Constrained Reinforcement Learning via Dissipative Saddle Flow Dynamics

Tianqi Zheng, Pengcheng You, and Enrique Mallada

***Abstract*— In constrained reinforcement learning (C-RL), an agent seeks to learn from the environment a policy that maximizes the expected cumulative reward while satisfying minimum requirements in secondary cumulative reward constraints. Several algorithms rooted in sampled-based primal-dual methods have been recently proposed to solve this problem in policy space. However, such methods are based on stochastic gradient descent-ascent algorithms whose trajectories are connected to the optimal policy only after a mixing output stage that depends on the algorithm's history. As a result, there is a mismatch between the behavioral policy and the optimal one. In this work, we propose a novel algorithm for constrained RL that does not suffer from these limitations. Leveraging recent results on regularized saddle-flow dynamics, we develop a novel stochastic gradient descent-ascent algorithm whose trajectories converge to the optimal policy almost surely.**

***Index Terms*— Constrained Reinforcement Learning, Stochastic Approximation, Stochastic Gradient Descent-Ascent**

I. Introduction

Reinforcement learning (RL) studies sequential decision-making problems where the agent aims to maximize its expected total reward by interacting with an unknown environment over time. However, in many applications such as electric grids and robotics, the agent often deals with conflicting requirements [1], or has safety constraints during the learning process [2]. The constrained reinforcement learning (C-RL) framework is a natural way to embed all conflicting requirements efficiently and incorporate safety [2]–[8].

There are two major approaches to finding the optimal policy of a C-RL problem, where the first approach solves it in the occupancy measure space. The constrained Markov Decision Process (CMDP) framework is a standard, and well-studied formulation for reinforcement learning with constraints [3]. The agent aims to maximize the total reward function while satisfying requirements in secondary cumulative reward constraints. The CMDP problem can be equivalently written as a linear programming problem in occupancy measure space, and the optimal policy could be recovered from the optimal occupancy measure [3]. However, this approach requires knowledge of the transition kernel of the underlying dynamical system explicitly, which is not always available in many realistic applications.

An alternative approach is to apply the Lagrangian duality and solve the C-RL problem in policy space [6]–[10]. These approaches solve the min-max optimization problem using a sampling-based primal-dual algorithm or stochastic gradient descent-ascent (SGDA) algorithm, where the Lagrangian function is augmented with a possible regularization term, e.g., a KL divergence regularization. The primal variables and dual variables are updated iteratively, either using gradient information or solving a sub-optimization problem. The outcome of primal-dual algorithms is often subject to two cases: in the first case, the output of the primal-dual algorithm is a mixing policy, which is a weighted average of history outputs [6]–[8]. In the second case, instead of showing the output policy converges to the optimal policy, they present a regret analysis for objective functions, and constraints [9], [10]. In summary, a key limitation is that the policy often oscillates and does not converge to the optimal policy, i.e., there is a mismatch between the behavioral policy and the optimal one. In this paper, we aim to tackle the above limitations by introducing a novel SGDA algorithm leveraging recent results on regularized saddle flow dynamics. Some of the proofs are omitted due to space constraints.

The key insight that the above sampling-based primal-dual algorithms do not converge is that the Lagrangian function for the C-RL problem does not possess sufficient convexity. The Lagrangian function is bilinear in occupancy measure space and is non-convex-concave in policy space. Our proposed method is rooted in the study of saddle flow dynamics [11], [12]. By adding a carefully crafted augmented regularization, the dissipative saddle flow proposed in [11] makes minimal requirements on convexity-concavity and yet still guarantees asymptotic convergence to a saddle point.

Leveraging tools from this dissipative saddle flow framework, we propose a novel algorithm to solve the C-RL problem in occupancy measure space, where the dynamics asymptotically converge to the optimal occupancy measure and optimal policy. We further extend the continuous-time algorithm in a model-free setting, where the discretized SGDA algorithm is shown to be the stochastic approximation of the continuous-time saddle flow dynamic. We prove that the SGDA algorithm almost surely converges to the optimal solution of the C-RL problem. To the best of our knowledge, this work is the first attempt to solve the C-RL problem to converge to the optimal occupancy measure and policy.

Notation: Let $\mathcal{K} \subset \mathbb{R}^n$ be a closed convex set. Given a point $y \in \mathbb{R}^n$, $\Psi_{\mathcal{K}}[y] = \operatorname{argmin}_{z\in\mathcal{K}} \|z-y\|$ denote the pointwise projection (nearest point) in $\mathcal{K}$ to y. Given $x \in \mathcal{K}$ and $v \in \mathbb{R}^n$, define the vector field projection of v at x with respect to $\mathcal{K}$ as: $\Pi_{\mathcal{K}}[x, v] = \lim_{\delta\to 0^+} \frac{\Psi_{\mathcal{K}}[x+\delta v]-x}{\delta}$

II. Problem Formulation

In the constrained reinforcement learning problem (C-RL), $\mathcal{S}$ denotes the finite state space, $\mathcal{A}$ denotes the finite action

T. Zheng and E. Mallada are with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA. Email: {tzheng8,mallada}@jhu.edu. P. You is with the Dept. of Industrial Engineering and Management, Peking University, Beijing, China. Email:pcyou@pku.edu.cn. This work was supported by NSF through grants CAREER 1752362, CPS 2136324, and TRIPODS 1934979.

space, and $P : \mathcal{S} \times \mathcal{A} \to \triangle^{|\mathcal{S}|}$ gives the transition dynamics of the CMDP, where $P(\cdot|s,a)$ denotes the probability distribution of next state conditioned on the current state s and action a. $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function, $g^i : \mathcal{S} \times \mathcal{A} \to [-1,1]$ denotes the i^{th} constraint cost function. The scalar γ denotes the discount factor, and q denotes the initial distribution of the states. A stationary policy is a map $\pi : \mathcal{S} \to \triangle^{|\mathcal{A}|}$ from states to a distribution in the action space. The value functions for both reward and constraints' cost following policy π are given by:

$$V_r^\pi(q) = (1-\gamma)\mathbf{E}_\pi[\textstyle\sum_{t=0}^\infty \gamma^t r(s_t,a_t) \,|\, s_0 \sim q],$$
$$V_{g^i}^\pi(q) = (1-\gamma)\mathbf{E}_\pi[\textstyle\sum_{t=0}^\infty \gamma^t g^i(s_t,a_t) \,|\, s_0 \sim q].$$

The standard C-RL problem aims to maximize the total reward function while satisfying requirements in secondary cumulative reward constraints:

$$\begin{aligned} \max_\pi \;& V_r^\pi(q) \\ s.t.\;& V_{g^i}^\pi(q) \geq h^i, \;\; \forall i \in [I]. \end{aligned} \tag{1}$$

There exist two classes of approaches to solving the optimal policy of a constrained reinforcement learning problem. The constrained Markov Decision Process (CMDP) framework equivalently expresses the C-RL problem as a linear programming problem in occupancy measure space [3]. Given a policy π, define $\lambda^\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$ as occupancy measure:

$$\lambda^\pi(s,a) = (1-\gamma)\textstyle\sum_{t=0}^\infty \gamma^t P_q^\pi(s_t = s, a_t = a),$$

where $s_0 \sim q$. By definition, the occupancy measure belongs to the probability simplex $\lambda^\pi \in \Delta$. Problem (1) can be equivalently written as a linear programming problem:

$$\begin{aligned} \max_{\lambda \in \Delta} \;& \textstyle\sum_a \lambda_a^T r_a \\ s.t.\;& \textstyle\sum_a \lambda_a^T g_a^i \geq h^i, \;\; i \in [I] \\ & \textstyle\sum_a (I - \gamma P_a^T)\lambda_a = (1-\gamma)q, \end{aligned} \tag{2}$$

where $\lambda_a = [\lambda(1,a), \ldots, \lambda(s,a)]^T \in \mathbb{R}^{|\mathcal{S}|}$ is the a^{th} column of λ^π, $r_a = [r(1,a), \ldots, r(s,a)]^T \in \mathbb{R}^{|\mathcal{S}|}$ denotes reward function associated with action a, P_a denotes the transition matrix associated with action a. The optimal policy could be recovered by finding the optimal occupancy measure

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

However, a key limitation in this approach is that it requires knowledge of the transition kernel of the underlying dynamical system explicitly, i.e., P_a, r_a, g_a^i.

Another approach is to apply the primal-dual algorithm to find the saddle points of the associated Lagrangian function of problem (1) in policy space:

$$L(\pi,\mu) = V_r^\pi + \textstyle\sum_{i=1}^I \mu_i (V_{g^i}^\pi - h^i).$$

Algorithms often augment Lagrangian function with a regularization term $\hat{L}(\pi,\mu) = L(\pi,\mu) + R(\pi,\mu)$, e.g., a KL divergence regularization, and update the policy and dual variable using one of the following rules:

$$\pi_{k+1} = \begin{cases} \pi_k + \eta \nabla_\pi \hat{L}(\pi, \mu_k) \\ \operatorname{argmax}_\pi \hat{L}(\pi, \mu_k) \end{cases} \quad \mu_{k+1} = \begin{cases} \mu_k - \eta \nabla_\pi \hat{L}(\pi_k, \mu) \\ \operatorname{argmin}_\mu \hat{L}(\pi_k, \mu) \end{cases}$$

Among the sampling-based primal-dual algorithms, several algorithms output a mixing policy of the form $\pi_T = \sum_{k=0}^{T-1} \eta_k \pi_k$, which is a weighted average of the history updates [6]–[8]. The output policy oscillates and does not converge to the optimal policy. On the other hand, several papers provide a regret analysis instead of showing the algorithm's convergence. To summarize, the CMDP approach could directly solve the optimal occupancy measure and the optimal policy while requiring knowledge of the transition kernel. The sampling-based primal-dual algorithms often output a mixing policy of history and do not converge to the optimal policy. The key limitation is that the Lagrangian function for the C-RL problem does not possess sufficient convexity. Specifically, the Lagrangian function is bilinear in occupancy measure space and is nonconvex in policy space. In this paper, we aim to provide a novel algorithm that tackles the above difficulties.

III. Key insight from saddle flow dynamics

Before introducing our algorithm, we would like to illustrate our key insight from saddle flow dynamics, which explains why the primal-dual algorithm oscillates and does not converge. For a min-max optimization problem, primal-dual algorithms require the Lagrangian $L(x,y)$ function to be strictly convex or concave on x or y, respectively, to converge. Consider the following motivating example with bilinear Lagrangian function:

$$\min_x \max_y L(x,y) := xy.$$

Our goal is to apply different dynamic laws that seek to converge to some saddle point $(x^*, y^*) = (0,0)$ of $L(x,y)$, which satisfies $L(x^*,y) \leq L(x^*,y^*) \leq L(x,y^*)$. In particular, consider the following classical primal-dual algorithm:

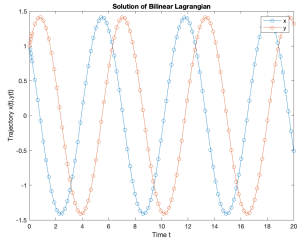
$$\dot{x} = -\nabla_x L(x,y) = -y,$$
$$\dot{y} = \nabla_y L(x,y) = x.$$

In Figure 1, (a) plots the time series trajectory of states x and y, and (b) plots the vector field and corresponding phase portrait. We observe that the dynamical system oscillates and does not converge to the saddle point (0,0).

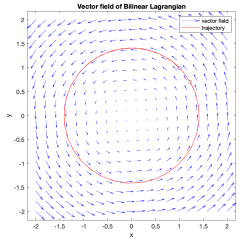
In [11], the authors introduce a regularization framework for saddle flow dynamics that guarantees asymptotic convergence to a saddle point based on mild assumptions. In this paper, we further extend the above framework to solve the C-RL problem. Specifically, consider the following constrained min-max optimization problem,

$$\min_{x \in \mathcal{K}} \max_{y \in \mathcal{V}} L(x,y)$$

where $\mathcal{K} \subset \mathbb{R}^n, \mathcal{V} \subset \mathbb{R}^m$ are bounded closed convex sets. We propose a regularized surrogate for $L(x,y)$ via the following



(a) time series trajectories



(b) phase portrait

Fig. 1: Primal-dual dynamics of bilinear Lagrangian function

augmentation:

$$L(x,y,z,w) := \frac{1}{2\rho}\|x-z\|^2 + L(x,y) - \frac{1}{2\rho}\|y-w\|^2$$

The following projected and regularized saddle flow dynamics aim to find the saddle points of the regularized Lagrangian, which contains the saddle point of the original Lagrangian. The regularized saddle flow dynamics still preserve the same distribution structure, which can be implemented in a fully distributed fashion, and requires the same gradient information as the classical primal-dual algorithm:

$$\begin{aligned}\dot{x} &= \Pi_{\mathcal{K}}\Big[x, -\nabla_x L(x,y) - \frac{1}{\rho}(x-z)\Big], \dot{z} = \Pi_{\mathcal{K}}\Big[z, \frac{1}{\rho}(x-z)\Big] \\ \dot{y} &= \Pi_{\mathcal{V}}\Big[y, -\nabla_y L(x,y) - \frac{1}{\rho}(y-w)\Big], \dot{w} = \Pi_{\mathcal{V}}\Big[w, \frac{1}{\rho}(y-w)\Big]\end{aligned} \tag{3}$$

Theorem 1. *Assume that $L(\cdot, y)$ is convex for $\forall y$ and $L(x,\cdot)$ is concave for $\forall x$, continuously differentiable, and there exists at least one saddle point $(x^* \in \mathcal{K}, y^* \in \mathcal{V})$, where $\mathcal{K} \subset \mathbb{R}^n, \mathcal{V} \subset \mathbb{R}^m$ are closed and convex. Then the projected saddle flow dynamics* (3) *asymptotically converge to some saddle point (x^*, y^*) of $L(x,y)$, while $x(t) \in \mathcal{K}, y(t) \in \mathcal{V}, \forall t$ with initialization $x(0) \in \mathcal{K}, y(0) \in \mathcal{V}$.*

Proof: See Appendix

The above theorem shows the projected and regularized saddle flow dynamics will asymptotically converge to the saddle point of the Lagrangian function, which requires mild assumptions on convexity. Additionally, the following result summarizes conditions under which the solutions of the projected system exist and are unique.

Proposition 2. *[12, Prop 2.2] Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be Lipschitz on a closed convex polyhedron $\mathcal{K} \in \mathbb{R}^n$. Then, for any $x_0 \in \mathcal{K}$, there exists a unique solution $t \to x(t)$ of the projected system $\dot{x} = \Pi_{\mathcal{K}}\Big[x, f(x)\Big]$ with $x(0) = x_0$.*

We now apply the regularized saddle flow dynamics to the bilinear Lagrangian function $L(x,y) = xy$.

$$\begin{aligned}\dot{x} &= -y - \frac{1}{\rho}(x-z), & \dot{z} &= \frac{1}{\rho}(x-z), \\ \dot{y} &= x - \frac{1}{\rho}(y-w), & \dot{w} &= \frac{1}{\rho}(y-w).\end{aligned}$$

According to Figure 2, the trajectories of the above saddle flow dynamics asymptotically converge to the saddle point $(0,0,0,0)$, even when the original Lagrangian function is bilinear.

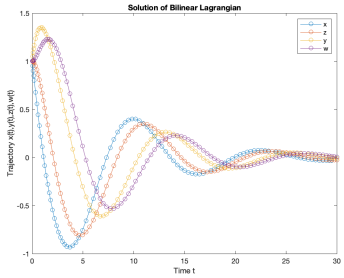


Fig. 2: Regularized saddle flow dynamics for $L(x,y) = xy$

A direct application of the above projected and regularized saddle flow dynamic is to solve the C-RL problem in occupancy measure space (2), where the Lagrangian function is also bilinear. Specifically, the Lagrangian function for (2) in occupancy measure space is:

$$\begin{aligned}L(\lambda,\mu,v) = &\sum_a \lambda_a^T r_a + \sum_i \mu_i \Big(\sum_a \lambda_a^T g_a^i - h^i\Big) \\ &+ (1-\gamma)\langle q, v\rangle - \sum_{a\in\mathcal{A}} \lambda_a^T (I - \gamma P_a) v,\end{aligned} \tag{5}$$

where $\mu_i \geq 0$ is the Lagrange multiplier associated with the i^{th} inequality constraint and v is the Lagrange multiplier associated with the equality constraint. Therefore, motivated by the projected and regularized saddle flow dynamics framework, we propose a regularized surrogate for (5) via the following augmentation:

$$\begin{aligned}L(v,\hat{v},\mu,\hat{\mu},\lambda,\hat{\lambda}) := &\frac{1}{2\rho}\|v-\hat{v}\|^2 + \frac{1}{2\rho}\|\mu-\hat{\mu}\|^2 \\ &+ L(v,\mu,\lambda) - \frac{1}{2\rho}\|\lambda-\hat{\lambda}\|^2\end{aligned} \tag{6}$$

Slater's condition for C-RL and the following Lemma establishes the boundedness of dual decision variables, which naturally provides a closed convex set for projection.

Assumption 1 (Slater's condition for C-RL)**.** *There exists a strictly feasible occupancy measure $\tilde{\lambda} \in \Delta$ of problem* (2)*, i.e., there exist some $\psi > 0$ such that*

$$\begin{aligned}&\sum_a \tilde{\lambda}_a^T g_a^i \geq h^i + \psi, \quad i \in [I] \\ &\sum_{a\in\mathcal{A}} (I - \gamma P_a^T)\tilde{\lambda}_a = (1-\gamma)q,\end{aligned}$$

Lemma 3. *[7, Lem.1][Bounded dual variable] Under the assumption 1, the optimal dual variables μ^*, v^* are bounded. Formally, it holds that $\|\mu^*\|_1 \leq \frac{2}{\psi}$ and $\|v^*\|_\infty \leq \frac{1}{1-\gamma} + \frac{2}{(1-\gamma)\psi}$*

Therefore, we propose the following projected saddle flow dynamics to find the saddle points of (6), where $\mathcal{U} := \{\mu | \mu \in \mathbb{R}^I_{\geq 0}, \|\mu\|_1 \leq \frac{2}{\psi}\}, \mathcal{V} := \{v | v \in \mathbb{R}^s, \|v^*\|_\infty \leq \frac{1}{1-\gamma} + \frac{2}{(1-\gamma)\psi}\}$

are both closed convex polyhedrons.

$$\begin{aligned}
\dot{v} &= \Pi_{\mathcal{V}}\Big[v, \sum_{a\in\mathcal{A}}(I-\gamma P_a^T)\lambda_a - (1-\gamma)q - \frac{1}{\rho}(v-\hat{v})\Big],\\
\dot{\hat{v}} &= \Pi_{\mathcal{V}}\Big[\hat{v}, \frac{1}{\rho}(v-\hat{v})\Big],\\
\dot{\mu}_i &= \Pi_{\mathcal{U}}\Big[\mu_i, h^i - \sum_a \lambda_a^T g_a^i - \frac{1}{\rho}(\mu_i - \hat{\mu}_i)\Big],\\
\dot{\hat{\mu}}_i &= \Pi_{\mathcal{U}}\Big[\hat{\mu}, \frac{1}{\rho}(\mu-\hat{\mu})\Big],\\
\dot{\lambda}_a &= \Pi_{\Delta}\Big[\lambda, r_a - (I-\gamma P_a)v + \sum_i \mu_i g_a^i - \frac{1}{\rho}(\lambda_a - \hat{\lambda}_a)\Big],\\
\dot{\hat{\lambda}}_a &= \Pi_{\Delta}\Big[\hat{\lambda}_a, \frac{1}{\rho}(\lambda-\hat{\lambda})\Big], \qquad (7)
\end{aligned}$$

The following theorem is a direct application of Theorem 1 and Proposition 2, which guarantees (7) asymptotically converge to the unique (optimal) saddle point of the C-RL problem (2). Then we could recover the optimal policy from the optimal occupancy measure λ^*.

Theorem 4. *Let Assumption 1 hold. Then the projected saddle flow dynamics* (7) *asymptotically converge to some saddle point* (λ^*, μ^*, v^*) *of* $L(\lambda, \mu, v)$*, while satisfying* $\lambda(t) \in \Delta, \mu(t) \in \mathcal{U}, \forall t$ *with proper initialization.*

IV. Stochastic Approximation for C-RL

In the following section, we aim to extend the proposed continuous-time saddle flow algorithm (7) to a model-free setting. Specifically, we propose a novel stochastic gradient descent-ascent algorithm, which does not require the knowledge of transition kernel. We show that the SGDA algorithm is a stochastic approximation of the continuous time saddle flow dynamics (7), which almost surely (w.p.1) converges to the unique saddle point of the C-RL problem.

In many optimization problems, the goal is to find some recursive numerical procedure that sequentially approximates a value of the decision variable x, which minimizes the objective function, e.g., $\dot{x} = h(x)$ or $x^{n+1} = x^n + \alpha^n h(x^n)$. Stochastic approximations attempt to solve the problem when one cannot actually observe $h(x)$, but rather $h(x)$ plus some error or noise. Consider the following projection algorithm:

$$x^{n+1} = \Psi_{\mathcal{G}}\Big[x^n + \alpha^n\Big(h(x^n) + \xi^n\Big)\Big], \qquad (8)$$

where $\mathcal{G} := \{x : q_i(x) \le 0, i \in [s]\}$ denotes the constraints and $\{\xi^n\}$ denotes a sequence of random variables. The goal is to generate a sequence $\{x^n\}$ estimate of the optimal value of x when the actual observation has random noise $h(x^n) + \xi^n$. In general, the projection $\Psi_{\mathcal{G}}[x]$ is easy to compute when the constraints are linear; i.e., when $\mathcal{G}$ is a polyhedron. We introduce the following list of standard assumptions for stochastic approximation

Assumption 2 (Stochastic Approximation)**.**

1.1 $h(\cdot)$ *is a continuous function.*

1.2 $\{\alpha^n\}$ *is a sequence of positive real numbers such that* $\alpha^n > 0, \sum_n \alpha^n = \infty, \sum_n (\alpha^n)^2 < \infty$*,*

1.3 G *is the closure of its interior and is bounded. The* $q_i(\cdot), i \in [s]$ *are continuously differentiable.*

1.4 There is a $T > 0$ *such that for each* $\epsilon > 0$

$$\lim_n P\{\sup_{j\ge n}\max_{t\le T}|\sum_{i=m(jT)}^{m(jT+t)-1}\alpha^i\xi^i| \ge \epsilon\} = 0,$$

where $t^n := \sum_{i=0}^{n-1}\alpha^i$ *and* $m(t) := \max_n\{t^n \le t\}$ *for* $t \ge 0$*.*

Under those standard assumptions for stochastic approximations, the sequence $\{x^n\}$ generated by the projection algorithm (8) will converge almost surely to a stable solution to the projected system.

Theorem 5. *[13, Theorem 5.3.1] Assume Assumption 2 hold. Consider the following ODE:*

$$\dot{x} = \Pi_{\mathcal{G}}\Big[x, h(x)\Big]. \qquad (9)$$

Let x^* *denotes an asymptotically stable point of* (9) *with domain of attraction* $DA(x^*)$ *and* x^n *generated by* (8)*. If* $A \in DA(x^*)$ *is compact and* $x^n \in A$ *infinitely often, then* x^n *converges to* x^* *almost surely as* $n \to \infty$*.*

Consider the following randomized primal-dual approach proposed in [7], [14], where we assume the presence of a generative model. For a given state action pair (s, a), the generative model provides the next state s' and the reward functions $r(s,a), g^i(s,a)$ to train the policy. Consider the following stochastic approximation for the Lagrangian function (5) for a distribution ξ:

$$L^{\xi}(\lambda,\mu,v) = (1-\gamma)v(s_0) - \sum_{i\in[I]}\mu_i h^i + \qquad (10)$$
$$\mathbf{1}_{\xi(s,a)>0}\frac{\lambda(s,a)\Big[r(s,a) - v(s) + \gamma v(s') + \sum_{i\in[I]}\mu_i g^i(s,a)\Big]}{\xi(s,a)}$$

where $s_0 \sim q, (s,a) \sim \xi$, and the next state $s' \sim P(\cdot|s,a)$. The stochastic approximation $L^{\xi}(\lambda,\mu,v)$ (10) is an unbiased estimator for the Lagrangian function (5), i.e., $\mathbf{E}_{\xi,P(\cdot|s,a),q}\Big[L^{\xi}(\lambda,\mu,v)\Big] = L(\lambda,\mu,v)$. Using the proposed stochastic approximation of the Lagrangian function, consider the following projection algorithm for solving the C-RL problem in a model-free setting:

$$\begin{aligned}
v^{n+1} &= \Psi_{\mathcal{V}}\Big[v^n + \alpha^n\Big(\mathbf{1}_{\xi(s,a)>0}\frac{\lambda(s,a)[e(s) - \gamma e(s')]}{\xi(s,a)}\\
&\quad - (1-\gamma)\mathbf{e}(s_0) - \frac{1}{\rho}(v^n - \hat{v}^n)\Big)\Big],\\
\hat{v}^{n+1} &= \Psi_{\mathcal{V}}\Big[\hat{v}^n + \alpha^n\frac{1}{\rho}(v^n - \hat{v}^n)\Big],\\
\mu_i^{n+1} &= \Psi_{\mathcal{U}}\Big[\mu_i^n + \alpha^n\Big(h^i - \mathbf{1}_{\xi(s,a)>0}\frac{\lambda(s,a)g^i(s,a)}{\xi(s,a)}\\
&\quad - \frac{1}{\rho}(\mu_i^n - \hat{\mu}_i^n)\Big)\Big],\\
\hat{\mu}_i^{n+1} &= \Psi_{\mathcal{U}}\Big[\hat{\mu}_i^n + \alpha^n\frac{1}{\rho}(\mu_i^n - \hat{\mu}_i^n)\Big],
\end{aligned}$$

$$\lambda_a^{n+1} = \Psi_\Delta\Big[\lambda_a^n + \alpha^n\Big(-\frac{1}{\rho}(\lambda_a^n - \hat{\lambda}_a^n) + \mathbf{1}_{\xi(s,a)>0}\frac{r(s,a) - v(s) + \gamma v(s') + \sum_i \mu_i^n g^i(s,a)}{\xi(s,a)}\Big)\Big],$$
$$\hat{\lambda}_a^{n+1} = \Psi_\Delta\Big[\hat{\lambda}_a^n + \frac{1}{\rho}(\lambda_a^n - \hat{\lambda}_a^n)\Big], \tag{11}$$

The following Theorem is a direct application of Theorem 5 and 4, which shows the sequence from (11) almost surely converges to the optimal solution to the C-RL problem.

Theorem 6. *Assume 1 and 2 hold, as $n \to \infty$, the sequence $\{\lambda^n, v^n, \mu^n\}$ generated by* (11) *almost surely (w.p.1) converge to the optimal solution of the C-RL problem* (2)*.*

V. Numerical Examples

In this section, we illustrate the effectiveness of our proposed approach using a classical CMDP problem: flow and service control problem in a single-server queue [3]. Specifically, we consider a discrete-time single-server queue with a buffer of finite size L. We assume that, at most, one customer may join the system in a time slot. The state s corresponds to the number of customers in the queue at the beginning of a time slot ($|\mathcal{S}| = L + 1$). The service action a is selected from a finite subset A, and the flow action b is selected from a finite subset B. Specifically, for two real numbers satisfying $0 < a_{\min} \leq a_{\max} < 1$, if the queue is non-empty and if the action of the server is $a \in A$, where A is a finite subset of $[a_{\min}, a_{\max}]$, then the service of a customer is successfully completed with probability a. Likewise, for two real numbers satisfying $0 \leq b_{\min} \leq b_{\max} < 1$, if the queue is not full and if the action of the server is $b \in B(s)$, where $B(s)$ is a finite subset of $[b_{\min}, b_{\max}]$, then the probability of having one arrival during this time slot is equal to b. We assume that $0 \in B(x)$ for all x; moreover, when the buffer is full, no arrivals are possible ($B(L) = 0$). The transition law $P(\cdot|s,a)$ is therefore given by:

$$\begin{cases} a(1-b) & \text{if } 1 \leq x \leq L, y = x-1; \\ ab + (1-a)(1-b) & \text{if } 1 \leq x \leq L, y = x; \\ (1-a)b & \text{if } 0 \leq x < L, y = x+1; \\ 1-(1-a)b & \text{if } y = x = 0; \end{cases}$$

The reward function $r(s,a,b)$ is a real-valued decreasing function that depends only on s, which can be interpreted as a holding cost. The reward function $g^1(s,a,b)$ corresponding to the service rate is assumed to be a decreasing function that depends only on a. It can be interpreted as a higher service success rate having a higher cost. The reward function $g^2(s,a,b)$ corresponding to the flow rate b is assumed to be an increasing function that depends only on b. It can be interpreted as a higher flow rate is more desired.

Suppose we want to solve the optimal policy for C-RL problem (1), while satisfying constraints for service and flow. In the following numerical example, we compare the result generated by (11) and the ground truth result by directly solving the linear programming 2, where we use the transition law stated above. Specifically, we choose $L = 4, A = [0.2, 0.3, 0.5, 0.6, 0.8], B = [0.1, 0.3, 0.5, 0.9, 0]$. The initial distribution q is set as uniform distribution. The reward functions are $r(s) = -s + 5, g^1(a) = -10a + 3, g^2(b) = 10b - 3$.

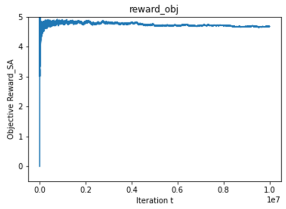


Fig. 3: objective function

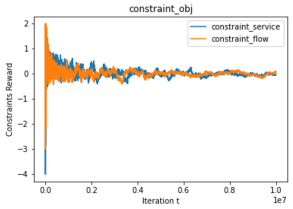


Fig. 4: constraint functions

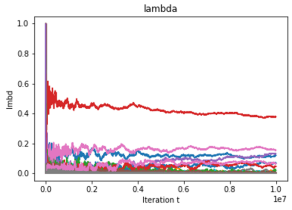


Fig. 5: occupancy measure λ

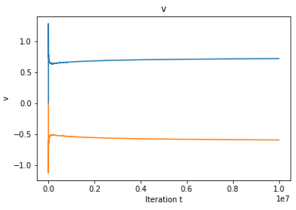


Fig. 6: dual variable v

We compare the cumulative reward function, constraint functions, and output decision variables λ, μ, v with the ground truth result by directly solving the linear programming problem (2). Results show that the decision variables converge to the optimal solution while satisfying the constraints for flow and service.

VI. Conclusion

In this work, we propose a novel SGDA algorithm to solve the C-RL problem in occupancy measure space leveraging tools from regularized saddle flow dynamics. Even when the Lagrangian function is bilinear, the continuous dynamics asymptotically converge to the optimal occupancy measure and policy. The discretized SGDA is a stochastic approximation of the continuous-time saddle flow dynamic. We further proved the SGDA algorithm almost surely converges to the optimal solution to the C-RL problem.

Appendix

A. Proof of Theorem 1

We will use the following technical Lemma:

Lemma 7. *For any closed convex set $\mathcal{K} \subset \mathbb{R}^n$ and $a, b \in \mathcal{K}, v \in \mathbb{R}^n$,the inner product*

$$\langle b - a, v - \Pi_{\mathcal{K}}[a, v]\rangle \leq 0$$

Proof: According to [15, Sec.0.6, Cor.1], we have the following variational inequality holds:

$$\langle b - \Psi_{\mathcal{K}}[c], c - \Psi_{\mathcal{K}}[c]\rangle \leq 0, \forall b \in \mathcal{K}, \forall c \in \mathbb{R}^n.$$

The rest follows from [16, Lem.4]

Using this lemma, the proof of Theorem 1 essentially follows from [11, Thm.9].

REFERENCES

[1] S. Mannor and N. Shimkin, "A geometric approach to multi-criterion reinforcement learning," *The Journal of Machine Learning Research*, vol. 5, pp. 325–360, 2004.

[2] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*. PMLR, 2017, pp. 22–31.

[3] E. Altman, *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

[4] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[5] A. Castellano, H. Min, J. Bazerque, and E. Mallada, "Reinforcement learning with almost sure constraints," *arXiv preprint arXiv:2112.05198*, 2021.

[6] Y. Chen, J. Dong, and Z. Wang, "A primal-dual approach to constrained markov decision processes," *arXiv preprint arXiv:2101.10895*, 2021.

[7] Q. Bai, A. S. Bedi, M. Agarwal, A. Koppel, and V. Aggarwal, "Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, 2022, pp. 3682–3689.

[8] M. Calvo-Fullana, S. Paternain, L. F. Chamon, and A. Ribeiro, "State augmented constrained reinforcement learning: Overcoming the limitations of learning with rewards," *arXiv preprint arXiv:2102.11941*, 2021.

[9] T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian, "Learning policies with zero or bounded constraint violation for constrained mdps," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 183–17 193, 2021.

[10] D. Ding, K. Zhang, T. Basar, and M. Jovanovic, "Natural policy gradient primal-dual method for constrained markov decision processes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8378–8390, 2020.

[11] P. You and E. Mallada, "Saddle flow dynamics: Observable certificates and separable regularization," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 4817–4823.

[12] A. Cherukuri, E. Mallada, and J. Cortés, "Asymptotic convergence of constrained primal–dual dynamics," *Systems & Control Letters*, vol. 87, pp. 10–15, 2016.

[13] H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems*. Springer Science & Business Media, 2012, vol. 26.

[14] M. Wang, "Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time," *Mathematics of Operations Research*, vol. 45, no. 2, pp. 517–546, 2020.

[15] J.-P. Aubin and A. Cellina, *Differential inclusions: set-valued maps and viability theory*. Springer Science & Business Media, 2012, vol. 264.

[16] E. Mallada and F. Paganini, "Stability of node-based multipath routing and dual congestion control," in *2008 47th IEEE Conference on Decision and Control*. IEEE, 2008, pp. 1398–1403.