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

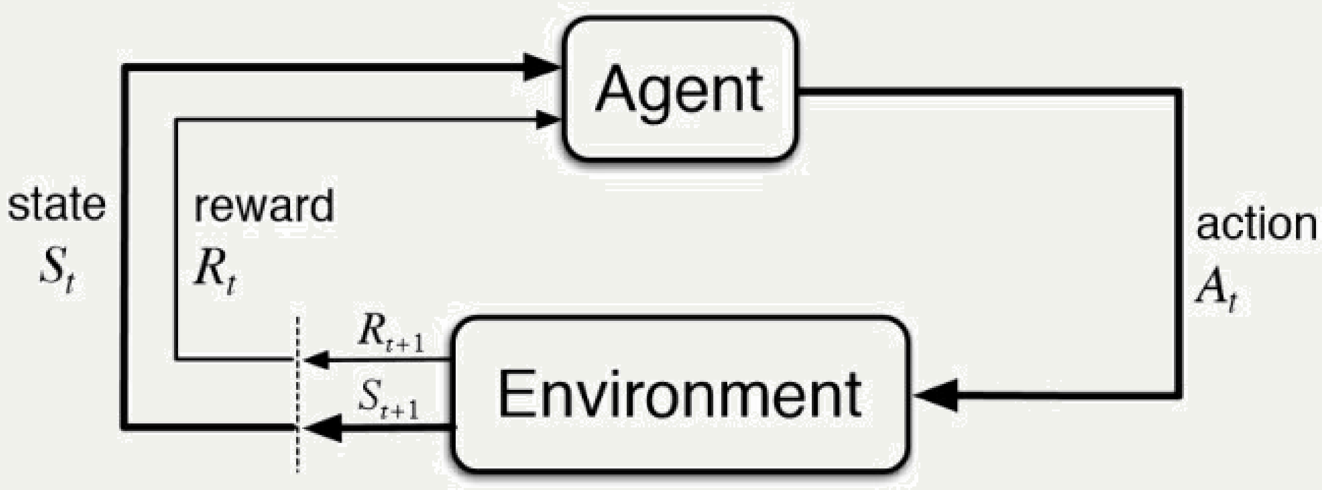
A. Agazzi and J. Lu

Duke University

Mathematical and Scientific Machine Learning

16-19.08.2021

Reinforcement Learning



Markov Decision Processes

Model the *environment* as a Markov Decision Process (MDP)

- A compact *state space* $\mathcal{S}$ and an *action space* $\mathcal{A}$
- A *transition kernel* $P \;:\; \mathcal{S} \times \mathcal{A} \to \mathcal{M}_+^1(\mathcal{S})$ (response of the environment)
- A bounded *reward* $R \;:\; \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (payoff of an action)

TicTacToe: $\mathcal{S} = \{0, 1, -1\}^9$, $\mathcal{A} \subseteq \{1, \ldots, 9\}$, $R(s, a) = \begin{cases} 1 & \text{if win} \\ -1 & \text{if lose} \end{cases}.$

Model the *agent* through its strategy:

- A *policy* $\pi \;:\; \mathcal{S} \to \mathcal{M}_+^1(\mathcal{A})$ (actions chosen by agent at state s)

For each π we have an effective kernel $P_\pi(s, ds') = \int P(s, a, ds')\pi(s, da)$

Value Functions

Objective: fixing $\gamma \in (0, 1)$ and a policy π learn the expected future reward

$$V_\pi^*(s) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R(s_k, a_k) \,\middle|\, s_0 = s \right] \qquad \text{(value function)}$$

For fixed π the value function must satisfy

$$\begin{aligned} V_\pi^*(s) &= \mathbb{E}_\pi \left[R(s_0, a_0) + \gamma (R(s_1, a_1) + \gamma R(s_2, a_2) + \ldots) \,\middle|\, s_0 = s \right] \\ &= \mathbb{E}_\pi \left[R(s_0, a_0) + \gamma \sum_{k=0}^{\infty} \gamma^k R(s_k, a_k) \middle| s_0 = s \right] = \mathbb{E}_\pi \left[R(s_0, a_0) + \gamma V_\pi^*(s_1) \,\middle|\, s_0 = s \right] \end{aligned}$$

In other words, $V_\pi^*(s)$ is a fixed point of the *Bellman* operator

$$T^\gamma V(s) = \mathbb{E}_\pi \left[R(s_0, a_0) + \gamma V(s_1) \,\middle|\, s_0 = s \right]$$

Temporal-difference learning

The operator

$$T^\gamma V(s) = \mathbb{E}_\pi \left[R(s_0, a_0) + \gamma V(s_1) \mid s_0 = s\right]$$

is a contraction in $L^2(\mu)$ where μ is the invariant measure of P_π (assumed unique and with full support)

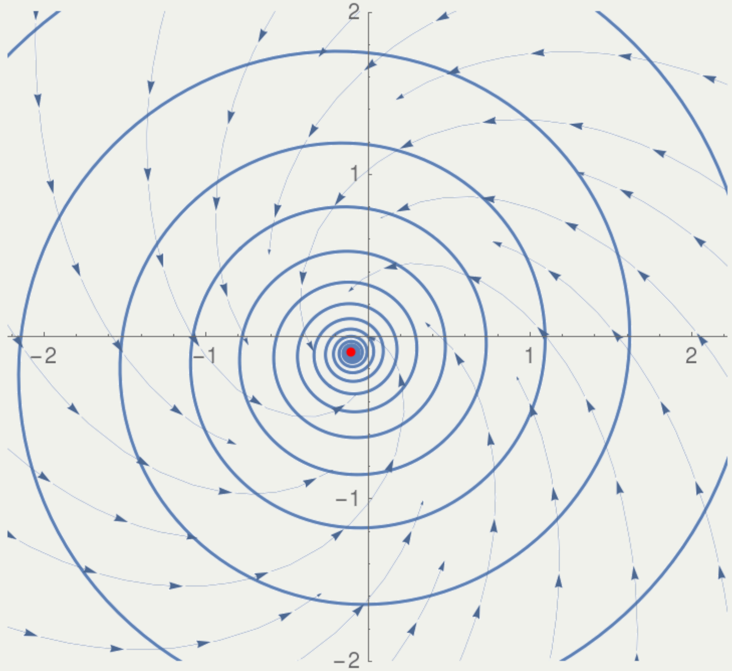
This suggests the *Temporal-Difference* (TD) update with stepsize β:

$$V(s) \leftarrow V(s) + \beta(T^\gamma V(s) - V(s))$$

For a parametric approximation V_w of V with $w \in \mathcal{W}$ the update becomes

$$\frac{d}{dt} w(t) = \mathbb{E}_\mu \left[DV_{w(t)}^\top(s) \left(T^\gamma V_{w(t)}(s) - V_{w(t)}(s)\right)\right]$$

Divergences in TD learning



$$\frac{d}{dt}w(t) = \mathbb{E}_\mu\left[DV_{w(t)}^\top(s)\left(T^\gamma V_{w(t)}(s) - V_{w(t)}(s)\right)\right]$$

(TsitsiklisVanRoy97)

Lazy training

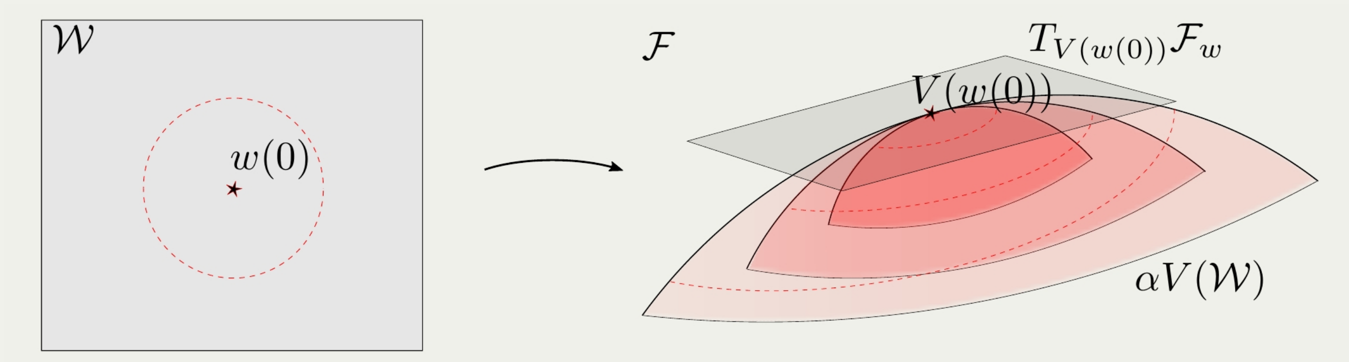
We scale the approximating function as $V_w \to \alpha V_w$ for large α

The parametric update becomes

$$\frac{d}{dt}w(t) = \frac{1}{\alpha}\mathbb{E}'_\mu \left[DV_w^\top(s')\,(T^\gamma \alpha V_w(s') - \alpha V_w(s'))\right]$$

And the functional update for large α is

$$\frac{d}{dt}\alpha V_{w(t)}(s) = \mathbb{E}'_\mu \left[DV_{w(t)}(s) \cdot DV_{w(t)}^\top(s')\,(T^\gamma \alpha V_{w(t)}(s') - \alpha V_{w(t)}(s'))\right]$$



(ChizatBachOyallon19)

Lazy training

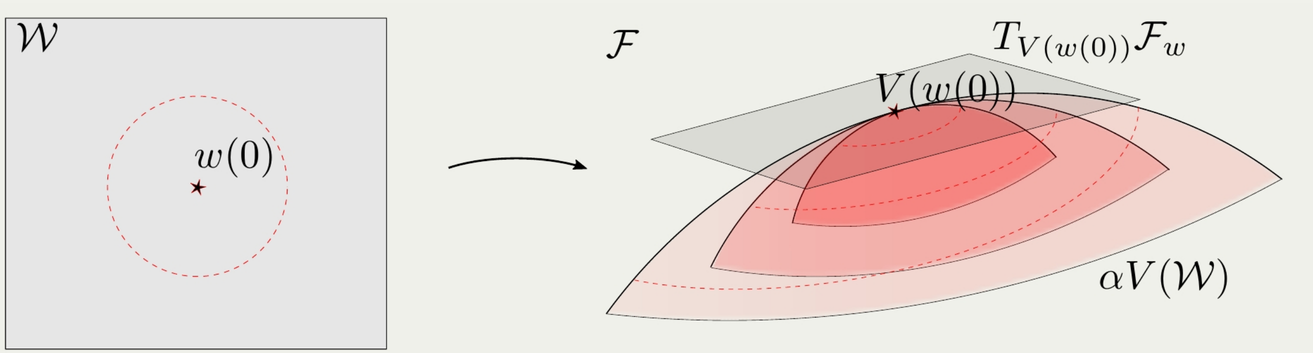
We scale the approximating function as $V_w \to \alpha V_w$ for large α

The parametric update becomes

$$\frac{d}{dt} w(t) = \frac{1}{\alpha} \mathbb{E}'_\mu \left[DV_w^\top(s') \left(T^\gamma \alpha V_w(s') - \alpha V_w(s') \right) \right]$$

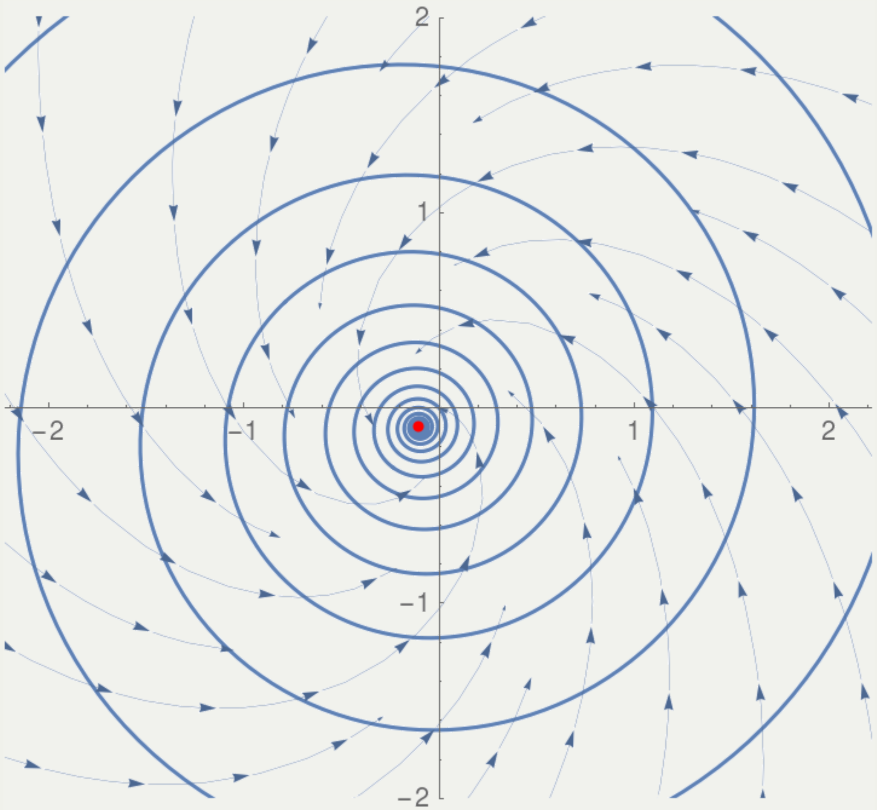
And the functional update for large α is

$$\frac{d}{dt} \alpha V_{w(t)}(s) \approx \mathbb{E}'_\mu \left[DV_{w(0)}(s) \cdot DV_{w(0)}^\top(s') \left(T^\gamma \alpha V_{w(t)}(s') - \alpha V_{w(t)}(s') \right) \right]$$

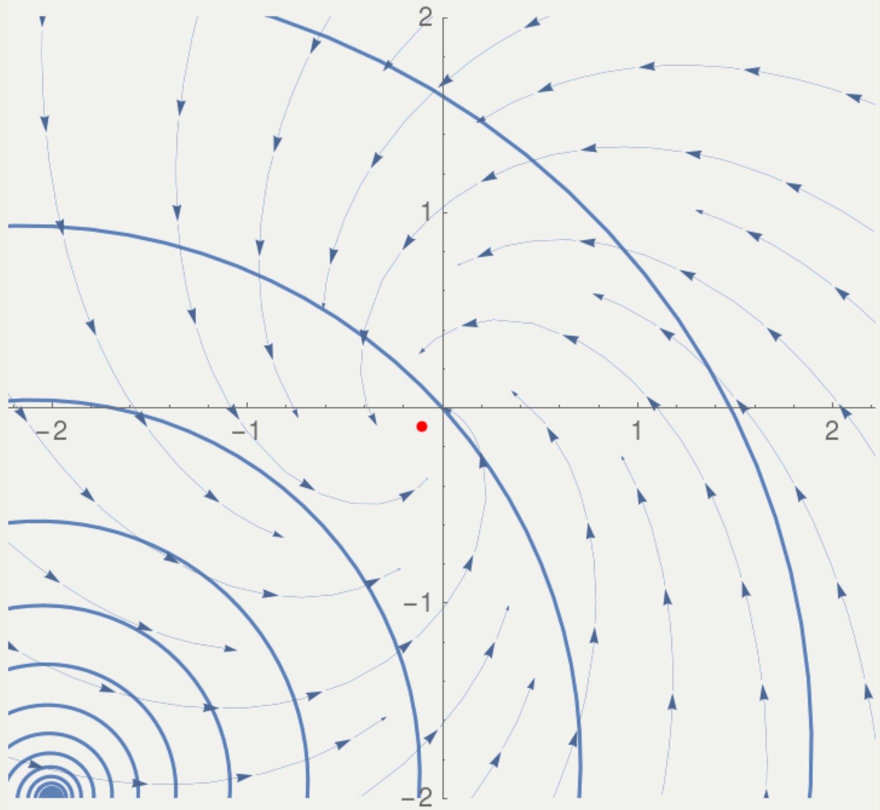


(ChizatBachOyallon19)

Fixing a divergent example



$\alpha = 1$



$\alpha \gg 1$

Convergence of lazy training

Let $\|\cdot\|_0$ be the RKHS norm induced by $DV_{w(0)}DV_{w(0)}^\top$, let Π_0 be the $L^2(\mu)$ projection on such RKHS and assume that $w(0)$ is s.t. $V_{w(0)} = 0$

Theorem 1a (Overparametrized, Informal):
There exist $\alpha_0, \lambda(\gamma) > 0$ s.t. for any $\alpha > \alpha_0$ we have for all $t \geq 0$ that

$$\|V_\pi^* - \alpha V_{w(t)}\|_0^2 \leq \|V_\pi^* - \alpha V_{w(0)}\|_0^2 e^{-\lambda(\gamma)t}$$

Theorem 1b (Underparametrized, Informal):
There exists $\alpha_0 > 0$ such that for any $\alpha > \alpha_0$ the approximation αV_w converges exponentially fast to a locally (in $\mathcal{W}$) attractive fixed point $\tilde{V}_\pi^*$, for which $\|\tilde{V}_\pi^* - V_\pi^*\|_\mu < \frac{1}{1-\gamma}\|\Pi_0 V_\pi^* - V_\pi^*\|_\mu$

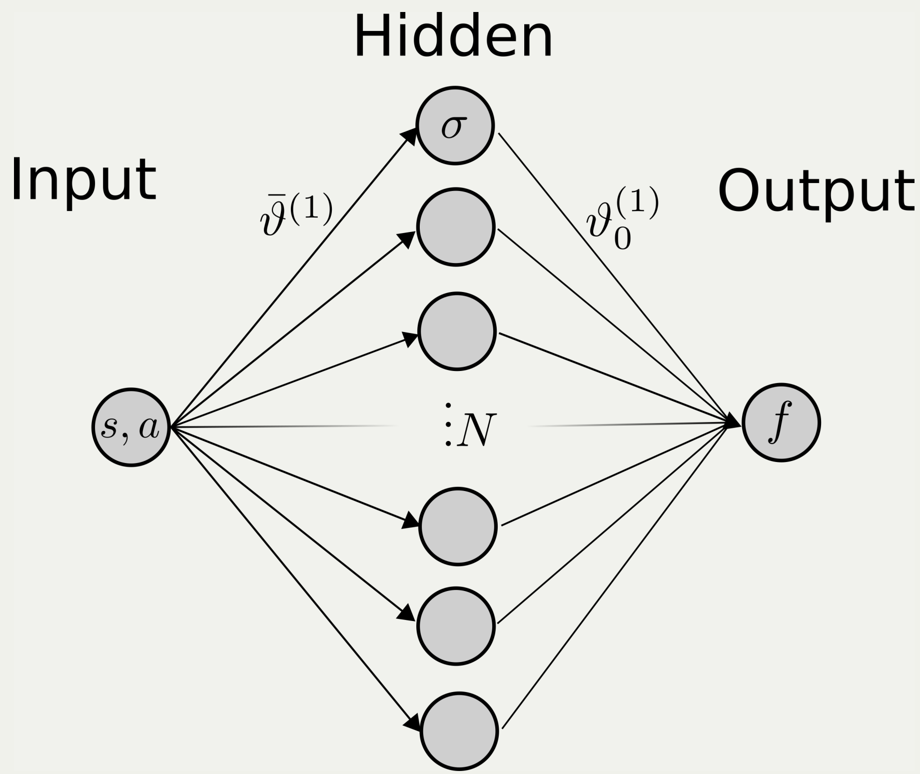
Neural Networks as function approximators

We consider single hidden layer neural networks: $w = (\vartheta^{(i)})_{i=1}^{N}$,

$$V_w(s) = \frac{1}{N}\sum_{i=1}^{N} \vartheta_0^{(i)} \sigma(s; \bar{\vartheta}^{(i)})$$

for $\vartheta^{(i)} = (\vartheta_0^{(i)}, \bar{\vartheta}^{(i)}) \in \Theta$ (weights)

Here the weights $\{\vartheta^{(i)}\}$ are initialized iid and σ is a Lipschitz smooth activation function (bounded, bounded derivative)



(ChizatBach18), (Chizat19), (RotskoffVanDenEijnden18), (MeiMontanariNguyen18), (NguyenPham20), (SirignanoSpiliopoulos18), (Wojtowytsch20), ...

Lazy vs mean-field initialization

The scaling (in N) of $\vartheta_0^{(i)}$ at initialization in $V_w(s) = \frac{1}{N}\sum_{i=1}^{N} \vartheta_0^{(i)}\sigma(s;\bar{\vartheta}^{(i)})$

determines if a network behaves like a lazy learner:

- **When $\boldsymbol{\vartheta_0^{(i)}(0) \sim \mathcal{N}(0, N)}$** (*e.g.* Xavier initialization) we have
$$V_w(s) = \alpha(N)\frac{1}{N}\sum_{i=1}^{N} \tilde{\vartheta}_0^{(i)}\sigma(s;\bar{\vartheta}^{(i)}) \qquad \text{for} \quad \alpha(N) = \sqrt{N}$$
for $\tilde{\vartheta}_0^{(i)}(0) \sim \mathcal{N}(0, 1)$, resulting in the *lazy* (or NTK) regime
with kernel
$K_{\nu_0}(s, s') = \mathbb{E}_{\nu_0}\left[\sigma(s\cdot\bar{\vartheta})\sigma(s'\cdot\bar{\vartheta})\right] + \mathbb{E}_{\nu_0}\left[(s\cdot s')\vartheta_0^2\sigma'(s\cdot\bar{\vartheta})\sigma'(s'\cdot\bar{\vartheta})\right]$
- **When $\boldsymbol{\vartheta_0^{(i)}(0) \sim \mathcal{N}(0, 1)}$** we are in the *mean-field* regime

(CaiYangLeeWang19), (ChizatBachOyallon19), (GhorbaniMeiMisiakiewiczMontanari20), (JacotGabrielHongler18), (SirignanoSpiliopoulos19) ...

Mean-field regime

We express the approximator through $\nu^{(N)}(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\vartheta^{(i)}}(\cdot) \in \mathcal{M}_+^1(\Theta)$

$$V_{\nu^{(N)}}(s) = \frac{1}{N}\sum_{i=1}^{N} \vartheta_0^{(i)}\sigma(s;\bar{\vartheta}^{(i)}) = \int_\Theta \vartheta_0\sigma(s;\bar{\vartheta})\nu^{(N)}(d\vartheta)$$

Then we can write the set of ODEs for the update of $\vartheta^{(i)}$

$$\frac{d}{dt}\vartheta^{(i)}(\tau) = \mathbb{E}_\mu\left[\nabla_{\vartheta^{(i)}} V_{w(\tau)}(s)\left(T^\gamma V_{w(\tau)}(s) - V_{w(\tau)}(s)\right)\right]$$

as a Vlasov PDE for the evolution of $\nu_t = \nu_t^{(N)}$:

$$\frac{d}{dt}\nu_t(\vartheta) = \mathrm{div}\left(\nu_t(\vartheta)\;\mathbb{E}_\mu\left[\nabla_\vartheta(\vartheta_0\sigma(s;\bar{\vartheta}))\;(T^\gamma V_{\nu_t}(s) - V_{\nu_t}(s))\right]\right)$$

(ChizatBach18), (Chizat19), (RotskoffVanDenEijnden18), (MeiMontanariNguyen18), (NguyenPham20), (SirignanoSpiliopoulos18), (Wojtowytsch20), ...

Mean-field regime: convergence

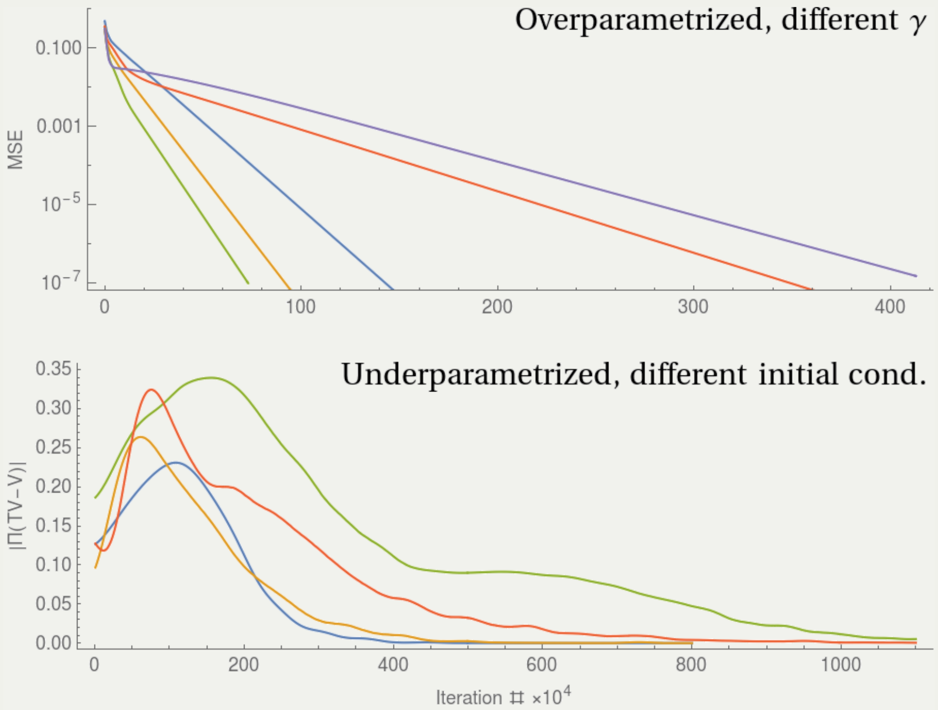
With $V_\nu(\cdot) = \int_\Theta \vartheta_0 \sigma(\cdot; \bar{\vartheta})\, \nu(d\vartheta)$ we write the evolution of ν as

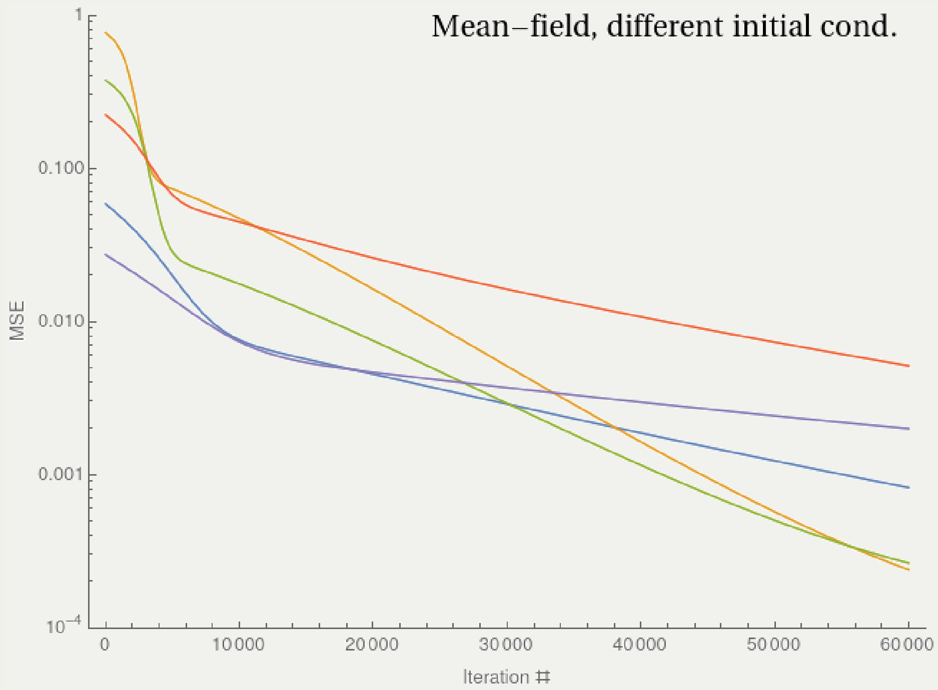
$$\frac{d}{dt}\nu_t(\vartheta) = \operatorname{div}\left(\nu_t(\vartheta)\ \mathbb{E}_\mu\left[\nabla_\vartheta(\vartheta_0\sigma(s;\bar{\vartheta}))\ (T^\gamma V_{\nu_t}(s) - V_{\nu_t}(s))\right]\right)$$

Proposition 2 ($N \to \infty$ convergence): Let $\{\vartheta_t^{(i)}\}_{i=1}^N$ obey the Temporal Difference ODEs and $\nu_0^{(N)} \to \nu_0 \in \mathcal{P}_2(\Theta)$ as $N \to \infty$ then for every $t > 0$ we have $\nu_t^{(N)} \to \nu_t$ solving the above PDE.

Theorem 2 (Optimality): Let $\operatorname{span}(\sigma(\cdot;\bar{\vartheta}))$ be dense in $L^2(\mathcal{S}, \mu)$, ν_0 have full support in Θ and assume that ν_t converges to ν^* as $t \to \infty$, then $V_{\nu^*} = V_\pi^*$ μ-a.e.

Numerical results





lazy

mean-field

Summary

The training dynamics of wide, single layer neural networks trained with Temporal-Difference learning are:

- Convergent (but not always optimal) in the lazy regime
- Optimal (but not provably convergent) in the mean-field regime

Open Questions:

- Convergence of the mean-field dynamics
- Finite-sample anaysis (stochastic approximation)
- Multilayer Neural Networks
- Other algorithms