GLOBAL OPTIMALITY OF SOFTMAX POLICY GRADIENT WITH SINGLE HIDDEN LAYER NEURAL NETWORKS IN THE MEAN-FIELD REGIME

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of policy optimization for infinite-horizon discounted Markov Decision Processes with softmax policy and nonlinear function approximation trained with policy gradient algorithms. We concentrate on the training dynamics in the *mean-field* regime, modeling *e.g.*, the behavior of wide single hidden layer neural networks, when exploration is encouraged through entropy regularization. The dynamics of these models is established as a Wasserstein gradient flow of distributions in parameter space. We further prove global optimality of the fixed points of this dynamics under mild conditions on their initialization.

1 INTRODUCTION

In recent years, deep reinforcement learning has revolutionized the world of Artificial Intelligence by outperforming humans in a multitude of highly complex tasks and achieving breakthroughs that were deemed unthinkable at least for the next decade. Spectacular examples of such revolutionary potential have appeared over the last few years, with reinforcement learning algorithms mastering games and tasks of increasing complexity, from learning to walk to the games of Go and Starcraft (Mnih et al., 2013; 2015; Silver et al., 2016; 2017; 2018; Haarnoja et al., 2018a; Vinyals et al., 2019).

In most cases, the main workhorse allowing artificial intelligence to pass such unprecedented milestones was a variation of a fundamental method to train reinforcement learning models: *policy gradient* (PG) algorithms (Sutton et al., 2000). This algorithm has a disarmingly simple approach to the optimization problem at hand: given a parametrization of the policy, it updates the parameters in the direction of steepest ascent of the associated integrated value function. Impressive progress has been made recently in the understanding of the convergence and optimization properties of this class of algorithms in the tabular setting (Agarwal et al., 2019; Cen et al., 2020; Bhandari & Russo, 2019), in particular leveraging the natural tradeoff between exploration and exploitation offered for entropy-regularized rewards by softmax policies (Haarnoja et al., 2018c; Mei et al., 2020). However, this simple algorithm alone is not sufficient to explain the multitude of recent breakthroughs in this field: in application domains such as Starcraft, robotics or movement planning, the space of possible states and actions are exceedingly large – or even continuous – and can therefore not be represented efficiently by tabular policies (Haarnoja et al., 2018b). Consequently, the recent impressive successes of artificial intelligence would be impossible without the natural choice of neural networks to approximate value functions and / or policy functions in reinforcement learning algorithms (Mnih et al., 2015; Sutton et al., 2000).

While neural networks, in particular deep neural networks, provide a powerful and versatile tool to approximate high dimensional functions on continuous spaces (Cybenko, 1989; Hornik, 1991; Barron, 1993), their intrinsic nonlinearity poses significant obstacles to the theoretical understanding of their training and optimization properties. For example, it is known that the optimization landscape of these models is highly nonconvex, preventing the use of most theoretical tools from classical optimization theory. For this reason, the unprecedented success of neural networks in artificial intelligence is still poorly understood from a theoretical perspective, with only a few recent works tackling the issue of their convergence. Indeed, even for supervised learning, which can be viewed as a special case of reinforcement learning, deep neural networks are still far from being understood despite having been an important and fashionable research focus in recent years. Recent progress has led to an emerging theory for neural network learning including recent works on mean-field point of view of training dynamics (Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Rotskoff et al., 2019; Wei et al., 2018; Chizat & Bach, 2018) and also the linearized training dynamics in the over-parametrized regime (Jacot et al., 2018; Allen-Zhu et al., 2018; Du et al., 2018; 2019; Zou et al., 2018; Allen-Zhu et al., 2019; Chizat et al., 2019;

Oymak & Soltanolkotabi, 2020; Ghorbani et al., 2019; Lee et al., 2019). More specifically to the context of reinforcement learning, some works focusing on value-based learning have recently appeared (Agazzi & Lu, 2019; Cai et al., 2019; Zhang et al., 2020), and others exploring the dynamics of policy gradient algorithms (Zhang et al., 2019). Despite this progress, the theoretical understanding of deep reinforcement learning still poses a significant challenge to the theoretical machine learning community, and it is of crucial importance to understand the convergence and optimization properties of such algorithms and to bridge the gap between theory and practice.

CONTRIBUTIONS.

The main goal of this work is to investigate entropy-regularized policy gradient dynamics for wide, single hidden layer neural networks. In particular, we give the following contributions:

- We give a mean-field formulation of policy gradient dynamics in parameter space, describing the evolution of neural network parameters in the form of a transport partial differential equation (PDE). We prove convergence of the particle dynamics to their mean-field counterpart. We further explore the structure of this problem by showing that such PDE is a gradient flow in the Wasserstein space for the appropriate energy functional.
- We investigate the convergence properties of the above dynamics in the space of measures. In particular, we prove that under some mild assumptions on the initialization of the neural network parameters and on the approximating power of the nonlinearity, all fixed points of the dynamics are global optima, *i.e.*, the approximate policy learned by the neural network is optimal,

RELATED WORKS.

Recent progress in the understanding of the parametric dynamics of simple neural networks trained with gradient descent in the supervised setting has been made in (Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Wei et al., 2018; Chizat, 2019; Chizat & Bach, 2020). These results have further been extended to the multilayer setting in (Nguyen & Pham, 2020). In particular, the paper (Chizat & Bach, 2018) proves optimality of fixed points for wide single layer neural networks leveraging a Wasserstein gradient flow structure and the strong convexity of the loss functional WRT the predictor. We extend these results to the reinforcement learning framework, where the convexity that is heavily leveraged in (Chizat & Bach, 2018) is lost. We bypass this issue by requiring a sufficient expressivity of the used nonlinear representation, allowing to characterize global minimizer as optimal approximators.

The convergence and optimality of policy gradient algorithms (including in the entropy-regularized setting) is investigated in the recent papers (Bhandari & Russo, 2019; Mei et al., 2020; Cen et al., 2020; Agarwal et al., 2019). These references establish convergence estimates through gradient domination bounds. In (Mei et al., 2020; Cen et al., 2020) such results are limited to the tabular case, while (Agarwal et al., 2019; 2020) also discuss neural softmax policy classes, but under a different algorithmic update and assuming certain well-conditioning assumptions along training. Furthermore, all these results heavily leverage the finiteness of action space. In contrast, this paper focuses on the continuous space and action setting with nonlinear function approximation.

Further recent works discussing convergence properties of reinforcement learning algorithms with function approximation via neural networks include (Zhang et al., 2019; Cai et al., 2019). These results only hold for finite action spaces, and are obtained in the regime where the network behaves essentially like a linear model (known as the neural or lazy training regime), in contrast to the results of this paper, which considers training in a nonlinear regime. We also note the work (Wang et al., 2019) where the action space is continuous but the training is again in an approximately linear regime.

2 MARKOV DECISION PROCESSES AND POLICY GRADIENTS

We denote a Markov Decision Process (MDP) by the 5-tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P = P(s, a, s')_{s,s' \in \mathcal{S}, a \in \mathcal{A}}$ an absolutely continuous transition kernel, $r(s, a, s')_{s,s' \in \mathcal{S}, a \in \mathcal{A}}$ is the real-valued, bounded and continuous immediate reward function and $\gamma \in (0, 1)$ is a discount factor. We will consider a probabilistic *policy*, mapping a state to a probability distribution on the action space, so that $\pi : \mathcal{S} \to \mathcal{M}_+^1(\mathcal{A})$, where $\mathcal{M}_+^1(\mathcal{A})$ denotes the space of probability measures on $\mathcal{A}$, and denote for any $s \in \mathcal{S}$ the corresponding density $\pi(s, \cdot) : \mathcal{A} \to \mathbb{R}_+$. We further encourage exploration by defining some (relative) *entropy-regularized* rewards (Williams & Peng, 1991)

$$R_\tau(s, a, s') = r(s, a, s') - \tau D_{\mathrm{KL}}(\pi(s, \cdot); \bar{\pi}(\,\cdot\,))\,,$$

where D_{KL} denotes the relative entropy, $\bar{\pi}$ is a reference measure and τ indicates the strength of regularization. Throughout, we choose $\bar{\pi}$ to be the Lebesgue measure on $\mathcal{A}$, which we assume, like $\mathcal{S}$, to be a compact set of the Euclidean space. This regularization encourages exploration and absolute continuity of the policy WRT Lebesgue measure. Consequently, with some abuse of notation, we use throughout the same notation for a distribution and its density in phase space. Note that the original, unregularized MDP can be recovered in the limit $\tau \to 0$.
In this context, given a policy π the associated *value function* $V_\pi : \mathcal{S} \to \mathbb{R}$ maps each state to the infinite-horizon expected discounted reward obtained by following the policy π and the Markov process defined by P:

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_\tau(s_t, a_t, s_{t+1}) \Big| s_0 = s \right] \\ &= \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t, a_t, s_{t+1}) - \tau D_{\mathrm{KL}}(\pi(s_t, \cdot); \bar{\pi}(\,\cdot\,)) \big) \Big| s_0 = s \right], \end{aligned} \tag{1}$$

where $\mathbb{E}_\pi\left[\,\cdot\,|s_0 = s\right]$ denotes the expectation of the stochastic process s_t starting at $s_0 = s$ and following the (stochastic) dynamics defined recursively by the transition operator $P_\pi(s, \mathrm{d}s') = \int P(s, a, \mathrm{d}s')\pi(s, \mathrm{d}a)$. Correspondingly, we define the Q-function $Q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$\begin{aligned} Q_\pi(s, a) &= \mathbb{E}_\pi \left[r(s_0, a_0, s_1) + \sum_{t=1}^{\infty} \gamma^t R_\tau(s_t, a_t, s_{t+1}) \Big| s_0 = s, a_0 = a \right] \\ &= \bar{r}(s, a) + \gamma \mathbb{E}_\pi \left[V_\pi(s_1) \Big| s_0 = s, a_0 = a \right], \end{aligned} \tag{2}$$

where $\bar{r}(s, a) = \mathbb{E}[r(s, a, s')]$ is the average reward from (s, a). Conversely, from the definition, we have the identity for $V_\pi(s) = \mathbb{E}_\pi\left[Q_\pi(s_0, a_0)|s_0 = s\right] - \tau D_{\mathrm{KL}}(\pi(s, \cdot); \bar{\pi}(\,\cdot\,))$.
We are interested in learning the optimal policy π^* of a given MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, which satisfies for all $s \in \mathcal{S}$

$$V_{\pi^*}(s) = \max_{\pi\,:\,\mathcal{S} \to \mathcal{M}_+^1(\mathcal{A})} V_\pi(s). \tag{3}$$

More specifically we would like to estimate this function through a family of approximators $\pi_w : \mathcal{S} \to \mathcal{M}_+^1(\mathcal{A})$ parametrized by a vector $w \in \mathcal{W} := \mathbb{R}^p$.
A popular algorithm to solve this problem is given by *policy gradient* algorithm (Sutton & Barto, 2018). Starting from an initial condition $w(0) \in \mathcal{W}$, this algorithm updates the parameters w of the predictor in the direction of steepest ascent of the average reward

$$w(t+1) := w(t) + \beta_t \nabla_w \tilde{\mathbb{E}}_{s \sim \varrho_0} V_\pi(s)\,, \tag{4}$$

for a *fixed* absolutely continuous initial distribution of initial states $\varrho_0 \in \mathcal{M}_+^1(\mathcal{S})$ and sequence of time steps $\{\beta_t\}_t$. Here $\tilde{\mathbb{E}}[\cdot]$ denotes an approximation of the expected value operator.
This work focuses on the asymptotic regime of small constant step-sizes $\beta_t \to 0$. In this adiabatic limit, the stochastic component of the dynamics is averaged out before the parameters of the model can undergo a significant change. This allows to consider the parametric update as a deterministic dynamical system emerging from the averaging of the underlying stochastic algorithm corresponding to the limit of infinite sample sizes. This is known as the ODE method (Borkar, 2009) for analyzing stochastic approximation. We focus on the analysis of this deterministic system to highlight the core dynamical properties of policy gradients with nonlinear function approximation. The averaged, deterministic dynamics is given by the set of ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t} w(t) = \mathbb{E}_{s \sim \varrho_0}\left[\nabla_w V_\pi(s)\right] = \mathbb{E}_{s \sim \varrho_\pi, a \sim \pi_w}\left[\nabla_w \log \pi_w(s, a)\left(Q_\pi(s, a) - \tau \log(\pi_w(s, a))\right)\right]\,, \tag{5}$$

where in the second equality we have applied the policy gradient theorem (Sutton et al., 2000; Sutton & Barto, 2018), defining for a fixed $\varrho_0 \in \mathcal{M}_+^1(\mathcal{S})$

$$\varrho^\pi(s_0, s) := \sum_{t=0}^{\infty} \gamma^t P_\pi^t(s_0, s)\,, \qquad \varrho_\pi(s) = \int_{\mathcal{S}} \varrho^\pi(s_0, s)\varrho_0(\mathrm{d}s_0)\,, \tag{6}$$

as the (improper) discounted empirical measure. For completeness, we include a derivation of (5) in Appendix A.

SOFTMAX POLICIES IN THE MEAN-FIELD REGIME

We choose to represent our policy as a softmax policy:

$$\pi_w(s,a) = \frac{\exp(f_w(s,a))}{\int_{\mathcal{A}} \exp(f_w(s,a))\mathrm{d}a}$$

and parametrize the *energy* f as a two-layer neural network in the *mean-field* regime, *i.e.*,

$$f_w(s,a) = \frac{1}{N}\sum_{i=1}^{N} \psi(s,a;w_i)$$

for a fixed, usually nonlinear function $\psi : \mathcal{S}\times\mathcal{A}\times\Omega \to \mathbb{R}$, where we have separated $w \in \mathcal{W}$ into N identical components $w_i \in \Omega$ so that $\mathcal{W} = \Omega^N$.

We can rewrite the above expression in terms of an empirical measure:

$$f_{\nu^{(N)}}(s,a) := \int_{\Omega} \psi(s,a;\omega)\nu^{(N)}(\mathrm{d}\omega) \qquad \text{where} \quad \nu^{(N)}(\mathrm{d}\omega) = \frac{1}{N}\sum_{i=1}^{N} \delta_{w^{(i)}}(\mathrm{d}\omega) \in \mathcal{M}_+^1(\Omega). \tag{7}$$

This empirical measure representation removes the symmetry of the approximating functions under permutations of parameters w_i. It also facilitates the limit $N \to \infty$, when $\nu^{(N)} \to \nu$ weakly, so that $f_{\nu^{(N)}} \to f_\nu$. Then, for a general distribution $\nu \in \mathcal{M}_+^1(\Omega)$ the softmax mean-field policy reads:

$$\pi_\nu(s,a) = \frac{\exp\left(\int_\Omega \psi(s,a;\omega)\nu(\mathrm{d}\omega)\right)}{\int_{\mathcal{A}} \exp\left(\int_\Omega \psi(s,a;\omega)\nu(\mathrm{d}\omega)\right)\mathrm{d}a}. \tag{8}$$

Note that by our choice of softmax policy and mean-field parametrization (7) we have

$$\begin{aligned}
\nabla_{w_i} \log \pi_{\nu^{(N)}}(s,a) &= \nabla_{w_i} f_{\nu^{(N)}}(s,a) - \nabla_{w_i} \log \int_{\mathcal{A}} \exp f_{\nu^{(N)}}(s,a)\mathrm{d}a \\
&= \nabla_{w_i} \frac{1}{N}\sum_{i=1}^{N} \psi(s,a;w_i) - \frac{\int_{\mathcal{A}} \nabla_{w_i} \exp\left[\frac{1}{N}\sum_{i=1}^{N}\psi(s,a;w_i)\right]\mathrm{d}a}{\int_{\mathcal{A}} \exp f_{\nu^{(N)}}(s,a)\mathrm{d}a} \\
&= \frac{1}{N}\left(\nabla_{w_i}\psi(s,a;w_i) - \int_{\mathcal{A}} \nabla_{w_i}\psi(s,a;w_i)\pi_{\nu^{(N)}}(s,\mathrm{d}a)\right).
\end{aligned}$$

Thus the training dynamics (5), after an appropriate rescaling of time ($t \mapsto t/N$, which is due to the mean-field parametrization for f_w), can be rewritten as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} w_i(t) = \int_{\mathcal{S}\times\mathcal{A}} &\left(\nabla_{w_i}\psi(s,a;w_i) - \mathbb{E}_{\pi_{\nu^{(N)}}}\left[\nabla_{w_i}\psi(s,\cdot;w_i)\right]\right) \times \\
&\times \left(Q_{\pi_{\nu^{(N)}}}(s,a) - \tau \log \pi_{\nu^{(N)}}(s,a)\right)\pi_{\nu^{(N)}}(s,\mathrm{d}a)\varrho_{\pi_{\nu^{(N)}}}(\mathrm{d}s).
\end{aligned} \tag{9}$$

The training dynamics can be more compactly represented by the evolution of the measure $\nu \in \mathcal{M}_+^1(\Omega)$ in parameter space, given by a *mean-field* transport partial differential equation of the Vlasov type as

$$\frac{\mathrm{d}}{\mathrm{d}t}\nu_t(\omega) = \mathrm{div}\left(\nu_t(\omega)\int_{\mathcal{S}\times\mathcal{A}} C_{\pi_\nu}[\nabla_\omega\psi(s,\cdot;\omega),\, Q_{\pi_\nu} - \tau\log\pi_\nu](s)\varrho_{\pi_\nu}(\mathrm{d}s)\right), \tag{10}$$

where $\omega \in \Omega$ and we have introduced the shorthand $C_\pi[f,g](s)$ to denote the covariance operator WRT the probability measure $\pi(s,\mathrm{d}a)$. Note that the above partial differential equation also captures the dynamics of the finite-width system, *i.e.*, of the empirical measure $\nu^{(N)}$ where each w_i follows (9).

We further note that the dynamics introduced above have a gradient flow structure in the probability space $\mathcal{M}_+^1(\Omega)$: defining the expected value function

$$\mathcal{E}[\nu] = \mathbb{E}_{s_0\sim\varrho_0}\left[V_{\pi_\nu}(s_0)\right] \tag{11}$$

the dynamics (10) is a gradient flow for $\mathcal{E}$ in the Wasserstein space (see e.g., (Santambrogio, 2017) for an introduction), as we prove in the appendix:

Proposition 2.1. *For a fixed initial distribution $\varrho_0 \in \mathcal{M}_+^1(\mathcal{S})$, the dynamics (10) is the Wasserstein gradient flow of the energy functional (11).*

Analogous dynamics equation for evolution of parameter space measure in the supervised learning case has been derived in (Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Chizat & Bach, 2018) and in the TD learning case in (Agazzi & Lu, 2019; Zhang et al., 2020). In particular, in the case of supervised learning, the resulting dynamics is a Wasserstein gradient flow, the structure of which is used to obtain the convergence of the particle system to the mean-field dynamics. In our case, however, the energy functional is not convex WRT the policy and moreover the softmax parametrization destroys the convexity of the approximator of the policy with respect to ν_t. Thus showing convergence of the dynamics becomes much more challenging.

3 SIMPLIFIED SETTING: THE BANDIT PROBLEM

We now introduce our results in the simple bandit setting in, where state space $\mathcal{S}$ is one point (and will be henceforth suppressed in the notation) and without loss of generality action space $\mathcal{A}$ is continuous. In this case, for a reward function r and a softmax policy

$$\pi_\nu(a) = \frac{\exp(f_\nu(a))}{\int_{\mathcal{A}} \exp(f_\nu(a))\mathrm{d}a}$$

we have that the value function for the regularized problem reads (we denote $V_\nu = V_{\pi_\nu}$ to simplify notation)

$$V_\nu = \int (r(a) - \tau \log \pi_\nu(a))\pi_\nu(\mathrm{d}a)\,,$$

while the Q function is simply $Q(a) = r(a)$. We further note that the optimal policy in the regularized case reads:

$$\pi^*(a) = Z^{-1}\exp(\tau^{-1}r(a)), \qquad Z = \int_{\mathcal{A}} \exp(\tau^{-1}r(a))\mathrm{d}a$$

Recalling the definition of the covariance operator $C_\pi[f,g](s)$ from (10), the expression for the policy gradient vector field in the latter case simplifies to

$$\begin{aligned}\partial_t\omega_t := F_t(\omega_t;\nu_t) &= \nabla_\omega D\pi_\nu DV_\nu = C_{\pi_\nu}[\nabla_\omega\psi(a;\omega), r - \tau\log(\pi_\nu)] \\ &= \int_{\mathcal{A}} \left(\nabla_\omega\psi(a;\omega) - \int \nabla_\omega\psi(a';\omega)\pi_\nu(\mathrm{d}a')\right)(r(a) - \tau f_\nu(a))\,\pi_\nu(\mathrm{d}a)\,, \qquad (12)\end{aligned}$$

where $D\pi_\nu, DV_\nu$ denote the Fréchét derivative of π_ν and V_π WRT ν and π respectively. Note that by the structure of the covariance operator C_π, adding a constant to the function $f_\nu(\cdot)$ does not affect the dynamics. This reflects the fact that the softmax policy is normalized by definition.

3.1 GLOBAL OPTIMALITY OF SOFTMAX POLICY GRADIENT

We now sketch the main steps in proving that the mean-field policy gradient dynamics converge, under appropriate assumptions, to global optimizers. The proof in this simpler setting is much more transparent than the general case to be discussed in the next section, and will provide some intuitions for the latter. The first part of the proof concerns the properties of fixed points of the dynamics (10), while the second part concerns the training dynamics.

STATICS

We first informally prove global optimality of any fixed point ν^* of the transport equation $\frac{\mathrm{d}}{\mathrm{d}t}\nu_t = -\mathrm{div}(\nu_t F_t)$ with F_t from (12) such that

a) ν^* has full support in Ω,

b) the nonlinearity is 1-homogeneous in the first component of its parameters, *i.e.*, that writing $\omega = (\omega_0, \bar\omega) \in \mathbb{R}\times\Theta$ one has $\psi(a;\omega) := \omega_0\phi(a;\bar\omega)$ for a regular enough $\phi : \mathcal{A}\times\Theta \to \mathbb{R}$,

c) the span of $\{\phi(a;\bar\omega)\}_{\bar\omega\in\Theta}$ is *dense* in $L^2(\mathcal{A})$,

so that *i.e.*, $\pi_{\nu^*}(a) = \pi^*(a) = Z^{-1} \exp\left[\tau^{-1} r(a)\right]$. Weaker assumptions and the general statement are given in the next section (Proposition 4.3) while the general proof appears in the appendix.

First, we note that by assumption a), $\operatorname{div}(\nu^* F(\,\cdot\,;\nu^*)) = 0$ directly implies that for all $\omega \in \Omega$

$$F(\omega;\nu^*) = \int_{\mathcal{A}} \nabla_\omega \psi(a;\omega)\,(r(a) - \tau f_{\nu^*}(a) - V_{\nu^*})\,\pi_{\nu^*}(\mathrm{d}a) = 0\,.$$

In particular, by homogeneity assumption b), the first component of the above vector field must vanish on Θ.

$$\int_{\mathcal{A}} \phi(a;\bar\omega)\,(r(a) - \tau f_{\nu^*}(a) - V_{\nu^*})\,\pi_{\nu^*}(\mathrm{d}a) = 0\,.$$

By assumption c) that span of ϕ is dense in $L^2(\mathcal{A})$ the above implies that

$$r(a) - \tau f_{\nu^*}(a) - V_{\nu^*} = 0 \qquad \pi_{\nu^*}\text{-a.e. in } \mathcal{A}\,. \tag{13}$$

Finally, recalling that by the softmax parametrization and by the boundedness of ϕ, $\pi_{\nu^*}(a) > 0$, we must have

$$f_{\nu^*}(a) = \tau^{-1} r(a) + C$$

which directly implies the optimality of the policy.

DYNAMICS

While it is clear that assumption b) and c) about the structure and approximating power of the nonlinearity ψ hold independently of t, we want to show that assumption a) also holds uniformly in time. In this sense, the continuity of the vector field (12) will preserve the full support properties of the measure ν_t for all $t > 0$, as we will prove in a more general framework in Lemma C.1. Consequently, any measure ν respecting assumption a) at initialization will do so for any finite positive time $t > 0$. However, the question remains of whether this property still holds at $t = \infty$. This is the object of Lemma C.2, where we prove that whenever the gradient approaches a fixed point in parameter space, if this fixed point is not a global minimizer, it must be avoided by the dynamics, and thus, the only possible fixed points of the dynamics are global minimizers.

4 RESULTS IN THE GENERAL SETTING

We now come back to the general MDP framework introduced in Section 2.

4.1 ASSUMPTIONS

To state the main result of this section, the optimality of fixed points of (10) we need the following

Assumption 1. *Assume that $\omega = (\omega_0, \bar\omega) \in \mathbb{R} \times \Theta$ for $\Theta = \mathbb{R}^{m-1}$ and $\psi(s,a;\omega) = \omega_0 \phi(s,a;\bar\omega)$ with*

a) Regularity of ϕ: *ϕ is bounded, differentiable and $D\phi_\omega$ is Lipschitz. Also, for all $f \in L^2(\mathcal{S}\times\mathcal{A})$ the regular values of the map $\bar\omega \mapsto g_f(\bar\omega) := \int f(s,a)\phi(s,a;\bar\omega)$ are dense in its range, and $g_f(r\bar\omega)$ converges in $C^1(\{\bar\omega \in \Theta : \|\bar\omega\|_2 = 1\})$ as $r \to \infty$ to a map $\bar g_f(\bar\omega)$ whose regular values are dense in its range.*

b) Universal approximation*: the span of $\{\phi(\cdot,\bar\omega) \,:\, \bar\omega \in \Theta\}$ is dense in $L^2(\mathcal{S}\times\mathcal{A})$;*

c) Support of the measure*: There exists $r > 0$ s.t. the support of the initial condition ν_0 is contained in $\mathcal{Q}_r := [-r,r] \times \Theta$ and separates $\{-r\}\times\Theta$ from $\{r\}\times\Theta$, i.e., any continuous path connecting $\{-r\}\times\Theta$ to $\{r\}\times\Theta$ intersects the support of ν_0.*

Assumption 1 a) is a common, technical regularity assumption ensuring that (10) is well behaved and controlling the growth, variation and regularity of ϕ. Alternative assumptions on the case $\Theta \neq \mathbb{R}^{m-1}$ are given in the appendix. Assumption 1 b) speaks to the approximating power of the nonlinearity, assumed to be expressive enough to approximate any function in $L^2(\mathcal{S}\times\mathcal{A})$. This condition replaces the convexity assumption from Chizat & Bach (2018), as the lack of convex structure in our setting prevents us from identifying the local and global minimization properties of a fixed point. Indeed, despite the one-point convexity of $\mathbb{E}_\varrho\left[V_\pi(s)\right]$ as a functional of π (Kakade & Langford, 2002) which can be leveraged in the tabular case, such property will be lost, in general, when restricting to policies through nonlinear function approximation. We bypass this issue by requiring sufficient expressivity of the approximating function class, guaranteeing that the optimal policy can be represented with arbitrary precision. Similar assumption on approximability of neural network representation was made in recent analysis of natural policy gradient algorithm (Agarwal et al., 2019). We note that this assumption is easily satisfied by widely used nonlinearities by the universal approximation theorem (Cybenko, 1989; Barron, 1993). Finally, Assumption 1 c) guarantees that the initial condition is such that the expressivity from b) can actually be exploited.

4.2 CONVERGENCE OF THE MANY-PARTICLE LIMIT

Before discussing the optimality properties of the dynamics (10), we show that this PDE accurately describes the policy gradient dynamics of a sufficiently wide, single layer neural network. To this aim, we let $\mathcal{P}_2(\Omega)$ be the space of probability distributions on Ω with finite second moment.

Theorem 4.1. *Let Assumption 1 hold and let $w_t^{(N)}$ be a solution of (5) with initial condition $w_0^{(N)} \in \mathcal{W} = \Omega^N$. If $\nu_0^{(N)}$ converges to $\nu_0 \in \mathcal{P}_2(\Omega)$ in Wasserstein distance W_2 then $\nu_t^{(N)}$ converges, for every $t > 0$, to the unique solution ν_t of (10).*

We note that by the law of large numbers for empirical distributions, the condition of convergence of $\nu_0^{(N)}$ to ν_0 is *e.g.*, satisfied when $w_0^{(i)}$ are drawn independently at random from ν_0.

The proof of this result is largely standard, under the given assumptions, and is provided in the appendix for completeness. The idea of the proof is a canonical propagation of chaos argument (Sznitman, 1991). In a nutshell the first step of the argument establishes sufficient regularity of the gradient dynamics, allowing to guarantee existence and uniqueness of the solution to (10). Then, one bounds the difference in differential updates for the particle system and the mean-field dynamics by comparing them with the evolution of the particle system according to a *linear, time-inhomogeneous* PDE using the drift term of the mean-field model. The proof is finally concluded by application of Gronwall inequality.

4.3 OPTIMALITY

After discussing the connection between particle dynamics and mean-field equations, we present the main convergence result of this paper:

Theorem 4.2. *Let Assumption 1 hold and ν_t given by* (10) *converge to ν^*, then $\pi_{\nu^*} = \pi^*$.*

Thus if the policy gradient dynamics (10) converges to a stationary point, it is a global minimizer. To prove this result, we first connect the optimality of a stationary point with the support of the underlying measure in parameter space.

Proposition 4.3. *Let ν^* be a stationary point of* (10) *such that the full support property of Assumption 1 c) holds at ν^*, then $\pi_{\nu^*} = \pi^*$.*

This intermediate result is obtained in two steps: we first show that, by homogeneity of ψ and Assumption 1 c), the velocity field of (10) must vanish a.e. in $\Omega = \mathbb{R} \times \Theta$. Then, by Assumption 1 b) we deduce that a stationary point ν^* cannot result from the orthogonality between Df_ν and the error estimate $\delta(\nu) = (Q_{\pi_\nu}(s,a) - \tau \log \pi_\nu(s,a) - V_{\pi_\nu}(s))$, implying $\delta(\nu) = 0$.

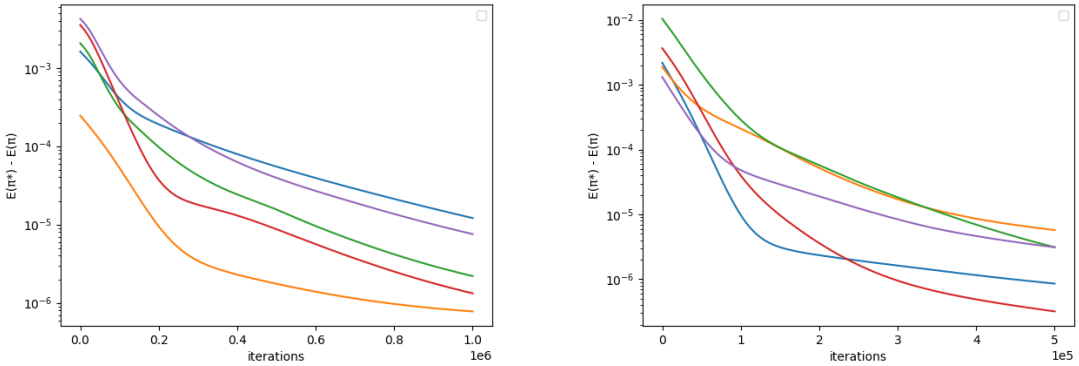
In light of Proposition 4.3, one important technical step in the proof of Theorem 4.2 is to show that the policy gradient dynamics in the mean-field regime preserves Assumption 1 c) on the support of the measure throughout training, as shown in Lemma C.1. This is true for any finite time by topological arguments: the separation property of the measure cannot be altered by the action of a continuous flow such as (10). To extend to infinite time (and thus to the stationary point) leveraging the above partial result, we finally prove in Lemma C.2, that spurious fixed points are avoided by the policy gradient dynamics (10) when initialized properly. To establish this we argue by contradiction: assuming that we are approaching such a spurious fixed point $\tilde{\nu}$ at time t_0, we show in Lemma C.4 that the velocity field will change little for any $t > t_0$. In particular, it follows that in this regime the dynamics of (10) can be approximated by the gradient descent dynamics (in particle space) of an approximately *fixed* potential. On the other hand, by the homogeneity of ψ and by Assumption 1 c), we show that a positive amount of measure $\tilde{\nu}$ will fall in a forward invariant region where its ω_0 component will grow linearly in t, thereby eventually contradicting the assumption that $\tilde{\nu}$ is a fixed point of (10).

5 NUMERICAL EXAMPLES

To test our theoretical results in a simple setting we train a wide, single hidden layer neural network with policy gradients to learn the optimal softmax policy (8) for entropy-regularized rewards with parametrization (7) and regularization parameter $\tau = 0.2$. We do so in two separate settings:

(a) $\mathcal{S} = \{0\}$, $\mathcal{A} = [0,1]$. This setting corresponds to bandits framework discussed in Section 3.

(b) $\mathcal{S} \times \mathcal{A}$ is a grid of size 100×100 in the set $[0,1]^2$. In this case, we have chosen a discount factor of $\gamma = 0.7$ and a transition process given by $P(s,a,s') = 0.9\delta(s'-a) + 0.1/100$ (*i.e.*, an action a leads to



(a) (b)

Figure 1: Evolution of $\mathcal{E}(\nu^*) - \mathcal{E}(\nu_t)$ as a function of training time, for experiments (a) and (b) as described in the main text. Different lines correspond to different random initializations of reward function and learning model. We see that the training error decreases monotonically (but with different rates) during trainig.

the corresponding state $s' = a$ with probability 0.9 and is uniformly distributed with probability 0.1). At each iteration we have computed the *exact* distribution ϱ_π by computing the resolvent of the (weighted) transition matrix.

In both cases, we defined the optimal Q function as $Q^*(s,a) = \tau f_w^*(s,a)$, where $f_w^*(s,a)$ is given by a single hidden layer neural network of width $n = 5$, ReLU nonlinearites and weights w drawn independently and identically distributed from a centered, normal distribution with variance $\sigma^2 = 4$, *i.e.*,

$$Q^*(s,a) = \tau f_w^*(s,a) \qquad \text{for } w_i \sim \mathcal{N}(0,4)\,.$$

We learn the optimal policy for the problem defined above using a $N = 800$-neurons wide single hidden layer neural network with ReLU nonlinearities in the mean-field regime (7) used as energy for a softmax policy (8). The initialization of the student network is as follows: first-layer weights are initialized at random drawn independently from a centered, normal distribution with variance $\sigma^2 = 4$, while output weights are initialized at 0. The model is trained according to (4) with fixed step-size $\beta_t = 10^{-3}$. We report the results of this training procedure in Fig. 1, where we notice that all the paths monotonically decrease the error $\mathcal{E}(\nu^*) - \mathcal{E}(\nu_t)$, as predicted by our results. Note that the convergence rate of the model varies across experiments, consistently with the purely qualitative nature of the convergence result we proved.

6 Conclusions and Future Work

This work addresses the problem of optimality of policy gradient algorithms, a workhorse of deep reinforcement learning, when combined with mean-field models such as neural networks. More specifically, we provide a mean-field formulation of the parametric dynamics of policy gradient algorithms for entropy-regularized MDPs and prove that, under mild assumptions, all fixed points of such dynamics are optimal. This extends similar results obtained in the "neural" or "lazy" regime to the mean-field one, which is known to be much more expressive (E et al., 2019; Ghorbani et al., 2020), but also highly nonlinear. The latter feature prevents, at present, from obtaining convergence results of these models, except in very specific settings (Chizat, 2019; Javanmard et al., 2019).

Interesting avenues or future research include the relaxing the adiabaticity assumption, *i.e.*, considering the stochastic approximation problem resulting from the finite number of samples and the finite gradient step-size, as well as establishing quantitative bounds for models with a large, but finite, number of parameters. Probably the most important open question, however, concerns establishing quantitative convergence of mean-field dynamics of neural networks: even in the supervised setting, despite recent results in specific settings (Chizat, 2019; Javanmard et al., 2019), these guarantees remain mainly out of reach.

REFERENCES

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020.

Andrea Agazzi and Jianfeng Lu. Temporal-difference learning for nonlinear value function approximation in the lazy training regime. *arXiv preprint arXiv:1905.10917*, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6155–6166, 2019.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, pp. 11315–11326, 2019.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 3040–3050, 2018.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pp. 2933–2943, 2019.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989.

S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks, 2018. preprint, arXiv:1810.02054.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685, 2019.

Weinan E, Chao Ma, and Lei Wu. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension, 2019. arXiv preprint arXiv:1904.12191.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *arXiv preprint arXiv:2006.13409*, 2020.

T. Haarnoja, S. Ha, A. Zhou, J. Tan, G. Tucker, and S. Levine. Learning to walk via deep reinforcement learning, 2018a. preprint, arXiv:1812.11103.

Tuomas Haarnoja, Sehoon Ha, Aurick Zhou, Jie Tan, George Tucker, and Sergey Levine. Learning to walk via deep reinforcement learning. *arXiv preprint arXiv:1812.11103*, 2018b.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870, 2018c.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991. ISSN 0893-6080.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *arXiv preprint arXiv:1901.01375*, 2019.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274. Morgan Kaufmann Publishers Inc., 2002.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pp. 8570–8581, 2019.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. *arXiv preprint arXiv:2005.06392*, 2020.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*. 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, (7540):529–533, 02 2015.

Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv preprint arXiv:2001.11443*, 2020.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 7146–7155. Curran Associates, Inc., 2018.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In *International Conference on Machine Learning*, pp. 5508–5517, 2019.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550:354, October 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pp. 165–251. Springer, 1991.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

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

C. Wei, J. D. Lee, Q. Liu, and T. Ma. On the margin theory of feedforward neural networks, 2018. preprint, arXiv:1810.05369.

Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

Yufeng Zhang, Qi Cai, Zhuoran Yang, Yongxin Chen, and Zhaoran Wang. Can temporal-difference and q-learning learn representation? a mean-field theory. *arXiv preprint arXiv:2006.04761*, 2020.

D. Zou, Y. Cao, D. Zhou, and Q. Gu. Stochastic gradient descent optimizes over-parametererized deep ReLU networks, 2018. preprint, arXiv:1811.08888.

A DERIVATION OF SOFTMAX POLICY GRADIENT DYNAMICS

Lemma A.1. *The gradient of the entropy-regularized value function can be written as*

$$\mathbb{E}_{s\sim\varrho_0}\left[\nabla_w V_\pi(s)\right] = \mathbb{E}_{s\sim\varrho^\pi, a\sim\pi_w}\left[\nabla_w \log\pi_w(s,a)\left(Q_\pi(s,a) - \tau\log\pi_w(s,a)\right)\right],$$

and thus the policy gradient dynamics (5).

Proof. We choose throughout $\bar\pi$ as the Lebesgue measure, and use that π_w is absolutely continuous wrt $\bar\pi$. Taking the gradient of (1) using a parametric policy π_w we obtain

$$\begin{aligned}
\nabla_w V_{\pi_w}(s) &= \nabla_w \mathbb{E}_{\pi_w}\left[\sum_{t=0}^\infty \gamma^t\left(r(s_t,a_t,s_{t+1}) - \tau D_{\mathrm{KL}}(\pi_w(s_t,\cdot);\bar\pi(\,\cdot\,))\right)\Big| s_0 = s\right] \\
&= \nabla_w \int_{\mathcal{S}\times\mathcal{A}}\left(r(s,a_1,s_1) - \tau\log\frac{\pi_w(s,a_1)}{\bar\pi(a_1)} + \gamma V_{\pi_w}(s_1)\right)P(s,a_1,\mathrm{d}s_1)\pi_w(s,\mathrm{d}a_1) \\
&= \int_{\mathcal{S}\times\mathcal{A}}\left(r(s,a_1,s_1) - \tau\log\frac{\pi_w(s,a_1)}{\bar\pi(a_1)} + \gamma V_{\pi_w}(s_1)\right)P(s,a_1,\mathrm{d}s_1)\nabla_w\pi_w(s,\mathrm{d}a_1) \\
&\qquad + \int_{\mathcal{S}\times\mathcal{A}}\left(-\tau\nabla_w\log\frac{\pi_w(s,a_1)}{\bar\pi(a_1)} + \gamma\nabla_w V_{\pi_w}(s_1)\right)P(s,a_1,\mathrm{d}s_1)\pi_w(s,\mathrm{d}a_1) \qquad \text{(A.1)}
\end{aligned}$$

Now, since $\int_{\mathcal{S}} P(s,a,\mathrm{d}s') = 1$ and $\int_{\mathcal{A}}\pi_w(s,\mathrm{d}a') = 1$ for all s,a,w we have for the first term in brackets in the last line

$$\int_{\mathcal{S}\times\mathcal{A}}\nabla_w\log\frac{\pi_w(s,a_1)}{\bar\pi(a_1)}P(s,a_1,\mathrm{d}s_1)\pi_w(s,\mathrm{d}a_1) = \int_{\mathcal{A}}\nabla_w\pi_w(s,a_1)\mathrm{d}a_1 = \nabla_w\int_{\mathcal{A}}\pi_w(s,a_1)\mathrm{d}a_1 = 0$$

On the other hand, we can rewrite the second term in brackets as $\mathbb{E}_{\pi_w}\left[\gamma\nabla V_{\pi_w}(s_1)|s_0 = s\right]$, and recognize the LHS of (A.1) evaluated at the next state s_1 in the expectation. Therefore, we can sequentially repeat the same computation as above, and recalling the definition of $\varrho^{\pi_w}(\,\cdot\,)$ and $Q_\pi(s,a)$ in (6) and (2) we obtain

$$\begin{aligned}
\nabla_w V_{\pi_w}(s) &= \int_{\mathcal{A}}\sum_{t=0}^\infty\gamma^t\left(r(s_t,a_t,s_{t+1}) - \tau\log\frac{\pi_w(s_t,a_t)}{\bar\pi(a_t)} + \gamma V_{\pi_w}(s_{t+1})\right)\nabla_w\pi_w(s_t,\mathrm{d}a_t)\Big|_{s_0=s} \\
&= \mathbb{E}_{\pi_w}\left[\sum_{t=0}^\infty\gamma^t\left(Q_{\pi_w}(s_t,a_t) - \tau\log\frac{\pi_w(s_t,a_t)}{\bar\pi(a_t)}\right)\nabla_w\log\pi_w(s_t,a_t)\Big| s_0 = s\right] \\
&= \int_{\mathcal{S}\times\mathcal{A}}\left(Q_{\pi_w}(s_t,a_t) - \tau\log\frac{\pi_w(s,a)}{\bar\pi(a)}\right)\nabla_w\log\pi_w(s,a)\,\varrho^{\pi_w}(\mathrm{d}s)\pi_w(s,\mathrm{d}a),
\end{aligned}$$

where in the second line we have used that, if $\pi_w > 0$ on $\mathcal{A}$, $\pi_w(s_t,\mathrm{d}a_t)\nabla_w\log\pi_w(s_t,a_t) = \nabla_w\pi_w(s_t,\mathrm{d}a_t)$. □

Proposition 2.1. *For a fixed initial distribution ϱ_0, the dynamics (10) is the Wasserstein gradient flow of the energy functional*

$$\mathcal{E}[\nu] = \mathbb{E}_{s_0\sim\varrho_0}\left[V_{\pi_\nu}(s_0)\right].$$

Proof. We find the potential of the gradient flow by functional differentiation of $\mathcal{E}$:

$$\frac{\delta}{\delta\nu}\mathcal{E}[\nu](\omega) = \int_{\mathcal{A}}\frac{\delta\mathcal{E}}{\delta\pi}(s,a)\frac{\delta\pi_\nu}{\delta\nu}(s,a;\omega)\mathrm{d}s\,\mathrm{d}a \qquad \text{(A.2)}$$

and consider the two terms in the integral separately, starting from the second:

$$\begin{aligned}
\frac{\delta\pi_\nu}{\delta\nu}(s,a;\omega) &= \frac{\delta}{\delta\nu}\frac{e^{\tau\int\psi(s,a;\omega)\nu(\mathrm{d}\omega)}}{\int_{\mathcal{A}}e^{\tau\int\psi(s,a;\omega)\nu(\mathrm{d}\omega)}\mathrm{d}a} \\
&= \frac{1}{\int_{\mathcal{A}}e^{\tau\int\psi(s,a;\omega)\nu(\mathrm{d}\omega)}\mathrm{d}a}\frac{\delta}{\delta\nu}e^{\tau\int\psi(s,a;\omega)\nu(\mathrm{d}\omega)}
\end{aligned}$$

$$- \frac{e^{\tau \int \psi(s,a;\omega)\nu(\mathrm{d}\omega)}}{\left(\int_{\mathcal{A}} e^{\tau \int \psi(s,a;\omega)\nu(\mathrm{d}\omega)} \mathrm{d}a\right)^2} \int_{\mathcal{A}} \frac{\delta}{\delta \nu} e^{\tau \int \psi(s,a;\omega)\nu(\mathrm{d}\omega)} \mathrm{d}a$$

$$= \tau \left(\psi(s,a;\omega) - \int_{\mathcal{A}} \psi(s,a';\omega)\pi_\nu(s,\mathrm{d}a') \right) \pi_\nu(s,a) \tag{A.3}$$

For the first term in the integrand of (A.2), we use $\pi(s,a)$ as a density and obtain

$$\frac{\delta \mathcal{E}}{\delta \pi}(s,a) = \frac{\delta}{\delta \pi} \left[\int_{\mathcal{S}\times\mathcal{A}} \left(\bar{r}(s,a) - \tau \log \pi(s,a)\right) \left(\varrho_\pi(\mathrm{d}s)\pi(s,\mathrm{d}a)\right) \right](s,a)$$

$$= \int_{\mathcal{S}} \Bigg[\left(\bar{r}(s,a) - \tau(\log \pi(s,a) + 1)\right) \varrho^\pi(s',s) \tag{A.4}$$

$$+ \int_{\mathcal{S}\times\mathcal{A}} \left(\bar{r}(s'',a'') - \tau \log \pi(s'',a'')\right) \frac{\delta \varrho^\pi(s',s'')}{\delta \pi}(s,a)\pi(s'',\mathrm{d}a'')\mathrm{d}s'' \Bigg] \varrho_0(\mathrm{d}s'), \tag{A.5}$$

where in the last line we have used that $\frac{\delta}{\delta \pi}[\pi \log \pi](s,a) = \log \pi(s,a) + 1$.
We evaluate the variational derivative of ϱ^π as

$$\frac{\delta \varrho^\pi(s',s'')}{\delta \pi}(s,a) = \sum_{t=0}^{\infty} \gamma^t \frac{\delta}{\delta \pi} \left[\int P_\pi^t(s',s'') \right](s,a)$$

$$= \gamma \int_{\mathcal{S}} \frac{\delta P_\pi(s',s''')}{\delta \pi} \left(\sum_{t=0}^{\infty} \gamma^t P_\pi^t(s''',s'') \right) + P_\pi(s',s''') \frac{\delta}{\delta \pi} \sum_{t=0}^{\infty} \gamma^t \frac{\delta}{\delta \pi} P_\pi^t(s''',s'')\mathrm{d}s'''$$

$$= \gamma \left[\int_{\mathcal{S}} \delta(s,s')P(s',a,s''')\varrho^\pi(s''',s'') + P_\pi(s',s''') \frac{\delta \varrho^\pi(s''',s'')}{\delta \pi}(s,a)\mathrm{d}s''' \right]$$

We further recognize the same derivative on the RHS of the above expression, allowing to write

$$\frac{\delta \varrho^\pi(s',s'')}{\delta \pi}(s,a) = \sum_{t=0}^{\infty} \gamma^t P_\pi^t(s',s) \int_{\mathcal{S}} P(s,a,s''')\varrho^\pi(s''',s'')\mathrm{d}s'''$$

Finally, we notice that the last term in (A.4) is constant in a and therefore vanishes when integrated against (A.3), so we only consider

$$\frac{\delta \mathcal{E}}{\delta \pi}(s,a) - \varrho_\pi(s) = \int_{\mathcal{S}^2\times\mathcal{A}} \left(\bar{r}(s'',a'') - \tau \log \pi(s'',a'')\right) \times$$

$$\times \left(\delta_{s,s''}\delta_{a,a''}\varrho^\pi(s',s) + \pi(s'',a'') \frac{\delta \varrho^\pi(s',s'')}{\delta \pi}(s,a) \right) \varrho_0(\mathrm{d}s')\mathrm{d}s''\mathrm{d}a''$$

$$= \int_{\mathcal{S}^2\times\mathcal{A}} \sum_{t=0}^{\infty} \gamma^t P_\pi^t(s',s) \left(\bar{r}(s'',a'') - \tau \log \pi(s'',a'')\right) \times$$

$$\times \left(\delta_{s,s''}\delta_{a,a''} + \gamma \int_{\mathcal{S}} P(s,a,s''')\varrho^\pi(s''',s'')\mathrm{d}s''' \right) \varrho_0(\mathrm{d}s')\mathrm{d}s''\mathrm{d}a''$$

$$= \sum_{t=0}^{\infty} \gamma^t \int_{\mathcal{S}} P_\pi^t(s',s) \left(Q_\pi(s,a) - \tau \log \pi(s,a)\right) \varrho_0(\mathrm{d}s')$$

$$= \left(Q_\pi(s,a) - \tau \log \pi(s,a)\right) \varrho_\pi(s) \tag{A.6}$$

We conclude by noting that combining (A.3) and (A.6) we obtain $C_\pi[\psi(\omega), Q_\pi - \tau \log \pi](s)$, and the Wasserstein gradient flow corresponding to this potential is (10). □

B PROOFS OF THE MANY-PARTICLE LIMIT

Theorem 4.1. *Let Assumption 1 hold and let $w_t^{(N)}$ be a solution of (5) with initial condition $w_0^{(N)} \in \mathcal{W} = \Omega^n$. If $\nu_0^{(N)}$ converges to $\nu_0 \in \mathcal{P}_2(\Omega)$ in Wasserstein distance W_2 then $\nu_t^{(N)}$ converges, for every $t > 0$, to the unique solution ν_t of (10).*

Proof. As anticipated in the main text, the proof is divided in two parts:

1. We prove sufficient regularity of the dynamics (10), allowing to establish existence and uniqueness of its solution.
2. We leverage the regularity proven above to establish a propagation of chaos result, showing that the system of interacting particles behaves asymptotically as its mean-field limit.

While carrying out this proof is needed in our context since the dependence of $\mathcal{E}(\nu)$ on ν is more involved than in *e.g.*, Mei et al. (2018); Rotskoff & Vanden-Eijnden (2018); Chizat & Bach (2018), the steps of this derivation are mainly standard, see e.g., (Sznitman, 1991).

B.1 Regularity

We prove existence and uniqueness of the gradient flow dynamics (10) through standard arguments from the optimal transportation literature (see *e.g.*, (Ambrosio et al., 2008)). More specifically, recalling that $\pi = \pi_\nu$ we leverage the Lipschitz continuity of the vector field

$$F_t(\omega, \nu) = C_\pi \left[\nabla_\omega \psi(s, a; \omega), Q_\pi(s, a) - \tau \log \pi\right]$$

with respect to ν. To prove such regularity result, decompose

$$\mathcal{E}(\nu) = R\left(\int \psi\nu\right) \qquad \text{for} \quad R(f) = S \circ \pi(f)\,. \tag{B.1}$$

and $S \,:\, (\mathcal{S} \to \mathcal{M}_+^1(\mathcal{A})) \to \mathbb{R}$ maps $\mu \mapsto \mathbb{E}_\mu\left[V_\mu(s) | s \sim \varrho_0\right]$ and $\pi \,:\, L^2(\mathcal{S} \times \mathcal{A}) \to (\mathcal{S} \to \mathcal{M}_+^1(\mathcal{A}))$ is the softmax policy parametrization (8) of its argument. Recalling the definition of $\mathcal{Q}_r$ from Assumption 1 and denoting $\mathcal{F}_r = \{\int \psi\nu \,:\, \operatorname{supp}\nu \in \mathcal{Q}_r\}$, we further define the norms and constants needed in the following proof as:

$$\|D\psi\|_{r,\infty} = \sup_{\omega \in \mathcal{Q}_r} \|D\psi_\omega\| \qquad L_{D\psi} = \sup_{\omega,\omega' \in \mathcal{Q}_r} \frac{\|D\psi_\omega - D\psi_{\omega'}\|}{\|\omega - \omega'\|_2}$$

$$\|DR\|_{r,\infty} = \sup_{\psi(\cdot;\omega)\,:\,\omega \in \mathcal{Q}_r} \|DR_\psi\| \qquad L_{DR} = \sup_{\psi,\psi' \in \mathcal{F}_r} \frac{\|DR_\psi - DR_{\psi'}\|}{\|\psi - \psi'\|_2}$$

where $\|\cdot\|$ denotes the operator norm. While for any $r > 0$ the boundedness of $\|D\psi\|_{r,\infty}$, $L_{D\psi}$ results directly from Assumption 1, more work is needed to prove that $\|DR\|_{r,\infty} < \infty$, $L_{DR} < \infty$. We prove this in Lemma C.4 below, and proceed with the proof of convergence of the particle dynamics.

For any r and corresponding $\mathcal{Q}_r$ from Assumption 1 we define the set of localized functionals

$$\mathcal{E}^{(r)}(\nu) = \begin{cases} \mathcal{E}(\nu) & \text{if } \operatorname{supp}(\nu) \subset \mathcal{Q}_r \\ \infty & \text{else} \end{cases}$$

Furthermore, we say that a coupling $\gamma \in \mathcal{M}_+^1(\Omega \times \Omega)$ is an *admissible transport plan* if both its marginals have support in $\mathcal{Q}_r$ and finite second moments. To every admissible transport plan, for $p \geq 1$ we associate a *transportation cost* $C_p(\gamma) = (\int_{\Omega^2} |\omega - \omega'|^p \mathrm{d}\gamma(\omega, \omega'))^{1/p}$.

We prove the following results: for every $r > 0$ we have

1. There exists $\lambda_r > 0$ such that for any admissible transport plan γ, defining the interpolation map $\nu_t^\gamma := (t\Pi_0 + (1-t)\Pi_1)_\# \gamma$, the function $t \mapsto \mathcal{E}(\nu_t^\gamma)$ is differentiable with Lipschitz continuous derivative with constant $\lambda_r C_2^2(\gamma)$.
2. Let ν_0 have support in $\mathcal{Q}_r$. Then for any given transport plan γ with first marginal given by ν_0, a velocity field F satisfies

$$\mathcal{E}((\Pi_1)_\# \gamma) \geq \mathcal{E}(\nu_0) + \int F(u) \cdot (u - u') \mathrm{d}\gamma(u, u') + o(C_2(\gamma)) \tag{B.2}$$

   if and only if $F(u)$ is in the subdifferential of $D\mathcal{E}_\nu(u) := C_\pi[\psi, Q_\pi - \log \pi](u)$ for $g \in L^2(\mathcal{S} \times \mathcal{A})$ (projected in the interior of $\mathcal{Q}_r$ when $u \in \partial\mathcal{Q}_r$) for ν_0 almost every $u \in \Omega$.

The proof of the two points above corresponds to (Chizat & Bach, 2018, Lemma B.2). We sketch the proof of these two points below, referring to the original reference for the details.

1. By the Lipschitz continuity of $\psi : \Omega \to L^2(\mathcal{S} \times \mathcal{A})$ and $R'(f) = DR_f = C_\pi[\cdot, Q_\pi - \tau \log \pi]$ in $\mathcal{Q}_r$, the energy $\mathcal{E}^{(r)}(\nu_t^\gamma)$ transported along an interpolating path ν_t^γ is differentiable and we can write its derivative as
$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}^{(r)}(\nu_t^\gamma) = \int R'(\int \psi \nu_t^\gamma) \int D\psi_{(1-t)\omega + t\omega'}(\omega' - \omega)\mathrm{d}\gamma(\omega, \omega')$$
Then, again by the Lipschitz continuity of $D\psi$ and DR we have for $0 \leq t' < t'' < 1$
$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}^{(r)}(\nu_{t'}^\gamma) - \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}^{(r)}(\nu_{t''}^\gamma)\right| &\leq \left|\int (R'(\int \psi \nu_{t'}^\gamma) - R'(\int \psi \nu_{t''}^\gamma)) \int D\psi_{(1-t')\omega + t'\omega'}(\omega' - \omega)\mathrm{d}\gamma(\omega, \omega')\right| \\ &\quad + \left|\int R'(\int \psi \nu_{t''}^\gamma) \int (D\psi_{(1-t')\omega + t'\omega'} - D\psi_{(1-t'')\omega + t''\omega'})(\omega' - \omega)\mathrm{d}\gamma(\omega, \omega')\right| \\ &\leq \lambda_r |t'' - t'|\end{aligned}$$
for a λ_r large enough, where in the last inequality we have used the uniform bounds on DR in $\mathcal{Q}_r$, that $|D\psi_{(1-t')\omega + t'\omega'} - D\psi_{(1-t'')\omega + t''\omega'}| \leq (t' - t'')L_{D\psi}|\omega - \omega'|$ and we applied Hölder's inequality to bound $C_1^2(\gamma) \leq C_2^2(\gamma)$.
2. The proof of this result leverages an expansion of the functionals R, ψ to the second order in their arguments:
$$\begin{aligned}\psi(\omega') &= \psi(\omega) + D\psi_\omega(\omega' - \omega) + \mathcal{R}_\psi(\omega, \omega') \\ R(g) &= R(f) + DR_f(g - f) + \mathcal{R}_R(f, g)\end{aligned}$$
Recalling the Lipschitz bounds on the remainders $\mathcal{R}_\psi(\omega, \omega') < \frac{1}{2}L_{D\psi}|\omega - \omega'|^2$, $\mathcal{R}_R(f, g) < \frac{1}{2}L_{DR}\|f - g\|_2^2$ and combining the two expansions above we have, for a transport plan γ with marginals $\nu_0 = (\Pi_0)_\#\gamma$, $\nu_1 = (\Pi_1)_\#\gamma$
$$\mathcal{E}^{(r)}(\nu_1) = \mathcal{E}^{(r)}(\nu_0) + \int R'(\int \psi \nu_0) D\psi_u(u' - u)\mathrm{d}\gamma(u, u')\mathrm{d}s\mathrm{d}a + \mathcal{R}$$
for a remainder term $\mathcal{R}$. We can bound such remainder term, again by the Lipschitz regularity of $D\psi$ and DR and by the boundedness of $D\psi, DR$ in $\mathcal{Q}_r$, by $C_2(\gamma)^2$ and $C_1(\gamma)^2 \leq C_2(\gamma)^2$, thereby obtaining that
$$\mathcal{E}^{(r)}(\nu_1) = \mathcal{E}^{(r)}(\nu_0) + \int R'(\int \psi \nu_0) D\psi_u(u' - u)\,\mathrm{d}s\mathrm{d}a\,\mathrm{d}\gamma(u, u') + o(C_2(\gamma))$$
Noting that the integrand against the coupling is the gradient flow vector field, the above uniquely characterizes the velocity field satisfying (B.2).

We note that point 1) above immediately implies that $\mathcal{E}(\cdot)$ is λ_r-semiconvex along geodesics, while by 2) $\mathcal{E}^{(r)}(\,\cdot\,)$ admits strong Wasserstein subdifferentials on its domain (Ambrosio et al., 2008, Definition 10.3.1). Combining the two results one obtains existence and uniqueness of the solutions of the Wasserstein gradient flow through (Ambrosio et al., 2008, Theorem 11.2.1).

B.2 Propagation of chaos

By the Lipschitz continuity of the transport field in (10) in ν_0 with $\operatorname{supp}\nu_0 \subset \mathcal{Q}_r$, there exists a time $t_r > 0$ such that, $\operatorname{supp}\nu_s^{(N)} \subset \mathcal{Q}_r$ for all $s \in [0, t_r]$, $N \in \mathbb{N}$. Consider now two times $0 \leq t_1 < t_2 \leq t_r$. To prove the existence of the limiting curve $(\nu_t)_t$, we show that the curves $\nu_t^{(N)}$ are uniformly in N equicontinuous in W_2 and as such, possess a converging subsequence by Arzela-Ascoli theorem. To show equicontinuity, we bound the the W_2 distance between distributions by coupling positions of the same particles at different times and using Cauchy-Schwartz inequality:
$$W_2(\nu_{t_1}^{(N)}, \nu_{t_2}^{(N)})^2 \leq \frac{1}{N}\sum_{i=1}^N \|w_{t_1}^{(i)} - w_{t_2}^{(i)}\|_2^2 \leq \frac{t_2 - t_1}{N}\sum_{i=1}^N \int_{t_1}^{t_2} \|\frac{\mathrm{d}}{\mathrm{d}s}w_s^{(i)}\|_2^2 \mathrm{d}s$$

Combining the above with the identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\nu_t^{(N)}) = \frac{1}{N}\sum_{i=1}^{N}\langle \nabla_{w_i}\mathcal{E}(\nu_t^{(N)}), \frac{\mathrm{d}}{\mathrm{d}t}w_t^{(i)}\rangle = \frac{1}{N}\sum_{i=1}^{N}\|\frac{\mathrm{d}}{\mathrm{d}t}w_t^{(i)}\|_2^2$$

we have

$$W_2(\nu_{t_1}^{(N)}, \nu_{t_2}^{(N)}) \leq \sqrt{t_2 - t_1}\sqrt{\int_{t_1}^{t_2}\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{E}(\nu_s^{(N)})\mathrm{d}s} \leq \sqrt{t_2 - t_1}\left(\sup_{\operatorname{supp}\nu\in Q_r}\mathcal{E}(\nu) - \inf_{\operatorname{supp}\nu\in\mathcal{Q}_r}\mathcal{E}(\nu)\right)^{1/2}$$

where we recall that $\mathcal{F}_r = \{\nu \in \mathcal{M}_+^1(\Omega) \ : \ \operatorname{supp}\nu \subset \mathcal{Q}_r\}$. In particular, the above continuity bound is independent on N, proving equicontinuity of $\nu^{(N)}$ in W_2.

We now prove that the limiting point of the converging subsequence whose existence was identified above must solve (10). To do so we compare both the differential for the mean-field and particle dynamics to the one of a linear inhomogneous PDE: for any bounded and continuous $f \ : \ \mathbb{R}\times\mathbb{R}^m \to \mathbb{R}^m$, denoting by $E = \nu_t F_t \mathrm{d}t$ and $E_N = \nu_t^{(N)} F_t^{(N)}\mathrm{d}t$ where $F_t, F_t^{(N)}$ are the vector fields of the mean-field and particle system respectively, we write

$$\left|\int f(\omega)\mathrm{d}(E - E_N)\right|_2 \leq \|f\|_\infty \int \left|F_t^{(N)} - F_t\right|_2 \mathrm{d}\nu_t^{(N)}\mathrm{d}t + \left|\int fF_t\mathrm{d}(\nu_t^{(N)} - \nu_t)\mathrm{d}t\right|.$$

By boundedness of fF_t over $\mathcal{Q}_r$, the second term converges by our choice of subsequence. For the first term, denoting throughout by $\|\cdot\|_{BL}$ the bounded Lipschitz norm, we leverage the Lipschitz continuity of the vector field F WRT the underlying parametric measure:

$$\|F_t^{(N)} - F_t\|_2 \leq C_r\|\nu_t - \nu_t^{(N)}\|_{BL}$$

for $C_r > 0$ large enough, again obtaining convergence by our choice of subsequence. This proves convergence of the particle model to the mean-field equation (10) on $[0, t_r]$. To extend the time interval on which we prove convergence, we use that $\mathcal{E}(\nu)$ (and thus R) decays along trajectories of (10). Consequently, by the boundedness of the differential DR on sublevel sets of R the Lipschitz constant of $\mathcal{E}(\nu_t)$ is uniformly bounded. Using again the Lipschitz continuity of F we can show that $\sup_{u\in\mathcal{Q}_r}\|F\| < A + Br$, *i.e.*, that particle velocities can grow at most linearly in r, and an application of Gronwall allows to find, for every $T > 0$ that there exists $r > 0$ such that $\operatorname{supp}\nu_t \in \mathcal{Q}_r$ for all $t \in [0, T]$ and propagation of chaos follows. □

C PROOFS OF OPTIMALITY

Theorem 4.2. *Let Assumption 1 hold and ν_t given by* (10) *converge to ν^*, then $\pi_{\nu^*} = \pi^*$ $\mathcal{S}\times\mathcal{A}$-a.e.*

We prove the above result as sketched in the bandit setting by first establishing optimality of full support stationary measures in parametric space through Proposition 4.3, and then investigating how the dynamics preserves full support property for any $t > 0$ in Lemma C.1 and avoids spurious minima in Lemma C.2. Before starting this program we introduce the equivalent of greedy policies in the entropy-regularized setting. For a given $Q(s, a)$, the associated *Boltzmann policy* $\pi_{\mathcal{B}}$ with respect to a reference measure $\bar{\pi}$ is given by

$$\pi_{\mathcal{B}}(s, a) := \exp\left[(Q(s, a) - V_Q(s))/\tau\right] \qquad \text{for} \quad V_Q(s) := \tau\log\mathbb{E}_{a\sim\bar{\pi}}\left[\exp\left[Q(s, a)/\tau\right]\right],$$

and satisfies

$$\pi_{\mathcal{B}}(s, \cdot) = \arg\max_{\pi\in\mathcal{M}_+^1(\mathcal{A})}\left(\mathbb{E}_{a\sim\pi}\left[Q(s, a)\right] - \tau D_{\mathrm{KL}}(\pi; \bar{\pi})(s)\right)$$

One can then define the *Boltzmann backup* or *soft Bellman backup* operator T^τ that, for a given Q and the associated Boltzmann policy $\pi_{\mathcal{B}}$, gives the action-value function $T^\tau Q$ associated to $\pi_{\mathcal{B}}$:

$$T^\tau Q(s, a) = \bar{r}(s, a) + \gamma\tau\mathbb{E}_\pi\left[\log\mathbb{E}_{a_1\sim\bar{\pi}}\left[\exp Q(s_1, a_1)/\tau\right]|s_0 = s\right] \tag{C.1}$$

It is known (Haarnoja et al., 2018c, Theorem 3) that the above operator is a contraction, and that it therefore has a unique fixed point, corresponding to the optimal policy $\pi_{\mathcal{B}}^* = \pi^*$.

With the above information at hand we now proceed to prove Proposition 4.3:

Proposition 4.3. *Let ν^* be a fixed point of* (10) *such that the full support property of Assumption 1 c) holds at ν^*, then $\pi_{\nu^*} = \pi^*$.*

Proof of Proposition 4.3. A fixed point $\nu = \nu^*$ of (10) must satisfy

$$\int_{\mathcal{S}\times\mathcal{A}} \nabla_\omega \psi(\omega'; s, a) \left(Q_{\pi_\nu}(s,a) - \tau \log \pi_\nu(s,a) - V_{\pi_\nu}(s)\right) \pi_\nu(s, \mathrm{d}a) \varrho_{\pi_\nu}(\mathrm{d}s) = 0\,. \tag{C.2}$$

for all $\omega' \in \operatorname{supp}(\nu)$. In particular, the first component (in ω_0 direction) of the above gradient must be 0:

$$\int_{\mathcal{S}\times\mathcal{A}} \phi(\bar{\omega}'; s, a) \left(Q_{\pi_\nu}(s,a) - \tau \log \pi_\nu(s,a) - V_{\pi_\nu}(s)\right) \pi_\nu(s, \mathrm{d}a) \varrho_{\pi_\nu}(\mathrm{d}s) = 0\,, \tag{C.3}$$

and by homogeneity of ψ the above identity must hold for all $\bar{\omega}'$ in the projection of $\operatorname{supp}(\nu)$ on Θ. Then, by the assumed full support property (Assumption 1 c)) of ν, (C.3) holds for all $\bar{\omega}' \in \Theta$.

To deduce from this the desired invariance in Ω, we apply Assumption 1 b) on the universal approximation property of the nonlinearity ϕ and the full support property of π_ν given by the chosen parametrization, which imply that

$$Q_{\pi_\nu}(s,a) - \tau \log \pi_\nu(s,a) = V_{\pi_\nu}(s) \quad (\mathcal{S}\times\mathcal{A})\text{-a.e.}\,.$$

Finally, we rewrite the above condition in compact notation as the fixed point equation

$$T^\tau Q_{\pi_\nu}(s,a) = Q_{\pi_\nu}(s,a)$$

for the *soft Q learning* or *Boltzmann backup* operator T^τ defined in (C.1). Since T^τ is a contraction in $\|\cdot\|_2$ for $\gamma < 1$ Haarnoja et al. (2018c), its fixed point is unique and hence $\pi_{\nu^*} = \pi_{\mathcal{B}}[Q^*] = \pi^*$ for almost every $(s,a) \in \mathcal{S}\times\mathcal{A}$. □

Before proceeding to prove Theorem 4.2, we state the alternative form of Assumption 1 a) in the case where $\Theta \neq \mathbb{R}^{m-1}$. Our proof of the theorem above can be easily generalized.

Assumption 2. *Assume that $\omega = (\omega_0, \bar{\omega}) \in \mathbb{R}\times\Theta$ for $\Theta \subset \mathbb{R}^{m-1}$ which is the closure of a bounded open convex set. Furthermore $\psi(s,a;\omega) = \omega_0\phi(s,a;\bar{\omega})$ where ϕ is bounded, differentiable and $D\phi_\omega$ is Lipschitz. Also, for all $f \in L^2(\mathcal{S}\times\mathcal{A})$ the regular values of the map $\bar{\omega} \mapsto g_f(\bar{\omega}) := \int f(s,a)\phi(s,a;\bar{\omega})\mathrm{d}a\mathrm{d}s$ are dense in its range and $g_f(\bar{\omega})$ satisfies Neumann boundary conditions (i.e., for all $\bar{\omega} \in \partial\Theta$ we have $dg_f(\bar{\omega})(n_{\bar{\omega}}) = 0$ where $n_{\bar{\omega}} \in \mathbb{R}^{m-1}$ is the normal of $\partial\Theta$ at $\bar{\omega}$).*

It is instrumental for the following proof to note that in light of the above spurious minima necessarily correspond to distributions ν such that the condition

$$\int_{\mathcal{S}\times\mathcal{A}} \phi(\bar{\omega}; s, a) \left(Q_{\pi_\nu}(s,a) - \tau \log \pi_\nu(s,a) - V_{\pi_\nu}(s)\right) \pi_\nu(s, \mathrm{d}a) \varrho_{\pi_\nu}(\mathrm{d}s) = 0\,, \tag{C.4}$$

does not hold $\bar{\omega}'$-almost everywhere in Θ. Indeed, if this is the case we must have that $Q_{\pi_\nu} \neq T^\tau Q_{\pi_\nu}$ on some subset of $\mathcal{A}\times\mathcal{S}$ with positive measure so that ultimately $Q_{\pi_\nu} \neq Q_{\pi^*}$.

C.1 Proof of Theorem 4.2

We prove below that spurious local minima that do not satisfy (C.4) Θ-a.e. are avoided by the dynamics. We do so by leveraging the approximate gradient structure of the policy gradient vector field when ν_t is close to one of such stationary points, as discussed in the main text. Combined with the assumed convergence to ν^* this proves Theorem 4.2.

By the chosen structure of the approximator we note that all measures with the same mean in the homogeneous component result in the same approximator, and therefore in the same policy. In other words, if ν, ν' are such that $\int \omega_0 \nu(\mathrm{d}\omega_0, \mathrm{d}\bar{\omega}) = \int \omega_0 \nu'(\mathrm{d}\omega_0, \mathrm{d}\bar{\omega})$ a.e. then clearly

$$f_\nu(\,\cdot\,) = \int \omega_0 \phi(\,\cdot\,;\bar{\omega})\nu(\mathrm{d}\omega_0, \mathrm{d}\bar{\omega}) = \int \omega_0 \phi(\,\cdot\,;\bar{\omega})\nu'(\mathrm{d}\omega_0, \mathrm{d}\bar{\omega}) = f_{\nu'}(\,\cdot\,)\,.$$

To remove this symmetry in our model, we define the quantity

$$h^1_\nu(\mathrm{d}\bar{\omega}) := \int \omega_0 \nu(\mathrm{d}\omega_0, \mathrm{d}\bar{\omega}) \tag{C.5}$$

which will play an important role in the proof of Proposition 4.3 below. This proof pivots on two important facts: on one hand that the solution to (10) does not lose (projected) support for any finite time, and on the other that stationary points $\tilde{\nu}$ with $V_{\tilde{\nu}} \neq V^*$ – which by Theorem 4.2 cannot have full support – are avoided by the dynamics. These facts are summarized by the following lemmas:

Lemma C.1. *Let Assumption 1 a) hold and let ν_0 satisfy Assumption 1 c), then ν_t solving (10) with initial condition ν_0 satisfies Assumption 1 c) for every $t > 0$.*

Denoting by $\| \cdot \|_{BL}$ the bounded Lipschitz norm, we further prove that

Lemma C.2. *Let Assumption 1 hold and $\tilde{\nu}$ be a fixed point of (10) such that (C.4) does not hold a.e.. There exists $\varepsilon > 0$ such that if $\|h^1_{\tilde{\nu}} - h^1_{\nu_{t_1}}\|_{BL} < \varepsilon$ for a $t_1 > 0$ there exists $t_2 > t_1$ such that $\|h^1_{\tilde{\nu}} - h^1_{\nu_{t_2}}\|_{BL} > \varepsilon$.*

Proof of Lemma C.1. This result parallels (Chizat & Bach, 2018, Lemma C.13) on the preservation of the separation property Assumption 1 c) under the evolution of ν_0 by the integrated gradient flow map X_t defined as

$$\partial_t X(t,u) = F_t(X(t,u);\nu_t), \tag{C.6}$$

where we recall thatF_t is the transport field in (10). We note that the analogous result from Chizat & Bach (2018) only relies on the *continuity* of the map X_t, proven in (Chizat & Bach, 2018, Lemma B.4) under Assumption 1 a) or, alternatively, Assumption 2. Hence, it is sufficient for our purposes to establish continuity of the map X_t in the present setting, *i.e.*, when F_t is the vector field of the transport equation (10). This, however results immediately from the one-sided Lipschitz property from Assumption 1 a) enjoyed by F_t on $\mathcal{Q}_r = [-r, r] \times \Theta$ uniformly on compact time intervals. This regularity is in turn guaranteed by the Lipschitz continuity and Lipschitz smoothness of ψ from Assumption 1 and boundedness of r. □

To simplify the notation in the following proof, we denote throughout

$$\delta(\nu) := Q_{\pi_\nu} - \tau \log \pi_\nu - V_{\pi_\nu} \qquad \text{and} \qquad \langle f, g\rangle_\pi := \int_{\mathcal{S}\times\mathcal{A}} f(s,a)g(s,a)\pi(s,\mathrm{d}a)\varrho_\pi(\mathrm{d}s)\,.$$

Proof of Lemma C.2. Since $Q_{\tilde{\nu}} \neq Q^*$, by Theorem 4.2 there must exist a subset of Θ where $\tilde{\nu}$ loses support and such that $\langle \nabla\psi, \delta(\tilde{\nu})\rangle_\pi \neq 0$. Indeed, reasoning by contradiction, if $\langle \nabla\psi, \delta(\tilde{\nu})\rangle_\pi = 0$ a.e. then following the proof of Theorem 4.2 from after (C.2) we have that $V_{\tilde{\nu}} = V^*$. In other words, the key quantity

$$g_{\tilde{\nu}}(\bar{\omega}) := \langle \partial_{\omega_0}\psi(\cdot;\omega), \delta(\tilde{\nu})\rangle_{\pi_\nu} = \langle \psi(\cdot;(1,\bar{\omega})), \delta(\tilde{\nu})\rangle_{\pi_\nu} = \langle \phi(\cdot;\bar{\omega}), \delta(\tilde{\nu})\rangle_{\pi_\nu} \tag{C.7}$$

cannot vanish a.e. on Θ. Then, by Assumption 1 there exists a nonzero regular value $-\eta$ of $g_{\tilde{\nu}}(\bar{\omega})$, which we assume without loss of generality to be negative, so that $\eta > 0$ (otherwise, it suffices to invert the signs of ω_0 in the rest of the proof). We conclude the proof by introducing the (nonempty) sublevel set $\mathcal{G} := \{(\omega_0, \bar{\omega}) \in \Omega \,:\, g_{\tilde{\nu}}(\bar{\omega}) < -\eta\}$ and define

$$\mathcal{G}_+ = \{(\omega_0, \bar{\omega}) \in \mathcal{G} \,:\, \omega_0 > 0\}\,. \tag{C.8}$$

Further denoting by $\bar{G} \subseteq \Theta$ the projection of $\mathcal{G}$ onto Θ, we note that the level set $\partial\bar{G}$ is an orientable manifold of dimension $m-2$ and is by definition orthogonal to the gradient field of $g_{\tilde{\nu}}(\bar{\omega})$. By continuity of $\nabla g_{\tilde{\nu}}(\bar{\omega})$ when $\bar{G}$ is compact (which is *e.g.*, automatic when Θ is bounded as in Assumption 2) there exists

$$\beta := \min_{\bar{\omega}\in\partial\bar{G}} n_{\bar{\omega}} \cdot \nabla_{\bar{\omega}} g_{\tilde{\nu}}(\bar{\omega}) > 0\,,$$

where $n_{\bar{\omega}}$ denotes the normal unit vector to $\partial\bar{G}$ in the outward direction.[1]

We now choose ε small enough that assuming

$$\|h^1_{\tilde{\nu}} - h^1_{\nu_t}\|_{BL} < \varepsilon \qquad \text{for all } t > t_1\,, \tag{C.9}$$

leads to a contradiction. More specifically, by Lemma C.4 we choose $\varepsilon(\alpha, \eta, \beta)$ small enough so that for all ν_t such that (C.9) holds we have $g_{\nu_t}(\bar{\omega}) < -\eta/2$ on $\bar{G}$ and $n_{\bar{\omega}} \cdot \nabla_{\bar{\omega}} g_{\nu_t} > \beta/2$ on $\partial\bar{G}$. Then, the two inequalities above combined with $\partial_{\omega_0}\psi(\omega_0, \bar{\omega}) = \psi(1, \bar{\omega})$ imply that the set $\mathcal{G}_+$ defined above is forward invariant and therefore that $\partial_t \nu_t(\mathcal{G}_+) \geq 0$ as long as (C.9) holds. Furthermore, by similar arguments we notice that no trajectories enter the set $\mathcal{G} \setminus \mathcal{G}_+$ after t_1.

We now distinguish two cases: either (i) $\nu_{t_1}(\mathcal{G}_+) > 0$ or (ii) $\nu_{t_1}(\mathcal{G}_+) = 0$. We treat these two cases separately by mimicking the arguments of the proofs of (Chizat & Bach, 2018, Lemmas C.4, C.18), respectively.

[1] when $\bar{G}$ is not compact we must choose η so that it is also a regular value of the function on $\{\bar{\omega} \in \mathbb{R}^{m-1} \,:\, \|\bar{\omega}\|_2 = 1\}$ to which g converges as $\bar{\omega}$ goes to infinity.

(i) Assume that $\nu_{t_1}(\mathcal{G}_+) > 0$. We note that, besides the forward invariance of the set $\mathcal{G}_+$, under our assumptions the first component of the velocity field in $\mathcal{G}$ is lower bounded by $\eta/2$, so that $\omega_0(t) = \omega_0(0) + t\eta/2$ is a subsolution to the ω_0-component of the trajectory of a test mass with initial condition with $\omega(0) \in \mathcal{G}$, as long as $\bar{\omega}(t) \in \bar{G}$. In particular, by the forward invariance of $\mathcal{G}_+$, if $\omega(0) \in \mathcal{G}_+$ we have $\omega_0(t) > t\eta/2$. Therefore, assuming that $\text{supp}(\nu_t) \subset (-M, M) \times \Theta$ we must have for every $t > t_1$

$$h^1_{\nu_t}(\bar{G}) \geq \eta/2(t - t_1)\nu_{t_1}(\mathcal{G}_+) + \min\{0, (t - t_1)\eta/2 - M\}\nu_{t_1}(\mathcal{G} \setminus \mathcal{G}_+)\,.$$

In particular, for $t > t_1 + 2M/\eta$ the above quantity grows linearly in time, in contradiction with the assumption made above that $\|h^1_{\tilde{\nu}} - h^1_{\nu_{t_1}}\|_{BL} < \varepsilon$ for all $t > t_1$.

(ii) Assume now that $\nu_{t_1}(\mathcal{G}_+) = 0$. We show that there exists $t_2 > t_1$ such that $\nu_{t_2}(\mathcal{G}_+) > 0$, so that the proof is completed by applying part i) above from t_2. Indeed let $\omega^* \in \text{supp}(\nu_{t_1})$ be such that $\bar{\omega}^* \in \bar{G}$ is a local minimum of $g_{\tilde{\nu}}$, *i.e.*, for which $\nabla g_{\tilde{\nu}} = 0$. Then choosing $\tilde{\varepsilon}$ such that $\mathcal{B}_{\tilde{\varepsilon}}(\bar{\omega}^*) \subset \bar{G}$, and choosing M large enough such that $\text{supp}\,(\nu_{t_1}) \subseteq [-M, M] \times \Theta$, we know by Lemma C.3 that there exists $t_2 > t_1$ such that the image at t_2 of $\omega(t_1) := \omega^*$ under the flow of the policy gradient vector field is contained in $\mathcal{G}_+$. By continuity of pushforward map of such vector field, this must also hold for a neighborhood of ω^*, to which ν_{t_1} assigns positive mass.

□

Defining by $\|\cdot\|_{C^1}$ the maximum of the supremum norm of a function and the supremum norm of its gradient and recalling the structure of the policy gradient vector field (recall the definition of g in (C.7)):

$$F_t(\omega, \nu_t) = -\nabla\langle\omega_0\phi(\bar{\omega}), \delta(\nu_t)\rangle_{\pi_\nu} = -\nabla(\omega_0\, g_{\nu_t}(\bar{\omega})) \tag{C.10}$$

where ν_t solves (10).

We recall the following result from Chizat & Bach (2018), proving that case ii) in the analysis above will ultimately reduce to case i). We prove this lemma again below for completeness.

Lemma C.3. *Let $\tilde{\nu} \in \mathcal{M}^1_+(\Omega)$ and $\bar{\omega}^*$ satisfy $|\nabla g_{\tilde{\nu}}(\bar{\omega}^*)| = 0$, $g_{\tilde{\nu}}(\bar{\omega}^*) < -\eta < 0$ for some $\eta > 0$. Then for every $M, \tilde{\varepsilon} > 0$ there exists $\varepsilon, t_2 > 0$ such that if for all $t \in (0, t_2)$ we have $\|g_{\tilde{\nu}} - g_{\nu_t}\|_{C^1} < \varepsilon$ and $\omega^*_0 \in [-M, 0]$, then at time t_2 the point ω^* is mapped, under the flow of the policy gradient vector field (C.10) to a subset of $\mathcal{B}_{\tilde{\varepsilon}}((1, \bar{\omega}^*))$.*

Proof of Lemma C.3. We can bound the components of the velocity of a particle $(\omega_0(t), \bar{\omega}(t))$ under (C.10) with initial condition $\omega(0) = \bar{\omega}^*$ in the homogeneous direction as:

$$\frac{\mathrm{d}}{\mathrm{d}t}\omega_0(t) = -g_{\nu_t}(\bar{\omega}(t)) \geq -g_{\tilde{\nu}}(\bar{\omega}^*) - |g_{\tilde{\nu}}(\bar{\omega}(t)) - g_{\tilde{\nu}}(\bar{\omega}^*)| - |g_{\nu_t}(\bar{\omega}(t)) - g_{\tilde{\nu}}(\bar{\omega}(t))|$$

while in the other directions, defining $q(t) := \|\bar{\omega}(t) - \bar{\omega}^*\|$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}q(t) &\leq |\omega_0(t)|\|\nabla_{\bar{\omega}} g_{\nu_t}(\bar{\omega}(t))\| \\ &\leq |\omega_0(t)|\,[\|\nabla_{\bar{\omega}} g_{\tilde{\nu}}(\bar{\omega}^*)\| + \|\nabla_{\bar{\omega}} g_{\tilde{\nu}}(\bar{\omega}(t)) - \nabla_{\bar{\omega}} g_{\tilde{\nu}}(\bar{\omega}^*)\| + \|\nabla_{\bar{\omega}} g_{\nu_t}(\bar{\omega}(t)) - \nabla_{\bar{\omega}} g_{\tilde{\nu}}(\bar{\omega}(t))\|]\end{aligned}$$

for all $t \in [0, \bar{\tau}]$ where $\bar{\tau} := \inf\{t \,:\, \omega_0(t) \notin [-M, 1]\}$. Moreover, by the Lipschitz continuity of $g_{\tilde{\nu}}(\,\cdot\,)$ and its Lipschitz smoothness there exists $L > 0$ such that $\max\{|g_{\tilde{\nu}}(\bar{\omega}) - g_{\tilde{\nu}}(\bar{\omega}^*)|, \|\nabla g_{\tilde{\nu}}(\bar{\omega}) - \nabla g_{\tilde{\nu}}(\bar{\omega}^*)\|\} \leq L\|\bar{\omega} - \bar{\omega}^*\|$. Since we have $\|g_{\tilde{\nu}} - g_{\nu_t}\|_{C^1} < \varepsilon$ by assumption, for $t \in [0, \bar{\tau}]$ we must also have

$$\begin{cases}\frac{\mathrm{d}}{\mathrm{d}t}\omega_0(t) \geq \eta - \varepsilon - Lq(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}q(t) \leq |\omega_0(t)|\,[\varepsilon + Lq(t)]\end{cases}$$

Upon increasing, if necessary, the value of L such that $\eta/4L < \tilde{\varepsilon}$, we define $\tau_q = \inf\{t \,:\, q(t) > \eta/4L\}$ and we proceed to show that one can choose $\varepsilon \in (0, \eta/4)$ such that $\tau_q > \bar{\tau}$, *i.e.*, that the forward dynamics of the point $\omega^* = (\omega^*_0, \bar{\omega}^*)$ will reach $\mathcal{G}_+$ before $q(t) > \tilde{\varepsilon}$. Recall that on $[0, \tau_q]$ and for $\varepsilon \in (0, \eta/4)$ we have

$$\omega_0(t) \geq \omega_0(0) + \frac{\eta}{2}t$$

and in particular $\omega_0(t) > \omega_0(0) \geq -M$. This implies that for all $t \in [0, \bar{\tau} \wedge \tau_q]$ we have $\frac{\mathrm{d}}{\mathrm{d}t}q(t) < q(0) + M\varepsilon + LMq(t)$, so that by Gronwall we obtain $q(t) \leq \varepsilon M \exp[LMt]$. Finally, setting $\tau_0 := 2(M + 1)/\eta \geq$

$-2(\omega_0(0)-1)/\eta > \bar{\tau}$ we note that we can choose ε small enough such that $\tau_q > \tau_0 > \bar{\tau}$, *i.e.*, by monotonicity of $q(t)$ under our bound, such that

$$q(\tau_0) \leq \varepsilon M \exp\left[2LM(M+1))/\eta\right] \leq \eta/4L\,.$$

Choosing $\varepsilon \in (0, \eta/4)$ such that the RHS inequality holds concludes the proof. □

We now proceed to show the needed regularity of the potential g_ν from (C.7) in terms of the signed measure h^1_ν defined in (C.5). In doing so, we also prove Lipschitz smoothness of the operator R defined in (B.1):

Lemma C.4. *For any $r > 0$, the operator R on $\mathcal{F}_r = \{\int \psi\nu \,:\, \operatorname{supp}\nu \in \mathcal{Q}_r\}$ is Lipschitz smooth, and DR_f is bounded in the supremum norm. Furthermore, for all $C_0 > 0$ there exists $\alpha > 0$ and $\varepsilon > 0$ such that for all ν, ν' satisfying $\|h^1_\nu\|_{BL}, \|h^1_{\nu'}\|_{BL} < C_0$, $\|h^1_\nu - h^1_{\nu'}\|_{BL} < \varepsilon$, one has*

$$\|g_\nu - g_{\nu'}\|_{C^1} \leq \alpha \|h^1_\nu - h^1_{\nu'}\|_{BL}\,. \tag{C.11}$$

To prove the above lemma, we first bound some relevant quantities.

Lemma C.5. *For all $f, f' \in \mathcal{F}_r$ there exists $\varepsilon > 0$, $C', C'', C''' > 0$ such that if $\|f - f'\|_2 < \varepsilon$ one has*

$$\|\pi_\nu(s,\cdot) - \pi_{\nu'}(s,\cdot)\|_2 \leq C'\|f - f'\|_2 \tag{C.12}$$

$$\|\varrho_\nu - \varrho_{\nu'}\|_2 \leq C''\|f - f'\|_2 \tag{C.13}$$

$$\|Q_{\pi_\nu} - Q_{\pi_{\nu'}}\|_2 \leq C'''\|f - f'\|_2 \tag{C.14}$$

Proof of Lemma C.5. Throughout this proof, for simplicity of notation, we will write $\pi = \pi_\nu$ and $\pi' = \pi_{\nu'}$. Furthermore, we use that for $f \in \mathcal{F}_r$ there exists $C_0 > 0$ so that

$$e^{-C_0\|\phi\|_{C^1}} \leq \exp[f(s,a)] \leq e^{C_0\|\phi\|_{C^1}}\,, \tag{C.15}$$

implying, together with the assumed absolute continuity of ϱ_0 that $\|Q_\pi\|_\infty, \|\pi\|_\infty, \|\varrho_\pi\|_\infty < \infty$. Setting throughout $\tau = 1$ to simplify the notation and combining the above with the pointwise upper bound $e^x < 1 + K_r|x|$ for $|x| < e^{C_0\|\phi\|_{C^1}}$ we obtain

$$\begin{aligned}
\|\pi - \pi'\|_2 &\leq \left\| \frac{\exp[f(s,a)]}{\int \exp[f(s,a')]\mathrm{d}a'} - \frac{\exp[f'(s,a)]}{\int \exp[f'(s,a')]\mathrm{d}a'} \right\|_2 \\
&\leq \left\| \frac{\exp[f(s,a)]}{\int \exp[f(s,a')]\mathrm{d}a'} - \frac{\exp[f'(s,a)]}{\int \exp[f(s,a')]\mathrm{d}a'} \right\|_2 + \left\| \frac{\exp[f'(s,a)]}{\int \exp[f(s,a')]\mathrm{d}a'} - \frac{\exp[f'(s,a)]}{\int \exp[f'(s,a')]\mathrm{d}a'} \right\|_2 \\
&\leq \left\| \frac{\exp[f'(s,a)]}{\int \exp[f(s,a')]\mathrm{d}a'} \right\|_\infty \left(\|1 - \exp[f(s,a) - f'(s,a)]\|_2 + \left\| \frac{\int \exp[f(s,a')]\mathrm{d}a'}{\int \exp[f'(s,a')]\mathrm{d}a'} - 1 \right\|_2 \right) \\
&\leq \frac{e^{2C_0\|\phi\|_{C^1}}}{|\mathcal{A}|} \left(K_r \left\| f(s,a) - f'(s,a) \right\|_2 + \frac{e^{2C_0\|\phi\|_{C^1}}}{|\mathcal{A}|} \left\| \int \big(\exp[f(s,a') - f'(s,a')] - 1\big)\mathrm{d}a' \right\|_2 \right) \\
&\leq \frac{e^{2C_0\|\phi\|_{C^1}}}{|\mathcal{A}|} K_r (1 + e^{4C_0\|\phi\|_{C^1}}) \|f - f'\|_2 =: C'\|f - f'\|_2,
\end{aligned} \tag{C.16}$$

where we have denoted by $|\mathcal{A}|$ the Lebesgue measure of the action space $\mathcal{A}$. We now proceed to establish the second bound in the statement of the lemma. In this case, denoting the t-steps transition probability as $P^t_\pi(s, \mathrm{d}s_t) = \int_{\mathcal{S}^{t-1}} P_\pi(s, \mathrm{d}s_1) P_\pi(s_1, \mathrm{d}s_2) \ldots P_\pi(s_{t-1}, \mathrm{d}s_t)$ we have

$$\begin{aligned}
\|\varrho_\pi - \varrho_{\pi'}\|_2 &= \left\| \sum_{t=1}^\infty \gamma^t \int_{\mathcal{S}} \varrho_0(\mathrm{d}s_0) \left(P^t_\pi(s_0, \mathrm{d}s) - P^t_{\pi'}(s_0, \mathrm{d}s) \right) \right\|_2 \\
&\leq \sum_{t=1}^\infty \gamma^t \sum_{j=0}^{t-1} \left\| \varrho_0 P^j_\pi \left(P_\pi - P_{\pi'} \right) P^{t-j-1}_{\pi'} \right\|_2 .
\end{aligned}$$

Observing that for any smooth $\varrho \in \mathcal{M}^1_+(\mathcal{S})$ for the operator norm of the difference in the above sum we have

$$\|\varrho(P_\pi - P_{\pi'})\|_2 = \left\| \int_{\mathcal{S}} \varrho(\mathrm{d}s) \int_{\mathcal{A}} P(s,a,\mathrm{d}s') \left(\pi(s,\mathrm{d}a) - \pi'(s,\mathrm{d}a) \right) \right\|_2$$

$$\leq \|\varrho\|_2 \left(\int_{\mathcal{S}\times\mathcal{S}'} \left(\int_{\mathcal{A}} |\pi(s,\mathrm{d}a) - \pi'(s,\mathrm{d}a)| P(s,a,s')\mathrm{d}a \right)^2 \mathrm{d}s\mathrm{d}s' \right)^{1/2}$$
$$\leq \|\varrho\|_2 \|\pi - \pi'\|_2 \|P\|_2 \leq \frac{(1-\gamma)^2}{\gamma} C'' \|\varrho\|_2 \|f - f'\|_2$$

for large enough C'', where we used (C.16) in the last line. From this we conclude

$$\|\varrho_\pi - \varrho_{\pi'}\|_2 \leq \frac{(1-\gamma)^2}{\gamma} \sum_{t=1}^{\infty} t\gamma^t C'' \|f - f'\|_2 = C'' \|f - f'\|_2$$

Finally, defining for notational convenience $R'_\tau := \bar{r} - \tau D_{KL}(\pi', \bar{\pi})$ we write:

$$\begin{aligned}
\|Q_\pi - Q_{\pi'}\|_2 &= \left\| \gamma \int P(s,a,\mathrm{d}s') \left(V_\pi(s') - V_{\pi'}(s') \right) \right\|_2 \\
&= \left\| \gamma \int P(s,a,\mathrm{d}s') \left(\int R_\tau(s'',a') \varrho^\pi(s',\mathrm{d}s'')\pi(\mathrm{d}a') - R'_\tau(s'',a')\varrho^{\pi'}(s',\mathrm{d}s'')\pi'(\mathrm{d}a') \right) \right\|_2 \\
&\leq \gamma \left\| \int R_\tau(s'',a')\varrho^\pi(s',\mathrm{d}s'')\pi(\mathrm{d}a') - R'_\tau(s'',a')\varrho^{\pi'}(s',\mathrm{d}s'')\pi'(\mathrm{d}a') \right\|_2 \\
&\leq \left\| \int \left(R_\tau - R'_\tau \right)(s'',a')\varrho^\pi(s',\mathrm{d}s'')\pi(\mathrm{d}a') \right\|_2 + \left\| \int R'_\tau(s'',a')(\varrho^\pi - \varrho^{\pi'})(s',\mathrm{d}s'')\pi(\mathrm{d}a') \right\|_2 \\
&\qquad + \left\| \int R'_\tau(s'',a')\varrho^{\pi'}(s',\mathrm{d}s'')(\pi - \pi')(\mathrm{d}a') \right\|_2
\end{aligned} \tag{C.17}$$

and bound each term separately letting $C'''_1, C'''_2, C'''_3 > 0$ be large enough constants. For the first, we have:

$$\begin{aligned}
\left\| \int \left(R_\tau - R'_\tau \right)(s'',a')\varrho_\pi(s',\mathrm{d}s'')\pi(s'',\mathrm{d}a') \right\|_2 &\leq \|\pi\varrho_\pi\|_\infty \|\log\pi - \log\pi'\|_2 \\
&\leq \|\pi\varrho_\pi\|_\infty \left(\|f - f'\|_2 + \|\log \frac{\int e^{f(s,a)}\mathrm{d}a}{\int e^{f'(s,a)}\mathrm{d}a}\|_2 \right) \\
&\leq C'''_1 \|f - f'\|_2
\end{aligned} \tag{C.18}$$

where we have bounded the $\log$ term as follows:

$$\begin{aligned}
\|\log \frac{\int e^{f(s,a)}\mathrm{d}a}{\int e^{f'(s,a)}\mathrm{d}a}\|_2 &= \|\log \frac{\int e^{f(s,a)} - e^{f'(s,a)}\mathrm{d}a}{\int e^{f'(s,a)}\mathrm{d}a} + 1\|_2 \\
&\leq \|\log \frac{\int e^{f'(s,a)} \left(e^{|f(s,a) - f'(s,a)|} - 1 \right)\mathrm{d}a}{\int e^{f'(s,a)}\mathrm{d}a} + 1\|_2 \\
&\leq \|\log \left(\frac{\|e^{-f'}\|_\infty \|e^{f'}\|_\infty}{|\mathcal{A}|} \int \left(e^{|f(s,a) - f'(s,a)|} - 1 \right)\mathrm{d}a + 1 \right)\|_2 \\
&\leq \|\log \left(\frac{\|e^{-f'}\|_\infty \|e^{f'}\|_\infty}{|\mathcal{A}|} K_r \int |f(s,a) - f'(s,a)|\mathrm{d}a + 1 \right)\|_2 \\
&\leq \|\log \left(\|e^{-f'}\|_\infty \|e^{f'}\|_\infty K_r \|f(s,a) - f'(s,a)\|_2 + 1 \right)\|_2 \\
&\leq \|e^{-f'}\|_\infty \|e^{f'}\|_\infty K_r \|f - f'\|_2
\end{aligned} \tag{C.19}$$

For the second term in (C.17), using the boundedness of $\|R\|_2 < C_R$ and that $\varrho^\pi - \varrho^{\pi'} = \gamma\left(\varrho_\pi - \varrho_{\pi'}\right)$ for a certain, ϱ_0 depending on s' we write

$$\left\| \int R'_\tau(s'',a')(\varrho^\pi - \varrho^{\pi'})(s',\mathrm{d}s'')\pi(s'',\mathrm{d}a') \right\|_2 \leq \|R'_\tau(s'',a')\|_\infty \|\varrho_\pi - \varrho_{\pi'}\|_2$$

$$\leq C_2'''\|f - f'\|_2. \tag{C.20}$$

We finally bound the third term by writing

$$\begin{aligned}\left\|\int R'_\tau(s'', a')\varrho_{\pi'}(s', \mathrm{d}s'')(\pi - \pi')(\mathrm{d}a')\right\|_2 &\leq \|\varrho_{\pi'} R'_\tau(s'', a')\|_\infty \|\pi - \pi'\|_2 \\ &\leq C_3'''\|f - f'\|_2 \end{aligned} \tag{C.21}$$

and obtain (C.14) by combining (C.18)-(C.21). □

Proof of Lemma C.4. We first establish the desired properties of the functional R. To do so we differentiate (8) at $f \in L^2(\mathcal{S} \times \mathcal{A})$

$$\begin{aligned}\frac{\delta \pi_f}{\delta f}(s, a) &= \frac{\delta}{\delta f}\frac{e^{\tau \int f(s,a)\nu(\mathrm{d}\omega)}}{\int_{\mathcal{A}} e^{\tau f(s,a)}\mathrm{d}a} = \frac{1}{\int_{\mathcal{A}} e^{\tau f}\mathrm{d}a}\frac{\delta}{\delta f}e^{\tau f} - \frac{e^{\tau f}}{\left(\int_{\mathcal{A}} e^{\tau f}\mathrm{d}a\right)^2}\int_{\mathcal{A}}\frac{\delta}{\delta \nu}e^{\tau f}\mathrm{d}a \\ &= \tau\left(f - \int_{\mathcal{A}} f\pi_f(s, \mathrm{d}a')\right)\pi_f(s, a)\,. \end{aligned} \tag{C.22}$$

Then, combining the above with (A.6) and Hölder inequality we obtain

$$\|DR_f\|_{\infty,r} = \sup_{f \in \mathcal{F}_r}\int (DR_f(s, a))^2 \mathrm{d}s\mathrm{d}a < \infty$$

where we have used that all the terms appearing in DR_f are bounded for every choice of $r > 0$.
To establish Lipschitz smoothness of R, letting $f, f' \in \mathcal{F}_r$ and denoting, to simplify notation, $\pi = \pi_f$ and $\pi' = \pi_{f'}$, we proceed to bound the operator norm by splitting the RHS as

$$\begin{aligned}\|DR_f - DR_{f'}\| &\leq \sup_{\ell \in L^2(\mathcal{S}\times\mathcal{A})\,:\,\|\ell\|=1} |C_\pi[\ell, Q_\pi - \tau\log\pi] - C_{\pi'}[\ell, Q_{\pi'} - \tau\log\pi']| \\ &\leq \sup_{\ell \in L^2(\mathcal{S}\times\mathcal{A})\,:\,\|\ell\|=1} [(I) + (II) + (III)]\end{aligned}$$

and considering the resulting terms separately.
First of all, defining throughout by slight abuse of notation $\delta(\pi) := Q_\pi - \tau\log\pi$, we have

$$\begin{aligned}(I) &:= |\langle\int \ell(s, a)\pi'(s, \mathrm{d}a) - \int \ell(s, a)\pi(s, \mathrm{d}a), \delta(\pi')\rangle_{\pi'}| \\ &\leq \int_{\mathcal{S}}\left(\int_{\mathcal{A}} \ell(s, a)(\pi(s, a) - \pi'(s, a))\mathrm{d}a\right)\left(\int \delta(\pi')\pi'(\mathrm{d}a)\right)\varrho_{\pi'}(s)\mathrm{d}s \\ &\leq \|\ell\|_2\|\pi - \pi'\|_2\|\delta(\pi')\varrho_{\pi'}\pi'\|_\infty \\ &\leq C\|\ell\|_2\|f - f'\|_2, \end{aligned} \tag{C.23}$$

where the last line was obtained using (C.12) and boundedness from above and below of π, Q_π in $\mathcal{F}_r$. We further write, for a $K < \infty$ large enough

$$\begin{aligned}(II) &:= |\langle\ell(s, a) - \int \phi(s, a', \cdot)\pi_\nu(s, \mathrm{d}a'), \delta(\pi')\rangle_{\pi_{\nu'}} - \langle\ell(s, a) - \int \phi(s, a', \cdot)\pi_\nu(s, \mathrm{d}a'), \delta(\pi')\rangle_\pi| \\ &\leq |\int \delta(\pi')\ell(s, a)\varrho_\pi(s)(\pi - \pi')(s, \mathrm{d}a)\mathrm{d}s| + |\int \ell(s, a)\delta(\pi')(\varrho_\pi - \varrho_{\pi'})(s)\pi'(s, \mathrm{d}a)\mathrm{d}s| \\ &\leq \|\delta(\pi')\varrho_{\pi'}\|_\infty\|\ell\|\left(\|\pi - \pi'\| + \|\varrho_\pi - \varrho_{\pi'}\|\right) \\ &\leq K\|\ell\|\|f - f'\|_2, \end{aligned} \tag{C.24}$$

where we have used Cauchy-Schwartz inequality together with (C.12) and (C.13). Finally, using (C.14) and (C.19), we bound for $K < \infty$ possibly larger than above

$$(III) := |\langle\ell(s, a) - \int \ell(s, a')\pi(s, \mathrm{d}a'), \delta(\pi) - \delta(\pi')\rangle_\pi|$$

$$
\begin{aligned}
&\leq \|\varrho_\pi \pi\|_\infty \|\ell\|_2 \|\delta(\pi) - \delta(\pi')\|_2 \\
&\leq \|\varrho_\pi \pi\|_\infty \|\ell\|_2 \left(\|Q_\pi - Q_{\pi'}\|_2 + \|V_\pi - V_{\pi'}\|_2 + \tau \| \log \frac{\pi}{\pi'} \|_2 \right) \\
&\leq \|\varrho_\pi \pi\|_\infty \|\ell\|_2 \left((1 + \|\pi'\|_\infty) \|Q_\pi - Q_{\pi'}\|_2 + \|Q_\pi\|_\infty \|\pi - \pi'\|_2 + \tau \| \log \frac{\pi}{\pi'} \|_2 \right) \\
&\leq K \|\ell\|_2 \|f - f'\|_2. \qquad \text{(C.25)}
\end{aligned}
$$

Combining (C.23), (C.24) and (C.25) we obtain that

$$
L_{DR} = \sup_{f, f' \in \mathcal{F}_r} \frac{\|DR_f - DR_{f'}\|}{\|f - f'\|_2} < \infty
$$

proving the Lipschitz smoothness claim.

Proceeding to the proof of (C.11), combining the Lipschitz smoothness of R on bounded sets, the identity $\psi(s, a; (1, \bar{\omega})) = \phi(s, a; \bar{\omega})$ and the boundedness of the set $\{\int \psi \nu \ : \ \nu \in \mathcal{P}_2(\Omega), |h^1_\nu| < C_0\}$ we obtain

$$
\begin{aligned}
\|f_\nu - f_{\nu'}\|_2 &= \left\| \int \psi(\cdot; \omega)(\nu - \nu')(\mathrm{d}\omega) \right\|_2 \qquad \text{(C.26)} \\
&= \left\| \int \phi(\cdot; \omega)(h^1_\nu - h^1_{\nu'})(\mathrm{d}\bar{\omega}) \right\|_2 \\
&\leq \sup_{\ell \in L^2(\mathcal{S} \times \mathcal{A}), \|\ell\| \leq 1} \int \int \ell(s, a) \phi(s, a; \bar{\omega}) \mathrm{d}s \mathrm{d}a (h^1_\nu - h^1_{\nu'})(\mathrm{d}\bar{\omega}) \\
&\leq \|\phi\|_{C^1} \|h^1_\nu - h^1_{\nu'}\|_{BL} \,,
\end{aligned}
$$

which combined with the $\|\phi\|_{C^1}$-Lipschitz continuity of the map $\int \ell(s, a) \phi(s, a; \bar{\omega}) \mathrm{d}s \mathrm{d}a$ and with the Lipschitz smoothness of R concludes the proof. □