Variance-reduced Q-learning is minimax optimal

Ziniu Li
liziniu1997@gmail.com

(incoming Ph.D. student at SDS)
The Chinese University of Hong Kong, Shenzhen, China

July 9, 2020

Summer Seminar of CUHK, Shenzhen
Mainly based on:
Wainwright, Martin J. "Variance-reduced Q-learning is minimax optimal." arXiv preprint arXiv:1906.04697 (2019).

Target

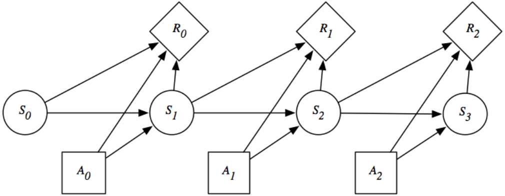
- We will briefly talk about the complexity of sequential decision-making, but mainly focus on the sample complexity under a generative model.
- We will illustrate the famous method called Q-learning and demonstrate the effectiveness of the variance-reduction technique.
- We will briefly explain the proof ideas for Q-learning and variance-reduced Q-learning.

Markov Decision Process

- Consider an infinite-horizon Markov Decision Process $\mathcal{M}^* = (\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$ [3].
  - $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, respectively.
  - P determines the transition probability of s_{t+1} conditioned on s_t and a_t.
  - R is the reward function, which is often assumed to be deterministic and is bounded within the range $[0, 1]$.
  - $\gamma \in [0, 1)$ is a discount factor.
  - d_0 specifies the initial state distribution.

Markov Decision Process

- The decision process is characterized as follows:
  - At the beginning of the epoch, the environment resets to some initial state s_0 according to d_0;
  - The agent observes the state s_0 and select an action a_0 to perform;
  - The environment transits to s_1 according to P and sends a reward signal r_0 to the agent.
  - This process repeats until some terminal signal is released, after which the environment resets to some initial state again.



Markov Decision Process

- The above action selection procedure can be described as a policy, which maps the state space to the action space.
- The goal of an intelligent agent is to maximize its payoff by searching the optimal policy π^* with maximal cumulative rewards.

$$\pi^* = \arg\max_{\pi} \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$$

- Though the above decision-making procedure seems endless, the effective planning horizon is $1/(1-\gamma)$.

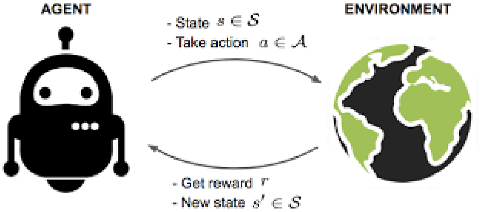
$$\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] \leq \frac{1}{1-\gamma} \cdot r_{\max}$$

Complexity of MDP

- With the knowledge of P and R, we can efficiently solve an (infinite-horizon) MDP with methods like value iteration, policy iteration, and linear programming [3].
- The computation complexity of the above methods mainly depends on $|\mathcal{S}|$ and $|\mathcal{A}|$ and $1/(1-\gamma)$.
  - The above methods often can find an ϵ-optimal solution with the speed of $\mathcal{O}(\gamma^t)$;
  - Thus, the number of iteration to find an ϵ-optimal solution is about $\mathcal{O}(\frac{\log(1/\epsilon)}{1-\gamma})$.
  - At each iteration, the above methods use P to perform the expected Bellman update (define later), and this computation complexity linearly scales up to the whole space size (i.e., $|\mathcal{S}| \times |\mathcal{A}|$).

Reinforcement Learning

- In reinforcement learning (RL), we cannot have access to the transition kernel P but we can interact with environments to collect information. Accordingly, we cannot directly apply the above methods since we cannot perform the expected Bellman update.



- Typically, we need exploration (e.g., take new actions) to discover potential high reward states and exploitation (e.g., take the best known action) to maintain a good performance.

Complexity of RL

- The PAC(provably approximation correct) complexity of RL is (informally) defined as: how many interactions/samples (m) do we need to find an good policy (with the optimality gap ϵ) with high probability (at least $1-\delta$)?
- Unfortunately, it's very challenging to analyze the complexity of RL methods, which does not only depend on $|\mathcal{S}|$, $|\mathcal{A}|$ and $1/(1-\gamma)$, but also the intrinsic difficulty of MDP.
  - For example, solving a motion planning task with many obstacles is much harder than the one with a simple structure even both MDPs have the same state and action spaces.
- Detailed analysis of the complexity of RL is beyond this talk. And we will focus on an intermediate problem defined later.

RL with a Generative Model

- Let us introduce the generative model $\mathcal{M}$. Importantly, we can directly reset it to any state s_t, after which we can take an action a_t and observe the next state $s_{t+1} \sim p_{a_t}(\cdot|s_t)$ and the reward $r(s_t, a_t)$.
  - Compared to the pure MDP problem, we still do not known P in advance.
  - Compared to the pure RL problem, we can go to any s_t without the planning from an initial state s_0.
- Example: a perfect simulator (e.g., some video game simulators), where we can load (reset) the state s_t from RAM.
- Luckily, the complexity of RL with a generative model is shown to only depend on $|\mathcal{S}|$, $|\mathcal{A}|$, and $1/(1-\gamma)$.

Bellman Optimality Equation

- The state-action value function (or Q-function) for an infinite-horizon MDP is defined as:

$$\theta^\pi(x,u) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r(x_k,u_k)|x_0 = x, u_0 = u] \qquad \text{where } u_k = \pi(x_k) \text{ for all } k \geq 1$$

  where we replace the state s_t with x_t and the action a_t with u_t.

- The Bellman Optimality Equation is defined as :

$$\theta^\pi(x,u) = r(x,u) + \mathbb{E}_{x'}[\max_{u' \in \mathcal{U}} \theta^\pi(x',u')] \qquad \text{where } x' \sim P_u(\cdot|x)$$

  where $P_u(\cdot|x)$ denotes the transition kernel based on current state x and current action u.

- Define the optimal state-value function $\theta^* = \max_\pi \theta^\pi$. It can be proved only θ^* is the solution to the above equation [3].

Bellman Operator

- The expected (population) Bellman operator $\mathcal{T}$ is a mapping from $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{U}|}$ to itself:

$$\mathcal{T}(\theta)(x,u) := r(x,u) + \gamma \mathbb{E}_{x'}[\max_{u' \in \mathcal{U}} \theta(x',u')] \qquad \text{where } x' \sim P_u(\cdot|x)$$

- Similarly, we can define the empirical (sampling-based) Bellman operator $\hat{\mathcal{T}}$:

$$\hat{\mathcal{T}}(\theta)(x,u) := r(x,u) + \gamma \max_{u' \in \mathcal{U}} \theta(x',u') \qquad \text{where } x' \sim P_u(\cdot|x)$$

- By construction, we have $\mathbb{E}[\hat{\mathcal{T}}(\theta)] = \mathcal{T}(\theta)$ and $\theta^* = \mathcal{T}(\theta^*)$

Properties of Bellman Operator

- (γ-contractive) For any $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{U}|}$ and define $||\theta||_\infty = \max_{(x,u)} |\theta(x,u)|$, we have

$$||\mathcal{T}(\theta_1) - \mathcal{T}(\theta_2)||_\infty \leq \gamma||\theta_1 - \theta_2||_\infty$$

- (orthant ordering) If $\theta_1 \preceq \theta_2$ (i.e., θ_1 is no larger than θ_2 elementwise), we have

$$\mathcal{T}(\theta_1) \preceq \mathcal{T}(\theta_2)$$

- Note the above properties also hold for $\hat{\mathcal{T}}$ (because $\hat{\mathcal{T}}$ is a special case of $\mathcal{T}$).

Properties of Bellman Operator

- Since $\mathcal{T}$ is γ-contractive, we can repeatedly apply on $\mathcal{T}$ on θ_k to get a contractive sequence $\{\theta_k\}$.
$$\theta_{k+1} := (1-\lambda_k)\theta_k + \lambda_k \mathcal{T}(\theta_k) \tag{1}$$
where $\{\lambda_k : \lambda_k \in (0,1]\}$ is some sequence of stepsize.
- By γ-contractive, we can show that the optimal gap $\Delta_k = \theta_k - \theta^*$ decays with a linear rate (i.e., $\mathcal{O}(\gamma^t)$). Thus $\theta \mapsto \theta^*$ if we know P to perform $\mathcal{T}$.
$$\Delta_{k+1} = (1-\lambda_k)\Delta_k + \lambda_k \left\{\mathcal{T}(\Delta_k + \theta^*) - \mathcal{T}(\theta^*)\right\}$$
$$||\Delta_{k+1}||_\infty \overset{(\lambda_k=1)}{\leq} \gamma||\Delta_k||_\infty \leq \gamma^t||\Delta_1||_\infty$$
- In the next part, we show the generative model only admits $\hat{\mathcal{T}}$, which results in sampling noise when updating.

Q-learning

- The (synchronous) Q-learning takes a stochastic approximation (SA) approach to the Bellman optimality equation with $\hat{\mathcal{T}}$:

$$\theta_{k+1} = (1 - \lambda_k)\theta_k + \lambda_k \hat{\mathcal{T}}_k(\theta_k) \tag{2}$$

- We can rewrite the above update rule as:

$$\theta_{k+1} = (1 - \lambda_k)\theta_k + \lambda_k \{\mathcal{T}(\theta_k) + E_k\}$$

  where $E_k = \hat{\mathcal{T}}(\theta_k) - \mathcal{T}(\theta_k)$ is a zero-mean noise matrix.
- Thus, we can view the above update rule as the expected Bellman update with some noise.

Noise in Q-learning

- Recall the Q-learning update rule (we will introduce θ^* and $\hat{\mathcal{T}}_k(\theta^*)$ to "center"):

$$\theta_{k+1} - \theta^* = (1 - \lambda_k)(\theta_k - \theta^*) + \lambda_k \hat{\mathcal{T}}_k(\theta_k) - \lambda_k \hat{\mathcal{T}}_k(\theta^*) + \lambda_k \hat{\mathcal{T}}_k(\theta^*) - \lambda_k \mathcal{T}(\theta^*)$$

- Similarly, let's consider the update rule from the view of the optimal gap $\Delta_k = \theta_k - \theta^*$:

$$\Delta_{k+1} = (1 - \lambda_k)\Delta_k + \underbrace{\lambda_k \{\hat{\mathcal{T}}_k(\theta^* + \Delta_k) - \hat{\mathcal{T}}_k(\theta^*)\}}_{\gamma\text{-contractive}} + \underbrace{\lambda_k W_k}_{\text{noise}} \quad (3)$$

  Here $W_k = \hat{\mathcal{T}}_k(\theta^*) - \mathcal{T}(\theta^*)$ is a zero-mean random (noise) matrix.
- In this way, Δ_k decays over iteration with the sampling noise.

Q-learning with Oracle Variance Reduction

- Let's consider the following update rule:

$$\theta_{k+1} = (1-\lambda_k)\theta_k + \lambda_k \left(\hat{\mathcal{T}}_k(\theta_k) - \hat{\mathcal{T}}_k(\theta^*) + \mathcal{T}(\theta^*)\right)$$

  Note that $\mathbb{E}[\hat{\mathcal{T}}_k(\theta^*)] = \mathcal{T}(\theta^*)$.

- Again, let's define the error matrix $\Delta_k = \theta_k - \theta^*$, we find that

$$\Delta_{k+1} = (1-\lambda_k)\Delta_k + \lambda_k \left\{\hat{\mathcal{T}}(\theta^* + \Delta_k) - \hat{\mathcal{T}}(\theta^*)\right\}$$

- Compared to the previous one (see Equation (3)), the noise term $W_k = \hat{\mathcal{T}}_k(\theta^*) - \mathcal{T}(\theta^*)$ vanishes.

Variance-reduced Q-learning

- Though the above method is not implementable because of the unknown θ^*, we can use a matrix $\bar{\theta}$ as a surrogate of θ^*.
- Let's consider the following control variate:

$$\tilde{\mathcal{T}}_N(\bar{\theta}) = \frac{1}{N} \sum_{i \in D} \hat{\mathcal{T}}_i(\bar{\theta})$$

  where D is a collection of N i.i.d samples.
- By construction, $\tilde{\mathcal{T}}_N(\bar{\theta})$ is an unbiased approximation to $\mathcal{T}(\bar{\theta})$, with the variance controlled by N.

Variance-reduced Q-learning

- Let's define an operator $\mathcal{V}_k$ on $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{U}|}$ via

$$\mathcal{V}_k(\theta; \lambda, \bar{\theta}, \tilde{\mathcal{T}}_N) = (1-\lambda)\theta + \lambda \left\{ \hat{\mathcal{T}}_k(\theta) - \hat{\mathcal{T}}_k(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta}) \right\}$$

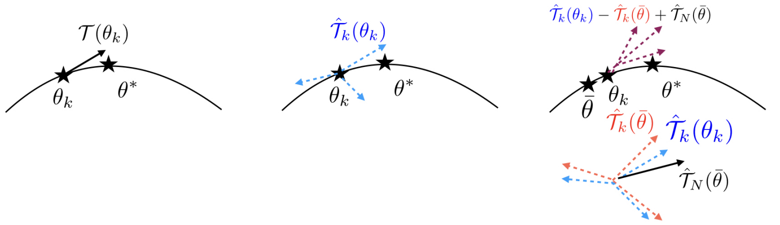
- By construction, we show that $\mathcal{V}_k$ is also unbiased:

$$\mathbb{E}\left[\hat{\mathcal{T}}_k(\theta) - \hat{\mathcal{T}}_k(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta}) \right] = \mathcal{T}(\theta)$$

- This variance-reduced operator is similar to the one used in SVRG [2].

Why variance-reduced?

- Why $\mathcal{V}_k(\theta; \lambda, \bar{\theta}, \tilde{\mathcal{T}}_N) = (1-\lambda)\theta + \lambda \left\{ \hat{\mathcal{T}}_k(\theta) - \hat{\mathcal{T}}_k(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta}) \right\}$ is variance-reduced?
- If $\bar{\theta}$ is close to θ and θ^*, $\hat{\mathcal{T}}_k(\theta)$ has the close direction with $\hat{\mathcal{T}}_k(\bar{\theta})$, and $\tilde{\mathcal{T}}_N(\bar{\theta})$ is very close to $\mathcal{T}(\theta)$ by choosing a large N. In this way, we "recover" the expected Bellman update.



Expected Bellman Update | **Q-learning** | **Variance-Reduced Q-learning**

Why variance-reduced?

- You may want to understand VRQL from the perspective of the optimality gap. If we follow the previous stepups, we have

$$\Delta_{k+1} = (1-\lambda_k)\Delta_k + \lambda_k \left\{\hat{\mathcal{T}}_k(\theta^* + \Delta_k) - \hat{\mathcal{T}}_k(\theta^*)\right\} + W_k$$

  where $W_k = \hat{\mathcal{T}}_k(\theta^*) - \mathcal{T}(\theta^*) - \hat{\mathcal{T}}_k(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta})$.
- However, note that $\mathbb{E}[W_k] \neq 0$ (the expectation is taken over the stochastic process of $\hat{\mathcal{T}}_k$).
- Correspondingly, W_k can not be viewed as a zero-mean noise term. In contrast, we also need to "center" $\hat{\mathcal{T}}_k(\bar{\theta})$ and consider the (shifted) fixed point by $\hat{\mathcal{V}}_k$ (we will formally analyze this later).

Sing Epoch of Variance-Reduced Q-learning

- Sing Epoch of variance-reduced Q-learning (VRQL) is outlined below:

Function `RunEpoch`$(\overline{\theta}; K, N)$
Inputs:
(a) Epoch length K (b) Recentering matrix $\overline{\theta}$ (c) Recentering sample size N

(1) Compute $\widetilde{\mathcal{T}}_N(\overline{\theta}) := \frac{1}{N}\sum_{i=1}^{N}\widehat{\mathcal{T}}_i(\overline{\theta})$.

(2) Initialize $\theta_1 = \overline{\theta}$.

(3) For $k = 1, \ldots, K$, compute the variance-reduced update (11):

$$\theta_{k+1} = \mathcal{V}_k(\theta_k; \lambda_k, \overline{\theta}, \widetilde{\mathcal{T}}_N) \qquad \text{with stepsize } \lambda_k = \frac{1}{1+(1-\gamma)k}. \tag{12}$$

Output: Return θ_{K+1}.

Overall Algorithm

- The overall algorithm runs by repeatedly calling the sub-procedure of RunEpoch.

Algorithm: Variance-reduced Q-learning
Inputs: (a) Number of epochs M (b) Epoch length K (c) Recentering sizes $\{N_m\}_{m=1}^M$

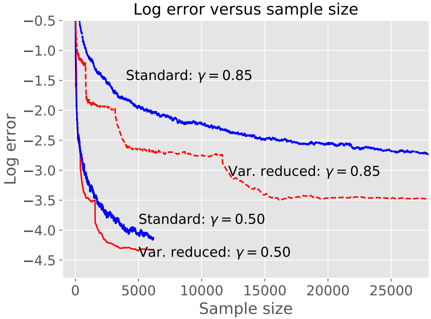
(1) Initialize $\overline{\theta}_0 = 0$.

(2) For epochs $m = 1, \ldots, M$: $\overline{\theta}_m = \texttt{RunEpoch}(\overline{\theta}_{m-1}; K, N_m)$.

- All input parameters: M-number of epochs, K-epoch length, $\{N_m\}_{m=1}^M$-centering sizes and $\{\lambda_k\}_{k=1}^K$-stepsizes.
- The total number of matrix samples required by VRQL is $KM + \sum_{m=1}^M N_m$.

Experimental Comparison

- We can compare VRQL (red line) and ordinary Q-learning (blue line) under two MDPs with different γ (this figure from [7]).



(a)

Parameter Choice

- Given a tolerance parameter $\delta \in (0, 1)$, let's choose the epoch length K and centering sizes $\{N_m\}_{m=1}^{M}$ so as to ensure that the final guarantees hold with probability as least $1 - \delta$.

$$K = c_1 \frac{\log\left(\frac{8MD}{(1-\gamma)\delta}\right)}{(1-\gamma)^3}$$
$$N_m = c_2 4^m \frac{\log(8MD/\delta)}{(1-\gamma)^2} \tag{4}$$

  where $D = |\mathcal{X}| \times |\mathcal{U}|$.

- The number of epoch M depends on the convergence rate and the desired accuracy, which will be decided later.

Linear Convergence Over Epochs

Theorem 1.
Given a γ-discounted MDP with optimal Q-function θ^* and a given error probability $\delta \in (0, 1)$, suppose that we run variance-reduced Q-learning from $\bar{\theta}_0 = 0$ for M epochs using parameters K and $\{N_m\}_{m=1}^{M}$ chosen according to the criteria (4). Then we have

$$||\bar{\theta}_M - \theta^*||_\infty \leq \frac{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty(1-\gamma)}{2^M}$$

with probability at least $1 - \delta$, where $||\sigma(\theta^*)||_\infty = \sqrt{\max_{(x,u)} \mathit{Var}\left(\hat{\mathcal{T}}(\theta^*)(x,u)\right)}$.

Sample Complexity of VRQL

Corollary 1.

Consider a γ-discounted MDP with optimal Q-function θ^*, a given error probability $\delta \in (0, 1)$ and ℓ_∞-error level $\epsilon > 0$. Then there are universal constants c, c' such that a total of

$$T(\theta^*, \delta, \epsilon) = \left\{ c \frac{\log\left(\frac{8MD}{(1-\gamma)\delta}\right)}{(1-\gamma)^3} \log\left(\frac{b_0}{\epsilon}\right) + c' \left(\frac{b_0}{\epsilon}\right)^2 \frac{\log(8MD/\delta)}{(1-\gamma)^2} \right\}$$

matrix samples in the generative model is sufficient to obtain an ϵ-accurate estimate with probability at least $1 - \delta$, where b_0 is defined as

$$b_0 = ||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty (1 - \gamma)$$

Proof of Corollary 1

- We first note that to obtain an ϵ-accurate estimate, the following number of epochs M is enough.

$$M = \left\lceil \log_2 \left(\frac{b_0}{\epsilon} \right) \right\rceil$$

- By construction, the total number of matrix samples of VRQL is $KM + \sum_{m=1}^{M} N_m$. Thus,

$$\begin{aligned} KM + \sum_{m=1}^{M} N_m &\leq MK + c4^M \frac{\log(8MD/\delta)}{(1-\gamma)^2} \\ &\leq c' \frac{\log\left(\frac{8MD}{(1-\gamma)\delta}\right)}{(1-\gamma)^3} \log\left(\frac{b_0}{\epsilon}\right) + c\left(\frac{b_0}{\epsilon}\right)^2 \frac{\log(8MD/\delta)}{(1-\gamma)^2} \end{aligned}$$

Worst Case Analysis

- Assume that reward function is bounded by $r_{\max}$, i.e., $\max_{(x,u)\in\mathcal{X}\times\mathcal{U}} |r(x,u)| \leq r_{\max}$.
- We can give a worst case bound for b_0:

$$\sup_{\mathcal{M}^*} b_0 = \sup_{\mathcal{M}^*} ||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty (1-\gamma) \leq r_{\max}\left(\frac{2}{1-\gamma}+1\right) \leq \frac{4r_{\max}}{1-\gamma}$$

- Applying this bound to Corollary 1, we have

$$\sup_{\mathcal{M}^*} T(\theta^*, \delta, \epsilon) \leq \left\lceil c\left(\frac{r_{\max}^2}{\epsilon^2}\right)\frac{\log\left(\frac{D}{(1-\gamma)\delta}\right)\log\left(\frac{1}{(1-\gamma)\epsilon}\right)}{(1-\gamma)^4}\right\rceil$$

and the total number of epochs required is $M = c\log\left(\frac{r_{\max}}{1-\gamma}\right)$ for some universal constant c.

Refine our analysis

- In the worst case, we require the following matrix samples:

$$\sup_{\mathcal{M}^*} T(\theta^*, \delta, \epsilon) \leq \left\lceil c \left(\frac{r_{\max}^2}{\epsilon^2} \right) \frac{\log\left(\frac{D}{(1-\gamma)\delta}\right) \log\left(\frac{1}{(1-\gamma)\epsilon}\right)}{(1-\gamma)^4} \right\rceil$$

- If we do not start with zero vector (zero vector is the worst one), we can further improve this result by a good initial point such that $\bar{\theta}_0$ with $||\bar{\theta}_0 - \theta^*||_\infty \leq \frac{r_{\max}}{\sqrt{1-\gamma}} \leq \frac{r_{\max}}{1-\gamma}$.

Refined Sample Complexity of VRQL

Proposition 1 (Minimax optimality).

Consider a γ-discounted MDP with optimal Q-function θ^*, a given error probability $\delta \in (0, 1)$, and a given error tolerance. Then running variance-reduced Q-learning from in initial point $\bar{\theta}_0$ such that $||\bar{\theta}_0 - \theta^*||_\infty \leq \frac{r_{\max}}{\sqrt{1-\gamma}}$ for a total of $M = c \log \left(\frac{r_{\max}}{\sqrt{(1-\gamma)\epsilon}} \right)$ epochs using K and $\{N_m\}_{m=1}^M$ chosen according to the criteria (4), yields a solution $\bar{\theta}_M$ such that $||\bar{\theta}_M - \theta^*|| \leq \epsilon$ with probability at least $1 - \delta$. And the total number of matrix samples is bounded by

$$T_{\max}(\theta^*, \delta, \epsilon) = c \left(\frac{r_{\max}^2}{\epsilon^2} \right) \frac{\log \left(\frac{D}{(1-\gamma)\delta} \right) \log \left(\frac{1}{(1-\gamma)\epsilon} \right)}{(1-\gamma)^3}$$

Lower Bound on Generative Model

Definition 1 ((ϵ, δ)-correct algorithm).

Let θ be the output of some RL algorithm $\mathbb{A}$. We say that $\mathbb{A}$ is (ϵ, δ)-correct on the class of MDPs $\mathbb{M} = \{\mathcal{M}_1^*, \mathcal{M}_2^*, \cdots\}$ if $||\theta^* - \theta||_\infty \leq \epsilon$ with probability at least $1 - \delta$ for all $\mathcal{M}^* \in \mathbb{M}$.

Theorem 2 (Lower bound on the sample complexity of RL with a generative model[1]).

There exist some constants $\epsilon_0, \delta_0, c_1, c_2$ and a class of MDPs $\mathbb{M}$ such that for all $\epsilon \in (0, \epsilon_0)$, $\delta \in (0, \delta_0/(|\mathcal{S}| \times |\mathcal{A}|))$, and every (ϵ, δ)-correct RL algorithm on the class of MDPs $\mathbb{M}$ the total number of state-transition samples need to be least

$$T = \left\lceil \frac{|\mathcal{S}| \times |\mathcal{A}|}{c_1 \epsilon^2 (1-\gamma)^3} \log \frac{|\mathcal{S}| \times |\mathcal{A}|}{c_2 \delta} \right\rceil$$

Sample Complexity of Ordinary Q-learning

Theorem 3 (Sublinear Convergence Rate of Q-learning).

Consider the stepsize $\lambda_k = \frac{1}{1+(1-\gamma)k}$. *Then there exist a universal constant* c *such that running the empirical Bellman update (see Equation* (2)*) yields*

$$\mathbb{E}\left[||\theta_{k+1} - \theta^*||\right] \leq \frac{||\theta_1 - \theta^*||_\infty}{1+(1-\gamma)k} + \frac{c}{1-\gamma}\left\{\frac{||\sigma(\theta^*)||_\infty\sqrt{\log(2D)}}{\sqrt{1+(1-\gamma)k}} + \frac{||\theta^*||_{span}\log(2eD(1+(1-\gamma)k))}{1+(1-\gamma)k}\right\}$$

where $||\theta^*||_{span} = \max_{(x.u)}\theta^*(x,u) - \min_{(x,u)}\theta^*(x,u)$, *and* $||\sigma(\theta^*)||_\infty = \sqrt{\max_{(x,u)} \mathit{Var}\left(\hat{\mathcal{T}}(\theta^*)(x,u)\right)}$.

(Remark) A high probability bound can also be derived by replacing $\log(D)$ with $c\log(Dk/\delta)$.

Sample Complexity of Ordinary Q-learning (worst case)

- Let's consider the worst case analysis.

$$\sup_{\mathcal{M}^*} ||\theta^*||_{\text{span}} \leq \frac{2r_{\max}}{1-\gamma}, \qquad \text{and} \ \sup_{\mathcal{M}^*} ||\sigma(\theta^*)||_\infty \leq \frac{r_{\max}}{1-\gamma}$$

- In this way, we claim that ordinary Q-learning requires a total of

$$\sup_{\mathcal{M}^*} T(\epsilon, \gamma, \theta^*) = \mathcal{O}\left(\frac{r_{\max}^2}{(1-\gamma)^5}\right)$$

matrix samples to find an ϵ-optimal solution in expectation.

Discussion

- VRQL ($\mathcal{O}(1/(1-\gamma)^4)$) improves the upper bound compared to ordinary Q-learning ($\mathcal{O}(1/(1-\gamma)^5)$) in the worst case .
- Note that model-free methods (e.g., value iteration and q-learning) with the variance-reduction technique can often get better performance [4].
- To match the lower bound $\mathcal{O}(1/(1-\gamma)^3)$, VRQL requires a good initial point. This is somewhat unsatisfying, because the same kind method of Variance-reduced Value Iteration [4] does not require this to match the lower bound.
- On the other hand, model-based methods do not require variance-reduction to match the lower bound [1].
  - Model-based methods first construct a virtual MDP $\hat{\mathcal{M}}$ with collected samples and then learns a (near-) optimal $\hat{\theta}^*$ on this recovered MDP.

Why variance-reduction is important for model-free methods?

- Intuitively, model-free methods iteratively interact with the environment to collect samples. As a result, we will waste samples if we do not use $\bar{\theta}$, which contains past information.
- Technically, both model-free and model-based approaches use samples to estimate the expected Bellman update.
  - Naive model-free methods require a union bound accuracy for all iterations.
  - Model-based methods only need the estimate is accuracy for the optimal $\hat{\theta}^*$ on recovered MDP.

Proof Idea of Q-learning

- We start with the simplest case: Q-learning, which will be insightful for analysis of VRQL.
- We can rewrite the update rule of Q-learning (ref to Equation (2)) as:

$$\theta_{k+1} - \theta^* = (1 - \lambda_k)(\theta_k - \theta^*) + \lambda_k \left\{ \hat{\mathcal{H}}_k(\theta_k) + W_k \right\}$$

$$\hat{\mathcal{H}}_k(\theta_k) = \hat{\mathcal{T}}_k(\theta_k) - \hat{\mathcal{T}}_k(\theta^*)$$

$$W_k = \hat{\mathcal{T}}_k(\theta^*) - \mathcal{T}(\theta^*)$$

- $\hat{\mathcal{H}}_k(\theta_k)$ is γ-contractive with respective to $||\theta_k - \theta^*||_\infty$.
- W_k is a θ_k-independent noise term, which is governed by the statistical features (e.g., bounded value and variance) of θ^*.

Proof Idea of Q-learning

- Note that W_k incurs a stochastic process, which is independent of θ_k,

$$P_k = (1 - \lambda_{k-1})P_{k-1} + \lambda_{k-1}W_{k-1}, \quad \text{with initialization } P_1 = 0$$

- Thanks to the linearity, by properly choosing two real-value series a_k (related to γ and $||P_k||$) and b_k (related to the initial value $||\theta_1 - \theta^*||_\infty$), we can show that (see [6] for details)

$$||\theta_k - \theta^*||_\infty \leq b_k + a_k + ||P_k||_\infty$$

Proof Idea of Q-learning

- Futhermore, when $\lambda_k = \frac{1}{1+(1-\gamma)k}$, we have (see [6] for details)

$$||\theta_{k+1} - \theta^*||_\infty \leq \lambda_k \left\{ \frac{||\theta_1 - \theta^*||_\infty}{\lambda_1} + \gamma \sum_{\ell=1}^{k} ||P_\ell||_\infty \right\} + ||P_{k+1}||_\ell$$

- Hence, for ordinary Q-learning, we need to bound $||P_k||_\infty$ to estimate the converge rate.

Proof Idea of Q-learning

- Recall that $W_k = \hat{\mathcal{T}}_k(\theta^*) - \mathcal{T}(\theta^*)$ is a zero-mean random matrix with bounded value $2||\theta^*||_\infty$ and the maximal variance $||\sigma(\theta^*)||_\infty^2$.
- Hence, we conclude that W_k satisfies Bernstein condition [5]. Using the inductive reasoning, we can show that $P_k(x, u)$ also satisfies certain Bernstein condition due to the linearity of the following stochastic process.

$$P_k = (1 - \lambda_{k-1})P_{k-1} + \lambda_{k-1}W_{k-1}, \quad \text{with initialization } P_1 = 0$$

- Finally, we can apply a union bound to derive high probability bound for $||P_k||_\infty$.

Proof Idea of VRQL

- The high-level proof procedure of VRQL is similar to the one of ordinary Q-learning.
- The main difference (difficulty) is that the noise term W_k is not a zero-mean random matrix!

$$\begin{aligned} \theta_{k+1} - \theta^* &= (1-\lambda_k)(\theta_k - \theta^*) + \lambda_k \left\{ \hat{\mathcal{H}}_k(\theta_k) + W_k \right\} \\ \hat{\mathcal{H}}_k(\theta_k) &= \hat{\mathcal{T}}_k(\theta_k) - \hat{\mathcal{T}}_k(\theta^*) \\ W_k &= -\hat{\mathcal{H}}_k(\bar{\theta}) - \mathcal{T}(\theta^*) + \tilde{\mathcal{T}}_N(\bar{\theta}) \end{aligned}$$

where $\hat{\mathcal{H}}_k(\bar{\theta}) = \hat{\mathcal{T}}_k(\bar{\theta}) - \hat{\mathcal{T}}_k(\theta^*)$ is a centered operator.

Proof Idea of VRQL

- To use concentration inequalities, we need to separately "center" each term in W_k.

$$
\begin{aligned}
{\color{blue}W_k} &= -\hat{\mathcal{H}}_k(\bar{\theta}) - \mathcal{T}(\theta^*) + \tilde{\mathcal{T}}_N(\bar{\theta}) \\
&= -\hat{\mathcal{H}}_k(\bar{\theta}) + \underbrace{\tilde{\mathcal{T}}_N(\bar{\theta}) - {\color{red}\tilde{\mathcal{T}}_N(\theta^*)}}_{\tilde{\mathcal{H}}_N(\bar{\theta})} + {\color{red}\tilde{\mathcal{T}}_N(\theta^*)} - \mathcal{T}(\theta^*) \\
&= -\hat{\mathcal{H}}_k(\bar{\theta}) + \tilde{\mathcal{H}}_N(\bar{\theta}) + \left\{\tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)\right\}
\end{aligned}
$$

  where we define $\tilde{\mathcal{H}}_N(\bar{\theta}) = \tilde{\mathcal{T}}_N(\bar{\theta}) - \tilde{\mathcal{T}}_N(\theta^*)$ as a centered operator.
- Note that only the first term depends on the iteration k, while the last two terms do not.

Proof Idea of VRQL

- To apply concentration inequalities, we need to introduce the population operator for each uncentered term that appeared in W_k.
- Let's define the population operator $\mathcal{H}(\theta) := \mathcal{T}(\theta) - \mathcal{T}(\theta^*)$, then

$$W_k = \underbrace{\left\{\mathcal{H}(\bar{\theta}) - \hat{\mathcal{H}}_k(\bar{\theta})\right\}}_{W_k'} + \underbrace{\left\{\tilde{\mathcal{H}}_N(\bar{\theta}) - \mathcal{H}(\bar{\theta})\right\}}_{W^o} + \underbrace{\left\{\tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)\right\}}_{W^\dagger}$$

- Again, we observe that only the first term W_k' is important for the induced stochastic process while the last two terms are independent over iteration k.
- Thus, we can similarly apply previous results by replacing W_k with W_k' to get P_k'.

Proof Idea of VRQL

- Now, our target becomes to separately bound $||P'_k||_\infty$ (induced by W'_k), $||W^o||_\infty$ and $||W^\dagger||_\infty$.

$$W_k = \underbrace{\left\{\mathcal{H}(\bar{\theta}) - \hat{\mathcal{H}}_k(\bar{\theta})\right\}}_{W'_k} + \underbrace{\left\{\tilde{\mathcal{H}}_N(\bar{\theta}) - \mathcal{H}(\bar{\theta})\right\}}_{W^o} + \underbrace{\left\{\tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)\right\}}_{W^\dagger}$$

  - Bounding $||P'_k||_\infty$ is also based on inductive reasoning of Bernstein inequalities.
  - Bounding $||W^o||_\infty$ can directly use Hoeffding's inequality.
  - Bounding $||W^\dagger||_\infty$ can smartly use Bernstein inequality since we know the variance.

Proof of Theorem 1

- At a high-level argument, we prove Theorem 1 via an inductive argument.

$$||\bar{\theta}_M - \theta^*||_\infty \leq \frac{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty(1-\gamma)}{2^M}$$

- **(Base case)** Given the initialization $\bar{\theta}_0 = 0$, we prove that $\bar{\theta}_1$ satisfies such a bound with probability at least $1 - \frac{\delta}{M}$.
- **(Inductive step)** In this step, we prove, with probability at least $1 - \frac{\delta}{M}$, $\bar{\theta}_{m+1}$ satisfies such a bound with the assumption that it holds for $\bar{\theta}_m$.
- **(Union bound)** Finally, by taking a union bound over all M epochs of the algorithm we guarantee the bound holds uniformly for all $m = 1, \cdots M$ with probability at least $1 - \delta$.

Proof of Theorem 1 - Base Case

- For the given initialization $\bar{\theta}_0 = 0$, we have $\hat{\mathcal{T}}_k(\bar{\theta}_0) = r$ and $\tilde{\mathcal{T}}_k(\bar{\theta}_0) = r$. Consequently, $\hat{\mathcal{T}}_k(\bar{\theta}_0) - \tilde{\mathcal{T}}_k(\bar{\theta}_0) = 0$, so that the update rule reduces to the case of ordinary Q-learning with stepsize $\lambda_k = \frac{1}{1+(1-\gamma)k}$.
- According to the prior work [6], there is a universal constant $c' > 0$ such that after M iterations, we have

$$||\theta_{K+1} - \theta^*||_\infty \leq \frac{||\theta^*||_\infty}{(1-\gamma)K} + c' \left\{ \frac{||\sigma(\theta^*)||_\infty \sqrt{\log(2DMK/\delta)}}{(1-\gamma)^{3/2}\sqrt{K}} + \frac{||\theta^*||_\infty \log\left(\frac{2eDMK}{\delta}(1+(1-\gamma)K)\right)}{(1-\gamma)^2 K} \right\}$$

- Choosing $K = c\frac{\log\left(\frac{8MD}{\delta(1-\gamma)}\right)}{(1-\gamma)^3}$ for a sufficient large constant c suffices to ensure that

$$||\theta_{K+1} - \theta^*|| \leq \frac{1}{2}\left\{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty(1-\gamma)\right\} \text{ with probability at least } 1 - \frac{\delta}{M}$$

Proof of Theorem 1 - Inductive Step

- For this step, we assume that the input $\bar{\theta}_m$ to epoch m satisfies the bound

$$||\bar{\theta}_m - \theta^*||_\infty \leq \underbrace{\frac{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty(1-\gamma)}{2^m}}_{=:b_m}$$

- Our target is to prove that $||\bar{\theta}_{m+1} - \theta^*||_\infty \leq b_{m+1} = \frac{b_m}{2}$.
- It turns out that if we can prove

$$||\bar{\theta}_{K+1} - \theta^*||_\infty \leq cb_m \left\{ \frac{1}{1+(1-\gamma)K} + \frac{1}{1-\gamma}\sqrt{\frac{\log(8MDK/\delta)}{1+(1-\gamma)K}} + \sqrt{4^m \frac{\log(8MD/\delta)}{(1-\gamma)^2 N_m}} \right\} \tag{5}$$

, K and N_m defined in Equation (4) are sufficient to prove the inductive step.

Proof of Theorem 1 - Inductive Step

- Recall the update rule of VRQL

$$\theta_{k+1} = (1-\lambda)\theta + \lambda_k\{\hat{\mathcal{T}}_k(\theta) - \hat{\mathcal{T}}_k(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta})\}$$

- Let's introduce the auxiliary recentered operators:

$$\hat{\mathcal{H}}_k(\theta) := \hat{\mathcal{T}}_k(\theta) - \hat{\mathcal{T}}_k(\theta^*)$$

- Thus, we can rewrite the VRQL update rule as

$$\begin{aligned}\theta_{k+1} - \theta_* &= (1-\lambda_k)(\theta_k - \theta^*) + \lambda_k\{\underbrace{\hat{\mathcal{T}}_k(\theta_k) - \hat{\mathcal{T}}_k(\theta^*)}_{\hat{\mathcal{H}}_k(\theta_k)} \underbrace{- \hat{\mathcal{T}}_k(\bar{\theta}) + \hat{\mathcal{T}}_k(\theta^*)}_{\hat{\mathcal{H}}_k(\bar{\theta})} + \tilde{\mathcal{T}}_N(\bar{\theta}) - \mathcal{T}(\theta^*)\} \\ &= (1-\lambda_k)(\theta_k - \theta^*) + \lambda_k\{\hat{\mathcal{H}}_k(\theta_k) - \hat{\mathcal{H}}_k(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta}) - \mathcal{T}(\theta^*)\}\end{aligned}$$

Proof of Theorem 1 - Inductive Step

- Continue to the last page, let $W_k = -\hat{\mathcal{H}}_k(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta}) - \mathcal{T}(\theta^*)$, we have

$$\begin{aligned} \theta_{k+1} - \theta_* &= (1-\lambda_k)(\theta_k - \theta^*) + \lambda_k \left\{ \hat{\mathcal{H}}_k(\theta_k) \underbrace{-\hat{\mathcal{H}}_k(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta}) - \mathcal{T}(\theta^*)}_{W_k} \right\} \\ &= (1-\lambda_k)(\theta_k - \theta^*) + \lambda_k \left\{ \hat{\mathcal{H}}_k(\theta_k) + W_k \right\} \end{aligned} \tag{6}$$

- We can view W_k as a random noise sequence, which defines the following auxiliary stochastic progress:

$$P_k := (1-\lambda_{k-1})P_{k-1} + \lambda_{k-1}W_{k-1}, \qquad \text{with initialization } P_1 = 0$$

Proof of Theorem 1 - Inductive Step

- Note that the operator $\hat{\mathcal{H}}_k(\theta) := \hat{\mathcal{T}}_k(\theta) - \hat{\mathcal{T}}_k(\theta^*)$ is monotonic respect to the orthant ordering and γ-contractive with respect to the ℓ_∞-norm.

Corollary 2.

[Adapted from the paper [6]] For all iterations $k = 1, 2, \cdots$, we have

$$||\theta_{k+1} - \theta^*||_\infty \leq \frac{2}{1 + (1-\gamma)k} \left\{ ||\theta_1 - \theta^*||_\infty + \sum_{\ell=1}^{k} ||P_\ell||_\infty \right\} + ||P_{k+1}||_\infty$$

Proof of Theorem 1 - Inductive Step

- In order to derive a concrete result based on Corollary 2, we need to obtain high-probability upper bounds on the terms $||P_\ell||_\infty$.
- Note that P_k relies on the stochastic process induced by W_k:

$$W_k = -\hat{\mathcal{H}}_k(\bar{\theta}) + \underbrace{\tilde{\mathcal{T}}_N(\bar{\theta}) - \tilde{\mathcal{T}}_N(\theta^*)}_{\tilde{\mathcal{H}}_N(\bar{\theta})} + \tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*) = -\hat{\mathcal{H}}_k(\bar{\theta}) + \tilde{\mathcal{H}}_N(\bar{\theta}) + \left\{\tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)\right\}$$

where $\tilde{H}_N(\theta) := \tilde{\mathcal{T}}_N(\theta) - \tilde{\mathcal{T}}_N(\theta^*)$.

- Let's define the population operator $\mathcal{H}(\theta) := \mathcal{T}(\theta) - \mathcal{T}(\theta^*)$ to center, then

$$W_k = \underbrace{\left\{\mathcal{H}(\bar{\theta}) - \hat{\mathcal{H}}_k(\bar{\theta})\right\}}_{W_k'} + \underbrace{\left\{\tilde{\mathcal{H}}_N(\bar{\theta}) - \mathcal{H}(\bar{\theta})\right\}}_{W^o} + \underbrace{\left\{\tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)\right\}}_{W^\dagger}$$

Proof of Theorem 1 - Inductive Step

- Continue to the last page,

$$W_k = \underbrace{\left\{\mathcal{H}(\bar{\theta}) - \hat{\mathcal{H}}_k(\bar{\theta})\right\}}_{W_k'} + \underbrace{\left\{\tilde{\mathcal{H}}_N(\bar{\theta}) - \mathcal{H}(\bar{\theta})\right\}}_{W^o} + \underbrace{\left\{\tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)\right\}}_{W^\dagger}$$

- We note that W^o and $W^\dagger$ are independent of k, thus using inductive reasoning, we can prove that (the original paper states that $P_k \preceq W^o + W^\dagger + P_k'$. However, this inequality is ill-conditioned for the base case $(k = 2)$.)

$$P_k \preceq W^o + W^\dagger + P_k'$$

Proof of Theorem 1 - Inductive Step

- Thus, we can decompose the error bound of $||P_\ell||_\infty$ in Corollary 2 into that (note that $||\theta_1 - \theta^*|| \leq b$)

$$||\theta_{K+1} - \theta^*||_\infty \leq \frac{2b}{1+(1-\gamma)K} + 3\left\{\frac{||W^o||_\infty + ||W^\dagger||_\infty}{1-\gamma}\right\} + \left\{\frac{2\sum_{\ell=1}^{K} ||P'_\ell||_\infty}{1+(1-\gamma)K} + ||P'_{K+1}||_\infty\right\} \tag{7}$$

- In the next, we will bound the noise terms W^o and $W^\dagger$, and the stochastic process $\{P'_k\}_{k\geq 1}$ separately.

Proof of Theorem 1 - Inductive Step: Bounding the recentering terms

Lemma 1 (High probability bounds on recentering terms).

Fix an arbitrary $\delta \in (0,1)$.
(a) If $||\bar{\theta} - \theta^*||_\infty \leq b_m$, *then there is a universal constant* c *such that (Note that the origin paper does not consider the constant* c*, but it should be! And this constant does not change the final result.)*

$$||W^o||_\infty \leq c4b_m\sqrt{\frac{\log(8MD/\delta)}{N}} \qquad \textit{with prob. at least } 1 - \frac{\delta}{3M}$$

(b) There is a universal constant c *such that*

$$||W^\dagger||_\infty \leq c\left\{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty(1-\gamma)\right\}\sqrt{\frac{\log(8MD/\delta)}{N}} \quad \textit{with prob. at least } 1 - \frac{\delta}{3M}$$

Proof of Lemma 1 - Bounding W^o

- Recall the definition of W^o:

$$W^o = \tilde{\mathcal{H}}_N(\bar{\theta}) - \mathcal{H}(\bar{\theta}) = \left\{\tilde{\mathcal{T}}_N(\bar{\theta}) - \tilde{\mathcal{T}}_N(\theta^*)\right\} - \left\{\mathcal{T}(\bar{\theta}) - \mathcal{T}(\theta^*)\right\}$$

- Thus, each entry of W^o is a zero mean, i.i.d. sum of N random variables bounded in absolute value by $2b_m$.
- By Hoeffding's inequality, we have

$$||W^o||_\infty \leq c4b_m\sqrt{\frac{\log(8MD/\delta)}{N}} \qquad \text{with prob. at least } 1 - \frac{\delta}{3M}$$

Proof of Lemma 1 - Bounding $W^\dagger$

- Recall the definition of $W^\dagger$:

$$W^\dagger = \tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)$$

- Note that $W^\dagger$ is a sum of N i.i.d. terms, each of which is bounded in absolute value by $||\theta^*||_\infty$ and has the variance $\sigma^2(\theta^*)$.
- By Bernstein's inequality, there is a universal constant c such that with prob. $1 - \frac{\delta}{3M}$, we have

$$||\tilde{\mathcal{T}}_N(\theta)^* - \mathcal{T}(\theta^*)||_\infty \leq c \left\{ ||\sigma(\theta^*)||_\infty \sqrt{\frac{\log(8MD/\delta)}{N}} + \frac{||\theta^*||_\infty \log(8MD/\delta)}{N} \right\}$$

Proof of Lemma 1 - Bounding $W^\dagger$

- Note that our choice of $N \geq c\frac{4^m \log(8MD/\delta)}{(1-\gamma)^2}$, we further have

$$
\begin{aligned}
||\tilde{\mathcal{T}}_N(\theta)^* - \mathcal{T}(\theta^*)||_\infty &\leq c\left\{||\sigma(\theta^*)||_\infty \sqrt{\frac{\log(8MD/\delta)}{N}} + \frac{||\theta^*||_\infty \log(8MD/\delta)}{N}\right\} \\
&= c\sqrt{\frac{\log(8MD/\delta)}{N}}\left\{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty \sqrt{\frac{\log(8MD/\delta)}{N}}\right\} \\
&\leq c\sqrt{\frac{\log(8MD/\delta)}{N}}\{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty(1-\gamma)\}
\end{aligned}
$$

Proof of Theorem 1 - Inductive Step: Bounding the stochastic process

Lemma 2 (High probability on noise).

There is a universal constant $c > 0$ *such that for any* $\delta \in (0,1)$

$$\left\{ \frac{2\sum_{\ell=1}^{K} ||P'_\ell||_\infty}{1+(1-\gamma)K} + ||P'_{K+1}||_\infty \right\} \leq \frac{cb_m}{1-\gamma}\sqrt{\frac{2\log(8MDK/\delta)}{1+(1-\gamma)K}}$$

with probability as least $1 - \frac{\delta}{3M}$.

Proof of Theorem 1 - Inductive Step

- Applying the bounds of Lemma 1 and 2 into Equation (7): there are universal constant c, c' such that

$$\frac{||\theta_{K+1} - \theta^*||_\infty}{b_m} \leq \frac{2}{1 + (1-\gamma)K} + c' \left\{ 1 + \frac{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty(1-\gamma)}{b_m} \right\} \sqrt{\frac{\log(8MD/\delta)}{(1-\gamma)^2 N}}$$
$$+ \frac{c}{1-\gamma} \sqrt{\frac{\log(8MDK/\delta)}{1 + (1-\gamma)K}}$$

with probability at least $1 - \frac{\delta}{M}$.

Proof of Theorem 1 - Inductive Step

- Recall that $b_m = \frac{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty (1-\gamma)}{2^m}$, we conclude that

$$\left\{1 + \frac{||\sigma(\theta^*)||_\infty + ||\theta^*||_\infty (1-\gamma)}{b_m}\right\} \sqrt{\frac{\log(8MD/\delta)}{(1-\gamma)^2 N}} \leq c'' \sqrt{\frac{4^m \log(8MD/\delta)}{(1-\gamma)^2 N}}$$

- Putting together the pieces, with probability at least $1 - \frac{\delta}{M}$, we have

$$\frac{||\theta_{K+1} - \theta^*||_\infty}{b_m} \leq c \left\{ \frac{1}{1+(1-\gamma)K} + \sqrt{\frac{4^m \log(8MD/\delta)}{(1-\gamma)^2 N}} + \frac{1}{1-\gamma} \sqrt{\frac{\log(8MDK/\delta)}{1+(1-\gamma)K}} \right\}$$

- By our choice of N_m and K, we complete the desired claim in Equation (5).

Proof of Lemma 2

- We prove Lemma 2 by two steps. In the first step, we prove by induction that the MGF of $P'_k(x, u)$ is bounded by

$$\log \mathbb{E}[e^{sP'_k(x,u)}] \leq \frac{b_m^2 s^2 \lambda_{k-1}}{8} \qquad \text{for all } s \in \mathbb{R} \tag{8}$$

- Combining the Chernoff bounding technique and the union bound, we find that there is a universal constant c such that

$$\Pr\left[||P'_\ell||_\infty \geq c b_m \sqrt{\lambda_{k-1}} \sqrt{\log 8KMD/\delta}\right] \leq \frac{\delta}{3KM}$$

Proof of Lemma 2

- Taking a union bound over all K iterations, we find that

$$\frac{2\sum_{\ell=1}^{K}||P'_\ell||_\infty}{1+(1-\gamma)K} + ||P'_{K+1}||_\infty \leq \frac{cb_m}{1+(1-\gamma)K}\sqrt{\log(8KMD/\delta)}\left\{\sum_{\ell=1}^{K}\sqrt{\lambda_{\ell-1}}+\sqrt{\lambda_K}\right\}$$

  with probability at least $1-\frac{\delta}{3M}$.
- From the proof of Corollary 3 in the paper [6], we have

$$\sum_{\ell=1}^{K}\sqrt{\lambda_{\ell-1}}+\sqrt{\lambda_K} \leq c\frac{\sqrt{1+(1-\gamma)k}}{1-\gamma}$$

- Putting together these pieces yields the claim bound Lemma 2.

Proof of Equation (8)

- Recall the stochastic process $\{P'_k\}_{k\geq 1}$ evolves the recursion $P'_{k+1} = (1-\lambda_k)P'_k + \lambda_k W'_k$, where
$$W'_k := \mathcal{H}(\bar{\theta}) - \hat{\mathcal{H}}_k(\bar{\theta}) = \{\mathcal{T}(\theta) - \mathcal{T}(\theta^*)\} - \left\{\hat{\mathcal{T}}_k(\bar{\theta}) - \hat{\mathcal{T}}_k(\theta^*)\right\}$$
- Similarly, we see that each entry of W'_k is a zero-mean random variable with the absolute value by $b_m := ||\bar{\theta} - \theta^*||$.
- Using the Hoeffding inequality, we have that
$$\log \mathbb{E}\left[e^{sW'_k(x,u)}\right] \leq \frac{s^2 b_m^2}{8} \qquad \text{for all } s \in \mathbb{R}$$

Proof of Equation (8) - Base case

- We will use the above bound to prove the following claim (ref to Equation (8)) by induction.

$$\log \mathbb{E}[e^{sP'_k(x,u)}] \leq \frac{b_m^2 s^2 \lambda_{k-1}}{8} \qquad \text{for all } s \in \mathbb{R}$$

- **Base case (k=1)**: The case $k = 1$ is trivial since $P'_1 = 0$ by definition.
- **Base case (k=2)**: When $k = 2$, we have $P'_2 = \lambda_1 W'_1$, and hence

$$\log \mathbb{E}[e^{sP'_2(x,u)}] = \log \mathbb{E}[e^{s\lambda_1 W'_1(x,u)}] \leq \frac{s^2 \lambda_1^2 b_m^2}{8} \leq \frac{s^2 \lambda_1 b_m^2}{8}$$

where the last inequality follows from the fact that $\lambda_k = \frac{1}{1+(1-\gamma)} \leq 1$.

Proof of Equation (8) - Inductive step

- Now we assume that Equation (8) holds for some iteration $k \geq 2$, and we verify that it holds for iteration $k+1$.

$$\log \mathbb{E}[e^{sP'_{k+1}(x,u)}] = \log \mathbb{E}[e^{s(1-\lambda_k)P'_k(x,u)}] + \log \mathbb{E}[e^{s\lambda_k P'_k(x,u)}]$$
$$\leq \frac{s^2(1-\lambda_k)^2\lambda_{k-1}b_m^2}{8} + \frac{s^2(1-\lambda_k)^2 b_m^2}{8}$$

- We can show that (details not given) based on the definition that $\lambda_k = \frac{1}{1+(1-\gamma)k}$

$$(1-\lambda_k)\lambda_{k-1} \leq \lambda_k$$

- Consequently, we can prove that

$$\frac{s^2(1-\lambda_k)^2\lambda_{k-1}b_m^2}{8} + \frac{s^2(1-\lambda_k)^2 b_m^2}{8} \leq \frac{s^2(1-\lambda_k)\lambda_k b_m^2}{8} + \frac{s^2(1-\lambda_k)^2 b_m^2}{8} = \frac{s^2\lambda_k b_m^2}{8}$$

Proof of Proposition 1 - Base case

- Again, at a high level, the proof is based on the stated condition ($||\theta_0 - \theta^*||_\infty \leq \frac{r_{\max}}{\sqrt{1-\gamma}}$) to show that

$$||\bar{\theta}_m - \theta^*||_\infty \leq \frac{1}{2^m}\frac{r_{\max}}{\sqrt{1-\gamma}} \qquad \text{for all } m = 1, \cdots, M \tag{9}$$

- The base case ($k = 0$) holds trivially and we will focus on the inductive step.
- By hypothesis, for $k \geq 1$ we have (with a little abuse of b_m)

$$||\bar{\theta} - \theta^*||_\infty \leq b_m := \frac{1}{2^m}\frac{r_{\max}}{\sqrt{1-\gamma}}$$

Proof of Proposition 1 - Inductive Step

- In this case, our analysis involves two operators

$$\hat{\mathcal{J}}_k(\theta) := \hat{\mathcal{T}}_k(\theta) - \hat{\mathcal{T}}_k(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta}) \text{ and } \mathcal{J}(\theta) := \mathcal{T}(\theta) - \mathcal{T}(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta})$$

- Note that the variance-reduced Q-learning updates can be written as

$$\theta_{k+1} = (1 - \lambda_k)\theta_k + \lambda_k \hat{\mathcal{J}}_k(\theta_k) \tag{10}$$

- Note that $\mathcal{J}$ is γ-contractive, thus it has a unique fixed point, which we denote by $\hat{\theta}$.
- Since $\mathcal{J}(\theta) = \mathbb{E}[\hat{\mathcal{J}}_k(\theta)]$ by construction, it is natural to analyze the convergence of θ_k to $\hat{\theta}$.

$$||\theta_{K+1} - \theta^*||_\infty \leq ||\theta_{K+1} - \hat{\theta}||_\infty + ||\hat{\theta} - \theta^*||_\infty$$

Proof of Proposition 1 - Inductive Step

Lemma 3.

After $K = c_1 \frac{\log\left(\frac{8MD}{(1-\gamma)\delta}\right)}{(1-\gamma)^3}$ *iterations, we are guaranteed that*

$$||\theta_{K+1} - \hat{\theta}||_\infty \leq \frac{b_m}{4} + \frac{1}{4}||\hat{\theta} - \theta^*||_\infty$$

with probability at least $1 - \frac{\delta}{2M}$.

Lemma 4.

Given a sample size $N_m = c_2 4^m \frac{\log(MD/\delta)}{(1-\gamma)^2}$, *we have*

$$||\hat{\theta} - \theta^*||_\infty \leq \frac{b_m}{5}$$

with probability at least $1 - \frac{\delta}{2M}$.

Proof of Proposition 1 - Inductive Step

- Combining Lemma 4 and Lemma 4, we have

$$
\begin{aligned}
||\theta_{K+1} - \theta^*||_\infty &\leq \left\{ \frac{b_m}{4} + \frac{1}{4}||\hat{\theta} - \theta^*||_\infty \right\} + ||\hat{\theta} - \theta^*||_\infty \\
&\leq \frac{b_m}{2}
\end{aligned}
$$

- Thus, we verify the claim of Equation (9). The computation of total samples is similar to what we have done:

$$
KM + \sum_{m=1}^{M} N_m
$$

- For VQRL, we have that the $K = c\log \frac{r_{\max}}{\epsilon\sqrt{1-\gamma}}$. It is clear that the discount complexity is reduced.

Proof of Lemma 3

- We rewrite Equation (9) as subtracting the fixed point of $\hat{\theta}$ of $\mathcal{J}$:

$$\theta_{k+1} - \hat{\theta} = (1 - \lambda_k)(\theta_k - \hat{\theta}) + \lambda_k \left(\hat{\mathcal{J}}_k(\theta_k) - \hat{\mathcal{J}}_k(\hat{\theta}) \right) + \lambda_k \underbrace{\left(\hat{\mathcal{J}}_k(\hat{\theta}) - \mathcal{J}(\hat{\theta}) \right)}_{E_k}$$

- We can similarly to apply Corollary 2 (see also Equation (6)). In this case, the noise term is given by (with a little abuse of notation, we previously use W_k to denote the noise term):

$$E_k := \hat{\mathcal{J}}_k(\hat{\theta}) - \mathcal{J}(\hat{\theta}) = \left\{ \hat{\mathcal{T}}_k(\hat{\theta}) - \hat{\mathcal{T}}_k(\bar{\theta}) \right\} - \left\{ \mathcal{T}_k(\hat{\theta}) - \mathcal{T}_k(\bar{\theta}) \right\}$$

- Consequently, we have $||E_k||_\infty \leq 2||\hat{\theta} - \bar{\theta}||_\infty$.

Proof of Lemma 3

- By applying Corollary 1 from the paper [6], we have

$$||\theta_{K+1} - \hat{\theta}||_\infty \leq \frac{2}{1 + (1-\gamma)K} \left\{ ||\bar{\theta} - \hat{\theta}||_\infty + \sum_{\ell=1}^{K} ||P_\ell||_\infty \right\} + ||P_{K+1}||_\ell$$

  where the auxiliary stochastic process evolves as $P_k = (1 - \lambda_{k-1})P_{k-1} + \lambda_{k-1}E_{k-1}$.

- Following the same line of argument as in the proof of Lemma 2, we find that

$$||\theta_{K+1} - \hat{\theta}||_\infty \leq c \left\{ \frac{||\bar{\theta} - \hat{\theta}||_\infty}{1 + (1-\gamma)K} + \frac{||\bar{\theta} - \hat{\theta}||_\infty}{(1-\gamma)^{3/2}\sqrt{K}} \right\} \sqrt{\log(8MD/\delta)}$$

  with probability at least $1 - \frac{\delta}{2M}$.

Proof of Lemma 3

- With the choice of $K = c_1 \frac{\log\left(\frac{8MD}{(1-\gamma)\delta}\right)}{(1-\gamma)^3}$, we are guaranteed that

$$||\theta_{K+1} - \hat{\theta}||_\infty \leq \frac{1}{4}||\bar{\theta} - \hat{\theta}||_\infty \leq \frac{1}{4}||\bar{\theta} - \theta^*||_\infty + \frac{1}{4}||\hat{\theta} - \theta^*||_\infty$$

Proof of Lemma 4

- Note that $\hat{\theta}$ is the fixed point of the operator $\mathcal{J}(\theta) := \mathcal{T}(\theta) - \mathcal{T}(\bar{\theta}) + \tilde{\mathcal{T}}_N(\bar{\theta})$, and hence can be viewed as a fixed point of the population Bellman operator defined with perturbed reward function $\tilde{r}$ with each entry $\tilde{r}(x,u) = r(x,u) + \left[\tilde{\mathcal{T}}(\bar{\theta}) - \mathcal{T}(\bar{\theta})\right](x,u)$.
- The following lemma guarantees that this perturbation is relatively small.

Lemma 5 (Bounds on perturbed reward).

For any matrix $\bar{\theta}$ such that $||\bar{\theta} - \theta^*||_\infty \leq b_m$, we have

$$|\tilde{r} - r| \preceq c(b_m \mathbf{1} + \sigma(\theta^*))\sqrt{\frac{\log(8MD/\delta)}{N}} + c'||\theta^*||_\infty \frac{\log(8MD/\delta)}{N}\mathbf{1}$$

with probability at least $1 - \frac{\delta}{8M}$, where $\mathbf{1}$ denotes the unit vector.

Proof of Lemma 4

- We still need a lemma that provides elementwise upper bounds on the absolute difference $|\theta^* - \hat{\theta}|$ in terms of the absolute difference $|\tilde{r} - r|$.
- Let's define $\mathbb{P}^{\pi^*}$ as the linear operator defined by the policy π^* that is optimal with respect to θ^*, and similarly let $P^{\hat{\pi}}$ be the linear operator defined by the policy $\hat{\pi}$ that is optimal with respect to $\hat{\theta}$.

Lemma 6 (Elementwise bounds).

We have the elementwise upper bound:

$$|\theta^* - \hat{\theta}| \preceq \max\left\{(\mathbb{I} - \gamma\mathbb{P}^{\pi^*})^{-1}|\tilde{r} - r|, (\mathbb{I} - \gamma\mathbb{P}^{\hat{\pi}})^{-1}|\tilde{r} - r|\right\}$$

Proof of Lemma 4 - Upper bounding $(\mathbb{I} - \gamma \mathbb{P}^{\pi^*})^{-1}|\tilde{r} - r|$

- Based on Lemma 5, we have

$$(\mathbb{I} - \gamma \mathbb{P}^{\pi^*})^{-1}|\tilde{r} - r| \preceq c \left(\frac{b_m}{1-\gamma} + ||\mathbb{I} - \gamma \mathbb{P}^{\pi^*})^{-1}\sigma(\theta^*)||_\infty \right) \sqrt{\frac{\log(8MD/\delta)}{N}} \mathbf{1}$$
$$+ c' \frac{||\theta^*||_\infty}{1-\gamma} \frac{\log(8MD/\delta)}{N} \mathbf{1}$$

where we have used the fact that $||(\mathbb{I} - \gamma \mathbb{P}^{\pi^*})^{-1} u||_\infty \leq \frac{||u||_\infty}{1-\gamma}$ for any vector u.

- According to Lemma 8 in [1], we have

$$||(\mathbb{I} - \gamma \mathbb{P}^{\pi^*})^{-1}\sigma(\theta^*)||_\infty \leq \frac{4}{(1-\gamma)^{3/2}} \leq \frac{4(2^m)}{1-\gamma} b_m$$

where the last step follows our notation that $b_m = \frac{1}{2^m} \frac{1}{\sqrt{1-\gamma}}$.

Proof of Lemma 4 - Upper bounding $(\mathbb{I} - \gamma \mathbb{P}^{\pi^*})^{-1}|\tilde{r} - r|$

- Similarly, we also have that

$$\frac{||\theta^*||_\infty}{1-\gamma} \leq \frac{1}{(1-\gamma)^2} \leq \frac{2^m b_m}{(1-\gamma)^{3/2}}$$

- Putting together pieces yields the elementwise bound

$$(\mathbb{I} - \gamma \mathbb{P}^{\pi^*})^{-1}|\tilde{r} - r| \preceq b_m \Phi(N, m, \gamma)\mathbf{1}$$

where we define the non-negative scalar

$$\Phi(N, m, \gamma) := c' \left\{ \frac{2^m}{1-\gamma}\sqrt{\frac{\log(8MD/\delta)}{N}} + \frac{2^m}{(1-\gamma)^{3/2}} \frac{\log(8MD/\delta)}{N} \right\}$$

Proof of Lemma 4 - Upper bounding $(\mathbb{I} - \gamma \mathbb{P}^{\hat{\pi}})^{-1} |\tilde{r} - r|$

- The only difference with the previous derivation is the term regarding $\sigma(\theta^*)$.
- Again, according to [1] we are guaranteed that

$$||\mathbb{I} - \gamma \mathbb{P}^{\hat{\pi}})^{-1} \sigma(\hat{\theta})||_\infty \leq \frac{4}{(1-\gamma)^{3/2}}.$$

- Moreover, we have $\sigma(\theta^*) \preceq \sigma(\hat{\theta}) + |\hat{\theta} - \theta^*|$.
- Combining the pieces, we are guaranteed to have the elementwise bound

$$(\mathbb{I} - \gamma \mathbb{P}^{\hat{\pi}})^{-1} |\tilde{r} - r| \preceq b_m \Phi(N, m, \gamma) \mathbf{1} + c \frac{|\hat{\theta} - \theta^*|}{1-\gamma} \sqrt{\frac{\log(8MD/\delta)}{N}}$$

Proof of Lemma 4 - Upper bounding $(\mathbb{I} - \gamma \mathbb{P}^{\hat{\pi}})^{-1}|\tilde{r} - r|$

- Combining the previous bounds with Lemma 6, we find

$$|\hat{\theta} - \theta^*| \preceq b_m \Phi(N, m, \gamma)\mathbf{1} + c\frac{|\hat{\theta} - \theta^*|}{1-\gamma}\sqrt{\frac{\log(8MD/\delta)}{N}}$$

- Our choice of N ensures that $\frac{c}{1-\gamma}\sqrt{\frac{\log(8MD/\delta)}{N}} \leq \frac{1}{2}$, so that we have established the upper bound $||\hat{\theta} - \theta^*||_\infty \leq 2b_m\Phi(N, m, \gamma)$.
- Finally, we see that our choice of N ensures that $||\Phi(N, m, \gamma)||_\infty \leq \frac{1}{10}$, so that we complete the proof of Lemma 6.

Proof of Lemma 5

- Starting with the definition of $\tilde{r}$ we have

$$
\begin{aligned}
|\tilde{r} - r| &= \left|\tilde{\mathcal{T}}_N(\bar{\theta}) - \mathcal{T}(\bar{\theta})\right| \\
&\leq \left|\left(\tilde{\mathcal{T}}_N(\bar{\theta}) - \tilde{\mathcal{T}}_N(\theta^*)\right) - \left(\mathcal{T}(\bar{\theta}) - \mathcal{T}(\theta^*)\right)\right| + \left|\tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)\right|
\end{aligned}
$$

- By definition, the random matrix $\left(\tilde{\mathcal{T}}_N(\bar{\theta}) - \tilde{\mathcal{T}}_N(\theta^*)\right)$ is the sum of N i.i.d terms, with each entry are uniformly bounded by $\gamma||\bar{\theta} - \theta^*||_\infty \leq b_m$. Consequently, with a combination of Hoeffding's inequality and the union bound, we find that

$$
\left\|\left(\tilde{\mathcal{T}}_N(\bar{\theta}) - \tilde{\mathcal{T}}_N(\theta^*)\right) - \left(\mathcal{T}(\bar{\theta}) - \mathcal{T}(\theta^*)\right)\right\|_\infty \leq 4b_m\sqrt{\frac{\log(8MD/\delta)}{N}}
$$

with probability at least $1 - \frac{\delta}{4M}$.

Proof of Lemma 5

- Turning to the term $|\tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)|$, by a Bernstein inequality, we have

$$|\tilde{\mathcal{T}}_N(\theta^*) - \mathcal{T}(\theta^*)| \leq c \left\{ \sigma(\theta^*) \sqrt{\frac{\log(8MD/\delta)}{N}} + ||\theta^*||_\infty \frac{\log(8MD/\delta)}{N} \right\}$$

- Combing the pieces yields the claim in Lemma 5.

Proof of Lemma 6

- In this proof, we make use of the function $|u|_+ = \max\{u, 0\}$, applied elementwise to a vector u.
- Note that we have $|u| = \max\{|u|_+, |-u|_+\}$ by definition, thus it suffices to prove that two elementwise bounds:

$$|\theta^* - \hat{\theta}|_+ \preceq (\mathbb{I} - \gamma \mathbb{P}^{\pi^*})^{-1} |\tilde{r} - r| \qquad \text{and } |\theta^* - \hat{\theta}|_+ \preceq (\mathbb{I} - \gamma \mathbb{P}^{\hat{\pi}})^{-1} |\tilde{r} - r|$$

- Recall that θ^* and $\hat{\theta}$ are the optimal Q-functions for the reward functions r and $\tilde{r}$, respectively. By this optimality, we have

$$\hat{\theta} = \tilde{r} + \gamma \mathbb{P}^{\hat{\pi}} \hat{\theta} \succeq \tilde{r} + \gamma \mathbb{P}^{\pi^*} \hat{\theta} \qquad \text{and } \theta^* = r + \gamma \mathbb{P}^{\pi^*} \theta^* \succeq r + \gamma \mathbb{P}^{\hat{\pi}} \theta^*$$

Proof of Lemma 6 - The first term

- Using these relations, we can rewrite that

$$\theta^* - \hat{\theta} = (r - \tilde{r}) + \gamma \mathbb{P}^{\pi^*}\theta^* - \mathbb{P}^{\hat{\pi}}\hat{\theta} \leq |\tilde{r} - r| + \gamma \mathbb{P}^{\pi^*}(\theta^* - \hat{\theta})$$
$$\leq |\tilde{r} - r| + \gamma \mathbb{P}^{\pi^*}|\theta^* - \hat{\theta}|_+$$

- Since the RHS is non-negative, the above inequality implies that

$$|\theta^* - \hat{\theta}|_+ \leq |\tilde{r} - r| + \gamma \mathbb{P}^{\pi^*}|\theta^* - \hat{\theta}|_+$$

- Rearranging, we have that

$$|\theta^* - \hat{\theta}|_+ \preceq (\mathbb{I} - \gamma \mathbb{P}^{\pi^*})^{-1}|\tilde{r} - r|$$

Proof of Lemma 6 - The second term

- Using the same reasoning, we have that

$$\begin{aligned}\hat{\theta} - \theta^* &= (r - \tilde{r}) + \gamma \mathbb{P}^{\hat{\pi}} \hat{\theta} - \gamma \mathbb{P}^{\pi^*} \theta^* \\ &\preceq |\tilde{r} - r| + \gamma \mathbb{P}^{\hat{\pi}} (\hat{\theta} - \theta^*) \\ &\preceq |\tilde{r} - r| + \gamma \mathbb{P}^{\hat{\pi}} |\hat{\theta} - \theta^*|_+\end{aligned}$$

- Therefore, we can prove that

$$|\hat{\theta} - \theta^*|_+ \preceq (\mathbb{I} - \gamma \mathbb{P}^{\hat{\pi}})^{-1} |\tilde{r} - r|$$

References I

[1] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. Machine Learning, 91(3):325–349, 2013.

[2] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems, pages 315–323, 2013.

[3] Martin L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley Series in Probability and Statistics. Wiley, 1994.

[4] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms, pages 770–787, 2018.

References II

[5] Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. Cambridge University Press, 2019.

[6] Martin J. Wainwright. Stochastic approximation with cone-contractive operators: sharp bounds for q-learning. arXiv, 1905.06265, 2019.

[7] Martin J. Wainwright. Variance-reduced q-learning is minimax optimal. arXiv, 1906.04697, 2019.

Acknowledgement

The presenter appreciates insightful comments and instructions from Yingru Li, Hao Liang, and Tian Xu.