Proceedings of Machine Learning Research vol 107:1–38, 2021 2nd Annual Conference on Mathematical and Scientific Machine Learning

Temporal-difference learning with nonlinear function approximation: lazy training and mean field regimes

Andrea Agazzi AGAZZI@MATH.DUKE.EDU
Department of Mathematics
Duke University
Durham, NC 27708

Jianfeng Lu JIANFENG@MATH.DUKE.EDU
Department of Mathematics
Department of Physics and Department of Chemistry
Duke University
Durham, NC 27708

Editors: Joan Bruna, Jan S Hesthaven, Lenka Zdeborova

Abstract

We discuss the approximation of the value function for infinite-horizon discounted Markov Reward Processes (MRP) with wide neural networks trained with the Temporal-Difference (TD) learning algorithm. We first consider this problem under a certain scaling of the approximating function, leading to a regime called *lazy* training. In this regime, which arises naturally, implicit in the initialization of the neural network, the parameters of the model vary only slightly during the learning process, resulting in approximately linear behavior of the model. Both in the under- and over-parametrized frameworks, we prove exponential convergence to local, respectively global minimizers of the TD learning algorithm in the lazy training regime. We then compare the above scaling with the alternative *mean-field* scaling, where the approximately linear behavior of the model is lost. In this nonlinear, mean-field regime we prove that all fixed points of the dynamics in parameter space are global minimizers. We finally give examples of our convergence results in the case of models that diverge if trained with non-lazy TD learning.

Keywords: Reinforcement learning, neural networks, temporal-difference learning, mean-field, lazy training

1. Introduction

In recent years, deep reinforcement learning has pushed the boundaries of Artificial Intelligence to an unprecedented level, achieving what was expected to be possible only in a decade and outperforming human intelligence in a number of highly complex tasks. Paramount examples of this potential have appeared over the past few years, with such algorithms mastering games and tasks of increasing complexity, from playing Atari to learning to walk and beating world grandmasters at the game of Go (Mnih et al., 2013; Mnih et al.; Silver et al., 2016, 2017, 2018; Haarnoja et al., 2018). Such impressive success would be impossible without using neural networks to approximate value functions and / or policy functions in reinforcement learning algorithms. While neural networks, in particular deep neural networks, provide a powerful and versatile tool to approximate high dimensional functions

(Cybenko, 1989; Hornik, 1991; Barron, 1993), their intrinsic nonlinearity might also lead to trouble in training, in particular in the context of reinforcement learning. For example, it is well known that nonlinear approximation of the value function might cause divergence in classical temporal-difference learning due to instability (Tsitsiklis and Van Roy, 1997). While several algorithms have been proposed in the literature to address the issue of non-convergence of nonlinear approximators (*e.g.*, Gradient Temporal Difference (Sutton et al., 2009a), GTD2, TD with gradient correction (Sutton et al., 2009b) and many others (Riedmiller, 2005; Bhatnagar et al., 2009; Maei and Sutton, 2010; Szepesvári, 2010)), practical deep reinforcement learning often employs and prefers basic algorithms such as temporal-difference (Sutton, 1988) and Q-learning (Watkins, 1989) due to their simplicity. It is thus crucial to understand the convergence of such algorithms and to bridge the gap between theory and practice.

The theoretical understanding of deep reinforcement learning is of course rather challenging, as even for supervised learning, which can be viewed as a special case of reinforcement learning, deep neural networks are still far from being understood despite the huge amount of research focus in recent years. On the other hand, recent progress has led to an emerging theory for neural network learning including recent works on the mean-field point of view of training dynamics (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Rotskoff et al., 2019; Wei et al., 2018; Chizat and Bach, 2018; Wojtowytsch, 2020) and also on the linearized training dynamics in the over-parametrized regime (Jacot et al., 2018; Allen-Zhu et al., 2018; Du et al., 2018, 2019; Zou et al., 2018; Allen-Zhu et al., 2019; Chizat et al., 2019; Oymak and Soltanolkotabi, 2020; Ghorbani et al., 2019a; Lee et al., 2019).

The main goal of this work is to analyze the dynamics of a prototypical reinforcement learning algorithm – Temporal-Difference (TD) learning – based on the recent progress in deep supervised learning. In particular, we first focus on the lazy training regime, (also known as the neural tangent kernel regime when training wide neural networks), and analyze TD learning in both over-parametrized and under-parametrized regimes with scaled value function approximations. We then compare such lazy models with their mean-field counterpart in terms of accuracy and convergence for TD learning.

Related Works. This work is closely related to the recent paper Chizat et al. (2019), addressing the problem of lazy training in the supervised learning framework when models are trained through (stochastic) gradient descent. In particular, that paper introduced the scaling that we consider in this work as an explanation, *e.g.*, of the small relative displacement of the weights of over- and under-parametrized neural networks for supervised learning. That work, however, leverages the gradient structure of the underlying vector field, which we lack in the present framework when the underlying policy is not reversible (Ollivier, 2018). The linear stability analysis is also considered in the recent work Achiam et al. (2019) based on the neural tangent kernel (Jacot et al., 2018) for off-policy deep Q-learning.

The groundbreaking paper Tsitsiklis and Van Roy (1997) proves convergence of TD learning for linear value function approximation, unifying the manifold interpretations of this convergence phenomenon that preceded it by highlighting that convergence of the algorithm is to be understood in the norm induced by the invariant measure of the underlying Markov process. Furthermore, the paper gives an illuminating counterexample for the extension of the linear result to the general, nonlinear setting. Our result shows that divergence does not occur in the lazy training regime.

The recent work Brandfonbrener and Bruna (2020) discusses the stabilizing effect homogeneity in the approximating function class on TD training dynamics. This work further shows convergence

and non-divergence of TD learning in the over-parameterized, respectively the under-parametrized regime, provided that the environment is sufficiently reversible. We note that working in the lazy training regime allows to ensure convergence independently on the reversibility of the environment and to quantify the error of the fitted model in the under-parametrized regime. Another recent work Cai et al. (2019) analyzes global convergence of a modified TD algorithm for two-layer neural networks with ReLu nonlinearity when the width of the hidden layer diverges. In contrast, in the present paper we focus on the *original* TD(λ) learning algorithm, and consider various training regimes. Furthermore, our convergence results in the lazy training regime apply to more general approximators (including, but not limited to, multiple activation functions). More recently, Zhang et al. (2020a) discuss the convergence of TD learning with neural network representations in the mean-field regime. The optimality bounds established in that paper, however, critically depend on a scaling parameter α which is assumed to be large. This scaling parameter corresponds to the lazy training parameter discussed in this paper. Therefore, the analysis of Zhang et al. (2020a) can be seen as a combination of the limits taken in this work (large width first and lazy training after). We also note that the results of Zhang et al. (2020a) are restricted to neural networks with bounded weights in the last layer. Further recent works discussing convergence properties of deep reinforcement learning algorithms both in the lazy and the mean-field regime include (Agazzi and Lu, 2020; Zhang et al., 2019, 2020b; Qiu et al., 2020; Wang et al., 2019; Sirignano and Spiliopoulos, 2019).

Finally, the mean-field analysis in our paper is tightly connected to the recent line of work (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Rotskoff et al., 2019; Wei et al., 2018; Chizat and Bach, 2018). In particular, our proof of global optimality of the fixed points extends the results from Chizat and Bach (2018) bypassing the lack of the gradient structure in the TD learning setting.

Contributions. This paper discusses the use of wide neural networks as nonlinear function approximators in value-based reinforcement learning. In particular, we consider on-policy TD learning for policy evaluation, a widely used algorithm for value function approximation tasks. We study convergence and optimality of wide neural networks trained with this algorithm under different scaling regimes of the parameters at their initialization, contrasting the results to highlight advantages and drawbacks of the various choices of initialization. More precisely:

1. We prove convergence of TD learning (asymptotically with probability one) in the lazy training regime. More specifically, we prove convergence of this algorithm in both the under- and over-parametrized regime to local and global minima, respectively, of a natural, weighted error function (the projected TD error), and illustrate such convergence properties through numerical examples.

2. To obtain the result summarized above, we adapt the contraction conditions developed in the framework of linear function approximations to a nonlinear, differential geometric setting. Furthermore, we extend existing results on the convergence in the lazy training regime of nonlinear models trained by gradient descent in the supervised learning framework to the context of reinforcement learning. This requires a generalization of the techniques developed in the gradient flow setting to non-gradient (*i.e.*, rotational) vector fields such as the ones encountered in the TD learning framework.

3. To put the result in 1. into perspective, we compare the lazy training regime to the alternative mean-field regime. In particular we show that under some technical assumptions, in the

mean-field case all the stationary points of the TD dynamics are global minimizers, *i.e.*, the model perfectly approximates the desired value function. This result provides evidence that models in the mean-field regime benefit of a far stronger approximating power.

Organization of the paper The paper is organized as follows: In Section 2 we introduce the framework of Markov reward processes and the TD learning algorithm, discussing the approximations made throughout the paper and introducing the lazy training regime. In Section 3 we discuss the convergence properties of the lazy training regime in the over-parametrized setting. In Section 4 extend the convergence results in the lazy setting to the under-parametrized case, and we compare this regime with the mean-field regime. We give some numerical examples in Section 5 and discuss our results in Section 6. Technical proof are deferred to the appendix.

2. Markov Reward Processes, TD learning and scaling limits

We denote a Markov Reward Process (MRP) by the 4-tuple $(\mathcal{S}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $P = P(s, s')_{s,s' \in \mathcal{S}}$ a transition kernel, $r(s, s')_{s,s' \in \mathcal{S}}$ is the real-valued, bounded immediate reward function and $\gamma \in (0, 1)$ is a discount factor. In this context, the *value function* $V : \mathcal{S} \to \mathbb{R}_+$ maps each state to the infinite-horizon, expected discounted reward obtained by following the Markov process defined by P. We assume that this Markov process satisfies the following assumption:

Assumption 1 *The Markov process with transition kernel P is ergodic and its stationary measure μ has full support in $\mathcal{S}$. Furthermore we assume that $\mathcal{S}$ is compact.*

We are interested in learning the value (or cost-to-go) function $V^*(s)$ of a given MRP $(\mathcal{S}, P, r, \gamma)$, which is given by

$$V^*(s) := \mathbb{E}_s \left[\sum_{k=0}^{\infty} \gamma^k r(s_k, s_{k+1}) \right], \tag{1}$$

where $\mathbb{E}_s [\,\cdot\,]$ denotes the expectation of the stochastic process s_k starting at $s_0 = s$. More specifically we would like to estimate this function through a set of predictors $V_w(s)$ in a Hilbert space $\mathcal{F}$ parametrized by a vector $w \in \mathcal{W} := \mathbb{R}^p$. We make the following assumption on such predictors:

Assumption 2 *The parametric model $V : \mathbb{R}^p \to \mathcal{F}$ mapping $w \mapsto V_w(\,\cdot\,)$ is differentiable with Lipschitz continuous derivative $DV : w \mapsto DV_w$ (where DV_w is a linear map from $\mathbb{R}^p \to \mathcal{F}$) with Lipschitz constant L_{DV} defined* WRT *the operator norm.*

A popular algorithm to solve this problem is given by value function approximation with *TD(λ)* updates (Sutton and Barto, 2018). Starting from an initial condition $w(0) \in \mathcal{W}$, for any $\lambda \in [0, 1)$, this learning algorithm updates the parameters w of the predictor along a path $\{s_k\}$ in $\mathcal{S}$ by the following rule:

$$w(t+1) := w(t) + \beta_t \, \delta(t, t) \, z_\lambda(t; \{s_k\}_0^t) \,, \tag{2}$$

for a *fixed* sequence of time steps $\{\beta_t\}$ to be specified later, where the *temporal-difference error* $\delta(\cdot, \cdot)$ and *eligibility vector* $z_\lambda(\cdot; \cdot)$ are given by

$$\delta(t, k) := r(s_k, s_{k+1}) + \gamma V_{w(t)}(s_{k+1}) - V_{w(t)}(s_k) \qquad z_\lambda(t; \{s_k\}_{k_0}^{k_1}) := \sum_{\tau=k_0}^{k_1} (\gamma\lambda)^{k_1 - \tau} \nabla_w V_{w(t)}(s_\tau) \,. \tag{3}$$

This work focuses on the asymptotic regime of small constant step-sizes $\beta_t \to 0$. In this adiabatic limit, the stochastic component of the dynamics is averaged out before the parameters of the model can undergo a significant change. This allows to consider the TD update as a deterministic dynamical system emerging from the averaging of the underlying stochastic algorithm. We focus on analysis of this deterministic system to highlight the aspect of nonlinear function approximation. The averaged, deterministic dynamics is given by the set of ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = \mathbb{E}_\mu \left[\big(r(s_0, s_1) + \gamma V_{w(t)}(s_1) - V_{w(t)}(s_0)\big) z_\lambda(t; \{s_k\}_{-\infty}^0) \right] , \tag{4}$$

where $\mathbb{E}_\mu$ denotes the expectation of the underlying dynamics at stationarity (started at $k_0 = -\infty$) and $z_\lambda(t; \{s_k\}_{k_0}^{k_1})$ is defined in (3). In the case of finite state space ($|\mathcal{S}| = d$) we can represent V_w as a vector in $\mathbb{R}^d$, while in general it is a function $\mathcal{S} \to \mathbb{R}$, which we will restrict to the Hilbert space $\mathcal{F} := L^2(\mathcal{S}, \mu)$ of square integrable functions WRT μ.

To streamline our analysis, we define the TD operator $T^\lambda : L^2(\mathcal{S}, \mu) \to L^2(\mathcal{S}, \mu)$:

$$T^\lambda V(s) := (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \mathbb{E}_s \left[\sum_{k=0}^{m} \gamma^k r(s_k, s_{k+1}) + \gamma^{m+1} V(s_{m+1}) \right] .$$

Note that when $\lambda = 0$ the above operator acquires the simple form $T^0 V := \bar{r} + \gamma P V$ for $\bar{r}(s) := \mathbb{E}_s [r(s, s')]$. Then, denoting throughout by DV_w the Fréchet derivative of V at w[1], it can be shown (Tsitsiklis and Van Roy, 1997, Lemma 8) (and is immediately verified when $\lambda = 0$) that the dynamics (4) for general $\lambda < 1$ can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = \langle T^\lambda V_{w(t)} - V_{w(t)}, DV_{w(t)} \rangle_\mu , \tag{5}$$

where we define throughout the inner product on $\mathcal{F}$ induced by the invariant measure μ (acting componentwise in expressions such as the one above) as

$$\langle a, b \rangle_\mu := \int_{\mathcal{S}} a(s) b(s) \mu(\mathrm{d}s) , \tag{6}$$

and denote throughout by $\| \cdot \|_\mu$ and Γ the corresponding norm and metric tensor respectively. Note that in the case $|\mathcal{S}| = d$, Γ reduces to the d-dimensional diagonal matrix whose entries are the (positive) values of the invariant measure $\mu(s)$, and one has $\langle a, b \rangle_\mu = a^\top \Gamma b$.

The extension of convergence results for the limiting, average dynamics we consider in this paper to convergence with probability one of the underlying, stochastic algorithm can be obtained through standard stochastic approximation arguments (Borkar and Meyn, 2000; Borkar, 2009). More details on this extension are given in Remark 4 in Section 2.1 and in the appendix.

2.1. Mean-field models and lazy training regime

In this paper, we consider models of the form

$$V_w(s) = \frac{1}{N} \sum_{i=1}^{N} \psi(s; w^{(i)}) \qquad \text{where } w^{(i)} \in \Omega \subseteq \mathbb{R}^m \, \forall \, i \in (1, \dots, N) , \psi : \mathcal{S} \times \Omega \mapsto \mathcal{F} , \tag{7}$$

1. in the finite-dimensional case $|\mathcal{S}| = d < \infty$, DV_w can be identified with the $d \times p$-dimensional Jacobian matrix of V_w

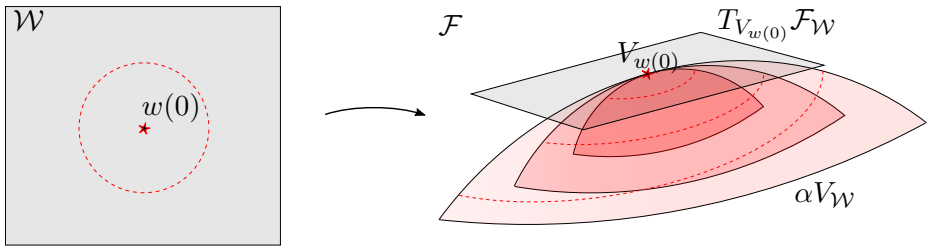


Figure 1: Schematic representation of the effect of the linear scaling of the approximating function (*e.g.*, in (12)) for different values of α in the *under-parametrized* setting. The space of parameters (left) is mapped to the space of predictors (right) by the parametric model V. The scaling $V \to \alpha V$ changes the manifold $\mathcal{F}_{\mathcal{W}}$ that the parameter space is mapped to (different surfaces on the right). In particular, this scaling "widens" the reach in the space of functions of the predictors within a ball of small radius in $\mathcal{W}$, but at the same time it "flattens" that space (locally in $\mathcal{W}$) bringing it closer to the tangential plane $\mathcal{T}_{V_{w(0)}}\mathcal{F}_{\mathcal{W}}$ to $\mathcal{F}_{\mathcal{W}}$ at the initial model $V_{w(0)}$. Choosing $V_{w(0)} = 0$ as in the picture above leaves the initial point of the dynamics (in predictor space) invariant under such transformation.

where the prediction of the model is obtained by averaging the output of N "simple", identical units ψ. In this case the TD update reads $\frac{\mathrm{d}}{\mathrm{d}t}w^{(i)}(t) = \langle T^{\lambda}V_{w(t)} - V_{w(t)}, N^{-1}\nabla_{\omega}\psi(\cdot; w^{(i)})\rangle_{\mu}$. This family of models includes single layer neural networks (Chizat and Bach, 2018), radial basis functions and linear models (Rotskoff and Vanden-Eijnden, 2018).

We further introduce a certain scaling of the TD learning update, which will be the central object of study of this paper. More specifically, we define the rescaled update

$$\frac{\mathrm{d}}{\mathrm{d}t}w(t) = \frac{1}{\alpha}\langle T^{\lambda}(\alpha V_{w(t)}) - \alpha V_{w(t)}, DV_{w(t)}\rangle_{\mu} \tag{8}$$

for a scaling parameter $\alpha \geq 1$. This update, resulting from a simultaneous rescaling of time and of V_w in (5), was designed to capture the effective training dynamics of neural network models under various initializations of their parameters, as we detail below.

The remainder of the paper discusses the large-N limit of (7), and contrast the effect of two possible choices of α when scaling such model as $V_w \to \alpha V_w$:

1. **Lazy training regime ($\alpha \to \infty$ as $N \to \infty$)** This regime is realized for *large* values of the scaling parameter α in (8) . One of the reasons why this scaling of the model is of practical interest is because it arises naturally when training neural networks, implicit in some widely applied choices of initial conditions, as we explain in Section 5.2 and as discussed in *e.g.*, Chizat et al. (2019); Cai et al. (2019). As we shall see below, under some mild assumptions for large values of α the parameters w of the model vary only slightly during training, justifying the name *lazy training* regime. A visual representation of the geometric effect of this scaling in the case where $p < d < \infty$ is given in Fig. 1. When the parameters $w^{(i)}$ are drawn IID from a common distribution, as $\alpha \to \infty$ together with $N \to \infty$ the model converges to a *deterministic* kernel, called *neural tangent kernel* (NTK) (Jacot et al., 2018). We investigate the effect of this scaling transformation on the training dynamics in Sections 3 and 4.1.

2. **Mean-field regime ($\alpha = \mathcal{O}_N(1)$ and $N \to \infty$)** This regime is realized, contrarily to the above, for *bounded* values of α, for example fixing $\alpha = 1$ while $N \to \infty$, and arises under a different scaling of the model parameters at initialization as discussed in Section 5.2. This setting was investigated in the context of supervised learning Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018); Chizat and Bach (2018) under the name of *mean-field* limit. A heuristic justification that the scaling in (7) is natural for the model V_w and does not lead to the lazy regime can be found *e.g.*, in Chizat et al. (2019). In Section 4.2 we investigate the optimality properties of neural networks in this regime and compare it to the lazy training from point 1.

3. Over-parametrized regime

In the over-parametrized setting we assume that $DV_{w(0)}$ is surjective, *i.e.*, its singular values are uniformly bounded away from 0. This is only possible in the finite state space setting and is automatically the case if the number of parameters p is larger than the size of the state space $\mathcal{S}$. Admittedly, in applications such as AlphaGo (Silver et al., 2016, 2017), it is unrealistic to over-parametrize, but we start with this regime, which parallels the study of over-parametrized supervised learning, to develop the intuition of the reader and introduce some important lemmas. Analysis of the under-parametrized regime will be discussed in the next section. In order to state our first result, we introduce the scalar product in $\mathcal{F}$ defined by $\langle a, b\rangle_0 = \langle a, g_{w(0)} b\rangle$ where $g_w := (DV_w \cdot DV_w^\top)^{-1}$, and denote by $\|\cdot\|_0$ the norm it induces. Note that g_w is the metric tensor of the pushforward metric induced by the parametric model $V : \mathbb{R}^p \to \mathcal{F}$. Furthermore, we note that if the singular values of $DV_{w(0)}$ are uniformly bounded away from 0, the norms $\|\cdot\|_\mu, \|\cdot\|_0$ are equivalent, *i.e.*, there exists $\kappa > 0$ such that $\kappa^{-1}\|f\|_0 < \|f\|_\mu < \kappa\|f\|_0$ for all $f \in \mathcal{F}$.

Theorem 1 (Lazy training, over-parametrized case) *Assume that $\sigma_{min} > 0$, where σ_{min} is the smallest singular value of $DV_{w(0)}$. Assume further that $w(0)$ is such that $\|V_{w(0)}\|_0 < M := (1-\gamma)^2\sigma_{min}^2/(192\kappa^2 L_{DV}\|DV_{w(0)}\|)$, then for $\alpha > \alpha_0 := \|V^*\|_0/M$ we have for all $t \geq 0$ that*

$$\|V^* - \alpha V_{w(t)}\|_0^2 \leq \|V^* - \alpha V_{w(0)}\|_0^2 e^{-\frac{1-\gamma}{2\kappa^2}t}. \tag{9}$$

Recall that V^* is the exact value function given by (1). Moreover, if $\|V_{w(0)}\|_0 \leq C\alpha^{-1}$ for a constant $C > 0$, then $\sup_{t>0}\|w(t) - w(0)\| = \mathcal{O}(\alpha^{-1})$.

Proof sketch of Theorem 1 Similarly to the proof in Chizat et al. (2019), we first show that DV_w and V_w do not change much assuming that w stays in a small ball of radius ϱ. Then, combining this result with the Lipschitz continuous character of DV in w, we show that w does indeed stay in the desired ball of radius ϱ. To bypass the absence of a strongly convex cost functional in our framework, which was crucial in the analysis of Chizat et al. (2019), we adopt a strategy based on the use of a local Lyapunov function

$$U(f) = \|f - V^*\|_0^2, \tag{10}$$

where V^* is the sought for value function (1). The theorem is based on two preparatory lemmas. The first one (Lemma 2) states that for large values of the scaling parameter α the pushforward metric g_w varies in a negligible way during training. To state the lemma, we denote throughout by $\mathbb{1}$ the identity map in the corresponding space and by $\mathcal{B}_\varrho^\mu(v)$, $\mathcal{B}_\varrho^0(v)$ and $\mathcal{B}_\varrho(v)$ the balls with radius ϱ around v in $\|\cdot\|_\mu$, $\|\cdot\|_0$ and $\|\cdot\|_2$ respectively.

Lemma 2 (Perturbation of the metric) *Let $\mathcal{G}_0$ be a compact subset of a linear space $\mathcal{G}$. For $v(0) \in \mathcal{G}_0$, let g_v be a continuous, self-adjoint linear operator that is positive definite in a neighborhood of $v(0)$ when restricted on $\mathcal{G}$. Then for all $\varepsilon > 0$ there exists $\delta > 0$ such that, for all $v \in \mathcal{B}_\delta(v(0)) \subseteq \mathcal{G}_0$*

$$g_{v(0)} = (\mathbb{1} + \tilde{g}_v) g_v \,, \tag{11}$$

for a linear operator $\tilde{g}_v$ with $\|\tilde{g}_v\| < \varepsilon$. More specifically, let σ_{min} be the smallest singular value of $DV_{w(0)}$. Then if $\varrho \le (1-\gamma)\sigma_{min}^2/(48 L_{DV})$, (11) holds with $\|\tilde{g}_{V(w)}\| < \frac{1-\gamma}{4}$ for all $w \in \mathcal{B}_\varrho(w(0))$.

We also recall from Tsitsiklis and Van Roy (1997) the following contraction property of the TD operator in the $\|\cdot\|_\mu$ norm. For the convenience of readers, we recall the proof in the appendix.

Lemma 3 *(Tsitsiklis and Van Roy, 1997, Lemmas 1, 3, 7) Under Assumption 1, for any $V, \tilde{V} \in \mathcal{F}$ we have that $\|T^\lambda V - T^\lambda \tilde{V}\|_\mu \le \gamma_\lambda \|V - \tilde{V}\|_\mu$ for $\gamma_\lambda := \gamma \frac{1-\lambda}{1-\gamma\lambda} \le \gamma < 1$. In particular there exists a unique fixed point of T^λ, $V^* \in \mathcal{F}$ given by (1).*

To prove Theorem 1, we then rely on the contraction T^λ (Lemma 3, from Tsitsiklis and Van Roy (1997)) to establish decay of the local Lyapunov function U as long as w stays within a ball. furthermore, by Lemma 2 the nonlinear effects become negligible when α is sufficiently large. The control of U in turn gives the bound of the change of w, which closes the argument. The full proofs of the theorem and the lemmas are given in the appendix. ■

Remark 4 *Our results can be extended to show stability and convergence in the stochastic approximation setting, similarly to Tsitsiklis and Van Roy (1997); Bhatnagar et al. (2009), under the additional assumption that the step size $\{\beta_t\}$ satisfies the Robbins-Monro condition (Robbins and Monro, 1951). For example, one can apply (Borkar and Meyn, 2000, Thms. 2.2, 2.4) guaranteeing almost sure convergence and exponential contraction of the expected error with probability one over the initial condition provided that the limiting vector field (in our case (8)) has a unique fixed point and is Lipschitz continuous. Lipschitz continuity is an immediate consequence of the linearity of T^λ and the boundedness of closed balls in $\mathcal{F}$ together with the Lipschitz continuity of the models Assumption 2. The existence of a fixed point (1) in $\mathcal{F}$ of the limiting vector field is trivial while its uniqueness is shown in the proof of Theorem 1 in the appendix.*

4. Under-parametrized regime

The underlying assumption in this section is that the size of state space is larger than the number of parameters, which in turn bounds the rank r of $DV_{w(0)}$ from above: $r < p < d$. In particular, in the case of wide neural networks of interest in this paper, this assumption is realized when the state space is not finite and, in particular, in the case of continuous state space $\mathcal{S}$.

4.1. Lazy training regime

In this regime, in general, there is no hope that TD will converge to the true value function V^*. In fact, the image of the operator T^λ might not even lie in the space $\mathcal{F}_\mathcal{W}$ of approximating functions. However, the derivative $DV_{w(t)}^\top$ in the TD update acts as a projection (WRT the product $\langle \cdot, \cdot \rangle_\mu$)

onto the tangent space of $\mathcal{F}_\mathcal{W}$ at $V_{w(t)}$ (more specifically, $DV_{w(t)}^\top$ projects the image of T^λ onto $\mathcal{W}$, which is then mapped back by $DV_{w(t)}$ to the tangent space $\mathcal{T}_{V(w(t))}\mathcal{F}_\mathcal{W}$ of $\mathcal{F}_\mathcal{W}$ at $V(w(t))$). We denote throughout by Π and Π_0 the projection operator under (6) onto $\mathcal{T}_{V(w(t))}\mathcal{F}_\mathcal{W}$ and $\mathcal{T}_{V(w(0))}\mathcal{F}_\mathcal{W}$ respectively. What one can hope for is that the TD algorithm converges to a locally "optimal" approximation $\tilde{V}^*$ of V^* on the manifold $\mathcal{F}_\mathcal{W}$, which is close to the best approximator $\Pi_0 V^*$ of V^* on the linear tangent space $\mathcal{T}_{V(w(0))}\mathcal{F}_\mathcal{W}$.

Theorem 5 (Lazy training, under-parametrized case) *Assume that $r := rank(DV_w)$ is constant in a neighborhood of $w(0)$ and $V_{w(0)} = 0$. Then there exists $\alpha_0 > 0$ such that for any $\alpha > \alpha_0$ the dynamics (8) (and the corresponding approximation V_w) converge exponentially fast to a locally (in $\mathcal{W}$) attractive fixed point $\tilde{V}^*$, for which $\|\Pi(T^\lambda\tilde{V}^* - \tilde{V}^*)\|_\mu = 0$ and $\|\tilde{V}^* - V^*\|_\mu < \frac{1-\lambda\gamma}{1-\gamma}\|\Pi_0 V^* - V^*\|_\mu + \mathcal{O}(\alpha^{-1})$.*

Note that for random initialization the constant rank assumption is satisfied almost surely. Indeed, the maximal rank property holds generically in $\mathcal{W}$ and thus with probability 1 at $w(0)$ when the model parameters are initialized randomly. Furthermore, by lower semicontinuity of the rank function the Jacobian DV will have maximal rank in an *open* subset of $\mathcal{W}$, ensuring the existence of the required neighborhood. The proof of Theorem 5 (sketched below) is given in the appendix.

The main difference in the proof of Theorem 5 WRT Theorem 1 is that $DV_w \cdot DV_w^\top$ does not have full rank anymore. This implies on one hand that the norms $\|\cdot\|_\mu$ and $\|\cdot\|_0$ are not equivalent in $\mathcal{F}$, even though we still have $\|\cdot\|_0 \leq \kappa\|\cdot\|_\mu$ for a $\kappa > 0$, provided that Assumption 1 holds. On the other hand, as mentioned above, this implies that the model V_w evolves on a submanifold $\mathcal{F}_\mathcal{W}$ of $\mathcal{F}$, and that T^λ does not, in general, map onto the tangential plane $\mathcal{T}_{V(w)}\mathcal{F}_\mathcal{W}$ of $\mathcal{F}_\mathcal{W}$ at V_w. The action of T^λ is then projected back onto $\mathcal{T}_{V(w)}\mathcal{F}_\mathcal{W}$ by the operator $DV_{w(t)}$. The nonlinear structure of the space $\mathcal{F}_\mathcal{W}$ complicates the proof WRT the over-parametrized case, and we apply standard differential geometric tools to map the problem back to a linear space.

Proof of Theorem 5 We apply the rank theorem (Boutaib, 2015; Lee, 2003) ((Abraham et al., 2012) for the ∞-dimensional setting) to show that there exist sets $\mathcal{W}_0, \overline{\mathcal{W}}_0 \subseteq \mathbb{R}^p$, $\mathcal{F}_0, \overline{\mathcal{F}}_0 \subseteq \mathcal{F}$ and diffeomorphic maps $\phi : \mathcal{W}_0 \to \overline{\mathcal{W}}_0$, $\psi : \mathcal{F}_0 \to \overline{\mathcal{F}}_0$ where $\psi \circ V \circ \phi^{-1} = \pi_r$, $\phi(w(0)) = 0$, $\psi(V_{w(0)}) = 0$ and, for an appropriate choice of bases, π_r maps the coordinates of $\overline{\mathcal{W}}_0$ to the *first* r coordinates of $\overline{\mathcal{F}}_0$, *i.e.*, $(x_1, \dots, x_p) \mapsto (x_1, \dots, x_r, 0, 0, \dots)$, where r is the rank of the operator $DV_{w(0)}$. See Figure 1 for an illustration. We denote by Π_r the hyperplane in $\mathcal{F}$ spanned by the first r vectors of the basis. We recall that by Boutaib (2015); Lee (2003); Abraham et al. (2012) the maps, ψ, ϕ, π_r are continuous with Lipschitz derivatives $D\psi, D\phi, D\pi_r$ respectively.

$$\begin{array}{ccc} \mathcal{W}_0 & \xrightarrow{V} & \mathcal{F}_0 \\ \downarrow{\scriptstyle \phi} & & \downarrow{\scriptstyle \psi} \\ \overline{\mathcal{W}}_0 & \xrightarrow{\pi_r} & \overline{\mathcal{F}}_0 \end{array}$$

Figure 2: Illustration of the spaces and maps used in the proof of Theorem 5

We consider the trajectory of $\overline{V}_{w(t)} := \pi_r \circ \phi(w(t)) = \psi(V_{w(t)})$. Denoting by $D\cdot$ the Fréchet derivative at the corresponding point of the dynamics and noting that $DV = D\psi^{-1}D\pi_r D\phi$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{V}_{w(t)} = -\frac{1}{\alpha}\langle D\psi DV DV^\top, T^\lambda \alpha\psi^{-1}(\overline{V}_{w(t)}) - \alpha\psi^{-1}(\overline{V}_{w(t)})\rangle_\pi$$

$$= -\frac{1}{\alpha}\langle D\pi_r D\phi D\phi^\top D\pi_r^\top (D\psi^{-1})^\top, T^\lambda \alpha\psi^{-1}(\overline{V}_{w(t)}) - \alpha\psi^{-1}(\overline{V}_{w(t)})\rangle_\pi \,, \tag{12}$$

so $\overline{V}$ remains in Π_r. As a consequence of the above we can naturally define a metric (the pushforward metric) on $\overline{\mathcal{F}}_0$ by the tensor $\bar{g}_{\bar{v}} = (D\pi_r D\phi D\phi^\top D\pi_r^\top)^{-1}$. In fact, by choosing the metric tensor to be constant on $\overline{\mathcal{F}}_0$, *i.e.*, equal to $\bar{g}_0$ for all $v \in \overline{\mathcal{F}}_0$, we equip the linear space $\overline{\mathcal{F}}_0$ with a scalar product $\langle \cdot, \cdot \rangle_0$. This, in turn, directly induces a norm $\| \cdot \|_0$ on the same space. We now proceed to use such simple metric structure to establish the existence and uniqueness of a fixed point of (12) in $\overline{\mathcal{F}}_0$ for α large enough.

The desired result follows from (Simpson-Porco and Bullo, 2014, Proposition 4.1), which establishes uniqueness and exponential contraction at rate $\ell > 0$ of a dynamical system evolving under the flow of a vector field X given by the RHS of (12) in a forward invariant set $\overline{\mathcal{F}}_0$ provided that for every geodesic $\gamma(s)$ in $\overline{\mathcal{F}}_0$ (13) holds. Therefore, the proof of convergence is concluded by applying Lemma 6 and Lemma 7. The proof of the optimality of the fixed point is postponed as Lemma 12. ■

Lemma 6 *There exists $\delta > 0$ and $\alpha_0 > 0$ such that the ball $\mathcal{B}_\delta^0(0) \subseteq \overline{\mathcal{F}}_0$ is forward invariant and forward complete* WRT *the dynamics of (8) for all $\alpha > \alpha_0$.*

Lemma 7 *There exists $\ell > 0$, $\delta > 0$ and $\alpha_0 > 0$ such that for all $\alpha > \alpha_0$ and all geodesics $\gamma(s)$ contained in the ball $\mathcal{B}_\delta^0(0) \subseteq \overline{\mathcal{F}}_0$, the function*

$$\langle \gamma'(s), X(\gamma(s))\rangle_0 - \ell s \langle \gamma'(0), \gamma'(0)\rangle_0 \,, \tag{13}$$

is strictly decreasing in s.

Remark 8 *The proof of Theorem 5 can be generalized to the case where V_0 is not identically 0 but within a $\| \cdot \|_\mu$-ball of radius $\varrho(\alpha)$ around 0 for $\varrho(\alpha)$ going to 0 with $\alpha \to \infty$. This generalization, however, requires the map V to be* uniformly *Lipschitz smooth for $w \in \mathcal{W}_0$. This extension allows to explicitly cover the training of randomly initialized, single layer neural networks.*

The *local* optimality in Theorem 5 is a consequence of the approximately linear behavior of lazy learners that we leverage in our proofs. Indeed, it is known (Jacot et al., 2018) that models in this regime behave asymptotically like kernel methods, and their expressive power is therefore limited to the corresponding Reproducing Kernel Hilbert Space (RKHS).

In this sense, in the case of randomly initialized, wide networks it is natural to compare these models to random feature models with an appropriate set of feature maps (Yehudai and Shamir, 2019). For a detailed comparison of random feature models and neural networks in the lazy training regime we refer the reader to Ghorbani et al. (2019b). Furthermore, we refer to Bach (2017) for a general discussion of the approximating power of neural network models in this lazy regime, highlighting their limits WRT their non-lazy counterparts.

4.2. Mean-field regime

To better understand the limitations of training wide neural networks in the lazy training regime, we contrast their behavior with the one of networks that do not display such approximately linear

behavior through training. These models, introduced below, correspond to a different scaling of the model parameters at initialization and are commonly referred to as mean-field models. While results on the convergence of such mean-field models are obtained in a more restricted setting detailed below, they provide an effective term of comparison to the linearized regime explored in previous sections.

In the following we set, for clarity of exposition, $\lambda = 0$ and we assume that $|\mathcal{S}| = \infty$ in order to be in the under-parametrized regime of interest. We further restrict to the setting where the function ψ is h-homogeneous for $h \geq 1$ in at least one of its parameters. A function f is called h-homogeneous if $f(\xi x) = \xi^h f(x)$ for all $\xi \in \mathbb{R}$. One simple case where this holds is when $\omega = (\omega_0, \bar{\omega}) \in \mathbb{R} \times \Theta$ for $\Theta \subseteq \mathbb{R}^{m-1}$ and $\psi(s;\omega) = \omega_0 \phi(x;\bar{\omega})$ for a certain, usually nonlinear function $\phi : \mathcal{S} \times \Theta \to \mathcal{F}$, so that ψ is 1-homogeneous in $\omega_0 \in \mathbb{R}$. This is for instance the case in the setting of single (hidden) layer neural networks, radial functions, and linear models, where ϕ represents, respectively, the nonlinear activation function, the radial basis function or the model features.

The mean-field regime By the assumed structure of the approximator, we represent V_w in (7) as $V_w(s) = V_{\nu^{(N)}}(s) := \int_\Omega \psi(s;\omega)\nu^{(N)}(\mathrm{d}\omega)$ where $\nu^{(N)}(\mathrm{d}\omega) = \frac{1}{N}\sum_{i=1}^N \delta_{w^{(i)}}(\mathrm{d}\omega) \in \mathcal{M}_+(\Omega)$. This empirical measure representation removes the symmetry of the approximating functions under permutations of the weight indices, while behaving naturally in the limit $N \to \infty$, when $\nu^{(N)} \to \nu$ weakly, so that $V_{\nu^{(N)}} \to V_\nu$. Upon rescaling time as $t \leftarrow Nt$, the evolution of the measure $\nu \in \mathcal{M}_+(\Omega)$ under (4) is then governed by a *mean-field* transport partial differential equation of the Vlasov type, given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\nu_t(\omega) = \mathrm{div}\left(\nu_t(\omega)\int_{\mathcal{S}\times\mathcal{S}} \nabla_\omega \psi(s;\omega)\delta(s,s',\nu_t)P(s,\mathrm{d}s')\mu(\mathrm{d}s)\right) \tag{14}$$

for the TD-error δ from (3), *i.e.*, $\delta(s,s',\nu) := r(s,s') + \gamma\int\psi(s';\omega')\nu(\mathrm{d}\omega') - \int\psi(s;\omega')\nu(\mathrm{d}\omega')$. Analogous dynamics equation in the supervised learning case has been derived in Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018); Chizat and Bach (2018).

To state the main result of this section, the optimality of fixed points of (14) we need the following

Assumption 3 *Assume that $\omega = (\omega_0, \bar{\omega}) \in \mathbb{R} \times \Theta = \Omega$ for $\Theta = \mathbb{R}^{m-1}$ and $\psi(s;\omega) = \omega_0\phi(s;\bar{\omega})$ with*

a) Regularity of ϕ: *ϕ is bounded, differentiable and $D\phi$ is Lipschitz. Also, for all $f \in \mathcal{F}$ the regular values of the map $\bar{\omega} \mapsto g_f(\bar{\omega}) := \langle f, \phi(\,\cdot\,;\bar{\omega})\rangle$ are dense in its range, and $g_f(r\bar{\omega})$ converges in $C^1(\{\bar{\omega}\in\Theta : \|\bar{\omega}\|_2 = 1\})$ as $r \to \infty$ to a map $\bar{g}_f(\bar{\omega})$ whose regular values are dense in its range.*

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

c) Support of the measure*: There exists $r_0 > 0$ s.t. the support of the initial condition ν_0 is contained in $\mathcal{Q}_{r_0} = [-r_0, r_0]\times\Theta$ and separates $\{-r_0\}\times\Theta$ from $\{r_0\}\times\Theta$, i.e., if any continuous path connecting $\{-r_0\}\times\Theta$ to $\{r_0\}\times\Theta$ intersects the support of ν_0.*

Assumption 3 a) is a common, technical regularity assumption (*e.g.*, (Chizat and Bach, 2018, Assumption 3.4)), ensuring that (14) is well behaved and controlling the growth, variation and regularity of ϕ. Alternative assumptions on the case $\Theta \neq \mathbb{R}^{m-1}$ are given in the appendix. Assumption 3 b) speaks to the approximating power of the nonlinearity, assumed to be powerful enough

to approximate any function in $L^2(\mathcal{S}, \mu)$, while c) guarantees that the initial condition is such that the expressivity from b) can actually be exploited. Universal approximation assumptions similar to Assumption 3 b) were made in Lu et al. (2020); Nguyen and Pham (2020).

Before discussing the optimality properties of the dynamics (14), we show that this PDE accurately describes the TD dynamics of a sufficiently wide, single layer neural network. To this aim, we let $\mathcal{P}_2(\Omega)$ be the space of probability distributions on Ω with finite second moment.

Proposition 9 *Let Assumption 3 hold and let $w_t^{(N)}$ be a solution of (4) with initial condition $w_0^{(N)} \in \mathcal{W} = \Omega^N$. If $\nu_0^{(N)}$ converges to $\nu_0 \in \mathcal{P}_2(\Omega)$ in Wasserstein distance W_2 as $N \to \infty$ then $\nu_t^{(N)}$ converges, for any $t > 0$, to the unique solution ν_t of (14).*

We note that by the law of large numbers for empirical distributions, the condition of convergence of $\nu_0^{(N)}$ to ν_0 is *e.g.*, satisfied when $w_0^{(i)}$ are drawn independently at random from ν_0. The proof of the above result is largely standard under the given assumptions, and was given in a setting similar to the one at hand in Zhang et al. (2020a). For completeness, we provide a sketch of the proof in the appendix. The idea of the proof is a canonical propagation of chaos argument (Sznitman, 1991), which we adapt from Chizat and Bach (2018) to the present context. In the absence of an energy functional in the TD setting, in order to establish the necessary regularity of the vector field we prove

Lemma 10 *For any ν_0 with $\int \omega_0^2 \nu_0(\mathrm{d}\omega) \leq \infty$ there exists $C_V > 0$ such that, for any $t > 0$ we have $\|V_{\nu_t}\|_\mu < C_V$.*

To prove the above result we leverage the homogeneity of the approximator, slightly adapting the proof of (Brandfonbrener and Bruna, 2020, Theorem 1) to the present setting. We note that the above result is of independent interest in that it rules out divergence in predictor space of the mean-field dynamics. We are now ready to state the main optimality result of this section:

Theorem 11 (Mean-field optimality) *Let Assumption 3 hold and ν_t given by (14) converge to ν^*, then $V_{\nu^*} = V^*$ μ-a.e.*

Thus if the TD-learning dynamics (14) converges to a fixed point, it is a global minimizer. To prove this result, we first connect the optimality of a fixed point with the support of the underlying measure in parameter space. More specifically, we show in Lemma 14 that by the expressivity of σ, the transport vector field of suboptimal fixed points of the dynamics (14) cannot vanish everywhere in parameter space, so that a measure with sufficient support cannot correspond to a suboptimal fixed point.

We then show in Lemma 15 that the TD dynamics in the mean-field regime preserves such sufficient notion of support of the measure (Assumption 3 c)) throughout training. This is true for any finite time by topological arguments: the separation property of the measure cannot be altered by the action of a continuous flow such as (14). Leveraging the above partial results, we finally prove in Lemma 16, that spurious fixed points are avoided by the TD dynamics (14) when initialized properly. To establish this we argue by contradiction: assuming that we are approaching such a spurious fixed point $\tilde{\nu}$ at time t_0, we show in Lemma 18 that the velocity field will change little for any $t > t_0$. On the other hand, by the homogeneity of ψ and by Assumption 3 c), we show that by Lemma 15 a positve amount of measure $\tilde{\nu}$ will fall in a forward invariant region – which exists by Lemma 14 – where its ω_0 component will grow linearly in t, thereby eventually contradicting the assumption that $\tilde{\nu}$ is a fixed point of (14).

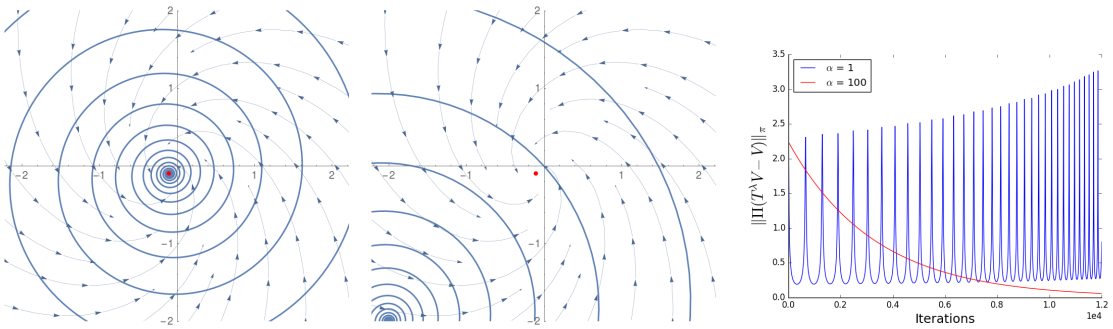
5. Numerical examples

5.1. A divergent nonlinear approximator

We illustrate the convergence properties of TD learning in the lazy training regime in the under-parametrized case by applying it to the classical framework of (Tsitsiklis and Van Roy, 1997, Section X). This reference gives an example of a family of nonlinear function approximators that diverge when trained with the TD method. The intuition behind this counterexample is that one can construct a manifold of approximating functions $\mathcal{F}_{\mathcal{W}}$ in the form of a spiral, with the same orientation as the rotation of the vector field induced by the TD update in the space of functions. By choosing the windings of the spiral to be dense enough, the projection of the TD vector field follows the spiral in the outward direction, leading to a divergence of the algorithm, as displayed schematically in Fig. 3a. More specifically, consistently with Tsitsiklis and Van Roy (1997), we parametrize the manifold $\mathcal{F}_{\mathcal{W}}$ as $V_\vartheta := e^{\hat{\varepsilon}\vartheta}(a\cos(\hat{\lambda}\vartheta) - b\sin(\hat{\lambda}\vartheta)) - V^*$ for $a = (10, -7, -3)$, $b = (2.3094, -9.815, 7.5056)$, $\hat{\varepsilon} = 0.01$, $\hat{\lambda} = 0.866$. We set $\gamma = 0.9$ and step-size $\beta_t \equiv 2 \times 10^{-3}$, while the underlying Markov chain is defined by the transition matrix $P_{ij} = (\delta_{j,\mathrm{mod}(i,3)+1} + \delta_{i,j})/2$, where $\delta_{i,j}$ is the Kronecker delta function and equals 1 if $i = j$ and 0 else. We note that the step-size does not affect the convergence properties of the algorithm, as argued in Tsitsiklis and Van Roy (1997), where the immediate reward was set to $\bar{r} = (0, 0, 0)$. Note that, as realizing the conditions of Theorem 5 starts the simulation at the solution $V^* = (0, 0, 0)$, we shift both the solution and the manifold of approximating functions by the same vector in the embedding space, leaving the new solution $V^* = -V_0 = -a$ at the center of the spiral, *i.e.*, realized at $\vartheta = -\infty$. This corresponds to choosing an average reward $\bar{r} = (-6.85, 8.35, -1.5)$. We note that by the affine nature of the TD update, this change in $\bar{r}$ results in a global shift of the TD vector field in $\mathcal{F}$ and does not affect the update of ϑ. In particular, this means that the TD update remains *divergent* for every initial condition different than the solution V^*. We run the TD update in the off-centered situation both for values of $\alpha = 1$ (the classical, divergent regime) and $\alpha = 100$. As explained in the previous sections, this scaling of the approximating function makes the TD update *convergent*, as displayed in Fig. 3c. The intuition behind the convergence of the algorithm is outlined in Fig. 3: when α is large we are in an almost linear regime where the TD update converges to a *local* minimum of the dynamics.

5.2. Single layer neural networks

We show that the regime of study arises naturally in one hidden layer neural networks for a certain family of initialization. We consider the example of ReLu activation, *i.e.*, when the model is given by $V_w(s) = \sum_{i=1}^N a_i \max(0, b_i \cdot s - c_i)$, for $s \in \mathbb{R}^m$ and N distinct $(m+2)$-dimensional vectors $w_i = (a_i, (b_i)_1, \dots, (b_i)_m, c_i)_{i\in(1,\dots,N)}$. Typical initialization of the weights of the above model is of the form $a_i \overset{iid}{\sim} \mathcal{N}(0, 1/\sqrt{N})$, $(b_i)_j \overset{iid}{\sim} \mathcal{N}(0, 1/\sqrt{m})$ for all j and $c_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. However, by the linearity of the activation function in a_i and by the rescaling property of normal distribution this is equivalent to writing $\alpha V_w(s) = \alpha \frac{1}{N} \sum_{i=1}^N a_i \max(0, b_i \cdot s - c_i)$ for an N-dependent $\alpha(N) = \sqrt{N}$ (diverging in N), $a_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, $(b_i)_j \overset{iid}{\sim} \mathcal{N}(0, 1/\sqrt{d})$ and $c_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. Therefore, this common choice of initial conditions implicitly starts the training of the above model in the lazy regime (Ghorbani et al., 2019b). We train the network by TD learning (8) with fixed step-size $\beta_t \equiv 10^{-3}$ both in the over- and under-parametrized regime. To do so, we draw an objective function V^* randomly with distribution $V^*(s) \overset{iid}{\sim} \mathcal{N}(0, 1)$ for all $s \in \mathcal{S}$ on a grid of d equally spaced points on



(*a*) $\alpha = 1$ (*b*) $\alpha = 100$ (*c*) Training dynamics

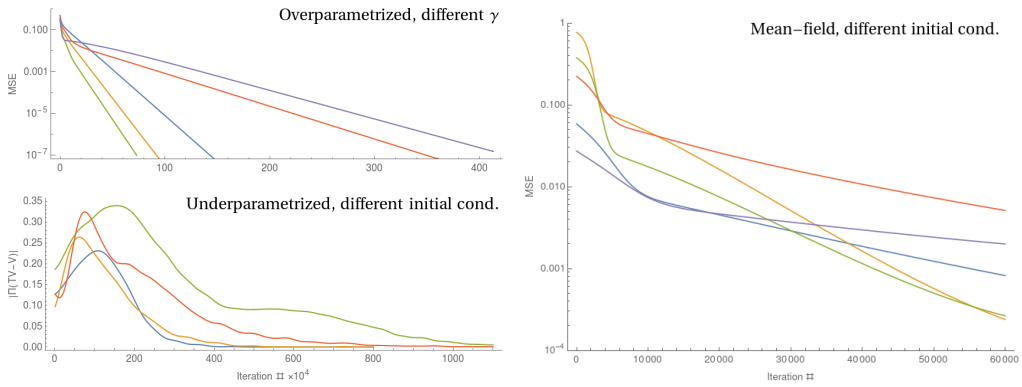
Figure 3: Schematic representation of the manifold $\mathcal{F}_{\mathcal{W}}$ for the example in Section 5.1 before (a) and after (b) scaling of α. The underlying vector field represents the TD error $\delta(V)$ from (3), whose projection on $\mathcal{T}_V\mathcal{F}_{\mathcal{W}}$ gives the dynamics of the TD update in $\mathcal{F}_{\mathcal{W}}$. In (a) this projection points "outwards" along the spiral, while (b) it has a fixed point close to 0. The scaling yields an effective "linearization" of the manifold around 0. The red point marks the global fixed point of the vector field. In (c), we plot the μ-norm of the projected TD error $\Pi(T^\lambda V - V)$. This quantity measures the increments of the model parameters during training and vanishes at a local minimum of the TD dynamics. We see that the algorithm diverges for $\alpha = 1$ (blue curve), but converges to a local minimum for $\alpha = 100$ (red curve).

the interval $[-1, 1]$. We then compute the corresponding average reward by solving $\bar{r} = (\mathbb{1} - \gamma P)V^*$, and train the model (8) for $\lambda = 0$, $\gamma = 0.9$ (when not specified otherwise) with transition matrix $P_{ij} = (\delta_{j,\mathrm{mod}(i,d)+1} + \delta_{i,j})/2$. To respect the conditions of Theorem 5, we initialize half of the model parameters as explained above, while the other half is obtained from the first by replicating the values of b_i, c_i and inverting the one of $a_i \to -a_i$. This "doubling trick" introduced in Chizat et al. (2019) produces a neural network with $V_{w(0)} \equiv 0$ and randomly initialized weights with the desired distribution. We consider situations where $N = 10$, $d = 50$ (under-parametrized, taking $\alpha = 100$) and $N = 100$, $d = 30$ (over-parametrized, with $\alpha = 500$), and plot the convergence to local, respectively global minima in Fig. 4a.

To illustrate the convergence properties of neural networks in the mean-field regime as presented in Theorem 11 we repeat the above experiment taking $\alpha = 1$. In this setting, we consider a model with parameters $N = 350$, $d = 700$, resulting in the under-parametrized regime. In this case, to ensure that the value function to be learned can be approximated by the finite-width network we consider a target value function $V^*(s)$ given by a neural network of the form (7) with $N^* = 4$, initialized as $a_i^* \overset{iid}{\sim} \mathcal{N}(0, 1)$, $(b_i^*)_j \overset{iid}{\sim} \mathcal{N}(0, 1/\sqrt{d})$ and $c_i^* \overset{iid}{\sim} \mathcal{N}(0, 1)$. The results of these simulations are displayed in Fig. 4b. We notice that, while the MSE converges to zero, the rate of decay of this quantity does not display the regularity observed in the over-parametrized, lazy regime, numerically corroborating the hypothesis that the convergence of mean-field models is not *uniformly* linear.

6. Conclusions and future work

In this work we have discussed the convergence and optimality properties of the TD learning algorithm with wide neural networks as function approximators, comparing the effect of different initialization



(*a*) Lazy training (*b*) Mean-field

Figure 4: Training results or the examples described Section 5.2 in the lazy (a) and mean-field (b) regimes. In (a, top) we plot the MSE of single layer neural network during training in the over-parametrized regime ($N = 100, d = 30, \alpha = 500$) for different choices of γ $(0.8, 0.83, 0.85, 0.87, 0.9)$, showing exponential convergence (at different rates) to the global minimum claimed in Theorem 1. In (a, bottom) we plot the norm of the projected TD error for a neural network in the under-parametrized regime ($N = 10, d = 50, \alpha = 100, \gamma = 0.9$) for different initial conditions, showing convergence to a local fixed point. In (b) we again plot the MSE in the mean-field regime ($N = 350, d = 700, \alpha = 1, \gamma = 0.9$) for different initial conditions. This plot displays qualitatively different convergence behaviors, with no apparent upper bound on the convergence rate, indicative of the nonlinear character of the model. Code available here: Agazzi and Lu (2021).

procedures on the dynamics of the algorithm and on the limiting approximator, both in the under- and over-parametrized setting.

In the lazy regime, the algorithm behaves essentially like a linear approximator spanning the tangential space of the approximating manifold (in function space) at initialization. As such, the training converges exponentially fast with probability one to the global minimum or a local fixed point depending on the codimension of the approximating manifold in the search space. This phenomenon can be understood as an effect of the linearized regime in which the neural networks are trained which reduces them, in the limit, to a kernel method (Jacot et al., 2018). This somewhat limited expressivity is reflected in the *local* optimality of the fixed points of the TD dynamics in the lazy regime.

We contrast this behavior by proving *global* optimality guarantees for the fixed points of the same models when trained in the mean-field scaling limit, the highly nonlinear dynamics of which make convergence (and convergence rate bounds) hard to establish. In this sense, we argue that convergence of lazy models comes at the expense of their expressivity. Nonetheless, the results proven in this work emphasize the interest of the lazy regime in the framework of deep reinforcement learning, where models often suffer from divergent behavior especially during early stages of training. Optimal results are expected given prior knowledge that the value function belongs to a known RKHS: using a nonlinearity whose NTK includes such RKHS would result in both convergence and optimality guarantees of the learned value function.

Future directions of research include the development of more refined, nonasymptotic versions of the above theorems and the extension of these results to more complex, nonlinear reinforcement

learning algorithms in the Markov Decision Processes setting. We note that such extension is immediate in settings such as SARSA, where the state-action space can be regarded as an extension of the state space in the TD case. On the contrary, the generalization of the results obtained above to off-policy algorithms such as Q-learning pose a more significant challenge. This is captured for instance in Melo et al. (2008), where convergence guarantees for linear approximators can only be obtained upon making relatively strong a-priori assumptions on the spectral properties of the feature matrix during training. Consequently this direction of research remains open. Furthermore, a more thorough exploration of the relationship between the limiting results in Chizat and Bach (2018) and the ones presented here and in Chizat et al. (2019) would be important for the understanding of the limiting dynamics of neural networks in this domain, in particular in terms of convergence of the nonlinear, mean-field limit in the context of reinforcement learning.

Acknowledgments

We thank a bunch of people.

References

Ralph Abraham, Jerrold E Marsden, and Tudor Ratiu. *Manifolds, tensor analysis, and applications*, volume 75. Springer Science & Business Media, 2012.

Joshua Achiam, Ethan Knight, and Pieter Abbeel. Towards Characterizing Divergence in Deep Q-Learning, 2019. preprint, arXiv:1903.08894.

Andrea Agazzi and Jianfeng Lu. Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. *arXiv e-prints*, art. arXiv:2010.11858, October 2020.

Andrea Agazzi and Jianfeng Lu. TD learning with wide, single layer neural networks. https://github.com/agazzian/deepTD, 2021.

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

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

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

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

Shalabh Bhatnagar, Doina Precup, David Silver, Richard S Sutton, Hamid R. Maei, and Csaba Szepesvári. Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems 22*, pages 1204–1212. 2009.

V. Borkar and S. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000. doi: 10.1137/S0363012997331639.

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

Youness Boutaib. On Lipschitz maps and their flows, 2015. preprint, arXiv:1510.07614.

David Brandfonbrener and Joan Bruna. Geometric insights into the convergence of nonlinear TD learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Qi Cai, Zhuoran Yang, Jason D. Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima, 2019. preprint, arXiv:1905.10027.

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

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

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

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

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

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

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 9108–9118, 2019b.

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

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

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

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

J.M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, 2003. ISBN 9780387954486.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean field analysis of deep resnet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pages 6426–6436. PMLR, 2020.

Hamid Reza Maei and Richard S. Sutton. GQ(lambda): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *3rd Conference on Artificial General Intelligence (AGI-2010)*. Atlantis Press, 2010. ISBN 978-90-78677-36-9.

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

Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671, 2008.

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

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

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

Yann Ollivier. Approximate temporal difference learning is a gradient descent for reversible policies, 2018. preprint, arXiv:1805.00869.

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

Shuang Qiu, Zhuoran Yang, Xiaohan Wei, Jieping Ye, and Zhaoran Wang. Single-Timescale Stochastic Nonconvex-Concave Optimization for Smooth Nonlinear TD Learning. *arXiv e-prints*, art. arXiv:2008.10103, August 2020.

Martin Riedmiller. Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method. In João Gama, Rui Camacho, Pavel B. Brazdil, Alípio Mário Jorge, and Luís Torgo, editors, *Machine Learning: ECML 2005*, pages 317–328, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22:400–407, 1951.

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

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

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

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

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

John W Simpson-Porco and Francesco Bullo. Contraction theory on Riemannian manifolds. *Systems & Control Letters*, 65:74–80, 2014.

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

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

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3 (1):9–44, Aug 1988.

Richard S Sutton, Hamid R Maei, and Csaba Szepesvári. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*, pages 1609–1616, 2009a.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000. ACM, 2009b.

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010.

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

John N Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems*, pages 1075–1081, 1997.

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

C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Oxford, 1989.

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

Stephan Wojtowytsch. On the convergence of gradient descent training for two-layer relu-networks in the mean field regime. *arXiv preprint arXiv:2005.13530*, 2020.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2019.

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

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

Yufeng Zhang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Generative adversarial imitation learning with neural network parameterization: Global optimality and convergence rate. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11044–11054, Virtual, 13–18 Jul 2020b. PMLR.

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

Appendix A. Supplementary proofs in the lazy training regime

To simplify the notation in the forthcoming analysis, we slightly abuse the notation used when the state space is finite-dimensional extending it, when necessary, to the infinite-dimensional setting. This naturally generalizes matrix multiplication to the action of linear operators. In particular the

action of Γ, which we recall in the finite-dimensional setting is a diagonal matrix with entries $\mu(s)$, is to be intended as

$$(a^\top \Gamma b)_{ij} = \int_{\mathcal{S}} a_i(s) b_j(s)\, \mu(\mathrm{d}s)\,.$$

Furthermore, we introduce the following decomposition of the TD operator:

$$T^\lambda V = \bar{r}^\lambda + \gamma P^\lambda V\,,$$

where

$$\bar{r}^\lambda(s) := (1-\lambda)\sum_{m=0}^{\infty}\lambda^m \mathbb{E}_s\left[\sum_{k=0}^{m}\gamma^k r(s_k, s_{k+1})\right], \quad P^\lambda V(s) := (1-\lambda)\sum_{m=0}^{\infty}(\lambda\gamma)^m \mathbb{E}_s\left[V(s_{m+1})\right],$$

or, in vector notation

$$\bar{r}^\lambda := (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{k=0}^{m}\gamma^k P^k r\,, \quad P^\lambda V(s) := (1-\lambda)\sum_{m=0}^{\infty}(\lambda\gamma)^m P^{m+1} V\,.$$

In the proofs below, we will use the above, simplified notation to obtain contraction estimates on the dynamical system (4). These estimates will leverage the fact that P^λ is nonexpansive and $\gamma < 1$, and from this notation contraction rates in terms of γ will arise naturally. However, by Lemma 3, we know that the contraction rate of T^λ is γ_λ. Rewriting the proofs with $\gamma \to \gamma_\lambda$ will show the stronger contraction.

A.1. Over-parametrized regime

Lemma 2 *Let $\mathcal{G}_0$ be a compact subset of a linear space $\mathcal{G}$. For $v(0) \in \mathcal{G}_0$, let g_v be a continuous, self-adjoint linear operator that is positive definite in a neighborhood of $v(0)$ when restricted on $\mathcal{G}$. Then for all $\varepsilon > 0$ there exists $\delta > 0$ such that, for all $v \in \mathcal{B}_\delta(v(0)) \subseteq \mathcal{G}_0$*

$$g_{v(0)} = (\mathbb{1} + \tilde{g}_v) g_v\,,$$

for a linear operator $\tilde{g}_v : \mathcal{F} \to \mathcal{F}$ with $\|\tilde{g}_v\| < \varepsilon$. More specifically, let σ_{min} be the smallest singular value of $DV_{w(0)}$. Then if $\varrho \le (1-\gamma)\sigma_{min}^2/(48 L_{DV})$, (11) holds with $\|\tilde{g}_{V(w)}\| < \frac{1-\gamma}{4}$ for all $w \in \mathcal{B}_\varrho(w(0))$.

Proof of Lemma 2 Let $B_w = DV_{w(0)}DV_{w(0)}^\top - DV_w DV_w^\top$. We carry out the proof for the case $\sigma_{\min} < 1$ (else the result holds with $\sigma_{\min} = 1$ in ϱ), in which case we have for all $w \in \mathcal{B}_\varrho(w(0))$ that

$$\|B_w\| \le 2L_{DV}\|w(0) - w\| + (L_{DV}\|w(0) - w\|)^2 \le 3L_{DV}\|w(0) - w\|\,.$$

Then we can write

$$\begin{aligned} g_{w(0)} &= (DV_{w(0)}DV_{w(0)}^\top)^{-1} = (DV_w DV_w^\top + B_w)^{-1} \\ &= (g_w^{-1}(1 + g_w B_w))^{-1} = (1 + g_w B_w)^{-1} g_w \\ &= \sum_{n=0}^{\infty}(-1)^n (g_w B_w)^n g_w = g_w + \sum_{n=1}^{\infty}(-1)^n (g_w B_w)^n g_w\,. \end{aligned}$$

Furthermore, by the assumptions on the regularity of V and on the initial condition $w(0)$ we have that $g_w \preceq 4/\sigma_{\min}^2 \mathbb{1}$, provided that $w \in \mathcal{B}_\varrho(w(0))$ for ϱ as in Lemma 2. Therefore, the perturbation $\tilde{g}_w := \sum_{n=1}^{\infty}(-1)^n (g_w B_w)^n$ satisfies

$$\|\tilde{g}_w\| = \|\sum_{n=1}^{\infty}(-1)^n (g_w B_w)^n\| \le \sum_{n=1}^{\infty} \|g_w B_w\|^n \le \sum_{n=1}^{\infty} \left(\frac{3L_{DV}}{\sigma_{\min}^2/4}\|w(0)-w\|\right)^n \le \frac{1-\gamma}{4}\,.$$

The same proof applies in the general case with different, implicit constants. ■

Lemma 3 *(Tsitsiklis and Van Roy, 1997, Lemmas 1, 3, 7) Under Assumption 1, for any $V, \tilde{V} \in \mathcal{F}$ we have that*

$$\|T^\lambda V - T^\lambda \tilde{V}\|_\mu \le \gamma_\lambda \|V - \tilde{V}\|_\mu \qquad \textit{for} \quad \gamma_\lambda := \gamma \frac{1-\lambda}{1-\gamma\lambda} \le \gamma < 1\,. \tag{15}$$

In particular there exists a unique fixed point of T^λ, $V^* \in \mathcal{F}$ given by (1).

Proof of Lemma 3 We first prove that $\|PV\|_\mu \le \|V\|_\mu$. This follows by Jensen inequality and by the invariance of μ:

$$\begin{aligned}\|PV\|_\mu^2 = V^\top P^\top \Gamma P V &= \int_{\mathcal{S}} \mu(\mathrm{d}s)\Big(\int_{\mathcal{S}} P(s,\mathrm{d}s')V(s')\Big)^2 \\ &\le \int_{\mathcal{S}^2} \mu(\mathrm{d}s)P(s,\mathrm{d}s')V(s')^2 = \int_{\mathcal{S}} \mu(\mathrm{d}s)V(s)^2 = \|V\|_\mu^2\,.\end{aligned} \tag{16}$$

Then, writing

$$\begin{aligned}T^\lambda V(s) &= (1-\lambda)\sum_{m=0}^{\infty}\lambda^m \mathbb{E}_s\left[\sum_{k=0}^{m}\gamma^k r(s_k,s_{k+1}) + \gamma^{m+1}V(s_{m+1})\right] \\ &= (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\left(\sum_{k=0}^{m}\gamma^k \mathbb{E}_s\left[\bar{r}(s_k)\right] + \gamma^{m+1}\mathbb{E}_s\left[V(s_{m+1})\right]\right) \\ &= (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\left(\sum_{k=0}^{m}\gamma^k P^k \bar{r}(s) + (\gamma P)^{m+1}V(s)\right)\,,\end{aligned}$$

where s_k is the process on $\mathcal{S}$ induced by P with initial condition s_0, we have contraction of the operator T^λ in $L^2(\mathcal{S},\mu)$ by

$$\begin{aligned}\|T^\lambda(V-\tilde{V})\|_\mu &= \left\|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\gamma P)^{m+1}\left(V(s)-\tilde{V}(s)\right)\right\|_\mu \\ &\le (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\left\|V(s)-\tilde{V}(s)\right\|_\mu \\ &= \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\left\|V(s)-\tilde{V}(s)\right\|_\mu\,,\end{aligned}$$

where in the inequality above we have used (16). This proves that T^λ is a contraction in $\mathcal{F}$, and as such it must have a unique fixed point. That this fixed point corresponds to (1) is immediately checked by direct computation. ■

Theorem 1 *Assume that $\sigma_{min} > 0$, where σ_{min} is the smallest singular value of $DV_{w(0)}$. Assume further that $w(0)$ is such that $\|V_{w(0)}\|_0 < M := (1-\gamma)^2\sigma_{min}^2/(192\kappa^2 L_{DV}\|DV_{w(0)}\|)$, then for $\alpha > \alpha_0 := \|V^*\|_0/M$ we have for all $t \geq 0$ that*

$$\|V^* - \alpha V_{w(t)}\|_0^2 \leq \|V^* - \alpha V_{w(0)}\|_0^2 e^{-\frac{1-\gamma}{2\kappa^2}t}\,. \tag{17}$$

Recall that V^* is the exact value function given by (1). Moreover, if $\|V_{w(0)}\|_0 \leq C\alpha^{-1}$ for a constant $C > 0$, then $\sup_{t>0}\|w(t) - w(0)\| = \mathcal{O}(\alpha^{-1})$.

Proof of Theorem 1 By setting $\varrho := (1-\gamma)\sigma_{\text{min}}^2/(48 L_{DV})$ and by the assumed Lipschitz smoothness of V, $DV_w \cdot DV_w^\top \succeq \sigma_{\text{min}}^2/4$ as long as $w \in \mathcal{B}_\varrho(w(0))$. We would like to check a local exponential contraction condition, *i.e.*, that for all $w(t) \in \mathcal{B}_\varrho(w(0))$ we have

$$\frac{\mathrm{d}}{\mathrm{d}\,t}U(\alpha V_{w(t)}) \leq \frac{\gamma-1}{2\kappa^2}U(\alpha V_{w(t)})\,, \qquad \text{for } t > 0\,. \tag{18}$$

To obtain the above result we apply the chain rule:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\,t}U(\alpha V_{w(t)}) &= \langle \partial_f U(\alpha V_{w(t)})\,,\, \frac{\mathrm{d}}{\mathrm{d}\,t}\alpha V_{w(t)}\rangle_0 \\ &= \alpha\langle \alpha V_{w(t)} - V^*\,,\, DV_{w(t)} \cdot \frac{\mathrm{d}}{\mathrm{d}\,t}w(t)\rangle_0 \\ &= \langle \alpha V_{w(t)} - V^*\,,\, DV_{w(t)} \cdot DV_{w(t)}^\top \Gamma(T^\lambda \alpha V_{w(t)} - \alpha V_{w(t)})\rangle_0\,. \end{aligned} \tag{19}$$

Throughout, we define $\tau_\varrho := \inf\{t < 0 \,:\, w(t) \notin \mathcal{B}_\varrho(w(0))\}$, $g_w := (DV_w \cdot DV_w^\top)^{-1}$ (recalling that the $DV_w \cdot DV_w^\top$ has full rank in $\mathcal{B}_\varrho(w(0))$) and write $g_0 = (\mathbb{1} + \tilde{g}_w)g_w$, where $\tilde{g}_w$ is defined in Lemma 2. Then, as long as $t < \tau_\varrho$ we have, for every $a, b \in \mathcal{F}$

$$\langle a, DV_{w(t)} \cdot DV_{w(t)}^\top \Gamma b\rangle_0 = \langle a, (\mathbb{1} + \tilde{g}_{w(t)})\Gamma b\rangle \leq \langle a, b\rangle_\mu + \|\tilde{g}_{w(t)}\|\|a\|_\mu\|b\|_\mu\,.$$

By the above result we can bound from above the RHS of (19) by

$$\frac{\mathrm{d}}{\mathrm{d}\,t}U(\alpha V_{w(t)}) \leq \langle \alpha V_{w(t)} - V^*, T^\lambda \alpha V_{w(t)} - \alpha V_{w(t)}\rangle_\mu + \|\tilde{g}_{w(t)}\|\|\alpha V_{w(t)} - V^*\|_\mu \|T^\lambda \alpha V_{w(t)} - \alpha V_{w(t)}\|_\mu\,. \tag{20}$$

Recalling that by Lemma 3 we have

$$\|T^\lambda \alpha V_{w(t)} - \alpha V_{w(t)}\|_\mu = \|T^\lambda \alpha V_{w(t)} - V^*\|_\mu + \|\alpha V_{w(t)} - V^*\|_\mu \leq 2\|\alpha V_{w(t)} - V^*\|_\mu\,, \tag{21}$$

and applying Lemma 2, we can bound the second term of (20) from above as

$$\|\tilde{g}_{w(t)}\|\|\alpha V_{w(t)} - V^*\|_\mu \|T^\lambda \alpha V_{w(t)} - \alpha V_{w(t)}\|_\mu \leq \frac{1-\gamma}{2}\|\alpha V_{w(t)} - V^*\|_\mu^2\,. \tag{22}$$

On the other hand, for the first term we have by Cauchy-Schwartz inequality and (15) that

$$\begin{aligned}\langle \alpha V_{w(t)} - V^*, T^\lambda \alpha V_{w(t)} - \alpha V_{w(t)} \rangle_\mu &= \langle \alpha V_{w(t)} - V^* \,,\, (T^\lambda \alpha V_{w(t)} - V^*) - (\alpha V_{w(t)} - V^*) \rangle_\mu \,, \\ &\leq \|\alpha V_{w(t)} - V^*\|_\mu \|T^\lambda \alpha V_{w(t)} - V^*\|_\mu - \|\alpha V_{w(t)} - V^*\|_\mu^2 \\ &\leq (\gamma - 1)\|\alpha V_{w(t)} - V^*\|_\mu^2 \,, \end{aligned} \tag{23}$$

where γ is the contraction rate of the TD difference in $\mathcal{F}$, see (15). Finally, combining (22) and (23) we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\, t} U(\alpha V_{w(t)}) \leq \frac{\gamma - 1}{2} \|\alpha V_{w(t)} - V^*\|_\mu^2 \leq \frac{\gamma - 1}{2\kappa^2} \|\alpha V_{w(t)} - V^*\|_0^2 \,, \tag{24}$$

and the last inequality results from the equivalence of norms $\|\cdot\|_0$ and $\|\cdot\|_\mu$ (both have full support on a finite set). The desired result (17) follows directly from the above by Grönwall's inequality for all $t < \tau_\varrho$.

It now only remains to show that under the given choice of α, we have $\tau_\varrho = \infty$. By the contraction of T^λ Lemma 3 and our choice of $\varrho < \sigma_{\min}/(2L_{DV})$ we write

$$\left\| \frac{\mathrm{d}}{\mathrm{d}\, t} w(t) \right\|_2 \leq \frac{1}{\alpha} \|DV_{w(t)}\| \|T^\lambda \alpha V_{w(t)} - \alpha V_{w(t)}\|_\mu \leq \frac{2}{\alpha} \|DV_{w(0)}\| \|\alpha V_{w(t)} - V^*\|_\mu \,.$$

Integrating the above and combining with the result from (24) in the previous paragraph we have

$$\begin{aligned}\|w(t) - w(0)\|_2 &\leq \frac{2}{\alpha} \|DV_{w(0)}\| \|\alpha V_{w(0)} - V^*\|_0 \int_0^t \exp\left[\frac{\gamma - 1}{2\kappa^2} s\right] \mathrm{d}s \\ &\leq \frac{4\kappa^2}{\alpha(1-\gamma)} \|DV_{w(0)}\| \|\alpha V_{w(0)} - V^*\|_0 \,. \end{aligned} \tag{25}$$

Given that $\|\alpha V_{w(0)} - V^*\|_0 \leq 2\alpha M$, the above quantity is bounded by ϱ and therefore $\tau_\varrho = \infty$, as desired.

Finally, from (25) we see that if $\|V_{w(0)}\|_0 \leq C\alpha^{-1}$ then $\|w(t) - w(0)\|_2 \leq \frac{4\kappa^2}{\alpha(1-\gamma)} \|DV_{w(0)}\|(C + M\alpha_0) = \mathcal{O}(\alpha^{-1})$ for all $t > 0$. ∎

A.2. Under-parametrized regime

Lemma 6 *There exists $\delta > 0$ and $\alpha_0 > 0$ such that the ball $\mathcal{B}_\delta^0(0) \subseteq \overline{\mathcal{F}}_0$ is forward invariant and forward complete* WRT *the dynamics of (8) for all $\alpha > \alpha_0$.*

Proof of Lemma 6 We define the Lyapunov function $\bar{U}(f) := \frac{1}{2}\|f\|_0^2$, whose sublevel sets are $\mathcal{B}_\delta^0(0)$. We prove forward invariance of such sets by showing that, on their boundary (*i.e.*, on the sphere $S_\delta^{r-1} \subset \bar{\mathcal{F}}_0$ of radius δ), $\bar{U}(f)$ decreases along trajectories of (8) for α large enough.

Noting that $S_\delta^{r-1} \subset \overline{\mathcal{F}}_0$ upon taking δ small enough, we differentiate $\bar{U}(\overline{V}_{w(t)})$ WRT time for $w(t)$ obeying (8) at points $\overline{V} := \overline{V}_{w(t)} \in S_\delta^{r-1}$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \bar{U}(\overline{V}) &= \frac{1}{\alpha} \langle \overline{V}, \bar{g}_{w(t)}^{-1} D\psi_{\overline{V}}^{-1} \Gamma(T^\lambda \alpha \psi^{-1}(\overline{V}) - \alpha \psi^{-1}(\overline{V})) \rangle_0 \\ &= \frac{1}{\alpha} \langle \overline{V}, (D\psi_{\overline{V}}^{-1})^\top \Gamma(T^\lambda \alpha \psi^{-1}(\overline{V}) - \alpha \psi^{-1}(\overline{V})) \rangle + R_g(\overline{V}) \end{aligned}$$

$$= \frac{1}{\alpha}\langle D\psi_{\overline{V}}^{-1}\overline{V}, \bar{r}^\lambda + \alpha(\gamma P^\lambda - \mathbb{1})\psi^{-1}(\overline{V})\rangle_\mu + R_g(\overline{V})$$
$$\leq \langle D\psi_{\overline{V}}^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})\psi^{-1}(\overline{V})\rangle_\mu + \frac{1}{\alpha}\|D\psi_{\overline{V}}^{-1}\overline{V}\|_\mu\|\bar{r}^\lambda\|_\mu + |R_g(\overline{V})|\,. \tag{26}$$

where we have defined $R_g(\overline{V}) := \frac{1}{\alpha}\langle \overline{V}, \tilde{g}_{w(t)}(D\psi_{\overline{V}}^{-1})^\top\Gamma(T^\lambda\alpha\psi^{-1}(\overline{V}) - \alpha\psi^{-1}(\overline{V}))\rangle$ for $\tilde{g}_w$ from Lemma 2. We now proceed to bound the last two terms on the RHS from above. The second term is of order α^{-1} and therefore goes to 0 for $\alpha \to \infty$ while for the last one we have that, by the equivalence of the norms $\|\cdot\|_\mu$ and $\|\cdot\|_2$,

$$|R_g(\overline{V})| \leq \frac{1}{\alpha}\|\overline{V}\|_2\|\tilde{g}_{w(t)}\|\|(D\psi_{\overline{V}}^{-1})^\top\Gamma\left[\bar{r}^\lambda + (\gamma P^\lambda - \mathbb{1})\alpha\psi^{-1}(\overline{V})\right]\|_2$$
$$\leq \frac{1}{\alpha}\|\overline{V}\|_2\|\tilde{g}_{w(t)}\|\|(D\psi_{\overline{V}}^{-1})^\top\Gamma\bar{r}^\lambda\| + \|\overline{V}\|_2\|\tilde{g}_{w(t)}\|\|(D\psi_{\overline{V}}^{-1})^\top\Gamma(\gamma P^\lambda - \mathbb{1})\psi^{-1}(\overline{V})\|_2$$
$$\leq \alpha^{-1}C + \varepsilon_R(\delta)\|\overline{V}\|_\mu^2\,. \tag{27}$$

for a constant C bounded by the norm of all operators and, by Lemma 2 a positive function $\varepsilon_R(\delta)$ with $\lim_{\delta\to 0}\varepsilon_R(\delta) = 0$. By the bounds established above and the fact that $\|\overline{V}\|_\mu \geq \kappa^{-1}\delta$ for $\overline{V} \in S_\delta^{r-1} \subset \overline{\mathcal{F}}_0$ it is sufficient to show that the first term in (26) satisfies

$$\langle D\psi_{\overline{V}}^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})\psi^{-1}(\overline{V})\rangle_\mu \leq -\varepsilon\|\overline{V}\|_\mu^2\,, \tag{28}$$

for δ small enough and a constant $\varepsilon > 0$ independent of δ. We Taylor-expand ψ^{-1} around the origin, denoting the second order remainder of that expansion by $R_2(\,\cdot\,,\,\cdot\,)$, and since $\psi^{-1}(\overline{V}_0) = 0$ we have,

$$\langle D\psi_{\overline{V}}^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})\psi^{-1}(\overline{V})\rangle_\mu = \langle D\psi_{\overline{V}}^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})D\psi_0^{-1}\overline{V}\rangle_\mu$$
$$+ \langle D\psi_{\overline{V}}^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})R_2(\overline{V}, \overline{V})\rangle_\mu\,, \tag{29}$$

where we have introduced the short hand notation $D\psi_0^{-1} = D\psi_{\overline{V}_0}^{-1}$. By the Lipschitz smoothness of $\psi^{-1}(\,\cdot\,)$ (Lee, 2003) we can bound the norm of the second term from above as

$$\langle D\psi_{\overline{V}}^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})R_2(\overline{V}, \overline{V})\rangle_\mu \leq 2\|D\psi_{\overline{V}}^{-1}\overline{V}\|_\mu\|R_2(\overline{V}, \overline{V})\|_\mu \leq 2L_{D\psi^{-1}}\|D\psi_{\overline{V}}^{-1}\|\|\overline{V}\|_\mu^3\,. \tag{30}$$

For the first term in (29) we can also expand $D\psi_{\overline{V}}^{-1} = D\psi_0^{-1} + \tilde{R}_2(\overline{V}, \cdot\,)$, and by applying a similar bound as (30) we obtain that

$$\langle D\psi_{\overline{V}}^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})D\psi_0^{-1}\overline{V}\rangle_\mu \leq \langle D\psi_0^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})D\psi_0^{-1}\overline{V}\rangle_\mu + 2L_{D\psi^{-1}}\|D\psi_0^{-1}\|\|\overline{V}\|_\mu^3\,. \tag{31}$$

The second term of the above equation being $\mathcal{O}(\|\overline{V}\|^3)$, we now consider the first one. By the nonexpansion of P in $\|\cdot\|_\mu$ proven in Lemma 3 we have

$$\langle D\psi_0^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})D\psi_0^{-1}\overline{V}\rangle_\mu \leq \gamma\|D\psi_0^{-1}\overline{V}\|_\mu\|P^\lambda D\psi_0^{-1}\overline{V}\|_\mu - \|D\psi_0^{-1}\overline{V}\|_\mu^2$$
$$\leq (\gamma - 1)\|D\psi_0^{-1}\overline{V}\|_\mu^2 \leq (\gamma - 1)(\sigma_{\min}^{D\psi^{-1}})^2\|\overline{V}\|_\mu^2\,, \tag{32}$$

where $\sigma_{\min}^{D\psi^{-1}}$ denotes the smallest singular value of $D\psi^{-1}$ in $\mathcal{B}_\delta^0(0)$. Combining (30), (31) and (32) we finally obtain

$$\langle D\psi_{\overline{V}}^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})\psi^{-1}(\overline{V})\rangle_\mu \leq \|\overline{V}\|_\mu^2((\gamma - 1)(\sigma_{\min}^{D\psi^{-1}})^2 + C'\kappa^{-1}\|\overline{V}\|_0)\,, \tag{33}$$

for $C' = 2L_{D\psi^{-1}}(\|D\psi_0^{-1}\| + \|D\psi_{\overline{V}}^{-1}\|)$ and recalling that κ is the equivalence constant between the norms $\|\cdot\|_\mu$ and $\|\cdot\|_0$ in $\overline{\mathcal{F}}_0$. [2] Therefore, choosing δ small enough we obtain (28) and conclude the proof of forward invariance.

By boundedness of $\mathcal{B}_\delta^0(0)$ in $\overline{\mathcal{F}}_0$, forward completeness follows directly from forward invariance. ■

Lemma 7 *There exists $\ell > 0$, $\delta > 0$ and $\alpha_0 > 0$ such that for all $\alpha > \alpha_0$ and all geodesics $\gamma(s)$ contained in the ball $\mathcal{B}_\delta^0(0) \subseteq \overline{\mathcal{F}}_0$, the function*

$$\langle \gamma'(s), X(\gamma(s))\rangle_0 - \ell s \langle \gamma'(0), \gamma'(0)\rangle_0 \,,$$

is strictly decreasing in s.

Proof of Lemma 7 To simplify the notation and the forthcoming computation, we prove the differential version of the desired result, *i.e.*, we show that there exists $\ell > 0$ such that

$$\frac{\mathrm{d}}{\mathrm{d}\, s}\left[\langle \gamma'(s), X(\gamma(s))\rangle_0 - \ell s \langle \gamma'(0), \gamma'(0)\rangle_0\right] < 0\,. \tag{34}$$

The above expression exists almost everywhere by Lipschitz continuity of the terms to be differentiated. When this is not the case, we must interpret this derivative in the sense of distributions. We will highlight the steps where this could be necessary as we go along the proof.

In our case, X is the RHS of (12) mapped through ψ onto $\overline{\mathcal{F}}_0$, *i.e.*,

$$X(\gamma(s)) = -\frac{1}{\alpha}\bar{g}^{-1}_{\gamma(s)}(D\psi^{-1}_{\gamma(s)})^\top \Gamma(T^\lambda \alpha\psi^{-1}(\gamma(s)) - \alpha\psi^{-1}(\gamma(s)))\,.$$

We are going to consider the "flattened" manifold obtained by the maps ϕ and ψ equipped with the metric $\bar{g}_0$. In this space, geodesics have the form $\gamma(s) = v_1 + s\Delta v$ where $\Delta v := v_2 - v_1$ for $v_1, v_2 \in \overline{\mathcal{F}}_0$ and their derivative is $\gamma'(s) = \Delta v$. Consequently (34) reads

$$\langle \Delta v, \frac{\mathrm{d}}{\mathrm{d}s} X(\gamma(s))\rangle_0 < \ell \|\Delta v\|_0^2\,, \tag{35}$$

where defining $\tilde{g}_{\gamma(s)} := \bar{g}_0 \bar{g}^{-1}_{\gamma(s)} - \mathbb{1}$ as in Lemma 2 we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s} X(\gamma(s)) &= \frac{\mathrm{d}}{\mathrm{d}s}\bar{g}_0\bar{g}^{-1}_{\gamma(s)}(D\psi^{-1}_{\gamma(s)})^\top \Gamma(T^\lambda(\alpha\psi^{-1}(\gamma(s))) - \alpha\psi^{-1}(\gamma(s))) \\ &= \frac{\mathrm{d}}{\mathrm{d}s}\bar{X}(\gamma(s)) + \tilde{g}_{\gamma(s)}\frac{\mathrm{d}}{\mathrm{d}s}\bar{X}(\gamma(s)) + D\tilde{g}_{\gamma(s)}(\bar{X}(\gamma(s)), \gamma'(s))\,. \end{aligned} \tag{36}$$

for

$$\bar{X}(\gamma(s)) := (D\psi^{-1}_{\gamma(s)})^\top \Gamma(T^\lambda(\alpha\psi^{-1}(\gamma(s))) - \alpha\psi^{-1}(\gamma(s)))\,.$$

2. We recall that by the construction of the mappings ψ, ϕ, π_r and by our assumption in Theorem 5 the metric tensor $\bar{g}_t$ has full rank on $\mathcal{F}_0$ and being the latter set compact its eigenvalues are uniformly bounded from below. At the same time, we can equip $\overline{\mathcal{F}}_0$ with the metric induced by Γ by restricting it to its first r elements, which are uniformly bounded from below. Hence, the two metrics are equivalent on this space for some equivalence constant κ.

We proceed by analyzing the first term in the above equation and leave the task of bounding the last two for later. Using $\partial_s \alpha\psi^{-1}(\gamma(s)) = \alpha D\psi^{-1}_{\gamma(s)}\gamma'(s) = \alpha D\psi^{-1}_{\gamma(s)}\Delta v$ we have that

$$\frac{\mathrm{d}}{\mathrm{d}s}\bar{X}(\gamma(s)) = \frac{1}{\alpha}(D^2\psi^{-1}_{\gamma(s)})^\top(\Gamma(T^\lambda\alpha\psi^{-1}(\gamma(s)) - \alpha\psi^{-1}(\gamma(s))), \Delta v) \\ + (D\psi^{-1}_{\gamma(s)})^\top\Gamma\big[DT^\lambda D\psi^{-1}_{\gamma(s)}\Delta v - D\psi^{-1}_{\gamma(s)}\Delta v\big]\,, \tag{37}$$

where $(D^2\psi^{-1}_{\gamma(s)})^\top$ denotes the inversion of the last two indices of the Hessian. We now proceed to consider the two terms in the sum above separately (multiplied by the scalar product of (35)), defining throughout $(TD)_s := \Gamma(T^\lambda\alpha\psi^{-1}(\gamma(s)) - \alpha\psi^{-1}(\gamma(s)))$. For the first term we have:

$$\frac{1}{\alpha}\langle\Delta v, D^2\psi^{-1}_{\gamma(s)}(TD_s, \Delta v)\rangle_0 \le \|\Delta v\|_0^2\|D^2\psi^{-1}_{\gamma(s)}\left(\alpha^{-1}\bar{r}^\lambda + (\gamma P^\lambda - \mathbb{1})\psi^{-1}\gamma(s)\right)\| \le \varepsilon'\|\Delta v\|_0^2\,, \tag{38}$$

for any $\varepsilon' > 0$ by using the linearity of the Hessian and bounding its operator norm of ψ^{-1} on a compact space in $\mathcal{F}_0$ while choosing α large enough and δ small enough, since $\gamma(s) \in \mathcal{B}^0_\delta(0)$. Note that if $D\psi^{-1}$ is not differentiable, the above computation is to be understood in the sense of distributions.

We now focus on the second term of (37). In this case we incorporate the operator Γ in the inner product and write this term as

$$\langle D\psi^{-1}_{\gamma(s)}\Delta v, DT^\lambda D\psi^{-1}_{\gamma(s)}\Delta v\rangle_\mu - \|D\psi^{-1}_{\gamma(s)}\Delta v\|_\mu^2\,.$$

Now, by the contraction property of T^λ onto the tangential space $\mathcal{T}_{\psi^{-1}_{\gamma(s)}}\mathcal{F}$ in the norm $\|\cdot\|_\mu$ we can write

$$\langle D\psi^{-1}_{\gamma(s)}\Delta v, DT^\lambda D\psi^{-1}_{\gamma(s)}\Delta v\rangle_\mu \le \|D\psi^{-1}_{\gamma(s)}\Delta v\|_\mu\|P^\lambda D\psi^{-1}_{\gamma(s)}\Delta v\|_\mu \le \gamma\|D\psi^{-1}_{\gamma(s)}\Delta v\|_\mu^2\,,$$

so that

$$\langle D\psi^{-1}_{\gamma(s)}\Delta v, DT^\lambda D\psi^{-1}_{\gamma(s)}\Delta v\rangle_\mu - \|D\psi^{-1}_{\gamma(s)}\Delta v\|_\mu^2 \le (\gamma - 1)\|D\psi^{-1}_{\gamma(s)}\Delta v\|_\mu^2\,. \tag{39}$$

Denoting by $\sigma^{D\psi^{-1}}_{\max}, \sigma^{D\psi^{-1}}_{\min}$ the largest and smallest, respectively, singular values of the map $D\psi^{-1}$ in $\mathcal{B}^0_\delta(0)$ (which are bounded away from 0 upon possibly making this set smaller), by nondegeneracy of $D\psi^{-1}$ and by the equivalence of the $\|\cdot\|_\mu$ and $\|\cdot\|_0$ norms on $\overline{\mathcal{F}}_0$ we have that

$$\kappa^{-1}\sigma^{D\psi^{-1}}_{\min}\|\Delta v\|_0 \le \|\Delta v\|_\mu\sigma^{D\psi^{-1}}_{\min} \le \|D\psi^{-1}_{\gamma(s)}\Delta v\|_\mu \le \|\Delta v\|_\mu\sigma^{D\psi^{-1}}_{\max} \le \kappa\|\Delta v\|_0\sigma^{D\psi^{-1}}_{\max}\,.$$

Thus we have

$$\|D\psi^{-1}_{\gamma(s)}\Delta v\|_\mu^2 \ge \kappa^{-2}\left(\sigma^{D\psi^{-1}}_{\min}\right)^2\|\Delta v\|_0^2\,. \tag{40}$$

Getting back to the last two terms in (26), we immediately see from Lemma 2 that $\tilde{g}_{\gamma(s)}$ is a small, Lipschitz continuous perturbation. Hence, the product

$$\langle\gamma'(s), \tilde{g}_{\gamma(s)}\bar{X}'(\gamma(s))\rangle$$

can be bounded from above similarly to (27), while the second order derivative in the third term of (36) can be dealt with analogously to what is done in (38), giving terms $\varepsilon''\|\Delta v\|_0^2$ and $\varepsilon^{(3)}\|\Delta v\|_0^2$ respectively, both going to 0 as $\delta \to 0$.

Therefore, combining the above with (38), (39) and (40) we have

$$\langle \Delta v, \frac{\mathrm{d}}{\mathrm{d}t}\bar{X}(\gamma(s))\rangle_0 \leq \frac{\gamma-1}{\kappa^2}\left(\sigma_{\min}^{D\psi^{-1}}\right)^2\|\Delta v\|_0^2 + \left(\sum_i^3 \varepsilon^{(i)}(\delta)\right)\|\Delta v\|_0^2 \\ \leq \frac{\gamma-1}{2\kappa^2}\left(\sigma_{\min}^{D\psi^{-1}}\right)^2\|\Delta v\|_0^2 .$$

This directly gives (35) by choosing ℓ large enough. ■

The next lemma estimates the distance between the fixed point $\tilde{V}^*$ of the dynamics (8) and V^* given by (1), showing that it is close, for large values of α to the best linear model in the tangent space of $\mathcal{F}_{\mathcal{W}}$ at $V_{w(0)}$, given by $\Pi_0 V^*$. We recall that the projection operator Π_0 onto the linear space spanned by the columns of DV is given by (Tsitsiklis and Van Roy, 1997, Eq. (1))

$$\Pi_0 W := \underset{\{DV_{w(0)}\Delta w\,:\,\Delta w\in\mathbb{R}^p\}}{\arg\min} \|DV_{w(0)}\Delta w - W\|_\mu = DV_{w(0)}(DV_{w(0)}^\top \Gamma DV_{w(0)})^{-1}DV_{w(0)}^\top \Gamma W ,$$

for all $W \in \mathcal{F}$ where, if necessary, we interpret $(DV_{w(0)}^\top \Gamma DV_{w(0)})^{-1}$ as a pseudo-inverse.

Lemma 12 *Let $\tilde{V}^*$ be the fixed point of (8) and V^* be the global fixed point of the* TD *operator, given by (1). Then under the assumptions of Theorem 5 there exists constants $\alpha_0 > 0$ and $C^* > 0$ (independent of α_0), such that*

$$\|\tilde{V}^* - V^*\|_\mu < \frac{1-\lambda\gamma}{1-\gamma}\|\Pi_0 V^* - V^*\|_\mu + C^*\alpha^{-1} , \tag{41}$$

where Π_0 is the projection operator onto $\mathcal{T}_{V(w(0))}\mathcal{F}_{\mathcal{W}}$.

To prove the above result we compare the dynamics (8) to the dynamics of the model V when *linearized* at $w(0)$. In this case, the dynamics of the parameters is given by

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\bar{w}(t) = DV_{w(0)}^\top \Gamma(T^\lambda \mathcal{V}_{\bar{w}(t)} - \mathcal{V}_{\bar{w}(t)}) , \tag{42}$$

where $\mathcal{V} \in \mathcal{F}$ is the linear, tangent model of V at $w(0)$ defined as

$$\mathcal{V}_w := V_{w(0)} + DV_{w(0)}(w - w(0)) . \tag{43}$$

We can also write the dynamics of the linear model as

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\mathcal{V}_{\bar{w}(t)} := DV_{w(0)} \cdot DV_{w(0)}^\top \Gamma(T^\lambda \mathcal{V}_{\bar{w}(t)} - \mathcal{V}_{\bar{w}(t)}) . \tag{44}$$

Scaling the model as $\mathcal{V} \to \alpha\mathcal{V}$ and $t \to \alpha^{-1}t$ we obtain the analogue of (8):

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\bar{w}(t) := \frac{1}{\alpha}DV_{w(0)}^\top \Gamma(T^\lambda \alpha\mathcal{V}_{\bar{w}(t)} - \alpha\mathcal{V}_{\bar{w}(t)}) . \tag{45}$$

which in $\mathcal{F}$ reads

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\alpha\mathcal{V}_{\bar{w}(t)} := DV_{w(0)} \cdot DV_{w(0)}^\top \Gamma(T^\lambda \alpha\mathcal{V}_{\bar{w}(t)} - \alpha\mathcal{V}_{\bar{w}(t)}) .$$

Proof of Lemma 12 Recall from (Tsitsiklis and Van Roy, 1997, Lemma 6) that for the linear value function approximation one has

$$\|\mathcal{V}^* - V^*\|_\mu < \frac{1-\lambda\gamma}{1-\gamma}\|\Pi_0 V^* - V^*\|_\mu\,, \tag{46}$$

where Π_0 is the projection on $\mathcal{T}_{V(w(0))}\mathcal{F}_\mathcal{W}$ and $\mathcal{V}^*$ is the unique fixed point of the dynamics (44) on that space. In light of this result, our task reduces to bounding the distance between the trajectories of the original (*i.e.*, dynamics (8)) and the linearized model (*i.e.*, dynamics (45)) by $C\alpha^{-1}$ for C large enough. We do so in 3 main steps. First of all, we bound the maximal excursion of the models $\mathcal{V}$ and V. Mapping both dynamics onto a common coordinate space, we then bound from above the distance between the two trajectories in this space by $\mathcal{O}(\alpha^{-1})$. Finally, we map the dynamics back to the embedding space and show that the correction is again of the same order $\mathcal{O}(\alpha^{-1})$.

Bounding the maximal excursion. To compare the dynamics of $\alpha V_{w(t)}$ and $\alpha\mathcal{V}_{\bar{w}(t)}$ we map them to a common space. Recalling the definition of the maps ϕ, π_r, ψ from the proof of Theorem 5 we note that the first order expansion of ψ, maps $\mathcal{T}_{V(w(0))}\mathcal{F}_\mathcal{W}$ to $\overline{\mathcal{F}}_0$. Explicitly, for $\overline{V} \in \overline{\mathcal{F}}_0$ and for $\Delta\mathcal{V} \in \mathcal{T}_{V(w(0))}\mathcal{F}_\mathcal{W}$ with $\|\Delta\mathcal{V}\|_0$ small enough we have

$$\bar{\psi}(V_{w(0)} + \Delta\mathcal{V}) := D\psi_0\Delta\mathcal{V} \qquad \text{and} \qquad \bar{\psi}^{-1}(\overline{V}) = V_{w(0)} + D\psi_0^{-1}\overline{V} \in \mathcal{T}_{V(w(0))}\mathcal{F}_\mathcal{W}\,. \tag{47}$$

Now, we proceed to show that the dynamics of (8) and (45), mapped to $\mathcal{F}_0$, do not exit a ball $\mathcal{B}_\delta^0(0)$, when choosing $\delta = C/\alpha$ for C large enough. We show this with the same strategy used for the proof of Lemma 6, *i.e.*, we show that $\bar{U}(f) := \frac{1}{2}\|f\|_0^2$ decreases on $S_\delta^{r-1}(0)$ along the trajectories of interest (note that δ is now much smaller than that used in Lemma 6). We will start with the curved dynamics (8) and will then show that the same result follows, in a simpler setting, for (45). For $\overline{V} := \overline{V}_{w(t)} \in S_\delta^{r-1}(0)$ we start by bounding, as in (26), the derivative

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\bar{U}(\overline{V}) \le \langle D\psi_{\overline{V}}^{-1}\overline{V}, (\gamma P^\lambda - \mathbb{1})\psi^{-1}(\overline{V})\rangle_\mu + \frac{1}{\alpha}\|D\psi_{\overline{V}}^{-1}\overline{V}\|_\mu\|\bar{r}^\lambda\|_\mu + |R_g(\overline{V})|\,. \tag{48}$$

Before bounding the above terms we recall that by Lipschitz smoothness of ψ we have that

$$\|\psi^{-1}(\overline{V})\| < \|V_{w(0)}\| + \|D\psi_0^{-1}\overline{V}\| + L_{D\psi^{-1}}\|\overline{V}\|^2\,. \tag{49}$$

Then, since $V_{w(0)} = 0$, similarly to (26) we have for the last term in (48) that, for α large enough,

$$|R_g(\overline{V})| \le \|\tilde{g}_w\|\|\overline{V}\|_2\Big(\|\overline{V}\|_2\|(D\psi_{\overline{V}}^{-1})^\top\Gamma(\gamma P^\lambda - \mathbb{1})\|(\|D\psi_0^{-1}\| + L_{DV}\|\overline{V}\|_2) + \frac{1}{\alpha}\|(D\psi_{\overline{V}}^{-1})^\top\Gamma\bar{r}^\lambda\|_2\Big)\,.$$

By the equivalence of the norms $\|\cdot\|_\mu$, $\|\cdot\|_2$ and $\|\cdot\|_0$ on Π_r and since $\delta = C/\alpha$ we have that

$$|R_g(\overline{V})| \le \|\tilde{g}_w\|\|\overline{V}\|_0^2(K+1) + \mathcal{O}(\alpha^{-3})\,, \tag{50}$$

upon increasing C if necessary and defining $K = \kappa_2^2\|(D\psi_{\overline{V}}^{-1})^\top\Gamma(\gamma P^\lambda - \mathbb{1})\|\|D\psi_0^{-1}\|$ for κ_2 the equivalence constant between $\|\cdot\|_2$ and $\|\cdot\|_0$ on Π_r. The second term in (48) can be bounded similarly to the above by the equivalence of norms:

$$\frac{1}{\alpha}\|D\psi_{\overline{V}}^{-1}\overline{V}\|_\mu\|\bar{r}^\lambda\|_\mu \le \|\overline{V}\|_0^2\frac{\kappa^2\|D\psi_{\overline{V}}^{-1}\|\|\bar{r}^\lambda\|_\mu}{C}\,. \tag{51}$$

The first term in (48) can be treated identically to the proof of Lemma 6 to obtain (33). Changing the norm in (33) and combining it with (50) and (51) gives

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\bar{U}(\overline{V}) \leq \|\overline{V}\|_0^2 \left(\frac{\gamma-1}{2\kappa^2}(\sigma_{\min}^{D\psi^{-1}})^2 + \frac{\kappa^2 \|D\psi_{\overline{V}}^{-1}\| \|\bar{r}^\lambda\|_\mu}{C} + \|\tilde{g}_w\|(K+1) \right) + \mathcal{O}(\alpha^{-3})\,.$$

Since $\gamma - 1 < 0$, we can choose C large enough to make the second term in brackets smaller than $(\gamma-1)/12\kappa^2(\sigma_{\min}^{D\psi^{-1}})^2$. The same holds for the third term in brackets by (11), and for the higher order term by taking α large enough, showing that

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\bar{U}(\overline{V}) \leq \frac{\gamma-1}{4\kappa^2}(\sigma_{\min}^{D\psi^{-1}})^2 \|\overline{V}\|_0^2 < 0\,,$$

as desired. We note that the same reasoning with $L_{DV} = 0$ and $D\psi_{\overline{V}}^{-1} \equiv D\psi_0^{-1}$ yields an identical conclusion for the dynamics of $\mathcal{V}$ in a ball of radius $\delta = C/\alpha$ for C, α large enough. Also, we note that combining the above computation with (30) yields

$$\begin{aligned}
\|D\psi_{\overline{V}}^{-1}\Gamma(T^\lambda \alpha\psi^{-1}(\overline{V}) - \alpha\psi^{-1}(\overline{V}))\| &\leq \|D\psi_{\overline{V}}^{-1}\Gamma\|(\|\bar{r}^\lambda\| + \alpha(\gamma+1)\|D\psi_0^{-1}\overline{V}\| + \alpha L_{D\psi_{\overline{V}}^{-1}}\|\overline{V}\|^2) \\
&\leq (\gamma+1)\|D\psi_{\overline{V}}^{-1}\Gamma\|(\|D\psi_0^{-1}\|C + \|\bar{r}^\lambda\| + \mathcal{O}(\alpha^{-1})) \\
&\leq C_0\,, \qquad (52)
\end{aligned}$$

for C_0 large enough, where $D\psi_{\overline{V}}^{-1}\Gamma$ is considered as an operator mapping $\mathcal{F}_0 \to \bar{\mathcal{F}}_0$.

Bounding the distance of trajectories. The distance between two trajectories with the same initial condition can be bounded by $\mathcal{O}(\alpha^{-2})$ using a similar argument as in (Chizat et al., 2019, Lemma B2) for the present framework. We include the proof of this lemma here as the assumptions are not identical and to make the paper self-contained, while we do not claim any improvement on that result. To enounce this result, we recall that $\sigma_{\min}^{D\psi^{-1}}$ denotes the smallest singular eigenvalue of $D\psi^{-1}$ in a ball $\mathcal{B}_\delta^0(0)$, which is bounded away from 0 for δ small enough. Similarly, we recall that $\bar{g}_t^{-1} \succeq \sigma_{\min}^g \mathbb{1}$ for $\sigma_{\min}^g > 0$ in $\mathcal{B}_\delta^0(0)$ for δ small enough.

Lemma 13 *Let $\overline{V}_t$, $\overline{\mathcal{V}}_t$ in $\overline{\mathcal{F}}_0$ be solutions of*

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\,t}\overline{V}_t &= \bar{g}_t^{-1}(D\psi_{\overline{V}_t}^{-1})^\top \Gamma(T^\lambda \alpha\psi^{-1}(\overline{V}_t) - \alpha\psi^{-1}(\overline{V}_t))\,, \\
\frac{\mathrm{d}}{\mathrm{d}\,t}\overline{\mathcal{V}}_t &= \bar{g}_0^{-1}(D\psi_0^{-1})^\top \Gamma(T^\lambda \alpha\bar{\psi}^{-1}(\overline{\mathcal{V}}_t) - \alpha\bar{\psi}^{-1}(\overline{\mathcal{V}}_t))\,.
\end{aligned}$$

Then defining $K := \sup_{t>0} \|(\bar{g}_t^{-1} - \bar{g}_0^{-1})(D\psi_{\overline{V}_t}^{-1})^\top \Gamma(T^\lambda \alpha\psi^{-1}(\overline{V}_t) - \alpha\psi^{-1}(\overline{V}_t))\|$ and $\beta := \frac{1-\gamma}{\kappa^2}(\sigma_{min}^{D\psi^{-1}})^2$ we have that

$$\sup_{t>0} \|\overline{V}_t - \overline{\mathcal{V}}_t\|_0 \leq \frac{1}{\alpha}\frac{2K}{\beta}\,.$$

Proof of Lemma 13 We define the function $h(t) := \frac{1}{2}\|\overline{V}_t - \overline{\mathcal{V}}_t\|_0^2$, take its time derivative

$$h'(t) = \langle \overline{V}_t' - \overline{\mathcal{V}}_t', \overline{V}_t - \overline{\mathcal{V}}_t \rangle_0\,,$$

and defining

$$(TD)_t := T^\lambda \alpha\psi^{-1}(\overline{V}_t) - \alpha\psi^{-1}(\overline{V}_t)\,,$$
$$(\mathcal{TD})_t := T^\lambda \alpha\bar{\psi}^{-1}(\overline{\mathcal{V}}_t) - \alpha\bar{\psi}^{-1}(\overline{\mathcal{V}}_t)\,,$$

we evaluate (for simplicity of notation, we introduce the short hand $D\psi_t^{-1} := D\psi_{\overline{V}_t}^{-1}$ for the rest of the proof)

$$\begin{aligned}\overline{V}'_t - \overline{\mathcal{V}}'_t &= \frac{1}{\alpha}\bar{g}_t^{-1}(D{\psi_t}^{-1})^\top\Gamma(TD)_t - \frac{1}{\alpha}\bar{g}_0^{-1}(D\psi_0^{-1})^\top\Gamma(\mathcal{TD})_t \\ &\le \frac{1}{\alpha}\left[\bar{g}_0^{-1}(D\psi_t^{-1})^\top\Gamma(TD)_t - \bar{g}_0^{-1}(D\psi_0^{-1})^\top\Gamma(\mathcal{TD})_t\right] \qquad (53)\\ &\quad + \frac{1}{\alpha}\left[\bar{g}_t^{-1}(D\psi_t^{-1})^\top\Gamma(TD)_t - \bar{g}_0^{-1}(D\psi_t^{-1})^\top\Gamma(TD)_t\right]. \qquad (54)\end{aligned}$$

We look at the two terms on the RHS separately and obtain, for (53)

$$\begin{aligned}&\frac{1}{\alpha}\langle\bar{g}_0^{-1}(D\psi_t^{-1})^\top\Gamma(TD)_t - \bar{g}_0^{-1}(D\psi_0^{-1})^\top\Gamma(\mathcal{TD})_t, \overline{V}_t - \overline{\mathcal{V}}_t\rangle_0 \qquad (55)\\ &\quad= \frac{1}{\alpha}\langle(D\psi_t^{-1})^\top\Gamma(TD)_t - (D\psi_0^{-1})^\top\Gamma(\mathcal{TD})_t, \overline{V}_t - \overline{\mathcal{V}}_t\rangle \\ &\quad= \frac{1}{\alpha}\langle(TD)_t - (\mathcal{TD})_t, D\psi_0^{-1}(\overline{V}_t - \overline{\mathcal{V}}_t)\rangle_\mu \qquad (56)\\ &\qquad+ \frac{1}{\alpha}\langle(D\psi_t^{-1} - D\psi_0^{-1})^\top\Gamma(TD)_t, \overline{V}_t - \overline{\mathcal{V}}_t\rangle\,. \qquad (57)\end{aligned}$$

We immediately see that by Lipschitz smoothness of ψ^{-1} and the equivalence of $\|\cdot\|_2$ and $\|\cdot\|_0$ norms on Π_r and (52), for (57) we have

$$\frac{1}{\alpha}\langle(D\psi_t^{-1} - D\psi_0^{-1})^\top\Gamma(TD)_t, \overline{V}_t - \overline{\mathcal{V}}_t\rangle \le \frac{1}{\alpha}L_{D\psi^{-1}}\|\overline{V}_t\|_2\|\Gamma(TD)_t\|\|\overline{V}_t - \overline{\mathcal{V}}_t\|_2 \le \frac{C_1}{\alpha^2}\sqrt{2h(t)}\,, \qquad (58)$$

by choosing C_1 large enough. For (56) by the definition of ψ we have

$$\begin{aligned}(TD)_t - (\mathcal{TD})_t &= T^\lambda\alpha\psi^{-1}(\overline{V}_t) - T^\lambda\alpha\bar{\psi}^{-1}(\overline{\mathcal{V}}_t) - \alpha(\psi^{-1}(\overline{V}_t) - \bar{\psi}^{-1}(\overline{\mathcal{V}}_t)) \\ &= \alpha(P^\lambda - \mathbb{1})(\psi^{-1}(\overline{V}_t) - \bar{\psi}^{-1}(\overline{\mathcal{V}}_t))\,,\end{aligned}$$

and hence, by (49) we have

$$\begin{aligned}\frac{1}{\alpha}\langle(TD)_t - (\mathcal{TD})_t, D\psi_0^{-1}(\overline{V}_t - \overline{\mathcal{V}}_t)\rangle_\mu &\le \langle(P^\lambda - \mathbb{1})(\psi^{-1}(\overline{V}_t) - \bar{\psi}^{-1}(\overline{\mathcal{V}}_t)), D\psi_0^{-1}(\overline{V}_t - \overline{\mathcal{V}}_t)\rangle_\mu \\ &\le \langle(P^\lambda - 1)D\psi_0^{-1}(\overline{V}_t - \overline{\mathcal{V}}_t), D\psi_0^{-1}(\overline{V}_t - \overline{\mathcal{V}}_t)\rangle_\mu \\ &\quad + L_{D\psi^{-1}}\|\overline{V}_t\|_\mu^2\|D\psi_0^{-1}\|\|\overline{V}_t - \overline{\mathcal{V}}_t\|_\mu\,.\end{aligned}$$

Defining $\beta := \frac{1-\gamma}{\kappa^2}(\sigma_{\min}^{D\psi^{-1}})^2$, the first term from above can be bounded as in (32) to obtain

$$\langle(P^\lambda - 1)D\psi_0^{-1}(\overline{V}_t - \overline{\mathcal{V}}_t), D\psi_0^{-1}(\overline{V}_t - \overline{\mathcal{V}}_t)\rangle_\mu \le -\beta h(t)\,, \qquad (59)$$

while for the second by our choice of $\delta = C/\alpha$ we have

$$L_{D\psi^{-1}}\|\overline{V}_t\|_\mu^2\|D\psi_0^{-1}\|\|\overline{V}_t - \overline{\mathcal{V}}_t\|_\mu \le \frac{C^2}{\alpha^2}\kappa L_{D\psi^{-1}}\|D\psi_0^{-1}\|\sqrt{2h(t)}\,. \qquad (60)$$

Finally, combining (58), (59) and (60) we have

$$(55) \leq -\beta h(t) + \frac{C_2}{\alpha^2}\sqrt{2h(t)}\,, \tag{61}$$

where $C_2 := C_1 + C^2\kappa L_{D\psi^{-1}}\|D\psi_0^{-1}\|$.

We now consider (54). Here by the definition of K we have

$$\frac{1}{\alpha}\langle(\bar{g}_t^{-1} - \bar{g}_0^{-1})D\psi_t^{-1}\Gamma(TD)_t, \overline{V}_t - \overline{\mathcal{V}}_t\rangle_0 \leq \frac{K}{\alpha}\|\overline{V}_t - \overline{\mathcal{V}}_t\|_0 = \frac{K}{\alpha}\sqrt{2h(t)}\,.$$

Combining the above with (61) we finally obtain

$$h'(t) \leq -\beta h(t) + \frac{K}{\alpha}\sqrt{2h(t)} + \frac{C_2}{\alpha^2}\sqrt{2h(t)} \leq -\beta h(t) + \frac{2K}{\alpha}\sqrt{h(t)}\,,$$

for α large enough. The above expression is negative as soon as $h(t) > 4K^2/(\alpha\beta)^2$. Therefore, because $h(0) = 0$, we must have that $h(t) \leq 4K^2/(\alpha\beta)^2$ for all $t > 0$, *i.e.*,

$$\|\overline{V}_t - \overline{\mathcal{V}}_t\|_0 < \frac{1}{\alpha}\frac{2K}{\beta} \qquad \text{for all } t > 0\,,$$

as claimed. ■

To achieve the claimed $\mathcal{O}(\alpha^{-2})$ bound, we observe that K in the above Lemma can be chosen $\mathcal{O}(\alpha^{-1})$ by the Lipschitz continuity of $\bar{g}_t^{-1}$. Indeed, since we chose $\|\overline{V}\|_0 = C/\alpha$, by (52) we have that

$$K \leq \sup_{t>0}\|\Gamma(TD)_t\|\|D\psi_{\overline{V}_t}^{-1}\|L_{\bar{g}_0^{-1}}\|\overline{V}\|_0 \leq C_0\sigma_{\max}^{D\psi^{-1}}L_{\bar{g}_0^{-1}}\frac{C}{\alpha} \leq \frac{\beta}{2}\frac{K'}{\alpha}\,,$$

for K' large enough, and therefore

$$\|\overline{V}_t - \overline{\mathcal{V}}_t\|_0 < \frac{K'}{\alpha^2} \qquad \text{for all } t > 0\,. \tag{62}$$

Mapping to the embedding space. We conclude the proof by mapping back to the original space, where we have

$$\begin{aligned}
\sup_{t>0}\|\mathcal{V}_t - V_t\|_\mu &= \sup_t \|\alpha\psi^{-1}(\overline{V}_t) - \alpha\bar{\psi}^{-1}(\overline{\mathcal{V}}_t)\|_\mu \\
&\leq \sup_t \alpha\left(\|D\psi_0^{-1}(\overline{V}_t - \overline{\mathcal{V}}_t)\|_\mu + L_{D\psi^{-1}}\|\overline{V}_t\|_\mu^2\right) \\
&\leq \alpha\left(\kappa\|D\psi_0^{-1}\|\sup_t\|\overline{V}_t - \overline{\mathcal{V}}_t\|_0 + \kappa^2 L_{D\psi^{-1}}\sup_t\|\overline{V}_t\|_0^2\right)\,.
\end{aligned}$$

Then, letting $\mathcal{V}^*$ be the fixed point of (44) (unique and attracting by Tsitsiklis and Van Roy (1997)), by our choice of $\delta = C/\alpha$, (46) and (62) we have that

$$\begin{aligned}
\|\tilde{V}^* - V^*\|_\mu &\leq \|\mathcal{V}^* - V^*\|_\mu + \sup_{t>0}\|\mathcal{V}_t - V_t\|_\mu \\
&\leq \frac{1-\gamma\lambda}{1-\gamma}\|\Pi_0 V^* - V^*\|_\mu + \frac{1}{\alpha}(\kappa\|D\psi_0^{-1}\|K' + \kappa^2 L_{D\psi^{-1}}C^2)\,,
\end{aligned}$$

as claimed. ■

Appendix B. Supplementary proofs in the mean field regime

B.1. Proof of convergence of the particle system

Proposition 9 *Let Assumption 3 hold and let $w_t^{(N)}$ be a solution of (4) with initial condition $w_0^{(N)} \in \mathcal{W} = \Omega^N$. If $\nu_0^{(N)}$ converges to $\nu_0 \in \mathcal{P}_2(\Omega)$ in Wasserstein distance W_2 as $N \to \infty$ then $\nu_t^{(N)}$ converges, for every $t > 0$, to the unique solution ν_t of (14).*

Proof As anticipated in the main text, the proof of this result is mainly standard. Therefore, we refer to (Chizat and Bach, 2018, Theorem 2.6) for a detailed proof of an analogous result in the supervised setting and we proceed to highlight the steps that need to be adapted in that proof to cover the given setting. The key estimate needed to prove propagation of chaos concern the Lipschitz continuity of the transport vector field in (14), which yields the desired result by Gronwall inequality. We note that, in order to establish such estimates, the authors combine the assumed regularity of the nonlinearity ψ with the ones of the energy functional $R : \mathcal{F} \to \mathbb{R}$. Therefore, since our assumptions on ψ coincide with those of Chizat and Bach (2018), it is sufficient to recover the desired result to show that the TD error $\delta(\nu)$, playing the role of dR from Chizat and Bach (2018) in the present context, is Lipschitz continuous on bounded sets and bounded on sets that are forward invariant with respect to the dynamics. The former property, however, follows immediately by the *linearity* of $\delta(\nu)$ in ν, and the assumed regularity of the nonlinearity ψ.

We conclude the proof by establishing the latter property, corresponding to "boundedness of dR on sublevel sets" in Chizat and Bach (2018) used to prove Lipschitz continuity of the vector field in ν. This directly results from Lemma 10 since boundedness of $\|V_\nu\|_\mu$ implies, by compactness of $\mathcal{S}$, the boundedness of $\delta(\nu)$ along the trajectories of (14) as required. ∎

Lemma 10 *For any ν_0 with $\int \omega_0^2 \nu_0(\mathrm{d}\omega) \leq \infty$ there exists $C_V > 0$ such that, for any $t > 0$ we have $\|V_{\nu_t}\|_\mu < C_V$.*

Proof We consider the evolution of the quantity

$$\|(\nu_t)_0\|_2^2 := \int_\Omega \omega_0^2 \nu_t(\mathrm{d}\omega)\,,$$

for which we have, by homogeneity of ψ and integrating by parts,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|(\nu_t)_0\|_2^2 = \frac{1}{2}\int_\Omega \omega_0^2 \frac{\mathrm{d}}{\mathrm{d}t}\nu_t(\mathrm{d}\omega) = \int_\Omega \int_\mathcal{S} \omega_0 \phi(s;\bar{\omega})\delta(\nu_t)\nu_t(\mathrm{d}\omega)\mu(\mathrm{d}s) = \int_\mathcal{S} V(s)\delta(\nu_t)\mu(\mathrm{d}s)\,.$$

Now, recalling the definition of the metric tensor Γ induced by the measure μ we define throughout the positive definite operator $A := \Gamma(1-\gamma P)$. Then, whenever

$$\|V_\nu\|_\mu > (1-\gamma)^{-1}\|(1-\gamma P)V^*\|_\mu + \varepsilon =: B + \varepsilon \tag{63}$$

we have

$$\begin{aligned} |V_\nu A V^*| &\leq \|V_\nu\|_\mu \|(1-\gamma P)V^*\|_\mu < (1-\gamma)\|V_\nu\|_\mu^2 - \varepsilon(1-\gamma)\|V_\nu\|_\mu \\ &< \|V_\nu\|_\mu^2 - \gamma\|V_\nu\|_\mu^2 - C_A \leq V_\nu A V_\nu - C_A \end{aligned}$$

for $C_A = \varepsilon(1-\gamma)((1-\gamma)^{-1}\|AV^*\|_\mu + \varepsilon)$, where in the last inequality we have combined $\|PV_\nu\|_\mu \leq \|V_\nu\|_\mu$ from (Tsitsiklis and Van Roy, 1997, Lemma 1) with Cauchy-Schwarz inequality. In light of the above, writing $\delta(t) = (r - (1-\gamma P)V_\nu) = (1-\gamma P)(V^* - V_\nu)$ we have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|(\nu_t)_0\|_2^2 < -\left(V_\nu A V_\nu - |V_\nu A V^*|\right) < -C_A < 0\,, \tag{64}$$

so that the norm $\|(\nu_t)_0\|_2^2$ must decrease whenever (63) holds, for any $\varepsilon > 0$. Furthermore, we note that by the boundedness of ϕ, *i.e.*, by $\phi(s;\bar{\omega}) < C_\phi$ we have

$$\|V_\nu\|_\mu^2 = \int_{\mathcal{S}} \left(\int_\Omega \omega_0 \phi(s;\bar{\omega})\nu(\mathrm{d}\omega)\right)^2 \mu(\mathrm{d}s) \leq C_\phi^2 \int_\Omega \omega_0^2 \nu(\mathrm{d}\omega) = C_\phi^2 \|(\nu_t)_0\|_2^2\,, \tag{65}$$

Combining (64) and (65) directly implies that the space $\{\nu \,:\, \|(\nu)_0\|_2 \leq (B+\varepsilon)/C_\phi\}$ (which contains $\{\nu \,:\, \|V_\nu\| \leq B + \varepsilon\}$) is globally attractive with respect to the dynamics (14). In other words, this means on one hand that if $\|(\nu_0)_0\|_2 > (B+\varepsilon)/C_\phi$ then we must have $\|V_{\nu_t}\|_\mu < C_\phi\|(\nu_t)_0\|_2 < C_\phi\|(\nu_0)_0\|_2$ for all $t \geq 0$. On the other hand if $\|(\nu_0)_0\|_2 \leq (B+\varepsilon)/C_\phi$ there cannot exist $t \geq 0$ such that $\|(\nu_t)_0\|_2 = 2(B+\varepsilon)/C_\phi$, directly implying that $\|V_{\nu_t}\|_\mu < 2(B+\varepsilon)$. Therefore, for any V^*, resulting from the chosen reward function r and transition operator P, setting $C_V = \max\{2(B+\varepsilon), C_\phi\|(\nu_0)_0\|_2\}$ we must have $\|V_{\nu_t}\|_\mu \leq C_V$ for all $t \geq 0$ as required. ■

B.2. Proof of Theorem 11

Theorem 11 *Let Assumption 3 hold and ν_t given by (14) converge to ν^*, then $V_{\nu^*} = V^*$ μ-a.e.*

Before proceeding to prove Theorem 11, we state the alternative form of Assumption 3 a) in the case where $\Theta \neq \mathbb{R}^{m-1}$.

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

The proof of Theorem 11 is carried out in two steps. We first show in Section B.2.1 that as a consequence of the expressivity of the activation function Assumption 3 b), the suboptimality of a fixed point is reflected in a nonvanishing transport vector field of (14)[3]. Then, in Section B.2.2 we use this partial result to bring the assumption of convergence towards a local minimizer to a contradiction.

B.2.1. The relation between vector field and TD error

To state the first lemma towards the proof of the above result, we observe that the ω_0-component of the transport vector field in (14) reads:

$$\left(\int_{\mathcal{S}\times\mathcal{S}} \nabla_\omega \psi(s;\omega)\delta(s,s',\nu_t)P(s,\mathrm{d}s')\mu(\mathrm{d}s)\right)_0 = \int_{\mathcal{S}} \phi(s;\omega)\int_{\mathcal{S}} \delta(s,s',\nu_t)P(s,\mathrm{d}s')\mu(\mathrm{d}s)\,. \tag{66}$$

3. We remind that this could still result in a fixed point if the measure ν loses support where the transport vector field is nonzero.

We note that the above is a function of $\bar{\omega}'$ only. In the following lemma we relate the optimality of a fixed point of (14) to (66).

Lemma 14 *Let Assumption 3 hold and ν be such that the first component of the vector field in (14) vanishes a.e., i.e., the condition*

$$\int_{\mathcal{S}^2} \left(r(s,s') + \gamma \int_\Omega \psi(s';\omega)\nu(\omega)\mathrm{d}\omega - \int_\Omega \psi(s;\omega)\nu(\omega)\mathrm{d}\omega \right) \phi(s,\bar{\omega}')P(s,\mathrm{d}s')\mu(\mathrm{d}s) = 0 \quad (67)$$

holds $\bar{\omega}'$-almost everywhere in Θ. Then we have $V_\nu = V^*$ μ-a.e..

Proof of Lemma 14 The result follows immediately by Assumption 3 a)-b): Assuming that (67) holds Lebesgue-a.e. in Θ, by the assumed continuity of ϕ in $\bar{\omega}'$ combined with the expressivity of ϕ Assumption 3 b) we must have that

$$\int_{\mathcal{S}} \left(r(s,s') + \gamma \int_\Omega \psi(s';\omega)\nu(\omega)\mathrm{d}\omega - \int_\Omega \psi(s;\omega)\nu(\omega)\mathrm{d}\omega \right) P(s,\mathrm{d}s') = 0 \qquad \mu\text{-a.e.}\,.$$

Because the operator T^λ is a contraction in $L^2(\mathcal{S},\mu)$, this condition can only be satisfied if $V_\nu = V^*$ μ-a.e.. ■

Consequently, suboptimal fixed points of the dynamics (14) cannot satisfy (67) Lebesgue-a.e. in Θ.

B.2.2. INSTABILITY OF LOCAL MINIMA

We prove below that spurious local minima are avoided by the dynamics as discussed in the main text. More specifically, we lead the assumption of convergence of the dynamics to a suboptimal fixed point $\tilde{\nu}$ to a contradiction by combining it with the approximate gradient structure of the TD vector field when ν_t is close to one of such stationary points. This proof leverages two important facts: that by Lemma 14 suboptimal fixed points $\tilde{\nu}$ with $V_{\tilde{\nu}} \neq V^*$ imply the existence of regions in parameter space where the transport vector field does not vanish and that the solution to (14) does not lose (projected) support for any finite time, as summarized in Lemma 15 below. These facts are finally combined in Lemma 16, leading the assumption of convergence to a suboptimal fixed point to the desired contradiction.

By the assumed structure of the approximator we note that all measures with the same expectation in the homogeneous component result in the same approximator, *i.e.*, if ν, ν' are such that $\int \omega_0 \nu(\mathrm{d}\omega_0, \mathrm{d}\bar{\omega}) = \int \omega_0 \nu'(\mathrm{d}\omega_0, \mathrm{d}\bar{\omega})$ a.e. then clearly

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

We therefore introduce the quantity

$$h^1_\nu(\bar{\omega}) := \int \omega_0 \nu(\mathrm{d}\omega_0, \mathrm{d}\bar{\omega}) \quad (68)$$

which will play an important role in the proof of Theorem 11 below.

We now proceed to state and prove the main lemmas of the section:

Lemma 15 *Let Assumption 3 a) hold and let ν_0 satisfy Assumption 3 c), then ν_t solving (14) with initial condition ν_0 satisfies Assumption 3 c) for every $t > 0$.*

Defining throughout $\|\cdot\|_{BL}$ as the bounded Lipschitz norm, we further prove that

Lemma 16 *Let Assumption 3 hold and $\tilde{\nu}$ be a fixed point of (14) such that (67) does not hold a.e.. There exists $\varepsilon > 0$ such that if $\|h^1_{\tilde{\nu}} - h^1_{\nu_{t_1}}\|_{BL} < \varepsilon$ for a $t_1 > 0$ there exists $t_2 > t_1$ such that $\|h^1_{\tilde{\nu}} - h^1_{\nu_{t_2}}\|_{BL} > \varepsilon$.*

Proof of Lemma 15 This result corresponds to (Chizat and Bach, 2018, Lemma C.13) about the stability of the separation property Assumption 3 c) under pushforwards of the initial condition ν_0 under the integrated gradient flow map X_t defined as

$$\partial_t X(t,u) = v_t(X(t,u)) \qquad \text{with} \qquad X(0,\cdot) = \mathrm{id} \tag{69}$$

where $\mathrm{id} : \Omega \to \Omega$ denotes the identity and v_t is the vector field of the transport partial differential equation (14). We note that in its original form Chizat and Bach (2018) this result does *not* leverage the gradient flow structure of the dynamics, but only the *continuity* of the map X_t, proven in (Chizat and Bach, 2018, Lemma B.4) under assumptions Assumption 3 a) or, alternatively, Assumption B.2. It is therefore sufficient for our purposes to establish such continuity property for the map X_t in the present setting, *i.e.*, when v_t is the vector field of the transport equation (14). The continuity of the above map results immediately from the one-sided Lipschitz property from Assumption 3 a) enjoyed by v_t on the sets $Q_r = [-r, r] \times \Theta$ uniformly on compact time intervals, as mentioned in (Chizat and Bach, 2018, Proof of Lemma B.4). This regularity is in turn guaranteed by the Lipschitz continuity and Lipschitz smoothness of ψ from Assumption 3 and bundedness of r. ■

Proof of Lemma 16 By Lemma 14, the key quantity

$$g_{\tilde{\nu}}(\bar{\omega}) := \langle \partial_{\omega_0}\psi(\cdot;\omega), \delta(\tilde{\nu})\rangle = \langle \psi(\cdot;(1,\bar{\omega})), \delta(\tilde{\nu})\rangle = \langle \phi(\cdot;\bar{\omega}), \delta(\tilde{\nu})\rangle \tag{70}$$

cannot vanish a.e. on Θ. Then, by Assumption 3 there exists a nonzero regular value $-\eta$ of $g_{\tilde{\nu}}(\bar{\omega})$, which without loss of generality we assume to be negative, so that $\eta > 0$ (else invert the signs of ω_0 in the rest of the proof). To conclude the proof we introduce the (nonempty) sublevel set $\mathcal{A} := \{(\omega_0, \bar{\omega}) \in \Omega \,:\, g_{\tilde{\nu}}(\bar{\omega}) < -\eta\}$ and define

$$A_+ = \{(\omega_0, \bar{\omega}) \in \mathcal{A} \,:\, \omega_0 > 0\}\,. \tag{71}$$

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

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

where $n_{\bar{\omega}}$ denotes the normal unit vector to $\partial\bar{A}$ in the outward direction.[4]

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

We now choose ε small enough that assuming

$$\|h^1_{\tilde{\nu}} - h^1_{\nu_t}\|_{BL} < \varepsilon \qquad \text{for all } t > t_1\,, \tag{72}$$

leads to a contradiction. More specifically, by Lemma 18 we choose $\varepsilon(\alpha, \eta, \beta)$ small enough so that for all ν_t such that (72) holds we have $g_{\nu_t}(\bar{\omega}) < -\eta/2$ on $\bar{A}$ and $n_{\bar{\omega}} \cdot \nabla_{\bar{\omega}} g_{\nu_t} > \beta/2$ on $\partial\bar{A}$. Then, the two inequalities above combined with $\partial_{\omega_0}\psi(\omega_0, \bar{\omega}) = \psi(1, \bar{\omega})$ imply that the set A_+ defined above is forward invariant and therefore hat $\partial_t \nu_t(A_+) \geq 0$ as long as (72) holds. Furthermore, by similar arguments we notice that no trajectories enter the set $\mathcal{A} \setminus A_+$ after t_1.

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

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

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

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

2. Assume now that $\nu_{t_1}(A_+) = 0$. We show that there exists $t_2 > t_1$ such that $\nu_{t_2}(A_+) > 0$, so that the proof is completed by applying part i) above from t_2. Indeed let $\omega^* \in \mathrm{supp}(\nu_{t_1})$ be such that $\bar{\omega}^* \in \bar{A}$ is a local minimum of $g_{\tilde{\nu}}$, *i.e.*, for which $\nabla g_{\tilde{\nu}} = 0$. Then choosing $\tilde{\varepsilon}$ such that $\mathcal{B}_{\tilde{\varepsilon}}(\bar{\omega}^*) \subset \bar{A}$, and choosing M large enough such that $\mathrm{supp}\,(\nu_{t_1}) \subseteq [-M, M] \times \Theta$, we know by Lemma 17 that there exists $t_2 > t_1$ such that the image at t_2 of $\omega(t_1) := \omega^*$ under the flow of the TD vector field is contained in A_+. By continuity of pushforward map of such vector field, this must also hold for a neighborhood of ω^*, to which ν_{t_1} assigns positive mass.

■

Defining by $\|\cdot\|_{C^1}$ the maximum of the supremum norm of a function and the supremum norm of its gradient and recalling the structure of the temporal difference vector field:

$$v_t(\omega) = -\langle \nabla\omega_0 \phi(\bar{\omega}), \delta(s, s', \nu_t) P(s, \mathrm{d}s')\mu(\mathrm{d}s)\rangle = -\nabla(\omega_0\, g_{\nu_t}(\bar{\omega})) \tag{73}$$

where ν_t solves (14) we are ready to state the lemma needed to prove part ii).

Lemma 17 *Let $\tilde{\nu} \in \mathcal{M}_+(\Omega)$ and $\bar{\omega}^*$ satisfy $|\nabla g_{\tilde{\nu}}(\bar{\omega}^*)| = 0$, $g_{\tilde{\nu}}(\bar{\omega}^*) < -\eta < 0$ for some $\eta > 0$. Then for every $M, \tilde{\varepsilon} > 0$ there exists $\varepsilon, t_2 > 0$ such that if for all $t \in (0, t_2)$ we have $\|g_{\tilde{\nu}} - g_{\nu_t}\|_{C^1} < \varepsilon$ and $\omega_0^* \in [-M, 0]$, then at time t_2 the point ω^* is mapped, under the flow of the TD vector field (73) to a subset of $\mathcal{B}_{\tilde{\varepsilon}}((1, \bar{\omega}^*))$.*

Proof of Lemma 17 Defining $q(t) := \|\bar{\omega}(t) - \bar{\omega}^*\|$, we see that the trajectory $(\omega_0(t), \bar{\omega}(t))$ of a particle under (73) with initial condition $\omega(0) = \bar{\omega}^*$ must satisfy:

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

and

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

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

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

Upon possibly increasing the value of L such that $\eta/4L < \tilde{\varepsilon}$, we define $\tau_q = \inf\{t \,:\, q(t) > \eta/4L\}$ and we now proceed to show that one can choose $\varepsilon \in (0, \eta/4)$ such that $\tau_q > \bar{\tau}$, *i.e.*, that the forward dynamics of the point $\omega^* = (\omega_0^*, \bar{\omega}^*)$ will reach A_+ before $q(t) > \tilde{\varepsilon}$. Recall that on $[0, \tau_q]$ and for $\varepsilon \in (0, \eta/4)$ we have

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

and in particular $\omega_0(t) > \omega_0(0) \geq -M$. This implies that for all $t \in [0, \bar{\tau} \wedge \tau_q]$ we have $\frac{\mathrm{d}}{\mathrm{d}t}q(t) < q(0) + M\varepsilon + LMq(t)$, so that by Gronwall we obtain $q(t) \leq \varepsilon M \exp[LMt]$. Finally, setting $\tau_0 := 2(M+1)/\eta \geq -2(\omega_0(0) - 1)/\eta > \bar{\tau}$ we note that we can choose ε small enough such that $\tau_q > \tau_0 > \bar{\tau}$, *i.e.*, by monotonicity of $q(t)$ under our bound, such that

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

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

Lemma 18 *Recalling the definitions of g_ν, h_ν^1 from (70), (68) respectively, for all $C_0 > 0$ there exists $\alpha > 0$ and for all ν, ν' satisfying $\|h_\nu^1\|_{BL}, \|h_{\nu'}^1\|_{BL} < C_0$, one has*

$$\|g_\nu - g_{\nu'}\|_{C^1} \leq \alpha \|\phi\|_{C^1}^2 \|h_\nu^1 - h_{\nu'}^1\|_{BL}\,.$$

Proof of Lemma 18 Recognizing that $\psi(s; (1, \bar{\omega})) = \phi(s; \bar{\omega})$ and letting α be the Lipschitz constant of $\langle\,\cdot\,, \delta\rangle$ on the bounded set $\{\int \psi\nu(\mathrm{d}\omega) \,:\, \|h_\nu^1\| < C_0\}$ we have

$$\begin{aligned}\|g_\nu - g_{\nu'}\|_{C^1} &= \|\langle\psi(s; (1, \,\cdot\,)), \delta(\nu)\rangle_{\mathcal{F}} - \langle\psi(s; (1, \,\cdot\,)), \delta(\nu')\rangle_{\mathcal{F}}\|_{C^1} \\ &= \|\langle\phi(s; \,\cdot\,), \delta(\nu) - \delta(\nu')\rangle_{\mathcal{F}}\|_{C^1} \leq \alpha\|\phi\|_{C^1}\|\delta(\nu) - \delta(\nu')\|_{\mathcal{F}} \\ &\leq \alpha\|\phi\|_{C^1} \sup_{f \in \mathcal{F}, \|f\| = 1} \int \langle f, \phi\rangle \mathrm{d}(h_\nu^1 - h_{\nu'}^1)(\bar{\omega}) \\ &\leq \alpha\|\phi\|_{C^1}^2 \|h_\nu^1 - h_{\nu'}^1\|_{BL}\end{aligned}$$

where we used that $\langle f, \phi\rangle_{\mathcal{F}}$ is $\|\,\phi\,\|_{C^1}$-Lipschitz and upper-bonded by $\|\phi\|_{C^1}$ when $\|f\|_{\mathcal{F}} < 1$. ∎