A Finite Sample Complexity Bound for Distributionally Robust Q-learning

Shengbo Wang
Stanford University

Nian Si
The University of Chicago

Jose Blanchet
Stanford University

Zhengyuan Zhou
New York University

Abstract

We consider a reinforcement learning setting in which the deployment environment is different from the training environment. Applying a robust Markov decision processes formulation, we extend the distributionally robust Q-learning framework studied in Liu et al. (2022). Further, we improve the design and analysis of their multi-level Monte Carlo estimator. Assuming access to a simulator, we prove that the worst-case expected sample complexity of our algorithm to learn the optimal robust Q-function within an ϵ error in the sup norm is upper bounded by $\tilde{O}(|S||A|(1-\gamma)^{-5}\epsilon^{-2}p_{\wedge}^{-6}\delta^{-4})$, where γ is the discount rate, $p_{\wedge}$ is the non-zero minimal support probability of the transition kernels and δ is the uncertainty size. This is the first sample complexity result for the model-free robust RL problem. Simulation studies further validate our theoretical results.

1 Introduction

Reinforcement learning (RL) (Powell, 2007; Bertsekas, 2011; Szepesvári, 2010; Sutton and Barto, 2018) has witnessed impressive empirical success in simulated environments, with applications spanning domains such as robotics (Kober et al., 2013; Gu et al., 2017), computer vision (Sadeghi and Levine, 2016; Huang et al., 2017), finance (Li et al., 2009; Choi et al., 2009; Deng et al., 2017) and achieving superhuman performance in well-known games such as Go and poker (Silver et al., 2016, 2018).

However, existing RL algorithms often make the implicit assumption that the training environment (i.e. a simulator) is the same as the deploying environment, thereby rendering the learned policy *fragile*. This fragility presents a significant impediment for carrying the remarkable success of RL into real environments, because in practice, such discrepancy between training and deploying environments is ubiquitous. On the one hand, simulator models often cannot capture the full complexity of the real environment, and hence will be mis-specified. On the other hand, even if a policy is trained directly in a real environment, the new deployment environment may not be the same and hence suffer from distributional shifts.

As an example of the latter, personalized promotions engine (learned from existing user browsing data collected in one region or market) may need to be deployed in a different region when the company intends to enter a new market. The new market may have similar but different population characteristics. Another example occurs in robotics, where, as articulated in Zhou et al. (2021) "a robot trained to perform certain maneuvers (such as walking Schulman et al. (2013) or folding laundry (Maitin-Shepard et al., 2010)) in an environment can fail catastrophically (Drew, 2015) in a slightly different environment, where the terrain landscape (in walking) is slightly altered or the laundry object (in laundry folding) is positioned differently".

Motivated by the necessity of policy robustness in RL applications, Zhou et al. (2021) adapted the distributionally robust (DR) Markov decision processes (MDPs) to a tabular RL setting and proposed a DR-RL paradigm. Subsequent works Yang et al. (2021) have improved on the sample complexity bounds, although the optimal bound is still unknown as of this writing. However, all these works all adopt a model-based approach which, as widely known, is computationally intensive, requires extensive memory storage, and does not generalize to function approximation settings. Motivated by this concern, the very recent work Liu et al. (2022) introduced the first distributionally robust Q-learning for robust MDPs, thus showing that Q-learning can indeed be made distributionally robust. However, an important issue is that the expected number of samples needed to run the algorithm in Liu et al. (2022) to converge to a fixed error distributionally robust optimal policy is infinite. As such, this naturally motivates the following

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

question:

Can we design a distributionally robust Q-learning that has finite sample complexity guarantee?

1.1 Our Contributions

In this paper, we extend the MLMC-based distributionally robust Bellman estimator in Liu et al. (2022) such that the expected sample size of constructing our estimator is of *constant order*. We establish unbiasedness and moment bounds for our estimator in Propositions 4.2 and 4.3 that are essential to the complexity analysis. Hinging on these properties, we prove that the expected sample complexity of our algorithm is $\tilde{O}\left(|S||A|(1-\gamma)^{-5}\epsilon^{-2}p_{\wedge}^{-6}\delta^{-4}\right)$ under rescaled-linear or constant stepsizes, where $|S|$ and $|A|$ are the number of states and actions, $\gamma \in (0,1)$ the discount factor, ϵ the target error in the infinity norm of the DR Q-function, $p_{\wedge}$ the minimal support probability, and δ the size of the (see Theorems 4.4 and 4.5). Our result is based on the finite sample analysis of stochastic approximations (SA) framework recently established by Chen et al. (2020). To our knowledge, this is the first model-free algorithm and analysis that guarantee solving the DR-RL problem with a finite expected sample complexity. Further, our complexity is tight in $|S||A|$ and nearly tight in the effective horizon $(1-\gamma)^{-1}$ at the same time. Finally, we numerically exhibit the validity of our theorem predictions and demonstrate the improvements of our algorithm over that in Liu et al. (2022).

1.2 Related Work

Distributionally robust optimization (DRO) is well-studied in the supervised learning setting; see, e.g., Bertsimas and Sim (2004); Delage and Ye (2010); Hu and Hong (2013a); Shafieezadeh-Abadeh et al. (2015); Bayraksan and Love (2015); Gao and Kleywegt (2016); Namkoong and Duchi (2016); Duchi et al. (2016); Staib and Jegelka (2017); Shapiro (2017); Lam and Zhou (2017); Volpi et al. (2018); Lee and Raginsky (2018); Nguyen et al. (2018); Yang (2020); Mohajerin Esfahani and Kuhn (2018); Zhao and Jiang (2017); Abadeh et al. (2018); Zhao and Guan (2018); Sinha et al. (2018); Gao et al. (2018); Chen et al. (2018); Ghosh and Lam (2019); Blanchet and Murthy (2019); Duchi and Namkoong (2018); Lam (2019); Duchi et al. (2019); Ho-Nguyen et al. (2020). Those works focus on the optimization formulation, algorithms, and statistical properties in settings where labeled data and a pre-specified loss are available. In those settings, vanilla empirical risk minimizers are outperformed by distributionally robust solutions because of either overfitting or distributional shifts.

In recent years, distributionally robust formulations also find applications in a wide range of research areas including dimensionality reduction under fairness Vu et al. (2022) and model selection Cisneros-Velarde et al. (2020).

Minimax sample complexities of standard tabular RL have been studied extensively in recent years. Azar et al. (2013); Sidford et al. (2018); Agarwal et al. (2020); Li et al. (2020) proposed algorithms and proved optimal upper bounds (the matching lower bound is proved in Azar et al. (2013)) $\tilde{\Theta}(|S||A|(1-\gamma)^{-3}\epsilon^{-2})$ of the sample complexity to achieve ϵ error in the model-based tabular RL setting. The complexity of model-free Q-learning has also been studied extensively (Even-Dar et al., 2003; Wainwright, 2019a; Li et al., 2021). It has been shown by Li et al. (2021) to have a minimax sample complexity $\tilde{\Theta}(|S||A|(1-\gamma)^{-4}\epsilon^{-2})$. Nevertheless, variance-reduced variants of the Q-learning, e.g., Wainwright (2019b), achieves the aforementioned model-based sample complexity lower bound $\tilde{\Theta}(|S||A|(1-\gamma)^{-3}\epsilon^{-2})$.

Recent advances in sample complexity theory of Q-learning and its variants are propelled by the breakthroughs in finite time analysis of SA. Wainwright (2019a) proved a sample path bound for the SA recursion. This enables variance reduction techniques that help to achieve optimal learning rate in Wainwright (2019b). In comparison, Chen et al. (2020) established finite sample guarantees of SA only under a second moment bound on the martingale difference noise sequence.

Our work uses the theoretical framework of the classical minimax control and robust MDPs; see, e.g., González-Trejo et al. (2002); Iyengar (2005); Wiesemann et al. (2013); Xu and Mannor (2010); Shapiro (2022), where those works establish the concept of distributional robustness in MDPs and derive the distributionally robust Bellman equation.

Recently, learning distributionally robust policies from data gains attention (Si et al., 2020; Zhou et al., 2021; Liu et al., 2022; Yang et al., 2021). Among those works, Si et al. (2020) studies the contextual bandit setting, while Zhou et al. (2021); Panaganti and Kalathil (2021); Liu et al. (2022); Yang et al. (2021) focus on the tabular RL regime.

2 Distributionally Robust Policy Learning Paradigm

2.1 Standard Policy Learning

Let $\mathcal{M}_0 = (S, A, R, \mathcal{P}_0, \mathcal{R}_0, \gamma)$ be an MDP, where S, A, and $R \subsetneq \mathbb{R}_{\geq 0}$ are finite state, action, and reward spaces respectively. $\mathcal{P}_0 = \{p_{s,a}, s \in S, a \in A\}$ and $\mathcal{R}_0 = \{\nu_{s,a}, s \in S, a \in A\}$ are the sets of the reward and transition distributions. $\gamma \in (0,1)$ is the discount factor. Define $r_{\max} = \max\{r \in R\}$ the maximum reward. We assume that the transition is Markovian, i.e., at each state $s \in S$, if action $a \in A$ is chosen, then the subsequent state is determined by the conditional distribution $p_{s,a}(\cdot) = p(\cdot|s,a)$. The decision maker will therefore receive a randomized reward $r \sim \nu_{s,a}$. Let Π be the history-dependent policy class (see Section 1 of the supplemental materials for a rigorous construction). For $\pi \in \Pi$, the value function $V^\pi(s)$ is defined as:

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s\right].$$

The optimal value function $V^*(s) := \max_{\pi\in\Pi} V^\pi(s)$, $\forall s \in S$. It is well known that the optimal value function is the unique solution of the following Bellman equation:

$$V^*(s) = \max_{a\in A}\left\{\mathbb{E}_{r\sim\nu_{s,a}}[r] + \gamma \mathbb{E}_{s'\sim p_{s,a}}\left[V^*(s')\right]\right\}.$$

An important implication of the Bellman equation is that it suffices to optimize within the stationary Markovian deterministic policy class.

The optimal Q-function and its Bellman equation:

$$\begin{aligned}Q^*(s,a) &:= \mathbb{E}_{r\sim\nu_{s,a}}[r] + \gamma \mathbb{E}_{s'\sim p_{s,a}}\left[V^*(s')\right]\\ &= \mathbb{E}_{r\sim\nu_{s,a}}[r] + \gamma \mathbb{E}_{s'\sim p_{s,a}}\left[\max_{b\in A} Q^*(s',b)\right].\end{aligned}$$

The optimal policy $\pi^*(s) = \arg\max_{a\in A} Q^*(s,a)$. Therefore, policy learning in RL environments can be achieved if we can learn a good estimate of Q^*.

2.2 Distributionally Robust Formulation

We consider a DR-RL setting, where both transition probabilities and rewards are perturbed based on the KL divergence $D_{\mathrm{KL}}(P\|Q) := \int_\Omega \log \frac{dP}{dQ} P(d\omega)$ when $P \ll Q$ (P is absolutely continuous w.r.t. Q). For each $(s,a) \in S \times A$, we define KL uncertainty set that are centered at $p_{s,a} \in \mathcal{P}_0$ and $\nu_{s,a} \in \mathcal{R}_0$ by $\mathcal{P}_{s,a}(\delta) := \{p : D_{\mathrm{KL}}(p\|p_{s,a}) \le \delta\}$ and $\mathcal{R}_{s,a}(\delta) := \{\nu : D_{\mathrm{KL}}(\nu\|\nu_{s,a}) \le \delta\}$. The parameter $\delta > 0$ controls the size of the uncertainty sets. These uncertainty sets quantify the possible distributional shifts from the reference model $\mathcal{P}_0, \mathcal{R}_0$.

Definition 1. The DR Bellman operator $\mathcal{B}_\delta$ for the value function is defined as the mapping

$$\mathcal{B}_\delta(v)(s) := \max_{a\in A} \inf_{\substack{p\in\mathcal{P}_{s,a}(\delta),\\ \nu\in\mathcal{R}_{s,a}(\delta)}} \left\{\mathbb{E}_{r\sim\nu}[r] + \gamma \mathbb{E}_{s'\sim p}\left[v(s')\right]\right\}. \tag{1}$$

Define the optimal DR value function V_δ^* as the solution of the DR Bellman equation:

$$V_\delta^* = \mathcal{B}_\delta(V_\delta^*) \tag{2}$$

Remark. The definition assumes the existence and uniqueness of a fixed point of the DR Bellman equation. This is a consequence of $\mathcal{B}_\delta$ being a contraction. Moreover, it turns out that under the notion of *rectangularity* (Iyengar, 2005; Wiesemann et al., 2013), this definition is equivalent to the minimax control optimal value

$$V_\delta^*(s) = \sup_{\pi\in\Pi} \inf_{P\in\mathcal{K}^\pi(\delta)} \mathbb{E}_P\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s\right]$$

for some appropriately defined *history-dependent* policy class Π and π-consistent uncertainty set of probability measures $\mathcal{K}^\pi(\delta)$ on the sample path space; cf. Iyengar (2005). Intuitively, this is the optimal value when the controller chooses a policy π, an adversary observes this policy and chooses a possibly history-dependent sequence of reward and transition measure within some uncertainty set indexed by a parameter δ that is consistent with this policy. Therefore, we can interpret $\delta > 0$ as the power of this adversary. The equivalence of minimax control optimal value and Definition 1 suggests the optimality of stationary deterministic Markov control policy and, under such policy, stationary Markovian adversarial distribution choice. We will rigorously discuss this equivalence in Section 1 of the supplemental materials.

2.3 Strong Duality

The r.h.s. of (1) could be hard to work with because the measure underlying the expectations are not directly accessible. To resolve this, we use strong duality:

Lemma 2.1 (Hu and Hong (2013a), Theorem 1)**.** *Suppose $H(X)$ has finite moment generating function in the neighborhood of zero. Then for any $\delta > 0$,*

$$\sup_{P: D_{\mathrm{KL}}(P\|P_0)\le\delta} \mathbb{E}_P\left[H(X)\right] = \inf_{\alpha\ge 0}\left\{\alpha \log \mathbb{E}_{P_0}\left[e^{H(X)/\alpha}\right] + \alpha\delta\right\}.$$

Boundedness of Q allow us to directly apply Lemma 2.1 to the r.h.s. of (2). The DR value function V_δ^* in fact satisfies the following *dual form* of the DR Bellman's equation.

$$\begin{aligned}&V_\delta^*(s) = \\ &\max_{a\in A}\left\{\sup_{\alpha\ge 0}\left\{-\alpha \log \mathbb{E}_{r\sim\nu_{s,a}}\left[e^{-r/\alpha}\right] - \alpha\delta\right\} + \right.\\ &\left.\gamma \sup_{\beta\ge 0}\left\{-\beta \log \mathbb{E}_{s'\sim p_{s,a}}\left[e^{-V_\delta^*(s')/\beta}\right] - \beta\delta\right\}\right\}.\end{aligned} \tag{3}$$

2.4 Distributionally Robust Q-function and its Bellman Equation

As in the classical policy learning paradigm, we make use of the optimal DR state-action value function, a.k.a. Q-function, for solving the DR control problem. The Q-function maps (s,a) pairs to the reals, thence can be identified with $Q \in \mathbb{R}^{S\times A}$. We will henceforth assume this identification. Let us define the notation $v(Q)(s) =$

$\max_{b\in A} Q(s,b)$. We proceed to rigorously define the optimal Q-function and its Bellman equation.

Definition 2. The optimal DR Q-function is defined as

$$Q^*_\delta(s,a) := \inf_{\substack{p\in\mathcal{P}_{s,a}(\delta),\\ \nu\in\mathcal{R}_{s,a}(\delta)}} \left\{\mathbb{E}_{r\sim\nu}[r] + \gamma\mathbb{E}_{s'\sim p}\left[V^*_\delta(s')\right]\right\} \quad (4)$$

where V^*_δ is the DR optimal value function in Definition 1.

By analogy with the Bellman operator, we can define the DR Bellman operator for the Q-function as follows:

Definition 3. Given $\delta > 0$ and $Q \in \mathbb{R}^{S\times A}$, the *primal form* of the DR Bellman operator $\mathcal{T}_\delta : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$ is defined as

$$\begin{aligned}&\mathcal{T}_\delta(Q)(s,a) := \\ &\inf_{\substack{p\in\mathcal{P}_{s,a}(\delta),\\ \nu\in\mathcal{R}_{s,a}(\delta)}} \left\{\mathbb{E}_{r\sim\nu}[r] + \gamma\mathbb{E}_{s'\sim p}\left[v(Q)(s')\right]\right\}\end{aligned} \quad (5)$$

The *dual form* of the DR Bellman operator is defined as

$$\begin{aligned}&\mathcal{T}_\delta(Q)(s,a) := \\ &\sup_{\alpha\geq 0}\left\{-\alpha\log\mathbb{E}_{r\sim\nu_{s,a}}\left[e^{-r/\alpha}\right] - \alpha\delta\right\} + \\ &\gamma\sup_{\beta\geq 0}\left\{-\beta\log\mathbb{E}_{s'\sim p_{s,a}}\left[e^{-v(Q)(s')/\beta}\right] - \beta\delta\right\}.\end{aligned} \quad (6)$$

The equivalence of the primal and dual form follows from Lemma 2.1. Note that by definition (4) and the Bellman equation (2), we have $v(Q^*_\delta) = V^*_\delta$. So, our definition implies that Q^*_δ is a fixed point of $\mathcal{T}_\delta$ and the Bellman equation $Q^*_\delta = \mathcal{T}_\delta(Q^*_\delta)$.

The optimal DR policy can be extracted from the optimal Q-function by $\pi^*_\delta(s) = \arg\max_{a\in A} Q^*_\delta(s,a)$. Hence the goal the DR-RL paradigm is to learn the δ-DR Q-function and extract the corresponding robust policy.

3 Q-Learning in Distributionally Robust RL

3.1 A Review of Synchronized Q-Learning and Stochastic Approximations

The synchronized Q-learning estimates the optimal Q-function using point samples. The classical synchronous Q-learning proceeds as follows. At iteration $k \in \mathbb{Z}_{\geq 0}$ and each $(s,a) \in S\times A$, we draw samples $r \sim \nu_{s,a}$ and $s' \sim p_{s,a}$. Then perform the Q-learning update

$$\begin{aligned}&Q_{k+1}(s,a) = \\ &(1-\alpha_k)Q_k(s,a) + \alpha_k(r + \gamma v(Q_k)(s'))\end{aligned} \quad (7)$$

for some chosen step-size sequence $\{\alpha_k\}$.

Stochastic approximations (SA) for the fixed point of a contraction operator $\mathcal{H}$ refers to the class of algorithms using the update

$$X_{k+1} = (1-\alpha_k)X_k + \alpha_k\mathcal{H}(X_k) + W_k. \quad (8)$$

$\{W_k\}$ is a sequence satisfying $\mathbb{E}[W_k|\mathcal{F}_{k-1}] = 0$, thence is known as the *martingale difference noise*. The asymptotics of the above recursion are well understood, cf. Kushner and Yin (2013); while finite time behavior is discussed in the literature review. The recursion representation of the Q-learning (7) fits into the SA framework: Note that $r + \gamma v(Q)(s')$ is an *unbiased* estimator of $\mathcal{T}(Q)$ where $\mathcal{T}$ is the Bellman operator for the Q-function. This representation motivates the DR Q-learning.

3.2 Distributionally Robust Q-learning

A foundation to the possibility of employing a Q-learning is the following result.

Proposition 3.1. *The DR Bellman operator $\mathcal{T}_\delta$ is a γ-contraction on the Banach space $(\mathbb{R}^{S\times A}, \|\cdot\|_\infty)$.*

Given a simulator, a natural estimator for $\mathcal{T}_\delta(Q)$ is the empirical dual Bellman operator: replace the population transition and reward measures in (6) with the empirical version. However, the nonlinearity in the underlying measure, which can be seen from the dual functional, makes this estimator biased in general. Instead, Liu et al. (2022) propose an alternative by employing the idea in Blanchet et al. (2019); i.e., producing unbiased estimate of nonlinear functional of a probability measure using multi-level randomization. Yet, the number of samples requested in every iteration in Liu et al. (2022) is infinite in expectation. We improve this by extending the construction to a regime where the expected number of samples used is constant.

Before moving forward, we introduce the following notation. Denote the empirical distribution on n samples with $\nu_{s,a,n}$ and $p_{s,a,n}$ respectively; i.e. for $f : U \to \mathbb{R}$, where U could be the S or R,

$$\mathbb{E}_{u\sim\mu_{s,a,n}} f(u) = \frac{1}{n}\sum_{j=1}^{n} f(u_i)$$

for $\mu = \nu, p$ and $u_i = r_i, s'_i$. Moreover, we use $\mu^O_{s,a,n}$ and $\mu^E_{s,a,n}$ to denote the empirical distribution formed by the odd and even samples in $\mu_{s,a,2n}$. With this notation, we defined our estimator:

Definition 4. For given $g \in (0,1)$ and $Q \in \mathbb{R}^{S\times A}$, define the MLMC-DR estimator:

$$\widehat{\mathcal{T}}_{\delta,g}(Q)(s,a) := \widehat{R}_\delta(s,a) + \gamma\widehat{V}_\delta(Q)(s,a). \quad (9)$$

For $\widehat{R}_\delta(s,a)$ and $\widehat{V}_\delta(s,a)$, we sample N_1, N_2 from a geometric distribution $\mathrm{Geo}(g)$ independently, i.e., $\mathbb{P}(N_j = n) = p_n := g(1-g)^n, n \in \mathbb{Z}_{\geq 0}, j = 1, 2$. Then, we draw 2^{N_1+1} samples $r_i \sim \nu_{s,a}$ and 2^{N_2+1} samples $s'_i \sim p_{s,a}$.

Finally, we compute

$$\widehat{R}_\delta(s,a) := r_1 + \frac{\Delta^R_{N_1,\delta}}{p_{N_1}}, . \tag{10}$$

$$\widehat{V}_\delta(Q)(s,a) := v(Q)(s_1') + \frac{\Delta^P_{N_2,\delta}(Q)}{p_{N_2}}. \tag{11}$$

where

$$\begin{aligned}
&\Delta^R_{n,\delta} := \\
&\sup_{\alpha\geq 0}\left\{-\alpha\log\mathbb{E}_{r\sim\nu_{s,a,2^{n+1}}}\left[e^{-r/\alpha}\right]-\alpha\delta\right\} - \\
&\frac{1}{2}\sup_{\alpha\geq 0}\left\{-\alpha\log\mathbb{E}_{r\sim\nu^E_{s,a,2^n}}\left[e^{-r/\alpha}\right]-\alpha\delta\right\} - \\
&\frac{1}{2}\sup_{\alpha\geq 0}\left\{-\alpha\log\mathbb{E}_{r\sim\nu^O_{s,a,2^n}}\left[e^{-r/\alpha}\right]-\alpha\delta\right\}
\end{aligned} \tag{12}$$

and

$$\begin{aligned}
&\Delta^P_{n,\delta}(Q) := \\
&\sup_{\beta\geq 0}\left\{-\beta\log\mathbb{E}_{s'\sim p_{s,a,2^{n+1}}}\left[e^{-v(Q)(s')/\beta}\right]-\beta\delta\right\} - \\
&\frac{1}{2}\sup_{\beta\geq 0}\left\{-\beta\log\mathbb{E}_{s'\sim p^E_{s,a,2^n}}\left[e^{-v(Q)(s')/\beta}\right]-\beta\delta\right\} - \\
&\frac{1}{2}\sup_{\beta\geq 0}\left\{-\beta\log\mathbb{E}_{s'\sim p^O_{s,a,2^n}}\left[e^{-v(Q)(s')/\beta}\right]-\beta\delta\right\}.
\end{aligned} \tag{13}$$

Let $\left\{\widehat{\mathcal{T}}_{\delta,g,k}; k\in\mathbb{Z}_{\geq 0}\right\}$ be i.i.d. copies of $\widehat{\mathcal{T}}_{\delta,g}$. We construct our DR Q-Learning algorithm in Algorithm 1.

Algorithm 1 Multi-level Monte Carlo Distributionally Robust Q-Learning (MLMCDR Q-learning)

```
Input: Uncertainty radius $\delta > 0$, parameter $g \in (0,1)$, step-size sequence $\{\alpha_k : k \in \mathbb{Z}_{\geq 0}\}$, termination time $T$ (could be random).
Initialization: $\widehat{Q}_{\delta,0} \equiv 0$, $k = 0$.
repeat
    for every $(s,a) \in S \times A$ do
        Sample independent $N_1, N_2 \sim \mathrm{Geo}(g)$.
        Independently draw $2^{N_1+1}$ samples $r_i \sim \nu_{s,a}$ and $2^{N_2+1}$ samples $s_i' \sim p_{s,a}$.
        Compute $\widehat{R}_\delta(s,a)$ and $\widehat{V}_\delta(\widehat{Q}_{\delta,k})(s,a)$ using Equation (10)-(13).
    end for
    Compute $\widehat{\mathcal{T}}_{\delta,g,k+1}(\widehat{Q}_{\delta,k}) = \widehat{R}_\delta + \gamma\widehat{V}_\delta(\widehat{Q}_{\delta,k})$.
    Perform synchronous $Q$-learning update:
        $$\widehat{Q}_{\delta,k+1} = (1-\alpha_t)\widehat{Q}_{\delta,k} + \alpha_k\widehat{\mathcal{T}}_{\delta,g,k+1}(\widehat{Q}_{\delta,k}).$$
    $k \leftarrow k + 1$.
until $k = T$
```

Remark. The specific algorithm used in Liu et al. (2022) only has asymptotic guarantees and requires an infinite number of samples to converge, whereas we propose a variant that yields finite-sample guarantees. More specifically, in the algorithm, we choose $g \in (1/2, 3/4)$, while they choose $g \in (0, 1/2)$. This has important consequences: each iteration of Liu et al. (2022)'s algorithm requires an infinite number of samples in expectation, while in our algorithm, the expected number of samples used until iteration k is $n(g)|S||A|k$, where $n(g)$ doesn't depend on γ and the MDP instance.

4 Algorithm Complexity

All proofs to the results in this section are relegated to Sections 2 - 6 in the supplementary materials.

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_{k\in\mathbb{Z}_{\geq 0}}, \mathbb{P})$ be the underlying filtered probability space, where $\mathcal{F}_{k-1}$ is the σ-algebra generated by the random variates used before iteration k. We motivate our analysis by making the following observations. If we define the noise sequence

$$W_{k+1}(\widehat{Q}_{\delta,k}) := \widehat{\mathcal{T}}_{\delta,g,k+1}(\widehat{Q}_{\delta,k}) - \mathcal{T}_\delta(\widehat{Q}_{\delta,k}),$$

then the update rule of Algorithm 1 can be written as

$$\widehat{Q}_{\delta,k+1} = (1-\alpha_k)\widehat{Q}_{\delta,k} + \alpha_k\left(\mathcal{T}_\delta(\widehat{Q}_{\delta,k}) + W_{k+1}(\widehat{Q}_{\delta,k})\right). \tag{14}$$

By construction, we expect that under some condition, $\widehat{\mathcal{T}}_{\delta,g,k+1}(Q)$ is an unbiased estimate of $\mathcal{T}_\delta(Q)$. Hence $\mathbb{E}[W_{k+1}(\widehat{Q}_{\delta,k})|\mathcal{F}_k] = 0$. Therefore, Algorithm 1 has update of the form (8), and hence can be analysed as a stochastic approximation.

We proceed to rigorously establish this. First we introduce the following complexity metric parameter:

Definition 5. Define the *minimal support probability* as

$$p_\wedge := \inf_{s,a\in S\times A}\left[\inf_{r\in R:\nu_{s,a}(r)>0}\nu_{s,a}(r)\wedge\inf_{s'\in S:p_{s,a}(s')>0}p_{s,a}(s')\right].$$

The intuition of why the complexity of the MDP should depend on this minimal support probability is that in order to estimate the DR Bellman operator accurately, in worst case one must know the entire support of transition and reward distributions. Therefore, at least $1/p_\wedge$ samples are necessary. See Si et al. (2020) for a detailed discussion.

Assumption 1. *Assume the following holds:*

1. *The uncertainty set size δ satisfies $\frac{1}{2}p_\wedge \geq 1 - e^{-\delta}$.*
2. *The geometric probability parameter $g \in (0, 3/4)$.*

Remark. The first entry is a technical assumption that ensures the differentiability of the dual form of the robust functional. We use this specific form just for cleanness of presentation. Moreover, we conjecture that such restriction is not necessary to get the same complexity bounds. See the supplement Subsection 6.1 for a detailed discussion.

With Assumption 1 in place, we are ready to state our key tools and results. The following propositions underly our iteration and expected sample complexity analysis:

Proposition 4.1. *Let Q_δ^* be the unique fixed point of the DR Bellman operator $\mathcal{T}_\delta$. Then $\|Q_\delta^*\|_\infty \leq r_{\max}(1-\gamma)^{-1}$.*

Proposition 4.2. *Suppose Assumption 1 is in force. For fixed $Q : S \times A \to \mathbb{R}$, $\widehat{\mathcal{T}}_\delta(Q)$ is an unbiased estimate of $\mathcal{T}_{\delta,g}(Q)$; i.e. $\mathbb{E}W(Q) = 0$.*

Proposition 4.2 guarantees the validity of our construction of the unbiased MLMC-DR Bellman estimator. As explained before, this enables us to establish Algorithm 1 as SA to the fixed point of $\mathcal{T}_\delta$.

For simplicity, define the log-order term

$$\begin{aligned}\tilde{l} =&(3+\log(|S||A||R|) \vee \log(|S|^2|A|))^2 \\ &\times \log(11/p_\wedge)^2 \frac{4(1-g)}{g(3-4g)}.\end{aligned} \tag{15}$$

Proposition 4.3. *Suppose Assumption 1 is in force. For fixed Q, there exists constant $c > 0$ s.t.*

$$\mathbb{E}\|W(Q)\|_\infty^2 \leq \frac{c\tilde{l}}{\delta^4 p_\wedge^6}\left(r_{\max}^2 + \gamma^2\|Q\|_\infty^2\right).$$

Proposition 4.3 bounds the infinity norm squared of the martingale difference noise. It is central to our complexity results in Theorems 4.4 and 4.5.

Theorem 4.4. *Suppose Assumption 1 is in force. Running Algorithm 1 until iteration k and obtain estimator $\widehat{Q}_{\delta,k}$, the following holds:*

Constant stepsize: there exists $c, c' > 0$ s.t. if we choose the stepsize sequence

$$\alpha_k \equiv \alpha \leq \frac{(1-\gamma)^2\delta^4 p_\wedge^6}{c'\gamma^2\tilde{l}\log(|S||A|)},$$

then we have

$$\begin{aligned}&\mathbb{E}\|\widehat{Q}_{\delta,k} - Q_\delta^*\|_\infty^2 \leq \\ &\frac{3r_{\max}^2}{2(1-\gamma)^2}\left(1 - \frac{(1-\gamma)\alpha}{2}\right)^k + \frac{c\alpha r_{\max}^2\log(|S||A|)\tilde{l}}{\delta^4 p_\wedge^6(1-\gamma)^4}.\end{aligned}$$

Rescaled linear stepsize: there exists $c, c' > 0$ s.t. if we choose the stepsize sequence

$$\alpha_k = \frac{4}{(1-\gamma)(k+K)},$$

where

$$K = \frac{c'\tilde{l}\log(|S||A|)}{\delta^4 p_\wedge^6(1-\gamma)^3},$$

then we have

$$\mathbb{E}\|\widehat{Q}_{\delta,k} - Q_\delta^*\|_\infty^2 \leq \frac{cr_{\max}^2\tilde{l}\log(|S||A|)\log(k+K)}{\delta^4 p_\wedge^6(1-\gamma)^5(k+K)}.$$

We define the iteration complexity as

$$k^*(\epsilon) := \inf\{k \geq 1 : E\|Q_k - Q_\delta^*\|_\infty^2 \leq \epsilon^2\}.$$

The proof of Theorem 4.4 is based on the recent advances of finite-time analysis of stochastic approximation algorithms (Chen et al., 2020). Theorem 4.4 bounds the algorithmic error by the current iteration completed. This implies an iteration complexity bound, which we will make clear afterwards.

We consider the expected number of samples we requested from the generator to compute the MLMC-DR estimator for one (s,a)-pair. It depends on the geometric parameter g. Denote this by $n(g)$, then

$$n(g) = \mathbb{E}\left[2^{N_1+1} + 2^{N_2+1}\right] = \frac{4g}{2g-1}.$$

We define the expected sample complexity $n^*(\epsilon)$ as the total expected number of samples used until $k^*(\epsilon)$ iterations: $n^*(\epsilon) := |S||A|n(g)k^*(\epsilon)$. Note that when $g > 1/2$, $n(g)$ is finite. This finiteness and Theorem 4.4 would imply a finite expected sample complexity bound.

Assumption 2. *In addition to Assumption 1, assume $g \in (1/2, 3/4)$.*

Theorem 4.5. *Suppose Assumptions 1 and 2 are enforced. The expected sample complexity of Algorithm 1 for both stepsizes specified in Theorem 4.4 satisfies*

$$n^*(\epsilon) \lesssim \frac{r_{\max}^2|S||A|}{\delta^4 p_\wedge^6(1-\gamma)^5\epsilon^2}.$$

Remark. Theorem 4.5 follows directly from Theorem 4.4 by choosing the stepsize in the constant stepsize case

$$\alpha \simeq \frac{(1-\gamma)^4\delta^4 p_\wedge^6}{\gamma^2\tilde{l}\log(|S||A|)},$$

where $\simeq$ and $\lesssim$ mean equal and less or equal up to a log factor and universal constants. The choice of the stepsizes, however, is dependent on $p_\wedge$, which is typically unknown a priori. As discussed before, this dependent on $p_\wedge$ is an intrinsic source of complexity of the DR-RL problem. So, one direction for future works is to come up with efficient procedure to consistently estimate $p_\wedge$. Also, our bound has a δ^{-4} dependence. However, we believe that it should be $O(1)$ as $\delta \downarrow 0$. Because the algorithmic behavior will converge to that of the classical Q-learning

The sample complexity bound in Theorem 4.5 is not uniform in $g \in (1/2, 3/4)$ as we think of g being a design parameter, not an inherit model parameter. The sample complexity dependence on g is $\frac{n(g)4(1-g)}{g(3-4g)}$, minimized at $g \approx 0.64645$. For convenience, we will use $g = 5/8 = 0.625$.

5 Numerical Experiments

In this section, we empirically validate our theories using two numerical experiments. Section 5.1 dedicates to hard MDP instances, which are constructed in Li et al. (2021) to prove the lower bound of the standard Q-learning. In Section 5.2, we use the same inventory control problems as the one used in Liu et al. (2022) to demonstrate the superiority of our algorithms to theirs. More details on the experiment setup are in Section 7 of the supplemental materials.

5.1 Hard MDPs for Q-learning

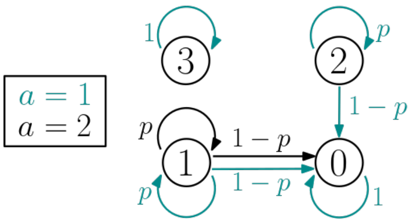
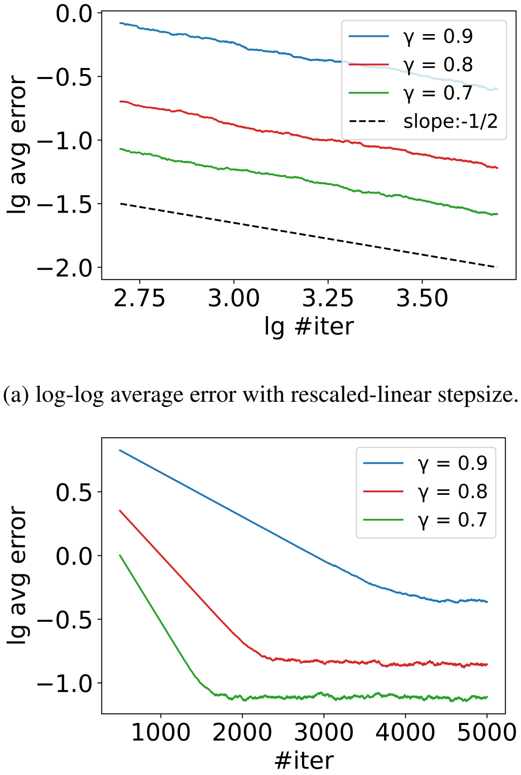


Figure 1: Hard MDP instances transition diagram.

First, we will test the convergence of our algorithm on the MDP in Figure 1. It has 4 states and 2 actions, the transition probabilities given action 1 and 2 are labeled on the arrows between states. Constructed in Li et al. (2021), it is shown that when $p = (4\gamma - 1)/(3\gamma)$ the standard non-robust Q-learning will have sample complexity $\tilde{\Theta}((1-\gamma)^{-4}\epsilon^{-2})$. We will use $\delta = 0.1$ in the proceeding experimentation.

Figures 2a and 2b show convergence properties of our algorithm with the rescaled-linear and constant step-size, respectively. Figure 2a is a log-log scale plot of the average (across 200 trajectories) error $\|Q_{\delta,k} - Q_\delta^*\|_\infty$ achieved at a given iteration k under rescaled-linear step-size $\alpha_k = 1/(1 + (1-\gamma)k)$. We see that the algorithm is indeed converging. Moreover, Theorem 4.4 predicts that the slope of each line should be close to $-1/2$, which corresponds to the canonical asymptotic convergence rate $n^{-1/2}$ of the stochastic approximations under the Robbins–Monro stepsize regime. This is confirmed in Figure 2a. The algorithm generates Figure 2b uses constant step-size $\alpha_k \equiv 0.008$. In Figure 2b, the horizontal axis is in linear scale. So, we observe that for all three choices of γ, the averaged errors first decay geometrically and then stay constant as the number of iterations increases. This is also consistent with the prediction of Theorem 4.4.

Next, we would like to visualize the γ-dependence of our algorithm with the rescaled-linear step-size on this hard MDP. Note that if we construct a sequence of hard MDPs with different $(\gamma, Q_\delta^*(\gamma), Q^*(\gamma))$, then for fixed iteration k, Theorem 4.4 (ignoring $p_\wedge$ as a function of γ) implies that $\log E\|Q_{\delta,k}(\gamma) - Q_\delta^*(\gamma)\|_\infty \lesssim -\frac{5}{2}\log(1-\gamma)$. On the other hand the standard Q-learning, from Li et al. (2021), $\log\|Q_k(\gamma) - Q^*(\gamma)\|_\infty \asymp -2\log(1-\gamma)$; corresponding to a $(1-\gamma)^{-4}$ dependence.



(a) log-log average error with rescaled-linear stepsize.

(b) log average error with constant stepsize.

Figure 2: Convergence of Algorithm 1 on MDP 1

Figure 3 plots the average error of the sequence of MLMCDR Q-learning at iteration $k = 500, 1000, 1500$ against $\log(1-\gamma)$, and performs a linear regression to extract the slope. We see that for all k, the slope is very close to -2, suggesting a $(1-\gamma)^{-4}$ dependence. Given that our analysis is based on the finite analysis of SA algorithms in Chen et al. (2020), which, if applied to the classical Q-learning, will also yield a $(1-\gamma)^{-5}$ dependence, we think the actual worst case sample complexity is $(1-\gamma)^{-4}$. However, the validity is less clear: the classical non-robust Q-learning employs the empirical bellman operator based on one sample each iteration, which is bounded; on the other hand, our estimator uses a random number of samples per iteration, which is only finite w.p.1. This distinction is not visible through the framework of Chen et al. (2020) and may result in a rate degradation.

As we point out eariler, we believe that the complexity dependence on δ should be $O(1)$ as $\delta \downarrow 0$. We also included some experimentation to confirm this in the appendix.

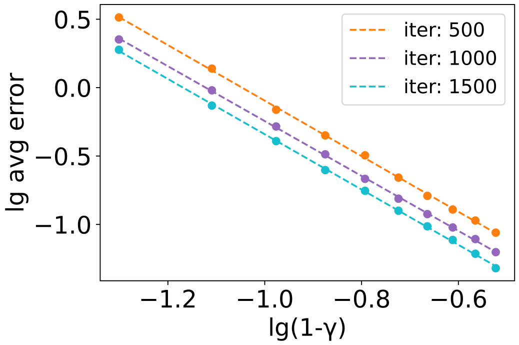
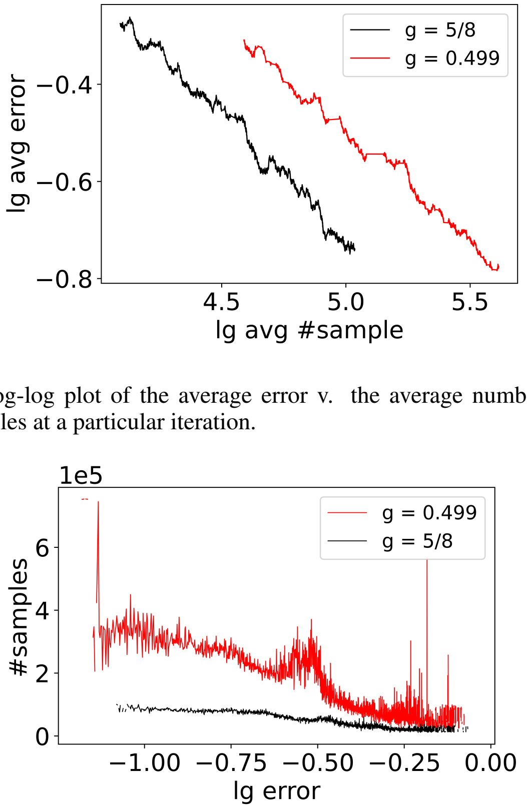


Figure 3: log averaged error against $\log(1-\gamma)$, the slopes of the regression line for iteration $k = 500, 1000, 1500$ are $-2.031, -2.007, -2.021$.



(a) log-log plot of the average error v. the average number of samples at a particular iteration.

(b) (Smoothed) algorithm error v. number of samples.

Figure 4: Algorithm comparison: inventory model.

5.2 Lost-sale Inventory Control

In this section, we apply Algorithm 1 to the lost-sale inventory control problem with i.i.d. demand, which is also used in Liu et al. (2022).

In this model, we consider state and action spaces $S = \{0, 1, \ldots, n_s\}$, $A = \{0, 1, \ldots, n_a\}$, $n_a \le n_s$; the state-action pairs $\{(s, a) \in S \times A : s + a \le n_s\}$. The demand has support $D = \{0, 1, \ldots, n_d\}$. We assume that at the beginning of the day t, we observe the inventory level s_t and place an order of a_t items which will arrive instantly. The cost incurred on day t is

$$\begin{aligned} C_t =& k \cdot 1\{A_t > 0\} + h \cdot (S_t + A_t - D_t)_+ \\ &+ p \cdot (S_t + A_t - D_t)_- \end{aligned}$$

where k is the ordering cost, h is the holding cost per unit of inventory, and p is the lost-sale price per unit of inventory.

For this numerical experiment, we use $\delta = 0.5$, $\gamma = 0.7$, $n_s = n_a = n_d = 7$, $k = 3$, $h = 1$, $p = 2$. Under the data-collection environment, we assume $D_t = \text{Unif}(D)$ and is i.i.d. across time,

Figure 4a and 4b compares the sample complexity of our algorithm to the one proposed in Liu et al. (2022) ($g \in (1/2, 3/4)$ verses. $g \in (0, 1/2)$). We run 300 independent trajectories and record the errors and number of samples used at iteration $1000 : 5000$. We clearly observe that our algorithm (black line) outperforms the one in Liu et al. (2022) (red line).

Specifically, Figure 4a, the log-log plot, shows a black line (our algorithm) of slope close to $-1/2$. This is consistent with our theory and also Figure 2a. The red line (the algorithm in Liu et al. (2022)) seems to be affine as well with a similar slope. However, there are visible jumps (horizontal segments) along the line. This is due to the MLMC-DR Bellman estimators having infinite mean when $g < 1/2$. Furthermore, the overall performance of our algorithm is significantly better. Figure 4b is a (smoothed) scatter plot of the (lg-error, number of sample used) pairs. We can clearly see that not only our algorithm (black line) has better sample complexity performance at every error value, but also the black line has significantly less variation compare to the red line (the algorithm in Liu et al. (2022)). Again, this is due to our MLMC estimator having a constant order expected sample size in contrast to the infinite expected sample size of the MLMC estimator in Liu et al. (2022) with the parameter $g < 1/2$.

6 Conclusion

We establish the first model-free finite sample complexity bound for the DR-RL problem: $\tilde{O}(|S||A|(1-\gamma)^{-5}\epsilon^{-2}\delta^{-4}p_\wedge^{-6})$. Though optimal in $|S||A|$, we believe that the dependence on other parameters are sub-optimal for Algorithm 1 and need research efforts for further improvements. Also, a minimax complexity lower bound of our algorithm will facilitate a better understanding of its performance. We leave these for future works.

Acknowledgments

Material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF grants 1915967 and 2118199.

This work is supported in part by National Science Foundation grant CCF-2106508, and the Digital Twin research grant from Bain & Company. Zhengyuan Zhou acknowledges the generous support from New York University's 2022-2023 Center for Global Economy and Business faculty research grant.

References

Abadeh, S. S., Nguyen, V. A., Kuhn, D., and Esfahani, P. M. (2018). Wasserstein distributionally robust kalman filtering. In *Advances in Neural Information Processing Systems*, pages 8483–8492. 2

Agarwal, A., Kakade, S., and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 67–83. PMLR. 2

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349. 2

Bayraksan, G. and Love, D. K. (2015). Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pages 1–19. Catonsville: Institute for Operations Research and the Management Sciences. 2

Bertsekas, D. P. (2011). Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*. 1

Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations Research*, 52(1):35–53. 2

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600. 2

Blanchet, J. H., Glynn, P. W., and Pei, Y. (2019). Unbiased multilevel monte carlo: Stochastic optimization, steady-state simulation, quantiles, and other applications. *arXiv preprint arXiv:1904.09929*. 4

Chen, Z., Kuhn, D., and Wiesemann, W. (2018). Data-driven chance constrained programs over wasserstein balls. *arXiv preprint arXiv:1809.00210*. 2

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8223–8234. Curran Associates, Inc. 2, 6, 7, 14

Choi, J. J., Laibson, D., Madrian, B. C., and Metrick, A. (2009). Reinforcement learning and savings behavior. *The Journal of finance*, 64(6):2515–2534. 1

Cisneros-Velarde, P., Petersen, A., and Oh, S.-Y. (2020). Distributionally robust formulation and model selection for the graphical lasso. In *International Conference on Artificial Intelligence and Statistics*, pages 756–765. PMLR. 2

Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612. 2

Deng, Y., Bao, F., Kong, Y., Ren, Z., and Dai, Q. (2017). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664. 1

Drew, K. (2015). California robot teaching itself to walk like a human toddler. *NBC News*. 1

Duchi, J., Glynn, P., and Namkoong, H. (2016). Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*. 2

Duchi, J., Hashimoto, T., and Namkoong, H. (2019). Distributionally robust losses against mixture covariate shifts. *arXiv preprint arXiv:2007.13982*. 2

Duchi, J. and Namkoong, H. (2018). Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*. 2

Even-Dar, E., Mansour, Y., and Bartlett, P. (2003). Learning rates for q-learning. *Journal of machine learning Research*, 5(1). 2

Gao, R. and Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*. 2

Gao, R., Xie, L., Xie, Y., and Xu, H. (2018). Robust hypothesis testing using wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pages 7902–7912. 2

Ghosh, S. and Lam, H. (2019). Robust analysis in stochastic simulation: Computation and performance guarantees. *Operations Research*. 2

González-Trejo, J. I., Hernández-Lerma, O., and Hoyos-Reyes, L. F. (2002). Minimax control of discrete-time stochastic systems. *SIAM Journal on Control and Optimization*, 41(5):1626–1659. 2

Gu, S., Holly, E., Lillicrap, T., and Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE. 1

Ho-Nguyen, N., Kılınç-Karzan, F., Küçükyavuz, S., and Lee, D. (2020). Distributionally robust chance-constrained programs with right-hand side uncertainty under wasserstein ambiguity. *arXiv preprint arXiv:2003.12685*. 2

Hu, Z. and Hong, L. J. (2013a). Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*. 2, 3

Hu, Z. and Hong, L. J. (2013b). Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*. 20

Huang, C., Lucey, S., and Ramanan, D. (2017). Learning policies for adaptive tracking with deep feature cascades. *ICCV*, pages 105–114. 1

Iyengar, G. (2005). Robust dynamic programming. *Math. Oper. Res.*, 30:257–280. 2, 3, 12, 13

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274. 1

Kushner, H. and Yin, G. (2013). *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer New York. 4

Lam, H. (2019). Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105. 2

Lam, H. and Zhou, E. (2017). The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4):301–307. 2

Lee, J. and Raginsky, M. (2018). Minimax statistical learning with wasserstein distances. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 2692–2701, USA. Curran Associates Inc. 2

Li, G., Cai, C., Chen, Y., Gu, Y., Wei, Y., and Chi, Y. (2021). Is q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*. 2, 7

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872. Curran Associates, Inc. 2

Li, Y., Szepesvari, C., and Schuurmans, D. (2009). Learning exercise policies for american options. In *Artificial Intelligence and Statistics*, pages 352–359. 1

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. (2022). Distributionally robust q-learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13623–13643. PMLR. 1, 2, 4, 5, 7, 8, 29

Luenberger, D. G., Ye, Y., et al. (2021). *Linear and nonlinear programming*. Springer. 23

Maitin-Shepard, J., Cusumano-Towner, M., Lei, J., and Abbeel, P. (2010). Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *2010 IEEE International Conference on Robotics and Automation*, pages 2308–2315. 1

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166. 2

Namkoong, H. and Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2216–2224. Red Hook: Curran Associates Inc. 2

Nguyen, V. A., Kuhn, D., and Esfahani, P. M. (2018). Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *arXiv preprint arXiv:1805.07194*. 2

Panaganti, K. and Kalathil, D. (2021). Sample complexity of robust reinforcement learning with a generative model. 2

Powell, W. B. (2007). *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons. 1

Sadeghi, F. and Levine, S. (2016). Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*. 1

Schulman, J., Ho, J., Lee, A. X., Awwal, I., Bradlow, H., and Abbeel, P. (2013). Finding locally optimal, collision-free trajectories with sequential convex optimization. In *Robotics: science and systems*, volume 9, pages 1–10. Citeseer. 1

Shafieezadeh-Abadeh, S., Esfahani, P., and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584. 2

Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275. 2

Shapiro, A. (2022). Distributionally robust modeling of optimal control. *Operations Research Letters*, 50(5):561–567. 2

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2014). *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA. 20

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. (2020). Distributional robust batch contextual bandits. 2, 5, 19

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018). Near-optimal time and sample complexities for solving markov decision processes with a generative model. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. 2

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489. 1

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144. 1

Sinha, A., Namkoong, H., and Duchi, J. (2018). Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*. 2

Staib, M. and Jegelka, S. (2017). Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*. 2

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press. 1

Szepesvári, C. (2010). Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103. 1

Volpi, R., Namkoong, H., Sener, O., Duchi, J., Murino, V., and Savarese, S. (2018). Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*. 2

Vu, H., Tran, T., Yue, M.-C., and Nguyen, V. A. (2022). Distributionally robust fair principal components via geodesic descents. *arXiv preprint arXiv:2202.03071*. 2

Wainwright, M. J. (2019a). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for q-learning. 2

Wainwright, M. J. (2019b). Variance-reduced q-learning is minimax optimal. 2

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183. 2, 3, 12

Xu, H. and Mannor, S. (2010). Distributionally robust markov decision processes. In *Advances in Neural Information Processing Systems*, pages 2505–2513. 2

Yang, I. (2020). Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*. 2

Yang, W., Zhang, L., and Zhang, Z. (2021). Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. 1, 2

Zhao, C. and Guan, Y. (2018). Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262 – 267. 2

Zhao, C. and Jiang, R. (2017). Distributionally robust contingency-constrained unit commitment. *IEEE Transactions on Power Systems*, 33(1):94–102. 2

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3331–3339. PMLR. 1, 2

Appendices

A Robust Markov Decision Processes and Reinforcement Learning

In this section, we rigorously introduce the robust Markov Decision Process formulation and its dynamic programming representation a.k.a. the Bellman equation. Our presentation here is a brief review of the constructions and results in Iyengar (2005).

We specialize to the infinite horizon, finite state-action-reward MDP setting. In particular recall that $\mathcal{M}_0 = (S, A, R, \mathcal{P}_0, \mathcal{R}_0, \gamma)$, where S, A, and $R \subsetneq \mathbb{R}_{\geq 0}$ are finite state, action, and reward spaces respectively. $\mathcal{P}_0 = \{p_{s,a} : s \in S, a \in A\}$, $\mathcal{R}_0 = \{\nu_{s,a} : s \in S, a \in A\}$ are the sets of the reward and transition distributions. We consider the KL uncertainty sets $\mathcal{R}_{s,a}(\delta) = \{\nu : D_{\mathrm{KL}}(\nu || \nu_{s,a}) \leq \delta\}$ and $\mathcal{P}_{s,a}(\delta) = \{p : D_{\mathrm{KL}}(p || p_{s,a}) \leq \delta\}$. Note that the robust MDP literature refers to the set $\mathcal{P}_{s,a}(\delta)$ as ambiguity set so as to distinguish it from randomness (uncertainty). Here, we will follow the convention in DRO literature and use the name uncertainty set. Before moving forward, we define some notations. For discrete set Y, let $\mathcal{P}(Y)$ denote the set of probability measures on $(Y, 2^Y)$. Given a probability measure $p \in \mathcal{P}(Y)$ and a function f on $(Y, 2^Y)$, let $p[f]$ denote the integral, i.e. $p[f] = \sum_{y \in Y} p_y f_y$.

We will employ a canonical construction using the product space $\Omega = (S \times A \times R)^{\mathbb{Z}_{\geq 0}}$ and $\mathcal{F}$ the σ-field of cylinder sets as the underlying measurable space. At time t the controller at state s_t using action a_t collects a randomized reward $r_t \sim \nu_{s,a}$.

Define the *history* of the controlled Markov chain as the sequence $h_t = (s_0, a_0, s_1, a_1, \ldots, s_t)$ and the collection $h_t \in H_t = (S \times A)^t \times S$. A history-dependent randomized decision rule at time t is a conditional measure $d_t : H_t \to \mathcal{P}(A)$. A decision rule d_t is Markovian if for any $h_t, h_t' \in H_t$ s.t. $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$ and $h_t = (s_0', a_0', \ldots, s_{t-1}', a_{t-1}', s_t)$, we have $d_t(h_t)[\cdot] = d_t(h_t')[\cdot]$. A decision rule is deterministic $d_t(h_t)[\mathbb{1}_a] = 1$ for some $a \in A$, where $\mathbb{1}_a$ is the indicator of $\{a\}$, seen as a vector of all 0 but 1 at a. The set of transition probabilities consistent with a history dependent decision rule within the uncertainty sets is defined as

$$\begin{aligned}\mathcal{K}^{d_t} = \{&p : H_t \to \mathcal{P}(A \times R \times S) : \forall h_t \in H_t, s \in S, a \in A,\\ &p(h_t)[\mathbb{1}_{(a,r,s')}] = d_t(h_t)[\mathbb{1}_a]\nu_{s_t,a}[\mathbb{1}_r]p_{s_t,a}[\mathbb{1}_{s'}] \text{ for some } \nu_{s_t,a} \in \mathcal{R}_{s_t,a}(\delta), p_{s_t,a} \in \mathcal{P}_{s_t,a}(\delta)\}.\end{aligned}$$

Note that this is equivalent to say that $\mathcal{R}_\delta = \{\nu_{s,a} : (s,a) \in S \times A\}$ is chosen from the product uncertainty set $\prod_{s,a} \mathcal{R}_{s,a}(\delta)$; same for p. This is known as s,a-rectangularity; c.f. Wiesemann et al. (2013). Intuitively, $\mathcal{K}^{d_t}$ is the set of history dependent measures generating a current action a_t from $d_t(h_t)$, and, condition on s_t, a_t, independently generate r_t and s_{t+1} from some $\nu \in \mathcal{R}_{s_t,a}(\delta)$ and $p \in \mathcal{P}_{s_t,a}(\delta)$.

A history dependent policy π is a sequence of decision rules $\pi = (d_t, t \in \mathbb{Z}_{\geq 0})$. We will denote the history-dependent policy class as $\Pi = \{\pi : (d_t, t \in \mathbb{Z}_{\geq 0})\}$. Given a policy π we naturally obtain a family of probability measures on $(\Omega, \mathcal{F})$; i.e. for initial distribution $\mu \in \mathcal{P}(S)$

$$\mathcal{K}_\mu^\pi = \left\{ P_\mu : \begin{array}{l} P(\{(s_0, a_0, r_0, s_1, a_1, r_1, \ldots, r_{T-1}, s_T)\}) = \mu(s_0) \displaystyle\prod_{t=0}^{T-1} p_k(h_t)[\mathbb{1}_{(a_t, r_t, s_{t+1})}] \\ \forall T \in \mathbb{Z}_{\geq 0}, s_i \in S, a_i \in A, r_i \in R, \text{ for some } \left\{p_t \in \mathcal{K}^{d_t}, t \leq T-1\right\} \end{array} \right\}$$

where the probabilities of a sample path until time T define the finite dimensional distributions, thence uniquely extend to probability measures on $(\Omega, \mathcal{F})$ by the Kolmogorov extension theorem. Formally, we can write $\mathcal{K}^\pi = \prod_{t \geq 0} \mathcal{K}^{d_t}$. Sometimes, we want to fix the initial state-action $(s_0 = s, a_0 = a)$. This can be done if we let $\mu = \delta_s$ and restrict π s.t. $d_0(h_0)[\mathbb{1}_a]$. In order to develop a dynamic programming theory, we assume that the sample path uncertainty set is the full $\mathcal{K}_\mu^\pi$. This is the notion of rectangularity in Iyengar (2005).

With these constructions, we can rigorously define the *pessimistic* value function $U_\delta^\pi : S \to \mathbb{R}_+$

$$U_\delta^\pi(s) := \inf_{P \in \mathcal{K}_{\delta_s}^\pi} E^P\left[\sum_{t=0}^\infty \gamma^t r_t\right]$$

and the optimal pessimistic value function as the following minimax value

$$U_\delta^*(s) := \sup_{\pi \in \Pi} U_\delta^\pi(s) = \sup_{\pi \in \Pi} \inf_{P \in \mathcal{K}_{\delta_s}^\pi} E^P\left[\sum_{t=0}^\infty \gamma^t r_t\right].$$

As mentioned before, the rectangularity assumptions allow us to develop a dynamic programming principle, a.k.a. Bellman equation, for the optimal pessimistic value. Before that, we first define and point out some important distinctions to some policy classes. Recall Π is the general history dependent policy class. The Markovian deterministic policy class is $\Pi_{\text{MD}} = \{\pi : d_t \in D_{\text{MD}}\}$ where D_{MD} is the set of Markovian and deterministic decision rule. The stationary Markovian deterministic policy class $\Pi_{\text{SMD}} = \Pi_{\text{MD}} \cap \{\pi : d_t = d, \forall t\}$. Next, we define the DR Bellman operator for the state value function:

$$\begin{aligned}\mathcal{B}_\delta(v)(s) &= \sup_{d_0\in D_{\text{MD}}} \inf_{p\in\mathcal{K}^{d_0}_{\delta_s}} E_{a,r,s'\sim p}\left[r+\gamma v(s')\right]\\ &= \sup_{a\in A} \inf_{\nu\in\mathcal{R}_{s,a}(\delta), p\in\mathcal{P}_{s,a}(\delta)} E_{r\sim\nu,s'\sim p}[r+\gamma v(s')]\end{aligned}$$

The second equality follows from noting that the first supremum is achieved by the greedy decision rule. This is our Definition 1; we also recall V^*_δ. Now, we state the dynamic programming principles for the robust MDP:

Theorem A.1 (Theorem 3.1, 3.2 of Iyengar (2005))**.** *The following statements hold:*

1. $U^*_\delta(s) = \sup_{\pi\in\Pi_{SMD}} U^\pi_\delta(s)$ *for all* $s \in S$.
2. $\mathcal{B}_\delta$ *is a contraction on* $(\mathbb{R}^{|S|}, \|\cdot\|_\infty)$*, hence* V^*_δ *is the unique solution to* $v = \mathcal{B}_\delta(v)$.
3. $U^*_\delta(s) = V^*_\delta(s)$.

Remark. Our construction allows a randomization of the reward in comparison to Iyengar (2005) and its results easily generalizes. The first entry is a consequence of S, A being finite (so that $|\Pi_{\text{SMD}}| < |S|^{|A|}$) and Iyengar (2005)'s Corollary 3.1: for any $\epsilon > 0$, $\exists\pi_\epsilon \in \Pi_{\text{SMD}}$ s.t. $\forall s \in S$, $U^*_\delta(s) \le U^{\pi_\epsilon}_\delta(s) + \epsilon$.

This establishes the validity of our approach to the robust control problem by staring from the DR Bellman equation. Of course, the DR Q-function, thence the entire DR-RL paradigm, can be interpreted analogous to the classical tabular RL problem.

B Notation and Formulation Remarks

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}, P)$ be a filtered probability space from which we can draw samples, where $\mathcal{F}_k$ is the smallest σ-algebra generated by the samples used until iteration k of the algorithm.

Before presenting our proof, we introduce the following notations. For finite discrete measureable space $(Y, 2^Y)$, fixed $u \in m2^Y$, and signed measure $m \in \mathcal{M}_\pm(Y, 2^Y)$, let $m[u]$ denote the integral. Let $w = w(\alpha) = e^{-u/\alpha}$ and

$$f(\mu, u, \alpha) = -\alpha \log \mu[e^{-u/\alpha}] - \alpha\delta. \tag{16}$$

We clarify that $f(\mu, u, 0) = \lim_{\alpha\downarrow 0} f(\mu, u, \alpha) = \operatorname{ess\,inf}_\mu u$. Sometimes, we only need to consider the perturbation analysis on the line of center measures $\left\{t\mu^O_{2^n} + (1-t)\mu^E_{2^n} : t \in [0,1]\right\}$. So, it is convenient to define

$$\begin{aligned}\mu_{s,a,n}(t) &= t\mu^O_{s,a,2^n} + (1-t)\mu^E_{s,a,2^n}\\ m_{s,a,n} &= \mu^O_{s,a,2^n} - \mu^E_{s,a,2^n}\\ g_{s,a,n}(t,\alpha) &= f(\mu_{s,a,2^n}(t), u, \alpha).\end{aligned} \tag{17}$$

Note that we will not explicitly indicate the dependence of u for the function g. We will also drop the dependence on (s, a) when clear.

Recall that the MLMC DR Bellman operator estimator has the form

$$\hat{\mathcal{T}}_\delta(Q)(s,a) = \hat{R}_\delta(s,a) + \gamma\hat{V}_\delta(Q)(s,a)$$

We want to pursue a unified analysis of $\hat{R}_\delta$ and $\hat{V}_\delta(Q)$. Define the Δ operator

$$\begin{aligned}\Delta(\mu^E_{s,a,2^n}, \mu^O_{s,a,2^n}, u) &= \sup_{\alpha\ge 0} f(\mu_{s,a,2^{n+1}}, u, \alpha) - \frac{1}{2}\sup_{\alpha\ge 0} f(\mu^E_{s,a,2^n}, u, \alpha) - \frac{1}{2}\sup_{\alpha\ge 0} f(\mu^O_{s,a,2^n}, u, \alpha)\\ &= \frac{1}{2}\left[\sup_{\alpha\ge 0} g_{s,a,n}(1/2,\alpha) - \sup_{\alpha\ge 0} g_{s,a,n}(0,\alpha)\right] + \frac{1}{2}\left[\sup_{\alpha\ge 0} g_{s,a,n}(1/2,\alpha) - \sup_{\alpha\ge 0} g_{s,a,n}(1,\alpha)\right].\end{aligned}$$

Sometimes, we will drop the dependence on $\mu_{s,a,2^{n+1}}$ and u in the following proof. Recall that $Q : S \times A \to \mathbb{R}$, we define $v(Q) : S \to \mathbb{R}$ by $v(Q)(s) = \max_{a \in A} Q(s,a)$. Let $N_1, N_2 \sim p_n = g(1-g)^n$ and $\left\{R_{s,a,i}, S_{s,a,j}, i = 0 \dots 2^{N_1+1}, j = 0 \dots 2^{N_2+1}\right\}$ generated from the reward and transition probabilities; let $\mu^R_{s,a,2^{N_1+1}}$ and $\mu^V_{s,a,2^{N_2+1}}$ be the empirical measures form by the samples $\left\{R_{s,a,i}, i = 1 \dots 2^{N_1+1}\right\}$ and $\left\{S_{s,a,i}, i = 1 \dots 2^{N_2+1}\right\}$ respectively, then under our new notation

$$\hat{R}_\delta(s,a) = R_{s,a,0} + \frac{\Delta(\mu^{R,E}_{s,a,2^{N_1}}, \mu^{R,O}_{s,a,2^{N_1}}, id)}{p_{N_1}}$$
$$\hat{V}_\delta(Q)(s,a) = v(Q)(S_{s,a,0}) + \frac{\Delta(\mu^{V,E}_{s,a,2^{N_2}}, \mu^{V,O}_{s,a,2^{N_2}}, v(Q))}{p_{N_2}}$$

where $id : \mathbb{R} \to \mathbb{R}$ is the identity function. Note that μ^R is supported on a finite subset of $\mathbb{R}$, but μ^V is supported on S. This construction suggests that one can employ almost identical analysis on $\hat{R}_\delta$ and $\hat{V}_\delta(Q)$. For notation simplicity, we will write $\Delta^R_{s,a,n}$, $\Delta^V_{s,a,n}$, or the generic $\Delta_{s,a,n}$ when the dependent empirical measures $\mu^E_{s,a,2^n}, \mu^O_{s,a,2^n}$ and function u are contextually clear.

Finally, we note that in this notation the minimal support probability definition 5 becomes

$$p_\wedge = \inf_{s,a \in S \times A} \min\left\{ \inf_{r \in R} \mu^R_{s,a}(r), \inf_{s' \in S} \mu^V_{s,a}(s') \right\}.$$

C Proof of Theorem 4.4

In this section to Theorem 4.4 assuming the propositions we state ealier.

Proof. We will denote $E_k[\cdot] = E[\cdot|\mathcal{F}_k]$. Proposition 4.2 and 4.3 implies that $E_k W_{k+1}(Q_k) = 0$ and

$$E_k \|W_{k+1}(Q_k)\|^2_\infty \le \frac{c\tilde{l}}{\delta^4 p^6_\wedge}\left(r^2_{\max} + \gamma^2 \|Q_k\|^2_\infty\right).$$

Apply Corollary 2.1.1. in Chen et al. (2020) under the condition of Corollary 2.1.3, we have that there exists constant $c, c', c'', c''' > 0$ s.t. when

$$\alpha_k \equiv \alpha \le \frac{(1-\gamma)^2 \delta^4 p^6_\wedge}{c' \gamma^2 \tilde{l} \log(|S||A|)},$$

we have

$$\begin{aligned} E\|Q_k - Q^*\|^2_\infty &\le \frac{3}{2}\|Q_0 - Q^*\|^2_\infty \left(1 - \frac{1}{2}(1-\gamma)\alpha\right)^k + \frac{c''\alpha \log(|S||A|) c\tilde{l}(r^2_{\max} + 2\gamma^2\|Q^*\|^2_\infty)}{\delta^4 p^6_\wedge (1-\gamma)^2} \\ &\le \frac{3r^2_{\max}}{2(1-\gamma)^2}\left(1 - \frac{1}{2}(1-\gamma)\alpha\right)^k + \frac{c'''\alpha r^2_{\max} \log(|S||A|)\tilde{l}}{\delta^4 p^6_\wedge (1-\gamma)^4}. \end{aligned}$$

where the last inequality follows from Proposition 4.1. Also there exists some other constant $c, c', c'', c''' > 0$ s.t. when

$$\alpha_k = \frac{4}{(1-\gamma)(k+K)},$$
$$K = \frac{c'\tilde{l}\log(|S||A|)}{\delta^4 p^6_\wedge (1-\gamma)^3},$$

we have

$$\begin{aligned} E\|Q_k - Q^*\|^2_\infty &\le \frac{3}{2}\|Q_0 - Q^*\|^2_\infty \frac{K}{k+K} + \frac{c'' \log(|S||A|)\tilde{l}(r^2_{\max} + 2\gamma^2\|Q^*\|^2_\infty)}{\delta^4 p^6_\wedge (1-\gamma)^3} \frac{\log(k+K)}{k+K} \\ &\le \frac{c''' r^2_{\max}\tilde{l}\log(|S||A|)\log(k+K)}{\delta^4 p^6_\wedge (1-\gamma)^5 (k+K)}. \end{aligned}$$

Renaming the constants gives the theorem. □

D Analysis of the Δ Operator

In this section, our analysis uses the compact notation defined in Section 2.

Before proving the propositions, we present some key identities of the operator $\Delta_{s,a,n}$. For each $(s,a) \in S \times A$, $\mu^E_{s,a,2^n}, \mu^O_{s,a,2^n}, \mu_{s,a,2^{n+1}}$ are sampled from $\mu_{s,a}$ the population measure (on finite discrete measureable space $(Y, 2^Y)$) and independent across (s,a). Define for $p > 0$

$$\Omega_{s,a,n}(p) = \left\{\omega : \sup_{y\in Y} |\mu^O_{s,a,2^n}(\omega)(y) - \mu_{s,a}(y)| \le p, \sup_{y\in Y} |\mu^E_{s,a,2^n}(\omega)(y) - \mu_{s,a}(y)| \le p\right\}$$

We will choose $p = \frac{1}{4}p_\wedge$ where

$$p_\wedge \le \inf_{\substack{(s,a)\in S\times A\\ y:\mu_{s,a}(y)>0}} \mu_{s,a}(y)$$

Note that on $\Omega_{s,a,n}(p)$, for any (s,a)

$$\mu_{s,a} \sim \mu_{s,a,2^{n+1}} \sim \mu^E_{s,a,2^n} \sim \mu^O_{s,a,2^n}. \tag{18}$$

Moreover, we have for all $t \in [0,1]$ that could depend on ω,

$$\begin{aligned}\sup_{y\in Y} |t\mu^E_{s,a,2^n} + (1-t)\mu^O_{s,a,2^n} - \mu| &\le t \sup_{y\in Y} |\mu^E_{s,a,2^n} - \mu| + (1-t) \sup_{y\in Y} |\mu^O_{s,a,2^n} - \mu| \\ &\le p.\end{aligned} \tag{19}$$

In this section, we want to bound

$$\begin{aligned}E \sup_{(s,a)\in S\times A} \Delta^2_{s,a,n} &= E \sup_{(s,a)\in S\times A} \Delta^2_{s,a,n}\mathbb{1}_{\Omega_{s,a,n}(p)} + E \sup_{(s,a)\in S\times A} \Delta^2_{s,a,n}\mathbb{1}_{\Omega_{s,a,n}(p)^c} \\ &=: E_1 + E_2.\end{aligned} \tag{20}$$

To bound two terms in equation (20), we introduce the following key results:

Lemma D.1. *There exists $t \in (0,1)$ s.t.*

$$\Delta^2_{s,a,n}\mathbb{1}_{\Omega_{s,a,n}(p)} \le \frac{1025\log(11/p_\wedge)^2\|u\|^2_{L^\infty(\mu_{s,a})}}{\delta^4 p_\wedge^2}\left\|\frac{dm_{s,a,n}}{d\mu_{s,a,n}(t)}\right\|^4_{L^\infty(\mu)}\mathbb{1}_{\Omega_{s,a,n}(p)}.$$

Lemma D.2. *Let $\{Y_i, i = 1\dots n\}$ be σ^2 sub-Gaussian, not necessarily independent, then*

$$EZ := E\left[\max_{i=1\dots n} Y_i^4\right] \le 16\sigma^4\left(3 + \log n\right)^2.$$

Lemma D.3. *For any $\nu \ll \mu$, we have*

$$-\|u\|_{L^\infty(\mu)} \le \sup_{\alpha\ge 0} f(\nu,\alpha) \le \|u\|_{L^\infty(\mu)}.$$

By (19),

$$\inf_{y\in Y} \mu_{s,a,n}(t)(y) \ge p_\wedge - p = \frac{3}{4}p_\wedge.$$

So, using (20) and Lemma D.1, we can bound

$$\begin{aligned}E_1 &\le \frac{1025\log(11/p_\wedge)^2\|u\|^2_{L^\infty(\mu_{s,a})}}{\delta^4 p_\wedge^2} E \sup_{(s,a)\in S\times A}\left(\operatorname{ess\,sup}_{\mu}\left|\frac{dm_{s,a,n}}{d\mu_{s,a,n}(t)}\right|\right)^4 \mathbb{1}_{\Omega_{u,n,\epsilon}} \\ &\le \frac{1025\log(11/p_\wedge)^2\|u\|^2_{L^\infty(\mu_{s,a})}}{\delta^4 p_\wedge^6}\left(\frac{4}{3}\right)^4 E \sup_{s,a\in S\times A}\sup_{y\in Y} m_{s,a,n}(y)^4 \\ &\le \frac{3240\log(11/p_\wedge)^2\|u\|^2_{L^\infty(\mu_{s,a})}}{\delta^4 p_\wedge^6}\frac{1}{2^{4n}} E \sup_{s,a,y}\left(\sum_{i=1}^{2^n}\mathbb{1}\left\{X^O_{s,a,i} = y\right\} - \mathbb{1}\left\{X^E_{s,a,i} = y\right\}\right)^4\end{aligned}$$

where $\left\{X^E_{s,a,i}, X^O_{s,a,i}, i = 1 \dots 2^n\right\}$ are independent samples from $\mu_{s,a}$ that forms $\mu_{s,a,n}(t)$. Recall that centered Bernoulli r.v.s. are $1/4$ sub-Gaussian; hence

$$Y_{s,a}(y) := \sum_{i=1}^{2^n} \mathbb{1}\left\{X^O_{s,a,i} = y\right\} - \mathbb{1}\left\{X^E_{s,a,i} = y\right\}$$

is $2^n/2$ sub-Gaussian. Using lemma D.2, we get that there exists constant c_1 s.t.

$$E_1 \leq \frac{c_1 \log(11/p_\wedge)^2 \|u\|^2_{L^\infty(\mu_{s,a})}}{\delta^4 p_\wedge^6 2^{2n}} (3 + \log(|S||A||Y|))^2 \tag{21}$$

Next, by Hölder's inequality we analyse separately

$$\begin{aligned} E_2 &\leq E\left[\sup_{s,a\in S\times A} \Delta^2_{s,a,n} \sup_{s,a\in S\times A} \mathbb{1}_{\Omega_{s,a,n}(p)^c}\right] \\ &\leq \left\|\sup_{s,a\in S\times A} \Delta^2_{s,a,n}\right\|_{L^\infty(P)} P\left(\bigcup_{s,a\in S\times A} \Omega_{s,a,n}(p)^c\right). \end{aligned}$$

Since the empirical measures are sampled from $\mu_{s,a}$, we always have that $\mu_{s,a,n}(t) \ll \mu_{s,a}$. By Lemma D.3, w.p.1.

$$\Delta^2_{s,a,n} \leq 2\|u\|^2_{L^\infty(\mu_{s,a})} + 2\|u\|^2_{L^\infty(\mu_{s,a})} \leq 4\|u\|^2_{L^\infty(\mu_{s,a})}$$

where we used Jensen's inequality $(a+b)^2 \leq 2a^2 + 2b^2$. Therefore the first term

$$\left\|\sup_{s,a\in S\times A} \Delta^2_{s,a,n}\right\|_{L^\infty(P)} \leq 4\|u\|^2_{L^\infty(\mu_{s,a})}.$$

For the second term

$$\begin{aligned} P\left(\bigcup_{s,a\in S\times A} \Omega_{s,a,n}(p)^c\right) &= P\left(\bigcup_{I=E,O} \bigcup_{s,a\in S\times A} \left\{\sup_{y\in Y} |\mu^I_{s,a,2^n}(y) - \mu_{s,a}(y)| > p\right\}\right) \\ &\leq 2P\left(\bigcup_{s,a\in S\times A} \left\{\sup_{y\in Y} |\mu^E_{s,a,2^n}(y) - \mu_{s,a}(y)| > p\right\}\right) \\ &= 2P\left(\sup_{s,a\in S\times A} \sup_{y\in Y} |\mu^E_{s,a,2^n}(y) - \mu_{s,a}(y)| > p\right) \\ &\leq \frac{2^9}{p_\wedge^4} E\left[\sup_{s,a,y} |\mu^E_{s,a,2^n}(y) - \mu_{s,a}(y)|^4\right] \\ &= \frac{2^9}{p_\wedge^4} \frac{1}{2^{4n}} E\left[\sup_{s,a,y} \left(\sum_{i=1}^{2^n} \mathbb{1}\left\{X^E_{s,a,i} = y\right\} - \mu_{s,a}(y)\right)^4\right] \end{aligned}$$

where the second last line follows from Markov's inequality. By Hoeffding's lemma, $\sum_{i=1}^{2^n} \mathbb{1}\left\{X^E_{s,a,i} = y\right\} - \mu_{s,a}(y)$ is $2^n/4$ sub-Gaussian. Therefore, by lemma D.2

$$P\left(\bigcup_{s,a\in S\times A} \Omega_{s,a,n}(p)^c\right) \leq \frac{2^9(3 + \log(|S||A||Y|))^2}{p_\wedge^4 2^{2n}}.$$

Recall (20). We conclude that there exists constant c s.t.

$$\begin{aligned} E \sup_{(s,a)\in S\times A} \Delta^2_{s,a,n} &\leq \frac{c_1 \log(11/p_\wedge)^2 \|u\|^2_{L^\infty(\mu_{s,a})}}{\delta^4 p_\wedge^6 2^{2n}} (3 + \log(|S||A||Y|))^2 + \frac{2^{11}(3 + \log(|S||A||Y|))^2 \|u\|_{L^\infty(\mu_{s,a})}}{p_\wedge^4 2^{2n}} \\ &\leq c \log(11/p_\wedge)^2 (3 + \log(|S||A||Y|))^2 \frac{\|u\|^2_{L^\infty(\mu_{s,a})}}{\delta^4 p_\wedge^6 2^{2n}} \end{aligned} \tag{22}$$

E Proof of Propositions

With the language and tools developed in the previous sections, we are ready to prove the claimed results

E.1 Proof of Proposition 3.1

Proof. We use the primal formulation of the distributionally robust Bellman operator

$$\mathcal{T}_\delta(Q)(s,a) = \inf_{\mu^R_{s,a},\mu^V_{s,a}\sim\delta} \mu^R_{s,a}[id] + \gamma\mu^V_{s,a}\left[v(Q)\right].$$

Then for $Q_1, Q_2 \in \mathbb{R}^{S\times A}$,

$$\begin{aligned}\mathcal{T}_\delta(Q_1)(s,a) - \mathcal{T}_\delta(Q_2)(s,a) &= \inf_{\mu^R_{s,a},\mu^V_{s,a}\sim\delta} \left(\mu^R_{s,a}[id] + \gamma\mu^V_{s,a}\left[v(Q_1)\right]\right)\\ &\quad + \sup_{\mu^R_{s,a},\mu^V_{s,a}\sim\delta}\left(-\mu^R_{s,a}[id] - \gamma\mu^V_{s,a}\left[v(Q_2)\right]\right)\\ &\le \sup_{\mu^V_{s,a}\sim\delta}\left(\gamma\mu^V_{s,a}\left[v(Q_1)\right] - \gamma\mu^V_{s,a}\left[v(Q_2)\right]\right)\\ &\le \gamma\sup_{s\in S}(v(Q_1)-v(Q_2)).\end{aligned}$$

Therefore,

$$\begin{aligned}\|\mathcal{T}_\delta(Q_1)-\mathcal{T}_\delta(Q_2)\|_\infty &\le \gamma \sup_{(s,a)\in S\times A}\max\left\{\mathcal{T}_\delta(Q_1)(s,a)-\mathcal{T}_\delta(Q_2)(s,a), \mathcal{T}_\delta(Q_2)(s,a)-\mathcal{T}_\delta(Q_1)(s,a)\right\}\\ &\le \sup_{(s,a)\in S\times A}\max\left\{\sup_{s\in S}(v(Q_1)-v(Q_2)), \sup_{s\in S}(v(Q_2)-v(Q_1))\right\}\\ &= \gamma\sup_{s\in S}|v(Q_1)-v(Q_2)|\\ &= \gamma\sup_{s\in S}\left|\sup_{b\in A}Q_1(s,b) - \sup_{b\in A}Q_2(s,b)\right|\\ &\le \gamma\sup_{s\in S}\sup_{b\in A}|Q_1(s,b)-Q_2(s,b)|\\ &\le \gamma\|Q_1-Q_2\|_\infty;\end{aligned}$$

i.e. $\mathcal{T}_\delta$ is a γ-contraction in $\|\cdot\|_\infty$. □

E.2 Proof of Proposition 4.1

Proof. Recall the dual formulation

$$\mathcal{T}_\delta(Q)(s,a) = \sup_{\alpha\ge 0} f(\mu^R_{s,a}, id, \alpha) + \gamma\sup_{\alpha\ge0} f(\mu^V_{s,a}, v(Q), \alpha)$$

Since Q^* is the fixed point of the distributionally robust Bellman operator. It follows from Lemma D.3 that

$$\begin{aligned}\|Q^*\|_\infty &= \|\mathcal{T}_\delta(Q^*)\|_\infty\\ &\le \|r\|_\infty + \gamma\|v(Q^*)\|_\infty\\ &\le r_{\max} + \gamma\|Q^*\|_\infty\end{aligned}$$

which implies the claimed result. □

E.3 Proof of Proposition 4.2

Proof. By construction,

$$\mathcal{T}_\delta(Q)(s,a) = E\left[R_{s,a,0}\right] + \sum_{n=0}^\infty E\left[\Delta^R_{s,a,n}\right] + \gamma E\left[v(Q)(S_{s,a,0})\right] + \gamma\sum_{n=0}^\infty E\left[\Delta^V_{s,a,n}\right]$$

If the sum and the integrals can be interchanged, then

$$\begin{aligned}\sum_{n=0}^{\infty} E\left[\Delta^R_{s,a,n}\right] &= E\left[\sum_{n=0}^{\infty}\frac{p_n}{p_n}\Delta^R_{s,a,n}\right]\\ &= E\left[\frac{\Delta^R_{s,a,N_1}}{p_{N_1}}\right].\end{aligned}$$

Similar results hold for $\hat{V}_\delta$, and hence the proposition. Therefore, it suffices to exchange the integrals. By Tonelli's theorem, a sufficient condition is that

$$E\left[\sum_{n=0}^{\infty}|\Delta^R_{s,a,n}|\right] < \infty.$$

Note that by Jensen's inequality,

$$\begin{aligned}E\left[\sum_{n=0}^{\infty}|\Delta^R_{s,a,n}|\right] &= E\left[\sum_{n=0}^{\infty}p_n\frac{|\Delta^R_{s,a,n}|}{p_n}\right]\\ &\le \sqrt{E\left[\sum_{n=0}^{\infty}p_n\frac{\left(\Delta^R_{s,a,n}\right)^2}{p_n^2}\right]}\\ &= \sqrt{E\left[\frac{\left(\Delta^R_{s,a,N_1}\right)^2}{p_{N_1}^2}\right]}\\ &\le \sqrt{E\sup_{s,a}\left(\frac{\Delta^R_{s,a,N_1}{}^2}{p_{N_1}^2}\right)}.\end{aligned}$$

We show that the quantity in the last line is indeed finite: by Tonelli's theorem, the definition of $\tilde{l}$ in (15), property (22), and the choose $g \in (0, 3/4)$:

$$\begin{aligned}E\left[\sup_{s,a}\left(\frac{\Delta^R_{s,a,N_1}}{p_{N_1}}\right)^2\right] &= \sum_{n=0}^{\infty}\frac{1}{p_n}E\sup_{s,a}\Delta^R_{s,a,N_1}{}^2\\ &\le c\log(11/p_\wedge)^2(3+\log(|S||A||R|))^2\frac{r_{\max}^2}{\delta^4 p_\wedge^6}\sum_{n=0}^{\infty}\frac{g}{(4-4g)^n}\\ &= c\log(11/p_\wedge)^2(3+\log(|S||A||R|))^2\frac{r_{\max}^2}{\delta^4 p_\wedge^6}\frac{4(1-g)}{g(3-4g)}\\ &\le \frac{c\tilde{l}r_{\max}^2}{\delta^4 p_\wedge^6}\end{aligned} \tag{23}$$

which is finite. Similarly, since $\|v(Q)\|_\infty \le \|Q\|_\infty$,

$$E\sup_{s,a}\left(\frac{\Delta^V_{s,a,N_2}}{p_{N_2}}\right)^2 \le \frac{c\tilde{l}\|Q\|_\infty^2}{\delta^4 p_\wedge^6} \tag{24}$$

is finite as well. This completes the proof □

E.4 Proof of Proposition 4.3

Proof. Recall bounds (23) and (24). We compute

$$\begin{aligned}
E\|W(Q)\|_\infty^2 &= E\left[\sup_{s,a}|\hat{\mathcal{T}}_\delta(Q)(s,a)-\mathcal{T}_\delta(Q)(s,a)|^2\right]\\
&\le 4r_{\max}^2+4E\sup_{s,a}\left(\frac{\Delta^R_{s,a,N_1}}{p_{N_1}}-E\frac{\Delta^R_{s,a,N_1}}{p_{N_1}}\right)^2\\
&\quad+8\gamma^2\|Q\|_\infty^2+4\gamma^2E\sup_{s,a}\left(\frac{\Delta^V_{s,a,N_2}}{p_{N_2}}-E\frac{\Delta^V_{s,a,N_2}}{p_{N_2}}\right)^2\\
&\le 4r_{\max}^2+8E\sup_{s,a}\left(\frac{\Delta^R_{s,a,N_1}}{p_{N_1}}\right)^2+8\sup_{s,a}\left(E\frac{\Delta^R_{s,a,N_1}}{p_{N_1}}\right)^2\\
&\quad+8\gamma^2\|Q\|_\infty^2+8\gamma^2E\sup_{s,a}\left(\frac{\Delta^V_{s,a,N_2}}{p_{N_2}}\right)^2+8\gamma^2\sup_{s,a}\left(E\frac{\Delta^V_{s,a,N_2}}{p_{N_2}}\right)^2\\
&\le 4r_{\max}^2+16E\left[\sup_{s,a}\left(\frac{\Delta^R_{s,a,N_1}}{p_{N_1}}\right)^2\right]+8\gamma^2\|Q\|_\infty^2+16\gamma^2E\sup_{s,a}\left(\frac{\Delta^V_{s,a,N_2}}{p_{N_2}}\right)^2\\
&\le\left(4+\frac{16c\tilde{l}}{\delta^4p_\wedge^6}\right)r_{\max}^2+\left(8+\frac{16c\tilde{l}}{\delta^4p_\wedge^6}\right)\gamma^2\|Q\|_\infty^2.
\end{aligned}$$

Since $\delta<\log 2$, $l>1$, and $p_\wedge\le 1/2$, we have

$$E\|W(Q)\|_\infty^2\le\frac{(16c+4)\tilde{l}}{\delta^4p_\wedge^6}r_{\max}^2+\frac{(16c+8)\tilde{l}}{\delta^4p_\wedge^6}\gamma^2\|Q\|_\infty^2.$$

replace $16c+8$ with c, we obtain the claimed result. □

F Proof of Technical Lemmas

F.1 Proof of Lemma D.1

Proof. Recall the definition (16) and (17). We fix (s,a) and write $\Delta_n=\Delta_{s,a,n}$, $\mu_n(t)=\mu_{s,a,n}(t)$. The following proof assumes that $\omega\in\Omega_{s,a,n}(p)$, so the equivalence (18) and the bound (19) hold.

From Si et al. (2020), it is sufficient to consider $\alpha\in[0,\delta^{-1}\|u\|_{L^\infty(\mu)}]=:K$. For $\alpha>0$ fixed,

$$\partial_t g_n(t,\alpha)=-\alpha\frac{m_n[w]}{\mu_n(t)[w]}.$$

Also, for $\alpha=0$, by (18), $g_n(t,0)\equiv\operatorname{ess\,inf}_\mu u$; hence $\partial_t g_n(t,0)\equiv 0$. Let $|m_n|(s)=|m_n(s)|$, again by (18), $\mu_n(t)(s)=0\iff\mu(s)=0\implies|m_n|(s)=0$; i.e. $|m_n|\ll\mu_n(t)$ and the Radon-Nikodym theorem applies. So, for fixed $t\in[0,1]$,

$$\begin{aligned}
\lim_{\alpha\downarrow 0}\sup_{s\in(t\pm\epsilon)\cap[0,1]}|\partial_t g_n(t,\alpha)| &\le\lim_{\alpha\downarrow 0}\sup_{t\in[0,1]}\alpha\left|\frac{m_n[w]}{\mu_n(t)[w]}\right|\\
&=\lim_{\alpha\downarrow 0}\sup_{t\in[0,1]}\alpha\left|\frac{1}{\mu_n(t)[w]}\mu_n(t)\left[\frac{dm_n}{d\mu_n(t)}w\right]\right|\\
&\le\lim_{\alpha\downarrow 0}\sup_{t\in[0,1]}\alpha\left\|\frac{dm_n}{d\mu_n(t)}\right\|_{L^\infty(\mu)}\\
&\le\lim_{\alpha\downarrow 0}\frac{\alpha}{p_\wedge}\\
&=0.
\end{aligned}$$

where we used Hölder's inequality to get the second last line. Therefore, $\partial_t g(\cdot,\cdot)$ is continuous on $[0,1]\times K$.

Next define

$$\Theta(t) := \arg\max_{\alpha\in K} g(t,\alpha).$$

We discuss two cases:

CASE 1: If u is μ-essentially constant, then

$$\sup_{\alpha\in K} -\alpha\log e^{-\bar{u}/\alpha}-\alpha\delta = \sup_{\alpha\in K}\bar{u}-\alpha\delta;$$

i.e. $\Theta(t)=\{0\}$.

CASE 2: u is not μ-essentially constant. Note that when $\alpha>0$, $w>0$; we can define a new measure

$$\mu_n^*(t)[\cdot]=\frac{\mu_n(t)[w\cdot]}{\mu_n(t)[w]}.$$

We have that

$$\begin{aligned}\partial_\alpha\partial_\alpha g_n(t,\alpha) &= -\frac{\mu_n(t)[u^2w]}{\alpha^3\mu_n(t)[w]}+\frac{\mu_n(t)[uw]^2}{\alpha^3\mu_n(t)[w]^2}\\ &= -\frac{\mu_n^*(t)[u^2]}{\alpha^3}+\frac{\mu_n^*(t)[u]^2}{\alpha^3}\\ &= -\frac{\mathsf{Var}_{\mu_n^*(t)}(u)}{\alpha^3}\\ &< 0;\end{aligned}$$

i.e. $g_n(t,\cdot)$ is strictly concave for $\alpha>0$. Also, recall that $g_n(t,\cdot)$ is continuous at 0. So, in this case either $\Theta(t)=\{0\}$ or $\Theta(t)=\{\alpha_n^*(t)\}$ where $\delta^{-1}\|u\|_{L^\infty(\mu)}\geq\alpha_n^*(t)>0$.

In particular, $\Theta(t)$ is a singleton which we will denote by $\alpha_n^*(t)$ in both cases. We conclude that by Shapiro et al. (2014) Theorem 7.21, the following derivative exists

$$d_t\sup_{\alpha\in K} g_n(t,\alpha)=\sup_{\alpha\in\Theta(t)}\partial_t g_n(t,\alpha)=\partial_t g_n(t,\alpha_n^*(t)).$$

Therefore, by the mean value theorem, there exists $t_1\in(0,1/2), t_2\in(1/2,1)$ depending on ω s.t.

$$\begin{aligned}\Delta_n &= \frac{1}{2}\left(\partial_t g_n(t_1,\alpha_n^*(t_1))-\partial_t g_n(t_2,\alpha_n^*(t_2))\right)\\ &= -\frac{1}{2}\left(\alpha_n^*(t_1)\frac{m_n[w_n^*(t_1)]}{\mu_n(t_1)[w_n^*(t_1)]}-\alpha_n^*(t_2)\frac{m_n[w_n^*(t_2)]}{\mu_n(t_2)[w_n^*(t_2)]}\right).\end{aligned}$$

where $w_n^*(t)=e^{-u/\alpha_n^*(t)}$. We will use w to denote $w_n^*(t)$ in the following derivations.

Again if u is μ-essentially constant, then $\Delta_n=0$. If not, then we consider the population optimizer, which, by the same reasoning, is also a singleton denoted by α^*. Let $\kappa_{s,a}=\mu_{s,a}(\{s:u(s)=\operatorname{ess\,inf}_{\mu_{s,a}}u\})$. There are two cases

1. $\alpha^*=0$. From Hu and Hong (2013b), $\alpha^*=0$ iff $\kappa_{s,a}\geq e^{-\delta}$. If we want $\alpha_n^*(t)=0$ for all $t\in[0,1]$, a sufficient condition is that $\kappa_{s,a,n}(t)\geq\kappa_{s,a}-p\geq e^{-\delta}$.
2. $\alpha^*>0$ iff $\kappa_{s,a}<e^{-\delta}$. If we want $\alpha_n^*(t)>0$ for all $t\in[0,1]$, a sufficient condition is that $\kappa_{s,a,n}(t)\leq\kappa_{s,a}+p<e^{-\delta}$.

Therefore, for any $e^{-\delta}\neq\{\kappa_{s,a}:(s,a)\in S\times A\}\subset\{\mu_{s,a}(y):(s,a,y)\in S\times A\times Y\}$. We can always choose p small enough s.t. for $\omega\in\Omega_{s,a,n}(p)$, $\alpha^*=0$ or $\alpha^*>0$ implies that $\alpha_n^*(t)=0$ or $\alpha_n^*(t)>0$ respectively.

Remark. While this generalizes to all but finitely many δ, for simplicity of presentation, we assume Assumption 1 that $p_\wedge/2\geq 1-e^{-\delta}$. This implies that if $\kappa_{s,a}\neq 1$, then $1-\kappa_{s,a}\geq p_\wedge>1-e^{-\delta}$; i.e. $\kappa_{s,a}<e^{-\delta}$ and case 1 cannot happen. Moreover, if we choose $p=\frac{1}{4}p_\wedge$, then

$$\kappa_{s,a}+p\leq 1-\frac{3}{4}p_\wedge<1-\frac{1}{2}p_\wedge\leq e^{-\delta}$$

satisfying the sufficient condition in case 2.

Therefore, for $\omega \in \Omega_{s,a,n}(p)$, there are two cases:

CASE 1: $\alpha^* = 0$, then $\Delta_n = 0$, Lemma D.1 holds trivially.

CASE 2: $\alpha^* > 0$, then $\alpha_n^*(t_1), \alpha_n^*(t_2) > 0$. Since $g_n(t, \cdot)$ is strictly convex, $\alpha_n^*(t)$ is the unique solution to the first order optimality condition

$$0 = \partial_\alpha g_n(t, \alpha_n^*(t)) = -\log \mu_n(t)[w] - \delta - \frac{\mu_n(t)[uw]}{\alpha_n^*(t)\mu_n(t)[w]}. \tag{25}$$

Note that $\partial_\alpha g_n \in C^\infty([0,1] \times \mathbb{R}_{++})$ and that $\partial_\alpha \partial_\alpha g_n(t, \alpha_n^*(t)) < 0$. The implicit function theorem implies that $\alpha_n^*(t) \in C^1((0,1))$ with derivative

$$\begin{aligned}
d_t\alpha_n^*(t) &= -\frac{\partial_t\partial_\alpha g_n(t,\alpha_n^*(t))}{\partial_\alpha\partial_\alpha g_n(t,\alpha_n^*(t))}\\
&= \left(\frac{\alpha_n^*(t)^3}{\mathsf{Var}_{\mu_n^*(t)}(u)}\right)\left(-\frac{m_n[w]}{\mu_n(t)[w]} + \frac{\mu_n(t)[uw]m_n[w]}{\alpha_n^*(t)\mu_n(t)[w]^2} - \frac{m_n[uw]}{\alpha_n^*(t)\mu_n(t)[w]}\right)\\
&= \left(\frac{\alpha_n^*(t)^3}{\mathsf{Var}_{\mu_n^*(t)}(u)}\right)\left(-\frac{m_n[w]}{\mu_n(t)[w]} + \frac{\mu_n(t)[uw/\alpha_n^*(t)]}{\mu_n(t)[w]^2}\mu_n(t)\left[\frac{dm_n}{d\mu_n(t)}w\right] - \frac{\mu_n(t)\left[\frac{dm_n}{d\mu_n(t)}uw/\alpha_n^*(t)\right]}{\mu_n(t)[w]}\right)\\
&= \left(\frac{\alpha_n^*(t)^3}{\mathsf{Var}_{\mu_n^*(t)}(u)}\right)\left(-\frac{m_n[w]}{\mu_n(t)[w]} + \mu_n^*(t)[u/\alpha_n^*(t)]\mu_n^*(t)\left[\frac{dm_n}{d\mu_n(t)}\right] - \mu_n^*(t)\left[\frac{dm_n}{d\mu_n(t)}u/\alpha_n^*(t)\right]\right)\\
&= -\left(\frac{\alpha_n^*(t)^3}{\mathsf{Var}_{\mu_n^*(t)}(u)}\right)\left(\frac{m_n[w]}{\mu_n(t)[w]} + \mathsf{Cov}_{\mu_n^*(t)}\left(\frac{u}{\alpha_n^*(t)}, \frac{dm_n}{d\mu_n(t)}\right)\right)
\end{aligned}$$

Therefore, we conclude that

$$\partial_t g_n(t, \alpha_n^*(t)) = -\alpha_n^*(t)\frac{m_n[w]}{\mu_n(t)[w]}$$

is $C^1((0,1))$ as a function of t with derivative

$$\begin{aligned}
-d_t\partial_t g_n(t,\alpha_n^*(t)) &= \alpha_n^*(t)\frac{m_n[w]^2}{\mu_n(t)[w]^2} - d_t\alpha_n^*(t)\left(\frac{m_n[w]}{\mu_n(t)[w]} + \frac{m_n[uw]}{\alpha_n^*(t)\mu_n(t)[w]} - \frac{m_n[w]\mu_n(t)[uw]}{\alpha_n^*(t)\mu_n(t)[w]^2}\right)\\
&= \alpha_n^*(t)\frac{m_n[w]^2}{\mu_n(t)[w]^2} + \left(\frac{\alpha_n^*(t)}{\mathsf{Var}_{\mu_n^*(t)}(u/\alpha_n^*(t))}\right)\left(\mu_n^*(t)\left[\frac{dm_n}{d\mu_n(t)}\right] + \mathsf{Cov}_{\mu_n^*(t)}\left(\frac{u}{\alpha_n^*(t)}, \frac{dm_n}{d\mu_n(t)}\right)\right)^2.
\end{aligned}$$

Therefore, by the mean value theorem, there exists $t_3 \in (t_1, t_2)$

$$\begin{aligned}
|\Delta_n| &= \frac{|t_1 - t_2|}{2}\left|d_t\partial_t g_n(t,\alpha_n^*(t))|_{t_3}\right|\\
&\le \left(\alpha_n^*(t) + \frac{\alpha_n^*(t)}{\mathsf{Var}_{\mu_n^*(t)}(u/\alpha_n^*(t))}\right)\frac{m_n[w]^2}{\mu_n(t)[w]^2} + \alpha_n^*(t)\mathsf{Var}_{\mu_n^*(t)}\left(\frac{dm_n}{d\mu_n(t)}\right)\Bigg|_{t_3}\\
&= \left(\alpha_n^*(t) + \frac{\alpha_n^*(t)}{\mathsf{Var}_{\mu_n^*(t)}(u/\alpha_n^*(t))}\right)\frac{m_n[w]^2}{\mu_n(t)[w]^2}\\
&\quad + \alpha_n^*(t)\frac{\mu_n(t)\left[\left(\frac{dm_n}{d\mu_n(t)}\right)^2 w\right]}{\mu_n(t)[w]} - \alpha_n^*(t)\frac{\mu_n(t)\left[\frac{dm_n}{d\mu_n(t)}w\right]^2}{\mu_n(t)[w]^2}\Bigg|_{t_3}\\
&= \frac{\alpha_n^*(t_3)}{\mathsf{Var}_{\mu_n^*(t_3)}(u/\alpha_n^*(t_3))}\frac{m_n[w]^2}{\mu_n(t_3)[w]^2} + \alpha_n^*(t_3)\mu_n^*(t_3)\left[\left(\frac{dm_n}{d\mu_n(t_3)}\right)^2\right]
\end{aligned} \tag{26}$$

Note that the log-likelihood ratio

$$\log\left(\frac{w}{\mu_n(t)[w]}\right) = -\frac{u}{\alpha_n^*(t)} - \log\mu_n(t)[w].$$

Moreover, by the optimality condition (25),

$$\begin{aligned}\mu_n^*(t)\left[\log\left(\frac{w}{\mu_n(t)[w]}\right)\right] &= \mu_n^*(t)\left[-\frac{u}{\alpha_n^*(t)} - \log\mu_n(t)[w]\right] \\ &= \mu_n^*(t)\left[-\frac{u}{\alpha_n^*(t)}\right] + \delta + \frac{\mu_n(t)[uw]}{\alpha_n^*(t)\mu_n(t)[w]} \\ &= -\mu_n^*(t)\left[\frac{u}{\alpha_n^*(t)}\right] + \delta + \mu_n^*(t)\left[\frac{u}{\alpha_n^*(t)}\right] \\ &= \delta.\end{aligned}$$

So, the variance:

$$\begin{aligned}\mathsf{Var}_{\mu_n^*(t)}(u/\alpha_n^*(t)) &= \mathsf{Var}_{\mu_n^*(t)}\left(\log\left(\frac{w}{\mu_n(t)[w]}\right)\right) \\ &= \mu_n^*(t)\left[\log\left(\frac{d\mu_n^*(t)}{d\mu_n(t)}\right)^2\right] - \delta^2.\end{aligned}$$

We bound this expression by the following lemma:

Lemma F.1. *For measures μ, μ' s.t. $D_{KL}(\mu'||\mu) = \delta$ and $\bar{p}_\wedge = \inf_{s\in S}\mu(s) = 1 - e^{-\delta-\psi_\wedge}$ for some $\psi_\wedge > 0$ we have that*

$$\mu'\left[\log\left(\frac{d\mu'}{d\mu}\right)^2\right] - \delta^2 \geq -\frac{\delta^2\psi_\wedge}{8\log(\psi_\wedge/8)}$$

Recall that we choose $p = \frac{1}{4}p_\wedge$, so we should choose $\psi_\wedge$

$$\bar{p}_\wedge = 1 - e^{-\delta-\psi_\wedge} = \inf_{s:\mu(s)>0}\mu_n(t)(s) \geq \frac{3}{4}p_\wedge.$$

We want the above bound to hold uniformly in δ:

$$\frac{3}{4}p_\wedge \leq \inf_{\delta\geq 0} 1 - e^{-\delta-\psi_\wedge} = 1 - e^{-\psi_\wedge}$$

So,

$$\psi_\wedge \geq -\log\left(1 - \frac{3}{4}p_\wedge\right) \geq \frac{3}{4}p_\wedge.$$

We conclude that by Lemma F.1,

$$\begin{aligned}\mathsf{Var}_{\mu_n^*(t)}(u/\alpha_n^*(t)) &= \mu_n^*(t)\left[\log\left(\frac{d\mu_n^*(t)}{d\mu_n(t)}\right)^2\right] - \delta^2 \\ &\geq -\frac{3}{32}\frac{\delta^2 p_\wedge}{\log(3p_\wedge/32)}\end{aligned}$$

Note that $-x/\log x = O(x^{1+\epsilon})$ as $x \downarrow 0$ for any $\epsilon > 0$.

Next we go back to bounding Δ_n in case 2.

Lemma F.2. *For $\delta \leq \log 2/2$ and $\omega \in \Omega_{s,a,n}(\frac{1}{4}p_\wedge)$, we have*

$$\sup_{\alpha\in K}\frac{\alpha m_n[w]^2}{\mu_n(t_3)[w]^2} \leq 3\|u\|_{L^\infty(\mu_{s,a})}\left\|\frac{dm_n}{d\mu_n(t_3)}\right\|^2_{L^\infty(\mu)}.$$

We conclude that from (26) and Lemma F.2

$$\begin{aligned}
\Delta_n^2 \mathbb{1}_{\Omega_{s,a,n}(p)} &\leq \left(\frac{2^{10}\log(\frac{32}{3p_\wedge})^2}{9\delta^4 p_\wedge^2} \frac{\alpha_n^*(t_3)^2 m_n[w]^4}{\mu_n(t_3)[w]^4} + \alpha_n^*(t_3)^2 \mu_n^*(t_3)\left[\left(\frac{dm_n}{d\mu_n(t_3)}\right)^2\right]^2 \right) \mathbb{1}_{\Omega_{s,a,n}(p)} \\
&\leq \left(\frac{2^{10}\log(11/p_\wedge)^2}{9\delta^4 p_\wedge^2} \left(\sup_{\alpha\in K} \frac{\alpha m_n[w]^2}{\mu_n(t_3)[w]^2}\right)^2 + \left(\sup_{\alpha\in K} \alpha\mu_n^*(t_3)\left[\left(\frac{dm_n}{d\mu_n(t_3)}\right)^2\right]\right)^2 \right) \mathbb{1}_{\Omega_{s,a,n}(p)} \\
&\leq \left(\frac{2^{10}\log(11/p_\wedge)^2}{\delta^4 p_\wedge^2} \|u\|_{L^\infty(\mu_{s,a})}^2 \left\|\frac{dm_n}{d\mu_n(t_3)}\right\|_{L^\infty(\mu)}^4 + \frac{\|u\|_\infty^2}{\delta^2}\left\|\frac{dm_n}{d\mu_n(t_3)}\right\|_{L^\infty(\mu)}^4 \right) \mathbb{1}_{\Omega_{s,a,n}(p)} \\
&\leq \frac{1025\log(11/p_\wedge)^2}{\delta^4 p_\wedge^2} \|u\|_{L^\infty(\mu_{s,a})}^2 \left\|\frac{dm_n}{d\mu_n(t_3)}\right\|_{L^\infty(\mu)}^4 \mathbb{1}_{\Omega_{s,a,n}(p)}.
\end{aligned}$$

This completes the proof of Lemma D.1. □

F.1.1 Proof of Lemma F.1

Proof. Consider the program for the probability measure ν defined on the positive entries of μ, i.e., $\nu = \{\nu(s) \geq 0 \text{ for } \mu(s) > 0, s \in S; \nu(s) = 0 \text{ for } \mu(s) = 0, s \in S; \sum_{s\in S}\nu(s) = 1\}$:

$$OPT_1 = \inf_{\nu: D_{\mathrm{KL}}(\nu||\mu) = \delta} \nu\left[\log\left(\frac{d\nu}{d\mu}\right)^2\right].$$

We first show that any feasible ν must satisfy $\mu \sim \nu$; i.e. if $\mu(s) > 0$ then $\nu(s) > 0$. Suppose, to the contrary that there exists $A \in \mathcal{S}$ s.t. $\mu(A^c) > 0$ but $\nu(A^c) = 0$. Also, since $\mu \gg \nu$, $\nu(A) = 1$ implies that $\mu(A) > 0$. We can define $\mu_A(\cdot) = \mu(A \cap \cdot)/\mu(A)$ the conditional measure.

$$\begin{aligned}
\delta &= D_{\mathrm{KL}}(\nu||\mu) \\
&= \nu\left[\log\left(\frac{d\nu}{d\mu}\right)\right] \\
&= \mu\left[\mathbb{1}_A \frac{d\nu}{d\mu}\log\left(\frac{d\nu}{d\mu}\right)\right] \\
&= \mu(A)\mu_A\left[\frac{d\nu}{d\mu}\log\left(\frac{d\nu}{d\mu}\right)\right]
\end{aligned}$$

Note that the function $x \to x\log x$ is convex for $x \geq 0$. We have that by Jensen's inequality

$$\begin{aligned}
\delta &\geq \mu(A)\mu_A\left[\frac{d\nu}{d\mu}\right]\log\left(\mu_A\left[\frac{d\nu}{d\mu}\right]\right) \\
&= \mu\left[\mathbb{1}_A\frac{d\nu}{d\mu}\right]\log\left(\frac{1}{\mu(A)}\mu\left[\mathbb{1}_A\frac{d\nu}{d\mu}\right]\right) \\
&= -\log\left(\mu(A)\right)
\end{aligned}$$

where the last inequality follows from the assumption that $\nu(A^c) = 0$. Since $\mu(A) \leq 1 - \bar{p}_\wedge$, the above inequality implies that $\delta \geq \delta + \psi_\wedge$, which is a contradiction.

This implies that the inequality constraint $\nu(s) \geq 0$ in OPT_1 is never active. Therefore, we can use the Lagrangian

$$\mathcal{L}(\nu, \lambda, \theta) = \nu\left[\log\left(\frac{d\nu}{d\mu}\right)^2\right] - \lambda\nu\left[\log\left(\frac{d\nu}{d\mu}\right)\right] + \lambda\delta - \theta\nu[1] + \theta.$$

Observe that for any feasible ν, $\partial_\nu D_{\mathrm{KL}}(\nu||\mu) = 1 + \log(d\nu/d\mu)$, $\partial_\nu\nu[1] = 1$ are never linearly dependent. So, any feasible point is a regular point of the equality constraints. Therefore by Chapter 11.3 in Luenberger et al. (2021), the KKT conditions are necessarily satisfied.

We take derivative

$$\partial_\nu \mathcal{L}(\nu, \lambda, \theta) = \log\left(\frac{d\nu}{d\mu}\right)^2 + 2\log\left(\frac{d\nu}{d\mu}\right) - \lambda\log\left(\frac{d\nu}{d\mu}\right) - \lambda - \theta.$$

If we define $l = \log\left(\frac{d\nu}{d\mu}\right)$, the KKT condition implies that

$$l^2 + (2 - \lambda^*)l - \lambda^* - \theta^* = 0 \implies l_\pm = \frac{1}{2}\left(\lambda^* - 2 \pm \sqrt{(\lambda^* - 2)^2 + 4(\lambda^* + \theta^*)}\right).$$

Since $\nu[1] = 1$, we must have that $l_+ \geq 0$ and $l_- \leq 0$. Define $S_+ = \{s \in S : l(s) = l_+\}$ We can write $\nu(s) = [e^{l_+}\mathbb{1}_{S_+}(s) + e^{l_-}\mathbb{1}_{S_+^c}(s)]\mu(s)$. Note that if $\mu(S_+) = 0, 1$, then $\mu = \nu$ violating the constraint. So, $\mu(S_+) \neq 0, 1$. Restrict μ, ν on $\mathcal{G} = \sigma(\{S_+, S_-\})$, then

$$l = \log\left(\frac{\nu(S_+)}{\mu(S_+)}\right)\mathbb{1}\{s \in S_+\} + \log\left(\frac{\nu(S_+^c)}{\mu(S_+^c)}\right)\mathbb{1}\left\{s \in S_+^c\right\} = \log\left(\frac{d\nu|_\mathcal{G}}{d\mu|_\mathcal{G}}\right).$$

and

$$OPT_1 = \nu|_\mathcal{G}\left[\log\left(\frac{d\nu|_\mathcal{G}}{d\mu|_\mathcal{G}}\right)^2\right].$$

Also under this notation,

$$D_{\mathrm{KL}}(\nu|_\mathcal{G}||\mu|_\mathcal{G}) = l_+\nu(S_+) + l_-\nu(S_-) = \nu\left[\log\left(\frac{d\nu}{d\mu}\right)\right] = \delta.$$

Therefore,

$$OPT_1 \geq \inf_{\mathcal{G}=\sigma(\{A,A^c\}):A\in\mathcal{S}} \quad \inf_{\eta|_\mathcal{G}:D_{\mathrm{KL}}(\eta|_\mathcal{G}||\mu|_\mathcal{G})=\delta} \eta|_\mathcal{G}\left[\log\left(\frac{d\eta|_\mathcal{G}}{d\mu|_\mathcal{G}}\right)^2\right].$$

We define:

$$OPT_2 := \inf_{\mathcal{G}=\sigma(\{A,A^c\}):A\in\mathcal{S}} \inf_{\eta|_\mathcal{G}:D_{\mathrm{KL}}(\eta|_\mathcal{G}||\mu|_\mathcal{G})=\delta} \eta|_\mathcal{G}\left[\log\left(\frac{d\eta|_\mathcal{G}}{d\mu|_\mathcal{G}}\right)^2\right].$$

As mentioned before, $A \neq S, \varnothing$ because of the constraint.

Next, we lower bound OPT_2. Notice that measureable strict subsets under $\mathcal{G}$ is only A and A^c. So, suppose $\eta|_\mathcal{G}(A) = q$, we consider the following program

$$\begin{aligned} \inf_{0<q<1} \quad & obj_2(q, b) := q\log\left(\frac{q}{b}\right)^2 + (1-q)\log\left(\frac{1-q}{1-b}\right)^2 \\ s.t. \quad & kl(q, b) := q\log\left(\frac{q}{b}\right) + (1-q)\log\left(\frac{1-q}{1-b}\right) = \delta \end{aligned}$$

where $b = \mu(S_+)$. This lower bounds OPT_2. The rest of the proof is denoted to compute the above program.

Note that w.l.o.g. we can assume $b \leq 1/2$ because if $b > 1/2$, we change to new variable $b' = 1 - b' < 1/2$ and $q' = 1 - q$. Compute the second derivatives

$$\begin{aligned} d_q d_q kl(q, b) &= \frac{1}{q} + \frac{1}{1-q} > 0; \\ d_b d_b kl(q, b) &= \frac{q}{b^2} + \frac{1-q}{(1-b)^2} > 0; \end{aligned}$$

i.e. $kl(\cdot, b), kl(q, \cdot)$ are convex on $[0, 1]$. So, its maximum is attained on the boundary: $kl(1, b) = -\log b$ and $kl(0, b) = -\log(1-b)$. By assumption, $1 - e^{-\delta} < \bar{p}_\wedge \leq b = \mu(S_+) \leq 1 - \bar{p}_\wedge < e^{-\delta}$. So, $kl(q, b) = \delta$ has 2 solutions $q_1 < b < q_2$. Moreover

$$\begin{aligned} \log\left(\frac{q_i}{b}\right) - \log\left(\frac{1-q_i}{1-b}\right) &= \frac{1}{q_i}\left(kl(q_i, b) - \log\left(\frac{1-q_i}{1-b}\right)\right) = \frac{1}{q_i}\left(\delta + \log\left(\frac{1-b}{1-q_i}\right)\right) \\ &= \frac{1}{1-q_i}\left(\log\left(\frac{q_i}{b}\right) - kl(q_i, b)\right) = -\frac{1}{1-q_i}\left(\delta + \log\left(\frac{b}{q_i}\right)\right) \end{aligned}$$

So, we have

$$\begin{aligned} OPT_2 - \delta^2 &\geq \inf_{i=1,2} obj_2(q_i) - kl(q_i,b)^2 \\ &= \inf_{i=1,2} q_i(1-q_i)\left(\log\left(\frac{q_i}{b}\right) - \log\left(\frac{1-q_i}{1-b}\right)\right)^2 \\ &= \inf_{i=1,2}\max\left(\frac{(1-q_i)}{q_i}\left(\delta+\log\left(\frac{1-b}{1-q_i}\right)\right)^2, \frac{q_i}{(1-q_i)}\left(\delta+\log\left(\frac{b}{q_i}\right)\right)^2\right) \end{aligned}$$

Observe that if $q > b$, then $1-q < 1-b$. So,

$$OPT_2 - \delta^2 \geq \delta^2 \inf_{i=1,2} \frac{1-q_i}{q_i}\mathbb{1}\{q > b\} + \frac{q_i}{1-q_i}\mathbb{1}\{q \leq b\}.$$

Now if we let $1/2 \geq b(\psi) = 1 - e^{-\delta-\psi}$. By convexity and $kl \geq 0$ while $kl(b(\psi), b(\psi)) = 0$, $kl(\cdot, b(\psi))$ is decreasing for $q \in [0, b]$.

$$\begin{aligned} kl(q, b(\psi)) &= (1-q)(\delta+\psi) - q\log(1-e^{-\delta-\psi}) + q\log q + (1-q)\log(1-q) \\ &= (1-q)(\delta+\psi) + q\xi\log(\delta+\psi) + q\log q + (1-q)\log(1-q) \end{aligned}$$

where we use $1 - e^{-\delta-\psi} = e^0 - e^{-\delta-\psi} = (\delta+\psi)e^{-\xi}$ for some $\xi \in (0, \delta+\psi)$. Also, $d_q(1-q)\log(1-q) = -1-\log(1-q) \geq -1$

$$\begin{aligned} kl(q,b) &\geq \delta + \psi + q(\eta\log(\delta+\psi) - \delta - \psi - 1) + q\log q \\ &\geq \delta + \psi + q((\delta+\psi)\log(\delta+\psi) - \delta - \psi - 1) + q\log q \\ &\geq \delta + \psi + q(-1-\delta-\psi-1) + q\log q \\ &\geq \delta + \psi - 4q + q\log q \end{aligned}$$

We let $q(\psi) = -(\psi/c)/\log(\psi/c)$ for some $c > 1$. Note that when $\psi = 0$, $\psi - 4q(\psi) + q(\psi)\log q(\psi) = 0$. We want $kl(q(\psi), b(\psi)) > \delta$; hence it suffices to show that

$$d_\psi(\psi - 4q(\psi) + q(\psi)\log q(\psi)) \geq 0.$$

Let $\phi = -\log(\psi/c)$ We compute

$$\begin{aligned} d_\psi(\psi - 4q(\psi) + q(\psi)\log q(\psi)) &= \frac{\log\left(\frac{\psi}{c\phi}\right)}{c\phi^2} - \frac{3}{c\phi^2} + \frac{\log\left(\frac{\psi}{c\phi}\right)}{c\phi} - \frac{3}{c\phi} + 1 \\ &= \frac{1}{c\phi^2}\left(c\phi^2 + (\phi+1)\log\left(\frac{\psi}{c\phi}\right) - 3\phi - 3\right) \\ &= \frac{1}{c\phi^2}\left((c-1)\phi^2 - 4\phi - (\phi+1)\log(\phi) - 3\right) \end{aligned}$$

Note that $c > e^{-1}$ and $\psi \in (0,1]$ implies that $\phi > 1$, $\log\phi \leq \phi - 1$. Therefore,

$$d_\psi(\psi - 4q + q\log q) \geq \frac{1}{c\phi^2}\left((c-1)\phi^2 - 4\phi - (\phi+1)(\phi-1) - 3\right) = \frac{1}{c\phi^2}\left((c-2)\phi^2 - 4\phi - 2\right)$$

We see that if we choose $c = 8$, $d_\psi(\psi - 4q + q\log q) > 0$. Hence $kl(q(\psi), b(\psi)) \geq \delta$ for all ψ. Continuity and monotonicity imply that $q(\psi) < b(\psi)$ and there exists $q_1(\psi) \in [q(\psi), b(\psi)]$ s.t. $kl(q_1(\psi), b(\psi)) = \delta$.

By the same argument, we can show that for $b'(\psi) = e^{-\delta-\psi}$, we have that $q_2(\psi) \in [b'(\psi), 1 - q(\psi)]$. Therefore, we conclude that for $1 - e^{-\delta-\psi} \leq b \leq e^{-\delta-\psi}$

$$OPT_2 - \delta^2 \geq -\delta^2\frac{\psi/8}{\log(\psi/8)}.$$

Recall that $\bar{p}_\wedge = 1 - e^{-\delta - \psi_\wedge}$ and that $p_\wedge \le \mu(S_+) \le 1 - \bar{p}_\wedge$ we conclude that

$$\begin{aligned} \mu'\left[\log\left(\frac{d\mu'}{d\mu}\right)^2\right] - \delta^2 &\ge \inf_{b\in[\bar{p}_\wedge, 1-\bar{p}_\wedge]} OPT_2 - \delta^2 \\ &\ge -\frac{\delta^2\psi_\wedge}{8\log(\psi_\wedge/8)}. \end{aligned}$$

□

F.1.2 Proof of Lemma F.2

Proof. Recall that for $\omega \in \Omega_{s,a,n}(p)$, we have (18). Write

$$\begin{aligned} \sup_{\alpha\in K}\frac{\alpha m_n[w]^2}{\mu_n(t_3)[w]^2} &= \max\left\{\sup_{\alpha\in[0,c\|u\|_\infty]}\frac{\alpha m_n[w]^2}{\mu_n(t_3)[w]^2}, \sup_{\alpha\in[c\|u\|_\infty, \delta^{-1}\|u\|_\infty]}\frac{\alpha m_n[w]^2}{\mu_n(t_3)[w]^2}\right\} \\ &=: \max\{J_1(c), J_2(c)\} \end{aligned}$$

We first bound $J_2(c)$

$$J_2(c) = \sup_{\alpha\in[c\|u\|_{L^\infty(\mu_{s,a})}, \delta^{-1}\|u\|_{L^\infty(\mu_{s,a})}]}\frac{\alpha m_n[e^{-(u+\|u\|_{L^\infty(\mu_{s,a})})/\alpha}]^2}{\mu_n(t_3)[e^{-(u+\|u\|_{L^\infty(\mu_{s,a})})/\alpha}]^2}$$

For simplicity, let $w' := e^{-(u+\|u\|_{L^\infty(\mu_{s,a})})/\alpha}$. Recall that $m_n = \mu^O - \mu^E$, so $m_n[1] = 0$ and

$$\alpha m_n[e^{-(u+\|u\|_{L^\infty(\mu_{s,a})})/\alpha}]^2 = (m_m[\alpha^{1/2}(e^{-(u+\|u\|_{L^\infty(\mu_{s,a})})/\alpha} - 1)])^2.$$

Define and note that $v := \alpha^{1/2}(e^{-(u+\|u\|_{L^\infty(\mu_{s,a})})/\alpha} - 1) < 0$. Then

$$\begin{aligned} \frac{\alpha m[w']^2}{\mu_n(t_3)[w']^2} &= \frac{m[v]^2}{\mu_n(t_3)[w']^2} \\ &= \frac{1}{\mu_n(t_3)[w']^2}\mu_n(t_3)\left[\frac{dm_n}{d\mu_n(t_3)}v\right]^2 \\ &\le \frac{\mu_n(t_3)\left[-v\right]^2}{\mu_n(t_3)[w']^2}\left\|\frac{dm_n}{d\mu_n(t_3)}\right\|_{L^\infty(\mu)}^2 \\ &\le \left\|\frac{v}{w'}\right\|_{L^\infty(\mu)}^2\left\|\frac{dm_n}{d\mu_n(t_3)}\right\|_{L^\infty(\mu)}^2 \end{aligned}$$

We defer the proof of the following claim:

Lemma F.3.

$$\sup_{\alpha\in[c\|u\|_{L^\infty(\mu_{s,a})}, \delta^{-1}\|u\|_{L^\infty(\mu_{s,a})}]}\left\|\frac{v}{w'}\right\|_{L^\infty(\mu)} \le (c\|u\|_{L^\infty(\mu_{s,a})})^{1/2}(e^{2/c} - 1)$$

Therefore,

$$J_2(c) \le c\|u\|_{L^\infty(\mu_{s,a})}(e^{2/c} - 1)^2\left\|\frac{dm_n}{d\mu_n(t_3)}\right\|_{L^\infty(\mu)}^2.$$

Assuming that $\delta \le \log 2/2$, choose $c = 2/\log 2$

$$\begin{aligned} \sup_{\alpha\in K}\frac{\alpha m_n[w]^2}{\mu_n(t_3)[w]^2} &= \max\{J_1(c), J_2(c)\} \\ &\le \max\left\{c\|u\|_\infty, c\|u\|_{L^\infty(\mu_{s,a})}(e^{2/c} - 1)^2\right\}\left\|\frac{dm_n}{d\mu_n(t_3)}\right\|_{L^\infty(\mu)}^2 \\ &\le 3\|u\|_{L^\infty(\mu_{s,a})}\left\|\frac{dm_n}{d\mu_n(t_3)}\right\|_{L^\infty(\mu)}^2 \end{aligned}$$

which completes the proof. □

F.1.3 Proof of Lemma F.3

Proof. We bound

$$\begin{aligned}\left\|\frac{v}{w'}\right\|_{L^\infty(\mu)} &= \operatorname*{ess\,sup}_{\mu} \alpha^{1/2}(e^{(u(s)+\|u\|_{L^\infty(\mu_{s,a})})/\alpha}-1)\\ &\le \alpha^{1/2}(e^{2\|u\|_{L^\infty(\mu_{s,a})}/\alpha}-1)\end{aligned}$$

Compute derivative: let $\beta = 2\|u\|_{L^\infty(\mu_{s,a})}/\alpha$

$$\begin{aligned}\frac{d}{d\alpha}\alpha^{1/2}(e^{2\|u\|_{L^\infty(\mu_{s,a})}/\alpha}-1) &= \frac{e^{2\|u\|_{L^\infty(\mu_{s,a})}/\alpha}-1}{2\alpha^{1/2}} - \frac{2\|u\|_{L^\infty(\mu_{s,a})}e^{2\|u\|_{L^\infty(\mu_{s,a})}/\alpha}}{\alpha^{3/2}}\\ &= \frac{1}{2}(e^{\beta}(1-2\beta)-1)\alpha^{-1/2}\end{aligned}$$

Notice that when $\beta = 0$, $e^\beta(1-2\beta)-1 = 0$. Moreover,

$$\frac{d}{d\beta}e^{\beta}(1-2\beta) = -e^{\beta}(1+2b) < 0;$$

i.e. $e^\beta(1-2\beta)$ decreasing. Therefore, for $\alpha > 0$

$$\frac{d}{d\alpha}\alpha^{1/2}(e^{2\|u\|_{L^\infty(\mu_{s,a})}/\alpha}-1) < 0;$$

i.e. $\alpha^{1/2}(e^{2\|u\|_{L^\infty(\mu_{s,a})}/\alpha}-1)$ is decreasing in α. Hence

$$\begin{aligned}\sup_{\alpha\in[c\|u\|_{L^\infty(\mu_{s,a})},\delta^{-1}\|u\|_{L^\infty(\mu_{s,a})}]}\left\|\frac{v}{w'}\right\|_{L^\infty(\mu)} &\le \sup_{\alpha\in[c\|u\|_{L^\infty(\mu_{s,a})},\delta^{-1}\|u\|_{L^\infty(\mu_{s,a})}]}\alpha^{1/2}(e^{2\|u\|_{L^\infty(\mu_{s,a})}/\alpha}-1)\\ &= (c\|u\|_{L^\infty(\mu_{s,a})})^{1/2}(e^{2/c}-1)\end{aligned}$$

establishing the claim. □

F.2 Proof of Lemma D.2

Proof. For any $\lambda > 0$, consider an increasing function $\phi_\lambda(z) = \exp(\lambda z^{1/4})$ for $z \ge 0$. Since $Z \ge 0$,

$$\begin{aligned}\phi_\lambda(EZ) &= \phi_\lambda(EZ\mathbb{1}\left\{Z > (3/\lambda)^4\right\} + EZ\mathbb{1}\left\{Z \le (3/\lambda)^4\right\})\\ &\le \phi_\lambda(EZ\mathbb{1}\left\{Z > (3/\lambda)^4\right\} + (3/\lambda)^4P(Z \le (3/\lambda)^4))\\ &\le \phi_\lambda(EZ + (3/\lambda)^4)\end{aligned}$$

By taking second derivatives, one can see that $\phi_\lambda(z)$ is convex for $z \ge (3/\lambda)^4$. Therefore, by Jensen's inequality

$$\begin{aligned}\phi_\lambda(EZ) &\le E\left[\phi_\lambda(Z+(3/\lambda)^4)\right]\\ &= e^3E\left[\exp(\lambda\mathbb{1}\max_{i=1\dots n}|Y_i|)\right]\\ &\le e^3\sum_{i=1}^{n}Ee^{\lambda|Y_i|}\end{aligned}$$

Since $\{Y_i\}$ are Sub-Gaussian,

$$P(|Y_i| > t) \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

which implies

$$\log Ee^{\lambda|Y_i|} \le 4\sigma^2\lambda^2.$$

Therefore,

$$\begin{aligned}\log\phi_\lambda(EZ) &= \lambda\left(E\max_{i=1\dots n}Y_i^4\right)^{1/4}\\ &\le 3+\log n+4\sigma^2\lambda^2.\end{aligned}$$

Rearrange and take infimum over $\lambda > 0$, we conclude

$$\begin{aligned} E \max_{i=1\dots n} Y_i^4 &\leq \left(\inf_{\lambda>0} \frac{3+\log n}{\lambda} + 4\sigma^2\lambda \right)^4 \\ &\leq 16\sigma^4 \left(3 + \log n\right)^2 \end{aligned}$$

□

F.3 Proof of Lemma D.3

Proof.

$$\sup_{\alpha\geq 0} f(\nu,\alpha) \geq \lim_{\alpha\downarrow 0} f(\nu,\alpha) = \operatorname*{ess\,inf}_{\nu} u \geq \operatorname*{ess\,inf}_{\mu} u \geq -\|u\|_{L^\infty(\mu)}$$

On the other hand, since the sup is achieved on compact K. For optimal $\alpha_\nu^* > 0$,

$$\begin{aligned} \sup_{\alpha\geq 0} f(\nu,\alpha) &\leq \|u\|_{L^\infty(\nu)} - \alpha_\nu^* \log \nu[e^{-(u-\|u\|_{L^\infty(\nu)})/\alpha_\nu^*}] \\ &\leq \|u\|_{L^\infty(\mu)} \end{aligned}$$

where the last line follows from that $\nu[e^{-(u-\|u\|_{L^\infty(\nu)})/\alpha_\nu^*}] > 0$ and $\nu \ll \mu$. Also, if $\alpha_\nu^* = 0$, the above holds trivially. □

G Numerical Experiment

G.1 Test of Convergence on the Hard MDP

For this numerical experiment using MDP 1, we run the algorithm to produce independent 200 trajectories of 5000 iterations of Algorithm 1 under the rescaled linear and constant stepsize. Denote the estimated Q-function under rescaled linear ($\alpha_k = 1/(1+(1-\gamma)k)$) and constant stepsize ($\alpha = 0.008$) with $\hat{Q}$ and $\bar{Q}$ respectively. For each $\gamma = 0.7, 0.8, 0.9$, we produce $\left\{\hat{Q}_{\delta,k}^{(i)}(\gamma), k = 0,\dots,5000\right\}$ and $\left\{\bar{Q}_{\delta,k}^{(i)}(\gamma), k = 0,\dots,5000\right\}$ for $i = 1,2,\dots,200$. We use the size of the uncertainty set $\delta = 0.1$.

Figure 2a plots the lines that linearly interpolates

$$\left\{ \left(\lg k, \lg \left(\frac{1}{200} \sum_{i=1}^{200} \left\| \hat{Q}_{\delta,k}^{(i)}(\gamma) - Q_\delta^*(\gamma) \right\|_\infty \right) \right); k = 0,1,\dots,5000 \right\}$$

for $\gamma = 0.7, 0.8, 0.9$ as well as a line of slope $-1/2$ in the lg-lg scale as a reference. Figure 2b plots the lines that linearly interpolates

$$\left\{ \left(k, \lg \left(\frac{1}{200} \sum_{i=1}^{200} \left\| \bar{Q}_{\delta,k}^{(i)}(\gamma) - Q_\delta^*(\gamma) \right\|_\infty \right) \right); k = 0,1,\dots,5000 \right\}.$$

G.2 Test of Convergence for different δ

We also numerically explore the complexity behavior when we change δ. We use the same plotting procedure as in Figure 2a only with different values of δ. This give us Figure 5. We observe that indeed as $\delta \downarrow 0$, the complexity is not sensitive to a change in δ. This confirms our conjecture that the dependence on δ should be $O(1)$ as $\delta \downarrow 0$. Also interestingly, as we increase δ, the complexity also becomes insensitive.

G.3 Test of γ Dependence on the Hard MDP

For Figure 3, we run 200 trajectories Algorithm 1 with rescaled linear stepsize $\alpha_k = 1/(1+(1-\gamma)k)$ for 10 evenly spaced $\gamma_j \in [0.7, 0.95], \gamma_1 = 0.7, \gamma_{10} = 0.95$ until a fixed iteration $k = 500, 1000, 1500$. For each This produce a data set $\left\{\hat{Q}_{\delta,k}^{(i)}(\gamma_j); i = 1,2\dots,200; j = 1,2,\dots,10\right\}$ for $k = 500, 1000, 1500$. We still use the size of the uncertainty set $\delta = 0.1$. Figure 3 plots the scattered pairs

$$\left\{ \left(\lg(1-\gamma_j), \lg \left(\frac{1}{200} \sum_{i=1}^{200} \left\| \hat{Q}_{\delta,k}^{(i)}(\gamma_j) - Q_\delta^*(\gamma_j) \right\|_\infty \right) \right); j = 1,2,\dots,10 \right\}$$

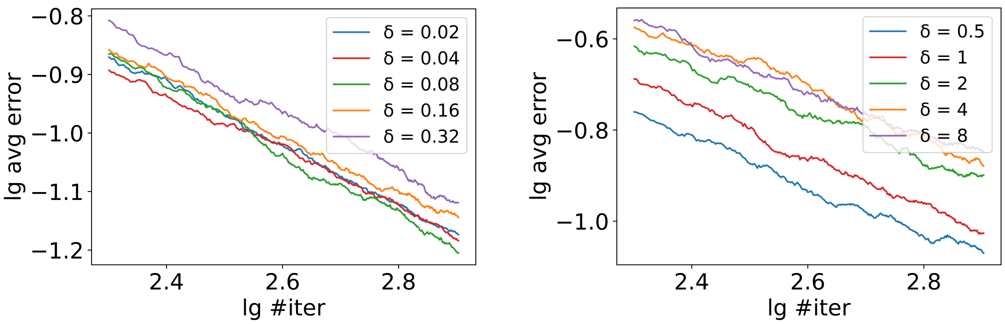


Figure 5: Test convergence for different δ

and the least square regression line for each $k = 500, 1000, 1500$.

G.4 Lost-Sale Inventory Control Model

In this experiment, we use $\delta = 0.5$, $\gamma = 0.7$. We run 300 trajectories of 5000 iterations of our Algorithm 1 ($g = 5/8$) and that in Liu et al. (2022) ($g = 0.499$). In both cases we use the rescaled-linear stepsize. We also record the number of sample used by trajectory i at iteration k (denote by $\hat{n}_{g,k}^{(i)}$) This produces data $\left\{\hat{n}_{g,k}^{(i)}, \hat{Q}_{\delta,g,k}^{(i)}; i = 1, \ldots, 300; k = 0, \ldots, 5000\right\}$ for $g = 5/8$ and $g = 0.499$.

Figure 4a plots the linear interpolation of points

$$\left\{\left(\lg\left(\frac{1}{300}\sum_{i=1}^{300}\hat{n}_{g,k}^{(i)}\right), \lg\left(\frac{1}{300}\sum_{i=1}^{300}\left\|\hat{Q}_{\delta,k}^{(i)}(\gamma) - Q_\delta^*(\gamma)\right\|_\infty\right)\right); k = 1000, 1001, \ldots, 5000\right\}$$

for $g = 5/8, 0.499$. For presentation clearness, Figure 4b plots the *smoothed* scattering of the data

$$\left\{\left(\hat{n}_{g,k}^{(i)}, \lg\left(\left\|\hat{Q}_{\delta,k}^{(i)}(\gamma) - Q_\delta^*(\gamma)\right\|_\infty\right)\right); i = 1, \ldots, 300; k = 1000, \ldots, 5000\right\}$$

The smoothing is in error, over a window of size $w = 0.0001$; i.e. if we define

$$\bar{n}_{g,k}^{(i)}(w) := \text{mean}\left\{n_{g,m}^{(n)} : 0 \le \hat{Q}_{\delta,g,m}^{(n)} - Q_{\delta,g,k}^{(i)} \le w\right\},$$

then Figure 4b plots the linear interpolation of points

$$\left\{\left(\hat{Q}_{\delta,g,k}^{(i)}, \bar{n}_{g,k}^{(i)}(w)\right); i = 1, \ldots, 300; k = 1000, \ldots, 5000\right\}$$

for $g = 5/8, 0.499$.