey.

Lavigne, P. and Pfeiffer, L. (2023). Generalized conditional gradient and learning in potential mean field games.

Neu, G., Gyorgy, A., and Szepesvari, C. (2012). The adversarial stochastic shortest path problem with unknown transition probabilities. In *International Conference on Artificial Intelligence and Statistics*, volume 22, pages 805–813, La Palma, Canary Islands.

Osband, I., Russo, D., and Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, pages 3003–3011.

Pasztor, B., Bogunovic, I., and Krause, A. (2021). Efficient model-based multi-agent mean-field reinforcement learning. *Trans. Mach. Learn. Res.*, 2023.

Pérolat, J., Perrin, S., Elie, R., Laurière, M., Piliouras, G., Geist, M., Tuyls, K., and Pietquin, O. (2022). Scaling mean field games by online mirror descent. In *International Conference on Autonomous Agents and Multiagent Systems*, page 1028–1037, Richland, SC.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition.

Rosenberg, A. and Mansour, Y. (2019). Online convex optimization in adversarial Markov decision processes. In *International Conference on Machine Learning*, volume 97, pages 5478–5486.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

Séguret, A., Wan, C., and Alasseur, C. (2021). A mean field control approach for smart charging with aggregate power demand constraints. In *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, pages 01–05.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020). Variational policy gradient method for reinforcement learning with general utilities. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583.

Zimin, A. and Neu, G. (2013). Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, volume 26, pages 1583–1591.

A Equivalence between policies and distributions in $\mathcal{M}_{\mu_0}$

Proposition A.1. *Let $\mu_0 \in \Delta_{\mathcal{X}\times\mathcal{A}}$. The application $\pi \mapsto \mu^\pi$ is a bijection from $(\Delta_\mathcal{A})^{\mathcal{X}\times N}$ to $\mathcal{M}_{\mu_0}$.*

Proof. Consider a fixed initial state-action distribution $\mu_0 \in \Delta_{\mathcal{X}\times\mathcal{A}}$. Let $\mu \in \mathcal{M}_{\mu_0}$ and define $\rho = (\rho_n)_{0\le n\le N}$ such that for all $x \in \mathcal{X}$, $\rho_n(x) = \sum_a \mu_n(x,a)$ (the associated state distribution). First, let us deal with the case where $\rho_n(x) \neq 0$. Define a policy sequence $\pi \in (\Delta_\mathcal{A})^{\mathcal{X}\times N}$ such that $\pi_n(a|x) = \frac{\mu_n(x,a)}{\rho_n(x)}$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$. We want to show that $\mu^\pi = \mu$ for this policy π. We reason by induction. For $n = 0$, $\mu_0^\pi = \mu_0$ by definition. Suppose $\mu_n^\pi = \mu_n$, thus for $n+1$ and for all $(x',a') \in \mathcal{X}\times\mathcal{A}$

$$\begin{aligned}
\mu_{n+1}^\pi(x',a') &= \sum_{x,a} p_{n+1}(x'|x,a)\mu_n^\pi(x,a)\pi_{n+1}(a'|x') \\
&= \sum_{x,a} p_{n+1}(x'|x,a)\mu_n(x,a)\frac{\mu_{n+1}(x',a')}{\rho_{n+1}(x')} \\
&= \sum_{a} \mu_{n+1}(x',a)\frac{\mu_{n+1}(x',a')}{\rho_{n+1}(x')} \\
&= \rho_{n+1}(x')\frac{\mu_{n+1}(x',a')}{\rho_{n+1}(x')} \\
&= \mu_{n+1}(x',a'),
\end{aligned}$$

where the first equality comes from Equation (4), the second equality comes from the induction assumption and the way we defined the strategy π, and the third comes from the assumption that $\mu \in \mathcal{M}_{\mu_0}$.

In the case $\rho_n(x) = 0$, we therefore have $\mu_n(x,a) = 0$ for all $a \in \mathcal{A}$, so any choice of $\pi_n(a|x)$ would work. Because we want to make sure that there is a unique mapping from each π to μ^π we agree to always set $\pi_n(a|x) = \frac{1}{|\mathcal{A}|}$ in this case, where $|\mathcal{A}|$ is the number of possible actions.

□

B Proofs of Section 4: algorithm 2 scheme and convergence rate

By abuse of notation, for any probability measure $\eta \in \Delta_E$ whatever the finite space E on which it is defined we introduce the neg-entropy function, with the convention $0\log(0) = 0$,

$$\phi(\eta) := \sum_{x\in E} \eta(x)\log\eta(x), \tag{16}$$

to which we associate the Bregman divergence D, also known as the KL divergence, such that for any pair $(\eta,\nu) \in \Delta_E \times \Delta_E$,

$$D(\eta,\nu) := \phi(\eta) - \phi(\nu) - \langle\nabla\phi(\nu), \eta - \nu\rangle.$$

Let ρ_n denote the marginal probability distribution on $\mathcal{X}$ associated with μ_n i.e., for all $x \in \mathcal{X}$

$$\rho_n(x) := \sum_{a\in\mathcal{A}} \mu_n(x,a)\ .$$

Observe that to any $\mu = (\mu_n)_{1\le n\le N} \in \mathcal{M}_{\mu_0}$ one can associate a unique probability mass function on $\Delta_{(\mathcal{X}\times\mathcal{A})^N}$ denoted by $\mu_{1:N}$ such that $\mu_{1:N}$ is *generated* by the strategy $\pi = (\pi_n)_{n\in[N]}$ associated with μ which is determined by

$$\pi_n(a|x) = \frac{\mu_n(x,a)}{\rho_n(x)}\ ,$$

when $\rho_n(x) \neq 0$, otherwise we fix an arbitrary strategy $\pi_n(a|x) = \frac{1}{|\mathcal{A}|}$.

Before proving Theorems 4.1 and 4.2 we state and prove a lemma which is key to proving both theorems.

Lemma B.1. *For any $\mu \in \mathcal{M}_{\mu_0}$ and $\mu' \in \mathcal{M}^*_{\mu_0}$, with associated probability mass functions $\mu_{1:N}, \mu'_{1:N} \in \Delta_{(\mathcal{X}\times\mathcal{A})^N}$ generated by π, π' respectively with the same initial state-action distribution, i.e. $\mu_0 = \mu'_0$, we have*

$$D(\mu_{1:N}, \mu'_{1:N}) = \Gamma(\mu, \mu') = \sum_{n=1}^{N} D(\mu_n, \mu'_n) - \sum_{n=1}^{N} D(\rho_n, \rho'_n), \tag{17}$$

where $\Gamma(\mu, \mu') := \sum_{n=1}^{N} \mathbb{E}_{(x,a)\sim\mu_n(\cdot)} \left[\log\left(\frac{\pi_n(a|x)}{\pi'_n(a|x)}\right)\right]$.

Proof. For each $n \in [N]$, let us define a transition matrix P^{π_n} for all $x, x' \in \mathcal{X}$ and $a, a' \in \mathcal{A}$,

$$P^{\pi_n}(x', a'|x, a) := p_n(x'|x, a)\pi_n(a'|x').$$

Given Definition 4, for any randomized policy the state-action distributions evolve according to linear dynamics

$$\mu_n(x', a') = \langle \mu_{n-1}(\cdot), P^{\pi_n}(x', a'|\cdot)\rangle.$$

Any randomized policy π induces a probability mass function $\mu_{1:N}$ that is Markovian:

$$\mu_{1:N}(\vec{y}) = \mu_0(y_0)P^{\pi_1}(y_1|y_0)\dots P^{\pi_N}(y_N|y_{N-1}), \tag{18}$$

where $\vec{y}$ represents the elements of $(\mathcal{X}\times\mathcal{A})^{N+1}$ such that $y_i = (x_i, a_i)$ for all $0 \le i \le N$. Note that $\mu_n(y_n)$ is the marginal probability mass function.

Consider $\mu, \mu' \in \mathcal{M}_{\mu_0}$ the state-action distribution sequences induced by π, π' respectively (i.e, $\mu = \mu^\pi$ and $\mu' = \mu^{\pi'}$). Thus, computing the relative entropy between the probability mass functions $\mu_{1:N}, \mu'_{1:N}$ gives

$$\begin{aligned}
D(\mu_{1:N}, \mu'_{1:N}) &= \sum_{\vec{y}} \mu_{1:N}(\vec{y}) \log\left(\frac{\mu_{1:N}(\vec{y})}{\mu'_{1:N}(\vec{y})}\right) \\
&= \sum_{y_0,\dots,y_N} \mu_{1:N}(\vec{y}) \log\left(\frac{\mu_0(y_0)P^{\pi_1}(y_1|y_0)\dots P^{\pi_N}(y_N|y_{N-1})}{\mu'_0(y_0)P^{\pi'_1}(y_1|y_0)\dots P^{\pi'_N}(y_N|y_{N-1})}\right) \\
&= \sum_{y_0,\dots,y_N} \mu_{1:N}(\vec{y}) \sum_{i=1}^{N} \log\left(\frac{P^{\pi_i}(y_i|y_{i-1})}{P^{\pi'_i}(y_i|y_{i-1})}\right).
\end{aligned}$$

Where

$$\begin{aligned}
\sum_{i=1}^{N} \log\left(\frac{P^{\pi_i}(y_i|y_{i-1})}{P^{\pi'_i}(y_i|y_{i-1})}\right) &= \sum_{i=1}^{N} \log\left(\frac{p_i(x_i|x_{i-1}, a_{i-1})\pi_i(a_i|x_i)}{p_i(x_i|x_{i-1}, a_{i-1})\pi'_i(a_i|x_i)}\right) \\
&= \sum_{i=1}^{N} \log\left(\frac{\pi_i(a_i|x_i)}{\pi'_i(a_i|x_i)}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
D(\mu_{1:N}, \mu'_{1:N}) &= \sum_{\vec{y}} \mu_{1:N}(\vec{y}) \sum_{i=1}^{N} \log\left(\frac{\pi_i(a_i|x_i)}{\pi'_i(a_i|x_i)}\right) \\
&= \sum_{\vec{y}} \mu_0(y_0)P^{\pi_1}(y_1|y_0)\dots P^{\pi_N}(y_N|y_{N-1}) \sum_{i=1}^{N} \log\left(\frac{\pi_i(a_i|x_i)}{\pi'_i(a_i|x_i)}\right) \\
&= \sum_{i=1}^{N} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \mu_i(x, a) \log\left(\frac{\pi_i(a|x)}{\pi'_i(a|x)}\right).
\end{aligned}$$

Where for the last equality we used that

$$\sum_{y_0,\dots,y_{i-1}} \mu_0(y_0)P^{\pi_1}(y_1|y_0)\dots P^{\pi_i}(y_i|y_{i-1}) = \sum_{y_i}\mu_i(y_i)$$

and for a fixed y_i,

$$\sum_{y_{i+1},\dots,y_N} P^{\pi_{i+1}}(y_{i+1}|y_i)\dots P^{\pi_N}(y_N|y_{N-1}) = 1.$$

This proves the first equality of the lemma. We now prove the second. For this, we recall that Proposition A.1 gives a unique relation between a state-action distribution sequence $\mu \in \mathcal{M}_{\mu_0}$ and the policy sequence $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$ inducing it by taking for all $1 \le i \le N$, $(x,a) \in \mathcal{X}\times\mathcal{A}$,

$$\pi_i(a|x) = \frac{\mu_i(x,a)}{\rho_i(x)},$$

where ρ is the marginal on the states of μ. Using this relation, we have then that

$$\begin{aligned}
D(\mu_{1:N}, \mu'_{1:N}) &= \sum_{i=1}^N \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \mu_i(x,a)\log\left(\frac{\pi_i(a|x)}{\pi'_i(a|x)}\right)\\
&= \sum_{i=1}^N \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \mu_i(x,a)\log\left(\frac{\mu_i(a|x)}{\rho_i(x)}\frac{\rho'_i(x)}{\mu'_i(a|x)}\right)\\
&= \sum_{i=1}^N \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \mu_i(x,a)\log\left(\frac{\mu_i(a|x)}{\mu'_i(a|x)}\right) - \sum_{i=1}^N \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \mu_i(x,a)\log\left(\frac{\rho_i(x)}{\rho'_i(x)}\right)\\
&= \sum_{i=1}^N \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \mu_i(x,a)\log\left(\frac{\mu_i(a|x)}{\mu'_i(a|x)}\right) - \sum_{i=1}^N \sum_{x\in\mathcal{X}} \rho_i(x)\log\left(\frac{\rho_i(x)}{\rho'_i(x)}\right)\\
&= \sum_{i=1}^N D(\mu_i,\mu'_i) - \sum_{i=1}^N D(\rho_i,\rho'_i)
\end{aligned}$$

which concludes the proof. □

B.1 Proof of Theorem 4.1: formulation of Algorithm 2

Theorem. *Let $k \ge 0$. The solution of Problem (8) is $\mu^{k+1} = \mu^{\pi^{k+1}}$, where for all $n \in [N]$, and $(x,a) \in \mathcal{X}\times\mathcal{A}$,*

$$\pi_n^{k+1}(a|x) := \frac{\pi_n^k(a|x)\exp\left(\tau_k \tilde{Q}_n^k(x,a)\right)}{\sum_{a'\in\mathcal{A}}\pi_n^k(a'|x)\exp\left(\tau_k\tilde{Q}_n^k(x,a')\right)},$$

where $\tilde{Q}$ is a regularized Q-function satisfying the following recursion

$$\begin{cases}
\tilde{Q}_N^k(x,a) = -\nabla f_N(\mu_N^k)(x,a)\\
\tilde{Q}_n^k(x,a) = \max\limits_{\pi_{n+1}\in(\Delta_{\mathcal{A}})^{\mathcal{X}}} \Big\{ -\nabla f_n(\mu_n^k)(x,a) + \\
\qquad \sum_{x'} p_{n+1}(x'|x,a)\sum_{a'}\pi_{n+1}(a'|x')\Big[-\frac{1}{\tau_k}\log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right) + \tilde{Q}_{n+1}^k(x',a')\Big]\Big\}.
\end{cases}$$

Proof. At each iteration we seek to find μ^{k+1} a minimizer of

$$\min_{\mu^\pi\in\mathcal{M}_{\mu_0}}\left\{\langle\nabla F(\mu^k),\mu^\pi\rangle + \frac{1}{\tau_k}\sum_{n=1}^N \mathbb{E}_{(x,a)\sim\mu_n(\cdot)}\left[\log\left(\frac{\pi_n(a|x)}{\pi_n^k(a|x)}\right)\right]\right\} \tag{19}$$

where recall that $\langle \nabla F(\mu^k), \mu^\pi \rangle := \sum_{n=1}^N \langle \nabla f_n(\mu_n^k), \mu_n^\pi \rangle$. We further use that $r_n(x_n, a_n, \mu_n) := -\nabla f_n(\mu_n)(x_n, a_n)$.

Now, we use the optimality principle to solve this optimization problem with an algorithm backwards in time. Remember that the initial distribution μ_0 is always fixed. The equivalence between solving a minimization problem on sequences of state-action distributions in $\mathcal{M}_{\mu_0}$ and on sequences of policies in $(\Delta_\mathcal{A})^{\mathcal{X}\times N}$ (see Proposition A.1), allows us to reformulate Problem (19) on $\mathcal{M}_{\mu_0}$ into a problem on $(\Delta_\mathcal{A})^{\mathcal{X}\times N}$, thus

$$
\begin{aligned}
(19) = \max_{\pi \in (\Delta_\mathcal{A})^{\mathcal{X}\times N}} \Bigg\{ & \sum_{n=0}^{N} \sum_{x,a} \mu_n^\pi(x,a) r_n(x,a,\mu_n^k) \\
& - \frac{1}{\tau_k} \sum_{n=1}^{N} \sum_{x,a} \mu_{n-1}^\pi(x,a) \sum_{x',a'} p_n(x'|x,a)\pi_n(a'|x') \log\left(\frac{\pi_n(a'|x')}{\pi_n^k(a'|x')}\right) \Bigg\} \\
= \max_{\pi \in (\Delta_\mathcal{A})^{\mathcal{X}\times N}} \Bigg\{ & \sum_{n=0}^{N} \sum_{x,a} \mu_n^\pi(x,a) \Bigg[r_n(x,a,\mu_n^k) \\
& - \frac{1}{\tau_k} \sum_{x',a'} p_{n+1}(x'|x,a)\pi_{n+1}(a'|x') \log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right) \Bigg] \Bigg\} \\
= \max_{\pi \in (\Delta_\mathcal{A})^{\mathcal{X}\times N}} \Bigg\{ & \mathbb{E}_\pi \Bigg[r_N(x_N, a_N, \mu_N^k) + \sum_{n=0}^{N-1} r_n(x_n, a_n, \mu_n^k) \\
& - \frac{1}{\tau_k} \sum_{x',a'} p_{n+1}(x'|x_n,a_n)\pi_{n+1}(a'|x') \log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right) \Bigg] \Bigg\}.
\end{aligned}
$$

Let us define a regularized version of the state-action value function that we denote by $\tilde{Q}^k$, such that for all $i \in [N]$, $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$
\begin{aligned}
\tilde{Q}_i^k(x,a) = \max_{\pi_{i+1:N} \in (\Delta_\mathcal{A})^{\mathcal{X}\times(N-i)}} \mathbb{E}_\pi \Bigg[& r_N(x_N, a_N, \mu_N^k) + \sum_{n=i}^{N-1} \Bigg\{ r_n(x_n, a_n, \mu_n^k) \\
& - \frac{1}{\tau_k} \sum_{x',a'} p_{n+1}(x'|x_n,a_n)\pi_{n+1}(a'|x') \log\left(\frac{\pi_{n+1}(a'|x')}{\pi_{n+1}^k(a'|x')}\right) \Bigg\} \Bigg| (x_i,a_i) = (x,a) \Bigg],
\end{aligned}
\tag{20}
$$

where $\pi_{i+1:N} = \{\pi_{i+1}, \dots, \pi_N\}$.

First, note that $\mathbb{E}_{(x,a)\sim\mu_0(\cdot)}[\tilde{Q}_0^k(x,a)] = (19)$. Moreover, the optimality principle states that this regularized state-action value function satisfies the following recursion

$$
\begin{cases}
\tilde{Q}_N(x,a) = r_N(x,a,\mu_N^k) \\
\tilde{Q}_i(x,a) = \max_{\pi_{i+1} \in (\Delta_\mathcal{A})^\mathcal{X}} \Bigg\{ r_i(x,a,\mu_i^k) + \\
\qquad \sum_{x'} p_{i+1}(x'|x,a) \sum_{a'} \pi_{i+1}(a'|x') \left[-\frac{1}{\tau_k} \log\left(\frac{\pi_{i+1}(a'|x')}{\pi_{i+1}^k(a'|x')}\right) + \tilde{Q}_{i+1}(x',a') \right] \Bigg\}.
\end{cases}
$$

Thus, to solve (19) we compute backwards in time, i.e. for $i = N-1, \dots, 0$, for all $x \in \mathcal{X}$,

$$
\pi_{i+1}^{k+1}(\cdot|x) \in \operatorname*{arg\,max}_{\pi(\cdot|x) \in \Delta_\mathcal{A}} \left\{ \left\langle \pi(\cdot|x), \tilde{Q}_{i+1}^k(x,\cdot) \right\rangle - \frac{1}{\tau_k} D\big(\pi(\cdot|x), \pi_{i+1}^k(\cdot|x)\big) \right\},
$$

where D is the KL divergence.

The solution of this optimisation problem for each time step i can be found by writing the associated Lagrangian function $\mathcal{L}$. Let λ be the Lagrangian multiplier associated with the simplex constraint. For simplicity, let

$\pi_x := \pi(\cdot|x)$, $\pi_x^k := \pi_{i+1}^k(\cdot|x)$ and $\tilde{Q}_x^k := \tilde{Q}_{i+1}^k(x,\cdot)$. Thus,

$$\mathcal{L}(\pi_x, \lambda) = \langle \pi_x, \tilde{Q}_x^k \rangle - \frac{1}{\tau_k} D(\pi_x, \pi_x^k) - \lambda \left(\sum_{a \in \mathcal{A}} \pi_x(a) - 1 \right).$$

Taking the gradient of the Lagrangian with respect to $\pi_x(a)$ for each $a \in \mathcal{A}$ gives

$$\frac{\partial \mathcal{L}(\pi_x, \lambda)}{\partial \pi_x(a)} = \tilde{Q}_x^k(a) - \frac{1}{\tau_k} \log\left(\frac{\pi_x(a)}{\pi_x^k(a)}\right) - \frac{1}{\tau_k} - \lambda,$$

and thus

$$\frac{\partial \mathcal{L}(\pi_x, \lambda)}{\partial \pi_x(a)} = 0 \implies \quad \pi_x(a) = \pi_x^k(a) \exp\Big(\tau_k \tilde{Q}_x^k(a) - 1 - \tau_k \lambda\Big).$$

Applying the simplex constraint, $\sum_{a \in \mathcal{A}} \pi_x(a) = 1$, we find the value of the Lagrangian multipler λ, and we get for all $a \in \mathcal{A}$

$$\pi_x(a) = \frac{\pi_x^k(a) \exp\Big(\tau_k \tilde{Q}_x^k(a)\Big)}{\sum_{a' \in \mathcal{A}} \pi_x^k(a') \exp\Big(\tau_k \tilde{Q}_x^k(a')\Big)},$$

which proves the theorem.

□

B.2 Proof of Proposition 4.3

Proposition. *Let $\mu \in \mathcal{M}_{\mu_0}$ with marginal given by $\rho \in (\Delta_{\mathcal{X}})^N$. The divergence Γ is a Bregman divergence induced by the function*

$$\psi(\mu) := \sum_{n=1}^N \phi(\mu_n) - \sum_{n=1}^N \phi(\rho_n).$$

Also, ψ is 1-strongly convex with respect to the $\|\cdot\|_{\infty,1}$ norm.

Proof. Lemma B.1 states that for any $\mu \in \mathcal{M}_{\mu_0}$ and $\mu' \in \mathcal{M}_{\mu_0}^*$, induced by π, π' respectively as in Equation 4, with the same initial state-action distribution, i.e. $\mu_0 = \mu_0'$, we have

$$\Gamma(\mu, \mu') := \sum_{n=1}^N \mathbb{E}_{(x,a)\sim\mu_n(\cdot)} \left[\log\left(\frac{\pi_n(a'|x')}{\pi_n^k(a'|x')}\right)\right] = \sum_{n=1}^N D(\mu_n', \mu_n) - \sum_{n=1}^N D(\rho_n', \rho_n).$$

Recall that ϕ is the negentropy and that D is the Bregman divergence induced by the negentropy. Define the function $\psi : (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ such that for any $\mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$

$$\psi(\mu) := \sum_{n=1}^N \phi(\mu_n) - \sum_{n=1}^N \phi(\rho_n).$$

To show Γ is a Bregman divergence induced by ψ we need to show that for any $\mu, \mu' \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$,

$$\psi(\mu) - \psi(\mu') - \langle \nabla\psi(\mu'), \mu - \mu' \rangle = \Gamma(\mu, \mu').$$

For that, first recall that the marginal ρ is such that for each $1 \le n \le N$, and for all $x \in \mathcal{X}$, $\rho_n(x) = \sum_{a \in \mathcal{A}} \mu_n(x,a)$. Thus,

$$\begin{aligned} \psi(\mu) &= \sum_n \left[\sum_{x,a} \mu_n(x,a) \log(\mu_n(x,a)) - \sum_x \rho_n(x) \log(\rho_n(x)) \right] \\ &= \sum_n \sum_{x,a} \mu_n(x,a) \log\left(\frac{\mu_n(x,a)}{\sum_{a'} \mu_n(x,a')}\right). \end{aligned} \tag{21}$$

Computing the first order partial derivative of ψ with respect to $\mu_n(x,a)$ for any $(x,a) \in \mathcal{X} \times \mathcal{A}$ and $1 \leq n \leq N$, we get

$$\begin{aligned}\frac{\partial \psi}{\partial \mu_n(x,a)}(\mu) &= \log\left(\frac{\mu_n(x,a)}{\sum_{a'} \mu_n(x,a')}\right) + \mu_n(x,a)\frac{1}{\mu_n(x,a)} - \sum_{a'} \mu_n(x,a')\frac{1}{\sum_{a'} \mu_n(x,a')} \\ &= \log\left(\frac{\mu_n(x,a)}{\sum_{a'} \mu_n(x,a')}\right) = \log\left(\frac{\mu_n(x,a)}{\rho_n(x)}\right).\end{aligned}$$

Hence, as $\phi(\mu_n) = \langle \mu_n, \log(\mu_n)\rangle$ and $\phi(\rho_n) = \langle \rho_n, \log(\rho_n)\rangle$, and $\pi_n = \mu_n/\rho_n$,

$$\begin{aligned}\psi(\mu) - \psi(\mu') - \langle \nabla\psi(\mu'), \mu - \mu'\rangle &= \sum_{n=1}^{N} \left[\phi(\mu_n) - \phi(\rho_n) - \big(\phi(\mu'_n) - \phi(\rho'_n) - \big\langle \mu_n - \mu'_n, \log(\mu_n) - \log(\rho_n)\big\rangle\big)\right] \\ &= \sum_{n=1}^{N} \left[\phi(\mu_n) - \phi(\rho_n) - \mu_n \log(\pi'_n)\right] \\ &= \sum_{n=1}^{N} \left[\mu_n\big(\log(\pi_n) - \log(\pi'_n)\big)\right] \\ &= \Gamma(\mu, \mu').\end{aligned}$$

Now we just need to show that ψ is strongly convex. For that, we apply the following convexity property (Boyd and Vandenberghe, 2004): ψ is 1-strongly convex with respect to a norm $\|\cdot\|$ if and only if for all $\mu, \mu' \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$, $\langle \nabla\psi(\mu) - \nabla\psi(\mu'), \mu - \mu'\rangle \geq \|\mu - \mu'\|^2$. Indeed,

$$\begin{aligned}\langle \nabla\psi(\mu) - \nabla\psi(\mu'), \mu - \mu'\rangle &= \sum_n \sum_{x,a} \left[\frac{\partial \psi}{\partial \mu_n(x,a)}(\mu) - \frac{\partial \psi}{\partial \mu_n(x,a)}(\mu')\right]\big(\mu_n(x,a) - \mu'_n(x,a)\big) \\ &= \sum_n \sum_{x,a} \left[\log\left(\frac{\mu_n(x,a)}{\rho_n(x)}\right) - \log\left(\frac{\mu'_n(x,a)}{\rho'_n(x)}\right)\right]\big(\mu_n(x,a) - \mu'_n(x,a)\big) \\ &\overset{(a)}{=} \sum_n D(\mu_n, \mu'_n) + D(\mu_n, \mu'_n) - D(\rho_n, \rho'_n) - D(\rho'_n, \rho_n) \\ &\overset{(b)}{=} \Gamma(\mu, \mu') + \Gamma(\mu', \mu), \end{aligned} \tag{22}$$

where (a) comes from the definition of the KL divergence D and (b) comes from the definition of Γ.

It remains to find a norm that lower bound the right-hand side. By Lemma B.1,

$$\begin{aligned}\Gamma(\mu, \mu') &= \sum_{n=1}^{N} D(\mu_n, \mu'_n) - \sum_{n=1}^{N} D(\rho_n, \rho'_n) = D(\mu_{1:N}, \mu'_{1:N}) \\ &\geq 2\|\mu_{1:N} - \mu_{1:N}\|_{\mathrm{TV}}^2 = \frac{1}{2}\|\mu_{1:N} - \mu'_{1:N}\|_1^2,\end{aligned}$$

the last inequality being a consequence of Pinsker's inequality. The norm $\|\cdot\|_{\mathrm{TV}}$ stands for the total variation norm. Let y represent an element of $(\mathcal{X}\times\mathcal{A})^{N+1}$ such that $y_i \in \mathcal{X}\times\mathcal{A}$ for all $0 \leq i \leq N$. Observe that

$$\begin{aligned}\|\mu_{1:N} - \mu'_{1:N}\|_1 &= \sum_{y \in (\mathcal{X}\times\mathcal{A})^{N+1}} |\mu_{1:N}(y) - \mu'_{1:N}(y)| \\ &\geq \sum_{y_n \in \mathcal{X}\times\mathcal{A}} \left| \sum_{y_s \in \mathcal{X}\times\mathcal{A},\, s\neq n} \big(\mu_{1:N}(y) - \mu'_{1:N}(y)\big)\right| \\ &= \sum_{y_n \in \mathcal{X}\times\mathcal{A}} |\mu_n(y_n) - \mu'_n(y_n)|, \quad \text{for all } n \in \{0, \cdots, N\}.\end{aligned}$$

Thus,

$$\|\mu_{1:N} - \mu'_{1:N}\|_1 \quad \geq \quad \|\mu - \mu'\|_{\infty,1}$$

which implies that

$$\Gamma(\mu,\mu') \geq \frac{1}{2}\|\mu-\mu'\|_{\infty,1}^2. \tag{23}$$

Finally, plugging back into Equation (22) shows that ψ is 1-strongly convex with respect to the norm $\|\cdot\|_{\infty,1}$.

□

B.3 Complements of the proof of Theorem 4.2

Lemma B.2. *Let $f_n : \Delta_{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}$ be convex and ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$ for all $n \in [N]$. If $F : (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ is defined for all $\mu := (\mu_n)_{n\in[N]} \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$ as $F(\mu) := \sum_{n=1}^N f_n(\mu_n)$, then F is also convex and L-Lipschitz with respect to the norm $\|\cdot\|_{\infty,1}$ for $L = \ell N$.*

Proof. **Convexity:** F is convex as the sum of convex functions.

Lipschitz: Let $\mu,\mu' \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$. As f_n is Lipschitz with respect to $\|\cdot\|_1$ with constant ℓ, then $|f_n(\mu_n) - f_n(\mu'_n)| \leq \ell\|\mu_n-\mu_n\|_1$ for all $1 \leq n \leq N$. Therefore,

$$\begin{aligned}|F(\mu)-F(\mu')| &= \left|\sum_{n=1}^N f_n(\mu_n) - f_n(\mu'_n)\right| \leq \sum_{n=1}^N |f_n(\mu_n)-f_n(\mu'_n)| \\ &\leq \ell\sum_{n=1}^N \|\mu_n-\mu'_n\|_1 \leq \ell N\|\mu-\mu'\|_{\infty,1}.\end{aligned}$$

□

C Proofs of Subsection 5.1: concentration results

C.1 Proof of Lemma 5.1

Lemma. *Let $\gamma > 0$. For any $0 < \delta < 1$, and for any function $\Lambda : \mathcal{X} \to \mathbb{R}$ such that $|\Lambda(x')| \leq \sqrt{\gamma}/2$ for all $x' \in \mathcal{X}$,*

$$\big(p_n - \hat{p}_n^t\big)(\Lambda)(x,a) \leq \sqrt{\frac{\gamma}{2M_n^t}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)}$$

holds with probability $1-\delta$ simultaneously for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, steps $n \in [N]$, and episodes $t \in [T]$.

Proof. Let $\gamma > 0$. Recall that, for all $n \in [N]$, and for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, $p_n(x|x,a) := \mathbb{P}\big(g_n(x,a,\varepsilon_n) = x\big)$ and $\hat{p}_n^t(x|x,a) = \frac{1}{M_n^t}\sum_{s=1}^{M_n^t}\delta_{g_n(x,a,\varepsilon_n^s)}$ where M_n^t is the number of times we observe step n until the start of episode t, and ε_n^s is the s-th noise observed at step n. Note that M_n^t is not random. Therefore,

$$\big(p_n-\hat{p}_n^t\big)(\Lambda)(x,a) := \sum_{x'\in\mathcal{X}}\big(p_n(x'|x,a)-\hat{p}_n^t(x'|x,a)\big)\Lambda(x') = \mathbb{E}_{\varepsilon_n\sim h_n(\cdot)}\big[\Lambda\big(g_n(x,a,\varepsilon_n)\big)\big] - \frac{1}{M_n^t}\sum_{s=1}^{M_n^t}\Lambda\big(g_n(x,a,\varepsilon_n^s)\big).$$

From the hypothesis on the bound of Λ, we have that almost surely $\Lambda\big(g_n(x,a,\varepsilon_i^s)\big) \in [-\sqrt{\gamma}/2, \sqrt{\gamma}/2]$ for all $s \in M_n^t$, therefore, applying Hoeffding's inequality to the sequence of random variables $\big(\Lambda\big(g_n(x,a,\varepsilon_i^s)\big)\big)_{s\in[M_n^t]}$ yields, for all $\xi > 0$,

$$\mathbb{P}\Big(\big(p_n-\hat{p}_n^t\big)(\Lambda)(x,a) \geq \xi\Big) \leq \exp\left(\frac{-2\xi^2 M_n^t}{\gamma}\right). \tag{24}$$

Applying the union bound we then get that simultaneously for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, steps $n \in [N]$ and episodes $t \in [T]$,

$$\big(p_n-\hat{p}_n^t\big)(\Lambda)(x,a) \leq \sqrt{\frac{\gamma}{2M_n^t}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)}$$

holds with probability $1-\delta$ for any $0 < \delta < 1$. □

C.2 Proof of Lemma 5.2

Lemma. *For any vector $v \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|}$, for any strategy π and for all $n \in [N]$,*

$$\langle v, \mu_n^{\pi,p} - \mu_n^{\pi,\hat{p}^t}\rangle = \sum_{i=1}^{n} \sum_{y\in\mathcal{X}\times\mathcal{A}} \mu_{i-1}^{\pi,\hat{p}^t}(y)\big(p_i - \hat{p}_i^t\big)(\Lambda_v^{i,n,\pi})(y),$$

where $\Lambda_v^{i,n,\pi} : \mathcal{X} \to \mathbb{R}$ is a function depending on v, i, n and π defined in Equation (25). Also, if $\|v\|_\infty := \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |v(x,a)| \le V$, then $\|\Lambda_v^{i,n,\pi}\|_\infty \le V$.

Proof. For $y \in \mathcal{X}\times\mathcal{A}$, we denote by $v(y)$ the element y of vector v.

For all $n \in [N]$, for all $y := (x,a) \in \mathcal{X}\times\mathcal{A}$ and $y' := (x',a') \in \mathcal{X}\times\mathcal{A}$, let

$$K_n(y,y') := p_n(x|x',a')\pi_n(a|x),$$
$$\hat{K}_n^t(y,y') := \hat{p}_n^t(x|x',a')\pi_n(a|x).$$

For η a vector over $\mathcal{X}\times\mathcal{A}$, we define for all $y \in \mathcal{X}\times\mathcal{A}$ and $y_0 \in \mathcal{X}\times\mathcal{A}$ the following notations

$$\eta K_{1:n}(y) := \sum_{y_0\in\mathcal{X}\times\mathcal{A}} \dots \sum_{y_{n-1}\in\mathcal{X}\times\mathcal{A}} \eta(y_0)K_1(y_0,y_1)\dots K_n(y,y_{n-1})$$
$$\eta(y_0)K_{1:n}(y) := \sum_{y_1\in\mathcal{X}\times\mathcal{A}} \dots \sum_{y_{n-1}\in\mathcal{X}\times\mathcal{A}} \eta(y_0)K_1(y_0,y_1)\dots K_n(y,y_{n-1}).$$

We can then rewrite the definition of a state-action distribution satisfying the Markovian dynamics and induced by a policy π stated in Equation (4) as $\mu_n^{\pi,p} = \mu_0 K_{1:n}$, and $\mu_n^{\pi,\hat{p}^t} = \mu_0\hat{K}_{1:n}^t$. With the convention that $K_{n+1:n} := \mathrm{Id}$ is the identity operator for all n, then

$$\begin{aligned}
\mu_n^{\pi,p} - \mu_n^{\pi,\hat{p}^t} &= \mu_0 K_{1:n} - \mu_0\hat{K}_{1:n}^t \\
&= (\mu_0 K_{1:n} - \mu_0\hat{K}_1^t K_{2:n}) + (\mu_0\hat{K}_1^t K_{2:n} - \mu_0\hat{K}_{1:2}^t K_{3:n}) + \dots + (\mu_0\hat{K}_{1:n-1}^t K_n - \mu_0\hat{K}_{1:n}^t) \\
&= \sum_{i=1}^{n} \mu_{i-1}^{\pi,\hat{p}^t}\big(K_i - \hat{K}_i^t\big)K_{i+1:n}.
\end{aligned}$$

Note that, for all $y \in \mathcal{X}\times\mathcal{A}$, and all $i \in \{0,\dots,n\}$,

$$\begin{aligned}
\mu_{i-1}^{\pi,\hat{p}^t}\big(K_i - \hat{K}_i^t\big)K_{i+1:n}(y) &= \sum_{y_{i-1}} \mu_{i-1}^{\pi,\hat{p}^t}(y_{i-1}) \sum_{y_i} \big(K_i(y_{i-1},y_i) - \hat{K}_i^t(y_{i-1},y_i)\big)K_{i+1:n}(y) \\
&= \sum_{y_{i-1}} \mu_{i-1}^{\pi,\hat{p}^t}(y_{i-1}) \sum_{x_i} \big(p_i(x_i|y_{i-1}) - \hat{p}_i^t(x_i|y_{i-1})\big) \sum_{a_i} \pi_i(a_i|x_i)K_{i+1:n}(y).
\end{aligned}$$

Hence,

$$\begin{aligned}
\langle v, \mu_n^{\pi,p} - \mu_n^{\pi,\hat{p}^t}\rangle &= \sum_y v(y) \sum_{i=1}^{n} \mu_{i-1}^{\pi,\hat{p}^t}\big(K_i - \hat{K}_i^t\big)K_{i+1:n}(y) \\
&= \sum_{i=1}^{n} \sum_{y_{i-1}\in\mathcal{X}\times\mathcal{A}} \mu_{i-1}^{\pi,\hat{p}^t}(y_{i-1}) \sum_{x_i\in\mathcal{X}} \big(p_i(x_i|y_{i-1}) - \hat{p}_i^t(x_i|y_{i-1})\big) \sum_{a_i\in\mathcal{A}} \pi_i(a_i|x_i) \sum_{y\in\mathcal{X}\times\mathcal{A}} K_{i+1:n}(y)v(y) \\
&= \sum_{i=1}^{n} \sum_{y_{i-1}\in\mathcal{X}\times\mathcal{A}} \mu_{i-1}^{\pi,\hat{p}^t}(y_{i-1}) \sum_{x_i\in\mathcal{X}} \big(p_i(x_i|y_{i-1}) - \hat{p}_i^t(x_i|y_{i-1})\big)\Lambda_v^{i,n,\pi}(x_i) \\
&:= \sum_{i=1}^{n} \sum_{y_{i-1}\in\mathcal{X}\times\mathcal{A}} \mu_{i-1}^{\pi,\hat{p}^t}(y_{i-1})\big(p_i - \hat{p}_i^t\big)(\Lambda_v^{i,n,\pi})(y_{i-1})
\end{aligned}$$

where we define the function $\Lambda_v^{i,n,\pi} : \mathcal{X} \to \mathbb{R}$ for any $v \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ as

$$\Lambda_v^{i,n,\pi}(x) := \sum_{a\in\mathcal{A}} \pi_i(a|x) \sum_{y\in\mathcal{X}\times\mathcal{A}} K_{i+1:n}(y)v(y). \tag{25}$$

If $\|v\|_\infty \le V$, then for all $x \in \mathcal{X}$,

$$\begin{aligned}
|\Lambda_v^{i,n,\pi}(x)| &\le \sum_{a\in\mathcal{A}} \pi_i\big(a|x\big) \sum_{y'\in\mathcal{X}\times\mathcal{A}} K_{i+1:n}(y')|v(y')| \\
&\le V \sum_{a\in\mathcal{A}} \pi_i\big(a|x\big) \sum_{y'\in\mathcal{X}\times\mathcal{A}} K_{i+1:n}(y') \\
&= V.
\end{aligned}$$

Therefore, $\|\Lambda_v^{i,n,\pi}\|_\infty \le V$. □

C.3 Proof of Proposition 5.5: upper bound on R_T^{MDP}

Proposition. *We consider an episodic MDP with finite state space $\mathcal{X}$, finite action space $\mathcal{A}$, episodes of length N, and probability kernel $p := (p_n)_{n\in[N]}$. We let $F^t := \sum_{n=1}^N f_n^t$ convex with f_n^t ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$ for all $n \in [N], t \in [T]$. Then, with probability $1-\delta$, Greedy MD-CURL obtains,*

$$R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) \le \ell N^2 \sqrt{\frac{2T}{M}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)}.$$

The exact same result being also valid for $R_T^{MDP}\big(\pi^*\big)$.

Proof. The proof steps are the same for both terms, hence we show only the steps for $R_T^{MDP}\big((\pi^t)_{t\in[T]}\big)$. Using the convexity of F^t we obtain

$$R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) \le \sum_{t=1}^T \langle \nabla F^t(\mu^{\pi^t,p}), \mu^{\pi^t,p} - \mu^{\pi^t,\hat{p}^t}\rangle = \sum_{t=1}^T\sum_{n=1}^N \langle \nabla f_n^t(\mu^{\pi^t,p}), \mu_n^{\pi^t,p} - \mu_n^{\pi^t,\hat{p}^t}\rangle.$$

To bound the inner product for each n, we first use the result of Lemma 5.2 to obtain that

$$\langle \nabla f_n^t(\mu^{\pi^t,p}), \mu_n^{\pi^t,p} - \mu_n^{\pi^t,\hat{p}^t}\rangle = \sum_{i=1}^n \sum_{y\in\mathcal{X}\times\mathcal{A}} \mu_{i-1}^{\pi,\hat{p}^t}(y)\big(p_i - \hat{p}_i^t\big)\big(\Lambda^{i,n,\pi^t}_{\nabla f_n^t(\mu^{\pi^t,p})}\big)(y).$$

As f_n^t is ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$ for all n and t, then for all state-action distribution $\mu_n \in \Delta_{\mathcal{X}\times\mathcal{A}}$, $\|\nabla f_n^t(\mu_n)\|_\infty := \sup_{(x,a)} |\nabla f_n^t(\mu_n)(x,a)| \le \ell$. Hence, from the second result of Lemma 5.2 we have $\big\|\Lambda^{i,n,\pi^t}_{\nabla f_n^t(\mu^{\pi^t,p})}\big\|_\infty \le \ell$. Therefore, all the conditions of Lemma 5.1 are satisfied with $\gamma = 4\ell^2$, meaning that

$$\begin{aligned}
\sum_{t=1}^T\sum_{n=1}^N \langle \nabla f_n^t(\mu_n), \mu_n^{\pi^t,p} - \mu_n^{\pi^t,\hat{p}^t}\rangle &= \sum_{t=1}^T\sum_{n=1}^N\sum_{i=1}^{n-1} \sum_{y\in\mathcal{X}\times\mathcal{A}} \mu_{i-1}^{\pi^t,\hat{p}^t}(y)\big(p_i - \hat{p}_i^t\big)\big(\Lambda^{i,n,\pi^t}_{\nabla f_n^t(\mu^{\pi^t,p})}\big)(y) \\
&\le \sum_{t=1}^T N^2\ell\sqrt{\frac{2}{Mt}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)} \\
&\le N^2\ell\sqrt{\frac{2T}{M}\log\left(\frac{N|\mathcal{X}||\mathcal{A}|T}{\delta}\right)}
\end{aligned}$$

holds with probability $1-\delta$, where we use that in Greedy MD-CURL, $M_n^t = M(t-1)$ for all $n \in [N]$. □

D Proofs of Theorem 5.4: upper bound on R_T

D.1 Auxiliary result: L_1 bound between distributions induced by the same policy but different probability kernels

The bound of Theorem 5.4 depends on the auxiliary lemma bellow stating that the L_1 deviation of two state-action distributions induced by the same policies but different probability kernels is bounded by the L_1 difference between the probability kernels.

Lemma D.1. *For any strategy $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{X}\times N}$, for any two probability kernels $p = (p_n)_{n\in[N]}$ and $q = (q_n)_{n\in[N]}$ such that $p_n, q_n : \mathcal{X}\times\mathcal{A}\times\mathcal{X} \to [0,1]$, and for all $n \in [N]$,*

$$\|\mu_n^{\pi,p} - \mu_n^{\pi,q}\|_1 \leq \sum_{i=0}^{n-1}\sum_{x,a} \mu_i^{\pi,p}(x,a)\|p_{i+1}(\cdot|x,a) - q_{i+1}(\cdot|x,a)\|_1.$$

Proof. From the definition of a state-action distribution sequence induced by a policy π in a probability kernel p in Equation (4), we have that for all $(x,a) \in \mathcal{X}\times\mathcal{A}$ and $n \in [N]$,

$$\mu_n^{\pi,p}(x,a) = \sum_{x',a'} \mu_{n-1}^{\pi,p}(x',a')p_n(x|x',a')\pi_n(a|x).$$

Thus,

$$\begin{aligned}
\|\mu_n^{\pi,p} - \mu_n^{\pi,q}\|_1 &= \sum_{x,a} \left|\mu_n^{\pi,p}(x,a) - \mu_n^{\pi,q}(x,a)\right| \\
&= \sum_{x,a}\sum_{x',a'} \left|\mu_{n-1}^{\pi,p}(x',a')p_n(x|x',a') - \mu_{n-1}^{\pi,q}(x',a')q_n(x|x',a')\right|\pi_n(a|x) \\
&= \sum_{x}\sum_{x',a'} \left|\mu_{n-1}^{\pi,p}(x',a')p_n(x|x',a') - \mu_{n-1}^{\pi,q}(x',a')q_n(x|x',a')\right| \\
&= \sum_{x}\sum_{x',a'} \Big|\mu_{n-1}^{\pi,p}(x',a')p_n(x|x',a') - \mu_{n-1}^{\pi,p}(x',a')q_n(x|x',a') \\
&\qquad + \mu_{n-1}^{\pi,p}(x',a')q_n(x|x',a') - \mu_{n-1}^{\pi,q}(x',a')q_n(x|x',a')\Big| \\
&\leq \sum_{x',a'} \mu_{n-1}^{\pi,p}(x',a')\|p_n(\cdot|x',a') - q_n(\cdot|x',a')\|_1 + \sum_{x',a'}\left|\mu_{n-1}^{\pi,p}(x',a') - \mu_{n-1}^{\pi,q}(x',a')\right| \\
&= \sum_{x',a'} \mu_{n-1}^{\pi,p}(x',a')\|p_n(\cdot|x',a') - q_n(\cdot|x',a')\|_1 + \|\mu_{n-1}^{\pi,p} - \mu_{n-1}^{\pi,q}\|_1.
\end{aligned}$$

Since for $n = 0$, $\|\mu_0^{\pi,p} - \mu_0^{\pi,q}\|_1 = 0$, by induction we get that

$$\|\mu_n^{\pi,p} - \mu_n^{\pi,q}\|_1 \leq \sum_{i=0}^{n-1}\sum_{x',a'} \mu_i^{\pi,p}(x',a')\|p_{i+1}(\cdot|x',a') - q_{i+1}(\cdot|x',a')\|_1.$$

□

D.2 Auxiliary result: upper bound on $-\psi$

Lemma D.2 shows that the function $-\psi$, where ψ is the function that induces the Bregman divergence Γ according to Proposition 4.3, is upper bounded. The definition of ψ is recalled in the lemma.

Lemma D.2. *Let ϕ be the neg-entropy function defined in Equation (16). Let $\psi : (\Delta_{\mathcal{X}\times\mathcal{A}})^N \to \mathbb{R}$ such that for all $\mu := (\mu_n)_{n\in[N]} \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$, where we let $\rho := (\rho_n)_{n\in[N]} \in (\Delta_{\mathcal{X}})^N$ be the associated sequence of marginals,*

$$\psi(\mu) := \sum_{n=1}^{N}\phi(\mu_n) - \sum_{n=1}^{N}\phi(\rho_n).$$

Then, $\sup_{\mu\in(\Delta_{\mathcal{X}\times\mathcal{A}})^N} -\psi(\mu) \leq N\log(|\mathcal{A}|)$.

Proof. For $\mu \in (\Delta_{\mathcal{X}\times\mathcal{A}})^N$ and for Lagrangian multipliers $\lambda_n \in \mathbb{R}$ associated to the constraints $\sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} \mu_n(x,a) = 1$ for all $n \in [N]$, consider the Lagrangian given by

$$\mathcal{L}(\mu,\lambda) = \psi(\mu) + \sum_n \lambda_n \left(1 - \sum_{x,a} \mu_n(x,a)\right).$$

For every $n \in [N]$, and $(x,a) \in \mathcal{X}\times\mathcal{A}$,

$$\frac{\partial \mathcal{L}(\mu,\lambda)}{\partial \mu_n(x,a)} = \log\left(\frac{\mu_n(x,a)}{\sum_{a'}\mu_n(x,a')}\right) - \lambda_n = 0,$$

thus

$$\frac{\mu_n(x,a)}{\sum_{a'}\mu_n(x,a')} = \exp(\lambda_n).$$

To satisfy the constraint for each n, we need

$$\sum_{x,a} \mu_n(x,a) = \sum_{x,a} \exp(\lambda_n) \sum_{a'} \mu_n(x,a') = \exp(\lambda_n)|\mathcal{A}| = 1.$$

Using the decomposition of ψ proved in Equation (21), we get

$$-\psi(\mu) = \sum_n \sum_{x,a} \mu_n(x,a) \log\left(\frac{\sum_{a'}\mu_n(x,a')}{\mu_n(x,a)}\right) \leq \sum_n \sum_{x,a} \mu_n(x,a)\log(|\mathcal{A}|) = N\log(|\mathcal{A}|).$$

□

D.3 Proof of Lemma 5.6

Lemma. *For any policy sequence π, the estimation of the probability kernel for two consecutive episodes done by Greedy MD-CURL satisfies, for all episodes $t \in [T-1]$, the following inequality*

$$\|\mu^{\pi,\hat{p}^{t+1}} - \mu^{\pi,\hat{p}^t}\|_{\infty,1} \leq \frac{2N}{t}.$$

Proof. For all $(x,a,x') \in \mathcal{X}\times\mathcal{A}\times\mathcal{X}$, and for all $i \in [N]$,

$$\begin{aligned}
\|\hat{p}_i^{t+1}(\cdot|x,a) - \hat{p}_i^t(\cdot|x,a)\|_1 &= \sum_{x'\in\mathcal{X}} |\hat{p}_i^{t+1}(x'|x,a) - \hat{p}_i^t(x'|x,a)| \\
&= \sum_{x'\in\mathcal{X}} \left| \frac{1}{Mt}\left(\sum_{j=1}^M \delta_{g_i(x,a,\varepsilon_i^{j,t})}(x') + M(t-1)\hat{p}_i^t(x'|x,a)\right) - \hat{p}_i^t(x'|x,a)\right| \qquad (26) \\
&= \frac{1}{Mt}\sum_{x'\in\mathcal{X}} \left|\sum_{j=1}^M \delta_{g_i(x,a,\varepsilon_i^{j,t})}(x') - \hat{p}_i^t(x'|x,a)\right| \leq \frac{2}{t}.
\end{aligned}$$

Therefore, using the result of Lemma D.1 with $\hat{p}_i^t$ and $\hat{p}_i^{t+1}$,

$$\begin{aligned}
\|\mu^{\pi,\hat{p}^{t+1}} - \mu^{\pi,\hat{p}^t}\|_{\infty,1} &= \sup_{n\in[N]} \|\mu_n^{\pi,\hat{p}^{t+1}} - \mu_n^{\pi,\hat{p}^t}\|_1 \\
&\leq \sup_{n\in[N]} \sum_{i=0}^{n-1} \sum_{x,a} \mu_i^{\pi,\hat{p}^t}(x,a)\|\hat{p}_i^{t+1}(\cdot|x,a) - \hat{p}_i^t(\cdot|x,a)\|_1 \\
&\leq \frac{2N}{t}.
\end{aligned}$$

□

D.4 Auxiliary result: bound between distributions induced by π^t and $\tilde{\pi}^t$

Lemma D.3. *For all episodes $t \in [T]$, where π^t is the strategy calculated using Greedy MD-CURL, $\tilde{\pi}^t = (1-\alpha_t)\pi^t + \frac{\alpha_t}{|\mathcal{A}|}$, and $\hat{p}^t, \hat{p}^{t+1}$ are two consecutive estimates of the probability kernel by Greedy MD-CURL, we have*

$$\|\mu^{\pi^t,\hat{p}^t} - \mu^{\tilde{\pi}^t,\hat{p}^{t+1}}\|_{\infty,1} \leq \sup_{n\in[N]} \left\{ \sum_{i=1}^{n-1} \sum_{x,a} \mu_i^{\pi^t,\hat{p}^t}(x,a) \|\hat{p}_{i+1}^t(\cdot|x,a) - \hat{p}_{i+1}^{t+1}(\cdot|x,a)\|_1 + 2n\alpha_t \right\}.$$

Proof. Using similar arguments as in Lemma D.1, we get that for all $n \in [N]$,

$$\begin{aligned}
\|\mu_n^{\pi^t,\hat{p}^t} - \mu_n^{\tilde{\pi}^t,\hat{p}^{t+1}}\|_1 &= \sum_{x,a} \left|\mu_n^{\pi^t,\hat{p}^t}(x,a) - \mu_n^{\tilde{\pi}^t,\hat{p}^{t+1}}(x,a)\right| \\
&\leq \sum_{x,a}\sum_{x',a'} \left| \mu_{n-1}^{\pi^t,\hat{p}^t}(x',a')\hat{p}_n^t(x|x',a')\pi_n^t(a|x) - \mu_{n-1}^{\tilde{\pi}^t,\hat{p}^{t+1}}(x',a')\hat{p}_n^{t+1}(x|x',a')\left((1-\alpha_t)\pi_n^t(a|x) + \alpha_t\frac{1}{|\mathcal{A}|}\right)\right| \\
&\leq \sum_{x,a}\sum_{x',a'} \left|\mu_{n-1}^{\pi^t,\hat{p}^t}(x',a')\hat{p}_n^t(x|x',a') - \mu_{n-1}^{\tilde{\pi}^t,\hat{p}^{t+1}}(x',a')\hat{p}_n^{t+1}(x|x',a')\right|\pi_n^t(a|x) \\
&\qquad + \alpha_t \sum_{x,a}\sum_{x',a'} \mu_{n-1}^{\tilde{\pi}^t,\hat{p}^{t+1}}(x',a')\hat{p}_n^{t+1}(x|x',a')\left|\pi_n^t(a|x) - \frac{1}{|\mathcal{A}|}\right| \\
&\leq \sum_{x}\sum_{x',a'} \left|\mu_{n-1}^{\pi^t,\hat{p}^t}(x',a')\hat{p}_n^t(x|x',a') - \mu_{n-1}^{\pi^t,\hat{p}^t}(x',a')\hat{p}_n^{t+1}(x|x',a')\right. \\
&\qquad \left. + \mu_{n-1}^{\pi^t,\hat{p}^t}(x',a')\hat{p}_n^{t+1}(x|x',a') - \mu_{n-1}^{\tilde{\pi}^t,\hat{p}^{t+1}}(x',a')\hat{p}_n^{t+1}(x|x',a')\right| + 2\alpha_t \\
&\leq \sum_{x',a'} \mu_{n-1}^{\pi^t,\hat{p}^t}(x',a')\|\hat{p}_n^t(\cdot|x',a') - \hat{p}_n^{t+1}(\cdot|x',a')\|_1 + \sum_{x',a'}\left|\mu_{n-1}^{\pi^t,\hat{p}^t}(x',a') - \mu_{n-1}^{\tilde{\pi}^t,\hat{p}^{t+1}}(x',a')\right| + 2\alpha_t \\
&\leq \sum_{x',a'} \mu_{n-1}^{\pi^t,\hat{p}^t}(x',a')\|\hat{p}_n^t(\cdot|x',a') - \hat{p}_n^{t+1}(\cdot|x',a')\|_1 + \|\mu_{n-1}^{\pi^t,\hat{p}^t} - \mu_{n-1}^{\tilde{\pi}^t,\hat{p}^{t+1}}\|_1 + 2\alpha_t \\
&\leq \sum_{i=0}^{n-1}\sum_{x,a} \mu_i^{\pi^t,\hat{p}^t}(x,a)\|\hat{p}_{i+1}^t(\cdot|x,a) - \hat{p}_{i+1}^{t+1}(\cdot|x,a)\|_1 + 2n\alpha_t,
\end{aligned}$$

where for the last inequality we use that $\mu_0^{\pi^t,\hat{p}^t} = \mu_0^{\tilde{\pi}^t,\hat{p}^{t+1}}$. To finish we just take the sup over $n \in [N]$. □

D.5 Proof of Proposition 5.7: upper bound on R_T^{policy}

Proposition. *Consider an episodic MDP with finite state space $\mathcal{X}$, finite action space $\mathcal{A}$, episodes of length N, and probability kernel $p := (p_n)_{n\in[N]}$. Let $F^t := \sum_{n=1}^N f_n^t$ convex with f_n^t ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$ for all $n \in [N], t \in [T]$. Let b be defined as in Equation (14). Then, Greedy MD-CURL obtains, for $\tau = \frac{b}{L\sqrt{T}}$,*

$$R_T^{policy} \leq 2\ell N b\sqrt{T}.$$

Proof. Using the convexity of F^t,

$$R_T^{policy} = \sum_{t=1}^T F^t(\mu^{\pi^t,\hat{p}^t}) - F^t(\mu^{\pi^*,\hat{p}^{t+1}}) \leq \sum_{t=1}^T \langle l^t, \mu^{\pi^t,\hat{p}^t} - \mu^{\pi^*,\hat{p}^{t+1}}\rangle$$

where $l^t := \nabla F^t(\mu^{\pi^t,t})$ to shorten notation, and we also use the notation introduced in the main paper $\mu^t := \mu^{\pi^t,\hat{p}^t}$ and $\tilde{\mu}^t := \mu^{\tilde{\pi}^t,\hat{p}^t}$, for all $t \in [T]$. We begin by examining Problem (12):

$$\mu^{t+1} \in \arg\min_{\mu\in\mathcal{M}_{\mu_0}^{t+1}} \{\tau\langle l^t, \mu\rangle + \Gamma(\mu, \tilde{\mu}^t)\}.$$

Since F^t is a convex function and $\mathcal{M}_{\mu_0}^{t+1}$ is a convex set, the optimality conditions imply that for all $\nu^{t+1} \in \mathcal{M}_{\mu_0}^{t+1}$,

$$\langle \tau l^t + \nabla\psi(\mu^{t+1}) - \nabla\psi(\tilde{\mu}^t), \nu^{t+1} - \mu^{t+1}\rangle \geq 0.$$

Recall that ψ is defined in Proposition 4.3 as the function inducing the Bregaman divergence Γ. Re-arranging the terms and using the three points inequality for Bregman divergences (Bubeck, 2015) we get that,

$$\tau\langle l^t, \mu^{t+1} - \nu^{t+1}\rangle \leq \langle \nabla\psi(\mu^{t+1}) - \nabla\psi(\tilde{\mu}^t), \nu^{t+1} - \mu^{t+1}\rangle = \Gamma(\nu^{t+1}, \tilde{\mu}^t) - \Gamma(\nu^{t+1}, \mu^{t+1}) - \Gamma(\mu^{t+1}, \tilde{\mu}^t).$$

This is in particular valid for $\nu^{t+1} := \mu^{\pi^*, \hat{p}^{t+1}}$. Therefore, by adding and subtracting $\tau\langle l^t, \mu^t\rangle$ on the left-hand side,

$$\begin{aligned}
&\tau\langle l^t, \mu^{t+1} - \nu^{t+1}\rangle + \tau\langle l^t, \mu^t\rangle - \tau\langle l^t, \mu^t\rangle \leq \Gamma(\nu^{t+1}, \tilde{\mu}^t) - \Gamma(\nu^{t+1}, \mu^{t+1}) - \Gamma(\mu^{t+1}, \tilde{\mu}^t) \\
\Rightarrow\ &\tau\langle l^t, \mu^t - \nu^{t+1}\rangle \leq \tau\langle l^t, \mu^t - \mu^{t+1}\rangle + \Gamma(\nu^{t+1}, \tilde{\mu}^t) - \Gamma(\nu^{t+1}, \mu^{t+1}) - \Gamma(\mu^{t+1}, \tilde{\mu}^t).
\end{aligned}$$

Then, by summing over $t \in [T]$, and by taking $\nu^{t+1} := \mu^{\pi^*, \hat{p}^{t+1}}$, we obtain

$$R_T^{policy} \leq \underbrace{\frac{1}{\tau}\sum_{t=1}^{T}\left[\tau\langle l^t, \mu^t - \mu^{t+1}\rangle - \Gamma(\mu^{t+1}, \tilde{\mu}^t)\right]}_{A} + \underbrace{\frac{1}{\tau}\sum_{t=1}^{T}\left[\Gamma(\nu^{t+1}, \tilde{\mu}^t) - \Gamma(\nu^{t+1}, \mu^{t+1})\right]}_{B}. \tag{27}$$

The term A appears due to our lack of knowledge of F^t at the beginning of episode t for all episodes. To remedy this, we use Young's inequality and the strong convexity of Γ. Note that if we were to consider the case where all F^t are known in advance, we would not have to deal with the A term. As for the term B, in the classic OMD proof (Shalev-Shwartz, 2012) where the set of constraints is fixed the sum of the difference between the Bregman divergences telescopes (as we would with a fixed ν). However, since we are considering time-varying constraint sets, this does not happen in our case. We now proceed to find an upper bound for each term.

Step 1: upper bound on B We begin by analyzing the second term of the sum in Equation (27). Recall that $\nu^t := \mu^{\pi^*, \hat{p}^t}$ for all $t \in [T]$. In order to make the Bregman divergence terms telescope we add and subtract $\Gamma(\nu^t, \mu^t) - \Gamma(\nu^t, \tilde{\mu}^t)$, obtaining

$$\sum_{t=1}^{T}\Gamma(\nu^{t+1}, \tilde{\mu}^t) - \Gamma(\nu^{t+1}, \mu^{t+1}) = \underbrace{\sum_{t=1}^{T}\Gamma(\nu^{t+1}, \tilde{\mu}^t) - \Gamma(\nu^t, \tilde{\mu}^t)}_{(i)} + \underbrace{\sum_{t=1}^{T}\Gamma(\nu^t, \tilde{\mu}^t) - \Gamma(\nu^t, \mu^t)}_{(ii)} + \underbrace{\sum_{t=1}^{T}\Gamma(\nu^t, \mu^t) - \Gamma(\nu^{t+1}, \mu^{t+1})}_{(iii)}.$$

We now analyze each term. Using the definition of a Bregman divergence induced by ψ we get that

$$\begin{aligned}
(i) &= \sum_{t=1}^{T}\psi(\nu^{t+1}) - \psi(\tilde{\mu}^t) - \langle\nabla\psi(\tilde{\mu}^t), \nu^{t+1} - \tilde{\mu}^t\rangle - \psi(\nu^t) + \psi(\tilde{\mu}^t) + \langle\nabla\psi(\tilde{\mu}^t), \nu^t - \tilde{\mu}^t\rangle \\
&= \sum_{t=1}^{T}\psi(\nu^{t+1}) - \psi(\nu^t) + \sum_{t=1}^{T}\langle\nabla\psi(\tilde{\mu}^t), \nu^t - \nu^{t+1}\rangle \\
&\leq -\psi(\nu^1) + \sum_{t=1}^{T}\|\nabla\psi(\tilde{\mu}^t)\|_{1,\infty}\|\nu^t - \nu^{t+1}\|_{\infty,1},
\end{aligned}$$

where in the last inequality we used that the first term telescopes and we apply Holder's inequality to the second term. Recall that for $v := (v_n)_{n\in[N]}$ such that $v_n \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$, we defined $\|v\|_{\infty,1} := \sup_{n\in[N]}\|v_n\|_1$. We now also define $\|\zeta\|_{1,\infty} := \sup_v\{|\langle\zeta, v\rangle|, \|v\|_{\infty,1} \leq 1\} = \sup_{n\in[N]}\|\zeta_n\|_1$ as the respective dual norm.

With our choice of Bregman divergence, and given the definition of $\tilde{\pi}$ in Equation (13), for each $n \in [N], (x, a) \in \mathcal{X}\times\mathcal{A}$, $|\nabla\psi(\tilde{\mu}^t)(n, x, a)| = |\log(\tilde{\pi}_n^t(a|x))| \leq \log(|\mathcal{A}|/\alpha_t)$. Plugging this result with the result of Lemma 5.6 into the bound of (i) we obtain that

$$(i) \leq -\psi(\nu^1) + \sum_{t=1}^{T} N\log\left(\frac{|\mathcal{A}|}{\alpha_t}\right)\|\mu^{\pi^*, \hat{p}^t} - \mu^{\pi^*, \hat{p}^{t+1}}\|_{\infty,1} \leq -\psi(\nu^1) + 2N^2\sum_{t=1}^{T}\log\left(\frac{|\mathcal{A}|}{\alpha_t}\right)\frac{1}{t}.$$

As for the second term, using our definition of Γ, we obtain that

$$\begin{aligned}(ii) &= \sum_{t=1}^{T}\sum_{n,x,a}\mu_n^{\pi^*,\hat{p}^t}(x,a)\log\left(\frac{\pi_n^*(a|x)}{\tilde{\pi}_n^t(a|x)}\right) - \sum_{n,x,a}\mu_n^{\pi^*,\hat{p}^t}(x,a)\log\left(\frac{\pi_n^*(a|x)}{\pi_n^t(a|x)}\right)\\ &= \sum_{t=1}^{T}\sum_{n,x,a}\mu_n^{\pi^*,\hat{p}^t}(x,a)\log\left(\frac{\pi_n^t(a|x)}{\tilde{\pi}_n^t(a|x)}\right)\\ &= \sum_{t=1}^{T}\sum_{n,x,a}\mu_n^{\pi^*,\hat{p}^t}(x,a)\log\left(\frac{\pi_n^t(a|x)}{(1-\alpha_t)\pi_n^t(a|x)+\alpha/|\mathcal{A}|}\right)\\ &\leq N\sum_{t=1}^{T}(-\log(1-\alpha_t)) \leq 2N\sum_{t=1}^{T}\alpha_t,\end{aligned}$$

where the last inequality is valid if $0 \leq \alpha_t \leq 0.5$.

It is easy to see that the third term telescopes, therefore, as $-\Gamma(\nu^{T+1},\mu^{T+1}) \leq 0$ as a Bregman divergence is always positive,

$$(iii) \leq \Gamma(\nu^1,\mu^1).$$

Before adding back the three terms, note that, for μ^1 initialized such that $\nabla\psi(\mu^1)=0$, we have $\Gamma(\nu^1,\mu^1)-\psi(\nu^1) = -\psi(\mu^1)$. Furthermore, from Lemma D.2, $-\psi(\mu^1) \leq N\log(|\mathcal{A}|)$. Therefore,

$$\Gamma(\nu^1,\mu^1) - \psi(\nu^1) \leq N\log(|\mathcal{A}|). \tag{28}$$

Summing over our bounds and using the Inequality (28), we get that B is upper bounded as

$$\frac{1}{\tau}\sum_{t=1}^{T}\left[\Gamma(\nu^{t+1},\tilde{\mu}^t)-\Gamma(\nu^{t+1},\mu^{t+1})\right] \leq \frac{1}{\tau}\left[(i)+(ii)+(iii)\right] \leq \frac{N}{\tau}\log(|\mathcal{A}|)+\frac{2N^2}{\tau}\sum_{t=1}^{T}\log\left(\frac{|\mathcal{A}|}{\alpha_t}\right)\frac{1}{t}+\frac{2N}{\tau}\sum_{t=1}^{T}\alpha_t. \tag{29}$$

Step 2: Upper bound on A It remains to delimit the first term of the bound in R_T^{policy} in Equation (27) given by

$$A = \frac{1}{\tau}\left[\sum_{t=1}^{T}\tau\langle l^t,\mu^t-\mu^{t+1}\rangle - \Gamma(\mu^{t+1},\tilde{\mu}^t)\right], \tag{30}$$

representing what we pay for not knowing the loss function in advance. For that we use Young's inequality (Beck and Teboulle, 2003).

Recall that Young's inequality states that for any $\sigma > 0$, for any dual norms,

$$\langle a,b\rangle \leq \frac{1}{2\sigma}\|a\|^2 + \frac{\sigma}{2}\|b\|_*^2.$$

Therefore, for any $\sigma > 0$ to be optimized later, and for each episode $t \in [T]$,

$$\tau\langle l^t,\mu^t-\mu^{t+1}\rangle - \Gamma(\mu^{t+1},\tilde{\mu}^t) \leq \frac{\tau^2\|l^t\|_{1,\infty}^2}{2\sigma} + \frac{\sigma}{2}\|\mu^t-\mu^{t+1}\|_{\infty,1}^2 - \Gamma(\mu^{t+1},\tilde{\mu}^t), \tag{31}$$

where recall that for $v := (v_n)_{n\in[N]}$ such that $v_n \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$, we defined $\|v\|_{\infty,1} := \sup_{n\in[N]}\|v_n\|_1$, and we let $\|\zeta\|_{1,\infty} := \sup_v\{|\langle\zeta,v\rangle|, \|v\|_{\infty,1}\leq 1\} = \sup_{n\in[N]}\|\zeta_n\|_1$ as the respective dual norm.

From Lemma B.1 and inequality (23) stating the strong convexity of ψ, we have that for all $t \in [T]$

$$\Gamma(\mu^{t+1},\tilde{\mu}^t) = D(\mu_{1:N}^{t+1},\mu_{1:N}^{\tilde{\pi}^t,\hat{p}^{t+1}}) \geq \frac{1}{2}\|\mu^{t+1}-\mu^{\tilde{\pi}^t,\hat{p}^{t+1}}\|_{\infty,1}^2 \tag{32}$$

where recall that $\mu_{1:N}$ is the joint state-action distribution while that $\mu := (\mu_n)_{n\in[N]}$ is the sequence of state-action distributions.

Using that for any vectors $a, b, c \in \mathbb{R}^d$, and that for any norm $\|\cdot\|$, $\|a-b\|^2 \leq 2\big(\|a-c\|^2 + \|b-c\|^2\big)$, we then have by Equation (32)

$$\begin{aligned}\frac{1}{4}\|\mu^t - \mu^{t+1}\|_{\infty,1}^2 - \Gamma(\mu^{t+1}, \tilde{\mu}^t) &\leq \frac{1}{4}\|\mu^t - \mu^{t+1}\|_{\infty,1}^2 - \frac{1}{2}\|\mu^{\tilde{\pi}^t, \hat{p}^{t+1}} - \mu^{t+1}\|_{\infty,1}^2 \\ &\leq \frac{1}{2}\big(\|\mu^t - \mu^{\tilde{\pi}^t, \hat{p}^{t+1}}\|_{\infty,1}^2 + \|\mu^{\tilde{\pi}^t, \hat{p}^{t+1}} - \mu^{t+1}\|_{\infty,1}^2\big) - \frac{1}{2}\|\mu^{\tilde{\pi}^t, \hat{p}^{t+1}} - \mu^{t+1}\|_{\infty,1}^2 \\ &= \frac{1}{2}\|\mu^t - \mu^{\tilde{\pi}^t, \hat{p}^{t+1}}\|_{\infty,1}^2. \end{aligned} \tag{33}$$

To bound $\|\mu^t - \mu^{\tilde{\pi}^t, \hat{p}^{t+1}}\|_{\infty,1}^2$ we first use Lemma D.3 which gives

$$\|\mu^t - \mu^{\tilde{\pi}^t, \hat{p}^{t+1}}\|_{\infty,1} \leq \sum_{n=1}^{N} \sum_{x,a} \mu_i^t(x,a) \|\hat{p}_{i+1}^t(\cdot|x,a) - \hat{p}_{i+1}^{t+1}(\cdot|x,a)\|_1 + 2N\alpha_t.$$

Then, by Equation (26), $\|\hat{p}_{i+1}^t(\cdot|x,a) - \hat{p}_{i+1}^{t+1}(\cdot|x,a)\|_1 \leq 2/t$ for all $t \in [T]$, therefore

$$\|\mu^t - \mu^{\tilde{\pi}^t, \hat{p}^{t+1}}\|_{\infty,1}^2 \leq \left(\frac{2N}{t} + 2N\alpha_t\right)^2.$$

Therefore, plugging into Equation (31) with $\sigma = 1/2$ yields,

$$\tau\langle l^t, \mu^t - \mu^{t+1}\rangle - \Gamma(\mu^{t+1}, \tilde{\mu}^t) \leq \tau^2 \|l^t\|_{1,\infty}^2 + \frac{1}{2}\left(\frac{2N}{t} + 2N\alpha_t\right)^2.$$

Summing over $t \in [T]$, and $\|l^t\|_{1,\infty} \leq L := lN$ as showed in Lemma B.2 then entails:

$$A \leq \tau L^2 T + \frac{1}{2\tau}\sum_{t=1}^{T}\left(\frac{2N}{t} + 2N\alpha_t\right)^2. \tag{34}$$

Conclusion Finally, by replacing the final bounds of Equations (29) and (34), we obtain

$$R_T^{policy} \leq A + B \leq \tau T L^2 + \frac{2N^2}{\tau}\sum_{t=1}^{T}\left(\frac{1}{t} + \alpha_t\right)^2 + \frac{N}{\tau}\log(|\mathcal{A}|) + \frac{2N^2}{\tau}\sum_{t=1}^{T}\log\left(\frac{|\mathcal{A}|}{\alpha_t}\right)\frac{1}{t} + \frac{2N}{\tau}\sum_{t=1}^{T}\alpha_t.$$

Let

$$b := \left(2N^2\left[\sum_{t=1}^{T}\left(\frac{1}{t} + \alpha_t\right)^2 + \sum_{t=1}^{T}\log\left(\frac{|\mathcal{A}|}{\alpha_t}\right)\frac{1}{t}\right] + 2N\sum_{t=1}^{T}\alpha_t + N\log(|\mathcal{A}|)\right)^{\frac{1}{2}}.$$

Optimising over $\tau = \frac{b}{L\sqrt{T}}$,

$$R_T^{policy} \leq 2Lb\sqrt{T} = 2lNb\sqrt{T},$$

concluding the proof.

In particular, if $\alpha_t = \frac{1}{T}$ for all $t \in [T]$, we have $R_T^{policy} \leq \sqrt{T}\log(T)$.

□

E Bounds with unknown g

Now suppose that g_n and h_n in the model of Equation (3) are unknown. In this case, we have no information about the probability kernel, and the exploration/exploitation dilemma arises.

In order to learn the complete probability kernel, we need to modify the learning model slightly. Let us suppose that, at each episode t the learner maintains the number of visit counts to each episode (x,a) at time step n,

denoted $N_n^t(x,a)$, and the number of times this event is followed by a transition to a state x', denoted $M_n^t(x'|x,a)$, that is

$$M_n^t(x'|x,a) = \sum_{s=1}^{t} \mathbb{1}_{\{x_{n+1}^s = x', x_n^s = x, a_n^s = a\}}$$
$$N_n^t(x,a) = \sum_{s=1}^{t} \mathbb{1}_{\{x_n^s = x, a_n^s = a\}}.$$

To ease notations, we take $M = 1$ in this section. We define $\hat{p}^t$, at each (x,a) and time step $n+1$ by

$$\hat{p}_{n+1}^t(x'|x,a) = \frac{M_n^t(x'|x,a)}{\max\{1, N_n^t(x,a)\}}. \tag{35}$$

The following lemma ensures the true probability kernel p lies at a certain distance from this estimation of $\hat{p}^t$ with high probability.

Lemma E.1 (Jaksch et al. (2008); Neu et al. (2012))**.** *For any $0 < \delta < 1$,*

$$\|p_n(\cdot|x,a) - \hat{p}_n^t(\cdot|x,a)\|_1 \leq \sqrt{\frac{4|\mathcal{X}|\log\left(\frac{|\mathcal{X}||\mathcal{A}|NT}{\delta}\right)}{\max\{1, N_n^t(x,a)\}}}$$

holds, with a probability of at least $1-\delta$, for simultaneously all $(x,a) \in \mathcal{X} \times \mathcal{A}$, all $n \in [N]$, and all episodes $t \in [T]$.

Recall that the regret R_T is decomposed as $R_T := R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) + R_T^{policy} + R_T^{MDP}(\pi^*)$, and we treat each term separately. The regret bound for R_T^{policy} follows the same procedure as in Proposition 5.7. However, the bound on the terms of R_T^{MDP} are different, as we must now ensure that we visit all necessary state-action pairs (x,a) sufficiently often. This also means that the bound for $R_T^{MDP}\big((\pi^t)_{t\in[T]}\big)$ is different from the bound of $R_T^{MDP}(\pi^*)$. For bounding both terms related to R_T^{MDP} we use a similar approach as in UC-O-REPS (Rosenberg and Mansour, 2019).

Lemma E.2. *For $0 < \delta < 1$,*

$$\sup_{n\in[N]} \sum_{t=1}^{T}\sum_{i=0}^{n-1}\sum_{x,a} \mu_i^{\pi^t,p}(x,a)\|p_{i+1}(\cdot|x,a) - \hat{p}_{i+1}^t(\cdot|x,a)\|_1$$
$$\leq (\sqrt{2}+1)N|\mathcal{X}|\sqrt{4|\mathcal{A}|T\log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)} + 2N|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}$$

with probability $1-2\delta$.

Proof. Using Lemma 19 from Jaksch et al. (2008), we have that

$$\sum_{t=1}^{T} \frac{\mathbb{1}_{\{x_n^t = x, a_n^t = a\}}}{N_n^t(x,a)} \leq (\sqrt{2}+1)\sqrt{N_n^T(x,a)},$$

and by Jensen's inequality,

$$\sum_{x,a}\sum_{t=1}^{T} \frac{\mathbb{1}_{\{x_n^t = x, a_n^t = a\}}}{N_n^t(x,a)} \leq (\sqrt{2}+1)\sum_{x,a}\sqrt{|\mathcal{X}||\mathcal{A}|T}. \tag{36}$$

Let $(x_n^t, a_n^t)_{n\in[N]}$ be the trajectory made by policy π^t for all $t \in [T]$. Therefore,

$$\sum_{i=0}^{n-1}\sum_{x,a} \mu_i^{\pi^t,p}(x,a)\|p_{i+1}(\cdot|x,a) - \hat{p}_{i+1}^t(\cdot|x,a)\|_1 \tag{37}$$
$$\leq \sum_{i=0}^{n-1} \|p_{i+1}(\cdot|x_i^t,a_i^t) - \hat{p}_{i+1}^t(\cdot|x_i^t,a_i^t)\|_1 + \sum_{i=0}^{n-1}\sum_{x,a}\left(\mu_i^{\pi^t,p}(x,a) - \mathbb{1}_{\{x_i^t=a, a_i^t=a\}}\right)\|p_{i+1}(\cdot|x,a) - \hat{p}_{i+1}^t(\cdot|x,a)\|_1$$

By Lemma E.1, with probability at least $1-\delta$, simultaneously for all $i \in [N]$ we have

$$\begin{aligned}\sum_{t=1}^{T} \|p_{i+1}(\cdot|x_i^t, a_i^t) - \hat{p}_{i+1}^t(\cdot|x_i^t, a_i^t)\|_1 &\leq \sum_{t=1}^{T} \sqrt{\frac{4|\mathcal{X}|\log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)}{\max\{1, N_i^t(x_i^t, a_i^t)\}}} \\ &\leq \sum_{x,a}\sum_{t=1}^{T} \mathbb{1}_{\{x_i^t=x, a_i^t=a\}} \sqrt{\frac{4|\mathcal{X}|\log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)}{\max\{1, N_i^t(x,a)\}}} \\ &\leq (\sqrt{2}+1)\sqrt{4|\mathcal{X}|^2|\mathcal{A}|T\log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)}, \end{aligned} \tag{38}$$

where, for the last inequality, we use the result of Equation (36).

As for the second term, note that for all $i \in [N]$ and $x \in \mathcal{X}$,

$$\left(\sum_a \left(\mu_i^{\pi^t, p}(x,a) - \mathbb{1}_{\{x_i^t=x, a_i^t=a\}}\right)\right)$$

forms a martingale difference with respect to the trajectory $(x_0^s, a_0^s, \ldots, x_N^s, a_N^s)_{s\in[T]}$ (the expectation of the term conditional on the past trajectory is zero). Therefore, by Azuma-Hoeffding inequality,

$$\mathbb{P}\left[\sum_{t=1}^{T}\sum_a \left(\mu_i^{\pi^t, p}(x,a) - \mathbb{1}_{\{x_i^t=x, a_i^t=a\}}\right) \geq \varepsilon\right] \leq \exp\left(\frac{-2\varepsilon^2}{4T}\right).$$

Taking the union bound over $i \in [N]$, we get that with probability $1-\delta$, simultaneously for all $i \in [N]$, and considering that $\|p_{i+1}(\cdot|x,a) - p_{i+1}^t(\cdot|x,a)\|_1 \leq 2$,

$$\sum_{t=1}^{T}\sum_{x,a}\left(\mu_i^{\pi^t, p}(x,a) - \mathbb{1}_{x_i^t=x, a_i^t=a}\right)\|p_{i+1}(\cdot|x,a) - \hat{p}_{i+1}^t(\cdot|x,a)\|_1 \leq 2|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}. \tag{39}$$

Plugging the bounds on Equation (38) and (39) into Equation (37), we get that with probability $1-2\delta$,

$$\begin{aligned}&\sup_{n\in[N]} \sum_{t=1}^{T}\sum_{i=0}^{n-1}\sum_{x,a} \mu_i^{\pi^t, p}\|p_{i+1}(\cdot|x,a) - \hat{p}_{i+1}^t(\cdot|x,a)\|_1 \\ &\leq \sup_{n\in[N]} \sum_{i=0}^{n-1}\left[(\sqrt{2}+1)\sqrt{4|\mathcal{X}|^2|\mathcal{A}|T\log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)} + 2|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}\right] \\ &\leq (\sqrt{2}+1)N|\mathcal{X}|\sqrt{4|\mathcal{A}|T\log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)} + 2N|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}.\end{aligned}$$

□

The result of Lemma E.2 allows us to state the following proposition bounding the term $R_T^{MDP}\big((\pi^t)_{t\in[T]}\big)$:

Proposition E.3. *We consider an episodic MDP with finite state space $\mathcal{X}$, finite action space $\mathcal{A}$, episodes of length N, and probability kernel $p := (p_n)_{n\in[N]}$. We let $F^t := \sum_{n=1}^{N} f_n^t$ convex with f_n^t ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$ for all $n \in [N], t \in [T]$. We consider the probability estimation per iteration as in Equation (35). Then, with probability $1-\delta$, Greedy MD-CURL obtains,*

$$R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) \leq (\sqrt{2}+1)\ell N^2|\mathcal{X}|\sqrt{4|\mathcal{A}|T\log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)} + 2\ell N^2|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}.$$

Proof. Recall that, given the convexity of F^t and by applying Holder's inequality using that if f_n^t are ℓ-Lipschitz with respect to $\|\cdot\|_1$ then F is L-Lipschitz with respect to $\|\cdot\|_{\infty,1}$ for $L = \ell N$ (see Lemma B.2), we obtain that

$$\begin{aligned} R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) &\leq \sum_{t=1}^{T} \langle \nabla F^t(\mu^{\pi^t,p}), \mu^{\pi^t,p} - \mu^{\pi^t,\hat{p}^t}\rangle \\ &\leq \sum_{t=1}^{T} \|\nabla F^t(\mu^{\pi^t,p})\|_* \sup_{n\in[N]} \|\mu_n^{\pi^t,p} - \mu_n^{\pi^t,\hat{p}^t}\|_1 \\ &\leq L \sup_{n\in[N]} \sum_{t=1}^{T} \|\mu_n^{\pi^t,p} - \mu_n^{\pi^t,\hat{p}^t}\|_1. \end{aligned}$$

The result then follows from the application of Lemma D.1 and Lemma E.2. □

To complete the bound on the regret R_T, we need to bound $R_T^{MDP}(\pi^*)$. For this, we need Lemma E.4, which states that the Greedy MD-CURL algorithm always computes policies that are lower bounded if $-\nabla f_n^t(x,a)(\mu_n) \in [0,1]$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, all $\mu_n \in \Delta_{\mathcal{X}\times\mathcal{A}}$, $n \in [N]$ and $t \in [T]$. Proposition E.5 states the bound for $R_T^{MDP}(\pi^*)$.

Lemma E.4. *Let $(\pi^t)_{t\in[T]}$ be the sequence of policies obtained after computing T episodes of Greedy MD-CURL with $\alpha_t \in (0,1/2)$ and objective functions $F^t = \sum_{n=1}^N f_n^t$ such that $-\nabla f_n^t(\mu_n)(x,a) \in [0,1]$. Consequently, there is $\xi \in (0,1)$ such that for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, for all $n \in [N]$, and for all episodes $t \in [T]$, $\pi_n^t(a|x) \geq \xi$.*

Proof. At each episode t, we compute $\pi^{t+1} := \text{MD-CURL}(1, \tilde{\pi}^t \backslash \pi^t, F^t, \hat{p}^{t+1}, \mu_0, \tau)$, where $\tilde{\pi}^t = (1-\alpha_t)\pi^t + \alpha_t \frac{1}{|\mathcal{A}|}$, and the other parameters are defined in Algorithm 3.

From its definition, we can see that $\tilde{\pi}_n^t(a|x) \geq \frac{\alpha_t}{|\mathcal{A}|}$. The closed form solution of one iteration of MD-CURL with the given parameters gives

$$\pi_n^{t+1}(a|x) = \frac{\tilde{\pi}_n^t(a|x)\exp\left(\tau\tilde{Q}_n^t(x,a)\right)}{\sum_{a'\in\mathcal{A}} \tilde{\pi}_n^t(a'|x)\exp\left(\tau\tilde{Q}_n^t(x,a')\right)},$$

where $\tilde{Q}_n^t(x,a)$ is defined in Equation (10) with $-\nabla f_n^t(x,a)(\mu_n^t)$ and $\tilde{\pi}^t$ at the place of π^k. As $-\nabla f_n^t(x,a)(\mu_n^t) \in [0,1]$, we have $1 \leq \exp\left(\tau\tilde{Q}_n^t(x,a')\right) \leq \exp(\tau(N-n))$. Therefore, we have $\pi_n^{t+1}(a|x) \geq \frac{\alpha_t}{|\mathcal{A}|\exp(\tau(N-n))}$ for all steps $n \in [N]$ and (x,a).

Taking $\xi := \min_{t\in[T],n\in[N]} \frac{\alpha_t}{|\mathcal{A}|\exp(\tau(N-n))} = \frac{1}{|\mathcal{A}|\exp(\tau N)} \min_{t\in[T]} \alpha_t$, we then have for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, for all $n \in [N]$, and for all episodes $t \in [T]$, $\pi_n^t(a|x) \geq \xi$. □

Proposition E.5. *We consider an episodic MDP with finite state space $\mathcal{X}$, finite action space $\mathcal{A}$, episodes of length N, and probability kernel $p := (p_n)_{n\in[N]}$. We let $F^t := \sum_{n=1}^N f_n^t$ convex with f_n^t ℓ-Lipschitz with respect to the norm $\|\cdot\|_1$ for all $n \in [N], t \in [T]$. We let ξ be the lower bound of $\pi_n^t(a|x)$ for all $n, t, (x,a)$ defined as in Lemma E.4. We consider the probability estimation per iteration as in Equation (35). Then, with probability $1-2\delta$, Greedy MD-CURL obtains,*

$$R_T^{MDP}(\pi^*) \leq \frac{1}{\xi^N}\left[(\sqrt{2}+1)\ell N^2|\mathcal{X}|\sqrt{4|\mathcal{A}|T\log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)} + 2\ell N^2|\mathcal{X}|\sqrt{2T\log\left(\frac{N}{\delta}\right)}\right].$$

Proof. Recall that, given the convexity of F^t and by applying Holder's inequality using that if f_n^t are ℓ-Lipschitz with respect to $\|\cdot\|_1$ then F is L-Lipschitz with respect to $\|\cdot\|_{\infty,1}$ for $L = \ell N$ (see Lemma B.2), we obtain that

$$\begin{aligned} R_T^{MDP}(\pi^*) &\leq \sum_{t=1}^{T} \langle \nabla F^t(\mu^{\pi^*,\hat{p}^t}), \mu^{\pi^*,\hat{p}^t} - \mu^{\pi^*,p}\rangle \\ &\leq \sum_{t=1}^{T} \|\nabla F^t(\mu^{\pi^t,p})\|_* \sup_{n\in[N]} \|\mu_n^{\pi^*,\hat{p}^t} - \mu_n^{\pi^*,p}\|_1 \\ &\leq L \sup_{n\in[N]} \sum_{t=1}^{T} \|\mu_n^{\pi^*,\hat{p}^t} - \mu_n^{\pi^*,p}\|_1. \end{aligned}$$

As $\pi_n^t(a|x) \geq \xi$ for all $n \in [N], t \in [T]$ and $(x,a) \in \mathcal{X} \times \mathcal{A}$, we get $\frac{\mu_n^{\pi^*,p}(x,a)}{\mu_n^{\pi^t,p}(x,a)} \leq \frac{1}{\xi^n}$. This can be demonstrated recursively: suppose it's true for n, then for $n+1$, by definition

$$\begin{aligned}
\mu_{n+1}^{\pi^t,p}(x,a) &= \sum_{x',a'} \mu_n^{\pi^t,p}(x',a') p_{n+1}(x|x',a') \pi_{n+1}^t(a|x) \\
&\leq \sum_{x',a'} \xi^n \mu_n^{\pi^*,p}(x',a') p_{n+1}(x|x',a') \pi_{n+1}^*(a|x) \xi \\
&= \mu_{n+1}^{\pi^*,p}(x,a) \xi^{n+1}.
\end{aligned}$$

Using Lemma D.1, and Proposition E.3, we get

$$\begin{aligned}
\sup_{n\in[N]} \sum_{t=1}^T \|\mu_n^{\pi^*,\hat{p}^t} - \mu_n^{\pi^*,p}\|_1 &\leq \sup_{n\in[N]} \sum_{t=1}^T \sum_{i=0}^{n-1} \sum_{x,a} \mu_i^{\pi^*,p}(x,a) \|p_{i+1}(\cdot|x,a) - p_{i+1}^t(\cdot|x,a)\|_1 \\
&\leq \sup_{n\in[N]} \sum_{t=1}^T \sum_{i=0}^{n-1} \frac{1}{\xi^i} \sum_{x,a} \mu_i^{\pi^t,p}(x,a) \|p_{i+1}(\cdot|x,a) - p_{i+1}^t(\cdot|x,a)\|_1 \\
&\leq \sup_{n\in[N]} \sum_{i=0}^{n-1} \frac{1}{\xi^i} \left[(\sqrt{2}+1) \sqrt{4|\mathcal{X}|^2|\mathcal{A}|T \log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)} + 2|\mathcal{X}| \sqrt{2T \log\left(\frac{N}{\delta}\right)} \right] \\
&\leq \frac{N}{\xi^N} \left[(\sqrt{2}+1) \sqrt{4|\mathcal{X}|^2|\mathcal{A}|T \log\left(\frac{T|\mathcal{X}||\mathcal{A}|N}{\delta}\right)} + 2|\mathcal{X}| \sqrt{2T \log\left(\frac{N}{\delta}\right)} \right]
\end{aligned}$$

where the third inequality is obtained by following the same steps of the proof from Lemma E.2. □

Conclusion: bounding R_T We join the propositions E.3 and E.5 bounding $R_T^{MDP}\big((\pi^t)_{t\in[T]}\big)$ and $R_T^{MDP}\big(\pi^*\big)$ when playing Greedy MD-CURL with g_n and h_n unknown, and Proposition 5.7 bounding R_T^{policy} which remains general regardless of prior knowledge of g_n and h_n.

Here, we show regret in terms of the number of episodes T and do not worry about other constant terms. We use $\lesssim$ to denote an inequality up to constant or logarithmic terms independent of T. To simplify, we take $\alpha_t = \alpha$ for all $t \in [T]$. Therefore, α and τ are the parameters to be optimized. We hypothesize that $\alpha < 1$, $T\alpha \geq \log\left(\frac{1}{\alpha}\right)\log(T)$, and $\tau \leq \frac{1}{N}$. We will later verify when the optimized α and τ satisfy these conditions.

From Proposition E.3, we have

$$R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) \lesssim \sqrt{T\log(T)}.$$

From Proposition 5.7,

$$R_T^{policy} \lesssim \tau T + \frac{T}{\tau}\alpha^2 + \frac{1}{\tau}\log\left(\frac{1}{\alpha}\right)\log(T) + \frac{T}{\tau}\alpha \lesssim \tau T + \frac{T}{\tau}\alpha.$$

From Proposition E.5, we have

$$R_T^{MDP}\big(\pi^*\big) \lesssim \xi^{-N}\sqrt{T\log(T)} \lesssim \alpha^{-N}\sqrt{T\log(T)},$$

where $\xi^{-1} = \left(\frac{\alpha}{|\mathcal{A}|\exp(\tau N)}\right) \lesssim \alpha^{-1}$.

Therefore,

$$R_T = R_T^{MDP}\big((\pi^t)_{t\in[T]}\big) + R_T^{policy} + R_T^{MDP}\big(\pi^*\big) \lesssim \tau T + \frac{T}{\tau}\alpha + \alpha^{-N}\sqrt{T\log(T)}.$$

We first optimize over τ. For $\tau = \alpha^{\frac{1}{2}}$, then

$$R_T \lesssim \alpha^{\frac{1}{2}}T + \alpha^{-N}\sqrt{T\log(T)}.$$

Then, we optimize over α. For $\alpha = T^{-\frac{1}{2N+1}}$,

$$R_T \lesssim T^{\frac{4N+1}{4N+2}}.$$

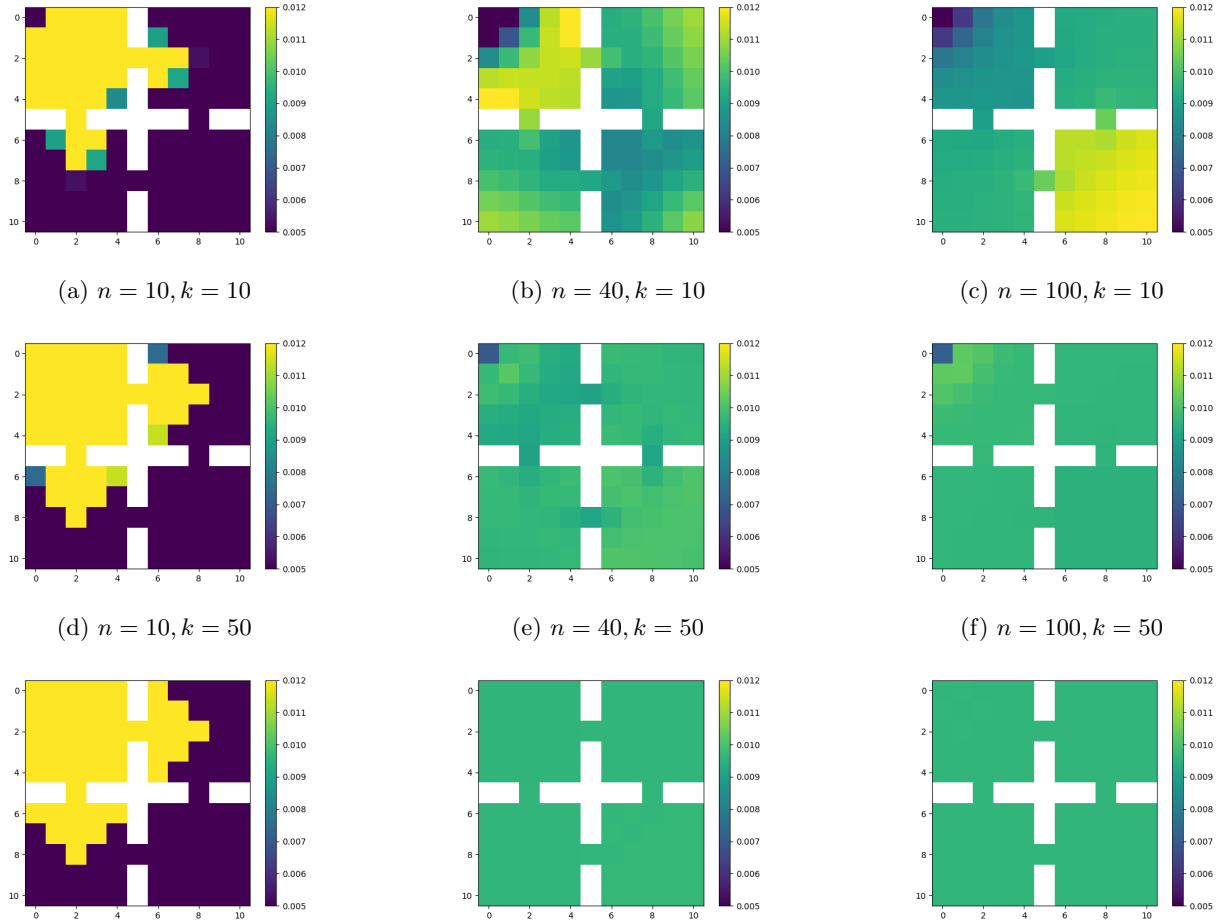
If $T > 1$, then the conditions $\alpha < 1$ and $T\alpha \geq \log\left(\frac{1}{\alpha}\right)\log(T)$ are satisfied. For $T \geq N^{4N+2}$, then $\tau \leq 1/N$.

In a classic non-episodic online learning scenario, or in an episodic MDP with stationary probability kernels, we would not incur the term on ξ^N but only ξ. This would reduce the final regret bound to $O\left(T^{\frac{5}{6}}\right)$ for any $T \geq 1$. That is for example the case of the showcase experiments in Section 6 and Appendix F.

F Additional experiments

F.1 MD-CURL known probability kernel

We present the state distribution induced by the policies computed with MD-CURL in the offline optimization scenario when both g_n and h_n are known for varying steps n and iterations k. The episode length is fixed to $N = 100$ for all experiments. We illustrate the Entropy Maximization problem in Figure 6 and the Multi-Objective problem in Figure 7, and show that MD-CURL achieves the main goal in both cases.

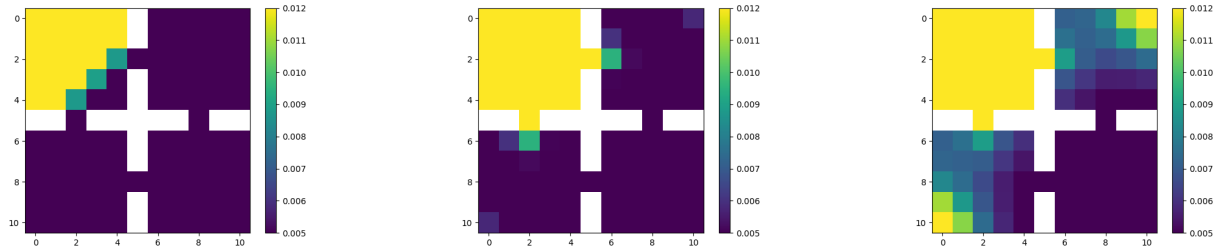


(a) $n = 10, k = 10$ (b) $n = 40, k = 10$ (c) $n = 100, k = 10$

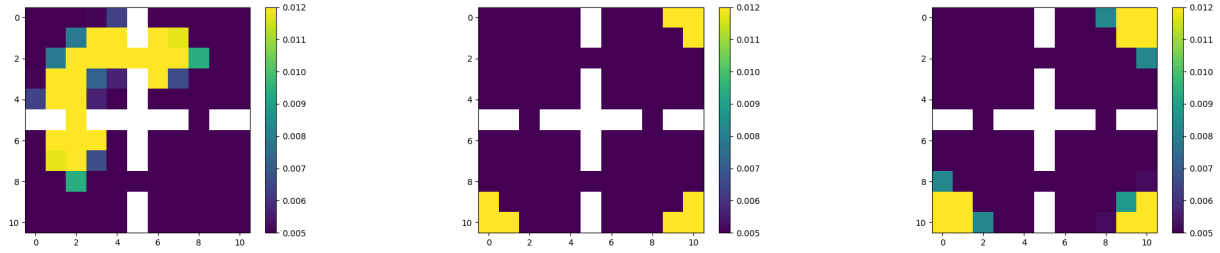
(d) $n = 10, k = 50$ (e) $n = 40, k = 50$ (f) $n = 100, k = 50$

(g) $n = 10, k = 500$ (h) $n = 40, k = 500$ (i) $n = 100, k = 500$

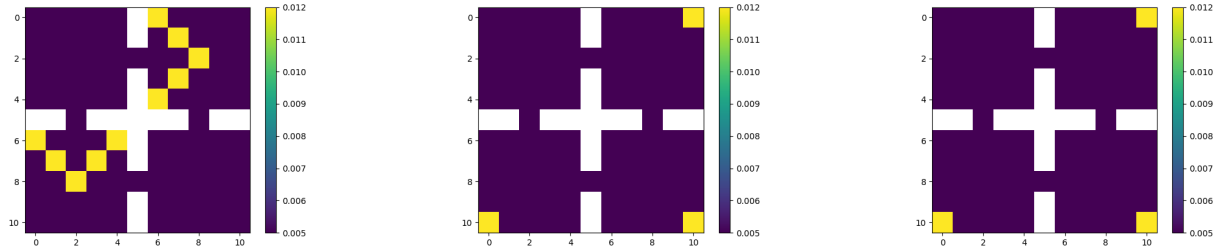
Figure 6: State distribution of MD-CURL applied to Entropy Maximisation for steps $n \in \{10, 40, 100\}$ and iterations $k \in \{10, 50, 500\}$.



(a) $n = 10, k = 10$ (b) $n = 40, k = 10$ (c) $n = 100, k = 10$



(d) $n = 10, k = 50$ (e) $n = 40, k = 50$ (f) $n = 100, k = 50$



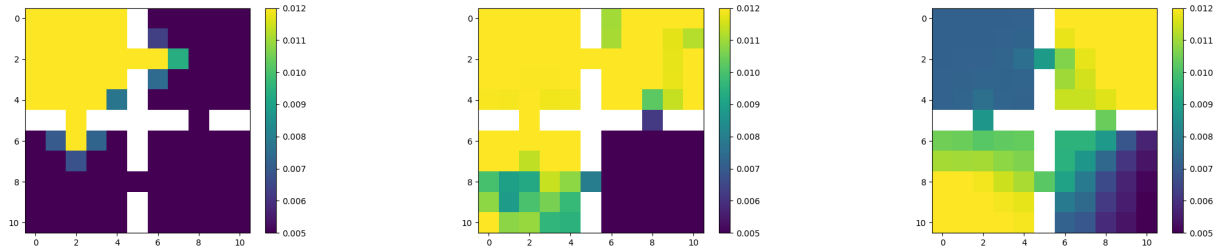
(g) $n = 10, k = 500$ (h) $n = 40, k = 500$ (i) $n = 100, k = 500$

Figure 7: State distribution of MD-CURL applied to Multi-Objectives for steps $n \in \{10, 40, 100\}$ and iterations $k \in \{10, 50, 500\}$.

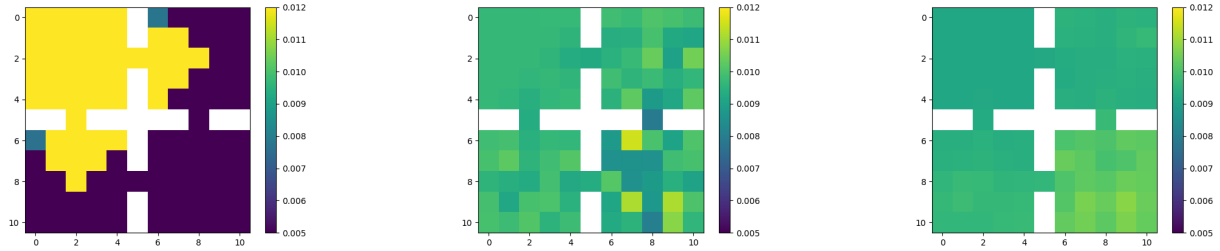
F.2 Greedy MD-CURL with completely unknown probability kernel

In this section, we present the state distribution induced by the policies computed with Greedy MD-CURL in the online learning scenario. We assume that both g_n and h_n are unknown, and we estimate the probability kernel $\hat{p}^t$ using Equation (35) at each episode. We vary the steps n and episodes t, and fix the episode length to $N = 100$ for all experiments.

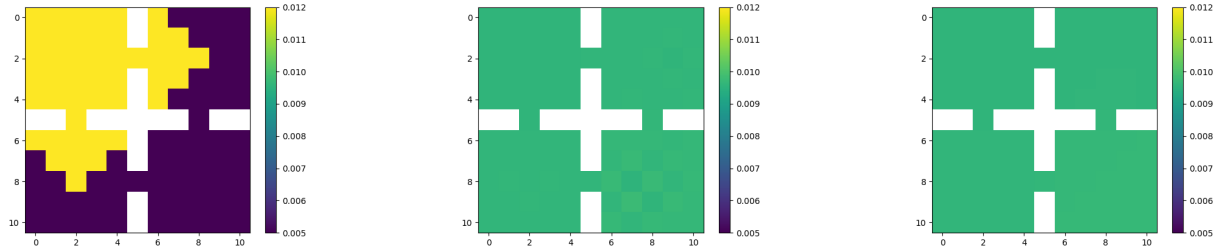
We illustrate the Entropy Maximization problem in Figure 8 and the Multi-Objective problem in Figure 9 with a central noise of a probability of 0.2. These results show that even when the full dynamics are unknown, Greedy MD-CURL can still achieve the main goal.



(a) $n = 10, t = 10$ (b) $n = 40, t = 10$ (c) $n = 100, t = 10$

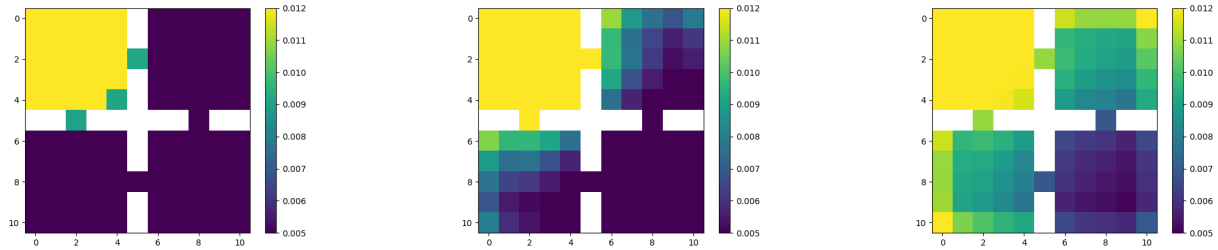


(d) $n = 10, t = 50$ (e) $n = 40, t = 50$ (f) $n = 100, t = 50$

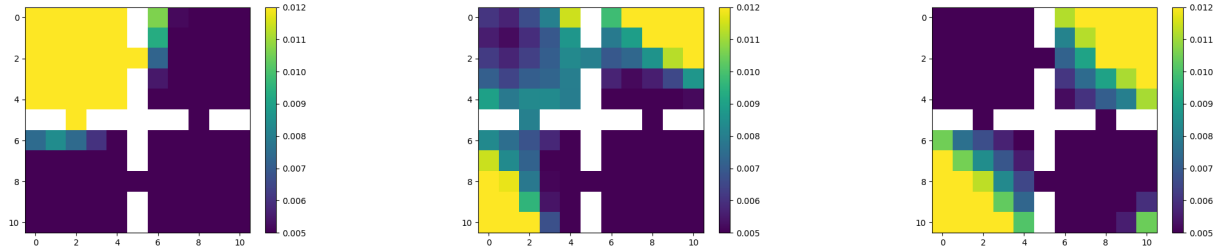


(g) $n = 10, t = 500$ (h) $n = 40, t = 500$ (i) $n = 100, t = 500$

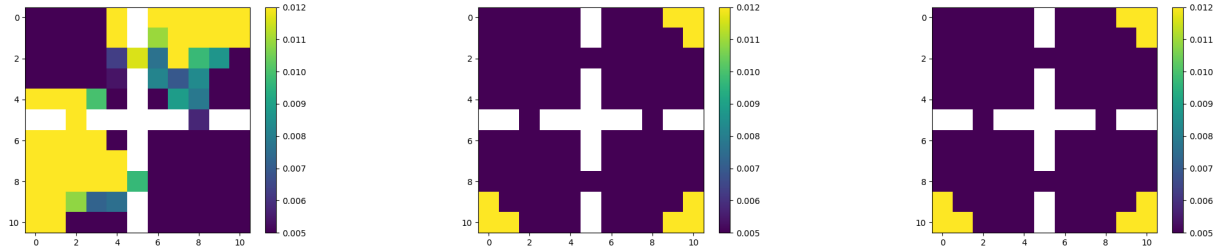
Figure 8: State distribution of Greedy MD-CURL applied to Entropy Maximisation for steps $n \in \{10, 40, 100\}$ and episodes $t \in \{10, 50, 500\}$.



(a) $n = 10, t = 10$ (b) $n = 40, t = 10$ (c) $n = 100, t = 10$



(d) $n = 10, t = 50$ (e) $n = 40, t = 50$ (f) $n = 100, t = 50$



(g) $n = 10, t = 500$ (h) $n = 40, t = 500$ (i) $n = 100, t = 500$

Figure 9: State distribution of Greedy MD-CURL applied to Multi-Objectives for steps $n \in \{10, 40, 100\}$ and episodes $t \in \{10, 50, 500\}$.