Towards Understanding Reinforcement Learning from Optimization Perspectives

Shaocong Ma

University of Utah

s.ma@utah.edu

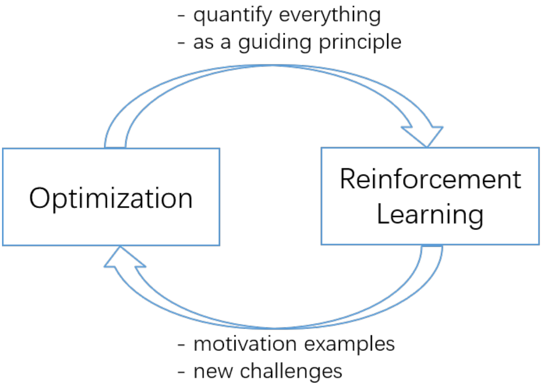
November 12, 2021

Shaocong Ma

Background

- Third-year Ph.D. student in EE at University of Utah.
- M.A. Degree and B.S. Degree in Statistics.

Research

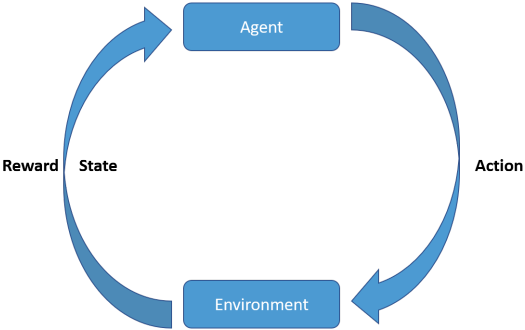


Overview

1. Challenge 1: Non-Independent Data
   - Reduce the influence of data dependence
   - Classical optimization techniques on dependent data
   - Critical thinking: is data dependence always bad?
2. Challenge 2: Exploration-Exploitation Trade-Off
   - Quantify the error caused by lacking of exploration
3. Reference

Challenges from RL: Non-Independent Data

Dataset in Reinforcement Learning

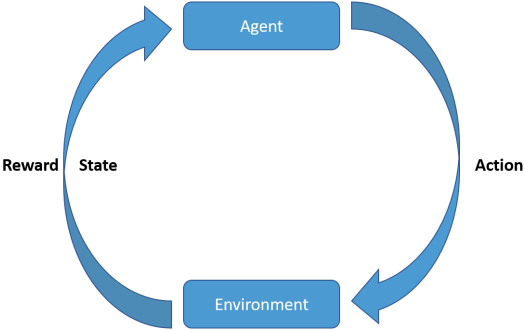


The data point (s_t, a_t, r_t, s_{t+1}) in RL comes from a trajectory:

$$s_1, a_1, r_1, s_2, a_2, r_2, \dots$$

Challenges from RL: Non-Independent Data

Dataset in Reinforcement Learning



The data point (s_t, a_t, r_t, s_{t+1}) in RL comes from a trajectory:

$$s_1, a_1, r_1, s_2, a_2, r_2, \ldots$$

$\{(s_i, a_i, r_i, s_{i+1})\}$ and $\{(s_j, a_j, r_j, s_{j+1})\}$ are non-independent!

Ultimate Goal of RL Find a strategy π of selecting action to maximize the future return:

$$\max_{\pi} Q^{\pi}(s,a) := \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t r_t | s, a]$$

Deep Q-Learning (DQN) with Target Network [DeepMind'13]

$$\theta_{k+1} \leftarrow \underbrace{\arg\min \mathbb{E}_{(s,a,r,s')\sim\mu} \|r + \gamma \max_{a'} Q_{\theta_k}(s',a') - Q_{\theta}(s,a)\|^2}_{\text{An optimization problem!}}$$

where μ is the stat. dist. of the stochastic process $\{(s_t, a_t, r_t, s_{t+1})\}$.

Ultimate Goal of RL Find a strategy π of selecting action to maximize the future return:

$$\max_{\pi} Q^{\pi}(s, a) := \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t r_t | s, a]$$

Deep Q-Learning (DQN) with Target Network [DeepMind'13]

$$\theta_{k+1} \leftarrow \underbrace{\arg\min \mathop{\mathbb{E}}_{(s,a,r,s')\sim\mu} \|r + \gamma \max_{a'} Q_{\theta_k}(s', a') - Q_{\theta}(s, a)\|^2}_{\text{An optimization problem!}}$$

where μ is the stat. dist. of the stochastic process $\{(s_t, a_t, r_t, s_{t+1})\}$.

Key difference: non-independent data

A general question Solve the optimization problem

$$\min_x \mathbb{E}_{\xi\sim\mu} f(x;\xi)$$

given a stochastic process $\{\xi_t\}$. How does it influence the optimization?

- RL applications: (double) Q-learning, Actor-Critic, PPO, and etc.

Existing work [Agarwal'12] With a high-probability,

$$\underbrace{\mathbb{E}_{\xi\sim\mu} f(\bar{x}_t;\xi) - \min_x \mathbb{E}_{\xi\sim\mu} f(x;\xi)}_{\text{opt. error}} \leq \mathcal{O}(\frac{1}{\sqrt{t}}) + \underbrace{\mathcal{O}(\sqrt{\frac{\tau}{t}} + \phi(\tau))}_{\text{data dependence}},$$

where $\phi(\tau) := \sup_k \sup_{A\in\mathcal{F}_k} d_{\mathrm{TV}}(\mathbb{P}(\xi_{\tau+k} \in \cdot|A), \mu)$.

Outline

1. Challenge 1: Non-Independent Data
   - Reduce the influence of data dependence
   - Classical optimization techniques on dependent data
   - Critical thinking: is data dependence always bad?

2. Challenge 2: Exploration-Exploitation Trade-Off
   - Quantify the error caused by lacking of exploration

3. Reference

Question How can we reduce the influence of data dependence?
Answer Just use a large batch size.

Our work [ICLR'22 - under review]

Data dependence level	$\phi(k)$	SGD	Mini-batch SGD
Geometric ϕ-mixing (Weakly dependent)	$\exp(-k^{\theta})$, $\theta > 0$	$\mathcal{O}(\epsilon^{-2}(\log \epsilon^{-1})^{\frac{2}{\theta}})$	$\mathcal{O}(\epsilon^{-2})$
Fast algebraic ϕ-mixing (Medium dependent)	$k^{-\theta}$, $\theta \geq 1$	$\mathcal{O}(\epsilon^{-2-\frac{2}{\theta}})$	$\widetilde{\mathcal{O}}(\epsilon^{-2})$
Slow algebraic ϕ-mixing (Highly dependent)	$k^{-\theta}$, $0 < \theta < 1$	$\mathcal{O}(\epsilon^{-2-\frac{2}{\theta}})$	$\mathcal{O}(\epsilon^{-1-\frac{1}{\theta}})$

How does this idea work?

- Reduce the variance:

$$
\begin{aligned}
&\text{(single)} \quad \mathbb{E}\|f(x;\xi_t) - \mathbb{E}_{\xi\sim\mu} f(x;\xi)\|^2 \approx \mathcal{O}(1) \\
&\text{(mini-batch)} \quad \mathbb{E}\|\frac{1}{B}\sum_{i=1}^{B} f(x;\xi_{t+i}) - \mathbb{E}_{\xi\sim\mu} f(x;\xi)\|^2 \approx \mathcal{O}(\frac{1}{B})
\end{aligned}
$$

- Reduce the bias:

$$
\begin{aligned}
&\text{(single)} \quad \mathbb{E}_{\xi_\tau} f(x;\xi_\tau) - \mathbb{E}_{\xi\sim\mu} f(x;\xi) \approx \phi(\tau) \\
&\text{(mini-batch)} \quad \frac{1}{B}\sum_{i=1}^{B} \mathbb{E}_{\xi_{\tau+i}} f(x;\xi_{\tau+i}) - \mathbb{E}_{\xi\sim\mu} f(x;\xi) \approx \frac{1}{B}\sum_{i=1}^{B}\phi(\tau+i)
\end{aligned}
$$

- Put them back to [Agarwal'12]:

$$
\text{opt. error} \le \mathcal{O}(\frac{1}{\sqrt{tB}}) + \underbrace{\mathcal{O}(\sqrt{\frac{\tau}{tB} + \frac{1}{B}\sum_{i=1}^{B}\phi(i)})}_{\text{data dependence}}.
$$

Many RL problems have highly dependent data!

- Markovian decision process admitting specific jump diffusion; e.g. financial market, self-driving car, and etc.
- Bad replay buffer; e.g.

$$\{\xi_1\}, \{\xi_1, \xi_2\}, \{\xi_1, \xi_2, \xi_3\}, \ldots.$$

- Exploration with a updating policy.

Outline

1. Challenge 1: Non-Independent Data
   - Reduce the influence of data dependence
   - Classical optimization techniques on dependent data
   - Critical thinking: is data dependence always bad?
2. Challenge 2: Exploration-Exploitation Trade-Off
   - Quantify the error caused by lacking of exploration
3. Reference

Question What is the influence of data dependence on those classical optimization techniques such as variance reduction?

Answer The performance of variance reduction is reduced.

Recap on Variance Reduction

$$\begin{aligned} &\text{(SGD)} \quad \nabla f(x;\xi) \\ &\text{(SVRG)} \quad \nabla f(x;\xi) - \nabla f(y;\xi) + \mathbb{E}_{\xi\sim\mu}\nabla f(y;\xi) \end{aligned}$$

- For IID data, they are both unbiased while SVRG has lower variance when $\|x-y\|^2$ is small.
- For Markovian data, the bias may dominates the error term.

We apply the variance reduction technique to two existing gradient-based RL algorithms: TD learning with gradient correction (TDC) and Greedy-GQ algorithm.

Our work [NeurIPS'20]

	TDC	VR-TDC
IID	$\tilde{\mathcal{O}}(\epsilon^{-1})$	$\tilde{\mathcal{O}}(\epsilon^{-\frac{3}{5}})$
Markovian	$\tilde{\mathcal{O}}(\epsilon^{-1})$	$\tilde{\mathcal{O}}(\epsilon^{-1})$

Our work [ICLR'21]

	Greedy-GQ	VR-Greedy-GQ	SVRG
Markovian	$\tilde{\mathcal{O}}(\epsilon^{-3})$	$\tilde{\mathcal{O}}(\epsilon^{-2})$	-
IID	-	-	$\mathcal{O}(\epsilon^{-\frac{5}{3}})$

Outline

1. Challenge 1: Non-Independent Data
   - Reduce the influence of data dependence
   - Classical optimization techniques on dependent data
   - Critical thinking: is data dependence always bad?

2. Challenge 2: Exploration-Exploitation Trade-Off
   - Quantify the error caused by lacking of exploration

3. Reference

Question Does the data dependence always make the algorithm perform worse?

Answer No. Sometimes, the dependence makes it better!

Our work [ICML'20]

- The empirical risk minimization problem:

$$\min_x \frac{1}{n} \sum_{i=1}^{n} \ell_i(x).$$

- We show that sampling with reshuffle is better than IID sampling.

Outline

1. Challenge 1: Non-Independent Data
   - Reduce the influence of data dependence
   - Classical optimization techniques on dependent data
   - Critical thinking: is data dependence always bad?

2. Challenge 2: Exploration-Exploitation Trade-Off
   - Quantify the error caused by lacking of exploration

3. Reference

Question How can we theoretically understand Exploration-Exploitation trade-off?

Answer We need to quantify the error caused by lacking of exploration.

Our work [ICML'22 - To be submitted]

- Given the off-line data $\mathcal{D}$, what is the best performance achieved by Q-learning?
- Bound the gap to optimal value function:

$$
\begin{aligned}
&(1-\gamma)\mathbb{E}_{s\sim\mu_0}[V^*(s) - V^{\pi^{(K)}}(s)] \\
&\le \underbrace{\frac{2}{1-\gamma}\sqrt{C\cdot(\epsilon_{\text{approx}} + \frac{1}{|\mathcal{D}|})} + 2\gamma^K \|Q^* - Q^{(0)}\|_{2,\tilde{\nu}}}_{\text{Standard error of off-line Q-learning}} \\
&+ M\cdot \underbrace{\sum_{k=0}^{K-1}\gamma^k\sqrt{\nu_{K-k}(\mathcal{D}^c)}}_{\text{Exploration error}} + M\cdot \underbrace{\sum_{k=0}^{K-1}\gamma^k\sqrt{\nu^*_{K-k}(\mathcal{D}^c)}}_{\text{Exploration error}}.
\end{aligned}
$$

- Greedy policy defined by a Q-function:

$$\pi(a|s) = \begin{cases} 1 & a = \arg\max_{a \in \mathcal{A}} Q(s,a) \\ 0 & \text{o.w.} \end{cases}.$$

$\pi^{(k)}$ is the greedy policy defined by the Q-function at k-th iteration.

- State visitation measure of a policy π:

$$d^{\pi} := (1-\gamma)\mathbb{E}\sum_{i=0}^{\infty} \gamma^t 1(s_t = s)$$

where $\{s_t\}$ is generated via the policy π. And

$$\nu_k := d^{\pi^{(k)}} \otimes \pi^{(k)}$$

is the greedy-policy state-action visitation measure;

$$\nu_k^* := d^{\pi^{(k)}} \otimes \pi^*$$

is the optimal policy state-action visitation measure.

Exploration error:

$$\epsilon_{\text{exploration}} = \sum_{k=0}^{K-1} \gamma^k \sqrt{\nu_{K-k}(\mathcal{D}^c)} + \sum_{k=0}^{K-1} \gamma^k \sqrt{\nu^*_{K-k}(\mathcal{D}^c)}.$$

- More efficient exploration strategy:
    - For each episode, it suffices to explore all possible state-action pairs generated by the target greedy policy AND one-step action taken by optimal policy.
    - Optimal exploration strategy: One-step Monte Carlo Tree Search.
- More reasonable replay buffer design:
    - All state-action pairs generated by greedy-policy are important. Don't delete them until the next epoch.
- ...

- [Agarwal'12] Agarwal, Alekh, and John C. Duchi. "The generalization ability of online algorithms for dependent data." IEEE Transactions on Information Theory 59.1 (2012): 573-587.
- [NeurIPS'20] Ma, Shaocong, Yi Zhou, and Shaofeng Zou. "Variance-Reduced Off-Policy TDC Learning: Non-Asymptotic Convergence Analysis." Advances in Neural Information Processing Systems 33 (2020).
- [ICLR'21] Ma, Shaocong, et al. "Greedy-GQ with Variance Reduction: Finite-time Analysis and Improved Complexity." International Conference on Learning Representations. 2020.
- [ICML'20] Ma, Shaocong, and Yi Zhou. "Understanding the Impact of Model Incoherence on Convergence of Incremental SGD with Random Reshuffle." International Conference on Machine Learning. PMLR, 2020.