Convergence of Policy Gradient for Entropy Regularized MDPs with Neural Network Approximation in the Mean-Field Regime

Bekzhan Kerimkulov [*1] **James-Michael Leahy** [*2] **David Šiška** [*13] **Łukasz Szpruch** [*14]

Abstract

We study the global convergence of policy gradient for infinite-horizon, continuous state and action space, and entropy-regularized Markov decision processes (MDPs). We consider a softmax policy with (one-hidden layer) neural network approximation in a mean-field regime. Additional entropic regularization in the associated mean-field probability measure is added, and the corresponding gradient flow is studied in the 2-Wasserstein metric. We show that the objective function is increasing along the gradient flow. Further, we prove that if the regularization in terms of the mean-field measure is sufficient, the gradient flow converges exponentially fast to the unique stationary solution, which is the unique maximizer of the regularized MDP objective. Lastly, we study the sensitivity of the value function along the gradient flow with respect to regularization parameters and the initial condition. Our results rely on the careful analysis of the nonlinear Fokker–Planck–Kolmogorov equation and extend the pioneering work of (Mei et al., 2020) and (Agarwal et al., 2020), which quantify the global convergence rate of policy gradient for entropy-regularized MDPs in the tabular setting.

1. Introduction

1.1. Overview

In the last decades, reinforcement learning (RL) algorithms with neural network approximation have demonstrated incredible performance. Notable successes have been reported in (Mnih et al., 2015; Silver et al., 2018; Vinyals et al., 2019). Various versions of the policy gradient algorithms have been demonstrated to be particularly effective. However, a mathematical theory that provides guarantees for the convergence of these algorithms has been elusive. In particular, we are not aware of any work that has established convergence of a policy gradient algorithm in the continuous state and action setting with neural network approximation. The motivation behind this work is to shed more light on this challenging open question in the setting of entropy regularized MDPs (Markov decision processes) where the policy is of softmax type and parameterized by a one-hidden layer neural network in the mean-field regime.

Entropy regularized MDPs have been widely studied due to their excellent empirical performance and desirable theoretical properties (Haarnoja et al., 2017; Geist et al., 2019). Convergence of policy gradient with softmax policies was first studied for the entropy regularized problem in (Agarwal et al., 2020). The convergence has been further quantified in (Mei et al., 2020), where the authors showed that softmax policy gradient converges at an exponential rate. Encouraged by the recent convergence results of (Mei et al., 2020) in the tabular setting, we consider the continuous state and action space setting with policies approximated mean-field neural networks. The mean-field setting was also considered in (Agazzi & Lu, 2020). The authors show that if the mean-field policy gradient flow converges, then, under appropriate assumptions, the limiting policy is optimal.

We consider an infinite horizon Markov decision model $\mathcal{M} = (S, A, P, r, \gamma)$, where S is the state space, A the action space with a fixed finite reference measure μ, P the transition probability kernel, r is a bounded reward function, and γ is the discount factor. For a given stochastic policy $\pi : S \to \mathcal{P}(A)$ and initial distribution $\rho \in \mathcal{P}(S)$, we consider the entropy-regularized value function

$$V_\tau^\pi(\rho) = \mathbb{E}_\rho^\pi \left[\sum_{n=0}^{\infty} \gamma^n \left(r(s_n, a_n) - \tau \ln \frac{d\pi}{d\mu}(a_n|s_n) \right) \right],$$

where $\tau \geq 0$ determines the intensity of the entropy regularization. Denoting $V^\pi(s) = V^\pi(\delta_s)$, we have

$$V_\tau^\pi(s) = V_0^\pi(s) - \frac{\tau}{1-\gamma} \int_S d^\pi(ds'|s) \mathrm{KL}(\pi(\cdot|s')|\mu), \quad (1)$$

[*]Equal contribution [1]School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom [2]Department of Mathematics, Imperial College London, London, United Kingdom [3]Vega Protocol, Gibraltar, Gibraltar [4]The Alan Turing Institute, London, United Kingdom. Correspondence to: James-Michael Leahy <j.leahy@imperial.ac.uk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

where d^π is the occupancy kernel under the policy π. For full details on our assumptions and notation, we refer to Sections A.1 and 2.1. There are two implications of having $\tau > 0$. The first is that the optimal policy satisfies

$$\pi_\tau^*(da|s) = \exp\left(\frac{1}{\tau}\left(Q_\tau^*(s,a) - V_\tau^*(s)\right)\right)\mu(da)\,,$$

where V_τ^* and Q_τ^* denote the optimal value and state-action value functions, respectively (see Theorem 2.1 or (Ziebart et al., 2010; Haarnoja et al., 2017; Geist et al., 2019)). Second, since the entropy term is strictly convex (Dupuis & Ellis, 1997)[Sec. 1.4], the addition (see, e.g., (1)) is expected to improve the convergence when optimizing $V_\tau^\pi(\rho)$ with respect π with a gradient-descent-type method such as softmax policy gradient. While the latter point may seem intuitive, the analysis is far from being straightforward even in the tabular case studied in (Mei et al., 2020), where the entropic regularization is shown to lead to exponential convergence of the softmax policy gradient algorithm. The difficulty arises chiefly from the fact that V_τ^π depends on π in a non-linear and, in general, non-convex way.

In the case where S and A are finite sets (i.e., the tabular setting), (Mei et al., 2020) considered softmax policies $\pi_F : S \to \mathcal{P}(A)$ of the form

$$\pi_F(a|s) \propto \exp(F(s,a)) \tag{2}$$

for $F : S \times A \to \mathbb{R}$, and where we have identified π_F with its probability mass function. Since F is equivalent to a parameter $\theta \in \mathbb{R}^{|S|\times|A|}$, policy gradient is a gradient flow in $\mathbb{R}^{|S|\times|A|}$. This approach becomes computationally intractable as the size of the sets S and A grow large. To overcome this and also to cover the continuous state-action setting, we use function approximation. More specifically, we consider policies approximated by a one hidden layer neural network with mean-field scaling:

$$\pi_\theta(da|s) \propto \exp\left(\frac{1}{m}\sum_{i=1}^m f(\theta^i, s, a)\right)\mu(da)\,, \theta \in \mathbb{R}^{d\times m}\,,$$

where $f : \mathbb{R}^d \times S \times A \to \mathbb{R}$. The continuous-time policy gradient is given by

$$\frac{d}{dt}\theta_t = \nabla V_\tau^{\pi_{\theta_t}}(\rho)\,,$$

where here and later ∇ means ∇_θ unless stated otherwise. One can equivalently view the policy as a function of an empirical measure over parameters $\nu_t^m := \frac{1}{m}\sum_{i=1}^m \delta_{\theta_t^i}$:

$$\pi_{\nu_t^m}(da|s) \propto \exp\left(\int_{\mathbb{R}^d} f(\theta, s, a)\nu_t^m(d\theta)\right)\mu(da)\,.$$

While the one-hidden-layer neural network is typically non-convex as a function of the parameters θ, once lifted to the space of measures it becomes linear as a function of measure.

The idea of using mean-field scaling to understand the convergence of gradient descent algorithms used to train neural networks has been extensively studied in a recent series of works, see (Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Chizat & Bach, 2018; Sirignano & Spiliopoulos, 2021; Hu et al., 2021). These works show that in the context of supervised learning, finding the optimal weights in deep neural networks can be viewed as a sampling problem. What emerges is the idea that the aim of the learning algorithm is to find optimal distribution over the parameter space rather than optimal values of the parameters. As a consequence, individual values of the parameters are not important in the sense that different sets of weights sampled from the correct (i.e., optimal) distribution are equally good.

The key feature of the supervised learning task studied by the aforementioned mentioned works is that a finite-dimensional non-convex optimization problem becomes convex when lifted to an infinite-dimensional space of measures; that is, working with the limiting measure $\nu^m \to \nu$ as $m \to \infty$. This observation, unfortunately, does not hold in RL problem with policies $\pi_\nu : S \to \mathcal{P}(A)$ indexed by probability measures $\nu \in \mathcal{P}(\mathbb{R}^d)$ of the form

$$\pi_\nu(da|s) \propto \exp\left(\int_{\mathbb{R}^d} f(\theta, s, a)\nu(d\theta)\right)\mu(da)$$

due to the non-linearity and non-convexity of the problem in π, and hence ν. Despite this, in (Agazzi & Lu, 2020), the authors show that if the policy gradient flow for the mean-field neural network parameters ν_t converges to ν^* with full support, then the resulting parameterized policy π_{ν^*} is globally optimal for the entropy regularized problem (i.e., $V^* = V^{\pi_{\nu^*}}$) under the assumption that the activation function used leads to a sufficiently expressive class of functions. However, the work of (Agazzi & Lu, 2020) does not establish that the flow converges.

In general, $\nu \in \mathcal{P}(\mathbb{R}^d) \mapsto V_\tau^{\pi_\nu}$ is not convex, and hence one should not expect convergence to the global optimum for the associated policy gradient algorithm. To alleviate this issue and establish global convergence rates for policy gradient, we introduce further entropic regularization of ν relative to a prior e^{-U} for some normalized potential U. For arbitrarily given $\tau \geq 0$, $\sigma \geq 0$, and $\rho \in \mathcal{P}(S)$, we define $J^{\tau,\sigma} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ by

$$J^{\tau,\sigma}(\nu) = V_\tau^{\pi_\nu}(\rho) - \frac{\sigma^2}{2}\mathrm{KL}(\nu|e^{-U})\,. \tag{3}$$

In Theorem A.10, we prove that $J^{\tau,\sigma}(\nu_t)$ increases as function of t along the gradient flow for any $\tau \geq 0$, $\sigma \geq 0$. This fact is missing in (Agazzi & Lu, 2020), which considered the case $\sigma = 0$. Moreover, for any $\tau \geq 0$ and for sufficiently

large $\sigma > 0$, we show that $\nu_t \to \nu^*$ exponentially fast and that ν^* is the optimal parameter measure, see Theorem 2.12. While having $\tau > 0$ motivates considering policies of softmax type (see Theorem 2.1 below), this condition is not needed to prove any of our results. How large σ needs to be taken depends on the bounds we establish in Theorem 2.4. In Corollary 2.14, we estimate the sensitivity of the value function along the gradient flow with respect to regularization parameters and the initial condition. A heuristic derivation of the gradient flow and conditions for optimality is provided in Section 1.3 for the readers convenience.

1.2. Literature Review

There is an enormous amount of research literature on RL and we cannot hope to do it justice here. For this reason, we focus on the subset of RL that we feel is most related to this work. Entropy regularized RL has demonstrated both good algorithmic performance and desirable theoretical properties (Haarnoja et al., 2017; 2018; Geist et al., 2019; Vieillard et al., 2020; Neu et al., 2017; Fox et al., 2015; Ziebart et al., 2010). It has been shown that softmax policies are optimal in the entropy regularized setting. In fact, many authors consider softmax parameterized policies and the policy gradient algorithm without adding entropic regularization. In the tabular and "compatible" function approximation case, (Agarwal et al., 2020) proved convergence of softmax policy gradient at rate $O(1/\sqrt{t})$. In the tabular case, (Mei et al., 2020) showed that softmax policy gradient converges with rate $O(1/t)$. Moreover, using Łojasiewicz inequalities, they showed that entropic regularization improves the rate of convergence to $O(e^{-t})$. We remark that the authors in (Li et al., 2021) establish lower bounds for the policy gradient algorithms studied in (Agarwal et al., 2020) and (Mei et al., 2020), and show that these may scale poorly with the cardinality of the state space. In (Liu et al., 2019) and (Wang et al., 2020), the authors consider consider over-parameterized neural-network policies and show that neural proximal/trust region and actor-critic policy optimization, respectively, converges with rate $O(1/\sqrt{t})$ in the idealized setting (exact action-value setting).

As explained above, the tabular softmax policy (2) gradient method becomes computationally intractable as the size of the sets S and A grow large. In this paper, we consider continuous state and action spaces and softmax policies approximated by an idealized infinitely-wide one hidden layer neural network (i.e., the mean-field regime). Working in the mean-field regime is motivated in part by the success of a series of recent works which prove the convergence of (noisy) stochastic gradient algorithms arising in the training of neural networks in the static and supervised learning setting, see (Jordan et al., 1998; Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Chizat & Bach, 2018; Sirignano & Spiliopoulos, 2021; Hu et al., 2021) and by the recent work (Agazzi & Lu, 2020) in RL. We will discuss the connection (Hu et al., 2021) and (Agazzi & Lu, 2020) in more detail.

In (Hu et al., 2021), the authors studied the supervised learning setting for a one-hidden layer neural network in the mean-field regime. The key feature of this setting is the convexity of the unregularized (e.g., $\sigma = 0$) objective. By taking $\sigma > 0$, the objective function becomes strictly convex, which allows them to establish convergence to the invariant measure for any $\sigma > 0$. In the RL setting, the objective function is highly non-linear and, in general, non-convex, and hence the analysis of the problem is significantly more involved. This can be seen in Theorem 2.4, where necessary boundedness and regularity results of the objective function are obtained. The analysis of the gradient flow (4) in the current work is done using PDE techniques that are different from the tools used in (Hu et al., 2021) and require less regularity. This enables us to establish that the objective function is decreasing along the flow even when $\sigma = 0$, which, even in a supervised setting, is a new result.

In (Agazzi & Lu, 2020), the authors consider mean-field softmax parameterized policies and derive the gradient flow for the parameter measure, see (4) with $\sigma = 0$. Their main theorem says that if ν^* is a stationary solution of (4) with full support and the family of functions $(s, a) \mapsto f(s, a, \theta)$ indexed by $\theta \in \mathbb{R}^d$ gives rise to a sufficiently expressive class of functions, then $V_\tau^{\pi_{\nu^*}} = V_\tau^*$, which implies that the parameterized policy is optimal. Whether there exists a solution to the stationary equation in the space of probability measures is unknown. Optimality is not established by directly showing $\nu^* \in \operatorname{argmax}_\nu J^{\tau,0}(\nu)$, but rather by showing $V_\tau^{\pi_{\nu^*}}$ satisfies the optimal Bellman equation (9) for V_τ^*, and hence that $\pi_{\nu^*} \in \operatorname{argmax}_\pi V_\tau^\pi(\rho)$. Since the gradient flow preserves the support of the initial condition ν_0, they are able to conclude that *if* the gradient flow ν_t converges to ν^*, then π_{ν^*} is optimal. Crucially, the work of (Agazzi & Lu, 2020) does not establish that the flow converges. Moreover, it does not establish that the gradient flow is non-decreasing; our Theorem 2.9 establishes this for $\tau, \sigma \geq 0$. The insight gained by (Agazzi & Lu, 2020) is that the value function plays a vital role as the solution to the Bellman equation. We no longer study the original value function $J^{\tau,0}(\nu) = V_\tau^{\pi_\nu}$, but rather $J^{\tau,\sigma}$, which with $\sigma > 0$ introduces regularization of the parameter measure, and does not solve a Bellman equation. Hence, we must employ different techniques to show optimality (see the proof of Theorem 2.12). In particular, we make use of Theorem 2.9, which is a step toward understanding what is happening at the level of the value function. Our other main theorem, Theorem 2.12, is meant to complement the work of (Agazzi & Lu, 2020) by identifying sufficient conditions under which one can conclude that $\nu_t \to \nu^*$ at a rate $O(e^{-\beta t})$, where $\beta > 0$ is a constant quantified in our analysis.

1.3. Heuristic Derivation of the Gradient Flow and Conditions for Optimality

Our aim is to maximize $J^{\tau,\sigma} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ defined in (3). In this section, we present a heuristic computation in the spirit of Otto calculus (Villani, 2009) through which we identify the gradient flow equation. For simplicity, in this section, we will restrict to absolutely continuous measures $\nu \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ and abuse notation and identify ν with its density $d\nu/d\theta$. At the end of this section, we explain how this abstract gradient flow relates to a familiar noisy gradient ascent. Let $E : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ denote a time-dependent vector field and consider the gradient flow

$$\partial_t \nu_t = \nabla \cdot (E_t \nu_t) \, .$$

Our aim is to identify the vector field E so that $J^{\tau,\sigma}(\nu_t)$ is increasing as a function of t. To do this, we will use differential calculus on $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the 2-Wasserstein distance, see (Ambrosio et al., 2008; Carmona & Delarue, 2018). Ignoring for the moment that entropy is only lower semi-continuous, and hence that the linear functional derivative $\frac{\delta J^{\tau,\sigma}}{\delta \nu} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ does not exists (see Definition A.1), we have

$$\begin{aligned}\partial_t J^{\tau,\sigma}(\nu_t) &= \lim_{h\to 0} \frac{J^{\tau,\sigma}(\nu_{t+h}) - J^{\tau,\sigma}(\nu_t)}{h} \\ &= \lim_{h\to 0} \frac{1}{h} \int_0^1 \int_{\mathbb{R}^d} \frac{\delta J^{\tau,\sigma}}{\delta \nu}(\nu_t^{\varepsilon,h})(\nu_{t+h} - \nu_t) \, d\theta \, d\varepsilon \, ,\end{aligned}$$

where $\nu_t^{\varepsilon,h} := \nu_t + \varepsilon(\nu_{t+h} - \nu_t)$. Since $\nu_t^{\varepsilon,h} \to \nu_t$ as $\varepsilon \to 0$, integrating by parts, we obtain

$$\begin{aligned}\partial_t J^{\tau,\sigma}(\nu_t) &= \int_{\mathbb{R}^d} \frac{\delta J^{\tau,\sigma}}{\delta \nu}(\nu_t) \nabla \cdot (E_t \nu_t) \, d\theta \\ &= -\int_{\mathbb{R}^d} \nabla \frac{\delta J^{\tau,\sigma}}{\delta \nu}(\nu_t) E_t \nu_t \, d\theta \, .\end{aligned}$$

To ensure that $J^{\tau,\sigma}(\nu_t)$ is increasing, one must take (up to a multiplicative constant) $E_t := -\nabla \frac{\delta J^{\tau,\sigma}}{\delta \nu}(\nu_t)$ so that

$$\partial_t J^{\tau,\sigma}(\nu_t) = \int_{\mathbb{R}^d} \left| \nabla \frac{\delta J^{\tau,\sigma}}{\delta \nu}(\nu_t) \right|^2 \nu_t \, d\theta \geq 0 \, .$$

It can be shown that along the gradient flow, we have

$$\frac{\delta J^{\tau,\sigma}}{\delta \nu}(\nu_t, \theta) = \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t, \theta) - \frac{\sigma^2}{2} \left(U(\theta) + \ln \nu_t(\theta) + 1 \right) \, ,$$

where we recall that $J^{\tau,0}(\nu) = V_\tau^{\pi_\nu}$. In Lemma 2.3, we compute $\frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t, \theta)$, and if $\tau = 0$, the expression is equivalent to the expression arising in policy gradient, see (Sutton & Barto, 2018)[Ch. 13], with the difference being that our policies are parameterized by measures, and so measure derivatives appear instead of classical gradients. Thus, the gradient flow is given by

$$\partial_t \nu_t = \nabla \cdot \left[\left(-\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t) + \frac{\sigma^2}{2} \nabla U \right) \nu_t \right] + \frac{\sigma^2}{2} \Delta \nu_t \, . \tag{4}$$

In Theorem 2.6 and Corollary 2.7, below, we show that any local maximum ν^* of $J^{\tau,\sigma}$ must satisfy the first order condition $\theta \mapsto \frac{\delta J^{\tau,\sigma}}{\delta \nu}(\nu^*, \theta)$ is constant, or equivalently that ν^* is a stationary solution of (4). This also implies that in the case $\sigma > 0$, ν^* satisfies

$$\nu^*(\theta) \propto \exp\left(\frac{2}{\sigma^2} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu^*, \theta) - U(\theta) \right) \, ,$$

which can be viewed as a posterior distribution over the parameters space given the prior e^{-U}.

On the other hand, if ν_t converges to a stationary solution of (4), then using the fact that $J^{\tau,\sigma}(\nu_t)$ is increasing, we can conclude that ν^* is a maximum as long as there is a unique stationary solution. Thus, we see that the question of the existence of the unique global minimizer is related to the existence of a unique stationary solution of (4). However, from the general theory of nonlinear Fokker–Planck–Kolmogorov equations, one only expects existence and uniqueness of solutions to the stationary equation if $\sigma > 0$ is large (see, e.g., (Bogachev et al., 2019)[Ex. 1.1], (Bogachev et al., 2018)[Ex. 4.3], (Manita et al., 2015)[Sec. 6]).

To implement the gradient flow (4) one needs to compute or estimate $\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta)$, where $\frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta)$ is specified in Lemma 2.3. The policy gradient method is a cornerstone of RL precisely because the terms $d_\rho^{\pi_\nu}$ and $Q_\tau^{\pi_\nu}$ can be estimated using Monte Carlo (e.g., Reinforce) from roll-outs without having to explicitly know the transition probability. Instead of working with the non-linear PDE (4), one can use the probabilistic representation $\nu = \mathrm{Law}(\theta)$, where $\theta : \Omega \times \mathbb{R}_+ \to \mathbb{R}^d$ is the solution of the McKean–Vlasov stochastic differential equation (SDE)

$$d\theta_t = \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\mathrm{Law}(\theta_t), \theta_t) - \frac{\sigma^2}{2} \nabla U(\theta_t) \right) dt + \sigma dW_t \, ,$$

and where W is a d-dimensional Wiener process. Approximating ν with its empirical measure and discretizing in time with a learning rate η, we arrive at the familiar noisy gradient ascent algorithm

$$\begin{cases} \theta_{k+1}^i = \theta_k^i + \eta \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_k^m, \theta_k^i) - \frac{\sigma^2}{2} \nabla U(\theta_k^i) \right) \\ \qquad\qquad + \sqrt{\eta} \sigma \zeta_{k+1}^i \, , \quad k \in \mathbb{N}_0 \, , \\ \nu_k^m := \frac{1}{m} \sum_{i=1}^m \delta_{\theta_k^i} \, , \end{cases} \tag{5}$$

where $\{\zeta_k^i\}_{1 \leq i \leq m, k \in \mathbb{N}_0} \overset{\text{i.i.d.}}{\sim} N(0,1)$. The convergence rates of (5) to the gradient flow (4) (or its probabilistic representation) are well understood under general conditions. For example, we refer a reader to (Jabir et al., 2019)[Thms. 8 and 9], in which such analysis has been carried out in the context of training recurrent neural networks, and to (Delarue & Tse, 2021), where uniform in time weak particles

approximation error has been studied. The approximations errors between ν_k^m given by (5) and ν_t, with $t = k\eta$, are of $O(1/m) + O(\eta)$ uniformly in time, and this can be seen as proxy for algorithmic complexity.

1.4. Outline of the Paper

The main contributions of this work are

1. Theorem 2.9, which proves that the objective $J^{\tau,\sigma}(\nu_t)$ increases along the gradient flow equation for the evolution of $(\nu_t)_{t\in\mathbb{R}_+}$ corresponding to policy gradient;
2. Theorem 2.12, which proves exponential convergence of ν_t to ν^* if $\sigma > 0$ is sufficiently large.
3. Theorem 2.13 and Corollary 2.14, where the dependence of the gradient flow and its stationary solutions and of the value function on τ and σ is quantified.

The notation and essential definitions are introduced in Section A.1. Entropy regularized MDPs are introduced in Section 2.1. The main results of the paper are stated Section 2.2, the proofs of which will be given in Section A.2. Conclusions and possible future research directions are discussed in Section 4. A heuristic derivation of our results is also provided in Section 1.3, which sheds more light on our work and the mean-field approach.

2. Formulation and the Statement of the Main Results

2.1. Entropy Regularized Markov Decision Processes

We refer readers to (Puterman, 2014; Bertsekas & Shreve, 2004; Hernández-Lerma & Lasserre, 2012) for thorough introduction Markov decision processes. Let S and A denote Polish spaces. Let $P \in \mathcal{P}(S|S \times A)$. Let $r \in B_b(S \times A)$. Let $\gamma \in [0, 1)$. Let $\tau \in \mathbb{R}_+$ and $\mu \in \mathcal{M}_+(A)$ denote a strictly positive finite measure. The seven-tuple $\mathcal{M} = (S, A, P, r, \gamma, \tau, \mu)$ determines an infinite horizon τ-entropy regularized γ-discounted Markov decision model. Here, we note, that $\tau = 0$ is considered, and if $\tau = 0$, then $\mathcal{M}$ is a infinite horizon Markov decision model $\mathcal{M} = (S, A, P, r, \gamma)$ since μ is not needed to formulate the control problem. If $\tau = 0$, we further assume that A is compact, $P(\cdot|\cdot, a) \in \mathcal{P}(S|S)$ is strongly Feller for all $a \in A$, and that $r(s, \cdot) : A \to \mathbb{R}$ is upper semi-continuous for every $s \in S$ so that Condition 3.3.3 in (Hernández-Lerma & Lasserre, 2012) holds, and thus measurable selection condition holds.

Let $((S\times A)^{\mathbb{N}}, \mathcal{F})$ denote the canonical sample space, where $\mathcal{F}$ is the corresponding σ-algebra. Elements of $(S \times A)^{\mathbb{N}}$ are of the form $(s_0, a_0, s_1, a_1, \ldots)$ with $s_n \in S$ and $a_n \in A$ denoting the projections and called the state and action variables, at time $n \in \mathbb{N}_0$, respectively. By Proposition 7.28 in (Bertsekas & Shreve, 2004), for an arbitrarily given initial distribution $\rho \in \mathcal{P}(S)$ and randomized stationary policy $\pi \in \mathcal{P}(A|S)$, there exists a unique product probability measure $\mathbb{P}_\rho^\pi$ on the canonical sample space with expectation denoted $\mathbb{E}_\rho^\pi$ such that for every time $n \in \mathbb{N}_0$, $\mathbb{P}_\rho^\pi(s_0 \in \mathcal{S}) = \rho(\mathcal{S})$, $\mathbb{P}^\pi(a_n \in \mathcal{A}|(s_0, a_0, \ldots, s_n)) = \pi(a_n|s_n)$, and

$$\mathbb{P}_\rho^\pi(s_{n+1} \in \mathcal{S}|(s_0, a_0, \ldots, s_n, a_n)) = \mathcal{P}(\mathcal{S}|s_n, a_n)$$

for all $\mathcal{S} \in \mathcal{B}(S)$ and $\mathcal{A} \in \mathcal{B}(A)$. In particular, $\{s_n\}_{n\in\mathbb{N}_0}$ is a Markov chain with kernel $P_\pi \in \mathcal{P}(S|S)$ given by

$$P_\pi(ds'|s) = \int_A P(ds'|s, a')\pi(da'|s)\,.$$

For given $s \in S$, we denote $\mathbb{E}_s^\pi = \mathbb{E}_{\delta_s}^\pi$, where $\delta_s \in \mathcal{P}(S)$ denotes the Dirac measure at $s \in S$.

Let $\Pi_0 = \mathcal{P}(A|S)$ and

$$\Pi_\tau = \{\exp(F(s, a))\mu(da) \in \mathcal{P}_\mu(A|S)|F \in B_b(S \times A)\}$$

if $\tau > 0$. For a given policy $\pi \in \Pi_\tau$, we define the τ-entropy regularized value function $V_\tau^\pi : S \to \mathbb{R}$ by

$$V_\tau^\pi(s) = \mathbb{E}_s^\pi\left[\sum_{n=0}^\infty \gamma^n \left(r(s_n, a_n) - \tau \ln \frac{d\pi}{d\mu}(a_n|s_n)\right)\right].$$

For a given policy $\pi \in \Pi_\tau$, we define the regularized state-action value function $Q_\tau^\pi \in B_b(S \times A)$ by

$$Q_\tau^\pi(s, a) = r(s, a) + \gamma \int_S V_\tau^\pi(s')P(ds'|s, a)\,. \quad (6)$$

The occupancy kernel $d^\pi \in \mathcal{P}(S|S)$ is defined by

$$d^\pi(ds'|s) = (1 - \gamma)\sum_{n=0}^\infty \gamma^n P_\pi^n(ds'|s)\,,$$

where $P_\pi^0(ds'|s) := \delta_s(ds')$, P_π^n is understood as a product of kernels, and convergence is understood in $b\mathcal{K}(S|S)$. It follows that for all $s \in S$,

$$\begin{aligned} V_\tau^\pi(s) &= \int_A \Big(r(s, a) - \tau \ln \frac{d\pi}{d\mu}(a|s) \\ &\qquad + \gamma \int_S V_\tau^\pi(s')P(ds'|s, a)\Big)\pi(da|s) \quad (7) \\ &= \frac{1}{1 - \gamma}\int_S\int_A (r(s', a') \\ &\qquad - \tau \ln \frac{d\pi}{d\mu}(a'|s')\Big)\pi(da'|s')d^\pi(ds'|s)\,. \quad (8) \end{aligned}$$

For a given initial distribution $\rho \in \mathcal{P}(S)$, we define

$$V_\tau^\pi(\rho) = \int_S V_\tau^\pi(s)\rho(ds),\ d_\rho^\pi(ds) = \int_S d^\pi(ds|s')\rho(ds')\,.$$

For each $(s,a) \in S \times A$, we define the measurable optimal value and state-action value functions by

$$V_\tau^*(s) = \sup_{\pi \in \Pi_\tau} V_\tau^\pi(s) \quad \text{and} \quad Q_\tau^*(s,a) = \sup_{\pi \in \Pi_\tau} Q_\tau^\pi(s,a)\,.$$

By virtue of (Hernández-Lerma & Lasserre, 2012)[Thm. 4.2.3] in the $\tau = 0$ case and (Geist et al., 2019)[Thm. 1] in the $\tau > 0$ case (using (Dupuis & Ellis, 1997)[Prop. 1.4.2.), we obtain the following dynamical programming principle. See, also, (Haarnoja et al., 2017)[Thm. 1 and 2].

Theorem 2.1 (Dynamic programming principle for $\tau \geq 0$)**.** *The optimal value function $V_\tau^* \in B_b(S)$ is the unique solution of the regularized Bellman equation given by*

$$V_\tau^*(s) = \max_{m \in \mathcal{P}(A)} \left[\int_A \left(r(s,a) - \tau \ln \frac{dm}{d\mu}(a) + \gamma \int_S V_\tau^*(s') P(ds'|s,a) \right) m(da) \right]. \tag{9}$$

Moreover, $Q_\tau^*(s,a) = r(s,a) + \gamma \int_S V_\tau^*(s') P(ds'|s,a)$.

If $\tau = 0$, then there exists a measurable function $f^* : S \to A$ called a selector such that $f^*(s) \in \operatorname{argmax}_{a \in A} Q_0^*(s,a)$ and the induced policy $\pi^* \in \Pi_0$ defined by $\pi^*(da|s) = \delta_{f^*(s)}(da)$ for all $s \in S$ satisfies $V_0^* = V_0^{\pi^*}$.

If $\tau > 0$, then for all $s \in S$,

$$V_\tau^*(s) = \tau \ln \int_A \exp\left\{ Q_\tau^*(s,a)/\tau \right\} \mu(da),$$

and $\pi_\tau^* \in \Pi_\tau$ defined by

$$\pi_\tau^*(da|s) = \exp\left((Q_\tau^*(s,a) - V_\tau^*(s))/\tau \right) \mu(da)$$

is the unique policy satisfying $V_\tau^* = V_\tau^{\pi_\tau^*}$.

2.2. Softmax Mean-Field Policy and the Entropy Regularized Objective

For arbitrarily given $k \in \mathbb{N}_0$, let $\mathcal{A}_k$ consist of jointly measurable functions $h : \mathbb{R}^d \times S \times A \to \mathbb{R}$ such that for all $s \in S$, μ-almost-all $a \in A$, h is k-times differentiable in θ, and satisfies

$$|h|_{\mathcal{A}_k} := \max_{0 \leq j \leq k} \operatorname*{ess\,sup}_{a \in A} \sup_{s \in S} \operatorname*{ess\,sup}_{\theta \in \mathbb{R}^d} |\nabla^j h(\theta, s, a)| < \infty\,,$$

where the essential supremum over A is defined relative to the reference measure μ, the essential supremum over $\mathbb{R}^d$ is defined relative to the Lebesgue measure λ, and here and henceforth $\nabla = \nabla_\theta$. For given $f \in \mathcal{A}_0$, define $\pi : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}_\mu(A|S)$ by

$$\pi_\nu(da|s) = \frac{\exp\left(\int_{\mathbb{R}^d} f(\theta, s, a) \nu(d\theta) \right)}{\int_A \exp\left(\int_{\mathbb{R}^d} f(\theta, s, a') \nu(d\theta) \right) \mu(da')} \mu(da)\,,$$

which we refer to as a mean-field softmax policy. It follows that $\ln(\frac{d\pi_\nu}{d\mu}) \in B_b(S \times A)$ (see Lemma A.4), and hence $\pi_\nu \in \Pi_\tau$ for $\tau > 0$.

If $S \times A$ is a compact subset of $\mathbb{R}^{d_S} \times \mathbb{R}^{d_A}$, then we may take f to be of the form

$$f(\theta, s, a) = f(\theta, x) = \psi(\theta_0) \cdot g(\theta_1 x)\,,$$

where $x = (s_1, \ldots, s_{d_S}, a_1, \ldots, a_{d_A}, 1)^\top \in \mathbb{R}^{d_S + d_A + 1}$, $\theta = (\theta_0, \theta_1) \in \mathbb{R}^{d'} \times \mathbb{R}^{d' \times (d_S + d_A + 1)}$, $g : \mathbb{R} \to \mathbb{R}$ is an activation function such as a sigmoid or hyperbolic tangent applied component-wise, and $\psi : \mathbb{R} \to [-C, C]$, $C > 0$, is a smooth rescaling function. We see that $f \in \mathcal{A}_k$ for any $k \in \mathbb{N}_0$. Here, we may also take the reference measure μ to be the Lebesgue measure.

We now introduce a measurable positive potential $U : \mathbb{R}^d \to \mathbb{R}$ satisfying $\int_{\mathbb{R}^d} e^{-U(\theta)} d\theta = 1$. The prototypical example one should have in mind is the quadratic potential $U(\theta) = \frac{d}{2} \ln(2\pi) + \frac{1}{2}|\theta|^2$. For given $\rho \in \mathcal{P}(S)$, $\tau \geq 0$, and $\sigma \geq 0$, we define the entropy regularized functional $J^{\tau,\sigma}(\rho) : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ by

$$J^{\tau,\sigma}(\rho)(\nu) = V_\tau^{\pi_\nu}(\rho) - \frac{\sigma^2}{2} \mathrm{KL}(\nu | e^{-U})\,.$$

Henceforth we fix $\rho \in \mathcal{P}(S)$ and write $J^{\tau,\sigma} = J^{\tau,\sigma}(\rho)$, unless otherwise specified. In the case $\sigma > 0$, we maximize $J^{\tau,\sigma}$ over $\mathcal{P}_2(\mathbb{R}^d)$ and, without loss of generality, it suffices to maximize over

$$\mathcal{P}_2^{\mathrm{fe}}(\mathbb{R}^d) := \{ \nu \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d) : \mathrm{KL}(\nu | e^{-U}) < \infty \}\,.$$

2.3. Statement of Main Results

Our main results are Theorems 2.6, 2.9, 2.12, and 2.13, which establish necessary conditions for optimality, the well-posedness of a gradient flow along which the objective increases, convergence of the gradient flow in a regularized regime, and upper bounds on W_2-distance between two gradient flows with different σ, τ, and initial conditions. Before stating these, we highlight our core auxiliary results, namely Lemmas 2.2 and 2.3 and Theorem 2.4, which enable us to apply methods from the analysis of non-linear Fokker–Planck PDEs and McKean–Vlasov dynamics to the MDP problem described in Section 2.1, and specifically the parameterized problem described in Section 2.2.

Lemma 2.2 (Functional derivative of π)**.** *If $f \in \mathcal{A}_0$, then the function $\pi : \mathcal{P}(\mathbb{R}^d) \to b\mathcal{K}_\mu(A|S)$ has a linear functional derivative $\frac{\delta \pi}{\delta \nu} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to b\mathcal{K}_\mu(A|S)$ given by*

$$\frac{\delta \pi}{\delta \nu}(\nu, \theta)(da|s) = \left(f(\theta, s, a) - \int_A f(\theta, s, a') \pi_\nu(da'|s) \right) \pi_\nu(da|s)\,. \tag{10}$$

Lemma 2.2 is proved in Section A.2.3.

Lemma 2.3 (Functional derivative of $J^{\tau,0}$)**.** *If $f \in \mathcal{A}_1$, then the function $J^{\tau,0} : \mathcal{P}_1(\mathbb{R}^d) \to \mathbb{R}$ has a linear functional derivative $\frac{\delta J^{\tau,0}}{\delta \nu} : \mathcal{P}_1(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ given by*

$$\frac{\delta J^{\tau,0}}{\delta \nu}(\nu,\theta) = \frac{1}{1-\gamma}\int_S \int_A \Big(Q_\tau^{\pi_\nu}(s,a) - \tau \ln \frac{d\pi_\nu}{d\mu}(a|s)\Big) \frac{\delta \pi}{\delta \nu}(\nu,\theta)(da|s) d\rho^{\pi_\nu}(ds)\,. \tag{11}$$

Lemma 2.3 is proved in Section A.2.4. If $\tau = 0$, then (11) is the policy gradient theorem in (Sutton & Barto, 2018).

Theorem 2.4 (Boundedness and regularity of $\frac{\delta J^{\tau,0}}{\delta \nu}$)**.** *There are constants $C_k = C(\gamma, |r|_{B_b(S\times A)}, \tau, \mu(A), |f|_{\mathcal{A}_k})$, $k \in \mathbb{N}$, $L = L(\gamma, |r|_{B_b(S\times A)}, \tau, \mu(A), |f|_{\mathcal{A}_1})$, and $D = D(\gamma, \mu(A), |f|_{\mathcal{A}_1})$ such that for all $\tau, \tau' \geq 0$, $\theta \in \mathbb{R}^d$, $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}^d)$, and $k \in \mathbb{N}$,*

$$\left|\nabla^k \frac{\delta J^{\tau,0}}{\delta \nu}(\nu,\theta)\right| \leq C_k\,, \tag{12}$$

$$|J^{\tau,0}(\nu') - J^{\tau,0}(\nu)| \leq C_1 W_1(\nu',\nu)\,,$$

$$\left|\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu',\theta) - \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu,\theta)\right| \leq L W_1(\nu',\nu)\,,$$

$$\text{and} \quad \left|\nabla \frac{\delta J^{\tau',0}}{\delta \nu}(\nu,\theta) - \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu,\theta)\right| \leq D|\tau' - \tau|\,.$$

Theorem 2.4 is proved in Section A.2.5. As a result of (12), we note $J^{\tau,\sigma}(\nu) = J^{\tau,\sigma}(\nu')$ if $\nu = \nu'$, λ-a.e.. We will always work under the following assumption on U.

Assumption 2.5 (Growth of ∇U)**.** There is a constant $C_U > 0$ such that for all $\theta \in \mathbb{R}^d$, $|\nabla U(\theta)| \leq C_U(1 + |\theta|)$.

Theorem 2.6 (Necessary condition for optimality)**.** *Let $f \in \mathcal{A}_1$ and Assumption 2.5 hold. If $\nu \in \mathcal{P}_1(\mathbb{R}^d)$ is a local maximum of $J^{\tau,\sigma}$, then*

$$\theta \mapsto \frac{\delta J^{\tau,0}}{\delta \nu}(\nu,\theta) - \frac{\sigma^2}{2} U(\theta) - \frac{\sigma^2}{2} \ln \nu(\theta)$$

is constant ν-a.e.. Moreover, if $\sigma > 0$, then ν is equivalent to the Lebesgue measure λ, and for λ-a.a. $\theta \in \mathbb{R}^d$,

$$\nu(\theta) = \mathcal{Z}^{-1} e^{\frac{2}{\sigma^2}\frac{\delta J^{\tau,0}}{\delta \nu}(\nu,\theta) - U(\theta)}\,, \tag{13}$$

$$\text{where} \quad \mathcal{Z} := \int_{\mathbb{R}^d} e^{\frac{2}{\sigma^2}\frac{\delta J^{\tau,0}}{\delta \nu}(\nu,\theta') - U(\theta')}\, d\theta'\,.$$

Theorem 2.6 is proved in Section A.2.6. Assume $f \in \mathcal{A}_1$ and for given $\nu \in \mathcal{P}_1(\mathbb{R}^d)$, define the linear operator L_ν on $C_c^\infty(\mathbb{R}^d)$ by

$$L_\nu \phi = \frac{\sigma^2}{2}\Delta\phi + \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu) - \frac{\sigma^2}{2}\nabla U\right) \cdot \nabla \phi\,.$$

We denote the adjoint of L_ν by L_ν^*, which acts on probability measures $\mathcal{P}(\mathbb{R}^d)$.

Corollary 2.7 (Local maximum and elliptic PDE)**.** *Let $f \in \mathcal{A}_1$ and Assumption 2.5 hold. If $\nu \in \mathcal{P}_1(\mathbb{R}^d)$ is a local maximum of $J^{\tau,\sigma}$ and has full support, then ν is a measure-valued solution of*

$$L_\nu^* \nu = 0\,. \tag{14}$$

Corollary 2.7 is proved in Section A.2.6. We will also require the following additional assumption on U.

Assumption 2.8 (Lipschitzness of ∇U)**.** There exists a constant $L_U > 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$, $|\nabla U(\theta) - \nabla U(\theta')| \leq L_U|\theta - \theta'|$.

The following theorem establishes precise conditions for the well-posedness of (4) and for the objective function to be increasing along the flow. The formula (16), even in $\sigma = 0$ case, is the first of its kind in the literature. In order to establish (16) for non-smooth and non-compactly-supported initial conditions ν_0, we have extended the argument in (Bogachev et al., 2016).

Theorem 2.9 (Gradient flow)**.** *Let $f \in \mathcal{A}_1$ and Assumption 2.5 hold. If $\nu_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p \in \mathbb{N}$, then there exists a measure-valued solution $\nu \in C(\mathbb{R}_+; \mathcal{P}_p(\mathbb{R}^d))$ of*

$$\partial_t \nu_t = L_{\nu_t}^* \nu_t\,, \quad \nu|_{t=0} = \nu_0\,. \tag{15}$$

If $\sigma = 0$ or $\nu_0 \in \mathcal{P}_2^{\text{fe}}(\mathbb{R}^d)$, then

$$J^{\tau,\sigma}(\nu_t) = J^{\tau,\sigma}(\nu_0) + \int_0^t \int_{\mathbb{R}^d} \left|\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s,\theta) - \frac{\sigma^2}{2}\nabla \ln \frac{\nu_s(\theta)}{e^{-U(\theta)}}\right|^2 \nu_s(d\theta)\, ds\,. \tag{16}$$

Moreover, if both $f \in \mathcal{A}_2$ and Assumption 2.8 holds or both $\sigma > 0$ and $p \geq 4$ holds, then the solution of (40) is unique.

Theorem 2.9 will be proved in Section A.2.7.

Theorem 2.10 (McKean–Vlasov SDE well-posedness)**.** *Let $f \in \mathcal{A}_2$ and Assumptions 2.5 and 2.8 hold. Let $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{R}_+}, \mathbb{P})$ denote a filtered probability triple supporting an $\mathbb{F}$-adapted Wiener process $(W_t)_{t\in\mathbb{R}_+}$ and $\mathcal{F}_0$-measurable random variable θ_0 independent of $(W_t)_{t\in\mathbb{R}_+}$. Then there exists a unique continuous $\mathbb{F}$-adapted solution $\theta : \Omega \times \mathbb{R}_+ \to \mathbb{R}^d$ of the McKean–Vlasov SDE*

$$d\theta_t = \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\text{Law}(\theta_t),\theta_t) - \frac{\sigma^2}{2}\nabla U(\theta_t)\right) dt + \sigma dW_t\,, \tag{17}$$

where $\theta|_{t=0} = \theta_0$. Moreover, $(\nu_t)_{t\geq 0} := (\text{Law}(\theta_t))_{t\geq 0}$ is the unique solution of (15) with $\nu_0 := \text{Law}(\theta_0)$.

If $f \in \mathcal{A}_2$ and Assumptions 2.5 and 2.8 hold, then by Theorem 2.4, we can conclude by (42), (43), and (44) that the drift of (17) has linear growth and is Lipschitz continuous in ν and θ (see Corollary A.9). Theorem 2.10 then follows from (Carmona & Delarue, 2018)[Thm. 4.21], Itô's formula ((Krylov, 2008)[Ch 2. Sec 10]) and Theorem 2.9.

The following theorem says that in the highly regularized regime, there is a unique solution of (14) which is a global optimizer of $J^{\tau,\sigma}$, and we have exponential convergence of the gradient flow (15) to this unique optimizer.

Assumption 2.11 (Dissipativity of ∇U)**.** There exists a constant $\kappa > 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$,

$$(\nabla U(\theta) - \nabla U(\theta')) \cdot (\theta - \theta') \geq \kappa |\theta - \theta'|^2.$$

Theorem 2.12 (Convergence in the regularized regime)**.** *Let $f \in \mathcal{A}_2$ and Assumptions 2.5, 2.8, and 2.11 hold. Assume further that $\beta := \frac{\sigma^2}{2}\kappa - C_2 - L > 0$, where C_2 and L are the constants given in Theorem 2.4. Then there exists a unique solution ν^* of (14) which is the global maximizer ν^* of $J^{\tau,\sigma}$ in $\mathcal{P}_2(\mathbb{R}^d)$. Moreover, if $(\nu_t)_{t\geq 0}$ is the solution of (15) for a given $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, then for all $t \in \mathbb{R}_+$,*

$$W_2(\nu_t, \nu^*) \leq e^{-\beta t} W_2(\nu_0, \nu^*)\,.$$

Theorem 2.12 will be proved in Section A.2.9. The exponential convergence holds also in the total variation norm (Butkovsky, 2014)[Thm. 3.1], in W_1 (Bogachev et al., 2018)[Remark 4.2], and in W_p using the method of Theorem 2.13 with the Itô formula applied to higher powers, see also (Šiška & Szpruch, 2020). From the general theory of nonlinear Fokker–Planck–Kolmogorov equations, one only expects existence and uniqueness of solutions to the stationary equation if $\sigma > 0$ is large relative to the constants appearing in Theorem 2.4 (see, e.g., (Bogachev et al., 2019)[Ex. 1.1], (Bogachev et al., 2018)[Ex. 4.3], (Manita et al., 2015)[Sec. 6]). The exponential convergence of the policy gradient flow for large σ can be regarded as an interpolation between the neural tangent kernel regime, where the neural network can be linearized around the initialization/prior, and the mean-field regime, where the distribution is evolving with the training time (see (Mei et al., 2019)).

The following theorem estimates the sensitivity of solutions of (15) and (14) on ν_0, τ, and σ.

Theorem 2.13 (Sensitivity in W_2)**.** *Let $f \in \mathcal{A}_2$ and Assumptions 2.5, 2.8, and 2.11 hold. Let $\sigma, \tau, \sigma', \tau' \geq 0$ and $\nu_0, \nu_0' \in \mathcal{P}_2(\mathbb{R}^d)$. Let $(\nu_t)_{t\geq 0}$ and $(\nu_t')_{t\geq 0}$ be the solutions of (15) with parameters and initial data σ, τ, ν_0 and σ', τ', ν_0', respectively. Then for all $\ell > 0$ and $t \in \mathbb{R}_+$,*

$$\begin{aligned} W_2^2(\nu_t, \nu_t') &\leq e^{-2\beta_\ell t} W_2^2(\nu_0, \nu_0') \\ &\quad + \frac{|\sigma^2 - \sigma'^2|}{8\ell} \int_0^t \int_{\mathbb{R}^d} e^{2\beta_\ell (s-t)} |\nabla U(\theta)|^2 \nu_s'(d\theta)\, ds \\ &\quad + \frac{1}{2\beta_\ell} \left(D|\tau - \tau'| + d|\sigma - \sigma'|^2\right)\left(1 - e^{-2\beta_\ell t}\right), \end{aligned} \tag{18}$$

where $\beta_\ell := \frac{\sigma^2}{2}\kappa - C_2(\tau) - L(\tau) - \ell|\sigma^2 - \sigma'^2|$ and $C_2(\tau)$ and $L(\tau)$ are the constants obtained in Theorem 2.4. Moreover, if $\beta := \frac{\sigma^2}{2}\kappa - C_2(\tau) - L(\tau) > 0$ and ν^* and ν'^* are solutions of (14) with σ, τ and σ', τ', respectively, then for all $\ell > 0$ such that $\beta_\ell = \beta - \ell|\sigma^2 - \sigma'^2| > 0$, we have

$$\begin{aligned} W_2^2(\nu^*, \nu'^*) &\leq \frac{|\sigma^2 - \sigma'^2|}{16\ell\beta_\ell} \int_{\mathbb{R}^d} |\nabla U(\theta)|^2 \nu'^*(d\theta) \\ &\quad + \frac{1}{2\beta_\ell}\left(D|\tau - \tau'| + d|\sigma - \sigma'|^2\right). \end{aligned} \tag{19}$$

Theorem 2.13 will be proved in Section A.2.8. Assumption 2.5 and Lemma A.11, proved in Section A.2.8, yield bounds on $\int_{\mathbb{R}^d} |\nabla U(\theta)|^2 \nu_t'(d\theta)$ for all $\sigma', \tau' \geq 0$ and $\nu_0' \in \mathcal{P}_2(\mathbb{R}^d)$ and bounds on $\int_{\mathbb{R}^d} |\nabla U(\theta)|^2 \nu'^*(d\theta)$ for all $\sigma', \tau' \geq 0$ such that either $\sigma'^2\kappa > 2C_2(\tau')$ or both $\nabla U(0) = 0$ and $\sigma' > 0$.

As a corollary of Theorems 2.4 and 2.13, we obtain the sensitivity with respect to the MDP value function.

Corollary 2.14 (Sensitivity of value function)**.** *Under the assumptions of Theorem 2.13 with $(\nu_t)_{t\geq 0}$, $(\nu_t')_{t\geq 0}$, ν^*, and ν'^* defined accordingly, for all $\hat{\tau} \geq 0$,*

$$|V_{\hat{\tau}}^{\pi_{\nu_t}}(\rho) - V_{\hat{\tau}}^{\pi_{\nu_t'}}(\rho)| \leq C_1(\hat{\tau}) W_2(\nu_t, \nu_t')$$

$$\text{and} \quad |V_{\hat{\tau}}^{\pi_{\nu^*}}(\rho) - V_{\hat{\tau}}^{\pi_{\nu'^*}}(\rho)| \leq C_1(\hat{\tau}) W_2(\nu^*, \nu'^*)\,,$$

where $W_2(\nu_t, \nu_t')$ and $W_2(\nu^*, \nu'^*)$ can be estimated by the square-roots of (18) and (19), respectively.

As discussed in the literature review, the authors of (Agazzi & Lu, 2020) showed that if $(\nu_t^0)_{t\in\mathbb{R}_+}$ is a solution of (15) with $\sigma = 0$ and $\tau > 0$ such that ν_t^0 converges to a solution $\nu^{0,*}$ of (14) in W_2 with full support, then $V_\tau^{\pi_{\nu^{0,*}}}(s) = V_\tau^*(s)$ for all $s \in S$, provided f is expressive enough (see (Agazzi & Lu, 2020)[Asm. 1]). Letting $\nu^{\sigma,*}$ be as in Theorem 2.12, by Corollary 2.14, we find that for all $\ell > 0$,

$$\begin{aligned} &|V_\tau^{\pi_{\nu^{\sigma,*}}}(\rho) - V_\tau^*(\rho)| \\ &\leq \frac{\sigma C_1}{4\sqrt{\ell\beta}} \left(\int_{\mathbb{R}^d} |\nabla U(\theta)|^2 \nu^{0,*}(d\theta)\right)^{\frac{1}{2}} + \frac{\sigma C_1}{\sqrt{2\beta}}\,, \end{aligned}$$

where $\beta := \frac{\sigma^2}{2}\kappa - C_2 - L - \ell|\sigma^2| > 0$. However, as discussed above, is not clear why, in the setting of (Agazzi & Lu, 2020), one would expect $\lim_{t\to\infty} \nu_t^0 = \nu^{0,*}$ in W_2.

3. Examples

3.1. The Effect of Regularization on Rate of Convergence

Consider the bandit setting. Let $S = \emptyset$, $A = \mathbb{R}^d$, $\gamma = 0$, $\mu(da) = da$, $\ell \in \mathbb{R}^d$, $\lambda > 0$, and $r(a) = \ell \cdot a - \lambda|a|^2$. While this example does not formally satisfy our assumptions due to unboundedness, we could extend our analysis to include this example with more cumbersome assumptions. The optimal policy in this setting is $\pi_\tau^*(da) \sim \mathcal{N}(\frac{\ell}{2\lambda}, \frac{\tau}{2\lambda} I_d)$ for

$\tau > 0$. We let $f(\theta, a) = -\frac{\lambda}{\tau}|a-\theta|^2$, which implies that $\pi_\nu \sim \mathcal{N}(\int_{\mathbb{R}^d} \theta\nu(d\theta), \frac{\tau}{2\lambda} I_d)$. For a given $m_U \in \mathbb{R}^d$ and $\sigma_U > 0$, we let $U(\theta) = \frac{d}{2}\ln 2\pi\sigma_U^2 + \frac{1}{2\sigma_U^2}|\theta - m_U|^2$. The first-order condition for the objective

$$J^{\tau,\sigma}(\nu) = \int_{\mathbb{R}^d} \theta^\top \ell \nu(d\theta) - \lambda\left(\frac{\tau}{2\lambda} + \left|\int_{\mathbb{R}^d} \theta\nu(d\theta)\right|^2\right) + \frac{\tau d}{2}\left(\ln 2\pi\sigma_f^2 + 1\right) - \frac{\sigma^2}{2}\mathrm{KL}(\nu|e^{-U})$$

is $\int_{\mathbb{R}^d} \theta\nu^*(d\theta) = \frac{\ell}{2\lambda}$ if $\sigma = 0$ and

$$\nu^* \sim e^{-\frac{1}{2\sigma_U^2}\left|\theta - \frac{m_U\sigma^2}{\sigma^2+4\lambda\sigma_U^2} - \frac{2\sigma_U^2\ell}{\sigma^2+4\lambda\sigma_U^2}\right|^2}$$

if $\sigma > 0$, where we have solved for the mean in the RHS of (13) using that the measure ν^* is Gaussian. We see that if $\sigma = 0$, then there are infinitely many critical points. Moreover, if $\sigma > 0$, then for any critical point, we have

$$\pi_{\nu^*} \sim \mathcal{N}\left(\frac{m_U\sigma^2}{\sigma^2+4\lambda\sigma_U^2} + \frac{2\sigma_U^2\ell}{\sigma^2+4\lambda\sigma_U^2}, \frac{\tau}{2\lambda} I_d\right).$$

We see that if $\sigma = 0$, then any critical point satisfies $\pi_{\nu^*} = \pi_\tau^*$, and thus $\nu^* \in \mathrm{argmax}_{\nu\in\mathcal{P}_2(\mathbb{R}^d)} V^{\pi_\nu}$. For $m_U = 0$ and small $\sigma > 0$, the mean of π_{ν^*} is approximately $\ell/2\lambda$, and hence $\pi_{\nu^*} \approx \pi_\tau^*$, which Theorem 2.13 and Corollary 2.14 illustrate. Furthermore, the probabilistic representation of the gradient flow (17) is given by

$$d\theta_t = \left(\ell - 2\lambda\mathbb{E}\theta_t - \frac{\sigma^2}{2\sigma_U^2}(\theta_t - m_U)\right)dt + \sigma dW_t$$

which is an Ornstein–Uhlenbeck-like process; the mean can be solved for explicitly and substituted back into the equation. For $\sigma > 0$, the process has a unique invariant measure and converges exponentially fast, which can be shown using the same proof as that of Theorem 2.12. We expect a similar phenomenon in the linear quadratic regulator setting, provided the function f is chosen appropriately so that its mean is Kx, where K is the control gain. For more complicated examples without any special structure (e.g., convexity) on the unregularized objective, we expect that one needs to take σ larger to obtain exponential convergence at the level of the parameterization.

3.2. Explicit Constants in Theorem 2.4

The following example, while a bit contrived, illustrates the general procedure to determine the constants in Theorem 2.4. The bandit setting is characterized by $S = \emptyset$ and $\gamma = 0$ with a bounded reward function $r : A \to \mathbb{R}$. We let $f(\theta, a) = \psi(\theta)\cdot\tanh(g(a))$, where $\psi : \mathbb{R} \to \mathbb{R}$ is a smooth rescaling function such that ψ and its derivatives up to order two are bounded by $|\psi|_\infty$, and where for simplicity we take $g : A \to \mathbb{R}^d$ to be a random feature map. For given $m \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ satisfying $\Sigma^{-1} \succeq \kappa I_d$, we let $U(\theta) := 2^{-1}\ln\det(2\pi\Sigma) + (\theta - m_U)^\top\Sigma^{-1}(\theta - m_U)$. In this setting, using Lemma 2.2 and 2.3, we find

$$\begin{aligned}&\nabla_\theta^k \frac{\delta J^{\tau,0}}{\delta\nu}(\nu,\theta)\\ &= \nabla_\theta^k\psi(\theta)\int_A\left(r(a) - \tau\tanh(g(a))\cdot\int_{\mathbb{R}^d}\psi(\theta')\nu(d\theta')\right)\\ &\times\left(\tanh(g(a)) - \int_A \tanh(g(a'))\pi_\nu(da')\right)\pi_\nu(da).\end{aligned}$$

Using (31), we obtain

$$C = C_k := 2|\psi|_\infty(|r|_\infty + \tau|\psi|_\infty), \quad k \in \mathbb{N},$$

$$\text{and} \qquad L := (6|r|_\infty + 2\tau)|\psi|_\infty^2 + 6\tau|\psi|_\infty^3,$$

and hence $\frac{\sigma^2}{2}\kappa > C + L$ in Theorem 2.12 is explicit.

4. Conclusion

We identified conditions that allow us to extend the work of (Mei et al., 2020), where exponential convergence of the policy gradient method has been established in the tabular case to the continuous state and action setting with policies parameterized by two-layer neural networks in the mean-field regime. This was enabled by introducing entropic regularization in the space of parameterizations ($\sigma > 0$) rather than just space of policies as is commonly done in entropy regularized in RL and careful analysis of the corresponding non-linear Fokker–Planck–Kolmogorov equations. The results and techniques of this paper open up many possible research directions of which we mention a few. It should be possible to extend the one-hidden-layer mean-field setting to recurrent neural network approximation, see (Weinan, 2017; Hu et al., 2019; Jabir et al., 2019). Moreover, it should be possible to extend the techniques identified here to actor-critic-type algorithms where the policy and the Q function are approximated by their own mean-field neural networks (Agarwal et al., 2020; Sirignano & Spiliopoulos, 2019). Furthermore, since the main focus of RL is in the regime where the model is not known, it would be interesting to explore the non-idealized setting, where regret bounds are proved in terms of the number of samples of state and action pairs.

Acknowledgements

We would like to thank the reviewers for their thoughtful comments and efforts toward improving our manuscript. JML is grateful for support from US AFOSR Grant FA8655-21-1-7034. LS is grateful for support from support from UKRI Prosperity Partnership Scheme (FAIR) under the EPSRC Grant EP/V056883/1.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Agazzi, A. and Lu, J. Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. *arXiv preprint arXiv:2010.11858*, 2020.

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

Bertsekas, D. P. and Shreve, S. *Stochastic optimal control: the discrete-time case*. Athena Scientific, 2004.

Bogachev, V. I., Krylov, N. V., Röckner, M., and Shaposhnikov, S. V. *Fokker–Planck–Kolmogorov Equations*, volume 207. American Mathematical Soc., 2015.

Bogachev, V. I., Röckner, M., and Shaposhnikov, S. V. Distances between transition probabilities of diffusions and applications to nonlinear Fokker–Planck–Kolmogorov equations. *Journal of Functional Analysis*, 271(5):1262–1300, 2016.

Bogachev, V. I., Kirillov, A. I., and Shaposhnikov, S. V. Distances between stationary distributions of diffusions and solvability of nonlinear Fokker–Planck–Kolmogorov equations. *Theory of Probability and its Applications*, 62 (1):12–34, 2018.

Bogachev, V. I., Röckner, M., and Shaposhnikov, S. V. Convergence in variation of solutions of nonlinear Fokker–Planck–Kolmogorov equations to stationary measures. *Journal of Functional Analysis*, 276(12):3681–3713, 2019.

Butkovsky, O. A. On ergodic properties of nonlinear Markov chains and stochastic McKean-Vlasov equations. *Theory of Probability and its Applications*, 58(4):661–674, 2014.

Carmona, R. and Delarue, F. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pp. 3040–3050, 2018.

Delarue, F. and Tse, A. Uniform in time weak propagation of chaos on the Torus. *arXiv preprint arXiv:2104.14973*, 2021.

Dupuis, P. and Ellis, R. S. *A weak convergence approach to the theory of large deviations*. John Wiley & Sons, Inc., New York, 1997.

Fox, R., Pakman, A., and Tishby, N. Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*, 2015.

Funaki, T. A certain class of diffusion processes associated with nonlinear parabolic equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 67 (3):331–348, 1984.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

Hammersley, W. R. P., Šiška, D., and Szpruch, Ł. McKean-Vlasov SDEs under measure dependent Lyapunov conditions. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 57(2):1032–1057, 2021. ISSN 0246-0203.

Hernández-Lerma, O. and Lasserre, J. B. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.

Hu, K., Kazeykina, A., and Ren, Z. Mean-field Langevin system, optimal control and deep neural networks. *arXiv:1909.07278*, 2019.

Hu, K., Ren, Z., Šiška, D., and Szpruch, Ł. Mean-field langevin dynamics and energy landscape of neural networks. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(4):2043–2065, 2021.

Jabir, J.-F., Šiška, D., and Szpruch, Ł. Mean-field neural ODEs via relaxed optimal control. *arXiv preprint arXiv:1912.05475*, 2019.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Kinderlehrer, D. and Stampacchia, G. *An introduction to variational inequalities and their applications*. SIAM, 2000.

Komorowski, T. and Walczuk, A. Central limit theorem for Markov processes with spectral gap in the Wasserstein metric. *Stochastic Processes and their Applications*, 122 (5):2155–2184, 2012.

Krylov, N. V. *Controlled diffusion processes*, volume 14. Springer Science & Business Media, 2008.

Kunze, M. A Pettis-type integral and applications to transition semigroups. *Czechoslovak mathematical journal*, 61 (2):437–459, 2011.

Lasota, A. and Mackey, M. C. *Chaos, fractals, and noise: stochastic aspects of dynamics*, volume 97. Springer Science & Business Media, 2013.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Softmax policy gradient methods can take exponential time to converge. *arXiv preprint arXiv:2102.11270*, 2021.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in Neural Information Processing Systems*, 32:10565–10576, 2019.

Manita, O. and Shaposhnikov, S. Nonlinear parabolic equations for measures. *St. Petersburg Mathematical Journal*, 25(1):43–62, 2014.

Manita, O. A., Romanov, M. S., and Shaposhnikov, S. V. On uniqueness of solutions to nonlinear Fokker–Planck–Kolmogorov equations. *Nonlinear Analysis*, 128:199–226, 2015.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pp. 2388–2464. PMLR, 2019.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Otto, F. and Villani, C. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

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

Rotskoff, G. M. and Vanden-Eijnden, E. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050:22, 2018.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Sirignano, J. and Spiliopoulos, K. Asymptotics of reinforcement learning with neural networks. *arXiv preprint arXiv:1911.07304*, 2019.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 2021.

Šiška, D. and Szpruch, Ł. Gradient flows for regularized stochastic control problems. *arXiv preprint arXiv:2006.05956*, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning*. MIT Press, 2018.

Vieillard, N., Kozuno, T., Scherrer, B., Pietquin, O., Munos, R., and Geist, M. Leverage the average: an analysis of KL regularization in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12163–12174, 2020.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in Starcraft II using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgQfkSYDS.

Weinan, E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. Modeling interaction via the principle of maximum causal entropy. In *ICML*, 2010.

A. Appendix

A.1. Notation and Definitions

Let $\mathbb{R}_+ := [0, \infty)$. For given $k \in \mathbb{N}_0 \cup \infty$, let $C^k(\mathbb{R}^d)$ denote the space of k-times continuously differentiable functions and $C_b^k(\mathbb{R}^d)$ the subspace of functions of $C^k(\mathbb{R}^d)$ for which the function and all the derivatives up to order k are bounded. Let $C_c^k(\mathbb{R}^d)$ denote the subspace of $C^k(\mathbb{R}^d)$ of functions with compact support. We will use standard notation L^p, $p \in [1, \infty]$, for Lebesgue spaces of integrable functions.

Let (E, d) denote a complete separable metric space (i.e., a Polish space). We always equip a Polish space with its Borel sigma-field $\mathcal{B}(E)$. Denote by $B_b(E)$ the space of bounded strongly measurable functions $f : E \to \mathbb{R}$ endowed with the supremum norm $|f|_{B_b(E)} = \sup_{x \in E} |f(x)|$. Denote by $\mathcal{M}(E)$ the Banach space of signed measures (finite) μ on E endowed with the total variation norm $|\mu|_{\mathcal{M}(A)} = |\mu|(E)$, where $|\mu|$ is the total-variation measure. We note that if $\mu = f d\rho$, where $\rho \in \mathcal{M}_+(E)$ is a non-negative measure and $f \in L^1(E, \rho)$, then $|\mu|_{\mathcal{M}(E)} = |f|_{L^1(E,\rho)}$. We denote by $\mathcal{P}(E) \subset \mathcal{M}(E)$ the convex subset of probability measures on E. For $\mu, \mu' \in \mathcal{P}(E)$ such that μ is absolutely continuous with respect to μ', the relative entropy of μ with respect to μ' (or Kullback-Liebler divergence of μ relative to μ') is defined by

$$\mathrm{KL}(\mu|\mu') = \int_E \ln \frac{d\mu}{d\mu'}(x)\mu(dx)\,.$$

For measure $\mu \in \mathcal{M}_+(E)$ and measurable function $f : E \to \mathbb{R}$, let

$$\mu - \operatorname*{ess\,sup}_{x \in E} f = \inf\{c \in \mathbb{R} : \mu\{x \in E : f(x) > c\} = 0\}\,.$$

It is convenient to have notation for measurable functions $k : E_1 \to \mathcal{M}(E_2)$ for given Polish spaces (E_1, d_1) and (E_2, d_2). For example, $P : S \to \mathcal{P}(S \times A)$ will denote a controlled transition probability and $\pi : S \to \mathcal{P}(A)$ will denote a stochastic policy. Denote by $b\mathcal{K}(E_1|E_2)$ the Banach space of bounded signed kernels $k : E_2 \to \mathcal{M}(E_1)$ endowed with the norm $|k|_{b\mathcal{K}(E_1|E_2)} = \sup_{x \in E_2} |k(x)|_{\mathcal{M}(E_1)}$; that is, $k(U|\cdot) : E_2 \to \mathbb{R}$ is measurable for all $U \in \mathcal{M}(E_1)$ and $k(\cdot|x) \in \mathcal{M}(E_1)$ for all $x \in E_2$. For a fixed positive reference measure $\mu \in \mathcal{M}(E_1)$, we denote by $b\mathcal{K}_\mu(E_1|E_2)$ the space of bounded kernels that are absolutely continuous with respect to μ.

Every kernel $k \in b\mathcal{K}(E_1|E_2)$ induces bounded linear operators $T_k \in \mathcal{L}(\mathcal{M}(E_2), \mathcal{M}(E_1))$ and $S_k \in \mathcal{L}(B_b(E_1), B_b(E_2))$ defined by

$$T_k\mu(dy) = \mu k(dy) = \int_{E_2} \mu(dx)k(dy|x) \quad \text{and} \quad S_k f(x) = \int_{E_1} k(dy|x)f(y)\,,$$

respectively. Moreover, by (Kunze, 2011)[Ex. 2.3 and Prop. 3.1], we have

$$|k|_{b\mathcal{K}(E_1|E_2)} = \sup_{x \in E_2} \sup_{\substack{h \in B_b(E_1) \\ |h|_{B_b(E_1)} \le 1}} \int_{E_1} h(y)k(dy|x) = |S_k|_{\mathcal{L}(B_b(E_1), B_b(E_2))} = |T_k|_{\mathcal{L}(\mathcal{M}(E_2), \mathcal{M}(E_1))}\,,$$

where the latter are operator norms. Thus, $b\mathcal{K}(E|E)$ is a Banach algebra with the product defined via composition of the corresponding linear operators; in particular, for a given $k \in b\mathcal{K}(E|E)$,

$$T_k^n\mu(dy) = \mu k^n(dy) = \int_{E^n} \mu(dx_0)k(dx_1|x_0)k(dx_2|x_1)\cdots k(dx_{n-1}|x_{n-2})k(dy|x_{n-1})\,.$$

Notice that if $f \in L^\infty(E_1, \mu)$ and $k \in b\mathcal{K}_\mu(E_1|E_2)$, then for all $x \in E_2$,

$$S_k f(x) = \int_{E_1} \mu(dy)\frac{dk}{d\mu}(y|x)f(y) \le |f|_{L^\infty(E_1,\mu)} \left|\frac{dk}{d\mu}(\cdot|x)\right|_{L^1(E_1,\mu)} \le |f|_{L^\infty(E_1,\mu)}|k|_{b\mathcal{K}(E_1|E_2)}\,. \tag{20}$$

We denote by $\mathcal{P}(E_1|E_2)$ the convex subspace of $P \in b\mathcal{K}(E_1|E_2)$ such that $P(\cdot|x) \in \mathcal{P}(E_1)$ for all $x \in E_2$; such kernels are referred to as stochastic kernels. A stochastic kernel $P \in \mathcal{P}(E_1|E_2)$ is said to be strongly Feller if $\int_{E_1} P(dy|x)f(y)$ is continuous in $x \in E_2$ for all $f \in B_b(E_1)$. For a fixed positive reference measure $\mu \in \mathcal{M}(E_1)$, we denote by $\mathcal{P}_\mu(E_1|E_2)$ the space of kernels that are absolutely continuous with respect to μ. A bounded kernel $k \in b\mathcal{K}(E_1|E_2)$ is thus strongly Feller if the range of S_k lies in the space of continuous functions on E_2.

For given $p \in \mathbb{N}$, let $\mathcal{P}_p(\mathbb{R}^d)$ denote the set of probability measures with finite p-th moments. The p-Wasserstein distance between $\nu, \nu' \in \mathcal{P}_p(\mathbb{R}^d)$ is given by

$$W_p(\nu, \nu') = \inf_{\pi \in \Pi(\nu, \nu')} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |\theta - \theta'|^p \pi(d\theta, d\theta') \right)^{\frac{1}{p}}$$

where $\Pi(\nu, \nu')$ denotes the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with ν and ν' as first and second marginals, respectively. For all $p, p' \in \mathbb{N}$ with $p' > p$ and $\nu, \nu' \in \mathcal{P}_{p'}(E)$, we have

$$W_p(\nu, \nu') \leq W_{p'}(\nu, \nu') \,. \tag{21}$$

The set $\mathcal{P}_p(\mathbb{R}^d)$ endowed with the topology induced by the p-Wasserstein distance is a complete separable metric space (Villani, 2009)[Thm. 6.18]. The Kantorovich–Rubinstein duality theorem (Villani, 2009)[Thm. 5.10] implies that for all $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}^d)$,

$$W_1(\nu, \nu') = \sup_{\phi \in \mathrm{Lip}_1(\mathbb{R}^d)} \int_{\mathbb{R}^d} \phi(\theta)(\nu - \nu')(d\theta) \,, \tag{22}$$

where $\mathrm{Lip}_1(\mathbb{R}^d)$ denotes the space of all functions $\phi : \mathbb{R}^d \to \mathbb{R}$ with a Lipschitz constant 1. Henceforth, we let $\mathcal{P}_0(\mathbb{R}^d) = \mathcal{P}(\mathbb{R}^d)$ and recall that this convex subset of $\mathcal{M}(\mathbb{R}^d)$ is endowed with the topology of total variation distance. We denote the Lebesgue measure on $\mathbb{R}^d$ by λ. Denote by $\mathcal{P}_p^{\mathrm{ac}}(\mathbb{R}^d)$ the subset of $\mathcal{P}_p(\mathbb{R}^d)$ consisting of measures that absolutely continuous with respect to λ. We abuse notation and identify $\nu = \frac{d\nu}{d\theta} := \frac{d\nu}{d\lambda}$ in $\mathcal{P}_p^{\mathrm{ac}}(\mathbb{R}^d)$. Let $C(\mathbb{R}_+; \mathcal{P}_p(\mathbb{R}^d))$ denote the space of continuous functions $\nu : \mathbb{R}^+ \to \mathcal{P}_p(\mathbb{R}^d)$.

Definition A.1. Let $p \in \mathbb{N}_0$. Let $(V, |\cdot|_V)$ denote a Banach space. We say a function $u : \mathcal{P}_p(\mathbb{R}^d) \to V$ has a linear functional derivative if there exists a continuous and bounded function $\frac{\delta u}{\delta \nu} : \mathcal{P}_p(\mathbb{R}^d) \times \mathbb{R}^d \to V$ such that for all $\nu, \nu' \in \mathcal{P}_p(\mathbb{R}^d)$,

$$\lim_{\substack{\varepsilon \to 0 \\ \varepsilon \in [0,1]}} \frac{u(\nu + \varepsilon(\nu' - \nu)) - u(\nu)}{\varepsilon} = \int_{\mathbb{R}^d} \frac{\delta u}{\delta \nu}(\nu, \theta)(\nu' - \nu)(d\theta) \,.$$

Owing to the fundamental theorem of calculus a function $u : \mathcal{P}_p(\mathbb{R}^d) \to V$ has a linear functional derivative if and only if there exists a continuous and bounded function $\frac{\delta u}{\delta \nu} : \mathcal{P}_p(\mathbb{R}^d) \times \mathbb{R}^d \to V$ such that for all $\nu, \nu' \in \mathcal{P}_p(\mathbb{R}^d)$,

$$u(\nu') - u(\nu) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta u}{\delta \nu}(\nu + \varepsilon(\nu' - \nu), \theta)(\nu' - \nu)(d\theta)\, d\varepsilon \,. \tag{23}$$

A.2. Proofs

A.2.1. Auxiliary results

We need the following version of the fundamental theorem of calculus of variations.

Lemma A.2 (c.f. Lem. 33 in (Jabir et al., 2019))**.** *Let $\nu \in \mathcal{P}(\mathbb{R}^d)$ and $u : \mathbb{R}^d \to \mathbb{R}$ be measurable. Assume that $\int_{\mathbb{R}^d} u(\theta)\nu(d\theta)$ exists and is finite. If for all $\nu' \in \mathcal{P}(\mathbb{R}^d)$*

$$\int_{\mathbb{R}^d} u(\theta)(\nu' - \nu)(d\theta) \geq 0 \,, \tag{24}$$

then u is a constant ν-a.e.. Moreover, if $\nu \in \mathcal{P}_2^{\mathrm{fe}}(\mathbb{R}^d)$ and (24) holds for all $\nu' \in \mathcal{P}_2^{\mathrm{fe}}(\mathbb{R}^d)$, then the conclusion still holds.

We will also require the following property of relative entropy.

Lemma A.3. *Let $U : \mathbb{R} \to \mathbb{R}_+$ be measurable such that $\int_{\mathbb{R}^d} e^{-U(\theta)} d\theta = 1$. Let $\gamma(d\theta) = e^{-U(\theta)} d\theta$. Moreover, assume that there exist constants $C > 0$ and $p \in \mathbb{N}$ such that $|U(\theta)| \leq C_U(1 + |\theta|^p)$ for all $\theta \in \mathbb{R}^d$. Then for all $\nu \in \mathcal{P}_p(\mathbb{R}^d)$,*

$$\int_{\mathbb{R}^d} \ln \frac{d\nu}{d\gamma}(\theta)\nu(d\theta) < \infty \quad \Leftrightarrow \quad \int_{\mathbb{R}^d} \left| \ln \frac{d\nu}{d\gamma}(\theta) \right| \nu(d\theta) < \infty \,,$$

$$\Leftrightarrow \quad \int_{\mathbb{R}^d} \ln \frac{d\nu}{d\theta}(\theta)\nu(d\theta) < \infty \quad \Leftrightarrow \quad \int_{\mathbb{R}^d} \left| \ln \frac{d\nu}{d\theta}(\theta) \right| \nu(d\theta) < \infty \,.$$

Proof. First notice that

$$\int_{\mathbb{R}^d} \ln \frac{d\nu}{d\gamma}(\theta)\nu(d\theta) = \int_{\mathbb{R}^d} \left(\ln \frac{d\nu}{d\gamma}(\theta)\right)^+ \nu(d\theta) - \int_{\mathbb{R}^d} \frac{d\nu}{d\gamma}(\theta) \left(\ln \frac{d\nu}{d\gamma}(\theta)\right)^- \gamma(d\theta)$$

and

$$\int_{\mathbb{R}^d} \left|\ln \frac{d\nu}{d\gamma}(\theta)\right| \nu(d\theta) = \int_{\mathbb{R}^d} \left(\ln \frac{d\nu}{d\gamma}(\theta)\right)^+ \nu(d\theta) + \int_{\mathbb{R}^d} \frac{d\nu}{d\gamma}(\theta) \left(\ln \frac{d\nu}{d\gamma}(\theta)\right)^- \gamma(d\theta),$$

where for all $x \in \mathbb{R}$, $x^+ = \max(x, 0)$ and $x^- = -\min(x, 0)$. Using the fact that the function $g(s) = s(\ln s)^-$ is bounded on $\mathbb{R}_+$, we find that

$$\int_{\mathbb{R}^d} \ln \frac{d\nu}{d\gamma}(\theta)\nu(d\theta) < \infty \quad \Leftrightarrow \quad \int_{\mathbb{R}^d} \left|\ln \frac{d\nu}{d\gamma}(\theta)\right| \nu(d\theta) < \infty\,.$$

The triangle and reverse triangle inequalities imply that

$$\int_{\mathbb{R}^d} \left|\ln \frac{d\nu}{d\theta}(\theta)\right| \nu(d\theta) - \int_{\mathbb{R}^d} U(\theta)\nu(d\theta) \le \int_{\mathbb{R}^d} \left|\ln \frac{d\nu}{d\gamma}(\theta)\right| \nu(d\theta) \le \int_{\mathbb{R}^d} \left|\ln \frac{d\nu}{d\theta}(\theta)\right| \nu(d\theta) + \int_{\mathbb{R}^d} U(\theta)\nu(d\theta)\,.$$

By assumption, we have

$$\int_{\mathbb{R}^d} U(\theta)\nu(d\theta) \le C \int_{\mathbb{R}^d} (1 + |\theta|)^p \nu(d\theta) < \infty$$

and hence

$$\int_{\mathbb{R}^d} \left|\ln \frac{d\nu}{d\gamma}(\theta)\right| \nu(d\theta) < \infty \quad \Leftrightarrow \quad \int_{\mathbb{R}^d} \nu(\theta) \left|\ln \frac{d\nu}{d\theta}(\theta)\right| \nu(d\theta) < \infty\,.$$

Combining the above equivalences, we complete the proof. □

A.2.2. AUXILIARY BOUNDS

In the proof of the main results, we will need the following auxiliary lemma, which concerns the boundedness of the i) log-density (relative to reference measure μ) of the class of mean-field policies, ii) value function and iii) state-action value function.

Lemma A.4. *Assume that $f \in \mathcal{A}_0$. For all $\tau, \tau' \ge 0$, $\nu \in \mathcal{P}(\mathbb{R}^d)$, $s \in S$, and $\mu - a.e.\, a \in A$, we have*

$$\left|\ln \frac{d\pi_\nu}{d\mu}(a|s)\right| \le 2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\,, \quad |V_\tau^{\pi_\nu}(s)| \le \frac{1}{1-\gamma}\left(|r|_{B_b(S\times A)} + \tau\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right)\,,$$

$$|Q_\tau^{\pi_\nu}(s,a)| \le \frac{1}{1-\gamma}\left(|r|_{B_b(S\times A)} + \gamma\tau\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right)\,,$$

$$|V_{\tau'}^{\pi_\nu}(s) - V_\tau^{\pi_\nu}(s)| \le \frac{|\tau' - \tau|}{1-\gamma}\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\,,$$

$$\textit{and} \quad |Q_{\tau'}^{\pi_\nu}(s,a) - Q_\tau^{\pi_\nu}(s,a)| \le \frac{\gamma|\tau' - \tau|}{1-\gamma}\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\,.$$

Proof. Let $a \in A$, $s \in S$, and $\nu \in \mathcal{P}(\mathbb{R}^d)$ be arbitrarily given. Estimating directly, we find

$$\begin{aligned}\left|\ln \frac{d\pi_\nu}{d\mu}(a|s)\right| &= \left|\int_{\mathbb{R}^d} f(\theta, s, a)\nu(d\theta) - \ln\left(\int_A \exp\left(\int_{\mathbb{R}^d} f(\theta, s, a')\nu(d\theta)\right)\mu(da')\right)\right| \\ &\le 2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\,.\end{aligned} \tag{25}$$

Using (8), (25), $|\pi_\nu|_{b\mathcal{K}(A|S)} = 1$ and $|d^{\pi_\nu}|_{b\mathcal{K}(S|S)} = 1$, we obtain

$$\begin{aligned}|V_\tau^{\pi_\nu}(s)| &= \frac{1}{1-\gamma}\left|\int_S \int_A \left(r(s', a) - \tau \ln \frac{d\pi_\nu}{d\mu}(a|s)\right) d^{\pi_\nu}(ds'|s)\pi_\nu(da|s')\right| \\ &\le \frac{1}{1-\gamma}\left(|r|_{B_b(S\times A)} + \tau\left|\ln \frac{d\pi_\nu}{d\mu}\right|_{B_b(S\times A)}\right)|\pi_\nu|_{b\mathcal{K}(A|S)}|d^{\pi_\nu}|_{b\mathcal{K}(S|S)} \\ &\le \frac{1}{1-\gamma}\left(|r|_{B_b(S\times A)} + \tau\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right)\,.\end{aligned} \tag{26}$$

To estimate the state-action value function, we use (6), (26), and $|P|_{b\mathcal{K}(S|S\times A)} = 1$ to get

$$\begin{aligned}|Q_\tau^{\pi_\nu}(s,a)| &\le |r|_{B_b(S\times A)} + \gamma |V_\tau^{\pi_\nu}|_{B_b(S)} |P|_{b\mathcal{K}(S|S\times A)} \\ &\le |r|_{B_b(S\times A)} + \frac{\gamma}{1-\gamma}\left(|r|_{B_b(S\times A)} + \tau\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right) \\ &= \frac{1}{1-\gamma}\left(|r|_{B_b(S\times A)} + \gamma\tau\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right).\end{aligned}$$

The remaining inequalities are derived similarly using

$$V_{\tau'}^{\pi_\nu}(s) - V_\tau^{\pi_\nu}(s) = \frac{\tau - \tau'}{1-\gamma}\int_S \int_A \ln \frac{d\pi_\nu}{d\mu}(a|s)\pi_\nu(da|s')d^\pi(ds'|s),$$

which follows from (8). □

A.2.3. Proof of Lemma 2.2 and a corollary

Proof. Let $\nu, \nu' \in \mathcal{P}(\mathbb{R}^d)$ and define $\nu^\varepsilon = \nu + \varepsilon(\nu' - \nu)$ for $\varepsilon \in [0,1]$. We must show that

$$\lim_{\substack{\varepsilon \to 0 \\ \varepsilon \in [0,1]}} \frac{\pi_{\nu^\varepsilon}(da|s) - \pi_\nu(da|s)}{\varepsilon} = \int_{\mathbb{R}^d} \frac{\delta \pi_\nu}{\delta \nu}(\nu,\theta)(da|s)(\nu' - \nu)(d\theta), \tag{27}$$

where $\frac{\delta \pi_\nu}{\delta \nu}(\nu,\theta)$ is given by the right-hand-side of (10) and the limit is understood in $b\mathcal{K}(A|S)$. Recall from Section A.1 that

$$\begin{aligned}&\left|\frac{\pi_{\nu^\varepsilon} - \pi_\nu}{\varepsilon} - \int_{\mathbb{R}^d} \frac{\delta \pi_\nu}{\delta \nu}(\nu,\theta)(da|s)(\nu' - \nu)(d\theta)\right|_{b\mathcal{K}(A|S)} \\ &= \sup_{s\in S}\int_A \left|\frac{\frac{d\pi_{\nu^\varepsilon}}{d\mu}(\cdot|s) - \frac{d\pi_\nu}{d\mu}(\cdot|s)}{\varepsilon} - \int_{\mathbb{R}^d} \frac{d\frac{\delta \pi_\nu}{\delta \nu}(\nu,\theta)}{d\mu}(a|s)(\nu' - \nu)(d\theta)\right| \mu(da). \end{aligned} \tag{28}$$

For convenience, we introduce the unnormalized policy $\tilde{\pi} : \mathcal{P}(\mathbb{R}^d) \to b\mathcal{K}_\mu(A|S)$ given by

$$\tilde{\pi}(\nu)(da|s) = \tilde{\pi}_\nu(da|s) = \exp\left(\int_{\mathbb{R}^d} f(\theta,s,a)\nu(d\theta)\right)\mu(da).$$

For all $(s,a) \in S \times A$, we have

$$\begin{aligned}&\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\pi_\nu}{d\mu}(a|s) \\ &= \frac{\frac{d\tilde{\pi}_{\nu^\varepsilon}}{d\mu}(a|s)}{\tilde{\pi}_{\nu^\varepsilon}(A|s)} - \frac{\frac{d\tilde{\pi}_\nu}{d\mu}(a|s)}{\tilde{\pi}_\nu(A|s)} = \frac{\frac{d\tilde{\pi}_{\nu^\varepsilon}}{d\mu}(a|s)\tilde{\pi}_\nu(A|s)}{\tilde{\pi}_{\nu^\varepsilon}(A|s)\tilde{\pi}_\nu(A|s)} - \frac{\frac{d\tilde{\pi}_\nu}{d\mu}(a|s)\tilde{\pi}_{\nu^\varepsilon}(A|s)}{\tilde{\pi}_{\nu^\varepsilon}(A|s)\tilde{\pi}_\nu(A|s)} \\ &= \frac{\frac{d\tilde{\pi}_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\tilde{\pi}_\nu}{d\mu}(a|s)}{\tilde{\pi}_\nu(A|s)}\frac{\tilde{\pi}_\nu(A|s)}{\tilde{\pi}_{\nu^\varepsilon}(A|s)} + \frac{\frac{d\tilde{\pi}_\nu}{d\mu}(a|s)}{\tilde{\pi}_{\nu^\varepsilon}(A|s)}\frac{\tilde{\pi}_\nu(A|s) - \tilde{\pi}_{\nu^\varepsilon}(A|s)}{\tilde{\pi}_\nu(A|s)} \\ &= \left[\frac{\frac{d\tilde{\pi}_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\tilde{\pi}_\nu}{d\mu}(a|s)}{\tilde{\pi}_\nu(A|s)} + \frac{d\pi_\nu}{d\mu}(a|s)\frac{\tilde{\pi}_\nu(A|s) - \tilde{\pi}_{\nu^\varepsilon}(A|s)}{\tilde{\pi}_\nu(A|s)}\right]\frac{\tilde{\pi}_\nu(A|s)}{\tilde{\pi}_{\nu^\varepsilon}(A|s)} \\ &= \left[\frac{\frac{d\tilde{\pi}_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\tilde{\pi}_\nu}{d\mu}(a|s)}{\tilde{\pi}_\nu(A|s)} + \frac{d\pi_\nu}{d\mu}(a|s)\int_A \left(\frac{\frac{d\tilde{\pi}_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\tilde{\pi}_\nu}{d\mu}(a|s)}{\tilde{\pi}_\nu(A|s)}\right)\mu(da)\right]\frac{\tilde{\pi}_\nu(A|s)}{\tilde{\pi}_{\nu^\varepsilon}(A|s)}.\end{aligned}$$

Simple manipulation yields

$$\frac{d\tilde{\pi}_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\tilde{\pi}_\nu}{d\mu}(a|s) = \exp\left(\int_{\mathbb{R}^d} f(\theta,s,a)\nu(d\theta)\right)\left(\exp\left(\varepsilon\int_{\mathbb{R}^d} f(\theta,s,a)(\nu' - \nu)(d\theta)\right) - 1\right).$$

Taylor expanding the exponential function, we find

$$\begin{aligned}&\exp\left(\varepsilon\int_{\mathbb{R}^d} f(\theta,s,a)(\nu'-\nu)(d\theta)\right)-1\\&=\varepsilon\int_{\mathbb{R}^d} f(\theta,s,a)(\nu'-\nu)(d\theta)+\varepsilon^2\sum_{n=2}^{\infty}\varepsilon^{n-2}\frac{\left(\int_{\mathbb{R}^d} f(\theta,s,a)(\nu'-\nu)(d\theta)\right)^n}{n!}\,,\end{aligned}$$

and hence

$$\begin{aligned}\varepsilon^{-1}\frac{\frac{d\tilde\pi_{\nu^\varepsilon}}{d\mu}(a|s)-\frac{d\tilde\pi_\nu}{d\mu}(a|s)}{\tilde\pi_\nu(A|s)}&=\frac{d\pi_\nu}{d\mu}(a|s)\int_{\mathbb{R}^d} f(\theta,s,a)(\nu'-\nu)(d\theta)\\&\quad+\varepsilon\frac{d\pi_\nu}{d\mu}(a|s)\sum_{n=2}^{\infty}\varepsilon^{n-2}\frac{\left(\int_{\mathbb{R}^d} f(\theta,s,a)(\nu'-\nu)(d\theta)\right)^n}{n!}\,.\end{aligned}\tag{29}$$

Thus, using $\pi_\nu(A|s)=1$, we obtain

$$\begin{aligned}&\int_A\left|\varepsilon^{-1}\frac{\frac{d\tilde\pi_{\nu^\varepsilon}}{d\mu}(a|s)-\frac{d\tilde\pi_\nu}{d\mu}(a|s)}{\tilde\pi_\nu(A|s)}-\frac{d\pi_\nu}{d\mu}(a|s)\int_{\mathbb{R}^d} f(\theta,s,a)(\nu'-\nu)(d\theta)\right|\mu(da)\\&\qquad\le\varepsilon\exp\left(|f|_{\mathcal{A}_0}|\nu'-\nu|_{\mathcal{M}(\mathbb{R}^d)}\right)\end{aligned}$$

and

$$\int_A\varepsilon^{-1}\left|\frac{\frac{d\tilde\pi_{\nu^\varepsilon}}{d\mu}(a|s)-\frac{d\tilde\pi_\nu}{d\mu}(a|s)}{\tilde\pi_\nu(A|s)}\right|\mu(da)\le|f|_{\mathcal{A}_0}|\nu'-\nu|_{\mathcal{M}(\mathbb{R}^d)}+\varepsilon\exp\left(|f|_{\mathcal{A}_0}|\nu'-\nu|_{\mathcal{M}(\mathbb{R}^d)}\right).$$

The dominated convergence theorem implies that

$$\lim_{\substack{\varepsilon\to0\\\varepsilon\in[0,1]}}|\pi_{\nu^\varepsilon}-\pi_\nu|_{b\mathcal{K}(A|S)}=\lim_{\substack{\varepsilon\to0\\\varepsilon\in[0,1]}}\sup_{s\in S}\int_A\left|\frac{d\pi_{\nu^\varepsilon}}{d\mu}(\cdot|s)-\frac{d\pi_\nu}{d\mu}(\cdot|s)\right|\mu(da)=0\,,\tag{30}$$

and hence

$$\lim_{\substack{\varepsilon\to0\\\varepsilon\in[0,1]}}\frac{\tilde\pi_\nu(A|s)}{\tilde\pi_{\nu^\varepsilon}(A|s)}=1\,.$$

We also have the bound

$$\sup_{s\in S}\frac{\tilde\pi_\nu(A|s)}{\tilde\pi_{\nu^\varepsilon}(A|s)}\le\exp(2|f|_{\mathcal{A}_0})\,.$$

Putting together the above bounds and limits, we find that limit of (28) as $\varepsilon\to0$ is zero, and hence we obtain (27). We must now show that $\frac{\delta\pi}{\delta\nu}$ is bounded and continuous. Boundedness follows from

$$\begin{aligned}\sup_{s\in S}\int_A\left|\frac{d\frac{\delta\pi}{\delta\nu}(\nu,\theta)(a|s)}{d\mu}\right|\mu(da)&=\sup_{s\in S}\int_A\left|\left(f(\theta,s,a)-\int_A f(\theta,s,a')\pi_\nu(da'|s)\right)\frac{d\pi_\nu(a|s)}{d\mu}\right|\mu(da)\\&\le2|f|_{\mathcal{A}_0}\,.\end{aligned}$$

Continuity of $\frac{\delta\pi}{\delta\nu}$ in the product topology then follows from the dominated convergence theorem along with the assumption $|f|_{\mathcal{A}_0}<\infty$. □

We will now use Lemma 2.2 to obtain Lipschitz continuity of the occupancy measure. First, we will show Lipschitz continuity of the occupancy measure with respect to stochastic policies.

Lemma A.5. *For given $\pi,\pi'\in\mathcal{P}(A|S)$, we have*

$$|d^{\pi'}-d^\pi|_{b\mathcal{K}(S|S)}\le\frac{\gamma}{1-\gamma}|\pi'-\pi|_{b\mathcal{K}(A|S)}\,.$$

Proof of Lemma A.5. It follows that

$$d^{\pi'} - d^{\pi} = (1-\gamma)\sum_{n=0}^{\infty}\sum_{i=0}^{n-1}\gamma^n (P^{\pi'})^{n-i-1}(P^{\pi'} - P^{\pi})(P^{\pi})^i\,,$$

where we understood both sides as elements of the Banach algebra $b\mathcal{K}(S|S)$, or equivalently as operators $\mathcal{L}(\mathcal{M}(S),\mathcal{M}(S))$. Thus,

$$\begin{aligned}
|d^{\pi'} - d^{\pi}|_{b\mathcal{K}(S|S)} &\le (1-\gamma)\sum_{n=0}^{\infty}\sum_{i=0}^{n-1}\gamma^n |P^{\pi'}|^{n-i-1}_{b\mathcal{K}(S|S)}|P^{\pi'} - P^{\pi}|_{b\mathcal{K}(S|S)}|P^{\pi}|^{i}_{b\mathcal{K}(S|S)}\\
&\le (1-\gamma)\sum_{n=0}^{\infty}\gamma^n |P|_{b\mathcal{K}(S|A\times S)}|\pi' - \pi|_{b\mathcal{K}(A|S)}\\
&= \frac{\gamma}{1-\gamma}|\pi' - \pi|_{b\mathcal{K}(A|S)}\,,
\end{aligned}$$

which completes the proof. □

Corollary A.6 (Lipschitz continuity of the occupancy measure in the parameter measure)**.** *For given $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}^d)$, we have*

$$|d^{\pi_{\nu'}} - d^{\pi_{\nu}}|_{b\mathcal{K}(S|S)} \le 2|f|_{\mathcal{A}_1}\frac{\gamma}{1-\gamma}W_1(\nu',\nu)\,.$$

Proof of Corollary A.6. By Lemma A.5 and (21), it is enough to show that

$$|\pi_{\nu'} - \pi_{\nu}|_{b\mathcal{K}(A|S)} \le 2|f|_{\mathcal{A}_1}W_1(\nu',\nu)\,. \tag{31}$$

We have

$$\begin{aligned}
|\pi_{\nu'} - \pi_{\nu}|_{b\mathcal{K}(A|S)} &= \sup_{s\in S}|\pi_{\nu'}(\cdot|s) - \pi_{\nu}(\cdot|s)|_{\mathcal{M}(A)}\\
&= \sup_{s\in S}\sup_{\substack{h\in B_b(A)\\ |h|_{B_b(A)}\le 1}}\int_A h(a)\,(\pi_{\nu'} - \pi_{\nu})\,(da|s)\,.
\end{aligned}$$

Let $\nu^{\varepsilon} = \varepsilon\nu' + (1-\varepsilon)\nu$, $\varepsilon\in[0,1]$. Let $s\in S$ and $h\in B_b(A)$ be arbitrarily given. Using Lemma 2.2, we find that

$$\int_A h(a)\,(\pi_{\nu'} - \pi_{\nu})\,(da|s) = \int_{\mathbb{R}^d} g_s(\theta)(\nu' - \nu)(d\theta)\,,$$

where

$$g_s(\theta) := \int_A h(a)\int_0^1\left(f(\theta,s,a) - \int_A f(\theta,s,a')\pi_{\nu^{\varepsilon}}(da'|s)\right)d\varepsilon\,\pi_{\nu^{\varepsilon}}(da|s)\,.$$

Applying (20), we get that for all $\theta,\theta'\in\mathbb{R}^d$ and $s\in S$,

$$\begin{aligned}
&|g_s(\theta') - g_s(\theta)|\\
&= \left|\int_A h(a)\int_0^1\left(f(\theta',s,a) - f(\theta,s,a) + \int_A\left(f(\theta,s,a') - f(\theta',s,a')\right)\pi_{\nu^{\varepsilon}}(da'|s)\right)d\varepsilon\,\pi_{\nu^{\varepsilon}}(da|s)\right|\\
&\le |h|_{B_b(A)}|\pi_{\nu^{\varepsilon}}|_{b\mathcal{K}(A|S)}\left(|f|_{\mathcal{A}_1} + |f|_{\mathcal{A}_1}|\pi_{\nu^{\varepsilon}}|_{b\mathcal{K}(A|S)}\right)|\theta' - \theta|\\
&\le 2|h|_{B_b(A)}|f|_{\mathcal{A}_1}|\theta' - \theta|\,.
\end{aligned}$$

Using (22), we obtain (31), which completes the proof. □

A.2.4. PROOF OF LEMMA 2.3

Proof. Let $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}^d)$ and define $\nu^\varepsilon = \nu + \varepsilon(\nu' - \nu)$ for $\varepsilon \in [0,1]$. We must show that

$$\lim_{\substack{\varepsilon \to 0 \\ \varepsilon \in [0,1]}} \frac{J^{\tau,0}(\nu^\varepsilon) - J^{\tau,0}(\nu)}{\varepsilon} = \int_{\mathbb{R}^d} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta)(\nu' - \nu)(d\theta)\,, \tag{32}$$

where $\frac{\delta J}{\delta \nu}(\nu, \theta)$ is as specified in the statement of the lemma. Noting that

$$\frac{J^{\tau,0}(\nu^\varepsilon) - J^{\tau,0}(\nu)}{\varepsilon} = \int_S \frac{V_\tau^{\pi_{\nu^\varepsilon}}(s) - V_\tau^{\pi_\nu}(s)}{\varepsilon} \rho(ds)\,, \tag{33}$$

we first study the difference quotient $F^\varepsilon(s) = \frac{V_\tau^{\pi_{\nu^\varepsilon}}(s) - V_\tau^{\pi_\nu}(s)}{\varepsilon}$. By (7), we have

$$\begin{aligned} F^\varepsilon(s) &= \frac{1}{\varepsilon}\left[\int_A \left(r(s,a) + \gamma \int_S P(ds'|s,a) V_\tau^{\pi_{\nu^\varepsilon}}(s') - \tau \ln \frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s)\right) \pi_{\nu^\varepsilon}(da|s)\right. \\ &\qquad \left. - \int_A \left(r(s,a) + \gamma \int_S P(ds'|s,a) V_\tau^{\pi_\nu}(s') - \tau \ln \frac{d\pi_\nu}{d\mu}(a|s)\right) \pi_\nu(da|s)\right] \\ &= I_1^\varepsilon(s) + I_2^\varepsilon(s) + I_3^\varepsilon(s) + I_4^\varepsilon(s) + I_5^\varepsilon(s)\,, \end{aligned} \tag{34}$$

where

$$\begin{aligned} I_1^\varepsilon(s) &:= \int_A r(s,a) \frac{\pi_{\nu^\varepsilon}(da|s) - \pi_\nu(da|s)}{\varepsilon}\,, \\ I_2^\varepsilon(s) &:= \gamma \int_A \int_S \frac{V_\tau^{\pi_{\nu^\varepsilon}}(s') - V_\tau^{\pi_\nu}(s')}{\varepsilon} P(ds'|s,a)\pi_\nu(da|s) = \gamma \int_S F^\varepsilon(s') P_{\pi_\nu}(ds'|s)\,, \\ I_3^\varepsilon(s) &:= \gamma \int_A \int_S V_\tau^{\pi_{\nu^\varepsilon}}(s') P(ds'|s,a) \frac{\pi_{\nu^\varepsilon}(da|s) - \pi_\nu(da|s)}{\varepsilon}\,, \\ I_4^\varepsilon(s) &:= -\tau \int_A \frac{\ln \frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \ln \frac{d\pi_\nu}{d\mu}(a|s)}{\varepsilon} \pi_{\nu^\varepsilon}(da|s)\,, \\ I_5^\varepsilon(s) &:= -\tau \int_A \ln \frac{d\pi_\nu}{d\mu}(a|s) \frac{\pi_{\nu^\varepsilon}(da|s) - \pi_\nu(da|s)}{\varepsilon}\,. \end{aligned}$$

Iterating (34) (i.e., applying the usual policy gradient proof technique), we obtain

$$\begin{aligned} F^\varepsilon(s) &= I_1^\varepsilon(s) + I_3^\varepsilon(s) + I_4^\varepsilon(s) + I_5^\varepsilon(s) + \gamma \int_S F^\varepsilon(s') P_{\pi_\nu}(ds'|s) \\ &= \frac{1}{1-\gamma} \int_S (I_1^\varepsilon(s') + I_3^\varepsilon(s') + I_4^\varepsilon(s') + I_5^\varepsilon(s')) d^{\pi_\nu}(ds'|s)\,, \end{aligned}$$

and hence

$$\frac{J^{\tau,0}(\nu^\varepsilon) - J^{\tau,0}(\nu)}{\varepsilon} = \frac{1}{1-\gamma} \int_S (I_1^\varepsilon(s) + I_3^\varepsilon(s) + I_4^\varepsilon(s) + I_5^\varepsilon(s)) d_\rho^{\pi_\nu}(ds)\,.$$

We now will pass to the limit as $\varepsilon \to 0$. Let us begin with the I_4^ε-term. Recalling (30), we have

$$\lim_{\substack{\varepsilon \to 0 \\ \varepsilon \in [0,1]}} |\pi_{\nu^\varepsilon} - \pi_\nu|_{b\mathcal{K}(A|S)} = \lim_{\substack{\varepsilon \to 0 \\ \varepsilon \in [0,1]}} \sup_{s \in S} \int_A \left| \frac{d\pi_{\nu^\varepsilon}}{d\mu}(\cdot|s) - \frac{d\pi_\nu}{d\mu}(\cdot|s) \right| \mu(da) = 0\,. \tag{35}$$

Since

$$\frac{d\pi_\nu}{d\mu}(a|s) = \tilde{\pi}_\nu(A)^{-1} \exp\left(\int_A f(\theta, s, a)\nu(d\theta)\right) \geq \frac{\exp(-2|f|_{\mathcal{A}_0})}{\mu(A)}\,,$$

there is an $\varepsilon_0 \in (0,1]$ such that for all $\varepsilon < \varepsilon_0$, $s \in S$, and $\mu - a.e.\, a \in A$,

$$\left| \frac{\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\pi_\nu}{d\mu}(a|s)}{\frac{d\pi_\nu}{d\mu}(a|s)} \right| < \frac{1}{2}\,.$$

Taylor expanding the logarithm, we get

$$\ln \frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \ln \frac{d\pi_\nu}{d\mu}(a|s) = \ln\left(1 + \frac{\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\pi_\nu}{d\mu}(a|s)}{\frac{d\pi_\nu}{d\mu}(a|s)}\right)$$
$$= \frac{1}{\frac{d\pi_\nu}{d\mu}(a|s)}\left(\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\pi_\nu}{d\mu}(a|s)\right) + \sum_{n=2}^{\infty}(-1)^{n+1}\frac{\left(\frac{1}{\frac{d\pi_\nu}{d\mu}(a|s)}\left(\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\pi_\nu}{d\mu}(a|s)\right)\right)^n}{n}.$$

Using (29), we find that

$$\frac{\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\pi_\nu}{d\mu}(a|s)}{\frac{d\pi_\nu}{d\mu}(a|s)} = \varepsilon\int_{\mathbb{R}^d} f(\theta,s,a)(\nu'-\nu)(d\theta) + \varepsilon^2\sum_{n=2}^{\infty}\varepsilon^{n-2}\frac{\left(\int_{\mathbb{R}^d} f(\theta,s,a)(\nu'-\nu)(d\theta)\right)^n}{n!},$$

which implies

$$\left|\frac{\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \frac{d\pi_\nu}{d\mu}(a|s)}{\frac{d\pi_\nu}{d\mu}(a|s)}\right| \le \varepsilon|f|_{\mathcal{A}_0}|\nu'-\nu|_{\mathcal{M}(\mathbb{R}^d)} + \varepsilon^2\exp\left(|f|_{\mathcal{A}_0}|\nu'-\nu|_{\mathcal{M}(\mathbb{R}^d)}\right),$$

Thus, by Lemma 2.2, we have

$$\lim_{\substack{\varepsilon\to 0\\ \varepsilon\in[0,1]}} \frac{\ln\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \ln\frac{d\pi_\nu}{d\mu}(a|s)}{\varepsilon} = \int_{\mathbb{R}^d}\left(f(\theta,s,a) - \int_A f(\theta,s,a')\pi_\nu(da'|s)\right)(\nu'-\nu)(d\theta), \tag{36}$$

and that there exists a constant $M>0$ such that for all $\varepsilon<\varepsilon_0$, $s\in S$, and $\mu-a.e.\ a\in A$,

$$\varepsilon^{-1}\left|\ln\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s) - \ln\frac{d\pi_\nu}{d\mu}(a|s)\right| \le M.$$

Therefore, owing to (35) and (36), we find

$$\lim_{\substack{\varepsilon\to 0\\ \varepsilon\in[0,1]}} I_4^\varepsilon(s) = -\tau\int_A\int_{\mathbb{R}^d}\left(f(\theta,s,a) - \int_A f(\theta,s,a')\pi_\nu(da'|s)\right)\pi_\nu(da|s)(\nu'-\nu)(d\theta) = 0.$$

We now turn our attention to I_3^ε. Recalling (8), we have

$$V_\tau^{\pi_{\nu^\varepsilon}}(s) = \frac{1}{1-\gamma}\int_S\int_A\left(r(s',a) - \tau\ln\frac{d\pi_{\nu^\varepsilon}}{d\mu}(a|s')\right)\pi_{\nu^\varepsilon}(da|s')d^{\pi_{\nu^\varepsilon}}(ds'|s).$$

It follows from Corollary A.6, (35), Lemma A.4, and the boundedness of the reward r that

$$\lim_{\substack{\varepsilon\to 0\\ \varepsilon\in[0,1]}} V_\tau^{\pi_{\nu^\varepsilon}}(s) = V_\tau^{\pi_\nu}(s).$$

Thus, by Lemmas A.4 and 2.2, we obtain

$$\lim_{\substack{\varepsilon\to 0\\ \varepsilon\in[0,1]}} I_3^\varepsilon(s) = \gamma\int_A\int_S V_\tau^{\pi_\nu}(s')P(ds'|s,a)\int_{\mathbb{R}^d}\frac{\delta\pi_\nu}{\delta\nu}(\nu,\theta)(da|s)(\nu'-\nu)(d\theta).$$

Using Lemmas A.4 and 2.2 and the boundedness of the reward, we get

$$\lim_{\substack{\varepsilon\to 0\\ \varepsilon\in[0,1]}} (I_1^\varepsilon(s) + I_5^\varepsilon(s)) = \int_A\left(r(s',a) - \tau\ln\frac{d\pi_\nu}{d\mu}(a|s)\right)\int_{\mathbb{R}^d}\frac{\delta\pi_\nu}{\delta\nu}(\nu,\theta)(da|s)(\nu'-\nu)(d\theta).$$

Putting it all together and using the definition of Q^π, we arrive at

$$\lim_{\substack{\varepsilon\to 0\\ \varepsilon\in[0,1]}}\sum_{j=1}^5 I_j^\varepsilon(s) = \int_A\left(Q_\tau^{\pi_\nu}(s,a) - \tau\ln\frac{d\pi_\nu}{d\mu}(a|s)\right)\int_{\mathbb{R}^d}\frac{\delta\pi_\nu}{\delta\nu}(\nu,\theta)(da|s)(\nu'-\nu)(d\theta).$$

Since $(I_j)_{1\le j\le 5}$ are bounded uniformly in ε, we may apply the bounded convergence theorem to pass to the limit in (33) to obtain (32). Continuity of $\frac{\delta J^{\tau,0}}{\delta\nu}$ in the product topology then follows from the dominated convergence theorem and the continuity of $\pi=\pi(\nu)$, d^{π_ν} (see Corollary A.6), and Q^{π_ν} in ν (see (8)), and the joint continuity of $\frac{\delta\pi}{\delta\nu}$, which completes the proof. □

A.2.5. PROOF OF THEOREM 2.4

Proof. For given $(s,a) \in S \times A$ and $\nu \in \mathcal{P}_1(\mathbb{R}^d)$, we denote

$$\bar{Q}_\tau^{\pi_\nu}(s,a) = \frac{1}{1-\gamma}\left(Q_\tau^{\pi_\nu}(s,a) - \tau \ln \frac{d\pi_\nu}{d\mu}(a|s)\right).$$

By Lemma A.4, for all $\tau \geq 0$, $\nu \in \mathcal{P}(\mathbb{R}^d)$, $s \in S$, and $\mu - a.e.\, a \in A$, we have

$$|\bar{Q}_\tau^{\pi_\nu}(s,a)| \leq \frac{1}{(1-\gamma)^2}\left(|r|_{B_b(S\times A)} + \tau\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right). \tag{37}$$

Thus, for all $\theta \in \mathbb{R}^d$, $\nu \in \mathcal{P}_1(\mathbb{R}^d)$ and $k \in \mathbb{N}$,

$$\begin{aligned}\left|\nabla^k \frac{\delta J^{\tau,0}}{\delta \nu}(\nu,\theta)\right| &\leq \int_S \int_A |\bar{Q}_\tau^{\pi_\nu}(s,a)| \left|\left(\nabla^k f(\theta,s,a) - \int_A \nabla^k f(\theta,s,a')\pi_\nu(da'|s)\right)\right| \pi_\nu(da|s) d_\rho^{\pi_\nu}(ds) \\ &\leq \frac{2}{(1-\gamma)^2}\left(|r|_{B_b(S\times A)} + \tau\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right)|f|_{\mathcal{A}_k}, \end{aligned} \tag{38}$$

which yields the first inequality in (12). For arbitrarily given $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}^d)$, define $\nu^\varepsilon = \nu + \varepsilon(\nu' - \nu)$, $\varepsilon \in [0,1]$. By Lemma 2.3 and (23) we have

$$J^{\tau,0}(\nu') - J^{\tau,0}(\nu) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu^\varepsilon, \theta)(\nu' - \nu)(d\theta)\, d\varepsilon.$$

Applying (38), we obtain the second inequality in (12)

$$|J^{\tau,0}(\nu') - J^{\tau,0}(\nu)| \leq \sup_{\varepsilon\in[0,1], \theta \in \mathbb{R}^d} \left|\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu^\varepsilon, \theta)\right| W_1(\nu', \nu).$$

For arbitrarily given $\theta \in \mathbb{R}^d$ and $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}^d)$, we have

$$\begin{aligned}&\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu',\theta) - \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu,\theta) = \int_S \int_A \bar{Q}_\tau^{\pi_{\nu'}}(s,a) \nabla \frac{\delta \pi}{\delta \nu}(\nu',\theta)(da|s)[d_\rho^{\pi_{\nu'}} - d_\rho^{\pi_\nu}](ds) (:= I_1) \\ &+ \int_S \int_A \left[\bar{Q}_\tau^{\pi_{\nu'}}(s,a) - \bar{Q}_\tau^{\pi_\nu}(s,a)\right] \nabla \frac{\delta \pi}{\delta \nu}(\nu',\theta')(da|s) d_\rho^{\pi_\nu}(ds) (:= I_2) \\ &+ \int_S \int_A \bar{Q}_\tau^{\pi_\nu}(s,a) \left[\nabla \frac{\delta \pi}{\delta \nu}(\nu',\theta)(da|s) - \nabla \frac{\delta \pi}{\delta \nu}(\nu,\theta)(da|s)\right] d_\rho^{\pi_\nu}(ds) \\ &= I_1 + I_2 + \int_S \int_A \bar{Q}_\tau^{\pi_\nu}(s,a) \int_A \nabla f(\theta,s,a')[\pi_{\nu'} - \pi_\nu](da'|s)\pi_{\nu'}(da|s) d_\rho^{\pi_\nu}(ds) (:= I_3) \\ &+ \int_S \int_A \bar{Q}_\tau^{\pi_\nu}(s,a) \left(\nabla f(\theta,s,a) - \int_A \nabla f(\theta,s,a')\pi_{\nu'}(da'|s)\right)[\pi_{\nu'} - \pi_\nu](da|s) d_\rho^{\pi_\nu}(ds) (:= I_4). \end{aligned}$$

We will now estimate $(I_j)_{1\leq j\leq 4}$. Applying Corollary A.6 and (37), we find

$$|I_1| \leq \frac{4\gamma}{(1-\gamma)^3}\left(|r|_{B_b(S\times A)} + \tau\left(2|f|_{\mathcal{A}_1} + |\ln \mu(A)|\right)\right)|f|_{\mathcal{A}_1}^2 W_1(\nu', \nu).$$

By virtue of (31) and the duality description of the $b\mathcal{K}(A|S)$-norm, for arbitrarily given measurable $h : \mathbb{R}^d \times S \times A \to \mathbb{R}$, we have

$$\int_A h(\theta,s,a)[\pi_{\nu'} - \pi_\nu](da|s) \leq 2|f|_{\mathcal{A}_1}|h|_{L^\infty} W_1(\nu', \nu).$$

Thus, applying (31) and (37), we deduce

$$|I_3 + I_4| \leq \frac{4}{(1-\gamma)^2}\left(|r|_{B_b(S\times A)} + \tau\left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right)|f|_{\mathcal{A}_1}^2 W_1(\nu', \nu).$$

It remains to estimate I_2, and, in particular, to study the Lipschitzness of $\bar{Q}_\tau^{\pi_\nu}$ in ν, and hence the Lipschitzness of $Q_\tau^{\pi_\nu}$ and $\ln \frac{d\pi_\nu}{d\mu}$ separately. By (6) and the fact that

$$V_\tau^{\pi_\nu}(s) = J^{\tau,0}(\delta_s)(\nu), \quad \forall s \in S\,,$$

using Lemma 2.3, (23), and (38) (with $\rho = \delta_s$), we find

$$\begin{aligned}
Q_\tau^{\pi_{\nu'}}(s,a) - Q_\tau^{\pi_\nu}(s,a) &= \gamma \int_S (V_\tau^{\pi_{\nu'}}(s') - V_\tau^{\pi_\nu}(s'))P(ds'|s,a) \\
&= \gamma \int_S \int_0^1 \int_{\mathbb{R}^d} \frac{\delta J^{\tau,0}(\delta_s)}{\delta \nu}(\nu^\varepsilon, \theta)(\nu' - \nu)(d\theta)P(ds'|s,a)\, d\varepsilon \\
&\leq \gamma \sup_{\nu \in \mathcal{P}_1(\mathbb{R}^d)} \sup_{\theta \in \mathbb{R}^d} \left| \nabla \frac{\delta J^{\tau,0}(\delta_s)}{\delta \nu}(\nu, \theta) \right| W_1(\nu', \nu) |P|_{bK(S|A\times S)} \\
&\leq \frac{2\gamma}{(1-\gamma)^2} \left(|r|_{B_b(S\times A)} + \tau \left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right) |f|_{\mathcal{A}_1} W_1(\nu', \nu)\,,
\end{aligned}$$

where in the first equality $\nu^\varepsilon = \nu + \varepsilon(\nu' - \nu)$ for $\varepsilon \in [0,1]$. Owing to (36) and the fundamental theorem of calculus (noting that π_ν is continuous in W_1), we find that for all $(s,a) \in S \times A$,

$$\begin{aligned}
\ln \frac{d\pi_{\nu'}}{d\mu}(a|s) - \ln \frac{d\pi_\nu}{d\mu}(a|s) &= \int_0^1 \int_{\mathbb{R}^d} \left(f(\theta, s, a) - \int_A f(\theta, s, a')\pi_{\nu^\varepsilon}(da'|s) \right) (\nu' - \nu)(d\theta)\, d\varepsilon \\
&\leq 2|f|_{\mathcal{A}_1} W_1(\nu', \nu)\,.
\end{aligned}$$

Thus,

$$|I_2| \leq \frac{4}{1-\gamma} \left(\frac{\gamma}{(1-\gamma)^2} \left(|r|_{B_b(S\times A)} + \tau \left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\right) + \tau \right) |f|_{\mathcal{A}_1}^2 W_1(\nu', \nu)\,.$$

Putting it all together, we find that there is a constant $L = L(\gamma, |r|_{B_b(S\times A)}, \tau, \mu(A), |f|_{\mathcal{A}_1})$ such that

$$\left| \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu', \theta) - \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) \right| \leq L W_1(\nu', \nu)\,.$$

By Lemma A.4, for all $\tau, \tau' \geq 0$, $\nu \in \mathcal{P}(\mathbb{R}^d)$, $s \in S$, and $\mu - a.e.\ a \in A$, we have

$$|\bar{Q}_{\tau'}^{\pi_\nu}(s,a) - \bar{Q}_\tau^{\pi_\nu}(s,a)| \leq \frac{|\tau' - \tau|}{(1-\gamma)^2} \left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right)\,,$$

and hence for all $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
\nabla \frac{\delta J^{\tau',0}}{\delta \nu}(\nu, \theta) - \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) &= \int_S \int_A \left(\bar{Q}_{\tau'}^{\pi_\nu}(s,a) - \bar{Q}_\tau^{\pi_\nu}(s,a) \right) \nabla \frac{\delta \pi}{\delta \nu}(\nu, \theta)(da|s) d_\rho^{\pi_\nu}(ds) \\
&\leq \frac{2|\tau' - \tau|}{(1-\gamma)^2} \left(2|f|_{\mathcal{A}_0} + |\ln \mu(A)|\right) |f|_{\mathcal{A}_1}\,,
\end{aligned}$$

which completes the proof. □

A.2.6. Proof of Theorem 2.6 and Corollary 2.7

Proof of Theorem 2.6. Let $\nu \in \mathcal{P}_1(\mathbb{R}^d)$ be a local maximizer of $J^{\tau,\sigma}$. Thus, there exists $\delta > 0$ such that $J^{\tau,\sigma}(\nu) \geq J^{\tau,\sigma}(\nu')$ for all $\nu' \in \mathcal{P}_1(\mathbb{R}^d)$ satisfying $W_1(\nu, \nu') < \delta$. Let $\nu' \in \mathcal{P}_1(\mathbb{R}^d)$ satisfy $W_1(\nu', \nu) < \delta$ and define $\nu^\varepsilon = \nu + \varepsilon(\nu' - \nu)$ for $\varepsilon \in [0,1]$. Note that for all $\varepsilon \in [0,1]$, we have $W_1(\nu, \nu^\varepsilon) < \delta$. Throughout the proof, we assume that $\nu \in \mathcal{P}_1^{\text{fe}}(\mathbb{R}^d)$ if $\sigma > 0$ since otherwise it would not be a local maximizer. Henceforth, we also restrict to $\nu' \in \mathcal{P}_1^{\text{fe}}(\mathbb{R}^d)$ if $\sigma > 0$. For brevity of notation, we drop the θ dependence in all integrands throughout the proof.

Using Lemma 2.2, we find

$$0 \leq \frac{J^{\tau,\sigma}(\nu) - J^{\tau,\sigma}(\nu^\varepsilon)}{\varepsilon} = -\frac{1}{\varepsilon} \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu^{\varepsilon'})(\nu' - \nu)(d\theta)\, d\varepsilon' + \frac{\sigma^2}{2\varepsilon} \left(\text{KL}(\nu^\varepsilon | e^{-U}) - \text{KL}(\nu | e^{-U}) \right)\,,$$

where we note that if $\sigma > 0$ then $\mathrm{KL}(\nu^\varepsilon|e^{-U}) < \infty$ since ν^ε is a convex combination of ν and ν', both of which belong to $\mathcal{P}_1^{\mathrm{fe}}(\mathbb{R}^d)$. It follows that

$$\begin{aligned}\mathrm{KL}(\nu^\varepsilon|e^{-U}) - \mathrm{KL}(\nu|e^{-U}) &= \int_{\mathbb{R}^d} \left(\nu^\varepsilon \ln \frac{\nu^\varepsilon}{e^{-U}} - \nu \ln \frac{\nu}{e^{-U}}\right) d\theta \\ &= \int_{\mathbb{R}^d} \left(\nu^\varepsilon \ln \nu^\varepsilon - \nu \ln \nu + U(\nu^\varepsilon - \nu)\right) d\theta \\ &= \int_{\mathbb{R}^d} \left(g(\nu^\varepsilon) - g(\nu) + \varepsilon U(\nu' - \nu)\right) d\theta\,,\end{aligned}$$

where $g(x) := x \ln x$. Since g is convex, we have that for all $\theta \in \mathbb{R}^d$,

$$\frac{1}{\varepsilon}\left(g(\nu^\varepsilon) - g(\nu) + \varepsilon U(\nu' - \nu)\right) \le g(\nu') - g(\nu) + U(\nu' - \nu) = \nu' \ln \frac{\nu'}{e^{-U}} - \nu \ln \frac{\nu}{e^{-U}}\,.$$

Since $\mathrm{KL}(\nu'|e^{-U})$ and $\mathrm{KL}(\nu|e^{-U})$ are both finite, the right hand side of the above inequality is integrable, and hence applying reverse Fatou's lemma, we get

$$\limsup_{\varepsilon \to 0} \frac{\mathrm{KL}(\nu^\varepsilon|e^{-U}) - \mathrm{KL}(\nu|e^{-U})}{\varepsilon} \le \int_{\mathbb{R}^d} \left(g'(\nu) + U\right)(\nu' - \nu)(d\theta)\,.$$

Thus, using the boundedness of $\frac{\delta J^{\tau,0}}{\delta \nu}$ (i.e., (12)) and Fatou's Lemma, we obtain

$$0 \le \int_{\mathbb{R}^d} \left(-\frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) + \frac{\sigma^2}{2} \ln \nu + \frac{\sigma^2}{2} U\right)(\nu' - \nu)(d\theta) \tag{39}$$

for all $\nu' \in \mathcal{P}_1(\mathbb{R}^d)$ if $\sigma = 0$ and for all $\nu' \in \mathcal{P}_1^{\mathrm{fe}}(\mathbb{R}^d)$ if $\sigma > 0$. Applying Lemma A.2, we find that the function

$$F(\theta) = \frac{\delta J^{\tau.0}}{\delta \nu}(\nu, \theta) - \frac{\sigma^2}{2} U(\theta) - \frac{\sigma^2}{2} \ln \nu(\theta)$$

is constant in θ, ν-a.e..

We would like to show that if $\sigma > 0$ then ν is equivalent to the Lebesgue measure λ and to that end, we follow the argument of the proof of (Hu et al., 2019)[Prop. 3.9]. Suppose, by contradiction, that ν is not equivalent to the Lebesgue measure. Then there is a $\mathcal{K} \in \mathcal{B}(\mathbb{R}^d)$ such that $\nu(\mathcal{K}) = 0$ and $\lambda(\mathcal{K}) > 0$. Note that on $\mathcal{K}$, we have $\ln \nu = -\infty$. Rearranging (39), we get

$$\frac{\sigma^2}{2} \int_{\mathcal{K}} \ln \frac{\nu}{e^{-U}} \nu'(d\theta) \ge \int_{\mathbb{R}^d} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu)(\nu' - \nu)(d\theta) - \frac{\sigma^2}{2} \mathrm{KL}(\nu|e^{-U}) - \frac{\sigma^2}{2} \int_{\mathbb{R}^d \setminus \mathcal{K}} \ln \frac{\nu}{e^{-U}} \nu'(d\theta)\,.$$

Choosing $\nu' = e^{-U} \in \mathcal{P}_1^{\mathrm{fe}}(\mathbb{R}^d)$, we get

$$\int_{\mathcal{K}} \ln \frac{\nu}{e^{-U}} \nu'(d\theta) = -\infty \quad \text{and} \quad \int_{\mathbb{R}^d \setminus \mathcal{K}} \ln \frac{\nu}{e^{-U}} \nu'(d\theta) < \infty\,.$$

Noting that $\mathrm{KL}(\nu|e^{-U}) < \infty$ is finite because $\nu \in \mathcal{P}_1^{\mathrm{fe}}(\mathbb{R}^d)$ and that $\int_{\mathbb{R}^d} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta)(\nu' - \nu) d\theta < \infty$ by (12), we get a contradiction, and hence ν is equivalent to the Lebesgue measure. Therefore, if $\sigma > 0$, we have that the function F is constant in θ, λ-a.e.. In particular, if $\sigma > 0$, then for λ-a.a. $\theta \in \mathbb{R}^d$,

$$\nu(\theta) = \mathcal{Z}^{-1} \exp\left(\frac{2}{\sigma^2} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) - U(\theta)\right), \quad \text{where} \quad \mathcal{Z} = \int_{\mathbb{R}^d} \exp\left(\frac{2}{\sigma^2} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta') - U(\theta')\right) d\theta'\,.$$

□

Proof of Corollary 2.7. Define $b : \mathcal{P}_1(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ by $b(\nu, \theta) = \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) - \frac{\sigma^2}{2} \nabla U(\theta)$. Recall that a measure $\tilde{\nu}$ is a solution of $L_{\tilde{\nu}}^* \tilde{\nu} = 0$ if for all $\phi \in C_c^\infty(\mathbb{R}^d)$, we have

$$\int_{\mathbb{R}^d} L_{\tilde{\nu}} \phi(\theta) \tilde{\nu}(d\theta) = \int_{\mathbb{R}^d} \left(\frac{\sigma^2}{2} \Delta \phi(\theta) + b(\tilde{\nu}, \theta) \cdot \nabla \phi(\theta)\right) \tilde{\nu}(d\theta) = 0\,.$$

Using Theorem 2.6, we have that

$$F(\theta) = \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) - \frac{\sigma^2}{2} U(\theta) - \frac{\sigma^2}{2} \ln \nu(\theta)$$

is constant in θ, ν-a.e.. By Theorem 2.4 and our assumptions on U, we have that $\frac{\delta J^{\tau,0}}{\delta \nu} + \frac{\sigma^2}{2} U$ is differentiable. Moreover, in the case $\sigma > 0$, by (13), we find that $\ln \nu$ is differentiable. Thus, $\nabla F = 0$, ν-a.e., which implies that for all $\phi \in C_c^\infty(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} \left(b(\nu, \theta) - \frac{\sigma^2}{2} \nabla \ln \nu(\theta) \right) \cdot \nabla \phi(\theta) \nu(d\theta) = 0 \,.$$

Applying the divergence theorem, we find

$$\int_{\mathbb{R}^d} L_\nu \phi(\theta) \nu(d\theta) = 0 \,,$$

which completes the proof. □

A.2.7. Proof of Theorem 2.9

Let $\sigma \geq 0$ and let $b : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ such that for all $\nu \in \mathcal{P}(\mathbb{R}^d)$ we have $b(\nu) : \mathbb{R}^d \to \mathbb{R}^d$ locally bounded. We consider the Cauchy problem for $\nu : \mathbb{R}_+ \to \mathcal{P}(\mathbb{R}^d)$ given by

$$\partial_t \nu_t = L^*_{\nu_t} \nu_t \,, \quad t \in (0, \infty) \,, \quad \nu|_{t=0} = \nu_0 \,, \tag{40}$$

where for all $\nu \in \mathcal{P}(\mathbb{R}^d)$, $\phi \in C_c^\infty(\mathbb{R}^d)$, and $\theta \in \mathbb{R}^d$,

$$L_\nu \phi(\theta) := \frac{\sigma^2}{2} \Delta \phi(\theta) + b(\nu, \theta) \cdot \nabla \phi(\theta) \,.$$

Note that if $b(\nu, \theta) = \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) - \frac{\sigma^2}{2} \nabla U(\theta)$, then this equation is the gradient flow (15) started at ν_0.

Definition A.7. A function $\nu : \mathbb{R}_+ \to \mathcal{P}(\mathbb{R}^d)$ is called a measure-valued solution of (40) if for all $\phi \in C_c^\infty(\mathbb{R}^d)$ and $t \in \mathbb{R}_+$,

$$\int_{\mathbb{R}^d} \phi(\theta) \nu_t(d\theta) = \int_{\mathbb{R}^d} \phi(\theta) \nu_0(d\theta) + \int_0^t \int_{\mathbb{R}^d} L_{\nu_s} \phi(\theta) \nu_s(d\theta) \, ds \,.$$

The next theorem recalls results concerning the existence and uniqueness of a solution of (40). Moreover, it documents important properties in the case $\sigma > 0$ that will be required in the proof of Theorem A.10. Let $W^{1,1}(\mathbb{R}^d) = \{u \in \mathcal{D}'(\mathbb{R}^d) : Du \in L^1(\mathbb{R}^d)\}$, where $\mathcal{D}'(\mathbb{R}^d)$ is the space of distributions on $\mathbb{R}^d$. Denote by $C^{2,1}(\mathbb{R}^d \times (0, \infty))$ the space of functions on $\mathbb{R}^d \times (0, \infty)$ that are twice continuously differentiable on $\mathbb{R}^d$ and once continuously differentiable on $(0, \infty)$.

Theorem A.8 (Existence, uniqueness, and properties)**.** *Assume that there exists a constants $C', L' > 0$ such that for all $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}^d)$ and $\theta \in \mathbb{R}^d$,*

$$|b(\nu, \theta)| \leq C'(1 + |\theta|) \quad \text{and} \quad |b(\nu', \theta) - b(\nu, \theta)| \leq L' W_1(\nu', \nu) \,.$$

If $\nu_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p \in \mathbb{N}$, then there exists a measure-valued solution $\nu \in C(\mathbb{R}_+; \mathcal{P}_p(\mathbb{R}^d))$ of (40). Moreover, if $\sigma > 0$, then we have that

i) (Bogachev et al., 2015)[Cor. 6.3.2] $\nu : [0, T] \to \mathcal{P}_p^{\text{ac}}(\mathbb{R}^d)$;

ii) (Bogachev et al., 2015)[Rem. 7.3.12] $\nu_t \to \nu_0$ in $L^1(\mathbb{R}^d)$ as $t \downarrow 0$;

iii) (Bogachev et al., 2015)[Thm. 7.4.1] If $\int_{\mathbb{R}^d} |\ln \nu_0(\theta)| \nu_0(d\theta) < \infty$, then for almost every $t \in (0, \infty)$, $\nu_t \in W^{1,1}(\mathbb{R}^d)$, and for all $t > 0$

$$\int_0^t \int_{\mathbb{R}^d} |\nabla \ln \nu_s(\theta)|^2 \nu_s(\theta) \, d\theta \, ds = \int_0^t \int_{\mathbb{R}^d} \frac{|\nabla \nu_s(\theta)|^2}{\nu_s(\theta)} \, d\theta \, ds < \infty \,.$$

In particular, for all $\phi \in C_b^1(\mathbb{R}^d)$ and $t_1, t_2 \in \mathbb{R}_+$, we have

$$\int_{\mathbb{R}^d} \phi(\theta) \nu_{t_2}(d\theta) = \int_{\mathbb{R}^d} \phi(\theta) \nu_{t_1}(d\theta) + \int_{t_1}^{t_2} \int_{\mathbb{R}^d} \nabla \phi(\theta) \cdot \left(b(\nu_s, \theta) - \frac{\sigma^2}{2} \nabla \ln \nu_s(\theta) \right) \nu_s(d\theta) \, ds \,;$$

iv) *(Bogachev et al., 2015)[Cor. 8.2.2] For each compact interval $[t_1, t_2] \subset (0, \infty)$, there exists a constant $K = K(p, t_1, t_2, C', \sigma)$ such that for all $t \in [t_1, t_2]$ and $\theta \in \mathbb{R}^d$, we have*

$$\exp\left[-K(1+|\theta|^2)\right] \leq \nu_t(\theta) \leq \exp\left[K(1+|\theta|^2)\right] ;$$

v) *(Bogachev et al., 2015)[Cor. 8.2.5] combined with (Bogachev et al., 2015)[Cor. 8.2.2] For every compact interval $[t_1, t_2] \subset (0, \infty)$, we have*

$$\int_{t_1}^{t_2} \int_{\mathbb{R}^d} |\ln \nu_s(\theta)| \nu_s(\theta)\, d\theta\, ds < \infty .$$

In particular, if $\nu_0 \in \mathcal{P}_2^{\text{fe}}(\mathbb{R}^d)$, then we can take $t_1 = 0$;

vi) *(Lasota & Mackey, 2013)[Thm. 11.7.1] Assume further that $b(\nu) \in C^3(\mathbb{R}^d)$ for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\nu_0 \in C_c^0(\mathbb{R}^d)$. Then $\nu \in C^{2,1}(\mathbb{R}^d \times (0, \infty))$ and for all $t > 0$*

$$|\nu_t(\theta)|, |\partial_t \nu_t(\theta)|, |\nabla \nu_t(\theta)|, |\nabla^2 \nu_t(\theta)| \leq K t^{(-p+2)/2} \exp\left(-\frac{1}{2t}\delta|\theta|^2\right) .$$

If, in addition, there exists an $L'' > 0$ such that for all $\nu \in \mathcal{P}_1(\mathbb{R}^d)$ and $\theta, \theta' \in \mathbb{R}^d$,

$$|b(\nu, \theta') - b(\nu, \theta)| \leq L''|\theta' - \theta| , \tag{41}$$

or both $\sigma > 0$ and $p \geq 4$, then the solution ν is the unique solution of (40).

Proof of Theorem A.8. By (Funaki, 1984)[Thm. 2.1], there exists a measure-valued solution of (40) (see, also, (Manita & Shaposhnikov, 2014)[Thm. 1.1] or (Hammersley et al., 2021)[Thm. 2.10]). The continuity in time of the solution in $\mathcal{P}_p(\mathbb{R}^d)$ follows from the weak continuity in time and (Villani, 2009)[Thm. 6.9].

Clearly, ν is a solution of the linear equation given by

$$\partial_t \mu_t = L^*_{\nu_t} \mu_t , \quad t \in (0, \infty) , \quad \mu|_{t=0} = \nu_0 .$$

Thus, if $\sigma > 0$ then, the statements i)-vi) hold by the references given in the statement of the theorem.

If (41) holds, then assumptions DH1-DH4 of (Manita et al., 2015)[Thm. 4.4] hold (i.e., taking the functions W, V, U and G in DH1-DH4 to be $W(\theta) = 1$, $V(\theta) = 1 + |\theta|^k$, $U(\theta) = \sqrt{1+|\theta|^2}$ (not our potential U), and $G(u) = u$), which implies that the solution of (40) is unique. If $\sigma > 0$ and $p \geq 4$, then the assumptions H1-H4 of (Manita et al., 2015)[Thm. 3.1, Ex. 3.2] hold, which implies that the solution of (40) is unique. □

The following lemma indicates sufficient conditions on the activation function f and potential U to satisfy the assumptions of Theorem A.8.

Corollary A.9. *Define $b_\sigma : \mathcal{P}_1(\mathbb{R}^p) \times \mathbb{R}^d \to \mathbb{R}^d$ by $b_\sigma(\nu, \theta) = \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) - \frac{\sigma^2}{2}\nabla U(\theta)$.*

i) If $f \in \mathcal{A}_1$ and Assumption 2.5 holds, then Theorem 2.4 implies that all $\nu, \nu' \in \mathcal{P}_1(\mathbb{R}^d)$ and $\theta \in \mathbb{R}^d$,

$$|b_\sigma(\nu, \theta)| \leq C_1 + \frac{\sigma^2}{2}|\nabla U(\theta)| \leq C_1 + \frac{\sigma^2 C_U}{2}(1 + |\theta|) \tag{42}$$

$$\text{and} \quad |b_\sigma(\nu', \theta) - b_\sigma(\nu, \theta)| \leq L W_1(\nu', \nu) \leq L W_2(\nu', \nu) . \tag{43}$$

ii) If $f \in \mathcal{A}_2$ and Assumption 2.8 holds, then Theorem 2.4 implies that for all $\nu \in \mathcal{P}_1(\mathbb{R}^d)$ and $\theta, \theta' \in \mathbb{R}^d$,

$$|b_\sigma(\nu, \theta') - b_\sigma(\nu, \theta)| \leq \left(C_2 + \frac{L_U \sigma^2}{2}\right)|\theta' - \theta| . \tag{44}$$

iii) If $f \in \mathcal{A}_4$ and $U \in C^4(\mathbb{R}^d)$, then by Theorem 2.4, $b_\sigma(\nu) \in C^3(\mathbb{R}^d)$ for all $\nu \in \mathcal{P}_1(\mathbb{R}^d)$.

Therefore, the relevant conclusions of Theorem A.8 hold.

In the following theorem, we prove that the objective function $J^{\tau,\sigma}(\nu) = J^{\tau,0}(\nu) - \mathrm{KL}(\nu|e^{-U})$ is increasing along the gradient flow $(\nu_t)_{t\in\mathbb{R}_+}$.

Theorem A.10. *Let $f \in \mathcal{A}_1$, Assumption 2.5 hold, and $\nu_0 \in \mathcal{P}_1(\mathbb{R}^d)$ if $\sigma = 0$, and $\nu_0 \in \mathcal{P}_2^{\mathrm{fe}}(\mathbb{R}^d)$ if $\sigma > 0$. Then for all $t \in \mathbb{R}_+$,*

$$J^{\tau,\sigma}(\nu_t) = J^{\tau,\sigma}(\nu_0) + \int_0^t \int_{\mathbb{R}^d} \left| \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s, \theta) - \frac{\sigma^2}{2} \nabla \ln \frac{\nu_s(\theta)}{e^{-U(\theta)}} \right|^2 \nu_s(d\theta)\, ds\,. \tag{45}$$

If we assume further that $f \in \mathcal{A}_4$, $U \in C^4(\mathbb{R}^d)$, and $\nu \in C_c^2(\mathbb{R}^d)$, then (45) can be proved through classical considerations using Corollary A.9 and Theorem A.8 (iiii-vi) and the fact that (40) holds classically for all $t \in \mathbb{R}_+$ (i.e., including $t = 0$), so one can appeal to the fundamental theorem of calculus. We omit the proof of this (see, for example, the proof of (Otto & Villani, 2000)[Lem. 1]).

Proof of Theorem A.10. We divide the proof into two steps. In the first step, we apply Lemma 2.3 to compute $\frac{d}{dt} J^{\tau,0}(\nu_t)$, and in the second step, we use ideas from (Bogachev et al., 2018)[Thm. 1.1] to compute $\frac{d}{dt}\mathrm{KL}(\nu_t|e^{-U})$. For brevity of notation, we drop the θ dependence in all integrands throughout the proof. Define $b : \mathcal{P}_1(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ by $b(\nu, \theta) = \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) - \frac{\sigma^2}{2}\nabla U(\theta)$.

Step I. By Corollary A.9, there exists a measure-valued solution $\nu \in C(\mathbb{R}_+; \mathcal{P}_1(\mathbb{R}^d))$ of (40). Using Lemma 2.3, we find that for all $t \in \mathbb{R}_+$,

$$\frac{d}{dt} J^{\tau,0}(\nu_t) = \lim_{h\to 0} \frac{J^{\tau,0}(\nu_{t+h}) - J^{\tau,0}(\nu_t)}{h} = \lim_{h\to 0} \int_0^1 \left[\int_{\mathbb{R}^d} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t^{\varepsilon,h}) \frac{1}{h}(\nu_{t+h} - \nu_t)(d\theta) \right] d\varepsilon\,,$$

where $\nu_t^{\varepsilon,h} := \nu_t + \varepsilon(\nu_{t+h} - \nu_t)$ and $h \geq -t$. In the case $t = 0$, we understand the derivative as a right-hand derivative. Our aim to calculate the limit on the right-hand-side. Using Lemma A.9 and Theorem A.8(iii) in the case $\sigma > 0$ and a simple density argument in the case $\sigma = 0$, we find that for all $\phi \in C_b^1(\mathbb{R}^d)$ and $t_1, t_2 \in \mathbb{R}_+$,

$$\int_{\mathbb{R}^d} \phi\, \nu_{t_2}(d\theta) = \int_{\mathbb{R}^d} \phi\, \nu_{t_1}(d\theta) + \int_{t_1}^{t_2} \int_{\mathbb{R}^d} \nabla\phi \cdot \left(b(\nu_s) - \frac{\sigma^2}{2} \nabla \ln \nu_s \right) \nu_s(d\theta)\, ds\,.$$

Owing to Theorem 2.4, we have that $\frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t^{\varepsilon,h}) \in C_b^1(\mathbb{R}^d)$ for all t, ε, and h. Thus, for all $t \in \mathbb{R}_+$,

$$\int_{\mathbb{R}^d} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t^{\varepsilon,h}) \frac{1}{h}(\nu_{t+h} - \nu_t)(d\theta) = \frac{1}{h} \int_t^{t+h} \int_{\mathbb{R}^d} \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s^{\varepsilon,h}) \cdot \left(b(\nu_s) - \frac{\sigma^2}{2} \nabla \ln \nu_s \right) \nu_s(d\theta)\, ds\,.$$

Owing to Corollary A.9 for all $t \in \mathbb{R}_+$, $\lim_{h\to 0} \nu_{t+h} = \nu_t$ in W_2. Thus, by Theorem 2.4 and (21), we obtain that for all $(\theta, t) \in \mathbb{R}^d \times \mathbb{R}_+$,

$$\lim_{h\downarrow 0} \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t^{\varepsilon,h}, \theta) = \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t, \theta)\,.$$

Therefore, by virtue of the Lebesgue differentiation theorem and Lebesgue dominated convergence theorem, we get that for all $t \in \mathbb{R}_+$,

$$\frac{d}{dt} J^{\tau,0}(\nu_t) = \int_{\mathbb{R}^d} \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t) \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t) - \frac{\sigma^2}{2} \nabla \ln \frac{\nu_t}{e^{-U}} \right) \nu_t(d\theta)\,.$$

An application of the fundamental theorem of calculus then implies that for all $t \in \mathbb{R}_+$,

$$J^{\tau,0}(\nu_t) = J^{\tau,0}(\nu_0) + \int_0^t \int_{\mathbb{R}^d} \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2} \nabla \ln \frac{\nu_s}{e^{-U}} \right) \nu_s(d\theta)\, ds\,. \tag{46}$$

The case $\sigma = 0$ of the theorem has now been proved (i.e. part i)).

Step II. From now on, we assume $\sigma > 0$ so that $\nu \in C(\mathbb{R}_+; \mathcal{P}^2_{\text{ac}}(\mathbb{R}^d))$. We aim to show that

$$\mathrm{KL}(\nu_t|e^{-U}) = \mathrm{KL}(\nu_0|e^{-U}) + \int_0^t \int_{\mathbb{R}^d} \nabla \ln \frac{\nu_s}{e^{-U}} \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2} \nabla \ln \frac{\nu_s}{e^{-U}} \right) \nu_s \, d\theta \, ds \,.$$

Once we have shown this, multiplying it by $\sigma^2/2$ and subtracting from (46) we obtain (45). Letting $h = e^U \nu$, we find that for all $t \in \mathbb{R}_+$,

$$\mathrm{KL}(\nu_t|e^{-U}) = \int_{\mathbb{R}^d} \nu_t \ln \frac{\nu_t}{e^{-U}} d\theta = \int_{\mathbb{R}^d} h_t \ln h_t e^{-U} d\theta \,.$$

In terms of h, we aim to show that

$$\int_{\mathbb{R}^d} h_t \ln h_t e^{-U} d\theta = \int_{\mathbb{R}^d} h_0 \ln h_0 e^{-U} d\theta + \int_0^t \int_{\mathbb{R}^d} \nabla \ln h_s \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2} \nabla \ln h_s \right) e^{-U} h_s \, d\theta \, ds \,. \tag{47}$$

Let us establish some preliminary bounds. By Corollary A.9, Theorem A.8 (iii) and (v) and Lemma A.3, we have

$$\begin{aligned}
\int_0^t \int_{\mathbb{R}^d} |\nabla \ln h_s|^2 e^{-U} h_s \, d\theta \, ds &= \int_0^t \int_{\mathbb{R}^d} |\nabla U + \nabla \ln \nu_s|^2 \nu_s \, d\theta \, ds \\
&\le 2 \int_0^t \int_{\mathbb{R}^d} \left(|\nabla U|^2 + |\nabla \ln \nu_s|^2 \right) \nu_s \, d\theta \, ds \\
&\le \int_0^t \int_{\mathbb{R}^d} \left(2C_U^2 (1 + |\theta|)^2 + |\nabla \ln \nu_s|^2 \right) \nu_s \, d\theta \, ds < \infty \,, \\
\int_0^t \int_{\mathbb{R}^d} h_s |\ln h_s| e^{-U} \, d\theta \, ds &= \int_0^t \nu_s \left(|\ln \nu_s| + |U| \right) \, d\theta \, ds \\
&\le \int_0^t \int_{\mathbb{R}^d} \nu_s \left(|\ln \nu_s| + C_U (1 + |\theta|^2) \right) \, d\theta \, ds < \infty \,.
\end{aligned} \tag{48, 49}$$

Notice that $\rho = e^{-U}$ satisfies for all $(\theta, t) \in \mathbb{R}^d \times \mathbb{R}_+$,

$$\partial_t \rho_t = 0 = \frac{\sigma^2}{2} \Delta \rho_t + \frac{\sigma^2}{2} \nabla \cdot (\nabla U \rho_t) \,.$$

Applying (Bogachev et al., 2018)[Lem. 2.4(i)], we find that for all $g \in C^2(\mathbb{R}^d)$, $\psi \in C_c^\infty(\mathbb{R}^d)$, and $t \ge \tau > 0$, [1]

$$\begin{aligned}
&\int_{\mathbb{R}^d} g(h_t) \psi e^{-U} \, d\theta + \frac{\sigma^2}{2} \int_\tau^t \int_{\mathbb{R}^d} |\nabla h_s|^2 g''(h_s) \psi e^{-U} \, d\theta \, ds \\
&\qquad = \int_{\mathbb{R}^d} g(h_\tau) \psi e^{-U} \, d\theta + \int_\tau^t \int_{\mathbb{R}^d} g(h_s) \left(\frac{\sigma^2}{2} \Delta \psi + \frac{\sigma^2}{2} \nabla U \cdot \nabla \psi \right) e^{-U} \, d\theta \, ds \\
&\qquad + \int_\tau^t \int_{\mathbb{R}^d} \left[g''(h_s) \nabla h_s \cdot \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) \psi + g'(h_s) \nabla \psi \cdot \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) \right] h_s e^{-U} \, d\theta \, ds \,.
\end{aligned}$$

Using the divergence theorem, the identity $h \nabla \ln h = \nabla h$ and Theorem A.8 (iii), we find that for all $t \ge \tau > 0$,

$$\begin{aligned}
\int_{\mathbb{R}^d} g(h_t) \psi e^{-U} \, d\theta &= \int_{\mathbb{R}^d} g(h_\tau) \psi e^{-U} \, d\theta + \int_\tau^t \int_{\mathbb{R}^d} g''(h_s) \nabla h_s \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2} \nabla \ln h_s \right) \psi h_s e^{-U} \, d\theta \, ds \\
&\quad + \int_\tau^t \int_{\mathbb{R}^d} g'(h_s) \nabla \psi \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2} \nabla \ln h_s \right) h_s e^{-U} \, d\theta \, ds \,.
\end{aligned} \tag{50}$$

[1] In (Bogachev et al., 2018), the authors introduce a probabilistic approximation to the identity $(\omega_\varepsilon)_{\varepsilon \in (0,1)}$ and the approximation $h_\varepsilon = \omega_\varepsilon * \nu / \omega_\varepsilon * U$, $\varepsilon \in (0,1)$. Then they write down the equation for h_ε, which, in particular, has smooth coefficients and holds pointwise. At this point, they work classically and easily derive (50). To pass to the limit as $\varepsilon \downarrow 0$, they appeal to the fact that $\omega_\varepsilon * \nu$ to ν uniformly on $\operatorname{supp} \psi \times [\tau, t]$ and $\nabla \omega_\varepsilon * \nu \to \nabla \nu$ in $L^2(\operatorname{supp} \psi \times [\tau, t])$ since ν is locally Hölder and Sobolev.

Consider $g \in C^1(\mathbb{R}_+)$ such that $g' \in C_b^0(\mathbb{R}_+)$, g' is continuously differentiable except at a finite number of points $D \subset \mathbb{R}_+$, and $\sup_{x\in\mathbb{R}_+\setminus D} xg''(x) < \infty$. The functions (52) we consider below satisfy this assumption. Mollifying g, we obtain a sequence $\{g_n\}_{n\in\mathbb{N}} \in C^\infty(\mathbb{R}_+)$ such that $g_n \to g$ and $g_n' \to g'$ pointwise on $\mathbb{R}_+$ as $n \to \infty$, $g_n'' \to g''$ pointwise on $\mathbb{R}_+ \setminus D$ as $n \to \infty$, $\sup_{n\in\mathbb{N},x\in\mathbb{R}_+} |g_n'(x)| < \infty$, and $\sup_{n\in\mathbb{N},x\in\mathbb{R}_+\setminus D} x|g_n''(x)| < \infty$. Let $\psi \in C_c^\infty(\mathbb{R}^d)$ be such that $\psi \geq 0$, $\psi = 1$ for $|\theta| < 1$ and $\psi = 0$ for $|\theta| \geq 2$. For $n \in \mathbb{N}$, define $\psi_n \in C_c^\infty(\mathbb{R}^d)$ by $\psi_n(x) = \psi(x/n)$. By (50) and the fact that the Lebesgue measure of the set $\{h_t \in D\}$ is zero by (Kinderlehrer & Stampacchia, 2000)[Ch. 2, Lem. A.4], we have that for all $n \in \mathbb{N}$ and $t \geq \tau > 0$,

$$\begin{aligned}\int_{\mathbb{R}^d} g_n(h_t)\psi_n e^{-U}\,d\theta = \int_{\mathbb{R}^d} g_n(h_\tau)\psi_n e^{-U}\,d\theta \\ + \int_\tau^t \int_{h_s\notin D} g_n''(h_s)\nabla h_s \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2}\nabla \ln h_s\right)\psi_n h_s e^{-U}\,d\theta\,ds \\ + \int_\tau^t \int_{\mathbb{R}^d} g_n'(h_s)\nabla\psi_n \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2}\nabla \ln h_s\right) h_s e^{-U}\,d\theta\,ds\,.\end{aligned}$$

For all $(\theta, t) \in \mathbb{R}^d \times \mathbb{R}_+$, we have

$$\begin{aligned}|g_n(h)|e^{-U} &\leq \sup_{n\in\mathbb{N}} |g_n(0)|e^{-U} + \sup_{n\in\mathbb{N},x\in\mathbb{R}_+} |g_n'(x)|he^{-U}, \\ \mathbf{1}_{h\notin D} g_n''(h)|\nabla h|\left|\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu) - \frac{\sigma^2}{2}\nabla \ln h\right| he^{-U} &\leq \sup_{n\in\mathbb{N},x\in\mathbb{R}_+\setminus D} |xg_n''(x)|\left(C_1|\nabla h|e^{-U} + \frac{\sigma^2}{2}|\nabla \ln h|^2 he^{-U}\right), \\ |g_n'(h)|\left|\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu) - \frac{\sigma^2}{2}\nabla \ln h\right| he^{-U} &\leq \sup_{n\in\mathbb{N},x\in\mathbb{R}_+} |g_n'(x)|\left(C_1 + \frac{\sigma^2}{2}|\nabla \ln h|he^{-U}\right).\end{aligned}$$

Owing to (48), we have that for all $t \in \mathbb{R}_+$,

$$\int_0^t \int_{\mathbb{R}^d} |\nabla \ln h_s| h_s e^{-U}\,d\theta\,ds \leq \sqrt{t}\left(\int_0^t \int_{\mathbb{R}^d} |\nabla \ln h_s|^2 h_s e^{-U}\,d\theta\,ds\right)^{1/2} < \infty$$

and

$$\int_0^t \int_{\mathbb{R}^d} |\nabla h_s| e^{-U}\,d\theta\,ds = \int_0^t \int_{\mathbb{R}^d} |\nabla \ln h_s| h_s e^{-U}\,d\theta\,ds < \infty\,.$$

Applying the dominated convergence theorem to pass to the limit as $n \to \infty$, we obtain that for all $t \geq \tau > 0$,

$$\int_{\mathbb{R}^d} g(h_t)e^{-U}\,d\theta = \int_{\mathbb{R}^d} g(h_\tau)e^{-U}\,d\theta + \int_\tau^t \int_{h_s\notin D} g''(h_s)\nabla h_s \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2}\nabla \ln h_s\right) h_s e^{-U}\,d\theta\,ds\,.$$

By Taylor's theorem and the fact that $g' \in C_b^0(\mathbb{R}^d)$, for all $t \in \mathbb{R}_+$, we have

$$\begin{aligned}\left|\int_{\mathbb{R}^d} g(h_t)e^{-U}d\theta - \int_{\mathbb{R}^d} g(h_0)e^{-U}d\theta\right| &\leq \int_0^1 \int_{\mathbb{R}^d} |g'(h_0 + \varepsilon(h_t - h_0))||h_t - h_0|e^{-U}\,d\varepsilon\,d\theta \\ &\leq |g'|_{L^\infty} \int_{\mathbb{R}^d} |\nu_t - \nu_0|\,d\theta\,,\end{aligned}$$

which tends to zero as $t \downarrow 0$ by Corollary A.9 and Theorem A.8 (i). Therefore, for all $t \in \mathbb{R}_+$,

$$\int_{\mathbb{R}^d} g(h_t)e^{-U}d\theta = \int_{\mathbb{R}^d} g(h_0)e^{-U}d\theta + \int_0^t \int_{h_s\notin D} g''(h_s)\nabla h_s \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2}\nabla \ln h_s\right) h_s e^{-U}\,d\theta\,ds\,. \tag{51}$$

We would like to take $g : \mathbb{R}_+ \to \mathbb{R}$ given by $g(x) = x\ln x - x$ in (51), but $g' \notin C_b^0(\mathbb{R}_+)$. Nevertheless, if we could, then for all $t \in \mathbb{R}_+$,

$$\begin{aligned}\int_{\mathbb{R}^d} (h_t \ln h_t - h_t)e^{-U}d\theta = \int_{\mathbb{R}^d} (h_0 \ln h_0 - h_0)e^{-U}d\theta \\ + \int_0^t \int_{\mathbb{R}^d} \nabla \ln h_s \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2}\nabla \ln h_s\right) h_s e^{-U}\,d\theta\,ds\,,\end{aligned}$$

which yields (47).

Instead, we will approximate $g(x) = x \ln x - x$ and pass to the limit. Towards this end, as in the proof of (Bogachev et al., 2018)[Thm. 1.1], we consider the sequence $\{g_{k,m}\}_{k,m\in\mathbb{N}} \subset C^1(\mathbb{R}_+)$ defined by

$$g_{k,m}(x) = \begin{cases} -x \ln k & x \le k^{-1}\,, \\ x \ln x - x + k^{-1} & k^{-1} < x \le m\,, \\ x \ln m - m + k^{-1} & x \ge m\,. \end{cases}$$

Clearly,

$$g'_{k,m}(x) = \begin{cases} -\ln k & x \le k^{-1}\,, \\ \ln x & k^{-1} < x \le m\,, \\ \ln m & x \ge m\,, \end{cases} \quad \text{and} \quad g''_{k,m}(x) = \begin{cases} 0 & x \le k^{-1}\,, \\ x^{-1} & k^{-1} < x < m\,, \\ 0 & x \ge m\,, \end{cases}$$

from which we deduce that $g'_{k,m} \in C^0_b(\mathbb{R}_+)$ and $\sup_{x\in\mathbb{R}_+-\{k,m\}} x g''_{k,m}(x) = 1$. Applying (51), we find that for all $t \in \mathbb{R}_+$,

$$\begin{aligned} \int_{\mathbb{R}^d} g_{k,m}(h_t) e^{-U} d\theta = \int_{\mathbb{R}^d} g_{k,m}(h_0) e^{-U} d\theta \\ + \int_0^t \int_{k^{-1}<h_s<m} \nabla \ln h_s \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2} \nabla \ln h_s \right) h_s e^{-U}\, d\theta\, ds\,. \end{aligned} \tag{52}$$

We will now pass to the limit each term of (52) as $k, m \to \infty$. For all $k, m \in \mathbb{N}$ and $t \in \mathbb{R}_+$, we have

$$\begin{aligned} \int_{\mathbb{R}^d} g_{k,m}(h_t) e^{-U} d\theta = -\ln k \int_{h_t \le k^{-1}} h_t e^{-U} d\theta + \int_{k^{-1}<h_t<m} (h_t \ln h_t - h_t)\, e^{-U} d\theta \\ + \int_{m \le h_t} (h_t \ln m - m)\, e^{-U} d\theta + k^{-1} \int_{k^{-1}<h_t} e^{-U} d\theta\,. \end{aligned}$$

Since $h \ge 0$, for all $k, m \in \mathbb{N}$ with $m \ge 3$ (so that $m \ln m \ge m$) and $t \in \mathbb{R}_+$,

$$\int_{\mathbb{R}^d} g_{k,m}(h_t) e^{-U} d\theta \le \int_{k^{-1}<h_t} h_t \ln h_t e^{-U} d\theta - \int_{k^{-1}<h_t<m} h_t e^{-U} d\theta + k^{-1} \int_{k^{-1}<h_t} e^{-U} d\theta$$

and

$$\int_{\mathbb{R}^d} g_{k,m}(h_t) e^{-U} d\theta \ge -\ln k \int_{h_t \le k^{-1}} h_t e^{-U} d\theta + \int_{k^{-1}<h_t<m} (h_t \ln h_t - h_t)\, e^{-U} d\theta\,.$$

The bound

$$\mathbf{1}_{h_t \le k^{-1}} h_t \ln k \le k^{-1} \ln k \le \sup_{x \ge 1} x^{-1} \ln x = e^{-1}\,, \quad \forall (\theta, t) \in \mathbb{R}^d \times \mathbb{R}_+\,,$$

allows us to apply the dominated convergence theorem to obtain that for all $t \in \mathbb{R}_+$,

$$\lim_{k\to\infty} \ln k \int_{h_t \le k^{-1}} h_t e^{-U} d\theta = 0\,.$$

Clearly, for all $t \in \mathbb{R}_+$,

$$\lim_{k\to\infty} k^{-1} \int_{k^{-1}<h_t} e^{-U} d\theta = 0\,.$$

The bound (49) and another application of the dominated convergence theorem imply that for all $t \in \mathbb{R}_+$,

$$\lim_{k,m\to\infty} \int_{k^{-1}<|h_t|<m} (h_t \ln h_t - h_t)\, e^{-U} d\theta = \int_{\mathbb{R}^d} h_t \ln h_t e^{-U} d\theta - 1$$

and

$$\lim_{k\to\infty} \int_{k^{-1}<|h_t|} h_t \ln h_t e^{-U} d\theta - \int_{k^{-1}<|h_t|<m} h_t e^{-U} d\theta = \int_{\mathbb{R}^d} h_t \ln h_t e^{-U} d\theta - 1\,.$$

Therefore, for all $t \in \mathbb{R}_+$,

$$\int_{\mathbb{R}^d} h_t \ln h_t e^{-U} d\theta - 1 \leq \lim_{k,m\to\infty} \int_{\mathbb{R}^d} g_{k,m}(h_t) e^{-U} d\theta \leq \int_{\mathbb{R}^d} h_t \ln h_t e^{-U} d\theta - 1\,,$$

which yields for all $t \in \mathbb{R}_+$,

$$\lim_{k,m\to\infty} \int_{\mathbb{R}^d} g_{k,m}(h_t) e^{-U} d\theta = \int_{\mathbb{R}^d} h_t \ln h_t e^{-U} d\theta - 1\,.$$

Using the bounds

$$\nabla \ln h_t \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_t) - \frac{\sigma^2}{2} \nabla \ln h_t \right) h_t e^{-U} \leq \left(C_1 |\nabla \ln h_t| + \frac{\sigma^2}{2} |\nabla \ln h_t|^2 \right) h_t e^{-U}\,, \quad \forall (\theta, t) \in \mathbb{R}^d \times \mathbb{R}_+\,,$$

and (48), and the dominated convergence theorem, we find that for all $t \in \mathbb{R}_+$,

$$\begin{aligned}
&\lim_{k,m\to\infty} \int_0^t \int_{k^{-1} < h_s < m} \nabla \ln h_s \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2} \nabla \ln h_s \right) h_s e^{-U} \, d\theta \, ds \\
&\quad = \int_0^t \int_{\mathbb{R}^d} \nabla \ln h_s \cdot \left(\nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu_s) - \frac{\sigma^2}{2} \nabla \ln h_s \right) h_s e^{-U} \, d\theta \, ds\,.
\end{aligned}$$

Combining the above, we pass to the limit in every term of (52) to obtain (47). □

A.2.8. Proof of Theorem 2.13 and moment estimates

Lemma A.11 (Moment estimates)**.** *Let $f \in \mathcal{A}_2$ and Assumptions 2.5, 2.8, and 2.11 hold. Let $(\nu_t)_{t\geq 0}$ be the solution of (15) with initial condition $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then for all $\ell > 0$ and $t \in \mathbb{R}_+$,*

$$\int_{\mathbb{R}^d} |\theta|^2 \nu_t(d\theta) \leq e^{-\alpha_{1,\ell} t} \int_{\mathbb{R}^d} |\theta|^2 \nu_0(d\theta) + \frac{1}{\alpha_{1,\ell}} \left(\frac{1}{4\ell} \left(2C_1 + \sigma^2 C_U \right) + d\sigma^2 \right) \left(1 - e^{-\alpha_{1,\ell} t} \right), \tag{53}$$

where $\alpha_{1,\ell} := \sigma^2 \kappa - 2C_2 - \ell$. Moreover, if $\sigma^2 \kappa > 2C_2$ and $\nu^* \in \mathcal{P}_2(\mathbb{R}^d)$ is a solution of (14), then for all $\ell > 0$ such that $\alpha_{1,\ell} > 0$,

$$\int_{\mathbb{R}^d} |\theta|^2 \nu^*(d\theta) \leq \frac{1}{\alpha_{1,\ell}} \left(\frac{1}{4\ell} \left(2C_1 + \sigma^2 C_U \right) + d\sigma^2 \right). \tag{54}$$

If, in addition, $\nabla U(0) = 0$, then for all $\ell > 0$ and $t \in \mathbb{R}_+$,

$$\int_{\mathbb{R}^d} |\theta|^2 \nu_t(d\theta) \leq e^{-\alpha_{2,\ell} t} \int_{\mathbb{R}^d} |\theta|^2 \nu_0(d\theta) + \frac{1}{\alpha_{2,\ell}} \left(\frac{C_1}{2\ell} + d\sigma^2 \right) \left(1 - e^{-\alpha_{2,\ell} t} \right),$$

where $\alpha_{2,\ell} := \sigma^2 \kappa - \ell$. An analog of (54) then holds for ν^* if $\sigma > 0$.

Proof. Define $b : \mathcal{P}_2(\mathbb{R}^p) \times \mathbb{R}^d \to \mathbb{R}^d$ by $b(\nu, \theta) = \nabla \frac{\delta J^{\tau,0}}{\delta \nu}(\nu, \theta) - \frac{\sigma^2}{2} \nabla U(\theta)$. Theorem 2.4 and Assumption 2.11 implies that for all $\theta, \theta' \in \mathbb{R}^d$ and $\nu, \nu' \in \mathcal{P}_2(\mathbb{R}^d)$,

$$2(b(\nu, \theta') - b(\nu, \theta)) \cdot (\theta - \theta') \leq \left(2C_2 - \sigma^2 \kappa \right) |\theta' - \theta|^2\,. \tag{55}$$

Using (42), (55), and Young's inequality, we find that for all $\theta \in \mathbb{R}^d$ and $\nu \in \mathcal{P}_1(\mathbb{R}^d)$ and $\ell > 0$,

$$\begin{aligned}
2b(\nu, \theta) \cdot \theta &\leq 2b(\nu, 0) \cdot \theta + \left(2C_2 - \sigma^2 \kappa \right) |\theta|^2 \leq \left(2C_1 + \sigma^2 C_U \right) |\theta| + \left(2C_2 - \sigma^2 \kappa \right) |\theta|^2 \\
&\leq \frac{1}{4\ell} \left(2C_1 + \sigma^2 C_U \right) + \left(2C_2 + \ell - \sigma^2 \kappa \right) |\theta|^2\,,
\end{aligned} \tag{56}$$

where C_1 is the constant from Theorem 2.4.

There exists a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ supporting a Wiener process $(W_t)_{t\geq 0}$ and a random variable θ_0 such that $\mathrm{Law}(\theta_0) = \nu_0$ and θ_0 is independent of W. Let $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ be the corresponding filtration generated by θ_0 and $(W_t)_{t\geq 0}$.

By Theorem 2.10, there exists a unique strong solution $(\theta_t)_{t\in\mathbb{R}_+}$ of (17) and $\nu_t = \mathrm{Law}(\theta_t)$ for all $t \in \mathbb{R}_+$. Applying Itô's formula and (56), we obtain that for all $\alpha' \in \mathbb{R}$, $\ell > 0$, and $t \in \mathbb{R}_+$,

$$\begin{aligned} d(e^{\alpha' t}|\theta_t|^2) &= e^{\alpha' t}\left(\alpha'|\theta_t|^2 + 2b(\mathrm{Law}(\theta_t),\theta_t)\cdot\theta_t + d\sigma^2\right)dt + e^{\alpha' t}2\sigma\theta_t\cdot dW_t \\ &\le e^{\alpha' t}\left((\alpha' + 2C_2 + \ell - \sigma^2\kappa)|\theta_t|^2 + \frac{1}{4\ell}\left(2C_1 + \sigma^2 C_U\right) + d\sigma^2\right)dt + e^{\alpha' t}2\sigma\theta_t\cdot dW_t\,. \end{aligned}$$

Setting $\alpha' = \alpha_\ell$, integrating the above, and taking the expectation (using a standard stopping time argument), we complete the proof of (53).

To obtain (54), notice that ν^* is a solution of (15), and hence by (53), for all $t \ge 0$ and $\ell > 0$,

$$\int_{\mathbb{R}^d}|\theta|^2\nu^*(d\theta) \le e^{-\alpha_{1,\ell}t}\int_{\mathbb{R}^d}|\theta|^2\nu^*(d\theta) + \frac{1}{\alpha_{1,\ell}}\left(\frac{1}{4\ell}\left(2C_1 + \sigma^2 C_U\right) + d\sigma^2\right)(1 - e^{-\alpha_{1,\ell}t})\,.$$

Under the assumption $\sigma^2\kappa > 2C_2$, choosing $\ell > 0$ such that $\alpha_{1,\ell} > 0$, we may pass to the limit in the right-hand-side as $t \to \infty$ to derive (54).

If $\nabla U(0) = 0$, then for all $\theta \in \mathbb{R}^d$ and $\nu \in \mathcal{P}_1(\mathbb{R}^d)$ and $\ell > 0$,

$$2b(\nu,\theta)\cdot\theta \le 2C_1|\theta| - \sigma^2\kappa|\theta|^2 \le \frac{C_1}{2\ell} + (\ell - \sigma^2\kappa)|\theta|^2\,,$$

from which the statements follow via a similar argument to the one given above. □

Proof of Theorem 2.13. Define $b_{\tau,\sigma} : \mathcal{P}_1(\mathbb{R}^p)\times\mathbb{R}^d \to \mathbb{R}^d$ by

$$b_{\tau,\sigma}(\nu,\theta) = \nabla\frac{\delta J^{\tau,0}}{\delta\nu}(\nu,\theta) - \frac{\sigma^2}{2}\nabla U(\theta)\,.$$

Theorem 2.4 implies that for all $\theta,\theta' \in \mathbb{R}^d$ and $\nu,\nu' \in \mathcal{P}_1(\mathbb{R}^d)$,

$$|b_{\sigma,\tau}(\nu,\theta) - b_{\sigma',\tau}(\nu,\theta)| \le \frac{|\sigma^2 - \sigma'^2|}{2}|\nabla U(\theta)|\,, \tag{57}$$

$$\text{and}\quad |b_{\sigma,\tau}(\nu,\theta) - b_{\sigma,\tau'}(\nu,\theta)| \le D|\tau - \tau'|\,, \tag{58}$$

where D is the constant from Theorem 2.4.

There exists a probability space $(\Omega,\mathcal{F},\mathbb{P})$ supporting a Wiener process $(W_t)_{t\ge0}$ and random variables θ_0, θ_0' such that $\mathrm{Law}(\theta_0) = \nu_0$, $\mathrm{Law}(\theta_0') = \nu_0'$, and (θ_0,θ_0') is independent of $(W_t)_{t\ge0}$. By Theorem 2.10, there exists a unique strong solution $(\theta_t)_{t\in\mathbb{R}_+}$ of (17), and $\nu_t = \mathrm{Law}(\theta_t)$ for all $t \in \mathbb{R}_+$. Similarly, there exists a unique strong solution $(\theta_t')_{t\in\mathbb{R}_+}$ of (17) with σ and τ replaced by σ' and τ', $\nu_t' = \mathrm{Law}(\theta_t')$ for all $t \in \mathbb{R}_+$.

By Lemma A.11, for all $\beta' \in \mathbb{R}$ and $t \in \mathbb{R}_+$,

$$\mathbb{E}\int_0^t e^{2\beta' s}\left(|\theta_s|^2 + |\theta_s'|^2\right)ds < \infty\,.$$

Thus, applying Itô's formula and taking the expectation, we find that for all $\beta' \in \mathbb{R}$ and $t \in \mathbb{R}_+$,

$$\begin{aligned} d\left(e^{2\beta' t}\mathbb{E}\left[|\theta_t - \theta_t'|^2\right]\right) \le{}& e^{2\beta' t}\mathbb{E}\left[2\beta'|\theta_t - \theta_t'|^2 + d|\sigma - \sigma'|^2\right]dt \\ &+ e^{2\beta' t}\mathbb{E}\left[2(\theta_t - \theta_t')\cdot(b_{\sigma,\tau}(\mathrm{Law}(\theta_t),\theta_t) - b_{\sigma,\tau}(\mathrm{Law}(\theta_t),\theta_t'))\right]dt \\ &+ e^{2\beta' t}\mathbb{E}\left[2|\theta_t - \theta_t'|\,|b_{\sigma,\tau}(\mathrm{Law}(\theta_t),\theta_t') - b_{\sigma,\tau}(\mathrm{Law}(\theta_t'),\theta_t')|\right]dt \\ &+ e^{2\beta' t}\mathbb{E}\left[2|\theta_t - \theta_t'|\,|b_{\sigma,\tau}(\mathrm{Law}(\theta_t'),\theta_t') - b_{\sigma',\tau}(\mathrm{Law}(\theta_t'),\theta_t')|\right]dt \\ &+ e^{2\beta' t}\mathbb{E}\left[2|\theta_t - \theta_t'|\,|b_{\sigma',\tau}(\mathrm{Law}(\theta_t'),\theta_t') - b_{\sigma',\tau'}(\mathrm{Law}(\theta_t'),\theta_t')|\right]dt\,. \end{aligned}$$

Making use of the identity

$$W_2(\mathrm{Law}(\theta_t), \mathrm{Law}(\theta'_t)) \le \mathbb{E}|\theta_t - \theta'_t|^2\,, \quad \forall t \in \mathbb{R}_+\,, \tag{59}$$

the bounds (55),(43),(57), (58), and Young's inequality, we get that for all $\beta' \in \mathbb{R}$, $\ell > 0$, and $t \in \mathbb{R}_+$,

$$\begin{aligned}
&d\left(e^{2\beta' t}\mathbb{E}\left[|\theta_t - \theta'_t|^2\right]\right)\\
&\le e^{2\beta' t}\mathbb{E}\left[(2\beta' - \sigma^2\kappa + 2C_2(\tau))|\theta_t - \theta'_t|^2 + 2L(\tau)|\theta_t - \theta'_t|W_2(\mathrm{Law}(\theta_t), \mathrm{Law}(\theta'_t))\right] dt\\
&\quad + e^{2\beta' t}\mathbb{E}\left[|\sigma^2 - \sigma'^2||\theta_t - \theta'_t||\nabla U(\theta'_t)| + D|\tau - \tau'| + d|\sigma - \sigma'|^2\right] dt\\
&\le (2\beta' - \sigma^2\kappa + 2C_2(\tau) + 2L(\tau) + 2\ell|\sigma^2 - \sigma'^2|)e^{2\beta' t}\mathbb{E}\left[|\theta_t - \theta'_t|^2\right] dt\\
&\quad + e^{2\beta' t}\mathbb{E}\left[\frac{|\sigma^2 - \sigma'^2|}{8\ell}|\nabla U(\theta'_t)|^2 + D|\tau - \tau'| + d|\sigma - \sigma'|^2\right] dt\,.
\end{aligned}$$

Setting $\beta' = \beta_\ell$, integrating the above, taking the expectation, and then applying (59), we complete the proof of (18).

To obtain (19), notice that ν^* and ν'^* are solutions of (15) with corresponding parameters σ, τ and σ', τ', respectively. Thus, by (18), for all $\ell > 0$ and $t \ge 0$, we have

$$\begin{aligned}
W_2^2(\nu^*, \nu'^*) \le e^{-2\beta_\ell t}W_2^2(\nu^*, \nu'^*) + \frac{|\sigma^2 - \sigma'^2|}{8\ell}\int_0^t e^{2\beta_\ell(s-t)}\,ds\int_{\mathbb{R}^d}|\nabla U(\theta)|^2\nu'^*(d\theta)\\
+ \frac{1}{2\beta_\ell}\left(D|\tau - \tau'| + d|\sigma - \sigma'|^2\right)\left(1 - e^{-2\beta_\ell t}\right),
\end{aligned}$$

Under the assumption $\beta := \frac{\sigma^2}{2}\kappa - C_2(\tau) - L(\tau) > 0$, choosing $\ell > 0$ such that $\beta_\ell > 0$, we may pass to the limit in the right-hand-side as $t \to \infty$ to derive (19).

□

A.2.9. Proof of Theorem 2.12

Proof of Theorem 2.12. Owing to (42), (55), (43), and $\beta := \frac{\sigma^2}{2}\kappa - C_2 - L > 0$, (Bogachev et al., 2018)[Thm. 4.1] there is a unique solution ν^* of (14). We could also argument employed in (Komorowski & Walczuk, 2012)[Thm. 2.1] combined with Theorem 2.13 to prove this result. Moreover, by Theorem 2.13, for any solution $\nu \in C(\mathbb{R}_+; \mathcal{P}_2(\mathbb{R}^d))$ of (15) and all $t \in \mathbb{R}_+$, we have

$$W_2(\nu_t, \nu^*) \le e^{-\beta t}W_2(\nu_0, \nu^*) \tag{60}$$

because $\tilde{\nu}_t = \nu^*$, $t \in \mathbb{R}_+$, is a solution of (15).

We must show that $J^{\tau,\sigma}(\nu^*) \ge J^{\tau,\sigma}(\nu_0)$ for an arbitrary $\nu_0 \in \mathcal{P}_2^{\mathrm{fe}}(\mathbb{R}^d)$. Let $\nu \in C(\mathbb{R}_+; \mathcal{P}_2(\mathbb{R}^d))$ be the unique solution of (15) such that $\nu|_{t=0} = \nu_0$. We have that $\lim_{t\to\infty}\nu_t = \nu^*$ in W_2 by (60). By the continuity of $J^{\tau,0}$ in W_2 (i.e., Theorem 2.4), the lower semi-continuity of relative entropy in the 2-Wasserstein topology ((Dupuis & Ellis, 1997)[Lem. 1.4.3]), and by Theorem 2.9, we have

$$\begin{aligned}
J^{\tau,\sigma}(\nu^*) - J^{\tau,\sigma}(\nu_0) &\ge \limsup_{t\to\infty} J^{\tau,\sigma}(\nu_t) - J^{\tau,\sigma}(\nu_0)\\
&= \limsup_{t\to\infty}\int_0^t\int_{\mathbb{R}^d}\left|\nabla\frac{\delta J^{\tau,0}}{\delta\nu}(\nu_s, \theta) - \frac{\sigma^2}{2}\nabla\ln\frac{\nu_s(\theta)}{e^{-U(\theta)}}\right|^2\nu_s(d\theta)\,ds \ge 0\,.
\end{aligned}$$

Hence, ν^* is a global maximizer. By Corollary 2.7, we know any local maximizer of $J^{\tau,\sigma}$ is a solution of (14). However, equation (14) only has one solution, so ν^* is the unique maximizer of $J^{\tau,\sigma}$. □