# On the Global Convergence of Policy Optimization in Deep Reinforcement Learning

Zhaoran Wang
(Joint Work with Qi Cai, Jason D. Lee, Boyi Liu, Zhuoran Yang)

Department of IEMS, Northwestern University

October 20, 2019

# Deep Reinforcement Learning: Success & Failure

- success: DL (representation) + RL (decision) = human-level AI
  - play: Atari, Texas hold'em, Go, Dota, Starcraft, Doom, ...
  - control: grasp, move, walk, swim, drive, ...
  - interact: recommend, personalize, chat, ...
  - explore: 3D scene, maze, ...
- failure: convergence, generalization, sample efficiency, reproducibility
  - "deep reinforcement learning that matters" (Henderson et al.)
  - "deep reinforcement learning doesn't work yet" (Irpan)
  - "RL never worked, and 'deep' only helped a bit" (Sahni)
  - "policy gradient is nothing more than random search" (Recht)
  - "are deep policy gradient algorithms truly policy gradient algorithms" (Iyas et al.)
- this talk: convergence & generalization of policy optimization

# Deep Reinforcement Learning: Success & Failure

- success: DL (representation) + RL (decision) = human-level AI
  - play: Atari, Texas hold'em, Go, Dota, Starcraft, Doom, ...
  - control: grasp, move, walk, swim, drive, ...
  - interact: recommend, personalize, chat, ...
  - explore: 3D scene, maze, ...
- failure: convergence, generalization, sample efficiency, reproducibility
  - "deep reinforcement learning that matters" (Henderson et al.)
  - "deep reinforcement learning doesn't work yet" (Irpan)
  - "RL never worked, and 'deep' only helped a bit" (Sahni)
  - "policy gradient is nothing more than random search" (Recht)
  - "are deep policy gradient algorithms truly policy gradient algorithms" (Iyas et al.)
- this talk: convergence & generalization of policy optimization

# Deep Reinforcement Learning: Success & Failure

- success: DL (representation) + RL (decision) = human-level AI
  - play: Atari, Texas hold'em, Go, Dota, Starcraft, Doom, ...
  - control: grasp, move, walk, swim, drive, ...
  - interact: recommend, personalize, chat, ...
  - explore: 3D scene, maze, ...
- failure: convergence, generalization, sample efficiency, reproducibility
  - "deep reinforcement learning that matters" (Henderson et al.)
  - "deep reinforcement learning doesn't work yet" (Irpan)
  - "RL never worked, and 'deep' only helped a bit" (Sahni)
  - "policy gradient is nothing more than random search" (Recht)
  - "are deep policy gradient algorithms truly policy gradient algorithms" (Iyas et al.)
- this talk: convergence & generalization of policy optimization

# Background on RL & DL

- Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $s' \sim \mathcal{P}(\cdot \,|\, s, a)$
  - actor: policy $a \sim \pi(\cdot \,|\, s)$
  - critic: action-value function

$$Q^\pi(s,a) := (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s,\ a_0 = a,\ a_t \sim \pi(\cdot \,|\, s_t)\right]$$

  - goal: $\max_\pi J(\pi) := \mathbb{E}_{s\sim\nu}[V^\pi(s)] = \mathbb{E}_{s\sim\nu}[\langle Q^\pi(s,\cdot), \pi(\cdot \,|\, s)\rangle]$

- two-layer neural networks $(b, \alpha, \sigma, m)$ serving as actor & critic

$$u_\alpha(s,a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \sigma([\alpha]_i^\top (s,a))$$

which is randomly initialized & "overparametrized"

- question: global convergence $J(\pi^*) - J(\pi^k) \to 0$?

# Background on RL & DL

- Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $s' \sim \mathcal{P}(\cdot \,|\, s, a)$
  - actor: policy $a \sim \pi(\cdot \,|\, s)$
  - critic: action-value function

$$Q^\pi(s,a) := (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s,\ a_0 = a,\ a_t \sim \pi(\cdot \,|\, s_t)\right]$$

  - goal: $\max_\pi J(\pi) := \mathbb{E}_{s\sim\nu}[V^\pi(s)] = \mathbb{E}_{s\sim\nu}[\langle Q^\pi(s,\cdot), \pi(\cdot \,|\, s)\rangle]$

- two-layer neural networks $(b, \alpha, \sigma, m)$ serving as actor & critic

$$u_\alpha(s,a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \sigma([\alpha]_i^\top (s,a))$$

  which is randomly initialized & “overparametrized”

- question: global convergence $J(\pi^*) - J(\pi^k) \to 0$?

# Background on RL & DL

- Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $s' \sim \mathcal{P}(\cdot \,|\, s, a)$
  - actor: policy $a \sim \pi(\cdot \,|\, s)$
  - critic: action-value function

$$Q^\pi(s,a) := (1-\gamma)\mathbb{E}\Bigg[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\Bigg|\, s_0 = s,\ a_0 = a,\ a_t \sim \pi(\cdot \,|\, s_t)\Bigg]$$

  - goal: $\max_\pi J(\pi) := \mathbb{E}_{s\sim\nu}[V^\pi(s)] = \mathbb{E}_{s\sim\nu}[\langle Q^\pi(s,\cdot), \pi(\cdot \,|\, s)\rangle]$

- two-layer neural networks $(b, \alpha, \sigma, m)$ serving as actor & critic

$$u_\alpha(s,a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \sigma([\alpha]_i^\top (s,a))$$

  which is randomly initialized & "overparametrized"

- question: global convergence $J(\pi^*) - J(\pi^k) \to 0$?

# Algorithm: TRPO/PPO (Actor) + TD (Critic)

- trust region/proximal policy optimization (Schulman et al.)

$$\theta_{k+1} \leftarrow \arg\max_{\theta} \widehat{\mathbb{E}}_{s\sim\nu_k}\Big[\underbrace{\langle Q^{\pi_{\theta_k}}(s,\cdot), \pi_\theta(\cdot\,|\,s)\rangle}_{\text{improvement}} - \beta_k \underbrace{\mathrm{KL}(\pi_\theta(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))}_{\text{regularization}}\Big]$$

which is related to REINFORCE, (S)AC, DPG, SBEED, ...

- temporal-difference learning (Sutton; Tsitsiklis & Van Roy)

$$\omega(t+1) \leftarrow \omega(t) - \eta \underbrace{\big(Q_{\omega(t)}(s,a) - r(s,a) - \gamma Q_{\omega(t)}(s',a')\big)}_{\text{Bellman residual}} \nabla_\omega Q_{\omega(t)}(s,a)$$

where $(s,a)\sim\nu_k$, $s'\sim\mathcal{P}(\cdot\,|\,s,a)$, $a'\sim\pi_{\theta_k}(\cdot\,|\,s')$

- combination: $Q_{\omega_k}$ from TD $\rightarrow Q^{\pi_{\theta_k}}$ in TRPO/PPO

# Algorithm: TRPO/PPO (Actor) + TD (Critic)

- trust region/proximal policy optimization (Schulman et al.)

$$\theta_{k+1} \leftarrow \arg\max_{\theta} \widehat{\mathbb{E}}_{s\sim\nu_k}\Big[\underbrace{\langle Q^{\pi_{\theta_k}}(s,\cdot), \pi_\theta(\cdot\,|\,s)\rangle}_{\text{improvement}} - \beta_k \underbrace{\mathrm{KL}(\pi_\theta(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))}_{\text{regularization}}\Big]$$

which is related to REINFORCE, (S)AC, DPG, SBEED, . . .

- temporal-difference learning (Sutton; Tsitsiklis & Van Roy)

$$\omega(t+1) \leftarrow \omega(t) - \eta \underbrace{\big(Q_{\omega(t)}(s,a) - r(s,a) - \gamma Q_{\omega(t)}(s',a')\big)}_{\text{Bellman residual}} \nabla_\omega Q_{\omega(t)}(s,a)$$

where $(s,a) \sim \nu_k$, $s' \sim \mathcal{P}(\cdot\,|\,s,a)$, $a' \sim \pi_{\theta_k}(\cdot\,|\,s')$

- combination: $Q_{\omega_k}$ from TD $\to Q^{\pi_{\theta_k}}$ in TRPO/PPO

# Algorithm: TRPO/PPO (Actor) + TD (Critic)

- trust region/proximal policy optimization (Schulman et al.)

$$\theta_{k+1} \leftarrow \arg\max_{\theta} \widehat{\mathbb{E}}_{s\sim\nu_k}\Big[\underbrace{\langle Q^{\pi_{\theta_k}}(s,\cdot), \pi_\theta(\cdot\,|\,s)\rangle}_{\text{improvement}} - \beta_k \underbrace{\mathrm{KL}(\pi_\theta(\cdot\,|\,s)\,\|\,\pi_{\theta_k}(\cdot\,|\,s))}_{\text{regularization}}\Big]$$

which is related to REINFORCE, (S)AC, DPG, SBEED, ...

- temporal-difference learning (Sutton; Tsitsiklis & Van Roy)

$$\omega(t+1) \leftarrow \omega(t) - \eta \underbrace{\big(Q_{\omega(t)}(s,a) - r(s,a) - \gamma Q_{\omega(t)}(s',a')\big)}_{\text{Bellman residual}} \nabla_\omega Q_{\omega(t)}(s,a)$$

where $(s,a) \sim \nu_k$, $s' \sim \mathcal{P}(\cdot\,|\,s,a)$, $a' \sim \pi_{\theta_k}(\cdot\,|\,s')$

- combination: $Q_{\omega_k}$ from TD $\rightarrow Q^{\pi_{\theta_k}}$ in TRPO/PPO

# Challenges & Agenda

- two of deadly triad: nonlinearity + bootstrapping + off-policy (Sutton & Barto)
- sources of nonconvexity: "bad" stationary points
  - $J(\pi_\theta)$ nonconvex in $\pi_\theta \in \Delta^{|\mathcal{S}|}$ (infinite dimensions)
  - critic: $Q_\omega$ nonlinear in $\omega$ (finite dimensions)
  - actor: $\pi_\theta$ nonlinear in $\theta$ (finite dimensions)
- causes of divergence: instability in practice
  - actor + critic: bilevel optimization (Pfau et al.; Heusel et al.)
  - critic: bias in TD update + nonlinearity of $Q_\omega$
  - actor: error in GD propagated from critic
- agenda: (i) TRPO/PPO (infinite dimensions) + (ii) TD/GD (finite dimensions)

# Challenges & Agenda

- two of deadly triad: nonlinearity + bootstrapping + off-policy (Sutton & Barto)
- sources of nonconvexity: "bad" stationary points
  - $J(\pi_\theta)$ nonconvex in $\pi_\theta \in \Delta^{|\mathcal{S}|}$ (infinite dimensions)
  - critic: $Q_\omega$ nonlinear in $\omega$ (finite dimensions)
  - actor: $\pi_\theta$ nonlinear in $\theta$ (finite dimensions)
- causes of divergence: instability in practice
  - actor + critic: bilevel optimization (Pfau et al.; Heusel et al.)
  - critic: bias in TD update + nonlinearity of $Q_\omega$
  - actor: error in GD propagated from critic
- agenda: (i) TRPO/PPO (infinite dimensions) + (ii) TD/GD (finite dimensions)

# Challenges & Agenda

- two of deadly triad: nonlinearity + bootstrapping + off-policy (Sutton & Barto)
- sources of nonconvexity: "bad" stationary points
  - $J(\pi_\theta)$ nonconvex in $\pi_\theta \in \Delta^{|\mathcal{S}|}$ (infinite dimensions)
  - critic: $Q_\omega$ nonlinear in $\omega$ (finite dimensions)
  - actor: $\pi_\theta$ nonlinear in $\theta$ (finite dimensions)
- causes of divergence: instability in practice
  - actor + critic: bilevel optimization (Pfau et al.; Heusel et al.)
  - critic: bias in TD update + nonlinearity of $Q_\omega$
  - actor: error in GD propagated from critic
- agenda: (i) TRPO/PPO (infinite dimensions) + (ii) TD/GD (finite dimensions)

# Challenges & Agenda

- two of deadly triad: nonlinearity + bootstrapping + off-policy (Sutton & Barto)
- sources of nonconvexity: "bad" stationary points
  - $J(\pi_\theta)$ nonconvex in $\pi_\theta \in \Delta^{|\mathcal{S}|}$ (infinite dimensions)
  - critic: $Q_\omega$ nonlinear in $\omega$ (finite dimensions)
  - actor: $\pi_\theta$ nonlinear in $\theta$ (finite dimensions)
- causes of divergence: instability in practice
  - actor + critic: bilevel optimization (Pfau et al.; Heusel et al.)
  - critic: bias in TD update + nonlinearity of $Q_\omega$
  - actor: error in GD propagated from critic
- agenda: (i) TRPO/PPO (infinite dimensions) + (ii) TD/GD (finite dimensions)

# Global Convergence of “Neural” TRPO/PPO

# An Infinite-Dimensional Optimization View

- ideal case: nonconvex infinite-dimensional mirror descent

$$\pi_{k+1} \leftarrow \arg\max_{\pi} \mathbb{E}_{s\sim\nu_k}\big[\langle Q^{\pi_k}(s,\cdot), \pi(\cdot,s)\rangle - \beta_k \mathrm{KL}(\pi(\cdot\,|\,s)\,\|\,\pi_k(\cdot\,|\,s))\big]$$

which factorizes across $\pi(\cdot\,|\,s) \in \Delta$ with $s \in \mathcal{S}$

- geometry via performance difference (Kakade & Langford)

$$0 \geq J(\pi) - J(\pi^*) = (1-\gamma)^{-1}\mathbb{E}_{s\sim\nu^*}\big[\langle \underbrace{Q^{\pi}(s,\cdot)}_{\text{dual}}, \underbrace{\pi(\cdot\,|\,s) - \pi^*(\cdot\,|\,s)}_{\text{primal}}\rangle\big]$$

where $J(\pi) := \mathbb{E}_{s\sim\nu^*}[V^{\pi}(s)] = \mathbb{E}_{s\sim\nu^*}[\langle Q^{\pi}(s,\cdot), \pi(\cdot\,|\,s)\rangle]$

- variational inequality view: $Q^{\pi}(s,\cdot)$ is one-point monotone (one-point convex/star convex/weakly quasiconvex)

# An Infinite-Dimensional Optimization View

- ideal case: nonconvex infinite-dimensional mirror descent

$$\pi_{k+1} \leftarrow \arg\max_{\pi} \mathbb{E}_{s\sim\nu_k}\big[\langle Q^{\pi_k}(s,\cdot), \pi(\cdot,s)\rangle - \beta_k \mathrm{KL}(\pi(\cdot\,|\,s)\,\|\,\pi_k(\cdot\,|\,s))\big]$$

which factorizes across $\pi(\cdot\,|\,s) \in \Delta$ with $s \in \mathcal{S}$

- geometry via performance difference (Kakade & Langford)

$$0 \geq J(\pi) - J(\pi^*) = (1-\gamma)^{-1}\mathbb{E}_{s\sim\nu^*}\big[\langle \underbrace{Q^{\pi}(s,\cdot)}_{\text{dual}}, \underbrace{\pi(\cdot\,|\,s) - \pi^*(\cdot\,|\,s)}_{\text{primal}}\rangle\big]$$

where $J(\pi) := \mathbb{E}_{s\sim\nu^*}[V^{\pi}(s)] = \mathbb{E}_{s\sim\nu^*}[\langle Q^{\pi}(s,\cdot), \pi(\cdot\,|\,s)\rangle]$

- variational inequality view: $Q^{\pi}(s,\cdot)$ is one-point monotone (one-point convex/star convex/weakly quasiconvex)

# An Infinite-Dimensional Optimization View

- ideal case: nonconvex infinite-dimensional mirror descent

$$\pi_{k+1} \leftarrow \arg\max_{\pi} \mathbb{E}_{s\sim\nu_k}\big[\langle Q^{\pi_k}(s,\cdot), \pi(\cdot,s)\rangle - \beta_k \mathrm{KL}(\pi(\cdot\,|\,s)\,\|\,\pi_k(\cdot\,|\,s))\big]$$

  which factorizes across $\pi(\cdot\,|\,s) \in \Delta$ with $s \in \mathcal{S}$

- geometry via performance difference (Kakade & Langford)

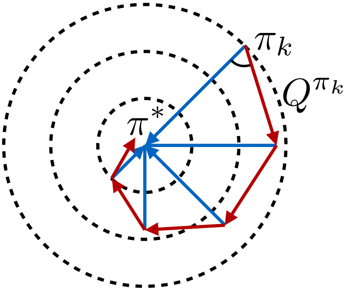
$$0 \geq J(\pi) - J(\pi^*) = (1-\gamma)^{-1}\mathbb{E}_{s\sim\nu^*}\big[\langle \underbrace{Q^{\pi}(s,\cdot)}_{\text{dual}}, \underbrace{\pi(\cdot\,|\,s) - \pi^*(\cdot\,|\,s)}_{\text{primal}}\rangle\big]$$

  where $J(\pi) := \mathbb{E}_{s\sim\nu^*}[V^{\pi}(s)] = \mathbb{E}_{s\sim\nu^*}[\langle Q^{\pi}(s,\cdot), \pi(\cdot\,|\,s)\rangle]$

- variational inequality view: $Q^{\pi}(s,\cdot)$ is one-point monotone (one-point convex/star convex/weakly quasiconvex)

# Primal-Dual Geometry: One-Point Monotone

$$0 \geq J(\pi) - J(\pi^*) = (1-\gamma)^{-1}\mathbb{E}_{s\sim\nu^*}\big[\langle \underbrace{Q^{\pi}(s,\cdot)}_{\text{dual}}, \underbrace{\pi(\cdot\,|\,s) - \pi^*(\cdot\,|\,s)}_{\text{primal}}\rangle\big]$$




infinite-dimensional mirror descent converges to global optimum!

# Proof Sketch: Sublinear Rate of Convergence

- primal-dual geometry via performance difference (Puterman)

$$0 \geq J(\pi) - J(\pi^*) = (1-\gamma)^{-1}\mathbb{E}_{s\sim\nu^*}\big[\langle Q^\pi(s,\cdot), \pi(\cdot\,|\,s) - \pi^*(\cdot\,|\,s)\rangle\big]$$

- one-step descent under one-point monotonicity

$$\begin{aligned} &2(1-\gamma)\beta_k^{-1}(J(\pi^*) - J(\pi_k)) \\ &\quad \leq 2\mathbb{E}_{s\sim\nu^*}\big[\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_k(\cdot\,|\,s)) - \mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{k+1}(\cdot\,|\,s))\big] \\ &\qquad + \beta_k^{-2}\mathbb{E}_{s\sim\nu^*}\big[\|Q^{\pi_k}(s,\cdot)\|_\infty^2\big] \end{aligned}$$

- telescoping one-step descent yields global convergence

$$\min_{k\in[K]}\{J(\pi^*) - J(\pi_k)\} \lesssim 1/\sqrt{K}$$

# Proof Sketch: Sublinear Rate of Convergence

- primal-dual geometry via performance difference (Puterman)

$$0 \geq J(\pi) - J(\pi^*) = (1-\gamma)^{-1}\mathbb{E}_{s\sim\nu^*}\big[\langle Q^\pi(s,\cdot), \pi(\cdot\,|\,s) - \pi^*(\cdot\,|\,s)\rangle\big]$$

- one-step descent under one-point monotonicity

$$\begin{aligned}
&2(1-\gamma)\beta_k^{-1}(J(\pi^*) - J(\pi_k)) \\
&\quad \leq 2\mathbb{E}_{s\sim\nu^*}\big[\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_k(\cdot\,|\,s)) - \mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{k+1}(\cdot\,|\,s))\big] \\
&\qquad + \beta_k^{-2}\mathbb{E}_{s\sim\nu^*}\big[\|Q^{\pi_k}(s,\cdot)\|_\infty^2\big]
\end{aligned}$$

- telescoping one-step descent yields global convergence

$$\min_{k\in[K]}\big\{J(\pi^*) - J(\pi_k)\big\} \lesssim 1/\sqrt{K}$$

# Proof Sketch: Sublinear Rate of Convergence

- primal-dual geometry via performance difference (Puterman)

$$0 \geq J(\pi) - J(\pi^*) = (1-\gamma)^{-1}\mathbb{E}_{s\sim\nu^*}\big[\langle Q^\pi(s,\cdot), \pi(\cdot\,|\,s) - \pi^*(\cdot\,|\,s)\rangle\big]$$

- one-step descent under one-point monotonicity

$$\begin{aligned} &2(1-\gamma)\beta_k^{-1}(J(\pi^*) - J(\pi_k)) \\ &\quad \leq 2\mathbb{E}_{s\sim\nu^*}\big[\mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_k(\cdot\,|\,s)) - \mathrm{KL}(\pi^*(\cdot\,|\,s)\,\|\,\pi_{k+1}(\cdot\,|\,s))\big] \\ &\qquad + \beta_k^{-2}\mathbb{E}_{s\sim\nu^*}\big[\|Q^{\pi_k}(s,\cdot)\|_\infty^2\big] \end{aligned}$$

- telescoping one-step descent yields global convergence

$$\min_{k\in[K]}\big\{J(\pi^*) - J(\pi_k)\big\} \lesssim 1/\sqrt{K}$$

# From Infinite Dimensions to Finite Dimensions

- parameterize actor $\pi_k$ (primal) & critic $Q^{\pi_k}$ (dual) as

$$\pi_{\theta_k}(a \,|\, s) \propto \exp(\tau^{-1} u_{\theta_k}(s,a)), \quad Q_{\omega_k} = u_{\omega_k}(s,a)$$

where $u_\alpha(s,a) = \dfrac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \sigma([\alpha]_i^\top (s,a))$

- incorporating primal & dual errors still gives $1/\sqrt{K}$ rate
  - policy improvement (actor): mean squared error of GD
  - policy evaluation (critic): mean squared Bellman error of TD

- unified analysis for GD & TD: focus on TD in the sequel

# From Infinite Dimensions to Finite Dimensions

- parameterize actor $\pi_k$ (primal) & critic $Q^{\pi_k}$ (dual) as

$$\pi_{\theta_k}(a \,|\, s) \propto \exp(\tau^{-1} u_{\theta_k}(s,a)), \quad Q_{\omega_k} = u_{\omega_k}(s,a)$$

where $u_\alpha(s,a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \sigma([\alpha]_i^\top (s,a))$

- incorporating primal & dual errors still gives $1/\sqrt{K}$ rate
  - policy improvement (actor): mean squared error of GD
  - policy evaluation (critic): mean squared Bellman error of TD

- unified analysis for GD & TD: focus on TD in the sequel

# From Infinite Dimensions to Finite Dimensions

- parameterize actor $\pi_k$ (primal) & critic $Q^{\pi_k}$ (dual) as

$$\pi_{\theta_k}(a \,|\, s) \propto \exp(\tau^{-1} u_{\theta_k}(s,a)), \quad Q_{\omega_k} = u_{\omega_k}(s,a)$$

$$\text{where } u_\alpha(s,a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \sigma([\alpha]_i^\top (s,a))$$

- incorporating primal & dual errors still gives $1/\sqrt{K}$ rate
  - policy improvement (actor): mean squared error of GD
  - policy evaluation (critic): mean squared Bellman error of TD

- unified analysis for GD & TD: focus on TD in the sequel

# Global Convergence of "Neural" TD

# TD: Bias & Nonlinearity Leads to Divergence

- policy evaluation by minimizing mean squared Bellman error

$$\min_{\omega} \mathrm{MSBE}(\omega) := \mathbb{E}_{(s,a)\sim\nu_k}\left[\left(Q_\omega(s,a) - \mathcal{T}^{\pi_{\theta_k}} Q_\omega(s,a)\right)^2\right]$$

where $\mathcal{T}^{\pi_{\theta_k}} Q(s,a) := \mathbb{E}_{s'\sim\mathcal{P}(\cdot\,|\,s,a),a'\sim\pi_{\theta_k}(\cdot\,|\,s')}\left[r(s,a) + \gamma Q(s',a')\right]$

- TD: stochastic semigradient descent (Sutton)

$$\omega(t{+}1) \leftarrow \omega(t) - \eta \underbrace{\left(Q_{\omega(t)}(s,a) - r(s,a) - \gamma Q_{\omega(t)}(s',a')\right)\nabla_\omega Q_{\omega(t)}(s,a)}_{\text{stochastic semigradient } g(t)}$$

where $(s,a)\sim\nu_k$, $s'\sim\mathcal{P}(\cdot\,|\,s,a)$, $a'\sim\pi_{\theta_k}(\cdot\,|\,s')$

- worse than nonconvex: bias + nonlinearity $\rightarrow$ divergence (Baird; Boyan & Moore; Tsitsiklis & Van Roy)

# TD: Bias & Nonlinearity Leads to Divergence

- policy evaluation by minimizing mean squared Bellman error

$$\min_{\omega} \mathrm{MSBE}(\omega) := \mathbb{E}_{(s,a)\sim\nu_k}\Big[\big(Q_\omega(s,a) - \mathcal{T}^{\pi_{\theta_k}} Q_\omega(s,a)\big)^2\Big]$$

where $\mathcal{T}^{\pi_{\theta_k}} Q(s,a) := \mathbb{E}_{s'\sim\mathcal{P}(\cdot\,|\,s,a),a'\sim\pi_{\theta_k}(\cdot\,|\,s')}\big[r(s,a) + \gamma Q(s',a')\big]$

- TD: stochastic semigradient descent (Sutton)

$$\omega(t{+}1) \leftarrow \omega(t) - \eta \underbrace{\big(Q_{\omega(t)}(s,a) - r(s,a) - \gamma Q_{\omega(t)}(s',a')\big)\nabla_\omega Q_{\omega(t)}(s,a)}_{\text{stochastic semigradient } g(t)}$$

where $(s,a)\sim\nu_k$, $s'\sim\mathcal{P}(\cdot\,|\,s,a)$, $a'\sim\pi_{\theta_k}(\cdot\,|\,s')$

- worse than nonconvex: bias + nonlinearity $\rightarrow$ divergence (Baird; Boyan & Moore; Tsitsiklis & Van Roy)

# TD: Bias & Nonlinearity Leads to Divergence

- policy evaluation by minimizing mean squared Bellman error

$$\min_{\omega} \mathrm{MSBE}(\omega) := \mathbb{E}_{(s,a)\sim\nu_k}\Big[\big(Q_\omega(s,a) - \mathcal{T}^{\pi_{\theta_k}} Q_\omega(s,a)\big)^2\Big]$$

where $\mathcal{T}^{\pi_{\theta_k}} Q(s,a) := \mathbb{E}_{s'\sim\mathcal{P}(\cdot\,|\,s,a),a'\sim\pi_{\theta_k}(\cdot\,|\,s')}\big[r(s,a) + \gamma Q(s',a')\big]$

- TD: stochastic semigradient descent (Sutton)

$$\omega(t+1) \leftarrow \omega(t) - \eta \underbrace{\big(Q_{\omega(t)}(s,a) - r(s,a) - \gamma Q_{\omega(t)}(s',a')\big)\nabla_\omega Q_{\omega(t)}(s,a)}_{\text{stochastic semigradient } g(t)}$$

where $(s,a)\sim\nu_k$, $s'\sim\mathcal{P}(\cdot\,|\,s,a)$, $a'\sim\pi_{\theta_k}(\cdot\,|\,s')$

- worse than nonconvex: bias + nonlinearity → divergence (Baird; Boyan & Moore; Tsitsiklis & Van Roy)

# Overparametrization Tames Divergence

- implicit linearization: linearize $u_\omega(s, a)$ as

$$v_\omega(s, a) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \, \mathbb{1}\Big\{[\omega(0)]_i^\top (s, a) > 0\Big\} [\omega]_i^\top (s, a)$$
$$\approx \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \, \mathbb{1}\Big\{[\omega]_i^\top (s, a) > 0\Big\} [\omega]_i^\top (s, a) := u_\omega(s, a)$$

- error of implicit linearization decreases in $m$

$$\mathbb{E}_{(s,a)\sim\nu_k, \omega(0)\sim N(0,I)} \Big[ \big| u_\omega(s, a) - v_\omega(s, a) \big|^2 \Big] \lesssim m^{-1/2}$$

  which implies error of implicit linearization in semigradient

- same role as explicit linearization in nonlinear gradient TD (Maei et al.) $\to$ TD converges to global optimum of MSBE!

# Overparametrization Tames Divergence

- implicit linearization: linearize $u_\omega(s,a)$ as

$$v_\omega(s,a) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \, \mathbb{1}\Big\{[\omega(0)]_i^\top(s,a) > 0\Big\} [\omega]_i^\top(s,a)$$
$$\approx \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \, \mathbb{1}\Big\{[\omega]_i^\top(s,a) > 0\Big\} [\omega]_i^\top(s,a) := u_\omega(s,a)$$

- error of implicit linearization decreases in $m$

$$\mathbb{E}_{(s,a)\sim\nu_k, \omega(0)\sim N(0,I)}\Big[\big|u_\omega(s,a) - v_\omega(s,a)\big|^2\Big] \lesssim m^{-1/2}$$

which implies error of implicit linearization in semigradient

- same role as explicit linearization in nonlinear gradient TD (Maei et al.) $\rightarrow$ TD converges to global optimum of MSBE!

# Overparametrization Tames Divergence

- implicit linearization: linearize $u_\omega(s,a)$ as

$$v_\omega(s,a) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \, \mathbb{1}\Big\{[\omega(0)]_i^\top (s,a) > 0\Big\} [\omega]_i^\top (s,a)$$

$$\approx \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \, \mathbb{1}\Big\{[\omega]_i^\top (s,a) > 0\Big\} [\omega]_i^\top (s,a) := u_\omega(s,a)$$

- error of implicit linearization decreases in $m$

$$\mathbb{E}_{(s,a)\sim\nu_k, \omega(0)\sim N(0,I)}\Big[\big|u_\omega(s,a) - v_\omega(s,a)\big|^2\Big] \lesssim m^{-1/2}$$

  which implies error of implicit linearization in semigradient

- same role as explicit linearization in nonlinear gradient TD (Maei et al.) $\rightarrow$ TD converges to global optimum of MSBE!

# Proof Sketch: Sublinear Rate of Convergence

- approximate one-point monotonicity via implicit linearization

$$\langle \mathbb{E}[g(t)], \omega(t) - \omega^* \rangle \geq (1-\gamma)\mathbb{E}\big[|u_{\omega(t)}(s,a) - u_{\omega^*}(s,a)|^2\big] + O(m^{-1/4})$$

- telescoping one-step descent yields global convergence of TD

$$\min_{t \in [T]} \mathrm{MSBE}(\omega(t)) \lesssim 1/\sqrt{T} + O(m^{-1/4})$$

with overparametrization: $m \to \infty$

- policy improvement (actor): same rate of convergence for MSE of GD $\to$ global convergence of TRPO/PPO

# Proof Sketch: Sublinear Rate of Convergence

- approximate one-point monotonicity via implicit linearization

$$\langle \mathbb{E}[g(t)], \omega(t) - \omega^* \rangle \geq (1-\gamma)\mathbb{E}\big[|u_{\omega(t)}(s,a) - u_{\omega^*}(s,a)|^2\big] + O(m^{-1/4})$$

- telescoping one-step descent yields global convergence of TD

$$\min_{t \in [T]} \mathrm{MSBE}(\omega(t)) \lesssim 1/\sqrt{T} + O(m^{-1/4})$$

with overparametrization: $m \to \infty$

- policy improvement (actor): same rate of convergence for MSE of GD $\to$ global convergence of TRPO/PPO

# Proof Sketch: Sublinear Rate of Convergence

- approximate one-point monotonicity via implicit linearization

$$\langle \mathbb{E}[g(t)], \omega(t) - \omega^* \rangle \geq (1-\gamma)\mathbb{E}\big[|u_{\omega(t)}(s,a) - u_{\omega^*}(s,a)|^2\big] + O(m^{-1/4})$$

- telescoping one-step descent yields global convergence of TD

$$\min_{t \in [T]} \mathrm{MSBE}(\omega(t)) \lesssim 1/\sqrt{T} + O(m^{-1/4})$$

with overparametrization: $m \to \infty$

- policy improvement (actor): same rate of convergence for MSE of GD $\rightarrow$ global convergence of TRPO/PPO

# Summary

# Summary: Global Convergence of TRPO/PPO

- ideal case: TRPO/PPO as infinite-dimensional mirror descent under one-point monotonicity
  - $1/\sqrt{K}$ rate of convergence to global optimal policy
- from infinite to finite dimensions: primal & dual errors of policy improvement (GD) & policy evaluation (TD)
  - TD: overparametrization tames divergence
  - $1/\sqrt{T}$ rate of convergence to global optimum of MSBE
  - similar rate of convergence for GD

$\rightarrow$ global convergence of TRPO/PPO

# Summary: Global Convergence of TRPO/PPO

- ideal case: TRPO/PPO as infinite-dimensional mirror descent under one-point monotonicity
  - $1/\sqrt{K}$ rate of convergence to global optimal policy
- from infinite to finite dimensions: primal & dual errors of policy improvement (GD) & policy evaluation (TD)
  - TD: overparametrization tames divergence
  - $1/\sqrt{T}$ rate of convergence to global optimum of MSBE
  - similar rate of convergence for GD

$\rightarrow$ global convergence of TRPO/PPO

# Summary: Global Convergence of TRPO/PPO

- ideal case: TRPO/PPO as infinite-dimensional mirror descent under one-point monotonicity
  - $1/\sqrt{K}$ rate of convergence to global optimal policy
- from infinite to finite dimensions: primal & dual errors of policy improvement (GD) & policy evaluation (TD)
  - TD: overparametrization tames divergence
  - $1/\sqrt{T}$ rate of convergence to global optimum of MSBE
  - similar rate of convergence for GD

$\rightarrow$ global convergence of TRPO/PPO

# Summary

- policy gradient and natural policy gradient?
  — Neural Policy Gradient Methods: Global Optimality and Rates of Convergence
  (joint work with Lingxiao Wang, Qi Cai, Zhuoran Yang)

- exploration for sample efficiency?
  — Provably Efficient Reinforcement Learning with Linear Function Approximation
  (joint work with Chi Jin, Zhuoran Yang, Micheal Jordan)