Regret Minimization Algorithms for Single-Controller Zero-Sum Stochastic Games

Peng Guan, Maxim Raginsky, Rebecca Willett, and Daphney-Stavroula Zois

***Abstract*— Two-player single-controller zero-sum stochastic games are a class of zero-sum dynamic games with Markovian state dynamics, where only one player controls the state transitions. Design of optimal strategies for such games with large state and action spaces relies on computationally demanding dynamic programming. Linear programming can also be used, but the number of constraints equals the number of states. This paper presents a class of simple suboptimal strategies that can be constructed by playing a certain repeated static game where neither player observes the specific mixed strategies used by the other player at each round. We quantify the suboptimality of the resulting strategies and show that, when the two players honestly follow the prescribed protocol, each player can exploit the regularity or predictability of the moves of the other player, and thus speed up convergence to the minimax value.**

I. Introduction

Game theory studies strategic interactions between rational agents. Stochastic games [1], [2] are a special type of dynamic games with a global state variable. At each stage of the game, the players choose their actions, and each receives a payoff that depends on the current state and on the chosen actions. Then, the game moves to a new random state, whose distribution depends on the previous state and on the most recent actions of the players. Herein, we are interested in two-player *single-controller* zero-sum stochastic games, where only one agent controls the state transitions, and the loss of one agent is the reward of the other and vice versa. We focus on single-controller games because of their natural connection with so-called *regret minimization* methods for sequential (or online) learning [3] (we explain this connection in more detail below). Existing work [1], [4]–[6] has mainly focused on finding and computing equilibria for such games. However, in practice, the associated computational burden severely limits the range of applications (e.g., the computational complexity of dynamic programming scales quadratically with the number of states; in linear programming formulations, the number of constraints equals the number of states). In this paper, we construct a class of suboptimal yet simple strategies, such that the expected payoff of the two players is close to the minimax value.

Single-controller stochastic games are closely related to regret minimization [3] in online Markov decision processes (MDPs) [7]–[15]. Just like its offline counterpart, the online MDP problem involves a controlled Markov chain, but the state-action cost function varies arbitrarily with time, and the agent finds out the current cost function only after having taken an action. An online MDP can be viewed as a two-player single-controller stochastic game, where the agent controls the state transitions, while an *oblivious* (i.e., open-loop) environment chooses the cost functions. In such a situation, the agent's objective is to minimize *regret* relative to the best stationary Markov policy that could have been selected with full knowledge of the cost function sequence over the horizon of interest. Earlier work [8], [9], [12], [14] has concentrated on developing algorithms that achieve sublinear regret when the environment is acting arbitrarily and likely sub-optimally. In contrast, we consider the environment to be a rational opponent who is adaptive to the agent's actions and has a specific goal: *maximize his long-term average reward.*

In the setting of zero-sum matrix games (without a state), the von Neumann minimax theorem guarantees the existence of a minimax equilibrium in mixed strategies, such that each player can minimize his maximum losses. There is an intimate connection between the minimax value of a zero-sum game and regret minimization: if two players in a repeated zero-sum matrix game (i.e., a stochastic game without a state) respond to each other's moves using regret minimization algorithms, then each player's average payoff converges to the minimax value, their average strategies constitute an approximate minimax equilibrium, and the rate of convergence is determined by the players' regret [3]. Moreover, recent work by Daskalakis et al. [16] and by Rakhlin and Sridharan [17] has shown that, if the two players are honest and do not deviate from the prescribed protocol, then their average payoff converges to the minimax value at fast rates. To the best of our knowledge, the possibility of using regret minimization algorithms to approach the horizon-dependent minimax value of *dynamic games* has not been previously investigated in the literature.

In this paper, we look at regret minimization and stochastic games from a different and novel prospective. We are not interested in minimizing the regret against the best stationary strategy; instead, we shift our focus toward using regret minimization strategies to derive simple suboptimal strategies for single-controller stochastic games. To the best of our knowledge, this topic has not been the subject of previous work. Note that in the case of zero-sum matrix games, the

This work was supported by NSF under grants CCF-1017564 and CCF-1302438, and by AFOSR under grant FA9550-10-1-0390.

P. Guan is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA; e-mail: peng.guan@duke.edu.

M. Raginsky and D.-S. Zois are with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA; e-mail: {maxim, zois}@illinois.edu.

R. Willett is with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53796, USA; e-mail: willett@discovery.wisc.edu.

minimax equilibrium is a static object referring to the one-stage game, contrary to a T-round stochastic game, where the minimax equilibrium is a dynamic object. In this paper, we show how to approximate the dynamic minimax equilibrium by simple stationary strategies using regret minimization. As a first step, we reduce the single-controller stochastic game to an online linear optimization problem. This enables us to use regret minimization strategies to generate a pair of stationary policies and prove that the average payoff converges to the average minimax payoff at fast rates. As a result, we are able to approximately solve two-player stochastic games in a simple and efficient way. Moreover, we precisely quantify the degree of sub-optimality of the proposed policies by characterizing the rate at which the associated payoff converges to the game's minimax value. Finally, inspired by the work of Rakhlin and Sridharan [17], we illustrate how we can achieve faster convergence rates when both players incorporate appropriate prediction models of their opponents' strategies into their decision loops.

Notation. The simplex of all probability distributions on a finite set Y is denoted by $\mathcal{P}(\mathsf{Y})$; the set of all *Markov* (or *row-stochastic*) matrices $P = [P(z|y)]_{y\in\mathsf{Y}, z\in\mathsf{Z}}$ with rows and columns indexed by the elements of Y and Z respectively is denoted by $\mathcal{M}(\mathsf{Z}|\mathsf{Y})$. The elements of $\mathcal{M}(\mathsf{Z}|\mathsf{Y})$ transform probability distributions on Y into probability distributions on Z by matrix multiplication: $\mu \mapsto \mu P$. Any $\mu \in \mathcal{P}(\mathsf{Y})$ and $P \in \mathcal{M}(\mathsf{Z}|\mathsf{Y})$ can be combined to form a probability distribution $\mu \otimes P \in \mathcal{P}(\mathsf{Y}\times\mathsf{Z})$: $\mu\otimes P(y,z) \triangleq \mu(y)P(z|y)$. The L_1 distance between $\mu, \nu \in \mathcal{P}(\mathsf{Y})$ is

$$\|\mu-\nu\|_1 \triangleq \sum_{y\in\mathsf{Y}} |\mu(y)-\nu(y)| \equiv \sup_{f:\,\|f\|_\infty\le 1} |\langle\mu,f\rangle - \langle\nu,f\rangle|,$$

where the supremum is over all real-valued functions on Y with absolute value bounded by 1, and we use the linear functional notation for expectations: $\langle\mu,f\rangle = \mathbb{E}_\mu[f] = \sum_{y\in\mathsf{Y}}\mu(y)f(y)$. The Kullback–Leibler divergence (or relative entropy) between μ and ν is denoted by $D(\mu\|\nu)$.

II. Problem formulation

We consider a single-controller stochastic game, i.e., only one player controls the state transitions [5]. The finite state space is denoted by X, and there are two finite action spaces for the two players, U_1 and U_2. The cost function is $c : \mathsf{X}\times\mathsf{U}_1\times\mathsf{U}_2 \to [0,1]$. Player 1's closed-loop behavioral strategy is denoted by the tuple $\boldsymbol{\gamma} = (\gamma_1,\ldots)$, where $\gamma_t : \mathsf{X}^t\times\mathsf{U}_1^{t-1}\times\mathsf{U}_2^{t-1} \to \mathcal{P}(\mathsf{U}_1)$, and $P_{1,t} \equiv \gamma_t(x^t,u_1^{t-1},u_2^{t-1})$ is his mixed strategy at time t given the history $(x^t,u_1^{t-1},u_2^{t-1})$ of past and present states and past actions by both players. Similarly, Player 2 also has a closed-loop behavioral strategy denoted by the tuple $\boldsymbol{\delta} = (\delta_1,\ldots)$ with $\delta_t : \mathsf{X}^t\times\mathsf{U}_1^{t-1}\times\mathsf{U}_2^{t-1} \to \mathcal{P}(\mathsf{U}_2)$, and his mixed strategy at time t is given by $P_{2,t} = \delta_t(x^t,u_1^{t-1},u_2^{t-1})$. Since Player 1 alone controls the state transitions, the controlled transition kernel is given by $K(y|x,u_1)$, which specifies the probability of moving to state y given the current state x and Player 1's action u_1. The game protocol is the following:

```
$X_1 = x$
for $t = 1, 2, \ldots$
  Players 1 and 2 observe the state $X_t$
  Player 1 chooses $P_{1,t}$ and draws $U_{1,t} \sim P_{1,t}$;
  Player 2 chooses $P_{2,t}$ and draws $U_{2,t} \sim P_{2,t}$;
  $U_{1,t}$ and $U_{2,t}$ are revealed to both players
  Player 1 incurs cost $c(X_t, U_{1,t}, U_{2,t})$,
  Player 2 incurs cost $-c(X_t, U_{1,t}, U_{2,t})$
  The state is updated to $X_{t+1} \sim K(\cdot|X_t, U_{1,t})$
end for
```

Given the initial state x and the strategy pair $(\boldsymbol{\gamma},\boldsymbol{\delta})$, we will denote by $\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}[\cdot]$ the expectation with respect to the process law of the game trajectory $\{(X_t,U_{1,t},U_{2,t})\}_{t=1}^\infty$ generated by the above protocol.

A basic result in the theory of stochastic games is that, for every initial state x, the game has a value given by

$$V_\infty(x) = \inf_{\boldsymbol{\gamma}}\sup_{\boldsymbol{\delta}}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^T c(X_t,U_{1,t},U_{2,t})\right].$$

This corresponds to an infinite-horizon stochastic game with optimal stationary strategies for both players [4], [5]. On the other hand, we can fix a time horizon T, and consider the corresponding value

$$V_T(x) = \inf_{\boldsymbol{\gamma}}\sup_{\boldsymbol{\delta}}\frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^T c(X_t,U_{1,t},U_{2,t})\right]. \quad (1)$$

In this case, instead, the optimal strategies of both players are non-stationary. Our goal is to explore whether there exist simple regret-minimization strategies for the two players, such that, if both players honestly follow *stationary* policies derived from the prescribed strategies, their average payoffs will approach the average minimax payoff at a fast rate.

III. An auxiliary static game and its value

Our eventual goal is to prove that the minimax value $V_T(x)$ of the dynamic game described in Section II can be tightly bounded from above by the minimax value of a certain *static* game, and that there exist *stationary* strategies for both players that can achieve this bound. We begin by describing this auxiliary static (one-shot) game, which we henceforth denote by G1.

We will denote the move spaces of Players 1 and 2 by $\mathcal{S}$ and $\mathcal{T}$, respectively, where $\mathcal{S}$ is the *state-action polytope*

$$\mathcal{S} \triangleq \Big\{\mu\in\mathcal{P}(\mathsf{X}\times\mathsf{U}_1) : \sum_{x,u_1} K(y|x,u_1)\mu(x,u_1) = \sum_{u_1}\mu(y,u_1), \forall y\in\mathsf{X}\Big\}, \quad (2)$$

associated to the Markov kernel K [18], while $\mathcal{T}$ is the probability simplex $\mathcal{P}(\mathsf{U}_2)$. Every element in $\mathcal{S}$ can be decomposed in the form

$$\mu(x,u_1) = \pi_P(x)\otimes P(u_1|x), \quad x\in\mathsf{X}, u_1\in\mathsf{U}_1$$

for some randomized Markov policy $P\in\mathcal{M}(\mathsf{U}_1|\mathsf{X})$, where π_P is the invariant distribution of the Markov kernel

$$K_P(x'|x) \triangleq \sum_{u_1\in\mathsf{U}_1} K(x'|x,u_1)P(u_1|x).$$

Conversely, any element $\mu \in \mathcal{S}$ induces a Markov policy

$$P_\mu(u_1|x) \triangleq \frac{\mu(x,u_1)}{\sum_{v\in\mathsf{U}_1}\mu(x,v)}. \tag{3}$$

Moreover, $\mathcal{S}$ is a closed convex subset of $\mathcal{P}(\mathsf{X}\times\mathsf{U}_1)$, whose extreme points are deterministic Markov policies. It plays a prominent role in the so-called *convex-analytic approach* to MDPs [19]. Specifically, consider an infinite-horizon MDP with state space X, action space U_1, and one-step cost $f : \mathsf{X}\times\mathsf{U}_1 \to \mathbb{R}$. Then it can be shown that, under mild ergodicity conditions [19]–[21], the optimal long-term average cost

$$v^f_\infty(x) \triangleq \inf_{P\in\mathcal{M}(\mathsf{U}_1|\mathsf{X})} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^P_x\left[\sum_{t=1}^T f(X_t,U_t)\right]$$

is equal to the value of the static optimization problem

$$\min_\mu \langle \mu, f\rangle \qquad \text{subject to } \mu\in\mathcal{S} \tag{4}$$

regardless of the initial condition $X_1 = x$. Thus, any $\mu^* \in \mathcal{S}$ that attains the optimum in (4) induces an optimal deterministic policy $P^* = P_{\mu^*}$ via (3).

We now return to our static game G1 with the payoff

$$G(\mu,\nu) \triangleq \langle \mu\otimes\nu, c\rangle, \qquad \forall \mu\in\mathcal{S}, \nu\in\mathcal{T}.$$

Since both $\mathcal{S}$ and $\mathcal{T}$ are compact convex subsets of finite-dimensional vector spaces, and the payoff $G(\mu,\nu)$ is affine in both μ and ν, the game G1 has a minimax value

$$G^* = \inf_{\mu\in\mathcal{S}}\sup_{\nu\in\mathcal{T}}\langle\mu\otimes\nu,c\rangle = \sup_{\nu\in\mathcal{T}}\inf_{\mu\in\mathcal{S}}\langle\mu\otimes\nu,c\rangle.$$

In view of our discussion of the convex-analytic approach to MDPs, we can interpret this minimax value as follows. For each $\nu\in\mathcal{T}$, consider an MDP with state space X, action space U_1, and one-step cost

$$f_\nu(x,u_1) \triangleq \langle \nu, c(x,u_1,\cdot)\rangle \equiv \sum_{u_2\in\mathsf{U}_2}\nu(u_2)c(x,u_1,u_2).$$

Then $\inf_{\mu\in\mathcal{S}}\langle\mu\otimes\nu,c\rangle = v^\nu_\infty(x) \equiv v^{f_\nu}_\infty(x)$ for any initial state $X_1 = x$, and therefore $G^* = \sup_{\nu\in\mathcal{T}} v^\nu_\infty(x)$.

A. Approaching G^* using regret minimization

We now show that Player 1 and Player 2 can approach the minimax value G^* in repeated play of G1, where we do not require the players to know the overall payoff structure of the game, i.e., the cost function c. The repeated game takes place as follows. At each time step $t\in\{1,\dots,T\}$, Player 1 selects an occupation measure $\mu_t\in\mathcal{S}$, while Player 2 selects a mixed strategy $\nu_t\in\mathcal{T}$. The pair (μ_t,ν_t) determines the functions $f_t : \mathsf{X}\times\mathsf{U}_1\to[0,1]$ and $g_t:\mathsf{U}_2\to[0,1]$ via

$$f_t(x,u_1) \triangleq \sum_{u_2\in\mathsf{U}_2}\nu_t(u_2)c(x,u_1,u_2),$$
$$g_t(u_2) \triangleq \sum_{x\in\mathsf{X}}\sum_{u_1\in\mathsf{U}_1}\mu_t(x,u_1)c(x,u_1,u_2).$$

Player 1 then observes f_t and incurs the cost $\langle\mu_t,f_t\rangle$, while Player 2 observes g_t and incurs the cost $-\langle\nu_t,g_t\rangle$. Under this observation structure, Player 1 may know f_t, but not the mixed strategy ν_t of Player 2, the overall cost structure c, or the size of U_2; similarly, even though Player 2 knows g_t, he may not know μ_t, c, or the size of U_1. However, from the definitions of f_t and g_t, it follows that

$$\langle\mu_t,f_t\rangle = \langle\nu_t,g_t\rangle = \langle\mu_t\otimes\nu_t,c\rangle = G(\mu_t,\nu_t), \tag{5}$$

which means that both players have enough knowledge to compute their one-step costs.

Once the problem is reduced to an online linear optimization, we let both players adopt regret minimization strategies, and look at their online learning regrets:

$$\sum_{t=1}^T\langle\mu_t,f_t\rangle - \inf_{\mu\in\mathcal{S}}\sum_{t=1}^T\langle\mu,f_t\rangle \le R^1(f_{1:T}) \tag{6}$$
$$\sum_{t=1}^T(-\langle\nu_t,g_t\rangle) - \inf_{\nu\in\mathcal{T}}\sum_{t=1}^T(-\langle\nu,g_t\rangle) \le R^2(g_{1:T}). \tag{7}$$

The term "regret" is motivated by the observation that the quantities on the left-hand sides of Eqs. (6) and (7) are the differences between the cumulative cost incurred by each player during game play and the cumulative cost of the best stationary strategy in hindsight. Here, we assume, for instance, that Player 1 produces a sequence $\mu_1,\dots,\mu_T\in\mathcal{S}$ in response to the observed expected payoffs $f_1,\dots,f_T$, and his regret is upper-bounded by $R^1(f_1,\dots,f_T)$ given Player 2's choice of actions. Similarly, Player 2 produces a sequence $\nu_1,\dots,\nu_T\in\mathcal{T}$ in response to the observed expected payoffs $g_1,\dots,g_T$, and his regret is denoted by $R^2(g_1,\dots,g_T)$ given Player 1's actions. Note that there is no Markov chain involved, and the strategies of both players depend only on the previous moves of their opponents. This online linear optimization problem refers to the online learning (steady-state) component of the game. Since this is a standard online learning problem, we know that there exist numerous regret minimization strategies for both players such that R^1 and R^2 are sublinear in the time horizon T [3].

We denote the averages of the sequences of the two players' actions by $\bar\mu_T = \frac{1}{T}\sum_{t=1}^T\mu_t$ and $\bar\nu_T = \frac{1}{T}\sum_{t=1}^T\nu_t$. Since the sets $\mathcal{S}$ and $\mathcal{T}$ are convex, $\bar\mu_T\in\mathcal{S}$ and $\bar\nu_T\in\mathcal{T}$. The following proposition will be exploited in the proofs of the next section:

Proposition 1. *Suppose both players adopt arbitrary regret minimization strategies in repeated play of G1. Let $\{\mu_t\}_{t=1}^T$ and $\{\nu_t\}_{t=1}^T$ denote the sequences of the players' choices, and let $\{f_t\}_{t=1}^T$ and $\{g_t\}_{t=1}^T$ denote the resulting sequences of observed payoff functions. Then*

$$\sup_{\nu\in\mathcal{T}} G(\bar\mu_T,\nu) - G^* \le \frac{R^1(f_{1:T})+R^2(g_{1:T})}{T}. \tag{8}$$

Proof. Adding up the regrets of the two players and using Eq. (5), we have

$$\sup_{\nu\in\mathcal{T}}\frac{1}{T}\sum_{t=1}^T\langle\mu_t\otimes\nu,c\rangle - \inf_{\mu\in\mathcal{S}}\frac{1}{T}\sum_{t=1}^T\langle\mu\otimes\nu_t,c\rangle$$
$$\le \frac{R^1(f_{1:T})+R^2(g_{1:T})}{T}. \tag{9}$$

Using Eq. (5) and linearity, we have

$$\sup_{\nu \in \mathcal{T}} \frac{1}{T} \sum_{t=1}^{T} \langle \mu_t \otimes \nu, c \rangle = \sup_{\nu \in \mathcal{T}} G(\bar{\mu}_T, \nu)$$

and

$$\begin{aligned} \inf_{\mu \in \mathcal{S}} \frac{1}{T} \sum_{t=1}^{T} \langle \mu \otimes \nu_t, c \rangle &= \inf_{\mu \in \mathcal{S}} G(\mu, \bar{\nu}_T) \\ &\le \sup_{\nu \in \mathcal{T}} \inf_{\mu \in \mathcal{S}} G(\mu, \nu) \\ &= \inf_{\mu \in \mathcal{S}} \sup_{\nu \in \mathcal{T}} G(\mu, \nu), \end{aligned}$$

where the last step is by the von Neumann minimax theorem. Using these facts in (9), we get (8). □

B. Incorporating prediction models

When both players use regret minimization algorithms to approach G^*, one can get performance guarantees (i.e., upper bounds on R^1 and R^2) that are uniform in all possible sequences $\{f_t\}$ and $\{g_t\}$, including the worst-case scenario. However, these regret bounds are often conservative. More optimistic results are desirable when both players have some side information about each other's strategies. For example, if they know their opponent's choices exhibit some form of "regularity," they may incorporate this information into their decision loops. This idea of *regret minimization with predictable sequences* was introduced recently by Rakhlin and Sridharan [17], [22], who have shown that prediction models can be used by two players in a finite zero-sum matrix game to converge to the minimax equilibrium at fast rates.

In this section, we follow the lead of [17] and introduce prediction models of the players' strategies. Formally, we assume that, at each time t, Player 1 and Player 2 construct history-dependent estimates of each other's next move by $\widehat{f_t} = M_t(f_1, \dots, f_{t-1})$ and $\widehat{g}_t = N_t(g_1, \dots, g_{t-1})$. Thus, $\widehat{f_t}$ is Player 1's estimate (or prediction) of f_t based on the past revealed realizations $f_1, \dots, f_{t-1}$, and similar considerations apply to Player 2. We refer to $\{M_t\}_{t=1}^T$ and $\{N_t\}_{t=1}^T$ as the *prediction models* of Player 1 and Player 2, respectively.

At each time t, Player 1 selects

$$\mu_t = \arg\min_{\mu \in \mathcal{S}} \left\{ \left\langle \mu, \frac{1}{\eta^1} \sum_{s=1}^{t-1} f_s + \frac{1}{\eta^1} \widehat{f_t} \right\rangle + \Phi(\mu) \right\}, \quad (10)$$

where $\eta^1 > 0$ is a tunable learning rate of Player 1, and where $\Phi(\mu)$ is the relative entropy regularization term $D(\mu \| \mu^\circ)$. Here, $\mu^\circ \in \mathcal{P}(\mathsf{X} \times \mathsf{U}_1)$ is the uniform measure over all state-action pairs: $\nu^\circ(\cdot, \cdot) = 1/|\mathsf{X} \times \mathsf{U}_1|$. Similarly, at time t, Player 2 selects

$$\nu_t = \arg\min_{\nu \in \mathcal{T}} \left\{ \left\langle \nu, \frac{1}{\eta^2} \sum_{s=1}^{t-1} g_s + \frac{1}{\eta^2} \widehat{g}_t \right\rangle + \Psi(\nu) \right\}, \quad (11)$$

where $\eta^2 > 0$ is Player 2's learning rate, and where $\Psi(\nu)$ is the relative entropy regularization term $D(\nu \| \nu^\circ)$ with $\nu^\circ \in \mathcal{T}$ denoting the uniform measure over the action space U_2.

We can now state the regret bound for the above algorithms (proofs are omitted due to space limitations):

Theorem 1. *For the repeated play of G1, the proposed algorithm* (10) *for Player 1 attains the regret*

$$R^1(f_{1:T}) \le \sum_{t=1}^{T} \frac{1}{\eta^1} \| f_t - \widehat{f}_t \|_\infty^2 + \eta^1 \log |\mathsf{X} \times \mathsf{U}_1|. \quad (12)$$

Similarly, the algorithm (11) *for Player 2 attains the regret*

$$R^2(g_{1:T}) \le \sum_{t=1}^{T} \frac{1}{\eta^2} \| g_t - \widehat{g}_t \|_\infty^2 + \eta^2 \log |\mathsf{U}_2| \quad (13)$$

If we cannot assume the two players follow the algorithms honestly, each player can set his learning rate conservatively. In particular, since $\|f_t\|_\infty \le 1$ for all t, we can assume without loss of generality that $\|\widehat{f_t}\|_\infty \le 1$ as well. This assumption gives the usual worst-case regret bound

$$R^1(f_{1:T}) \le \frac{4T}{\eta^1} + \eta^1 \log |\mathsf{X}_1 \times \mathsf{U}|,$$

which can be optimized w.r.t. the learning rate η^1 to yield the usual $O(\sqrt{T \log |\mathsf{X}_1 \times \mathsf{U}|})$ regret bound. However, if the players are honest and do cooperate, the learning rates η^1 and η^2 can be tuned adaptively using prior knowledge. For instance, if Player 1 has prior knowledge about the term $\sum_{t=1}^T \|f_t - \widehat{f_t}\|_\infty$, he can optimize R^1 by choosing $\eta^1 = \sqrt{\sum_{t=1}^T \frac{\|f_t - \widehat{f}_t\|_\infty^2}{\log |\mathsf{X} \times \mathsf{U}_1|}}$. This leads to the regret bound

$$R^1(f_{1:T}) \le 2 \sqrt{\log |\mathsf{X} \times \mathsf{U}_1| \sum_{t=1}^{T} \|f_t - \widehat{f}_t\|_\infty^2}.$$

For example, if Player 1's prediction of f_t is simply the previously revealed f_{t-1}, i.e., $\widehat{f_t} = M_t(f_1, \dots, f_{t-1}) = f_{t-1}$, then the bound becomes

$$R^1(f_{1:T}) \le 2 \sqrt{\log |\mathsf{X} \times \mathsf{U}_1| \sum_{t=1}^{T} \|f_t - f_{t-1}\|_\infty^2},$$

which is known in the literature as a path-length bound [23], [24]. In situations where Player 2 gradually changes his moves, i.e., when the previous move of Player 2 is a good proxy for his next move, such bounds can be tighter than the worst-case pessimistic $O(\sqrt{T})$ bound.

If both players adopt the algorithms of Eqs. (10) and (11), then R^1 and R^2 diminish as the prediction models become more accurate. As we show in the next section, smaller R^1 and R^2 make the average payoffs of the players converge faster to the minimax payoff in dynamic stochastic games.

IV. Regret minimization in stochastic games

In the preceding section, we have analyzed a certain static (one-shot) game G1 and showed that the two players can approach its minimax payoff using regret-minimization algorithms in repeated play of G1. This naturally leads to the following question: *How can we relate the static game G1 to the original dynamic game that has multiple rounds?* In this section, we will answer this question by relating the minimax value G^* to V_T and the quantity $\sup_{\nu \in \mathcal{T}} G(\bar{\mu}_T, \nu)$

to the actual realized payoff of the dynamic game. We will then propose certain stationary strategies for the two players in a single-controller zero-sum stochastic games and quantify the gap between the resulting expected payoff and V_T.

A. Stationary strategies for finite-horizon stochastic games

We first relate the static game G1 to the dynamic game with the value given by Eq. (1):

Lemma 1. *Suppose both players adopt regret minimization strategies in repeated play of G1. Let $\{\mu_t\}_{t=1}^T$ and $\{f_t\}_{t=1}^T$ be the resulting sequences of the players' choices. If Player 1 uses the average occupation measure $\bar{\mu}_T$ in G1, then, regardless of Player 2's choice, Player 1's payoff approaches the minimax payoff of the T-step finite-horizon stochastic game starting at any initial state x. Specifically, let $R^1(f_{1:T})$ and $R^2(g_{1:T})$ be the online learning regrets of the corresponding regret minimization algorithms used in repeated play of G1. Then, there exists some constant C, such that*

$$\sup_{\nu\in\mathcal{T}} G(\bar{\mu}_T,\nu) \le V_T(x) + \frac{C}{T} + \frac{R^1(f_{1:T})+R^2(g_{1:T})}{T}.$$

Proof. Let $\Delta_{\rm iid}$ denote the subset of behavioral strategies of Player 2, where the actions $U_{2,t}$ are drawn i.i.d. from some fixed distribution in $\mathcal{T}$, regardless of the state and of Player 1's actions. Thus, there is an obvious one-to-one correspondence between $\Delta_{\rm iid}$ and $\mathcal{T}$, so we will use the notation $\mathbb{E}_x^{\boldsymbol{\gamma},\nu}[\cdot]$ for $\boldsymbol{\delta}\in\Delta_{\rm iid}$ induced by $\nu\in\mathcal{T}$. Then

$$\begin{aligned}
V_T(x) &= \inf_{\boldsymbol{\gamma}}\sup_{\boldsymbol{\delta}} \frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^T c(X_t,U_{1,t},U_{2,t})\right]\\
&\ge \inf_{\boldsymbol{\gamma}}\sup_{\boldsymbol{\delta}\in\Delta_{\rm iid}} \frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^T c(X_t,U_{1,t},U_{2,t})\right]\\
&= \inf_{\boldsymbol{\gamma}}\sup_{\nu\in\mathcal{T}} \frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\nu}\left[\sum_{t=1}^T f_\nu(X_t,U_{1,t})\right]\\
&\ge \sup_{\nu\in\mathcal{T}}\inf_{\boldsymbol{\gamma}} \frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\nu}\left[\sum_{t=1}^T f_\nu(X_t,U_{1,t})\right],
\end{aligned}$$

where

$$f_\nu(x,u_1) \triangleq \sum_{u_2\in\mathsf{U}_2}\nu(u_2)c(x,u_1,u_2).$$

For a fixed choice of $\nu\in\mathcal{T}$, Player 1 faces a T-step MDP with one-step state-action cost f_ν. Let us denote the optimal payoff of that MDP starting from initial state $x\in\mathsf{X}$ by $v_T^\nu(x) = \inf_{\boldsymbol{\gamma}}(1/T)\mathbb{E}_x^{\boldsymbol{\gamma},\nu}\left[\sum_{t=1}^T f_\nu(X_t,U_{1,t})\right]$ and denote $v_\infty^\nu(x) \triangleq \lim_{T\to\infty} v_T^\nu(x)$. When ν is a Dirac measure centered at some $u_2\in\mathsf{U}_2$, we will write $v_T^{u_2}$ and $v_\infty^{u_2}$ instead.

By [2, Prop. 5.21], $T\|v_T^\nu - v_\infty^\nu\|_\infty$ is uniformly bounded: for each $\nu\in\mathcal{T}$, there exists a constant $C(\nu)$, such that

$$\|v_T^\nu - v_\infty^\nu\|_\infty \le \frac{C(\nu)}{T}.$$

Since the functional $\nu\mapsto\|v_T^\nu - v_\infty^\nu\|_\infty$ is convex, we have

$$\sup_{\nu\in\mathcal{T}}\|v_T^\nu - v_\infty^\nu\|_\infty = \max_{u_2\in\mathsf{U}_2}\|v_T^{u_2} - v_\infty^{u_2}\|_\infty \le \frac{C}{T},$$

where $C \triangleq \max_{u_2\in\mathsf{U}_2} C(u_2)$ is some finite constant. Moreover, from the convex-analytic approach to MDP's [25], we know that $v_\infty^\nu(x) = \inf_{\mu\in\mathcal{S}}\langle\mu,f_\nu\rangle = \inf_{\mu\in\mathcal{S}}\langle\mu\otimes\nu,c\rangle$.

Consequently,

$$\begin{aligned}
V_T(x) &\ge \sup_{\nu\in\mathcal{T}}\inf_{\boldsymbol{\gamma}}\frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\nu}\left[\sum_{t=1}^T f_\nu(X_t,U_{1,t})\right]\\
&= \sup_{\nu\in\mathcal{T}} v_T^\nu(x)\\
&\ge \sup_{\nu\in\mathcal{T}}\left\{\inf_{\mu\in\mathcal{S}}\langle\mu\otimes\nu,c\rangle - \frac{C}{T}\right\}\\
&= \inf_{\mu\in\mathcal{S}}\sup_{\nu\in\mathcal{T}}\langle\mu\otimes\nu,c\rangle - \frac{C}{T},
\end{aligned}$$

where the last step is by the von Neumann minimax theorem. Combining this with (8), we complete the proof. □

Next we relate the quantity $\sup_{\nu\in\mathcal{T}}\langle\bar{\mu}_T\otimes\nu,c\rangle$ to the payoff of several different dynamic games. We start with the repeated play of G1. Recall that Player 1 uses a regret-minimization algorithm to produce a sequence of T occupation measures on the space of state-action pairs $\mathsf{X}\times\mathsf{U}_1$, and computes the average of these measures $\bar{\mu}_T$. Player 2 also adopts a regret minimization algorithm to generate a sequence of $\{\nu_t\}_{t=1}^T$ and computes the average $\bar{\nu}_T$. Now the two players can use these objects to construct their Markov randomized stationary strategies $\boldsymbol{\gamma}=(\gamma,\gamma,\ldots)$ and $\boldsymbol{\delta}=(\delta,\delta,\ldots)$, where

$$\gamma(x) = P_{\bar{\mu}_T}(\cdot|x) \qquad \text{and} \qquad \delta(x) = \bar{\nu}_T. \tag{14}$$

Here, $P_{\bar{\mu}_T}$ is the Markov policy induced by the occupation measure $\bar{\mu}_T$. Note, by the way, that Player 2's strategy ignores the state variable. Now, we consider a T-round stochastic game, throughout which Player 1 uses the stationary strategy $\boldsymbol{\gamma}$ and Player 2 uses $\boldsymbol{\delta}$.

We impose the following *uniform mixing condition* [8], [9], [14]: There exists a finite constant $\tau>0$ such that for all $P\in\mathcal{M}(\mathsf{U}_1|\mathsf{X})$ and $\mu_1,\mu_2\in\mathcal{P}(\mathsf{X})$,

$$\|\mu_1 K_P - \mu_2 K_P\|_1 \le e^{-1/\tau}\|\mu_1-\mu_2\|_1. \tag{15}$$

In other words, the collection of all state transition laws induced by all Markov policies P of Player 1 is *uniformly mixing*. Here we assume, without loss of generality, that $\tau\ge 1$. This uniform mixing property guarantees that every Markov policy P has a unique invariant state distribution $\pi_P\in\mathcal{P}(\mathsf{X})$, i.e., $\pi_P = \pi_P K(\cdot|P)$.

Theorem 2. *Assume the uniform mixing condition is satisfied by the controlled transition kernel K. When both players follow stationary strategies $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ derived according to* (14) *from their regret minimization strategies in G1, we have*

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^T c(X_t,U_{1,t},U_{2,t})\right] - V_T(x)\\
&\qquad\le \frac{R^1(f_{1:T})+R^2(g_{1:T})}{T} + \frac{C+2(\tau+1)}{T}.
\end{aligned}$$

Proof. Let π_t be the marginal state distribution at time t, and let $\bar{\pi}$ denote the (unique) invariant distribution of the

Markov policy $P_{\bar{\mu}_T}$. Then $\bar{\mu}_T = \bar{\pi} \otimes P_{\bar{\mu}_T}$. Denoting by $\tilde{\mu}_t$ the state-action distribution $\pi_t \otimes P_{\bar{\mu}_T}$, we have

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}_x^{\gamma,\boldsymbol{\delta}}\left[\sum_{t=1}^T c(X_t, U_{1,t}, U_{2,t})\right] = \frac{1}{T}\sum_{t=1}^T \langle \tilde{\mu}_t \otimes \bar{\nu}_T, c\rangle \\
&\le \frac{1}{T}\sum_{t=1}^T \langle \bar{\mu}_T \otimes \bar{\nu}_T, c\rangle + \frac{1}{T}\sum_{t=1}^T \|\pi_t - \bar{\pi}\|_1 \\
&\le \langle \bar{\mu}_T \otimes \bar{\nu}_T, c\rangle + \frac{1}{T}\sum_{t=1}^T 2e^{-t/\tau} \\
&\le \sup_{\nu \in \mathcal{T}} \langle \bar{\mu}_T \otimes \nu, c\rangle + \frac{2(\tau+1)}{T} \\
&\le V_T(x) + \frac{R^1(f_{1:T}) + R^2(g_{1:T}) + C + 2(\tau+1)}{T},
\end{aligned}$$

where the second inequality is by the uniform mixing condition, and the last inequality by Lemma 1. □

This result quantifies the sub-optimality of the stationary policies $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ for the T-round stochastic game. When the two players use these near-optimal stationary strategies, the gap between the actual payoff and the average minimax payoff $V_T(x)$ decays to zero as $T \to \infty$. In fact, the policy $\boldsymbol{\delta}$ is near-minimax for Player 1:

Corollary 1. *Under the same assumptions as in Theorem 2,*

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}_x^{\gamma,\boldsymbol{\delta}}\left[\sum_{t=1}^T c(X_t, U_{1,t}, U_{2,t})\right] \\
&\qquad - \sup_{\boldsymbol{\delta}} \frac{1}{T}\mathbb{E}_x^{\gamma,\boldsymbol{\delta}}\left[\sum_{t=1}^T c(X_t, U_{1,t}, U_{2,t})\right] \\
&\qquad \le \frac{R^1(f_{1:T}) + R^2(g_{1:T})}{T} + \frac{C + 2(\tau+1)}{T}.
\end{aligned}$$

V. Conclusion

In this paper, we designed simple and efficient strategies for two-player single-controller zero-sum stochastic games and quantified the gap between their expected payoff and the dynamic equilibrium value of the game. In particular, we developed stationary policies based on regret minimization strategies and quantified their rate of convergence to the minimax value of an auxiliary static game. This suggests that we can achieve sublinear convergence to the minimax payoff of the dynamic game if the associated regret minimization strategies exhibit sublinear regret. Finally, we investigated the case where both players of the game exploit side information regarding the strategies of their opponents and devised appropriate algorithms that achieve more optimistic convergence rates to the minimax payoff.

The main limitation of our results is that we restrict Player 2 to open-loop strategies, whereas even in the case when only Player 1 controls the state transitions, both players observe the state of the game and can use it to guide their actions. Imposing this restriction on Player 2 has allowed us to introduce an auxiliary static game, which was used to design the strategies for the two players in the original dynamic game. Removing this limitation is a major direction for future research.

References

[1] L. S. Shapley, "Stochastic games," *Proceedings of the National Academy of Sciences*, vol. 39, pp. 1095–1100, 1953.

[2] S. Sorin, *A first course on zero-sum repeated games*. Springer, 2002.

[3] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning and Games*. Cambridge Univ. Press, 2006.

[4] J. F. Mertens and A. Neyman, "Stochastic games," *International Journal of Game Theory*, vol. 10, pp. 53–66, 1981.

[5] T. Parthasarathy and T. Raghavan, "An order field property for stochastic games when one player controls transition probabilities," *J. Opt. Theory. Appl*, pp. 375–392, 1981.

[6] O. J. Vrieze, "Linear programming and undercounted stochastic game in which one player controls transition," *OR Spektrum*, no. 3, pp. 15–24, 1981.

[7] H. McMahan, "Planning in the presence of cost functions controlled by an adversary," *The 20th International Conference on Machine Learning*, pp. 536–543, 2003.

[8] E. Even-Dar, S. M. Kakade, and Y. Mansour, "Online Markov decision processes," *Math. Oper. Res.*, vol. 34, no. 3, pp. 726–736, 2009.

[9] J. Y. Yu, S. Mannor, and N. Shimkin, "Markov decision processes with arbitrary reward processes," *Math. Oper. Res.*, vol. 34, no. 3, pp. 737–757, 2009.

[10] G. Neu, A. György, C. Szepesvári, and A. Antos, "Online Markov decision processes under bandit feedback," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010, pp. 1804–1812.

[11] R. Arora, O. Dekel, and A. Tewari, "Deterministic MDPs with adversarial rewards and bandit feedback," in *Proceedings of the 28th Annual Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 2012, pp. 93–101.

[12] P. Guan, M. Raginsky, and R. Willett, "Online Markov decision processes with Kullback-Leibler control cost," *IEEE Trans. Automat. Control*, vol. 59, no. 6, pp. 1423–1438, 2014.

[13] Y. Abbasi-Yadkori, P. L. Bartlett, and C. Szepesvári, "Online learning in Markov decision processes with adversarially chosen transition probability distributions," *http://arxiv.org/1303.3055*, 2013.

[14] T. Dick, A. György, and C. Szepesvári, "Online learning in Markov decision processes with changing cost sequences," *ICML*, 2014.

[15] P. Guan, M. Raginsky, and R. Willett, "From minimax value to low-regret algorithms for online Markov decision processes," *In Proceedings of American Control Conference*, 2014.

[16] C. Daskalakis, A. Deckelbaum, and A. Kim, "Near-optimal no-regret algorithms for zero-sum games," *In Proceedings of the 22nd Annual ACM-SIAM symposium on Discrete Algorithms*, pp. 235–254, 2011.

[17] A. Rakhlin and K. Sridharan, "Optimization, learning, and games with predictable sequences," *Adv. Neural Inform. Processing Systems*, 2013.

[18] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

[19] V. S. Borkar, "Convex analytic methods in Markov decision processes," in *Handbook of Markov decision processes*. Kluwer Academic Publishers, 2002.

[20] A. S. Manne, "Linear programming and sequential decisions," *Management Science*, vol. 6, no. 3, pp. 259–267, 1960.

[21] S. Meyn, *Control techniques for complex networks*. Cambrdige Univ. Press, 2008.

[22] A. Rakhlin and K. Sridharan, "Online learning with predictable sequences," *In proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013.

[23] C. K. Chiang, T. Yang, C. J. Lee, M. Mahdavi, C. J. Lu, R. Jin, and S. Zhu, "Online optimization with gradual variations," *In proceedings of the 25nd Annual Conference on Learning Theory (COLT)*, 2012.

[24] A. Rakhlin, K. Sridharan, and A. Tewari, "Online learning: Stochastic, constrained, and smoothed adversaries," *Adv. Neural Inform. Processing Systems*, 2011.

[25] V. S. Borkar, "A convex analytic approach to Markov decision processes," *Probab. Th. Rel. Fields*, vol. 78, pp. 583–602, 1988.