Reinforcement Learning in Non-Stationary Environments[1]

Pranay (pranaysh@iitb.ac.in)
C-MInDS, IIT Bombay

CNI Seminar, IISc Bengaluru

February 14, 2026

[1] J, P, **S**, Q, and J. "Natural Policy Gradient for Average Reward Non-Stationary RL." accepted in TMLR (Jan, 2026).

Neharika Jali

Eeshika Pathak

Guannan Qu

Gauri Joshi

Reinforcement Learning (RL)

Sequential decision-making under uncertainty

Non-stationarity in RL

Non-stationarity in RL

We want to do well in a time-varying environment, in the long run

Non-stationarity in RL

We want to **do well** in a **time-varying environment**, in the **long run**

Outline

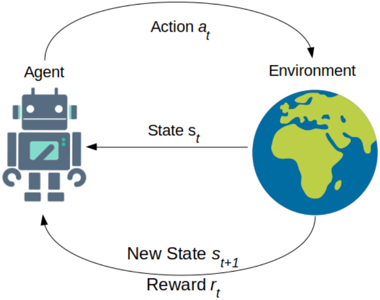
1. Some Background
2. Natural Actor-Critic
3. Non-stationary Natural Actor-Critic (NS-NAC) Algorithm
4. Regret Bounds
5. Proof Sketch
6. Concluding Remarks

Some Background

Markov Decision Processes (MDPs)

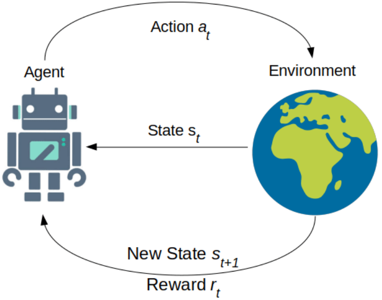
$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$$

- $\mathcal{S}$, $\mathcal{A}$ - states and actions sets
- Trajectory (s_t, a_t, r_t, s_{t+1})
- Action $a_t \sim \pi(\cdot|s_t)$, with policy $\boldsymbol{\pi}$
- Reward r_t
- Next state $s_{t+1} \sim P(\cdot|s, a)$



Markov Decision Processes (MDPs)

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$$



- $\mathcal{S}$, $\mathcal{A}$ - states and actions sets
- Trajectory (s_t, a_t, r_t, s_{t+1})
- Action $a_t \sim \pi(\cdot|s_t)$, with policy $\boldsymbol{\pi}$
- Reward r_t
- Next state $s_{t+1} \sim P(\cdot|s, a)$

$\mathbf{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ - transition probability matrix

$\mathbf{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ - reward vector

In stationary MDP - **P** and **r** are *time-invariant*

Our non-stationary setting - $\{\mathbf{P}_t, \mathbf{r}_t\}_t$

Discounted vs Average Reward

- Usually - maximize the **cumulative discounted reward**

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\middle| s_0 \sim \eta, a_t \sim \pi(\cdot \mid s_t)\right]$$

η - some initial state distribution; π - policy

Discounted vs Average Reward

- Usually - maximize the **cumulative discounted reward**

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 \sim \eta, a_t \sim \pi(\cdot \mid s_t)\right]$$

  η - some initial state distribution; π - policy

- Here - **average reward**

$$J^{\pi} := \lim_{T \to \infty} \frac{\mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right]}{T}$$

Discounted vs Average Reward

- Usually - maximize the **cumulative discounted reward**

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\middle| s_0 \sim \eta, a_t \sim \pi(\cdot \mid s_t)\right]$$

  η - some initial state distribution; π - policy

- Here - **average reward**

$$J^{\boldsymbol{\pi}} := \lim_{T\to\infty} \frac{\mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right]}{T}$$

- Under **ergodicity** assumption

$$J^{\boldsymbol{\pi}} = \mathbb{E}_{s\sim d^{\boldsymbol{\pi},\mathbf{P}}(\cdot), a\sim\pi(\cdot|s)}\left[r(s, a)\right],$$

  $d^{\boldsymbol{\pi},\mathbf{P}}$ - stationary distribution over states induced by policy $\boldsymbol{\pi}$ and transition probabilities $\mathbf{P}$

Why Average Reward?

- Discounting introduces time preference: reward *now* is more important than reward later

Why Average Reward?

- Discounting introduces time preference: reward *now* is more important than reward later

"Yes, the planet got destroyed. But for a beautiful moment in time we created a lot of value for shareholders."

Tom Toro for the New Yorker, Nov 2012

Why Average Reward?

- Discounting introduces time preference: reward *now* is more important than reward later

Tom Toro for the New Yorker, Nov 2012

- Average reward emphasizes steady-state behavior

Why Average Reward?

- Right notion when the system runs "forever"
  - Queueing systems - maximize throughput or minimize average latency over an indefinite period
  - Communication networks
  - Power grids
  - Recommendation systems with continuous users

(Relative) Value Functions

- *Relative* state-value function

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty}\left(r(s_t, a_t) - J^{\pi}\right)\Big| s_0 = s\right],$$

- *Relative* state-action value function

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

- How much better than average a state (state-action pair) is

(Relative) Value Functions

- *Relative* state-value function

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty}(r(s_t, a_t) - J^{\pi})\,\Big|\, s_0 = s\right],$$

- *Relative* state-action value function

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

- How much better than average a state (state-action pair) is
- Bellman equations

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) Q^{\pi}(s,a)$$

$$Q^{\pi}(s,a) = r(s,a) - J^{\pi} + \sum_{s' \in \mathcal{S}} P(s'|s,a) V^{\pi}(s').$$

Value-based vs Policy-based Methods

Goal: find π that maximizes $V^{\boldsymbol{\pi}}$ or $Q^{\boldsymbol{\pi}}$

- **Value-based Methods** iteratively update value function $\implies$ use it to select actions
  - Can diverge with function approximation under continuous setting

Value-based vs Policy-based Methods

Goal: find π that maximizes $V^{\boldsymbol{\pi}}$ or $Q^{\boldsymbol{\pi}}$

- **Value-based Methods** iteratively update value function $\Longrightarrow$ use it to select actions
  - Can diverge with function approximation under continuous setting
- **Policy-based Methods** directly learn policy. E.g.,
  - Given policy parameter $\boldsymbol{\theta}$, and some performance measure $J(\boldsymbol{\theta})$

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \beta_t \widehat{\nabla J(\boldsymbol{\theta}_t)}$$

Value-based vs Policy-based Methods

Goal: find π that maximizes V^{π} or Q^{π}

- **Value-based Methods** iteratively update value function $\Longrightarrow$ use it to select actions
  - Can diverge with function approximation under continuous setting
- **Policy-based Methods** directly learn policy. E.g.,
  - Given policy parameter $\boldsymbol{\theta}$, and some performance measure $J(\boldsymbol{\theta})$

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \beta_t \widehat{\nabla J(\boldsymbol{\theta}_t)}$$

- Advantages
  - Often policies are easier to approximate
  - Can inject prior knowledge about policy

Policy-based Methods: Actor-only[1] vs Actor-Critic[2]

Actor-only methods

- Can be naturally applied to continuous settings
- Suffer from high variance when estimating the policy gradient

[1] Agarwal, et al. "On the theory of policy gradient methods: Optimality, approximation, and distribution shift." JMLR (2021).

[2] Bhatnagar, et al. "Natural actor–critic algorithms." Automatica (2009).

Policy-based Methods: Actor-only[1] vs Actor-Critic[2]

Actor-only methods

- Can be naturally applied to continuous settings
- Suffer from high variance when estimating the policy gradient

Actor-critic methods

- Critic tries to learn the value function, given actor's policy
- Actor estimates the policy gradient based on approximate value function provided by the critic

[1] Agarwal, et al. "On the theory of policy gradient methods: Optimality, approximation, and distribution shift." JMLR (2021).

[2] Bhatnagar, et al. "Natural actor–critic algorithms." Automatica (2009).

Rationale

Goal: find $\boldsymbol{\pi}^\star = \max_{\boldsymbol{\pi}} J^{\boldsymbol{\pi}}$

- Critic-only (Value-function-based) methods might diverge

[1] Wu, et al. "A finite-time analysis of two time-scale actor-critic methods." NeurIPS (2020).
[2] Bhatnagar, et al. "Natural actor–critic algorithms." Automatica (2009).
[3] Khodadadian, et al. "Finite-sample analysis of two-time-scale natural actor–critic algorithm." IEEE TAC (2022).

Rationale

Goal: find $\boldsymbol{\pi}^\star = \max_{\boldsymbol{\pi}} J^{\boldsymbol{\pi}}$

- Critic-only (Value-function-based) methods might diverge
- Actor-only methods - sample inefficient, high variance

[1] Wu, et al. "A finite-time analysis of two time-scale actor-critic methods." NeurIPS (2020).

[2] Bhatnagar, et al. "Natural actor–critic algorithms." Automatica (2009).

[3] Khodadadian, et al. "Finite-sample analysis of two-time-scale natural actor–critic algorithm." IEEE TAC (2022).

Rationale

Goal: find $\boldsymbol{\pi}^\star = \max_{\boldsymbol{\pi}} J^{\boldsymbol{\pi}}$

- Critic-only (Value-function-based) methods might diverge
- Actor-only methods - sample inefficient, high variance
- Actor-critic - best of both worlds[1]
- Natural actor-critic[2]
  - Leverages the second-order Natural Gradient method
  - Guarantees global optimality[3]

---

[1] Wu, et al. "A finite-time analysis of two time-scale actor-critic methods." NeurIPS (2020).

[2] Bhatnagar, et al. "Natural actor–critic algorithms." Automatica (2009).

[3] Khodadadian, et al. "Finite-sample analysis of two-time-scale natural actor–critic algorithm." IEEE TAC (2022).

Problem Statement

Problem. Infinite-horizon Average-reward RL in Non-stationary Environments

Approach. Natural Actor-Critic type method

Outline

1. Some Background
2. Natural Actor-Critic
3. Non-stationary Natural Actor-Critic (NS-NAC) Algorithm
4. Regret Bounds
5. Proof Sketch
6. Concluding Remarks

Natural Actor-Critic

Natural Actor-Critic - Actor Update

With $J(\boldsymbol{\theta}) \triangleq J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$, the *actor* updates the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ via a natural gradient step[1]

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta F_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{-1} \nabla J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$$

[1] Martens, "New insights and perspectives on the natural gradient method." JMLR (2020).

[2] Sutton, and Barto. "Reinforcement learning: An introduction." Cambridge: MIT press, (1998).

Natural Actor-Critic - Actor Update

With $J(\boldsymbol{\theta}) \triangleq J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$, the *actor* updates the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ via a natural gradient step[1]

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta F_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{-1} \nabla J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$$

- $\nabla J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ - given by policy gradient theorem[2]

$$\nabla J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}} = \mathbb{E}_{\underbrace{s \sim d^{\boldsymbol{\pi}_{\boldsymbol{\theta}}, \mathbf{P}}(\cdot)}_{\text{Stationary distribution}}, \underbrace{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}_{\text{policy}}} \Big[\underbrace{Q^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(s, a)}_{\text{State-action value function}} \nabla \log \boldsymbol{\pi}_{\boldsymbol{\theta}}(a|s) \Big]$$

[1] Martens, "New insights and perspectives on the natural gradient method." JMLR (2020).

[2] Sutton, and Barto. "Reinforcement learning: An introduction." Cambridge: MIT press, (1998).

Natural Actor-Critic - Actor Update

With $J(\boldsymbol{\theta}) \triangleq J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$, the *actor* updates the policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ via a natural gradient step[1]

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta F_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{-1} \nabla J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$$

- $\nabla J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ - given by policy gradient theorem[2]

$$\nabla J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}} = \mathbb{E}_{\underbrace{s \sim d^{\boldsymbol{\pi}_{\boldsymbol{\theta}}, \mathbf{P}}(\cdot)}_{\text{Stationary distribution}}, \underbrace{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}_{\text{policy}}} \Big[\underbrace{Q^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(s, a)}_{\text{State-action value function}} \nabla \log \boldsymbol{\pi}_{\boldsymbol{\theta}}(a|s) \Big]$$

- $F_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$ is the Fisher Information matrix

$$F_{\boldsymbol{\pi}_{\boldsymbol{\theta}}} := \mathbb{E}_{s \sim d^{\boldsymbol{\pi}_{\boldsymbol{\theta}}, \mathbf{P}}(\cdot), a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\nabla \log \boldsymbol{\pi}_{\boldsymbol{\theta}}(a|s) \left(\nabla \log \boldsymbol{\pi}_{\boldsymbol{\theta}}(a|s) \right)^{\top} \right]$$

[1] Martens, "New insights and perspectives on the natural gradient method." JMLR (2020).

[2] Sutton, and Barto. "Reinforcement learning: An introduction." Cambridge: MIT press, (1998).

Natural Actor-Critic - Actor Update

With softmax parameterization, i.e., with $\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$

$$\pi_{\boldsymbol{\theta}}(a \mid s) = \frac{\exp[\boldsymbol{\theta}]_{s,a}}{\sum_{a' \in \mathcal{A}} \exp[\boldsymbol{\theta}]_{s,a'}}, \text{ for all } a \in \mathcal{A}, s \in \mathcal{S}$$

the actor update

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta F_{\boldsymbol{\pi_\theta}}^{-1} \nabla J^{\boldsymbol{\pi_\theta}}$$

simplifies to

Natural Actor-Critic - Actor Update

With softmax parameterization, i.e., with $\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$

$$\pi_{\boldsymbol{\theta}}(a \mid s) = \frac{\exp[\boldsymbol{\theta}]_{s,a}}{\sum_{a' \in \mathcal{A}} \exp[\boldsymbol{\theta}]_{s,a'}}, \text{ for all } a \in \mathcal{A}, s \in \mathcal{S}$$

the actor update

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta F_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}^{-1} \nabla J^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$$

simplifies to

$$\pi_{\boldsymbol{\theta}}(a|s) \leftarrow \frac{\pi_{\boldsymbol{\theta}}(a|s) \exp(\beta Q^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(s,a))}{\sum_{a' \in \mathcal{A}} \pi_{\boldsymbol{\theta}}(a'|s) \exp(\beta Q^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}(s,a'))}, \text{ for all } a \in \mathcal{A}, s \in \mathcal{S}$$

- Does this update remind you of something?

Natural Actor-Critic - Actor Update

With softmax parameterization, i.e., with $\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$

$$\pi_{\boldsymbol{\theta}}(a \mid s) = \frac{\exp[\boldsymbol{\theta}]_{s,a}}{\sum_{a' \in \mathcal{A}} \exp[\boldsymbol{\theta}]_{s,a'}}, \text{ for all } a \in \mathcal{A}, s \in \mathcal{S}$$

the actor update

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \beta F_{\boldsymbol{\pi_\theta}}^{-1} \nabla J^{\boldsymbol{\pi_\theta}}$$

simplifies to

$$\pi_{\boldsymbol{\theta}}(a|s) \leftarrow \frac{\pi_{\boldsymbol{\theta}}(a|s) \exp(\beta Q^{\boldsymbol{\pi_\theta}}(s,a))}{\sum_{a' \in \mathcal{A}} \pi_{\boldsymbol{\theta}}(a'|s) \exp(\beta Q^{\boldsymbol{\pi_\theta}}(s,a'))}, \text{ for all } a \in \mathcal{A}, s \in \mathcal{S}$$

- Does this update remind you of something?
- We don't have $Q^{\boldsymbol{\pi_\theta}}$

Natural Actor-Critic - Critic Update

Critic estimates Q-Value function $Q^{\pi}(s, a)$ using TD-learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r(s,a) - \eta + Q(s',a') - Q(s,a) \right],$$

- $s' \sim P(\cdot|s,a)$, $a' \sim \pi(\cdot|s')$
- η is an estimate of average reward J^{π_θ}

Natural Actor-Critic - Critic Update

Critic estimates Q-Value function $Q^{\pi}(s, a)$ using TD-learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r(s,a) - \eta + Q(s',a') - Q(s,a) \right],$$

- $s' \sim P(\cdot|s,a)$, $a' \sim \pi(\cdot|s')$
- η is an estimate of average reward J^{π_θ}

Average reward: $\eta_{t+1} = \eta_t + \gamma \left(r(s_t, a_t) - \eta_t \right)$

Natural Actor-Critic

Initialize: $\pi_0(a|s) = \frac{1}{|\mathcal{A}|}$, value function $Q_0(s, a) = 0$, $\forall s, a$, average reward estimate $\eta_0 = 0$

- **Trajectory:** Observe reward $r(s_t, a_t)$, next state $s_{t+1} \sim P(\cdot|s_t, a_t)$, take action $a_{t+1} \sim \pi_t(\cdot|s_{t+1})$
- **Average reward:** $\eta_{t+1} = \eta_t + \gamma \left(r(s_t, a_t) - \eta_t \right)$
- **Critic:**

$$Q_{t+1}(s_t, a_t) = \Pi_R \left[Q_t(s_t, a_t) + \alpha \left(r_t(s_t, a_t) - \eta_t + Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \right) \right]$$

- **Actor:** $\pi_{t+1}(a|s) = \frac{\pi_t(a|s) \exp(\beta Q_t(s,a))}{\sum_{a' \in \mathcal{A}} \pi_t(a'|s) \exp(\beta Q_t(s,a'))}$, $\forall s, a$
- Repeat

Natural Actor-Critic

Initialize: $\pi_0(a|s) = \frac{1}{|\mathcal{A}|}$, value function $Q_0(s,a) = 0$, $\forall s, a$, average reward estimate $\eta_0 = 0$

- **Trajectory:** Observe reward $r(s_t, a_t)$, next state $s_{t+1} \sim P(\cdot|s_t, a_t)$, take action $a_{t+1} \sim \pi_t(\cdot|s_{t+1})$
- **Average reward:** $\eta_{t+1} = \eta_t + \gamma\left(r(s_t, a_t) - \eta_t\right)$
- **Critic:**

$$Q_{t+1}(s_t, a_t) = \Pi_R\left[Q_t(s_t, a_t) + \alpha\left(r_t(s_t, a_t) - \eta_t + Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)\right)\right]$$

- **Actor:** $\pi_{t+1}(a|s) = \frac{\pi_t(a|s)\exp(\beta Q_t(s,a))}{\sum_{a' \in \mathcal{A}} \pi_t(a'|s)\exp(\beta Q_t(s,a'))}$, $\forall s, a$
- Repeat

Q. What if rewards and transition probabilities change over time?

Outline

1. Some Background

2. Natural Actor-Critic

3. Non-stationary Natural Actor-Critic (NS-NAC) Algorithm

4. Regret Bounds

5. Proof Sketch

6. Concluding Remarks

Non-stationary Natural Actor-Critic (NS-NAC) Algorithm

NAC under Stationarity

- Critic estimates value function $\mathbf{Q}_t$ of the current policy $\boldsymbol{\pi}_t$
- However, policy $\boldsymbol{\pi}_t$ also evolves constantly
- If Critic step-size $\alpha \gg$ Actor step-size β, critic achieves *good enough* estimates of $\mathbf{Q}_t$[1]

[1] Not necessary. See Wang et al. ICML (2024).

NAC under Non-stationarity

- MDP is modeled by a sequence of environments

$$\mathcal{M} = \{\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \mathbf{P}_t, \mathbf{r}_t)\}_{t=0}^{T-1}$$

Time-varying rewards $\{\mathbf{r}_t\}$ and transition probabilities $\{\mathbf{P}_t\}$

NAC under Non-stationarity

- MDP is modeled by a sequence of environments

$$\mathcal{M} = \{\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \mathbf{P}_t, \mathbf{r}_t)\}_{t=0}^{T-1}$$

  Time-varying rewards $\{\mathbf{r}_t\}$ and transition probabilities $\{\mathbf{P}_t\}$

- Actor chases a moving target

$$\boldsymbol{\pi}_t^\star = \arg\max_{\boldsymbol{\pi}} \left\{ J_t^{\boldsymbol{\pi}} \triangleq \mathbb{E}_{s\sim d^{\boldsymbol{\pi},\mathbf{P}_t}(\cdot),a\sim\pi(\cdot|s)} \left[r_t(s,a)\right] \right\}$$

  Time-varying optimal policy in the environment $\mathcal{M}_t$ at time t

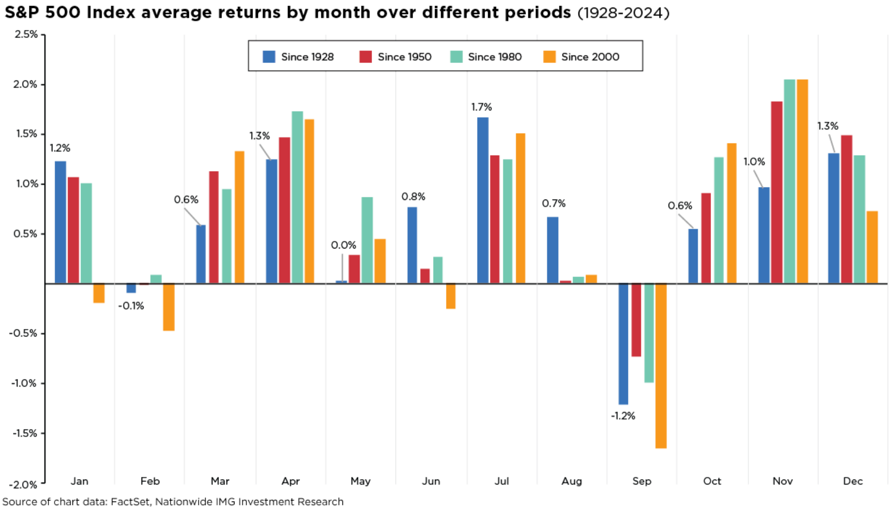


Image from nationwide.com

NAC under Non-stationarity

- MDP is modeled by a sequence of environments

$$\mathcal{M} = \{\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \mathbf{P}_t, \mathbf{r}_t)\}_{t=0}^{T-1}$$

  Time-varying rewards $\{\mathbf{r}_t\}$ and transition probabilities $\{\mathbf{P}_t\}$

- Actor chases a moving target

$$\boldsymbol{\pi}_t^\star = \arg\max_{\boldsymbol{\pi}} \left\{ J_t^{\boldsymbol{\pi}} \triangleq \mathbb{E}_{s \sim d^{\boldsymbol{\pi}, \mathbf{P}_t}(\cdot), a \sim \pi(\cdot|s)} \left[r_t(s, a) \right] \right\}$$

  Time-varying optimal policy in the environment $\mathcal{M}_t$ at time t

- **Challenges**
  - Need to explore more aggressively than in the stationary setting
  - Balance forgetting old environments versus learning new ones

NS-NAC

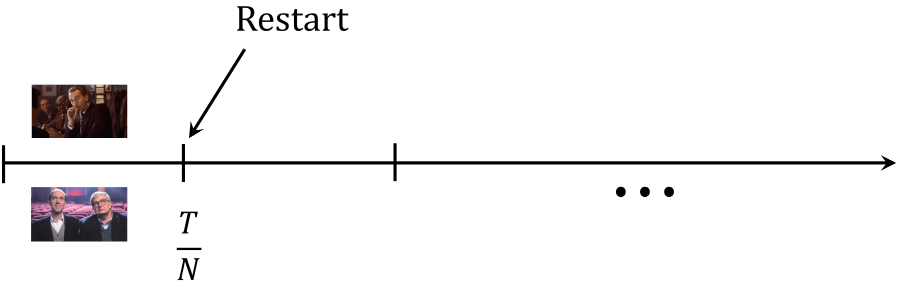
- The horizon T is divided into N segments of length T/N

NS-NAC

- The horizon T is divided into N segments of length T/N
- Restarts at the beginning of each segment (for sufficient exploration)

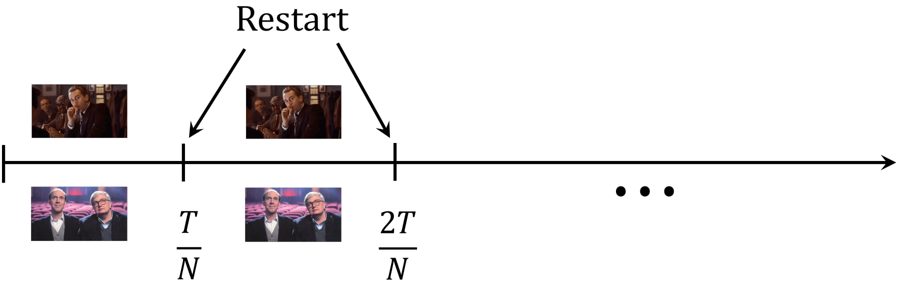
NS-NAC

- The horizon T is divided into N segments of length T/N
- Restarts at the beginning of each segment (for sufficient exploration)



NS-NAC

- The horizon T is divided into N segments of length T/N
- Restarts at the beginning of each segment (for sufficient exploration)



NS-NAC

- The horizon T is divided into N segments of length T/N
- Restarts at the beginning of each segment (for sufficient exploration)
- Within each segment
  Initialize: $\pi(a|s) = \frac{1}{|\mathcal{A}|}$, value function $Q(s,a) = 0$, $\forall s, a$, average reward estimate $\eta = 0$
  - **Trajectory:** Observe reward $r_t(s_t, a_t)$, next state $s_{t+1} \sim P_t(\cdot|s_t, a_t)$, take action $a_{t+1} \sim \pi_t(\cdot|s_{t+1})$

NS-NAC

- The horizon T is divided into N segments of length T/N
- Restarts at the beginning of each segment (for sufficient exploration)
- Within each segment
  Initialize: $\pi(a|s) = \frac{1}{|\mathcal{A}|}$, value function $Q(s,a) = 0$, $\forall s, a$, average reward estimate $\eta = 0$
  - **Trajectory:** Observe reward $r_t(s_t, a_t)$, next state $s_{t+1} \sim P_t(\cdot|s_t, a_t)$, take action $a_{t+1} \sim \pi_t(\cdot|s_{t+1})$
  - **Average reward:** estimates $J_t^{\pi_t}$

$$\eta_{t+1} = \eta_t + \gamma \left(r_t(s_t, a_t) - \eta_t \right)$$

  - **Critic:** estimates $Q_t^{\pi_t}$

$$Q_{t+1}(s_t, a_t) = \Pi_R \left[Q_t(s_t, a_t) + \alpha \left(r_t(s_t, a_t) - \eta_t + Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \right) \right]$$

  - **Actor:** estimates $\pi_t^\star$

$$\pi_{t+1}(a|s) = \frac{\pi_t(a|s) \exp(\beta Q_t(s,a))}{\sum_{a' \in \mathcal{A}} \pi_t(a'|s) \exp(\beta Q_t(s,a'))}, \qquad \forall s, a$$

Outline

1. Some Background
2. Natural Actor-Critic
3. Non-stationary Natural Actor-Critic (NS-NAC) Algorithm
4. Regret Bounds
5. Proof Sketch
6. Concluding Remarks

Regret Bounds

Dynamic Regret

- **Goal:** maximize the time-averaged reward $\frac{1}{T}\sum_{t=0}^{T-1} r_t(s_t, a_t)$

[1] Cheung, et al. "Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism." ICML (2020).

[2] Mao, et al. "Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control." Management Science (2025).

[3] Feng, et al. "Non-stationary reinforcement learning under general function approximation." ICML (2023).

[4] Even-Dar, et al. "Online Markov decision processes." Math of OR (2009).

Dynamic Regret

- **Goal:** maximize the time-averaged reward $\frac{1}{T}\sum_{t=0}^{T-1} r_t(s_t, a_t)$
- Our performance metric - *dynamic regret*[1][2][3]

$$\text{Dyn-Reg}(\mathcal{M}, T) := \mathbb{E}\left[\sum_{t=0}^{T-1} J_t^{\boldsymbol{\pi}_t^\star} - r_t(s_t, a_t)\right]$$

$\boldsymbol{\pi}_t^\star = \arg\max_{\boldsymbol{\pi}} J_t^{\boldsymbol{\pi}}$ is the optimal policy in the environment $\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \mathbf{P}_t, \mathbf{r}_t)$ at time t

[1] Cheung, et al. "Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism." ICML (2020).

[2] Mao, et al. "Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control." Management Science (2025).

[3] Feng, et al. "Non-stationary reinforcement learning under general function approximation." ICML (2023).

[4] Even-Dar, et al. "Online Markov decision processes." Math of OR (2009).

Dynamic Regret

- **Goal:** maximize the time-averaged reward $\frac{1}{T}\sum_{t=0}^{T-1} r_t(s_t, a_t)$
- Our performance metric - *dynamic regret*[1][2][3]

$$\text{Dyn-Reg}(\mathcal{M}, T) := \mathbb{E}\left[\sum_{t=0}^{T-1} J_t^{\boldsymbol{\pi}_t^\star} - r_t(s_t, a_t)\right]$$

$\boldsymbol{\pi}_t^\star = \arg\max_{\boldsymbol{\pi}} J_t^{\boldsymbol{\pi}}$ is the optimal policy in the environment $\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \mathbf{P}_t, \mathbf{r}_t)$ at time t

- Compare to **static regret** - cumulative reward relative to a *single* stationary optimal policy[4]

[1] Cheung, et al. "Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism." ICML (2020).

[2] Mao, et al. "Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control." Management Science (2025).

[3] Feng, et al. "Non-stationary reinforcement learning under general function approximation." ICML (2023).

[4] Even-Dar, et al. "Online Markov decision processes." Math of OR (2009).

Variation Budgets

Given the sequence of environments

$$\mathcal{M} = \{\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \mathbf{P}_t, \mathbf{r}_t)\}_{t=0}^{T-1}$$

with time-varying rewards $\{\mathbf{r}_t\}$ and transition probabilities $\{\mathbf{P}_t\}$

Variation Budgets

Given the sequence of environments

$$\mathcal{M} = \{\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \mathbf{P}_t, \mathbf{r}_t)\}_{t=0}^{T-1}$$

with time-varying rewards $\{\mathbf{r}_t\}$ and transition probabilities $\{\mathbf{P}_t\}$

- Cumulative change in the reward and transition probabilities

$$\Delta_{R,T} = \sum_{t=0}^{T-1} \|\mathbf{r}_{t+1} - \mathbf{r}_t\|_\infty, \quad \Delta_{P,T} = \sum_{t=0}^{T-1} \|\mathbf{P}_{t+1} - \mathbf{P}_t\|_\infty$$

  with $\Delta_T = \Delta_{R,T} + \Delta_{P,T}$

Variation Budgets

Given the sequence of environments

$$\mathcal{M} = \{\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \mathbf{P}_t, \mathbf{r}_t)\}_{t=0}^{T-1}$$

with time-varying rewards $\{\mathbf{r}_t\}$ and transition probabilities $\{\mathbf{P}_t\}$

- Cumulative change in the reward and transition probabilities

$$\Delta_{R,T} = \sum_{t=0}^{T-1} \|\mathbf{r}_{t+1} - \mathbf{r}_t\|_\infty, \quad \Delta_{P,T} = \sum_{t=0}^{T-1} \|\mathbf{P}_{t+1} - \mathbf{P}_t\|_\infty$$

  with $\Delta_T = \Delta_{R,T} + \Delta_{P,T}$

- Variations at time t, $\|\mathbf{r}_{t+1} - \mathbf{r}_t\|_\infty$ and $\|\mathbf{P}_{t+1} - \mathbf{P}_t\|_\infty$, are unknown

Assumption: Uniform Ergodicity

Markov chain generated by implementing policy π in environment with transition probabilities $\mathbf{P}$ is **uniformly ergodic**

Assumption: Uniform Ergodicity

Markov chain generated by implementing policy $\boldsymbol{\pi}$ in environment with transition probabilities $\mathbf{P}$ is **uniformly ergodic**

- There exists $m > 0$ and $\rho \in (0, 1)$ such that

$$d_{TV}\left(P(s_\tau \in \cdot | s_0 = s), d^{\boldsymbol{\pi},\mathbf{P}}\right) \leq m\rho^\tau \ \forall \tau \geq 0, s \in \mathcal{S}$$

- $P(s_\tau \in \cdot | s_0 = s)$ - Markov chain state distribution at time τ
- $d^{\boldsymbol{\pi},\mathbf{P}}$ - Stationary distribution

Assumption: Uniform Ergodicity

Markov chain generated by implementing policy $\boldsymbol{\pi}$ in environment with transition probabilities $\mathbf{P}$ is **uniformly ergodic**

- There exists $m > 0$ and $\rho \in (0, 1)$ such that

$$d_{TV}\left(P(s_\tau \in \cdot | s_0 = s), d^{\boldsymbol{\pi}, \mathbf{P}}\right) \leq m\rho^\tau \ \forall \tau \geq 0, s \in \mathcal{S}$$

  - $P(s_\tau \in \cdot | s_0 = s)$ - Markov chain state distribution at time τ
  - $d^{\boldsymbol{\pi}, \mathbf{P}}$ - Stationary distribution

- Assume Markov chains induced by *all* potential policies $\boldsymbol{\pi}_t$ in *all* environments $\mathbf{P}_t$, $t \in [T]$, are uniformly ergodic with m, ρ

Dynamic Regret Bound

Under **uniform ergodicity** assumption, with variation budget $\Delta_T = \Delta_{R,T} + \Delta_{P,T}$

Dynamic Regret Bound

Under **uniform ergodicity** assumption, with variation budget $\Delta_T = \Delta_{R,T} + \Delta_{P,T}$

- Step-size choices
  - Average reward $\gamma^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$
  - Critic $\alpha^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$
  - Actor $\beta^\star = \left(\frac{\Delta_T}{T}\right)^{1/2}$

Dynamic Regret Bound

Under **uniform ergodicity** assumption, with variation budget $\Delta_T = \Delta_{R,T} + \Delta_{P,T}$

- Step-size choices
  - Average reward $\gamma^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$
  - Critic $\alpha^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$
  - Actor $\beta^\star = \left(\frac{\Delta_T}{T}\right)^{1/2}$
- Number of restarts $N^\star = \Delta_T^{5/6} T^{1/6}$

Trajectory: Observe reward $r_t(s_t, a_t)$, next state $s_{t+1} \sim P_t(\cdot|s_t, a_t)$, take action $a_{t+1} \sim \pi_t(\cdot|s_{t+1})$

Average reward: estimates $J_t^{\pi_t}$

$$\eta_{t+1} = \eta_t + \gamma\left(r_t(s_t, a_t) - \eta_t\right)$$

Critic: estimates $Q_t^{\pi_t}$

$$Q_{t+1}(s_t, a_t) = \Pi_R\left[Q_t(s_t, a_t) + \alpha\left(r_t(s_t, a_t) - \eta_t + Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)\right)\right]$$

Actor: estimates $\pi_t^\star$

$$\pi_{t+1}(a|s) = \frac{\pi_t(a|s)\exp(\beta Q_t(s,a))}{\sum_{a' \in \mathcal{A}} \pi_t(a'|s)\exp(\beta Q_t(s,a'))}, \qquad \forall s, a$$

NS-NAC algorithm

Dynamic Regret Bound

Under **uniform ergodicity** assumption, with variation budget $\Delta_T = \Delta_{R,T} + \Delta_{P,T}$

- Step-size choices
  - Average reward $\gamma^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$
  - Critic $\alpha^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$
  - Actor $\beta^\star = \left(\frac{\Delta_T}{T}\right)^{1/2}$
- Number of restarts $N^\star = \Delta_T^{5/6} T^{1/6}$

Trajectory: Observe reward $r_t(s_t, a_t)$, next state $s_{t+1} \sim P_t(\cdot|s_t, a_t)$, take action $a_{t+1} \sim \pi_t(\cdot|s_{t+1})$

Average reward: estimates $J_t^{\pi_t}$

$$\eta_{t+1} = \eta_t + \gamma\left(r_t(s_t, a_t) - \eta_t\right)$$

Critic: estimates $Q_t^{\pi_t}$

$$Q_{t+1}(s_t, a_t) = \Pi_R\left[Q_t(s_t, a_t) + \alpha\left(r_t(s_t, a_t) - \eta_t + Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)\right)\right]$$

Actor: estimates $\pi_t^\star$

$$\pi_{t+1}(a|s) = \frac{\pi_t(a|s)\exp(\beta Q_t(s,a))}{\sum_{a' \in \mathcal{A}} \pi_t(a'|s)\exp(\beta Q_t(s,a'))}, \qquad \forall s, a$$

NS-NAC algorithm

NS-NAC achieves the regret bound

$$\text{Dyn-Reg}(\mathcal{M}, T) \leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/2}\Delta_T^{1/6}T^{5/6}\right)$$

Effect of Non-Stationarity

With average reward $\gamma^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$, critic $\alpha^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$, actor $\beta^\star = \left(\frac{\Delta_T}{T}\right)^{1/2}$ and $N^\star = \Delta_T^{5/6} T^{1/6}$ restarts

$$\text{Dyn-Reg}(\mathcal{M}, T) \leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/2}\Delta_T^{1/6} T^{5/6}\right)$$

The variation budget Δ_T represents the extent of non-stationarity

Effect of Non-Stationarity

With average reward $\gamma^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$, critic $\alpha^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$, actor $\beta^\star = \left(\frac{\Delta_T}{T}\right)^{1/2}$ and $N^\star = \Delta_T^{5/6} T^{1/6}$ restarts

$$\text{Dyn-Reg}(\mathcal{M}, T) \leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/2}\Delta_T^{1/6} T^{5/6}\right)$$

The variation budget Δ_T represents the extent of non-stationarity

- $\Delta_T \uparrow$ - worse regret

Effect of Non-Stationarity

With average reward $\gamma^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$, critic $\alpha^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$, actor $\beta^\star = \left(\frac{\Delta_T}{T}\right)^{1/2}$ and $N^\star = \Delta_T^{5/6} T^{1/6}$ restarts

$$\text{Dyn-Reg}(\mathcal{M}, T) \leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/2}\Delta_T^{1/6} T^{5/6}\right)$$

The variation budget Δ_T represents the extent of non-stationarity

- $\Delta_T \uparrow$ - worse regret
- Rapidly changing environment - large Δ_T
  - Must adapt quickly - larger step-sizes
  - Must explore more - more restarts

Effect of Non-Stationarity

With average reward $\gamma^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$, critic $\alpha^\star = \left(\frac{\Delta_T}{T}\right)^{1/3}$, actor $\beta^\star = \left(\frac{\Delta_T}{T}\right)^{1/2}$ and $N^\star = \Delta_T^{5/6} T^{1/6}$ restarts

$$\text{Dyn-Reg}(\mathcal{M}, T) \leq \tilde{\mathcal{O}}\left(|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/2}\Delta_T^{1/6} T^{5/6}\right)$$

The variation budget Δ_T represents the extent of non-stationarity

- $\Delta_T \uparrow$ - worse regret
- Rapidly changing environment - large Δ_T
  - Must adapt quickly - larger step-sizes
  - Must explore more - more restarts
- Larger state/action spaces $(|\mathcal{S}|, |\mathcal{A}|)$ - need more samples to detect changes and learn a good policy

Proof Sketch

Regret Decomposition

$$\text{Dyn-Reg}(\mathcal{M}, T) \triangleq \mathbb{E}\left[\sum_{t=0}^{T-1} J_t^{\boldsymbol{\pi}_t^\star} - r_t(s_t, a_t)\right]$$

Regret Decomposition

$$\text{Dyn-Reg}(\mathcal{M}, T) \triangleq \mathbb{E}\left[\sum_{t=0}^{T-1} J_t^{\boldsymbol{\pi}_t^\star} - r_t(s_t, a_t)\right]$$

$$= \sum_{t=0}^{T-1} \underbrace{\mathbb{E}\left[J_t^{\boldsymbol{\pi}_t^\star} - J_t^{\boldsymbol{\pi}_t}\right]}_{I_1:\substack{\text{optimal versus}\\ \text{actual avg reward}}} + \underbrace{\mathbb{E}\left[J_t^{\boldsymbol{\pi}_t} - r_t(s_t, a_t)\right]}_{I_2:\substack{\text{actual avg versus}\\ \text{instantaneous reward}}}$$

Regret Decomposition

$$\text{Dyn-Reg}(\mathcal{M}, T) \triangleq \mathbb{E}\left[\sum_{t=0}^{T-1} J_t^{\pi_t^\star} - r_t(s_t, a_t)\right]$$

$$= \sum_{t=0}^{T-1} \underbrace{\mathbb{E}\left[J_t^{\pi_t^\star} - J_t^{\pi_t}\right]}_{I_1:\text{ optimal versus actual avg reward}} + \underbrace{\mathbb{E}\left[J_t^{\pi_t} - r_t(s_t, a_t)\right]}_{I_2:\text{ actual avg versus instantaneous reward}}$$

- I_1 - average reward of the actual policy π_t at time t relative to the optimal policy $\pi_t^\star$

Regret Decomposition

$$\text{Dyn-Reg}(\mathcal{M}, T) \triangleq \mathbb{E}\left[\sum_{t=0}^{T-1} J_t^{\boldsymbol{\pi}_t^\star} - r_t(s_t, a_t)\right]$$

$$= \sum_{t=0}^{T-1} \underbrace{\mathbb{E}\left[J_t^{\boldsymbol{\pi}_t^\star} - J_t^{\boldsymbol{\pi}_t}\right]}_{I_1:\text{ optimal versus actual avg reward}} + \underbrace{\mathbb{E}\left[J_t^{\boldsymbol{\pi}_t} - r_t(s_t, a_t)\right]}_{I_2:\text{ actual avg versus instantaneous reward}}$$

- I_1 - average reward of the actual policy $\boldsymbol{\pi}_t$ at time t relative to the optimal policy $\boldsymbol{\pi}_t^\star$
- I_2 - average reward vs actual rewards received

Actor Error

Divide total horizon T into N restarted segments of length H each

$$I_1 = \sum_{t=0}^{T-1} \mathbb{E}\left[J_t^{\boldsymbol{\pi}_t^\star} - J_t^{\boldsymbol{\pi}_t}\right]$$

$$= \mathbb{E}\Bigg[\sum_{n=0}^{N-1}\sum_{h=0}^{H-1} \underbrace{\left(J_{nH+h}^{\boldsymbol{\pi}^\star_{nH+h}} - J_{nH}^{\boldsymbol{\pi}^\star_{nH}}\right)}_{I_3:\text{ optimal avg. reward across two environments}} + \underbrace{\left(J_{nH}^{\boldsymbol{\pi}^\star_{nH}} - J_{nH}^{\boldsymbol{\pi}_{nH+h}}\right)}_{I_4:\text{ avg. reward sub-optimality}} + \underbrace{\left(J_{nH}^{\boldsymbol{\pi}_{nH+h}} - J_{nH+h}^{\boldsymbol{\pi}_{nH+h}}\right)}_{I_5:\text{ avg. reward with same policy in two environments}}\Bigg]$$

Actor Error

Divide total horizon T into N restarted segments of length H each

$$
\begin{aligned}
I_1 &= \sum_{t=0}^{T-1} \mathbb{E}\left[J_t^{\pi_t^\star} - J_t^{\pi_t} \right] \\
&= \mathbb{E}\Bigg[\sum_{n=0}^{N-1} \sum_{h=0}^{H-1} \underbrace{\left(J_{nH+h}^{\pi_{nH+h}^\star} - J_{nH}^{\pi_{nH}^\star} \right)}_{I_3:\ \substack{\text{optimal avg. reward} \\ \text{across two environments}}} + \underbrace{\left(J_{nH}^{\pi_{nH}^\star} - J_{nH}^{\pi_{nH+h}} \right)}_{I_4:\ \substack{\text{avg. reward} \\ \text{sub-optimality}}} + \underbrace{\left(J_{nH}^{\pi_{nH+h}} - J_{nH+h}^{\pi_{nH+h}} \right)}_{I_5:\ \substack{\text{avg. reward with same} \\ \text{policy in two environments}}} \Bigg]
\end{aligned}
$$

- I_3 should depend on changes in the environment

Actor Error

Divide total horizon T into N restarted segments of length H each

$$I_1 = \sum_{t=0}^{T-1} \mathbb{E}\left[J_t^{\pi_t^\star} - J_t^{\pi_t}\right]$$

$$= \mathbb{E}\Bigg[\sum_{n=0}^{N-1}\sum_{h=0}^{H-1} \underbrace{\left(J_{nH+h}^{\pi^\star_{nH+h}} - J_{nH}^{\pi^\star_{nH}}\right)}_{I_3:\ \text{optimal avg. reward across two environments}} + \underbrace{\left(J_{nH}^{\pi^\star_{nH}} - J_{nH}^{\pi_{nH+h}}\right)}_{I_4:\ \text{avg. reward sub-optimality}} + \underbrace{\left(J_{nH}^{\pi_{nH+h}} - J_{nH+h}^{\pi_{nH+h}}\right)}_{I_5:\ \text{avg. reward with same policy in two environments}}\Bigg]$$

- I_3 should depend on changes in the environment
- I_5 should also depend on changes in the environment

Actor Error

Divide total horizon T into N restarted segments of length H each

$$I_1 = \sum_{t=0}^{T-1} \mathbb{E}\left[J_t^{\pi_t^\star} - J_t^{\pi_t}\right]$$

$$= \mathbb{E}\Bigg[\sum_{n=0}^{N-1}\sum_{h=0}^{H-1} \underbrace{\left(J_{nH+h}^{\pi_{nH+h}^\star} - J_{nH}^{\pi_{nH}^\star}\right)}_{I_3:\ \substack{\text{optimal avg. reward} \\ \text{across two environments}}} + \underbrace{\left(J_{nH}^{\pi_{nH}^\star} - J_{nH}^{\pi_{nH+h}}\right)}_{I_4:\ \substack{\text{avg. reward} \\ \text{sub-optimality}}} + \underbrace{\left(J_{nH}^{\pi_{nH+h}} - J_{nH+h}^{\pi_{nH+h}}\right)}_{I_5:\ \substack{\text{avg. reward with same} \\ \text{policy in two environments}}}\Bigg]$$

- I_3 should depend on changes in the environment
- I_5 should also depend on changes in the environment
- I_4 - how suboptimal is policy π_{nH+h}; depends on our algorithm

Actor Error

Divide total horizon T into N restarted segments of length H each

$$I_1 = \sum_{t=0}^{T-1} \mathbb{E}\left[J_t^{\pi_t^\star} - J_t^{\pi_t}\right]$$

$$= \mathbb{E}\Bigg[\sum_{n=0}^{N-1}\sum_{h=0}^{H-1} \underbrace{\left(J_{nH+h}^{\pi_{nH+h}^\star} - J_{nH}^{\pi_{nH}^\star}\right)}_{I_3:\text{ optimal avg. reward across two environments}} + \underbrace{\left(J_{nH}^{\pi_{nH}^\star} - J_{nH}^{\pi_{nH+h}}\right)}_{I_4:\text{ avg. reward sub-optimality}} + \underbrace{\left(J_{nH}^{\pi_{nH+h}} - J_{nH+h}^{\pi_{nH+h}}\right)}_{I_5:\text{ avg. reward with same policy in two environments}}\Bigg]$$

- I_3 should depend on changes in the environment
- I_5 should also depend on changes in the environment
- I_4 - how suboptimal is policy π_{nH+h}; depends on our algorithm
- N balances exploration-for-change and learning a good policy

Actor Error

$$I_1 = \mathbb{E}\Big[\sum_{n=0}^{N-1}\sum_{h=0}^{H-1} \underbrace{\left(J_{nH+h}^{\pi^{\star}_{nH+h}} - J_{nH}^{\pi^{\star}_{nH}}\right)}_{I_3:\ \text{optimal avg. reward across two environments}} + \underbrace{\left(J_{nH}^{\pi^{\star}_{nH}} - J_{nH}^{\pi_{nH+h}}\right)}_{I_4:\ \text{avg. reward sub-optimality}} + \underbrace{\left(J_{nH}^{\pi_{nH+h}} - J_{nH+h}^{\pi_{nH+h}}\right)}_{I_5:\ \text{avg. reward with same policy in two environments}}\Big]$$

- I_3 should depend on changes in the environment

$$I_3 = J_{nH+h}^{\pi^{\star}_{nH+h}} - J_{nH}^{\pi^{\star}_{nH}} \lesssim \|\mathbf{r}_{nH+h} - \mathbf{r}_{nH}\|_\infty + \|\mathbf{P}_{nH+h} - \mathbf{P}_{nH}\|_\infty$$

- I_5 should also depend on changes in the environment

$$I_5 = J_{nH}^{\pi_{nH+h}} - J_{nH+h}^{\pi_{nH+h}} \lesssim \|\mathbf{r}_{nH+h} - \mathbf{r}_{nH}\|_\infty + \|\mathbf{P}_{nH+h} - \mathbf{P}_{nH}\|_\infty$$

Actor Error

$$I_1 = \mathbb{E}\Big[\sum_{n=0}^{N-1}\sum_{h=0}^{H-1} \underbrace{\left(J_{nH+h}^{\pi^\star_{nH+h}} - J_{nH}^{\pi^\star_{nH}}\right)}_{I_3:\text{ optimal avg. reward across two environments}} + \underbrace{\left(J_{nH}^{\pi^\star_{nH}} - J_{nH}^{\pi_{nH+h}}\right)}_{I_4:\text{ avg. reward sub-optimality}} + \underbrace{\left(J_{nH}^{\pi_{nH+h}} - J_{nH+h}^{\pi_{nH+h}}\right)}_{I_5:\text{ avg. reward with same policy in two environments}}\Big]$$

- I_4 - how suboptimal is policy π_{nH+h}; depends on our algorithm
- **Average-Reward Performance Difference Lemma**

$$J_t^{\pi} - J_t^{\pi'} = \sum_{s\in\mathcal{S}} \underbrace{d^{\pi,\mathbf{P}_t}(s)}_{\text{stationary distribution}} \sum_{a\in\mathcal{A}} \pi(a|s)\Big[\underbrace{Q_t^{\pi'}(s,a) - V_t^{\pi'}(s)}_{\text{advantage}}\Big]$$

- Adapting to $J_{nH}^{\pi^\star_{nH}} - J_{nH}^{\pi_{nH+h}}$

$$Q_{nH}^{\pi_{nH+h}}(s,a) - V_{nH}^{\pi_{nH+h}}(s) \lesssim \underbrace{\|\mathbf{Q}_{nH+h}^{\pi_{nH+h}} - \mathbf{Q}_{nH+h}\|_\infty}_{\text{Critic error}}$$

$$+ \underbrace{\|\mathbf{r}_{nH+h+1} - \mathbf{r}_{nH+h}\|_\infty + \|\mathbf{P}_{nH+h+1} - \mathbf{P}_{nH+h}\|_\infty}_{\text{Change in environment}} + O(1)$$

Critic

- Critic update

$$Q_{t+1}(s_t, a_t) = \Pi_R\left[Q_t(s_t, a_t) + \alpha\left(r_t(s_t, a_t) - \eta_t + Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)\right)\right]$$

- In vector form

$$\mathbf{Q}_{t+1} = \Pi_{R_Q}\left[\mathbf{Q}_t + \alpha\left(\mathbf{r}_t(O_t) - \boldsymbol{\eta}_t(O_t) + \mathbf{A}(O_t)\mathbf{Q}_t\right)\right]$$

$O_t = (s_t, a_t, s_{t+1}, a_{t+1})$

- $\boldsymbol{\eta}_t(O_t)$ tracks average reward $\mathbf{J}_t^{\boldsymbol{\pi}_t}$
- $\mathbf{A}(O_t)$ is a random matrix

- We can bound $\boldsymbol{\psi}_t = \mathbf{Q}_t - \mathbf{Q}_t^{\boldsymbol{\pi}_t}$ recursively

$$\|\boldsymbol{\psi}_{t+1}\|_2^2 \lesssim (1-\alpha)\|\boldsymbol{\psi}_t\|_2^2 + \alpha\underbrace{\left(\mathbf{J}_t^{\boldsymbol{\pi}_t}(O_t) - \boldsymbol{\eta}_t(O_t)\right)^2}_{\text{avg. reward error}}$$

$$+\frac{1}{\alpha}\underbrace{\|\mathbf{Q}_t^{\boldsymbol{\pi}_t} - \mathbf{Q}_{t+1}^{\boldsymbol{\pi}_{t+1}}\|_2^2}_{\text{value function drift}} + \underbrace{\alpha^2\|\mathbf{r}_t(O_t) - \boldsymbol{\eta}_t(O_t) + \mathbf{A}(O_t)\mathbf{Q}_t\|_2^2}_{\text{higher-order term}}$$

$$+\alpha\underbrace{\boldsymbol{\psi}_t^\top\left[\left(\mathbf{r}_t(O_t) - \mathbf{J}_t^{\boldsymbol{\pi}_t}(O_t) + \mathbf{A}(O_t)\mathbf{Q}_t^{\boldsymbol{\pi}_t}\right) + \left(\mathbf{A}(O_t) - \mathbb{E}[\mathbf{A}(O_t)]\right)\boldsymbol{\psi}_t\right]}_{\text{error due to Markov noise}}$$

Critic

$$\|\boldsymbol{\psi}_{t+1}\|_2^2 \lesssim (1-\alpha)\|\boldsymbol{\psi}_t\|_2^2 + \alpha\underbrace{(\mathbf{J}_t^{\boldsymbol{\pi}_t}(O_t) - \boldsymbol{\eta}_t(O_t))^2}_{\text{avg. reward error}}$$

$$+ \frac{1}{\alpha}\underbrace{\|\mathbf{Q}_t^{\boldsymbol{\pi}_t} - \mathbf{Q}_{t+1}^{\boldsymbol{\pi}_{t+1}}\|_2^2}_{\text{value function drift}} + \underbrace{\alpha^2\|\mathbf{r}_t(O_t) - \boldsymbol{\eta}_t(O_t) + \mathbf{A}(O_t)\mathbf{Q}_t\|_2^2}_{\text{higher-order term}}$$

$$+ \underbrace{\alpha\psi_t^\top\left[(\mathbf{r}_t(O_t) - \mathbf{J}_t^{\boldsymbol{\pi}_t}(O_t) + \mathbf{A}(O_t)\mathbf{Q}_t^{\boldsymbol{\pi}_t}) + \left(\mathbf{A}(O_t) - \bar{\mathbf{A}}^{\boldsymbol{\pi}_t,\mathbf{P}_t}\right)\psi_t\right]}_{\text{error due to Markov noise}}$$

- Value function drift

$$\|\mathbf{Q}_t^{\boldsymbol{\pi}_t} - \mathbf{Q}_{t+1}^{\boldsymbol{\pi}_{t+1}}\|_2 \lesssim \|\boldsymbol{\pi}_{t+1} - \boldsymbol{\pi}_t\|_2 + \|\mathbf{r}_{t+1} - \mathbf{r}_t\|_\infty + \|\mathbf{P}_{t+1} - \mathbf{P}_t\|_\infty$$

- Insignificant for small enough α
- Bounded using Markov chain mixing
- η_t tracks average reward $\mathbf{J}_t^{\boldsymbol{\pi}_t}$

$$\eta_{t+1} = \eta_t + \gamma\left(r_t(s_t, a_t) - \eta_t\right)$$

Average Reward Estimation Error

The error $\phi_t = \eta_t - J_t^{\pi_t}$ decomposes as

$$\phi_{t+1}^2 \lesssim (1-\gamma)\phi_t^2 + \underbrace{\gamma(r_t(O_t) - J_t^{\pi_t})^2}_{\text{error due to Markov noise}} + \frac{1}{\gamma}\underbrace{(J_t^{\pi_t} - J_{t+1}^{\pi_{t+1}})^2}_{\text{avg reward at consecutive timesteps}}$$

$$+ \underbrace{\gamma^2(r_t(O_t) - \eta_t)^2}_{\text{higher order}}$$

- Insignificant for small enough γ
- Bounded using Markov chain mixing
- Can be bounded in terms of changes in policy and environment

$$J_t^{\pi_t} - J_{t+1}^{\pi_{t+1}} \lesssim \|\pi_{t+1} - \pi_t\|_2 + \|\mathbf{r}_{t+1} - \mathbf{r}_t\|_\infty + \|\mathbf{P}_{t+1} - \mathbf{P}_t\|_\infty$$

Bound on Markovian Noise

Original Markov chain

$$s_{t-\tau} \xrightarrow{\boldsymbol{\pi}_{t-\tau-1}} a_{t-\tau} \xrightarrow{\mathbf{P}_{t-\tau}} s_{t-\tau+1} \xrightarrow{\boldsymbol{\pi}_{t-\tau}} a_{t-\tau+1} \xrightarrow{\cdots} s_t \xrightarrow{\boldsymbol{\pi}_{t-1}} a_t \xrightarrow{\mathbf{P}_t} s_{t+1} \xrightarrow{\boldsymbol{\pi}_t} a_{t+1}.$$

Auxiliary Markov chain

$$s_{t-\tau} \xrightarrow{\boldsymbol{\pi}_{t-\tau-1}} a_{t-\tau} \xrightarrow{\mathbf{P}_{t-\tau}} \tilde{s}_{t-\tau+1} \xrightarrow{\boldsymbol{\pi}_{t-\tau-1}} \tilde{a}_{t-\tau+1} \xrightarrow{\cdots} \tilde{s}_t \xrightarrow{\boldsymbol{\pi}_{t-\tau-1}} \tilde{a}_t \xrightarrow{\mathbf{P}_{t-\tau}} \tilde{s}_{t+1} \xrightarrow{\boldsymbol{\pi}_{t-\tau-1}} \tilde{a}_{t+1}.$$

- Characterize the distance between the two chains

$$d_{TV}(P_{\text{original}}(\cdot|\mathcal{F}_{t-\tau}), P_{\text{aux}}(\cdot|\mathcal{F}_{t-\tau}))$$

  where $\mathcal{F}_{t-\tau} = \{s_{t-\tau}, \boldsymbol{\pi}_{t-\tau-1}, \mathbf{P}_{t-\tau}\}$
- Prior works use auxiliary Markov chains for stationary environments[1]
- Non-stationarity adds extra complexity - time-varying transition probabilities $\mathbf{P}_t$

[1] Wang, et al. "Non-asymptotic analysis for single-loop (natural) actor-critic with compatible function approximation." ICML (2024).

Bounding I_2

$$\text{Dyn-Reg}(\mathcal{M}, T) = \sum_{t=0}^{T-1} \underbrace{\mathbb{E}\left[J_t^{\boldsymbol{\pi}_t^\star} - J_t^{\boldsymbol{\pi}_t}\right]}_{I_1: \substack{\text{optimal versus} \\ \text{actual avg reward}}} + \underbrace{\mathbb{E}\left[J_t^{\boldsymbol{\pi}_t} - r_t(s_t, a_t)\right]}_{I_2: \substack{\text{actual avg versus} \\ \text{instantaneous reward}}}$$

Using auxiliary Markov chain

$$I_2 = \mathbb{E}\left[J_t^{\boldsymbol{\pi}_t} - r_t(s_t, a_t)\right] = \mathbb{E}\left[J_t^{\boldsymbol{\pi}_t} - r_t(s_t, a_t)\right]$$
$$\lesssim \sum_{i=t-\tau}^{t-1} \left(\|\mathbf{r}_{i+1} - \mathbf{r}_t\|_\infty + \|\mathbf{P}_{i+1} - \mathbf{P}_t\|_\infty\right) + m\rho^\tau$$

Summary of Proof Sketch

$$\text{Dyn-Reg}(\mathcal{M}, T) \triangleq \mathbb{E}\left[\sum_{t=0}^{T-1} J_t^{\pi_t^\star} - r_t(s_t, a_t)\right]$$

$$= \sum_{t=0}^{T-1} \underbrace{\mathbb{E}\left[J_t^{\pi_t^\star} - J_t^{\pi_t}\right]}_{I_1:\ \substack{\text{optimal versus} \\ \text{actual avg reward}}} + \underbrace{\mathbb{E}\left[J_t^{\pi_t} - r_t(s_t, a_t)\right]}_{I_2:\ \substack{\text{actual avg versus} \\ \text{instantaneous reward}}}$$

$$I_1 \lesssim \Delta_{\text{Environment}}(\Delta_R, \Delta_P)^1 + \ \text{Error}_{\text{Critic}}$$

$$\text{Error}_{\text{Critic}} \lesssim \Delta_{\text{Environment}} + \Delta_{\text{Policy}} + \text{Error}_{\text{Avg. Reward}} + m\rho^\tau$$

$$\text{Error}_{\text{Avg. Reward}} \lesssim \Delta_{\text{Environment}} + \Delta_{\text{Policy}} + m\rho^\tau$$

$$I_2 \lesssim \Delta_{\text{Environment}} + m\rho^\tau$$

---

[1] $\sum_{t=0}^{T-1} \|\mathbf{r}_{t+1} - \mathbf{r}_t\|_\infty, \quad \sum_{t=0}^{T-1} \|\mathbf{P}_{t+1} - \mathbf{P}_t\|_\infty$

Concluding Remarks

Lower Bound[1]

For any learning algorithm, there exists a non-stationary MDP such that the dynamic regret of the algorithm is at least

$$\Omega(|\mathcal{S}|^{1/3}|\mathcal{A}|^{1/3}D^{2/3}\Delta_T^{1/3}T^{2/3})$$

D is the diameter of the MDP

[1] Mao, et al. "Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control." Management Science (2025).

Gap between Bounds

- The gap results from a slack in Natural Actor-Critic (NAC) analysis

[1] Khodadadian, et al. "Finite-sample analysis of two-time-scale natural actor–critic algorithm." IEEE TAC (2022).

Gap between Bounds

- The gap results from a slack in Natural Actor-Critic (NAC) analysis
- We are forced to use a single-loop two-timescale algorithm

[1] Khodadadian, et al. "Finite-sample analysis of two-time-scale natural actor–critic algorithm." IEEE TAC (2022).

Gap between Bounds

- The gap results from a slack in Natural Actor-Critic (NAC) analysis
- We are forced to use a single-loop two-timescale algorithm
  - Single-loop - necessary due to time-varying environment

[1] Khodadadian, et al. "Finite-sample analysis of two-time-scale natural actor–critic algorithm." IEEE TAC (2022).

Gap between Bounds

- The gap results from a slack in Natural Actor-Critic (NAC) analysis
- We are forced to use a single-loop two-timescale algorithm
  - Single-loop - necessary due to time-varying environment
  - Two-timescale - our analysis forces us

[1] Khodadadian, et al. "Finite-sample analysis of two-time-scale natural actor–critic algorithm." IEEE TAC (2022).

Gap between Bounds

- The gap results from a slack in Natural Actor-Critic (NAC) analysis
- We are forced to use a single-loop two-timescale algorithm
  - Single-loop - necessary due to time-varying environment
  - Two-timescale - our analysis forces us
  - The best-known regret bounds for NAC for an infinite horizon *stationary* MDP with two-timescale algorithm is $\tilde{\mathcal{O}}(T^{3/4})$[1]

[1] Khodadadian, et al. "Finite-sample analysis of two-time-scale natural actor–critic algorithm." IEEE TAC (2022).

Potential Next Steps

- Matching the lower bound

Potential Next Steps

- Matching the lower bound
- Constrained settings - regret *and* constraint violation
  - Autonomous driving
  - Might need stronger notions of constraint violations

Potential Next Steps

- Matching the lower bound
- Constrained settings - regret *and* constraint violation
  - Autonomous driving
  - Might need stronger notions of constraint violations
- We study exogenous changes to environment - can we say more if changes are endogenous/performative?
  - Endogenous environment changes - Self-driving cars, financial trading
  - Endogenous reward changes - recommendation systems

Potential Next Steps

- Matching the lower bound
- Constrained settings - regret *and* constraint violation
  - Autonomous driving
  - Might need stronger notions of constraint violations
- We study exogenous changes to environment - can we say more if changes are endogenous/performative?
  - Endogenous environment changes - Self-driving cars, financial trading
  - Endogenous reward changes - recommendation systems
- How do these methods actually perform in practice?

Summary

- Model-free policy-based algorithm in the infinite-horizon average-reward setting

Summary

- Model-free policy-based algorithm in the infinite-horizon average-reward setting
- Natural Actor-critic with restarts

Summary

- Model-free policy-based algorithm in the infinite-horizon average-reward setting
- Natural Actor-critic with restarts
- First dynamic regret bound in this setting

Summary

- Model-free policy-based algorithm in the infinite-horizon average-reward setting
- Natural Actor-critic with restarts
- First dynamic regret bound in this setting
- Gap with lower bound - natural actor critic analysis
- What if we don't know the variation budget Δ_T?

Summary

- Model-free policy-based algorithm in the infinite-horizon average-reward setting
- Natural Actor-critic with restarts
- First dynamic regret bound in this setting
- Gap with lower bound - natural actor critic analysis
- What if we don't know the variation budget Δ_T?
    - We have an adaptive algorithm

Summary

- Model-free policy-based algorithm in the infinite-horizon average-reward setting
- Natural Actor-critic with restarts
- First dynamic regret bound in this setting
- Gap with lower bound - natural actor critic analysis
- What if we don't know the variation budget Δ_T?
  - We have an adaptive algorithm
- What about function approximation?

Summary

- Model-free policy-based algorithm in the infinite-horizon average-reward setting
- Natural Actor-critic with restarts
- First dynamic regret bound in this setting
- Gap with lower bound - natural actor critic analysis
- What if we don't know the variation budget Δ_T?
  - We have an adaptive algorithm
- What about function approximation?
  - We have result for linear function approximation

arXiv:2504.16415. On TMLR soon.

Shameless Self-Promotion - C-MInDS (IIT-B)

Parthe Pandit - High dim. stats, Kernel machines (AI2050 Early Career Fellowship from Schmidt Sciences)

Arjun Bhagoji - Robust and Reliable ML, ML for society

Pratik Jawanpuria - Optimization and optimal transport (was principal researcher at Microsoft)

- 2 more joining very soon
- 60+ associate faculty from 15 departments. **We are hiring!**
- 100+ graduate students. **Admissions in March-April.**
- PhD, MS(R), pre-doc and e-PG Diploma in DS and AI

Thank You
Questions?