Robust Satisficing MDPs

Haolin Ruan [1] **Siyu Zhou** [2] **Zhi Chen** [3] **Chin Pang Ho** [1]

Abstract

Despite being a fundamental building block for reinforcement learning, Markov decision processes (MDPs) often suffer from ambiguity in model parameters. Robust MDPs are proposed to overcome this challenge by optimizing the worst-case performance under ambiguity. While robust MDPs can provide reliable policies with limited data, their worst-case performances are often overly conservative, and so they do not offer practical insights into the actual performance of these reliable policies. This paper proposes robust satisficing MDPs (RSMDPs), where the expected returns of feasible policies are softly-constrained to achieve a user-specified target under ambiguity. We derive a tractable reformulation for RSMDPs and develop a first-order method for solving large instances. Experimental results demonstrate that RSMDPs can prescribe policies to achieve their targets, which are much higher than the optimal worst-case returns computed by robust MDPs. Moreover, the average and percentile performances of our model are competitive among other models. We also demonstrate the scalability of the proposed algorithm compared with a state-of-the-art commercial solver.

1. Introduction

Markov decision processes (MDPs) have emerged as a powerful modeling framework for sequential decision-making problems under uncertainty (Ashok et al., 2019; Puterman, 2014; Sutton & Barto, 2018). Successful employments of MDPs largely rely on the perfect estimation of model parameters (Petrik & Russel, 2019), which, unfortunately, is not always the case. A common situation is when the true parameters are estimated from a limited amount of samples, which may lead to non-negligible estimation deviation (Mannor et al., 2007). Sometimes, these true parameters themselves may be uncertain or even time-dependent, yet they are mistreated as fixed ones in the modeling process (Mannor et al., 2016; Suilen et al., 2022). Due to the sequential nature of MDPs, these estimation errors accumulate quickly (Behzadian et al., 2021b; Xu & Mannor, 2009), and so the output policies of MDPs could be disappointing in practice.

As an encouraging framework to mitigate or resolve these issues, a robust MDP (RMDP) assumes the uncertain reward function and/or transition kernel to reside in an ambiguity set, which includes the possible candidates of the unknown true parameters with high confidence (Delgado et al., 2016; Ghavamzadeh et al., 2016; Hanasusanto & Kuhn, 2013; Petrik, 2012; Tamar et al., 2014; Xu & Mannor, 2006). By optimizing against the most adversarial value within the ambiguity set, RMDPs can provide policies that are robust in practice (Auer et al., 2008; Goyal & Grand-Clement, 2022; Hansen et al., 2013; Ho et al., 2018; Iyengar, 2005; Kaufman & Schaefer, 2013; Taleghan et al., 2015).

However, since RMDPs optimize the worst cases (which probably are unusual in most cases), the optimized worst-case performances are often too pessimistic and do not offer insights into the actual performance of the obtained policies. Moreover, one major drawback of using ambiguity sets to account for parameter ambiguity in RMDPs (resp., DR-MDPs) is that the model may perform poorly when the true parameter (resp., true probability distribution of parameter) is outside the ambiguity set, which could be catastrophic in high-risk applications such as healthcare and robotics (Brown et al., 2020). Another potential issue is the difficulty of determining the right "size" of the ambiguity set: for example, when using a Wasserstein ambiguity set (Derman & Mannor, 2020; Yang, 2017), choosing an appropriate size/radius that provides satisfying out-of-sample performances can be challenging (Mohajerin Esfahani & Kuhn, 2018).

In this paper, we focus on the case of ambiguous transition kernel and we propose a novel framework, named robust satisficing MDPs (RSMDPs), to handle the ambiguity. Robust satisficing is an alternative framework of optimization under uncertainty (Long et al., 2023). As opposed to RMDPs,

[1]School of Data Science, City University of Hong Kong [2]The City University of Hong Kong Shenzhen Research Institute [3]CUHK Business School, The Chinese University of Hong Kong. Correspondence to: Chin Pang Ho <clint.ho@cityu.edu.hk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

the proposed RSMDPs consider all possible transition kernels, therefore one does not need to specify an ambiguity set. Unlike nominal MDPs where only the hard constraints that correspond to the empirical transition kernel are considered, we also impose soft constraints for all other transition kernels in RSMDPs. The magnitudes of violation of these soft constraints depend on the distance between the associated transition kernel and the empirical one. This is one notable feature of our RSMDPs that prevents the model from performing too poorly when the empirical transition kernel is not close to the true one, which RMDPs may fail to achieve when the true kernel is not included in the ambiguity set. In particular, to achieve robustness, RSMDPs directly optimize/minimize the magnitude of constraint violations, and the level of robustness can be controlled elastically by articulating the targeted average return. Compared to the radius/size of RMDPs, the targeted return is a more tangible parameter, where a lower target corresponds to a higher level of robustness. Moreover, parameter selection approaches are much easier to use for choosing the one-dimensional targeted return, rather than choosing the possibly multi-dimensional size of an ambiguity set (*e.g.*, (Delage & Ye, 2010)). More details of RSMDPs are provided in Section 3.

Our contributions may be summarized as follows.

(*i*). We propose a novel framework of RSMDPs that allows the uncertain transition probabilities to vary within the entire support set and optimizes/minimizes the constraint violation directly while attaining an intuitive and tangible target on the expected return articulated by the decision-maker.

(*ii*). We derive a tractable reformulation of our RSMDPs as a conic program. To solve RSMDPs, we design a first-order method that is more scalable than the Gurobi solver (Gurobi Optimization, LLC, 2022), and thus is advantageous in large-scale problems.

(*iv*). Via data-driven experiments, we compare RSMDPs with nominal MDPs (NMDPs), RMDPs and distributionally robust MDPs (DRMDPs). Results show that RSMDPs have better percentile performances and target-oriented feature.

The remainder of the paper is organized as follows. Preliminaries are introduced in Section 2. We study RSMDPs and derive their tractable reformulation in Section 3, and we propose a first-order method to solve RSMDPs efficiently in Section 4. We conduct numerical experiments in Section 5.

Notations. We denote vectors (resp., matrices) by boldface lowercase (resp., uppercase) letters. The sets of non-negative and strictly positive real numbers are denoted as $\mathbb{R}_+$ and $\mathbb{R}_{++}$, respectively. The symbols $\mathbf{0}$ and $\mathbf{e}$ stand for the vectors of all 0's and all 1's of a size that is clear from the context, respectively. We use $\mathbf{e}_s,\ s \in \{1, \cdots, S\}$ to denote the s-th standard basis vector in $\mathbb{R}^S$. A probability simplex is denoted as $\Delta^S = \{\boldsymbol{p} \in \mathbb{R}^S_+ \mid \mathbf{e}^\top \boldsymbol{p} = 1\}$. The matrix $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{a}) \in \mathbb{R}^{S\times S}$ is diagonal with its diagonal entries being defined by the vector $\boldsymbol{a} \in \mathbb{R}^S$.

Related Work

RMDPs Solving RMDPs is generally an NP-hard problem (Ho et al., 2021; Iyengar, 2005; Wiesemann et al., 2013). The rectangularity assumption is crucial in obtaining tractability, which ensures that the optimal policy can be computed via robust variants of value or policy iteration in polynomial time (Hansen et al., 2013; Iyengar, 2005); many RMDPs are equipped with rectangular ambiguity sets: the sa-rectangularity is the most common assumption where the uncertain transition probabilities at each state-action pair are independently distributed (Iyengar, 2005; Nilim & El Ghaoui, 2005; Strehl et al., 2009). However, the volumes of sa-rectangular ambiguity sets are usually unnecessarily large, which implies that the output policies are often too conservative. To this end, s-rectangular sets are proposed to be a less conservative alternative where only independence between states are assumed (Goyal & Grand-Clement, 2022; Ho et al., 2021; 2022; Wiesemann et al., 2013). Fast algorithms for solving RMDPs are also an active research topic, where the RMDPs are equipped with polyhedral ambiguity sets that are defined using L_1 or L_∞-norm (Behzadian et al., 2021a; Derman et al., 2021; Ho et al., 2021) or with nonlinear ambiguity sets such as the spherical ambiguity set and the KL uncertainty set (Grand-Clément & Kroer, 2021b; Ho et al., 2022). On another note, it is worth mentioning that DRMDPs (Chen et al., 2019; Clement & Kroer, 2021; Xu & Mannor, 2010) are closely related to RMDPs, where the transition kernels are assumed to be random and subject to some unknown probability distributions that reside in ambiguity sets.

Dual formulation of NMDPs As we will demonstrate in Section 3, since the interpretation of the value function will become unclear when applying the robust satisficing framework to the primal formulation of NMDPs, the proposed RSMDPs are motivated by the dual formulation. In recent years, various new models have been proposed based on the dual formulation of NMDPs due to their interpretability. For example, (Lobo et al., 2020) optimize a weighted average of expectation and conditional value-at-risk (CVaR) of return under model ambiguity, (Brown et al., 2020) and (Delage & Mannor, 2010) optimize CVaR and the value-at-risk (VaR) of average return, respectively; by considering random rewards $\tilde{\boldsymbol{r}} \sim \mathbb{P}$ and a risk threshold $\varepsilon \in (0, 1)$, (Brown et al., 2020) and (Delage & Mannor, 2010) replace the objective function $\boldsymbol{r}^\top \boldsymbol{u}$ in (2) by $\max_y \left\{y - (1/(1-\varepsilon)) \cdot \mathbb{E}_{\mathbb{P}}[(y - \tilde{\boldsymbol{r}}^\top \boldsymbol{u})^+]\right\}$ and $\max_y\{y \mid \mathbb{P}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x} \geq y] \geq 1 - \varepsilon\}$, respectively. Our proposed RSMDPs, as opposed to the aforementioned models, do not optimize the expected return or the risk of return. Instead, we optimize/minimize the constraint violation while

specifying a target value for the expected return, reflecting their target-oriented feature.

Model-free approaches for robust reinforcement learning Our RSMDPs are motivated by the linear programming formulation of NMDPs, both of which are model-based approaches. We remark that, beyond model-based methods, there is an abundance of inspiring research on robust reinforcement learning, such as robust policy gradient (Wang & Zou, 2022), sample complexity analysis (Panaganti & Kalathil, 2022), least-squares policy iteration (Lagoudakis & Parr, 2003) and robust Q-learning (Roy et al., 2017; Wang & Zou, 2021). While model-free methods often require a large number of interactions with the environment, model-based learning is known for high sample efficiency (Sutton & Barto, 2018), which are especially preferred for those applications with limited data such as medicine (Imani et al., 2018) and manufacturing (Doltsinis et al., 2014).

2. Preliminaries

Consider an infinite-horizon MDP denoted by a tuple $\langle \mathcal{S}, \mathcal{A}, \boldsymbol{p}, \boldsymbol{r}, \gamma, \boldsymbol{d} \rangle$ with a finite state space $\mathcal{S} = \{1, \cdots, S\}$ and a finite action space $\mathcal{A} = \{1, \cdots, A\}$. When an action $a \in \mathcal{A}$ is chosen at state $s \in \mathcal{S}$, transition to a new state $s' \in \mathcal{S}$ follows a distribution $\boldsymbol{p}_{s,a} \in \Delta^S$ and a non-negative reward $r_{s,a} \in \mathbb{R}_+$ materializes. We condense the transition probabilities and rewards to $\boldsymbol{p} = (\boldsymbol{p}_{s,a})_{s\in\mathcal{S},a\in\mathcal{A}} \in (\Delta^S)^{S\cdot A}$ and $\boldsymbol{r} = (r_{s,a})_{s\in\mathcal{S},a\in\mathcal{A}}$, respectively. The discount factor is $\gamma \in (0,1)$ and the initial state distribution is $\boldsymbol{d} \in \mathbb{R}^S_{++}$. We optimize a policy $\boldsymbol{\pi} \in \Pi = (\Delta^A)^S$ that takes an action $a \in \mathcal{A}$ with probability $\pi_{s,a}$ at state $s \in \mathcal{S}$, where Π is the set of all (stationary) randomized policies. For a nominal MDP (NMDP), to obtain the optimal policy that maximizes our total expected discounted reward, we solve $\max_{\boldsymbol{\pi}\in\Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r_{S_t,A_t}\right]$, where the initial state S_0 follows the distribution $\boldsymbol{d}$, and for any time step t, we use S_t to denote the state at time step t while A_t is a random action governed by the distribution $\boldsymbol{\pi}_{S_t} \in \Delta^A$.

To obtain the optimal policy of an NMDP, under the dynamic programming principle, we let $\boldsymbol{v}^{\boldsymbol{\pi}} \in \mathbb{R}^S_+$ be a vector where $v_s^{\boldsymbol{\pi}}$ denotes the total expected discounted reward starting at state $s \in \mathcal{S}$ when applying policy $\boldsymbol{\pi}$. The optimal value function $v_s^\star$ (achieved by the optimal policy $\boldsymbol{\pi}^\star$) can be expressed via the Bellman optimality equation

$$v_s^\star = \max_{a\in\mathcal{A}} \left\{ r_{s,a} + \gamma \cdot \boldsymbol{p}_{s,a}^\top \boldsymbol{v}^\star \right\} \quad \forall s \in \mathcal{S}, \tag{1}$$

and the optimal solution can be retrieved by value iteration or policy iteration; see, *e.g.*, (Sutton & Barto, 2018). Alternatively, an NMDP can be formulated as a linear program (in primal and dual)[1]

$$\begin{aligned} \mathrm{Z_N}(\boldsymbol{p}) = \min\ & \boldsymbol{d}^\top \boldsymbol{v} \\ \text{s.t.}\ & v_s \geq r_{s,a} + \gamma \cdot \boldsymbol{p}_{s,a}^\top \boldsymbol{v}\ \forall (s,a) \in \mathcal{S}\times\mathcal{A} \\ & \boldsymbol{v} \in \mathbb{R}^S_+ \\ = \max\ & \boldsymbol{r}^\top \boldsymbol{u} \\ \text{s.t.}\ & \mathbf{e}^\top \boldsymbol{u}_s - \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u} - d_s \leq 0\ \forall s \in \mathcal{S} \\ & \boldsymbol{u} \in \mathbb{R}^{S\cdot A}_+, \end{aligned} \tag{2}$$

where for each $s \in \mathcal{S}$, $\boldsymbol{Q}_s = \gamma \cdot \mathrm{diag}(\mathbf{e}_s, \cdots, \mathbf{e}_s)$ such that $\gamma \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_{s',a,s} u_{s',a} = \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u}$. Here, we focus on the dual formulation (*i.e.*, the maximization problem) in (2), where the feasible solution $\boldsymbol{u}$ is interpreted as the discounted probability of executing action a at state s when employing a (stationary randomized) policy $\pi_{s,a} = u_{s,a} / \sum_{a'\in\mathcal{A}} u_{s,a'}\ \ \forall (s,a) \in \mathcal{S}\times\mathcal{A}$. Therefore, the dual formulation in essence, chooses the optimal policy that maximizes the total expected discounted reward $\boldsymbol{r}^\top \boldsymbol{u}$ (Puterman, 2014).

Robust optimization (*e.g.*, (Ben-Tal et al., 2009)) is a classic paradigm to account for parameter uncertainty:

$$\begin{aligned} \max\ & \boldsymbol{r}^\top \boldsymbol{u} \\ \text{s.t.}\ & \mathbf{e}^\top \boldsymbol{u}_s - \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u} - d_s \leq 0\ \forall \boldsymbol{p} \in \mathcal{F}, s \in \mathcal{S} \\ & \boldsymbol{u} \in \mathbb{R}^{S\cdot A}_+. \end{aligned} \tag{3}$$

Here, the ambiguity set $\mathcal{F} = \{\boldsymbol{p} \in \mathcal{P} \mid \ell(\boldsymbol{p}, \hat{\boldsymbol{p}}) \leq \kappa\}$ is a κ-neighbourhood (measured by a distance function ℓ) around the empirical $\hat{\boldsymbol{p}}$. Compare the first set of constraints of (3) to that of the dual formulation (2), one can observe that, for all $s \in \mathcal{S}$, the solution of (3) is robust against all transition kernels in the ambiguity set $\mathcal{F}$, *i.e.*, the solution of (3) always remains feasible for all $\boldsymbol{p} \in \mathcal{F}$, while the solution of $\mathrm{Z_N}(\hat{\boldsymbol{p}})$ is only guaranteed to be feasible when the true transition kernel is the same as the empirical one (*i.e.*, $\hat{\boldsymbol{p}}$).

3. Robust Satisficing MDPs

3.1. Model

Equipped with the dual formulation of NMDPs in (2), our target-oriented robust satisficing MDP is formulated as

$$\begin{aligned} \mathrm{Z_{RS}}(\hat{\boldsymbol{p}}) = \min\ & \boldsymbol{w}^\top \boldsymbol{k} \\ \text{s.t.}\ & \mathbf{e}^\top \boldsymbol{u}_s - \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u} - d_s \leq k_s \cdot \ell(\boldsymbol{p}, \hat{\boldsymbol{p}}) \\ & \qquad\qquad \forall \boldsymbol{p} \in \mathcal{P},\ s \in \mathcal{S} \\ & \boldsymbol{r}^\top \boldsymbol{u} \geq \tau \\ & \boldsymbol{u} \in \mathbb{R}^{S\cdot A}_+,\ \boldsymbol{k} \in \mathbb{R}^S_+, \end{aligned} \tag{4}$$

where $\hat{\boldsymbol{p}}$ is the empirical transition kernel and the support set is $\mathcal{P} = \left\{\boldsymbol{p} \in \mathbb{R}^{S\cdot A\cdot S}_+ \mid \mathbf{e}^\top \boldsymbol{p}_{s,a} = 1\ \forall s \in \mathcal{S},\ a \in \mathcal{A}\right\}$.

[1] Throughout this paper, the last line of constraints of an optimization problem indicates its decision variables (and their dimensions). For example, $\boldsymbol{v} \in \mathbb{R}^S_+$ and $\boldsymbol{u} \in \mathbb{R}^{S\cdot A}_+$ herein are the decision variables of the primal and dual NMDPs, respectively.

While $\boldsymbol{w}$ could be set to any non-negative vector, we commonly set $\boldsymbol{w} = \mathbf{e}$ in (4). If there is no additional information, the first set of constraints appear to be symmetric to each other and should be equally important. The term $\ell(\boldsymbol{p}, \hat{\boldsymbol{p}})$, which can be a general norm $\ell(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \|\boldsymbol{p} - \hat{\boldsymbol{p}}\|$ or other distances such as the KL divergence, measures the proximity between $\boldsymbol{p}$ and the empirical $\hat{\boldsymbol{p}}$.

Comparing the first collection of constraints in RSMDP (4) to that of (the dual formulation of) NMDP (2), we observe that the decision variables $\{k_s\}_{s\in\mathcal{S}}$ in RSMDPs reflect the magnitude of constraint violation incurred by the distance between the ambiguous transition kernel and the empirical one: when the values of $\{k_s\}_{s\in\mathcal{S}}$ are small, only mild violation will occur even when the true transition kernel is far away from the empirical estimation $\hat{\boldsymbol{p}}$, which is often unlikely to happen. RSMDPs then minimize a weighted sum of $\{k_s\}_{s\in\mathcal{S}}$ under the promise that the average return $\boldsymbol{r}^\top \boldsymbol{u}$ is not smaller than the pre-specified target $\tau > 0$, highlighting the notion of satisficing. Notice that when $\boldsymbol{p} = \hat{\boldsymbol{p}}$, there will be no violation in our RSMDPs (4), thus the average return achieved should reach the target τ.

Note that the feasible region of (4) will become smaller with a larger τ, where the optimal $k_s,\ s \in \mathcal{S}$ tend to be larger. Therefore, one can interpret τ as, in addition to the targeted expected return, the controller of the robustness of (4). That is, a smaller τ corresponds to higher robustness. By setting $\tau = \mathrm{Z_N}(\hat{\boldsymbol{p}})$, one can recover the optimal policy of the corresponding NMDP.

Proposition 3.1. *(i) Any optimal $\boldsymbol{u}^\star$ of $\mathrm{Z_{RS}}(\hat{\boldsymbol{p}})$ with $\tau = \mathrm{Z_N}(\hat{\boldsymbol{p}})$ is also optimal in $\mathrm{Z_N}(\hat{\boldsymbol{p}})$, while (ii) $\mathrm{Z_{RS}}(\hat{\boldsymbol{p}})$ is infeasible when $\tau > \mathrm{Z_N}(\hat{\boldsymbol{p}})$.*

Since we consider non-negative rewards, when we set $\tau \leq 0$, the feasible region of (4) is unchanged with its second set of constraints (*i.e.*, $\boldsymbol{r}^\top \boldsymbol{u} \geq \tau$) being eliminated. This observation, together with Proposition 3.1, provide the interval $[0, \mathrm{Z_N}(\boldsymbol{p})]$ within which we calibrate τ for RSMDPs. As opposed to sizing the ambiguity sets in RMDPs, setting the target τ is another distinguished feature of RSMDPs. To a decision-maker, the target τ is more directly related to her objective (*i.e.*, return) and is thus more tangible.

The robust optimization model (3) is another important benchmark for RSMDPs (4). Compare their first set of constraints, one can observe that for all $s \in \mathcal{S}$, the robust optimization model only hedges against transition kernels in the ambiguity set $\mathcal{F}$, while giving no guarantee about the constraint violation when the transition kernel is outside $\mathcal{F}$. In contrast, the RSMDP (4) minimizes the magnitude of the constraint violation for all possible transition kernels from the whole support set $\mathcal{P}$, where the decision variable k_s measures the maximal violation of the s-th constraint among all $\boldsymbol{p} \in \mathcal{P}$ which is to be optimized. To further distinguish between these two frameworks, let $\boldsymbol{u}^{\mathrm{RO}}$ and $(\boldsymbol{u}^{\mathrm{RS}}, \boldsymbol{k})$ be the feasible solutions of the robust optimization model (3) and RSMDP (4), respectively. Comparing the s-th constraints of (3) and (4), for the former we have

$$\begin{cases} \mathbf{e}^\top \boldsymbol{u}_s^{\mathrm{RO}} - \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u}^{\mathrm{RO}} - d_s \leq 0 & \forall \boldsymbol{p} \in \mathcal{F} \\ \mathbf{e}^\top \boldsymbol{u}_s^{\mathrm{RO}} - \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u}^{\mathrm{RO}} - d_s \leq +\infty & \forall \boldsymbol{p} \in \mathcal{P}\backslash\mathcal{F}, \end{cases} \tag{5}$$

while for the latter, we have

$$\begin{cases} \mathbf{e}^\top \boldsymbol{u}_s^{\mathrm{RS}} - \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u}^{\mathrm{RS}} - d_s \leq k_s \cdot \ell(\boldsymbol{p}, \hat{\boldsymbol{p}}) & \forall \boldsymbol{p} \in \mathcal{F} \\ \mathbf{e}^\top \boldsymbol{u}_s^{\mathrm{RS}} - \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u}^{\mathrm{RS}} - d_s \leq k_s \cdot \ell(\boldsymbol{p}, \hat{\boldsymbol{p}}) & \forall \boldsymbol{p} \in \mathcal{P}\backslash\mathcal{F}. \end{cases} \tag{6}$$

It is clearly indicated by (5) and (6) that, though RSMDPs may allow some additional violation (*i.e.*, $k_s \cdot \ell(\boldsymbol{p}, \hat{\boldsymbol{p}})$) when the true transition kernel is inside the ambiguity set $\mathcal{F}$, it can protect one from disastrous situations when $\boldsymbol{p} \in \mathcal{P}\backslash\mathcal{F}$. This is one advantage of RSMDPs because the ambiguity set $\mathcal{F}$ is often much smaller compared to the support set $\mathcal{P}$. One can also consider $\mathcal{P}$ as the ambiguity set for RSMDPs, which is a "special" one in the way that the magnitude of violation is proportional to the proximity of the unknown true transition kernel to its nominal value (*i.e.*, $\ell(\boldsymbol{p}, \hat{\boldsymbol{p}})$); for the robust optimization framework, the violation is equal for all $\boldsymbol{p} \in \mathcal{F}$ (and is unpredictable and ignored for $\boldsymbol{p} \notin \mathcal{F}$).

3.2. Reformulation

RSMDP (4) is an infinitely constrained optimization problem, where each possible $\boldsymbol{p} \in \mathcal{P}$ corresponds to one constraint. Quite notably, it can be reformulated as a conic program when equipped with a general norm.

Theorem 3.2. *Equipped with a general norm $\ell(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \|\boldsymbol{p} - \hat{\boldsymbol{p}}\|$, RSMDP (4) is equivalent to the conic program*

$$\begin{aligned} & \mathrm{Z_{RS}}(\hat{\boldsymbol{p}}) \\ = \ & \min \quad \sum_{s\in\mathcal{S}} w_s \cdot \|\boldsymbol{\beta}_s - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}_s\|_* \\ & \text{s.t.} \quad \mathbf{e}^\top \boldsymbol{u}_s - d_s \leq -\hat{\boldsymbol{p}}^\top \boldsymbol{\beta}_s + \hat{\boldsymbol{p}}^\top \boldsymbol{Q}_s \boldsymbol{u} \qquad \forall s \in \mathcal{S} \\ & \qquad \boldsymbol{r}^\top \boldsymbol{u} \geq \tau \\ & \qquad \boldsymbol{u} \in \mathbb{R}_+^{S\cdot A},\ \boldsymbol{\alpha}_s \in \mathbb{R}^{S\cdot A},\ \boldsymbol{\beta}_s \in \mathbb{R}_+^{S\cdot A\cdot S}\ \forall s \in \mathcal{S} \end{aligned} \tag{7}$$

where $\boldsymbol{B} = \mathrm{diag}(\mathbf{e}^\top, \cdots, \mathbf{e}^\top) \in \mathbb{R}^{S\cdot A \times S\cdot A\cdot S}$ and $\mathbf{e} \in \mathbb{R}^S$.

We remark that, due to the choice of a general norm in (4), the conic program (7) admits an equivalent reformulation as a minimax optimization problem (as we will show in the coming section), to which an efficient primal-dual algorithm can be applied. Besides a general norm, choosing the KL divergence in (4) also allows a reformulation as a convex optimization problem (see more details in Appendix B). We also remark that, after solving (7), the optimal policy of RSMDPs can be retrieved as in NMDPs (2).

4. First-Order Method

By Theorem 3.2, we are already able to compute the optimal policy of an RSMDP (equipped with a general norm) by solving its equivalent reformulation (7) via the state-of-the-art commercial solvers; however, the computation time may be quite long when the problem size becomes large. To this end, we will apply the first-order primal-dual algorithm (PDA) (*e.g.*, (Chambolle & Pock, 2016; Esser et al., 2010; Grand-Clément & Kroer, 2021b; He & Yuan, 2012)) that aims to solve problems in a convex-concave min-max form at a convergence rate $\mathcal{O}(1/N)$ (Chambolle & Pock, 2016), where N is the number of iterations. First-order methods are known for their computationally efficient updates, by which large-scale problems can be solved with moderate accuracy efficiently. To apply PDA, we will transform problem (7) into a min-max form. A necessary preceding result for the transformation is relegated to Lemma A.1 in Appendix A.2, and we provide the min-max reformulation in the following proposition.

Proposition 4.1. *When $\boldsymbol{w} \in \mathbb{R}^S_{++}$, problem (7) has an equivalent minimax reformulation:*

$$\min_{\boldsymbol{u}\in\mathcal{U}} \max_{(\lambda_s,\boldsymbol{\theta}_s)\in\mathcal{V}_q(w_s):s\in\mathcal{S}} \sum_{s\in\mathcal{S}} (\lambda_s \cdot (\mathbf{e}^\top \boldsymbol{u}_s - d_s) - \boldsymbol{\theta}_s^\top \boldsymbol{Q}_s \boldsymbol{u}), \tag{8}$$

where $\mathcal{U} = \{\boldsymbol{u} \in \mathbb{R}^{S\cdot A}_+ \mid \boldsymbol{r}^\top \boldsymbol{u} \geq \tau\}$ and

$$\mathcal{V}_q(w) = \left\{ (\lambda, \boldsymbol{\theta}) \in \mathbb{R}_+ \times \mathbb{R}^{S\cdot A\cdot S}_+ \;\middle|\; \begin{array}{l} \|\boldsymbol{\theta} - \lambda \cdot \hat{\boldsymbol{p}}\|_q \leq w \\ \lambda \cdot \mathbf{e} = \boldsymbol{B}\boldsymbol{\theta} \end{array} \right\}.$$

In the remainder, we focus on solving the min-max problem (8) (thus RSMDPs (4)) equipped with an L_∞-norm:

$$\min_{\boldsymbol{u}\in\mathcal{U}} \max_{(\lambda_s,\boldsymbol{\theta}_s)\in\mathcal{V}_\infty(w_s):s\in\mathcal{S}} \sum_{s\in\mathcal{S}} (\lambda_s \cdot (\mathbf{e}^\top \boldsymbol{u}_s - d_s) - \boldsymbol{\theta}_s^\top \boldsymbol{Q}_s \boldsymbol{u}). \tag{9}$$

As we demonstrate later in this section, the specific problem structure of this min-max formulation, together with the computational scheme of PDA, allows us to decompose the min-max formulation into several subproblems that can be solved efficiently in each iteration by our designed strategy. In particular, to update the decision variable of the outer minimization problem, we solve a minimization problem by an efficient procedure with time complexity $\mathcal{O}(SA \cdot \log(SA))$; while to update the inner problem, we design a strategy with time complexity $\mathcal{O}(S^3 \log(S) A \log(1/\delta))$, where δ is the desired precision of the golden section search that will be introduced in Section 4.2.

Observe that, for any fixed $\boldsymbol{u} \in \mathbb{R}^{S\cdot A}_+$, the inner maximization problem of (9) is decomposable into S subproblems, which allows an equivalent reformulation of (9) as

$$\min_{\boldsymbol{u}\in\mathcal{U}} \sum_{s\in\mathcal{S}} \max_{(\lambda_s,\boldsymbol{\theta}_s)\in\mathcal{V}_\infty(w_s)} \lambda_s \cdot (\mathbf{e}^\top \boldsymbol{u}_s - d_s) - \boldsymbol{\theta}_s^\top \boldsymbol{Q}_s \boldsymbol{u}.$$

Hence, for any fixed $\boldsymbol{u}$, it is sufficient to solve each of these S subproblems separately. This decomposition is beneficial to our PDA since it can remarkably lower the time complexity for the update of the decision variables $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ for the inner maximization problem in (9).

Now we introduce our PDA in Algorithm 1. In every iteration, our PDA updates the primal (resp., dual) variable by solving a minimization problem with the dual (resp., primal) variable fixed at a value related to its last update. Here, the primal update operator is defined as

$$\begin{aligned} &\mathfrak{P}(\boldsymbol{\lambda}, \boldsymbol{\theta}; \hat{\boldsymbol{u}}) \\ &= \operatorname*{argmin}_{\boldsymbol{u}\in\mathcal{U}} \sum_{s\in\mathcal{S}} \left(\lambda_s \mathbf{e}^\top \boldsymbol{u}_s - \boldsymbol{\theta}_s^\top \boldsymbol{Q}_s \boldsymbol{u}\right) + \frac{1}{2\nu} \cdot \|\boldsymbol{u} - \hat{\boldsymbol{u}}\|_2^2 \end{aligned}$$

for $\hat{\boldsymbol{u}} \in \mathbb{R}^{S\cdot A}_+$ and $(\boldsymbol{\lambda}, \boldsymbol{\theta}) \in \mathbb{R}^S_+ \times \mathbb{R}^{S\cdot S\cdot A\cdot S}_+$, while for any $\hat{s} \in \mathcal{S}$, the dual update operator is

$$\begin{aligned} \mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}}) = \operatorname{argmin}\; & \lambda(d_{\hat{s}} - \mathbf{e}^\top \boldsymbol{u}_{\hat{s}}) + \boldsymbol{\theta}^\top \boldsymbol{Q}_{\hat{s}} \boldsymbol{u} \\ & + \tfrac{1}{2\sigma} \cdot ((\lambda - \hat{\lambda})^2 + \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2) \\ \text{s.t.}\quad & (\lambda, \boldsymbol{\theta}) \in \mathcal{V}_\infty(w_{\hat{s}}) \end{aligned}$$

for any $\boldsymbol{u} \in \mathbb{R}^{S\cdot A}_+$ and $(\hat{\lambda}, \hat{\boldsymbol{\theta}}) \in \mathbb{R}_+ \times \mathbb{R}^{S\cdot A\cdot S}_+$, where $\nu > 0$ and $\sigma > 0$ are, respectively, the stepsizes of the primal and dual updates. The input coefficient matrix $\boldsymbol{C} = (\operatorname{diag}(\mathbf{e}, \cdots, \mathbf{e}), -\boldsymbol{Q}_1^\top, \cdots, -\boldsymbol{Q}_S^\top)^\top \in \mathbb{R}^{(S+S\cdot S\cdot A\cdot S)\times S\cdot A}$ with S all-ones vectors $\mathbf{e} \in \mathbb{R}^A$ satisfies $\langle \boldsymbol{C}\boldsymbol{u}, (\boldsymbol{\lambda}^\top, \boldsymbol{\theta}^\top)^\top \rangle = \sum_{s\in\mathcal{S}} \{\lambda_s(\mathbf{e}^\top \boldsymbol{u}_s - d_s) - \boldsymbol{\theta}_s^\top \boldsymbol{Q}_s \boldsymbol{u}\}$. Note that one needs to choose the Bregman divergence in PDA, and herein we choose the convex function in the definition of the Bregman divergence as $(1/2) \cdot \|\cdot\|_2^2$ for both the primal and the dual updates (Chambolle & Pock, 2016). We provide a simplified result for the convergence rate of Algorithm 1, which is based on a stronger but more technical convergence result in Theorem 1 in (Chambolle & Pock, 2016). Specifically, our result is obtained by specifying the sufficient condition $\nu\sigma \leq 1/L^2$. We refer interested readers to Theorem A.2 in Appendix A.2 for the original convergence result in (Chambolle & Pock, 2016).

Theorem 4.2. *Let $(\boldsymbol{u}^k, (\boldsymbol{\lambda}^k, \boldsymbol{\theta}^k))$, $k = 0, 1, 2, \cdots, K$ be a sequence generated by Algorithm 1. If the stepsizes ν, $\sigma > 0$ are chosen such that $\nu\sigma \leq 1/L^2$. Then for any feasible solution $(\boldsymbol{u}, (\boldsymbol{\lambda}, \boldsymbol{\theta}))$ of problem (9) it holds that*

$$\sum_{s\in\mathcal{S}} \phi_s(\bar{\boldsymbol{u}}^K, \lambda_s, \boldsymbol{\theta}_s) - \sum_{s\in\mathcal{S}} \phi_s(\boldsymbol{u}, \bar{\lambda}_s^K, \bar{\boldsymbol{\theta}}_s^K) = \mathcal{O}\left(\frac{1}{K}\right),$$

where $\bar{\boldsymbol{u}}^K = (\sum_{k\in[K]} \boldsymbol{u}^k)/K$, $(\bar{\boldsymbol{\lambda}}^K, \bar{\boldsymbol{\theta}}^K) = (\sum_{k\in[K]} (\boldsymbol{\lambda}^k, \boldsymbol{\theta}^k))/K$, and for all $s \in \mathcal{S}$, $\phi_s(\boldsymbol{u}, \boldsymbol{\theta}, \lambda) = \lambda \cdot (\mathbf{e}^\top \boldsymbol{u}_s - d_s) - \boldsymbol{\theta}^\top \boldsymbol{Q}_s \boldsymbol{u}$.

4.1. Solving $\mathfrak{P}(\boldsymbol{\lambda}, \boldsymbol{\theta}; \hat{\boldsymbol{u}})$ via Interval Search

To solve $\mathfrak{P}(\boldsymbol{\lambda}, \boldsymbol{\theta}; \hat{\boldsymbol{u}})$ in Step 1, note that it is a quadratic program with no cross term in the objective function and with

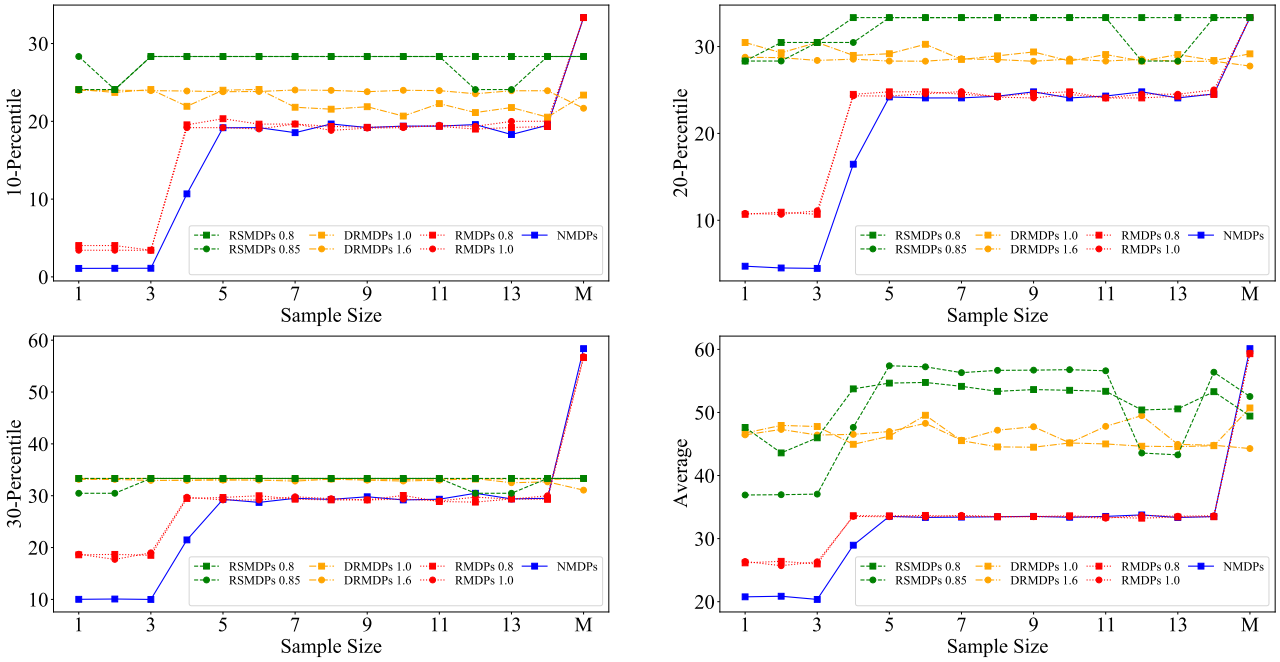

Figure 1. Average and percentile performances over 5000 out-of-sample testing trajectories in the *river swim* application.

Algorithm 1 Primal-Dual Algorithm (PDA) for Problem (9)

Input: Operator norm $L = \|C\|$, initial feasible solution $(\boldsymbol{u}^0, (\boldsymbol{\lambda}^0, \boldsymbol{\theta}^0))$ of problem (9), stepsizes ν, $\sigma > 0$, desired precision ε, $k \leftarrow 0$
repeat
  // Step 1 : Primal update
  $\boldsymbol{u}^{k+1} \leftarrow \mathfrak{P}(\boldsymbol{\lambda}^k, \boldsymbol{\theta}^k; \boldsymbol{u}^k)$;
  // Step 2 : Dual update
  for $\hat{s} = 1$ **to** S **do**
    $(\lambda_{\hat{s}}^{k+1}, \boldsymbol{\theta}_{\hat{s}}^{k+1}) \leftarrow \mathfrak{D}_{\hat{s}}(2\boldsymbol{u}^{k+1} - \boldsymbol{u}^k; \lambda_{\hat{s}}^k, \boldsymbol{\theta}_{\hat{s}}^k)$;
  end for
  $k \leftarrow k + 1$;
until $\|\boldsymbol{u}^k - \boldsymbol{u}^{k-1}\|_\infty < \varepsilon$
Output: Solution $\bar{\boldsymbol{u}}^k = \frac{1}{k}\sum_{i\in[k]} \boldsymbol{u}^i$ and $(\bar{\boldsymbol{\lambda}}^k, \bar{\boldsymbol{\theta}}^k) = \frac{1}{k}\sum_{i\in[k]}(\boldsymbol{\lambda}^i, \boldsymbol{\theta}^i)$

only linear constraints. Based on its structure, we develop an efficient algorithm to solve this problem. In particular, we show that solving $\mathfrak{P}(\boldsymbol{\lambda}, \boldsymbol{\theta}; \hat{\boldsymbol{u}})$ is equivalent to finding the root of a non-decreasing piecewise linear function, which can be done efficiently via *Interval Search* (where we search the intervals between breakpoints of the piecewise linear function). The time complexity of the proposed algorithm is provided in the statement of the following proposition, while the details of the algorithm are provided in the proof and pseudocode that are relegated to the appendix.

Proposition 4.3. *Problem $\mathfrak{P}(\boldsymbol{\lambda}, \boldsymbol{\theta}; \hat{\boldsymbol{u}})$ can be solved in time $\mathcal{O}(SA \cdot \log(SA))$.*

4.2. Solving $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$ via Golden Section Search

For Step 2 in our PDA, for each $\hat{s} \in \mathcal{S}$, we solve the subproblem $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$ by a well-known golden section search algorithm (*e.g.*, (Truhar & Veselić, 2009)). For any fixed $\lambda \in \mathbb{R}_+$, the function

$$f_{\hat{s}}(\lambda) = \min \quad \lambda(d_{\hat{s}} - \mathbf{e}^\top \boldsymbol{u}_{\hat{s}}) + \boldsymbol{\theta}^\top \boldsymbol{Q}_{\hat{s}} \boldsymbol{u} + \frac{1}{2\sigma} \cdot \left\| \begin{matrix} \lambda - \hat{\lambda} \\ \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \end{matrix} \right\|_2^2$$
$$\text{s.t.} \quad (\lambda, \boldsymbol{\theta}) \in \mathcal{V}_\infty(w_{\hat{s}})$$
$$\boldsymbol{\theta} \in \mathbb{R}_+^{S\times A\times S} \tag{10}$$

defined on $\mathbb{R}_+$ involves the inner minimization problem that we need to solve. Notice that, herein we treat the problem $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$ as a min-min problem where we optimize $\lambda \in \mathbb{R}_+$ in the outer minimization problem and $\boldsymbol{\theta} \in \mathbb{R}_+^{S\cdot A\cdot S}$ in the inner one. The golden section search is used to locate the optimal $\lambda^\star$ for the outer problem: we initialize the interval $[\underline{\lambda}, \overline{\lambda}]$ for the search, where $\underline{\lambda} = 0$ and we provide Lemma A.3 in Appendix A.2 for selecting $\overline{\lambda}$. At the initialization phase and in each iteration of the search, we need to solve the inner problem (10) once.

To implement the golden section search, we need to prove that $f_{\hat{s}}$ is well-defined on $\mathbb{R}_+$. We provide this proof in Lemma A.4 in Appendix A.2. In general, the golden section search converges to a local minimizer of the problem; fortunately, this is a global minimizer in our case.

Proposition 4.4. *The golden section search converges to a global minimizer $\lambda^\star$ of $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$.*

Table 1. Predicted returns and the corresponding differences (in median) between sample returns and predicted returns over 1000 samples in the *river swim* application.

	τ / $\mathrm{Z_N}(\hat{\boldsymbol{p}})$	1.0	0.9	0.8	0.7	0.6	0.5
RSMDPs	Predicted Return	58.6	52.7	46.9	41.0	35.2	29.3
	Difference (in median)	-4.4	**0.9**	**6.1**	**12.0**	**17.9**	**23.6**
DRMDPs	r	0.0	0.3	0.6	0.9	1.2	1.5
	Predicted Return	58.6	41.0	32.4	26.5	21.5	17.3
	Difference (in median)	-4.4	**12.9**	**20.7**	**26.6**	**31.7**	**35.9**
RMDPs	r	0.0	0.3	0.6	0.9	1.2	1.5
	Predicted Return	58.6	38.6	27.2	20.2	15.0	11.3
	Difference (in median)	-4.4	**15.6**	**27.0**	**34.1**	**39.2**	**43.0**

It remains to solve problem (10) to obtain the optimal solution $\boldsymbol{\theta}^\star$. Observe that the problem can be further decomposed into SA subproblems, and for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned} \min\ & \frac{1}{2\sigma} \cdot \boldsymbol{\theta}_{s,a}^\top \boldsymbol{\theta}_{s,a} + \boldsymbol{\theta}_{s,a}^\top (\boldsymbol{z}_{\hat{s},s,a} - \frac{1}{\sigma} \cdot \hat{\boldsymbol{\theta}}_{s,a}) \\ \text{s.t. } & [\lambda \hat{p}_{s,a,s'} - w_{\hat{s}}]_+ \le \theta_{s,a,s'} \le \lambda \hat{p}_{s,a,s'} + w_{\hat{s}}\ \forall s' \in \mathcal{S} \\ & \mathbf{e}^\top \boldsymbol{\theta}_{s,a} = \lambda \\ & \boldsymbol{\theta}_{s,a} \in \mathbb{R}^S, \end{aligned} \tag{11}$$

where $\boldsymbol{z}_{\hat{s}} = \boldsymbol{Q}_{\hat{s}} \boldsymbol{u} \in \mathbb{R}^{S \cdot A \cdot S}$. The only difference between (11) and $\mathfrak{P}(\boldsymbol{\lambda}, \boldsymbol{\theta}; \hat{\boldsymbol{u}})$ is that the former has upper bounds for decision variables, while the latter has not. Therefore, we can develop a similar efficient strategy to solve problem (11), whose time complexity is provided in the following proposition.

Proposition 4.5. *Problem* (11) *can be solved in time* $\mathcal{O}(S \log S)$.

Now we provide the total time complexity of our strategy described in this section to compute Step 2 in Algorithm 1.

Proposition 4.6. *The output of Step* 2 *in Algorithm 1 can be computed in time* $\mathcal{O}(S^3 \log(S) A \log(\delta^{-1}))$*, where* $\delta > 0$ *is the desired precision of the golden section search.*

4.3. Randomized Block Coordinate Gradient Descent for Dual Updates

Although Step 2 in Algorithm 1 can be computed in time complexity that is almost linear in the number of dual variables $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, the dual updates in Step 2 remain as the computational bottleneck of the algorithm. This is because the number of dual variables is in $\mathcal{O}(S^3 A)$, as opposed to the number of primal variables $\boldsymbol{u}$, which is in $\mathcal{O}(SA)$. Moreover, as mentioned above, our policy is computed via normalization step, $\pi_{s,a} = u_{s,a} / \sum_{a' \in \mathcal{A}} u_{s,a'}\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$; therefore, it would be desirable to update the primal variables $\boldsymbol{u}$ frequently toward their optimal values. To achieve this, one possible direction is to update only part of the dual variables at every Step 2. In particular, inspired by the randomized block coordinate gradient descent, one may modify the Step 2 in Algorithm 1 to be

$$(\lambda_s^{k+1}, \boldsymbol{\theta}_s^{k+1}) \leftarrow \mathfrak{D}_s(2\boldsymbol{u}^{k+1} - \boldsymbol{u}^k; \lambda_s^k, \boldsymbol{\theta}_s^k)\ \forall s \in \mathcal{S}^k. \tag{12}$$

Here $\mathcal{S}^k$ with $|\mathcal{S}^k| = M \ll S\ \forall k$ is an index set whose elements are sampled uniformly from $\mathcal{S}$ without replacement in each iteration. Therefore, only M dual update operators would be applied, and the time complexity of this modified Step 2 is reduced from $\mathcal{O}(S^3 \log(S) A \log(\delta^{-1}))$ to $\mathcal{O}(MS^2 \log(S) A \log(\delta^{-1}))$. In the same spirit, a more radical approach is to uniformly sample $(\hat{s}, s, a)$ from $\mathcal{S} \times \mathcal{S} \times \mathcal{A}$ at every Step 2 and update the dual variables that are only associated to the corresponding problem (11). The complexity of such updates would be reduced to $\mathcal{O}(S \log S)$. This second type of updates with problem (11) doesn't update $\boldsymbol{\lambda}$. Hence, one needs to apply both aforementioned updates. We relegate the discussions on the computation complexities of NMDPs, RMDPs and RSMDPs to Appendix C.

5. Numerical Results

We compare the performances of the proposed RSMDPs with NMDPs, RMDPs and DRMDPs in three applications: *river swim* (Strehl & Littman, 2008), *machine replacement* (Delage & Mannor, 2010) and *grid world* (Ghavamzadeh et al., 2016), where the results of the latter two applications are provided in Appendices D.5 and D.6, and detailed settings are provided in Appendices D, D.1 and D.4. We also compare our proposed algorithms with the Gurobi solver in terms of scalability; see more details in Appendix D.7. The code and data to reproduce our experiments is available online at `https://github.com/RUANHaolin/RSMDPs`.

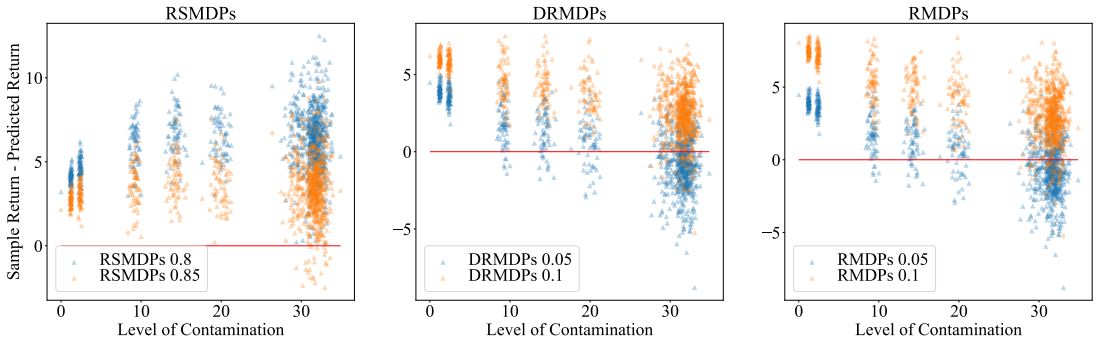

Figure 2. Differences between sample returns and predicted returns over 1000 samples in the *river swim* environment. Due to the page limit, we only plot the parameters (of the three models) where the sample returns are close to predicted returns.

Table 2. Average of computation times (in ms) of different algorithms for RSMDPs and Gurobi for RMDPs, and the ratios of Gurobi's computation times of RSMDPs to those of PDA, $\text{PDA}_{\text{block}}$, and $\text{PDA}_{\text{block+}}$.

	Computation times					Ratios of computation times		
$S = A$	Gurobi	PDA	$\text{PDA}_{\text{block}}$	$\text{PDA}_{\text{block+}}$	RMDP	Gurobi/ PDA	Gurobi/ $\text{PDA}_{\text{block}}$	Gurobi/ $\text{PDA}_{\text{block+}}$
10	8459.3	1521.5	1325.2	341.4	2818.6	5.6	6.4	24.8
13	60684.0	3104.8	2158.5	1132.7	4900.6	19.5	28.1	53.6
15	194114.5	4813.1	2241.7	2049.5	6604.5	40.3	86.6	94.7
17	552354.5	9305.4	2825.4	5286.8	9332.2	59.4	195.5	104.5

5.1. Improvements on Percentiles

In this section we compare the average and percentile performances of the four MDP models. In Figure 1, we present the case where the models are equipped with the "best" two parameters selected by cross validation. In most cases, RSMDPs outperform the other three in terms of both average and percentile performances.

5.2. Target-Oriented Feature

In this section, we contaminate the true transition kernel $\boldsymbol{p}^\star$ to obtain the "polluted transition kernel" $\boldsymbol{p}$; in particular, we quantify the level of contamination by $\|\boldsymbol{p} - \boldsymbol{p}^\star\|_1$. We first input the true transition kernel to the four MDP models, then use the polluted kernels to test the policies they yield. To test the models on their abilities of reaching their "predicted" (average) returns (with $\boldsymbol{p}^\star$ input) under contamination, we compute the difference between their sample returns and predicted returns, so that a non-negative difference indicates reaching the prediction.

Table 1 compares the differences between sample returns and predicted returns with 1000 samples for each value of parameters (*i.e.*, τ for RSMDPs, r for RMDPs and DRMDPs). Results show that RSMDPs have predicted returns that are much higher than those of RMDPs and DRMDPs, while their sample returns can still reach their predicted/target returns in most times (*i.e.*, have "high prediction accuracy"). Notice that the three models have the same predicted and sample returns in the third column of Table 1 because RSMDPs with $\tau = \text{Z}_{\text{N}}(\hat{\boldsymbol{p}})$ (by Proposition 3.1), and RMDPs and DRMDPs with $r = 0$ all degenerate to NMDPs. Figure 2 also illustrates that even with transition kernel samples that are far away from the true one, in contrast to RMDPs and DRMDPs where nearly half and even most of the sample returns fall below the prediction, RSMDPs still remain highly accurate in prediction, reflecting its target-oriented feature.

5.3. Scalability of Different Algorithms

In this section, we compare the computation times of our proposed first-order algorithms with the state-of-the-art solver Gurobi (academic license) (Gurobi Optimization, LLC, 2022). Table 2 reports the computation times of Gurobi when solving RMDPs (see Appendix D.2 for details of the model) and RSMDPs, as well as the proposed algorithms when solving RSMDPs. Results show that directly solving RSMDPs could be very challenging: the computation time increases rapidly with even a small increase in problem size, and is much larger than that of RMDP. This observation confirms our motivation on developing a tailored first-order method for this problem. Compared to Gurobi, the proposed algorithms remain scalable as the problem size increases. In particular, the computation time of PDA scales similarly to that of RMDP, and $\text{PDA}_{\text{block}}$ (PDA with dual

updates (12)) and PDA$_{\text{block}+}$ (PDA where the dual updating step follows the second strategy mentioned in Section 4.3) provide computationally cheap updates on the policy. This matches the advantages of first-order methods, which are used to solve large problems to moderate accuracy with high efficiency.

6. Conclusion and Future Work

We propose RSMDPs to compute satisficing policies under model ambiguity. In particular, the expected return is constrained to meet a user-specified target which is strictly imposed under the empirical transition kernel and softly imposed under all other possible transition kernels. RSMDPs minimize the magnitude of violation of those soft constraints with additional tolerance that depends on the distance of the associated transition kernel to the empirical one. We reformulate the RSMDP model into a min-max form where a scalable PDA algorithm is applicable. Experimental results showcase the robustness and target-oriented feature of RSMDPs as well as the scalability of the algorithm. A promising future work would be extending RSMDPs to the setting with continuous state and action spaces, for which discretization of the state and action spaces, as well as the Approximate Linear Programming (ALP) method (*e.g.*, (Abbasi-Yadkori et al., 2019)) may be needed.

Acknowledgements

We thank the anonymous reviewers for their comments. This work was supported, in part, by the General Research Fund Grant (Project No. 9043424) from the Hong Kong Research Grants Council, the NSFC/RGC Joint Research Scheme N_CityU105/21 from the Hong Kong Research Grants Council, the CityU Start-Up Grant (Project No. 9610481), the National Natural Science Foundation of China (Project No. 72032005), and Chow Sang Sang Group Research Fund sponsored by Chow Sang Sang Holdings International Limited (Project No. 9229076).

References

Abbasi-Yadkori, Y., Bartlett, P. L., Chen, X., and Malek, A. Large-scale Markov decision problems via the linear programming dual. *arXiv preprint arXiv:1901.01992*, 2019.

Ashok, P., Křetínskỳ, J., and Weininger, M. Pac statistical model checking for Markov decision processes and stochastic games. In *Computer Aided Verification: 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I 31*, pp. 497–519. Springer, 2019.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Bayraksan, G. and Love, D. K. Data-driven stochastic programming using phi-divergences. In *The operations research revolution*, pp. 1–19. INFORMS, 2015.

Behzadian, B., Petrik, M., and Ho, C. P. Fast algorithms for L_∞-constrained S-rectangular robust MDPs. *Advances in Neural Information Processing Systems*, 34, 2021a.

Behzadian, B., Russel, R. H., Petrik, M., and Ho, C. P. Optimizing percentile criterion using robust MDPs. In *International Conference on Artificial Intelligence and Statistics*, pp. 1009–1017. PMLR, 2021b.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*, volume 28. Princeton University Press, 2009.

Bertsimas, D., Shtern, S., and Sturt, B. A data-driven approach for multi-stage linear optimization. *Available at Optimization Online*, 2018.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

Brown, D., Niekum, S., and Petrik, M. Bayesian robust optimization for imitation learning. *Advances in Neural Information Processing Systems*, 33:2479–2491, 2020.

Chambolle, A. and Pock, T. On the ergodic convergence rates of a first-order primal-dual algorithm. *Mathematical Programming*, 159(1):253–287, 2016.

Chen, Z., Yu, P., and Haskell, W. Distributionally robust optimization for sequential decision-making. *Optimization*, 68(12):2397–2426, 2019.

Clement, J. G. and Kroer, C. First-order methods for Wasserstein distributionally robust MDP. In *International Conference on Machine Learning*, pp. 2010–2019. PMLR, 2021.

Delage, E. and Mannor, S. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.

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

Delgado, K., De Barros, L., Dias, D., and Sanner, S. Real-time dynamic programming for Markov decision processes with imprecise probabilities. *Artificial Intelligence*, 230:192–223, 2016.

Derman, E. and Mannor, S. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

Derman, E., Geist, M., and Mannor, S. Twice regularized MDPs and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

Doltsinis, S., Ferreira, P., and Lohse, N. An MDP model-based reinforcement learning approach for production station ramp-up optimization: Q-learning analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(9):1125–1138, 2014.

Esser, E., Zhang, X., and Chan, T. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

Ghavamzadeh, M., Petrik, M., and Chow, Y. Safe policy improvement by minimizing robust baseline regret. *Advances in Neural Information Processing Systems*, 29, 2016.

Goyal, V. and Grand-Clement, J. Robust Markov decision processes: beyond rectangularity. *Mathematics of Operations Research*, 2022.

Grand-Clément, J. and Kroer, C. First-order methods for Wasserstein distributionally robust MDPs. In *Proceedings of Machine Learning Research*, volume 139, pp. 2010–2019, 2021a.

Grand-Clément, J. and Kroer, C. Scalable first-order methods for robust MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12086–12094, 2021b.

Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual*, 2022. URL https://www.gurobi.com.

Hanasusanto, G. and Kuhn, D. Robust data-driven dynamic programming. *Advances in Neural Information Processing Systems*, 26, 2013.

Hansen, T., Miltersen, P., and Zwick, U. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.

He, B. and Yuan, X. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1): 119–149, 2012.

Ho, C. P., Petrik, M., and Wiesemann, W. Fast Bellman updates for robust MDPs. In *International Conference on Machine Learning*, pp. 1979–1988. PMLR, 2018.

Ho, C. P., Petrik, M., and Wiesemann, W. Partial policy iteration for L_1-robust Markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

Ho, C. P., Petrik, M., and Wiesemann, W. Robust phi-divergence MDPs. *arXiv preprint arXiv:2205.14202*, 2022.

Imani, M., Ghoreishi, S. F., and Braga-Neto, U. M. Bayesian control of large MDPs with unknown dynamics in data-poor environments. *Advances in Neural Information Processing Systems*, 31, 2018.

Iyengar, G. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

Kaufman, D. and Schaefer, A. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

Khanh, P. Q. and Quan, N. H. Versions of the Weierstrass theorem for bifunctions and solution existence in optimization. *SIAM Journal on Optimization*, 29(2):1502–1523, 2019.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4: 1107–1149, 2003.

Lobo, E. A., Ghavamzadeh, M., and Petrik, M. Soft-robust algorithms for batch reinforcement learning. *arXiv preprint arXiv:2011.14495*, 2020.

Long, D. Z., Sim, M., and Zhou, M. Robust satisficing. *Operations Research*, 71(1):61–82, 2023.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

Mannor, S., Mebel, O., and Xu, H. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

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

MOSEK ApS. *MOSEK Optimizer API for Python 9.3.20*, 2022. URL https://docs.mosek.com/latest/pythonapi/index.html.

Nilim, A. and El Ghaoui, L. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Petrik, M. Approximate dynamic programming by minimizing distributionally robust bounds. *arXiv preprint arXiv:1205.1782*, 2012.

Petrik, M. and Russel, R. Beyond confidence regions: tight Bayesian ambiguity sets for robust MDPs. *Advances in Neural Information Processing Systems*, 32, 2019.

Petrik, M. and Subramanian, D. RAAM: the benefits of robustness in approximating aggregated MDPs in reinforcement learning. *Advances in Neural Information Processing Systems*, 27, 2014.

Puterman, M. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Roy, A., Xu, H., and Pokutta, S. Reinforcement learning under model mismatch. *Advances in Neural Information Processing Systems*, 30, 2017.

Russel, R., Behzadian, B., and Petrik, M. Optimizing norm-bounded weighted ambiguity sets for robust MDPs. *arXiv preprint arXiv:1912.02696*, 2019.

Strehl, A. and Littman, M. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Strehl, A., Li, L., and Littman, M. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10(11), 2009.

Suilen, M., Simão, T. D., Parker, D., and Jansen, N. Robust anytime learning of Markov decision processes. *Advances in Neural Information Processing Systems*, 35:28790–28802, 2022.

Sutton, R. and Barto, A. *Reinforcement learning: an introduction*. MIT Press, 2018.

Taleghan, M., Dietterich, T., Crowley, M., Hall, K., and Albers, J. PAC optimal MDP planning with application to invasive species management. *Journal of Machine Learning Research*, 16, 2015.

Tamar, A., Mannor, S., and Xu, H. Scaling up robust MDPs using function approximation. In *International Conference on Machine Learning*, pp. 181–189. PMLR, 2014.

Truhar, N. and Veselić, K. An efficient method for estimating the optimal dampers' viscosity for linear vibrating systems using Lyapunov equation. *SIAM Journal on Matrix Analysis and Applications*, 31(1):18–39, 2009.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Wang, Y. and Zou, S. Policy gradient method for robust reinforcement learning. *arXiv preprint arXiv:2205.07344*, 2022.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. Inequalities for the L_1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

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

Xie, W. Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ Wasserstein ball. *Operations Research Letters*, 48(4):513–523, 2020.

Xu, H. and Mannor, S. The robustness-performance tradeoff in Markov decision processes. *Advances in Neural Information Processing Systems*, 19, 2006.

Xu, H. and Mannor, S. Parametric regret in uncertain Markov decision processes. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pp. 3606–3613. IEEE, 2009.

Xu, H. and Mannor, S. Distributionally robust Markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

Yang, I. A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169, 2017.

A. Technical Results and Proofs

A.1. Proofs of Results in Section 3

Proof of Proposition 3.1 By letting $\boldsymbol{p} = \hat{\boldsymbol{p}}$ in the first S constraints in (4), on the one hand, one can observe that the feasible set of (4) is a subset of the optimal solution set of (2), where conclusion (*i*) follows; on the other hand, one can also observe that for all feasible $\boldsymbol{u}$ of (4), the maximal reachable value of $\boldsymbol{r}^\top \boldsymbol{u}$ is at most $\mathrm{Z_N}(\hat{\boldsymbol{p}})$, which implies the infeasibility when $\tau > \mathrm{Z_N}(\hat{\boldsymbol{p}})$. □

Proof of Theorem 3.2 Focus on the first S constraints in problem (4). For every $s \in \mathcal{S}$, the s-th one equivalent to

$$\mathbf{e}^\top \boldsymbol{u}_s - d_s \leq \min_{\boldsymbol{p} \in \mathcal{P}} \left\{ \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u} + k_s \cdot \|\boldsymbol{p} - \hat{\boldsymbol{p}}\| \right\}.$$

The optimization problem on the right-hand side is

$$\begin{array}{ll} \min & \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u} + k_s \cdot \|\boldsymbol{p} - \hat{\boldsymbol{p}}\| \\ \text{s.t.} & \boldsymbol{B}\boldsymbol{p} = \mathbf{e} \\ & \boldsymbol{p} \in \mathbb{R}_+^{S \cdot A \cdot S}, \end{array}$$

where $\boldsymbol{B}\boldsymbol{p} = \mathbf{e}$ is a compact form of $\sum_{s' \in \mathcal{S}} p_{s,a,s'} = 1 \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$. The equivalent min-max form of this problem is

$$\min_{\boldsymbol{p}} \max_{\boldsymbol{\beta} \geq \mathbf{0}, \boldsymbol{\alpha}} \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u} + k_s \cdot \|\boldsymbol{p} - \hat{\boldsymbol{p}}\| + \boldsymbol{\alpha}^\top (\boldsymbol{B}\boldsymbol{p} - \mathbf{e}) - \boldsymbol{\beta}^\top \boldsymbol{p},$$

whose dual problem[2] is

$$\begin{array}{l} \max\limits_{\boldsymbol{\beta} \geq \mathbf{0}, \boldsymbol{\alpha}} \min\limits_{\boldsymbol{p}} \boldsymbol{p}^\top (\boldsymbol{Q}_s \boldsymbol{u} + \boldsymbol{B}^\top \boldsymbol{\alpha} - \boldsymbol{\beta}) + k_s \cdot \|\boldsymbol{p} - \hat{\boldsymbol{p}}\| - \boldsymbol{\alpha}^\top \mathbf{e} \\ = \max\limits_{\boldsymbol{\beta} \geq \mathbf{0}, \boldsymbol{\alpha}} -\boldsymbol{\alpha}^\top \mathbf{e} - \max\limits_{\boldsymbol{p}} \left\{ \boldsymbol{p}^\top (\boldsymbol{\beta} - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}) - k_s \cdot \|\boldsymbol{p} - \hat{\boldsymbol{p}}\| \right\}. \end{array}$$

By the technique of convex conjugate, we have its equivalent reformulation as

$$\begin{array}{ll} \max & -\boldsymbol{\alpha}^\top \mathbf{e} - \hat{\boldsymbol{p}}^\top (\boldsymbol{\beta} - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}) \\ \text{s.t.} & \|\boldsymbol{\beta} - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}\|_* \leq k_s \\ & \boldsymbol{\alpha} \in \mathbb{R}^{S \cdot A}, \ \boldsymbol{\beta} \in \mathbb{R}_+^{S \cdot A \cdot S}. \end{array}$$

Therefore, for all $s \in \mathcal{S}$, the s-th constraint in (4) can be reformulated as

$$\left\{ \begin{array}{l} \mathbf{e}^\top \boldsymbol{u}_s - d_s \leq -\boldsymbol{\alpha}_s^\top \mathbf{e} - \hat{\boldsymbol{p}}^\top (\boldsymbol{\beta}_s - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}_s) \\ \|\boldsymbol{\beta}_s - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}_s\|_* \leq k_s \\ \boldsymbol{\alpha}_s \in \mathbb{R}^{S \cdot A}, \ \boldsymbol{\beta}_s \in \mathbb{R}_+^{S \cdot A \cdot S}. \end{array} \right.$$

Now we can re-express problem (4) as

$$\begin{array}{lll} \min & \boldsymbol{w}^\top \boldsymbol{k} & \\ \text{s.t.} & \mathbf{e}^\top \boldsymbol{u}_s - d_s \leq -\boldsymbol{\alpha}_s^\top \mathbf{e} - \hat{\boldsymbol{p}}^\top (\boldsymbol{\beta}_s - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}_s) & \forall s \in \mathcal{S} \\ & \|\boldsymbol{\beta}_s - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}_s\|_* \leq k_s & \forall s \in \mathcal{S} \\ & \boldsymbol{r}^\top \boldsymbol{u} \geq \tau & \\ & \boldsymbol{u} \in \mathbb{R}_+^{S \cdot A}, \ \boldsymbol{k} \in \mathbb{R}_+^S, \boldsymbol{\alpha}_s \in \mathbb{R}^{S \cdot A}, \ \boldsymbol{\beta}_s \in \mathbb{R}_+^{S \cdot A \cdot S} & \forall s \in \mathcal{S}, \end{array}$$

which is equivalent to

$$\begin{array}{lll} \min & \sum\limits_{s \in \mathcal{S}} w_s \cdot \|\boldsymbol{\beta}_s - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}_s\|_* & \\ \text{s.t.} & \mathbf{e}^\top \boldsymbol{u}_s - d_s \leq -\boldsymbol{\alpha}_s^\top \mathbf{e} - \hat{\boldsymbol{p}}^\top (\boldsymbol{\beta}_s - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}_s) & \forall s \in \mathcal{S} \\ & \boldsymbol{r}^\top \boldsymbol{u} \geq \tau & \\ & \boldsymbol{u} \in \mathbb{R}_+^{S \cdot A}, \ \boldsymbol{\alpha}_s \in \mathbb{R}^{S \cdot A}, \ \boldsymbol{\beta}_s \in \mathbb{R}_+^{S \cdot A \cdot S} & \forall s \in \mathcal{S}. \end{array}$$

Our conclusion then follows from the fact that $\boldsymbol{B}\hat{\boldsymbol{p}} = \mathbf{e}$. □

[2]Here, strong duality follows from the fact that the primal problem has only the collection of constraints $\boldsymbol{p}_{s,a} \in \Delta^S \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.

A.2. Proofs of Results in Section 4

Lemma A.1. *When $w \in \mathbb{R}_{++}$, strong duality holds for the following convex optimization problem:*

$$\max_{(\lambda,\boldsymbol{\theta})\in\mathcal{V}_q(w)} \lambda(\mathbf{e}^\top \boldsymbol{u} - d) - \boldsymbol{\theta}^\top \boldsymbol{Q}\boldsymbol{u}, \tag{13}$$

where $\boldsymbol{u} \in \mathbb{R}^A,\ d \in \mathbb{R}$, $\boldsymbol{Q} \in \mathbb{R}^{S\cdot A\cdot S\times S\cdot A}$ and $q \geq 1$ are constants, and

$$\mathcal{V}_q(w) := \left\{ (\lambda,\boldsymbol{\theta}) \in \mathbb{R}_+ \times \mathbb{R}_+^{S\cdot A\cdot S} \;\middle|\; \begin{array}{l} \|\boldsymbol{\theta} - \lambda\cdot\hat{\boldsymbol{p}}\|_q \leq w \\ \lambda\cdot\mathbf{e} = \boldsymbol{B}\boldsymbol{\theta} \end{array} \right\}.$$

Proof of Lemma A.1 Notice that, since we have $\boldsymbol{B}\hat{\boldsymbol{p}} = \mathbf{e}$ by the definition of $\boldsymbol{B}$, by letting λ be any nonnegative number and $\boldsymbol{\theta} = \lambda\cdot\hat{\boldsymbol{p}}$, we can have $(\lambda,\boldsymbol{\theta})$ as the feasible solution of (13) (which is also strictly feasible in the first inequality constraint in $\mathcal{V}_q(w)$). If $\hat{\boldsymbol{p}} > \mathbf{0}$, then $(\lambda, \lambda\cdot\hat{\boldsymbol{p}})$ with $\lambda > 0$ is the strictly feasible solution we want. Otherwise, if $\hat{p}_{\bar{s},\bar{a},\bar{s}'} = 0$ for some $(\bar{s},\bar{a},\bar{s}') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, then $(\lambda,\boldsymbol{\theta})$ with $\lambda > 0$ and $\boldsymbol{\theta} = \lambda\cdot\hat{\boldsymbol{p}}$ will be a solution of (13) that is not strictly feasible because $\theta_{\bar{s},\bar{a},\bar{s}'} = 0$.

Now we demonstrate how we can construct a strictly feasible solution of (13). First, observe that the second set of constraints in $\mathcal{V}_q(w)$ of (13) is equivalent to

$$\mathbf{e}^\top\boldsymbol{\theta}_{s,a} = \lambda\ \forall(s,a) \in \mathcal{S}\times\mathcal{A}. \tag{14}$$

Since $\boldsymbol{\theta} \geq \mathbf{0}$ and $\lambda > 0$, there must exists some $\bar{s}'' \in \mathcal{S}$ such that $\theta_{\bar{s},\bar{a},\bar{s}''} > 0$. Let

$$\varepsilon = \min\{\theta_{\bar{s},\bar{a},\bar{s}''}/2,\ w/(2\cdot\|\mathbf{e}_{\bar{s},\bar{a},\bar{s}'} - \mathbf{e}_{\bar{s},\bar{a},\bar{s}''}\|)\},$$

where $\mathbf{e}_{\bar{s},\bar{a},\bar{s}'}$ and $\mathbf{e}_{\bar{s},\bar{a},\bar{s}'}$ are two standard bases of $\mathbb{R}^{S\cdot A\cdot S}$. One can then easily verify that $(\lambda,\bar{\boldsymbol{\theta}})$ with

$$\bar{\boldsymbol{\theta}} = \boldsymbol{\theta} + \varepsilon\cdot(\mathbf{e}_{\bar{s},\bar{a},\bar{s}'} - \mathbf{e}_{\bar{s},\bar{a},\bar{s}''})$$

is a feasible solution of (13) which remains strictly feasible in the first inequality constraint, while $\theta_{\bar{s},\bar{a},\bar{s}''}$ remains strictly positive and $\theta_{\bar{s},\bar{a},\bar{s}'}$ becomes strictly positive. By going through a similar procedure iteratively, we can finally construct a strictly feasible solution of (13). □

Proof of Proposition 4.1 Notice that, by Theorem 3.2, the RSMDP model (4) is equivalent to

$$\begin{array}{lll} \min & \displaystyle\sum_{s\in\mathcal{S}} w_s\cdot\|\boldsymbol{y}_s\|_* & \\ \text{s.t.} & \mathbf{e}^\top\boldsymbol{u}_s - d_s \leq -\boldsymbol{\alpha}_s^\top\mathbf{e} - \hat{\boldsymbol{p}}^\top\boldsymbol{y}_s & \forall s\in\mathcal{S} \\ & \boldsymbol{y}_s = \boldsymbol{\beta}_s - \boldsymbol{Q}_s\boldsymbol{u} - \boldsymbol{B}^\top\boldsymbol{\alpha}_s & \forall s\in\mathcal{S} \\ & \boldsymbol{r}^\top\boldsymbol{u} \geq \tau & \\ & \boldsymbol{u}\in\mathbb{R}_+^{S\cdot A},\ \boldsymbol{\alpha}_s\in\mathbb{R}^{S\cdot A},\ \boldsymbol{\beta}_s\in\mathbb{R}_+^{S\cdot A\cdot S},\ \boldsymbol{y}_s\in\mathbb{R}^{S\cdot A\cdot S} & \forall s\in\mathcal{S}. \end{array} \tag{15}$$

Since the first set of constraints in (4) implies that $\mathbf{e}^\top\boldsymbol{u}_s - \hat{\boldsymbol{p}}^\top\boldsymbol{Q}_s\boldsymbol{u} - d_s \leq 0\ \forall s\in\mathcal{S}$, we have an equivalent reformulation of (15):

$$\begin{array}{lll} \min & \displaystyle\sum_{s\in\mathcal{S}} w_s\cdot\|\boldsymbol{y}_s\|_* & \\ \text{s.t.} & \mathbf{e}^\top\boldsymbol{u}_s - d_s \leq -\boldsymbol{\alpha}_s^\top\mathbf{e} - \hat{\boldsymbol{p}}^\top\boldsymbol{y}_s & \forall s\in\mathcal{S} \\ & \boldsymbol{y}_s = \boldsymbol{\beta}_s - \boldsymbol{Q}_s\boldsymbol{u} - \boldsymbol{B}^\top\boldsymbol{\alpha}_s & \forall s\in\mathcal{S} \\ & \boldsymbol{r}^\top\boldsymbol{u} \geq \tau & \\ & \mathbf{e}^\top\boldsymbol{u}_s - \hat{\boldsymbol{p}}^\top\boldsymbol{Q}_s\boldsymbol{u} - d_s \leq 0 & \forall s\in\mathcal{S} \\ & \boldsymbol{u}\in\mathbb{R}_+^{S\cdot A},\ \boldsymbol{\alpha}_s\in\mathbb{R}^{S\cdot A},\ \boldsymbol{\beta}_s\in\mathbb{R}_+^{S\cdot A\cdot S},\ \boldsymbol{y}_s\in\mathbb{R}^{S\cdot A\cdot S} & \forall s\in\mathcal{S}. \end{array} \tag{16}$$

Notice that for any $\boldsymbol{u} \in \mathbb{R}_+^{S\cdot A}$ that satisfies $\boldsymbol{r}^\top\boldsymbol{u} \geq \tau$ and $\mathbf{e}^\top\boldsymbol{u}_s - \hat{\boldsymbol{p}}^\top\boldsymbol{Q}_s\boldsymbol{u} - d_s \leq 0\ \forall s\in\mathcal{S}$, $(\boldsymbol{\alpha}_s,\boldsymbol{\beta}_s,\boldsymbol{y}_s) = (\mathbf{0},\mathbf{0},-\boldsymbol{Q}_s\boldsymbol{u}),\ s\in\mathcal{S}$ is feasible in (16), whose objective value has a lower bound 0. We thus can re-express (16) in

a min-min form:

$$\begin{aligned}
\min_{\boldsymbol{u}} \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{y}} \ & \sum_{s\in\mathcal{S}} w_s \cdot \|\boldsymbol{y}_s\|_* \\
\text{s.t.} \quad & \mathbf{e}^\top \boldsymbol{u}_s - d_s \le -\boldsymbol{\alpha}_s^\top \mathbf{e} - \hat{\boldsymbol{p}}^\top \boldsymbol{y}_s && \forall s \in \mathcal{S} \\
& \boldsymbol{y}_s = \boldsymbol{\beta}_s - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}_s && \forall s \in \mathcal{S} \\
& \boldsymbol{r}^\top \boldsymbol{u} \ge \tau \\
& \mathbf{e}^\top \boldsymbol{u}_s - \hat{\boldsymbol{p}}^\top \boldsymbol{Q}_s \boldsymbol{u} - d_s \le 0 && \forall s \in \mathcal{S} \\
& \boldsymbol{u} \in \mathbb{R}_+^{S\cdot A},\ \boldsymbol{\alpha}_s \in \mathbb{R}^{S\cdot A},\ \boldsymbol{\beta}_s \in \mathbb{R}_+^{S\cdot A\cdot S},\ \boldsymbol{y}_s \in \mathbb{R}^{S\cdot A\cdot S} && \forall s \in \mathcal{S},
\end{aligned} \tag{17}$$

Fix any $\boldsymbol{u} \in \mathbb{R}_+^{S\cdot A}$ that satisfies $\boldsymbol{r}^\top \boldsymbol{u} \ge \tau$ and $\mathbf{e}^\top \boldsymbol{u}_s - \hat{\boldsymbol{p}}^\top \boldsymbol{Q}_s \boldsymbol{u} - d_s \le 0\ \forall s \in \mathcal{S}$, we have the dual of the inner minimization problem of (17) as follows:

$$\begin{aligned}
\max_{\boldsymbol{\lambda},\boldsymbol{\theta}} \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{y}} \ & \sum_{s\in\mathcal{S}} w_s \cdot \|\boldsymbol{y}_s\|_* + \sum_{s\in\mathcal{S}} \lambda_s (\mathbf{e}^\top \boldsymbol{u}_s - d_s + \boldsymbol{\alpha}_s^\top \mathbf{e} + \hat{\boldsymbol{p}}^\top \boldsymbol{y}_s) + \sum_{s\in\mathcal{S}} \boldsymbol{\theta}_s^\top (\boldsymbol{\beta}_s - \boldsymbol{Q}_s \boldsymbol{u} - \boldsymbol{B}^\top \boldsymbol{\alpha}_s - \boldsymbol{y}_s) \\
\text{s.t.} \quad & \boldsymbol{\lambda} \in \mathbb{R}_+^S,\ \boldsymbol{\theta} \in \mathbb{R}_+^{S\cdot S\cdot A\cdot S},\ \boldsymbol{\alpha}_s \in \mathbb{R}^{S\cdot A},\ \boldsymbol{\beta}_s \in \mathbb{R}_+^{S\cdot A\cdot S},\ \boldsymbol{y}_s \in \mathbb{R}^{S\cdot A\cdot S} && \forall s \in \mathcal{S}.
\end{aligned} \tag{18}$$

Here, for the inner minimization problem of (18), we have

$$\min_{\boldsymbol{y}} \sum_{s\in\mathcal{S}} \left\{ w_s \cdot \|\boldsymbol{y}_s\|_* + \boldsymbol{y}_s^\top (\lambda_s \cdot \hat{\boldsymbol{p}} - \boldsymbol{\theta}_s) \right\} = \begin{cases} 0 & \|\boldsymbol{\theta}_s - \lambda_s \cdot \hat{\boldsymbol{p}}\| \le w_s\ \forall s \in \mathcal{S} \\ -\infty \text{ otherwise,} \end{cases}$$

$$\min_{\boldsymbol{\alpha}} \sum_{s\in\mathcal{S}} \boldsymbol{\alpha}_s^\top (\lambda_s \cdot \mathbf{e} - \boldsymbol{B}\boldsymbol{\theta}_s) = \begin{cases} 0 & \lambda_s \cdot \mathbf{e} - \boldsymbol{B}\boldsymbol{\theta}_s = \mathbf{0}\ \forall s \in \mathcal{S} \\ -\infty \text{ otherwise,} \end{cases}$$

and

$$\min_{\boldsymbol{\beta}} \sum_{s\in\mathcal{S}} \boldsymbol{\theta}_s^\top \boldsymbol{\beta}_s = \begin{cases} 0 & \boldsymbol{\theta}_s \ge \mathbf{0}\ \forall s \in \mathcal{S} \\ -\infty \text{ otherwise.} \end{cases}$$

Therefore we have the equivalent reformulation of (18) as follows:

$$\begin{aligned}
\max \ & \sum_{s\in\mathcal{S}} \left\{ \lambda_s (\mathbf{e}^\top \boldsymbol{u}_s - d_s) - \boldsymbol{\theta}_s^\top \boldsymbol{Q}_s \boldsymbol{u} \right\} \\
\text{s.t.} \ & \|\boldsymbol{\theta}_s - \lambda_s \hat{\boldsymbol{p}}\| \le w_s && \forall s \in \mathcal{S} \\
& \lambda_s \cdot \mathbf{e} = \boldsymbol{B}\boldsymbol{\theta}_s && \forall s \in \mathcal{S} \\
& \boldsymbol{\lambda} \in \mathbb{R}_+^S,\ \boldsymbol{\theta}_s \in \mathbb{R}_+^{S\cdot A\cdot S} && \forall s \in \mathcal{S},
\end{aligned} \tag{19}$$

where a strictly feasible solution exists by Lemma A.1. Therefore we claim that strong duality between the inner minimization problem of (17) and its dual problem (19) holds by Slater's condition. We thus have an equivalent minimax formulation of (17) is as follows:

$$\begin{aligned}
\min_{\boldsymbol{u}} \max_{\boldsymbol{\lambda},\boldsymbol{\theta}} \ & \sum_{s\in\mathcal{S}} \left\{ \lambda_s (\mathbf{e}^\top \boldsymbol{u}_s - d_s) - \boldsymbol{\theta}_s^\top \boldsymbol{Q}_s \boldsymbol{u} \right\} \\
\text{s.t.} \quad & \|\boldsymbol{\theta}_s - \lambda_s \hat{\boldsymbol{p}}\| \le w_s && \forall s \in \mathcal{S} \\
& \lambda_s \cdot \mathbf{e} = \boldsymbol{B}\boldsymbol{\theta}_s && \forall s \in \mathcal{S} \\
& \mathbf{e}^\top \boldsymbol{u}_s - \hat{\boldsymbol{p}}^\top \boldsymbol{Q}_s \boldsymbol{u} - d_s \le 0 && \forall s \in \mathcal{S} \\
& \boldsymbol{r}^\top \boldsymbol{u} \ge \tau \\
& \boldsymbol{u} \in \mathbb{R}_+^{S\cdot A},\ \boldsymbol{\lambda} \in \mathbb{R}_+^S,\ \boldsymbol{\theta}_s \in \mathbb{R}_+^{S\cdot A\cdot S} && \forall s \in \mathcal{S}.
\end{aligned}$$

It remains to argue that it is free to eliminate the third collection of constraints. To this end, notice that, if we choose a feasible $\boldsymbol{u}$ in (8) that satisfies $\mathbf{e}^\top \boldsymbol{u}_s - \hat{\boldsymbol{p}}^\top \boldsymbol{Q}_s \boldsymbol{u} - d_s > 0$ for some $s \in \mathcal{S}$, the inner maximum will approach $+\infty$ since we can choose a feasible solution $(\lambda_s, \boldsymbol{\theta}_s)$ that satisfies $\boldsymbol{\theta}_s = \lambda_s \cdot \hat{\boldsymbol{p}}$ with arbitrarily large $\lambda_s > 0$, by which the objective value is expanded to $+\infty$. □

Theorem A.2. *[theorem 1, (Chambolle & Pock, 2016)] Let $(\boldsymbol{u}^k, (\boldsymbol{\lambda}^k, \boldsymbol{\theta}^k)),\ k = 0, 1, 2, \cdots, K$ be a sequence generated by Algorithm 1. If the stepsize parameters $\sigma,\ \nu > 0$ satisfy*

$$\frac{1}{2\nu} \cdot \|\boldsymbol{u} - \boldsymbol{u}'\|_2^2 + \frac{1}{2\sigma} \cdot \left\| \begin{matrix} \boldsymbol{\lambda} - \boldsymbol{\lambda}' \\ \boldsymbol{\theta} - \boldsymbol{\theta}' \end{matrix} \right\|_2^2 - \langle \boldsymbol{C}(\boldsymbol{u} - \boldsymbol{u}'), ((\boldsymbol{\lambda} - \boldsymbol{\lambda}')^\top, (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top)^\top \rangle \ge 0 \tag{20}$$

for any $\boldsymbol{u},\ \boldsymbol{u}' \in \mathbb{R}^{S\cdot A}$ *and* $(\boldsymbol{\lambda},\boldsymbol{\theta}),\ (\boldsymbol{\lambda}',\boldsymbol{\theta}') \in \mathbb{R}^S \times \mathbb{R}^{S\cdot S\cdot A\cdot S}$. *Then it holds for any feasible solution* $(\boldsymbol{u},(\boldsymbol{\lambda},\boldsymbol{\theta}))$ *of problem* (9) *that*

$$\begin{aligned}&\sum_{s\in\mathcal{S}}\{(\lambda_s\cdot(\mathbf{e}^\top\bar{\boldsymbol{u}}_s^K-d_s)-\boldsymbol{\theta}_s^\top\boldsymbol{Q}_s\boldsymbol{u})-(\bar{\lambda}_s^K\cdot(\mathbf{e}^\top\boldsymbol{u}_s-d_s)-\bar{\boldsymbol{\theta}}_s^{K\top}\boldsymbol{Q}_s\boldsymbol{u}))\}\\ \le\ &\frac{1}{K}\cdot\left(\frac{1}{2\nu}\cdot\|\boldsymbol{u}-\boldsymbol{u}_0\|_2^2+\frac{1}{2\sigma}\cdot\left\|\begin{matrix}\boldsymbol{\lambda}-\boldsymbol{\lambda}_0\\ \boldsymbol{\theta}-\boldsymbol{\theta}_0\end{matrix}\right\|_2^2-\langle\boldsymbol{C}(\boldsymbol{u}-\boldsymbol{u}_0),((\boldsymbol{\lambda}-\boldsymbol{\lambda}_0)^\top,(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top)^\top\rangle\right),\end{aligned}$$

where $\bar{\boldsymbol{u}}^K=\frac{1}{K}\sum_{k\in[K]}\boldsymbol{u}^k$, $\bar{\boldsymbol{\lambda}}^K=\frac{1}{K}\sum_{k\in[K]}\boldsymbol{\lambda}^k$ *and* $\bar{\boldsymbol{\theta}}^K=\frac{1}{K}\sum_{k\in[K]}\boldsymbol{\theta}^k$.

Proof of Proposition 4.3 We first rewrite $\mathfrak{P}(\boldsymbol{\lambda},\boldsymbol{\theta};\hat{\boldsymbol{u}})$ as

$$\begin{array}{ll}\arg\min & \boldsymbol{a}^\top\boldsymbol{u}+\dfrac{1}{2\nu}\cdot\boldsymbol{u}^\top\boldsymbol{u}\\ \text{s.t.} & \boldsymbol{r}^\top\boldsymbol{u}\ge\tau\\ & \boldsymbol{u}\in\mathbb{R}_+^{S\cdot A},\end{array}\tag{21}$$

where the coefficient vector $\boldsymbol{a}\in\mathbb{R}^{S\cdot A}$ satisfies $\boldsymbol{a}^\top\boldsymbol{u}=\sum_{s\in\mathcal{S}}\left(\lambda_s\cdot\mathbf{e}^\top\boldsymbol{u}_s-\boldsymbol{\theta}_s^\top\boldsymbol{Q}_s\boldsymbol{u}\right)-\frac{1}{\nu}\cdot\hat{\boldsymbol{u}}^\top\boldsymbol{u}$. Introducing dual variables $\zeta\in\mathbb{R}_+$ and $\boldsymbol{\kappa}\in\mathbb{R}_+^{S\cdot A}$, the Lagrangian dual function of this problem is

$$L(\boldsymbol{u},\zeta,\boldsymbol{\kappa})=\boldsymbol{a}^\top\boldsymbol{u}+\frac{1}{2\nu}\cdot\boldsymbol{u}^\top\boldsymbol{u}+\zeta\cdot(\tau-\boldsymbol{r}^\top\boldsymbol{u})-\boldsymbol{\kappa}^\top\boldsymbol{u}.$$

Since $\nu>0$, problem (21) is a convex optimization problem. The KKT conditions (*e.g.*, (Boyd & Vandenberghe, 2004)) thus are sufficient conditions of the optimality for the primal and dual solutions, which are

$$\left\{\begin{array}{ll}\boldsymbol{r}^\top\boldsymbol{u}\ge\tau & \\ \boldsymbol{u}\ge\mathbf{0} & \\ \zeta\ge0 & \\ \boldsymbol{\kappa}\ge\mathbf{0} & \\ \zeta\cdot(\tau-\boldsymbol{r}^\top\boldsymbol{u})=0 & \\ \kappa_iu_i=0 & \forall i\in[SA]\\ \nabla_{\boldsymbol{u}}L(\boldsymbol{u},\zeta,\boldsymbol{\kappa})=\dfrac{1}{\nu}\cdot\boldsymbol{u}+\boldsymbol{a}-\zeta\cdot\boldsymbol{r}-\boldsymbol{\kappa}=\mathbf{0}. & \end{array}\right.$$

By these conditions, for $\zeta=0$ we have

$$\left\{\begin{array}{l}\boldsymbol{r}^\top\boldsymbol{u}\ge\tau\\ u_i=\left\{\begin{array}{ll}-\nu a_i & \forall i\in[SA]:\kappa_i=0\\ 0 & \forall i\in[SA]:\kappa_i\ne0\end{array}\right.\\ \boldsymbol{u}\ge\mathbf{0};\end{array}\right.$$

and for $\zeta>0$ we have

$$\left\{\begin{array}{l}\boldsymbol{r}^\top\boldsymbol{u}=\tau\\ u_i=\left\{\begin{array}{ll}\nu\cdot(\zeta r_i-a_i) & \forall i\in[SA]:\kappa_i=0\\ 0 & \forall i\in[SA]:\kappa_i\ne0\end{array}\right.\\ \boldsymbol{u}\ge\mathbf{0},\end{array}\right.$$

where it is sufficient to find the optimal $\zeta^\star\in\mathbb{R}_+$ which is the solution to the equation $\Phi(\zeta):=\sum_{i\in[SA]}r_i\nu\cdot[\zeta r_i-a_i]_+=\tau$, and then obtain optimal $u_i^\star=[\nu\cdot(\zeta^\star r_i-a_i)]_+\ \forall i\in[SA]$. Notice that $\Phi(\zeta)$ is a piecewise linear function so that we can locate all its breakpoints $\alpha_i=\frac{a_i}{r_i}:i\in\mathcal{I}$ with $\mathcal{I}=\{i\in[SA]\mid a_i>0,r_i>0\}$ and then sort them from smallest to largest as $\zeta_{i_1}\le\zeta_{i_2}\le\cdots\le\zeta_{i_{|\mathcal{I}|}}$. As Φ is non-decreasing on $[0,+\infty)$, by searching the intervals $[0,\zeta_{i_1}],[\zeta_{i_1},\zeta_{i_2}],\cdots,[\zeta_{i_I},+\infty)$ sequentially in an ascending order, we can obtain optimal $\zeta^\star$ and $\boldsymbol{u}^\star$.

The time complexity is dominated by the breakpoint sorting, which is $\mathcal{O}(SA\cdot\log(SA))$. □

We provide the pseudocode for solving problem (21) in Algorithm 2.

Algorithm 2 Interval Search Algorithm for Problem (21)

if $\sum_{i\in[SA]} r_i[-\nu a_i]_+ \geq \tau$ **then**
  return $\zeta^\star = 0$ and $u_i^\star = [-\nu a_i]_+ \ \forall i \in [SA]$
else
  Compute all the breakpoints $\zeta_i = \frac{a_i}{r_i} : i \in \mathcal{I}$ with $\mathcal{I} = \{i \in [SA] \mid a_i > 0, r_i > 0\}$;
  Sort the breakpoints from smallest to largest as $\zeta_{i_1} \leq \zeta_{i_2} \leq \cdots \leq \zeta_{i_{|\mathcal{I}|}}$;
  Compute the initial index set $\bar{\mathcal{I}} \leftarrow [SA] \setminus \mathcal{I}$;
  for $k = 1$ **to** $|\mathcal{I}|$ **do**
    if $\tau \leq \sum_{j\in\bar{\mathcal{I}}} \nu(\zeta_{i_k} r_j - a_j) \cdot r_j$ **then**
      return $\zeta^\star = \frac{\frac{\tau}{\nu} + \sum_{j\in\bar{\mathcal{I}}} a_j r_j}{\sum_{j\in\bar{\mathcal{I}}} r_j r_j}$ and $u_i^\star = [\nu(\zeta^\star r_i - a_i)]_+ \ \forall i \in [SA]$
    else
      $\bar{\mathcal{I}} \leftarrow \bar{\mathcal{I}} \cup \{i_k\}$;
    end if
  end for
  return $\zeta^\star = \frac{\frac{\tau}{\nu} + \boldsymbol{a}^\top \boldsymbol{r}}{\boldsymbol{r}^\top \boldsymbol{r}}$ and $u_i^\star = [\nu(\zeta^\star r_i - a_i)]_+ \ \forall i \in [SA]$
end if
Output: Solutions $\zeta^\star$ and $\boldsymbol{u}^\star$

Lemma A.3. *For the optimal solution $(\lambda^\star, \boldsymbol{\theta}^\star)$ of $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$, the upper bound for $\lambda^\star$ is*

$$\lambda^\star \leq \hat{\lambda} + \sigma \cdot \bigg(-(d_{\hat{s}} - \mathbf{e}^\top \boldsymbol{u}_{\hat{s}}) + \bigg\{ (d_{\hat{s}} - \mathbf{e}^\top \boldsymbol{u}_{\hat{s}})^2 + \frac{2}{\sigma} \cdot \bigg(\frac{\sigma}{2} \cdot \bigg(\boldsymbol{z}^\top \boldsymbol{z} - \frac{2}{\sigma} \cdot \boldsymbol{z}^\top \hat{\boldsymbol{\theta}} \bigg) + \sum_{(s,a,s')\in\mathcal{Z}_+} z_{s,a,s'} (\hat{\lambda} \hat{p}_{s,a,s'} + w_{\hat{s}}) + \sum_{(s,a,s')\in\mathcal{Z}_-} z_{s,a,s'} [\hat{\lambda} \hat{p}_{s,a,s'} - w_{\hat{s}}]_+ \bigg) \bigg\}^{\frac{1}{2}} \bigg),$$

where $\boldsymbol{z} = \boldsymbol{Q}_{\hat{s}} \boldsymbol{u}$, $\mathcal{Z}_+ = \{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mid z_{s,a,s'} > 0\}$ and $\mathcal{Z}_- = \{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mid z_{s,a,s'} < 0\}$.

Proof. By Algorithm 1, $(\hat{\lambda}, \hat{\boldsymbol{\theta}})$ is a feasible solution of $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$. Thus by the first constraint in $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$, we have $[\hat{\lambda} \hat{p}_{s,a,s'} - w_{\hat{s}}]_+ \leq \hat{\theta}_{s,a,s'} \leq \hat{\lambda} \hat{p}_{s,a,s'} + w_{\hat{s}} \ \forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Plugging $(\hat{\lambda}, \hat{\boldsymbol{\theta}})$ in the objective function of $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$, we obtain an upper bound of its optimal value

$$\begin{aligned} &(d_{\hat{s}} - \mathbf{e}^\top \boldsymbol{u}_{\hat{s}}) \cdot \hat{\lambda} + \boldsymbol{z}^\top \hat{\boldsymbol{\theta}} \\ \leq\ &(d_{\hat{s}} - \mathbf{e}^\top \boldsymbol{u}_{\hat{s}}) \cdot \hat{\lambda} + \sum_{(s,a,s')\in\mathcal{Z}_+} z_{s,a,s'} (\hat{\lambda} \hat{p}_{s,a,s'} + w_{\hat{s}}) + \sum_{(s,a,s')\in\mathcal{Z}_-} z_{s,a,s'} [\hat{\lambda} \hat{p}_{s,a,s'} - w_{\hat{s}}]_+. \end{aligned}$$

Let $(\lambda^\star, \boldsymbol{\theta}^\star)$ with $\lambda^\star = \hat{\lambda} + \Delta\lambda$ be the optimal solution of $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$ and $(\hat{\lambda} + \Delta\lambda, \boldsymbol{\theta})$ be a feasible solution of $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$. Plugging $(\hat{\lambda} + \Delta\lambda, \boldsymbol{\theta})$ in $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$, we consider the problem:

$$\begin{aligned} \min\ & (d_{\hat{s}} - \mathbf{e}^\top \boldsymbol{u}_{\hat{s}}) \cdot (\hat{\lambda} + \Delta\lambda) + \boldsymbol{z}^\top \boldsymbol{\theta} + \frac{1}{2\sigma} \cdot ((\Delta\lambda)^2 + \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2) \\ \text{s.t.}\ & (\hat{\lambda} + \Delta\lambda) \hat{p}_{s,a,s'} - w_{\hat{s}} \leq \theta_{s,a,s'} && \forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \\ & (\hat{\lambda} + \Delta\lambda) \hat{p}_{s,a,s'} + w_{\hat{s}} \geq \theta_{s,a,s'} && \forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \\ & \mathbf{e}^\top \boldsymbol{\theta}_{s,a} = \hat{\lambda} + \Delta\lambda && \forall s \in \mathcal{S},\ a \in \mathcal{A} \\ & \boldsymbol{\theta} \in \mathbb{R}_+^{S \cdot A \cdot S}. \end{aligned} \tag{22}$$

Notice that problem (22) and $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$ share the same optimal value. The equivalent minimax representation of

problem (22) is:

$$\begin{aligned}\min_{\boldsymbol{\theta}\in\mathbb{R}^{S\cdot A\cdot S}} \max_{\substack{\boldsymbol{\chi}\in\mathbb{R}_+^{S\cdot A\cdot S},\\ \boldsymbol{\psi}\in\mathbb{R}_+^{S\cdot A\cdot S},\\ \boldsymbol{\xi}\in\mathbb{R}^{S\cdot A},\\ \boldsymbol{\mu}\in\mathbb{R}_+^{S\cdot A\cdot S}}} & (d_{\hat{s}}-\mathbf{e}^\top\boldsymbol{u}_{\hat{s}})\cdot(\hat{\lambda}+\Delta\lambda)+\boldsymbol{z}^\top\boldsymbol{\theta}+\frac{1}{2\sigma}\cdot((\Delta\lambda)^2+\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_2^2)\\ &+\sum_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}\chi_{s,a,s'}\cdot((\hat{\lambda}+\Delta\lambda)\hat{p}_{s,a,s'}-w_{\hat{s}}-\theta_{s,a,s'})\\ &+\sum_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}\psi_{s,a,s'}\cdot(\theta_{s,a,s'}-(\hat{\lambda}+\Delta\lambda)\hat{p}_{s,a,s'}-w_{\hat{s}})\\ &+\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\xi_{s,a}\cdot(\mathbf{e}^\top\boldsymbol{\theta}_{s,a}-\hat{\lambda}-\Delta\lambda)\\ &-\boldsymbol{\mu}^\top\boldsymbol{\theta}.\end{aligned}$$

Taking its dual, we have

$$\begin{aligned}\max\ & (d_{\hat{s}}-\mathbf{e}^\top\boldsymbol{u}_{\hat{s}})\cdot(\hat{\lambda}+\Delta\lambda)+\frac{1}{2\sigma}\cdot((\Delta\lambda)^2+\hat{\boldsymbol{\theta}}^\top\hat{\boldsymbol{\theta}})\\ &+\sum_{(s,a,s')\in\mathcal{S}\cdot\mathcal{A}\cdot\mathcal{S}}\chi_{s,a,s'}\cdot((\hat{\lambda}+\Delta\lambda)\hat{p}_{s,a,s'}-w_{\hat{s}})\\ &-\sum_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}\psi_{s,a,s'}\cdot((\hat{\lambda}+\Delta\lambda)\hat{p}_{s,a,s'}+w_{\hat{s}})\\ &-\sum_{(s,a)\in\mathcal{S}\cdot\mathcal{A}}\xi_{s,a}\cdot(\hat{\lambda}+\Delta\lambda)-\frac{\sigma}{2}\left\|\boldsymbol{z}-\frac{1}{\sigma}\hat{\boldsymbol{\theta}}-\boldsymbol{\chi}+\boldsymbol{\psi}+\boldsymbol{B}^\top\boldsymbol{\xi}-\boldsymbol{\mu}\right\|_2^2\\ \text{s.t.}\ & \boldsymbol{\chi}\in\mathbb{R}_+^{S\cdot A\cdot S},\ \boldsymbol{\psi}\in\mathbb{R}_+^{S\cdot A\cdot S},\ \boldsymbol{\xi}\in\mathbb{R}^{S\cdot A},\ \boldsymbol{\mu}\in\mathbb{R}_+^{S\cdot A\cdot S}.\end{aligned}\tag{23}$$

Considering a feasible solution $(\boldsymbol{\chi},\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\mu})=\mathbf{0}$ in (23), by weak duality, we have a lower bound:

$$\mathfrak{D}_{\hat{s}}(\boldsymbol{u};\hat{\lambda},\hat{\boldsymbol{\theta}})\geq(d_{\hat{s}}-\mathbf{e}^\top\boldsymbol{u}_{\hat{s}})\cdot(\hat{\lambda}+\Delta\lambda)+\frac{1}{2\sigma}\cdot((\Delta\lambda)^2+\hat{\boldsymbol{\theta}}^\top\hat{\boldsymbol{\theta}})-\frac{\sigma}{2}\cdot\left\|\boldsymbol{z}-\frac{1}{\sigma}\cdot\hat{\boldsymbol{\theta}}\right\|_2^2.$$

Hence, the existence of the optimal value requires that the following inequality must hold:

$$\begin{aligned}&(d_{\hat{s}}-\mathbf{e}^\top\boldsymbol{u}_{\hat{s}})\cdot(\hat{\lambda}+\Delta\lambda)+\frac{1}{2\sigma}\cdot((\Delta\lambda)^2+\hat{\boldsymbol{\theta}}^\top\hat{\boldsymbol{\theta}})-\frac{\sigma}{2}\cdot\left\|\boldsymbol{z}-\frac{1}{\sigma}\cdot\hat{\boldsymbol{\theta}}\right\|_2^2\\ \leq\ &(d_{\hat{s}}-\mathbf{e}^\top\boldsymbol{u}_{\hat{s}})\cdot\hat{\lambda}+\sum_{(s,a,s')\in\mathcal{Z}_+}z_{s,a,s'}(\hat{\lambda}\hat{p}_{s,a,s'}+w_{\hat{s}})+\sum_{(s,a,s')\in\mathcal{Z}_-}z_{s,a,s'}[\hat{\lambda}\hat{p}_{s,a,s'}-w_{\hat{s}}]_+,\end{aligned}$$

which is

$$\begin{aligned}\frac{1}{2\sigma}\cdot(\Delta\lambda)^2+(d_{\hat{s}}-\mathbf{e}^\top\boldsymbol{u}_{\hat{s}})\cdot\Delta\lambda-\Bigg[&\frac{\sigma}{2}\cdot(\boldsymbol{z}^\top\boldsymbol{z}-\frac{2}{\sigma}\cdot\boldsymbol{z}^\top\hat{\boldsymbol{\theta}})\\ &+\sum_{(s,a,s')\in\mathcal{Z}_+}z_{s,a,s'}(\hat{\lambda}\hat{p}_{s,a,s'}+w_{\hat{s}})+\sum_{(s,a,s')\in\mathcal{Z}_-}z_{s,a,s'}[\hat{\lambda}\hat{p}_{s,a,s'}-w_{\hat{s}}]_+\Bigg]\leq 0.\end{aligned}$$

Hence, we have an upper bound

$$\begin{aligned}\Delta\lambda\leq\sigma\cdot\Bigg(&-(d_{\hat{s}}-\mathbf{e}^\top\boldsymbol{u}_{\hat{s}})+\Bigg\{(d_{\hat{s}}-\mathbf{e}^\top\boldsymbol{u}_{\hat{s}})^2+\frac{2}{\sigma}\cdot\Bigg(\frac{\sigma}{2}\cdot\Bigg(\boldsymbol{z}^\top\boldsymbol{z}-\frac{2}{\sigma}\cdot\boldsymbol{z}^\top\hat{\boldsymbol{\theta}}\Bigg)\\ &+\sum_{(s,a,s')\in\mathcal{Z}_+}z_{s,a,s'}(\hat{\lambda}\hat{p}_{s,a,s'}+w_{\hat{s}})+\sum_{(s,a,s')\in\mathcal{Z}_-}z_{s,a,s'}[\hat{\lambda}\hat{p}_{s,a,s'}-w_{\hat{s}}]_+\Bigg)\Bigg\}^{\frac{1}{2}}\Bigg),\end{aligned}$$

from which our conclusion follows immediately. □

By Proposition 4.3, we have in Algorithm 1 that for all $k\in\mathbb{N}_+$, $\boldsymbol{u}^{k+1}$ is bounded if $\boldsymbol{u}^k$ and $(\boldsymbol{\lambda},\boldsymbol{\theta})$ are bounded; while by Lemma A.3, λ^{k+1} is bounded if $\boldsymbol{u}^{k+1}$, $\boldsymbol{u}^k$ and $(\boldsymbol{\lambda}^k,\boldsymbol{\theta}^k)$ are bounded. Since we initialize $(\boldsymbol{u}^0,(\boldsymbol{\lambda}^0,\boldsymbol{\theta}^0))$ with some real numbers, Lemma A.3 can iteratively provide us the upper bounds for the golden section search.

Lemma A.4. *The function $f_{\hat{s}}$ is well-defined on $\mathbb{R}_+$.*

Proof It is sufficient to prove that problem (10) can achieve its minimum for all $\lambda \in \mathbb{R}_+$. Since for all $\lambda \in \mathbb{R}_+$, the feasible region of (10),

$$C(\lambda) := \left\{ \boldsymbol{\theta} \in \mathbb{R}_+^{S\cdot A\cdot S} \;\middle|\; \begin{array}{ll} \|\boldsymbol{\theta} - \lambda \cdot \hat{\boldsymbol{p}}\|_\infty \le w_{\hat{s}} & \\ \mathbf{e}^\top \boldsymbol{\theta}_{s,a} = \lambda & \forall (s,a) \in \mathcal{S}\times\mathcal{A} \end{array} \right\}$$

is compact and the objective function

$$h_{\hat{s}}(\lambda, \boldsymbol{\theta}) := (d_{\hat{s}} - \mathbf{e}^\top \boldsymbol{u}_{\hat{s}}) \cdot \lambda + \boldsymbol{\theta}^\top \boldsymbol{Q}_{\hat{s}} \boldsymbol{u} + \frac{1}{2\sigma} \cdot \left\| \begin{matrix} \lambda - \hat{\lambda} \\ \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \end{matrix} \right\|_2^2$$

is continuous in $\boldsymbol{\theta}$, by the Weierstrass extreme-value theorem (see, *e.g.*, (Khanh & Quan, 2019)), problem (10) can obtain its minimum. □

Proof of Proposition 4.4 Let us fix $\lambda, \lambda' \in \mathbb{R}_+$ and $\omega \in [0,1]$. By Lemma A.4, $f_{\hat{s}}$ is well-defined, which means that there exists $\boldsymbol{\theta} \in C(\lambda),\ \boldsymbol{\theta}' \in C(\lambda')$ such that

$$h_{\hat{s}}(\lambda, \boldsymbol{\theta}) = f_{\hat{s}}(\lambda) \text{ and } h_{\hat{s}}(\lambda', \boldsymbol{\theta}') = f_{\hat{s}}(\lambda'),$$

where we denote

$$C(\lambda) := \left\{ \boldsymbol{\theta} \in \mathbb{R}_+^{S\cdot A\cdot S} \;\middle|\; \begin{array}{ll} \|\boldsymbol{\theta} - \lambda \cdot \hat{\boldsymbol{p}}\|_\infty \le w_{\hat{s}} & \\ \mathbf{e}^\top \boldsymbol{\theta}_{s,a} = \lambda & \forall (s,a) \in \mathcal{S}\times\mathcal{A} \end{array} \right\}$$

and

$$h_{\hat{s}}(\lambda, \boldsymbol{\theta}) := (d_{\hat{s}} - \mathbf{e}^\top \boldsymbol{u}_{\hat{s}}) \cdot \lambda + \boldsymbol{\theta}^\top \boldsymbol{Q}_{\hat{s}} \boldsymbol{u} + \frac{1}{2\sigma} \cdot \left\| \begin{matrix} \lambda - \hat{\lambda} \\ \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \end{matrix} \right\|_2^2$$

(as in the proof of Lemma A.4). By the convexity of

$$\mathcal{C} := \left\{ (\lambda, \boldsymbol{\theta}) \in \mathbb{R}_+ \times \mathbb{R}_+^{S\cdot A\cdot S} \mid \boldsymbol{\theta} \in C(\lambda) \right\}$$

(which is, the feasible region of $\mathfrak{D}_{\hat{s}}(\boldsymbol{u}; \hat{\lambda}, \hat{\boldsymbol{\theta}})$), we have $((1-\omega)\cdot\lambda + \omega\lambda', (1-\omega)\cdot\boldsymbol{\theta} + \omega\cdot\boldsymbol{\theta}') \in \mathcal{C}$. Thus we have $(1-\omega)\cdot\boldsymbol{\theta} + \omega\cdot\boldsymbol{\theta}' \in C((1-\omega)\cdot\lambda + \omega\lambda')$ and

$$\begin{aligned} f_{\hat{s}}((1-\omega)\cdot\lambda + \omega\lambda') &\le h_{\hat{s}}((1-\omega)\cdot\lambda + \omega\lambda', (1-\omega)\cdot\boldsymbol{\theta} + \omega\cdot\boldsymbol{\theta}') \\ &\le (1-\omega)\cdot h_{\hat{s}}(\lambda, \boldsymbol{\theta}) + \omega \cdot h_{\hat{s}}(\lambda', \boldsymbol{\theta}') \\ &= (1+\omega)\cdot f_{\hat{s}}(\lambda) + \omega \cdot f_{\hat{s}}(\lambda'), \end{aligned}$$

where the first inequality follows from the definition of $f_{\hat{s}}$ and the second inequality holds because of the convexity of h. Since λ, λ' and ω are arbitrary, we have proved the convexity of $f_{\hat{s}}$, where our conclusion follows. □

Proof of Proposition 4.5 Rearranging the constraints in (11), we rewrite the problem as

$$\begin{array}{lll} \min & \dfrac{1}{2\sigma}\cdot\boldsymbol{\theta}_{s,a}^\top\boldsymbol{\theta}_{s,a} + \boldsymbol{\theta}_{s,a}^\top(\boldsymbol{z}_{\hat{s},s,a} - \dfrac{1}{\sigma}\cdot\hat{\boldsymbol{\theta}}_{s,a}) & \\ \text{s.t.} & \theta_{s,a,s'} \ge [\lambda\hat{p}_{s,a,s'} - w_{\hat{s}}]_+ & \forall s' \in \mathcal{S} \\ & \theta_{s,a,s'} \le \lambda\hat{p}_{s,a,s'} + w_{\hat{s}} & \forall s' \in \mathcal{S} \\ & \mathbf{e}^\top\boldsymbol{\theta}_{s,a} = \lambda & \\ & \boldsymbol{\theta}_{s,a} \in \mathbb{R}^S. & \end{array}$$

Introducing dual variables $\boldsymbol{\eta} \in \mathbb{R}_+^S, \boldsymbol{\varphi} \in \mathbb{R}_+^S$ and $\iota \in \mathbb{R}$, we have its Lagrangian dual function as

$$\begin{aligned} L(\boldsymbol{\theta}_{s,a}, \boldsymbol{\eta}, \boldsymbol{\varphi}, \iota) = &\frac{1}{2\sigma}\cdot\boldsymbol{\theta}_{s,a}^\top\boldsymbol{\theta}_{s,a} + \boldsymbol{\theta}_{s,a}^\top(\boldsymbol{z}_{\hat{s},s,a} - \frac{1}{\sigma}\hat{\boldsymbol{\theta}}_{s,a}) \\ &+ \sum_{s'\in\mathcal{S}} \eta_{s'}\cdot([\lambda\hat{p}_{s,a,s'} - w_{\hat{s}}]_+ - \theta_{s,a,s'}) \\ &+ \sum_{s'\in\mathcal{S}} \varphi_{s'}\cdot(\theta_{s,a,s'} - (\lambda\hat{p}_{s,a,s'} + w_{\hat{s}})) + \iota\cdot(\lambda - \mathbf{e}^\top\boldsymbol{\theta}_{s,a}). \end{aligned}$$

Thus we have the KKT conditions of problem (11) as

$$\begin{cases} \theta_{s,a,s'} \geq [\lambda \hat{p}_{s,a,s'} - w_{\hat{s}}]_+ & \forall s' \in \mathcal{S} \\ \theta_{s,a,s'} \leq \lambda \hat{p}_{s,a,s'} + w_{\hat{s}} & \forall s' \in \mathcal{S} \\ \mathbf{e}^\top \boldsymbol{\theta}_{s,a} = \lambda & \\ \boldsymbol{\eta} \geq \mathbf{0} & \\ \boldsymbol{\varphi} \geq \mathbf{0} & \\ \eta_{s'} \cdot ([\lambda \hat{p}_{s,a,s'} - w_{\hat{s}}]_+ - \theta_{s,a,s'}) = 0 & \forall s' \in \mathcal{S} \\ \varphi_{s'} \cdot (\theta_{s,a,s'} - (\lambda \hat{p}_{s,a,s'} + w_{\hat{s}})) = 0 & \forall s' \in \mathcal{S} \\ \nabla_{\boldsymbol{\theta}_{s,a}} L(\boldsymbol{\theta}_{s,a}, \boldsymbol{\eta}, \boldsymbol{\varphi}, \iota) = \frac{1}{\sigma} \cdot \boldsymbol{\theta}_{s,a} + (\boldsymbol{z}_{\hat{s},s,a} - \frac{1}{\sigma} \cdot \hat{\boldsymbol{\theta}}_{s,a}) - \boldsymbol{\eta} + \boldsymbol{\varphi} - \iota \cdot \mathbf{e} = \mathbf{0}, & \end{cases}$$

from which we have

$$\theta_{s,a,s'} = \begin{cases} \lambda \hat{p}_{s,a,s'} + w_{\hat{s}} & \forall s' \in \mathcal{S} : \varphi_{s'} \neq 0 \\ \sigma \cdot (\iota + \frac{1}{\sigma} \hat{\theta}_{s,a,s'} - z_{\hat{s},s,a,s'}) & \forall s' \in \mathcal{S} : \eta_{s'} = 0 \text{ and } \varphi_{s'} = 0 \\ [\lambda \hat{p}_{s,a,s'} - w_{\hat{s}}]_+ & \forall s' \in \mathcal{S} : \eta_{s'} \neq 0. \end{cases}$$

where it is sufficient to solve the equation $W_{s,a}(\iota) = \lambda$ to find $\iota^\star$ then obtain the optimal solution $\theta^\star_{s,a,s'} = W_{s,a,s'}(\iota^\star)\ \forall s' \in \mathcal{S}$, where $W_{s,a}(\iota) = \sum_{s' \in \mathcal{S}} W_{s,a,s'}(\iota)$ with

$$\begin{aligned} & W_{s,a,s'}(\iota) \\ = & \begin{cases} \lambda \hat{p}_{s,a,s'} + w_{\hat{s}} & \text{if } \iota \geq \frac{1}{\sigma} \cdot (\lambda \hat{p}_{s,a,s'} + w_{\hat{s}}) + z_{\hat{s},s,a,s'} - \frac{1}{\sigma} \hat{\theta}_{s,a,s'} \\ [\lambda \hat{p}_{s,a,s'} - w_{\hat{s}}]_+ & \text{if } \iota < \frac{1}{\sigma} \cdot [\lambda \hat{p}_{s,a,s'} - w_{\hat{s}}]_+ + z_{\hat{s},s,a,s'} - \frac{1}{\sigma} \hat{\theta}_{s,a,s'} \\ \sigma \cdot (\iota + \frac{1}{\sigma} \theta^k_{s,a,s'} - z_{\hat{s},s,a,s'}) & \text{otherwise,} \end{cases} \end{aligned}$$

for all $s' \in \mathcal{S}$. Since for all $s' \in \mathcal{S}$, the function $W_{s,a,s'}$ is clearly piecewise linear and non-decreasing, the function $W_{s,a} = \sum_{s' \in \mathcal{S}} W_{s,a,s'}$ is thus also piecewise linear and non-decreasing with $2S$ breakpoints: the upper breakpoints $\frac{1}{\sigma} \cdot (\lambda \hat{p}_{s,a,s'} + w_{\hat{s}}) - \frac{1}{\sigma} \hat{\theta}_{s,a,s'} + z_{\hat{s},s,a,s'},\ s' \in \mathcal{S}$ and the lower breakpoints $\frac{1}{\sigma} \cdot [\lambda \hat{p}_{s,a,s'} - w_{\hat{s}}]_+ - \frac{1}{\sigma} \hat{\theta}_{s,a,s'} + z_{\hat{s},s,a,s'},\ s' \in \mathcal{S}$. Sort them from smallest to largest as $\iota_1 \leq \cdots \leq \iota_{2S}$ and search the intervals $[\iota_1, \iota_2], [\iota_2, \iota_3], \cdots, [\iota_{2S-1}, \iota_{2S}]$ in an ascending order, we can locate the optimal $\iota^\star$ and $\theta^\star_{s,a,s'} = W_{s,a,s'}(\iota^\star)\ \forall s' \in \mathcal{S}$.

The time complexity is dominated by sorting the breakpoints, which is achieved in time $\mathcal{O}(S \log(S))$. □

We provide the pseudocode for solving problem (11) in Algorithm 3. Here the functions $p_1(\cdot) : [2S] \mapsto \mathcal{S}$ and $p_2(\cdot) : [2S] \mapsto \{\text{“lower”}, \text{“upper”}\}$ map the indices of the non-decreasing breakpoint sequence to the indices and types of lower/upper breakpoints, respectively; *e.g.*, if ι_3 corresponds to $\underline{\iota}_5$, then we have $p_1(3) = 5$ and $p_2(3) =$ “lower”.

Proof of Proposition 4.6 By the strategy described in Section 4.2, to obtain the output of Step 2 in Algorithm 1, we need to solve S subproblems, each of which solved by a golden section search with time complexity $\log(\frac{1}{\delta})$. By the analysis in Section 4.2, each function valuation of $f_{\hat{s}}$ in the search requires to solve SA subproblems in the form of (11), and each of these subproblems is, by Proposition 4.5, solvable in time $\mathcal{O}(S \log(S))$. Our conclusion thus follows. □

B. An Equivalent Convex Optimization Problem for RSMDPs with KL Divergence

Lemma B.1. *Let* $\mathcal{U} := \{\boldsymbol{u} \in \mathbb{R}^{S \cdot A}_+ \mid \boldsymbol{r}^\top \boldsymbol{u} \geq \tau\}$. *For all* $s \in \mathcal{S}$ *and* $(\boldsymbol{\alpha}, \boldsymbol{u}) \in \mathcal{R}^{S \cdot A} \times \mathcal{U}$, *if*

$$\lim_{k \to 0^+} -\boldsymbol{\alpha}^\top \mathbf{e} - \sum_{i \in [SAS]} \hat{p}_i k \left(\exp\left(-\frac{\boldsymbol{Q}^\top_{s,i} \boldsymbol{u} + \boldsymbol{B}^\top_{:,i} \boldsymbol{\alpha}}{k} \right) - 1 \right) \geq \mathbf{e}^\top \boldsymbol{u}_s - d_s, \tag{24}$$

then it holds that $\boldsymbol{Q}_s \hat{\boldsymbol{u}} + \boldsymbol{B}^\top \hat{\boldsymbol{\alpha}} \geq \mathbf{0}$.

Proof Suppose to the contrary that there exists some $i \in [SAS]$ such that $\boldsymbol{Q}^\top_{s,i} \boldsymbol{u} + \boldsymbol{B}^\top_{:,i} \boldsymbol{\alpha} < 0$. It follows that

$$\lim_{k \to 0^+} k \exp\left(-\frac{\boldsymbol{Q}^\top_{s,i} \boldsymbol{u} + \boldsymbol{B}^\top_{:,i} \boldsymbol{\alpha}}{k} \right) = \infty,$$

Algorithm 3 Interval Search Algorithm for Problem (11)

Compute all the upper breakpoints $\bar{\iota}_{s'} \leftarrow \frac{1}{\sigma}(\lambda\hat{p}_{s,a,s'} + w_{\hat{s}}) - \frac{1}{\sigma}\theta_{s,a,s'} + z_{\hat{s},s,a,s'}\ \forall s' \in \mathcal{S}$ and lower breakpoints $\underline{\iota}_{s'} \leftarrow \frac{1}{\sigma}[\lambda\hat{p}_{s,a,s'} - w_{\hat{s}}]_+ - \frac{1}{\sigma}\theta_{s,a,s'} + z_{\hat{s},s,a,s'}\ \forall s' \in \mathcal{S}$;
Sort the breakpoints from smallest to largest as $\iota_1 \leq \cdots \leq \iota_{2S}$;
Initialize $\chi \leftarrow \sigma$ and $\psi \leftarrow \sum_{s' \in \mathcal{S}: s' \neq p_1(1)} [\lambda\hat{p}_{s,a,s'} - w_{\hat{s}}]_+ + \sigma \cdot (\frac{1}{\sigma}\theta_{s,a,p_1(1)} - z_{\hat{s},s,a,p_1(1)})$;
Initialize upper breakpoint index set $\mathcal{U} \leftarrow \emptyset$ and lower breakpoint index set $\mathcal{L} \leftarrow \mathcal{S} \setminus p_1(1)$;
for $k = 1$ **to** $2S - 1$ **do**
  if $\chi \cdot \iota_{k+1} + \psi \geq \lambda$ **then**
    $\iota^\star \leftarrow \frac{\lambda - \psi}{\chi}$;
    for $s' = 1$ **to** S **do**

$$\theta^\star_{s,a,s'} \leftarrow \begin{cases} \lambda\hat{p}_{s,a,s'} + w_{\hat{s}} & \forall s' \in \mathcal{U} \\ [\lambda\hat{p}_{s,a,s'} - w_{\hat{s}}]_+ & \forall s' \in \mathcal{L} \\ \sigma \cdot (\iota^\star + \frac{1}{\sigma}\theta_{s,a,s'} - z_{\hat{s},s,a,s'}) & \forall s' \in \mathcal{S} \setminus (\mathcal{U} \cup \mathcal{L}); \end{cases}$$

    end for
    break
  else if $p_2(k+1) =$ "upper" **then**
    $\chi \leftarrow \chi - \sigma$;
    $\psi \leftarrow \psi - \sigma \cdot (\frac{1}{\sigma}\theta_{s,a,p_1(k+1)} - z_{\hat{s},s,a,p_1(k+1)}) + \lambda\hat{p}_{s,a,p_1(k+1)} + w_{\hat{s}}$;
  else
    $\chi \leftarrow \chi + \sigma$;
    $\psi \leftarrow \psi + \sigma \cdot (\frac{1}{\sigma}\theta_{s,a,p_1(k+1)} - z_{\hat{s},s,a,p_1(k+1)}) - [\lambda\hat{p}_{s,a,p_1(k+1)} - w_{\hat{s}}]_+$;
  end if
end for
Output: Solution $\boldsymbol{\theta}^\star_{s,a}$

by which the left-hand side of (24) will be driven to $-\infty$, yielding a contradiction. □

Lemma B.2. *Let* $\mathcal{B}_s(\boldsymbol{u}) := \{\boldsymbol{\alpha} \in \mathbb{R}^{S \cdot A} \mid -\boldsymbol{\alpha}^\top \mathbf{e} \geq \mathbf{e}^\top \boldsymbol{u}_s - d_s \text{ and } \boldsymbol{Q}_s \boldsymbol{u} + \boldsymbol{B}^\top \boldsymbol{\alpha} \geq \mathbf{0}\}$ *and*

$$\mathcal{B}'_s(\boldsymbol{u}) := \left\{ \boldsymbol{\alpha} \in \mathbb{R}^{S \cdot A} \;\middle|\; \lim_{k \to 0^+} -\boldsymbol{\alpha}^\top \mathbf{e} - \sum_{i \in [SAS]} \hat{p}_i k \left(\exp\left(-\frac{\boldsymbol{Q}^\top_{s,i} \boldsymbol{u} + \boldsymbol{B}^\top_{:,i} \boldsymbol{\alpha}}{k} \right) - 1 \right) \geq \mathbf{e}^\top \boldsymbol{u}_s - d_s \right\}$$

for all $s \in \mathcal{S}$*, where* $\boldsymbol{Q}_{s,i}$ *is the* i*-th row of* $\boldsymbol{Q}_s$ *and* $\boldsymbol{B}_{:,i}$ *is the* i*-th column of* $\boldsymbol{B}$*, both of them are column vectors. When* $\hat{\boldsymbol{p}} \in \mathbb{R}^{S \cdot A \cdot S}_{++}$*, it holds that* $\mathcal{B}_s(\boldsymbol{u}) = \mathcal{B}'_s(\boldsymbol{u})\ \forall s \in \mathcal{S}$ *for all* $\boldsymbol{u} \in \mathcal{U} := \{\boldsymbol{u} \in \mathbb{R}^{S \cdot A}_+ \mid \boldsymbol{r}^\top \boldsymbol{u} \geq \tau\}$.

Proof Fix some $s \in \mathcal{S}$. We will first prove that $\mathcal{B}_s(\boldsymbol{u}) \subseteq \mathcal{B}'_s(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathcal{U}$. To this end, let $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{u}}) \in \mathbb{R}^{S \cdot A} \times \mathcal{U}$ be arbitrarily taken such that $\hat{\boldsymbol{\alpha}} \in \mathcal{B}_s(\hat{\boldsymbol{u}})$, *i.e.*, $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{u}})$ satisfies

$$\begin{cases} \boldsymbol{r}^\top \hat{\boldsymbol{u}} \geq \tau \\ -\hat{\boldsymbol{\alpha}}^\top \mathbf{e} \geq \mathbf{e}^\top \hat{\boldsymbol{u}}_s - d_s \\ \boldsymbol{Q}_s \hat{\boldsymbol{u}} + \boldsymbol{B}^\top \hat{\boldsymbol{\alpha}} \geq \mathbf{0}. \end{cases}$$

We then have

$$\lim_{k \to 0^+} -\hat{\boldsymbol{\alpha}}^\top \mathbf{e} - \sum_{i \in [SAS]} \hat{p}_i k \left(\exp\left(-\frac{\boldsymbol{Q}^\top_{s,i} \hat{\boldsymbol{u}} + \boldsymbol{B}^\top_{:,i} \hat{\boldsymbol{\alpha}}}{k} \right) - 1 \right) = -\hat{\boldsymbol{\alpha}}^\top \mathbf{e},$$

which implies that $\hat{\boldsymbol{\alpha}} \in \mathcal{B}'_s(\hat{\boldsymbol{u}})$.

Next, we prove that $\mathcal{B}'_s(\boldsymbol{u}) \subseteq \mathcal{B}_s(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathcal{U}$. To this end, let $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{u}}) \in \mathbb{R}^{S \cdot A} \times \mathcal{U}$ be arbitrarily taken such that $\hat{\boldsymbol{\alpha}} \in \mathcal{B}'_s(\hat{\boldsymbol{u}})$. By Lemma B.1, it holds that $\boldsymbol{Q}_s \hat{\boldsymbol{u}} + \boldsymbol{B}^\top \hat{\boldsymbol{\alpha}} \geq \mathbf{0}$. In this case, we have

$$\lim_{k \to 0^+} -\hat{\boldsymbol{\alpha}}^\top \mathbf{e} - \sum_{i \in [SAS]} \hat{p}_i k \left(\exp\left(-\frac{\boldsymbol{Q}^\top_{s,i} \hat{\boldsymbol{u}} + \boldsymbol{B}^\top_{:,i} \hat{\boldsymbol{\alpha}}}{k} \right) - 1 \right) = -\hat{\boldsymbol{\alpha}}^\top \mathbf{e},$$

which means that $\hat{\boldsymbol{\alpha}} \in \mathcal{B}'_s(\hat{\boldsymbol{u}})$ if and only if $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{u}})$ satisfies $-\hat{\boldsymbol{\alpha}}^\top \mathbf{e} \geq \mathbf{e}^\top \hat{\boldsymbol{u}}_s - d_s$. Hence, $\hat{\boldsymbol{\alpha}} \in \mathcal{B}_s(\hat{\boldsymbol{u}})$ follows from the definition of $\mathcal{B}_s(\hat{\boldsymbol{u}})$. □

Proposition B.3. *When $\hat{\boldsymbol{p}} \in \mathbb{R}^{S\cdot A\cdot S}_{++}$, the RSMDPs (4) equipped with the KL divergence[3] $\ell(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \sum_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} \hat{p}_{s,a,s'} \phi(p_{s,a,s'}/\hat{p}_{s,a,s'})$, where the phi-divergence function $\phi(t) = t\log t - t + 1$, is equivalent to the following convex optimization problem:*

$$
\begin{aligned}
\mathrm{Z}_{\mathrm{RS}} = \min\ & \boldsymbol{w}^\top \boldsymbol{k} \\
\text{s.t. } & \mathbf{e}^\top \boldsymbol{u}_s - d_s \leq -\boldsymbol{\alpha}_s^\top \mathbf{e} - \sum_{i\in[SAS]} \hat{p}_i k_s \left(\exp\left(-\frac{\boldsymbol{Q}_{s,i}^\top \boldsymbol{u} + \boldsymbol{B}_i^\top \boldsymbol{\alpha}_s}{k_s} \right) - 1 \right) && \forall s \in \mathcal{S} \\
& \boldsymbol{r}^\top \boldsymbol{u} \geq \tau \\
& \boldsymbol{u} \in \mathbb{R}^{S\cdot A}_+,\ \boldsymbol{k} \in \mathbb{R}^S_+,\ \boldsymbol{\alpha}_s \in \mathbb{R}^{S\cdot A} && \forall s \in \mathcal{S},
\end{aligned}
$$

where $\boldsymbol{Q}_{s,i}$ is the i-th row of $\boldsymbol{Q}_s$ and $\boldsymbol{B}_{:,i}$ is the i-th column of $\boldsymbol{B}$, both of them are column vectors.

Proof The s-th constraint of (4) is equivalent to

$$
\mathbf{e}^\top \boldsymbol{u}_s - d_s \leq \min_{\boldsymbol{p}\in\mathcal{P}} \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u} + k_s \cdot \ell(\boldsymbol{p}, \hat{\boldsymbol{p}}),
$$

where the minimization problem on the right-hand side is:

$$
\begin{aligned}
\min\ & \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u} + k_s \cdot \ell(\boldsymbol{p}, \hat{\boldsymbol{p}}) \\
\text{s.t. } & \boldsymbol{B}\boldsymbol{p} = \mathbf{e} \\
& \boldsymbol{p} \in \mathbb{R}^{S\cdot A\cdot S}_+.
\end{aligned}
$$

The dual of this problem is:

$$
\begin{aligned}
& \max_{\boldsymbol{\alpha}} \min_{\boldsymbol{p}\geq\mathbf{0}}\ \boldsymbol{p}^\top \boldsymbol{Q}_s \boldsymbol{u} + k_s \cdot \ell(\boldsymbol{p}, \hat{\boldsymbol{p}}) + \boldsymbol{\alpha}^\top (\boldsymbol{B}\boldsymbol{p} - \mathbf{e}) \\
&= \max_{\boldsymbol{\alpha}} \min_{\boldsymbol{p}\geq\mathbf{0}}\ (\boldsymbol{Q}_s \boldsymbol{u} + \boldsymbol{B}^\top \boldsymbol{\alpha})^\top \boldsymbol{p} + k_s \cdot \sum_{i\in[SAS]} \hat{p}_i \cdot \phi\left(\frac{p_i}{\hat{p}_i}\right) - \boldsymbol{\alpha}^\top \mathbf{e} \\
&= \max_{\boldsymbol{\alpha}} -\boldsymbol{\alpha}^\top \mathbf{e} + \min_{\boldsymbol{p}\geq\mathbf{0}}\ (\boldsymbol{Q}_s \boldsymbol{u} + \boldsymbol{B}^\top \boldsymbol{\alpha})^\top \boldsymbol{p} + k_s \cdot \sum_{i\in[SAS]} \hat{p}_i \cdot \phi\left(\frac{p_i}{\hat{p}_i}\right) \\
&= \max_{\boldsymbol{\alpha}} -\boldsymbol{\alpha}^\top \mathbf{e} - \sum_{i\in[SAS]} \max_{p\geq 0} -(\boldsymbol{Q}_{s,i}^\top \boldsymbol{u} + \boldsymbol{B}_{:,i}^\top \boldsymbol{\alpha}) p - k_s \hat{p}_i \phi\left(\frac{p}{\hat{p}_i}\right) \\
&= \max_{\boldsymbol{\alpha}} -\boldsymbol{\alpha}^\top \mathbf{e} - \sum_{i\in[SAS]} \max_{q\geq 0} -(\boldsymbol{Q}_{s,i}^\top \boldsymbol{u} + \boldsymbol{B}_{:,i}^\top \boldsymbol{\alpha}) \hat{p}_i q - k_s \hat{p}_i \phi(q).
\end{aligned}
$$

Here, strong duality holds for $\{\boldsymbol{p} \in \mathbb{R}^{S\cdot A\cdot S}_+ \mid \boldsymbol{B}\boldsymbol{p} = \mathbf{e}\} = (\Delta^S)^{SA}$, where $\boldsymbol{p} = (1/S)\cdot\mathbf{e}$ is a strictly feasible solution. The last equality holds by the substitution $q = p/\hat{p}_i$. We can further have

$$
\begin{aligned}
& \max_{\boldsymbol{\alpha}} -\boldsymbol{\alpha}^\top \mathbf{e} - \sum_{i\in[SAS]} \max_{q\geq 0} -(\boldsymbol{Q}_{s,i}^\top \boldsymbol{u} + \boldsymbol{B}_{:,i}^\top \boldsymbol{\alpha}) \hat{p}_i q - k_s \hat{p}_i \phi(q) \\
&= \begin{cases}
\left\{ \begin{array}{ll} \max & -\boldsymbol{\alpha}^\top \mathbf{e} \\ \text{s.t.} & \boldsymbol{Q}_s \boldsymbol{u} + \boldsymbol{B}^\top \boldsymbol{\alpha} \geq \mathbf{0} \\ & \boldsymbol{\alpha} \in \mathbb{R}^{S\cdot A} \end{array} \right. & \text{if } k_s = 0 \\
\max\limits_{\boldsymbol{\alpha}} -\boldsymbol{\alpha}^\top \mathbf{e} - \sum\limits_{i\in[SAS]} \hat{p}_i k_s \left(\exp\left(-\dfrac{\boldsymbol{Q}_{s,i}^\top \boldsymbol{u} + \boldsymbol{B}_{:,i}^\top \boldsymbol{\alpha}}{k_s} \right) - 1 \right) & \text{if } k_s > 0,
\end{cases}
\end{aligned}
$$

where the equality for the case $k_s > 0$ follows from the convex conjugate of $\phi(\cdot)$; see (Bayraksan & Love, 2015). Therefore,

[3] We refer intereted readers to (Bayraksan & Love, 2015) for more details about the KL divergence.

we have that RSMDPs (4) is equivalent to

$$
\begin{aligned}
\mathrm{Z}_{\mathrm{RS}}(\hat{\boldsymbol{p}}) = \min\ & \boldsymbol{w}^\top \boldsymbol{k} \\
\text{s.t.}\ & -\boldsymbol{\alpha}_s^\top \mathbf{e} \geq \mathbf{e}^\top \boldsymbol{u}_s - d_s && \forall s : k_s = 0 \\
& \boldsymbol{Q}_s \boldsymbol{u} + \boldsymbol{B}^\top \boldsymbol{\alpha}_s \geq \mathbf{0} && \forall s : k_s = 0 \\
& -\boldsymbol{\alpha}_s^\top \mathbf{e} - \sum_{i \in [SAS]} \hat{p}_i k_s \left(\exp\left(-\frac{\boldsymbol{Q}_{s,i}^\top \boldsymbol{u} + \boldsymbol{B}_{:,i}^\top \boldsymbol{\alpha}_s}{k_s} \right) - 1 \right) \geq \mathbf{e}^\top \boldsymbol{u}_s - d_s && \forall s : k_s > 0 \\
& \boldsymbol{r}^\top \boldsymbol{u} \geq \tau \\
& \boldsymbol{u} \in \mathbb{R}_+^{S \cdot A},\ \boldsymbol{k} \in \mathbb{R}_+^S, \boldsymbol{\alpha}_s \in \mathbb{R}^{S \cdot A} && \forall s \in \mathcal{S}.
\end{aligned}
$$

Our conclusion then follows from Lemma B.2. □

C. Discussions on Computation Complexities

For value iteration, the time complexity per iteration for solving NMDPs is $\mathcal{O}(S^2 A)$. For robust value iteration, the time complexity per iteration for solving RMDPs differs for different ambiguity sets. In general, for s-rectangular ambiguity set, the practical time complexity per iteration could be at least $\mathcal{O}(S^4 A^3)$ (Boyd & Vandenberghe, 2004). There are recent algorithmic developments for specific ambiguity sets; for example, the time complexity per iteration for solving RMDPs with an unweighted L_1-norm ambiguity set could be reduced to $\mathcal{O}(S^2 A \log S)$ (Ho et al., 2021). As mentioned in Section 4.3, the proposed $\mathrm{PDA}_{\mathrm{block}}$ and $\mathrm{PDA}_{\mathrm{block+}}$ have time complexities of $\mathcal{O}(MS^2 \log(S) A \log(\delta^{-1}))$ and $\mathcal{O}(S \log S)$ (with high probability as $\mathrm{PDA}_{\mathrm{block+}}$ is a randomized algorithm), respectively.

D. Additional Details and Results on the Experiments

In our experiments, the reward functions are deterministic and known to the agent while the transition kernel is uncertain. We adopt data-driven setups and evaluate models based on their out-of-sample performances. All optimization problems are solved on an Intel 3.6 GHz processor with 32GB RAM.

In the experiment in Section 5.1, we vary the training sample size [4] among $\{10, 20, \cdots, 140, \mathrm{M}\}$ with some sufficiently large M such that the empirical transition kernel is close enough to the true one, and we generate 5000 out-of-sample testing trajectories of length 100. The parameters of RMDPs and DRMDPs (*i.e.*, the radius r of the ambiguity set) are selected from $r \in [0.0, 1.8]$ and those of RSMDPs (*i.e.*, the target τ) are selected from $\tau \in [0.5\mathrm{Z}_{\mathrm{N}}, 1.0\mathrm{Z}_{\mathrm{N}}]$, via cross validation.

In the experiment in Section 5.2, τ is the predicted return for RSMDPs, while $\mathrm{Z}_{\mathrm{N}} = \boldsymbol{d}^\top \boldsymbol{v}_{\mathrm{N}}^\star$, $\boldsymbol{d}^\top \boldsymbol{v}_{\mathrm{R}}^\star$ and $\boldsymbol{d}^\top \boldsymbol{v}_{\mathrm{DR}}^\star$ are the predicted returns of NMDPs, RMDPs and DRMDPs, respectively, where the optimal value function $\boldsymbol{v}_{\mathrm{N}}^\star$ is computed via (1), and $\boldsymbol{v}_{\mathrm{R}}^\star$ and $\boldsymbol{v}_{\mathrm{DR}}^\star$ are via robust Bellman optimality equation, respectively (see details in Appendices D.2 and D.3, respectively), all with the true $\boldsymbol{p}^\star$. All sample returns are computed by Bellman equation with the polluted transition kernel.

In the experiment in Section 5.3, we implemented our algorithms in C++, whereas Gurobi is also called from C++. In our experiments, we synthetically generate random RSMDP instances, and the details of the experiments and parameters can be found in the Appendix D.7, which also includes additional experimental results with $\mathrm{PDA}_{\mathrm{block+}}$, a variant of PDA where the dual updating step follows the second strategy mentioned in Section 4.3.

D.1. Setup for Initial Transition Kernel

For the experiments in Section 5.1, the next state space for each state-action pair is given to the agent at the beginning when the state space and action space are big (*e.g.*, *grid world*), but such information is unknown to the agent when the state space and action space are small (*e.g.*, *machine replacement*, *river swim*), otherwise the optimal policy can be obtained immediately with the initial transition kernel using any methods.

[4] Here "sample size" is the number of transitions that an agent experiences (with a uniform initial distribution).

Algorithm 4 Solve the inner minimization problem in (26)

Input: Value function $\boldsymbol{v}$, transition kernel $\boldsymbol{p} = \hat{\boldsymbol{p}}_{s,a}$
Sort $\boldsymbol{p}$ such that its corresponding $\boldsymbol{v}$ is non-decreasing, indexed as $1, \cdots, S$;
$\boldsymbol{o} \leftarrow \text{copy}(\boldsymbol{p})$;
$i \leftarrow S$;
$b \leftarrow \min\left\{1 - p_1, \frac{r}{2}\right\}$;
$o_1 \leftarrow b + p_1$;
repeat
  $o_i \leftarrow o_i - \min\{b, o_i\}$;
  $b \leftarrow b - \min\{b, o_i\}$;
  $i \leftarrow i - 1$;
until $b \leq 0$ **or** $i < 0$
$\boldsymbol{p}^\star \leftarrow \boldsymbol{o}$;
Output: Optimal solution $\boldsymbol{p}^\star$

D.2. Additional Details on Robust MDPs

Robust MDPs (RMDPs) maximize the total expected return considering the worst-case realization of the uncertain parameter within a predefined ambiguity set $\bar{\mathcal{P}}$:

$$\mathrm{Z_R} = \max_{\boldsymbol{\pi} \in \Pi} \min_{\boldsymbol{p} \in \bar{\mathcal{P}}} \boldsymbol{d}^\top \boldsymbol{v}(\boldsymbol{\pi}, \boldsymbol{p}). \tag{25}$$

In the experiments, we adopt the popular sa-rectangular ambiguity set (Behzadian et al., 2021a; Iyengar, 2005; Nilim & El Ghaoui, 2005; Strehl et al., 2009; Weissman et al., 2003):

$$\bar{\mathcal{P}} = \left\{\boldsymbol{p} \in \mathbb{R}_+^{S \cdot A \cdot S} \mid \boldsymbol{p}_{s,a} \in \bar{\mathcal{P}}_{s,a},\ \mathbf{e}^\top \boldsymbol{p}_{s,a} = 1\ \forall s \in \mathcal{S},\ a \in \mathcal{A}\right\},$$

where $\bar{\mathcal{P}}_{s,a} = \left\{\boldsymbol{p} \in \Delta^S \mid \|\boldsymbol{p} - \hat{\boldsymbol{p}}_{s,a}\|_1 \leq r\right\}$ with $\hat{\boldsymbol{p}}_{s,a} \in \Delta^S$ being the subvector of empirical transition kernel $\hat{\boldsymbol{p}} \in \bar{\mathcal{P}}$; see, *e.g.*, (Petrik & Subramanian, 2014; Russel et al., 2019). To calculate the optimal policy in RMDP, we utilise the value iteration method, where in each iteration, the robust Bellman optimality equation:

$$v_s^\star = \max_{a \in \mathcal{A}} \min_{\boldsymbol{p} \in \bar{\mathcal{P}}_{s,a}} \left\{r_{s,a} + \gamma \cdot \boldsymbol{p}^\top \boldsymbol{v}^\star\right\} = \max_{a \in \mathcal{A}} \min_{\boldsymbol{p} \in \Delta^S} \left\{r_{s,a} + \gamma \cdot \boldsymbol{p}^\top \boldsymbol{v}^\star \mid \|\boldsymbol{p} - \hat{\boldsymbol{p}}_{s,a}\|_1 \leq r\right\} \quad s \in \mathcal{S} \tag{26}$$

is solved to update our value function. We adopt Algorithm 4 to solve the inner minimization problem in (26) (Petrik & Subramanian, 2014).

We remark the the input $\boldsymbol{v}$ in Algorithm 4 is given by the value function at the last iteration in the value iteration process, where it is initialized as $\mathbf{0}$ at the first iteration.

D.3. Additional Details on Distributionally Robust MDPs

We follow the settings in (Grand-Clément & Kroer, 2021a) where the ambiguity set is defined using Wasserstein distance with L_p-norm $\|\cdot\|_p$ (for some $p \in \mathbb{R} \cup \{\infty\}$). In particular, we solve the following distributionally robust Bellman equation

$$v_s^\star = \max_{\boldsymbol{\pi}_s \in \Delta^A} \min_{\boldsymbol{p}^1, \cdots, \boldsymbol{p}^N \in \hat{\mathcal{P}}} \left\{\sum_{a \in \mathcal{A}} \pi_{s,a} \cdot \left(r_{s,a} + \gamma \cdot \frac{1}{N} \sum_{i=1}^N \boldsymbol{p}_{s,a}^{i\top} \boldsymbol{v}^\star\right) \;\middle|\; \frac{1}{N} \sum_{i=1}^N \|\boldsymbol{p}_s^i - \hat{\boldsymbol{p}}_s^i\|_p \leq \mu_{\mathrm{W}}^p \right\},$$

where μ_{W} is the radius and $\hat{\boldsymbol{p}}^1, \ldots, \hat{\boldsymbol{p}}^N \in \mathcal{P}$ are N samples of transition kernels (Bertsimas et al., 2018; Xie, 2020; Yang, 2017). In the experiments, we set p to be 1 and Mosek (academic license) is utilized to solve the inner minimization problem in each iteration (MOSEK ApS, 2022).

D.4. Additional Details on Environments

We use a discounted factor $\gamma = 0.85$ for all environments, and the objective is always maximizing (total discounted) rewards.

Machine Replacement: we have 2 repair options constituting our action set [“repair”, “do nothing”] and 10 states. The rewards relate only to the states, which are [20, 20, 20, 20, 20, 20, 20, 0, 18, 10].

River Swim: we have 2 swimming directions constituting our action set [“move left”, “move right”] and 10 states, and the rewards relate only to the state, which are [5, 0, 0, 10, 10, 10, 10, 10, 10, 15].

Grid World: the grid world has two rows and 12 columns, and the rewards relate to the column indices only, which are [0, 3, 21, 27, 6, 0, 0, 0, 0, 0, 15, 24]. There are four available actions, “move up” and “move down” for vertical moves (that decreases and increases the column index, respectively), as well as “move left”, and “move right” for horizontal moves (that decreases and increases the row index, respectively). Horizontal moves have a chance of failure that only related to row indices (0.9 for the first row and 0.2 for the second). Failing a transfer or selecting a vertical move would generate the column index of the next state according to a Dirichlet distribution. After selecting a horizontal move, the agent will randomly go up, go down, or stay with probabilities 0.35, 0.35 and 0.3, respectively.

D.5. Additional Results on the Improvements on Percentiles

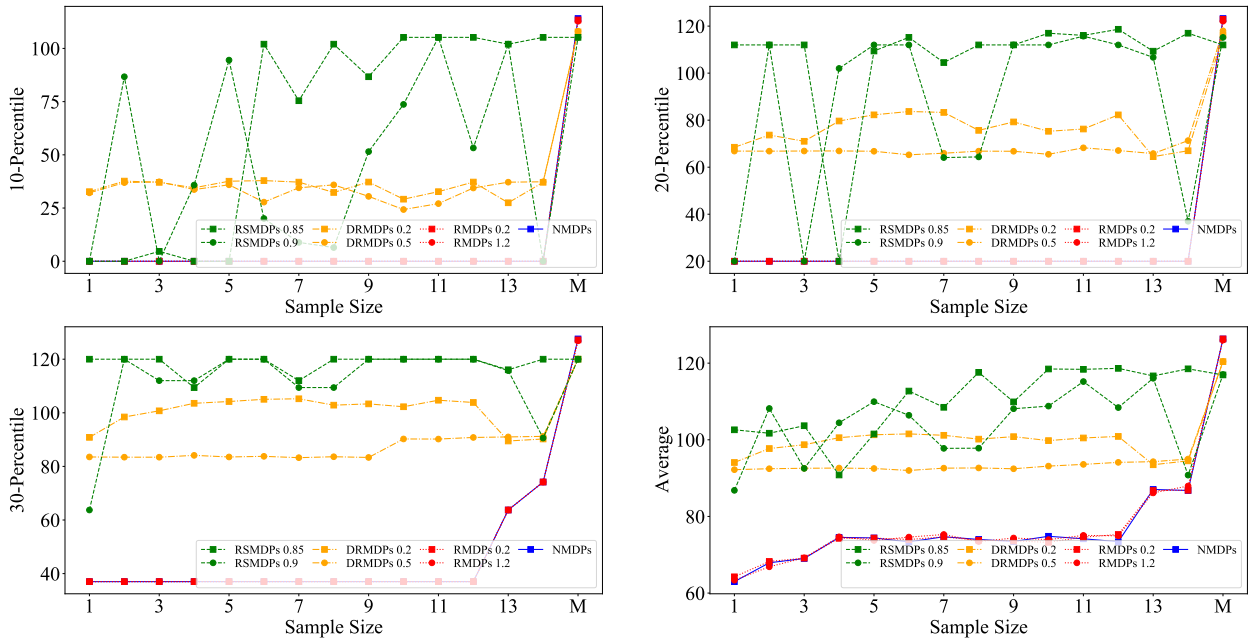

Figure 3. Average and percentile performances over 5000 out-of-sample testing trajectories in the *machine replacement* application.

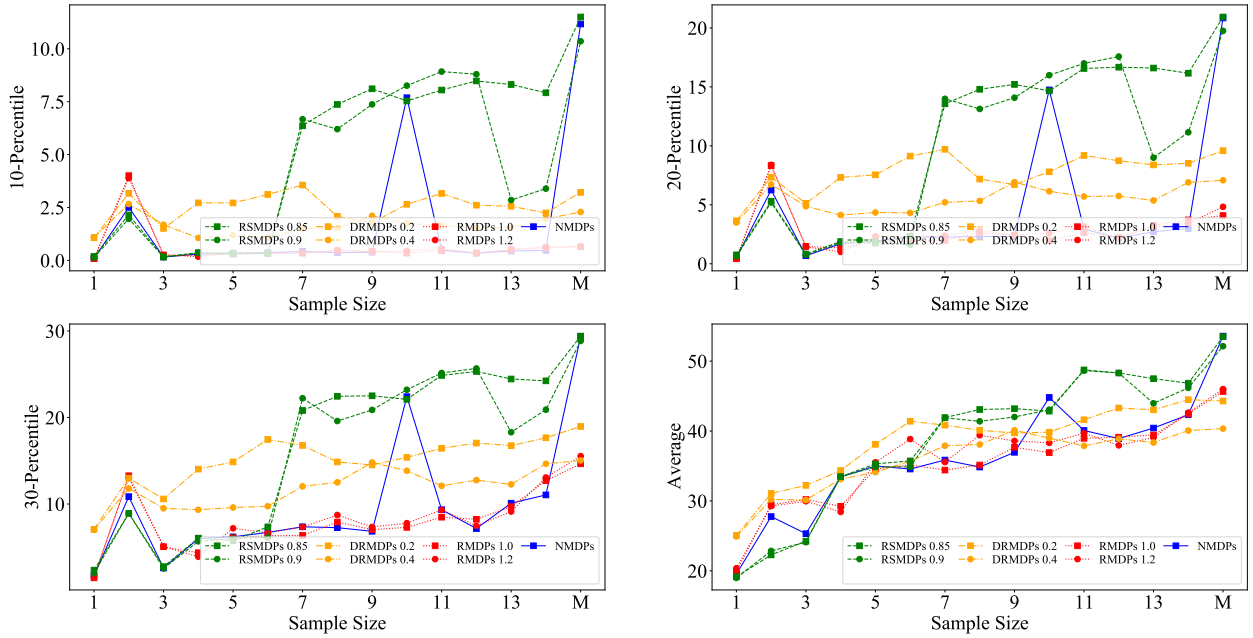

Figure 4. Average and percentile performances over 5000 out-of-sample testing trajectories in the *grid world* application.

D.6. Additional Results on the Target-Oriented Feature

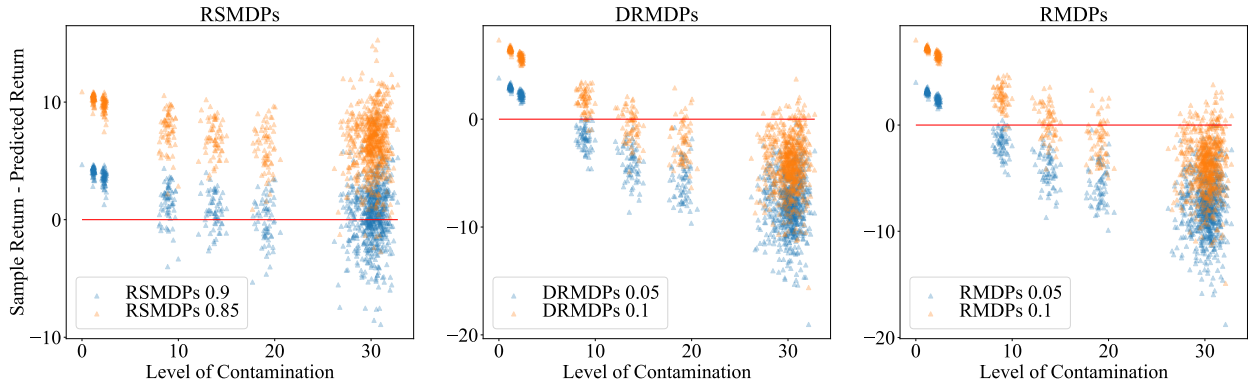

Figure 5. Differences between sample returns and predicted returns over 1000 samples in the *machine replacement* environment. Due to the page limit, we only plot the parameters (of the three models) that the sample returns are close to predicted returns.

Table 3. Predicted returns and the corresponding differences (in median) between sample returns and predicted returns over 1000 samples in the *machine replacement* application.

RSMDPs	τ / $\mathrm{Z_N}(\hat{\boldsymbol{p}})$	1.0	0.9	0.8	0.7	0.6	0.5
	Predicted Return	123.9	111.5	99.1	86.7	74.4	62.0
	SamRet-PreRet (Median)	-10.5	**1.0**	**13.4**	**25.8**	**38.2**	**50.9**
DRMDPs	r	0.0	0.3	0.6	0.9	1.2	1.5
	Predicted Return	123.9	105.4	88.2	68.5	45.2	31.5
	Difference (in median)	-10.5	**7.6**	**24.8**	**44.0**	**66.9**	**79.1**
RMDPs	r	0.0	0.3	0.6	0.9	1.2	1.5
	Predicted Return	123.9	99.9	76.1	52.8	30.2	20.7
	SamRet-PreRet (Median)	-10.5	**13.7**	**37.5**	**60.8**	**83.4**	**92.9**

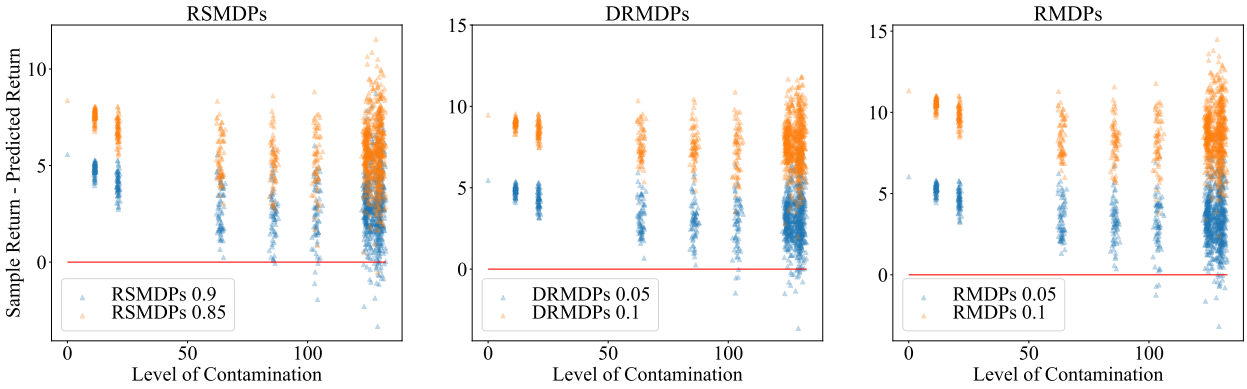

Figure 6. Differences between sample returns and predicted returns over 1000 samples in the *grid world* environment. Due to the page limit, we only plot the parameters (of the three models) that the sample returns are close to predicted returns.

D.7. Detailed Numerical Results on PDAs

In our experiments, we generate random instances as follows. The entries in reward function $\boldsymbol{r}$, initial distribution $\boldsymbol{d}$, and transitional kernel $\hat{\boldsymbol{p}}$ are sampled from an uniform distribution in $[0, 1]$. The entries in $\boldsymbol{d}$ and $\hat{\boldsymbol{p}}$ are then normalized so that $\boldsymbol{d}$ and $\hat{\boldsymbol{p}}_{s,a}$ are elements in a probability simplex, for all $s \in \mathcal{S},\ a \in \mathcal{A}$. We set the discount factor $\gamma = 0.95$, and the target $\tau = 0.85\mathrm{Z_N}$.

We denote PDA as the proposed Algorithm 1. $\mathrm{PDA_{block}}$ is a variant of PDA where the dual updating step (*i.e.* Step 2 in Algorithm 1) follows the first strategy mentioned in Section 4.3 in which the dual updating step becomes

$$(\lambda_s^{k+1}, \boldsymbol{\theta}_s^{k+1}) \leftarrow \mathfrak{D}_s(2\boldsymbol{u}^{k+1} - \boldsymbol{u}^k; \lambda_s^k, \boldsymbol{\theta}_s^k)\ \forall s \in \mathcal{S}^k. \tag{27}$$

Here $\mathcal{S}^k$ with $|\mathcal{S}^k| = M \ll S$ is an index set where its elements are sampled uniformly from $\mathcal{S}$ without replacement in each iteration, and we set $M = 2$. On the other hand, $\mathrm{PDA_{block+}}$ is a PDA where the dual updating step follows the second strategy mentioned in Section 4.3. In particular, at each iteration, $(\hat{s}, s, a)$ is sampled uniformly, we perform the update (27) with probability P and otherwise we update $\boldsymbol{\theta}_{s,a}$ by solving problem (11), which is the subproblem corresponded to problem (10). We set $P = 1/(SA)$ in this experiment.

In the experiment, we also compare our PDAs to Gurobi solver for RMDPs. Since the overhead (in terms of computation time) can be dominating, we only report the runtime of the Gurobi solver.

Figure 7 reports the average of computation times and per-iteration computational times (relative to Gurobi) over 20 generated test instances. The vertical bars indicate the average $\pm$ standard deviations. Since PDA is a first-order method, it is impractical to expect solution with extreme precision with optimally. Hence, we terminate PDA when $|f_{\mathrm{PDA}} - f_{\mathrm{Gurobi}}|/f_{\mathrm{Gurobi}} < 5\%$, where f_{PDA} and f_{Gurobi} are the objective values computed by PDA and Gurobi, respectively. The same stopping criteria is

Table 4. Predicted returns and the corresponding differences (in median) between sample returns and predicted returns over 1000 samples in the *grid world* application.

RSMDPs	τ / $Z_N(\hat{\boldsymbol{p}})$	1.0	0.9	0.8	0.7	0.6	0.5
	Predicted Return	55.9	50.3	44.7	39.1	33.5	27.9
	SamRet-PreRet (Median)	-2.4	**3.1**	**8.7**	**14.3**	**19.8**	**25.4**
DRMDPs	r	0.0	0.3	0.6	0.9	1.2	1.5
	Predicted Return	55.9	31.8	21.1	14.7	11.0	9.4
	Difference (in median)	-2.5	**21.1**	**31.6**	**38.0**	**41.6**	**42.8**
RMDPs	r	0.0	0.3	0.6	0.9	1.2	1.5
	Predicted Return	55.9	26.6	15.5	11.1	9.4	8.5
	SamRet-PreRet (Median)	-2.5	**26.7**	**38.2**	**42.6**	**44.3**	**45.2**

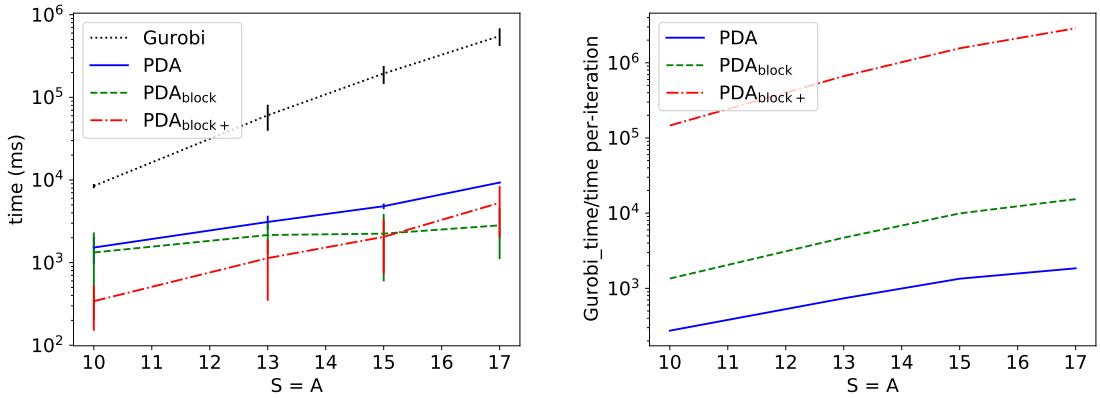

Figure 7. Left: Computation times (in ms) of different algorithms and Gurobi. *Right*: The ratio of Gurobi's computation time to the per-iteration computation times of PDA, $\text{PDA}_{\text{block}}$, and $\text{PDA}_{\text{block+}}$.

used for $\text{PDA}_{\text{block}}$ and $\text{PDA}_{\text{block+}}$. To ensure algorithms will terminate within a reasonable amount of time, we also set their maximal numbers of iterations to be 2000, 20000, and 400000 for PDA, $\text{PDA}_{\text{block}}$, and $\text{PDA}_{\text{block+}}$, respectively.

Figure 7 shows that as the problem size increases, the computational time of Gurobi increases rapidly compared to PDA, $\text{PDA}_{\text{block}}$, and $\text{PDA}_{\text{block+}}$. PDA also exhibits similar performance, but with a slower rate compared to Gurobi. The right figure of Figure 7 demonstrates the scalability of the proposed algorithms. As the problem size increase, the per-iteration computation times are remarkably cheaper compared to Gurobi. This phenomenon identifies the advantage of the proposed PDAs, and it matches the expected algorithmic behaviors for first-order methods, which are often proposed to efficiently solve large problems to moderate accuracy by computationally cheaper updates.

E. Limitations and Potential Negative Societal Impact

One limitation of this work is that the RSMDPs are not (yet) solvable by a dynamic programming approach. Another limitation would be the lack of further exploration about how the optimal policy and robust value function of RMDPs can indicate the range of the targeted return of RSMDPs. Potential negative societal impact is not applicable to this work.