Cautious Reinforcement Learning via Distributional Risk in the Dual Domain

Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, *Member, IEEE*, and Alec Koppel, *Member, IEEE*

***Abstract*—We study the estimation of risk-sensitive policies in reinforcement learning (RL) 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, and hence modifications of Bellman's equations are required, which often do not permit efficient fixed point schemes. To ameliorate this issue, we propose a new definition of risk, which we call *caution*, as a penalty function added to the dual of the linear programming (LP) formulation of tabular RL. Caution quantifies the distributional risk of a policy, and is a function of the policy's long-term state occupancy measure. When the risk is a convex function of the occupancy measure, we propose to solve this problem in an online model-free manner with a stochastic variant of primal-dual method that uses Kullback-Lieber (KL) divergence as its proximal term. We establish that the sample complexity required to attain near-optimal solutions matches tight dependencies on the cardinality of the spaces, but also depends on the infinity norm of the risk measure's gradient. Moreover, when the risk is non-convex, we propose a block-coordinate augmentation of the aforementioned approach, and establish its convergence to KKT points. Experiments demonstrate that this approach improves the reliability of reward accumulation without additional computation as compared to risk-neutral LP solvers, both for convex and non-convex risks, respectively, for the cases of KL divergence to a prior, and the variance.**

***Index Terms*—Markov decision process (MDP), reinforcement learning, non-convex optimization, stochastic optimization.**

I. Introduction

IN REINFORCEMENT learning (RL) [2], 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, at which point 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 [3], videogames [4], [5], control amidst complicated physics [6], and management applications [7].

The canonical performance metric for RL is the expected long-term accumulation of rewards, or value. Unfortunately, restricting focus to expected returns fails to encapsulate many well-documented aspects of reasoning under uncertainty such as anticipation [8], inattention [9], and risk-aversion [10]. In this work, we focus on goals beyond expected rewards, specifically, risk sensitivity, both due to its inherent value in behavioral science and in pursuit of improving the reliability of RL in safety-critical applications [11]. Risk-awareness broadens the focus of decision making from expected outcomes to other quantifiers of uncertainty. Definitions of risk, originally quantified using the variance in portfolio management [12], have broaden to higher-order moments or quantiles [13], and gave rise to a rich theory of coherent risk [14], which has gained attention in RL in recent years [15], [16] 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* [17], that is, Bellman's principle of optimality does not hold. This issue has necessitated modified Bellman equations [18], [19], [20], multi-stage schemes [16], or policy search [21], 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 [15], [22], [23], [24], [25] in the spirit of chance-constrained programming [26] common in model predictive control. However, chance-constraint satisfaction may require additional Monte Carlo sampling as well as Lagrangian relaxation in settings where strong duality does not hold [15], [23]. A risk of specific interest in risk-aware reinforcement learning is the mean-variance trade-off, see [27], [28], [29], [30] and references therein. Recently, [31] proposed a block coordinate ascent approach to solve the mean-variance trade-off in RL, in which a nonassymptotic convergence to a stationary point was provided.

An additional approach is Bayesian [32] and distributional RL [33], 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) [19], or other quantiles of the return distribution [34], [35]. However, designing efficient and convergent distributional models, an active research thrust

Manuscript received September 29, 2020; revised April 18, 2021; accepted May 8, 2021. Date of publication May 17, 2021; date of current version June 21, 2021. This work was supported in part by the Army Research Laboratory under Agreement W911NF-16-2-0008, in part by the SMART Scholarship for Service, in part by ARL DCIST CRA, and in part by the ORAU Journeyman Fellowship. A part of this work was presented in American Control Conference (ACC), New Orleans, USA, 2021 [1]. *(Junyu Zhang and Amrit Singh Bedi contributed equally to this work.) (Corresponding author: Alec Koppel.)*

Junyu Zhang and Mengdi Wang are with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: junyuz@princeton.edu; mengdiw@princeton.edu).

Amrit Singh Bedi and Alec Koppel are with CISD, U.S. Army Research Laboratory, Adelphi, MD 20783 USA (e-mail: amrit0714@gmail.com; alec.e.koppel.civ@mail.mil).

Digital Object Identifier 10.1109/JSAIT.2021.3081108

within approximate Bayesian inference, is hardly a settled question in its own right [34], [35]. Specifically, since the prior and posterior are non-conjugate for this setting, one must set up an approximate variational inference problem, whose statistical properties depend on solutions to global optimization problems [36], and hence are in general nontrivial to quantify.

For these reasons, in this work we put forth a new definition of 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 (Section 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 (Section IV).
- establish that the number of sample transitions required to attain approximately optimal solutions of this scheme matches tight dependencies on the cardinality of the state and action spaces, as compared to the typical risk-neutral setting (Section V), specifically, Theorems 1–2 and Corollary 1.
- extend the method to nonconvex caution risks by using a block coordinate ascent (BCA) scheme (Section V-B), and establish that this augmented version of the algorithm converges to a local Karush-Kuhn-Tucker (KKT) point, i.e., it nearly satisfies the constraints and the objective reaches near-stationarity (Theorem 3).
- demonstrate the experimental merits of this approach for improving the reliability of reward accumulation without additional computational burdens (Section VI), specifically for the convex risk of KL divergence to a prior (Section VI-B) and the non-convex risk as defined by the variance (Section VI-A)

Additional Context: We expand upon some related notions of risk in RL, and the associated algorithmic approaches to impose it in practice. In particular, value/policy iteration schemes based upon a risk surrogate derived from the dual LP formulation of RL is considered in [37, Appendix A.2]. We note that the resulting solution methodology hinges upon policy/value iteration schemes which require access to error-free value function estimates, and performance is guaranteed only asymptotically. Imposing constraints directly on the space of occupancy measures in the primal-domain is considered in [38], which results in an infinite-dimensional state space augmentation. Discretization together with approximate value iteration is developed, which yields convergence rates that depend on the discretization error, although its exact sample complexity is unaddressed.

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 $r_{sa} \coloneqq \mathbb{E}[\hat{r}_{ss'a}|s, a] = \sum_{s' \in \mathcal{S}} P_a(s, s') \cdot \hat{r}_{ss'a}$ and the vector of rewards for each action a as $r_a = [r_{1a}, \ldots, r_{|\mathcal{S}|a}]^T \in \mathbb{R}^{|\mathcal{S}|}$.

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

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

B. Bellman Equation and Duality

The optimal value function v^* (1) satisfies Bellman's optimality principle [39]:

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

for all $s \in \mathcal{S}$. Let v be a column vector in $\mathbb{R}^{|\mathcal{S}|}$, then $v \geq v^*$ as long as $v \geq \gamma P_a v + r_a, \forall a$, see [40, Sec. 6.2]. This property can be illustrated by a multi-armed bandit (single-state MDP) problem, where $v \geq \gamma v + \max_a\{r_a\}$ directly implies $v \geq \frac{\max_a\{r_a\}}{1-\gamma} = v^*$. Consequently, the Bellman optimality equation (2) may be reformulated as a linear program (LP) (see [40, Sec. 6.9] and [41]):

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

where ξ is an arbitrary positive vector. The dual of (3) is given as (see [40, Sec. 6.9])

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

where $\lambda_a = [\lambda_{1a}, \ldots, \lambda_{|\mathcal{S}|a}]^\top \in \mathbb{R}^{|\mathcal{A}|}$ 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 (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 1: Suppose the variable $\lambda \in \mathbb{R}_+^{|\mathcal{S}|\times|\mathcal{A}|}$ satisfies the conditions

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

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

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

Furthermore, it satisfies

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

and

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

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 1. Forms of Proposition 1 have been floating around in the operations research community for some time [42], [43], although they percolated into modern RL beginning with [44].

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 peak risk and variance.

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., (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 (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 (4) then takes the form:

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

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. When ρ is convex, problem (9) is a convex optimization problem that facilitates computationally efficient solutions.

Denote by λ^* the optimal solution to the cautious policy optimization problem (9). 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 (9), the corresponding primal is no longer the LP problem (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 (9) is no longer the risk-neutral value vector, meaning that we are solving a *different problem* than (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 (9).

Example 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 $\rho(\lambda) = -\log(\lambda(\bar{S}) - (1-\delta))$, 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 $\rho(\lambda) = -\sum_{j=1}^m \log(\langle \lambda, r_j\rangle - b_j)$. In this way, the optimal caution-sensitive policy will meet all the job requirements for large enough penalty c.

Example 2 (Peak Risk): Let $f_1, \ldots, f_m$ be risk functions of λ. Consider the peak risk defined as $\rho(\lambda) = \sup_{j\in[m]} f_j(\lambda)$.

- *Worst-Case Exposure to Danger Areas:* For example, let $S_1, \ldots S_m$ be known "danger" sets. If we let $f_j(\lambda) = \lambda(S_j)$

quantify the long-term exposure to S_j, the peak risk ρ measures the long-run exposure to the most acute danger.

• *Worst-Case Multitask Performance:* For another example, suppose there are m different tasks defined in the same environment with reward functions $r_1, \ldots, r_m$. Let $f_j(\lambda) = -\langle \lambda, r_j \rangle$ be the negative cumulative return for task j, and an agent has to do well in all the tasks and be evaluated based on her worst performance. Then the objective is a peak risk: $-\rho(\lambda) = \sup_{j\in[m]} -\langle \lambda, r_j \rangle$. In the preceding examples, f_j's are linear, therefore ρ is always convex.

Example 3 (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 $\hat{\lambda} \coloneqq (1-\gamma)\lambda$. Then it holds that $\langle \hat{\lambda}, r \rangle$ is the expected reward accumulation. Then, the variance of return per timestep takes the form

$$\rho(\lambda) = Var\big(\hat{r}_{ss'a}|\lambda\big) = \mathbb{E}^{\hat{\lambda}}\left[\left(\mathbb{E}^{\hat{\lambda}}\big[\hat{r}_{ss'a}\big] - \hat{r}_{ss'a}\right)^2\right] \tag{9}$$

where $\mathbb{E}^{\hat{\lambda}} \coloneqq \mathbb{E}_{(s,a,s')\sim\hat{\lambda}\times\mathcal{P}(\cdot|a,s)}$. 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

$$\rho(\lambda) = \left\langle \hat{\lambda}, R \right\rangle - \left\langle \hat{\lambda}, r \right\rangle^2, \tag{10}$$

which is a quadratic function of the variable λ. Note that the variance risk $\rho(\lambda)$ is non-convex with respect to λ.

Example 4 (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. Given access to such a prior policy $\bar{\pi}$, one may construct an associated occupancy state-action distribution $\bar{\lambda}$ by tracking state-action visitation counts across trajectories associated with actions selected by $\bar{\pi}$. This is because the probability of visiting a state-action pair may be written as the expectation of an indicator at state-action pair (s,a). As an example, under the environments such as video games, the prior policy $\bar{\pi}$ and the associated trajectories can be collected as the game replays of human players. Maintaining baseline performance with respect to this prior $\bar{\lambda}$, 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

$$\rho(\lambda) = \mathrm{KL}\big((1-\gamma)\lambda||\bar{\lambda}\big) \tag{11}$$

which is substituted into (9) 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 $\rho(\lambda) = \mathrm{KL}((1-\gamma)\sum_a \lambda_a||\mu)$, 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 (9). 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, (9) admits an equivalent formulation as a saddle point problem:

$$\begin{aligned}\max_{\lambda\in\mathcal{L}} \min_{v\in\mathcal{V}} L(v,\lambda) &= \langle \lambda, r\rangle - c\rho(\lambda) + \langle \xi, v\rangle \\ &\quad + \sum_{a\in\mathcal{A}} \lambda_a^\top(\gamma P_a - I)v,\end{aligned} \tag{12}$$

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

$$\mathcal{L} \coloneqq \left\{\lambda : \lambda \geq 0, \|\lambda\|_1 = (1-\gamma)^{-1}\right\}. \tag{13}$$

Given distribution ζ over $\mathcal{S}\times\mathcal{A}$, define the stochastic approximation of the risk-neutral component of the Lagrangian that is, the Lagrangian in (12) with $\rho(\lambda)$ omitted:

$$L^{\zeta}_{(s,a,s'),\bar{s}}(v,\lambda) \coloneqq v_{\bar{s}} + \mathbf{1}_{\{\zeta_{sa}>0\}} \cdot \frac{\lambda_{sa}\big(\hat{r}_{ss'a} + \gamma v_{s'} - v_s\big)}{\zeta_{sa}} \tag{14}$$

where $\bar{s} \sim \mathcal{P}(\xi)$ is a sample from the discrete distribution defined by probability vector ξ. Here ζ may be viewed as an importance sampling distribution that has positive mass over all state-action pairs, i.e., $\zeta(s,a) > 0$ for $\forall(s,a)$. By properly choosing the sampling distribution ζ, one will be able to construct a stochastic gradient estimator that has small variance. Then by direct computation, since the support of ζ contains that of λ, i.e., $\mathrm{supp}(\lambda) \subset \mathrm{supp}(\zeta)$, we may write

$$L(v,\lambda) = \mathbb{E}_{(s,a,s')\sim\zeta\times\mathcal{P}(\cdot|a,s),\bar{s}\sim\xi}\left[L^{\zeta}_{(s,a,s'),\bar{s}}(v,\lambda)\right] - c\rho(\lambda). \tag{15}$$

Thus, we view (12) as a stochastic saddle point problem.

We propose variants of stochastic primal-dual method applied to (12). To obtain the primal descent direction, we note that if $\mathrm{supp}(\lambda) \subset \mathrm{supp}(\zeta)$, an unbiased estimator of the gradient of L w.r.t. $v \in \mathcal{V}$ is

$$\begin{aligned}\hat{\nabla}_v L(v,\lambda) &\coloneqq \nabla_v L^{\zeta}_{(s,a,s'),\bar{s}}(v,\lambda) \\ &= \mathbf{e}_{\bar{s}} + \mathbf{1}_{\{\zeta_{sa}>0\}} \cdot \frac{\lambda_{sa}}{\zeta_{sa}}(\gamma\mathbf{e}_{s'} - \mathbf{e}_s),\end{aligned} \tag{19}$$

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

$$\begin{aligned}\hat{\partial}_\lambda L(v,\lambda) &\coloneqq \mathbf{1}_{\{\zeta_{sa}>0\}} \cdot \frac{\hat{r}_{ss'a} + \gamma v_{s'} - v_s - M_1}{\zeta_{sa}} \cdot \mathbf{E}_{s,a} \\ &\quad - c\hat{\partial}\rho(\lambda) - M_2 \cdot \mathbf{1},\end{aligned} \tag{20}$$

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 $s_t' \sim \mathcal{P}(\cdot|a_t, s_t)$ & $\hat{r}_{s_t s_t' a_t}$ from generative model.
Construct $\hat{\nabla}_v L(v^t, \lambda^t)$ [see (19)] and $\hat{\partial}_\lambda L(v^t, \lambda^t)$ [see (20)] .
Update v and λ as

$$v^{t+1} = \Pi_{\mathcal{V}}(v^t - \alpha \hat{\nabla}_v L(v^t, \lambda^t)) \tag{16}$$

and

$$\lambda^{t+\frac{1}{2}} = \operatorname*{argmax}_{\lambda} \langle \hat{\partial}_\lambda L(v^t, \lambda^t), \lambda - \lambda^t \rangle - \frac{1}{(1-\gamma)\beta} KL\big((1-\gamma)\lambda \,||\, (1-\gamma)\lambda^t\big). \tag{17}$$

$$\lambda^{t+1} = \lambda^{t+\frac{1}{2}} / \Big((1-\gamma)\|\lambda^{t+\frac{1}{2}}\|_1\Big). \tag{18}$$

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 (20), M_1 and M_2 are the "shift" parameters specified in Theorem 1 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:

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

With these estimates for the primal gradient and dual subgradient of the Lagrangian (15), we propose executing primal-dual stochastic subgradient iteration [45], [46] with the KL divergence in the dual domain. KL divergence is employed in the dual domain as a specialization of choice of Bregman divergence for a mirror ascent-style update (also known as entropic/exponentiated ascent). This specification attunes the metric to the (unnormalized) probability simplex [47], [48], which is well-known to result in refined constants in the convergence rate. 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 establish the convergence of Algorithm 1 when the caution (risk) ρ in (9) is convex in λ, after which we present extensions of Algorithm 1 to address the non-convex variance risk, with its associated convergence presented thereafter. 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 1: The caution function $\rho(\lambda)$ is convex but possibly non-smooth, and it has bounded subgradients as

$$\sup_{\lambda\in\mathcal{L}} \sup_{u\in\partial\rho(\lambda)} \|u\|_\infty \le \sigma < \infty. \tag{21}$$

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

$$\mathbb{E}\Big[\hat{\partial}\rho(\lambda)\Big] \in \partial\rho(\lambda), \quad \sup_{\lambda\in\mathcal{L}} \Big\|\hat{\partial}\rho(\lambda)\Big\|_\infty \le \sigma. \tag{22}$$

In our subsequent analysis, we treat σ as a known constant. In all of Examples 1-4, the caution function ρ is explicitly known, which yields $\hat{\partial}\rho(\lambda) \in \partial\rho(\lambda)$. For an instance, in Example 2, if we let $\rho(\lambda) = \sup_{j\in[m]}\langle c_j, \lambda\rangle$, then any subgradient is bounded by $|\hat{\partial}\rho(\lambda)| \le \sup_j \|c_j\|_\infty = \mathcal{O}(1)$. For another instance, in Example 4, $\rho(\lambda) = KL(\hat{\lambda} \,||\, \mu)$ for some fixed μ [see (11)], the gradient takes the form

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

for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Then, we can ensure Assumption 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

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

A. The Case of Convex Caution Risk

In this subsection, we characterize the performance of Algorithm 1 when the caution ρ is convex. We begin by noting that the saddle point problem (12) 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 1: If $\xi > 0$, then the primal optimizer v^* satisfies

$$\|v^*\|_\infty \le (1-\gamma)^{-1}(1 + c\sigma). \tag{23}$$

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

$$\mathcal{V} := \left\{v \in \mathbb{R}^{|\mathcal{S}|} : \|v\|_\infty \le 2\frac{1 + c\sigma}{1-\gamma}\right\}. \tag{24}$$

The proof of Lemma 1 is provided in Appendix VII. 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 (12) (and the equivalently (9)). 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 (12).

Theorem 1 (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 $\bar{\nu}$ be the output of Algorithm 1 and let λ^* be the optimum. Then for the output of Algorithm 1, we have

$$\mathbb{E}\left[L(\bar{\nu}, \lambda^*) - \min_{v\in\mathcal{V}} L(v, \bar{\lambda})\right] \leq \mathcal{O}\left(\sqrt{\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{T}} \cdot \frac{1+2c\sigma}{(1-\gamma)^2}\right). \tag{25}$$

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

$$T = \Theta\left(\frac{|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)(1+2c\sigma)^2}{(1-\gamma)^4\epsilon^2}\right). \tag{26}$$

The proof of this Theorem is provided in Appendix C. 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 (9).

Theorem 2 (Convergence to Optimal Caution-Sensitive Policies): Let the parameters M_1, M_2, δ, β, and α, as defined in Theorem 1, if $\bar{\lambda}$ is the output of Algorithm 1 after T iterations, then the constraint violation of the original problem (9) satisfies

$$\begin{cases} \bar{\lambda} \geq 0, \quad \|\bar{\lambda}\|_1 = (1-\gamma)^{-1} \\ \left\|\sum_{a\in\mathcal{A}}(I - \gamma P_a^\top)\bar{\lambda}_a - \xi\right\|_1 \leq \frac{(1-\gamma)\epsilon}{1+c\sigma} \leq (1-\gamma)\epsilon. \end{cases} \tag{27}$$

Moreover, the sub-optimality of (9) is given as

$$\mathbb{E}\left[\left(\langle\lambda^*, r\rangle - c\rho(\lambda^*)\right) - \left(\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda})\right)\right] \leq \epsilon \tag{28}$$

Eqs. (27) and (28) showed the output solution is ϵ-feasible and ϵ-optimal. Note that ϵ determines the number of samples T as given in (26). The proof is provided in Appendix D.

Theorem 2 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., [46], [49], [50] which proves that Algorithm 1 is sample-efficient.

Further, consider the case where ρ is a KL divergence as in Example 4. This ρ acts as a regularization term to keep λ close to a prior long-term behavior. In this case, we can show that the primal-dual algorithm enjoys better convergence rates. In particular, we show a tighter KL divergence bound between the estimated $\bar{\lambda}$ and the optimal state-action distribution.

Corollary 1 (The Case When ρ Is a KL Divergence): If $\bar{\lambda}$ is the output of Algorithm 1, with parameters M_1, M_2, δ, β, α and T chosen as in Theorem 1, with $\rho(\lambda) \coloneqq KL((1-\gamma)\lambda \,||\, \mu)$ is the KL divergence from given prior μ, we have $\mathbb{E}[KL((1-\gamma)\bar{\lambda} \,||\, (1-\gamma)\lambda^*)] \leq \frac{\epsilon}{c}$.

Corollary 1 explicitly gives a dependence of the risk in terms of penalty parameter c, which may be made small as the penalty parameter grows large. This is mainly due to the strong convexity of the KL divergence. When $c = 0$, the strong convexity of the problem is lost and an ϵ-duality gap will longer provide an $\mathcal{O}(\epsilon)$ upper bound on the KL distance to the optimal solution. Consequently, Corollary V.5 is only able to provide a trivial upper bound of $+\infty$. Next, we discuss the case where the caution ρ is not necessarily convex.

Algorithm 2 A Block Coordinate Ascent (BCA) Framework for Policy Optimization With Nonconvex Caution

Initialize: λ^0, μ^0.
for $k = 0, 1, \ldots, K-1$
Update μ^{k+1} by solving

$$\mu^{k+1} = \arg\max \Phi(\lambda^k, \mu) \quad \text{s.t.} \quad \mu \geq 0, \ \|\mu\|_1 = 1 \tag{29}$$

with a known closed form solution.
Update λ^{k+1} by solving the following to ϵ-sub-optimality

$$\max_{\lambda} \ \Phi(\lambda, \mu^{k-1}) \text{ s.t. } \sum_{a\in\mathcal{A}}(I - \gamma P_a^\top)\lambda_a = \xi, \ \lambda \geq 0 \tag{30}$$

using Algorithm 1.
Output: Select $(\lambda^{k^*}, \mu^{k^*})$ randomly from $(\lambda^1, \mu^1), \ldots, (\lambda^K, \mu^K)$.

B. Extension to Nonconvex Variance Risk

In this subsection, we specify the caution as variance as in Example 3, which is nonconvex unlike other examples. For this instance, our strategy of addressing the constraints of problem (9) via Lagrangian relaxation fails here due to the nonconvexity of the variance in terms of λ. Therefore, we propose to approximately solve the nonconvex saddle point problem by solving a blockwise convex surrogate problem. Consider the following surrogate problem for some $M > 0$:

$$\max_{\lambda\in\Lambda}\max_{\mu\in U} \ \Phi(\lambda, \mu) \coloneqq \langle\lambda, r\rangle - c\rho(\lambda, \mu) - \frac{M}{2}\left\|\mu - \hat{\lambda}\right\|^2 \tag{31}$$

where $\Lambda \coloneqq \{\lambda : \sum_{a\in\mathcal{A}}(I - \gamma P_a^\top)\lambda_a = \xi, \lambda_a \geq 0, a \in \mathcal{A}\}$,

$$U \coloneqq \{\mu : \mu \geq 0, \|\mu\|_1 = 1\}, \ \hat{\lambda} = (1-\gamma)\lambda,$$
$$\rho(\lambda, \mu) = \left\langle\hat{\lambda}, r\right\rangle^2 - 2\langle\mu, r\rangle\left\langle\hat{\lambda}, r\right\rangle + \langle\mu, R\rangle.$$

Note that the surrogate problem (31) is not equivalent to the original problem (9) when risk ρ is chosen to be variance. However, in this alternative formulation a quadratic penalty is applied to push distribution μ towards $\hat{\lambda}$. Observe that when $\mu = \hat{\lambda}$, we have $\rho(\lambda, \mu) = \langle\hat{\lambda}, R\rangle - \langle\hat{\lambda}, r\rangle^2$, which equals exactly to the variance function. Therefore, the problem (31) will be close to the original problem (9) when the penalty parameter M is reasonably large. In what follows we propose algorithmic solution to the surrogate problem, assuming that a sufficiently large M is chosen.

The surrogate objective Φ is strongly concave in λ for any fixed μ and is strongly concave in μ for any fixed λ. But Φ is not jointly concave in λ and μ. Therefore, we can employ block-coordinate ascent (BCA) to solve problem (31).

The BCA method alternates between the two steps: First we fix λ and optimize the problem over μ - this subproblem is a projection onto a simplex and has a closed form solution (see [51]); Second we fix μ and optimize over λ, which is a convex problem and can be solved by using Algorithm 1. The full scheme is presented in Algorithm 2. Finally, we establish the sample complexity of Algorithm 2 to find a first-order stationary point of (31).

Theorem 3 (Convergence to Approximate Stationarity): Suppose we apply Algorithm 1 to solve the subproblem (30) with T satisfying

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

And we solve the subproblem (29) with a closed form solution [51]. Let the number of outer iterations be

$$K \geq \frac{\max_{\lambda,\mu}\Phi(\lambda,\mu)-\Phi\left(\lambda^0,\mu^0\right)}{\epsilon}.$$

Then the output $(\lambda^{k^*},\mu^{k^*})$ of Algorithm 2 is an approximate-stationary solution to problem (31), which satisfies

$$\mathbb{E}\left[\|\Pi_\Lambda(\nabla_\lambda\Phi(\lambda,\mu))\|^2+\left\|\Pi_U\left(\nabla_\mu\Phi(\lambda,\mu)\right)\right\|^2\right] \leq \mathcal{O}\left((1-\gamma)^2\left(\frac{c^2|\mathcal{S}|^2|\mathcal{A}|^2}{M}+M\right)\epsilon\right) \tag{32}$$

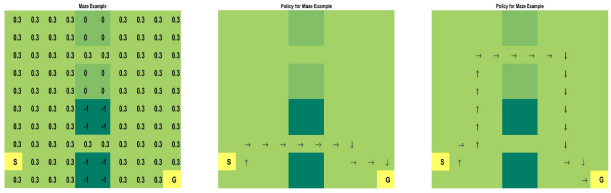
and

$$\begin{cases}\mathbb{E}\left[\left\|\sum_{a\in\mathcal{A}}\left(I-\gamma P_a^\top\right)\lambda_a^{k^*}-\xi\right\|_1\right]\leq(1-\gamma)\epsilon,\\ \lambda^{k^*}\geq 0,\quad \left\|\lambda^{k^*}\right\|_1=(1-\gamma)^{-1},\\ \mu^{k^*}\geq 0,\quad \left\|\mu^{k^*}\right\|_1=1.\end{cases} \tag{33}$$

Eqs. (32), (33) suggest that both the projected gradient norm and the level of constraint violation are $\mathcal{O}(\epsilon)$ small. They imply that the output solution is nearly feasible and nearly stationary. See Appendix F for the proof of theorem. We note that the results in Theorem V.6 are presented for the specialized case of variance as the risk objective. Here, we contrast these convergence rate results to other mean-variance optimization methods for MDP such as in [31], [52]. The algorithm in [52] is shown to converge asymptotically with no finite-time performance given. On the other hand, the algorithm in [31] achieves $\mathcal{O}(1/\epsilon^2)$ convergence but an explicit dependence on the cardinality of state actions space is absent. By contrast, Theorem V.6, establishes an $\mathcal{O}(1/\epsilon^2)$ convergence rate as well as explicit dependence upon the cardinality of state and action spaces.

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 (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 $\mathcal{A} :=$ {`up`, `down`, `left`, and `right`}). For the transition model, given the direction of the previous action selection, the agent moves 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 {`left`, or `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 and 4 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]) to the submission which visualize the safety of risk-awareness during training.



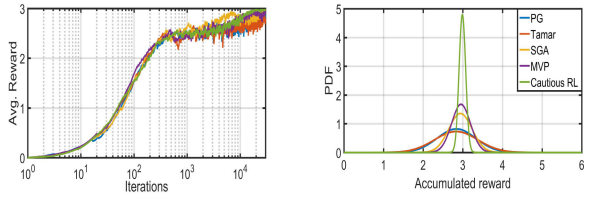
(a) Reward dist. (b) Risk neutral (c) Risk averse

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. Variance-Sensitive Policy Optimization

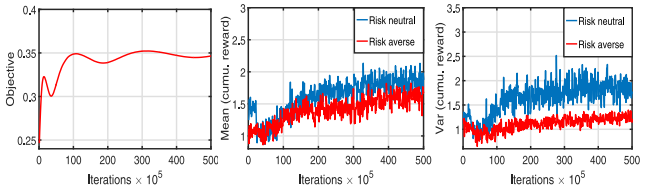
The variance risk given in Example 3 characterizes the statistical robustness of the rewards from a policy. To evaluate the merit of this definition, consider the maze example with the rewards distribution as described in Fig. 1(a). There are two ways to go from start to destination. The reward of dark 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. 3(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. 3(b) and 3(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. Further, for the grid world problem mentioned in Fig. 1(a), we compare the proposed technique against that which is given in [31] and [52]. We implemented policy gradient (PG), stochastic gradient ascent (SGA), and mean-variance policy improvement (MVP) all described in [31]. For clarity,

[1]https://tinyurl.com/sk4lddb, https://tinyurl.com/tlcl3m2



(a) Average reward return. (b) PDF for reward return.

Fig. 2. We compare the mean and variance of the cumulative return at the end of training visualized according to Laplace's approximation, i.e., a Gaussian distribution parameterized by the associated mean and variance. Observe that the proposed cautious RL algorithm yields higher returns on average with far less variance for the navigation task over the grid world (see Fig. 1).



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

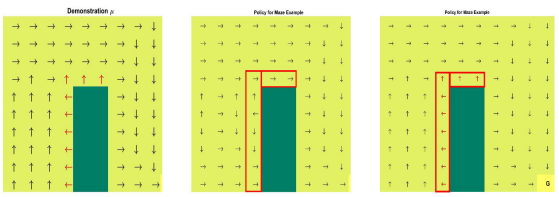
Fig. 3. (a) Convergence of the dual objective [see (9)]; Sample mean return (b) and variance (c) over 100 simulated trajectories when the risk is the (non-convex) variance. Observe the expected reward return is comparable while the risk-averse policy attains lower variance, and is thus more reliable.

we denote the one proposed in [52] as 'Tamar' in the legends of figure. We note that the proposed cautious RL algorithm yields higher reward returns as well as lower variance for the navigation task – see Fig. 2.

B. Caution as Proximity to a Prior

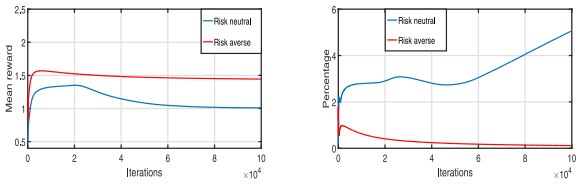
When a prior is available in the form of some baseline state-action distribution μ , KL divergence to the baseline makes sense as a measure of caution [see (11)] as stated in Example 4. To evaluate this definition, consider the setting where the baseline μ is a risk-neutral policy (shown in Fig. 4(a)) learned by solving (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. 4(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 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 4(b) and 4(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 (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) μ (b) Risk neutral (c) Risk averse

Fig. 4. 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. 5. 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).

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. 5. Overall, then, the incorporation of prior demonstrations results in the faster learning [see Fig. 5(a)] and reduces the proportion of time spent in unrewarding states as evidenced by Fig. 5(b).

VII. Conclusion

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 (9), generalizations to continuous spaces, and broadening caution to encapsulate other aspects of decision-making such as inattention and anticipation.

Appendix A
Proof of Proposition 1

Proof: Under the initial distribution ξ and the randomized policy $\pi : \mathcal{S} \mapsto \Delta_{|\mathcal{A}|}$ defined in (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

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

First, let us prove that (5) is equivalent to (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 (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

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

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

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

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

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

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

$$\begin{aligned}\lambda^\top r &= \hat{\xi}^\top r + \gamma \hat{\xi}^\top \hat{P}_\pi r + \gamma^2 \hat{\xi}^\top \hat{P}_\pi^2 r + \cdots \\ &= \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \hat{r}_{i_t i_{t+1} a_t} \middle| i_0 \sim \xi, a_t \sim \pi(\cdot|i_t)\right],\end{aligned}$$

which is as stated in (8) ∎

Appendix B
Proof of Lemma 1

Proof: Consider the min-max saddle point problem,

$$\begin{aligned}\max_{\lambda \ge 0} \min_{v \in \mathbf{R}^{|\mathcal{S}|}} L(v, \lambda) &= \langle \lambda, r\rangle - c\rho(\lambda) + \langle \xi, v\rangle \\ &\quad + \sum_{a \in \mathcal{A}} \lambda_a^\top (\gamma P_a - I)v, \qquad (34)\end{aligned}$$

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

$$\lambda^* = \operatorname*{argmax}_{\lambda \ge 0} L(v^*, \lambda) \quad \text{and} \quad \sum_{a \in \mathcal{A}} \Big(I - \gamma P_a^\top\Big)\lambda_a^* - \xi = 0. \qquad (35)$$

A remark is that, this is also the KKT condition for the original convex problem (9). Due to the concavity of $L(v^*, \lambda)$ for any fixed v^*, the condition $\lambda^* = \operatorname*{argmax}_{\lambda \ge 0} L(v^*, \lambda)$ is equivalent to the existence of a subgradient $w^* \in \partial_\lambda L(v^*, \lambda^*)$ s.t.

$$\langle w^*, \lambda - \lambda^*\rangle \le 0 \quad \text{for} \quad \forall \lambda \ge 0. \qquad (36)$$

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 $\forall (s', a') \ne (s, a)$, (36) further implies

$$\Big((e_s - \gamma P_{as})^\top v^* - r_{sa} + cu_{sa}^*\Big)\big(\lambda_{sa} - \lambda_{sa}^*\big) \ge 0,$$

where P_{as} is a column vector, with $P_{as}(s') = \mathcal{P}(s'|a, s)$. Combine this inequality with (35), we can formally write the final optimality condition as follows.

$$\begin{aligned}&\exists u^* \in \partial\rho\big(\lambda^*\big) \\ &\text{s.t.} \begin{cases} \sum_{a \in \mathcal{A}} \big(I - \gamma P_a^\top\big)\lambda_a^* = \xi, \quad \lambda^* \ge 0, \\ \big((e_s - \gamma P_{as})^\top v^* - r_{sa} + cu_{sa}^*\big)\big(\lambda_{sa} - \lambda_{sa}^*\big) \ge 0, \\ \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \forall \lambda_{sa} \ge 0. \end{cases} \qquad (37)\end{aligned}$$

By (7) of Proposition 1, we know that $\sum_{a \in \mathcal{A}} \lambda_{sa}^* \ge \sum_{a \in \mathcal{A}} \mathbf{Prob}(i_0 = s, a_0 = a | i_0 \sim \xi, a_0 \sim \pi(\cdot|i_0)) = \xi_s > 0 \quad \text{for} \quad \forall s \in \mathcal{S}$. 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 (37) implies that,

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

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

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

As a result,

$$\begin{aligned}1 + c\sigma &\ge \|\tilde{r} - c\tilde{u}\|_\infty \\ &= \left\|\Big(I - \gamma \tilde{P}^\top\Big)v^*\right\|_\infty \\ &\ge \|v^*\|_\infty - \left\|\gamma \tilde{P}^\top v^*\right\|_\infty \\ &\ge (1 - \gamma)\|v^*\|_\infty, \qquad (38)\end{aligned}$$

which implies the statement of Lemma 1. ∎

Appendix C
Proof of Theorem 1

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

Lemma 2: Suppose the iterate sequence $\{v^t\}$ is updated according to the rule (16) in Algorithm 1. Then for any t,

$$\begin{aligned}\big\langle \nabla_v L\big(v^t, \lambda^t\big), v^t - v\big\rangle &\le \frac{1}{2\alpha}\left(\big\|v^t - v\big\|^2 - \left\|v^{t+1} - v\right\|^2\right) \\ &\quad + \frac{\alpha}{2}\left\|\hat{\nabla}_v L\big(v^t, \lambda^t\big)\right\|^2 \\ &\quad + \Big\langle \nabla_v L\big(v^t, \lambda^t\big) - \hat{\nabla}_v L\big(v^t, \lambda^t\big), v^t - v\Big\rangle. \qquad (39)\end{aligned}$$

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

Lemma 3: Suppose the iterate sequence $\{\lambda^t\}$ is updated according to the rule (17) and (18) in Algorithm 1. For $\forall t$,

$$\begin{aligned}-\big\langle w^t, \lambda^t - \lambda\big\rangle \le \frac{1}{(1-\gamma)\beta}\Big(&KL\big((1-\gamma)\lambda\|(1-\gamma)\lambda^t\big) \\ &- KL\Big((1-\gamma)\lambda\|(1-\gamma)\lambda^{t+1}\Big)\Big)\end{aligned}$$

$$+\frac{\beta}{2}\sum_{s,a}\lambda_{sa}^{t}\big(\Delta_{sa}^{t}\big)^2\rangle,$$
$$+\Big\langle\hat{\partial}_{\lambda}L\big(v^t,\lambda^t\big)-w^t,\lambda^t-\lambda\Big\rangle, \tag{40}$$

where $w^t \coloneqq \mathbb{E}[\hat{\partial}_{\lambda}L(v^t,\lambda^t)|\lambda^t,v^t]+(M_1+M_2)\cdot\mathbf{1}\in\partial_{\lambda}L(v^t,\lambda^t)$ is a subgradient vector.

The proof of this lemma is provided in Appendix C-B. Based on these two lemmas, we start the proof of Theorem 1. 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 $\bar{v}^*\coloneqq\operatorname{argmin}_{v\in\mathcal{V}}L(v,\bar{\lambda})$. Then by the convex-concave structure of L we have

$$\begin{aligned}L\big(\bar{v},\lambda^*\big)-L\big(\bar{v}^*,\bar{\lambda}\big)&\le\frac{1}{T}\sum_{t=1}^{T}\big(L\big(v^t,\lambda^*\big)-L\big(\bar{v}^*,\lambda^t\big)\big)\\&=\frac{1}{T}\sum_{t=1}^{T}\big(L\big(v^t,\lambda^*\big)-L\big(v^t,\lambda^t\big)+L\big(v^t,\lambda^t\big)\\&\qquad-L\big(\bar{v}^*,\lambda^t\big)\big)\\&\le\frac{1}{T}\sum_{t=1}^{T}\big(-\big\langle w^t,\lambda^t-\lambda^*\big\rangle\\&\qquad+\big\langle\nabla_v L\big(v^t,\lambda^t\big),v^t-\bar{v}^*\big\rangle\big),\end{aligned} \tag{41}$$

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 (39) and $\lambda=\lambda^*$ in (40), we can sum up the inequlities (39) and (40) for $t=1,\dots,T$ to yield

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

Substitute this inequality into (41) and take the expectation on both sides, we get

$$\mathbb{E}\Big[L\big(\bar{v},\lambda^*\big)-\min_{v\in\mathcal{V}}L\big(v,\bar{\lambda}\big)\Big]\le\sum_{i=1}^{6}\mathbb{E}[T_i]. \tag{42}$$

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

$$\begin{aligned}&\mathbb{E}[T_1]\le\frac{\log(|\mathcal{S}\|\mathcal{A}|)}{T(1-\gamma)\beta},\mathbb{E}[T_2]\le\frac{4\beta c^2\sigma^2}{1-\gamma}\\&\qquad+\frac{128\beta|\mathcal{S}\|\mathcal{A}|(1+c\sigma)^2}{(1-\gamma)^3},\\&\mathbb{E}[T_3]=0,\mathbb{E}[T_4]\le\frac{8|\mathcal{S}|(1+c\sigma)^2}{T\alpha(1-\gamma)^2},\mathbb{E}[T_5]\le\frac{27\alpha}{2(1-\gamma)^2}\\&\mathbb{E}[T_6]\le\frac{3\sqrt{3|\mathcal{S}|}(1+c\sigma)}{\sqrt{T}(1-\gamma)^2}.\end{aligned}$$

Substitute these bounds for $\mathbb{E}[T_i]$'s into inequality (42) we get

$$\begin{aligned}\mathbb{E}\Big[L\big(\bar{v},\lambda^*\big)-\min_{v\in\mathcal{V}}L\big(v,\bar{\lambda}\big)\Big]&\le\frac{\log(|\mathcal{S}\|\mathcal{A}|)}{T(1-\gamma)\beta}+\frac{4\beta c^2\sigma^2}{1-\gamma}\\&\quad+\frac{128\beta|\mathcal{S}\|\mathcal{A}|(1+c\sigma)^2}{(1-\gamma)^3}\\&\quad+\frac{8|\mathcal{S}|(1+c\sigma)^2}{T\alpha(1-\gamma)^2}+\frac{27\alpha}{2(1-\gamma)^2}\\&\quad+\frac{3\sqrt{3|\mathcal{S}|}(1+c\sigma)}{\sqrt{T}(1-\gamma)^2}.\end{aligned} \tag{43}$$

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 $\mathbb{E}[L(\bar{v},\lambda^*)-\min_{v\in\mathcal{V}}L(v,\bar{\lambda})]\le\mathcal{O}(\sqrt{\frac{|\mathcal{S}\|\mathcal{A}|\log(|\mathcal{S}\|\mathcal{A}|)}{T}}\cdot\frac{1+c\sigma}{(1-\gamma)^2})$, which completes the proof. ∎

A. Proof of Lemma 2

Proof: Consider the update rule of v provided in (16). For any $v\in\mathcal{V}$, it holds that

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

Rearranging the above inequality yields

$$\begin{aligned}2\alpha\big\langle\nabla_v L\big(v^t,\lambda^t\big),v^t-v\big\rangle&\le\big\|v^t-v\big\|^2-\Big\|v^{t+1}-v\Big\|^2\\&\quad+\alpha^2\Big\|\hat{\nabla}_v L\big(v^t,\lambda^t\big)\Big\|^2\\&\quad-2\alpha\Big\langle\hat{\nabla}_v L\big(v^t,\lambda^t\big)\\&\qquad-\nabla_v L\big(v^t,\lambda^t\big),v^t-v\Big\rangle.\end{aligned} \tag{44}$$

Dividing both sides by 2α proves lemma. ∎

B. Proof of Lemma 3

Proof: Now let us consider the update rule of λ given by (17) and (18). Note that in the subproblem (17), the problem is separable for each component of λ and allows for a closed form solution, i.e.,

$$\begin{aligned}\lambda_{sa}^{t+\frac{1}{2}} &= \operatorname*{argmax}_{\lambda_{sa}} \Delta_{sa}^t \lambda_{sa} - \frac{1}{(1-\gamma)\beta}(1-\gamma)\lambda_{sa}\log\left(\frac{\lambda_{sa}}{\lambda_{sa}^t}\right) \\ &= \lambda_{sa}^t \cdot \exp\{\beta\Delta_{sa}^t\}, \end{aligned} \tag{45}$$

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

$$\lambda_{sa}^{t+1} = \frac{\lambda_{sa}^t \cdot \exp\{\beta\Delta_{sa}^t\}}{(1-\gamma)\sum_{s',a'}\lambda_{s'a'}^t \cdot \exp\{\beta\Delta_{s'a'}^t\}}. \tag{46}$$

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

$$\begin{aligned} &KL\Big((1-\gamma)\lambda\|(1-\gamma)\lambda^{t+1}\Big) - KL\big((1-\gamma)\lambda\|(1-\gamma)\lambda^t\big) \\ &= (1-\gamma)\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(\lambda_{sa}\log\left(\frac{\lambda_{sa}}{\lambda_{sa}^{t+1}}\right) - \lambda_{sa}\log\left(\frac{\lambda_{sa}}{\lambda_{sa}^t}\right)\right) \\ &= (1-\gamma)\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\lambda_{sa}\log\left(\frac{\lambda_{sa}^t}{\lambda_{sa}^{t+1}}\right) \\ &= (1-\gamma)\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\lambda_{sa}\left(\log\left((1-\gamma)\sum_{s',a'}\lambda_{s'a'}^t\exp\{\beta\Delta_{s'a'}^t\}\right) - \beta\Delta_{sa}^t\right) \\ &= \log\left((1-\gamma)\sum_{s',a'}\lambda_{s'a'}^t\exp\{\beta\Delta_{s'a'}^t\}\right) \\ &\quad - (1-\gamma)\beta\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\lambda_{sa}\Delta_{sa}^t \\ &= \log\left((1-\gamma)\sum_{s,a}\lambda_{sa}^t\exp\{\beta\Delta_{sa}^t\}\right) \\ &\quad - (1-\gamma)\beta\Big\langle\hat{\partial}_\lambda L\big(v^t,\lambda^t\big),\lambda\Big\rangle. \end{aligned} \tag{47, 48}$$

The equality in (47) is obtained by using the elementary definition of λ_{sa}^{t+1} in (46); The last equality of (48) is obtained by applying the definition of Δ_{sa}^t. Note that

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

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

$$\begin{aligned} &\log\left((1-\gamma)\sum_{s,a}\lambda_{sa}^t\cdot\exp\{\beta\Delta_{sa}^t\}\right) \\ &\le \log\left((1-\gamma)\sum_{s,a}\lambda_{sa}^t\cdot\left(1+\beta\Delta_{sa}^t+\frac{\beta^2}{2}\big(\Delta_{sa}^t\big)^2\right)\right) \\ &= \log\left(1+(1-\gamma)\beta\Big\langle\hat{\partial}_\lambda L(v^t,\lambda^t),\lambda^t\Big\rangle + \frac{(1-\gamma)\beta^2}{2}\sum_{s,a}\lambda_{sa}^t\big(\Delta_{sa}^t\big)^2\right) \\ &\le (1-\gamma)\beta\Big\langle\hat{\partial}_\lambda L\big(v^t,\lambda^t\big),\lambda^t\Big\rangle + \frac{(1-\gamma)\beta^2}{2}\sum_{s,a}\lambda_{sa}^t\big(\Delta_{sa}^t\big)^2. \end{aligned} \tag{49}$$

Utilizing the upper bound of (49) into the right hand side of (48) results in

$$\begin{aligned} &KL\Big((1-\gamma)\lambda\|(1-\gamma)\lambda^{t+1}\Big) - KL\big((1-\gamma)\lambda\|(1-\gamma)\lambda^t\big) \\ &\le \frac{(1-\gamma)\beta^2}{2}\sum_{s,a}\lambda_{sa}^t\big(\Delta_{sa}^t\big)^2 \\ &\quad + (1-\gamma)\beta\Big\langle\hat{\partial}_\lambda L\big(v^t,\lambda^t\big) - w^t + w^t, \lambda^t - \lambda\Big\rangle. \end{aligned} \tag{50}$$

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

$$\begin{aligned} \mathbb{E}[T_1] &= \frac{1}{T(1-\gamma)\beta}\sum_{s,a}(1-\gamma)\lambda_{sa}^* \\ &\quad\times\Big(\log\big(\lambda_{sa}^*\big) - \log\Big(|\mathcal{S}|^{-1}|\mathcal{A}|^{-1}\Big)\Big) \\ &\le \frac{1}{T(1-\gamma)\beta}\sum_{s,a}(1-\gamma)\lambda_{sa}^*\log(|\mathcal{S}\|\mathcal{A}|) \\ &= \frac{\log(|\mathcal{S}\|\mathcal{A}|)}{T(1-\gamma)\beta}. \end{aligned} \tag{51}$$

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

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

$$\le 8(1-\gamma)^{-1}c^2\sigma^2 + 2\sum_{s,a}\frac{64\lambda_{sa}^t(1-\gamma)^{-2}(1+c\sigma)^2}{(1-\delta)(1-\gamma)\lambda_{sa}^t+\frac{\delta}{|\mathcal{S}\|\mathcal{A}|}}$$
$$\le 8(1-\gamma)^{-1}c^2\sigma^2 + \frac{128|\mathcal{S}\|\mathcal{A}|(1+c\sigma)^2}{(1-\delta)(1-\gamma)^3}$$
$$\le 8(1-\gamma)^{-1}c^2\sigma^2 + \frac{256|\mathcal{S}\|\mathcal{A}|(1+c\sigma)^2}{(1-\gamma)^3}.$$

The second row follows the definition of Δ_{sa}^t; The 4-th row is due to the assumption that $\|\hat{\partial}\rho\|_\infty \le \sigma$; 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

$$\mathbb{E}[T_2] = \frac{\beta}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{s,a}\lambda_{sa}^t\left(\Delta_{sa}^t\right)^2\right] \le \frac{4\beta c^2\sigma^2}{1-\gamma} + \frac{128\beta|\mathcal{S}\|\mathcal{A}|(1+c\sigma)^2}{(1-\gamma)^3}. \tag{52}$$

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

$$\mathbb{E}\left[\left\langle\hat{\partial}_\lambda L\left(v^t,\lambda^t\right)-w^t,\lambda^t-\lambda\right\rangle|v^t,\lambda^t\right] \tag{53}$$
$$= -\left\langle(M_1+M_2)\cdot\mathbf{1},\lambda^t-\lambda^*\right\rangle = 0, \tag{54}$$

where we have applied the fact that $\sum_{s,a}\lambda_{sa}^t = \sum_{s,a}\lambda_{sa}^*$, 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,

$$\mathbb{E}[T_3] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\langle\hat{\partial}_\lambda L\left(v^t,\lambda^t\right)-w^t,\lambda^t-\lambda^*\right\rangle\right] = 0. \tag{55}$$

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

$$\mathbb{E}[T_4] = \frac{1}{2T\alpha}\mathbb{E}\left[\left\|v^1-\bar{v}^*\right\|^2\right] \le \frac{8|\mathcal{S}|(1+c\sigma)^2}{T\alpha(1-\gamma)^2}. \tag{56}$$

Step 5: Bounding $\mathbb{E}[T_5]$, applying the expression (19) yields

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

Consequently,

$$\mathbb{E}[T_5] = \frac{\alpha}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\hat{\nabla}_v L\left(v^t,\lambda^t\right)\right\|^2\right] \le \frac{27\alpha}{2(1-\gamma)^2}. \tag{57}$$

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 [53]: 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 $\mathbb{E}[\langle\hat{\nabla}_v L(v^t,\lambda^t)|v^t,\lambda^t] = \nabla_v L(v^t,\lambda^t)$, we have

$$\mathbb{E}[T_6] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\langle\nabla_v L\left(v^t,\lambda^t\right)-\hat{\nabla}_v L\left(v^t,\lambda^t\right),v^t-\bar{v}^*\right\rangle\right]$$
$$= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\langle\nabla_v L\left(v^t,\lambda^t\right)-\hat{\nabla}_v L\left(v^t,\lambda^t\right),\bar{v}^*\right\rangle\right]$$
$$\le \frac{\sqrt{|\mathcal{S}|}(1+c\sigma)}{T(1-\gamma)} \times\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla_v L(v^t,\lambda^t)-\hat{\nabla}_v L(v^t,\lambda^t)\right\|^2\right]}$$
$$\le \frac{\sqrt{|\mathcal{S}|}(1+c\sigma)}{T(1-\gamma)}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left\|\hat{\nabla}_v L(v^t,\lambda^t)\right\|^2\right]}$$
$$\le \frac{\sqrt{|\mathcal{S}|}(1+c\sigma)}{T(1-\gamma)}\sqrt{\frac{2T}{\alpha}\mathbb{E}[T_5]}$$
$$\le \frac{3\sqrt{3|\mathcal{S}|}(1+c\sigma)}{\sqrt{T}(1-\gamma)^2}. \tag{58}$$

APPENDIX D
PROOF OF THEOREM 2

Proof: The first row of (27) is directly satisfied due to the feasibility of $\bar{\lambda} \in \mathcal{L}$. Now we prove the second row of (27). When the parameters are chosen according to Theorem 1, we know

$$\epsilon \ge \mathbb{E}\left[L\left(\bar{v},\lambda^*\right)-\min_{v\in\mathcal{V}}L\left(v,\bar{\lambda}\right)\right]. \tag{59}$$

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

$$\min_{v\in\mathcal{V}}L\left(v,\bar{\lambda}\right) = \min_{\|v\|_\infty\le 2C}\left\langle\bar{\lambda},r\right\rangle - c\rho\left(\bar{\lambda}\right)+\langle\xi,v\rangle + \sum_{a\in\mathcal{A}}\bar{\lambda}_a(\gamma P_a - I)v$$
$$= \left\langle\bar{\lambda},r\right\rangle - c\rho\left(\bar{\lambda}\right) - 2C\left\|\sum_{a\in\mathcal{A}}\left(I-\gamma P_a^\top\right)\bar{\lambda}_a-\xi\right\|_1. \tag{60}$$

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

$$L\left(\bar{v},\lambda^*\right) = \left\langle\lambda^*,r\right\rangle - c\rho\left(\lambda^*\right)+\langle\xi,\bar{v}\rangle + \sum_{a\in\mathcal{A}}\left(\lambda_a^*\right)^\top(\gamma P_a - I)\bar{v} = \left\langle\lambda^*,r\right\rangle - c\rho\left(\lambda^*\right). \tag{61}$$

Substituting (60) and (61) into (59) yields

$$\mathbb{E}\Big[\big(\langle\lambda^*, r\rangle - c\rho(\lambda^*)\big) - \big(\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda})\big) + 2C\left\|\sum_{a\in\mathcal{A}}\Big(I - \gamma P_a^\top\Big)\bar{\lambda}_a - \xi\right\|_1\Big] \le \epsilon. \tag{62}$$

Actually, this inequality has already proved the bound (28) in terms of the objective value of problem (9). Also, by the feasibility of λ^*, the convexity of ρ and the optimality condition (37), we have

$$\begin{aligned}
&\big(\langle\lambda^*, r\rangle - c\rho(\lambda^*)\big) - \big(\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda})\big) + \left\langle v^*, \sum_{a\in\mathcal{A}}\Big(I - \gamma P_a^\top\Big)\bar{\lambda}_a - \xi\right\rangle \\
&= \big(\langle\lambda^*, r\rangle - c\rho(\lambda^*)\big) - \big(\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda})\big) + \left\langle v^*, \sum_{a\in\mathcal{A}}\Big(I - \gamma P_a^\top\Big)\bar{\lambda}_a - \sum_{a\in\mathcal{A}}\Big(I - \gamma P_a^\top\Big)\lambda_a^*\right\rangle \\
&\ge \sum_{a\in\mathcal{A}}\big\langle(I - \gamma P_a)v^* - r_a + cu_a^*, \bar{\lambda}_a - \lambda_a^*\big\rangle \ge 0,
\end{aligned}$$

where $u^* \in \partial\rho(\lambda^*)$ is defined in (37), and $u_a^* \coloneqq [u_{1a}^*, \ldots, u_{|\mathcal{S}|a}^*]^\top$ is column vector. Immediately, this implies

$$\begin{aligned}
&\big(\langle\lambda^*, r\rangle - c\rho(\lambda^*)\big) - \big(\langle\bar{\lambda}, r\rangle - c\rho(\bar{\lambda})\big) \\
&\ge -\left\langle v^*, \sum_{a\in\mathcal{A}}\Big(I - \gamma P_a^\top\Big)\bar{\lambda}_a - \xi\right\rangle \\
&\ge -\|v^*\|_\infty\left\|\sum_{a\in\mathcal{A}}\Big(I - \gamma P_a^\top\Big)\bar{\lambda}_a - \xi\right\|_1 \\
&\ge -C\left\|\sum_{a\in\mathcal{A}}\Big(I - \gamma P_a^\top\Big)\bar{\lambda}_a - \xi\right\|_1.
\end{aligned} \tag{63}$$

where we used the fact that $\|v^*\|_\infty \le C$ proved in Lemma 1. Substitute (63) into (62) gives

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

Divide both sides by $C = (1-\gamma)^{-1}(1 + c\sigma)$ proves inequality (27). ∎

Appendix E
Proof of Proposition 1

Proof: Let v^* be the optimal for the saddle point problem (12). Then the duality gap also guarantees that

$$\mathbb{E}\big[L(\bar{v}, \lambda^*) - L(v^*, \bar{\lambda})\big] \le \epsilon.$$

Substituting the detailed form of $L(\cdot,\cdot)$ we get $\mathbb{E}[T_1 + T_2] \le \epsilon$ where

$$T_1 = \langle\xi, \bar{v}\rangle - c\rho(\lambda^*, r) - \big[\langle\xi, v^*\rangle - c\rho(\bar{\lambda}, r)\big] \tag{64}$$

and

$$T_2 = \sum_{a\in\mathcal{A}}\langle\bar{\lambda}_a, (I - \gamma P_a)\bar{v} - r_a\rangle - \sum_{a\in\mathcal{A}}\langle\lambda_a^*, (I - \gamma P_a)v^* - r_a\rangle.$$

Note that for λ^*, $\sum_{a\in\mathcal{A}}(I - \gamma P_a^\top)\lambda_a^* = \xi$. Consequently,

$$\begin{aligned}
T_2 &= \sum_{a\in\mathcal{A}}\big\langle(I - \gamma P_a)v^* - r_a, \bar{\lambda}_a - \lambda_a^*\big\rangle + \sum_{a\in\mathcal{A}}\big\langle\lambda_a^*, (I - \gamma P_a)\big(v^* - \bar{v}\big)\big\rangle \\
&= \sum_{a\in\mathcal{A}}\big\langle(I - \gamma P_a)v^* - r_a, \bar{\lambda}_a - \lambda_a^*\big\rangle + \xi^\top\big(v^* - \bar{v}\big).
\end{aligned}$$

Let us define $D(\bar{\lambda}, \lambda^*) \coloneqq \rho(\bar{\lambda}) - \rho(\lambda^*) - \langle\nabla\rho(\lambda^*), \bar{\lambda} - \lambda^*\rangle \ge 0$. Combine this equality for T_2 and the definition of T_1 in (64), we get

$$\begin{aligned}
\epsilon &\ge \mathbb{E}[T_1 + T_2] = c\mathbb{E}\big[D(\bar{\lambda}, \lambda^*)\big] \\
&\quad + \mathbb{E}\left[\underbrace{\sum_{a\in\mathcal{A}}\big\langle(I - \gamma P_a)v^* - r_a + c\nabla_{\lambda_a}\rho(\lambda^*), \bar{\lambda}_a - \lambda_a^*\big\rangle}_{\ge 0 \text{ due to optimality condition (37)}}\right] \\
&\ge c\mathbb{E}\big[D(\bar{\lambda}, \lambda^*)\big].
\end{aligned}$$

Now let us denote $\theta = (1 - \gamma)$ for the ease of notation. By direct calculation, we compute the function D as follows

$$\begin{aligned}
D(\bar{\lambda}, \lambda^*) &= \rho(\bar{\lambda}) - \rho(\lambda^*) - \big\langle\nabla\rho(\lambda^*), \bar{\lambda} - \lambda^*\big\rangle \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(\theta\bar{\lambda}_{sa}\log\left(\frac{\theta\bar{\lambda}_{sa}}{\mu_{sa}}\right) - \theta\lambda_{sa}^*\log\left(\frac{\theta\lambda_{sa}^*}{\mu_{sa}}\right)\right. \\
&\qquad \left. - \left(\theta\log\left(\frac{\theta\lambda_{sa}^*}{\mu_{sa}}\right) + \theta\right)\big(\bar{\lambda}_{sa} - \lambda_{sa}^*\big)\right) \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(\theta\bar{\lambda}_{sa}\log\left(\frac{\theta\bar{\lambda}_{sa}}{\mu_{sa}}\right) - \theta\bar{\lambda}_{sa}\log\left(\frac{\theta\lambda_{sa}^*}{\mu_{sa}}\right)\right. \\
&\qquad \left. - \theta\big(\bar{\lambda}_{sa} - \lambda_{sa}^*\big)\right) \\
&= KL\big(\theta\bar{\lambda}\|\theta\lambda^*\big),
\end{aligned}$$

where the last inequality use the fact that $\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\theta\bar{\lambda}_{sa} = 1 = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\theta\lambda_{sa}^*$. This completes the proof. ∎

Appendix F
Proof of Theorem 3

Proof: For the ease of presentation, let us first prove two lemmas.

Lemma 4: For each subproblem (29), there is a closed form solution. For (30), we apply Algorithm 1 with the number of iterations T set according to Theorem 3. Then for all λ^k and μ^k with $k \in \{1, \ldots, K\}$, the approximate feasibility condition (33) is satisfied. Furthermore we have

$$\mathbb{E}\left[\Phi\Big(\lambda^{k+1}, \mu^{k+1}\Big) - \max_{\lambda\in\Lambda}\Phi\Big(\lambda, \mu^{k+1}\Big)\right] \ge -\epsilon.$$

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

Lemma 5: Suppose the sequence $\{(\lambda^k, \mu^k)\}$ is generated by Algorithm 2 with the parameters set by Theorem 3. For each iteration of Algorithm 2, let us define

$$\lambda_*^{k+1} = \arg\max_{\lambda\in\Lambda}\Phi\Big(\lambda, \mu^{k+1}\Big).$$

Then the output $(\lambda^{k^*}, \mu^{k^*})$ solution satisfies

$$\mathbb{E}\left[\left\|\lambda^{k^*} - \lambda_*^{k^*}\right\|^2\right] \le \frac{2\epsilon}{M} \quad \text{and} \quad \mathbb{E}\left[\left\|\lambda_*^{k^*} - \lambda^{k^*-1}\right\|^2\right] \le \frac{4\epsilon}{M}.$$

The proof of this lemma is provided in Appendix F-B. Based on this result, let us bound the expected squared projected gradients. By the optimality of μ^{k^*} for subproblem (29),

$$\Pi_U\left(\nabla_\mu \Phi\left(\lambda^{k^*-1}, \mu^{k^*}\right)\right) = 0.$$

Consequently,

$$\begin{aligned}
&\left\|\Pi_U\left(\nabla_\mu \Phi\left(\lambda^{k^*}, \mu^{k^*}\right)\right)\right\|^2 \\
&= \left\|\Pi_U\left(\nabla_\mu \Phi\left(\lambda^{k^*}, \mu^{k^*}\right)\right) - \Pi_U\left(\nabla_\mu \Phi\left(\lambda^{k^*-1}, \mu^{k^*}\right)\right)\right\|^2 \\
&\le \left\|\nabla_\mu \Phi\left(\lambda^{k^*}, \mu^{k^*}\right) - \nabla_\mu \Phi\left(\lambda^{k^*-1}, \mu^{k^*}\right)\right\|^2 \\
&= (1-\gamma)^2 \left\|4c\left\langle \lambda^{k^*} - \lambda^{k^*-1}, r\right\rangle r + 2M\left(\lambda^{k^*} - \lambda^{k^*-1}\right)\right\|^2 \\
&\le (1-\gamma)^2 \left\|4crr^\top + 2MI\right\|_2^2 \left\|\lambda^{k^*} - \lambda^{k^*-1}\right\|^2 \\
&= 4(1-\gamma)^2 (2c|\mathcal{S}\|\mathcal{A}| + M)^2 \left\|\lambda^{k^*} - \lambda^{k^*-1}\right\|^2 \\
&\le 8(1-\gamma)^2 (2c|\mathcal{S}\|\mathcal{A}| + M)^2 \\
&\times \left(\left\|\lambda_*^{k^*} - \lambda^{k^*-1}\right\|^2 + \left\|\lambda^{k^*} - \lambda_*^{k^*}\right\|^2\right).
\end{aligned}$$

In the above arguments, the fourth row is yielded by directly substituting the formulas of $\nabla_\mu \Phi(\lambda^{k^*}, \mu^{k^*})$ and $\nabla_\mu \Phi(\lambda^{k^*-1}, \mu^{k^*})$; the sixth row is due to $\|rr^\top\|_2 = \|r\|^2 \le |\mathcal{S}\|\mathcal{A}|$. Substituting the bounds provided in Lemma 5 yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|\Pi_U\left(\nabla_\mu \Phi\left(\lambda^{k^*}, \mu^{k^*}\right)\right)\right\|^2\right] \\
&\quad \le \frac{48\epsilon}{M}(1-\gamma)^2 (2c|\mathcal{S}\|\mathcal{A}| + M)^2. \qquad (65)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left\|\Pi_\Lambda\left(\nabla_\lambda \Phi\left(\lambda^{k^*}, \mu^{k^*}\right)\right)\right\|^2 \\
&= \left\|\Pi_\Lambda\left(\nabla_\lambda \Phi\left(\lambda^{k^*}, \mu^{k^*}\right)\right) - \Pi_\Lambda\left(\nabla_\lambda \Phi\left(\lambda_*^{k^*}, \mu^{k^*}\right)\right)\right\|^2 \\
&= \left\|\nabla_\lambda \Phi\left(\lambda^{k^*}, \mu^{k^*}\right) - \nabla_\lambda \Phi\left(\lambda_*^{k^*}, \mu^{k^*}\right)\right\|^2 \\
&= (1-\gamma)^4 \left\|2c\left\langle \lambda_*^{k^*} - \lambda^{k^*}, r\right\rangle \cdot r + M\left(\lambda_*^{k^*} - \lambda^{k^*}\right)\right\|^2 \\
&\le (1-\gamma)^4 \left\|2crr^\top + MI\right\|_2^2 \left\|\lambda_*^{k^*} - \lambda^{k^*}\right\|^2 \\
&\le (1-\gamma)^4 (2c|\mathcal{S}\|\mathcal{A}| + M)^2 \left\|\lambda_*^{k^*} - \lambda^{k^*}\right\|^2 \\
&\le \frac{16\epsilon}{M}(1-\gamma)^4 (2c|\mathcal{S}\|\mathcal{A}| + M)^2.
\end{aligned}$$

Combine the above inequality with (65), we get

$$\begin{aligned}
&\mathbb{E}\left[\|\Pi_\Lambda(\nabla_\lambda \Phi(\lambda, \mu))\|^2 + \left\|\Pi_U\left(\nabla_\mu \Phi(\lambda, \mu)\right)\right\|^2\right] \\
&\quad \le \mathcal{O}\left((1-\gamma)^2\left(\frac{c^2|\mathcal{S}|^2|\mathcal{A}|^2}{M} + M\right)\epsilon\right). \qquad (66)
\end{aligned}$$

■

A. Proof of Lemma 4

Proof: First, let us consider the subproblem (29) for updating μ. Note that for any fixed λ, we rewrite the subproblem as follows (constants omitted)

$$\begin{aligned}
\max_\mu \ & 2c\left\langle \hat{\lambda}, r\right\rangle \langle \mu, r\rangle - c\langle \mu, R\rangle - \frac{M}{2}\left\|\mu - \hat{\lambda}\right\|^2 \\
\text{s.t.} \ & \mu \ge 0, \|\mu\|_1 = 1. \qquad (67)
\end{aligned}$$

With a few more reformulation, this is actually a projection problem:

$$\begin{aligned}
\min_\mu \ & \left\|\mu - \left(\hat{\lambda} + \frac{2}{M}\left(2\left\langle \hat{\lambda}, r\right\rangle \cdot r - R\right)\right)\right\|^2 \\
\text{s.t.} \ & \mu \ge 0, \|\mu\|_1 = 1. \qquad (68)
\end{aligned}$$

This problem has a closed form solution and can be implemented within $\mathcal{O}(|\mathcal{S}\|\mathcal{A}| \log(|\mathcal{S}\|\mathcal{A}|))$ cost, see [51].

Second, we consider the subproblem (30). As a special case of problem (9) whose sample complexity is fully characterized by Theorem 1 and Theorem 2, all we need to do here is to specify the constant σ with the following "ρ function":

$$\rho(\lambda) = \left\langle \hat{\lambda}, r\right\rangle^2 - 2\langle \mu, r\rangle\left\langle \hat{\lambda}, r\right\rangle + \langle \mu, R\rangle + \frac{M}{2c}\left\|\hat{\lambda} - \mu\right\|^2,$$

where $\hat{\lambda} = (1-\gamma)\lambda$. Note that $\nabla\rho(\lambda) = (1-\gamma)\cdot(2(\langle \hat{\lambda}, r\rangle - \langle \mu, r\rangle)\cdot r + \frac{M}{c}(\hat{\lambda} - \mu))$. Then, if we directly set $\hat{\partial}\rho(\lambda) \coloneqq \nabla\rho(\lambda)$, we have

$$\sigma = \sup_{\lambda \in \mathcal{L}} \|\nabla\rho(\lambda)\|_\infty \le (1-\gamma)(1 + M/c).$$

Therefore, Theorem 1 and Theorem 2 tell us that the required sample complexity is

$$\begin{aligned}
T &= \Theta\left(\frac{|\mathcal{S}\|\mathcal{A}| \log(|\mathcal{S}\|\mathcal{A}|)(1+c\sigma)^2}{(1-\gamma)^4\epsilon^2}\right) \\
&= \Theta\left(\frac{|\mathcal{S}\|\mathcal{A}| \log(|\mathcal{S}\|\mathcal{A}|)}{(1-\gamma)^4\epsilon^2}\left(1 + (1-\gamma)^2\left(c^2 + M^2\right)\right)\right).
\end{aligned}$$

Thus we complete the proof. ■

B. Proof of Lemma 5

Proof: By Lemma 4, we have

$$\mathbb{E}\left[\Phi\left(\lambda^{k+1}, \mu^{k+1}\right) - \Phi\left(\lambda_*^{k+1}, \mu^{k+1}\right)\right] \ge -\epsilon. \qquad (69)$$

By M-strongly concavity of $\Phi(\cdot, \mu^k)$, we know

$$\begin{aligned}
\frac{M}{2}\left\|\lambda^{k+1} - \lambda_*^{k+1}\right\|^2 &\le \mathbb{E}\left[\Phi\left(\lambda_*^{k+1}, \mu^{k+1}\right) - \Phi\left(\lambda^{k+1}, \mu^{k+1}\right)\right] \\
&\le \epsilon.
\end{aligned}$$

Dividing both sides by $M/2$ proves the first inequality of the lemma. Again, by M-strongly concavity of $\Phi(\cdot, \mu^k)$, we also have

$$\Phi\left(\lambda_*^{k+1}, \mu^{k+1}\right) \ge \Phi\left(\lambda^k, \mu^{k+1}\right) + \frac{M}{2}\left\|\lambda_*^{k+1} - \lambda^k\right\|^2. \qquad (70)$$

Because $\mu^{k+1} = \arg\max_\mu \Phi(\lambda^k, \mu)$ s.t. $\mu \ge 0, \|\mu\|_1 = 1$. We know

$$\Phi\left(\lambda^k, \mu^{k+1}\right) \ge \Phi\left(\lambda^k, \mu^k\right). \qquad (71)$$

Combining the above three inequalities, we have

$$\begin{aligned}
&\mathbb{E}\Big[\Phi\Big(\lambda^{k+1},\mu^{k+1}\Big)-\Phi\Big(\lambda^{k},\mu^{k}\Big)\Big]\\
&\geq \mathbb{E}\Big[\Phi\Big(\lambda^{k+1},\mu^{k+1}\Big)-\Phi\Big(\lambda^{k},\mu^{k+1}\Big)\Big]\\
&= \mathbb{E}\Big[\Phi\Big(\lambda^{k+1},\mu^{k+1}\Big)-\Phi\Big(\lambda_*^{k+1},\mu^{k+1}\Big)\Big]\\
&\quad+\mathbb{E}\Big[\Phi\Big(\lambda_*^{k+1},\mu^{k+1}\Big)-\Phi\Big(\lambda^{k},\mu^{k+1}\Big)\Big]\\
&\geq \frac{M}{2}\left\|\lambda_*^{k+1}-\lambda^{k}\right\|^2-\epsilon,
\end{aligned}$$

where the second row is due to (71), the last row is due to (69) and (70). Summing up the above inequalities over all k and then take the average, then we know

$$\begin{aligned}
\frac{M}{2K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\lambda_*^{k}-\lambda^{k-1}\right\|^2\right] &\leq \frac{\max_{\lambda,\mu}\Phi(\lambda,\mu)-\Phi\big(\lambda^0,\mu^0\big)}{K}+\epsilon\\
&\leq 2\epsilon.
\end{aligned}$$

Because k^* is randomly chosen from $\{1,\ldots,K\}$, we get

$$\mathbb{E}\left[\left\|\lambda_*^{k^*}-\lambda^{k^*-1}\right\|^2\right]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\lambda_*^{k}-\lambda^{k-1}\right\|^2\right]\leq\frac{4\epsilon}{M}.$$

This proves the second inequality of this lemma. ∎

ACKNOWLEDGMENT

The views and conclusions contained in this article are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. government. The U.S. government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] J. Zhang, A. S. Bedi, M. Wang, and A. Koppel, "'Beyond cumulative returns via reinforcement learning over state-action occupancy measures"' in *Proc. IEEE Amer. Control Conf. (ACC)*, 2021.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998. [Online]. Available: http://www.cs.ualberta.ca/

[3] A. Karatzoglou, L. Baltrunas, and Y. Shi, "Learning to rank for recommender systems," in *Proc. 7th ACM Conf. Recommender Syst. (RecSys)*, 2013, pp. 493–494.

[4] V. Mnih *et al.*, "Playing Atari with deep reinforcement learning," 2013, arXiv:1312.5602.

[5] O. Vinyals *et al.*, *Alphastar: Mastering the Real-Time Strategy Game Starcraft II*, DeepMind Blog, London, U.K., 2019, p. 2.

[6] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," 2016. [Online]. Available: https://arxiv.org/abs/1506.02438

[7] D. Peidro, J. Mula, R. Poler, and F.-C. Lario, "Quantitative models for supply chain planning under uncertainty: A review," *Int. J. Adv. Manuf. Syst.*, vol. 43, nos. 3–4, pp. 400–420, 2009.

[8] 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," *Cogn. Process.*, vol. 12, no. 3, pp. 301–310, 2011.

[9] C. A. Sims, "Implications of rational inattention," *J. Monetary Econ.*, vol. 50, no. 3, pp. 665–690, 2003.

[10] 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.

[11] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 22–31.

[12] H. Markowitz, "Portfolio selection," *J. Financ.*, vol. 7, no. 1, pp. 77–91, 1952.

[13] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *J. Bank. Financ.*, vol. 26, no. 7, pp. 1443–1471, 2002.

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

[15] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6070–6120, 2017.

[16] D. R. Jiang and W. B. Powell, "Risk-averse approximate dynamic programming with quantile-based risk measures," *Math. Oper. Res.*, vol. 43, no. 2, pp. 554–579, 2018.

[17] T. Björk and A. Murgoci, "A theory of Markovian time-inconsistent stochastic control in discrete time," *Financ. Stoch.*, vol. 18, pp. 545–592, 2014. [Online]. Available: https://doi.org/10.1007/s00780-014-0234-y

[18] A. Ruszczyński, "Risk-averse dynamic programming for Markov decision processes," *Math. Program.*, vol. 125, no. 2, pp. 235–261, 2010.

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

[20] U. Kose and A. Ruszczynski, "Risk-averse learning by temporal difference methods," 2020, arXiv:2003.00780.

[21] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy gradient for coherent risk measures," in *Advances in Neural Information Processing Systems (NIPS)*. Red Hook, NY, USA: Cuuran, 2015, pp. 1468–1476.

[22] V. Krishnamurthy, K. Martin, and F. V. Abad, "Implementation of gradient estimation to a constrained Markov decision problem," in *Proc. 42nd IEEE Int. Conf. Decis. Control (CDC)*, vol. 5. Maui, HI, USA, 2003, pp. 4841–4846.

[23] L. A. Prashanth, "Policy gradients for CVaR-constrained MDPs," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2014, pp. 155–169.

[24] S. Paternain, M. Calvo-Fullana, L. F. O. Chamon, and A. Ribeiro, "Learning safe policies via primal–dual methods," in *Proc. 58th IEEE Conf. Decis. Control*, Nice, France, 2019, pp. 6491–6467.

[25] M. Yu, Z. Yang, M. Kolar, and Z. Wang, "Convergent policy optimization for safe reinforcement learning," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 32, 2019, pp. 3121–3133.

[26] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM J. Opt.*, vol. 17, no. 4, pp. 969–996, 2007.

[27] M. J. Sobel, "The variance of discounted Markov decision processes," *J. Appl. Probab.*, vol. 19, pp. 794–802, Dec. 1982.

[28] S. Mannor and J. N. Tsitsiklis, "Mean-variance optimization in Markov decision processes," 2011, arXiv:1104.5601.

[29] D. Di Castro, A. Tamar, and S. Mannor, "Policy gradients with variance related risk criteria," 2012, arXiv:1206.6404.

[30] L. A. Prashanth and M. Ghavamzadeh, "Variance-constrained actor-critic algorithms for discounted and average reward MDPs," *Mach. Learn.*, vol. 105, no. 3, pp. 367–417, 2016.

[31] T. Xie *et al.*, "A block coordinate ascent algorithm for mean-variance optimization," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1073–1083.

[32] M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar, "Bayesian reinforcement learning: A survey," *Found. Trends Mach. Learn.*, vol. 8, nos. 5–6, pp. 359–483, 2015.

[33] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. 34th Int Conf Mach. Learn. (ICML)*, 2017, pp. 449–458.

[34] W. Dabney, G. Ostrovski, D. Silver, and R. Munos, "Implicit quantile networks for distributional reinforcement learning," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 80, Jul. 2018, pp. 1096–1105.

[35] 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 (NIPS)*, 2019, pp. 6190–6199.

[36] C. Qu, S. Mannor, and H. Xu, "Nonlinear distributional gradient temporal-difference learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 5251–5260.

[37] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A Lyapunov-based approach to safe reinforcement learning," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8103–8112.

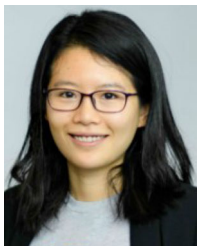
[38] W. B. Haskell and R. Jain, "A convex analytic approach to risk-aware Markov decision processes," *SIAM J. Control Optim.*, vol. 53, no. 3, pp. 1569–1598, 2015.

[39] D. P. Bertsekas and S. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*, 2004.
[40] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 2014.
[41] D. P. De Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Oper. Res.*, vol. 51, no. 6, pp. 850–865, 2003.
[42] A. S. Manne, "Linear programming and sequential decisions," *Manage. Sci.*, vol. 6, no. 3, pp. 259–267, 1960.
[43] F. d'Epenoux, "A probabilistic production and inventory problem," *Manage. Sci.*, vol. 10, no. 1, pp. 98–108, 1963.
[44] T. Wang, M. Bowling, D. Schuurmans, and D. J. Lizotte, "Stable dual dynamic programming," in *Advances in Neural Information Processing Systems*, 2008, pp. 1569–1576.
[45] Y. Chen and M. Wang, "Stochastic primal–dual methods and sample complexity of reinforcement learning," 2016, arXiv:1612.02516.
[46] Y. Chen, L. Li, and M. Wang, "Scalable bilinear *pi* learning using state and action features," 2018, arXiv:1804.10328.
[47] D. Precup and R. S. Sutton, "Exponentiated gradient methods for reinforcement learning," in *Proc. Int. Conf. Mech. Learn. (ICML)*, 1997, pp. 272–277.
[48] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Oper. Res. Lett.*, vol. 31, no. 3, pp. 167–175, 2003.
[49] M. Wang, "Primal-dual *pi* learning: Sample complexity and sublinear run time for ergodic Markov decision problems," 2017, arXiv:1710.06100.
[50] M. Wang, "Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time," *Math. Oper. Res.*, vol. 45, no. 2, pp. 517–546, 2020.
[51] W. Wang and M. Á. Carreira-Perpinán, "Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application," 2013, arXiv:1309.1541.
[52] A. Tamar, D. Di Castro, and S. Mannor, "Policy gradients with variance related risk criteria," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1651–1658.
[53] F. Bach and K. Y. Levy, "A universal algorithm for variational inequalities adaptive to smoothness and noise," in *Proc. 32nd Conf. Learn. Theory (COLT)*, 2019, pp. 164–194.

Junyu Zhang received the B.Sc. degree in computational mathematics from Peking University, China, in 2015, and the Ph.D. degree in industrial engineering from the University of Minnesota, Twin Cities, MN, USA, in 2020. He is currently a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Princeton University, USA. He will move to the Department of Industrial Systems Engineering and Management, National University of Singapore, as an Assistant Professor. His research interests are mainly in the theory of nonconvex optimization, Riemannian optimization, stochastic programming as well as their various applications in reinforcement learning and other machine learning and statistics problems.

Amrit Singh Bedi received the Diploma degree in electronics and communication engineering (ECE) from Guru Nanak Dev Polytechnic, Ludhiana, India, in 2009, the B.Tech. degree in ECE from the Rayat and Bahra Institute of Engineering and Bio-Technology, Kharar, India, in 2012, and the M.Tech. and Ph.D. degrees in electrical engineering from IIT Kanpur, in 2016 and 2018, respectively.

He is currently working as a Research Associate with the U.S. Army Research Laboratory, Adelphi, MD, USA. His research interests include optimization, signal processing, and machine learning. He is currently involved in developing risk-aware reinforcement learning and scalable online Bayesian and nonparametric methods. His paper was among the Best Paper Finalist at the 2017 IEEE Asilomar Conference on Signals, Systems, and Computers.



Mengdi Wang (Member, IEEE) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology in 2013. She is an Associate Professor with the Department of Electrical Engineering and Center for Statistics and Machine Learning, Princeton University. She is also affiliated with the Department of Computer Science and a visiting Research Scientist with DeepMind. At MIT, she was affiliated with the Laboratory for Information and Decision Systems and was advised by Dr. D. P. Bertsekas. Her research is supported by NSF, AFOSR, NIH, ONR, Google, Microsoft C3.ai DTI, and FinUP. Her research focuses on data-driven stochastic optimization and applications in machine and reinforcement learning. She received the Young Researcher Prize in Continuous Optimization of the Mathematical Optimization Society in 2016 (awarded once every three years), the Princeton SEAS Innovation Award in 2016, the NSF Career Award in 2017, the Google Faculty Award in 2017, and the MIT Tech Review 35-Under-35 Innovation Award (China region) in 2018. She serves as an Associate Editor for *Operations Research* and *Mathematics of Operations Research*, as an Area Chair for ICML, NeurIPS, AISTATS, and is on the editorial board of *Journal of Machine Learning Research*.

Alec Koppel (Member, IEEE) received the bachelor's degree in mathematics and the master's degree in systems science and mathematics from Washington University in St. Louis, St. Louis, MO, USA, and the master's degree in statistics and the Doctoral degree in electrical and systems engineering from the University of Pennsylvania (Penn) in August 2017. He has been a Research Scientist with the U.S. Army Research Laboratory in the Computational and Information Sciences Directorate since September 2017. His research interests are in optimization and machine learning. He currently focuses on scalable Bayesian learning, reinforcement learning, and decentralized optimization, with an emphasis on problems arising in robotics and autonomy. He is a recipient of the 2016 UPenn ESE Department Award for Exceptional Service, an Awardee of the Science, Mathematics, and Research for Transformation Scholarship, a coauthor of Best Paper Finalist at the 2017 IEEE Asilomar Conference on Signals, Systems, and Computers, and a Finalist for the ARL Honorable Scientist Award 2019, and an Awardee of the 2020 Director's Research Award–Translational Research Challenge.