Beyond Cumulative Returns via Reinforcement Learning over State-Action Occupancy Measures

Junyu Zhang*[1], Amrit Singh Bedi*[2], Mengdi Wang[1] and Alec Koppel[2]

***Abstract*— We study the estimation of risk-sensitive policies in reinforcement learning problems defined by a Markov Decision Process (MDPs) whose state and action spaces are countably finite. Prior efforts are predominately afflicted by computational challenges associated with the fact that risk-sensitive MDPs are time-inconsistent. To ameliorate this issue, we propose a new definition of risk, which we call *caution*, as a penalty function added to the dual objective of the linear programming (LP) formulation of reinforcement learning. The caution measures the distributional risk of a policy, which is a function of the policy's long-term state occupancy distribution. To solve this problem in an online model-free manner, we propose a stochastic variant of primal-dual method that uses Kullback-Lieber (KL) divergence as its proximal term. We establish that the number of iterations/samples required to attain approximately optimal solutions of this scheme matches tight dependencies on the cardinality of the state and action spaces, but differs in its dependence on the infinity norm of the gradient of the risk measure. Experiments demonstrate the merits of this approach for improving the reliability of reward accumulation without additional computational burdens.**

I. Introduction

In reinforcement learning (RL) [1], an autonomous agent in a given state selects an action and then transitions to a new state randomly depending on its current state and action, and then the environment reveals a reward. This framework for sequential decision making has gained traction in recent years due to its ability to effectively describe problems where the long-term merit of decisions does not have an analytical form and is instead observed only in increments, as in recommender systems [2], videogames [3], [4], control amidst complicated physics [5], and management applications [6].

The canonical performance metric for RL is the expected value of long-term accumulation of rewards also called as expected returns. Unfortunately, restricting focus to expected returns fails to encapsulate many well-documented aspects of reasoning under uncertainty such as anticipation [7], inattention [8], and risk-aversion [9]. In this work, we focus on risk (objective beyond expected rewards), both due to its inherent value in behavioral science and in pursuit of improving the reliability of RL in safety-critical applications [10].

Risk-awareness broadens the focus of decision making from expected outcomes to other quantifiers of uncertainty. Risk, originally quantified using the variance in portfolio management [11], has broaden to higher-order moments or quantiles [12], and gave rise to a rich theory of coherent risk [13], which has gained attention in RL in recent years [14], [15] as a frequentist way to define uncertainty-aware decision-making.

Incorporating risk gives rise to computational challenges in RL. In particular, if one replaces the expectation in the value function by a risk measure, the MDP becomes *time-inconsistent* [16], that is, Bellman's principle of optimality does not hold. This issue has necessitated modified Bellman equations [17], multi-stage schemes [15], or policy search [18], all of which do not attain near-optimal solutions in polynomial time, even for finite MDPs. Alternatively, one may impose risk as a probabilistic constraint [19], [20], [14], [21], [22] in the spirit of chance-constrained programming [23] common in model predictive control.

An additional approach is Bayesian [24] and distributional RL [25], which seeks to track a full posterior over returns. These approaches benefit from the fact that with access to a full distribution, one may define risk specifically, with, e.g., conditional value at risk (CVaR) [26]. One limitation is that succinctly parameterizing the value distribution intersects with approximate Bayesian computation, an active area of research [27].

In this paper, we seek to define risk in sequential decision making that (1) provides a tunable tradeoff between the mean return and uncertainty of a decision; (2) captures long-term behaviors of policies that cannot be modeled using cumulative functions; (3) can be solved efficiently in polynomial time, depending on the choice of risk. To do so, we formulate a class of distributional risk-averse policy optimization problems to address risks involving the long-term behaviors that permit the derivation of efficient algorithms. More specifically, we:

- propose a new definition of the risk of a policy, which we call *caution*, as a function of the policy's long-term state-action occupancy distribution. We formulate a caution-sensitive policy optimization problem by adding the caution risk as a penalty function to the dual objective of the linear programming (LP) formulation of RL. The caution-sensitive optimization problem is often convex, allowing us to directly design the policy's long-term occupancy distribution (Sec. III).
- derive an online model-free algorithm based on a stochastic variant of primal-dual policy gradient method that uses Kullback-Lieber (KL) divergence as its proximal term (Sec. IV).
- establish that the number of sample transitions required to attain approximately optimal solutions of this scheme

*Equal contribution

[1]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 {junyuz, mengdiw}@princeton.edu

[2]CISD, US Army Research Laboratory, Adelphi, MD, USA alec.e.koppel.civ@mail.mil

matches tight dependencies on the cardinality of the state and action spaces, as compared to the typical risk-neutral setting (Sec. V).

Further, we demonstrate the experimental merits of this approach for improving the reliability of reward accumulation without additional computational burdens (Sec. VI)

II. Preliminaries

A. Discounted Markov Decision Process

We consider the problem of reinforcement learning (RL) with finitely many states and actions as mathematically described by a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$. For each state $s \in \mathcal{S}$, a transition to state $s' \in \mathcal{S}$ occurs when selecting action $a \in \mathcal{A}$ according to a conditional probability distribution $s' \sim \mathcal{P}(\cdot|a, s)$, for which we define the short-hand notation $P_a(s, s')$. Moreover, a reward $\hat{r} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is revealed and is denoted as $\hat{r}_{ss'a}$. Without loss of generality, we assume $\hat{r}_{ss'a} \in [0, 1]$ with probability 1 for $\forall s, s' \in \mathcal{S}$ and $\forall a \in \mathcal{A}$ throughout the paper. For future reference, we denote the expected reward with respect to transition dynamics as $r_{sa} := \mathbb{E}[\hat{r}_{ss'a}|s, a] = \sum_{s' \in \mathcal{S}} P_a(s, s') \cdot \hat{r}_{ss'a}$ and the vector of rewards for each action a as $r_a = [r_{1a}, \cdots, r_{|\mathcal{S}|a}]^T \in \mathbb{R}^{|\mathcal{S}|}$.

In standard (risk-neutral) RL, the goal is to find the action sequence which yields the most long-term reward, or value:

$$v^*(s) := \max_{\{a_t \in \mathcal{A}\}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \hat{r}_{s_t s_{t+1} a_t} \;\middle|\; s_0 = s\right], \quad \forall s \in \mathcal{S}. \tag{II.1}$$

B. Bellman Equation and Duality

The optimal value function v^* (II.1) satisfies Bellman's optimality principle [28]:

$$v^*(s) = \max_{a \in \mathcal{A}} \left\{\gamma \sum_{s' \in \mathcal{S}} P_a(s, s') v^*(s) + \sum_{s' \in \mathcal{S}} P_a(s, s') \hat{r}_{ss'a}\right\} \tag{II.2}$$

for all $s \in \mathcal{S}$. Then, due to [29], the Bellman optimality equation (II.2) may be reformulated as a linear program (LP)

$$\begin{array}{ll} \min_{v \geq 0} & \langle \xi, v \rangle \\ \text{s.t.} & (I - \gamma P_a)v - r_a \geq 0, \quad \forall a \in \mathcal{A} \end{array} \tag{II.3}$$

where ξ is an arbitrary positive vector. The dual of (II.3) is given as

$$\begin{array}{ll} \max_{\lambda \geq 0} & \sum_{a \in \mathcal{A}} \langle \lambda_a, r_a \rangle \\ \text{s.t.} & \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top)\lambda_a = \xi, \quad \forall a \in \mathcal{A} \end{array} \tag{II.4}$$

where $\lambda_a = [\lambda_{1a}, \cdots, \lambda_{|\mathcal{S}|a}]^\top \in \mathbb{R}^{|\mathcal{A}|}$ is the a-th column of λ. To the subsequent development, an essential fact is that λ is an unnormalized state-action occupancy measure and

$$\sum_{a \in \mathcal{A}} \langle \lambda_a, r_a \rangle = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_{s_t s_{t+1} a_t} \;\middle|\; s_0 \sim \xi, a_t \sim \pi(\cdot|s_t)\right]$$

when ξ belongs to the probability simplex. The dual LP formulation (II.4) has a clear physical meaning. Suppose $\xi \geq 0$ and $\|\xi\|_1 = 1$ is a distribution over the state space $\mathcal{S}$. Then the following proposition explains the meaning of the dual problem.

Proposition 2.1: Suppose the variable $\lambda \in \mathbb{R}_+^{|\mathcal{S}| \times |\mathcal{A}|}$ satisfies the conditions

$$\lambda \geq 0 \quad \text{and} \quad \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top)\lambda_a = \xi, \tag{II.5}$$

Then λ is an **unnormalized distribution**, or **flux**, under the randomized policy π:

$$\pi(a|s) = \frac{\lambda_{sa}}{\sum_{a' \in \mathcal{A}} \lambda_{sa'}}, \quad \text{for} \quad \forall a \in \mathcal{A}, \forall s \in \mathcal{S}. \tag{II.6}$$

Furthermore, it satisfies

$$\lambda_{sa} = \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\left(s_t = s, a_t = a \;\middle|\; s_0 \sim \xi, a_t \sim \pi(\cdot|s_t)\right) \tag{II.7}$$

and

$$\langle \lambda, r \rangle = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_{s_t s_{t+1} a_t} \;\middle|\; s_0 \sim \xi, a_t \sim \pi(\cdot|s_t)\right]. \tag{II.8}$$

Proof: Under the initial distribution ξ and the randomized policy $\pi : \mathcal{S} \mapsto \Delta_{|\mathcal{A}|}$ defined in (II.6), we define a new initial distribution $\hat{\xi}$ as

$$\hat{\xi}_{sa} = \xi_s \cdot \pi(a|s) \quad \text{for} \quad \forall s \in \mathcal{S}, a \in \mathcal{A}$$

as the distribution of the initial state-action pair (s_0, a_0). Therefore the dynamics of the state-action pairs (s_t, a_t) form another Markov chain with transition matrix $\hat{P} \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{A}||\mathcal{S}|}$ defined as

$$\hat{P}_\pi(s, a; s', a') = P_a(s, s') \cdot \pi(a'|s').$$

First, let us prove that (II.5) is equivalent to (II.7). For the ease of notation, we used the multi-indices. Let us view both r and λ as vectors with s, a being a multi-index. Note that (II.5) implies that for all $s \in \mathcal{S}$

$$\xi_s = \sum_{a' \in \mathcal{A}} \lambda_{sa'} - \gamma \sum_{a' \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_{a'}(s', s) \lambda_{s'a'}.$$

Multiplying both sides by $\pi(a|s) = \frac{\lambda_{sa}}{\sum_{a' \in \mathcal{A}} \lambda_{sa'}}$, we get

$$\begin{aligned} \hat{\xi}_{sa} &= \lambda_{sa} - \gamma \sum_{a' \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_{a'}(s', s) \cdot \pi(a|s) \cdot \lambda_{s'a'} \\ &= \lambda_{sa} - \gamma \sum_{a' \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \hat{P}_\pi(s', a'; s, a) \lambda_{s'a'} \end{aligned}$$

for any $s \in \mathcal{S}, a \in \mathcal{A}$. If we write this equation in a compact matrix form, we get

$$\hat{\xi} = (I - \gamma \hat{P}_\pi^\top)\lambda.$$

Note that $\|\gamma \hat{P}_\pi^\top\|_2 \leq \gamma < 1$, we know $(I - \gamma \hat{P}_\pi^\top)^{-1} = \sum_{i=0}^{\infty} \gamma^i (\hat{P}_\pi^i)^\top$. Consequently,

$$\lambda^\top = \hat{\xi}^\top (I - \gamma \hat{P}_\pi)^{-1} = \hat{\xi}^\top + \gamma \hat{\xi}^\top \hat{P}_\pi + \gamma^2 \hat{\xi}^\top \hat{P}_\pi^2 + \cdots$$

If we write the above equation in an elementwise way, we get (II.7). Consequently, we also have

$$\begin{aligned}\lambda^\top r &= \hat{\xi}^\top r + \gamma \hat{\xi}^\top \hat{P}_\pi r + \gamma^2 \hat{\xi}^\top \hat{P}_\pi^2 r + \cdots \\ &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \hat{r}_{s_t s_{t+1} a_t} \;\middle|\; s_0 \sim \xi, a_t \sim \pi(\cdot|s_t)\right],\end{aligned}$$

which is as stated in (II.8) ∎

Hence, one can recover the policy parameters through normalization of the dual variable as $\pi(a|s) = \lambda_{sa} / \sum_{a' \in \mathcal{A}} \lambda_{sa'}$ for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, as detailed in Proposition 2.1.

III. Caution-Sensitive Policy Optimization

In this work, we prioritize definitions of risk in MDPs that capture long-term behavior of the policy and permit the derivation of computationally efficient algorithms. We focus on optimizing the policy's long-run behaviors that cannot be described by any cumulative sum of rewards, for examples the barrier risk and variance (Sec. III-B).

A. Problem Formulation

We focus on directly designing the long-term state-action occupancy distribution, whose unnormalized version is the dual variable $\lambda := \{\lambda_a\}_{a \in \mathcal{A}}$. Rather than only maximizing the expected cumulative return, i.e., the typical objective in risk-neutral MDP (e.g., (II.4)), we seek policies that incorporate risk functions concerning the full distribution λ.

We propose a non-standard notion of risk: in standard definitions, such as those previously mentioned, they are typically risk measures of the cumulative rewards; by contrast, here we augment the risk to be defined over the *long-term state-action occupancy distributions*, which we dub *caution* measures. Specifically, denote as $\rho(\lambda)$ a caution function that takes as input dual variables λ (unnormalized state-action distributions) feasible to (II.4) and maps to the reals $\mathbb{R}$. The caution risk measures the fitness of the entire state path, rather than just a cumulative sum over the path.

In pursuit of computationally efficient solutions, we hone in on properties of the dual LP formulation of RL. The caution-sensitive variant of (II.4) then takes the form:

$$\begin{aligned}\max_{\lambda \geq 0} \quad & \langle \lambda, r \rangle - c\rho(\lambda) \\ \text{s.t.} \quad & \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top) \lambda_a = \xi, \\ & \|\lambda\|_1 = (1-\gamma)^{-1},\end{aligned} \tag{III.1}$$

where c is a positive penalty parameter and we take ξ to be the vector of uniform distribution without loss of generality, i.e., $\xi = \frac{1}{|\mathcal{S}||\mathcal{A}|} \cdot \mathbf{1}$; and $\|\lambda\|_1 := \sum_{s,a} |\lambda_{sa}|$. The constraints require that λ be the unnormalized state-action distribution corresponding to *some* policy. The last constraint is implied by $\sum_{a \in \mathcal{A}} (I - \gamma P_a^\top) \lambda_a = \xi$, but we include it for clarity. In this work, we consider the scenarios when ρ is convex, which implies that the problem (III.1) is a convex optimization problem that facilitates computationally efficient solutions.

Let us denote the optimal solution to the cautious policy optimization problem (III.1) by λ^*. This λ^* gives the optimal long-term state-action occupancy distribution under the caution risk. Let π^* be the mixed policy given by

$$\pi^*(a|s) = \frac{\lambda^*(s,a)}{\sum_{a'} \lambda^*(s,a')}.$$

We call this π^* the *optimal caution-sensitive policy.* We remark that with the introduction of the risk measure into the dual form (III.1), the corresponding primal is no longer the LP problem (II.3) but changes to one that incorporates risk. The optimal caution-sensitive policy π^* differs from the optimal policy in the typical risk-neutral setting. Since the LP structure is lost, the optimal risk-sensitive policy π^* is not guaranteed to be deterministic. Moreover, the Lagrangian multipliers, denoted by v^*, for the risk-sensitive problem (III.1) is no longer the risk-neutral value vector, meaning that we are solving a *different problem* than (II.1). Indeed, by defining caution in this way, we incorporate long-term distributional risk into the dual domain of Bellman equation, while sidestepping the computational challenges of time-inconsistency.

B. Examples of Caution Risk

Next, we discuss several examples of the caution risk ρ to clarify the problem setting (III.1).

*Example 3.1 (***Barrier risks):** Caution risk can take the form of barriers to guarantee that a policy's long-term behavior meets certain expectations. Two examples follow:

• *Staying in safety set.* Suppose we want to keep the state trajectory within a safety set $\bar{S} \subset \mathcal{S}$ for more than $1 - \delta$ fraction of all time. In light of the typical barrier risk used in constrained optimization, we define

$$\rho(\lambda) = -\log\left(\lambda(\bar{S}) - (1-\delta)\right),$$

where $\lambda(\bar{S}) = (1-\gamma) \sum_{s,a} \lambda(s,a) \mathbf{1}_{s \in \bar{S}}$. Since $\lambda(\bar{S})$ is linear in λ, we can verify that the log barrier risk ρ is convex.

• *Meeting multiple job requirements.* Further, suppose there are multiple tasks with strict requirements on their expected returns $\langle \lambda, r_j \rangle \geq b_j$, $j = 1, \ldots, m$. One can transform these return constraints into a log barrier given by

$$\rho(\lambda) = -\sum_{j=1}^{m} \log\left(\langle \lambda, r_j \rangle - b_j\right).$$

In this way, the optimal caution-sensitive policy will meet all the job requirements for large enough penalty c.

*Example 3.2 (***Variance risk):** In finance applications, one canonical risk concern is the variance of return. To formulate risk as variance, we first note that λ is an unnormalized distribution, whose normalized counterpart is denoted as $\hat{\lambda} := (1-\gamma)\lambda$. Then it holds that $\langle \hat{\lambda}, r \rangle$ is the expected reward accumulation. Then, the variance of return per timestep takes the form

$$\rho(\lambda) = Var(\hat{r}_{ss'a}|\lambda) = \mathbb{E}^{\hat{\lambda}}\left[\left(\mathbb{E}^{\hat{\lambda}}\left[\hat{r}_{ss'a}\right] - \hat{r}_{ss'a}\right)^2\right] \tag{III.2}$$

where $\mathbb{E}^{\hat{\lambda}} := \mathbb{E}_{(s,a,s') \sim \hat{\lambda} \times \mathcal{P}(\cdot|a,s)}$. For ease of notation, denote $R \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ with $R(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|a,s)}[\hat{r}^2_{ss'a}]$. Substituting in these definitions, we may write

$$\rho(\lambda) = \langle \hat{\lambda}, R \rangle - \langle \hat{\lambda}, r \rangle^2, \tag{III.3}$$

which is a quadratic function of the variable λ. Note that the variance risk $\rho(\lambda)$ is non-convex with respect to λ. Alternatively, we consider a surrogate for the variance function in (III.3) as

$$\begin{aligned}\rho(\lambda) &= \mathbb{E}^{\mu}\left[\left(\mathbb{E}^{\hat{\lambda}}\left[\hat{r}_{ss'a}\right]-\hat{r}_{ss'a}\right)^2\right] \\ &= \langle\hat{\lambda},r\rangle^2-2\langle\mu,r\rangle\langle\hat{\lambda},r\rangle+\langle\mu,R\rangle \quad \text{(III.4)}\end{aligned}$$

where μ is some predetermined distribution and $\mathbb{E}^{\mu} := \mathbb{E}_{(s,a,s')\sim\mu\times\mathcal{P}(\cdot|a,s)}$. Note that this risk function in (III.4) is convex in λ. We investigate the merit of choosing (III.3) and (III.4) in later sections.

*Example 3.3 (***Divergence for incorporating priors):** Often in applications, we have access to demonstrations, which can be used to learn a prior on the policy for ensuring baseline performance. Let $\bar{\lambda}$ be a prior state-action distribution learned from demonstrations. Maintaining baseline performance with respect to this prior, or demonstration distribution, then can be encoded as the Kullback-Liebler (KL) divergence between the normalized distribution $\hat{\lambda}=(1-\gamma)\lambda$ and the prior $\bar{\lambda}$ stated as

$$\rho(\lambda)=\text{KL}\left((1-\gamma)\lambda||\bar{\lambda}\right) \quad \text{(III.5)}$$

which is substituted into (III.1) to obtain a framework for efficiently incorporating a baseline policy. In some scenarios, existing demonstrations are only state trajectories without revealing the actions taken. Then one may estimate the long-term state-only distribution μ and define the risk as

$$\rho(\lambda)=\text{KL}\left((1-\gamma)\sum_a\lambda_a||\mu\right),$$

which measures the divergence between the marginalized state occupancy distribution and the prior. In addition to KL, one can also use other convex distances such as Wasserstein, total variation, or even a simple quadratic.

IV. Stochastic Primal-Dual Policy Gradient

We shift focus to developing an algorithmic solution to the caution-sensitive policy optimization problem (III.1). While the problem upon first glance appears deterministic, the transition matrices P_a are a priori unknown and we assume the presence of a generative model. Such a generative model is fairly common in control/RL applications where a system simulator is available. For a given state action pair (s,a), the generative model provides the next state s' and the stochastic reward $\hat{r}_{ss'a}$ according to the unknown transition dynamics.

Thus, we propose methodologies based on Lagrangian duality together with stochastic approximation. Given the convexity of ρ, by virtue of duality, (III.1) admits an equivalent formulation as a saddle point problem:

$$\max_{\lambda\in\mathcal{L}}\min_{v\in\mathcal{V}} L(v,\lambda)=\langle\lambda,r\rangle-c\rho(\lambda)+\langle\xi,v\rangle+\sum_{a\in\mathcal{A}}\lambda_a^{\top}(\gamma P_a-I)v, \quad \text{(IV.1)}$$

where $\mathcal{V}$ should be $\mathbb{R}^{|\mathcal{S}|}$ in principle. However, we can later on find a large enough compact set to replace the whole space without loss of optimality. By choosing ξ to satisfy $\xi\geq 0$ and $\|\xi\|_1=1$, we define the dual feasible set $\mathcal{L}$ as

$$\mathcal{L}:=\{\lambda:\lambda\geq 0,\|\lambda\|_1=(1-\gamma)^{-1}\}. \quad \text{(IV.2)}$$

Given distribution ζ over $\mathcal{S}\times\mathcal{A}$, define the stochastic approximation of the risk-neutral component of the Lagrangian:

$$L^{\zeta}_{(s,a,s'),\bar{s}}(v,\lambda):=v_{\bar{s}}+\mathbf{1}_{\{\zeta_{sa}>0\}}\cdot\frac{\lambda_{sa}(\hat{r}_{ss'a}+\gamma v_{s'}-v_s)}{\zeta_{sa}} \quad \text{(IV.3)}$$

where $\bar{s}\sim\mathcal{P}(\xi)$ is a sample from the discrete distribution defined by probability vector ξ. Then by direct computation, when the support of ζ contains that of λ, i.e., $\text{supp}(\lambda)\subset\text{supp}(\zeta)$, we may write

$$L(v,\lambda)=\mathbb{E}_{(s,a,s')\sim\zeta\times\mathcal{P}(\cdot|a,s),\bar{s}\sim\xi}\left[L^{\zeta}_{(s,a,s'),\bar{s}}(v,\lambda)\right]-c\rho(\lambda). \quad \text{(IV.4)}$$

Thus, we view (IV.1) as a stochastic saddle point problem.

We propose variants of stochastic primal-dual method applied to (IV.1). To obtain the primal descent direction, we note that if ζ is chosen such that $\text{supp}(\lambda)\subset\text{supp}(\zeta)$, an unbiased estimator of the gradient of L w.r.t. $v\in\mathcal{V}$ is

$$\begin{aligned}\hat{\nabla}_v L(v,\lambda) &:= \nabla_v L^{\zeta}_{(s,a,s'),\bar{s}}(v,\lambda) \quad \text{(IV.8)}\\ &= \mathbf{e}_{\bar{s}}+\mathbf{1}_{\{\zeta_{sa}>0\}}\cdot\frac{\lambda_{sa}}{\zeta_{sa}}(\gamma\mathbf{e}_{s'}-\mathbf{e}_s),\end{aligned}$$

where $\mathbf{e}_s\in\mathbb{R}^{|\mathcal{S}|}$ is a column vector with only the s-th entry equaling to 1 and all other entries being 0. Moreover, a dual subgradient of the instantaneous Lagrangian is given as

$$\begin{aligned}\hat{\partial}_\lambda L(v,\lambda) &:= \mathbf{1}_{\{\zeta_{sa}>0\}}\cdot\frac{\hat{r}_{ss'a}+\gamma v_{s'}-v_s-M_1}{\zeta_{sa}}\cdot\mathbf{E}_{s,a}\\ &\quad -c\hat{\partial}\rho(\lambda)-M_2\cdot\mathbf{1}, \quad \text{(IV.9)}\end{aligned}$$

Algorithm 1 Stochastic Risk-Averse (Cautious) RL

Input: Sample size T. Parameter $\xi=\frac{1}{|\mathcal{S}|}\cdot\mathbf{1}$. Stepsizes $\alpha,\beta>0$. Discount $\gamma\in(0,1)$. Constants $M_1,M_2>0$, $\delta\in(0,1)$.
Initialize: Arbitrary $v^1\in\mathcal{V}$ and $\lambda^1:=\frac{1}{|\mathcal{S}||\mathcal{A}|(1-\gamma)}\cdot\mathbf{1}\in\mathcal{L}$.
for $t=1,2,\cdots,T$
Set $\zeta^t:=(1-\delta)(1-\gamma)\lambda^t+\frac{\delta}{|\mathcal{S}||\mathcal{A}|}\mathbf{1}$.
Sample $(s_t,a_t)\sim\zeta^t$ and $\bar{s}_t\sim\xi$.
Generate $s'_t\sim\mathcal{P}(\cdot|a_t,s_t)$ & $\hat{r}_{s_ts'_ta_t}$ from generative model.
Construct $\hat{\nabla}_v L(v^t,\lambda^t)$ [cf. (IV.8)] and $\hat{\partial}_\lambda L(v^t,\lambda^t)$ [cf. (IV.9)]
Update v and λ as

$$v^{t+1}=\Pi_{\mathcal{V}}(v^t-\alpha\hat{\nabla}_v L(v^t,\lambda^t)) \quad \text{(IV.5)}$$

and

$$\begin{aligned}\lambda^{t+\frac{1}{2}} =& \operatorname*{argmax}_{\lambda}\langle\hat{\partial}_\lambda L(v^t,\lambda^t),\lambda-\lambda^t\rangle \quad \text{(IV.6)}\\ &-\frac{1}{(1-\gamma)\beta}KL\big((1-\gamma)\lambda||(1-\gamma)\lambda^t\big).\\ \lambda^{t+1} =& \frac{\lambda^{t+\frac{1}{2}}}{(1-\gamma)\|\lambda^{t+\frac{1}{2}}\|_1}. \quad \text{(IV.7)}\end{aligned}$$

Output: $\bar{\lambda}:=\frac{1}{T}\sum_{t=1}^T\lambda^t$ and $\bar{v}:=\frac{1}{T}\sum_{t=1}^T v^t$.

where $\mathbf{E}_{s,a} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ is a matrix with (s,a)-th entry equal to 1 and all other entries equal to 0. $\hat{\partial}\rho(\lambda)$ is an unbiased subgradient estimate of the convex but possibly non-smooth function ρ, i.e. $\mathbb{E}[\hat{\partial}\rho(\lambda)] \in \partial\rho(\lambda)$. In (IV.9), M_1 and M_2 are the "shift" parameters specified in Lemma 5.3 by the convergence analysis in Section V. Note that since the function ρ is often known in practice, a full subgradient $u \in \partial\rho(\lambda)$ may be used instead of an instantaneous approximate $\hat{\partial}\rho(\lambda)$. With appropriately defined shift parameters M_1, M_2 in the subgradient estimator, if $\zeta > 0$, then the dual subgradient is biased with a constant shift:

$$\mathbb{E}[\hat{\partial}_\lambda L(v,\lambda)] \in \partial_\lambda L(v,\lambda) - (M_1+M_2)\cdot\mathbf{1}.$$

With these estimates for the primal gradient and dual subgradient of the Lagrangian (IV.4), we propose executing primal-dual stochastic subgradient iteration [30], [31] with the KL divergence in the dual domain. The detailed steps are summarized in Algorithm 1. Employing KL divergence in defining the dual update permits us to leverage the structure of λ as a distribution to derive tighter convergence rates, as detailed in Section V.

Algorithm 1 provides a model-free method for learning cautious-optimal policies from transition samples. Each primal and dual update can be computed easily based on a single observation. Although Algorithm 1 is given in the tabular form, its spirit of primal-dual stochastic approximation can be generalized to work with function approximations in the primal and dual spaces as the subject of future work.

V. Convergence Analysis

In this section, we provide sample complexity results for finding near-optimal solutions whose dependence on the size of the state and action spaces is tight. Before delving into these details, we state a technical condition on the caution function ρ required for the subsequent analysis, which is that we have access to a first-order oracle providing noisy samples of its subgradient, and that the infinity norm of these samples is bounded.

Assumption 5.1: The caution function $\rho(\lambda)$ is convex but possibly non-smooth, and it has bounded subgradients as

$$\sup_{\lambda\in\mathcal{L}}\sup_{u\in\partial\rho(\lambda)}\|u\|_\infty \le \sigma < \infty. \tag{V.1}$$

Further, samples $\hat{\partial}\rho(\lambda)$ of its subgradients are unbiased and have finite infinity norm:

$$\mathbb{E}[\hat{\partial}\rho(\lambda)] \in \partial\rho(\lambda)\ , \quad \sup_{\lambda\in\mathcal{L}}\|\hat{\partial}\rho(\lambda)\|_\infty \le \sigma. \tag{V.2}$$

In our subsequent analysis, we treat σ as a known constant. In all of Examples 3.1-3.3, the caution function ρ is explicitly known, which yields $\hat{\partial}\rho(\lambda) \in \partial\rho(\lambda)$. For instance, in Example 3.3, $\rho(\lambda) = KL(\hat{\lambda}\,||\,\mu)$ for some fixed μ [cf. (III.5)], the gradient takes the form

$$|\nabla_{\lambda_{sa}}\rho(\lambda)| = \left|(1-\gamma)\left(1+\log\left(\hat{\lambda}_{sa}/\mu_{sa}\right)\right)\right|$$

for any $s\in\mathcal{S}$ and $a\in\mathcal{A}$. Then, we can ensure Assumption 5.1 by imposing an elementwise lower bound δ_0 on μ and λ s.t. $\mu \ge \delta_0\cdot\mathbf{1}$ and $\lambda \ge \delta_0\cdot\mathbf{1}$. The constant δ_0 may be chosen extremely small, for instance, $\delta_0 = \min\{10^{-15}; |\mathcal{S}|^{-1}|\mathcal{A}|^{-1}\}$. Consequently, we have

$$\sigma \le \mathcal{O}\left((1-\gamma)\left(1+\log\left(\delta_0^{-1}\right)\right)\right) = \mathcal{O}(1).$$

Next, we begin the analysis by noting that the saddle point problem (IV.1) does not specify the feasible region $\mathcal{V}$ for the variable v. However, the convergence necessitates $\mathcal{V}$ to be a compact set rather than the entire $\mathbb{R}^{|\mathcal{S}|}$. To disambiguate the domain of v, next we derive a bounded region that contains the primal optimizer v^*.

Lemma 5.2: If $\xi > 0$, then the primal optimizer v^* satisfies

$$\|v^*\|_\infty \le (1-\gamma)^{-1}(1+c\sigma). \tag{V.3}$$

Therefore, we can define the feasible region $\mathcal{V}$ to be the compact set

$$\mathcal{V} := \left\{v\in\mathbb{R}^{|\mathcal{S}|} : \|v\|_\infty \le 2\frac{1+c\sigma}{1-\gamma}\right\}. \tag{V.4}$$

The proof of Lemma 5.2 is provided in Appendix I of the supplementary material [32]. We note that the factor of 2 is incorporated to simplify the analysis.

Subsequently, we analyze the primal-dual convergence of Algorithm 1 for solving (IV.1) (and the equivalently (III.1)). Before providing the main theorem, we introduce a technical result which defines convergence in terms of a form of duality gap. The duality gap measures the distance of the Lagrangian evaluations to a saddle point as defined by (IV.1).

*Lemma 5.3 (***Convergence of duality gap):** For Algorithm 1, select shift parameters $M_1 = \frac{4(1+c\sigma)}{1-\gamma}$ and $M_2 = c\sigma$, $\delta \in (0,\frac{1}{2})$, $\beta = \frac{1-\gamma}{1+c\sigma}\sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}|)}{T|\mathcal{S}||\mathcal{A}|}}$, and $\alpha = \sqrt{\frac{|\mathcal{S}|}{T}}(1+c\sigma)$. Let $\bar{\lambda}$ and $\bar{v}$ be the output of Algorithm 1 and let λ^* be the optimum. Then for the output of Algorithm 1, we have

$$\begin{aligned}\mathbb{E}[L(\bar{v},\lambda^*) - \min_{v\in\mathcal{V}} L(v,\bar{\lambda})] \\ \le \mathcal{O}\left(\sqrt{\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}}\cdot\frac{1+2c\sigma}{(1-\gamma)^2}\right).\end{aligned} \tag{V.5}$$

As a result, to guarantee $\mathbb{E}[L(\bar{v},\lambda^*) - \min_{v\in\mathcal{V}} L(v,\bar{\lambda})] \le \epsilon$, the amount of samples needed is

$$T = \Theta\left(\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)(1+2c\sigma)^2}{(1-\gamma)^4\epsilon^2}\right). \tag{V.6}$$

The proof of this Lemma is provided in Appendix II of the supplementary material [32].

We may then use the convergence of duality gap to characterize the sub-optimality and constraint violation attained by the output of Algorithm 1 for the problem (III.1). Hence, the main result is summarized in Theorem 5.4 next.

Theorem 5.4: (**Convergence to optimal caution-sensitive policies**): Let the parameters M_1, M_2, δ, β, and α, as defined in Theorem 5.3, if $\bar{\lambda}$ is the output of Algorithm 1 after T iterations, then the constraint violation of the original problem (III.1) satisfies

$$\begin{cases}\bar{\lambda}\ge 0, \quad \left\|\bar{\lambda}\right\|_1 = (1-\gamma)^{-1} \\ \left\|\sum_{a\in\mathcal{A}}(I-\gamma P_a^\top)\bar{\lambda}_a - \xi\right\|_1 \le \frac{(1-\gamma)\epsilon}{1+c\sigma} \le (1-\gamma)\epsilon.\end{cases} \tag{V.7}$$

Moreover, the sub-optimality of (III.1) is given as

$$\mathbb{E}[(\langle\lambda^*, r\rangle - c\rho(\lambda^*)) - (\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda}))] \leq \epsilon \tag{V.8}$$

Eqs. (V.7) and (V.8) showed the output solution is ϵ-feasible and ϵ-optimal. Note that ϵ determines the number of samples T as given in (V.6).

Proof: The first row of (V.7) is directly satisfied due to the feasibility of $\bar{\lambda} \in \mathcal{L}$. Now we prove the second row of (V.7). When the parameters are chosen according to Lemma 5.3, we know

$$\epsilon \geq \mathbb{E}[L(\bar{v}, \lambda^*) - \min_{v\in\mathcal{V}} L(v, \bar{\lambda})]. \tag{V.9}$$

For the ease of notation, denote $C := (1-\gamma)^{-1}(1+c\sigma)$. Then substitute the details of L we get

$$\begin{aligned}\min_{v\in\mathcal{V}} L(v,\bar{\lambda}) =& \min_{\|v\|_\infty \leq 2C} \langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda}) + \langle\xi, v\rangle \\ &+ \sum_{a\in\mathcal{A}} \bar{\lambda}_a(\gamma P_a - I)v \qquad \text{(V.10)} \\ =& \langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda}) - 2C\Big\| \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\bar{\lambda}_a - \xi \Big\|_1.\end{aligned}$$

By the feasibility of λ^*, namely, $\sum_{a\in\mathcal{A}}(I-\gamma P_a^\top)\lambda_a^* - \xi = 0$, we have

$$\begin{aligned}L(\bar{v}, \lambda^*) =& \langle\lambda^*, r\rangle - c\rho(\lambda^*) + \langle\xi, \bar{v}\rangle \qquad \text{(V.11)} \\ &+ \sum_{a\in\mathcal{A}} (\lambda_a^*)^\top(\gamma P_a - I)\bar{v} = \langle\lambda^*, r\rangle - c\rho(\lambda^*).\end{aligned}$$

Substituting (V.10) and (V.11) into (V.9) yields

$$\begin{aligned}\mathbb{E}\Bigg[&(\langle\lambda^*, r\rangle - c\rho(\lambda^*)) - \big(\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda})\big) \\ &+ 2C\Big\| \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\bar{\lambda}_a - \xi \Big\|_1 \Bigg] \leq \epsilon. \qquad \text{(V.12)}\end{aligned}$$

Actually, this inequality has already proved the bound (V.8) in terms of the objective value of problem (III.1). Also, by the feasibility of λ^*, the convexity of ρ, and the optimality condition (I.4), we have

$$\begin{aligned}&(\langle\lambda^*, r\rangle - c\rho(\lambda^*)) - \big(\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda})\big) \\ &\quad + \big\langle v^*, \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\bar{\lambda}_a - \xi \big\rangle \\ &= (\langle\lambda^*, r\rangle - c\rho(\lambda^*)) - \big(\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda})\big) \\ &\quad + \big\langle v^*, \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\bar{\lambda}_a - \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\lambda_a^* \big\rangle \\ &\geq \sum_{a\in\mathcal{A}} \big\langle (I - \gamma P_a)v^* - r_a + cu_a^*, \bar{\lambda}_a - \lambda_a^* \big\rangle \\ &\geq 0, \qquad \text{(V.13)}\end{aligned}$$

where $u^* \in \partial\rho(\lambda^*)$ is defined in (I.4), and $u_a^* := [u_{1a}^*, ..., u_{|\mathcal{S}|a}^*]^\top$ is column vector. Immediately, this implies

$$\begin{aligned}&(\langle\lambda^*, r\rangle - c\rho(\lambda^*)) - \big(\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda})\big) \qquad \text{(V.14)} \\ &\qquad \geq -\big\langle v^*, \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\bar{\lambda}_a - \xi \big\rangle \\ &\qquad \geq -\|v^*\|_\infty \Big\| \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\bar{\lambda}_a - \xi \Big\|_1 \\ &\qquad \geq -C\Big\| \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\bar{\lambda}_a - \xi \Big\|_1.\end{aligned}$$

where we used the fact that $\|v^*\|_\infty \leq C$ proved in Lemma 5.2. Substitute (V.14) into (V.12) gives

$$\mathbb{E}\big[C\Big\| \sum_{a\in\mathcal{A}} (I - \gamma P_a^\top)\bar{\lambda}_a - \xi \Big\|_1\big] \leq \epsilon.$$

Divide both sides by $C = (1-\gamma)^{-1}(1+c\sigma)$ proves inequality (V.7). ■

Theorem 5.4 suggests that to get ϵ-optimal policy and its corresponding state-action distribution, the sample complexity has near-linear dependence (up to logarithmic factors) on the sizes of $\mathcal{S}$ and $\mathcal{A}$. This matches the optimal dependence in the risk-neutral case, see e.g. [31], [33], [34] which proves that Algorithm 1 is sample-efficient.

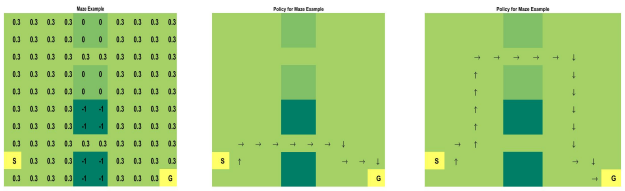
VI. Experimental Results

In this section, we experimentally evaluate the proposed technique for incorporating risk or other sources of exogenous information into RL training. In particular, we consider a setting in which an agent originally learns in the risk-neutral sense of (II.2), i.e., focusing on expected returns. The MDP we focus on is a 10×10 grid with each state permitting for four possible actions (moving $\mathcal{A} := \{\texttt{up}, \texttt{down}, \texttt{left}, \text{ and } \texttt{right}\}$). For the transition model, given the direction of the previous action selection, the agent movies in the same direction with probability p and moves in the different direction with probability $1-p$, and moves backwards with null probability. For instance, in a given state action pair (s, a), suppose the action a selected is `up`. Then, the next action will be `up` with prob p and $\{\texttt{left}, \text{ or } \texttt{right}\}$ with prob $1-p$, and `down` with null probability. Overall, this means that the transition matrix has four nonzero sequences of likelihoods along the main diagonal, i.e., it is quad-diagonal. For the experiments, we consider the caution-sensitive formulation presented in Examples 3.2 and 3.3 which respectively correspond to quantifying risk via the variance and the KL divergence to a previously learned policy which serves as a prior. We append videos (links in the footnote[1][2]) to the submission which visualize the safety of risk-awareness during training.

A. Variance-Sensitive Policy Optimization

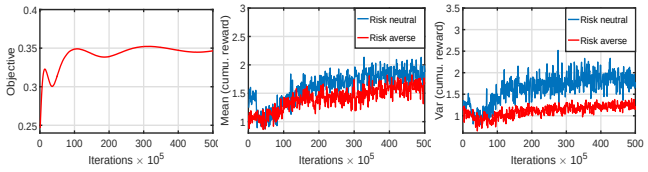
The variance risk given in Example 3.2 characterizes the statistical robustness of the rewards from a policy. To evaluate the merit of this definition, consider the maze example with

[1] `https://tinyurl.com/sk4lddb`
[2] `https://tinyurl.com/tlcl3m2`



(a) Reward dist. (b) Risk neutral (c) Risk averse

Fig. 1. Experiment on grid world with variance as the risk. (a) Reward distribution for the Maze environment; (b) Risk neutral and (c) Risk averse trajectories, respectively, from start to goal. The trajectory resulting from greedily following the risk-averse policy avoids negative reward states.



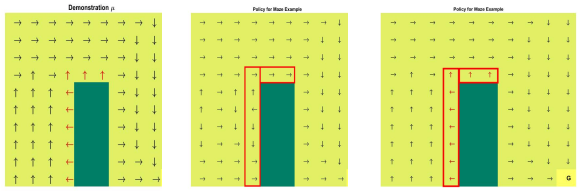
(a) Objective (b) Reward mean (c) Reward Variance

Fig. 2. (a) Convergence of the dual objective [cf. (III.1)]; Sample mean return (b) and variance (c) over 100 simulated trajectories. Observe the expected reward return is comparable while the risk-averse policy attains lower variance, and is thus more reliable.

the rewards distribution as described in Fig. 1(a). There are two ways to go from start to destination. The reward of dark green areas is more negative than lighter shades of green, and thus it is riskier to be near darker green in terms of the returns of a trajectory. We display a sample path of the Markov chain obtained by solving the variance-sensitive policy optimization problem as Fig. 1(c), whereas the one based on the risk-neutral (classical) formulation is shown in Fig. 1(b). Clearly, the risk-averse one avoids the dark green areas and collects a sequence of more robust rewards, yet still reaches the goal. The convergence of objective is plotted in Fig. 2(a) for the proposed algorithm. Further, we plot the associated sample mean and variance of the discounted return over number of training indices in Figs. 2(b) and 2(c), respectively. Observe that the risk-averse policy yields comparable mean reward accumulation with reduced variance, meaning it more reliably reaches the goal without visiting unwanted states whose rewards are negative.

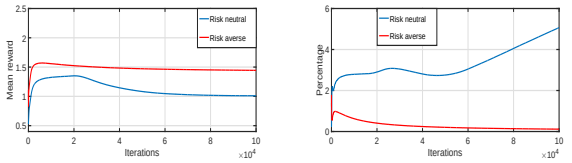
B. Caution as Proximity to a Prior

When a prior is available in the form of some baseline state-action distribution μ , KL divergence to the baseline makes sense as a measure of caution [cf. (III.5)] as stated in Example 3.3. To evaluate this definition, consider the setting where the baseline μ is a risk-neutral policy (shown in Fig. 3(a)) learned by solving (II.4) with a reward that is highly negative $r = -5$ in the dark green area, strictly positive $r = 0.3$ in the light green area, and $r = 1$ at the goal in the bottom right denoted by G in Fig. 3(a). The transition probabilities are defined by $p = 0.4$. Then, the resulting risk-neutral policy is used as a baseline policy for a drifted MDP whose reward is $r = 0$ for the dark green area



(a) μ (b) Risk neutral (c) Risk averse

Fig. 3. Results for the learning with demonstration μ. We have used KL divergence as the risk function for these results. (a) The given demonstration, (b) Risk neutral solution, (c) Risk averse solution. Note that incorporating KL divergence yields a policy that avoids unrewarding states (red block in (b) and (c)).



(a) Comparison (b) Time in unrewarding states

Fig. 4. We plot the running average of (a) Expected reward return, (b) percentage of time we visit the unrewarding states. Note that the prior demonstration helps in the faster convergence as clear from (a). Further, the KL divergence based risk helps to avoid the visitation of the unrewarding states as clear from the result in (b).

while identical elsewhere, and whose transition dynamics are defined by likelihood parameter $p = 0.6$. The overarching purpose is that although the reward landscape and transition dynamics changed, the "lessons" of past learning may still be incorporated.

The resulting policy learned from this procedure, as compared with the risk-neutral policy, are visualized in Figures 3(b) and 3(c), respectively. Observe that the policy associated with incorporating past experience in the form of policy μ has explicitly pushed avoidance of the dark green region, whereas the risk-neutral policy resulting from (II.4) does not. Thus, past (negative) experiences may be incorporated into the learned policy. This hearkens back to psychological experiments on mice: if its food supply is electrified, then a mouse will refuse to eat, even after the electricity is shut off, a form of fear conditioning. Further, we plot the associated discounted return and empirical occupancy of negative reward states with the iteration index of the optimization procedure in Algorithm 1 in Fig. 4. Overall, then, the incorporation of prior demonstrations results in the faster learning (see Fig. 4(a)) and reduces the proportion of time spent in unrewarding states as evidenced by Fig. 4(b).

VII. Conclusions

In this work, we proposed a new definition of risk named caution which takes as input unnormalized state-action occupancy distributions, motivated by the dual of the LP formulation of the MDP. To solve the resulting risk-aware RL in an online model-free manner, we proposed a variant of stochastic primal-dual method to solve it, whose

sample complexity matches optimal dependencies of risk-neutral problem. Experiments illuminated the usefulness of this definition in practice. Future work includes deriving the Bellman equations associated with cautious policy optimization (III.1), generalizations to continuous spaces, and broadening caution to encapsulate other aspects of decision-making such as inattention and anticipation.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, 2018.

[2] A. Karatzoglou, L. Baltrunas, and Y. Shi, "Learning to rank for recommender systems," in *Proceedings of the 7th ACM conference on Recommender systems*, 2013, pp. 493–494.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[4] O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. M. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell *et al.*, "Alphastar: Mastering the real-time strategy game starcraft ii," *DeepMind blog*, p. 2, 2019.

[5] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[6] D. Peidro, J. Mula, R. Poler, and F.-C. Lario, "Quantitative models for supply chain planning under uncertainty: a review," *The International Journal of Advanced Manufacturing Technology*, vol. 43, no. 3-4, pp. 400–420, 2009.

[7] A. Roca, P. R. Ford, A. P. McRobert, and A. M. Williams, "Identifying the processes underpinning anticipation and decision-making in a dynamic time-constrained task," *Cognitive processing*, vol. 12, no. 3, pp. 301–310, 2011.

[8] C. A. Sims, "Implications of rational inattention," *Journal of monetary Economics*, vol. 50, no. 3, pp. 665–690, 2003.

[9] S. M. Tom, C. R. Fox, C. Trepel, and R. A. Poldrack, "The neural basis of loss aversion in decision-making under risk," *Science*, vol. 315, no. 5811, pp. 515–518, 2007.

[10] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 22–31.

[11] H. Markowitz, "Portfolio selection," *The journal of finance*, vol. 7, no. 1, pp. 77–91, 1952.

[12] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of banking & finance*, vol. 26, no. 7, pp. 1443–1471, 2002.

[13] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical finance*, vol. 9, no. 3, pp. 203–228, 1999.

[14] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6070–6120, 2017.

[15] D. R. Jiang and W. B. Powell, "Risk-averse approximate dynamic programming with quantile-based risk measures," *Mathematics of Operations Research*, vol. 43, no. 2, pp. 554–579, 2018.

[16] T. Bjork and A. Murgoci, "A general theory of markovian time inconsistent stochastic control problems," *Available at SSRN 1694759*, 2010.

[17] A. Ruszczyński, "Risk-averse dynamic programming for markov decision processes," *Mathematical programming*, vol. 125, no. 2, pp. 235–261, 2010.

[18] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy gradient for coherent risk measures," in *Advances in Neural Information Processing Systems*, 2015, pp. 1468–1476.

[19] V. Krishnamurthy, K. Martin, and F. V. Abad, "Implementation of gradient estimation to a constrained markov decision problem," in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, vol. 5. IEEE, 2003, pp. 4841–4846.

[20] L. Prashanth, "Policy gradients for cvar-constrained mdps," in *International Conference on Algorithmic Learning Theory*. Springer, 2014, pp. 155–169.

[21] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, "Learning safe policies via primal-dual methods," in *Proceedings of the 58th IEEE Conference on Decision and Control, IEEE*, 2019.

[22] M. Yu, Z. Yang, M. Kolar, and Z. Wang, "Convergent policy optimization for safe reinforcement learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 3121–3133.

[23] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 969–996, 2007.

[24] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar *et al.*, "Bayesian reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 8, no. 5-6, pp. 359–483, 2015.

[25] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 449–458.

[26] R. Keramati, C. Dann, A. Tamkin, and E. Brunskill, "Being optimistic to be conservative: Quickly learning a cvar policy," *arXiv preprint arXiv:1911.01546*, 2019.

[27] D. Yang, L. Zhao, Z. Lin, T. Qin, J. Bian, and T.-Y. Liu, "Fully parameterized quantile function for distributional reinforcement learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 6190–6199.

[28] D. P. Bertsekas and S. Shreve, *Stochastic optimal control: the discrete-time case*, 2004.

[29] D. P. De Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Operations research*, vol. 51, no. 6, pp. 850–865, 2003.

[30] Y. Chen and M. Wang, "Stochastic primal-dual methods and sample complexity of reinforcement learning," *arXiv preprint arXiv:1612.02516*, 2016.

[31] Y. Chen, L. Li, and M. Wang, "Scalable bilinear pi learning using state and action features," *arXiv preprint arXiv:1804.10328*, 2018.

[32] J. Zhang, A. S. Bedi, M. Wang, and A. Koppel, "beyond cumulative returns via reinforcement learning over state-action occupancy measures," Sep 2020. [Online]. Available: https://tinyurl.com/y2rxft8n

[33] M. Wang, "Primal-dual pi learning: Sample complexity and sublinear run time for ergodic markov decision problems," *arXiv preprint arXiv:1710.06100*, 2017.

[34] ——, "Randomized linear programming solves the discounted markov decision problem in nearly-linear (sometimes sublinear) running time," *arXiv preprint arXiv:1704.01869*, 2017.

[35] F. Bach and K. Y. Levy, "A universal algorithm for variational inequalities adaptive to smoothness and noise," *arXiv preprint arXiv:1902.01637*, 2019.

SUPPLEMENTARY MATERIAL FOR "BEYOND CUMULATIVE RETURNS VIA REINFORCEMENT LEARNING OVER STATE-ACTION OCCUPANCY MEASURES"

APPENDIX I
PROOF OF LEMMA 5.2

Proof: Consider the min-max saddle point problem,

$$\max_{\lambda\geq 0}\min_{v\in\mathbf{R}^{|\mathcal{S}|}} L(v,\lambda)=\langle\lambda,r\rangle-c\rho(\lambda)+\langle\xi,v\rangle+\sum_{a\in\mathcal{A}}\lambda_a^\top(\gamma P_a-I)v, \tag{I.1}$$

Then (λ^*,v^*) solves this saddle point problem if and only if

$$\lambda^*=\operatorname*{argmax}_{\lambda\geq 0} L(v^*,\lambda) \qquad \text{and} \qquad \sum_{a\in\mathcal{A}}(I-\gamma P_a^\top)\lambda_a^*-\xi=0. \tag{I.2}$$

A remark is that, this is also the KKT condition for the original convex problem (III.1). Due to the concavity of $L(v^*,\lambda)$ for any fixed v^*, the condition $\lambda^*=\operatorname*{argmax}_{\lambda\geq 0} L(v^*,\lambda)$ is equivalent to the existence of a subgradient $w^*\in\partial_\lambda L(v^*,\lambda^*)$ s.t.

$$\langle w^*,\lambda-\lambda^*\rangle\leq 0 \quad \text{for} \quad \forall\lambda\geq 0. \tag{I.3}$$

If we use u^* to denote the specific subgradient in $\partial\rho(\lambda^*)$ that consists w^*. For any fixed s,a, we know $w^*_{sa}=-(e_s-\gamma P_{as})^\top v^*+r_{sa}-cu^*_{sa}$. If we choose $\lambda_{s'a'}=\lambda^*_{s'a'}$ for $\forall(s',a')\neq(s,a)$, (I.3) further implies

$$\big((e_s-\gamma P_{as})^\top v^*-r_{sa}+cu^*_{sa}\big)(\lambda_{sa}-\lambda^*_{sa})\geq 0,$$

where P_{as} is a column vector, with $P_{as}(s')=\mathcal{P}(s'|a,s)$. Combine this inequality with (I.2), we can formally write the final optimality condition as follows.

$$\exists u^*\in\partial\rho(\lambda^*)\ \text{ s.t. }\ \begin{cases}\sum_{a\in\mathcal{A}}(I-\gamma P_a^\top)\lambda_a^*=\xi, \quad \lambda^*\geq 0,\\ \big((e_s-\gamma P_{as})^\top v^*-r_{sa}+cu^*_{sa}\big)(\lambda_{sa}-\lambda^*_{sa})\geq 0, \quad \forall s\in\mathcal{S},\forall a\in\mathcal{A},\forall\lambda_{sa}\geq 0.\end{cases} \tag{I.4}$$

By (II.7) of Proposition 2.1, we know that

$$\sum_{a\in\mathcal{A}}\lambda^*_{sa}\geq\sum_{a\in\mathcal{A}}\mathbf{Prob}\big(i_0=s,a_0=a|i_0\sim\xi,a_0\sim\pi(\cdot|i_0)\big)=\xi_s>0 \quad \text{for} \quad \forall s\in\mathcal{S}.$$

Therefore, for any $s\in\mathcal{S}$, there exists an a_s such that $\lambda^*_{sa_s}>0$. Therefore, the second inequality of the optimality condition (I.4) implies that,

$$(e_s-\gamma P_{a_s s})^\top v^*-r_{sa_s}+cu^*_{sa_s}=0 \quad \text{for} \quad \forall s\in\mathcal{S}.$$

Let us denote $\tilde{r}:=[r_{1a_1},\cdots,r_{|\mathcal{S}|a_{|\mathcal{S}|}}]^\top\in\mathbb{R}^{|\mathcal{S}|}$, $\tilde{u}:=[u^*_{1a_1},\cdots,u^*_{|\mathcal{S}|a_{|\mathcal{S}|}}]^\top\in\mathbb{R}^{|\mathcal{S}|}$ and $\tilde{P}:=[P_{a_1 1},\cdots,P_{a_{|\mathcal{S}|}|\mathcal{S}|}]\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$. Then we can write

$$\left(I-\gamma\tilde{P}^\top\right)v^*=\tilde{r}-c\tilde{u}.$$

As a result,

$$1+c\sigma\geq\|\tilde{r}-c\tilde{u}\|_\infty=\|(I-\gamma\tilde{P}^\top)v^*\|_\infty\geq\|v^*\|_\infty-\|\gamma\tilde{P}^\top v^*\|_\infty\geq(1-\gamma)\|v^*\|_\infty,$$

which implies the statement of Lemma 5.2. ∎

APPENDIX II
PROOF OF LEMMA 5.3

Proof: To make the proof of this result clearer, we will separate part of the major steps into several different lemmas.

Lemma 2.1: Suppose the iterate sequence $\{v^t\}$ is updated according to the rule (IV.5) in Algorithm 1. Then for any t,

$$\begin{aligned}\langle\nabla_v L(v^t,\lambda^t),v^t-v\rangle\leq&\frac{1}{2\alpha}(\|v^t-v\|^2-\|v^{t+1}-v\|^2)+\frac{\alpha}{2}\|\hat{\nabla}_v L(v^t,\lambda^t)\|^2\\&+\langle\nabla_v L(v^t,\lambda^t)-\hat{\nabla}_v L(v^t,\lambda^t),v^t-v\rangle.\end{aligned} \tag{II.1}$$

The proof of this lemma is provided in Appendix II-A.

Lemma 2.2: Suppose the iterate sequence $\{\lambda^t\}$ is updated according to the rule (IV.6) and (IV.7) in Algorithm 1. For $\forall t$,

$$\begin{aligned}-\langle w^t,\lambda^t-\lambda\rangle\leq&\frac{1}{(1-\gamma)\beta}\big(KL\big((1-\gamma)\lambda\,||\,(1-\gamma)\lambda^t\big)-KL\big((1-\gamma)\lambda\,||\,(1-\gamma)\lambda^{t+1}\big)\big)\\&+\frac{\beta}{2}\sum_{s,a}\lambda^t_{sa}(\Delta^t_{sa})^2+\langle\hat{\partial}_\lambda L(v^t,\lambda^t)-w^t,\lambda^t-\lambda\rangle,\end{aligned} \tag{II.2}$$

where $w^t := \mathbb{E}\left[\hat{\partial}_\lambda L(v^t,\lambda^t)\middle|\lambda^t, v^t\right] + (M_1+M_2)\cdot\mathbf{1} \in \partial_\lambda L(v^t,\lambda^t)$ is a subgradient vector.
The proof of this lemma is provided in Appendix II-B. Based on these two lemmas, we start the proof of Theorem 5.3. Note that by definition, $\bar{v} = \frac{1}{T}\sum_{t=1}^T v^t$ and $\bar{\lambda} = \frac{1}{T}\sum_{t=1}^T \lambda^t$. Define $\bar{v}^* := \operatorname{argmin}_{v\in\mathcal{V}} L(v,\bar{\lambda})$. Then by the convex-concave structure of L we have

$$\begin{aligned} L(\bar{v},\lambda^*) - L(\bar{v}^*,\bar{\lambda}) &\leq \frac{1}{T}\sum_{t=1}^T \left(L(v^t,\lambda^*) - L(\bar{v}^*,\lambda^t)\right) \\ &= \frac{1}{T}\sum_{t=1}^T \left(L(v^t,\lambda^*) - L(v^t,\lambda^t) + L(v^t,\lambda^t) - L(\bar{v}^*,\lambda^t)\right) \\ &\leq \frac{1}{T}\sum_{t=1}^T \left(-\langle w^t, \lambda^t - \lambda^*\rangle + \langle \nabla_v L(v^t,\lambda^t), v^t - \bar{v}^*\rangle\right), \end{aligned} \tag{II.3}$$

where the first line applies Jensen's inequality and last line is due to the convexity of $L(\cdot,\lambda^t)$ and the concavity of $L(v^t,\cdot)$. Note that by specifying $v = \bar{v}^*$ in (II.1) and $\lambda = \lambda^*$ in (II.2), we can sum up the inequlities (II.1) and (II.2) for $t = 1, ..., T$ to yield

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T \left(-\langle w^t, \lambda^t - \lambda^*\rangle + \langle \nabla_v L(v^t,\lambda^t), v^t - \bar{v}^*\rangle\right) \\ \leq &\underbrace{\frac{KL\big((1-\gamma)\lambda^*||(1-\gamma)\lambda^1\big)}{T(1-\gamma)\beta}}_{T_1} + \underbrace{\frac{\beta}{2T}\sum_{t=1}^T\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2}_{T_2} + \underbrace{\frac{1}{T}\sum_{t=1}^T\langle\hat{\partial}_\lambda L(v^t,\lambda^t) - w^t, \lambda^t - \lambda^*\rangle}_{T_3} \\ &+ \underbrace{\frac{\|v^1 - \bar{v}^*\|^2}{2T\alpha}}_{T_4} + \underbrace{\frac{\alpha}{2T}\sum_{t=1}^T \|\hat{\nabla}_v L(v^t,\lambda^t)\|^2}_{T_5} + \underbrace{\frac{1}{T}\sum_{t=1}^T\langle\nabla_v L(v^t,\lambda^t) - \hat{\nabla}_v L(v^t,\lambda^t), v^t - \bar{v}^*\rangle}_{T_6}. \end{aligned}$$

Substitute this inequality into (II.3) and take the expectation on both sides, we get

$$\mathbb{E}[L(\bar{v},\lambda^*) - \min_{v\in\mathcal{V}} L(v,\bar{\lambda})] \leq \sum_{i=1}^6 \mathbb{E}[T_i]. \tag{II.4}$$

For the $\mathbb{E}[T_i]$'s, the following bounds hold with detailed derivation provided in Appendix II-C:

$$\mathbb{E}[T_1] \leq \frac{\log(|\mathcal{S}||\mathcal{A}|)}{T(1-\gamma)\beta}, \qquad \mathbb{E}[T_2] \leq \frac{4\beta c^2\sigma^2}{1-\gamma} + \frac{128\beta|\mathcal{S}||\mathcal{A}|(1+c\sigma)^2}{(1-\gamma)^3}, \qquad \mathbb{E}[T_3] = 0,$$

$$\mathbb{E}[T_4] \leq \frac{8|\mathcal{S}|(1+c\sigma)^2}{T\alpha(1-\gamma)^2}, \qquad \mathbb{E}[T_5] \leq \frac{27\alpha}{2(1-\gamma)^2}, \qquad \mathbb{E}[T_6] \leq \frac{3\sqrt{3|\mathcal{S}|}(1+c\sigma)}{\sqrt{T}(1-\gamma)^2}.$$

Substitute these bounds for $\mathbb{E}[T_i]$'s into inequality (II.4) we get

$$\begin{aligned} \mathbb{E}[L(\bar{v},\lambda^*) - \min_{v\in\mathcal{V}} L(v,\bar{\lambda})] \leq &\frac{\log(|\mathcal{S}||\mathcal{A}|)}{T(1-\gamma)\beta} + \frac{4\beta c^2\sigma^2}{1-\gamma} + \frac{128\beta|\mathcal{S}||\mathcal{A}|(1+c\sigma)^2}{(1-\gamma)^3} \\ &+ \frac{8|\mathcal{S}|(1+c\sigma)^2}{T\alpha(1-\gamma)^2} + \frac{27\alpha}{2(1-\gamma)^2} + \frac{3\sqrt{3|\mathcal{S}|}(1+c\sigma)}{\sqrt{T}(1-\gamma)^2}. \end{aligned} \tag{II.5}$$

If we choose $\beta = \frac{1-\gamma}{1+c\sigma}\sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}|)}{T|\mathcal{S}||\mathcal{A}|}}$ and $\alpha = \sqrt{\frac{|\mathcal{S}|}{T}}(1+c\sigma)$, we have

$$\mathbb{E}[L(\bar{v},\lambda^*) - \min_{v\in\mathcal{V}} L(v,\bar{\lambda})] \leq \mathcal{O}\left(\sqrt{\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}}\cdot\frac{1+c\sigma}{(1-\gamma)^2}\right),$$

which completes the proof. ■

A. Proof of Lemma 2.1

Proof: Consider the update rule of v provided in (IV.5). For any $v \in \mathcal{V}$, it holds that

$$\begin{aligned}\|v^{t+1} - v\|^2 &= \|\Pi_{\mathcal{V}}(v^t - \alpha\hat{\nabla}_v L(v^t, \lambda^t)) - v\|^2 \\ &\leq \|v^t - \alpha\hat{\nabla}_v L(v^t, \lambda^t) - v\|^2 \\ &= \|v^t - v\|^2 + \alpha^2\|\hat{\nabla}_v L(v^t, \lambda^t)\|^2 - 2\alpha\langle\hat{\nabla}_v L(v^t, \lambda^t), v^t - v\rangle \\ &= \|v^t - v\|^2 + \alpha^2\|\hat{\nabla}_v L(v^t, \lambda^t)\|^2 - 2\alpha\langle\hat{\nabla}_v L(v^t, \lambda^t) - \nabla_v L(v^t, \lambda^t) + \nabla_v L(v^t, \lambda^t), v^t - v\rangle.\end{aligned}$$

Rearranging the above inequality yields

$$2\alpha\langle\nabla_v L(v^t, \lambda^t), v^t - v\rangle \leq \|v^t - v\|^2 - \|v^{t+1} - v\|^2 + \alpha^2\|\hat{\nabla}_v L(v^t, \lambda^t)\|^2 - 2\alpha\langle\hat{\nabla}_v L(v^t, \lambda^t) - \nabla_v L(v^t, \lambda^t), v^t - v\rangle.$$

Deviding both sides by 2α proves lemma. ■

B. Proof of Lemma 2.2

Proof: Now let us consider the update rule of λ given by (IV.6) and (IV.7). Note that in the subproblem (IV.6), the problem is separable for each component of λ and allows for a closed form solution, i.e.,

$$\begin{aligned}\lambda_{sa}^{t+\frac{1}{2}} =& \underset{\lambda_{sa}}{\operatorname{argmax}}\ \Delta_{sa}^t\lambda_{sa} - \frac{1}{(1-\gamma)\beta}(1-\gamma)\lambda_{sa}\log\left(\frac{(1-\gamma)\lambda_{sa}}{(1-\gamma)\lambda_{sa}^t}\right) \\ =& \lambda_{sa}^t \cdot \exp\{\beta\Delta_{sa}^t\},\end{aligned} \tag{II.6}$$

where we denote Δ_{sa}^t to be the (s,a)-th component of $\hat{\partial}_\lambda L(v^t, \lambda^t)$. Then the next iterate is constructed as

$$\lambda^{t+1} = \frac{\lambda^{t+\frac{1}{2}}}{(1-\gamma)\|\lambda^{t+\frac{1}{2}}\|_1}.$$

Or in a more elementary way, we define

$$\lambda_{sa}^{t+1} = \frac{\lambda_{sa}^t \cdot \exp\{\beta\Delta_{sa}^t\}}{(1-\gamma)\sum_{s',a'}\lambda_{s'a'}^t \cdot \exp\{\beta\Delta_{s'a'}^t\}}. \tag{II.7}$$

It is straightforward that $\lambda^{t+1} \in \mathcal{L}$. As a result, for any $\lambda \in \mathcal{L}$,

$$KL\big((1-\gamma)\lambda \,||\, (1-\gamma)\lambda^{t+1}\big) - KL\big((1-\gamma)\lambda \,||\, (1-\gamma)\lambda^t\big) \tag{II.8}$$

$$=(1-\gamma)\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(\lambda_{sa}\log\left(\frac{\lambda_{sa}}{\lambda_{sa}^{t+1}}\right) - \lambda_{sa}\log\left(\frac{\lambda_{sa}}{\lambda_{sa}^t}\right)\right)$$

$$=(1-\gamma)\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\lambda_{sa}\log\left(\frac{\lambda_{sa}^t}{\lambda_{sa}^{t+1}}\right)$$

$$=(1-\gamma)\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\lambda_{sa}\left(\log\left((1-\gamma)\sum_{s',a'}\lambda_{s'a'}^t \cdot \exp\{\beta\Delta_{s'a'}^t\}\right) - \beta\Delta_{sa}^t\right) \tag{II.9}$$

$$=\log\left((1-\gamma)\sum_{s',a'}\lambda_{s'a'}^t \cdot \exp\{\beta\Delta_{s'a'}^t\}\right) - (1-\gamma)\beta\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\lambda_{sa}\Delta_{sa}^t$$

$$=\log\left((1-\gamma)\sum_{s,a}\lambda_{sa}^t \cdot \exp\{\beta\Delta_{sa}^t\}\right) - (1-\gamma)\beta\langle\hat{\partial}_\lambda L(v^t, \lambda^t), \lambda\rangle. \tag{II.10}$$

The equality in (II.9) is obtained by using the elementary definition of λ_{sa}^{t+1} in (II.7); The last equality of (II.10) is obtained by applying the definition of Δ_{sa}^t. Note that

$$\Delta_{sa}^t = \begin{cases}\frac{\hat{r}_{s_t s_t' a_t} + \gamma v_{s_t'} - v_{s_t} - M_1}{\zeta_{s_t a_t}^t} - c\left(\hat{\partial}\rho(\lambda^t)\right)_{s_t a_t} - M_2, & \text{if } (s,a) = (s_t, a_t), \\ -c\left(\hat{\partial}\rho(\lambda^t)\right)_{s_t a_t} - M_2, & \text{if } (s,a) \neq (s_t, a_t).\end{cases}$$

When we choose $M_1 = 4(1-\gamma)^{-1}(1+c\sigma)$ and $M_2 = c\sigma$, we can guarantee that $\Delta_{sa}^t \leq 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. Therefore, by the fact that $e^x \leq 1 + x + \frac{x^2}{2}$ for all $x \leq 0$ and $\log(1+x) \leq x$ for all $x > -1$, we have

$$\begin{aligned}\log\left((1-\gamma)\sum_{s,a}\lambda_{sa}^t \cdot \exp\{\beta\Delta_{sa}^t\}\right) \leq & \log\left((1-\gamma)\sum_{s,a}\lambda_{sa}^t \cdot \big(1+\beta\Delta_{sa}^t + \frac{\beta^2}{2}(\Delta_{sa}^t)^2\big)\right) \\ = & \log\left(1+(1-\gamma)\beta\langle\hat{\partial}_\lambda L(v^t,\lambda^t),\lambda^t\rangle + \frac{(1-\gamma)\beta^2}{2}\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2\right) \\ \leq & (1-\gamma)\beta\langle\hat{\partial}_\lambda L(v^t,\lambda^t),\lambda^t\rangle + \frac{(1-\gamma)\beta^2}{2}\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2. \end{aligned} \tag{II.11}$$

Utilizing the upper bound of (II.11) into the right hand side of (II.8) results in

$$\begin{aligned} & KL\big((1-\gamma)\lambda \,||\, (1-\gamma)\lambda^{t+1}\big) - KL\big((1-\gamma)\lambda \,||\, (1-\gamma)\lambda^t\big) \\ \leq & \frac{(1-\gamma)\beta^2}{2}\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2 + (1-\gamma)\beta\langle\hat{\partial}_\lambda L(v^t,\lambda^t) - w^t + w^t, \lambda^t - \lambda\rangle. \end{aligned}$$

Rearranging the terms and deviding both sides by $(1-\gamma)\beta$ proves this lemma. ■

C. Bounding the $\mathbb{E}[T_i]$'s

Step 1. Bounding $\mathbb{E}[T_1]$. Note that $\lambda^1 = \frac{\mathbf{1}}{(1-\gamma)|\mathcal{S}||\mathcal{A}|}$, we know

$$\begin{aligned}\mathbb{E}[T_1] &= \frac{1}{T(1-\gamma)\beta}\sum_{s,a}(1-\gamma)\lambda_{sa}^*\left(\log(\lambda_{sa}^*) - \log(|\mathcal{S}|^{-1}|\mathcal{A}|^{-1})\right) \\ &\leq \frac{1}{T(1-\gamma)\beta}\sum_{s,a}(1-\gamma)\lambda_{sa}^*\log(|\mathcal{S}||\mathcal{A}|) \\ &= \frac{\log(|\mathcal{S}||\mathcal{A}|)}{T(1-\gamma)\beta}. \end{aligned} \tag{II.12}$$

Step 2. Bounding $\mathbb{E}[T_2]$. For each t, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2\big|v_t,\lambda_t\right] &= \mathbb{E}_{s_t,a_t}\left[\sum_{s,a}\lambda_{sa}^t\left(\frac{\hat{r}_{ss'a}+\gamma v_{s'}-v_s-M_1}{\zeta_{sa}^t}\cdot\mathbf{1}_{(s,a)=(s_t,a_t)} - c\left(\hat{\partial}\rho(\lambda^t)\right)_{sa} - M_2\right)^2\Big|v_t,\lambda_t\right] \\ &\leq 2\mathbb{E}_{s_t,a_t}\left[\sum_{s,a}\lambda_{sa}^t\left(c\left(\hat{\partial}\rho(\lambda^t)\right)_{sa}+M_2\right)^2 + \lambda_{s_t,a_t}^t\left(\frac{\hat{r}_{s_ts_t'a_t}+\gamma v_{s_t'}-v_{s_t}-M_1}{\zeta_{s_ta_t}^t}\right)^2\Big|v_t,\lambda_t\right] \\ &\leq 8(1-\gamma)^{-1}c^2\sigma^2 + 2\sum_{s,a}\lambda_{sa}^t\zeta_{sa}^t\left(\frac{\hat{r}_{ss'a}+\gamma v_{s'}-v_s-M_1}{\zeta_{sa}^t}\right)^2 \\ &\leq 8(1-\gamma)^{-1}c^2\sigma^2 + 2\sum_{s,a}\frac{\lambda_{sa}^t\left(\hat{r}_{ss'a}+\gamma v_{s'}-v_s-M_1\right)^2}{(1-\delta)(1-\gamma)\lambda_{sa}^t + \frac{\delta}{|\mathcal{S}||\mathcal{A}|}} \\ &\leq 8(1-\gamma)^{-1}c^2\sigma^2 + 2\sum_{s,a}\frac{64\lambda_{sa}^t(1-\gamma)^{-2}(1+c\sigma)^2}{(1-\delta)(1-\gamma)\lambda_{sa}^t + \frac{\delta}{|\mathcal{S}||\mathcal{A}|}} \\ &\leq 8(1-\gamma)^{-1}c^2\sigma^2 + \frac{128|\mathcal{S}||\mathcal{A}|(1+c\sigma)^2}{(1-\delta)(1-\gamma)^3} \\ &\leq 8(1-\gamma)^{-1}c^2\sigma^2 + \frac{256|\mathcal{S}||\mathcal{A}|(1+c\sigma)^2}{(1-\gamma)^3}. \end{aligned}$$

The second row follows the definition of Δ_{sa}^t; The 4-th row is due to the assumption that $\|\hat{\partial}\rho\|_\infty \leq \sigma$; In the 5-th we substitute the definition of ζ_{sa}^t provided in Algorithm 1; In the 6-th row we substitute the detailed value of M_1; The 8-th row is because $\delta \in (0, \frac{1}{2})$. As a result, we have

$$\mathbb{E}[T_2] = \frac{\beta}{2T}\sum_{t=1}^T\mathbb{E}\left[\sum_{s,a}\lambda_{sa}^t(\Delta_{sa}^t)^2\right] \leq \frac{4\beta c^2\sigma^2}{1-\gamma} + \frac{128\beta|\mathcal{S}||\mathcal{A}|(1+c\sigma)^2}{(1-\gamma)^3}. \tag{II.13}$$

Step 3. Bounding $\mathbb{E}[T_3]$, because λ^* is a constant, for each t, we have

$$\mathbb{E}[\langle \hat{\partial}_\lambda L(v^t,\lambda^t) - w^t, \lambda^t - \lambda\rangle | v^t,\lambda^t] = -\langle (M_1+M_2)\cdot \mathbf{1}, \lambda^t - \lambda^*\rangle = 0,$$

where we have applied the fact that $\sum_{s,a}\lambda^t_{sa} = \sum_{s,a}\lambda^*_{sa}$, and $w^t = \mathbb{E}[\hat{\partial}_\lambda L(v^t,\lambda^t)|v^t,\lambda^t] + (M_1+M_2)\cdot\mathbf{1}$ when $\zeta^t > 0$. As a result,

$$\mathbb{E}[T_3] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\langle \hat{\partial}_\lambda L(v^t,\lambda^t) - w^t, \lambda^t - \lambda^*\rangle\right] = 0. \tag{II.14}$$

Step 4. Bounding $\mathbb{E}[T_4]$, we have

$$\mathbb{E}[T_4] = \frac{1}{2T\alpha}\mathbb{E}\left[\|v^1 - \bar{v}^*\|^2\right] \le \frac{8|\mathcal{S}|(1+c\sigma)^2}{T\alpha(1-\gamma)^2}. \tag{II.15}$$

Step 5. Bounding $\mathbb{E}[T_5]$, applying the expression (IV.8) yields

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\nabla}_v L(v^t,\lambda^t)\|^2 \big| v^t,\lambda^t\right] &= \mathbb{E}_{s_t,a_t,s'_t,\bar{s}_t}\left[\left\|\mathbf{e}_{\bar{s}_t} + \frac{\lambda^t_{s_ta_t}}{\zeta^t_{s_ta_t}}(\gamma\mathbf{e}_{s'_t} - \mathbf{e}_{s_t})\right\|^2 \Big| v_t,\lambda^t\right] \\
&= \mathbb{E}_{s_t,a_t,s'_t,\bar{s}_t}\left[\left\|\mathbf{e}_{\bar{s}_t} + \frac{\lambda^t_{s_ta_t}}{(1-\delta)(1-\gamma)\lambda^t_{s_ta_t} + \frac{\delta}{|\mathcal{S}||\mathcal{A}|}}(\gamma\mathbf{e}_{s'_t} - \mathbf{e}_{s_t})\right\|^2 \Big| v_t,\lambda^t\right] \\
&\le \mathbb{E}_{s_t,a_t,s'_t,\bar{s}_t}\left[3 + \frac{3\gamma^2+3}{(1-\delta)^2(1-\gamma)^2}\Big| v_t,\lambda^t\right] \\
&\le \frac{27}{(1-\gamma)^2}.
\end{aligned}$$

Consequently,

$$\mathbb{E}[T_5] = \frac{\alpha}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{\nabla}_v L(v^t,\lambda^t)\|^2\right] \le \frac{27\alpha}{2(1-\gamma)^2}. \tag{II.16}$$

Step 6. Bounding $\mathbb{E}[T_6]$. Because $\bar{v}^*$ is a random variable dependent on $\hat{\nabla}_v L(v^t,\lambda^t)$ we will need the following proposition.

Proposition 2.3 ([35]): Let $\mathcal{Z}\subseteq\mathbb{R}^d$ be a convex set and $w:\mathcal{Z}\to\mathbb{R}$ be a 1 strongly convex function with respect to norm $\|\cdot\|$ over $\mathcal{Z}$. With the assumption that for all $x\in\mathcal{Z}$ we have $w(x) - \min_{x\in\mathcal{Z}} w(x) \le \frac{1}{2}D^2$, then for any martingale difference sequence $\{Z_k\}_{k=1}^K \in \mathbb{R}^d$ and any random vector $z\in\mathcal{Z}$, it holds that

$$\mathbb{E}\left[\sum_{k=1}^{K}\langle Z_k, x\rangle\right] \le \frac{D}{2}\sqrt{\sum_{k=1}^{K}\mathbb{E}\left[\|Z_k\|_*^2\right]},$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.
With this proposition, and note that $\mathbb{E}\left[\langle\hat{\nabla}_v L(v^t,\lambda^t)|v^t,\lambda^t\right] = \nabla_v L(v^t,\lambda^t)$, we have

$$\begin{aligned}
\mathbb{E}[T_6] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\langle \nabla_v L(v^t,\lambda^t) - \hat{\nabla}_v L(v^t,\lambda^t), v^t - \bar{v}^*\rangle\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\langle \nabla_v L(v^t,\lambda^t) - \hat{\nabla}_v L(v^t,\lambda^t), \bar{v}^*\rangle\right] \\
&\le \frac{\sqrt{|\mathcal{S}|}(1+c\sigma)}{T(1-\gamma)}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla_v L(v^t,\lambda^t) - \hat{\nabla}_v L(v^t,\lambda^t)\|^2\right]} \\
&\le \frac{\sqrt{|\mathcal{S}|}(1+c\sigma)}{T(1-\gamma)}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|\hat{\nabla}_v L(v^t,\lambda^t)\|^2\right]} \\
&\le \frac{\sqrt{|\mathcal{S}|}(1+c\sigma)}{T(1-\gamma)}\sqrt{\frac{2T}{\alpha}\mathbb{E}[T_5]} \\
&\le \frac{3\sqrt{3|\mathcal{S}|}(1+c\sigma)}{\sqrt{T}(1-\gamma)^2}.
\end{aligned} \tag{II.17}$$