Matryoshka Policy Gradient for Entropy-Regularized RL: Convergence and Global Optimality

François G. Ged[1,2] FGED.MATH@GMAIL.COM
[1] *Chair of Statistical Field Theory*
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
[2] *Dynamical Systems in Biomathematics*
University of Vienna
Vienna, Austria

Maria Han Veiga HANVEIGA.1@OSU.EDU
Department of Mathematics
The Ohio State University
Columbus, USA

Editor: Martha White

Abstract

A novel Policy Gradient (PG) algorithm, called *Matryoshka Policy Gradient* (MPG), is introduced and studied, in the context of fixed-horizon max-entropy reinforcement learning, where an agent aims at maximizing entropy bonuses additional to its cumulative rewards. In the linear function approximation setting with softmax policies, we prove uniqueness and characterize the optimal policy of the entropy regularized objective, together with global convergence of MPG. These results are proved in the case of continuous state and action space. MPG is intuitive, theoretically sound and we furthermore show that the optimal policy of the infinite horizon max-entropy objective can be approximated arbitrarily well by the optimal policy of the MPG framework. Finally, we provide a criterion for global optimality when the policy is parametrized by a neural network in terms of the neural tangent kernel at convergence. As a proof of concept, we evaluate numerically MPG on standard test benchmarks.

Keywords: reinforcement learning, policy gradient, entropy regularization, global convergence, neural networks

1. Introduction

The family of Policy Gradient algorithms (PG) in Reinforcement Learning (RL) originated several decades ago with the algorithm REINFORCE (Williams, 1992), the name *Policy Gradient* appearing only in 2000 in Sutton et al. (1999), they recently regained interest thanks to many remarkable achievements, to name a few: in continuous control (Lillicrap et al., 2015; Schulman et al., 2015, 2017b) and natural language processing such as GPT-3 (Brown et al., 2020)[1], and more generally in the fine-tuning from human feedback stage of

[1] instructGPT and chatGPT are trained with Proximal Policy Optimization, see https://openai.com/blog/chatgpt/.

large language models (Ziegler et al., 2019). See the blog post of Weng (2018) that lists important PG methods and provides a concise introduction to each of them.

PG methods are considered more suitable for large (possibly continuous) state and action spaces than other nonetheless important methods such as Q-learning and its variations. However, for large spaces, the exploitation-exploration dilemma becomes more challenging. In order to enhance exploration, it has become standard to use a regularization to the objective, as in *max-entropy RL* (Nachum et al., 2017; O'Donoghue et al., 2016; Schulman et al., 2017a; Mnih et al., 2016; Haarnoja et al., 2018a), where the agent maximizes the sum of its rewards plus a bonus for the entropy of its policy[2]. In particular Ahmed et al. (2019) study specifically the impact of entropy on policy optimization. Not only does max-entropy RL boost exploration, it also yields an optimal policy that is stochastic, in the form of a Boltzmann measure, such that the agent keeps taking actions at random while maximizing the regularized objective. This is sometimes preferable than deterministic policies. In particular, Eysenbach and Levine (2021) show that the max-entropy RL optimal policy is robust to adversarial change of the reward function (their Theorem 4.1) and transition probabilities (their Theorem 4.2); see also references therein for more details on that topic. Finally, max-entropy RL is appealing from a theoretical perspective. For example, soft Q-learning, introduced by Haarnoja et al. (2017) (see also Haarnoja et al. (2018a,b) for implementations of soft Q-learning with an actor-critic scheme), strongly resembles PG in max-entropy RL (Schulman et al., 2017a); max-entropy RL has also been linked to variational inference (Levine, 2018). Other appealing features of max-entropy RL are discussed by Eysenbach and Levine (2019) and references therein.

A vast number of works on RL have focused on either infinite horizon tasks, or episodic tasks where the length of an episode is random. In both these cases, policies only depend on the current state of the agent. In Ernst et al. (2003), the fixed, finite horizon optimal policy is used as an approximation, as the horizon grows to infinity, to approximate the infinite-horizon optimal policy. Nonetheless and even though the fixed (deterministic) horizon setting has received less attention, the benefits of fixing the horizon are multiple and have been investigated in recent relevant works, such as Asis et al. (2019); Guin and Bhatnagar (2023); VP and Bhatnagar (2021). Asis et al. (2019) study a *Temporal Difference* algorithm (which is not PG), typically involving bootstrapping. When it is used offline (off-policy) together with function approximation, it encounters the well-known stability issue called the *deadly triad*, see Sutton and Barto (2018) Section 11.3. By using horizon-dependent value functions, they do not rely on bootstrapping, getting rid of one element of the triad, thus ensuring more stability. It is worth noting that thanks to the fixed-horizon setting, they empirically overcome the specific Baird's example of divergence. Guin and Bhatnagar (2023) defines an actor-critic algorithm for constrained RL, where the agent aims at maximizing the cumulative rewards while satisfying some given constraints. They prove the convergence for finite state and action spaces but do not study global convergence.

In the same spirit, van Seijen et al. (2019) investigate the impact of the discount factor when optimizing a discounted infinite-horizon objective evaluated on a finite-horizon undiscounted objective. They empirically found that for some tasks, lower discount factors (thus closer to a fixed-horizon objective) lead to better performance. With a fixed horizon,

[2] Other regularization techniques are used and studied in the literature, we focus on entropy regularized RL in this paper.

policies are time-dependent, and are usually called *non-stationary* policies, as in dynamic programming (Bertsekas, 1995).

Regarding PG methods specifically in the fixed-horizon setting with tabular softmax parametrization, the preprint by Klein et al. (2023) proves global convergence using the *gradient domination property*, which generally does not holds in the infinite state-action space, see our discussion on previous approaches below. PG with fixed horizon has also recently been studied outside of the MDP setting. Global convergence of some algorithms is established for some class of continuous time problems, see e.g. Hambly et al. (2021); Giegrich et al. (2022).

Contributions. We consider the function approximation setting with log-linear parametric policies, that are constructed as the softmax of linear models. For convergence results, we assume perfect gradient updates, i.e. we have access to the exact gradient of the objective. The main contributions of this work are:

(i) We define the fixed-horizon max-entropy RL objective and introduce a new algorithm (Equation (8)), named *Matryoshka Policy Gradient* (MPG).

(ii) We establish global convergence for continuous state and action space: under the realizability assumption, MPG converges to the unique optimal policy (Theorem 2). When the realizability assumption does not hold, we prove uniqueness of the optimal policy and prove global convergence of MPG (Theorem 3).

(iii) We approximate arbitrarily well the optimal policy for the infinite horizon objective by the optimal policy of the MPG objective (Proposition 2).

(iv) In the case where the policy is parametrized as the softmax of a (deep) neural network's output, we describe the limit of MPG training in terms of the *neural tangent kernel* and the *conjugate kernel* of the neural network at the end of training (Corollary 1). In particular, MPG globally converges in the *lazy regime*.

(v) Numerically, we successfully train agents on standard simple tasks without relying on RL tricks, and confirm our theoretical findings (see Section 4).

In our numerical experiments described in Section 4, we first consider an analytical task and verify the global convergence property of the MPG: MPG consistently finds the unique global optimum, which satisfies the projectional consistency property. Then, we study two benchmarks from OpenAI: the Frozen Lake game and the Cart Pole. We obtain successful policies for both benchmarks with the MPG algorithm, comparing also to vanilla PG method (Sutton et al., 1999). Rather than competing with the state-of-the-art algorithms, our aim is to provide a proof of concept by showing that successful training can be obtained with a straightforward implementation of MPG. We hope that more general and bigger scale experiments implementing variations of MPG will follow the present work.

Comparison with previous results and approaches. Besides the well-known *Policy Gradient Theorem* (see Chapter 13 in Sutton and Barto (2018)) that can imply convergence of PG (provided good learning rate and other assumptions), for many years, not much more was known about the global convergence of PG (i.e. convergence to an optimal policy) until recently. Despite the numerous remaining gaps, some important progress have already been made. In particular, the global convergence of PG methods has been studied and proved in specific settings, see for instance Fazel et al. (2018); Agarwal et al. (2022); Bhandari and Russo (2019); Mei et al. (2020); Zhang et al. (2020, 2019); Cen et al. (2020); Ding et al. (2021); Wang et al. (2019); Agazzi and Lu (2020); Bhandari and Russo (2020); Leahy et al. (2022); Guin and Bhatnagar (2023). Convergence guarantees often come with convergence rates (with or without perfect gradient estimates). Though strengthening the trust in PG methods for practical tasks, most of the theoretical guarantees obtained in the literature so far require rather restrictive assumptions, and often assume that the action-state space of the MDP is finite (but not always, e.g. Agazzi and Lu (2020) address continuous action-state space for neural policies in the mean-field regime and Leahy et al. (2022) prove global convergence when adding a strong enough regularization on the parameters.) In particular, in the context of tabular softmax policies, Li et al. (2021) study the dependency of the number of iterations of the (perfect) gradient PG update on the size of the state space, and construct environments with only three actions per state requiring the algorithm to make $\frac{1}{\eta}|\mathcal{S}|^{2^{\Omega(1/(1-\gamma))}}$ iterations to converge, where γ is the discount factor and η the learning rate.

We now highlight the key differences in proof techniques between our work and previous works. Agarwal et al. (2022) give many convergence guarantees for different policy gradient algorithms. In particular, in the tabular case with finite state and action spaces, global convergence is obtained thanks to the *gradient domination property*, stated in their Lemma 4.1 as follows: for every probability distributions μ, ρ on $\mathcal{S}$, for every policy π, the difference between the value function of the optimal policy π_* and the value funtion of π satisfies

$$\sum_{s\in\mathcal{S}}(V_{\pi_*}(s) - V_\pi(s))\rho(s) \leq \frac{1}{1-\gamma}\left|\left|\frac{d_\rho^{\pi_*}}{\mu}\right|\right|_\infty \max_{\pi'}(\pi' - \pi)\nabla_\pi \sum_{s\in\mathcal{S}} V_\pi(s)\mu(s),$$

where $d_\rho^{\pi_*}$ is the state-visitation distribution induced by the initial state distribution ρ and π_*, and the max is taken over the set of all policies. The factor $||d_\rho^{\pi_*}/\mu||_\infty$ is called the *distribution mismatch* between $d_\rho^{\pi_*}$ and μ. Since $d_\rho^{\pi_*}(s)$ is proportional to the discounted time spent in state s under the optimal policy π_*, the gradient domination property ensures that if the policy is trained with respect to a state distribution μ whose support contains that of $d_\rho^{\pi_*}$, then the gradient vanishes only at the optimum.

One advantage is that in the tabular setting, the rate of convergence can be deduced, involving the distribution mismatch, see e.g. Section 4 in Agarwal et al. (2022). In the function approximation setting with infinite state space and finite action space, they obtain convergence results for the *natural policy gradient* algorithm, which uses the Fisher information matrix induced by the policy in the update, but do not obtain the optimality of the limit.

Similarly, Mei et al. (2020) study convergence rates towards global optimum but rely on tabular parametrization with finite state and action spaces to guarantee finite concentration coefficient and to obtain a Łojasiewicz type of inequality (also lower bound on the gradient), which is vacuous in infinite state space. See also Ding et al. (2021) for global convergence of an entropy-regularized PG method with softmax policies with sample-based updates, assuming finite state and action spaces and tabular parametrization. Convergence rates for general modified policy iteration approaches are also obtained in Geist et al. (2019) for finite state and action spaces with finite concentration coefficient. In the parametric case, with log-linear policies, Yuan et al. (2022) study convergence rates for natural gradient descent. However, they do not show global convergence and also require finite state and action space. In Mei et al. (2021), a convergence rate to the global optimum is obtained for Geometry-aware Normalized PG, but it requires finite state and action spaces to guarantee finite concentrability coefficient, while their non-uniform Łojasiewicz inequality is vacuous as the number of states grows to infinity.

The main takeaway is that standard methods, including the works mentioned above, use bounds on the gradient of the objective (gradient domination property, Łojasiewicz inequality, ...) to deduce convergence rates, which often need finite state and action space, while global optimality usually requires tabular parametrization. Besides, none of them concerns non-stationary policies. On the other hand, our proof for the convergence of MPG does not rely on a lower bound of the gradient, which is why we do not obtain convergence rates. However, we obtain global convergence as follows:

i) The gradient of the objective remains Lipschitz continuous along the training trajectory, which ensures that the objective converges (Theorem 1).

ii) To ensure that the policy converges too, we show that the sequence generated during training is *relatively compact* (Lemma 9), which essentially guarantees that the sequence of policies generated during training converge (more precisely, any of its subsequence has a converging subsequence).

iii) The parameters remain bounded during training (Lemma 10), which entails that any limiting policy of training is a critical point that belongs to the parametric space.

iv) Using tools from information geometry (Appendix C), we then show that the only critical point of the objective inside the parametric space is the unique projection of the global optimal policy onto the parametric space (see proofs of Theorem 2 and Theorem 3). This projection is globally optimal in the parametric space.

The main strength of MPG

- State space and action space can be infinite, continuous and even unbounded.
- Global convergence is guaranteed in the function approximation setting with log-linear policies.
- Even when the realizable assumption does not hold, MPG converges to the unique optimal policy in the parametric policy space. This global optimum can be characterized as the unique policy satisfying the *projectional consistency property* in the parametric space.

2. Fixed-horizon max-entropy RL

In this section, we introduce the fixed-horizon max-entropy RL, describe its optimal policy and establish some of its properties.

2.1 Definitions

Markov Decision Process The *Markov Decision Process* (MDP) setup is a very standard and important setup in RL (Puterman, 2014; Sutton and Barto, 2018). It is suited for sequential decision making, where the dynamics are Markovian, i.e. depend on the past decisions only through the current state of the agent, making it mathematically tractable.

The agent evolves according to a MDP characterized by the tuple $(\mathcal{S}, \mathcal{A}, p, p_{\text{rew}})$ modelling the *environment*, a map called a *policy* $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$, and an *initial state distribution* ν on $\mathcal{S}$.[3] The action space can be state dependent $\mathcal{A}_s$, nonetheless we assume for simplicity that it is the same regardless of the current state of the agent. We assume that the action and state spaces $\mathcal{A}, \mathcal{S} \subset \mathbb{R}^d$ are closed sets. Let $s' \mapsto p(s, a, s')$ be the probability (the density if $\mathcal{S}$ is continuous) that the agent moves from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ after taking action $a \in \mathcal{A}$. When $p(s, a, s') = \delta_{s', f(s,a)}$ for some $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, then we say that the transitions are deterministic. The reward depends on the action and on the current state, its law is denoted by $p_{\text{rew}}(\cdot|s, a)$. Throughout, we assume that the rewards are uniformly bounded and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we denote by $r(a, s)$ the mean reward after taking action a at state s.

A stationary policy $\pi : \mathcal{A} \times \mathcal{S} \to [0\,, 1]$ is a map such that for all $s \in \mathcal{S}$, $\pi(\cdot|s)$ is a probability distribution on $\mathcal{A}$ that describes the law of the action taken by the agent at state s. Let $\mathcal{P}$ denote the set of stationary policies. Let $n \in \mathbb{N}$ be some fixed horizon, we denote by $\mathcal{P}_n$ the set of non-stationary policies $\pi = (\pi^{(1)}, \ldots, \pi^{(n)})$ where for all $i = 1, \ldots, n$, $\pi^{(i)} \in \mathcal{P}$. Henceforth, we use the term "policy" for non-stationary policies. We say that the agent follows a policy $\pi \in \mathcal{P}_n$ if and only if it chooses its actions sequentially according to $\pi^{(n)}$, then $\pi^{(n-1)}$, and so on until $\pi^{(1)}$ and the end of the episode. That is for each episode of fixed length n, starting from a given state S_0, the agent generates a path $S_0, A_0, S_1, A_1, \ldots, A_{n-1}, S_n$, where $A_i \sim \pi^{(n-i)}(\cdot|S_i)$ and $S_{i+1} \sim p(S_i, A_i, \cdot)$. Note that in the standard infinite horizon setting, a policy corresponds to an infinite sequence of the same stationary policy $\{(\pi, \pi, \pi, \cdots); \pi \in \mathcal{P}\} \subset \mathcal{P}_\infty$. All random variables are such that the process is Markovian.

Henceforth, we assume that $\mathcal{A}$ and $\mathcal{S}$ are continuous, the results identically holding true when they are countable. We also assume that

- The sets $\mathcal{A}, \mathcal{S} \subset \mathbb{R}^d$ are closed Borel sets.
- The transition probability function $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ and the reward function $r : \mathcal{A} \times \mathcal{S}$ are measurable.
- For all Borel set $B \subset \mathcal{S}$, the map $(s, a) \mapsto p(s, a, B)$ is continuous.
- The so-called *measurable selection assumption* holds.

The measurable selection assumption is a technical statement that ensures that the MDP has a well-defined optimal solution. Several conditions ensuring that it holds can be found in Hernández-Lerma and Lasserre (1999) Section 3.3. To avoid technical discussions that are not relevant to the present work, the reader can replace the measurable selection assumption by assuming the following simple sufficient and not too restrictive condition:

[3] Implicitly assumed in the MDP definition is the fact that all variables such that actions, visited states and rewards are measurable, so that they are well-defined random variables.

- The action space $\mathcal{A}$ is compact and the reward function and the transition probability function are continuous with respect to the Euclidean metric.

Value and Q functions. For every $s \in \mathcal{S}$ and $\pi, \pi' \in \mathcal{P}$, we denote by $\mathrm{D}_{\mathrm{KL}}(\pi||\pi')(s) = \mathrm{D}_{\mathrm{KL}}(\pi(\cdot|s)||\pi'(\cdot|s))$ the Kullback-Leibler divergence between $\pi(\cdot|s)$ and $\pi(\cdot|s')$, defined as

$$D_{\mathrm{KL}}(\pi||\pi')(s) := \int_{\mathcal{A}} \log \frac{\pi}{\pi'}(a|s)\pi(\mathrm{d}a|s),$$

and is set to ∞ if $\pi'(\cdot|s)$ is not absolutely continuous with respect to $\pi(\cdot|s)$.

To regularize the rewards, we add a penalty term that corresponds to the Kullback-Leibler (KL) divergence of the agent's policy and a baseline policy. In practice, the baseline policy can be used to encode a priori knowledge of the environment; a uniform baseline policy corresponds to adding entropy bonuses to the rewards. Regularizing with the KL divergence is thus more general than with entropy bonuses and this is the regularization that we consider in this paper, akin to Schulman et al. (2017a).

We denote by $\overline{\pi}$ the arbitrary baseline policy and let us assume for conciseness that $\overline{\pi} \in \mathcal{P}$. Let $\tau > 0$ be the so-called *temperature parameter* governing the strength the of the regularization. Similar to Ding et al. (2021); Mei et al. (2020); Geist et al. (2019) in the stationary setting, we define the n-step value function $V_\pi^{(n)} : \mathcal{S} \to \mathbb{R}$ induced by a policy $\pi \in \mathcal{P}_n$ as

$$V_\pi^{(n)} : s \mapsto \mathbb{E}_\pi \left[\sum_{k=0}^{n-1} (R_k - \tau D_{\mathrm{KL}}(\pi^{(n-k)}||\overline{\pi})(S_k)) \Big| S_0 = s \right],$$

where the expectation is along the trajectory of length n sampled under policy $\pi = (\pi^{(1)}, \dots, \pi^{(n)})$. Note that we have

$$V_\pi^{(n)}(s) = \mathbb{E}_{\pi^{(n)}}[R_0] - \tau D_{\mathrm{KL}}(\pi^{(n)}||\overline{\pi})(s) + \mathbb{E}_{\pi^{(n)}}[V_{\pi'}^{(n-1)}(S_1)], \tag{1}$$

where $\pi' = (\pi^{(1)}, \dots, \pi^{(n-1)}) \in \mathcal{P}_{n-1}$, and $S_1 \sim \int_{\mathcal{A}} p(s, a, \cdot)\pi^{(n)}(\mathrm{d}a|s)$. It is common to add a discount factor $\gamma \in (0, 1]$ to the rewards to favor more the quickly obtainable rewards. In the infinite horizon case ($n = \infty$), $\gamma < 1$ ensures that the cumulative reward is finite a.s. (provided finite first moment). Our study trivially applies to the case where the rewards are discounted.

The n-step entropy regularized Q-function induced by π is defined as

$$Q_\pi^{(n)} : (a, s) \mapsto r(a, s) + \int_{\mathcal{S}} p(s, a, \mathrm{d}s') V_{\pi'}^{(n-1)}(s'). \tag{2}$$

Notation: Henceforth, for a policy $\pi \in \mathcal{P}_n$, we use the abuse of notation $V_\pi^{(i)}$ for $i < n$ for the i-step value function associated with $(\pi^{(1)}, \dots, \pi^{(i)})$, and similarly for the Q functions and other quantities of interest, when the context makes it clear which policy is used.

2.2 Objective and optimal policy

The standard discounted max-entropy RL objective is defined for stationary policies $\pi \in \mathcal{P}$ by

$$J(\pi) := \int_{\mathcal{S}} \mathbb{E}_{\pi_t}\left[\sum_{k=0}^{T} \gamma^k \left(R_k - D_{\mathrm{KL}}(\pi_t||\overline{\pi})(S_k)\right) \Big| S_0 = s\right] \nu(\mathrm{d}s), \tag{3}$$

where $T \in \mathbb{N} \cup \{\infty\}$ is the horizon and ν is the initial state distribution, see e.g. Eysenbach and Levine (2021) and references therein. It is often assumed that T is random and therefore π is stationary.

Instead of the above objective, we consider the following objective function for non-stationary policy π:

$$J_n(\pi) := \int_{\mathcal{S}} V_\pi^{(n)}(s)\nu(\mathrm{d}s). \tag{4}$$

Since we assume that the rewards are bounded and since the Kullback-Leibler divergence is non-negative, the objective function above is bounded from above by $n||r||_\infty$.

We say that a policy $\pi \in \mathcal{P}_n$ is optimal if and only if $J_n(\pi) \geq J_n(\pi')$ for all $\pi' \in \mathcal{P}_n$. Note that in general, uniqueness is not guaranteed, since for example $\pi^{(n)}$ only sees states in the support of ν, which can be strictly smaller than $\mathcal{S}$. If a policy $\pi \in \mathcal{P}_n$ is such that $V_\pi^{(i)}(s) \geq V_{\pi'}^{(i)}(s)$ for all $s \in \mathcal{S}$, all $i = 1, \ldots, n$ and all $\pi' \in \mathcal{P}_n$, we say that π is *uniformly optimal*. It is clear that a uniformly optimal policy is in particular optimal. The existence and unicity of the uniformly optimal policy is established by the next proposition, providing in passing its explicit expression.

Proposition 1 *There exists a unique uniformly optimal policy, denoted by $\pi_* = (\pi_*^{(1)}, \ldots, \pi_*^{(n)}) \in \mathcal{P}_n$. The i-step optimal policies, $i = 1, \ldots, n$, can be obtained as follows: for all $a \in \mathcal{A}$, $s \in \mathcal{S}$,*

$$\pi_*^{(1)}(a|s) = \frac{\overline{\pi}(a|s)\exp(r(a,s)/\tau)}{\mathbb{E}_{\overline{\pi}}[\exp(r(A,s)/\tau)]}, \quad \pi_*^{(i+1)}(a|s) = \frac{\overline{\pi}(a|s)\exp\left(Q_*^{(i+1)}(a,s)/\tau\right)}{\mathbb{E}_{\overline{\pi}}\left[\exp\left(Q_*^{(i+1)}(A,s)/\tau\right)\right]},$$

where $Q_*^{(i+1)}$ is a short-hand notation for $Q_{\pi_*}^{(i+1)}$ recursively defined as in (2).

For $i = 1, \ldots, n$, let $\mathbf{m}_\pi^{(i)}$ be the law of S_{n-i} under π and with given initial state distribution ν. Note that $\mathbf{m}_\pi^{(n)} = \nu$. It is readily seen that if π is optimal for J_n, then necessarily, $\pi^{(n)}(\cdot|s) = \pi_*^{(n)}(\cdot|s)$ for ν-almost every s. In particular, $m_\pi^{(n-1)} = m_{\pi_*}^{(n-1)}$ and then $\pi^{(n-1)}(\cdot|s) = \pi_*^{(n-1)}(\cdot|s)$ for $m_{\pi_*}^{(n-1)}$-almost every $s \in \mathcal{S}$. Reasoning by induction shows that $\pi^{(i)}(\cdot|s) = \pi_*^{(i)}(\cdot|s)$ for $m_{\pi_*}^{(i)}$-almost every $s \in \mathcal{S}$, for all $i = 1, \ldots, n$. Hence, the optimal policy is unique over the support of the state distributions induced by the uniformly optimal policy. Since $\pi_*(a|s) > 0$ for all $(a,s) \in \mathcal{A} \times \mathcal{S}$, the supports of these state distributions consist of all reachable states from the support of ν.

Lemma 1 *For all $s \in \mathcal{S}$ and $n \geq 1$, it holds that*

$$V_*^{(n)}(s) = \tau \log \mathbb{E}_{\overline{\pi}} \left[\exp\left(Q_*^{(n)}(A, s)/\tau \right) \right],$$

where $V_*^{(0)}(s') = 0$.

Thanks to Lemma 1, we can write more concisely

$$\pi_*^{(i)}(a|s) = \overline{\pi}(a|s) \exp\left(\left(Q_*^{(i)}(a, s) - V_*^{(i)}(s) \right) / \tau \right). \tag{5}$$

For all $n, m \in \mathbb{N}$ such that $n > m$, we define the operator $T_{n,m} : \mathcal{P}_n \to \mathcal{P}_m$ by

$$T_{n,m} : (\pi^{(1)}, \ldots, \pi^{(n)}) \mapsto (\pi^{(1)}, \ldots, \pi^{(m)}). \tag{6}$$

In Proposition 2 below, for all $n \in \mathbb{N}$, we denote by $\pi_{*,n} \in \mathcal{P}_n$ the uniformly optimal policy with maximum horizon n and with discounted rewards. The infinite horizon entropy-regularized RL objective is defined in (3) with $T = \infty$, and we denote by $\pi_{*,\infty}$ the corresponding uniformly optimal policy.

Proposition 2 *Suppose that the initial state distribution ν has full support and that the MDP is ergodic. We have:*

(i) As $n \to \infty$, the policy $\pi_{*,n}^{(n)}$ converges to $\pi_{*,\infty}$, in the sense that for ν-almost all $s \in \mathcal{S}$,

$$\lim_{n\to\infty} \int_{\mathcal{A}} \left| \pi_{*,n}^{(n)}(a|s) - \pi_{*,\infty}(a|s) \right| \mathrm{d}a = 0.$$

(ii) for all $n, m \in \mathbb{N}$ such that $n > m$, it holds that $T_{n,m}(\pi_{*,n}) = \pi_{*,m}$.

The above Proposition 2 is intuitive when ν has full support so that the unique optimal policy is the uniformly optimal policy: item (i) shows that one can learn the standard discounted entropy-regularized RL objective by incrementally extending the agent's horizon; item (ii) goes the other way and shows that the uniformly optimal policy for large horizon is built of shorter horizons uniformly optimal policies in a consistent manner.

3. Matryoshka Policy Gradient

3.1 Policy parametrization

For $i \in \{1, \ldots, n\}$, let $\theta^{(i)} \in \mathbb{R}^{P_i}$ be the parameters of a linear model $h_{\theta^{(i)}}^{(i)} : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, that outputs for all $(a, s) \in \mathcal{A} \times \mathcal{S}$ the i-step preference $h_{\theta^{(i)}}^{(i)}(a, s)$ for action a at state s, that is,

$$h_{\theta^{(i)}}^{(i)}(a, s) := \theta^{(i)} \cdot \psi^{(i)}(a, s),$$

where $\psi^{(i)} : \mathcal{A} \times \mathcal{S} \to \mathbb{R}^{P_i}$ is a feature map. We assume throughout the paper that $\psi^{(i)}$ is a continuous and bounded map, such that for all $s \in \mathcal{S}$ and all $\theta^{(i)}$ with $||\theta^{(i)}|| \neq 0$, the map

$a \mapsto h_\theta^{(i)}(a,s)$ is not constant. The i-step policy $\pi_{\theta^{(i)}}^{(i)}$ is defined as the Boltzmann policy according to $h^{(i)}$, that is, for all $(a,s) \in \mathcal{A} \times \mathcal{S}$,

$$\pi_{\theta^{(i)}}^{(i)}(a|s) := \overline{\pi}(a|s) \frac{\exp(h_{\theta^{(i)}}^{(i)}(a,s)/\tau)}{\int_{\mathcal{A}} \exp(h_{\theta^{(i)}}^{(i)}(a',s)/\tau)\overline{\pi}(\mathrm{d}a'|s)}.$$

The gradient of the policy thus reads as

$$\nabla \pi_{\theta^{(i)}}^{(i)}(a|s) = \pi_{\theta^{(i)}}^{(i)}(a|s) \int_{\mathcal{A}} \left(\delta_{a,\mathrm{d}a'} - \pi_{\theta^{(i)}}^{(i)}(\mathrm{d}a'|s)\right) \nabla h_{\theta^{(i)}}^{(i)}(a',s)/\tau. \tag{7}$$

3.2 Definition of the MPG update

Policy gradient (PG) for max-entropy RL consists in following $\nabla_\theta J(\pi_\theta)$ for the standard objective (3). In our setting, the ideal PG update would be such that $\theta_{t+1} - \theta_t = \eta \nabla_\theta J_n(\pi_t)$. We introduce Matryoshka Policy Gradient (MPG) as a practical algorithm that produces unbiased estimates of the gradient (see Lemma 8 in Appendix).

Suppose that at time $t \in \mathbb{N}$ of training, the agent starts at a state $S_0 \sim \nu_0$. To lighten the notation, we write $\pi_t^{(i)} := \pi_{\theta_t^{(i)}}^{(i)}$. We assume that the agent samples a trajectory according to the policy π_t, defined as follows:

- sample action A_0 according to $\pi_t^{(n)}(\cdot|S_0)$,
- collect reward $R_0 \sim p_{\mathrm{rew}}(\cdot|S_0, A_0)$ and move to next state $S_1 \sim p(S_0, A_0, \cdot)$,
- sample action A_1 according to $\pi_t^{(n-1)}(\cdot|S_1)$,
- $\cdots$
- stop at state S_n.

The MPG update is as follows: for $i = 1, \dots, n$,

$$\begin{aligned}\theta_{t+1}^{(i)} &= \theta_t^{(i)} + \eta \sum_{\ell=n-i}^{n-1} \left(R_\ell - \tau \log \frac{\pi_t^{(n-\ell)}}{\overline{\pi}}(A_\ell|S_\ell) \right) \nabla \log \pi_t^{(i)}(A_{n-i}|S_{n-i}) \\ &= \theta_t^{(i)} + \eta C_i \nabla \log \pi_t^{(i)}(A_{n-i}|S_{n-i}), \end{aligned} \tag{8}$$

where we just introduced the shorthand notation C_i. We see that the i-step policy $\pi^{(i)}$ is updated using the $(i-\ell)$-step policies.

3.3 Global convergence: the realizable case

Recall that $\mathbf{m}_\pi^{(i)}$ denotes the law of S_{n-i} when following policy π from a starting state with distribution ν. We say that a sequence of policies $(\pi_t)_{t\in\mathbb{N}} \subset \mathcal{P}_n$ converges to $\pi \in \mathcal{P}_n$ if and only if for every $i = 1, \dots, n$, for π-almost every $a \in \mathcal{A}$ and for $\mathbf{m}_\pi^{(i)}$-almost every $s \in \mathcal{S}$, it holds that

$$\pi_t^{(i)}(a|s) \underset{t\to\infty}{\longrightarrow} \pi^{(i)}(a|s). \tag{9}$$

To be concise, the states on which the convergence holds are called *reachable state*, where we keep the dependence in i implicit. In particular, we will write "$\pi_1 = \pi_2$ on reachable states" to mean that for all $i = 1, \dots, n$, it holds that $\pi_1^{(i)}(\cdot|s) = \pi_2^{(2)}(\cdot|s)$ for $\mathbf{m}_{\pi_2}^{(i)}$-almost every s.

With PG, the so-called *Policy Gradient Theorem* (see Section 13.2 in Sutton and Barto (2018)) provides a direct way to guarantee convergence of the algorithm. The analog holds in our setup, that is, if θ_{t+1} is obtained as in (8), then $\mathbb{E}[\theta_{t+1} - \theta_t] = \eta \nabla_\theta J_n(\pi_t)$; it is proven in Appendix D.3.

Importantly, training with true gradient update converges, as stated in the following theorem:

Theorem 1 *Suppose that* $||\psi|| := \sup_{a \in \mathcal{A}, s \in \mathcal{S}, i=1,\dots,n} ||\psi^{(i)}(a,s)||_2 < \infty$, *and that* $\theta_{t+1} = \theta_t + \eta \nabla_\theta J_n(\pi_t)$ *for all* $t \geq 0$. *For all initial parameters* θ_0, *if* $\eta < \frac{2}{L(\theta_0)}$, *then it holds that* $J_n(\pi_t)$ *converges as* $t \to \infty$, *where*

$$L(\theta_0) := 4(n^2 + n^3)\Big(2 + \frac{P}{\tau}\Big)\frac{\|\psi\|^2}{\tau^2}(J_n(\pi_*) - J_n(\pi_{\theta_0}) + 3^n\|r\|_\infty) + 4n^2\frac{\|\psi\|^2}{\tau^2}.$$

Moreover, $||\nabla_\theta J_n(\pi_t)||_2 \to 0$ *as* $t \to \infty$.

Note that Theorem 1 does not show that the policy π_t converges as $t \to \infty$, only that the objective does. Furthermore, even if π_t converges, its limit could be outside of the parametric space, if the parameters during training are such that $||\theta_t||_2 \to \infty$ as $t \to \infty$.

For the Theorem below, we assume the following:

A1. Realizability assumption There exists $\theta_* \in \mathbb{R}^P$ such that $\pi_{\theta_*} = \pi_*$.

Theorem 2 *Under* ***A1.*** *and the same assumptions as in Theorem 1,* $\lim_{t\to\infty} \pi_t = \pi_*$ *in the sense of* (9).

Intuition of the proof: the bandit case. To prove Theorem 1, we bound the 2-norm of the Hessian of the objective to show that $\nabla_\theta J_n(\pi_\theta)$ is Lipschitz, with a constant that only decreases along training trajectories, as long as the learning rate is chosen small enough. However, the objective is non-concave, and it is not obvious that its critical points all correspond to policy π_*. We present the heuristics on the illustrative bandit case $|\mathcal{S}| = 1$ with horizon $n = 1$. We thus keep the state and horizon implicit below.

We know from Theorem 1 that $J(\pi_t)$ converges as $t \to \infty$, but the limit could be reached as $\|\theta_t\| \to \infty$. However, since there is an optimal $\theta_* \in \mathbb{R}^P$ by **A1.**, for very large $\|\theta_t\|_2$, the vectors $-\theta_t$ and $\theta_* - \theta_t$ must be almost colinear. Based on this observation, we show that if the norm of θ_t is very large, then moving the parameters slightly in the direction $-\theta_t$ improves the performance, showing that $\|\theta_t\|_2$ remains bounded.

The second step is to show that for all $\theta \in \mathbb{R}^P$ if $\pi_\theta \neq \pi_*$, then θ is not a critical point. The objective is given by $J(\pi_\theta) = J(\pi_*) - \tau D_{\mathrm{KL}}(\pi_\theta || \pi_*)$. Without loss of generality, assume that $\tau = 1$. In particular,

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= -\int_{\mathcal{A}} \pi_\theta(\mathrm{d}a)\left(\log\frac{\pi_\theta}{\pi_*}(a) + 1\right)\nabla_\theta \log \pi_\theta(a) \\ &= -\int_{\mathcal{A}} \pi_\theta(\mathrm{d}a) \log\frac{\pi_\theta}{\pi_*}(a) \nabla_\theta \log \pi_\theta(a),\end{aligned}$$

where we used that $\mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(A)] = 0$. For the case $|\mathcal{A}| < \infty$ and tabular parametrisation $\pi_\theta(a) = e^{\theta_a}/\mathbb{E}_{\overline{\pi}}[e^{\theta_A}]$ and $\nabla_\theta \log \pi_\theta(a) = (\delta_a(a') - \pi_\theta(a'))_{a' \in \mathcal{A}}$. Recall that $\pi_*(a) = e^{r(a)}/\mathbb{E}_{\overline{\pi}}[e^{r(A)}]$. We have

$$\begin{aligned}
\nabla_\theta J(\pi_\theta) &= -\sum_{a \in \mathcal{A}} \pi_\theta(a)\,(\theta_a - r(a) + \text{Const})\,\nabla_\theta \log \pi_\theta(a) \\
&= -\sum_{a \in \mathcal{A}} \pi_\theta(a)\,(\theta_a - r(a))\,(\delta_a(a') - \pi_\theta(a'))_{a' \in \mathcal{A}} \\
&= -\left(\pi_\theta(a')(\theta_{a'} - r(a') - \mathbb{E}_{\pi_\theta}[\theta_A - r(A)])\right)_{a' \in \mathcal{A}}.
\end{aligned}$$

Hence, the gradient is null if and only if $\theta_{a'} - r(a') = \mathbb{E}_{\pi_\theta}[\theta_A - r(A)]$ for all $a' \in \mathcal{A}$, that is, if and only if $\theta_{a'} - r(a')$ is constant in a'. This is equivalent to having $\pi_\theta = \pi_*$, which proves that all critical points of $\theta \mapsto J(\pi_\theta)$ encode the optimal policy π_* in the bandit case with tabular softmax parametrisation.

For the more general log-linear parametrisation $h_\theta = \theta \cdot \psi$, the gradient times itself yields

$$\begin{aligned}
\nabla_\theta J(\pi_\theta) \cdot \nabla_\theta J(\pi_\theta) &= \int_{\mathcal{A}} \pi_\theta(\mathrm{d}a) \log \frac{\pi_\theta}{\pi_*}(a) \int_{\mathcal{A}} \pi_\theta(\mathrm{d}a') \log \frac{\pi_\theta}{\pi_*}(a') \nabla_\theta \log \pi_\theta(a) \cdot \nabla_\theta \log \pi_\theta(a') \\
&= \int_{\mathcal{A}} \pi_\theta(\mathrm{d}a)\,(h_\theta(a) - r(a)) \int_{\mathcal{A}} \pi_\theta(\mathrm{d}a')\,\left(h_\theta(a') - r(a')\right) \widetilde{\Theta}(a, a'),
\end{aligned}$$

where we introduced the matrix $\widetilde{\Theta}(a, a') = \nabla_\theta \log \pi_\theta(a) \cdot \nabla_\theta \log \pi_\theta(a')$, depending implicitly on the parameters. It is linked to the other matrix $\Theta(a, a') := \psi(a) \cdot \psi(a')$ since one can check that

$$\widetilde{\Theta}(a, a') = \Theta(a, a') - \mathbb{E}_{\pi_\theta}[\Theta(A, a')] - \mathbb{E}_{\pi_\theta}[\Theta(a, A')] + \mathbb{E}_{\pi_\theta}[\Theta(A, A')].$$

The gradient $\nabla_\theta J(\pi_\theta)$ is null if and only if $(h_\theta(a) - r(a))_{a \in \mathcal{A}}$ belongs to the null space of $\widetilde{\Theta}$. Note, however, that $(h_\theta(a) - r(a))_{a \in \mathcal{A}}$ belongs to the image of Θ, since we assume **A1.**. One can show through standard facts on matrices that the relation between Θ and $\widetilde{\Theta}$ implies that $\nabla_\theta J(\pi_\theta)$ is null if and only if $h_\theta - r$ is constant, or equivalently, if and only if $\pi_\theta = \pi_*$.

This idea works for infinite or continuous $\mathcal{A}$ using kernels and their reproducible kernel Hilbert space (details are provided in Appendix B). We also stress that using non-stationary policies is crucial in extending the proof to larger horizons $n > 1$, by using that fixing the parameters of the policies with horizon less or equal to $n - 1$, the horizon n objective can be seen as a horizon 1 objective where the rewards are determined by r and the fixed subsequent policies.

3.4 Global convergence: beyond the realizability assumption

Let $\mathscr{P}_n = \{\pi_\theta; \theta \in \mathbb{R}^P\} \subset \mathcal{P}_n$ be the set of parametric policies. We now address the case $\pi_* \notin \mathscr{P}_n$, that is, Assumption **A1.** does not hold. We give a sketch of the main ideas to extend Theorem 2 to the non-realizable case, showing global convergence and providing a characterisation of the limit. All details are provided in Appendix D

We focus on the 1-step policy. Suppose that $\vartheta \in \mathbb{R}^P$ is a critical point of $\theta \mapsto J_n(\pi_\theta)$. Since $Q_*^{(1)}(a, s) = r(a, s)$, one can show that

$$0 = \nabla_{\theta^{(1)}} J_n(\pi_\vartheta) = -\mathbb{E}_{\pi_\vartheta}\left[\nabla_{\theta^{(1)}} D_{\mathrm{KL}}(\pi_\vartheta^{(1)} || \pi_*^{(1)})(S_{n-1})\right], \tag{10}$$

where the law of S_{n-1} only depends on $\pi_\vartheta^{(n)}, \dots, \pi_\vartheta^{(2)}$. Since this law is fixed (ϑ is a critical point), the right-hand side above corresponds to the gradient of a *Bregman divergence* on the set of 1-step policies, which we denote by $D(\pi_\vartheta^{(1)}, \pi_*^{(1)})$. Let $\pi_{\theta_*}^{(1)} \in \operatorname{argmin}_{\pi_\theta^{(1)} \in \mathscr{P}_1} D(\pi_\theta^{(1)}, \pi_*^{(1)})$. Bregman divergences satisfy a Pythagorean identity, which in particular implies that

$$D(\pi_\vartheta^{(1)}, \pi_*^{(1)}) = D(\pi_\vartheta^{(1)}, \pi_{\theta_*}^{(1)}) + D(\pi_{\theta_*}^{(1)}, \pi_*^{(1)}).$$

Hence, we have by (10) that

$$0 = -\nabla_{\theta^{(1)}} D(\pi_\vartheta^{(1)}, \pi_{\theta_*}^{(1)}).$$

We deduce that $\pi_{\vartheta^{(1)}}$ is a critical point of the 1-step MPG objective, where the initial state distribution is prescribed by $\pi_\vartheta^{(i)}$ for $i = 2, \dots, n$, and where the rewards are given by $r_{\theta_*} := h_{\theta_*}^{(1)}$. In particular, Theorem 2 applies and shows that $\pi_\vartheta^{(1)}(\cdot|s) = \pi_{\theta_*}^{(1)}(\cdot|s)$ for all reachable states s. This also proves the uniqueness of $\pi_{\theta_*}^{(1)}$ on reachable states.

The argument propagates to larger horizons, by using that maxima can be taken in any order, which proves that the unique critical point π_ϑ of J_n is globally optimal. Formally, the following theorem completes the picture of the global convergence guarantees of MPG:

Theorem 3 *Under the same assumptions as in Theorem 1, it holds that* $\lim_{t\to\infty} \pi_t = \pi_{\theta_*}$ *in the sense of* (9)*, where* $\pi_{\theta_*} = \operatorname{argmax}_{\pi_\theta \in \mathscr{P}_n} J_n(\pi_\theta)$ *is unique on reachable states.*

Clearly, when $\pi_* \in \mathscr{P}_n$, then $\pi_\infty = \pi_*$ on reachable states and we retrieve Theorem 2.

3.5 Projectional consistency property

Let $\Theta^{(i)} : (\mathcal{A} \times \mathcal{S})^2 \to \mathbb{R}$ be the positive-semidefinite kernel given by the dot product of the feature map, that is

$$\Theta^{(i)}((a,s),(a',s')) := \psi^{(i)}(a,s) \cdot \psi^{(i)}(a',s').$$

The function space $\{f : (a,s) \mapsto \theta^{(i)} \cdot \psi^{(i)}(a,s); \theta^{(i)} \in \mathbb{R}^{P_i}\}$ from which we chose the preferences of our parametric Boltzmann policies corresponds to the *reproducible kernel Hilbert space* (RKHS) associated with $\Theta^{(i)}$, that we denote by $\mathcal{H}_{\Theta^{(i)}}$. Note that when $\mathcal{A}, \mathcal{S}$ are finite, with Kronecker delta kernels $\Theta^{(i)}((a,s),(a',s')) = \delta_{a,a'}\delta_{s,s'}$, we retrieve the so-called *tabular case* with one parameter $\theta_{s,a}$ per state-action pair (s,a). Since we assume that the $\psi^{(i)}$'s are continuous and bounded, it is also the case for the kernels.

The realizability assumption **A1.** can be equivalently written as: for all $i = 1, \dots, n$, there exists a map $C_i : \mathcal{S} \to \mathbb{R}$ such that $(a,s) \mapsto Q_*^{(i)}(a,s) + C_i(s) \in \mathcal{H}_{\Theta^{(i)}}$, where the maps C_i are constant in a. The C_i's play no role in the policies encoded by functions in the RKHS, since for a fixed s, shifting the preferences by a constant keeps the policy unchanged.

It turns out that the global optimum π_{θ_*} from Theorem 2 can be characterised by a property of independent interest. Let $\theta \in \mathbb{R}^P$ and for all $i = 1, \dots, n$, let $\mathbf{m}_\theta^{(i)}$ be the law of state S_{n-i} under policy π_θ with $\mathbf{m}_\theta^{(n)} = \nu_0$ by assumption. Define $P_i : L^2(\mathbf{m}_\theta^{(i)}(\mathrm{d}s)\pi_\theta^{(i)}(\mathrm{d}a|s)) \to \mathcal{H}_{\Theta^{(i)}}$ to be the orthogonal projection onto $\mathcal{H}_{\Theta^{(i)}}$ in the $L^2(\mathbf{m}_\theta^{(i)}(\mathrm{d}s)\pi_\theta^{(i)}(\mathrm{d}a|s))$ sense. We say that π_θ satisfies the *projectional consistency property* if and only if for all $i = 1, \ldots, n$, it holds that

$$\pi_\theta^{(i)}(a|s) = \overline{\pi}(a|s)\frac{\exp\left(P_i Q^{(i)}_{\pi_\theta}(a,s)/\tau\right)}{\int_{\mathcal{A}} \overline{\pi}(\mathrm{d}a'|s)\exp\left(P_i Q^{(i)}_{\pi_\theta}(a',s)/\tau\right)}. \tag{11}$$

Proposition 3 *The global optimum π_{θ_*} from Theorem 3 is the only policy in $\mathscr{P}_n$, up to non-reachable states, that satisfies the projectional consistency property* (11)

3.6 Neural MPG

Suppose that instead of a linear model, the policy's preferences $h_\theta^{(i)}$, $i = 1, \ldots, n$, are parametrized by deep neural networks. It is immediate from the proofs that the policy gradient theorem holds true, that is, $\theta_{t+1} - \theta_t = \eta \nabla_\theta J_n(\pi_t)$ for the ideal MPG update. We describe the limit of training in terms of the *Neural Tangent Kernels* (NTKs) of the neural networks and the *conjugate kernels* (CKs). The NTK of the i-step policy (or rather, of the i-step preference) at time t of training is defined for all $(a,s), (a',s') \in \mathcal{A} \times \mathcal{S}$ as

$$\Theta_t^{(i)}((a,s),(a',s')) := \nabla_{\theta^{(i)}} h_t^{(i)}(a,s) \cdot \nabla_{\theta^{(i)}} h_t^{(i)}(a',s').$$

The CK of the i-step policy is defined as the inner product of the last hidden layer, that we denote by α, that is

$$\Sigma_t((a,s),(a',s')) := \alpha_t(a,s) \cdot \alpha_t(a',s').$$

Moreover, letting $\mathcal{H}_K$ be the induced RKHS of a kernel K, it holds that $\mathcal{H}_{\Sigma_t} \subset \mathcal{H}_{\Theta_t}$, see Appendix A.2 for more details.

For the trained policy π_∞, let $\mathscr{P}_n^\Theta$ be the space of log-linear policies whose i-step preference belongs to $\mathscr{H}_{\Theta_\infty^{(i)}}$, $i = 1, \ldots, n$, and similarly for $\mathscr{P}_n^\Sigma$ and $\Sigma_\infty^{(i)}$.

Corollary 1 *Let $\pi_t \in \mathcal{P}_n$ be parametrized by neural networks. Suppose that $\theta_{t+1} - \theta_t = \eta \nabla_\theta J_n(\pi_t)$ with $\eta > 0$ small enough and that $\pi_\infty = \lim_{t\to\infty} \pi_t$ with parameters $||\theta_\infty||_2 < \infty$. Then, it holds that*

$$\pi_\infty = \underset{\pi \in \mathscr{P}_n^\Theta}{\operatorname{argmax}}\ J_n(\pi) = \underset{\pi \in \mathscr{P}_n^\Sigma}{\operatorname{argmax}}\ J_n(\pi).$$

In particular, if $\pi_* \in \mathcal{P}_n^\Theta$ (equivalently $\mathcal{P}_n^\Sigma$), then $\pi_\infty = \pi_*$ on reachable states.

A direct consequence of Corollary 1 is global convergence of MPG in the NTK regime, see the forthcoming Remark 1 in Appendix.

4. Numerical experiments

This section describes the performance of the MPG framework. In the first experiment, we evaluate the MPG framework on an analytical task and show that we converge to the the optimal policy when the optimal policy is realisable, and to the optimal policy in the parameter space when the true optimal policy is not representable by the policy parameterisation. Then, we compare the performance of MPG against REINFORCE (Sutton et al., 1999) (denoted as PG), REINFORCE with entropy regularisation (softPG) and a non-stationary policy gradient method (nsPG), which is the MPG method without entropy regularisation in two simple control tasks. More details on the implementation, experimental setups and additional results can be found in Appendix F.

4.1 Analytical task

To numerically evaluate the consequences of Theorem 2, we devise the following analytical problem: consider a state-space consisting of $\mathcal{S} = \{0, 1, 2, 3, 4\}$, an action space $\mathcal{A} = \{1, 2\}$ with horizon $n = 2$. At each state s, the agent performs action a, taking the agent to the next state $(s + a) \mod 5$ (see appendix F.1 for fully specified $Q^{(1)}_*$ function, the linear basis $\{e_i; i = 1, \ldots, 5\}$ considered and experimental setup for the presented experiments).

We consider the preference function to be represented by a linear model and we consider a true gradient update. Then, we investigate the first two step policies obtained using MPG are when assumption **A1.** holds and when it does not. The results are shown in Figure 1. Namely, on the left, we use the full basis $\{e_i; i = 1, \ldots, 5\}$ for the parametric model, and we are able to find the 1-step and 2-step policies which maximize the objective J, and converge towards the optimal 1-step and 2-step policies. On the right of Figure 1, we performed the same experiment using an incomplete basis, that cannot express $Q^{(1)}_*$ nor $Q^{(2)}_*$. Namely, we used $\{e_i; i = 1, \ldots, 4\}$ for both the 1-step and the 2-step policies. In this case, we check that the limit is the only policy satisfying the projectional consistency property within the parametric policy space. In both cases, the L_∞-error between the obtained policies and optimal policies go to zero as more episodes are used.

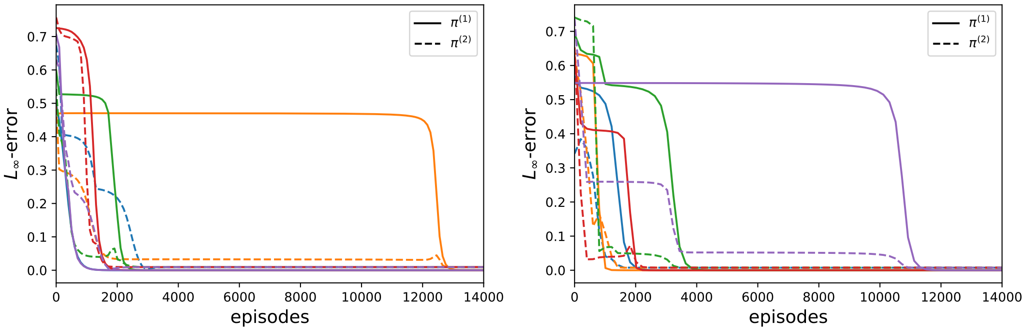

Figure 1: Analytical task. Convergence of 5 agents with random initialisation during training; the errors are measured through the L_∞-norm and $\pi^{(1)}$ denotes the one-step policy and $\pi^{(2)}$ the two-step policy. On the left, the convergence of the found 1-step and 2-step policies towards the optimal policies when the parametric space can represent the policy (i.e. when assumption **A1.** holds) is shown. On the right, the convergence of the 1-step and 2-step policies towards the optimal projected policies (i.e. when assumption **A1.** does not hold) is shown.

4.2 Control problems

In this section we present a summary of the performance of the MPG algorithm on two standard RL problems, comparing it to REINFORCE (Sutton et al., 1999) (denoted as PG), REINFORCE with entropy regularisation (softPG) and a non-stationary policy gradient method (nsPG). In the following experiments, we use a deep neural network to represent the policy (see appendix F for architecture) and we estimate the gradient update based on one trajectory as in (8).

For both tasks, we follow the experimental protocol as in Patterson et al. (2023), where we first make a sweep over the hyper-parameter spaces considered, evaluating the performance over 3 agents per set of hyper-parameters, which are initial temperature τ_0 and initial learning rate η_0 because we decay both the learning rate and temperature – namely, starting from a higher temperature encourages the exploration of the environment in early stages of training. After this initial stage, we select the best performing sets of hyper-parameters and run more throughout experiments (over 50 agents) to compare the performance of the different algorithms. For those experiments, confidence intervals around the mean are computed using bootstrap and considering $m = 1000$ resampled samples of the mean. In this section we present the results for the the extended experiments, while the preliminary experiments with the hyper-parameter exploration can be found in the appendix F.2.

Frozen Lake: The Frozen Lake benchmark (see Brockman et al. (2016)) features a 4×4 grid composed of cells, holes and one treasure, and a discrete action space, namely, the agent can move in four directions (up, down, left, right). The episode terminates when the agent reaches the treasure or falls down holes.

For all three algorithms, we considered the hyper-parameter space over initial learning rates and initial temperature, as specified in table 1. Then, for each algorithm, we augmented the search space if we found best performing agents at the boundary of the considered initial hyper-parameter space, also denoted in table 1. In addition, we considered a horizon length of $N = 20$, a terminal $\tau_T = 0.01$, terminal learning rate $\eta_T = 1 \times 10^{-6}$ and 1000 episodes. From these initial runs (results can be found in the appendix), we found the best sets of hyper-parameters for each of the algorithms, denoted on table 2[4]. Using those hyper-parameters, we ran more extended experiments, now considering 50 independent agents. In figure 2, on the left, we present the training curves, showing the accumulated reward per episode, the shaded regions bound the mean using the 2.5th percentile and 97.5th percentile means using bootstrap to compute the confidence interval; on the right, we present the histogram of the cumulative rewards at test time, after training: each agent attempts to solve the task 100 times. While the performance between MPG and PG was relatively similar: either the agent finds the treasure consistently or fails to find the treasure, training with nsPG or softPG did not yield a good performance in this task; namely, using softPG the agent often gets stuck at moving around the map (which yielded a +0.01 reward at each step, and a cumulative reward of 0.2) instead of finding the treasure. We conducted more experiments using different terminal temperature τ_T and terminal learning rate η_T with little success. Possible ways to address this could be to reshape the reward function further but this was beyond the scope of this current paper.

[4] In the cases where there were several sets of good hyper-parameters, we ran an intermediate step with 15 agents and selected the best set of hyper-parameters.

Table 1: Hyper-parameters for Frozen lake

	Initial	PG	softPG	nsPG	MPG
η_0	$\{0.01, 5\times 10^{-3}, 1\times 10^{-3},$ $5\times 10^{-4}, 1\times 10^{-4}\}$	$\{\}$	$\{0.1, 0.05\}$	$\{0.5, 0.1, 0.05\}$	$\{\}$
τ_0	$\{0.15, 0.2, 0.25, 0.3, 0.35,$ $0.4, 0.45, 0.5, 0.55, 0.6\}$	NA	$\{0.6, 0.65, 0.7\}$	NA	$\{\}$

Table 2: Best set of hyper-parameters for Frozen lake

	PG	softPG	nsPG	MPG
η_0	0.01	0.01	0.05	5×10^{-3}
τ_0	NA	0.6	NA	0.4

Furthermore, we noted that when the horizon was decreased (for $N = 10$ and $N = 15$), we were not able to find the treasure with PG, softPG nor nsPG, whereas training with MPG, the agents would still consistently find the treasure and train successfully.

Cart Pole: The Cart Pole benchmark is a classical control problem where a pole is attached by an un-actuated joint to a cart, which moves along a frictionless track. The pole is placed upright on the cart, and the goal is to balance the pole by moving the cart to the left or right for some finite horizon time. It features a continuous environment and a discrete action space.

We considered a horizon length of $N = 100$, a terminal $\tau_T = 0.01$, terminal learning rate $\eta_T = 5\times 10^{-8}$ and 1000 episodes. For all three algorithms, we considered the hyper-parameter space over initial learning rates and initial temperature, as specified in table 3, if the best parameter was found at the boundary of the hyper-parameter space, we added another value to the hyper-parameter search. From these initial runs, we found the best sets of hyper-parameters for each of the algorithms, denoted on table 4. Using these hyper-parameters, we ran more extended experiments, now considering 50 independent agents. In figure 3, on the left, we present the training curves, showing the accumulated reward per episode, the shaded regions bound the mean using the 2.5th percentile and 97.5th percentile mean using bootstrap to compute the confidence interval; on the right, we present the histogram of the cumulative rewards after training. We note that all algorithms attain quite similar performance, with the entropy regularised ones (MPG and softPG) requiring more episodes to reach the same cumulative reward. This is not surprising, as the agent spends more time exploring the environment at the early stages of training because τ is larger. We observe that once trained, the testing performance of PG, MPG and nsPG is quite similar, whereas softPG has a larger spread in the cumulative reward, which appears consistent with the larger confidence intervals observed during training.

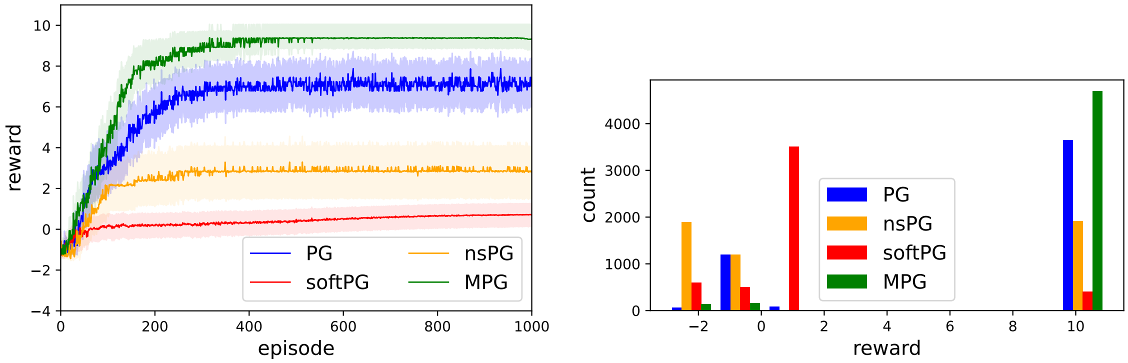

Figure 2: Frozen Lake. Left: Cumulative rewards per episode during training time when training using different RL algorithms with the best found set of hyper-parameters. Right: Cumulative rewards per episode after training, each trained agent attempts to solve the task 100 times.

Table 3: Hyper-parameters for balancing cart pole task

	Initial	PG	softPG	nsPG	MPG
η_0	$\{5 \times 10^{-5},$ $1 \times 10^{-5},$ $5 \times 10^{-6},$ $1 \times 10^{-6}\}$	$\{0.005, 0.001,$ $5 \times 10^{-4},$ $1 \times 10^{-4}\}$	$\{1 \times 10^{-4}\}$	$\{0.005, 0.001,$ $5 \times 10^{-4},$ $1 \times 10^{-4}\}$	$\{1 \times 10^{-4}\}$
τ_0	$\{0.1, 0.15,$ $0.20, 0.25,$ $0.3\}$	NA	$\{0.35, 0.4, 0.45\}$	NA	$\{\}$

5. Conclusion

In this paper, we have studied a framework combining fixed-horizon RL and max-entropy RL. We have introduced the Matryoshka Policy Gradient algorithm in the function approximation setting, with log-linear parametric policies. We proved that the global optimum of the MPG objective is unique, and that MPG converges to this global optimum, including for continuous state and action space. Furthermore, we proved that these results hold true even when the true optimal policy does not belong to the parametric space (that is

Table 4: Best set of hyper-parameters for the balancing cart pole task

	PG	softPG	nsPG	MPG
η_0	0.001	5×10^{-5}	0.001	5×10^{-5}
τ_0	NA	0.3	NA	0.3

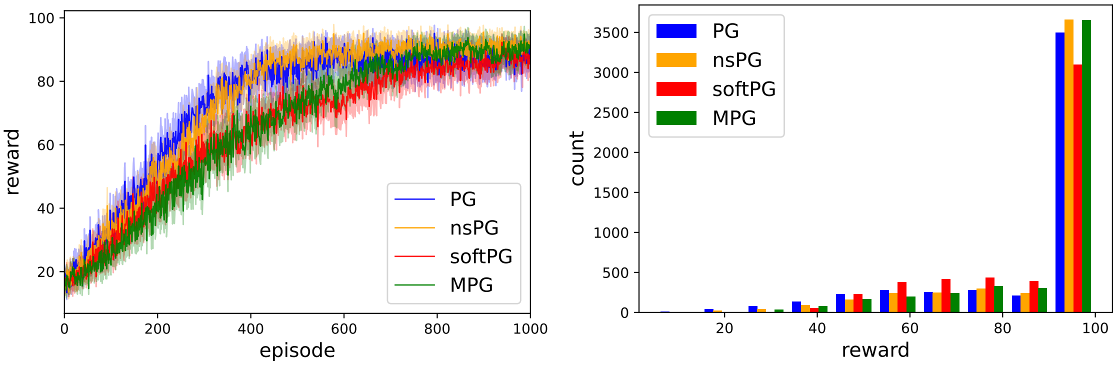

Figure 3: Cart Pole. Left: Cumulative rewards per episode during training time when training using different RL algorithms with the best found set of hyper-parameters. Right: Cumulative rewards per episode after training, each trained agent attempts to solve the task 100 times.

when Assumption **A1.** does not hold). The limit – globally optimal within the parametric space – corresponds to a projection of the optimal policy onto the parametric space. It is written as the softmax of orthogonal projections of the optimal preferences onto the RKHSs of the parametrization, with respect to the state-visitation measures induced by the policy, see (11). Finally, letting the horizon tend to infinity, the optimal policy of MPG retrieves the optimal policy of the standard infinite-horizon max-entropy objective, when the initial state distribution has full support. For neural policies, we prove that the limit is optimal within the RKHSs of the NTK (equivalently of the CK) at the end of training, and can be written in terms of orthogonal projections of optimal preferences onto these RKHSs, yielding criterion for global optimality in terms of the NTK (equivalently the CK). In particular it establishes the global convergence of neural MPG in the NTK regime. The MPG framework is intuitive, theoretically sound and it is easy to implement. Furthermore, as verified in the numerical experiments, there appears to be an slight advantage to using entropy regularisation and non-stationary policies over the compared PG algorithms. More challenging experiments will be considered in future work.

Limitations. The main limitations of our work are the following: (a) we have not studied the rate of convergence of MPG (typically more assumptions on the environment, the horizon, are needed), (b) we assumed that perfect gradient updates whereas in practice, one uses the estimate (8), (c) as a theoretical paper, our numerical experiments are rather simple. We hope to address these limitations in future work, as well a other perspectives such as:

- Additionally to MPG as defined in this paper, we expect to have nice theoretical properties of variations of MPG that are used for other PG algorithms. E.g. one can think of natural MPG, actor-critic MPG, path consistency MPG (see Nachum et al. (2017) for path consistency learning).

- We motivated the use of MPG with neural softmax policies by some theoretical, practical, and heuristic arguments; we believe that more can be said on the use of neural policies with MPG, in particular by studying the spectra of the NTK and the CK of neural networks along specific geodesics in the parametric space.
- How does the fixed-horizon max-entropy framework compares to the standard max-entropy RL framework in terms of exploration, adversarial robustness, sample efficiency, and so on?

Appendix

The appendix is organized as follows:

- A: we recall basic properties of softmax policies, then discuss the potential benefits to using a single neural network for the preferences of all *i*-step policies. This section ends with an explanation on how to approximate a kernel with finitely many features.
- B: we state and derive some basic facts on RKHS.
- C: We use concepts from Information Geometry to show that critical points of the MPG objective correspond to critical points of a *Bregman divergence*; this fact is useful when the realizable assumption does not hold to ensure that MPG converges to the unique global optimum.
- D: we prove the Matryoshka Policy Gradient Theorem (Theorem 1), Proposition 2 that shows that the infinite horizon optimal policy can approximated arbitrarily well by finite horizon optimal policies, Theorem 2 and Theorem 3 that shows global convergence of MPG.
- E: we list and discuss our main assumptions.
- F: we provide more detailed numerical experiments implementing MPG.

Appendix A. More on the parametrization

A.1 Softmax policy

As long as the map ψ is uniformly bounded, softmax policies enjoy the two following properties:

- For all $s \in \mathcal{S}$, it holds that

$$\mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(A|s)\right] = \int_{\mathcal{A}} \nabla_\theta \pi_\theta(\mathrm{d}a|s) = 0. \tag{12}$$

- It holds that $\pi_\theta(a|s) > 0$ for all $(a, s) \in \mathcal{A} \times \mathcal{S}$ such that $\overline{\pi}(a|s) > 0$.

A.2 Neural networks

Neural Tangent Kernel. For a measurable nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$, we recursively define a neural network of depth $L \geq 1$, with trainable parameters $W^\ell \in \mathbb{R}^{d_\ell} \times \mathbb{R}^{d_{\ell+1}}$ as $f : x \in \mathbb{R}^{d_0} \mapsto \widetilde{\alpha}^L(x) \in \mathbb{R}^{d_L}$, with $\alpha^0(x) := x$ and

$$\begin{aligned}\widetilde{\alpha}^{\ell+1}(x) &:= W^\ell \alpha^\ell(x),\\ \alpha^{\ell+1}(x) &:= \sigma\left(\widetilde{\alpha}^{\ell+1}(x)\right),\end{aligned}$$

where σ is applied element-wise.

Note that the connection between the last hidden layer and the output layer is linear, since $f = W^L \alpha^{L-1}$. In particular, f belongs to the RKHS of the *conjugate kernel* (CK) associated with the neural network, defined as

$$\Sigma(x, x') := \alpha^{L-1}(x) \cdot \alpha^{L-1}(x').$$

On the other hand, the training of the neural network is governed by the *neural tangent kernel* (NTK), which is defined as

$$\Theta(x, x') := \nabla f(x) \cdot \nabla f(x') = \sum_{p=1}^{P} \partial_{\theta_p} f(x) \partial_{\theta_p} f(x'),$$

where $\theta \in \mathbb{R}^P$ is the vector of all the trainable parameters of the neural network. It is important to note that both the CK and the NTK depend on the parameters and as such, move during training. Moreover, isolating the derivatives with respect to parameters W^L of the last linear layer from the others $\widetilde{\theta}$, we have that

$$\begin{aligned}\Theta(x, x') &= \alpha^{L-1}(x)\alpha^{L-1}(x') + \nabla_{\widetilde{\theta}} f(x) \nabla_{\widetilde{\theta}} f(x') \\ &= \Sigma(x, x') + K(x, x'),\end{aligned}$$

and K is another positive semidefinite kernel. We therefore have that

$$\mathcal{H}_\Sigma \subset \mathcal{H}_\Theta, \qquad \forall \theta \in \mathbb{R}^P. \tag{13}$$

Remark 1 *For infinitely wide neural networks in the NTK regime (Jacot et al., 2018) (a.k.a. lazy regime (Chizat and Bach, 2018), kernel regime), the NTK is fixed during training and is strictly positive definite, therefore convergence to the optimal policy is guaranteed.*

Non-stationary policy parametrized by a single neural network. One of the assumptions of MPG is that for any $i \neq j$, the policies $\pi^{(i)}_{\theta^{(i)}}$ and $\pi^{(j)}_{\theta^{(j)}}$ do not share parameters. Using one neural network per horizon becomes quickly costly as the maximal horizon increases. In order to avoid this issue, one can use a single neural network h_θ to parametrize all i-step policies by using i as an input such that $\pi^{(i)}_\theta(a|s) \propto \overline{\pi}(a|s) \exp(h_\theta(a, s, i)/\tau)$. By deviating from the theory, we nonetheless expect the performance of the model to be enhanced: as i grows large, the i-step optimal policy gets closer to the $i+1$-step policy. One could also use $1 - \frac{1}{i}$ as an input to the network (or any increasing map $g : \mathbb{N} \mapsto [0\,, 1]$ such that $i \mapsto g(i+1) - g(i)$ is decreasing).

A.3 Kernel methods

Suppose that Θ is a strictly pd kernel with P positive eigenvalues. Recall the linear model $a \mapsto h_\theta(a) = \theta \cdot \psi(a)$, with parameters $\theta \in \mathbb{R}^P$, such that ψ is a feature map associated with Θ. Then if $P = \infty$, one can use random features, i.e. sample $g_1, \ldots, g_{P'}$ i.i.d. Gaussian processes with covariance kernel Θ, then $h_\theta := \frac{1}{\sqrt{P'}} \sum_{i=1}^{P'} \theta_i g_i$. One can thus approximate the true kernel predictor using a finite number of features, see Jacot et al. (2020).

Another way to approximate the kernel predictor with finitely many features is to use the spectral truncated kernel $\widehat{\Theta}$ of rank $P' \in \mathbb{N}$, by cutting off the smallest eigenvalues. If

$(e_i, \lambda_i)_{i\geq 1}$ are the eigenfunction/eigenvalue pairs of Θ ranked in the non-increasing order of λ_i, one can use

$$\widehat{\Theta}(x, x') := \sum_{i=1}^{P'} \lambda_i e_i(x) e_i(x'),$$

and the predictor $h_\theta := \sum_{i=1}^{P'} \theta_i e_i$.

Appendix B. Reproducible kernel Hilbert spaces

In this section, we recall and provide some basic facts on RKHSs that we use throughout the proofs. Given some RKHS $\mathcal{H}$, we write $\mathcal{H}^\perp$ for its orthogonal complement; it is also an RKHS.

Lemma 2 *Let $\mathcal{H}_1, \mathcal{H}_2$ be two RKHSs on $\mathcal{A} \times \mathcal{S}$,*

(i) The intersection $\mathcal{H}_1 \cap \mathcal{H}_2$ is an RKHS.

(ii) For any element $f \in \mathcal{H}_1$, there exists a unique decomposition $f = g_\bullet + g_\perp$ such that $g_\bullet \in \mathcal{H}_1 \cap \mathcal{H}_2$ and $g_\perp \in \mathcal{H}_1 \cap (\mathcal{H}_2)^\perp$.

For a probability measure of the form $\mu(\mathrm{d}s)\pi(\mathrm{d}a|s)$ on $\mathcal{A} \times \mathcal{S}$, where π is a policy, and for a positive-semidefinite kernel K on $\mathcal{A} \times \mathcal{S}$, we define the integral operator $I_K(\mu, \pi)$: $L^2(\mu(\mathrm{d}s)\pi(\mathrm{d}a|s))$ by

$$I_K(f; \mu, \pi) : (a, s) \mapsto \int_{\mathcal{A}\times\mathcal{S}} \mu(\mathrm{d}s')\pi(\mathrm{d}a'|s') f(a', s') K((a, s), (a', s')).$$

Mercer's Theorem states that if $\mathcal{A} \times \mathcal{S}$ is closed (in a real space), and if K is continuous and satisfies $\int_{(\mathcal{A}\times\mathcal{S})^2} K((a, s), (a', s'))^2 \pi(\mathrm{d}a|s)\mu(\mathrm{d}s)\pi(\mathrm{d}a'|s')\mu(\mathrm{d}s') < \infty$, then there exists eigenfunction/eigenvalue pairs $(e_i, \lambda_i)_{i\geq 1}$ associated with $I_K(\mu, \pi)$, ranked in the non-increasing order of $\lambda_i \geq 0$ such that

$$K((a, s), (a', s')) = \sum_{i\geq 1} \lambda_i e_i(a, s) e_i(a', s').$$

Moreover, $\{e_i; i \geq 1\}$ is an orthonormal basis of $L^2(\mu(\mathrm{d}s)\pi(\mathrm{d}a|s))$ and the RKHS $\mathcal{H}_K$ has orthonormal basis $\{\sqrt{\lambda_i} e_i; \lambda_i > 0\}$ with respect to the RKHS inner product. We refer the reader to Minh et al. (2006) for more details.

We stress that the notion of orthogonality **depends** on the measure $\mu(\mathrm{d}s)\pi(\mathrm{d}a|s)$. Henceforth, we write $\mathcal{H}^\perp$ for the orthogonal space of the RKHS, where this measure is implicit but given by the context.

In the rest of the current section, we use the notation introduced above and assume that Mercer's Theorem applies.

Lemma 3 *Let $f \in L^2(\mu(\mathrm{d}s)\pi(\mathrm{d}a|s))$. It holds that $I_K(f; \mu, \pi)(a, s) = 0$ for all $a \in \mathcal{A}, s \in \mathcal{S}$ if and only if $f \in (\mathcal{H}_K)^\perp$.*

Proof We write

$$\begin{aligned}
&\int_{\mathcal{A}\times\mathcal{S}} \mu(\mathrm{d}s)\pi(\mathrm{d}a|s)f(a,s)I_K(f;\mu,\pi)(a,s) \\
&\qquad= \int_{\mathcal{A}\times\mathcal{S}} \mu(\mathrm{d}s)\pi(\mathrm{d}a|s)(a,s)\int_{\mathcal{A}\times\mathcal{S}} \mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')f(a,s)f(a',s')K((a,s),(a',s')) \\
&\qquad= \sum_{i\geq 1}\lambda_i\left(\int_{\mathcal{A}\times\mathcal{S}} \mu(\mathrm{d}s)\pi(\mathrm{d}a|s)f(a,s)e_i(a,s)\right)^2,
\end{aligned}$$

where we used Mercer's Theorem to write $K((a,s),(a',s')) = \sum_{i\geq 1}\lambda_i e_i(a,s)e_i(a',s')$. The claim follows. ■

Lemma 4 *It holds that*

$$\begin{aligned}
\widetilde{K}((a,s),(a',s')) := K((a,s),(a',s')) &- \int_{\mathcal{A}} K((b,s),(a',s'))\pi(\mathrm{d}b|s) \\
&- \int_{\mathcal{A}} K((a,s),(b',s'))\pi(\mathrm{d}b'|s') + \int_{\mathcal{A}^2} K((b,s),(b',s'))\pi(\mathrm{d}b|s)\pi(\mathrm{d}b'|s')
\end{aligned}$$

is a positive-semidefinite kernel. Furthermore, any map $g \in \mathcal{H}_K \cap (\mathcal{H}_{\widetilde{K}})^{\perp}$ is such that for μ-almost every $s \in \mathcal{S}$, the map $a \mapsto g(a,s)$ is constant.

Proof Let $d := \sup\{i \geq 1 : \lambda_i > 0\}$ where the λ_i's are the eigenvalues of I_K. To prove the first part of the claim, it suffices to show that for all $g \in L^2(\mu(\mathrm{d}s)\pi(\mathrm{d}a|s))$, we have

$$\int_{(\mathcal{S}\times\mathcal{A})^2} \mu(\mathrm{d}s)\pi(\mathrm{d}a|s)\mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')g(a,s)g(a',s')\widetilde{K}((a,s),(a',s')) \geq 0. \tag{14}$$

To ease the notation, for any maps $f, g \in L^2(\mu(\mathrm{d}s)\pi(\mathrm{d}a|s))$ we write

$$\begin{aligned}
\langle f, g\rangle &:= \int_{\mathcal{S}\times\mathcal{A}} \mu(\mathrm{d}s)\pi(\mathrm{d}a|s)f(a,s)g(a,s), \\
\overline{f}(s) &:= \int_{\mathcal{A}} \pi(\mathrm{d}a|s)f(a,s).
\end{aligned}$$

We now establish (14). Using that $K((a,s),(a',s')) = \sum_{j\leq d}\lambda_j e_j(a,s)e_j(a',s')$, we get

$$\widetilde{K}((a,s),(a',s')) = \sum_{j\leq d}\lambda_j(e_j(a,s) - \overline{e}_j(s))(e_j(a',s') - \overline{e}_j(s')).$$

The left-hand side of (14) thus reads as

$$\begin{aligned}
\sum_{j\leq d}\lambda_j\left(\langle g, e_j\rangle^2 - 2\langle g, e_j\rangle\langle \overline{g}, \overline{e}_j\rangle + \langle \overline{g}, \overline{e}_j\rangle^2\right) &= \sum_{j\leq d}\lambda_j\left(\alpha_j^2 - 2\alpha_j\langle \overline{g}, \overline{e}_j\rangle + \langle \overline{g}, \overline{e}_j\rangle^2\right) \\
&= \sum_{j\leq d}\lambda_j\left(\alpha_j - \langle \overline{g}, \overline{e}_j\rangle\right)^2.
\end{aligned}$$

The right-hand side above being clearly non-negative, this shows that $\widetilde{K}$ is positive-semidefinite.

We now turn our attention to the last part of the claim. Suppose that $g \in \mathcal{H}_K \cap (\mathcal{H}_{\widetilde{K}})^\perp$, so that we can write $g = \sum_{j\leq d} \alpha_j e_j$, with $\alpha_j = \langle g, e_j\rangle$. Moreover, by Lemma 3, we have an equality in (14), and we get

$$\sum_{j\leq d} \lambda_j(\langle g, e_j\rangle - \langle \overline{g}, \overline{e}_j\rangle)^2 = 0.$$

We thus necessarily have $\langle g, e_j\rangle = \langle \overline{g}, \overline{e}_j\rangle$ for all $j \leq d$. In particular,

$$\langle g, g\rangle = \sum_{j\leq d} \alpha_j^2 = \sum_{j\leq d} \alpha_j \langle \overline{g}, \overline{e}_j\rangle = \langle \overline{g}, \overline{g}\rangle.$$

On the other hand, Cauchy-Schwarz Inequality shows that if $s \mapsto g(a, s)$ is not constant in a for all s, then

$$\begin{aligned}\langle g, g\rangle &= \int_{\mathcal{S}} \mu(\mathrm{d}s) \int_{\mathcal{A}} \pi(\mathrm{d}a|s) g(a, s)^2 \\ &> \int_{\mathcal{S}} \mu(\mathrm{d}s) \left(\int_{\mathcal{A}} \pi(\mathrm{d}a|s) |g(a, s)|\right)^2 \\ &\geq \langle \overline{g}, \overline{g}\rangle.\end{aligned}$$

This is a contradiction and thus implies that g must be constant in a. ■

Since the feature maps we consider in this work are continuous, necessarily, the map g in the above is constant for μ-almost every $s \in \mathcal{S}$ if and only if it is constant for all s in the support of μ.

Appendix C. Information Geometry

The goal of this section is to show that a Pythagorean identity that is used in the forthcoming proof of Theorem 3. We use it in the case of a 1-step policy, and for a fixed state distribution, that we denote by ν in this section. Without loss of generality, we also assume to ease the notation that $\tau = 1$.

Consider the parametric space of preferences $\mathcal{H}_\Theta := \{h_\theta = \sum_{k=1}^d \theta_k \psi_k; \theta \in \mathbb{R}^d\}$, where ψ is the feature map of a positive definite kernel Θ that we assume to be continuous and bounded. The space $\mathcal{H}_\Theta$ is the RKHS associated with Θ. Fix $\vartheta \in \mathbb{R}^d$ and let π_ϑ be the 1-step policy induced by the preference h_ϑ (with baseline policy $\overline{\pi}$ as usual).

Up to reparametrization (potentially decreasing the value of d), we can assume without loss of generality that $\mathcal{H}_\Theta := \{h_\theta = \sum_{k=1}^d \theta_k \varphi_k; \theta \in \mathbb{R}^d\}$ where $\{\varphi_k; k = 1, \ldots, d\}$ is an orthonormal basis of $\mathcal{H}_\Theta$ in $L^2(\nu(\mathrm{d}s)\pi_\vartheta(\mathrm{d}a|s))$. For $k \geq 1$, define φ_k such that $\{\varphi_k; k \geq d+1\}$ is an orthonormal basis of $(\mathcal{H}_\Theta)^\perp$, the orthogonal complement of $\mathcal{H}_\Theta$ in $L^2(\nu(\mathrm{d}s)\pi_\vartheta(\mathrm{d}a|s))$.

The map

$$\begin{aligned}F : \theta &\mapsto \int_{\mathcal{S}} \nu(\mathrm{d}s) \log \int_{\mathcal{A}} \overline{\pi}(\mathrm{d}a|s) e^{h_\theta(a,s)} \\ \mathbb{R}^{d'} &\to \mathbb{R}\end{aligned}$$

is strictly convex. Indeed, it is straightforward to compute

$$\partial_{\theta_i} F(\theta) = \int_{\mathcal{S}} \nu(\mathrm{d}s) \int_{\mathcal{A}} \pi_\theta(\mathrm{d}a|s)\varphi_i(a,s),$$

and then

$$\begin{aligned}\nabla_\theta \nabla_\theta F(\theta) &= \left(\int_{\mathcal{S}} \nu(\mathrm{d}s) \int_{\mathcal{A}} \pi_\theta(\mathrm{d}a|s)\varphi_i(a,s)(\varphi_j(a,s) - \mathbb{E}_{\pi_\theta}[\varphi_j(A,s)])\right)_{i,j\leq d'} \\ &= \int_{\mathcal{S}} \nu(\mathrm{d}s)\mathrm{Var}_{\pi_\theta}\left[\varphi(A,s)\right],\end{aligned}$$

where $\mathrm{Var}_{\pi_\theta}\left[\varphi(A,s)\right]$ is the covariance matrix of $\varphi(A,s)$ for $A \sim \pi_\theta(\cdot|s)$. We thus have that

$$\theta^T \nabla_\theta \nabla_\theta F(\theta)\theta = \int_{\mathcal{S}} \nu(\mathrm{d}s) \int_{\mathcal{A}} \pi_\theta(\mathrm{d}a|s)\left(h_\theta(a,s) - \mathbb{E}_{\pi_\theta}[h_\theta(A,s)]\right)^2,$$

which is non-negative, and null if and only if $h_\theta(a,s) = \mathbb{E}_{\pi_\theta}[h(A,s)]$ for all $a \in \mathcal{A}$, that is, if and only if $\theta = 0$. This shows that $\nabla_\theta \nabla_\theta F(\theta)$ is strictly convex on $\mathbb{R}^{d'}$.

As a strictly convex map, F induces a *Bregman divergence* on the quotient space $\mathscr{P}(d')$, where $\mathscr{P}(d')$ is the set of softmax policies with preferences parametrized by $\theta \in \mathbb{R}^{d'}$. The Bregman divergence is defined as

$$\begin{aligned}D_F(\theta,\theta') :&= F(\theta) - F(\theta') - \nabla F(\theta')\cdot(\theta - \theta') \\ &= \int_{\mathcal{S}} \nu(\mathrm{d}s) D_{\mathrm{KL}}(\pi_{\theta'}||\pi_\theta)(s).\end{aligned}$$

More generally, $D_F(\pi,\pi') := \int_{\mathcal{S}} \nu(\mathrm{d}s)D_{\mathrm{KL}}(\pi'||\pi)(s)$ is well defined for any policies $\pi,\pi' \in \mathcal{P}$. One can define a dual coordinate system $\xi(\theta) := \nabla F(\theta)$, and the manifold $\mathscr{P}(d')$ is said to be *dually flat*, as each coordinate system induces a notion of flatness.

Recall that $\vartheta \in \mathbb{R}^d$ is fixed and let $\widehat{Q} \in L^2(\nu(\mathrm{d}s)\pi_\vartheta(\mathrm{d}a|s))$, with $\widehat{\pi}$ the induced policy. In particular, we can write $\widehat{Q} = \sum_{k=1}^\infty \widehat{\theta}_k\varphi_k$. For all $d' \geq 1$, let $\widehat{\pi}_{d'} \in \mathscr{P}(d')$ be the policy induced by $\widehat{h}_{d'} := \sum_{k=1}^{d'} \widehat{\theta}_k\varphi_k \in \mathcal{F}_{d'}$. In particular, by Theorem 1.5 in Amari (2016) p.27, we have that

$$\widehat{\pi}_d^{d'} := \mathrm{argmin}_{\pi_\theta\in\mathscr{P}(d)} D_F(\widehat{\pi}_{d'},\pi_\theta)$$

is unique in $\mathscr{P}(d)$, and moreover,

$$D_F(\widehat{\pi}_{d'},\pi_\vartheta) = D_F(\widehat{\pi}_{d'},\widehat{\pi}_d^{d'}) + D_F(\widehat{\pi}_d^{d'},\pi_\vartheta).$$

We now extend this identity to the infinite dimensional case, that is, with $\widehat{\pi}$ in place of $\widehat{\pi}_{d'}$.

Firstly, it is clear that $\widehat{\pi}_{d'} \to \widehat{\pi}$ as $d' \to \infty$. Since D_F is continuous, the Maximum Theorem (see p.116 of Berge (1963)) entails that

$$\widehat{\pi}_d^\infty := \lim_{d'\to\infty} \widehat{\pi}_d^{d'} = \mathrm{argmin}_{\pi_\theta\in\mathscr{P}(d)} D_F(\widehat{\pi},\pi_\theta),$$

and then

$$D_F(\widehat{\pi},\pi_\vartheta) = D_F(\widehat{\pi},\widehat{\pi}_d^\infty) + D_F(\widehat{\pi}_d^\infty,\pi_\vartheta).$$

(Alternatively, one can show the above as Equation (4) in Fukumizu (2005))

From the above equation we easily deduce the following Lemma:

Lemma 5 *With the notation introduced above, the vector $\vartheta \in \mathbb{R}^d$ is a critical point of $\theta \mapsto D_F(\widehat{\pi},\pi_\theta)$ if and only if it is a critical point of $\theta \mapsto D_F(\widehat{\pi}_d^\infty,\pi_\theta)$.*

Appendix D. Proofs

D.1 Matryoshka Policy Gradient Theorem and convergence of the objective

The following property is simple yet useful for later computations.

Lemma 6 *For all $n \geq 1$, all $\pi \in \mathcal{P}_n$ and all $s \in \mathcal{S}$, it holds that*

$$V_\pi^{(n)}(s) - V_*^{(n)}(s) = -\tau \mathbb{E}_\pi \left[\sum_{i=0}^{n-1} D_{\mathrm{KL}}(\pi^{(n-i)} || \pi_*^{(n-i)})(S_i) \middle| S_0 = s \right].$$

Proof Recall (1) and write

$$\begin{aligned} V_\pi^{(n)}(s) &= \int_{\mathcal{A}} \pi^{(n)}(\mathrm{d}a|s) \bigg(r(a,s) - \tau \log \frac{\pi^{(n)}}{\overline{\pi}}(a|s) + \int_{\mathcal{S}} p(s,a,\mathrm{d}s') V_\pi^{(n-1)}(s') \bigg) \\ &= \int_{\mathcal{A}} \pi^{(n)}(\mathrm{d}a|s) \bigg(V_*^{(n)}(s) - \tau \log \frac{\pi^{(n)}}{\pi_*}(a|s) + \int_{\mathcal{S}} p(s,a,\mathrm{d}s') (V_\pi^{(n-1)}(s') - V_*^{(n-1)}(s')) \bigg), \end{aligned}$$

where we plugged in the expression of the optimal policy (5). We can rewrite the above as

$$V_\pi^{(n)}(s) - V_*^{(n)}(s) = -\tau D_{\mathrm{KL}}(\pi^{(n)} || \pi_*^{(n)}) + \mathbb{E}_\pi \left[V_\pi^{(n-1)}(S_1) - V_*^{(n-1)}(S_1) \middle| S_0 = s \right].$$

The claim follows by induction. ■

The next lemma provides bounds that are needed to take derivatives of the objective for the proof of Theorem 1 later on.

Lemma 7 *For all $a \in \mathcal{A}$, $s \in \mathcal{S}$, $\theta, \theta' \in \mathbb{R}^P$ and all $i \in \{1, \ldots, n\}$, it holds that*

(i) $\|\nabla_\theta \pi_\theta^{(i)}(a|s)\|_2 \leq \frac{2}{\tau} \|\psi\| \pi_\theta^{(i)}(a|s)$*;*

(ii) $\|\nabla_\theta^2 \log \pi_\theta^{(i)}(a|s)\|_2 \leq \frac{2P_i}{\tau^2} \|\psi\|^2$*;*

(iii) $\int_{\mathcal{A}} \pi_\theta^{(i)}(\mathrm{d}a|s) \left| \log \frac{\pi_\theta^{(i)}}{\pi_*^{(i)}}(a|s) \right| \leq D_{\mathrm{KL}}(\pi_\theta^{(i)} || \pi_*^{(i)})(s) + \frac{3^n}{\tau} \|r\|_\infty$*;*

(iv) Let $f_\theta : \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ be differentiable with respect to $\theta \in \mathbb{R}^P$ such that $\|f\|_\infty \leq C$ and $\|\nabla_\theta f_\theta(a,s)\|_2 \leq C(1 + \|\theta\|_2)$ for all $(a,s) \in \mathcal{A} \times \mathcal{S}$ for some constant $C > 0$, then it holds that

$$\nabla_\theta \mathbb{E}_{\pi_\theta} [f_\theta(A_i, S_i)] = \mathbb{E}_{\pi_\theta} \left[\bigg(\sum_{j=0}^{i} \nabla_\theta \log \pi_\theta^{(n-j)}(A_j|S_j) \bigg) f_\theta(A_i, S_i) \right] + \mathbb{E}_{\pi_\theta} [\nabla_\theta f_\theta(A_i, S_i)].$$

Proof (*i*) Recall (7). We have $\nabla_\theta \pi_\theta^{(i)}(a|s) = \pi_\theta^{(i)}(a|s)(\psi^{(i)}(a,s) - \mathbb{E}_{\pi_\theta}[\psi^{(i)}(A,s)])/\tau$. In particular, $\|\nabla_\theta \pi_\theta^{(i)}(a|s)\|_2 \leq 2\pi_\theta^{(i)}(a|s)\|\psi\|/\tau$, where we recall that $\|\psi\| = \sup_{a,s,i} \|\psi^{(i)}(a,s)\|_2$. This proves (*i*).

(ii) We compute

$$\begin{aligned}\nabla^2_\theta \log \pi^{(i)}_\theta(a|s) &= \nabla_\theta\big(\psi^{(i)}(a, s - \mathbb{E}_{\pi_\theta}[\psi^{(i)}(A,s)])\big)/\tau \\ &= -\frac{1}{\tau^2}\mathbb{E}_{\pi^{(i)}_\theta}\Big[\psi^{(i)}(A,s)\big(\psi^{(i)}(A,s) - \mathbb{E}_{\pi_\theta}[\psi^{(i)}(A',s)]\big)^T\Big].\end{aligned}$$

Since the 2-norm of a matrix is upper-bounded by its Frobenius norm, the above entails that $\|\nabla^2_\theta \log \pi_\theta(a|s)\|_2^2 \le \sum_{k,\ell\le P}(\nabla^2_\theta \log \pi_\theta(a|s))^2_{k,\ell} \le \frac{4P^2}{\tau^4}\|\psi\|^4$, which proves *(ii)*.

(iii) Let $i \in \{1,\dots,n\}$. We note that

$$\int_{\mathcal{A}} \pi^{(i)}_\theta(\mathrm{d}a|s)\left|\log\frac{\pi^{(i)}_\theta}{\pi^{(i)}_*}(a|s)\right| = D_{\mathrm{KL}}(\pi^{(i)}_\theta||\pi^{(i)}_*)(s) + 2\int_{\mathcal{A}}\pi^{(i)}_\theta(\mathrm{d}a|s)\log\frac{\pi^{(i)}_*}{\pi^{(i)}_\theta}(a|s)\mathbb{1}_{\{\pi^{(i)}_\theta(a|s)<\pi^{(i)}_*(a|s)\}}.$$

We claim that for all $\theta \in \mathbb{R}^P$ and all $s \in \mathcal{S}$, it holds that

$$\int_{\mathcal{A}}\pi^{(i)}_\theta(\mathrm{d}a|s)\log\frac{\pi^{(i)}_*}{\pi^{(i)}_\theta}(a|s)\mathbb{1}_{\{\pi^{(i)}_*(a|s)>\pi^{(i)}_\theta(a|s)\}} \le e^{-1} + \frac{3^n}{2\tau}||r||_\infty. \tag{15}$$

To lighten the notation, let us keep the variables a and s implicit in the calculations. To establish (15), we write by definition that

$$\begin{aligned} 0 &\le \int_{\mathcal{A}}\mathrm{d}\pi^{(i)}_\theta \log\frac{\pi^{(i)}_*}{\pi^{(i)}_\theta}\mathbb{1}_{\{\pi^{(i)}_*>\pi^{(i)}_\theta\}} \\ &= \frac{1}{\tau}\int_{\mathcal{A}}\mathrm{d}\pi^{(i)}_\theta\left(Q^{(i)}_* - V^{(i)}_*\right)\mathbb{1}_{\{\pi^{(i)}_*>\pi^{(i)}_\theta\}} \\ &\qquad\qquad - \int_{\mathcal{A}}\mathrm{d}\overline{\pi}e^{h^{(i)}_\theta/\tau - \log\mathbb{E}_{\overline{\pi}}[e^{h^{(i)}_\theta/\tau}]}\left(\frac{1}{\tau}h^{(i)}_\theta - \log\mathbb{E}_{\overline{\pi}}[e^{h^{(i)}_\theta/\tau}]\right)\mathbb{1}_{\{\pi^{(i)}_*>\pi^{(i)}_\theta\}} \\ &\le \frac{||Q^{(i)}_* - V^{(i)}_*||_\infty}{\tau} - \int_{\mathcal{A}}\mathrm{d}\overline{\pi}e^{h^{(i)}_\theta/\tau - \log\mathbb{E}_{\overline{\pi}}[e^{h^{(i)}_\theta/\tau}]}\left(\frac{1}{\tau}h^{(i)}_\theta - \log\mathbb{E}_{\overline{\pi}}[e^{h^{(i)}_\theta/\tau}]\right)\mathbb{1}_{\{\pi^{(i)}_*>\pi^{(i)}_\theta\}}.\end{aligned}$$

To lower bound the second term, note that the integral is of the form $\int_{\mathcal{A}}\mathrm{d}\overline{\pi}e^{f_\theta}f_\theta$, and one can easily check that $xe^x \ge e^{-1}$ for all $x \in \mathbb{R}$, so that $\int_{\mathcal{A}}\mathrm{d}\overline{\pi}e^{f_\theta}f_\theta \ge e^{-1}$. We thus have proved that

$$\int_{\mathcal{A}}\mathrm{d}\pi^{(i)}_\theta\log\frac{\pi^{(i)}_*}{\pi^{(i)}_\theta}\mathbb{1}_{\{\pi^{(i)}_*>\pi^{(i)}_\theta\}} \le e^{-1} + \frac{||Q^{(i)}_* - V^{(i)}_*||_\infty}{\tau}.$$

We now bound $||Q^{(i)}_* - V^{(i)}_*||_\infty$. Note that $V^{(1)}_*(s) = \tau\log\mathbb{E}_{\overline{\pi}}[e^{r(A,s)/\tau}]$, so that $||V^{(1)}_*||_\infty \le ||r||_\infty$, and then,

$$||Q^{(1)}_*||_\infty, ||V^{(1)}_*||_\infty \le ||r||_\infty.$$

We reason by induction. We have

$$\begin{aligned}\left|Q_*^{(i+1)}(a,s)-V_*^{(i+1)}(s)\right| &= \left|r(a,s)+\int_{\mathcal{S}}p(s,a,\mathrm{d}s')V_*^{(i)}(s')-\tau D_{\mathrm{KL}}(\pi_*^{(i+1)}||\overline{\pi})(s)-V_*^{(i+1)}(s)\right|\\ &\leq \left|r(a,s)+\int_{\mathcal{S}}p(s,a,\mathrm{d}s')V_*^{(i)}(s')+\mathbb{E}_{\pi_*}[Q_*^{(i+1)}(A,s)]\right|\\ &\leq ||r||_\infty+||V_*^{(i)}||_\infty+||Q_*^{(i+1)}||_\infty\\ &\leq 2||r||_\infty+2||V_*^{(i)}||_\infty.\end{aligned}$$

In particular, $||V_*^{(i+1)}||_\infty \leq ||Q_*^{(i+1)}||_\infty + 2||r||_\infty + 2||V_*^{(i)}||_\infty \leq 3||r||_\infty + 3||V_*^{(i)}||_\infty$. By induction, we get for all $i=1,\dots,n$ that

$$\begin{aligned}||V_*^{(i)}||_\infty &\leq \sum_{j=1}^{i-1}3^j r = \frac{3^{i-1}}{2}||r||_\infty \leq \frac{3^{n-1}-1}{2}||r||_\infty,\\ ||Q_*^{(i)}||_\infty &\leq ||r||_\infty+||V_*^{(i-1)}||_\infty \leq \frac{3^{n-2}+1}{2}||r||_\infty.\end{aligned}\tag{16}$$

This proves (15), which in turns proves (*iii*).

(*iv*) Let f_θ satisfy the conditions of the statement. For all $i\in\{0,\dots,n-1\}$, the state distribution satisfies

$$\begin{aligned}\mathbf{m}_{\pi_\theta}^{(n-i)}(\mathrm{d}s) &= \int_{\mathcal{S}}\mathbf{m}_{\pi_\theta}^{(n-i+1)}(\mathrm{d}s')\int_{\mathcal{A}}\pi_\theta^{(n-i+1)}(\mathrm{d}a|s')p(s',a,\mathrm{d}s)\\ &= \int_{\mathcal{S}}\nu(\mathrm{d}s_0)\int_{\mathcal{A}}\pi_\theta^{(n)}(\mathrm{d}a_0|s_0)\int_{\mathcal{S}}p(s_0,a_0,\mathrm{d}s_1)\dots\\ &\qquad\dots\times\int_{\mathcal{A}}\pi_\theta^{(n-i+1)}(\mathrm{d}a_{i-1}|s_{i-1})\int_{\mathcal{S}}p(s_{i-1},a_{i-1},\mathrm{d}s).\end{aligned}$$

Note that $\pi_\theta^{(i)}=\overline{\pi}\frac{e^{h_\theta^{(i)}/\tau}}{\mathbb{E}_{\overline{\pi}}[e^{h_\theta^{(i)}/\tau}]}$, with $\|h_\theta^{(i)}\|_\infty \leq \sup_{a,s}\|\theta^{(i)}\|_2\|\psi^{(i)}(a,s)\|_2 < \|\psi\|$, which entails that for all $\theta\in\mathbb{R}^P$, being integrable with respect to $\mathbf{m}_{\pi_\theta}^{(i)}(\mathrm{d}s)\pi_\theta^{(i)}(\mathrm{d}a|s)$ is equivalent to being integrable with respect to $\mathbf{m}_{\overline{\pi}}^{(i)}(\mathrm{d}s)\overline{\pi}^{(i)}(\mathrm{d}a|s)$. On the other hand, for any $s_0,a_0,\dots,s_i,a_i$, we have that

$$\nabla_\theta\prod_{\ell=0}^{i}\pi_\theta^{(n-\ell)}(a_\ell|s_\ell)=\left(\prod_{j=0}^{i}\pi_\theta^{(n-j)}(a_j|s_j)\right)\sum_{j=0}^{i}\nabla_\theta\log\pi_\theta^{(n-j)}(a_j|s_j)$$

Since $\|\log\pi_\theta^{(n-j)}(a_j|s_j)\|_2 \leq \frac{2}{\tau}\|\psi\|$ by (i), from Measure Theory, we know that we can differentiate inside the integral and write

$$
\begin{aligned}
&\nabla_\theta \int_{\mathcal{S}} \mathbf{m}_{\pi_\theta}^{(\ell)}(\mathrm{d}s) \int_{\mathcal{A}} \pi_\theta^{(\ell)}(\mathrm{d}a|s) f_\theta(a,s) \\
&= \int\cdots\int \nu(\mathrm{d}s_0) \left(\prod_{i=0}^{\ell} \pi_\theta^{(n-i)}(\mathrm{d}a_i|s_i) p(s_i, a_i \mathrm{d}s_{i+1}) \right) \left(\sum_{j=0}^{\ell} \nabla_\theta \log \pi_\theta^{(n-j)}(a_j|s_j) \right) f_\theta(a_\ell, s_\ell) \\
&\qquad + \int_{\mathcal{S}} \mathbf{m}_{\pi_\theta}^{(\ell)}(\mathrm{d}s) \int_{\mathcal{A}} \pi_\theta^{(\ell)}(\mathrm{d}a|s) \nabla_\theta f_\theta(a,s),
\end{aligned}
$$

which is equivalent to the claim and thus concludes the proof. ■

We show below that in expectation, the MPG update (8) is proportional to the gradient of the objective. It is the only statement where we do not assume perfect gradient update; everywhere else, we assume that $\theta_{t+1} = \theta_t + \eta\nabla_\theta J_n(\pi_t)$.

Lemma 8 *For θ_t constructed as in* (8)*, it holds that* $\mathbb{E}[\theta_{t+1} - \theta_t] \propto \nabla_\theta J_n(\pi_t)$.

Proof Recall that $\|r\|_\infty < \infty$ and note that

$$
\begin{aligned}
\left|\log \frac{\pi_t^{(i)}}{\overline{\pi}}(a|s)\right| &\leq \left|h_\theta^{(i)}(a,s)/\tau - \log \mathbb{E}_{\overline{\pi}}[e^{h_\theta^{(i)}(A,s)/\tau}]\right| \\
&\leq 2\|\theta\|_2\|\psi\|,
\end{aligned}
$$

where we recall that $\|\psi\| = \sup_{a,s,i}\|\psi^{(i)}(a,s)\|_2$. Hence, Lemma 7$(iv)$ applies in the computations below.

Let $\mathbf{m}_\pi^{(i)}$ denote the law of S_{n-i}, that is, the $(n-i)$-th visited state under π. The distribution of the sequence $S_0, A_0, \ldots, A_{n-i-1}, S_{n-i}$ is not influenced by the parameters $\theta^{(i)}$, thus we can write

$$
\begin{aligned}
\nabla_{\theta^{(i)}} J_n(\pi_t) &= \int_{\mathcal{S}} \nabla_{\theta^{(i)}} V_{\pi_t}^{(n)}(s) \nu_0(\mathrm{d}s) \\
&= \nabla_{\theta^{(i)}} \Bigg(\mathbb{E}_{\pi_t}\Bigg[\sum_{\ell=0}^{n-i-1} R_\ell - \tau \log \frac{\pi_t^{(n-\ell)}}{\overline{\pi}}(A_\ell|S_\ell) \Bigg] \\
&\qquad\qquad + \mathbb{E}_{S_{n-i}\sim \mathbf{m}_{\pi_t}^{(i)}} \Bigg[\mathbb{E}_{\pi_t}\Bigg[\sum_{\ell=n-i}^{n} R_\ell - \tau \log \frac{\pi_t^{(n-\ell)}}{\overline{\pi}}(A_\ell|S_\ell) \Bigg| S_{n-i} \Bigg]\Bigg]\Bigg) \\
&= 0 + \mathbb{E}_{S_{n-i}\sim \mathbf{m}_{\pi_t}^{(i)}} \Big[\nabla_{\theta^{(i)}} V_{\pi_t}^{(i)}(S_{n-i}) \Big],
\end{aligned}
$$

where we have used the Markov property. We then have that

$$\begin{aligned}
\nabla_{\theta^{(i)}} V_{\pi_t}^{(i)}(S_{n-i}) &= \nabla_{\theta^{(i)}} \mathbb{E}_{T_{n,i}(\pi_t)} \Big[\sum_{\ell=n-i}^{n} R_\ell - \tau \log \frac{\pi_t^{(n-\ell)}}{\overline{\pi}}(A_\ell|S_\ell) \Big| S_{n-i} \Big] \\
&= \mathbb{E}_{\pi_t} \Big[\Big(\sum_{\ell=n-i}^{n} \Big(R_\ell - \tau \log \frac{\pi_t^{(n-\ell)}}{\overline{\pi}}(A_\ell|S_\ell) \Big) - \tau \Big) \nabla \log \pi_t^{(i)}(A_{n-i}|S_{n-i}) \Big| S_{n-i} \Big] \\
&= \mathbb{E}_{\pi_t} \Big[\sum_{\ell=n-i}^{n} \Big(R_\ell - \tau \log \frac{\pi_t^{(n-\ell)}}{\overline{\pi}}(A_\ell|S_\ell) \Big) \nabla \log \pi_t^{(i)}(A_{n-i}|S_{n-i}) \Big| S_{n-i} \Big],
\end{aligned}$$

where we have used (12) to get rid of τ.

Recalling the MPG update (8), we thus have proved that $\mathbb{E}[\theta_{t+1} - \theta_t] = \eta \nabla_\theta J_n(\pi_t)$. ■

Proof (*Theorem 1*) The strategy is to show that $\theta \mapsto \nabla_\theta J_n(\pi_\theta)$ is Lipschitz by bounding the 2-norm of the Hessian of J_n along the training trajectory. It is standard in Optimisation that this implies that for η smaller than 2 over the Lipschitz constant, the objective is monotonically increasing during gradient ascent, to finally deduce the convergence of $J_n(\pi_t)$ as $t \to \infty$.

Recall that by Lemma 7(i), for all $i \in \{0, \dots, n-1\}$, for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, we have

$$\|\nabla_\theta \log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(a|s)\|_2 \le \frac{2}{\tau}\|\psi\|.$$

Moreover, for all $(a, s) \in \mathcal{A} \times \mathcal{S}$,

$$\begin{aligned}
\Big| \log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(a|s) \Big| &\le \big| h_\theta^{((n-i))}(a,s)/\tau - \log \mathbb{E}_{\overline{\pi}}[e^{h_\theta^{(n-i)}(A,s)/\tau}] \big| + \frac{1}{\tau}\|Q_*^{(n-i)} - V_*^{(n-i)}\|_\infty \\
&\le \frac{1}{\tau}\left(\|\theta\|_2 \|\psi\| + 3^n \|r\|_\infty \right), \qquad (17)
\end{aligned}$$

where we used (16). Hence, we can apply Lemma 7(iv) to differentiate $J_n(\pi_\theta)$, thus obtaining

$$\begin{aligned}
\nabla_\theta J_n(\pi_\theta) &= -\tau \sum_{i=0}^{n-1} \nabla_\theta \mathbb{E}_{\pi_\theta} \left[\log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i) \right] \\
&= -\tau \sum_{i=0}^{n-1} \mathbb{E}_{\pi_\theta} \Big[\Big(\sum_{j=0}^{i} \nabla_\theta \log \pi_\theta^{(n-j)}(A_j|S_j) \Big) \log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i) \Big] \\
&\qquad\qquad + \mathbb{E}_{\pi_\theta} \left[\nabla_\theta \log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i) \right] \\
&= -\tau \sum_{i=0}^{n-1} \sum_{j=0}^{i} \mathbb{E}_{\pi_\theta} \Big[\nabla_\theta \log \pi_\theta^{(n-j)}(A_j|S_j) \log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i) \Big], \qquad (18)
\end{aligned}$$

where we used that $\mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta^{(n-i)}(A_i|S_i)] = 0$. Exchanging the order of summation and focusing on the components of the gradient $\nabla_{\theta^{(n-j)}}$ for $j \in \{0, \dots, n-1\}$ fixed, we get

$$\nabla_{\theta^{(n-j)}} J_n(\pi_\theta) = -\tau \sum_{i=j}^{n-1} \mathbb{E}_{\pi_\theta}\Big[\nabla_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(A_j|S_j) \log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i)\Big]. \tag{19}$$

To compute the Hessian of $J_n(\pi_\theta)$, we need to differentiate once more the expectation. Fix $j \geq j' \in \{0, \dots, n-1\}$, we compute the components of the Hessian of the form $\nabla_{\theta^{(n-j')}} \nabla_{\theta^{(n-j)}} J_n(\pi_\theta)$. For $j' < j$, the terms inside the expectation do not depend on $\theta^{(n-j')}$, so that Lemma 7(iv) trivially applies and yields

$$\begin{aligned}&\nabla^2_{\theta^{(n-j')},\theta^{(n-j)}} J_n(\pi_\theta)\\ &\quad= -\tau \sum_{i=j}^{n-1} \mathbb{E}_{\pi_\theta}\Big[\nabla_{\theta^{(n-j')}} \log \pi_\theta^{(n-j')}(A_{j'}|S_{j'})(\nabla_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(A_j|S_j))^T \log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i)\Big].\end{aligned}$$

Lemma 7 shows that

$$\begin{aligned}\|\nabla^2_{\theta^{(n-j')},\theta^{(n-j)}} J_n(\pi_\theta)\|_2 &\leq \sum_{i=j}^{n-1} \frac{4\|\psi\|^2}{\tau} \mathbb{E}_{\pi_\theta}\Big[\Big|\log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i)\Big|\Big]\\ &\leq n\frac{4\|\psi\|^2}{\tau^2}(J_n(\pi_*) - J_n(\pi_\theta) + 3^n\|r\|_\infty),\end{aligned}$$

where we used Lemma 6 to bound the expectation of the Kullback-Leibler divergences by the performance gap. For $j' = j$, to apply Lemma 7(iv), we check that the gradient of the terms in the expectation of (19) is bounded, that is,

$$\begin{aligned}(a,s) \mapsto \nabla^2_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(a|s) \mathbb{E}_{\pi_\theta}\Big[\log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i)\Big| A_j = a, S_j = s\Big]\\ + \nabla_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(a|s)(\nabla_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(a|s))^T.\end{aligned}$$

By (17) and Lemma 7(i) and (ii), each coordinate of the above matrix is bounded by $C\|\theta\|_2$ for some constant $C > 0$. Hence Lemma 7(iv) applies and the Hessian for $j = j'$ has the additional terms

$$\begin{aligned}\sum_{i=j}^{n-1} \mathbb{E}_{\pi_\theta}\Big[\nabla^2_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(A_j|S_j) \log \frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i)\Big]\\ + \mathbb{E}_{\pi_\theta}\Big[\nabla_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(A_j|S_j)(\nabla_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(A_j|S_j))^T\Big]\end{aligned}$$

We get by Lemma 7

$$\begin{aligned}
\|\nabla^2_{\theta^{(n-j)},\theta^{(n-j)}} J_n(\pi_\theta)\|_2 &\leq 4n\frac{\|\psi\|^2}{\tau^2}(J_n(\pi_*) - J_n(\pi_\theta) + 3^n\|r\|_\infty) \\
&\quad + \sum_{i=j}^{n-1}\Bigg(\mathbb{E}\left[\|\nabla^2_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(A_j|S_j)\|_2 \Bigg| \log\frac{\pi_\theta^{(n-i)}}{\pi_*^{(n-i)}}(A_i|S_i)\Bigg|\right] \\
&\quad + \mathbb{E}_{\pi_\theta}\left[\left\|\nabla_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(A_j|S_j)(\nabla_{\theta^{(n-j)}} \log \pi_\theta^{(n-j)}(A_j|S_j))^T\right\|_2\right]\Bigg) \\
&\leq 4n\frac{\|\psi\|^2}{\tau^2}(J_n(\pi_*) - J_n(\pi_\theta) + 3^n\|r\|_\infty) \\
&\quad + 2n\frac{P}{\tau^3}\|\psi\|^2(J_n(\pi_*) - J_n(\pi_\theta) + 3^n\|r\|_\infty) + 4n\frac{\|\psi\|^2}{\tau^2} \\
&= 2n\frac{\|\psi\|^2}{\tau^2}\Big(2 + \frac{P}{\tau}\Big)(J_n(\pi_*) - J_n(\pi_\theta) + 3^n\|r\|_\infty) + 4n\frac{\|\psi\|^2}{\tau^2}.
\end{aligned}$$

Finally, to obtain a bound on $\|\nabla^2_\theta J_n(\pi_\theta)\|_2$, we only need to sum over $j, j' \in \{0, \dots, n-1\}$. This yields

$$\begin{aligned}
\nabla^2_\theta J_n(\pi_\theta)\|_2 &\leq 2n^2\frac{\|\psi\|^2}{\tau^2}\Big(2 + \frac{P}{\tau}\Big)(J_n(\pi_*) - J_n(\pi_\theta) + 3^n\|r\|_\infty) + 4n^2\frac{\|\psi\|^2}{\tau^2} \\
&\quad + (n-1)4n^2\frac{\|\psi\|^2}{\tau^2}(J_n(\pi_*) - J_n(\pi_\theta) + 3^n\|r\|_\infty) \\
&\leq 4(n^2 + n^3)\Big(2 + \frac{P}{\tau}\Big)\frac{\|\psi\|^2}{\tau^2}(J_n(\pi_*) - J_n(\pi_\theta) + 3^n\|r\|_\infty) + 4n^2\frac{\|\psi\|^2}{\tau^2} \\
&= L(\theta).
\end{aligned}$$

We thus have shown that $\nabla_\theta J_n(\pi_\theta)$ is locally Lipschitz with constant $L(\theta)$. Since $\theta \mapsto L(\theta)$ is monotonically decreasing as $J_n(\pi_*) - J_n(\pi_\theta)$ decreases, if $\eta < 2/L(\theta_0)$ at the start of training, as explained at the beginning of the proof, it implies by induction that $\eta < 2/L(\theta_t)$ for all $t \geq 0$, which entails the claim and concludes the proof. ■

D.2 On the optimal policy

Proof (*Lemma 1*) By definition, we write

$$\begin{aligned}
V_*^{(n)}(s) &= \tau \int_{\mathcal{A}} \pi_*^{(n)}(\mathrm{d}a|s)\left(Q_*^{(n)}(a,s) - \tau\log\frac{\pi_*^{(n)}}{\overline{\pi}}(a|s)\right) \\
&= \tau \log \mathbb{E}_{\overline{\pi}}\left[\exp(Q_*^{(n)}(A,s)/\tau)\right] \int_{\mathcal{A}} \overline{\pi}(\mathrm{d}a|s)\frac{\exp\left(Q_*^{(n)}(a,s)/\tau\right)}{\mathbb{E}_{\overline{\pi}}\left[\exp\left(Q_*^{(n)}(A,s)/\tau\right)\right]} \\
&= \tau \log \mathbb{E}_{\overline{\pi}}\left[\exp\left(Q_*^{(n)}(A,s)/\tau\right)\right],
\end{aligned}$$

as claimed, which concludes the proof. ■

Proof (*Proposition 2*) Assume that ν has full support.

(i) Let $\pi \in \mathcal{P}$ be any standard policy, and let $\pi_n = (\pi, \dots, \pi) \in \mathcal{P}_n$. By definition of the standard infinite-horizon discounted objective J_∞, using the dominated convergence theorem (rewards are bounded), we have that $J_n(\pi_n) \to J_\infty(\pi)$. In particular, we get that $\pi_{*,n}^{(n)}$ achieves a performance arbitrarily close to that of $\pi_{*,\infty}$ in the infinite horizon discounted setting, and since the optimal policy of J_∞ is unique (ν-almost everywhere), we deduce that $\pi_{*,n}^{(n)} \to \pi_{*,\infty}$ as $n \to \infty$.

(ii) Suppose that $J_1(\pi_{*,1}) > J_1(T_{n,1}(\pi_{*,n}))$, that is

$$\int_{\mathcal{S}} V_{\pi_{*,1}}^{(1)}(s)\nu(\mathrm{d}s) > \int_{\mathcal{S}} V_{\pi_{*,n}}^{(1)}(s)\nu(\mathrm{d}s).$$

In particular, the set $\widetilde{\mathcal{S}} := \{s \in \mathcal{S} : \; V_{\pi_{*,1}}^{(1)}(s) > V_{\pi_{*,n}}^{(1)}(s)\}$ is non-empty and $\nu(\widetilde{\mathcal{S}})$. Furthermore, by optimality, $s \in \mathcal{S} \setminus \widetilde{\mathcal{S}}$ if and only if $V_{\pi_{*,1}}^{(1)}(s) = V_{\pi_{*,n}}^{(1)}(s)$. Let $\widetilde{\pi}_{*,n} \in \mathcal{P}_n$ be identical to $\pi_{*,n}$ except for the 1-step policy where $\pi_{*,n}^{(1)}$ is replaced by $\pi_{*,1}$. Then, the recursive structure of the value function (1) entails that $J_n(\widetilde{\pi}_{*,n}) > J_n(\pi_{*,n})$, which is a contradiction. Therefore, $T_{n,1}(\pi_{*,n}) = \pi_{*,1}$.

Then, by induction and using the recursive structure of the value function, the same argument shows that $T_{n,m}(\pi_{*,n}) = \pi_{*,m}$ for all $m = 2, \dots, n-1$, which concludes the proof. ■

D.3 On the convergence of training

By Theorem 1, we know that $J_n(\pi_t)$ converges monotonically as $t \to \infty$. However, this does not ensure that π_t converges, and a fortiori that θ_t converges. In this section, we prove two results of importance in establishing the global convergences of Theorem 2 and Theorem 3

Below, we show in Lemma 9 that the sequence of policies visited during training is *relatively compact*, that is, any of its subsequences admits a *weakly converging* subsequence. A sequence of measure $(\mu_k)_{k\geq 0}$ is said to converge weakly if and only if $\lim_{k\to\infty} \int f \mathrm{d}\mu_k = \int f \mathrm{d}\mu$ for every continuous bounded map f.

Then, we show in Lemma 10 that the parameters of a converging subsequence $(\pi_{t_k})_{k\geq 0}$ remain uniformly bounded, which implies that any limit of the parameters θ_{t_k} belongs to $\mathbb{R}^P$. In particular, the subsequences of policies converging weakly during training actually have their parameters converging inside $\mathbb{R}^P$, so that the convergence is in the stronger sense of (9).

Lemma 9 *Under the assumptions of Theorem 1, for all $i = 1, \dots, n$, the sequence of probability measures $(\mathbf{m}_{\pi_t}^{(i)}(\mathrm{d}s)\pi_t^{(i)}(\mathrm{d}a))_{t\geq 0}$ on $\mathcal{S} \times \mathcal{A}$ is relatively compact.*

Proof By Prohorov's theorem (Theorem 5.1 in Billingsley (2013)), it suffices to show that $(\mathbf{m}_{\pi_t}^{(i)}(\mathrm{d}s)\pi_t^{(i)}(\mathrm{d}a))_{t\geq 0}$ is *tight* for all $i = 1, \dots, n$. We say that a sequence of probability measures μ_t is tight if and only if for all $\epsilon > 0$, there exists a compact set K_ϵ such that $\mu_t(K_\epsilon) > 1 - \epsilon$. Roughly speaking, this ensures that no mass escapes at infinity.

Starting with $i = n$, we first show that for every $\epsilon > 0$, there exists a compact set $K_\epsilon \subset \mathcal{S} \times \mathcal{A}$ such that $\limsup_{t\to\infty} \int_{K_\epsilon} \mathrm{d}\nu \mathrm{d}\pi_t^{(n)} > 1 - \epsilon$. By contradiction, suppose that this is not the case, then there exists $\epsilon > 0$ such that $\limsup_{t\to\infty} \int_{K^c} \mathrm{d}\nu \mathrm{d}\pi_t^{(n)} \geq \epsilon$ for all compact $K \subset \mathcal{S} \times \mathcal{A}$, where K^c is the complement of K. Let $\delta > 0$ be arbitrarily smaller that ϵ and consider a compact $K_\delta \subset \mathcal{S} \times \mathcal{A}$ such that $\int_{K_\delta^c} \mathrm{d}\nu \mathrm{d}\overline{\pi} < \delta$, then necessarily, we have

$$\begin{aligned}\limsup_{t\to\infty} \int_{K_\epsilon^c} \nu(\mathrm{d}s)\pi_t^{(n)}(\mathrm{d}a|s) \log \frac{\pi_t^{(n)}}{\overline{\pi}}(a|s) &= -\limsup_{t\to\infty} \int_{K_\epsilon^c} \nu(\mathrm{d}s)\pi_t^{(n)}(\mathrm{d}a|s) \log \frac{\overline{\pi}}{\pi_t^{(n)}}(a|s) \\ &\geq -\limsup_{t\to\infty} \log \frac{\int_{K_\epsilon^c} \nu(\mathrm{d}s)\overline{\pi}(\mathrm{d}a|s)}{\int_{K_\epsilon^c} \nu(\mathrm{d}s)\pi_t^{(n)}(\mathrm{d}a|s)} \\ &\geq -\log \frac{\delta}{\epsilon},\end{aligned}$$

where we used Jensen's inequality by concavity of the logarithm. Since $\delta > 0$ is arbitrary, this shows that $\limsup_{t\to\infty} \int_{\mathcal{S}} D_{\mathrm{KL}}(\pi_t^{(n)}||\overline{\pi})(s)\nu(\mathrm{d}s) = \infty$, which contradicts the fact that $J_n(\pi_t) = \sum_{i=0}^n \mathbb{E}_{\pi_t}[R_i - D_{\mathrm{KL}}(\pi_t^{(n-i)}||\overline{\pi})(S_i)]$ converges to a finite value.

Hence $(\nu(\mathrm{d}s)\pi_t^{(n)}(\mathrm{d}a))_{t\geq 0}$ is tight and then relatively compact.

To end the proof, we reason by induction as follows: let $1 < i \leq n$ and consider a subsequence $(\pi_{t_k})_{k\geq 0}$ such that $\mathbf{m}_{\pi_t}^{(j)}(\mathrm{d}s)\pi_{t_k}^{(j)}(\mathrm{d}a)$ converges weakly toward $\mathbf{m}_{\pi_\infty}^{(j)}(\mathrm{d}s)\pi_\infty^{(j)}(\mathrm{d}a)$ for all $j = i, \ldots, n$, for some policy π_∞. Note that $\mathbf{m}_{\pi_{t_k}}^{(i-1)}$ only depends on $\pi_{t_k}^{(j)}$ for $j \in \{i, \ldots, n\}$, so that for any Borel subset $B \subset \mathcal{S}$,

$$\begin{aligned}\mathbf{m}_{\pi_{t_k}}^{(i-1)}(B) &= \int_{\mathcal{S}} \mathbf{m}_{\pi_{t_k}}^{(i)}(\mathrm{d}s') \int_{\mathcal{A}} \pi_{t_k}^{(i)}(\mathrm{d}a)p(s', a, B) \\ &\xrightarrow[k\to\infty]{} \int_{\mathcal{S}} \mathbf{m}_{\pi_\infty}^{(i)}(\mathrm{d}s') \int_{\mathcal{A}} \pi_\infty^{(i)}(\mathrm{d}a)p(s', a, B),\end{aligned}$$

where the convergence is in the weak sense and where we used the fact that $(s, a) \mapsto p(s, a, B)$ is continuous (and obviously bounded). Then, the same reasoning by contradiction as for the case $i = n$ applies, by letting $K_\delta \subset \mathcal{S} \times \mathcal{A}$ be a compact subset such that $\int_{K_\delta^c} \mathrm{d}\mathbf{m}_{\pi_{t_k}}^{(i-1)} \mathrm{d}\overline{\pi} < \delta$. This concludes the proof. ■

Lemma 10 *Under the assumptions of Theorem 1, it holds that* $\sup_{t\geq 0} ||\theta_t|| < \infty$.

Proof Let θ_{t_k} be a subsequence of θ_t such that π_{t_k} converges weakly to some π_∞, which exists thanks to Lemma 9. Assume that there exists $i \in \{1, \ldots, n\}$ such that $||\theta_{t_k}^{(i)}|| \to \infty$ as $k \to \infty$, and let i be the smallest such integers. Let $\underline{\theta}_{t_k}^{(i)} := \frac{\theta_{t_k}^{(i)}}{||\theta_{t_k}^{(i)}||_2}$ and $\underline{\theta}_\infty^{(i)} := \lim_{k\to\infty} \underline{\theta}_{t_k}^{(i)}$ (to ensure convergence, one can always take subsequences since $\underline{\theta}_{t_k}$ lives in a compact sphere.) Without loss of generality, we choose the subsequence such that

$$\underline{\theta}_\infty^{(i)} \cdot \nabla_{\theta^{(i)}} J_n(\pi_{t_k}) > 0. \tag{20}$$

Indeed, since $||\theta_{t_k}^{(i)}||_2 \to \infty$ and $\underline{\theta}_{t_k} \to \underline{\theta}_\infty$, necessarily, $\nabla_{\theta^{(i)}} J_n(\pi_{t_k})$ must point inside the half-plane $\{v \in \mathbb{R}^{P_i} : v \cdot \underline{\theta}_\infty^{(i)} > 0\}$ infinitely many times. We now show that this leads to a contradiction.

Firstly, there exists a constant C independent of t_k such that for all $j < i$, it holds that $||\theta_{t_k}^{(j)}||_2 \leq C$. We thus have for all $s \in \mathcal{S}$ that

$$\begin{aligned} D_{\mathrm{KL}}(\pi_{t_k}^{(j)}||\overline{\pi})(s) &= \int_{\mathcal{A}} \pi_{t_k}^{(j}(\mathrm{d}a|s) \left(h_{t_k}^{(j)}(a,s)/\tau - \log \mathbb{E}_{\overline{\pi}}[e^{h_{t_k}^{(j)}(A,s)/\tau}] \right) \\ &\leq \frac{2}{\tau}||\theta_{t_k}^{(j)}||_2||\psi^{(j)}||_\infty \\ &\leq \frac{2}{\tau}C||\psi^{(j)}||_\infty \end{aligned}$$

In particular, by definition of Q-functions, this yields

$$||Q_{\pi_{t_k}}^{(i)}||_\infty \leq i\left(||r||_\infty + 2C||\psi||_\infty\right). \tag{21}$$

On the other hand, by Lemma 6, it holds that

$$\begin{aligned} Q_{\pi_t}^{(i)}(a,s) &= r(a,s) + \int_{\mathcal{S}} p(s,a,\mathrm{d}s')V_{\pi_t}^{(i-1)}(s') \\ &= \tau \log \frac{\pi_*^{(i)}}{\overline{\pi}}(a|s) - \int_{\mathcal{S}} p(s,a,\mathrm{d}s')(V_{\pi_t}^{(i-1)}(s') - V_*^{(i)}(s')) \\ &= \tau \log \frac{\pi_*^{(i)}}{\overline{\pi}}(a|s) - \tau \sum_{k=i+1}^{n-1} \mathbb{E}_{\pi_t}\left[D_{\mathrm{KL}}(\pi_t^{(n-k)}||\pi_*^{(n-k)})(S_k)|S_i = s, A_i = a \right]. \end{aligned}$$

Note that by compactness, we can take a subsequence such that $\underline{\theta}_{t_k}^{(j)} \to \underline{\theta}_\infty^{(j)}$ simultaneously for all $j \leq i$. We still denote by $\underline{\theta}_{t_k}^{(j)}$ such subsequences. A computation similar to Equation (18) gives

$$\begin{aligned} -\underline{\theta}_{t_k}^{(i)} \cdot \nabla_{\theta^{(i)}} J_n(\pi_{t_k}) = \int_{\mathcal{S}} \mathbf{m}_{\pi_{t_k}}^{(i)}(\mathrm{d}s) \int_{\mathcal{A}} \pi_{t_k}^{(i)}(\mathrm{d}a|s) \log \frac{\pi_{\underline{\theta}_{t_k}}^{(i)}}{\overline{\pi}}(a|s) \\ \times \left(\tau \log \frac{\pi_{t_k}^{(i)}}{\overline{\pi}}(a|s) - \tau D_{\mathrm{KL}}(\pi_{t_k}^{(i)}||\overline{\pi})(s) - Q_{\pi_{t_k}}^{(i)}(a,s) + \mathbb{E}_{\pi_{t_k}}\left[Q_{\pi_{t_k}^{(i)}}(A|s) \right] \right). \end{aligned}$$

Using that $h_{\theta_{t_k}^{(i)}}^{(i)} = ||\theta_t^{(i)}||_2 h_{\underline{\theta}_{t_k}^{(i)}}^{(i)}$ by definition, we have that

$$\begin{aligned} \tau\left(\log \frac{\pi_{t_k}^{(i)}}{\overline{\pi}}(a|s) - D_{\mathrm{KL}}(\pi_{t_k}^{(i)}||\overline{\pi})(s) \right) &= h_{\theta_{t_k}}^{(i)}(a,s) - \mathbb{E}_{\pi_{t_k}}\left[h_{\theta_{t_k}}^{(i)}(A,s) \right] \\ &= ||\theta_{t_k}^{(i)}||_2 \left(h_{\underline{\theta}_{t_k}}^{(i)}(a,s) - \mathbb{E}_{\pi_{t_k}}\left[h_{\underline{\theta}_{t_k}}^{(i)}(A,s) \right] \right) \\ &= \tau||\theta_{t_k}^{(i)}||_2 \left(\log \frac{\pi_{\underline{\theta}_{t_k}}^{(i)}}{\overline{\pi}}(a|s) - \mathbb{E}_{\pi_{t_k}}\left[\log \frac{\pi_{\underline{\theta}_{t_k}}^{(i)}}{\overline{\pi}}(A|s) \right] \right) \end{aligned}$$

Hence, we obtain

$$-\underline{\theta}^{(i)}_{t_k}\cdot\nabla_{\theta^{(i)}}J_n(\pi_{t_k}) = ||\theta^{(i)}_{t_k}||_2\int_{\mathcal{S}}\mathbf{m}^{(i)}_{\pi_{t_k}}(\mathrm{d}s)\int_{\mathcal{A}}\pi^{(i)}_{t_k}(\mathrm{d}a|s)\log\frac{\pi^{(i)}_{\underline{\theta}_{t_k}}}{\overline{\pi}}(a|s)$$
$$\times\left(\tau\log\frac{\pi^{(i)}_{\underline{\theta}_{t_k}}}{\overline{\pi}}(a|s)-\tau\mathbb{E}_{\pi_{t_k}}\left[\log\frac{\pi^{(i)}_{\underline{\theta}_{t_k}}}{\overline{\pi}}(A|s)\right]-\frac{Q^{(i)}_{\pi_{t_k}}(a,s)-\mathbb{E}_{\pi_{t_k}}\left[Q^{(i)}_{\pi_{t_k}}(A|s)\right]}{||\theta_{t_k}||_2}\right).$$

By Theorem 1, the right-hand side above converges to 0 as $k\to\infty$. In particular, since $||\theta^{(i)}_{t_k}||_2\to\infty$, the weak convergence of $\mathbf{m}^{(i)}_{\pi_{t_k}}(\mathrm{d}s)\times\pi^{(i)}_{t_k}(\mathrm{d}a|s)$ and the uniform boundedness of $\log\frac{\pi^{(i)}_{\underline{\theta}_{t_k}}}{\overline{\pi}}$ and $Q^{(i)}_{\pi_{t_k}}$ entail that

$$0=\int_{\mathcal{S}}\mathbf{m}^{(i)}_{\pi_\infty}(\mathrm{d}s)\int_{\mathcal{A}}\pi^{(i)}_{\infty}(\mathrm{d}a|s)\left(\log\frac{\pi_{\underline{\theta}^{(i)}_\infty}}{\overline{\pi}}(a|s)-\mathbb{E}_{\pi_\infty}\left[\log\frac{\pi^{(i)}_{\underline{\theta}_\infty}}{\overline{\pi}}(A|s)\right]\right)^2.$$

This shows that for $\mathbf{m}^{(i)}_{\pi_\infty}$-almost every s, the map $a\mapsto\log\frac{\pi^{(i)}_{\underline{\theta}_\infty}}{\overline{\pi}}(a|s)$ is constant on the support of $\pi^{(i)}_\infty$. If $\pi^{(i)}_\infty(\cdot|s)$ has full support for $\mathbf{m}^{(i)}_{\pi_{t_k}}$-almost every s, then $a\mapsto h_{\underline{\theta}_\infty}(a,s)$ is constant for such s, which contradicts the fact that $||\underline{\theta}_\infty||_2=1\neq0$.

Suppose instead that $\pi^{(i)}_\infty(\cdot|s)$ does not have full support for a subset of $\mathcal{S}$ with positive $\mathbf{m}^{(i)}_{\pi_\infty}$-measure. For all $s\in\mathcal{S}$, let $E_s:=\mathrm{Supp}(\pi_\infty(\cdot|s))$ and denote by $E^c_{\mathcal{S}}$ its complementary set. One can show by contradiction that since $h^{(i)}_{t_k}=||\theta^{(i)}_{t_k}||_2h^{(i)}_{\underline{\theta}_{t_k}}$ and $h^{(i)}_{\underline{\theta}_{t_k}}\to h^{(i)}_{\underline{\theta}_\infty}$ pointwise as $k\to\infty$, it holds that $h^{(i)}_{\underline{\theta}_\infty}(a,s)=C_s:=\sup_{a'\in\mathcal{A}}h^{(i)}_{\underline{\theta}_\infty}(a',s)$, for all $a\in E_s$, and similarly, $h^{(i)}_{\underline{\theta}_\infty}(a,s)\le C_s$ for all $a\in E^c_s$. We therefore see that for all $s\in\mathcal{S}$, it holds that

$$\int_{\mathcal{A}}\pi^{(i)}_{t_k}(\mathrm{d}a|s)\left(\log\frac{\pi^{(i)}_{t_k}}{\overline{\pi}}(a|s)-D_{\mathrm{KL}}(\pi^{(i)}_{t_k}||\overline{\pi})(s)-Q^{(i)}_{\pi_{t_k}}(a,s)+\mathbb{E}_{\pi_{t_k}}\left[Q^{(i)}_{\pi_{t_k}}(A,s)\right]\right)\log\frac{\pi^{(i)}_{\underline{\theta}_\infty}}{\overline{\pi}}(s)$$
$$=\int_{E^c_s}\pi^{(i)}_{t_k}(\mathrm{d}a|s)\left(\log\frac{\pi^{(i)}_{t_k}}{\overline{\pi}}(a|s)-D_{\mathrm{KL}}(\pi^{(i)}_{t_k}||\overline{\pi})(s)-Q^{(i)}_{\pi_{t_k}}(a,s)+\mathbb{E}_{\pi_{t_k}}\left[Q^{(i)}_{\pi_{t_k}}(A,s)\right]\right)$$
$$\times\left(\log\frac{\pi^{(i)}_{\underline{\theta}_\infty}}{\overline{\pi}}(a|s)-\log\frac{C_s}{\overline{\pi}}\right),\quad(22)$$

where we used that the integral over $\mathcal{A}$ of the terms inside the first parentheses is null. As explained above, the second factor in the integral is non-positive, and strictly negative for some a since $\pi^{(i)}_{\underline{\theta}}\neq\overline{\pi}$. We claim that the first term is negative, which is seen as follows: we note that for all s, for all $a\in E_s$ and $a'\in E^c_s$, we have for all k large enough that $\log\frac{\pi^{(i)}_{t_k}}{\overline{\pi}}(a|s)+Q^{(i)}_{\pi_{t_k}}(a,s)>\log\frac{\pi^{(i)}_{t_k}}{\overline{\pi}}(a'|s)+Q^{(i)}_{\pi_{t_k}}(a',s)$, since $Q^{(i)}_{\pi_{t_k}}$ is uniformly bounded and $\log\frac{\pi^{(i)}_{t_k}}{\overline{\pi}}(a'|s)\to-\infty$ for all $a'\in E^c_s$. Moreover, $\pi^{(i)}_{t_k}(E^c_s|s)\to0$ as $k\to\infty$. Hence, for all $a'\in E^c_s$, we have $\log\frac{\pi^{(i)}_{t_k}}{\overline{\pi}}(a'|s)+Q^{(i)}_{\pi_{t_k}}(a',s)>\mathbb{E}_{\pi_{t_k}}[\log\frac{\pi^{(i)}_{t_k}}{\overline{\pi}}(A|s)+Q^{(i)}_{\pi_{t_k}}(A,s)]$ for all k large enough. This shows that for all $s \in \mathcal{S}$, for all k large enough, the left-hand side of (22) is positive.

Hence, we now see that

$$-\underline{\theta}^{(i)}_{\infty} \cdot \nabla_{\theta^{(i)}} J_n(\pi_{t_k}) = \int_{\mathcal{S}} \mathbf{m}^{(i)}_{\pi_{t_k}}(\mathrm{d}s) \int_{\mathcal{A}} \pi^{(i)}_{t_k}(\mathrm{d}a|s) \log \frac{\pi^{(i)}_{\underline{\theta}_\infty}}{\overline{\pi}}(a|s)$$
$$\times \left(\tau \log \frac{\pi^{(i)}_{t_k}}{\overline{\pi}}(a|s) - \tau D_{\mathrm{KL}}(\pi^{(i)}_{t_k} || \overline{\pi})(s) - Q^{(i)}_{\pi_{t_k}}(a,s) + \mathbb{E}_{\pi_{t_k}}\left[Q_{\pi^{(i)}_{t_k}}(A|s) \right] \right)$$

is positive for all k large enough. This contradicts (20) and concludes the proof. ■

D.4 Global optimality of MPG: realizable case

Proof (*Proposition 1*) The Kullback-Leibler divergence being non-negative, it is readily seen that for all $s \in \mathcal{S}$, the maximal value of $\pi \mapsto V^{(n)}_{\pi}(s)$ is obtained for $\pi = \pi_*$. It is then immediate that π_* is the unique uniformly optimal policy for the objective J_n given in (4). ■

Recall that $\mathbf{m}^{(i)}_{\pi}$ denotes the law of S_{n-i} when following policiy π from initial state $S_0 \sim \nu$.

Lemma 11 *Let $t \in \mathbb{N}$ and $m \in \{1, \dots, n\}$. Suppose that $\pi^{(k)}_t(\cdot|s) = \pi^{(k)}_*(\cdot|s)$ for $\mathbf{m}^{(k)}_{\pi}$-almost every $s \in \mathcal{S}$, for all $k = 1, \dots, m-1$. For all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, it holds that*

$$\log \pi^{(m)}_{t+1}(a|s) - \log \pi^{(m)}_t(a|s)$$
$$= -\eta\tau \int_{\mathcal{S}} \mathbf{m}^{(m)}_{\pi_t}(\mathrm{d}s') \int_{\mathcal{A}} \pi^{(m)}_t(\mathrm{d}a'|s') \left(\log \frac{\pi^{(m)}_t}{\pi^{(m)}_*}(a'|s') - D_{\mathrm{KL}}(\pi^{(m)}_t || \pi^{(m)}_*)(s') \right)$$
$$\times \left(\Theta^{(m)}((a,s),(a',s')) - \mathbb{E}_{\pi^{(m)}_t}[\Theta^{(m)}((A,s),(a',s'))] \right) + o\left(\eta C(\theta_t)\right),$$

where the constant $C(\theta_t)$ does not depend on η.

Proof The gradient of the policy reads as

$$\nabla_\theta \pi^{(m)}_t(a|s) = \frac{1}{\tau} \pi^{(m)}_t(a|s) \int_{\mathcal{A}} \left(\delta_{a,\mathrm{d}a'} - \pi^{(m)}_t(\mathrm{d}a'|s) \right) \nabla_\theta h^{(m)}_t(a,s). \tag{23}$$

Let $(a,s) \in \mathcal{A} \times \mathcal{S}$. Using (8) and a first order Taylor approximation, we write

$$\log \pi^{(m)}_{t+1}(a|s) - \log \pi^{(m)}_t(a|s) = (\theta^{(m)}_{t+1} - \theta^{(m)}_t) \cdot \frac{\nabla_\theta \pi^{(m)}_t(a|s)}{\pi^{(m)}_t(a|s)} + o\left(\eta C(\theta_t)\right)$$
$$= \frac{\eta}{\tau^2} \mathbb{E}_{\pi_t}\Big[C_m \int_{\mathcal{A}\times\mathcal{A}} \left(\delta_{a,\mathrm{d}a'} - \pi^{(m)}_t(\mathrm{d}a'|s) \right)$$
$$\times \left(\delta_{A_{n-m},\mathrm{d}a''} - \pi^{(m)}_t(\mathrm{d}a''|S_{n-m}) \right) \Theta^{(m)}((a',s),(a'',S_{n-m})) \Big]$$
$$+ o\left(\eta C(\theta_t)\right). \tag{24}$$

We focus on the expectation. It is equal to

$$\mathbb{E}_{\pi_t}\Bigg[C_m\Bigg(\Theta^{(m)}((a,s),(A_{n-m},S_{n-m})) - \mathbb{E}_A\left[\Theta^{(m)}((A,s),(A_{n-m},S_{n-m}))\right] \\ - \mathbb{E}_A\left[\Theta^{(m)}((a,s),(A',S_{n-m}))\right] + \mathbb{E}_{A,A'}\left[\Theta^{(m)}((A,s),(A',S_{n-m}))\right]\Bigg)\Bigg],$$

where A, A' have respective laws $\pi_t^{(m)}(\cdot|s)$ and $\pi_t^{(m)}(\cdot|S_{n-m})$ and are mutually independent of all other variables (conditionally given S_{n-m} for A'). Using the trick $\mathbb{E}[X(Y-\mathbb{E}[Y])] = \mathbb{E}[(X-\mathbb{E}[X])Y]$, we obtain

$$\mathbb{E}_{\pi_t}\Big[\left(C_m - \mathbb{E}\left[C_m|S_{n-m}\right]\right)\Big(\Theta^{(m)}((a,s),(A_{n-m},S_{n-m})) - \mathbb{E}_A\left[\Theta^{(m)}((A,s),(A_{n-m},S_{n-m}))\right]\Big)\Big]. \tag{25}$$

We write

$$\begin{aligned}\mathbb{E}\left[C_m|S_{n-m}\right] &= \mathbb{E}\left[\sum_{\ell=n-m}^{n}\left(R_\ell - \tau\log\frac{\pi_t^{(n-\ell)}}{\overline{\pi}}(A_\ell|S_\ell)\right)\Bigg|S_{n-m}\right]\\ &= V_{\pi_t}^{(m)}(S_{n-m})\\ &= V_*^{(m)}(S_{n-m}) - D_{\mathrm{KL}}(\pi_t^{(m)}||\pi_*^{(m)})(S_{n-m}),\end{aligned}$$

where we used Lemma 6 and the fact that $\pi_t^{(i)}(\cdot|s) = \pi_*^{(i)}(\cdot|s)$ for $\mathbf{m}_{\pi_t^{(i)}}$-almost every $s\in\mathcal{S}$, for all $i = 1,\dots,m-1$. Similarly and using the expression (5) of the optimal policy, we have

$$\begin{aligned}\mathbb{E}[C_m|S_{n-m},A_{n-m}] &= R_{n-m} - \tau\log\frac{\pi_t^{(m)}}{\overline{\pi}}(A_{n-m}|S_{n-m}) + \mathbb{E}\left[V_{\pi_t}^{(m-1)}(S_{n-m})\Big|S_{n-m},A_{n-m}\right]\\ &= \mathbb{E}\Big[V_{\pi_t}^{(m-1)}(S_{n-m+1}) - V_*^{(m-1)}(S_{n-m+1})\Big|S_{n-m},A_{n-m}\Big]\\ &\qquad - \tau\log\frac{\pi_t^{(m)}}{\pi_*^{(m)}}(A_{n-m}|S_{n-m}) + V_*^{(m)}(S_{n-m})\\ &= -\tau\log\frac{\pi_t^{(m)}}{\pi_*^{(m)}}(A_{n-m}|S_{n-m}) + V_*^{(m)}(S_{n-m}).\end{aligned}$$

Hence, the expression in (25) becomes

$$\tau\mathbb{E}_{\pi_t}\Bigg[\Bigg(D_{\mathrm{KL}}(\pi_t^{(m)}||\pi_*^{(m)})(A_{n-m}|S_{n-m}) - \log\frac{\pi_t^{(m)}}{\pi_*^{(m)}}(A_{n-m}|S_{n-m})\Bigg)\Big(\Theta^{(m)}((a,s),(A_{n-m},S_{n-m})) \\ - \mathbb{E}_A\left[\Theta^{(m)}((A,s),(A_{n-m},S_{n-m}))\right]\Big)\Bigg],$$

which corresponds to the first order term in right-hand side of the equation in the Lemma. Coming back to (24), this concludes the proof. ■

Proof (*Theorem 2*) The idea of the proof is rather simple: by Lemmas 9 and 10, we know that any subsequence of π_t has a converging subsequence, and that the limits have finite norm parameters. Hence, we only need to show the unicity of the limit, namely, that $\theta \in \mathbb{R}^P$ is a critical point if and only $\pi_\theta = \pi_*$. We follow the intuition given after Theorem 2.

We reason by induction. Let $m \leq n$, suppose that $\pi_t^{(i)}(\cdot|s) = \pi_*^{(i)}(\cdot|s)$ for $\mathbf{m}_{\pi_t}^{(i)}$-almost every $s \in \mathcal{S}$, for all $i = 1, \dots, m-1$, and that $\pi_t^{(i)} = \pi_\infty^{(i)}$ for all $i = m, \dots, n$. In particular, we are at a critical point $(\theta_t^{(1)}, \dots, \theta_t^{(n)})$ of $(\theta^{(1)}, \dots, \theta^{(n)}) \mapsto J_n(\pi_\theta)$. Let $a \in \mathcal{A}, s \in \mathcal{S}$. By Lemma 11, we have that

$$\begin{aligned}
0 &= \log \pi_{t+1}^{(m)}(a|s) - \log \pi_t^{(m)}(a|s) \\
&= -\eta\tau \int_{\mathcal{A}\times\mathcal{S}} \mathbf{m}_{\pi_t}^{(m)}(\mathrm{d}s')\pi_t^{(m)}(\mathrm{d}a'|s') \left(\log \frac{\pi_t^{(m)}}{\pi_*^{(m)}}(a'|s') - D_{\mathrm{KL}}(\pi_t^{(m)} || \pi_*^{(m)})(s') \right) \\
&\qquad\qquad \times \left(\Theta^{(m)}((a,s),(a',s')) - \mathbb{E}_{\pi_t^{(m)}}[\Theta^{(m)}((A,s),(a',s'))] \right) + o\left(\eta C(\theta_t)\right),
\end{aligned}$$

Since the above must be true for all $\eta > 0$, we deduce that

$$\begin{aligned}
&\int_{\mathcal{A}\times\mathcal{S}} \mathbf{m}_{\pi_t}^{(m)}(\mathrm{d}s')\pi_t^{(m)}(\mathrm{d}a'|s') \left(\log \frac{\pi_t^{(m)}}{\pi_*^{(m)}}(a'|s') - D_{\mathrm{KL}}(\pi_t^{(m)} || \pi_*^{(m)})(s') \right) \\
&\qquad\qquad \times \left(\Theta^{(m)}((a,s),(a',s')) - \mathbb{E}_{\pi_t^{(m)}}[\Theta^{(m)}((A,s),(a',s'))] \right) = 0. \qquad (26)
\end{aligned}$$

Let $\widetilde{\Theta}^{(m)}$ be the positive-semidefinite kernel constructed from $\Theta^{(m)}$ and $\pi_t^{(m)}$ as in Lemma 4. One can easily check that

$$\log \frac{\pi_t^{(m)}}{\pi_*^{(m)}}(a|s) - D_{\mathrm{KL}}(\pi_t^{(m)} || \pi_*^{(m)}) = h_t^{(m)}(a,s) - Q_*^{(m)}(a,s) - \mathbb{E}_{\pi_t^{(m)}}\left[h_t^{(m)}(A,s) - Q_*^{(m)}(A,s) \right].$$

In particular, using the trick $\mathbb{E}[X(Y - \mathbb{E}[Y])] = \mathbb{E}[(X - \mathbb{E}[X])Y]$, we can rewrite (26) as

$$\int_{\mathcal{A}\times\mathcal{S}} \mathbf{m}_{\pi_t}^{(m)}(\mathrm{d}s')\pi_t^{(m)}(\mathrm{d}a'|s') \left(h_t^{(m)}(a',s') - Q_*^{(m)}(a',s') \right) \widetilde{\Theta}^{(m)}((a,s),(a',s')) = 0. \qquad (27)$$

Since the above is true for all $(a,s) \in \mathcal{A}\times\mathcal{S}$, we see by Lemma 3 that $h_t^{(m)} - Q_*^{(m)} \in (\mathcal{H}_{\widetilde{\Theta}^{(m)}})^\perp$, that is the orthogonal complement of $\mathcal{H}_{\widetilde{\Theta}^{(m)}}$ in $L^2(\mathbf{m}_{\pi_t}^{(m)}(\mathrm{d}s')\pi_t^{(m)}(\mathrm{d}a'|s'))$. By Assumption **A1.**, we get $h_t^{(m)} - Q_*^{(m)} \in \mathcal{H}_{\Theta^{(m)}} \cap (\mathcal{H}_{\widetilde{\Theta}^{(m)}})^\perp$, and Lemma 4 entails that for $\mathbf{m}_{\pi_t}^{(m)}$-almost every $s \in \mathcal{S}$, the map $a \mapsto h_t^{(m)}(a,s) - Q_*(a,s)$ is constant. This implies in turn that $\pi_t^{(m)}(\cdot|s) = \pi_*^{(m)}(\cdot|s)$ for $\mathbf{m}_{\pi_t}^{(m)}$-almost every $s \in \mathcal{S}$, which concludes the proof. ■

D.5 Global optimality of MPG: non-realizable case

In order to extend the global optimality from the case where π_* belongs to the parametric space $\mathscr{P}_n$ to the case where π_* is outside of $\mathscr{P}_n$, we use tools from information geometry and apply the strategy outlined in Section 3.4.

We use the following notation in the proof: the set of parametric 1-step policies whose preference h_θ belongs to $\mathcal{H}_{\Theta^{(i)}}$ is denoted by $\mathscr{P}^{(i)}$.

Proof (*Theorem 3*) Let $\vartheta \in \mathbb{R}^P$ be a critical point of $\theta \mapsto J_n(\pi_\theta)$. Consider a fixed $i \in \{1, \ldots, n\}$. Recall that $Q^{(i)}_{\pi_\vartheta}(a, s) = r(a, s) + \int_{\mathcal{S}} p(s, a, \mathrm{d}s')V^{(i-1)}_{\pi_\vartheta}(s')$, which does not depend on $\pi^{(j)}_\vartheta$, $j \geq i$. Let $\widehat{\pi}^{(i)}$ be the policy with preference $Q^{(i)}_{\pi_\vartheta}$. Note that $Q^{(i)}_{\pi_\vartheta}$ does not necessarily belong to $\mathcal{H}_{\Theta^{(i)}}$, hence we do not make the dependence on ϑ (which is fixed) explicit in $\widehat{\pi}^{(i)}$. This is the optimal policy given that the shorter j-step policies, $j < i$, are fixed. Indeed, we always have that

$$\begin{aligned}\widehat{J}^{(i)}(\pi^{(i)}, \vartheta) &:= \int_{\mathcal{S}} \mathbf{m}^{(i)}_{\pi_\vartheta}(\mathrm{d}s)\left(\mathbb{E}_{\pi^{(i)}}[Q^{(i)}_{\pi_\vartheta}(A, s)] - \tau D_{\mathrm{KL}}(\pi^{(i)}||\overline{\pi})(s)\right)\\ &= \tau \int_{\mathcal{S}} \mathbf{m}^{(i)}_{\pi_\vartheta}(\mathrm{d}s)\left(\log\left(\int_{\mathcal{A}} \overline{\pi}(\mathrm{d}a|s)e^{Q^{(i)}_{\pi_\vartheta}(a,s)/\tau}\right) - D_{\mathrm{KL}}(\pi^{(i)}||\widehat{\pi}^{(i)})(s)\right).\end{aligned}$$

The first term of the right-hand side depends on $\pi^{(j)}_\vartheta$ through $Q^{(i)}_{\pi_\vartheta}$ for $j < i$, whereas it depends on $\pi^{(j)}_\vartheta$ through $\mathbf{m}^{(i)}_{\pi_\vartheta}$ for $j > i$, but it does not depend on $\pi^{(i)}_\vartheta$. Therefore, we see that $\widehat{\pi}^{(i)} = \mathrm{argmax}_{\pi^{(i)} \in \mathcal{P}_1} \widehat{J}^{(i)}(\pi^{(i)}, \vartheta)$.

Similar to what was done in appendix C, let $\mathscr{P}^{(i)}_\star$ be the quotient space of $\mathscr{P}^{(i)}$ and its subspace of policies whose preferences are constant in a for all $s \in \mathrm{Supp}(\mathbf{m}_{\pi^{(i)}_\vartheta})$. In this quotient space, policies that are equal to each other on the support of $\mathbf{m}_{\pi^{(i)}_\vartheta}$ are identified as the same policy, since states outside of this set are never visited with probability one. Define

$$\pi^{(i)}_{\theta_*} = \underset{\pi^{(i)}_\theta \in \mathscr{P}^{(i)}_\star}{\mathrm{argmin}}\ D^{(i)}(\widehat{\pi}^{(i)}, \pi^{(i)}_\theta) := \underset{\pi^{(i)}_\theta \in \mathscr{P}^{(i)}_\star}{\mathrm{argmin}} \int_{\mathcal{S}} \mathbf{m}^{(i)}_{\pi_\vartheta}(\mathrm{d}s) D_{\mathrm{KL}}(\pi^{(i)}_\theta||\widehat{\pi}^{(i)})(s).$$

It turns out that the map $D^{(i)}$ defined above is a Bregman divergence on $\mathcal{P}_\star$ (denoting the space where policies that coincide on the support of $\mathbf{m}^{(i)}_{\pi_\vartheta}$ are identified together). Using the fact that ϑ is a critical point combined with Lemma 5, we have that

$$0 = \nabla_{\theta^{(i)}} J_n(\pi_\vartheta) = -\nabla_{\theta^{(i)}} D(\pi^{(i)}_\vartheta, \pi^{(i)}_{\theta_*}).$$

We stress once more that $\pi^{(i)}_{\theta_*}$ only depends on $\widehat{\pi}^{(i)}$, which in turn only depends on $\pi^{(1)}_\vartheta, \ldots, \pi^{(i-1)}_\vartheta$ through $Q^{(i)}_{\pi_\vartheta}$ and on $\pi^{(i+1)}_\vartheta, \ldots, \pi^{(n)}_\vartheta$ through $\mathbf{m}^{(i)}_{\pi_\vartheta}$. Therefore, the equation above corresponds to the gradient of the objective of 1-step MPG with optimal policy $\pi^{(i)}_{\theta_*}$. This observation brings us back to the realizable case, for which Theorem 2 applies. This implies that necessarily, $\pi^{(i)}_\vartheta(\cdot|s) = \pi^{(i)}_{\theta_*}(\cdot|s)$ for $\mathbf{m}^{(i)}_{\pi_\vartheta}$-almost every $s \in \mathcal{S}$. In particular, this shows the uniqueness of the argmin for reachable states.

The above argument proves that if $\vartheta \in \mathbb{R}^P$ is a critical point, then

$$J_n(\pi_\vartheta) = \max_{\theta^{(i)} \in \mathbb{R}^{P_i}} J_n(\pi^{(1)}_\vartheta, \ldots, \pi^{(i)}_{\theta^{(i)}}, \ldots, \pi^{(n)}_{\vartheta^{(n)}}).$$

Since this is true for every $i = 1, \ldots, n$ and since maxima can be taken in any order, we have that

$$J_n(\pi_\vartheta) = \max_{\theta \in \mathbb{R}^P} J_n(\pi_\theta)$$

We have thus proved that any critical point is a global maximum of the objective. As in the proof of Theorem 2, the argument using Lemmas 9 and 10 applies to show global convergence, concluding the proof. ■

Proof (**Proposition 3**) Suppose that $\theta_t = (\theta_t^{(1)}, \dots, \theta_t^{(n)})$ satisfies the projectional consistency property (11). We thus have that $h_t^{(1)} - Q_*^{(1)} \in (\mathcal{H}_{\widetilde{\Theta}^{(1)}})^\perp$, the orthogonal space of $\mathcal{H}_{\widetilde{\Theta}^{(1)}}$ in $L^2(\mathbf{m}^{(1)}(\mathrm{d}s)\pi_t^{(1)}(\mathrm{d}a))$. In particular, using Lemma 3, one can show that Equation (27) is satisfied, entailing that $\nabla_{\theta^{(1)}} J_n(\pi_\theta) = 0$. The same reasoning applies for all steps $i = 1, \dots, n$, showing that θ_t is a critical point, and therefore, the unique global optimum by Theorem 3. This concludes the proof. ■

Appendix E. Assumptions

We now list the assumptions and briefly mention their roles in this work:

- In Proposition 2, ν has full support in $\mathcal{S}$ and the MDP is ergodic: it is not restrictive, as its role is to ensure that the optimal policies for all horizons visit Lebesgue almost all states, thus avoiding considerations about reachable states. Ergodicity ensures the existence of a stationary state distribution. In particular, the optimal policy π_* does not depend on ν.

- Continuous closed $\mathcal{A}, \mathcal{S}$: to apply Mercer's Theorem.

- Continuous and bounded kernels $\Theta^{(i)}$: to apply Mercer's Theorem.

- Measurable selection assumption and measurability of p and r: ensures the measurability of the variables generated by the MDP, Lebesgue integrability and avoid pathological cases.

- For all Borel set $B \subset \mathcal{S}$, the map $(s, a) \mapsto p(s, a, B)$ is continuous: is used in the proof of Lemma 9 to guarantee convergence of a subsequence of the state distributions and policies.

- Rewards are bounded: ensures that value functions are well defined. It is also used to prove the convergence of the objective and of the parameters to finite values.

- For all $s \in \mathcal{S}$ and all $\theta^{(i)} \in \mathbb{R}^{P_i}$, the map $a \mapsto h_\theta^{(i)}(a, s)$ is constant if and only if $||\theta^{(i)}|| = 0$: this guarantees that $\pi_\theta = \overline{\pi}$ if and only if $||\theta|| = 0$ and avoid pathological cases, such as divergence of parameters.

Appendix F. Numerical experiments

We apply MPG on a number of numerical experiments as detailed in section 4. The MPG is implemented as in algorithm 1.

Algorithm 1 MPG implementation for N horizon task

```
**Input:** initial temperature $\tau_0$, initial learning rate $\eta_0$, final temperature $\tau_T$, final learning rate $\eta_T$
$\tau \leftarrow \tau_0$
$\eta \leftarrow \eta_0$
**for** t = 1, ..., episodes **do**
    generate trajectory from policies $\{\pi_t^n, \pi_t^{n-1}, ..., \pi_t^1\}$: $\{(s_i, s_{i+1}, a_i, r_i)\}_{i=0}^{n-1}$
    **for** i = 1, $\cdots$ , $n$ **do**
        $C_i = \sum_{\ell=n-i}^{n-1} \left( r_\ell - \tau \log \frac{\pi_t^{(n-\ell)}}{\bar{\pi}}(a_\ell|s_\ell) \right)$
        $\theta_{t+1}^{(i)} = \theta_t^{(i)} + \eta C_i \nabla \log \pi_t^{(i)}(a_{n-i}|s_{n-i})$
    **end for**
    decay $\tau, \eta$ using $d_\tau = \left(\frac{\tau_T}{\tau_0}\right)^{1/\text{episodes}}$ and $d_\eta = \left(\frac{\eta_T}{\eta_0}\right)^{1/\text{episodes}}$
**end for**
```

F.1 Analytical task

Set-up: We consider a state-space consisting of $\mathcal{S} = \{0, 1, 2, 3, 4\}$, an action space $\mathcal{A} = \{1, 2\}$. At each state s, the agent performs action a, taking the agent to the next state $(s + a) \mod 5$.

We define an orthonormal basis (in $\ell^2(\mathcal{A} \times \mathcal{S})$) of the space of functions $\{f : \mathcal{A} \times \mathcal{S} \to \mathbb{R} : f(1, s) + f(2, s) = 0, \ \forall s \in \mathcal{S}\}$. Note that one can always recenter any map g on $\mathcal{A} \times \mathcal{S}$ so that $g(1, s) + g(2, s) = 0$, without changing the policy obtained as the softmax of g, in particular, any policy can be written as the softmax of such a function. The basis is defined as

$$e_1 = \sqrt{6} \begin{pmatrix} 1 & -1 \\ 0 & 0 \\ 1 & -1 \\ 0 & 0 \\ 1 & -1 \end{pmatrix}, \quad e_2 = \sqrt{4} \begin{pmatrix} 0 & 0 \\ 1 & -1 \\ 0 & 0 \\ 1 & -1 \\ 0 & 0 \end{pmatrix}, \quad e_3 = \sqrt{4} \begin{pmatrix} 0 & 0 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ 0 & 0 \end{pmatrix},$$

$$e_4 = \sqrt{8} \begin{pmatrix} -2 & 2 \\ 0 & 0 \\ 1 & -1 \\ 0 & 0 \\ 1 & -1 \end{pmatrix}, \quad e_5 = \sqrt{4} \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \end{pmatrix}.$$

Recall that $Q_*^{(1)}(a, s) = r(a, s)$, which can be represented by

$$Q_*^{(1)}(a, s) = \sum_{j=1}^{5} \theta_j^* e_j(a, s), \quad \theta_j^* \in \mathbb{R}.$$

Experiments:

1. Obtaining the first two step policies with assumption **A1.**, namely, that the optimal policy's parameters can be represented by our parametric space, and when the assumption **A1.** does not hold.

Setup:

- preference function $h(a, s)$ is expressed by a linear model
- θ_0 randomly initialised with i.i.d. centered Gaussian with standard deviation 1;
- $\theta^* = (0, 0.1, -0.15, 0.05, -0.1)$;
- Initial learning rate $\eta_0 = 0.001$, terminal learning rate $\eta_T = 0.001$ (no decay);
- Temperature $\tau = 1.0$ remains fixed during training (no decau);
- True gradient update
- Number of episodes: 14000

F.2 Control problems

Setup:

- Preference function $h(a, s)$ is expressed by a fully connected neural network with 3 hidden layers, each with width 100 and ReLU activation function. The output layer has a softmax activation;
- Parameters are initialised with He initialisation;
- Initial learning rate η_0 and initial temperature τ_0 are hyper-parameters;
- Final learning rate η_T and final temperature τ_T are fixed and problem dependent;
- Number of episodes is task dependent;
- Both the learning rate and temperature decay during training, using the following decay rates $d_\eta = \left(\frac{\eta_T}{\eta_0}\right)^{1/episodes}$ and $d_\tau = \left(\frac{\tau_T}{\tau_0}\right)^{1/episodes}$, respectively;
- Gradient update estimated using one trajectory as in (8).

Frozen lake: Aside from the details specified above, further details of the set-up for Frozen lake are:

- Reward: it is well-known that reshaping the reward function can change the performance of the algorithm. The original reward function does not discriminate between falling into a hole, not moving and moving, so we used a reshaped reward function: falling (-1), moving against a wall (-0.1), moving successfully ($+0.01$) and reaching the treasure ($+10.0$);
- Final learning rate $\eta_T = 1 \times 10^{-6}$;
- Final temperature $\tau_T = 0.01$ (when applicable);
- Number of episodes: 1000.

We train sets of 3 agents to explore the hyper-parameter space as denoted in 1. In table 5, we show the hyper-parameter exploration for the different considered algorithms.

Table 5: Hyper-parameter search using 3 agents for Frozen lake task. The '**' symbol denotes that no runs were made for that set of hyper-parameters.

η_0	softPG	MPG
$\tau_0 = 0.15$		
0.1	-1.89	**
0.05	-1.66	**
0.01	0.20	2.13
0.005	0.20	2.06
0.001	0.20	3.04
0.0005	0.20	9.36
0.0001	0.20	-1.38
$\tau_0 = 0.20$		
0.1	-0.53	**
0.05	-0.19	**
0.01	0.20	6.53
0.005	0.20	6.69
0.001	0.20	10.05
0.0005	0.20	10.05
0.0001	0.20	1.66
$\tau_0 = 0.25$		
0.1	-0.93	**
0.05	-1.27	**
0.01	0.20	10.05
0.005	0.20	10.05
0.001	0.20	10.05
0.0005	0.20	9.99
0.0001	0.20	1.39
$\tau_0 = 0.30$		
0.1	-1.89	**
0.05	-1.89	**
0.01	0.18	6.33
0.005	0.20	10.05
0.001	0.20	10.05
0.0005	0.20	10.05
0.0001	0.19	-1.13
$\tau_0 = 0.35$		
0.1	-1.27	**
0.05	-1.05	**
0.01	0.20	10.05
0.005	0.20	10.05
0.001	0.20	10.05
0.0005	0.20	6.04
0.0001	0.20	-1.65
$\tau_0 = 0.40$		
0.1	-1.78	**
0.05	-1.66	**
0.01	0.19	10.05
0.005	0.19	10.05
0.001	0.20	10.05
0.0005	0.20	6.72
0.0001	0.20	0.30
$\tau_0 = 0.45$		
0.1	-1.55	**
0.05	-1.26	**
0.01	-0.42	10.05
0.005	0.20	6.40
0.001	0.20	4.09
0.0005	0.20	2.95
0.0001	0.20	0.25
$\tau_0 = 0.50$		
0.1	-2.00	**
0.05	0.20	**
0.01	-0.15	10.05
0.005	0.20	6.61
0.001	0.20	9.75
0.0005	0.20	2.49
0.0001	0.20	-1.22
$\tau_0 = 0.55$		
0.1	-1.78	**
0.05	-1.27	**
0.01	3.48	10.05
0.005	0.19	9.82
0.001	0.20	7.02
0.0005	0.20	2.59
0.0001	0.18	-1.29
$\tau_0 = 0.60$		
0.1	-1.55	**
0.05	-1.27	**
0.01	3.48	10.05
0.005	0.20	10.05
0.001	0.20	6.71
0.0005	0.20	-0.92
0.0001	0.20	-1.09
$\tau_0 = 0.65$		
0.1	-2.00	**
0.05	-0.53	**
0.01	3.14	**
0.005	0.20	**
0.001	0.20	**
0.0005	-0.53	**
0.0001	-0.44	**
$\tau_0 = 0.70$		
0.1	-1.89	**
0.05	-1.27	**
0.01	-0.55	**
0.005	-0.54	**
0.001	0.20	**
0.0005	-0.34	**
0.0001	-0.56	**

η_0	PG	nsPG
0.5	**	-1.89
0.1	**	-1.76
0.05	2.24	6.03
0.01	10.05	2.91
0.005	7.69	2.88
0.001	-0.83	-1.00
0.0005	-0.88	-1.06
0.0001	-1.12	-1.15

F.3 Cart Pole

Aside from the details specified above, further details of the set-up for Frozen lake are:

- Reward: The original reward function gives a +1 reward for each time that the pole stays upright, and the task finishes if the cart leaves the domain or if the pole is far enough from being upright. We reshape the reward function to yield a penalty $-10\exp(-0.05(i-1))$ when the task concludes, where i is the length of time the pole stayed upright. Namely, if the pole falls early in the task, the penalisation is larger.
- Final learning rate $\eta_T = 1 \times 10^{-8}$;
- Final temperature $\tau_T = 0.01$ (when applicable);
- Number of episodes: 1000.

We train sets of 3 agents to explore the hyper-parameter space as denoted in 3. In table 6, we show the hyper-parameter exploration for the different considered algorithms.

Table 6: Hyper-parameter search using 3 agents for balancing the cart pole task. The '**' symbol denotes that no runs were made for that set of hyper-parameters.

η_0	softPG	MPG
$\tau_0 = 0.05$		
0.001	**	1.87
0.0005	**	1.80
0.0001	**	1.70
5×10^{-5}	**	25.21
1×10^{-5}	**	73.22
1×10^{-6}	**	41.37
5×10^{-6}	**	58.85
$\tau_0 = 0.10$		
0.001	**	1.82
0.0005	1.70	7.50
0.0001	50.95	48.21
5×10^{-5}	51.72	41.21
1×10^{-5}	81.99	95.32
1×10^{-6}	55.40	33.45
5×10^{-6}	85.41	70.08
$\tau_0 = 0.15$		
0.001	**	1.69
0.0005	14.86	1.83
0.0001	59.52	68.65
5×10^{-5}	79.77	86.28
1×10^{-5}	75.65	88.94
1×10^{-6}	64.15	36.35
5×10^{-6}	83.88	81.07
$\tau_0 = 0.20$		
0.001	**	29.14
0.0005	1.84	1.79
0.0001	51.66	77.32
5×10^{-5}	79.13	88.69
1×10^{-5}	79.07	77.21
1×10^{-6}	37.01	19.39
5×10^{-6}	66.27	74.18
$\tau_0 = 0.25$		
0.001	**	6.01
0.0005	66.09	20.18
0.0001	62.15	71.22
5×10^{-5}	91.59	91.37
1×10^{-5}	64.18	77.80
1×10^{-6}	26.29	18.45
5×10^{-6}	57.06	73.08
$\tau_0 = 0.30$		
0.001	**	13.95
0.0005	61.24	32.84
0.0001	67.95	97.10
5×10^{-5}	93.45	99.37
1×10^{-5}	67.59	74.31
1×10^{-6}	61.09	28.72
5×10^{-6}	65.20	80.48
$\tau_0 = 0.35$		
0.001	**	37.08
0.0005	**	54.98
0.0001	84.55	75.19
5×10^{-5}	72.55	90.18
1×10^{-5}	86.74	76.82
1×10^{-6}	33.44	37.75
5×10^{-6}	59.14	65.92
$\tau_0 = 0.40$		
0.001	**	17.66
0.0005	**	46.27
0.0001	75.47	98.04
5×10^{-5}	65.54	95.11
1×10^{-5}	64.47	66.66
1×10^{-6}	27.15	20.33
5×10^{-6}	75.12	63.17
$\tau_0 = 0.45$		
0.001	**	**
0.0005	**	**
0.0001	88.47	**
5×10^{-5}	76.74	**
1×10^{-5}	64.13	**
1×10^{-6}	3.08	**
5×10^{-6}	52.08	**

η_0	PG	nsPG
0.005	2.18	38.68
0.001	80.75	98.94
0.0005	31.99	34.39
0.0001	18.62	19.18
5×10^{-5}	17.81	18.83
1×10^{-5}	16.66	16.96
5×10^{-6}	16.60	18.25
1×10^{-6}	17.62	16.28

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(1), jul 2022. ISSN 1532-4435.

Andréa Agazzi and Jianfeng Lu. Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. *ArXiv*, abs/2010.11858, 2020.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pages 151–160. PMLR, 2019.

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

Kristopher De Asis, Alan Chan, Silviu Pitis, Richard S. Sutton, and Daniel Graves. Fixed-horizon temporal difference methods for stable reinforcement learning. *ArXiv*, abs/1909.03906, 2019.

Claude Berge. *Topological spaces: Including a treatment of multi-valued functions, vector spaces and convexity.* Oliver & Boyd, 1963.

Dimitri P. Bertsekas. Dynamic programming and optimal control. 1995.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *ArXiv*, abs/1906.01786, 2019.

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Patrick Billingsley. *Convergence of probability measures.* John Wiley & Sons, 2013.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Oper. Res.*, 70:2563–2578, 2020.

Lénaïc Chizat and Francis R. Bach. A note on lazy training in supervised differentiable programming. *ArXiv*, abs/1812.07956, 2018.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *ArXiv*, abs/2110.10117, 2021.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Iteratively extending time horizon reinforcement learning. In *Proceedings of the 14th European Conference on Machine Learning*, ECML'03, page 96–107, Berlin, Heidelberg, 2003. Springer-Verlag. ISBN 3540201211. doi: 10.1007/978-3-540-39857-8_11. URL https://doi.org/10.1007/978-3-540-39857-8_11.

Benjamin Eysenbach and Sergey Levine. If maxent rl is the answer, what is the question? *ArXiv*, abs/1910.01913, 2019.

Benjamin Eysenbach and Sergey Levine. Maximum entropy rl (provably) solves some robust rl problems. *ArXiv*, abs/2103.06257, 2021.

Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, 2018.

KENJI Fukumizu. Infinite dimensional exponential families by reproducing kernel hilbert spaces. In *2nd International Symposium on Information Geometry and its Applications (IGAIA 2005)*, pages 324–333, 2005.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

Michael Giegrich, Christoph Reisinger, and Yufei Zhang. Convergence of policy gradient methods for finite-horizon stochastic linear-quadratic control problems. *arXiv preprint arXiv:2211.00617*, 2022.

Soumyajit Guin and Shalabh Bhatnagar. A policy gradient approach for finite horizon constrained markov decision processes. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 3353–3359. IEEE, 2023.

Tuomas Haarnoja, Haoran Tang, P. Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, 2017.

Tuomas Haarnoja, Aurick Zhou, P. Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 2018a.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, G. Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, P. Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *ArXiv*, abs/1812.05905, 2018b.

Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59 (5):3359–3391, 2021.

Onésimo Hernández-Lerma and Jean Bernard Lasserre. Discrete-time markov control processes. 1999.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8580–8589, Red Hook, NY, USA, 2018. Curran Associates Inc.

Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. *ArXiv*, abs/2002.08404, 2020.

Sara Klein, Simon Weissmann, and Leif Döring. Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods. *arXiv preprint arXiv:2310.02671*, 2023.

James-Michael Leahy, Bekzhan Kerimkulov, David Siska, and Lukasz Szpruch. Convergence of policy gradient for entropy regularized MDPs with neural network approximation in the mean-field regime. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12222–12252. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/leahy22a.html`.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *ArXiv*, abs/1805.00909, 2018.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. *ArXiv*, abs/2102.11270, 2021.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2015.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, 2020.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pages 7555–7564. PMLR, 2021.

Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings 19*, pages 154–168. Springer, 2006.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 2772–2782, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Brendan O'Donoghue, Rémi Munos, Koray Kavukcuoglu, and Volodymyr Mnih. Pgq: Combining policy gradient and q-learning. *ArXiv*, abs/1611.01626, 2016.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning, 2023.

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

John Schulman, Sergey Levine, P. Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. *ArXiv*, abs/1502.05477, 2015.

John Schulman, P. Abbeel, and Xi Chen. Equivalence between policy gradients and soft q-learning. *ArXiv*, abs/1704.06440, 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017b.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Richard S. Sutton, David A. McAllester, Satinder Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 1999.

Harm van Seijen, Mehdi Fatemi, and Arash Tavakoli. Using a logarithmic mapping to enable lower discount factors in reinforcement learning. In *Neural Information Processing Systems*, 2019.

Vivek VP and Dr Shalabh Bhatnagar. Finite horizon q-learning: Stability, convergence, simulations and an application on smart grids. *arXiv preprint arXiv:2110.15093*, 2021.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *ArXiv*, abs/1909.01150, 2019.

Lilian Weng. Policy gradient algorithms. *lilianweng.github.io*, 2018. URL `https://lilianweng.github.io/posts/2018-04-08-policy-gradient/`.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with reinforce. In *AAAI Conference on Artificial Intelligence*, 2020.

K. Zhang, Alec Koppel, Haoqi Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM J. Control. Optim.*, 58:3586–3612, 2019.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.