Homotopic Policy Mirror Descent

Policy Convergence, Implicit Regularization, and Improved Sample Complexity

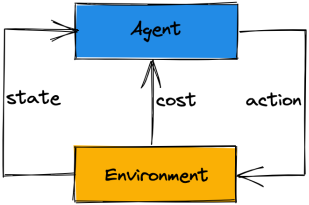
Yan Li

Georgia Institute of Technology

Joint work with Tuo Zhao, Guanghui (George) Lan

ICCOPT 2022

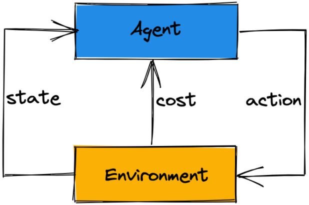
Markov decision process



Key elements:

- $\mathcal{S}$: state space, finite
- $\mathcal{A}$: action space, finite
- $\mathbb{P}$: transition kernel
- γ: discount factor
- c: costs

Markov decision process



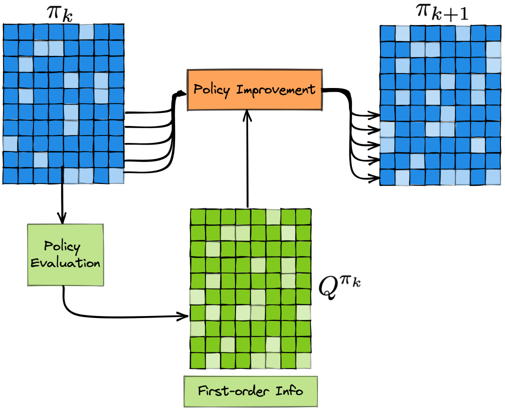
Key elements:

- $\mathcal{S}$: state space, finite
- $\mathcal{A}$: action space, finite
- $\mathbb{P}$: transition kernel
- γ: discount factor
- c: costs

- **Planning in $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathbb{P}, \gamma, c, h)$**:

$$\min_{\pi} V^{\pi}(s) := \mathbb{E}_{\pi}[\textstyle\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \big| s_0 = s], \quad \forall s \in \mathcal{S}$$

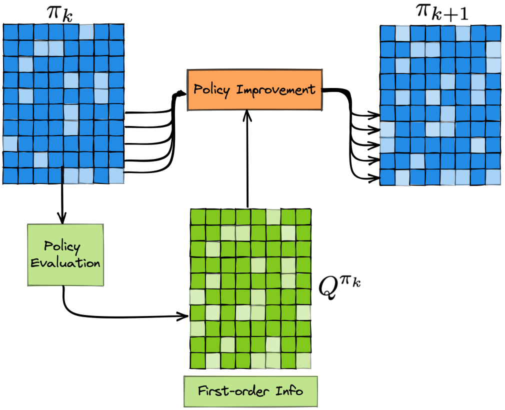
A Conceptual Recap on Policy Gradient Methods



- Q-function table: $Q^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ defined as

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t \left(c(s_t,a_t)\right) \middle| s_0 = s, a_0 = a\right]$$

A Conceptual Recap on Policy Gradient Methods



- **Single-objective:**

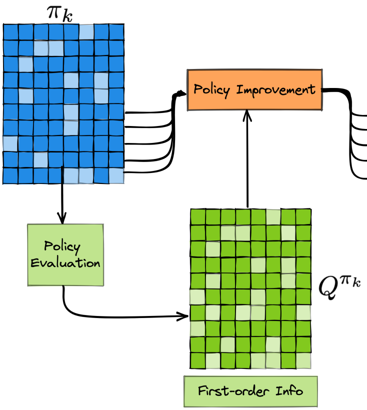
$$f(\pi) = \sum_{s \in \mathcal{S}} \nu^*(s) V^\pi(s)$$

$\star$ nonconvex

- Q-function table: $Q^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ defined as

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(c(s_t, a_t)\right) \middle| s_0 = s, a_0 = a \right]$$

A Conceptual Recap on Policy Gradient Methods



- **Single-objective:**

$$f(\pi) = \sum_{s \in \mathcal{S}} \nu^*(s) V^\pi(s)$$

  ⋆ nonconvex

- Policy evaluation:
  - ⋆ matrix inversion / fixed point iter.
  - ⋆ TD / simulator
- Policy improvement:
  - ⋆ policy gradient
  - ⋆ natural policy gradient

- Q-function table: $Q^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ defined as

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(c(s_t, a_t) \right) \middle| s_0 = s, a_0 = a \right]$$

Recent developments on Policy Gradient

- Possibly even earlier ..
- Even-Dar, Kakade, Mansour '09: $\mathcal{O}(1/\sqrt{T})$ regret of NPG
- Agarwal, Kakade, Lee, Mahajan '19: $\mathcal{O}(1/T)$ of NPG
  - technique inspired by Even-Dar, Kakade, Mansour '09
- Cen, Cheng, Chen, Wei, Chi '20: linear convergence of NPG for entropy regularized MDPs
- Lan '21: (approximate) policy mirror descent
  - linear convergence of NPG/PMD for entropy regularized MDPs
  - linear convergence of APMD for standard MDPs
  - linear convergence of stochastic variants and optimal sample complexity
- Khodadadian, Jhunjhunwala, Varma, Maguluri '21: linear convergence of NPG with adaptive stepsize for standard MDPs

More recently ..

- Xiao '22: linear convergence of NPG/PMD with increasing stepsize

And many more ...

What can be overlooked?

- Empirically, PG converges superlinearly at later stage:
    - With algorithmic-dependent assumptions: Khodadadian et al. '21, Xiao '22.
    - General arguments still missing.

What can be overlooked?

- Empirically, PG converges superlinearly at later stage:
  - With algorithmic-dependent assumptions: Khodadadian et al. '21, Xiao '22.
  - General arguments still missing.
- No clear understanding on the policy convergence:

What can be overlooked?

- Empirically, PG converges superlinearly at later stage:
  - With algorithmic-dependent assumptions: Khodadadian et al. '21, Xiao '22.
  - General arguments still missing.
- No clear understanding on the policy convergence:
  - Value convergence only implies subsequence convergence of policies.

What can be overlooked?

- Empirically, PG converges superlinearly at later stage:
  - With algorithmic-dependent assumptions: Khodadadian et al. '21, Xiao '22.
  - General arguments still missing.
- No clear understanding on the policy convergence:
  - Value convergence only implies subsequence convergence of policies.
  - Does $\{\pi_k\}$ even converge at all?

What can be overlooked?

- Empirically, PG converges superlinearly at later stage:
  - With algorithmic-dependent assumptions: Khodadadian et al. '21, Xiao '22.
  - General arguments still missing.
- No clear understanding on the policy convergence:
  - Value convergence only implies subsequence convergence of policies.
  - Does $\{\pi_k\}$ even converge at all?
  - There can be infinitely many optimal stochastic polices.

What can be overlooked?

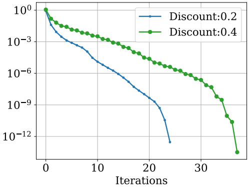
- Empirically, PG converges superlinearly at later stage:
  - With algorithmic-dependent assumptions: Khodadadian et al. '21, Xiao '22.
  - General arguments still missing.
- No clear understanding on the policy convergence:
  - Value convergence only implies subsequence convergence of policies.
  - Does $\{\pi_k\}$ even converge at all?
  - There can be infinitely many optimal stochastic polices.

$$\forall s \in \mathcal{S}, \mathrm{supp}(\pi(\cdot|s)) \subset \operatorname*{Argmin}_{a \in \mathcal{A}} Q^*(s,a) := \mathcal{A}^*(s)$$

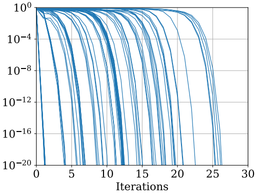
$$\Downarrow$$

$$\pi \in \Pi^*$$

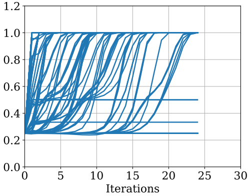
An Empirical Preview with GridWorld



(a) Avg Optimality Gap.



(b) $\sum_{a \notin \mathcal{A}^*(s)} \pi_k(a|s)$.



(c) $\min_{a \in \mathcal{A}^*(s)} \pi_k(a|s)$.

- Two-phase convergence? [Fig. (a), (b)]
  - Linear $\rightarrow$ Something even faster (perhaps superlinear)
- Implicit exploration? [Fig. (c)]
  - Probability strictly greater than 0 for any $a \in \mathcal{A}^*(s)$.

Presentation Outline

1. **Homotopic Policy Mirror Descent, and its Local Acceleration**
   - Method
   - Global linear convergence
   - Local super-linear convergence
2. **Policy Convergence**
   - With Kullback–Leibler divergence
   - Generalization to decomposable Bregman divergences
3. **Improved Sample Complexity**
4. **Conclusion**

Part I: HPMD and its Local Acceleration

Homotopic Policy Mirror Descent

Idea: diminishing entropy regularization in policy updates

Homotopic Policy Mirror Descent

Idea: diminishing entropy regularization in policy updates

Algorithm The homotopic policy mirror descent (HPMD) method

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$
for $k = 0, 1, \ldots$ **do**
  Update policy:

$$\pi_{k+1}(\cdot|s) = \underset{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \ \eta_k \left[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle - \tau_k \mathrm{Ent}(p)\right] + D^p_{\pi_k}(s)$$

end for

Homotopic Policy Mirror Descent

Idea: diminishing entropy regularization in policy updates

Algorithm The homotopic policy mirror descent (HPMD) method

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$

for $k = 0, 1, \ldots$ **do**

Update policy:

$$\pi_{k+1}(\cdot|s) = \underset{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \ \eta_k \left[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle - \tau_k \mathrm{Ent}(p)\right] + D^{p}_{\pi_k}(s)$$

end for

- $D^{\pi}_{\pi'}(s) := \mathrm{KL}(\pi(\cdot|s)\|\pi'(\cdot|s))$
- $\mathrm{Ent}(q) := -\sum_i q_i \log q_i$

Homotopic Policy Mirror Descent

Idea: diminishing entropy regularization in policy updates

Algorithm The homotopic policy mirror descent (HPMD) method

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$

for $k = 0, 1, \ldots$ **do**

Update policy:

$$\pi_{k+1}(\cdot|s) = \underset{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \ \eta_k \left[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle - \tau_k \mathrm{Ent}(p)\right] + D^{p}_{\pi_k}(s)$$

end for

- $D^{\pi}_{\pi'}(s) := \mathrm{KL}(\pi(\cdot|s)\|\pi'(\cdot|s))$
- $\mathrm{Ent}(q) := -\sum_i q_i \log q_i$
- Natural policy gradient (a.k.a. policy mirror descent) when $\tau_k = 0$.

Homotopic Policy Mirror Descent

Idea: diminishing entropy regularization in policy updates

Algorithm The homotopic policy mirror descent (HPMD) method

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k \geq 0}$

for $k = 0, 1, \ldots$ **do**

Update policy:

$$\pi_{k+1}(\cdot|s) = \underset{p(\cdot|s) \in \Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \; \eta_k \left[\langle Q^{\pi_k}(s, \cdot), p(\cdot|s) \rangle - \tau_k \mathrm{Ent}(p) \right] + D^{p}_{\pi_k}(s)$$

end for

- $D^{\pi}_{\pi'}(s) := \mathrm{KL}(\pi(\cdot|s) \| \pi'(\cdot|s))$
- $\mathrm{Ent}(q) := -\sum_i q_i \log q_i$
- Natural policy gradient (a.k.a. policy mirror descent) when $\tau_k = 0$.
- Still solves the original MDP ($\tau_k \to 0$)

Global Linear Convergence

Theorem (Li, Zhao, Lan '22)

By choosing $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$, *then for any iteration* $k \geq 1$,

$$f(\pi_k) - f(\pi^*) \leq \gamma^k \left(f(\pi_0) - f(\pi^*) + \frac{4 \log|\mathcal{A}|}{1-\gamma} \right).$$

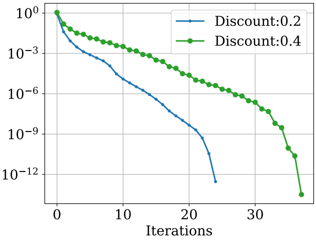


Figure: Avg Optimality Gap.

Global Linear Convergence

Theorem (Li, Zhao, Lan '22)

By choosing $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$, *then for any iteration* $k \geq 1$,

$$f(\pi_k) - f(\pi^*) \leq \gamma^k \left(f(\pi_0) - f(\pi^*) + \frac{4 \log |\mathcal{A}|}{1-\gamma} \right).$$

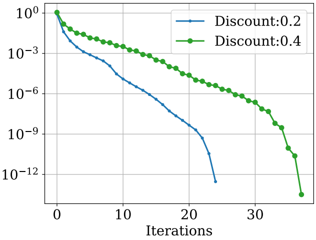


Figure: Avg Optimality Gap.

- Simplification to APMD (Lan '21).
  - regularization only in the update.

Global Linear Convergence

Theorem (Li, Zhao, Lan '22)

By choosing $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$, *then for any iteration* $k \geq 1$,

$$f(\pi_k) - f(\pi^*) \leq \gamma^k \left(f(\pi_0) - f(\pi^*) + \frac{4 \log|\mathcal{A}|}{1-\gamma} \right).$$

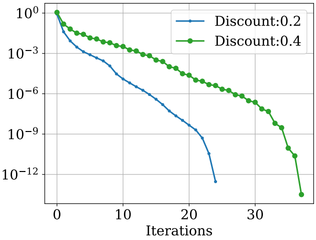


Figure: Avg Optimality Gap.

- Simplification to APMD (Lan '21).
  - regularization only in the update.
- Simple exponential stepsize scaling.

Global Linear Convergence

Theorem (Li, Zhao, Lan '22)

By choosing $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$, *then for any iteration* $k \geq 1$,

$$f(\pi_k) - f(\pi^*) \leq \gamma^k \left(f(\pi_0) - f(\pi^*) + \frac{4 \log |\mathcal{A}|}{1-\gamma} \right).$$

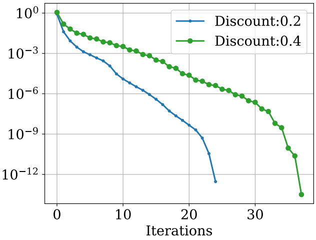


Figure: Avg Optimality Gap.

- Simplification to APMD (Lan '21).
  - regularization only in the update.
- Simple exponential stepsize scaling.
- $\mathcal{O}(\log k/k)$ rate with constant η_k and $\tau_k = 1/k$.

Conceptual Preview

Interactions between Value and Policy Convergence

$$\begin{tikzcd}
f(\pi_k) \xrightarrow{\text{linear}} \inf_{\pi} f(\pi) \arrow[r] & \pi_k \xrightarrow{\text{local sup. linear}} \Pi^* \arrow[dl] \\
f(\pi_k) \xrightarrow{\text{local sup. linear}} \inf_{\pi} f(\pi) \arrow[r] & \pi_k \to \pi^* \in \Pi^*
\end{tikzcd}$$

Local Superlinear Convergence - Policy

Theorem (Li, Zhao, Lan '22)

Suppose $\Delta^*(\mathcal{M}) < \infty$, *then with* $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$,

$$\mathrm{dist}_{\ell_1}(\pi_k, \Pi^*) = \mathcal{O}\left(\exp\left(-\frac{\Delta^*(\mathcal{M})}{2}\gamma^{-2k-1}\right)\right),$$

for any iteration $k \geq K_1 = \mathcal{O}\left(\log_\gamma \Delta^*(\mathcal{M})\right)$.

Local Superlinear Convergence - Policy

Theorem (Li, Zhao, Lan '22)

Suppose $\Delta^*(\mathcal{M}) < \infty$*, then with* $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$*,*

$$\mathrm{dist}_{\ell_1}(\pi_k, \Pi^*) = \mathcal{O}\left(\exp\left(-\frac{\Delta^*(\mathcal{M})}{2}\gamma^{-2k-1}\right)\right),$$

for any iteration $k \geq K_1 = \mathcal{O}\left(\log_\gamma \Delta^*(\mathcal{M})\right)$.

- $\|\pi - \pi'\|_1 := \max_{s \in \mathcal{S}} \|\pi(\cdot|s) - \pi'(\cdot|s)\|_1$.

Local Superlinear Convergence - Policy

Theorem (Li, Zhao, Lan '22)

Suppose $\Delta^*(\mathcal{M}) < \infty$*, then with* $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$,

$$\mathrm{dist}_{\ell_1}(\pi_k, \Pi^*) = \mathcal{O}\left(\exp\left(-\frac{\Delta^*(\mathcal{M})}{2}\gamma^{-2k-1}\right)\right),$$

for any iteration $k \geq K_1 = \mathcal{O}\left(\log_\gamma \Delta^*(\mathcal{M})\right)$.

- $\|\pi - \pi'\|_1 := \max_{s \in \mathcal{S}} \|\pi(\cdot|s) - \pi'(\cdot|s)\|_1$.
- $\Delta^*(\mathcal{M}) = \min_{s \in \mathcal{S}} \left\{\min_{a \notin \mathcal{A}^*(s)} Q^*(s,a) - \min_{a \in \mathcal{A}} Q^*(s,a)\right\}$.
    - Hardness of MDP.
- $\Delta^*(\mathcal{M}) = \infty \Rightarrow$ Any policy is optimal.

Proof Idea

What happens when $Q^*(s,i) < Q^*(s,j)$?

⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.

Proof Idea

What happens when $Q^*(s,i) < Q^*(s,j)$?

- ⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.
- ⋆ Look at the logit space:

Proof Idea

What happens when $Q^*(s,i) < Q^*(s,j)$?

- ⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.
- ⋆ Look at the logit space:

$$z_i^{k+1} - z_j^{k+1} = \underbrace{\gamma^{k+1}(z_i^0 - z_j^0)}_{\text{initialization does not matter}} - \sum_{t=0}^{k} \underbrace{\gamma^{k+1-t}\eta_t(Q_i^t - Q_j^t)}_{\text{Only recent history matters}}.$$

- ⋆ What happens in recent history?

Proof Idea

What happens when $Q^*(s,i) < Q^*(s,j)$?

- ⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.
- ⋆ Look at the logit space:

$$z_i^{k+1} - z_j^{k+1} = \underbrace{\gamma^{k+1}(z_i^0 - z_j^0)}_{\text{initialization does not matter}} - \underbrace{\textstyle\sum_{t=0}^{k} \gamma^{k+1-t}\eta_t(Q_i^t - Q_j^t)}_{\text{Only recent history matters}}.$$

- ⋆ What happens in recent history?

$$\underbrace{Q_i^t - Q_j^t \leq (Q^*(s,i) - Q^*(s,j))/2, t \geq K}_{\text{implied by linear convergence}}$$

Proof Idea

What happens when $Q^*(s,i) < Q^*(s,j)$?

⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.

⋆ Look at the logit space:

$$z_i^{k+1} - z_j^{k+1} = \underbrace{\gamma^{k+1}(z_i^0 - z_j^0)}_{\text{initialization does not matter}} - \sum_{t=0}^{k} \underbrace{\gamma^{k+1-t}\eta_t(Q_i^t - Q_j^t)}_{\text{Only recent history matters}}.$$

⋆ What happens in recent history?

$$\underbrace{Q_i^t - Q_j^t \leq (Q^*(s,i) - Q^*(s,j))/2, t \geq K}_{\text{implied by linear convergence}}$$

⋆ Recent history amplified by $\{\eta_t\}$:

$$z_i^{k+1} - z_j^{k+1} \geq \underbrace{-\frac{2C\gamma^2}{(1-\gamma^3)(1-\gamma)}}_{\text{old history}} + \underbrace{\frac{\gamma^{-2k-1}}{2}[Q^*(s,j) - Q^*(s,i)]}_{\text{recent history}}, \ k \geq K$$

Proof Idea

What happens when $Q^*(s,i) < Q^*(s,j)$?

- ⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.
- ⋆ Look at the logit space:

$$z_i^{k+1} - z_j^{k+1} = \underbrace{\gamma^{k+1}(z_i^0 - z_j^0)}_{\text{initialization does not matter}} - \sum_{t=0}^{k} \underbrace{\gamma^{k+1-t}\eta_t(Q_i^t - Q_j^t)}_{\text{Only recent history matters}}.$$

- ⋆ What happens in recent history?

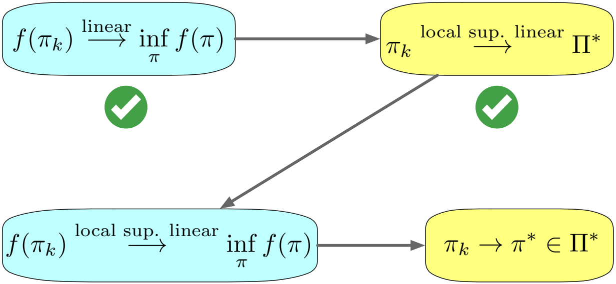
$$\underbrace{Q_i^t - Q_j^t \le (Q^*(s,i) - Q^*(s,j))/2, t \ge K}_{\text{implied by linear convergence}}$$

- ⋆ Recent history amplified by $\{\eta_t\}$:

$$z_i^{k+1} - z_j^{k+1} \ge \underbrace{-\frac{2C\gamma^2}{(1-\gamma^3)(1-\gamma)}}_{\text{old history}} + \underbrace{\frac{\gamma^{-2k-1}}{2}[Q^*(s,j) - Q^*(s,i)]}_{\text{recent history}}, \; k \ge K$$

$$\pi_k(j|s) = \mathcal{O}(\exp(-\gamma^{-2k}))$$

Interactions between Value and Policy Convergence



Local Superlinear Convergence - Value

Theorem (Li, Zhao, Lan '22)

Suppose $\Delta^*(\mathcal{M}) < \infty$, *then with* $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$,

$$f(\pi_k) - \inf_{\pi \in \Pi} f(\pi) = \mathcal{O}\left(\exp\left(-\frac{\Delta^*(\mathcal{M})}{2}\gamma^{-2k-1}\right)\right)$$

for any iteration $k \geq K_1$.

Local Superlinear Convergence - Value

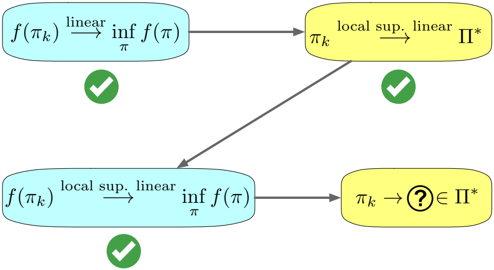
Theorem (Li, Zhao, Lan '22)

Suppose $\Delta^*(\mathcal{M}) < \infty$, *then with* $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$,

$$f(\pi_k) - \inf_{\pi \in \Pi} f(\pi) = \mathcal{O}\left(\exp\left(-\frac{\Delta^*(\mathcal{M})}{2}\gamma^{-2k-1}\right)\right)$$

for any iteration $k \geq K_1$.

- One-line proof by performance difference lemma.



Part II: Policy Convergence in HPMD

Policy Convergence with KL-divergence

Theorem (Li, Zhao, Lan '22)

Suppose $\Delta^*(\mathcal{M}) < \infty$, *then with* $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$,

$$\lim_{k \to \infty} \pi_k(a|s) = \pi_U^*(a|s) := \begin{cases} 1/\left|\mathcal{A}^*(s)\right|, & a \in \mathcal{A}^*(s), \\ 0, & a \notin \mathcal{A}^*(s). \end{cases}$$

Policy Convergence with KL-divergence

Theorem (Li, Zhao, Lan '22)

Suppose $\Delta^*(\mathcal{M}) < \infty$, *then with* $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$,

$$\lim_{k \to \infty} \pi_k(a|s) = \pi_U^*(a|s) := \begin{cases} 1/\left|\mathcal{A}^*(s)\right|, & a \in \mathcal{A}^*(s), \\ 0, & a \notin \mathcal{A}^*(s). \end{cases}$$

- First policy convergence result for PG methods.
  - Intuition: Entropy encourages exploration.

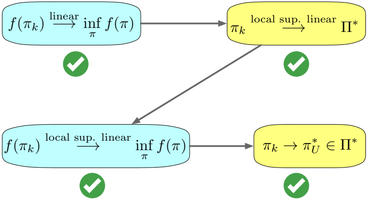
Policy Convergence with KL-divergence

Theorem (Li, Zhao, Lan '22)

Suppose $\Delta^*(\mathcal{M}) < \infty$, *then with* $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$,

$$\lim_{k\to\infty} \pi_k(a|s) = \pi_U^*(a|s) := \begin{cases} 1/\left|\mathcal{A}^*(s)\right|, & a \in \mathcal{A}^*(s), \\ 0, & a \notin \mathcal{A}^*(s). \end{cases}$$

- First policy convergence result for PG methods.
  - Intuition: Entropy encourages exploration.
- Implicit regularization: HPMD still solves the unregularized MDP.



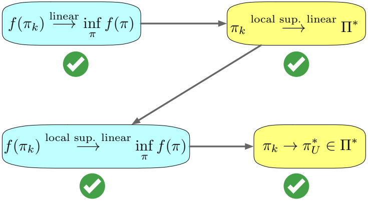
Policy Convergence with KL-divergence

Theorem (Li, Zhao, Lan '22)

Suppose $\Delta^*(\mathcal{M}) < \infty$, *then with* $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$,

$$\lim_{k \to \infty} \pi_k(a|s) = \pi_U^*(a|s) := \begin{cases} 1/\left|\mathcal{A}^*(s)\right|, & a \in \mathcal{A}^*(s), \\ 0, & a \notin \mathcal{A}^*(s). \end{cases}$$

- First policy convergence result for PG methods.
  - Intuition: Entropy encourages exploration.
- Implicit regularization: HPMD still solves the unregularized MDP.



⋆ **General Scenario**: Holds for constant stepsize HPMD ($\eta_k = \eta$, $\tau_k = 1/k$).

Proof Idea

What happens when $Q^*(s,i) = Q^*(s,j)$?

Proof Idea

What happens when $Q^*(s,i) = Q^*(s,j)$?

- ⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.
- ⋆ Look at the logit space:

$$z_i^{k+1} - z_j^{k+1} = \underbrace{\gamma^{k+1}(z_i^0 - z_j^0)}_{\text{initialization does not matter}} - \sum_{t=0}^{k} \underbrace{\gamma^{k+1-t}\eta_t(Q_i^t - Q_j^t)}_{\text{Only recent history matters}}.$$

- ⋆ What happens in recent history?

Proof Idea

What happens when $Q^*(s,i) = Q^*(s,j)$?

- ⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.
- ⋆ Look at the logit space:

$$z_i^{k+1} - z_j^{k+1} = \underbrace{\gamma^{k+1}(z_i^0 - z_j^0)}_{\text{initialization does not matter}} - \sum_{t=0}^{k} \underbrace{\gamma^{k+1-t}\eta_t(Q_i^t - Q_j^t)}_{\text{Only recent history matters}}.$$

- ⋆ What happens in recent history?

$$\underbrace{\left|Q_i^t - Q_j^t\right| \leq \left|Q_i^t - Q_i^*\right| + \left|Q_j^* - Q_j^t\right| = \mathcal{O}\left(\exp(-\gamma^{-2t})\right),\ t \geq K'}_{\text{implied by local sup. linear convergence}}$$

Proof Idea

What happens when $Q^*(s,i) = Q^*(s,j)$?

⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.

⋆ Look at the logit space:

$$z_i^{k+1} - z_j^{k+1} = \underbrace{\gamma^{k+1}(z_i^0 - z_j^0)}_{\text{initialization does not matter}} - \sum_{t=0}^{k} \underbrace{\gamma^{k+1-t}\eta_t(Q_i^t - Q_j^t)}_{\text{Only recent history matters}}.$$

⋆ What happens in recent history?

$$\underbrace{\left|Q_i^t - Q_j^t\right| \leq \left|Q_i^t - Q_i^*\right| + \left|Q_j^* - Q_j^t\right| = \mathcal{O}\left(\exp(-\gamma^{-2t})\right),\ t \geq K'}_{\text{implied by local sup. linear convergence}}$$

⋆ Recent history tampered by $\{\eta_t\}$?

Proof Idea

What happens when $Q^*(s,i) = Q^*(s,j)$?

- ⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.
- ⋆ Look at the logit space:

$$z_i^{k+1} - z_j^{k+1} = \underbrace{\gamma^{k+1}(z_i^0 - z_j^0)}_{\text{initialization does not matter}} - \sum_{t=0}^{k} \underbrace{\gamma^{k+1-t}\eta_t(Q_i^t - Q_j^t)}_{\text{Only recent history matters}}.$$

- ⋆ What happens in recent history?

$$\underbrace{\left|Q_i^t - Q_j^t\right| \le \left|Q_i^t - Q_i^*\right| + \left|Q_j^* - Q_j^t\right| = \mathcal{O}\left(\exp(-\gamma^{-2t})\right),\ t \ge K'}_{\text{implied by local sup. linear convergence}}$$

- ⋆ Recent history tampered by $\{\eta_t\}$?

$$\left|z_i^{k+1} - z_j^{k+1}\right| = \underbrace{\mathcal{O}(\gamma^k)}_{\text{initialization}} + \underbrace{\mathcal{O}(\gamma^k \cdot M)}_{\text{old history}} + \underbrace{\mathcal{O}(\gamma^k)}_{\text{recent history}},\ \forall k \ge K'$$

Proof Idea

What happens when $Q^*(s,i) = Q^*(s,j)$?

- ⋆ Notation shorthand: $z_i^k = \log \pi_k(i|s)$, $Q_i^k = Q^{\pi_k}(s,i)$.
- ⋆ Look at the logit space:

$$z_i^{k+1} - z_j^{k+1} = \underbrace{\gamma^{k+1}(z_i^0 - z_j^0)}_{\text{initialization does not matter}} - \sum_{t=0}^{k} \underbrace{\gamma^{k+1-t}\eta_t(Q_i^t - Q_j^t)}_{\text{Only recent history matters}}.$$

- ⋆ What happens in recent history?

$$\underbrace{\left|Q_i^t - Q_j^t\right| \leq \left|Q_i^t - Q_i^*\right| + \left|Q_j^* - Q_j^t\right| = \mathcal{O}\left(\exp(-\gamma^{-2t})\right),\ t \geq K'}_{\text{implied by local sup. linear convergence}}$$

- ⋆ Recent history tampered by $\{\eta_t\}$?

$$\left|z_i^{k+1} - z_j^{k+1}\right| = \underbrace{\mathcal{O}(\gamma^k)}_{\text{initialization}} + \underbrace{\mathcal{O}(\gamma^k \cdot M)}_{\text{old history}} + \underbrace{\mathcal{O}(\gamma^k)}_{\text{recent history}},\ \forall k \geq K'$$

$$\pi_k(j|s)/\pi_k(i|s) \to 1$$

Policy Convergence with Decomposable Divergences

HPMD with Decomposable Divergence

$$\pi_{k+1}(\cdot|s) = \operatorname*{argmin}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \eta_k \left[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + \tau_k w(p)\right] + D^p_{\pi_k}(s), \ \forall s \in \mathcal{S}.$$

Policy Convergence with Decomposable Divergences

HPMD with Decomposable Divergence

$$\pi_{k+1}(\cdot|s) = \operatorname*{argmin}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \eta_k \left[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + \tau_k w(p)\right] + D^{p}_{\pi_k}(s), \ \forall s \in \mathcal{S}.$$

- $D^{\pi}_{\pi'}(s)$ - Bregman divergence indued by w.
- Separable w: $w(p) = \sum_{i=1}^{|\mathcal{A}|} v(p_i)$, $v : \mathbb{R} \to \mathbb{R}$ is strictly convex, $\mathrm{dom}(v) \supset \mathbb{R}_+$, differentiable inside $\mathrm{dom}(v)$.

Policy Convergence with Decomposable Divergences

Theorem (Li, Zhao, Lan '22)

Suppose $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$, *and*

1. *Growth condition:* $\lim_{x \to \infty} v(x)/x = \infty$;
2. *Light-tail conjugate:* $\lim_{x \to \infty} \nabla \widehat{v}^*(-x)x = 0$, $\widehat{v}$ *is the restriction of* v *on* $\mathbb{R}_+$.

Then for any initial policy π_0 *satisfying* $\pi_0(a|s) > 0$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\lim_{k \to \infty} \pi_k(a|s) = \pi^*_U(a|s), \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Furthermore, if $\partial v(0) \neq \infty$, *then the above claim holds with any* π_0.

Policy Convergence with Decomposable Divergences

Theorem (Li, Zhao, Lan '22)

Suppose $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$, *and*

1. *Growth condition:* $\lim_{x \to \infty} v(x)/x = \infty$;
2. *Light-tail conjugate:* $\lim_{x \to \infty} \nabla \widehat{v}^*(-x)x = 0$, $\widehat{v}$ *is the restriction of* v *on* $\mathbb{R}_+$.

Then for any initial policy π_0 *satisfying* $\pi_0(a|s) > 0$ *for all* $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\lim_{k \to \infty} \pi_k(a|s) = \pi_U^*(a|s), \ \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Furthermore, if $\partial v(0) \neq \infty$, *then the above claim holds with any* π_0.

- Includes KL-divergence as a special case.

Policy Convergence with Decomposable Divergences

Theorem (Li, Zhao, Lan '22)

Suppose $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$, *and*

1. *Growth condition:* $\lim_{x \to \infty} v(x)/x = \infty$;
2. *Light-tail conjugate:* $\lim_{x \to \infty} \nabla \widehat{v}^*(-x)x = 0$, $\widehat{v}$ *is the restriction of* v *on* $\mathbb{R}_+$.

Then for any initial policy π_0 *satisfying* $\pi_0(a|s) > 0$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\lim_{k \to \infty} \pi_k(a|s) = \pi_U^*(a|s), \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Furthermore, if $\partial v(0) \neq \infty$, *then the above claim holds with any* π_0.

- Includes KL-divergence as a special case.
- Both conditions satisfied by many common regularizers: p-th power of ℓ_p-norm ($p > 1$), Tsallis entropy.

Policy Convergence with Decomposable Divergences

Theorem (Li, Zhao, Lan '22)

Suppose $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$*, and*

1. *Growth condition:* $\lim_{x \to \infty} v(x)/x = \infty$;
2. *Light-tail conjugate:* $\lim_{x \to \infty} \nabla \widehat{v}^*(-x)x = 0$, $\widehat{v}$ *is the restriction of* v *on* $\mathbb{R}_+$.

Then for any initial policy π_0 *satisfying* $\pi_0(a|s) > 0$ *for all* $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\lim_{k \to \infty} \pi_k(a|s) = \pi_U^*(a|s), \ \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Furthermore, if $\partial v(0) \neq \infty$*, then the above claim holds with any* π_0.

- Includes KL-divergence as a special case.
- Both conditions satisfied by many common regularizers: p-th power of ℓ_p-norm ($p > 1$), Tsallis entropy.
- Condition 1 can be removed with additional care.

Policy Convergence with Decomposable Divergences

Theorem (Li, Zhao, Lan '22)

Suppose $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-2(k+1)}$, *and*

1. *Growth condition:* $\lim_{x \to \infty} v(x)/x = \infty$;
2. *Light-tail conjugate:* $\lim_{x \to \infty} \nabla \widehat{v}^*(-x)x = 0$, $\widehat{v}$ *is the restriction of* v *on* $\mathbb{R}_+$.

Then for any initial policy π_0 *satisfying* $\pi_0(a|s) > 0$ *for all* $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\lim_{k \to \infty} \pi_k(a|s) = \pi^*_U(a|s), \ \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Furthermore, if $\partial v(0) \neq \infty$, *then the above claim holds with any* π_0.

- Includes KL-divergence as a special case.
- Both conditions satisfied by many common regularizers: p-th power of ℓ_p-norm ($p > 1$), Tsallis entropy.
- Condition 1 can be removed with additional care.
- The same limiting policy as KL? Why?

Understanding the Limiting Policy

Revisiting HPMD Update

$$\pi_{k+1}(\cdot|s) = \underset{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \ \eta_k \left[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + \tau_k w(p)\right] + D^{p}_{\pi_k}(s), \ \forall s \in \mathcal{S}.$$

⋆ Consider the following problem

$$\min_{\pi\in\Pi} \sum\nolimits_{s\in\mathcal{S}} w(\pi(\cdot|s)), \text{s.t. } f(\pi) \le f(\pi'), \forall \pi' \in \Pi.$$

Understanding the Limiting Policy

Revisiting HPMD Update

$$\pi_{k+1}(\cdot|s) = \underset{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \; \eta_k \left[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + \tau_k w(p)\right] + D^p_{\pi_k}(s), \;\; \forall s \in \mathcal{S}.$$

⋆ Consider the following problem

$$\min_{\pi\in\Pi} \textstyle\sum_{s\in\mathcal{S}} w(\pi(\cdot|s)), \text{s.t. } f(\pi) \leq f(\pi'), \forall \pi' \in \Pi.$$

- Constraint: π is optimal.
- Objective: minimize the complexity of π, measured by w.
- Minimizer(s)?

Understanding the Limiting Policy

Revisiting HPMD Update

$$\pi_{k+1}(\cdot|s) = \underset{p(\cdot|s)\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \ \eta_k \left[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + \tau_k w(p)\right] + D^p_{\pi_k}(s), \ \forall s \in \mathcal{S}.$$

⋆ Consider the following problem

$$\min_{\pi\in\Pi} \textstyle\sum_{s\in\mathcal{S}} w(\pi(\cdot|s)), \text{s.t. } f(\pi) \le f(\pi'), \forall \pi' \in \Pi.$$

- Constraint: π is optimal.
- Objective: minimize the complexity of π, measured by w.
- Minimizer(s)?
  - Unique minimizer $= \pi^*_U$ if w is decomposable and strictly convex.

Understanding the Limiting Policy

Revisiting HPMD Update

$$\pi_{k+1}(\cdot|s) = \operatorname*{argmin}_{p(\cdot|s)\in\Delta_{|\mathcal{A}|}} \eta_k \left[\langle Q^{\pi_k}(s,\cdot), p(\cdot|s)\rangle + \tau_k w(p)\right] + D^p_{\pi_k}(s), \ \forall s \in \mathcal{S}.$$

⋆ Consider the following problem

$$\min_{\pi\in\Pi} \textstyle\sum_{s\in\mathcal{S}} w(\pi(\cdot|s)), \text{s.t. } f(\pi) \leq f(\pi'), \forall \pi' \in \Pi.$$

- Constraint: π is optimal.
- Objective: minimize the complexity of π, measured by w.
- Minimizer(s)?
  - Unique minimizer $= \pi^*_U$ if w is decomposable and strictly convex.
- Analogous to homotopy methods (regularization path) in statistics (separable linear classification, Rosset et al. '04)

Specialization to Common Regularizers

Corollary (p-th power of ℓ_p-norm)

For any $p \in (1, \infty)$, let $v(x) = |x|^p$, then for any π_0, we have

- $\lim_{k\to\infty} \pi_k = \pi_U^*$.
- *There exists $K > 0$ such that* $f(\pi_k) = \inf_{\pi\in\Pi} f(\pi)$, $\forall k \geq K$.

Corollary (Negative Tsallis entropy)

Let $v(x) = \frac{k}{q-1}(x^q - 1/|\mathcal{A}|)$ for $x \geq 0$ and ∞ elsewhere. For any $q > 1$ and any π_0, we have

- $\lim_{k\to\infty} \pi_k = \pi_U^*$.
- *There exists $K > 0$ such that* $f(\pi_k) = \inf_{\pi\in\Pi} f(\pi)$, $\forall k \geq K$.

Specialization to Common Regularizers

Corollary (p-th power of ℓ_p-norm)

For any $p \in (1, \infty)$*, let* $v(x) = |x|^p$*, then for any* π_0*, we have*

- $\lim_{k \to \infty} \pi_k = \pi_U^*$.
- *There exists* $K > 0$ *such that* $f(\pi_k) = \inf_{\pi \in \Pi} f(\pi)$, $\forall k \geq K$.

Corollary (Negative Tsallis entropy)

Let $v(x) = \frac{k}{q-1}(x^q - 1/|\mathcal{A}|)$ *for* $x \geq 0$ *and* ∞ *elsewhere. For any* $q > 1$ *and any* π_0*, we have*

- $\lim_{k \to \infty} \pi_k = \pi_U^*$.
- *There exists* $K > 0$ *such that* $f(\pi_k) = \inf_{\pi \in \Pi} f(\pi)$, $\forall k \geq K$.

Specialization to Common Regularizers

Corollary (p-th power of ℓ_p-norm)

For any $p \in (1, \infty)$, let $v(x) = |x|^p$, then for any π_0, we have

- $\lim_{k\to\infty} \pi_k = \pi_U^*$.
- *There exists $K > 0$ such that $f(\pi_k) = \inf_{\pi\in\Pi} f(\pi)$, $\forall k \geq K$.*

Corollary (Negative Tsallis entropy)

Let $v(x) = \frac{k}{q-1}(x^q - 1/|\mathcal{A}|)$ for $x \geq 0$ and ∞ elsewhere. For any $q > 1$ and any π_0, we have

- $\lim_{k\to\infty} \pi_k = \pi_U^*$.
- *There exists $K > 0$ such that $f(\pi_k) = \inf_{\pi\in\Pi} f(\pi)$, $\forall k \geq K$.*

- The first finite-time convergence of PG methods.

Specialization to Common Regularizers

Corollary (p-th power of ℓ_p-norm)

For any $p \in (1, \infty)$, let $v(x) = |x|^p$, then for any π_0, we have

- $\lim_{k\to\infty} \pi_k = \pi_U^*$.
- *There exists $K > 0$ such that* $f(\pi_k) = \inf_{\pi\in\Pi} f(\pi)$, $\forall k \geq K$.

Corollary (Negative Tsallis entropy)

Let $v(x) = \frac{k}{q-1}(x^q - 1/|\mathcal{A}|)$ for $x \geq 0$ and ∞ elsewhere. For any $q > 1$ and any π_0, we have

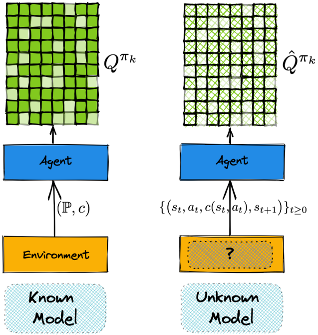
- $\lim_{k\to\infty} \pi_k = \pi_U^*$.
- *There exists $K > 0$ such that* $f(\pi_k) = \inf_{\pi\in\Pi} f(\pi)$, $\forall k \geq K$.

- The first finite-time convergence of PG methods.
- Policy moves towards π_U^* even π_k is already optimal.

Part III: Improved Sample Complexity of Stochastic HPMD

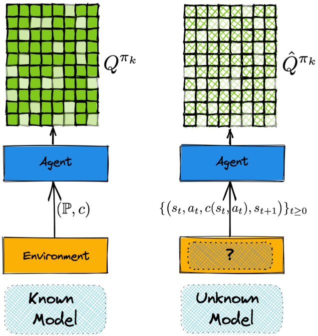
The Stochastic HPMD

Unknown Environment: obtaining exact Q^{π} can be impractical



The Stochastic HPMD

Unknown Environment: obtaining exact Q^π can be impractical



Independent Trajectories:

$$\xi_k = \{\zeta_k^i(s,a), s \in \mathcal{S}, a \in \mathcal{A}, i \in [M_k]\}$$

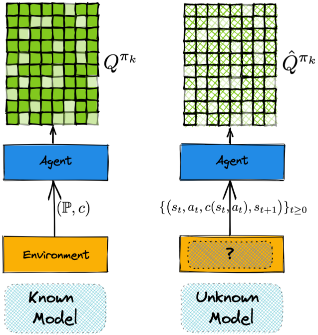
$$\zeta_k^i(s,a) = \{(s_0^i = s, a_0^i = a), \ldots, (s_{T_k-1}^i, a_{T_k-1}^i)\}$$

$$\Downarrow \text{ (Monte-Carlo)}$$

$$Q^{\pi_k, \xi_k}(s,a) = \frac{1}{M_k} \sum_{i=1}^{M_k} \sum_{t=0}^{T_k-1} \gamma^t c(s_t^i, a_t^i)$$

The Stochastic HPMD

Unknown Environment: obtaining exact Q^π can be impractical



Independent Trajectories:

$$\xi_k = \{\zeta_k^i(s,a), s \in \mathcal{S}, a \in \mathcal{A}, i \in [M_k]\}$$

$$\zeta_k^i(s,a) = \{(s_0^i = s, a_0^i = a), \ldots, (s_{T_k-1}^i, a_{T_k-1}^i)\}$$

$$\Downarrow \text{ (Monte-Carlo)}$$

$$Q^{\pi_k, \xi_k}(s,a) = \frac{1}{M_k} \sum_{i=1}^{M_k} \sum_{t=0}^{T_k-1} \gamma^t c(s_t^i, a_t^i)$$

Policy update: replace Q^π with sample estimate $Q^{\pi,\xi}$

Conditions on the Noisy Estimate

$$
\begin{aligned}
\mathbb{E}_{\xi_k} Q^{\pi_k,\xi_k} &= \overline{Q}^{\pi_k} \\
\|\overline{Q}^{\pi_k} - Q^{\pi_k}\|_\infty &\leq \varepsilon_k = \widetilde{\mathcal{O}}(\gamma^{T_k}), \quad \text{[bias]} \\
\mathbb{E}\|Q^{\pi_k,\xi_k} - Q^{\pi_k}\|_\infty^2 &\leq \sigma_k^2 = \widetilde{\mathcal{O}}(1/M_k), \quad \text{[variance]}
\end{aligned}
$$

Conditions on the Noisy Estimate

$$\mathbb{E}_{\xi_k} Q^{\pi_k,\xi_k} = \overline{Q}^{\pi_k}$$
$$\|\overline{Q}^{\pi_k} - Q^{\pi_k}\|_\infty \leq \varepsilon_k = \widetilde{\mathcal{O}}(\gamma^{T_k}), \quad \text{[bias]}$$
$$\mathbb{E}\|Q^{\pi_k,\xi_k} - Q^{\pi_k}\|_\infty^2 \leq \sigma_k^2 = \widetilde{\mathcal{O}}(1/M_k), \quad \text{[variance]}$$

Theorem (Li, Zhao, Lan '22)

Take $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-(k+1)/2}\sqrt{\log|\mathcal{A}|}$. *If*

$$\sigma_k = \gamma^{(k+1)/2}, \ \varepsilon_k = \gamma^{3(k+1)/4},$$

then

$$\mathbb{E}\left[f(\pi_k) - f(\pi^*)\right] \leq \gamma^{k/2}\frac{6\sqrt{\log|\mathcal{A}|}+C}{(1-\gamma)(1-\gamma^{1/2})\gamma}, \ \forall k \geq 1.$$

Conditions on the Noisy Estimate

$$\mathbb{E}_{\xi_k} Q^{\pi_k,\xi_k} = \overline{Q}^{\pi_k}$$

$$\|\overline{Q}^{\pi_k} - Q^{\pi_k}\|_\infty \le \varepsilon_k = \widetilde{\mathcal{O}}(\gamma^{T_k}), \quad \text{[bias]}$$

$$\mathbb{E}\|Q^{\pi_k,\xi_k} - Q^{\pi_k}\|_\infty^2 \le \sigma_k^2 = \widetilde{\mathcal{O}}(1/M_k), \quad \text{[variance]}$$

Theorem (Li, Zhao, Lan '22)

Take $1 + \eta_k \tau_k = 1/\gamma$ *and* $\eta_k = \gamma^{-(k+1)/2}\sqrt{\log|\mathcal{A}|}$. *If*

$$\sigma_k = \gamma^{(k+1)/2}, \ \varepsilon_k = \gamma^{3(k+1)/4},$$

then

$$\mathbb{E}\left[f(\pi_k) - f(\pi^*)\right] \le \gamma^{k/2}\frac{6\sqrt{\log|\mathcal{A}|}+C}{(1-\gamma)(1-\gamma^{1/2})\gamma}, \ \forall k \ge 1.$$

$\widetilde{\mathcal{O}}(|\mathcal{S}|\,|\mathcal{A}|\,/\epsilon^2)$ sample complexity.

Improved Sample Complexity with High Prob.

Theorem (Li, Zhao, Lan '22)

There exists ϵ_0, such that if $\epsilon < \epsilon_0$, then SHPMD outputs $\pi_{k(\epsilon)}$ satisfying $f(\pi_{k(\epsilon)}) - f(\pi^*) \leq \epsilon$ with probability $p(\epsilon)$, where

$$\epsilon_0 = \mathcal{O}(\Delta^*(\mathcal{M})^3)$$

$$k(\epsilon) = \mathcal{O}\Big(\underbrace{\log_\gamma(\epsilon_0)}_{\text{linear convergence}} + \underbrace{\log_\gamma\left(\frac{\Delta^*(\mathcal{M})\sqrt{\log|\mathcal{A}|}}{\log(C_\gamma^n|\mathcal{A}|/\epsilon(1-\gamma))}\right)}_{\text{local sup. linear}}\Big)$$

$$p(\epsilon) \geq 1 - \gamma^{k(\epsilon)/6}/(1-\gamma^{1/4})$$

The number of samples are bounded by

$$\widetilde{\mathcal{O}}\left(|\mathcal{S}|\,|\mathcal{A}|\,(1 + \log^2(1/\epsilon))\,/\epsilon_0^2\right)$$

Improved Sample Complexity with High Prob.

Theorem (Li, Zhao, Lan '22)

There exists ϵ_0*, such that if* $\epsilon < \epsilon_0$*, then SHPMD outputs* $\pi_{k(\epsilon)}$ *satisfying* $f(\pi_{k(\epsilon)}) - f(\pi^*) \leq \epsilon$ *with probability* $p(\epsilon)$*, where*

$$\epsilon_0 = \mathcal{O}(\Delta^*(\mathcal{M})^3)$$

$$k(\epsilon) = \mathcal{O}\Big(\underbrace{\log_\gamma(\epsilon_0)}_{\text{linear convergence}} + \underbrace{\log_\gamma\left(\frac{\Delta^*(\mathcal{M})\sqrt{\log|\mathcal{A}|}}{\log(C_\gamma^n|\mathcal{A}|/\epsilon(1-\gamma))}\right)}_{\text{local sup. linear}} \Big)$$

$$p(\epsilon) \geq 1 - \gamma^{k(\epsilon)/6}/(1-\gamma^{1/4})$$

The number of samples are bounded by

$$\widetilde{\mathcal{O}}\left(|\mathcal{S}|\,|\mathcal{A}|\,(1 + \log^2(1/\epsilon))\,/\epsilon_0^2 \right)$$

- The sample complexity grows logarithmically after a threshold.

Improved Sample Complexity with High Prob.

Theorem (Li, Zhao, Lan '22)

There exists ϵ_0, such that if $\epsilon < \epsilon_0$, then SHPMD outputs $\pi_{k(\epsilon)}$ satisfying $f(\pi_{k(\epsilon)}) - f(\pi^*) \leq \epsilon$ with probability $p(\epsilon)$, where

$$\epsilon_0 = \mathcal{O}(\Delta^*(\mathcal{M})^3)$$

$$k(\epsilon) = \mathcal{O}\Big(\underbrace{\log_\gamma (\epsilon_0)}_{\text{linear convergence}} \underbrace{+ \log_\gamma \left(\frac{\Delta^*(\mathcal{M})\sqrt{\log|\mathcal{A}|}}{\log(C_\gamma^n |\mathcal{A}|/\epsilon(1-\gamma))} \right)}_{\text{local sup. linear}} \Big)$$

$$p(\epsilon) \geq 1 - \gamma^{k(\epsilon)/6}/(1-\gamma^{1/4})$$

The number of samples are bounded by

$$\widetilde{\mathcal{O}}\big(|\mathcal{S}|\,|\mathcal{A}|\,(1 + \log^2(1/\epsilon))\,/\epsilon_0^2\big)$$

- The sample complexity grows logarithmically after a threshold.
- Local acceleration carries to the stochastic setting, but, only with high probability.

Conclusion

- HPMD (KL-divergence): global linear and local superlinear convergence.

Conclusion

- HPMD (KL-divergence): global linear and local superlinear convergence.
- Characterizes the last-iterate convergence of the policy.
  - With exponentially increasing stepsize – linear convergence.
  - With constant stepsize – sublinear convergence.

Conclusion

- HPMD (KL-divergence): global linear and local superlinear convergence.
- Characterizes the last-iterate convergence of the policy.
  - With exponentially increasing stepsize – linear convergence.
  - With constant stepsize – sublinear convergence.
- Generalization to common divergences.
  - Policy convergence.
  - Finite-time exact value convergence.

Conclusion

- HPMD (KL-divergence): global linear and local superlinear convergence.
- Characterizes the last-iterate convergence of the policy.
  - With exponentially increasing stepsize – linear convergence.
  - With constant stepsize – sublinear convergence.
- Generalization to common divergences.
  - Policy convergence.
  - Finite-time exact value convergence.
- Improved sample complexity for the stochastic variant.

Presentation based on Preprint

- Li, Y., Zhao, T. and Lan, G., 2022. Homotopic Policy Mirror Descent: Policy Convergence, Implicit Regularization, and Improved Sample Complexity. arXiv preprint arXiv:2201.09457.