Quantifying the stochasticity of policy parameters in reinforcement learning problems

Vahe Galstyan[1,2] and David B. Saakian[2,*]
[1]*AMOLF, Science Park 104, 1098 XG Amsterdam, Netherlands*
[2]*A.I. Alikhanyan National Science Laboratory (Yerevan Physics Institute) Foundation, 2 Alikhanian Brothers Street, Yerevan 375036, Armenia*

(Received 7 August 2022; revised 3 January 2023; accepted 16 February 2023; published 8 March 2023)

The stochastic dynamics of reinforcement learning is studied using a master equation formalism. We consider two different problems—Q learning for a two-agent game and the multiarmed bandit problem with policy gradient as the learning method. The master equation is constructed by introducing a probability distribution over continuous policy parameters or over both continuous policy parameters and discrete state variables (a more advanced case). We use a version of the moment closure approximation to solve for the stochastic dynamics of the models. Our method gives accurate estimates for the mean and the (co)variance of policy variables. For the case of the two-agent game, we find that the variance terms are finite at steady state and derive a system of algebraic equations for computing them directly.

DOI: 10.1103/PhysRevE.107.034112

I. INTRODUCTION

The emergence of complex cognition requires the development of ways to gather and respond to information from a noisy environment [1]. Here we consider this problem in the framework of reinforcement learning [2], with a focus on the statistical physics aspects [3–14]. In this framework, learning agents interact with the environment by taking actions and receiving rewards from the environment that weight the relative success of the actions taken. Agents learn from these interactions by updating their strategy of making actions (policy) in such a way that maximizes the reward received from the environment in the long run.

Due to the stochasticity in the received environmental cues, the process of policy optimization will also involve fluctuations. While the dynamics of average policy parameters has been the subject of a number of previous studies [3,5,8], our understanding of the stochastic aspects of reinforcement learning is far from being complete. In our work, we study the stochasticity of learning dynamics on an example of two different problems using a master equation formalism [15], having as the main goal the accurate estimation of fluctuations in policy parameters.

In both cases, probabilities of the agent's actions are specified by a vector, called Q vector, that is updated at every learning step. The master equation is constructed for the probability distribution $P(\mathbf{q}, t)$ of Q values ($\mathbf{q}$ from now on). Our aim is to calculate the covariance matrix of this distribution which will give a notion about the likelihood of whether the learning algorithm has properly converged or not.

The paper is organized as follows. In Sec. II, we introduce and study the first problem—Q learning in the context of a two-player two-action game. Similar learning schemes were considered in Refs. [3,4,8] where dynamic equations for the policy were written and analytic results for average values were derived. Here, we use the master equation method to estimate the noise in policy over time. We also derive a system of equations for directly computing the covariance terms of Q values at steady state. Then, in Sec. III we apply the master equation formalism to the K-armed bandit problem where learning dynamics follows the policy gradient algorithm [10,11]. In this setting, the rewards received from the environment are continuously distributed random variables, in contrast to those in the first problem. Applying a proper version of the moment-closure approximation [16–18], we obtain an iterative analytical method for estimating the variance of policy values over time and verify its predictions against simulations. We discuss our results and their implications for the field of reinforcement learning and beyond in Sec. IV.

II. Q LEARNING IN A TWO-AGENT GAME

A. The model

The schematic of the two-agent game studied in our work is shown in Fig. 1(a). Each player receives a reward from the environment after taking an action. The reward of a player not only depends on the action he takes, but also on the other player's action at the earlier time step. We have chosen this version of reward allocation to get an advanced master equation where policy update depends on the previous state (discrete variable). If the reward depends only on the current choice of actions, then the master equation is written for continuous Q values alone, resulting in rather simple update equations (see Appendix A for the discuss of this scenario).

The reward amounts in the Fig. 1(a) setting are specified through payoff matrices A and B for the two players. The matrix element A_{ln} represents the reward that the first player receives after taking action l, provided the second player performed action n at the earlier step. Elements of the matrix B are defined similarly for the second player.

*saakian@yerphi.am

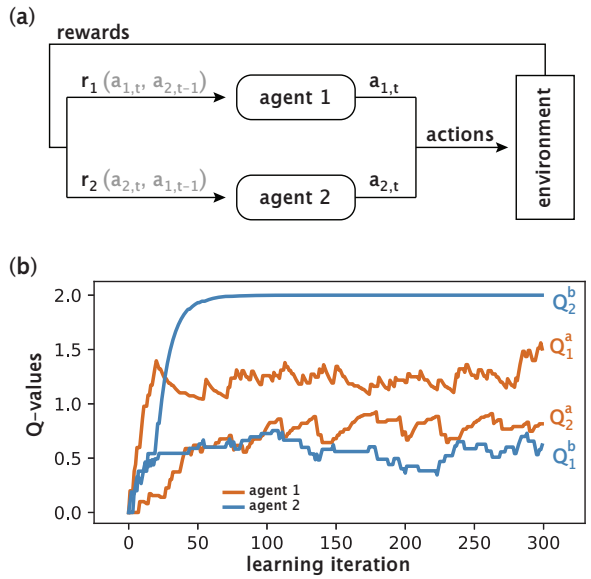

FIG. 1. Reinforcement learning for a two-agent game. (a) Schematic of the learning algorithm. (b) Example stochastic dynamics of Q values. Parameters used in the study were $\alpha = 0.1$, $T = 1.0$, $A = [(2, 1),\ (0, 1)]$, and $B = [(1, 0),\ (2, 2)]$.

Players choose their actions probabilistically based on the Q values. Specifically, actions l and n are taken by the first and second players with probabilities

$$x_l^a = \frac{\exp\left(Q_l^a/T\right)}{\sum_{l'}\exp\left(Q_{l'}^a/T\right)}, \tag{1a}$$

$$x_n^b = \frac{\exp\left(Q_n^b/T\right)}{\sum_{n'}\exp\left(Q_{n'}^b/T\right)}, \tag{1b}$$

respectively, where summation is performed over the two possible actions of each player ($l \in \{1, 2\}$, $n \in \{1, 2\}$). Note that here the Q values for one player depend only on the actions he can take. In Appendix B, we demonstrate how our method is generalized to the case where the policy parameters also depend on the actions of the other player (i.e., where we have $Q_{l,n'}^a$ and $Q_{n,l'}^b$).

After taking an action and receiving a reward, each player updates the Q value corresponding to the action taken. If the first and second players take actions l and n, respectively, then the corresponding Q values are updated via

$$Q_l^a(t+\alpha) = Q_l^a(t) + \alpha\left[A_{l\hat{n}} - Q_l^a(t)\right], \tag{2a}$$

$$Q_n^b(t+\alpha) = Q_n^b(t) + \alpha\left[B_{n\hat{l}} - Q_n^b(t)\right], \tag{2b}$$

where $\hat{l}$ and $\hat{n}$ are the actions taken by the first and second players at the previous time step. The parameter $\alpha \in [0, 1]$ here represents the learning rate that defines the extent to which new information (received reward) contributes to the updated policy.

An example learning dynamics of Q values is shown in Fig. 1(b), where the stochastic nature of the process can be observed. Analysis of the model with differential equations performed in Ref. [8] yields only the dynamics of the mean values and provides no information about the inherent stochasticity, motivating the development of the master equation formulation of the problem.

B. Master equation formulation and stochastic dynamics of learning

Our goal is to approximate the probability distribution $P_t(\mathbf{q})$ over the course of the learning dynamics. We first introduce several convenient notations, namely,

$$x_1 = x_1^a, \quad x_2 = x_2^a, \quad x_3 = x_1^b, \quad x_4 = x_2^b, \tag{3}$$

$$X_1 = x_1^a x_1^b, \quad X_2 = x_1^a x_2^b, \quad X_3 = x_2^a x_1^b, \quad X_4 = x_2^a x_2^b, \tag{4}$$

where X_s stands for the probability of the joint action $s \in \{1, 2, 3, 4\}$. For compactness of equations, we also introduce a four-element vector $\mathbf{d}_{s\hat{s}}$, the kth element of which represents the difference between the received reward and Q_k if the joint action s together with the previous joint action $\hat{s}$ lead to a change in Q_k; otherwise, the element is zero. For example, if $s = 1$ (a:1, b:1) and $\hat{s} = 3$ (a:2, b:1), then the changing Q values are Q_1^a and Q_1^b and the corresponding vector is $\mathbf{d}_{13} = \{A_{11} - Q_1^a,\ 0,\ B_{12} - Q_1^b,\ 0\}$.

We write a master equation for the joint probability distribution of Q values and joint actions as

$$P_{t+1}(\mathbf{Q}, s) = \int d\mathbf{q}\, X_s(\mathbf{q}) \sum_{\hat{s}=1}^{4} \delta(-\mathbf{Q} + \mathbf{q} + \alpha\mathbf{d}_{s\hat{s}})\, P_t(\mathbf{q}, \hat{s}). \tag{5}$$

Here, $P_{t+1}(\mathbf{Q}, s)$ is the probability that the joint action with index s is taken and the Q values are updated to $\mathbf{Q}$ at time $t + 1$. The joint action s is taken based on the current Q values given by $\mathbf{q}$; hence, the term $X_s(\mathbf{q})$ in the equation. The updated vector $\mathbf{Q}$ is specified by the vector $\mathbf{q}$ in the earlier time step and the received rewards that depend on the joint action taken earlier ($\hat{s}$). Integrating by $\mathbf{Q}$ and summing over s yields 1 on both sides, verifying the consistency of the master equation.

Next, we introduce summary metrics $E[Q_k|s]$ and $E[Q_kQ_l|s]$ to stand for the expectation of Q values and their pairwise products, conditional on the joint action taken being s. By definition, these metrics are

$$\begin{aligned} E[Q_k|s] &= \int d\mathbf{Q}\, Q_k P_{t+1}(\mathbf{Q}|s) \\ &= \frac{1}{P_{t+1}(s)} \int d\mathbf{q}\, X_s(\mathbf{q}) \sum_{\hat{s}} \left(q_k + \alpha d_{s\hat{s}}^{(k)}\right) P_t(\mathbf{q}|\hat{s}) P_t(\hat{s}), \end{aligned} \tag{6}$$

$$\begin{aligned} E[Q_kQ_l|s] &= \int d\mathbf{Q}\, Q_kQ_l P_{t+1}(\mathbf{Q}|s) \\ &= \frac{1}{P_{t+1}(s)} \int d\mathbf{q}\, X_s(\mathbf{q}) \\ &\quad \times \sum_{\hat{s}} \left(q_k + \alpha d_{s\hat{s}}^{(k)}\right)\left(q_l + \alpha d_{s\hat{s}}^{(l)}\right) P_t(\mathbf{q}|\hat{s}) P_t(\hat{s}), \end{aligned} \tag{7}$$

where $P_{t+1}(s) = \int d\mathbf{Q}\, P_{t+1}(\mathbf{Q}, s)$ is the probability of taking the joint action s, $P_t(\hat{s})$ is the same probability at the earlier time step, and $d_{s\hat{s}}^{(k)}$ is the kth component of the vector $\mathbf{d}_{s\hat{s}}$.

We want to estimate $E[Q_k|s]$ and $E[Q_kQ_l|s]$ together with $P_{t+1}(s)$ using their values in the previous iteration. To that end, we introduce the covariance matrix of $\mathbf{q}$, conditional on the joint action, namely,

$$v_{mn}^{(\hat{s})} = E[q_m|\hat{s}]\,E[q_n|\hat{s}] - E[q_mq_n|\hat{s}]. \tag{8}$$

Using it, we approximate integrals involving the probability distribution $P_t(\mathbf{q}|\hat{s})$ as

$$\int d\mathbf{q}\, f(\mathbf{q})P_t(\mathbf{q}|\hat{s}) \approx f(E[\mathbf{q}|\hat{s}]) + \frac{1}{2}\sum_{mn} f''_{mn}(E[\mathbf{q}|\hat{s}])v_{mn}^{(\hat{s})}. \tag{9}$$

Here $f(\mathbf{q})$ is any smooth function of $\mathbf{q}$. When writing this approximation, we are assuming that the third moments are much lower (by a factor of α) than the variance.

Applying Eq. (9) to Eqs. (6) and (7), as well as to the definition of $P_{t+1}(s)$, we obtain the update equations as

$$\begin{aligned} E[Q_k|s]P_{t+1}(s) \approx& \sum_{\hat{s}} \hat{X}_s\big(\hat{q}_k + \alpha d_{s\hat{s}}^{(k)}\big)P_t(\hat{s}) \\ &+ \frac{1}{2}\sum_{\hat{s}}\sum_{mn} \hat{X}''_{s;mn}\big(\hat{q}_k + \alpha d_{s\hat{s}}^{(k)}\big)v_{mn}^{(\hat{s})}\,P_t(\hat{s}) \\ &+ \sum_{\hat{s}}\sum_{m} \hat{X}'_{s;m}(1-\alpha I_{sk})\,v_{mk}^{(\hat{s})}\,P_t(\hat{s}), \end{aligned} \tag{10}$$

$$\begin{aligned} &E[Q_kQ_l|s]P_{t+1}(s) \\ &\approx \sum_{\hat{s}} \hat{X}_s\big(\hat{q}_k + \alpha d_{s\hat{s}}^{(k)}\big)\big(\hat{q}_l + \alpha d_{s\hat{s}}^{(l)}\big)P_t(\hat{s}) \\ &\quad + \frac{1}{2}\sum_{\hat{s}}\sum_{mn} \hat{X}''_{s;mn}\big(\hat{q}_k + \alpha d_{s\hat{s}}^{(k)}\big)\big(\hat{q}_l + \alpha d_{s\hat{s}}^{(l)}\big)v_{mn}^{(\hat{s})}P_t(\hat{s}) \\ &\quad + \sum_{\hat{s}} \hat{X}_s(1-\alpha I_{sk})(1-\alpha I_{sl})v_{kl}^{(\hat{s})}P_t(\hat{s}) \\ &\quad + \sum_{\hat{s}}\sum_{m} \hat{X}'_{s;m}(1-\alpha I_{sk})\big(\hat{q}_l + \alpha d_{s\hat{s}}^{(l)}\big)v_{mk}^{(\hat{s})}P_t(\hat{s}) \\ &\quad + \sum_{\hat{s}}\sum_{m} \hat{X}'_{s;m}\big(\hat{q}_k + \alpha d_{s\hat{s}}^{(k)}\big)(1-\alpha I_{sl})v_{ml}^{(\hat{s})}P_t(\hat{s}), \end{aligned} \tag{11}$$

$$P_{t+1}(s) \approx \sum_{\hat{s}} \hat{X}_sP_t(\hat{s}) + \frac{1}{2}\sum_{\hat{s}}\sum_{mn} \hat{X}''_{s;mn}v_{mn}^{(\hat{s})}P_t(\hat{s}), \tag{12}$$

where $\hat{q}_k \equiv E[q_k|\hat{s}]$, $\hat{X}_s \equiv X_s(\hat{\mathbf{q}})$, and $I_{sk} = 1$ if the joint action s involves a change in q_k; otherwise, $I_{sk} = 0$. These results yield the conditional expectation and covariance terms in the next learning iteration. Specifically, $V_{kl}^{(s)}$ at time $t+1$ can be calculated as

$$V_{kl}^{(s)} = E[Q_kQ_l|s] - E[Q_k|s] \times E[Q_l|s]. \tag{13}$$

We also consider the variance of Q vector components irrespective of the previous state, defined via

$$V_{kl} = E[Q_kQ_l] - E[Q_k]\,E[Q_l]. \tag{14}$$

It is expressed in terms of the conditional metrics as

$$\begin{aligned} V_{kl} &= \sum_s E[Q_kQ_l|s]P_t(s) - \sum_s E[Q_k|s]P_t(s)\sum_s E[Q_l|s]P_t(s) \\ &= E\big[V_{kl}^{(s)}\big] + \mathrm{Cov}(E[Q_k|s], E[Q_l|s]). \end{aligned} \tag{15}$$

We note two contributions to V_{kl}. The first is the state-average of conditional variance values $V_{kl}^{(s)}$ and the second is the covariance of conditional mean Q values. The dominant contribution after a short transient comes from the first term, while the second term decays quickly.

To verify the accuracy of our iterative analytical method [Eqs. (10)–(12)], we simulated the learning process multiple times and calculated the means and covariance terms from the sampled trajectories. Comparison of estimates from the simulation and our method is shown in Fig. 2. As can be seen, the dynamics of both mean Q values and the different (co)variance terms V_{kl} is captured accurately.

C. Steady-state fluctuations

It if often of practical interest to know the stationary behavior of the learning process. While the iterative approach derived in the previous section can be applied repeatedly until convergence is observed, it is more practical to have a system of algebraic equations that would directly yield fluctuations in policy at steady state. Such system of equations can be obtained by requiring the covariance terms at consecutive iterations to be equal to each other ($V_{kl} = v_{kl}$) and ignoring higher-order terms [$O(\alpha^3)$] in the learning parameter α. The steady-state solution of our equations gives

$$\begin{aligned} &(x_k + x_l)v_{kl} - x_k\sum_s\sum_m X'_{s;m}v_{ml}\,r_{ks} \\ &\quad - x_l\sum_s\sum_m X'_{s;m}v_{mk}\,r_{ls} = \alpha\chi_{kl}\sum_s (r_{ks} - \bar{q}_k)(r_{ls} - \bar{q}_l)X_s \end{aligned} \tag{16}$$

for $1 \leqslant k, l \leqslant 4$. Here, $\chi_{kl} = \sum_s X_sI_{sk}I_{sl}$, r_{ks} represents the reward associated with the change in q_k when the joint action in the previous step is s, and $\bar{q}_k$ represent the mean-field Q values at steady state. As shown in Fig. 3, predictions of covariance terms via Eq. (16) accurately match the numerical estimates from the simulated learning process. The presence of the factor of α on the right-hand side of Eq. (16) also suggest a linear scaling of the steady-state covariance terms with the learning rate, as is demonstrated numerically in the inset of Fig. 3.

III. *K*-ARMED BANDIT PROBLEM

The reinforcement learning model in Sec. II involved a discrete set of actions with deterministic rewards. In this section, we study the stochasticity of policy dynamics for the K-armed bandit problem [10,11] where the rewards are now random and sampled from a continuous distribution. We limit our discussion to single-agent learning performed via the natural policy gradient algorithm [12].

A. Two-armed case

We begin with a case where the agent performs one of two possible actions and receives an action-dependent reward sampled from a normal distribution $\mathcal{N}(r_i, s_i)$. Here, r_i and s_i are the mean and the standard division of the reward received after performing action $i \in \{1, 2\}$.

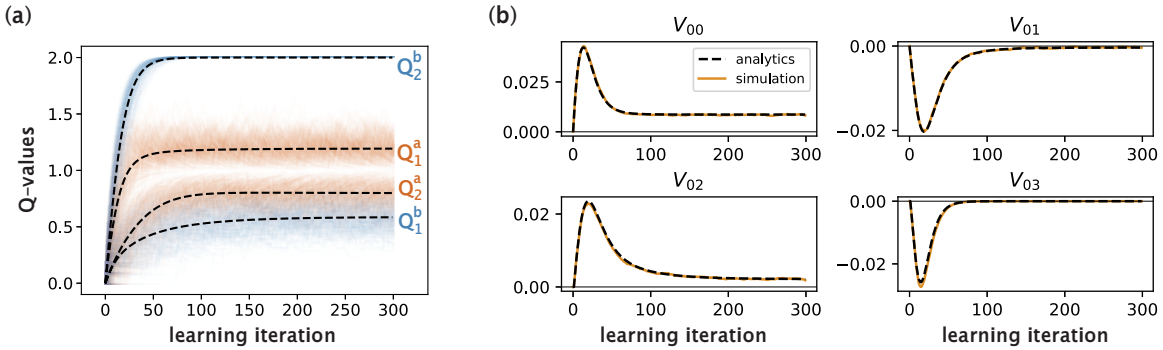

FIG. 2. Comparison of simulation results with those of the iterative analytical method. (a) Dynamics of Q values. Transparent trajectories represent independent learning simulations. Dotted lines represent estimates from the iterative analytical approach. (b) Dynamics of covariance terms (4 of the 10 independent elements are shown for clarity). Parameters used are identical to those in the caption of Fig. 1(b). Q values were initialized at $q^a_{\text{init},l} = q^b_{\text{init},n} = 0$.

Action probabilities are defined via the policy parameter Q. Specifically, action 1 is performed with probability

$$x(Q) = \frac{1}{1+e^{-Q}}. \tag{17}$$

The probability of the second action is then $1 - x(Q)$. With the natural policy gradient method, if action 1 is performed, then the Q value is updated via

$$Q = q + \frac{\alpha R_1}{F}\frac{d\ln x(q)}{dq} = q + \frac{\alpha R_1}{x(q)}, \tag{18}$$

where α is the learning rate, $R_1 \sim \mathcal{N}(r_1, s_1)$ is the corresponding reward for action 1, and F is the Fisher information metric [10]. When action 2 is made instead, the update equation becomes

$$Q = q + \frac{\alpha R_2}{F}\frac{d\ln[1-x(q)]}{dq} = q - \frac{\alpha R_2}{1-x(q)}. \tag{19}$$

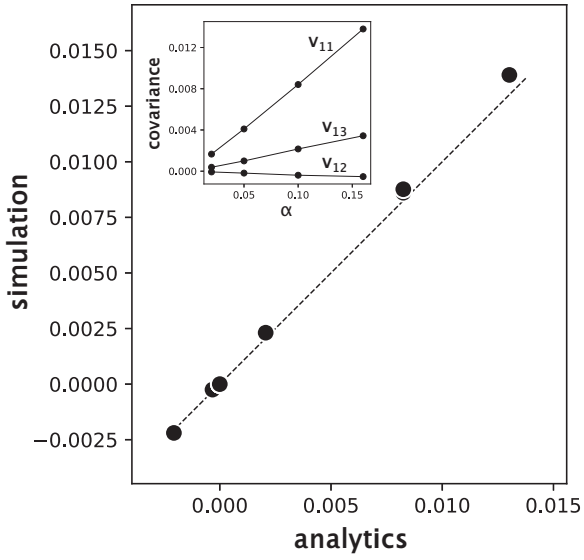

FIG. 3. Comparison of steady-state covariance elements estimated using simulation and analytical methods [Eq. (16)]. Inset: Linear scaling of covariance terms with the learning rate α (obtained from simulation).

We are interested in studying the stochastic dynamics of Q values which are updated according to the above rules.

To that end, we start off my writing the master equation for the probability distribution $P_{t+1}(Q)$, namely,

$$\begin{aligned} P_{t+1}(Q) = &\int dq \int dR_1\, x(q)\,\rho_1(R_1) \\ &\times \delta\left(-Q + q + \frac{\alpha R_1}{x(q)}\right) P_t(q) \\ &+ \int dq \int dR_2\,(1-x(q))\,\rho_2(R_2) \\ &\times \delta\left(-Q + q - \frac{\alpha R_2}{1-x(q)}\right) P_t(q). \end{aligned} \tag{20}$$

Here, $\rho_i(R_i)$ is the normal distribution of rewards received after action i. Implementing the methodology for evaluating the moments outlined in the previous section, we immediately obtain the update equations for the mean and the variance of Q:

$$E[Q] = E[q] + \alpha(r_1 - r_2), \tag{21}$$

$$\begin{aligned} V = v + \alpha^2\left[\frac{\lambda_1^2}{x} + \frac{\lambda_2^2}{1-x} - (r_1-r_2)^2\right] \\ + \frac{\alpha^2}{2}\left(\lambda_1^2\frac{1-x}{x} + \lambda_2^2\frac{x}{1-x}\right)v, \end{aligned} \tag{22}$$

where x is evaluated at $q = E[q]$, and $\lambda_i^2 = r_i^2 + s_i^2$ is introduced for convenience.

Additionally, using the moments of q, the mean and the variance of the action 1 probability $x(q)$ can be approximated via

$$E[x(q)] = x + \tfrac{1}{2}x''v, \tag{23}$$

$$\mathrm{Var}[x(q)] = (x')^2 v - \tfrac{1}{4}(x'')^2 v^2, \tag{24}$$

where $x' = x(1-x)$ and $x'' = x(1-x)(1-2x)$, both evaluated at the expected value $q = E[q]$.

To test the validity of our method, we performed numerical simulations of the learning process and compared the statistics with the predictions of the analytical update equations (see

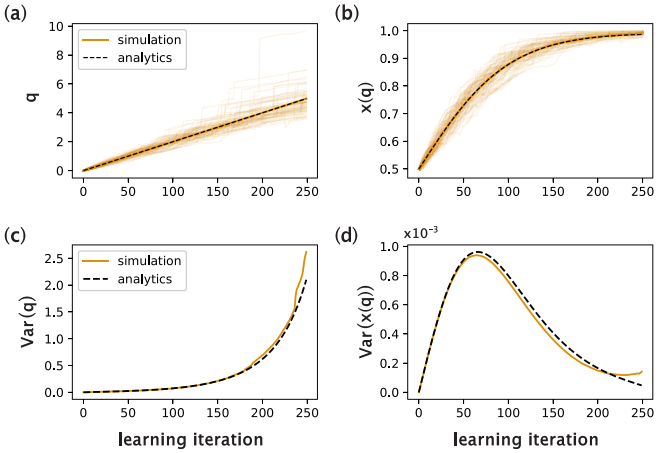

FIG. 4. Estimating the stochastic learning dynamics in the two-armed bandit problem. (a) Policy parameter q and (b) action 1 probability $x(q)$ as a function of the learning iteration. Transparent trajectories represent example stochastic realizations of the learning process. (c) Variance of the policy parameter and (c) variance of the action 1 probability estimated via simulation (solid lines) vs our iterative analytical approach (dashed lines). Parameters and initial conditions used: $r_1 = 1$, $r_2 = -1$, $s_1 = s_2 = 1$, and $q(0) = 0$.

Fig. 4). We observe a close match between them up to $V \sim 1$. Our iterative analytical estimate begins to deviate from simulation results beyond that point where the growth of the variance in Q is exponentially fast.

B. General case

We next consider the more general scenario where the agent can perform K different actions, with the probability of action i given by

$$x_i(\mathbf{Q}) = \frac{\exp(Q_i)}{\sum_{k=1}^{K} \exp(Q_k)}. \tag{25}$$

Here, $\mathbf{Q}$ is a K-element vector of policy parameters. If action i is performed, the corresponding Q value is updated via

$$Q_i = q_i + \frac{\alpha R_i}{x_i(\mathbf{q})}, \tag{26}$$

where $R_i \sim \mathcal{N}(r_i, s_i)$ represents the stochastic reward received for action i.

As in the two-armed case, we first write the master equation characterizing the exact stochastic dynamics of the learning process, namely,

$$P_{t+1}(\mathbf{Q}) = \int d\mathbf{q} \int d\mathbf{R} \sum_k x_k(\mathbf{q})\, \rho_k(R_k) \times \delta(-\mathbf{Q} + \mathbf{q} + \alpha \mathbf{d}_k) P_t(\mathbf{q}), \tag{27}$$

where $d_k^i = \delta_{ik} R_k / x_k$. In this multidimensional case, we want to estimate the updated means of the $\mathbf{Q}$-vector components, together with the $K \times K$ covariance matrix. During our derivations, we use the following identities for the different

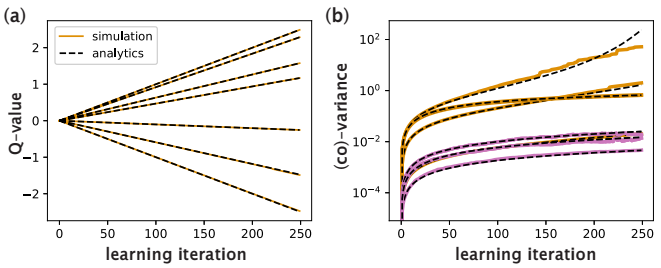

FIG. 5. Stochastic learning dynamics in the K-armed bandit model with $K = 50$. (a) Mean policy parameters $E[q_i]$ as a function of the learning iteration. (b) Dynamics of the variance terms V_{ii} (orange) and covariance terms V_{ij} (pink). Dashed lines are obtained from our iterative analytical approach, while the solid lines are generated from numerical simulations. Due to the large number of different summary statistics, only a subset of the terms are shown. Parameters used in the study: $r_i = 1 - 2(i-1)/(K-1)$, $s_i = 0.1$, $\alpha = 0.01$.

derivatives of action probabilities:

$$x'_{i,i} = x_i(1 - x_i), \quad x'_{i;j} = -x_i x_j, \quad \left(\frac{1}{x_i}\right)''_{ii} = \frac{1 - x_i}{x_i},$$
$$\left(\frac{1}{x_i}\right)''_{jk} = 0, \quad \left(\frac{1}{x_i}\right)''_{ij} = -\frac{x_j}{x_i}, \quad \left(\frac{1}{x_i}\right)''_{jj} = \frac{x_j}{x_i}. \tag{28}$$

The indices i, j, and k in the above identities are all different from each other.

The update rule for the mean Q values is simply

$$E[Q_i] = E[q_i] + \alpha r_i. \tag{29}$$

The diagonal and nondiagonal entries of the covariance matrix have distinct update rules. The exact expression for the diagonal terms is

$$V_{ii} = v_{ii} + \alpha^2 \left(\lambda_i^2 \int d\mathbf{q}\, \frac{1}{x_i(\mathbf{q})} P_t(\mathbf{q}) - r_i^2 \right). \tag{30}$$

Using the identities in Eq. (28), we apply the moment closure technique to find an approximate iterative equation, namely,

$$V_{ii} = v_{ii} + \alpha^2 \left(\frac{\lambda_i^2}{x_i} - r_i^2 \right) + \frac{\alpha^2 \lambda_i^2}{2 x_i} \left(v_{ii} + \sum_m x_m v_{mm} - 2 \sum_m x_m v_{im} \right). \tag{31}$$

Here, $\lambda_i^2 = r_i^2 + s_i^2$ as before, and x_i are the action probabilities evaluated at $\mathbf{q} = E[\mathbf{q}]$. Last, we find a succinct update rule for the nondiagonal entries:

$$V_{ij} = v_{ij} - \alpha^2 r_i r_j. \tag{32}$$

The accuracy of our iterative analytical method is illustrated in Fig. 5 for a K-armed bandit problem with a large number of actions ($K = 50$). We note that the accuracy starts to drop when the variance becomes so high that $V_{ii} \gg E[Q_i]^2$ for some i.

IV. DISCUSSION

In this paper, we studied the stochastic dynamics of reinforcement learning processes using a master equation for the

probability distribution of value functions. The formulation of the master equation as a first step in the investigation of the stochastic process is nontrivial, as the probability distribution may depend on both discrete (state index) and continuous (Q values) variables. To solve the master equation approximately, we used a method previously elaborated for the solution of the finite population problem in evolutionary games [15], but modified it for our more involved case.

We derived bulk equations for the dynamics of the average Q values (known before), as well us an iterative equation scheme for estimating the variance of the distribution (derived for the first time). The variance is a key statistical characteristic of the model. Having it, one can estimate the reliability of the algorithm, i.e., find the probability of matching the optimal solution with some accuracy in a given period of time. For the Boltzmann Q-learning problem with two agents, we derived a system of equation for the steady-state values of covariance terms, offering a quick way of assessing fluctuations in the stationary policy. We also applied our method for estimating the fluctuations in the K-armed bandit problem [10,13] where diffusion models have been applied till now, despite them being not the most accurate models [19]. We assume that solving iterative algebraic equations is much easier than solving a system of partial differential equations.

An interesting extension to the K-armed bandit problem studied in our work would be to consider it in an environment that changes its state randomly. Here the problem is very closely related to the subject of correlated random matrix products [20] (the random matrix is mapped to the state dynamics of the environment), where a transition has been found between localized and delocalized phases. For the proper working of the algorithm, the iteration dynamics should follow the localized phase. Rigorous mathematical considerations reveal that there are infinite singularities in such problems [21].

For a fruitful application of statistical physics to reinforcement learning, the phase structure should also be investigated. It can be rather rich due to the analogy to spin-glasses proposed in Ref. [10] for the case where the agent performs multiple actions simultaneously. The point is that now there are several phases in the model: the paramagnetic and spin glass phase [22], as well as the ferromagnetic one [23,24]. For the algorithm to work efficiently, it should operate in the ferromagnetic phase and maximally avoid the spin-glass phase with slow and chaotic dynamics.

The moment closure approximation, for the first time suggested in the current article, could also be applied to reinforcement learning problems with different schemes of discount [14] and iterated games in the general case. As the method has worked successfully for both Q-learning and policy gradient algorithms, we hope to combine it with the approach proposed in Ref. [25] and apply to the deep deterministic policy gradient (DDPG) algorithm as well in future work.

Besides reinforcement learning, our method can be applied to problems of evolution theory with stochastic transitions in the environment [26–30]. Evolution problems on dynamic environments are one of the most actively investigated topics in modern evolution theory. One application, for example, is to the dynamics of the Wright-Fisher model in the case of stochastically changing fitness landscape with a goal of avoiding the fixation of the allele. Such a task is interesting for finding optimal cancer therapies.

ACKNOWLEDGMENTS

We thank Armen Allahverdyan, Erik Arakelyan, Andranik Khachatryan, Ricard Solé, and Edgar Vardanyan for fruitful discussions. This work was supported by State Committee of Science Republic of Armenia, Grants No. 20TTAT-QTa003 and No. 21T-1C037.

APPENDIX A: STANDARD REWARD CASE FOR THE TWO-AGENT GAME

When the player rewards depend only on the actions taken in the current time step (action l for player "a" and action n for player "b"), the update equations for the Q values are

$$Q_l^a(t+\alpha) = Q_l^a(t) + \alpha\big[A_{ln} - Q_l^a(t)\big], \tag{A1a}$$

$$Q_n^b(t+\alpha) = Q_n^b(t) + \alpha\big[B_{nl} - Q_n^b(t)\big]. \tag{A1b}$$

This is in contrast to Eq. (2) where the actions in the previous time step entered the reward expressions. In this section, we derive the iterative update equations for the first and second moments of Q values which evolve according to Eq. (A1b).

In this setting, the master equation simplifies into

$$P_{t+1}(\mathbf{Q}) = \int d\mathbf{q} \sum_s X_s(\mathbf{q})\, \delta(-\mathbf{Q} + \mathbf{q} + \alpha \mathbf{d}_s) P_t(\mathbf{q}), \tag{A2}$$

where now it is written for the distribution of Q values. Here, $\mathbf{d}_s$ is a four-component vector for given s, with two of its entries given by the differences between the rewards and the q values corresponding to the joint action s, while the other two are zero. Components of $\mathbf{d}_s$ for all choices of s are

$$\begin{aligned}
&\text{(a:1, b:1)} \quad \mathbf{d}_1 = \{A_{11} - q_1 \quad 0 \quad B_{11} - q_3 \quad 0 \quad\},\\
&\text{(a:1, b:2)} \quad \mathbf{d}_2 = \{A_{12} - q_1 \quad 0 \quad 0 \quad B_{21} - q_4\},\\
&\text{(a:2, b:1)} \quad \mathbf{d}_3 = \{\quad 0 \quad A_{21} - q_2 \quad B_{12} - q_3 \quad 0 \quad\},\\
&\text{(a:2, b:2)} \quad \mathbf{d}_4 = \{\quad 0 \quad A_{22} - q_2 \quad 0 \quad B_{22} - q_4\}.
\end{aligned} \tag{A3}$$

We also introduce the matrix I_{sk} that indicates the nonzero elements of the vector $\mathbf{d}_s$, i.e., $I_{sk} = 1$ if $d_{sk} \neq 0$ and is zero otherwise.

By definition, the moments of Q values are given by

$$\begin{aligned}
E[Q_k] &= \int d\mathbf{Q}\, Q_k P_{t+1}(\mathbf{Q})\\
&= \int d\mathbf{q} \sum_s X_s(\mathbf{q}) \big(q_k + \alpha d_s^{(k)}\big) P_t(\mathbf{q}),
\end{aligned} \tag{A4}$$

$$\begin{aligned}
E[Q_k Q_l] &= \int d\mathbf{Q}\, Q_k Q_l P_{t+1}(\mathbf{Q})\\
&= \int d\mathbf{q} \sum_s X_s(\mathbf{q}) \big(q_k + \alpha d_s^{(k)}\big)\big(q_l + \alpha d_s^{(l)}\big) P_t(\mathbf{q}).
\end{aligned} \tag{A5}$$

We introduce the variance v_{kl} as

$$v_{kl} = E[q_k q_l] - E[q_l] E[q_l]. \tag{A6}$$

The updated variance V_{kl} is similarly defined as

$$V_{kl} = E[Q_k Q_l] - E[Q_l]E[Q_l]. \tag{A7}$$

Using our approximation method [Eq. (9)], we write the update equation for the means as

$$E[Q_k] = E[q_k] + \alpha \sum_s X_s d_s^{(k)} - \frac{\alpha}{2}\left(\sum_s \sum_{mn} X''_{s;mn} d_s^{(k)} v_{mn} - \sum_s \sum_m X'_{s;m} I_{sk} v_{mk}\right). \tag{A8}$$

Applying the same method on $E[Q_k Q_l]$ and substituting the resulting expression together with the above result for $E[Q_k]$ into Eq. (A7), we obtain the update equation for the variance, namely,

$$V_{kl} = v_{kl} - \alpha(x_l + x_k)v_{kl} + \alpha \sum_s \sum_m X'_{s;m}\left(d_s^{(k)} v_{ml} + d_s^{(l)} v_{mk}\right) + \alpha^2\left(\sum_s X_s d_s^{(k)} d_s^{(l)} - \sum_s X_s d_s^{(k)} \sum_s X_s d_s^{(l)}\right), \tag{A9}$$

where $x(\mathbf{q})$, $X(\mathbf{q})$, and $\mathbf{d}(\mathbf{q})$ are all evaluated at $\mathbf{q} = E[\mathbf{q}]$.

APPENDIX B: TWO-AGENT GAME WITH STATE-DEPENDENT Q VALUES

Here we consider a generalization of the Q-learning problem framed in Sec. II where now the Q values of one player depend also on the previous action taken by the other player (the state variable). If $\hat{l}$ and $\hat{n}$ are the actions taken by players "a" and "b," respectively, in the previous time step, and l and n are their current actions, then the update rules for the Q values are

$$Q^a_{l,\hat{n}}(t+\alpha) = Q^a_{l,\hat{n}}(t) + \alpha\left[A_{l\hat{n}} - Q^a_{l,\hat{n}}(t)\right],$$
$$Q^b_{n,\hat{l}}(t+\alpha) = Q^b_{n,\hat{l}}(t) + \alpha\left[B_{n\hat{l}} - Q^b_{n,\hat{l}}(t)\right]. \tag{B1}$$

Due to the added dimension, $\mathbf{Q}$ is now an 8-component vector (2 players × 2 actions × 2 states):

$$\mathbf{Q} = \left(Q^a_{1,1}, Q^a_{2,1}, Q^a_{1,2}, Q^a_{2,2}, Q^b_{1,1}, Q^b_{2,1}, Q^b_{1,2}, Q^b_{2,2}\right). \tag{B2}$$

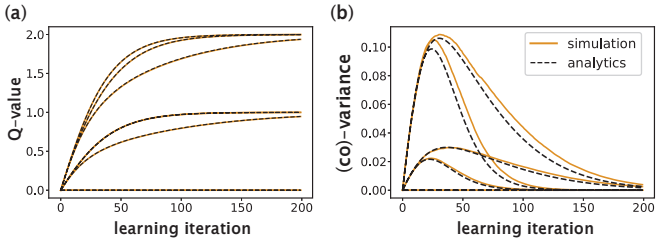

FIG. 6. Comparison of summary statistics obtained from simulation vs analytics in the state-dependent two-agent game. (a) Dynamics of Q values. (b) Dynamics of covariance element (a subset is shown for clarity). Same parameters as those in Fig. 1(b) were used.

The master equation for the joint distribution of Q values ($\mathbf{Q}$) and joint actions (s) has a form very similar to Eq. (5), namely,

$$P_{t+1}(\mathbf{Q}, s) = \int d\mathbf{q} \sum_{\hat{s}=1}^{4} X_{s\hat{s}}(\mathbf{q})\, \delta(-\mathbf{Q} + \mathbf{q} + \alpha \mathbf{d}_{s\hat{s}}) P_t(\mathbf{q}, \hat{s}). \tag{B3}$$

The main difference is that now the probability $X_{s\hat{s}}$ of taking the joint action s depends on the joint action in the previous step ($\hat{s}$). For example, the probability that players "a" and "b" perform actions 2 and 1 ($s = 3$), respectively, when their earlier actions were 1 and 1 ($\hat{s} = 1$), is given by

$$X_{31} = x^a_{2,1} x^b_{1,1}, \tag{B4}$$

where

$$x^a_{2,1} = \frac{\exp\left(Q^a_{2,1}/T\right)}{\sum_l \exp\left(Q^a_{l,1}/T\right)}, \quad x^b_{1,1} = \frac{\exp\left(Q^b_{1,1}/T\right)}{\sum_n \exp\left(Q^b_{n,1}/T\right)}. \tag{B5}$$

The update equations for the first and second moments of Q values are identical in form to the ones in the main text [Eqs. (10) and (11)]. The only difference is in the $\hat{s}$ dependence of joint action probabilities $X_{s\hat{s}}$ as well as of the indicator function $I_{s\hat{s}k}$, which takes the value 1 if the current joint action s and the previous joint action $\hat{s}$ lead to a change in the kth $\mathbf{Q}$-vector element. The set of iterative update equations gives the dynamics of 8 Q values and 256 covariance terms (Q_k and $V^{(s)}_{kl}$, with $k, l \in \{1, 2, ..., 8\}$, $s \in \{1, 2, 3, 4\}$), respectively). The agreement of this iterative analytical approach with the results of extensive simulations of the learning process is demonstrated in Fig. 6.

[1] E. Jablonka and M. J. Lamb, The evolution of information in the major transitions, J. Theor. Biol. **239**, 236 (2006).

[2] R. Sutton and A. Barto, *Reinforcement: An Introduction* (MIT Press, Cambridge, MA, 1998).

[3] Y. Sato and J. P. Crutchfield, Coupled replicator equations for the dynamics of learning in multiagent systems, Phys. Rev. E **67**, 015206(R) (2003).

[4] Y. Sato, E. Akiyama, and J. P. Crutchfield, Stability and diversity in collective adaptation, Physica D **210**, 21 (2005).

[5] K. Tuyls, P. J. T. Hoen, and B. Vanschoenwinkel, An evolutionary dynamical analysis of multiagent learning in iterated games, Auton. Agent. Multi-Agent Syst. **12**, 115 (2006).

[6] D. Bloembergen, K. Tuyls, D. Hennes, and M. Kaisers, Evolutionary dynamics of multiagent learning: A survey, J. Artif. Intell. Res. **53**, 659 (2015).

[7] A. Lipowski, K. Gontarek, and M. Ausloos, Statistical mechanics approach to a reinforcement learning model with memory, Physica A **388**, 1849 (2009).

[8] A. Kianercy and A. Galstyan, Dynamics of Boltzmann Q learning in two-player two-action games, Phys. Rev. E **85**, 041145 (2012).

[9] J. Rahme and R. P. Adams, A theoretical connection between statistical physics and reinforcement learning, arXiv:1906.10228.

[10] R. Fabbricatore and V. V. Palyulin, Gradient dynamics in reinforcement learning, Phys. Rev. E **106**, 025315 (2022).
[11] T. Lattimore and C. Szepesvári, *Bandit Algorithms* (Cambridge University Press, Cambridge, UK, 2020).
[12] S. Kakade, A natural policy gradient, Advances in Neural Information Processing Systems **14**, 1531 (2001).
[13] B. Li and C. H. Yeung, Understanding the stochastic dynamics of sequential decision-making processes: A path-integral analysis of Multi-armed Bandits, arXiv:2208.06245.
[14] M. Schultheis, C. A. Rothkopf, and H. Koeppl, Reinforcement learning with non-exponential discounting, arXiv:2209.13413.
[15] E. Vardanyan and D. B. Saakian, The analytical dynamics of the finite population evolution games, Physica A **553**, 124233 (2020).
[16] R. Grima, A study of the accuracy of moment-closure approximations for stochastic chemical kinetics, J. Chem. Phys. **136**, 154105 (2012).
[17] D. Schnoerr, G. Sanguinetti, and R. Grima, Validity conditions for moment closure approximations in stochastic chemical kinetics, J. Chem. Phys. **141**, 084103 (2014).
[18] D. Schnoerr, G. Sanguinetti, and R. Grima, Comparison of different moment-closure approximations for stochastic chemical kinetics, J. Chem. Phys. **143**, 185101 (2015).
[19] D. B. Saakian and C. K. Hu, Solution of classical evolutionary models in the limit when the diffusion approximation breaks down, Phys. Rev. E **94**, 042422 (2016).
[20] D. B. Saakian, Semianalytical solution of the random-product problem of matrices and discrete-time random evolution, Phys. Rev. E **98**, 062115 (2018).
[21] R. Poghosyan and D. B. Saakian, Infinite series of singularities in the correlated random matrices product, Front. Phys. **9**, 678805 (2021).
[22] A. Crisanti and H. J. Sommers, The sphericalp-spin interaction spin glass model: The statics, Z. Phys. B: Condens. Matter **87**, 341 (1992).
[23] D. B. Saakyan, Spherical P-spin glass in the limit $P \to \infty$ and information storage by means of continuous spins, J. Exp. Theor. Phys. Lett. **64**, 479 (1996).
[24] J. A. Hertz, D. Sherrington, and T. M. Nieuwenhuizen, Competition between glassiness and order in a multispin glass, Phys. Rev. E **60**, R2460 (1999).
[25] A. Crisanti and H. Sompolinsky, Path integral approach to random neural networks, Phys. Rev. E **98**, 062120 (2018).
[26] V. Mustonen and M. Lässig, From fitness landscapes to seascapes: Nonequilibrium dynamics of selection and adaptation, Trends Genet. **25**, 111 (2009).
[27] V. Mustonen and M. Lässig, Fitness flux and ubiquity of adaptive evolution, Proc. Natl. Acad. Sci. USA **107**, 4248 (2010).
[28] A. Mayer, T. Mora, O. Rivoire, and A. M. Walczak, Diversity of immune strategies explained by adaptation to pathogen statistics, Proc. Natl. Acad. Sci. USA **113**, 8630 (2016).
[29] D. B. Saakian, T. Yakushkina, and E. V. Koonin, Allele fixation probability in a Moran model with fluctuating fitness landscapes, Phys. Rev. E **99**, 022407 (2019).
[30] I. Cvijović, B. H. Good, E. R. Jerison, and M. M. Desai, Fate of a mutation in a fluctuating environment, Proc. Natl. Acad. Sci. USA **112**, E5021 (2015).