# Recent Advances in Reinforcement Learning Theory

**Yingbin Liang**, The Ohio State University
**Shaofeng Zou**, University at Buffalo, SUNY
**Yi Zhou**, University of Utah

2021 IEEE International Symposium on Information Theory

July 18, 2021

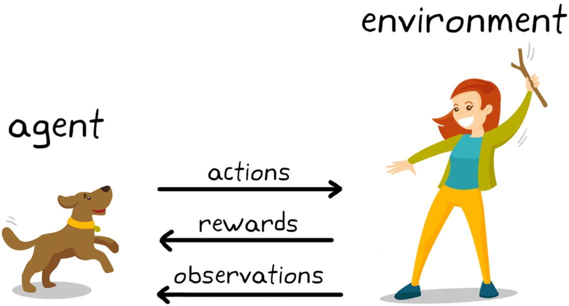
# Outline

# Outline

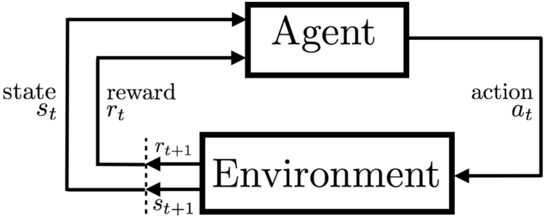
1. Introduction to Reinforcement Learning and Applications
2. Policy Evaluation and TD Learning
3. Value-based Method for Optimal Control
4. Policy Gradient Algorithms
5. Advanced Topics on RL and Open Directions

# Reinforcement Learning

- An agent learns to interact with environment in the best way
  - Agent observes state, and takes an action based on a policy
  - Environment changes the state
  - Agent receives a reward
  - Agent finds a policy to maximize reward



# Markov Decision Process (MDP)




- Markov decision process (MDP): $(\mathcal{S}, \mathcal{A}, r, \mathrm{P})$
  - $\mathcal{S}$ and $\mathcal{A}$: state and action spaces
  - $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: reward function
  - $\mathrm{P}(s'|s, a)$: transition kernel; prob of $s \to s'$ given action $a$
- Agent's policy $\pi(a|s)$: prob of selecting action $a$ in state $s$

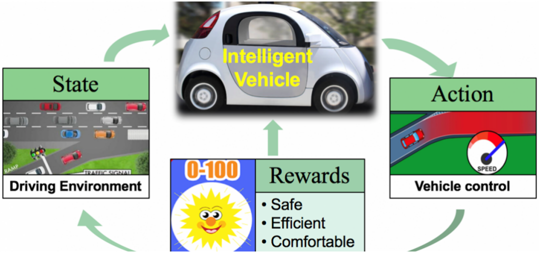
MDP trajectory $\{s_t, a_t, r_t, s_{t+1}\}_{t=0}^{\infty}$ defined by

$$s_0 \xrightarrow{\pi(\cdot|s_0)} a_0 \xrightarrow{\mathrm{P}(\cdot|s_0,a_0)} (s_1, r_0) \xrightarrow{\pi(\cdot|s_1)} a_1 \cdots$$

- Randomness: actions, state transitions

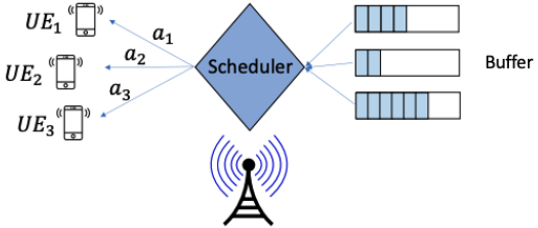
# Application: Autonomous Driving

- Collects driving data
- AI agent trained to optimize driving control
- Specification of MDP
  - State: driving environment (distance to nearby cars, weather, etc)
  - Action: turn left/right, accelerate, brake
  - Reward: stay safe, drive smoothly
  - Policy: vehicle control in a state



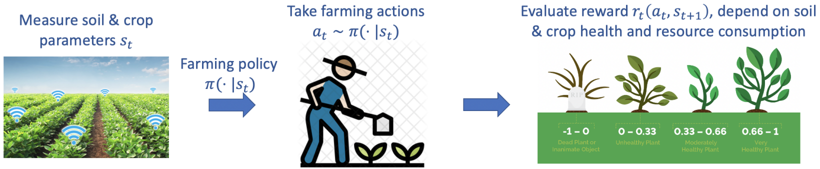
# Application: Wireless Communication

- Downlink Scheduling [1]
- Learn optimal scheduling to minimize average queuing delay
- Specification of MDP
  - State: buffer status and channel state
  - Action: assign resource block, determine number of transmitted bits
  - Reward: buffer cost
  - Policy: determine action in a given state



# Application: Agricultural Farming

- Collect data on crop & soil health
- Learn good farming policy to maximize yield
- Specification of MDP
  - State: crop & soil health
  - Action: apply amount of water & fertilizer
  - Reward: expected yield, crop & soil health
  - Farming policy: guide farming action in a state



# Formulation of RL

- MDP trajectory $\{s_t, a_t, r_t, s_{t+1}\}_t$ with $r_t := r(s_t, a_t, s_{t+1})$
- Quality of $s, a$: discount factor $\gamma \in (0, 1)$

$$\text{(State value):} \quad V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, \pi\right]$$
$$\text{(State-action value):} \quad Q_\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a, \pi\right]$$

- Expected long-term accumulated reward start with $s, a$

**RL Goal: find the best policy $\pi^*$**

$$\text{(Criterion I):} \quad V_{\pi^*}(s) \geq V_\pi(s), \quad \forall \pi, \forall s$$
$$\text{(Criterion II):} \quad \max_\pi J(\pi) := \mathbb{E}_{s \sim \xi}[V_\pi(s)]$$

Tutorial will not cover all the RL formulations

- Finite-time horizon, Average reward, Regret analysis

# Outline

1. Introduction to Reinforcement Learning and Applications
2. Policy Evaluation and TD Learning
3. Value-based Method for Optimal Control
4. Policy Gradient Algorithms
5. Advanced Topics on RL and Open Directions

# Formulation of Policy Evaluation

- Recall Markov Decision Process: $\{s_t, a_t, r_t, s_{t+1}\}_t$

$$s_0 \xrightarrow{\pi(\cdot|s_0)} a_0 \xrightarrow{\mathrm{P}(\cdot|s_0,a_0)} (s_1, r_0) \xrightarrow{\pi(\cdot|s_1)} a_1 \cdots$$

- State value function:

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, \pi\right]$$

  - Expected accumulated reward, start with $s$ follow $\pi$.

**Policy Evaluation Problem:**

Given a fixed policy $\pi$, how to evaluate its state value function $V_\pi$?

- Foundation for policy optimization

# Summary of Policy Evaluation Approaches

- Known transition kernel $P(\cdot|s,a)$
  - Solving Bellman equation
- Unknown transition kernel $P(\cdot|s,a)$ (Model-free)
  - On-policy TD learning
  - Off-policy TD learning

Our focus is model-free approaches.

# Known P: Bellman Equation

Transition kernel $\mathrm{P}(\cdot|s,a)$ is known

- By definition of $V_\pi(s)$:

$$\begin{aligned} V_\pi(s) &= \mathbb{E}[r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots | s_0 = s, \pi] \\ &= \mathbb{E}[r_0 | s_0 = s, \pi] + \gamma \mathbb{E}[r_1 + \gamma r_2 + \cdots | s_0 = s, \pi] \end{aligned}$$

- Note that

$$\begin{aligned} &\mathbb{E}[r_1 + \gamma r_2 + \cdots | s_0 = s, \pi] \\ &= \mathbb{E}_{s_1}\Big[\mathbb{E}[r_1 + \gamma r_2 + \cdots | s_0 = s, s_1 = s', \pi]\Big] \\ &= \mathbb{E}_{s_1}[V_\pi(s')] \end{aligned}$$

$$V_\pi(s) = \sum_{a,s'} \mathrm{P}(s'|s,a)\pi(a|s)\Big(r(s,a,s') + \gamma V_\pi(s')\Big)$$

$$V_\pi(s) = \sum_{a,s'} \mathrm{P}(s'|s,a)\pi(a|s)\Big(r(s,a,s') + \gamma V_\pi(s')\Big)$$

- Define Bellman operator

(**Bellman operator**):

$$\mathrm{T}_\pi V_\pi(s) = \sum_{a,s'} \mathrm{P}(s'|s,a)\pi(a|s)\Big(r(s,a,s') + \gamma V_\pi(s')\Big)$$

**Bellman Equation for Value Function**

$$V_\pi(s) = \mathrm{T}_\pi V_\pi(s)$$

- Linear programming: Directly solve the linear equation
  - High computation complexity
- Value iteration: fixed point update

$$V_{t+1}(s) = \mathrm{T}_\pi V_t(s)$$

  - $\mathrm{T}_\pi$ is contraction $\Rightarrow V_t \to V_\pi$.

# Model-Free: On-Policy TD Learning

## Model-Free

- Transition kernel $\mathrm{P}(\cdot|s, a)$ is unknown

## On-Policy Data

- Collect Markovian data $\{s_t, a_t, r_t, s_{t+1}\}_t$ following target policy $\pi$

# On-Policy TD(0) Algorithm

- Recall Bellman equation

$$V_\pi(s) = \mathbb{E}[r(s,a,s') + \gamma V_\pi(s')]$$

- Idea: update $V_\pi(s)$ using $r(s,a,s') + \gamma V_\pi(s')$
- Formally: collect $\{s_t, a_t, r_t, s_{t+1}\}_t$ and do

$$V(s_t) = \underbrace{r_{t+1} + \gamma V(s_{t+1})}_{\text{Target (one-step bootstrap)}}, \qquad (*)$$

- TD learning is a damped version of (*): $0 < \eta < 1$,

$$V(s_t) \leftarrow (1-\eta)V(s_t) + \eta\big(r_{t+1} + \gamma V(s_{t+1})\big), \qquad \text{(TD)}$$

**TD(0) Algorithm [2]**

$$V(s_t) \leftarrow V(s_t) + \eta\big(\underbrace{r_{t+1} + \gamma V(s_{t+1}) - V(s_t)}_{\text{temporal difference}}\big)$$

# TD($\lambda$) Algorithm

## TD(0) Algorithm

$$V(s_t) \leftarrow V(s_t) + \eta\big(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\big)$$

- In TD(0), target $r_{t+1} + \gamma V(s_{t+1})$ is one-step bootstrap
- Extension: $n$-step bootstrap

$$G_t^{(n)} := r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n V(s_{t+n})$$

- Define $\lambda$-return: $G_t^{\lambda} := (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$.

## TD($\lambda$) Algorithm [3]

$$V(s_t) \leftarrow V(s_t) + \eta\big(G_t^{\lambda} - V(s_t)\big)$$

- Reduce the variance of TD target

# Value Function Approximation

- **Curse of dimensionality:** state space is often large or infinite
- **Solution:** approximate $V_\pi$ using parameterized model $V_\theta$
  - Linear model: $V_\theta(s) := \phi_s^\top \theta$, where $\phi_s$ is feature vector of $s$
  - Neural model: $V_\theta(s) := \mathrm{NN}_\theta(s)$, where $\mathrm{NN}_\theta$ is neural network

### TD(0) learning with function approximation

- Initialize model $\theta_0$.
- Observe sample $\{s_t, a_t, r_t, s_{t+1}\}$, define target $G_t = r_t + \gamma V_{\theta_t}(s_{t+1})$
- Define loss $\ell_t(\theta) := \frac{1}{2}(V_\theta(s_t) - G_t)^2$, compute $g_t(\theta_t) = -\frac{\partial \ell_t(\theta)}{\partial \theta}|_{\theta=\theta_t}$
- TD update:

$$\theta_{t+1} = \theta_t + \eta g_t(\theta_t),$$

where $g_t(\theta_t) = (r_t + \gamma V_{\theta_t}(s_{t+1}) - V_{\theta_t}(s_t))\nabla V_{\theta_t}(s_t)$

# Analysis of TD(0) with Linear Approximation

**TD(0) with linear approximation $V_\theta(s) := \phi_s^\top \theta$**

$$\theta_{t+1} = \mathrm{Proj}_R\big(\theta_t + \eta g_t(\theta_t)\big),$$

$$\text{where } g_t(\theta_t) = (r_t + \gamma\phi_{s_{t+1}}^\top \theta_t - \phi_{s_t}^\top \theta_t)\phi_{s_t}$$

- **Challenge:** $g_t(\theta_t)$ is gradient of time-varying function $\ell_t$
- **Challenge:** Samples $\{s_t, a_t, r_t, s_{t+1}\}_t$ are Markovian and correlated

Non-exhaustive summary of existing work:

- Asymptotic convergence: [4, 5, 6, 7]
- Non-asymptotic (finite-time) convergence
  - I.I.D. samples: [8]
  - Markovian samples: [9], [10] (will be presented)

# Finite-Time Convergence of TD(0)

**Key Assumption: Geometric Mixing**

State stationary distribution $\mu$. There exist $\kappa > 0$, $\rho \in (0,1)$ such that

$$\sup_{s \in \mathcal{S}} \mathrm{d}_{TV}\big(\mathsf{P}(s_t | s_0 = s), \mu\big) \leq \kappa \rho^t, \quad \forall t \in \mathbb{N}_0$$

- Hold for irreducible and aperiodic Markov chains
- Given $s_0$ and large $t$, $s_t$ is almost like being sampled from $\mu$

- Feature matrix $\Phi = [\phi_{s_1}^\top; \dots; \phi_{s_n}^\top]$ full column rank, $V_\theta = \Phi\theta$
- Solution point $\theta^*$ satisfies [4]

$$V_{\theta^*} = \Pi_{\mathcal{L}} \mathsf{T}_\pi V_{\theta^*}, \quad \text{where } \mathcal{L} = \{\Phi x | x \in \mathbb{R}^d\}$$

**Theorem: finite-time convergence [10]**

Set learning rate $\eta \leq \mathcal{O}(\frac{1}{1-\gamma})$. After $T$ iterations,

$$\mathbb{E}\left[\|\theta_T - \theta^*\|^2\right] \leq \mathcal{O}\Big(\exp(-c\eta T)\|\theta_0 - \theta^*\|^2 + \eta\frac{\tau_{\text{mix}}(\eta)}{1-\gamma}\Big),$$

where $\tau_{\text{mix}}(\eta) := \min\{t \mid \kappa\rho^t \leq \eta\}$ is the mixing time of Markov chain.

- A faster mixing implies smaller convergence error

# Outline of Proof

- Recall TD(0): $\theta_{t+1} = \mathrm{Proj}_R\big(\theta_t + \eta g_t(\theta_t)\big)$
  - $g_t(\cdot)$ depends on sample $O_t = \{s_t, a_t, r_t, s_{t+1}\}$
- Define $\bar{g}(\theta) = \mathbb{E}[g_t(\theta)]$, where $\mathbb{E}$ over $O_t \sim \mathbb{P}(O_t)$
- Using the update rule yields

$$\mathbb{E}\big[\|\theta_{t+1} - \theta^*\|^2\big] \le \mathbb{E}\big[\|\theta_t - \theta^*\|^2\big] - 2\eta(1-\gamma)\mathbb{E}\big[\|V_{\theta_t} - V_{\theta^*}\|_D^2\big] + \eta\mathbb{E}\big[\underbrace{\langle g_t(\theta_t) - \bar{g}(\theta_t), \theta_t - \theta^*\rangle}_{\text{Bias } \zeta(\theta_t, O_t)}\big] + \mathcal{O}(\eta^2)$$

- can show $\mathbb{E}\big[\|V_{\theta_t} - V_{\theta^*}\|_D^2\big] \ge \sigma\|\theta_t - \theta^*\|^2$
- The key is to bound the bias term $\zeta(\theta_t, O_t)$
  - If all $O_t$ are i.i.d from $\mu$, then $\mathbb{P}(O_t|\theta_t) = \mathbb{P}(O_t) = \mu$ and

$$\mathbb{E}[g_t(\theta_t)|\theta_t] = \bar{g}(\theta_t) \quad \Rightarrow \quad \mathbb{E}[\zeta(\theta_t, O_t)] = 0$$

  - However, now samples are correlated. $\mathbb{P}(O_t|\theta_t) \neq \mathbb{P}(O_t)$

# Bounding the Bias

- Idea: $\mathbb{P}(O_t|\theta_{t-\tau})$ is close to $\mu$ due to geometric mixing

$$\begin{aligned}\zeta(\theta_t, O_t) &= \zeta(\theta_{t-\tau}, O_t) + \sum_{i=t-\tau}^{t-1} \zeta(\theta_{i+1}, O_t) - \zeta(\theta_i, O_t) \\ &\leq \zeta(\theta_{t-\tau}, O_t) + G^2\eta\tau\end{aligned}$$

- $\eta\tau$ can be controlled by using small learning rate $\eta$
- $\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)]$ is small due to geometric mixing

$$\begin{aligned}\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)] &\leq 2\|\zeta\|_\infty \sup_s \mathrm{d}_{TV}\big(\mathrm{P}(s_t|s_{t-\tau} = s), \mu\big) \\ &\leq 4G^2\kappa\rho^\tau\end{aligned}$$

# Putting Things Together

$$\mathbb{E}\left[\|\theta_{t+1} - \theta^*\|^2\right] \leq \mathbb{E}\left[\|\theta_t - \theta^*\|^2\right] - 2\eta(1-\gamma)\mathbb{E}\left[\|V_{\theta_t} - V_{\theta^*}\|^2\right] + \eta\mathbb{E}[\zeta(\theta_t, O_t)] + \eta^2 G^2$$

- $\mathbb{E}\left[\|V_{\theta_t} - V_{\theta^*}\|_D^2\right] \geq \sigma\|\theta_t - \theta^*\|^2$
- $\zeta(\theta_t, O_t) \leq \zeta(\theta_{t-\tau}, O_t) + G^2\eta\tau$
- $\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)] \leq 4G^2\kappa\rho^\tau$

# Connection to Linear SA

**Linear stochastic approximation (SA)**

$$\theta_{t+1} = \theta_t + \eta(A(O_t)\theta_t + b(O_t))$$

- $\{O_t\}_t$ forms a Markov chain
- $A(O_t), b(O_t)$ are matrix and vector
- TD(0) with linear approximation can be rewritten using

$$\begin{aligned} O_t &= (s_t, s_{t+1})^\top \\ A(O_t) &= -\phi_{s_t}(\phi_{s_t}^\top - \gamma\phi_{s_{t+1}}^\top) \\ b(O_t) &= r_t\phi_{s_t} \end{aligned}$$

- Convergence established using Lyapunov-type analysis [11]

# TD Learning for Off-Policy Evaluation

- Previous TD(0) uses on-policy data

**On-Policy Data**

Collect Markovian data $\{s_t, a_t, r_t, s_{t+1}\}_t$ following target policy $\pi$

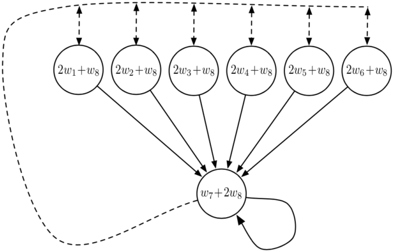
- Limitation: requires executing the target policy
- Limitation: in practice may not have sufficient on-policy data

**Off-policy data**

Collect Markovian data $\{s_t, a_t, r_t, s_{t+1}\}_t$ following behavior policy $\pi_b$. The goal is to evaluate $V_\pi$ of the target policy $\pi$.

# Divergence of Off-Policy TD(0)

**Key message:** TD(0) with linear approximation may diverge in the off-policy setting [12]

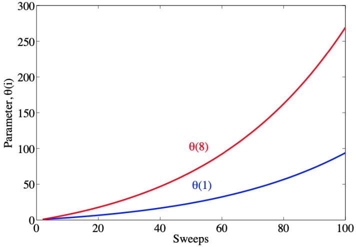



$\pi(\text{solid}|\cdot) = 1$

$b(\text{dashed}|\cdot) = 6/7$

$b(\text{solid}|\cdot) = 1/7$

$\gamma = 0.99$




- Zero reward, function approximation

$$V(s) = 2\theta(s) + \theta_0, \quad s = 1, ..., 6$$
$$V(7) = \theta(7) + 2\theta_0$$

- Under certain initialization, parameter diverges

# Gradient TD for Off-Policy Evaluation

- Recall $V_\theta(s) = \phi_s^\top \theta$. Optimal $\theta^*$ satisfies

$$V_{\theta^*} = \Pi_{\mathcal{L}} \mathsf{T}^\pi V_{\theta^*}$$

- Data sampled by behavior policy $\pi_b$, stationary distribution $\mu_b$

**Mean-square projected Bellman error (MSPBE) [13]**

$$\text{(MSPBE)}: \ J(\theta) := \mathbb{E}_{s\sim\mu_b}\left[V_\theta(s) - \Pi_{\mathcal{L}} \mathsf{T}^\pi V_\theta(s)\right]^2$$

- Error $V_\theta(s) - \Pi_{\mathcal{L}} \mathsf{T}^\pi V_\theta(s)$ based on target policy
- $\mathbb{E}_{s\sim\mu_b}$: stationary state distribution induced by behavior policy

# Idea of Importance Sampling

- Denote TD error $\delta_t(\theta) = r_t + \gamma\phi_{s_{t+1}}^\top\theta - \phi_{s_t}^\top\theta$
- MSPBE can be rewritten as

$$J(\theta) = \mathbb{E}_{\mu_b,\pi}[\delta_t(\theta)\phi_{s_t}]^\top \mathbb{E}_{\mu_b}[\phi_{s_t}\phi_{s_t}^\top]^{-1}\mathbb{E}_{\mu_b,\pi}[\delta_t(\theta)\phi_{s_t}]$$

## Importance Sampling Lemma

$$\mathbb{E}_{\mu_b,\pi}[\delta_t(\theta)\phi_{s_t}] = \mathbb{E}_{\mu_b,\pi_b}\left[\frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}\delta_t(\theta)\phi_{s_t}\right],$$

where $\rho_t = \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}$ is the importance sampling ratio. Then, we have

$$-\frac{1}{2}\nabla J(\theta) = \mathbb{E}\left[\rho_t(\phi_{s_t} - \gamma\phi_{s_{t+1}})\phi_{s_t}^\top\right]\mathbb{E}\left[\phi_{s_t}\phi_{s_t}^\top\right]^{-1}\mathbb{E}[\rho_t\delta_t(\theta)\phi_{s_t}]$$

# GTD2 Algorithm

$$-\frac{1}{2}\nabla J(\theta) = \mathbb{E}\left[\rho_t(\phi_{s_t} - \gamma\phi_{s_{t+1}})\phi_{s_t}^\top\right] \underbrace{\mathbb{E}\left[\phi_{s_t}\phi_{s_t}^\top\right]^{-1}\mathbb{E}[\rho_t\delta_t(\theta)\phi_{s_t}]}_{\omega^*(\theta)}$$

- $\omega^*(\theta)$ can be viewed as solution to the LMS

$$\text{(LMS): } \omega^*(\theta) = \operatorname*{argmin}_u \mathbb{E}\left[\phi_{s_t}^\top u - \rho_t\delta_t(\theta)\right]^2$$

**GTD2 algorithm [13]**

$$\theta_{t+1} = \theta_t + \alpha_t\rho_t(\phi_{s_t} - \gamma\phi_{s_{t+1}})\phi_{s_t}^\top\omega_t$$
$$\omega_{t+1} = \omega_t + \beta_t(\rho_t\delta_t(\theta_t)\phi_{s_t} - \phi_{s_t}\phi_{s_t}^\top\omega_t)$$

- Two timescale updates
- $\omega$ update is one-step SGD applied to LMS

# TDC Algorithm

$$-\frac{1}{2}\nabla J(\theta) = \mathbb{E}\big[\rho_t(\phi_{s_t} - \gamma\phi_{s_{t+1}})\phi_{s_t}^\top\big]\underbrace{\mathbb{E}\big[\phi_{s_t}\phi_{s_t}^\top\big]^{-1}\mathbb{E}[\rho_t\delta_t(\theta)\phi_{s_t}]}_{\omega^*(\theta)}$$

$$= \mathbb{E}\big[\rho_t\delta_t(\theta)\phi_{s_t}\big] - \gamma\mathbb{E}\big[\rho_t\phi_{s_{t+1}}\phi_{s_t}^\top\big]\omega^*(\theta)$$

**TDC algorithm [13]**

$$\theta_{t+1} = \theta_t + \alpha_t\rho_t(\delta_t(\theta_t)\phi_{s_t} - \gamma\phi_{s_{t+1}}\phi_{s_t}^\top\omega_t)$$
$$\omega_{t+1} = \omega_t + \beta_t(\rho_t\delta_t(\theta_t)\phi_{s_t} - \phi_{s_t}\phi_{s_t}^\top\omega_t)$$

- $\theta$ update is different from GTD2
- $\omega$ update is the same as GTD2

# Analysis of TDC with Linear Approximation

**TDC with linear approximation**

$$\theta_{t+1} = \Pi_{R_\theta}\left(\theta_t + \alpha_t \rho_t(\delta_t(\theta_t)\phi_{s_t} - \gamma \phi_{s_{t+1}} \phi_{s_t}^\top \omega_t)\right)$$
$$\omega_{t+1} = \Pi_{R_\omega}\left(\omega_t + \beta_t(\rho_t \delta_t(\theta_t)\phi_{s_t} - \phi_{s_t}\phi_{s_t}^\top \omega_t)\right)$$

- Challenge: Correlated Markovian samples
- Challenge: Correlated two timescale updates

Non-exhaustive of existing work:

- Asymptotic convergence: [13, 14, 15]
- Non-asymptotic (finite-time) convergence
  - I.I.D. samples: [8]
  - Markovian samples: [16], [17] (will be presented)

# Finite-Time Convergence of TDC

**Key Assumptions:**

- (Geometric mixing): There exist $\kappa > 0$, $\rho \in (0, 1)$ such that

$$\sup_{s \in \mathcal{S}} \mathrm{d}_{TV}\big(\mathrm{P}(s_t | s_0 = s), \mu\big) \leq \kappa \rho^t, \quad \forall t \in \mathbb{N}_0$$

- (Non-singularity): The following matrices are non-singular

$$A := \mathbb{E}_{\mu_b}[\rho_{s,a}(\gamma \phi_s \phi_{s'}^\top - \phi_s \phi_s^\top)], \quad C := -\mathbb{E}_{\mu_b}[\phi_s \phi_s^\top]$$

# Finite-Time Convergence of TDC

**Theorem: finite-time convergence [17]**

Set learning rates $\alpha < \frac{1}{|\lambda_{\max}(2A^\top C^{-1}A)|}, \beta < \frac{1}{|\lambda_{\max}(2C)|}$. After $T$ iterations,

$$\mathbb{E}\big[\|\theta_T - \theta^*\|^2\big] \le \mathcal{O}\Big((1-c\alpha)^t + \alpha \log \alpha^{-1} + \sqrt{\beta \log \beta^{-1} + \frac{\alpha}{\beta}}\Big)$$

- Need small $\alpha, \beta$ and $\frac{\alpha}{\beta}$
- Small $\frac{\alpha}{\beta}$: $\omega_t$ takes faster update than $\theta_t$, because it needs to approximate the double expectation in $\theta$ update

# Outline of Proof: Step 1

Rewrite TDC Update

- Recall that $\omega_t$ is used to approximate

$$\omega_t \to \omega^*(\theta) := \underbrace{\mathbb{E}\left[\phi_{s_t}\phi_{s_t}^\top\right]^{-1}\mathbb{E}[\rho_t\delta_t(\theta)\phi_{s_t}]}_{-C^{-1}(b+A\theta)}$$

- Define tracking error $z_t = \omega_t - \omega^*(\theta) = \omega_t + C^{-1}(b + A\theta)$
- TDC can be rewritten as: $O_t = (s_t, a_t, r_t, s_{t+1})$

$$\theta_{t+1} = \Pi_R\big(\theta_t + \alpha(f_1(\theta_t; O_t) + g_1(z_t; O_t))\big)$$
$$z_{t+1} = z_t + \beta(f_2(\theta_t; O_t) + g_2(z_t; O_t)) - \omega^*(\theta_t) + \omega^*(\theta_{t+1})$$

# Outline of Proof: Step 2

Develop bound of $\mathbb{E}[\|z_t\|^2]$

$$z_{t+1} = \Pi_R\big(z_t + \beta(f_2(\theta_t; O_t) + g_2(z_t; O_t)) - \omega^*(\theta_t) + \omega^*(\theta_{t+1})\big)$$

- Use $z_t$ update to develop a preliminary bound of $\mathbb{E}[\|z_t\|^2]$
  - Linear converging term, variance, bias, slow drift term
  - The proof uses constant bound of $\|z_t\|$
- Further use preliminary bound to develop refined bound
  - The proof uses preliminary bound of $\|z_t\|$

# Outline of Proof: Step 3

Develop bound of $\mathbb{E}[\|\theta_t - \theta^*\|^2]$

$$\theta_{t+1} = \Pi_R\big(\theta_t + \alpha(f_1(\theta_t; O_t) + g_1(z_t; O_t))\big)$$

- Use $\theta_t$ update and the refined bound of $\mathbb{E}[\|z_t\|^2]$

$$\leq (1-\alpha)\mathbb{E}[\|\theta_t - \theta^*\|^2] + 2\alpha\mathbb{E}[\zeta_{f_1}(\theta_t, O_t)] + \alpha^2 + 2\alpha\mathbb{E}\big[\|z_t\|^2 + \|\theta_t - \theta^*\|^2\big]$$

# Extension: Mini-batch TDC [18]

**Mini-batch TDC with linear approximation**

$$\theta_{t+1} = \theta_t + \frac{\alpha_t}{M} \sum_{i=tM}^{(t+1)M-1} \rho_i(\delta_i(\theta_t)\phi_{s_i} - \gamma\phi_{s_{i+1}}\phi_{s_i}^\top\omega_t)$$

$$\omega_{t+1} = \omega_t + \frac{\beta_t}{M} \sum_{i=tM}^{(t+1)M-1} (\rho_i\delta_i(\theta_t)\phi_{s_i} - \phi_{s_i}\phi_{s_i}^\top\omega_t)$$

- No need to use bounded projection
- Allow large constant learning rates
- Reduce variance of two timescale stochastic updates

# Outline

# Optimal Value/State-Action Value Function

- Recall definition of value and state-action value functions:

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \middle| s_0 = s, \pi\right]$$

$$Q_\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \middle| s_0 = s, a_0 = a, \pi\right]$$

- Goal: to find an optimal policy that maximizes the value function from any initial state $s_0$
- Optimal value function:

$$V^*(s) = \sup_\pi V_\pi(s), \forall s \in \mathcal{S}$$

- Optimal state-action value function:

$$Q^*(s, a) = \sup_\pi Q_\pi(s, a), \forall (s, a) \in \mathcal{S} \times \mathcal{A}$$

# Bellman Operator and Contraction

- Optimal policy $\pi^*$: take action $\arg\max_{a \in \mathcal{A}} Q^*(s, a)$ at state $s \in \mathcal{S}$
- $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a), \forall s \in \mathcal{S}$
- The Bellman operator T is defined as

$$(\mathrm{T}V)(s) = \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim \mathrm{P}(\cdot|s,a)} \left[ r(s, a, s') + \gamma V(s') \right]$$

- T is contraction: for any $V_1$ and $V_2$

$$\|\mathrm{T}V_1 - \mathrm{T}V_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$$

- $V^*$ is the fixed point of T: $V^* = \mathrm{T}V^*$

# Value Iteration

- Assume known reward $r$ and transition kernel $\mathrm{P}$

## Value Iteration

- Initialize $V(s)$ arbitrarily for any $s \in \mathcal{S}$
- Repeat until convergence
  - $V(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \mathrm{P}(s'|s,a)(r(s,a,s') + \gamma V(s'))$, for all $s \in \mathcal{S}$

- Repeatedly update $V(s)$ using Bellman operator, i.e, $V \leftarrow \mathrm{T}V$
- Convergence can be proved using contraction of $\mathrm{T}$
  - $\|\mathrm{T}V - V^*\|_\infty = \|\mathrm{T}V - \mathrm{T}V^*\|_\infty \leq \gamma \|V - V^*\|_\infty$
  - $\|\underbrace{\mathrm{T}\cdots\mathrm{T}}_{t \text{ times}} V - V^*\|_\infty \leq \gamma^t \|V - V^*\|_\infty \rightarrow 0$, as $t \rightarrow \infty$

# Policy Iteration

- Assume known reward $r$ and transition kernel P

## Policy Iteration

- Initialize $\pi$ arbitrarily
- Repeat until convergence
  - Evaluate $Q_\pi$
  - $\pi'(s) \leftarrow \arg\max_{a \in \mathcal{A}} Q_\pi(s, a)$ for all $s \in \mathcal{S}$
  - $\pi \leftarrow \pi'$

- **Policy improvement theorem**: Let $\pi$ and $\pi'$ be any pair of deterministic policies such that for all $s \in \mathcal{S}$, $Q_\pi(s, \pi'(s)) \geq V_\pi(s)$, then $\pi'$ is no worse than $\pi$: $V_{\pi'}(s) \geq V_\pi(s), \forall s \in \mathcal{S}$
- Policy from policy iteration has higher or same value than before

# SARSA: On-Policy TD Control

- Finite $\mathcal{S}$ and $\mathcal{A}$, unknown reward $r$ and transition kernel P

**SARSA**

- Parameter: step size $\alpha \in (0, 1]$, small $\epsilon > 0$
- Initialize $Q(s, a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ arbitrarily
- Initialize $s_0$ and $a_0$, $t = 0$
- Repeat until convergence
  - Observe state $s_{t+1}$, receive reward $r(s_t, a_t, s_{t+1})$
  - Take action $a_{t+1}$ using target policy derived from $Q$ (e.g., $\epsilon$-greedy)
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(\underbrace{r(s_t, a_t, s_{t+1}) + \gamma Q(s_{t+1}, a_{t+1})}_{\text{target}} - Q(s_t, a_t))$
  - $t \leftarrow t + 1$

- SARSA converges to $Q^*$ if
  - All state-action pairs are visited infinitely often
  - The policy converges to the greedy policy (e.g., $\epsilon$-greedy with $\epsilon = 1/t$)

# SARSA with Linear Function Approximation

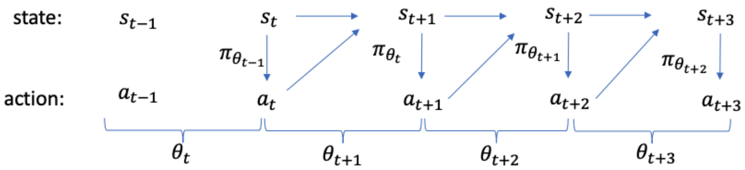
- Large $\mathcal{S}$ and $\mathcal{A}$, unknown $r$ and P

**SARSA**

- Initialization: $\theta_0$, $s_0$, $\phi_i$, for $i = 1, 2, ..., N$
- $\pi_{\theta_0} \leftarrow \Gamma(\phi^\top \theta_0)$ (e.g., $\epsilon$-greedy, softmax w.r.t. $\phi^\top \theta_0$)
- Choose $a_0$ according to $\pi_{\theta_0}$
- For $t = 0, 1, 2, \ldots$
  - Observe $s_{t+1}$ and $r(s_t, a_t, s_{t+1})$
  - Choose $a_{t+1}$ according to $\pi_{\theta_t}$
  - $\theta_{t+1} \leftarrow \theta_t + \alpha_t g_t(\theta_t)$
  - **Policy improvement**: $\pi_{\theta_{t+1}} \leftarrow \Gamma(\phi^\top \theta_{t+1})$

- $g_t(\theta_t) = \nabla_\theta Q_\theta(s_t, a_t)\Delta_t = \phi(s_t, a_t)\Delta_t$: "gradient"
- $\Delta_t$ denotes the temporal difference error at time t: $\Delta_t = r(s_t, a_t, s_{t+1}) + \gamma\phi^\top(s_{t+1}, a_{t+1})\theta_t - \phi^\top(s_t, a_t)\theta_t$,

# SARSA Sample Path




- As $\theta_t$ is updated, $\pi_{\theta_t}$ changes with time
- On-policy algorithm, time-varying policy
- Non-i.i.d. data

# Finite-Sample Analysis [20]

- The limit point $\theta^*$ of the projected SARSA [19]: $A_{\theta^*}\theta^* + b_{\theta^*} = 0$, where $A_{\theta^*} = \mathbb{E}_{\theta^*}[\phi(s,a)(\gamma\phi^T(s',a') - \phi^\top(s,a)]$ and $b_{\theta^*} = \mathbb{E}_{\theta^*}[\phi(s,a)r(s,a,s')]$
- The limiting point $\theta^*$ is the one such that $\mathbb{E}_{\theta^*}[g(\theta^*)] = 0$, where $s \sim \mu_{\pi_{\theta^*}}$, $a \sim \pi_{\theta^*}(\cdot|s)$

**Theorem**

- Finite-sample bound on convergence of SARSA with diminishing step-size: $\mathbb{E}\|\theta_T - \theta^*\|_2^2 \leq \mathcal{O}\left(\frac{\log T}{T}\right)$
- Finite-sample bound on convergence of SARSA with constant step-size: $\mathbb{E}\|\theta_T - \theta^*\|_2^2 \leq \mathcal{O}\left(e^{-cT}\right) + \mathcal{O}(\alpha)$

- With diminishing step-size, SARSA converges exactly to optimal $\theta^*$
- With constant step-size, SARSA converges exponentially fast to a small neighborhood of $\theta^*$

# Challenges in Technical Analysis

- Non-i.i.d. samples
  - Strong coupling between sample path $\{s_t, a_t\}_{t\geq 0}$ and $\{\theta_t\}_{t\geq 0}$
  - Samples are used to compute gradient $g_t$, and $\theta_{t+1}$, which introduce bias in $g_t$
  - $\theta_t$ is further used (as in policy $\pi_{\theta_t}$) to generate subsequent actions
- Convergence can be established using O.D.E approach [19]
- For finite-time bound, stochastic bias in $g_t$ needs to be explicitly characterized
- Dynamically changing learning policy
  - Analysis in [10] relies on the fact that the learning policy is fixed so that the Markov process reaches its stationary distribution quickly
  - Episodic SARSA in [21], with each episode, the learning policy is fixed, and the Markov process reaches its stationary distribution within each episode
  - No such nice properties for SARSA!

# Proof Sketch

- Step 1. Error decomposition
- Step 2. Gradient descent type analysis
- Step 3. Stochastic bias analysis
- Step 4. Putting the first three steps together and recursively apply step 1 completes the proof

**Key idea:**

Design an auxiliary uniformly ergodic Markov chain to approximate original Markov chain induced by SARSA

# Step 1. Error Decomposition

- Some notations
  - $\bar{g}(\theta) = \mathbb{E}_{\theta}[g_t(\theta)]$: noiseless gradient at $\theta$
  - $\Lambda_t(\theta) = \langle \theta - \theta^*, g_t(\theta) - \bar{g}(\theta) \rangle$: bias caused by using non-i.i.d. samples to estimate gradient
- Decompose error recursively:

$$
\begin{aligned}
&\mathbb{E}[\|\theta_{t+1} - \theta^*\|_2^2] \\
&\leq \underbrace{\mathbb{E}[\|\theta_t - \theta^*\|_2^2] + 2\alpha_t \mathbb{E}[\langle \theta_t - \theta^*, \bar{g}(\theta_t) - \bar{g}(\theta^*) \rangle] + \alpha_t^2 \mathbb{E}[\|g_t(\theta_t)\|_2^2]}_{\text{Gradient descent type analysis}} \\
&+ 2\alpha_t \underbrace{\mathbb{E}[\Lambda_t(\theta_t)]}_{\text{Stochastic bias}}
\end{aligned}
$$

# Step 2. Gradient Descent Type Analysis

- $\mathbb{E}[\|\theta_t - \theta^*\|_2^2] + 2\alpha_t \underbrace{\mathbb{E}[\langle \theta_t - \theta^*, \bar{g}(\theta_t) - \bar{g}(\theta^*)\rangle]}_{\text{term1}} + \underbrace{\alpha_t^2 \mathbb{E}[\|g_t(\theta_t)\|_2^2]}_{\sim\mathcal{O}(\alpha_t^2)}$
- True gradient $\bar{g}(\theta_t)$ is used, term 1 can be bounded:

$$\begin{aligned}\mathbb{E}[\langle \theta_t - \theta^*, \bar{g}(\theta_t) - \bar{g}(\theta^*)\rangle] &\leq (\theta_t - \theta^*)^T (A_{\theta^*} + C\lambda I)(\theta_t - \theta^*) \\ &\leq -w_s \mathbb{E}[\|\theta_t - \theta^*\|_2^2]\end{aligned}$$

  where $-w_s$ is the largest eigenvalue of $A_{\theta^*} + C\lambda I$
- $A_\theta$ is negative definite for all $\theta$

# Step 2. Some Assumptions

- Assumption (smooth policy): $\pi_\theta$ is Lipschitz with respect to $\theta$: $\forall (s,a) \in \mathcal{S} \times \mathcal{A}\ |\pi_{\theta_1}(a|s) - \pi_{\theta_2}(a|s)| \leq C\|\theta_1 - \theta_2\|_2$
- Assumption (non-singularity): C is small enough so that $A_{\theta^*} + C\lambda I$ is negative definite
- Assumption (geometric mixing): for fixed $\theta$, the Markov chain induced by $\pi_\theta$ and P is uniformly ergodic with invariant measure $\mu_\theta$, and there are constants $\kappa > 0$ and $\rho \in (0,1)$ such that

$$\sup_{s \in \mathcal{S}} \mathrm{d}_{TV}\big(\mathrm{P}(s_t|s_0 = s), \mu_\theta\big) \leq \kappa \rho^t, \quad \forall t \geq 0$$

# Step 3. Stochastic Bias Analysis

- $\mathbb{E}[\Lambda_t(\theta_t)]$: Bias caused by using a single sample path with non-i.i.d. data and dynamically changing learning policy $\pi_{\theta_t}$
- Define $O_t = (s_t, a_t, s_{t+1}, a_{t+1})$
- Recall stochastic bias: $\Lambda_t(\theta_t) = \langle \theta_t - \theta^*, g_t(\theta_t) - \bar{g}(\theta_t) \rangle$ and $g_t(\theta_t) = \phi(s_t, a_t) \left( r(s_t, a_t, s_{t+1}) + \gamma \phi^T(s_{t+1}, a_{t+1})\theta_t - \phi^T(s_t, a_t)\theta_t \right)$
- Complicated dependency between $O_t$ and $\theta_t$
- Rewrite $\Lambda_t(\theta_t)$ as $\Lambda_t(\theta_t, O_t)$

# Step 3. Stochastic Bias Analysis

- First, we show that $\Lambda_t(\theta)$ is Lipschitz in $\theta$
- Second, $\theta_t$ changes slowly with $t$
- Then for any $\tau > 0$,
  $\Lambda_t(\theta_t, O_t) \leq \Lambda_t(\theta_{t-\tau}, O_t) + (6 + \lambda C)G^2 \sum_{i=t-\tau}^{t-1} \alpha_i$ (part a)
- Intend to decouple the dependency between $\theta_t$ and $O_t$ by considering $\theta_{t-\tau}$ and $O_t$
- If the Markov chain induced by SARSA is uniformly ergodic, then given any $\theta_{t-\tau}$, $O_t$ would reach its stationary distribution quickly for large $\tau$
- However, this argument is not necessarily true since the policy $\pi_{\theta_t}$ changes with time

# Step 3. Stochastic Bias Analysis

- Key idea: design an auxiliary Markov chain that is uniformly ergodic to assist proof
- Auxiliary Markov chain design:
  - Before time $t-\tau+1$, everything is the same as SARSA
  - After that, fix learning policy as $\pi_{\theta_{t-\tau}}$ to generate all subsequent actions
  - Denote new observations as $\tilde{O}_t = (\tilde{s}_t, \tilde{a}_t, \tilde{s}_{t+1}, \tilde{a}_{t+1})$
- Since $\pi_{\theta_{t-\tau}}$ is kept fixed, for large $\tau$, $\tilde{O}_t$ reaches stationary distribution induced by $\pi_{\theta_{t-\tau}}$ by geometric mixing assumption for large $\tau$
- Thus, $\mathbb{E}[\Lambda_t(\theta_{t-\tau}, \tilde{O}_t)] \leq 4G^2\kappa\rho^{\tau-1}$ (part b)

# Step 3. Stochastic Bias Analysis

- Bound different between SARSA Markov chain and auxiliary Markov chain
- $\theta_t$ changes slowly due to small stepsize
- Due to Lipschitz property of $\pi_\theta$, the two Markov chains should not deviate from each other too much
- We can show $\mathbb{E}[\Lambda_t(\theta_{t-\tau}, O_t)] - \mathbb{E}[\Lambda_t(\theta_{t-\tau}, \tilde{O}_t)] \leq \frac{C|\mathcal{A}|G^3\tau}{w} \log \frac{t}{t-\tau}$ (part c)
- Combining parts a, b and c yields an upper bound on $\mathbb{E}[\Lambda_t(\theta_t)]$

# Step 4.

- Putting the first three steps together
- Recursively applying Step 1 completes the proof

# Q-Learning: Off-Policy TD Control

- Finite $\mathcal{S}$ and $\mathcal{A}$, unknown $r$ and P

## Q-Learning

- Parameter: step size $\alpha \in (0, 1]$
- Initialize $Q(s, a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ arbitrarily
- Initialize $s_0$, behavior policy $\pi_b$, $t = 0$
- Repeat until convergence
  - Take action $a_t$ following fixed $\pi_b$, observe next state $s_{t+1}$, receive reward $r(s_t, a_t, s_{t+1})$
  - $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r(s_t, a_t, s_{t+1}) + \gamma \max_{a' \in \mathcal{A}} Q(s_{t+1}, a') - Q(s_t, a_t))$
  - $t \leftarrow t + 1$

- Q-learning converges to $Q^*$ if
  - All state-action pairs are visited infinitely often
- Q-learning sample complexity studies, e.g., [22], [23] and [24]

# Gradient TD Method for Optimal Control

- Q-learning with function approximation suffers from divergence issue
- Greedy-Gradient Q-learning (Greedy-GQ) with linear function approximation [25]
- Consider mean squared projected Bellman error (MSPBE):

$$J(\theta) \triangleq \|\Pi \mathsf{T} Q_\theta - Q_\theta\|_\mu^2$$

  - $\mu$: stationary distribution induced by behavior policy $\pi_b$
  - $\|Q(\cdot,\cdot)\|_\mu \triangleq \int_{s\in\mathcal{S},a\in\mathcal{A}} d\mu_{s,a} Q(s,a)$
  - $\Pi$: projection operator $\Pi\hat{Q} = \arg\min_{Q\in\mathcal{Q}} \|Q - \hat{Q}\|_\mu$
  - $\mathcal{Q} = \left\{Q_\theta = \phi^\top\theta : \theta \in \mathbb{R}^N\right\}$

**Goal:** $\min_\theta J(\theta)$

# Two Time-Scale Update Rule

- Define $\bar{V}_{s'}(\theta) = \max_{a' \in \mathcal{A}} \theta^\top \phi_{s',a'}$
- TD error: $\delta_{s,a,s'}(\theta) = r(s,a,s') + \gamma \bar{V}_{s'}(\theta) - \theta^\top \phi_{s,a}$
- Let $\hat{\phi}_{s'}(\theta) = \nabla \bar{V}_{s'}(\theta)$. Then gradient of MSPBE is

$$\frac{\nabla J(\theta)}{2} = -\mathbb{E}_\mu[\delta_{s,a,s'}(\theta)\phi_{s,a}] + \gamma \mathbb{E}_\mu[\hat{\phi}_{s'}(\theta)\phi_{s,a}^\top]\omega^*(\theta),$$

  where $\omega^*(\theta) = \mathbb{E}_\mu[\phi_{s,a}\phi_{s,a}^\top]^{-1}\mathbb{E}_\mu[\delta_{s,a,s'}(\theta)\phi_{s,a}]$.
- Double-sampling issue for estimating $\mathbb{E}_\mu[\hat{\phi}_{s'}(\theta)\phi_{s,a}^\top]\omega^*(\theta)$: it involves product of two expectations
- Weight doubling trick [13]:

$$\text{Slow time-scale: } \theta_{t+1} = \theta_t + \alpha(\delta_{t+1}(\theta_t)\phi_t - \gamma(\omega_t^\top \phi_t)\hat{\phi}_{t+1}(\theta_t)),$$
$$\text{Fast time-scale: } \omega_{t+1} = \omega_t + \beta(\delta_{t+1}(\theta_t) - \phi_t^\top \omega_t)\phi_t,$$

# Finite-Sample Analysis [26, 27]

**Challenges:**

- Non-convex objective $J(\theta)$ with two time-scale update rule
- Non-smooth due to max in $\bar{V}_{s'}(\theta) = \max_{a' \in \mathcal{A}} \theta^\top \phi_{s',a'}$
  - Approximate max with a smooth approximation, e.g., softmax
- Biased gradient estimate due to two time-scale update and Markovian noise

**Theorem**

Finite-sample bound on convergence of Greedy-GQ with linear function approximation: $\mathbb{E}[\|\nabla J(\theta_W)\|^2] = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$

- Gradient norm converges to 0 implies convergence to stationary points

# Finite-Sample Analysis [26, 27]

- $\omega^*(\theta)$: limit of fast time-scale if $\theta_t$ is fixed to be $\theta$
- Define tracking error: $z_t = \omega_t - \omega^*(\theta_t)$: how fast the fast time-scale tracks its limit
- Denote estimate of $\frac{\nabla J(\theta)}{2}$ by $G_{t+1}(\theta, \omega) = \delta_{t+1}(\theta)\phi_t - \gamma(\omega^\top \phi_t)\hat{\phi}_{t+1}(\theta)$
- Slow time-scale can be written as $\theta_{t+1} = \theta_t + \alpha G_{t+1}(\theta_t, \omega_t)$

# Stochastic Bias in Gradient Estimate

- Bias in the gradient estimate can be decomposed as follows:

$$\mathbb{E}\left[G_{t+1}(\theta_t, \omega_t) + \frac{\nabla J(\theta)}{2}\right]$$

$$= \underbrace{\mathbb{E}\left[G_{t+1}(\theta_t, \omega^*(\theta_t)) + \frac{\nabla J(\theta)}{2}\right]}_{\text{Bias (a): due to Markovian noise}} + \underbrace{\mathbb{E}\left[G_{t+1}(\theta_t, \omega_t) - G_{t+1}(\theta_t, \omega^*(\theta_t))\right]}_{\text{Bias (b): due to tracking error}}$$

- For bias (a), under the i.i.d. setting, it is zero. Under the Markovian setting, it can be bounded similarly to proof of TDC.
- For bias (b), $\|G_{t+1}(\theta_t, \omega_t) - G_{t+1}(\theta_t, \omega^*(\theta_t))\| \leq L\|\underbrace{\omega_t - \omega^*(\theta_t)}_{z_t}\|$,

  for some Lipschitz constant $L > 0$. Thus, a tight bound on the tracking error $\|z_t\|$ is needed.

# Tracking Error Bound

- $z_t$ can be recursively written as
  $z_{t+1} = z_t + \beta((\delta_{t+1}(\theta_t) - \phi_t^\top \omega^*(\theta_t))\phi_t - \phi_t^\top z_t \phi_t + \omega^*(\theta_t) - \omega^*(\theta_{t+1})$
- Then the recursion of $\|z_t\|^2$ naturally involves a term $\langle z_t, \omega^*(\theta_t) - \omega^*(\theta_{t+1})\rangle$, to bound which, the Taylor expansion of $\omega^*(\theta)$ at $\theta_t$ is used:

$$
\begin{aligned}
&\omega^*(\theta_{i+1}) - \omega^*(\theta_i) \\
&= \nabla\omega^*(\theta_i)^\top(\theta_{i+1} - \theta_i) + \mathcal{O}(\alpha^2) \\
&= \alpha\nabla\omega^*(\theta_i)^\top \underbrace{G_{i+1}(\theta_i, \omega_i)}_{\text{should also converge to 0}} + \mathcal{O}(\alpha^2)
\end{aligned}
$$

- Basic idea: bound tracking error $z_t$ in terms of $\nabla J(\theta_t)$, which shall also converges to zero, instead of a constant bound

# Variance Reduced Greedy-GQ [29]

- Greedy-GQ update: denote $O_t = (s_t, a_t, r_t, s_{t+1})$

$$\theta_{t+1} = \theta_t - \alpha G_{O_t}(\theta_t, \omega_t), \quad \omega_{t+1} = \omega_t - \beta H_{O_t}(\theta_t, \omega_t)$$

- Variance reduction [28]: reference parameters $\widetilde{\theta}$, $\widetilde{\omega}$

(Reference updates) $\widetilde{G} := \frac{1}{M}\sum_{i=1}^{M} G_{O_i}(\widetilde{\theta}, \widetilde{\omega}), \ \widetilde{H} := \frac{1}{M}\sum_{i=1}^{M} H_{O_i}(\widetilde{\theta}, \widetilde{\omega})$

(Variance-reduced Greedy-GQ):

$$\theta_{t+1} = \theta_t - \alpha\big(G_{O_t}(\theta_t, \omega_t) - G_{O_t}(\widetilde{\theta}, \widetilde{\omega}) + \widetilde{G}\big)$$

$$\omega_{t+1} = \omega_t - \beta\big(H_{O_t}(\theta_t, \omega_t) - H_{O_t}(\widetilde{\theta}, \widetilde{\omega}) + \widetilde{H}\big)$$

- Periodically update $\widetilde{\theta}, \widetilde{\omega}, \widetilde{G}, \widetilde{H}$
- Improved sample complexity

# Outline

# Formulation of RL

- State value function:

$$V_\pi(s) = \mathbb{E}[\textstyle\sum_{t=0}^\infty \gamma^t r(s_t, a_t, s_{t+1}) | s_0 = s, \pi]$$

- State-action value function:

$$Q_\pi(s, a) = \mathbb{E}[\textstyle\sum_{t=0}^\infty \gamma^t r(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a, \pi]$$

  where $a_t \sim \pi(\cdot|s_t)$ for all $t \geq 0$.

- Average value function:

$$J(\pi) = (1-\gamma)\mathbb{E}[\textstyle\sum_{t=0}^\infty \gamma^t r(s_t, a_t, s_{t+1})] = \mathbb{E}_{s\sim\xi}[V_\pi(s)]$$

  where $\xi(\cdot)$ denotes initial distribution.

**RL Goal: find the best policy $\pi^*$**

$$\text{(Criterion I):} \quad V_{\pi^*}(s) \geq V_\pi(s), \quad \forall \pi, \forall s$$

$$\text{(Criterion II):} \quad \max_\pi J(\pi) := \mathbb{E}_{s\sim\xi}[V_\pi(s)]$$

# Parameterization of Policy

- Central idea:
  - Parameterize the policy as $\{\pi_w, w \in \mathcal{W}\}$
  - $J(\pi) = J(\pi_w) := J(w)$

Goal of Policy-Based RL: $\max_{w\in\mathcal{W}} J(\pi_w) := J(w)$

- Example parameterizations of policy
  - Direct parameterization: $\pi_w(a|s) = w_{s,a}$, where $w \in \triangle(\mathcal{A})^{|\mathcal{S}|}$, i.e., $w_{s,a} \geq 0$, and $\sum_{a\in\mathcal{A}} w_{s,a} = 1$ for all $(s, a)$
  - Tabular softmax parameterization:

$$\pi_w(a|s) = \frac{\exp(w_{s,a})}{\sum_{a'\in\mathcal{A}} \exp(w_{s,a'})}$$

  - Linear softmax parameterization:

$$\pi_w(a|s) \propto \exp(\phi(s, a)^T w)$$

  - Gaussian policy: $\pi_w(a|s) = \mathcal{N}(\phi(s)^T w, \sigma^2)$

# Policy Gradient Algorithm

Goal of Policy-Based RL: $\max_{w \in \mathcal{W}} J(\pi_w) := J(w)$

- Policy gradient $\nabla J(w)$ [30]

$$\nabla_w J(w) = \mathbb{E}_{\nu_{\pi_w}} \left[ Q_{\pi_w}(s,a) \nabla_w \log \pi_w(a|s) \right]$$

  - Visitation distribution: $\nu_\pi(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a)$
  - Define score function $\psi_w(s,a) := \nabla_w \log \pi_w(a|s)$
  - Define advantage function: $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$

$$\nabla_w J(w) = \mathbb{E}_{\nu_{\pi_w}} \left[ Q_{\pi_w}(s,a) \psi_w(s,a) \right] = \mathbb{E}_{\nu_{\pi_w}} \left[ A_{\pi_w}(s,a) \psi_w(s,a) \right]$$

**Policy gradient algorithm [30, 31]**

- update the parameter $w$ via gradient ascent

$$w_{t+1} = w_t + \alpha_t \nabla_w J(w_t)$$

where $\alpha_t > 0$ is the stepsize.

# TRPO/PPO Algorithm

## Trusted Region Policy Optimization (TRPO) [32]

- Update the parameter $w$ under KL constraint

$$w_{t+1} = \operatorname*{argmax}_{w}[J(w_t) + (w - w_t)^T \nabla_w J(w_t)]$$
$$\text{subject to} \quad \mathbb{E}_{\nu(s)}[KL(\pi_{w_t} || \pi_w)] \leq c$$

where $c > 0$ is a hyperparameter.

## Proximal Policy Optimization (PPO) [33]

- Update the parameter $w$ via KL-regularized gradient ascent

$$w_{t+1} = \operatorname*{argmax}_{w}[J(w_t) + (w - w_t)^T \nabla_w J(w_t) - \alpha \mathbb{E}_{\nu_w(s)}[KL(\pi_{w_t} || \pi_w)]]$$

where $c > 0$ is a hyperparameter.

# Natural Policy Gradient (NPG) Algorithm

- Second-order Taylor approximation to KL distance

$$KL(\pi_{w_t}||\pi_w) \approx \frac{1}{2}(w - w_t)^T F(w)(w - w_t)$$

  - Fisher information matrix $F(w) = \mathbb{E}_{\nu_{\pi_w}}[\nabla_w \log \pi_{w_t} \nabla_w \log \pi_{w_t}^T]$
- KL-regularized update: at time $t$

$$\begin{aligned}
&\operatorname*{argmax}_w \left[J(w_t) + (w - w_t)^T \nabla_w J(w_t) - \alpha \mathbb{E}_{\nu_w(s)}[KL(\pi_{w_t}||\pi_w)]\right] \\
&\approx \operatorname*{argmax}_w \left[J(w_t) + (w - w_t)^T \nabla_w J(w_t) - \frac{\alpha}{2}(w - w_t)^T F(w_t)(w - w_t)\right] \\
&= w_t + \alpha F(w_t)^\dagger \nabla_w J(w_t)
\end{aligned}$$

where $F(w_t)^\dagger$ denotes the pseudo-inverse of $F(w_t)$.

### Natural Policy Gradient (NPG) [34]

- Update parameter $w$ via KL approximator based regularizer

$$w_{t+1} = w_t + \alpha F(w_t)^\dagger \nabla_w J(w_t)$$

# Convergence with Exact Policy Gradient

- Policy gradient
  - Direct and tabular softmax policy: global sublinear convergence [35]
  - Direct policy: global linear convergence via regularized MDP [36]
  - Direct policy: global linear convergence via line search [37]
- TRPO/PPO
  - Direct policy: global sublinear convergence via adaptivity [38]
  - Direct policy: global linear convergence via regularized MDP [36]
  - Direct policy: global convergence via line search [37]
- NPG
  - Tabular softmax policy: global sublinear convergence [35]
  - Tabular softmax policy: global linear convergence via regularized MDP [39]

# Policy Gradient Algorithms under Unknown MDP

$$\nabla J(w) = \mathbb{E}_{\nu_{\pi_w}}\left[Q_{\pi_w}(s,a)\psi_w(s,a)\right] = \mathbb{E}_{\nu_{\pi_w}}\left[A_{\pi_w}(s,a)\psi_w(s,a)\right]$$

- Let $\hat{P}(\cdot|s_t,a_t) = \gamma\mathbb{P}(\cdot|s_t,a_t) + (1-\gamma)\xi(\cdot)$ [40]
  - $\xi(\cdot)$: initial distribution
  - Samples drawn from $\hat{P}(\cdot|s_t,a_t)$ converge to visitation distribution $\nu_{\pi_w}$

## Model-free Policy Gradient

- Sample $s_t \sim \hat{P}(\cdot|s_{t-1},a_{t-1}), a_t \sim \pi_{w_t}(\cdot|s_t)$
- Unbiased estimation of $A_{\pi_{w_t}}(s_t,a_t)$
  - Sample a length-$K$ trajectory starting at $(s_t,a_t)$, $K \sim \text{Geom}(1-\gamma)$
  - Estimate $\hat{Q}(s_t,a_t)$ by adding rewards over the sample path
  - Sample a length-$K$ trajectory starting at $(s_t)$, $K \sim \text{Geom}(1-\gamma)$
  - Estimate $\hat{V}(s_t)$ by adding rewards over the sample path
  - $\hat{A}_{\pi_{w_t}}(s_t,a_t) = \hat{Q}(s_t,a_t) - \hat{V}(s_t)$
- Estimate policy gradient $g_t = \hat{A}_{\pi_{w_t}}(s_t,a_t)\nabla_{w_t}\log(\pi_{w_t}(a_t|s_t))$
- Update $w_{t+1} = w_t + \alpha_t g_t$

# Analysis of Model-free PG Algorithms

- Parameterization: general *nonlinear* policy $\{\pi_w : w \in \mathcal{W}\}$
- Sampling is over a single trajectory path

### Assumption 1 (Smoothness of policy)

*For any* $(w, w')$ *and* $(s, a)$*, there exist positive* $L_\psi$*,* $C_\psi$*, and* $C_\pi$ *such that:*

- $\|\psi_w(s,a) - \psi_{w'}(s,a)\|_2 \leq L_\psi \|w - w'\|_2$
- $\|\psi_w(s,a)\|_2 \leq C_\psi$
- $\mathrm{d}_{TV}(\pi_w(\cdot|s), \pi_{w'}(\cdot|s)) \leq C_\pi \|w - w'\|_2$

### Assumption 2 (Geometric Mixing)

*For any policy* $\pi_w$ *and transition kernel* $\mathrm{P}(\cdot|s,a)$ *or* $\hat{\mathrm{P}}(\cdot|s,a)$*, let* $\mu_{\pi_w}$ *be stationary distribution. There exist* $\kappa > 0$*,* $\rho \in (0,1)$ *such that*

$$\sup_{s \in \mathcal{S}} \mathrm{d}_{TV}\left(\mathbb{P}(s_t|s_0 = s), \mu_{\pi_w}\right) \leq \kappa \rho^t, \quad \forall t \geq 0$$

# Convergence of Model-free PG Algorithms

**Theorem ([41])**

*Suppose Assumptions 1 and 2 hold. Under a diminishing stepsize* $\alpha_t = \frac{1}{\sqrt{t}}$ *for* $t = 1, \dots, T$*, the output of model-free PG satisfies*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla_{w_t} J(w_t)\|^2\right] \leq \mathcal{O}\left(\frac{\log T}{T}\right) + \mathcal{O}\left(\frac{\log^2 T}{\sqrt{T}}\right).$$

*Furthermore, under constant stepsize* $\alpha_t = \alpha$*:*

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla_{w_t} J(w_t)\|^2\right] \leq \mathcal{O}\left(\frac{1}{\alpha T}\right) + \mathcal{O}(\alpha \log^2 \tfrac{1}{\alpha}).$$

- Under constant stepsize, PG converges to a neighborhood of a stationary point at a rate of $\mathcal{O}\left(\frac{1}{T}\right)$.
  - $\alpha$ controls a tradeoff between convergence rate and accuracy
  - Decreasing $\alpha$ improves accuracy, but slows down convergence
  - Let $\alpha_t = \frac{1}{\sqrt{T}}$, PG converges with a rate of $\mathcal{O}\left(\frac{\log^2 T}{\sqrt{T}}\right)$

# Actor-Critic Algorithms [42]

## Actor-Critic Algorithm

- Critic
  - Estimates $V_\theta(s)$ by linear function approximation $\phi(s)^\top\theta$
  - Takes $T_c$ length-$M$ minibatch TD learning updates and outputs $\theta_t$
- Actor
  - Approximates $A_{\pi_w}(s,a)$ by temporal difference error $\delta_\theta(s,a,s')$

$$\hat{A}_{\pi_w}(s,a) = \delta_\theta(s,a,s') = r(s,a,s') + \gamma\phi(s')^\top\theta - \phi(s)^\top\theta$$

  - Estimate policy gradient $v_t(\theta_t)$ by averaging $\delta_{\theta_t}(s_t,a_t,s_{t+1})\psi_{w_t}(s_t,a_t)$ over a length-$B$ sample trajectory
  - Updates $w_{t+1} = w_t + \alpha_t v_t(\theta_t)$

- Parameterization: general *nonlinear* policy $\{\pi_w : w \in \mathcal{W}\}$
- Sampling is over a single trajectory path

# Convergence Rate of Actor-Critic Algorithm

**Theorem ([43])**

*Suppose Assum. 1 and 2 hold, and $\hat{T}$ is chosen uniformly from $\{1, \cdots, T\}$.*

$$\mathbb{E}[\|\nabla_w J(w_{\hat{T}})\|_2^2] \leq \mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}\left(\frac{1}{B}\right) + (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c} + \mathcal{O}\left(\frac{\beta}{M}\right) + \mathcal{O}(\zeta^{critic}_{approx}).$$

*With total sample complexity $\mathcal{O}(\epsilon^{-2}\log(1/\epsilon))$*

$$\mathbb{E}[\|\nabla_w J(w_{\hat{T}})\|_2^2] \leq \epsilon + \mathcal{O}(\zeta^{critic}_{approx}).$$

- Actor has sublinear convergence, and critic has linear convergence
- Actor's bias and variance $\mathcal{O}\left(\frac{1}{B}\right)$; Critic's bias and variance $\mathcal{O}\left(\frac{\beta}{M}\right)$
- Critic's approximation error: $\zeta^{\text{critic}}_{\text{approx}} = \max_{w\in\mathcal{W}} \mathbb{E}_{\nu_w}[|V_{\pi_w}(s) - V_{\theta^*_{\pi_w}}(s)|^2]$

# Convergence Rate of Actor-Critic Algorithm

### Theorem ([43])

*Suppose Assum. 1 and 2 hold, and* $\hat{T}$ *is chosen uniformly from* $\{1, \cdots, T\}$.

$$\mathbb{E}[\|\nabla_w J(w_{\hat{T}})\|_2^2] \leq \mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}\left(\frac{1}{B}\right) + (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c} + \mathcal{O}\left(\frac{\beta}{M}\right) + \mathcal{O}(\zeta_{approx}^{critic}).$$

*With total sample complexity* $\mathcal{O}(\epsilon^{-2}\log(1/\epsilon))$

$$\mathbb{E}[\|\nabla_w J(w_{\hat{T}})\|_2^2] \leq \epsilon + \mathcal{O}(\zeta_{approx}^{critic}).$$

- Actor's mini-batch yields faster convergence rate of $\mathcal{O}(1/T)$ rather than $\mathcal{O}(1/\sqrt{T})$
- This further yields better overall sample complexity

# Proof of Convergence

- Let $v_t(\theta)$ denote estimator of $g(\theta, w) = \mathbb{E}_{\nu_w}[A_\theta(s,a)\psi_w(s,a)]$
- Decompose error terms

$$\begin{aligned}
&\Big(\frac{1}{2}\alpha - L_J\alpha^2\Big)\mathbb{E}[\|\nabla_w J(w_t)\|_2^2 \,|\mathcal{F}_t] \\
&\leq \mathbb{E}[J(w_{t+1})|\mathcal{F}_t] - J(w_t) + 3\Big(\frac{1}{2}\alpha + L_J\alpha^2\Big)\mathbb{E}\Big[\,\big\|v_t(\theta_t) - v_t(\theta^*_{w_t})\big\|_2^2 \\
&\quad + \big\|v_t(\theta^*_{w_t}) - g(\theta^*_{w_t}, w_t)\big\|_2^2 + \big\|g(\theta^*_{w_t}, w_t) - \nabla_w J(w_t)\big\|_2^2 \,\big|\mathcal{F}_t\Big].
\end{aligned}$$

- Error due to TD learning

$$\begin{aligned}
&\mathbb{E}[\big\|v_t(\theta_t) - v_t(\theta^*_{w_t})\big\|_2^2 \,|\mathcal{F}_t] \\
&\leq 4\mathbb{E}[\big\|\theta_t - \theta^*_{w_t}\big\|_2^2 \,|\mathcal{F}_t] \leq (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c} + \mathcal{O}(\beta/M)
\end{aligned}$$

# Proof of Convergence (Cont.)

- Gradient estimation error under Markovian minibatch sampling

$$\mathbb{E}\left[\left\|v_t(\theta^*_{w_t}) - g(\theta^*_{w_t}, w_t)\right\|_2^2 |\mathcal{F}_t\right] \le \mathcal{O}\left(\frac{1}{B}\right).$$

- Critic's approximation error

$$\left\|g(\theta^*_{w_t}, w_t) - \nabla_w J(w_t)\right\|_2^2 \le \mathcal{O}\left(\zeta^{\text{critic}}_{\text{approx}}\right)$$

- Combine error bounds and take summarization over iteration path

$$\mathbb{E}[\left\|\nabla_w J(w_{\hat{T}})\right\|_2^2] \le \mathcal{O}\left(\tfrac{1}{T}\right) + (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c} + \mathcal{O}\left(\tfrac{\beta}{M}\right) + \mathcal{O}\left(\tfrac{1}{B}\right) + \mathcal{O}(\zeta^{\text{critic}}_{\text{approx}}).$$

# Natural Policy Gradient under Unknown MDP

- Natural policy gradient (NPG) [34, 44],

$$w_{t+1} = w_t + \alpha_t F(w_t)^{\dagger} \nabla J(w_t)$$

- Consider $\min_{\theta \in \mathbb{R}^d} L_w(\theta) = \mathbb{E}_{\nu_{\pi_w}} [A_{\pi_w}(s,a) - \psi(s,a)^\top \theta]^2$
  - Minimum norm solution satisfies $\theta_w = F(w)^{\dagger} \nabla J(w)$
- NPG update [35]: $w_{t+1} = w_t + \alpha_t \theta_t$

**Model-free NPG [35]**

- At step $t$, solve least square problem via $K$ iterations
  - Obtain unbiased estimator $\hat{A}_{\pi_{w_t}}(s_k, a_k)$ (same as PG)
  - Update $\theta_{k+1} = \theta_k - \beta \nabla_\theta L_{w_t}(\theta_k)$
- Update $w_{t+1} = w_t + \alpha_t \theta_K$

- NPG with general nonlinear policy converges globally as $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ [35]
- Can achieve $\mathcal{O}\left(\frac{1}{T}\right)$ by self-variance reduction of gradient norm [43]

# Natural Actor-Critic Algorithm

$$J(w) = \mathbb{E}_{\nu_{\pi_w}}\left[Q_{\pi_w}(s,a)\psi_w(s,a)\right] = \mathbb{E}_{\nu_{\pi_w}}\left[A_{\pi_w}(s,a)\psi_w(s,a)\right]$$
$$w_{t+1} = w_t + \alpha_t F(w_t)^{\dagger}\nabla J(w_t)$$

## Natural Actor-Critic Algorithm

- Critic (same as critic in actor-critic algorithm)
  - Estimates $V_\theta(s)$ by linear function approximation $\phi(s)^\top\theta$
  - Takes $T_c$ length-$M$ minibatch TD learning updates and outputs $\theta_t$
- Actor
  - Computes policy gradient estimator $v_t(\theta_t)$ as in actor-critic algorithm
  - Computes Fisher information estimator $F_t(w_t)$ by averaging over a length-$B$ sample trajectory
  - Updates $w_{t+1} = w_t + \alpha_t F_t(w_t)^{\dagger} v_t(\theta_t)$

- Parameterization: general *nonlinear* policy $\{\pi_w : w \in \mathcal{W}\}$
- Sampling is over a single trajectory path

# Convergence Rate of Natural Actor-Critic Algorithm

**Theorem ([43])**

*Let Assum. 1 and 2 hold and $\hat{T}$ is chosen uniformly from $\{1, \cdots, T\}$.*

$$J(\pi^*) - \mathbb{E}\big[J(\pi_{w_{\hat{T}}})\big] \le \mathcal{O}\left(\tfrac{1}{T}\right) + \mathcal{O}\left(\tfrac{1}{\sqrt{B}}\right) + (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c/2} + \mathcal{O}\left(\tfrac{1}{\sqrt{M}}\right)$$
$$+ \mathcal{O}\left(\sqrt{\zeta^{\text{critic}}_{\text{approx}}}\right) + \mathcal{O}\left(\tfrac{1}{B}\right) + (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c} + \mathcal{O}\left(\tfrac{\beta}{M}\right) + \mathcal{O}(\zeta^{\text{critic}}_{\text{approx}}) + \mathcal{O}\left(\sqrt{\zeta^{\text{actor}}_{\text{approx}}}\right)$$

*With total sample complexity $\mathcal{O}(\epsilon^{-3}\log(1/\epsilon))$, we achieve*

$$J(\pi^*) - \mathbb{E}\big[J(\pi_{w_{\hat{T}}})\big] \le \epsilon + \mathcal{O}\left(\sqrt{\zeta^{\text{actor}}_{\text{approx}}}\right) + \mathcal{O}\left(\sqrt{\zeta^{\text{critic}}_{\text{approx}}}\right) + \mathcal{O}\left(\zeta^{\text{critic}}_{\text{approx}}\right).$$

- Actor has sublinear convergence, and critic has linear convergence
- Critic's approx. error: $\zeta^{\text{critic}}_{\text{approx}} = \max_{w\in\mathcal{W}} \mathbb{E}_{\nu_w}[|V_{\pi_w}(s) - V_{\theta^*_{\pi_w}}(s)|^2]$
- Actor's approx. error: $\zeta^{\text{actor}}_{\text{approx}} = \max_{w\in\mathcal{W}} \min_{p\in\mathbb{R}^{d_2}} \mathbb{E}_{\nu_{\pi_w}}\left[\psi_w(s,a)^\top p - A_{\pi_w}(s,a)\right]^2$

# Convergence Rate of Natural Actor-Critic Algorithm

**Theorem ([43])**

*Let Assum. 1 and 2 hold and $\hat{T}$ is chosen uniformly from $\{1, \cdots, T\}$.*

$$J(\pi^*) - \mathbb{E}\big[J(\pi_{w_{\hat{T}}})\big] \le \mathcal{O}\left(\tfrac{1}{T}\right) + \mathcal{O}\left(\tfrac{1}{\sqrt{B}}\right) + (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c/2} + \mathcal{O}\left(\tfrac{1}{\sqrt{M}}\right)$$
$$+ \mathcal{O}(\sqrt{\zeta^{\text{critic}}_{\text{approx}}}) + \mathcal{O}\left(\tfrac{1}{B}\right) + (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c} + \mathcal{O}\left(\tfrac{\beta}{M}\right) + \mathcal{O}(\zeta^{\text{critic}}_{\text{approx}}) + \mathcal{O}(\sqrt{\zeta^{\text{actor}}_{\text{approx}}})$$

*With total sample complexity $\mathcal{O}(\epsilon^{-3}\log(1/\epsilon))$, we achieve*

$$J(\pi^*) - \mathbb{E}\big[J(\pi_{w_{\hat{T}}})\big] \le \epsilon + \mathcal{O}\left(\sqrt{\zeta^{\text{actor}}_{\text{approx}}}\right) + \mathcal{O}\left(\sqrt{\zeta^{\text{critic}}_{\text{approx}}}\right) + \mathcal{O}\left(\zeta^{\text{critic}}_{\text{approx}}\right).$$

- Diminishing variance in actor's update yields a faster convergence rate of $\mathcal{O}(1/T)$ than $\mathcal{O}(1/\sqrt{T})$
- Performance difference lemma [35] of NAC yields global convergence

# Proof of Convergence (Part I)

- Define $u_{w_t} = F(w_t)^{-1}\nabla_w J(w_t)$ and $u_t(\theta) = F_t(w_t)^{-1}v_t(\theta)$, where $F_t(w_t)$ is assumed to be nonsingular.
- Bound the norm of policy gradient

$$\begin{aligned}&\mathbb{E}[\|\nabla_w J(w_t)\|_2^2 \,|\mathcal{F}_t] \\ &\leq \mathcal{O}\left(\mathbb{E}[J(w_{t+1})|\mathcal{F}_t] - J(w_t)\right) + \mathcal{O}\left(\mathbb{E}[\left\|u_t(\theta_t) - F(w_t)^{-1}\nabla_w J(w_t)\right\|_2^2 \,|\mathcal{F}_t]\right)\end{aligned}$$

- Bound estimation error of natural policy gradient

$$\begin{aligned}&\mathbb{E}[\left\|u_t(\theta_t) - F(w_t)^{-1}\nabla_w J(w_t)\right\|_2^2 \,|\mathcal{F}_t] \\ &\leq \mathcal{O}\left(\mathbb{E}[\|v_t(\theta_t) - \nabla_w J(w_t)\|_2^2 \,|\mathcal{F}_t]\right) + \mathcal{O}\left(\mathbb{E}[\|F(w_t) - F_t(w_t)\|_2^2 \,|\mathcal{F}_t]\right) \\ &\leq (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c} + \mathcal{O}\left(\frac{\beta}{M}\right) + \mathcal{O}\left(\frac{1}{B}\right) + \mathcal{O}(\zeta_{\text{approx}}^{\text{critic}})\end{aligned}$$

# Proof of Convergence (Part I Cont.)

- Policy gradient estimation error due to TD learning (same as AC)

$$\mathbb{E}[\|v_t(\theta_t) - \nabla_w J(w_t)\|_2^2 \,|\mathcal{F}_t] \le (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c} + \mathcal{O}\left(\frac{\beta}{M}\right) + \mathcal{O}\left(\frac{1}{B}\right) + \mathcal{O}(\zeta_{\text{approx}}^{\text{critic}})$$

- Fisher information estimation error

$$\mathbb{E}[\|F(w_t) - F_t(w_t)\|_2^2 \,|\mathcal{F}_t] \le \mathcal{O}\left(\frac{1}{B}\right)$$

- Overall convergence of gradient norm

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_w J(w_t)\|_2^2] \le \mathcal{O}\left(\frac{\mathbb{E}[J(w_T)] - J(w_0)}{T}\right) + (1 - \mathcal{O}(\lambda_{A_\pi}\beta))^{T_c} + \mathcal{O}\left(\frac{\beta}{M}\right) + \mathcal{O}\left(\frac{1}{B}\right) + \mathcal{O}(\zeta_{\text{approx}}^{\text{critic}})$$

# Proof of Convergence (Part II)

- Define $D(w) = KL\left(\pi^*(\cdot|s) \| \pi_w(\cdot|s)\right) = \mathbb{E}_{\nu_{\pi^*}}\left[\log \frac{\pi^*(a|s)}{\pi_w(a|s)}\right]$
- Bound function value gap (global convergence)

$$\begin{aligned} & D(w_t) - D(w_{t+1}) \\ & \geq \alpha \mathbb{E}_{\nu_{\pi^*}}\left[A_{\pi_{w_t}}(s,a)\right] - \alpha \left\| u_t(\theta_t) - F(w_t)^{-1} \nabla_w J(w_t) \right\|_2 \\ & \quad + \alpha \mathbb{E}_{\nu_{\pi^*}}\left[\psi_{w_t}(s,a)^\top F(w_t)^{-1} \nabla_w J(w_t) - A_{\pi_{w_t}}(s,a)\right] - \frac{L_\psi}{2} \alpha^2 \left\| u_t(\theta_t) \right\|_2^2 \end{aligned}$$

- Performance difference lemma (central for global convergence) [35]

$$\mathbb{E}_{\nu_{\pi^*}}[A_{\pi_{w_t}}(s,a)] = (1-\gamma)[J(\pi^*) - J(\pi_{w_t})]$$

- Natural policy gradient estimation error

$$\mathbb{E}[\left\| u_t(\theta_t) - F(w_t)^{-1} \nabla_w J(w_t) \right\|_2] \leq \sqrt{\mathbb{E}[\| u_t(\theta_t) - F(w_t)^{-1} \nabla_w J(w_t) \|_2^2]}$$

which is bounded in Part I.

# Proof of Convergence (Part II Cont.)

- Actor's approximation error

$$\mathbb{E}_{\nu_{\pi^*}}\left[\psi_{w_t}(s,a)^\top F(w_t)^{-1}\nabla_w J(w_t) - A_{\pi_{w_t}}(s,a)\right] \geq -\sqrt{\frac{1}{1-\gamma}\left\|\frac{\nu_{\pi^*}}{\nu_{\pi_{w_0}}}\right\|_\infty}\sqrt{\zeta^{\text{actor}}_{\text{approx}}}$$

- Second moment of policy gradient

$$\begin{aligned}&\mathbb{E}[\,\|u_t(\theta_t)\|_2^2] \\ &\quad\leq \mathcal{O}(\mathbb{E}[\left\|u_t(\theta_t) - F(w_t)^{-1}\nabla_w J(w_t)\right\|_2^2]) + \mathcal{O}(\mathbb{E}[\|\nabla_w J(w_t)\|_2^2])\end{aligned}$$

where both terms are bounded in Part I.

- Substitute all bounds into the first step of Part II, rearrange terms, and take summation over $t = 0$ to $T-1$.

# Extension I: Policy Gradient Algorithm with Adam

## PG-AMSGrad [41]

- Sample $s_t \sim \hat{P}(\cdot|s_{t-1}, a_{t-1}), a_t \sim \pi_{w_t}(\cdot|s_t)$
- Estimate Q-function $\hat{Q}_{\pi_{w_t}}(s_t, a_t)$ as in PG
- Estimate policy gradient $g_t = \hat{Q}_{\pi_{w_t}}(s_t, a_t)\nabla_{w_t}\log(\pi_{w_t}(a_t|s_t))$
- $m_t = (1-\beta_1)m_{t-1} + \beta_1 g_t$ momentum
- $v_t = (1-\beta_2)\hat{v}_{t-1} + \beta_2 g_t^2$ stepsize adaptation
- $\hat{v}_t = \max(\hat{v}_{t-1}, v_t),\ \hat{V}_t = diag(\hat{v}_{t,1}, \ldots, \hat{v}_{t,d})$
- Update policy parameter $w_{t+1} = w_t - \alpha_t \hat{V}_t^{-\frac{1}{2}} m_t$

- Convergence rate of PG-AMSGrad [41]
- In practice, PG with Adam converges much faster

# Extension II: Off-Policy Policy Gradient Algorithms

- Off-policy policy gradient
  - On-policy sampling with target policy is not possible
  - Off-policy sampling under behavior policy: $(s_i, a_i, s_i') \sim \mathcal{D}$
  - Estimate $\nabla_w J(w)$ with off-policy samples

**Actor-critic with distribution correction (AC-DC)**

$$g(w) = \hat{\rho}(s,a)\hat{Q}_{\pi_w}(s,a)\nabla_w \log(\pi_w(s,a))$$

where $\hat{\rho}$ and $\hat{Q}_{\pi_w}$ are approximation of $\rho = \nu_{\pi_w}/\mathcal{D}$ and $Q_{\pi_w}$, respectively.

- Bias error of AC-DC suffers substantially from estimation errors

$$\Delta_g = \mathbb{E}_{\mathcal{D}}[g(w)] - \nabla_w J(\pi_w) = \Theta(\mathbb{E}[\varepsilon_\rho(s,a) + \varepsilon_Q(s,a)])$$

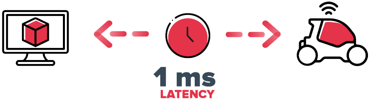
where $\varepsilon_\rho = \rho - \hat{\rho}$ and $\varepsilon_Q = Q - \hat{Q}$

- Doubly robust off-policy PG estimation [45] reduces bias error

# Outline

# Safe Reinforcement Learning

- Practical RL applications involve various safety/resource constraints
  - Left: Power constraint on battery powered devices
  - Right: Safety constraints on autonomous robotics and vehicles
  - Bottom: Delay constraint in communication system



# Constrained Markov Decision Process (CMDP)

- Same dynamics as general MDP
- Agent receives reward $R$ and cost $C$
- Value function w.r.t. reward $R$:

$$V_R^\pi(\rho) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t R(s_t, a_t, s_{t+1}) \middle| S_0 \sim \rho\right]$$

- Value function w.r.t. cost $C$:

$$V_C^\pi(\rho) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t C(s_t, a_t, s_{t+1}) \middle| S_0 \sim \rho\right]$$

**Goal of CMDP**

$$\max_\pi \quad V_R^\pi(\rho) \quad \text{subject to} \quad V_C^\pi(\rho) \leq c$$

# Two Popular Approaches for CMDP

- Primal-Dual Approach: e.g. CPO [46], PDO [47]
  - Define Lagrangian: let $\lambda > 0$ be Lagrangian multiplier

$$\mathcal{L}(\pi, \lambda) = -V_R^{\pi}(\rho) + \lambda(V_C^{\pi}(\rho) - c).$$

  - Solve a minimax problem over augmented Lagrangian function

$$\max_{\lambda \in \mathbb{R}_+} \min_{\pi} \mathcal{L}(\pi, \lambda)$$

  - Zero duality gap [48, 49]; Convergence rate [49, 50]
- Primal Approach: CRPO [51]
  - If constraint is violated, take one step NPG update to reduce ${V_C}^{\pi_t}(\rho)$
  - If constraint is satisfied, take one step NPG update to enlarge ${V_R}^{\pi_t}(\rho)$
  - Convergence rate [51]

# Imitation Learning

- Imitation Learning
  - Reward function is unknown
  - Some expert demonstrations are available
  - Goal: find a learner's policy that produces behaviors as close as possible to expert demonstrations
- Two major approaches
  - Behavioral cloning [52]
    - ★ Directly provides a mapping from state to action based on supervised learning to match expert demonstrations
  - Inverse Reinforcement Learning [53, 54]
    - ★ First recovers unknown reward function based on expert's trajectories, and then find an optimal policy using such a reward function
    - ★ Generative adversarial imitation learning (GAIL) framework [55]

# Generative Adversarial Imitation Learning (GAIL)

- Parameterize reward function as $r_\alpha(s, a)$ where $\alpha \in \Lambda \subset \mathbb{R}^q$
- $\pi_E$: expert policy; demonstration samples under $\pi_E$ are available
- $\pi_L$: learner's policy to be optimized
- $J(\pi_E, r_\alpha)$: average value function under expert policy
- $J(\pi_L, r_\alpha)$: average value function under learner's policy
- $\psi(\alpha)$: regularizer of reward parameter

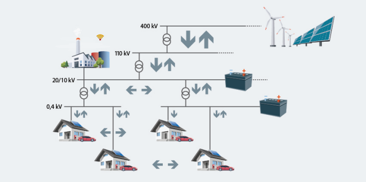
**GAIL Framework [55]**

$$\min_{\pi_L} \max_{\alpha \in \Lambda} F(\pi_L, \alpha) := J(\pi_E, r_\alpha) - J(\pi_L, r_\alpha) - \psi(\alpha)$$

- Maximization: find reward function that best distinguishes between expert's and learner's policies
- Minimization: find learner's policy that matches expert's policy as close as possible

# Multi-Agent Reinforcement Learning (MARL)

- Many RL applications involve multiple agents
  - Left: stock market with numerous investors
  - Middle: multi-drone control
  - Bottom: multi-agent power network



# Formulation of MARL

- State value function (of joint policy $\pi$):

$$V_\pi(s) = \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t \frac{1}{M} \sum_{m=1}^{M} r_t^{(m)} | s_0 = s, \pi\big]$$

- Average value function:

$$J(\pi) = (1-\gamma)\mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t \frac{1}{M} \sum_{m=1}^{M} r_t^{(m)}\big] = \mathbb{E}_\xi[V_\pi(s)]$$

**MARL Problem:**

$$\max_{\{\pi^{(m)}\}_m} J(\pi)$$

- MARL algorithms are similar to single-agent RL algorithms
- Agents need synchronize information (local state observations, actions, rewards, etc)
- Tradeoff between communication & computation complexities

# Open Problems in Reinforcement Learning

- Multi-task reinforcement learning
  - Tasks can share similar but different transition kernels
  - Meta-learning can be applied to achieve sampling efficiency
  - Open issues in theory: characterization of sample complexity improvement due to meta-learning
- Off-policy/Offline reinforcement learning
  - No access to online interaction with environment, but access only to a given set of data samples
  - Dataset has limited coverage over state-action space, and is sampled under behavior policy, not target policy
  - Open issues in design: how to design desirable algorithms to address overestimation and distribution shift
  - Open issues in theory: what is the minimum requirement to achieve polynomial sample complexity efficiency

# Open Problems (Cont.)

- Partially observable MDP
  - No access to full state information
  - Optimal policy is not stationary
  - Markovian structure does not hold anymore
  - Open issues in design: how to design efficient model-free and model-based methods
  - Open issues in theory: how to characterize sample complexity
- Multi-agent RL
  - Agents need to jointly achieve a design goal
  - Decentralized algorithms under partial observations of environments
  - Challenges in design: delayed communication; communication depends on network topology
  - Open issues in theory: tradeoff among communications, computations, privacy

Questions?

# References


[1] N. Sharma, S. Zhang, S. R. S. Venkata, F. Malandra, N. Mastronarde, and J. Chakareski, "Deep reinforcement learning for delay-sensitive lte downlink scheduling," in *IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–6, IEEE, 2020.

[2] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

[3] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*.
2018.

[4] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE transactions on automatic control*, vol. 42, no. 5, pp. 674–690, 1997.

[5] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*, vol. 48.
2009.

[6] A. Benveniste, P. Priouret, and M. Métivier, *Adaptive Algorithms and Stochastic Approximations*.
Springer-Verlag, 1990.

[7] V. Tadić, "On the convergence of temporal-difference learning with linear function approximation," *Machine learning*, vol. 42, no. 3, pp. 241–267, 2001.

[8] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor, "Finite sample analyses for td (0) with function approximation," in *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, vol. 32, 2018.

[9] R. Srikant and L. Ying, "Finite-time error bounds for linear stochastic approximation andtd learning," in *Proc. Conference on Learning Theory (COLT)*, pp. 2803–2830, 2019.

[10] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Proc. Conference on Learning Theory (COLT)*, vol. 75, pp. 1691–1692, 2018.

[11] R. Srikant and L. Ying, "Finite-time error bounds for linear stochastic approximation andtd learning," in *Proc. Conference on Learning Theory*, vol. 99, pp. 2803–2830, 25–28 Jun 2019.

[12] L. Baird, "Residual algorithms: Reinforcement learning with function approximation," in *Proc. International Conference on Machine Learning (ICML)*, pp. 30–37, 1995.

[13] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, "Fast gradient-descent methods for temporal-difference learning with linear function approximation," in *Proc. International Conference on Machine Learning (ICML)*, pp. 993–1000, 2009.

[14] H. R. Maei, *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta, 2011.

[15] H. Yu, "On convergence of some gradient-based temporal-differences algorithms for off-policy learning," *arXiv1712.09652*, 2018.

[16] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai, "Finite time analysis of linear two-timescale stochastic approximation with markovian noise," in *Proc. Conference on Learning Theory (COLT)*, pp. 2144–2203, 2020.

[17] T. Xu, S. Zou, and Y. Liang, "Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10634–10644, 2019.

[18] T. Xu and Y. Liang, "Sample complexity bounds for two timescale value-based reinforcement learning algorithms," in *Proc. International Conference on Artificial Intelligence and Statistics*, vol. 130, pp. 811–819, 13–15 Apr 2021.

[19] F. S. Melo, S. P. Meyn, and M. I. Ribeiro, "An analysis of reinforcement learning with function approximation," in *Proc. International Conference on Machine Learning (ICML)*, pp. 664–671, ACM, 2008.

[20] S. Zou, T. Xu, and Y. Liang, "Finite-sample analysis for sarsa with linear function approximation," in *Proc. Advances in Neural Information Processing Systems*, pp. 8665–8675, 2019.

[21] T. J. Perkins and D. Precup, "A convergent form of approximate policy iteration," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1627–1634, 2003.

[22] G. Li, C. Cai, Y. Chen, Y. Gu, Y. Wei, and Y. Chi, "Is q-learning minimax optimal? a tight sample complexity analysis," *arXiv preprint arXiv:2102.06548*, 2021.

[23] M. J. Wainwright, "Variance-reduced $q$-learning is minimax optimal," *arXiv preprint arXiv:1906.04697*, 2019.

[24] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, "Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction," *arXiv preprint arXiv:2006.03041*, 2020.

[25] H. R. Maei, C. Szepesvári, S. Bhatnagar, and R. S. Sutton, "Toward off-policy learning control with function approximation," in *Proc. International Conference on Machine Learning (ICML)*, 2010.

[26] Y. Wang and S. Zou, "Finite-sample analysis of Greedy-GQ with linear function approximation under Markovian noise," in *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, vol. 124, pp. 11–20, 2020.

[27] T. Xu and Y. Liang, "Sample complexity bounds for two timescale value-based reinforcement learning algorithms," *ArXiv:2011.05053*, 2020.

[28] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Advances in Neural Information Processing Systems*, vol. 26, 2013.

[29] S. Ma, Z. Chen, Y. Zhou, and S. Zou, "Greedy-GQ with variance reduction: Finite-time analysis and improved complexity," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[30] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1057–1063, 2000.

[31] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[32] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *The 32nd International Conference on Machine Learning (ICML)*, pp. 1889–1897, 2015.

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[34] S. M. Kakade, "A natural policy gradient," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1531–1538, 2002.

[35] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, “Optimality and approximation with policy gradient methods in Markov decision processes,” *arXiv preprint arXiv:1908.00261*, 2019.

[36] G. Lan, “Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes,” *ArXiv:2102.00135*, 2021.

[37] J. Bhandari and D. Russo, “Global optimality guarantees for policy gradient methods,” *arXiv preprint arXiv:1906.01786*, 2019.

[38] L. Shani, Y. Efroni, and S. Mannor, “Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs,” *arXiv preprint arXiv:1909.02769*, 2019.

[39] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, “Fast global convergence of natural policy gradient methods with entropy regularization,” *arXiv:2007.06558*, 2020.

[40] V. Konda, "Actor-critic algorithms (ph.d. thesis)," *Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology*, 2002.

[41] H. Xiong, T. Xu, YingbinLiang, and W. Zhang, "Non-asymptotic convergence of Adam-type reinforcement learning algorithms under Markovian sampling," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[42] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1008–1014, 2000.

[43] T. Xu, Z. Wang, and Y. Liang, "Improving sample complexity bounds for (natural) actor-critic algorithms," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, also available as *arXiv preprint arXiv:2004.12956*, 2020.

[44] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[45] T. Xu, Z. Yang, Z. Wang, and Y. Liang, "Doubly robust off-policy actor-critic: Convergence and optimality," in *Proc. International Conference on Machine Learning (ICML)*, 2021.

[46] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. International Conference on Machine Learning (ICML)*, pp. 22–31, 2017.

[47] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6070–6120, 2017.

[48] E. Altman, *Constrained Markov Decision Processes*, vol. 7. CRC Press, 1999.

[49] S. Paternain, L. F. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2019.

[50] D. Ding, K. Zhang, T. Basar, and M. Jovanovic, “Natural policy gradient primal-dual method for constrained markov decision processes,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[51] T. Xu, Y. Liang, and G. Lan, “CRPO: A new approach for safe reinforcement learning with convergence guarantee,” in *Proc. International Conference on Machine Learning (ICML)*, 2021.

[52] D. A. Pomerleau, “Efficient training of artificial neural networks for autonomous navigation.,” *Neural Computation*, vol. 3, no. 1, pp. 88–97, 1991.

[53] S. Russell, “Learning agents for uncertain environments,” in *Proc. Eleventh Annual Conference on Computational Learning Theory*, 1998.

[54] A. Y. Ng and S. Russell, “Algorithms for inverse reinforcement learning,” in *Proc. International Conference on Machine Learning (ICML)*, 2000.

[55] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2016.