A Novel Framework for Policy Mirror Descent with General Parameterization and Linear Convergence

Carlo Alfano[1], Rui Yuan[2], Patrick Rebeschini[1]

[1] University of Oxford, [2]Stellantis*

*The work was done prior to joining Stellantis,
while the author was at Télécom Paris

In Neural Information Processing Systems (Neurips), 2023.

Context

Objective: maximize the value function

Context

Objective: maximize the value function

- Natural policy gradient (NPG) [Kakade, 2001]

Context

Objective: maximize the value function

- Natural policy gradient (NPG) [Kakade, 2001]
- NPG is the building block of several state-of-the-art algorithms (TRPO, PPO)

Context

Objective: maximize the value function

- Natural policy gradient (NPG) [Kakade, 2001]
- NPG is the building block of several state-of-the-art algorithms (TRPO, PPO)
- Linear convergence of NPG is established for tabular case [Xiao, 2022]

Context

Objective: maximize the value function

- Natural policy gradient (NPG) [Kakade, 2001]
- NPG is the building block of several state-of-the-art algorithms (TRPO, PPO)
- Linear convergence of NPG is established for tabular case [Xiao, 2022]

Context

Objective: maximize the value function

- Natural policy gradient (NPG) [Kakade, 2001]
- NPG is the building block of several state-of-the-art algorithms (TRPO, PPO)
- Linear convergence of NPG is established for $\boxed{\text{tabular case}}$ [Xiao, 2022]
- Linear convergence of NPG is established for log-linear policy [Yuan et al., 2023]

Context

Objective: maximize the value function

- Natural policy gradient (NPG) [Kakade, 2001]
- NPG is the building block of several state-of-the-art algorithms (TRPO, PPO)
- Linear convergence of NPG is established for tabular case [Xiao, 2022]
- Linear convergence of NPG is established for log-linear policy [Yuan et al., 2023]

Context

Objective: maximize the value function

- Natural policy gradient (NPG) [Kakade, 2001]
- NPG is the building block of several state-of-the-art algorithms (TRPO, PPO)
- Linear convergence of NPG is established for $\boxed{\text{tabular case}}$ [Xiao, 2022]
- Linear convergence of NPG is established for $\boxed{\text{log-linear policy}}$ [Yuan et al., 2023]

Motivations

- Extend the linear convergence analysis of NPG from tabular and linear parametrization to general parametrization, including the neural network parametrization.

NPG with tabular as policy mirror descent

Linear convergence analysis of NPG [Xiao, 2022]

Lin Xiao
On the convergence rates of policy gradient methods. Journal of Machine Learning Research, 2022.

NPG with tabular as policy mirror descent

Linear convergence analysis of NPG [Xiao, 2022]

- Softmax tabular policies ($\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$) : no approximation

$$\pi^{(k)}(a \mid s) = \frac{\exp \theta^{(k)}(s,a)}{\sum_{a'\in\mathcal{A}} \exp \theta^{(k)}(s,a')}$$

Lin Xiao
On the convergence rates of policy gradient methods. Journal of Machine Learning Research, 2022.

NPG with tabular as policy mirror descent

Linear convergence analysis of NPG [Xiao, 2022]

- Softmax tabular policies ($\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$) : no approximation

$$\pi^{(k)}(a \mid s) = \frac{\exp \theta^{(k)}(s,a)}{\sum_{a' \in \mathcal{A}} \exp \theta^{(k)}(s,a')}$$

- NPG with softmax tabular policies as policy mirror descent

$$\pi^{(k+1)}(\,\cdot \mid s) \in \arg \max_{p \in \Delta(\mathcal{A})} \left\{ \boxed{\eta_k} \mathbb{E}_{a\sim p}\left[Q^{(k)}(s,a)\right] - \mathrm{KL}\left(p, \pi^{(k)}(\,\cdot \mid s)\right)\right\}, \quad \forall s \in \mathcal{S}$$

Step size

Lin Xiao
On the convergence rates of policy gradient methods. Journal of Machine Learning Research, 2022.

NPG with tabular as policy mirror descent

Linear convergence analysis of NPG [Xiao, 2022]

- Softmax tabular policies ($\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$) : no approximation

$$\pi^{(k)}(a \mid s) = \frac{\exp \theta^{(k)}(s,a)}{\sum_{a' \in \mathcal{A}} \exp \theta^{(k)}(s,a')}$$

- NPG with softmax tabular policies as policy mirror descent

$$\pi^{(k+1)}(\,\cdot \mid s) \in \arg\max_{p \in \Delta(\mathcal{A})} \left\{ \boxed{\eta_k} \mathbb{E}_{a\sim p}\left[Q^{(k)}(s,a)\right] - \mathrm{KL}\left(p, \pi^{(k)}(\,\cdot \mid s)\right)\right\}, \quad \forall s \in \mathcal{S}$$

Step size

Lin Xiao
On the convergence rates of policy gradient methods. Journal of Machine Learning Research, 2022.

NPG with tabular as policy mirror descent

Linear convergence analysis of NPG [Xiao, 2022]

- Softmax tabular policies ($\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$) : no approximation

$$\pi^{(k)}(a \mid s) = \frac{\exp \theta^{(k)}(s,a)}{\sum_{a' \in \mathcal{A}} \exp \theta^{(k)}(s,a')}$$

⚠ Not for large scale RL

- NPG with softmax tabular policies as policy mirror descent

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathcal{A})} \left\{ \boxed{\eta_k} \mathbb{E}_{a \sim p}\left[Q^{(k)}(s,a)\right] - \mathrm{KL}\left(p, \pi^{(k)}(\cdot \mid s)\right) \right\}, \quad \forall s \in \mathcal{S}$$

Step size

Lin Xiao
On the convergence rates of policy gradient methods. Journal of Machine Learning Research, 2022.

NPG with log-linear as policy mirror descent

Linear convergence analysis of NPG [Yuan et al., 2023]

Rui Yuan, Simon S. Du, Robert M. Gower, Alessandro Lazaric, Lin Xiao
Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies, ICLR, 2023.

NPG with log-linear as policy mirror descent

Linear convergence analysis of NPG [Yuan et al., 2023]

Feature map $\phi(s, a) \in \mathbb{R}^d$ over $\mathcal{S} \times \mathcal{A}$

- Log-linear policy ($\theta \in \mathbb{R}^d$)

$$\pi^{(k)}(a \mid s) = \frac{\exp \boxed{\phi(s, a)}^\top \theta^{(k)}}{\sum_{a' \in \mathcal{A}} \exp \phi(s, a')^\top \theta^{(k)}}$$

Rui Yuan, Simon S. Du, Robert M. Gower, Alessandro Lazaric, Lin Xiao
Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies, ICLR, 2023.

NPG with log-linear as policy mirror descent

Linear convergence analysis of NPG [Yuan et al., 2023]

Feature map $\phi(s, a) \in \mathbb{R}^d$ over $\mathcal{S} \times \mathcal{A}$

- Log-linear policy ($\theta \in \mathbb{R}^d$)

$$\pi^{(k)}(a \mid s) = \frac{\exp \boxed{\phi(s, a)}^\top \theta^{(k)}}{\sum_{a' \in \mathcal{A}} \exp \phi(s, a')^\top \theta^{(k)}}$$

- Q-NPG [Agarwal et al., 2021] with log-linear policies as policy mirror descent

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[\phi(s, a)^\top w_\star^{(k)}\right] - \mathrm{KL}\left(p, \pi^{(k)}(\cdot \mid s)\right) \right\}, \quad \forall s \in \mathcal{S}$$

Rui Yuan, Simon S. Du, Robert M. Gower, Alessandro Lazaric, Lin Xiao
Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies, ICLR, 2023.

NPG with log-linear as policy mirror descent

Linear convergence analysis of NPG [Yuan et al., 2023]

Feature map $\phi(s, a) \in \mathbb{R}^d$ over $\mathcal{S} \times \mathcal{A}$

- Log-linear policy ($\theta \in \mathbb{R}^d$)

$$\pi^{(k)}(a \mid s) = \frac{\exp \boxed{\phi(s,a)}^\top \theta^{(k)}}{\sum_{a' \in \mathcal{A}} \exp \phi(s,a')^\top \theta^{(k)}}$$

- Q-NPG [Agarwal et al., 2021] with log-linear policies as policy mirror descent

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[\phi(s,a)^\top w_\star^{(k)}\right] - \mathrm{KL}\left(p, \pi^{(k)}(\cdot \mid s)\right) \right\}, \quad \forall s \in \mathcal{S}$$

$w_\star^{(k)} \in \arg\min_{w \in \mathbb{R}^d} \mathbb{E}_{(s,a) \sim \mathcal{D}^{(k)}}\left[\left(\phi(s,a)^\top w - Q^{(k)}(s,a)\right)^2\right]$: compatible function approximation [Agarwal et al., 2021]

Rui Yuan, Simon S. Du, Robert M. Gower, Alessandro Lazaric, Lin Xiao
Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies, ICLR, 2023.

NPG with log-linear as policy mirror descent

Linear convergence analysis of NPG [Yuan et al., 2023]

Feature map $\phi(s,a) \in \mathbb{R}^d$ over $\mathcal{S} \times \mathcal{A}$

- Log-linear policy ($\theta \in \mathbb{R}^d$)

$$\pi^{(k)}(a \mid s) = \frac{\exp \boxed{\phi(s,a)}^\top \theta^{(k)}}{\sum_{a' \in \mathcal{A}} \exp \phi(s,a')^\top \theta^{(k)}}$$

- Q-NPG [Agarwal et al., 2021] with log-linear policies as policy mirror descent

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[\boxed{\phi(s,a)^\top w_\star^{(k)}}\right] - \mathrm{KL}\left(p, \pi^{(k)}(\cdot \mid s)\right) \right\}, \quad \forall s \in \mathcal{S}$$

Linear approximation

$$w_\star^{(k)} \in \arg\min_{w \in \mathbb{R}^d} \mathbb{E}_{(s,a) \sim \mathcal{D}^{(k)}}\left[\left(\boxed{\phi(s,a)^\top w - Q^{(k)}(s,a)}\right)^2\right]$$: compatible function approximation [Agarwal et al., 2021]

Rui Yuan, Simon S. Du, Robert M. Gower, Alessandro Lazaric, Lin Xiao
Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies, ICLR, 2023.

NPG with log-linear as policy mirror descent

Linear convergence analysis of NPG [Yuan et al., 2023]

Feature map $\phi(s,a) \in \mathbb{R}^d$ over $\mathcal{S} \times \mathcal{A}$

- Log-linear policy ($\theta \in \mathbb{R}^d$)

$$\pi^{(k)}(a \mid s) = \frac{\exp \boxed{\phi(s,a)}^\top \theta^{(k)}}{\sum_{a' \in \mathcal{A}} \exp \phi(s,a')^\top \theta^{(k)}}$$

- Q-NPG [Agarwal et al., 2021] with log-linear policies as policy mirror descent

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[\boxed{\phi(s,a)^\top w_\star^{(k)}}\right] - \mathrm{KL}\left(p, \pi^{(k)}(\cdot \mid s)\right) \right\}, \quad \forall s \in \mathcal{S}$$

Linear approximation : not sufficiently expressive ⚠

$$w_\star^{(k)} \in \arg\min_{w \in \mathbb{R}^d} \mathbb{E}_{(s,a) \sim \mathcal{D}^{(k)}}\left[\left(\boxed{\phi(s,a)^\top w - Q^{(k)}(s,a)}\right)^2\right]$$: compatible function approximation [Agarwal et al., 2021]

Rui Yuan, Simon S. Du, Robert M. Gower, Alessandro Lazaric, Lin Xiao
Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies, ICLR, 2023.

Approximate Mirror Policy Optimization (AMPO)

Approximate Mirror Policy Optimization (AMPO)

- General function parametrization ($\theta \in \Theta$)

$$\pi^{(k)}(a \mid s) = \frac{\exp\left(f^{\theta^{(k)}}(s, a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(f^{\theta^{(k)}}(s, a')\right)}$$

Approximate Mirror Policy Optimization (AMPO)

- General function parametrization ($\theta \in \Theta$)

$$\pi^{(k)}(a \mid s) = \frac{\exp\left(f^{\theta^{(k)}}(s, a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(f^{\theta^{(k)}}(s, a')\right)}$$

$f^{\theta}(s, a) = \theta(s, a)$:

$$\pi^{(k)}(a \mid s) = \frac{\exp \theta^{(k)}(s, a)}{\sum_{a' \in \mathcal{A}} \exp \theta^{(k)}(s, a')}$$

Softmax tabular policies

Approximate Mirror Policy Optimization (AMPO)

- General function parametrization ($\theta \in \Theta$)

$$\pi^{(k)}(a \mid s) = \frac{\exp\left(f^{\theta^{(k)}}(s,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(f^{\theta^{(k)}}(s,a')\right)}$$

$f^\theta(s,a) = \theta(s,a)$:

$$\pi^{(k)}(a \mid s) = \frac{\exp \theta^{(k)}(s,a)}{\sum_{a' \in \mathcal{A}} \exp \theta^{(k)}(s,a')}$$

Softmax tabular policies

$f^\theta(s,a) = \phi(s,a)^\top \theta$:

$$\pi^{(k)}(a \mid s) = \frac{\exp \phi(s,a)^\top \theta^{(k)}}{\sum_{a' \in \mathcal{A}} \exp \phi(s,a')^\top \theta^{(k)}}$$

Log-linear policies

Approximate Mirror Policy Optimization (AMPO)

Approximate Mirror Policy Optimization (AMPO)

- Step I: Generalized compatible function approximation

$$\theta^{(k+1)} \in \arg\min_{\theta \in \Theta} \mathbb{E}_{(s,a)\sim\mathcal{D}^{(k)}}\left[\left(f^{\theta}(s,a) - Q^{(k)}(s,a) - \eta_k^{-1}(\log \pi^{(k)}(a\,|\,s) + 1)\right)^2\right]$$

Approximate Mirror Policy Optimization (AMPO)

- Step I: Generalized compatible function approximation

$$\theta^{(k+1)} \in \arg\min_{\theta \in \Theta} \mathbb{E}_{(s,a)\sim\mathcal{D}^{(k)}}\left[\left(f^{\theta}(s,a) - Q^{(k)}(s,a) - \eta_k^{-1}(\log \pi^{(k)}(a \mid s) + 1)\right)^2\right]$$

- Step II: Policy mirror descent update

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \mathbb{E}_{a\sim p}\left[f^{\theta^{(k+1)}}(s,a) - \eta_k^{-1} \log \pi^{(k)}(a \mid s)\right] - \mathrm{KL}\left(p, \pi^{(k)}(\cdot \mid s)\right) \right\}$$

Convergence theory

Convergence theory

- Approximation error

$$\mathbb{E}_{(s,a)\sim\mathcal{D}^{(k)}}\left[\left(f^{\theta^{(k+1)}}(s,a) - Q^{(k)}(s,a) - \eta_k^{-1}(\log \pi^{(k)}(a\,|\,s) + 1)\right)^2\right] \le \epsilon_{\text{approx}}$$

Convergence theory

- Approximation error

$$\mathbb{E}_{(s,a)\sim\mathcal{D}^{(k)}}\left[\left(f^{\theta^{(k+1)}}(s,a) - Q^{(k)}(s,a) - \eta_k^{-1}(\log \pi^{(k)}(a\,|\,s) + 1)\right)^2\right] \le \epsilon_{\text{approx}}$$

- Linear convergence to the global optimum by increasing step size by $1/\gamma$

$$V^\star - \mathbb{E}[V(\theta^{(K)})] \le \mathcal{O}((1-c)^K) + \mathcal{O}(\epsilon_{\text{approx}}) \quad \text{with} \quad c \in (0,1)$$

Convergence theory

- Approximation error

$$\mathbb{E}_{(s,a)\sim\mathscr{D}^{(k)}}\left[\left(f^{\theta^{(k+1)}}(s,a) - Q^{(k)}(s,a) - \eta_k^{-1}(\log \pi^{(k)}(a\,|\,s) + 1)\right)^2\right] \le \epsilon_{\text{approx}}$$

- Linear convergence to the global optimum by increasing step size by $1/\gamma$

$$\boxed{V^\star} - \mathbb{E}[V(\theta^{(K)})] \le \mathcal{O}((1-c)^K) + \mathcal{O}(\epsilon_{\text{approx}}) \quad \text{with} \quad c \in (0,1)$$

Optimal

value function

Convergence theory

- Approximation error

$$\mathbb{E}_{(s,a)\sim\mathcal{D}^{(k)}}\Big[\Big(f^{\theta^{(k+1)}}(s,a) - Q^{(k)}(s,a) - \eta_k^{-1}(\log \pi^{(k)}(a\,|\,s) + 1)\Big)^2\Big] \leq \epsilon_{\text{approx}}$$

Neural networks: universal approximators, $\epsilon_{\text{approx}} \longrightarrow 0$

- Linear convergence to the global optimum by increasing step size by $1/\gamma$

$$\boxed{V^{\star}} - \mathbb{E}[V(\theta^{(K)})] \leq \mathcal{O}((1-c)^K) + \mathcal{O}(\epsilon_{\text{approx}}) \quad \text{with} \quad c \in (0,1)$$

Optimal

value function

Connection with Policy Iteration

Connection with Policy Iteration

- Connection with Policy Iteration

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathscr{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[Q^{(k)}(s, a)\right] \right\}$$

Connection with Policy Iteration

- Connection with Policy Iteration

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p\in\Delta(\mathscr{A})} \left\{\eta_k \mathbb{E}_{a\sim p}\left[Q^{(k)}(s,a)\right]\right\}$$

- AMPO with geometrically increasing step sizes

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p\in\Delta(\mathscr{A})} \left\{\eta_k \mathbb{E}_{a\sim p}\left[f^{\theta^{(k+1)}}(s,a) - \eta_k^{-1}\log \pi^{(k)}(a \mid s)\right] - \mathrm{KL}(p, \pi^{(k)}(\cdot \mid s))\right\}$$

Connection with Policy Iteration

- Connection with Policy Iteration

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathscr{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[Q^{(k)}(s,a) \right] \right\}$$

- AMPO with geometrically increasing step sizes

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathscr{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[f^{\theta^{(k+1)}}(s,a) - \eta_k^{-1} \log \pi^{(k)}(a \mid s) \right] - \mathrm{KL}(p, \pi^{(k)}(\cdot \mid s)) \right\}$$

$$\eta_k \longrightarrow \infty$$

Connection with Policy Iteration

- Connection with Policy Iteration

$$\pi^{(k+1)}(\,\cdot \mid s) \in \arg\max_{p \in \Delta(\mathscr{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[Q^{(k)}(s,a) \right] \right\}$$

- AMPO with geometrically increasing step sizes

$$\pi^{(k+1)}(\,\cdot \mid s) \in \arg\max_{p \in \Delta(\mathscr{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[f^{\theta^{(k+1)}}(s,a) - \eta_k^{-1} \log \pi^{(k)}(a \mid s) \right] - \mathrm{KL}(p, \cancel{\pi^{(k)}}(\,\cdot \mid s)) \right\}$$

$$\eta_k \longrightarrow \infty$$

Connection with Policy Iteration

- Connection with Policy Iteration

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathscr{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[Q^{(k)}(s,a)\right] \right\}$$

- AMPO with geometrically increasing step sizes

Function approximation

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathscr{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\Big[\underbrace{f^{\theta^{(k+1)}}(s,a) - \eta_k^{-1} \log \pi^{(k)}(a \mid s)}_{\approx Q^{(k)}(s,a)}\Big] - \cancel{\mathrm{KL}(p, \pi^{(k)}(\cdot \mid s))} \right\}$$

$\eta_k \longrightarrow \infty$

Connection with Policy Iteration

- Connection with Policy Iteration

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathscr{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\left[Q^{(k)}(s,a) \right] \right\}$$

- AMPO with geometrically increasing step sizes

Function approximation

$$\pi^{(k+1)}(\cdot \mid s) \in \arg\max_{p \in \Delta(\mathscr{A})} \left\{ \eta_k \mathbb{E}_{a \sim p}\Big[\underbrace{f^{\theta^{(k+1)}}(s,a) - \eta_k^{-1} \log \pi^{(k)}(a \mid s)}_{\approx Q^{(k)}(s,a)} \Big] - \cancel{\mathrm{KL}(p, \pi^{(k)}(\cdot \mid s))} \right\}$$

$\eta_k \longrightarrow \infty$

Behave more and more like policy iteration and enjoy fast linear convergence

Experimental results for AMPO

Classic control environment: Cart Pole & Acrobot

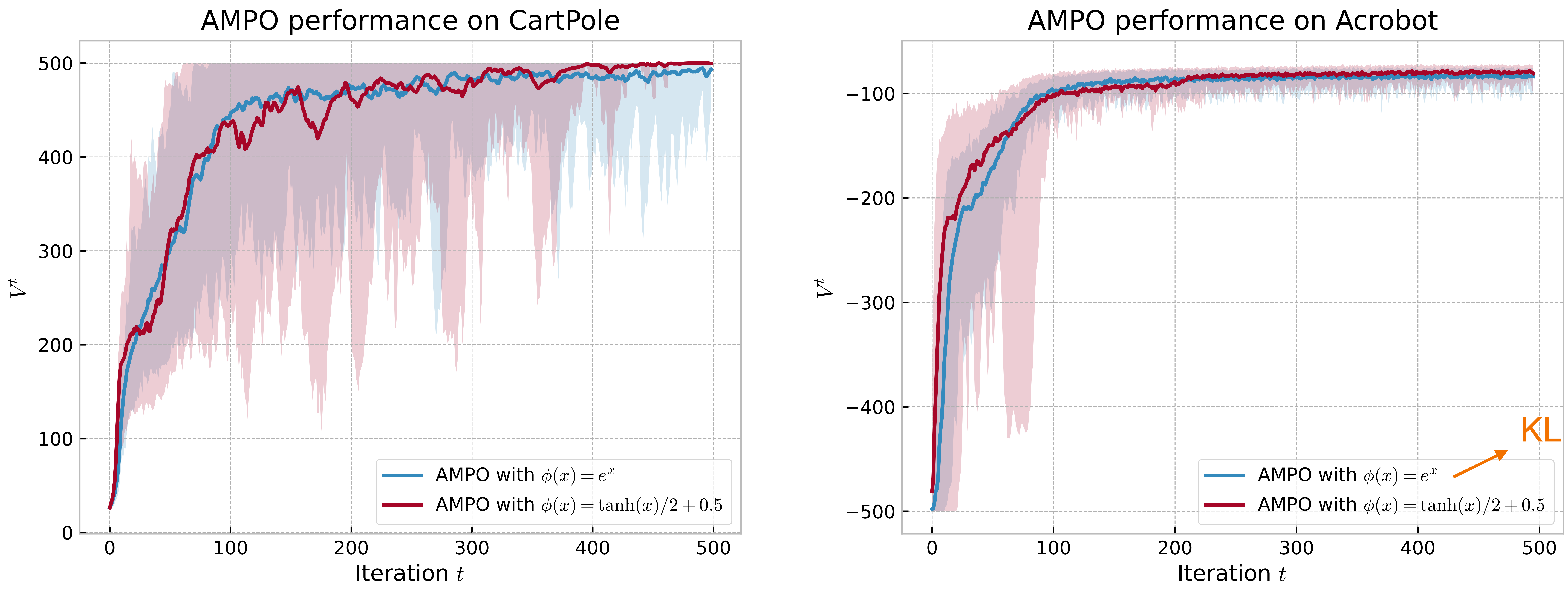

Figure: Experiments for AMPO with constant step size.

Discussion & Conclusion

Discussion & Conclusion

- A novel policy optimization framework AMPO that naturally accommodates general parametrization and enjoys linear convergence

Discussion & Conclusion

- A novel policy optimization framework AMPO that naturally accommodates general parametrization and enjoys linear convergence
- Apply AMPO to the offline setting

Discussion & Conclusion

- A novel policy optimization framework AMPO that naturally accommodates general parametrization and enjoys linear convergence
- Apply AMPO to the offline setting
- Design efficient policy evaluation algorithms and construct adaptive representation learning

Thank you !

References

- Sham M Kakade. A natural policy gradient. In Advances in Neural Information Processing Systems, volume 14. MIT Press, 2001.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 1889–1897, Lille, France, 07–09 Jul 2015.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.
- Lin Xiao. On the convergence rates of policy gradient methods. Journal of Machine Learning Research, 23(282):1–36, 2022.
- Rui Yuan, Simon S. Du, Robert M. Gower, Alessandro Lazaric, and Lin Xiao. Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies. In International Conference on Learning Representations, 2023.
- Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. Journal of Machine Learning Research 22.98, pp. 1–76, 2021.