Dynamic social learning under graph constraints

Konstantin Avrachenkov, Vivek S Borkar, Sharayu Moharir, Suhail Mohmad Shah

Dynamic social learning under graph constraints

Konstantin Avrachenkov, Vivek S. Borkar, *Fellow, IEEE*, Sharayu Moharir, and Suhail Mohmad Shah

***Abstract*— We introduce a model of graph-constrained dynamic choice with reinforcement modeled by positively α-homogeneous rewards. We show that its empirical process, which can be written as a stochastic approximation recursion with Markov noise, has the same probability law as a certain vertex reinforced random walk. We use this equivalence to show that for $\alpha > 0$, the asymptotic outcome concentrates around the optimum in a certain limiting sense when 'annealed' by letting $\alpha \uparrow \infty$ slowly.**

***Index Terms*— dynamic choice with reinforcement, optimal choice, graphical constraints, annealed dynamics, vertex reinforced random walk**

I. Introduction

DYNAMIC choice models, wherein the subsequent choice of one among finitely many alternatives depends upon the relative frequency with which it has been selected in past, have found many applications. This is so particularly in the scenario when the higher the frequency, the higher the probability of an alternative being chosen again. Such 'positive reinforcement' is seen in models of herding behavior [17], evolution of conventions [32], 'increasing returns' economics [2], etc. Similar dynamics also arise in other disciplines, e.g., population algorithms for optimization [13] and more recently, for service requests in web based platforms for search, e-commerce, etc. [30]. One common caveat in all these is what is already the concern of the aforementioned models of herding and increasing returns economics, viz., the risk of some initial randomness leading to the process getting eventually trapped in an undesirable or suboptimal equilibrium behavior. In this work we present a different take on this issue. Firstly, we introduce what we call a graph-constrained framework, wherein the choice at any instant is restricted by the choice during the previous instant. This is a realistic scenario that reduces to the classical case when the constraint graph is fully connected. Some examples are:

1. Consider buyers buying a product on an e-commerce portal. They are influenced by both the average rating (assumed to be stable) and the number of people who bought the product, as reflected in the number of reviews. In this application the graph constraints come from suggestions from the e-commerce portal for purchase of items from the same or related categories.

2. Consider the task of locating an object in a large image using crowdsourced agents. Typically, the image is split into multiple sub-images and each agent is asked to examine a few sub-images for the desired object. Since the image is large, it is desirable to determine which sub-image to examine next based on partial information of the current state. One way to do this is to constrain the next sub-image to be one of the neighbors of the sub-image examined most recently, chosen randomly according to a probability distribution based on the current information from the crowd about these sub-images. See, e.g., [19] for one potential real application.

3. A graphical constraint may also arise in a scenario where a mobile sensing unit (e.g., a robot or a UAV) covers an area repeatedly. It has to plan its trajectory according to certain objectives which prioritize dynamically the preferred regions or 'hot spots'. The movement, however, can only be to neighboring positions. If there is no central coordinator, then one is faced with the kind of problem we have.

4. Online video sharing platforms such as YouTube make yet another application case. Typically, after a user has seen a video, he or she is recommended a list of suggested videos. The videos are recommended based on semantic similarity and the number of views. In this case, the graphical constraints come from the physical limitation of the screen (typically no one scrolls down more than one or two screens) and semantic similarities. Furthermore, the system is more likely to recommend a content with a large number of views and the user is also more likely to click on a content with a significant number of views. Our model not only confirms that this leads to the effect of social bubbles [26], but also proposes a way of tuning the recommendation mechanism to break such bubbles.

As indicated above, optimality is not guaranteed in many of the aforementioned models because of the dynamics getting trapped in a suboptimal limit, the so called 'trapping' phenomenon [2]. We show here that by suitably tuning or 'annealing' the choice probabilities, the asymptotic profile

Authors listed alphabetically. This is the author version of the paper accepted in IEEE Trans. on Control of Network Systems.

This work was supported by the grant '*Machine Learning for Network Analytics*' from the Indo-French Centre for Promotion of Advanced Scientific Research. The work of VB was also supported in part by a J. C. Bose Fellowship from the Government of India. The work of KA was also supported in part by grant "Distributed Learning and Control for Network Analysis" from Nokia Bell Labs.

Konstantin Avrachenkov is with INRIA Sophia Antipolis, 2004 Route des Lucioles, Valbonne 06902, France (e-mail: K.Avrachenkov@inria.fr).

VB and SM are and SMS was with the Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India. SMS is now with the Department of Electrical Communications Engineering, Hong Kong Uni. of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. (e-mail: borkar.vs@gmail.com; sharayu.moharir@gmail.com; suhailshah2005@gmail.com).

can be made to concentrate on the optimal behavior. The tuning scheme increases the concentration of probability on the current front runner and corresponds to the natural phenomenon whereby the agents' confidence in their choices increases with increasing adoption thereof by their peers. Our agents are autonomous, though influenced by the past. Thus the final outcome is *emergent* and not *engineered*. In the basic model (i.e., without the aforementioned 'annealing'), we get convergence to a common decision, but not necessarily to an optimal one. The 'annealed' variant on the other hand ensures the latter, i.e., asymptotic optimality. It should be emphasized that while we borrow terminology from simulated annealing (SA), our annealing scheme modulates the net drift, i.e., the driving vector field of the stochastic approximation iteration and *ipso facto* its limiting o.d.e., in order to achieve optimality, its effect on the noise component is unimportant. This is unlike classical SA where it is the noise variance that is being tuned. We do add extraneous noise to the choice probabilities (see (2) below) just as in SA, but its aim is to ensure that unstable equilibria are avoided almost surely, not to ensure avoidance of *stable* suboptimal equilibria as in SA. The former is an easier objective as it entails only some 'persistent excitation' (to borrow a phrase from control theory) to push the iterates away from unstable equillibria and their stable manifolds, and does not call for 'hill climbing' with noise as in SA. The slow morphing of the drift is tantamount to morphing of the landscape itself to make it more 'peaked' while retaining the same optima. (That is, ratio of the function value at a global maximum to that at a local maximum which is not a global maximum progressively increases, but their locations don't change.) The dynamics in question is closely related to similar dynamics arising in connection with vertex reinforced random walks [7], a fact we exploit.

We give brief comparisons with some related works in multiarmed bandits in order to highlight the differences. In [30], a related model is considered and it is observed that the process may get locked into suboptimal equilibria. The remedy they propose is to randomize the rewards for a fixed time window in a clever manner (dubbed a 'balanced' exploration) before the aforementioned dynamic choice process takes over. We eschew any such modification and instead take recourse to the above scheme which is indeed optimal in the limit. This result is of a distinct flavor compared to [30]. Also, our techniques are different, as are our objectives: we seek asymptotic optimality and do not consider regret. In [18], which is methodologically closer to our work, a full fledged game problem is considered wherein many agents are concurrently exercising their choices with their payoffs depending on others' choices as well. Their focus is on ε-Nash equilibria and not on optimal behavior as in our (non–game theoretic) work. While the core technique, viz., use of the multiplicative weight rule, is common between this work and [18], they use a different choice thereof. Graphical constraints analogous to ours are used in [29] in a bandit framework, but they are motivated by how communication among agents can be factored into the analysis. In general, bandit algorithms do not involve graphical constraints and their focus is on non-asymptotic behavior unlike ours. However, graphical restrictions in bandit context do arise in a number of practical applications and have important implications. The standard algorithms deployed to solve bandit problems such as the ϵ-greedy strategy or *UCB* algorithm [24] may fail to achieve optimal behavior under graph constraints, as one may get stuck with a choice with a sub-optimal reward. We substantiate this claim in Section VI with a simple example.

We draw upon the framework of [7] substantially. (See [8], [9], [10] for extensions.) The key contribution of *ibid.* is the analysis of a general vertex reinforced random walk using the 'o.d.e.' approach to stochastic approximation. It derives very broad results about their asymptotic behavior, and then narrows these down to concrete examples with linear reinforcement to obtain stronger claims. Our model is pitched in between - it is a nonlinear model, but a very specific one and allows for more specific claims to be established. Use of annealing ideas in this context is another novelty of our work.

Such graphically constrained choice models can also be posed as stochastic combinatorial optimization problems. A well known heuristic for solving such problems is simulated annealing. However, SA *with noisy observations* is well known to be sample inefficient [14], [20], [21]. In fact, the best possible sample complexity results that have been obtained (Theorem 3, [14]) require that the number of samples required per iteration increase to infinity with the iteration count. This

Key Notation	
μ_i	Reward associated with object i.
m	Number of objects.
$S_i(n)$	Number of times i was picked.
$x_i(n)$	Relative frequency $S_i(n)/n$.
$\mathcal{S}_m$	Unit simplex in $\mathbb{R}^m$.
$\text{int}(\mathcal{S}_m)$	Interior of $\mathcal{S}_m$.
$\mathcal{G}$	Directed graph.
$\mathcal{V}$	Node set of $\mathcal{G}$.
$\mathcal{E}$	Edge set of $\mathcal{G}$.
$\mathcal{N}(i)$	Neighbourhood of i.
$\zeta(n)$	Noise in reward vector.
$\mathcal{F}_n$	$\sigma\big(\xi(k), \zeta_i(k), 1 \leq i \leq m, k \leq n\big)$.
$\hat{\mu}_i(n)$	Empirical mean, see (4).
$\epsilon(n)$	Exploration time step (see (5)).
$c(n), a(n)$	See (5) and (7).
$f_i^\alpha(x)$	Reinforcement function, $(\mu_i x_i)^\alpha$.
α	Reinforcement exponent, see above.
$\chi_\cdot(i)$	Uniform distribution on $\mathcal{N}(i)$.
$p_{ij}^\alpha(x)$	Transition prob. of $\{\xi(n)\}$, see (8).
$\pi^\alpha(x)$	Stationary distribution of $p_{ij}^\alpha(x)$.
$\varphi^\alpha(x)$	See (11).
$\iota_i(n)$	See (10).
A	Adjacency matrix, $A := [[a_{ij}]]_{i,j\in\mathcal{V}}$.
T	Temperature, defined as $1/\alpha$.
$b(n)$	Time step in T, see (18).
D	$\{i \in \mathcal{V} : \mu_i = \max_j \mu_j\}$.
$f(n) = O(g(n))$	$\limsup_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} < \infty$.
$f(n) = \Omega(g(n))$	$g(n) = O(f(n))$.
$f(n) = o(g(n))$	$\lim_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} = 0$.
$f(n) = \omega(g(n))$	$g(n) = o(f(n))$.
$f(n) = \Theta(g(n))$	$f(n) = O(g(n))$ and $g(n) = O(f(n))$.

makes deploying SA with noisy observations quite difficult, particularly for applications where obtaining samples may entail time consuming simulations. In contrast, our algorithm needs one sample per iteration under i.i.d. bounded variance noise, which makes it much more sample efficient as compared to SA with noisy observations.

We describe our model in the next section and demonstrate its connection with the vertex reinforced random walk. Section 3 provides convergence analysis of the basic scheme. In section 4 we analyze its 'annealed' counterpart, leading to the desired result. Section 5 specializes the problem to complete graph where we can say more. Section 6 provides some numerical experiments. Three appendices sketch some technical issues left out of the main text for ease of reading.

Notation: For ease of reference, we list the key notation used in the paper in the above table. This includes the standard Big-O notation used throughout the paper.

II. Problem formulation

In this section we set up our model of choice dynamics and the key notation.

Model: Consider a stream of agents arriving one at a time[1] and choosing one of $m > 1$ distinct objects, with a reward $\mu_i > 0$ associated with the ith object. The $(n+1)$-st agent picks the jth object with conditional probability (conditioned on past history) $p_j(n)$, which we shall soon specify. Let $\xi(n) = i$ if the nth agent picks object i. Let $S_i(n) :=$ the number of times object i was picked till time n and $x_i(n) := \frac{S_i(n)}{n}$, $n \geq 1$, its relative frequency. Then a simple calculation leads to the recursion

$$x_i(n+1) = x_i(n) + \frac{1}{n+1}\left(\mathbb{I}\{\xi(n+1) = i\} - x_i(n)\right),\ n \geq 0. \tag{1}$$

Here $\mathbb{I}\{\cdots\}$ is the 'indicator function' which is 1 if its argument is true and 0 otherwise. For specificity, we arbitrarily set $x_i(0) = \frac{1}{m}\ \forall i$, suggestive of a uniform prior. This will not affect our conclusions. Throughout, we use the convention $\frac{0}{0} = 0$. The vector $x(n) := [x_1(n), \cdots, x_m(n)]^T$ takes values in the simplex of probability vectors,

$$\mathcal{S}_m := \left\{ x = [x_1, \cdots, x_m]^T : x_i \geq 0\ \forall i,\ \sum_j x_j = 1 \right\}.$$

We shall denote by $\text{int}(\mathcal{S}_m)$ the interior of $\mathcal{S}_m$. We assume that the observed reward at time n for choice i is not μ_i, but $\tilde{\mu}_i(n) = \mu_i + \zeta_i(n)$ where $\{\zeta_i(n), n \geq 0\}$ is i.i.d. zero mean noise with bounded variance.

Graphical Constraints: We assume that the choice in the $(n+1)$-st time slot is constrained by the choice made in the nth slot, e.g., when, given the present choice, only some selected 'nearby' or 'related' choices are offered or preferred (see examples in the introduction). We model this as follows. Consider a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}, \mathcal{E}$ are resp., its node and edge sets, with $|\mathcal{V}| = m$. Assume that $\mathcal{G}$ is irreducible, i.e., there is a directed path from any node to any other node. Let $\mathcal{N}(i) := \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$ denote the set of successors of i in $\mathcal{G}$. If i is chosen at any instant n, the next choice must come from $\mathcal{N}(i)$. We assume:

(A1) *For each i, $i \in \mathcal{N}(i)$. This implies a self-loop at each node, i.e., $(i,i) \in \mathcal{E}\ \forall\ i \in \mathcal{V}$. (Thus, in particular, $|\mathcal{N}(i)| \geq 2\ \forall i$.) We also assume that the neighborhood structure is bidirectional, i.e., $i \in \mathcal{N}(j) \iff j \in \mathcal{N}(i)$.*

Selection Policy: Let $\mathcal{F}_n :=$ the σ-field $\sigma(\xi(t), \zeta_i(t), 1 \leq i \leq m, t \leq n)$. Then the vector process $x(n) \in \mathcal{S}_m$, whose i*'th* component $x_i(n) := \frac{S_i(n)}{n}$, is assumed to satisfy (1) with

$$\mathbb{P}(\xi(n+1) = j | \mathcal{F}_n) = (1 - \varepsilon(n))\tilde{p}^{\alpha}_{\xi(n)j}(x(n)) + \varepsilon(n)\chi_j(\xi(n)). \tag{2}$$

Here:

- $$\tilde{p}^{\alpha}_{ij}(x) := \mathbb{I}\{j \in \mathcal{N}(i)\}\frac{\hat{f}^{\alpha,n}_j(x)}{\sum_{l \in \mathcal{N}(i)} \hat{f}^{\alpha,n}_l(x)}, \tag{3}$$

  for $\hat{f}^{\alpha,n}_i(x) := (\hat{\mu}_i(n) x_i(n))^{\alpha}$, where

  $$\hat{\mu}_i(n) := \frac{\sum_{k=0}^{n} \mathbb{I}\{\xi(k) = i\}\tilde{\mu}_i(k)}{\sum_{k=0}^{n} \mathbb{I}\{\xi(k) = i\}}$$

  is the empirical estimate of μ_i at time n recursively computed by

  $$\begin{aligned} \hat{\mu}_i(n+1) &= \left(1 - \frac{1}{S_i(n+1)}\right)\hat{\mu}_i(n) + \frac{\tilde{\mu}_i(n+1)}{S_i(n+1)}, \\ &\qquad\qquad \text{if } \xi(n+1) = i, \\ &= \hat{\mu}_i(n), \qquad\qquad \text{otherwise,} \end{aligned} \tag{4}$$

  with $\hat{\mu}_i(0) := 0$.
- $\{\varepsilon(n)\}$ satisfy the recursion

  $$\varepsilon(n+1) = (1 - c(n))\varepsilon(n), \tag{5}$$

  where $0 < c(n) \downarrow 0$, $\sum_n c(n) = \infty$, $nc(n) \overset{n\uparrow\infty}{\to} 0$. The last condition implies that $\sum_n c(n)^2 < \infty$. We also assume that for $a(n) := \frac{1}{n+1}$,

  $$\sum_n \varepsilon(n)^m = \infty,\ \sum_n a(n)\varepsilon(n) = \infty, \tag{6}$$

  $$\varepsilon(n) = \omega\left(\frac{1}{\sqrt{n}}\right). \tag{7}$$

  One example is $c(n) = \frac{1}{1+(n+1)\log(n+1)}$, which results in $\varepsilon(n) = \Theta\left(\frac{1}{\log n}\right)$, see Appendix III for details.
- $\chi_\cdot(i)$ is the uniform distribution on $\mathcal{N}(i), i \in \mathcal{V}$.

That is, with probability $1 - \varepsilon(n)$, we pick $\xi(n+1) = j$ with probability $p^{\alpha}_{\xi(n)j}(x(n))$, and with probability $\varepsilon(n)$, we pick it uniformly from $\mathcal{N}(\xi(n))$. As $\alpha \downarrow 0$, the process approaches a simple random walk on the graph that picks a neighbor with equal probability. As $\alpha \uparrow \infty$, the process at i will (asymptotically) pick the $j \in \mathcal{N}(i)$ for which $\mu_j x_j = \max_{k \in \mathcal{N}(i)} \mu_k x_k$, uniformly. An immediate observation is the following, proved in Appendix I.

[1]This is for convenience. The identity of agents is irrelevant here and they may repeat as long as the choice mechanism remains the same.

Lemma 2.1: $\hat{\mu}_i(n) \to \mu_i$ a.s. $\forall i$.

Thus a.s., $\lim_{n\uparrow\infty} \hat{f}_i^{\alpha,n}(x) = f_i^\alpha(x) := (\mu_i x_i)^\alpha \ \forall\ i, x, \alpha$ and

$$\lim_{n\uparrow\infty} \tilde{p}_{ij}^\alpha(x) = p_{ij}^\alpha(x) := \mathbb{I}\{j \in \mathcal{N}(i)\} \frac{f_j^\alpha(x)}{\sum_{l\in\mathcal{N}(i)} f_l^\alpha(x)} \ . \quad (8)$$

The functions f_i^α are monotone increasing, which captures the 'positive reinforcement', i.e., the fact that increased choice of a particular object i increases its probability of being chosen in future, all else remaining the same. Each f_j^α is a locally Lipschitz function in $\text{int}(\mathcal{S}_m)$, strictly increasing in x_j and satisfying *positive α-homogeneity*: $f_j^\alpha(ax_j) = a^\alpha f_j^\alpha(x_j)$ for $a \geq 0$. Then $\mu_i x_i$ can be viewed as the fraction of the total reward accrued by the fraction of population that chose i. Thus, e.g., in example 1 in the introduction, it is the average rating of i times the fraction of the customers who bought i from among all who bought similar products. (In fact, it can be the *number* thereof rather than the *fraction*, because the normalization factor cancels out in the transition probability defined in (8).) Its homogeneity property renders the choice probabilities defined in (8) scale-independent, as it should. Since our selection probability for i will be proportional to $f_i^\alpha(x_i)$, a higher value of α makes the preference more peaked in the sense already described: it concentrates the probability mass further near global maxima, thereby putting higher weight on 'exploitation' than on 'exploration'. Smaller α do the opposite. The 'annealed' scheme we propose later slowly increases α to capture the trade-off between the two.

III. Convergence analysis

This section analyzes the convergence of the above scheme for fixed α using the theory of stochastic approximation [12]. The standard stochastic approximation algorithm is

$$y(n+1) = y(n) + a(n)[F(y(n), Y(n+1)) + \iota(n) + W(n+1)] \quad (9)$$

where the possibly random positive stepsizes $\{a(n)\}$ satisfy $\sum_n a(n) = \infty$, $\sum_n a(n)^2 < \infty$, the 'martingale noise' $\{W(n)\}$ satisfies $E[W(n+1)|\mathcal{F}'_n] =$ the zero vector for $\mathcal{F}'_n := \sigma(y(t), a(t), Y(t), \iota(t), W(t), t \leq n)$, $\iota(n) \to 0$ componentwise a.s., and the 'Markov noise' $\{Y(n)\}$ satisfies $P(Y(n+1) \in \cdot|\mathcal{F}'_n) = \hat{p}_{y(n)}(\cdot|Y(n))$ for a suitable transition probability $\hat{p}_y(\cdot|\cdot)$ parametrized by y. Then (1) has this form with $y(n) = x(n), a(n) = \frac{1}{n+1}$,

$$\begin{aligned} W_i(n) \ &= \ I\{\xi(n+1) = i\} - (1 - \varepsilon(n))p_{\xi(n)i}^\alpha(x(n)) \\ &\quad - \varepsilon(n) I\{i \in \mathcal{N}(\xi(n))\}/m_i, \end{aligned}$$

$Y(n) = \xi(n)$, $\hat{p}_y(j|i) = p_{ij}^\alpha(x)$. Also, $\iota(n)$ is a vector whose ith component is

$$\varepsilon(n)(m_i^{-1} - \tilde{p}_{\xi(n)i}^\alpha(x(n))) + (\tilde{p}_{\xi(n)i}^\alpha(x(n)) - p_{\xi(n)i}^\alpha(x(n))) \to 0. \quad (10)$$

(The presence of $\iota(n)$ does not affect the convergence, see the third 'extension' in section 2.2, [12] which applies to the stochastic approximation with Markov noise as well.) The stochastic matrix $[[p_{ij}^\alpha(x)]]_{i,j\in\mathcal{V}}$ is parametrized by the probability vector $x \in \mathcal{S}_m$. For fixed x, let $\pi^\alpha(x)$ denote its stationary distribution, whose existence and uniqueness is ensured for each fixed $x \in \text{int}(\mathcal{S}_m)$ by our irreducibility assumption for $\mathcal{G}$ (see e.g., [6, Section 6.1]). A direct calculation shows that

$$\tilde{\pi}_i^\alpha(x) := \frac{f_i^\alpha(x) \sum_{k\in\mathcal{N}(i)} f_k^\alpha(x)}{\sum_\ell (f_\ell^\alpha(x) \sum_{k\in\mathcal{N}(\ell)} f_k^\alpha(x))}, \ i \in \mathcal{V},$$

satisfies the local balance condition $\tilde{\pi}_i^\alpha(x) p_{ij}^\alpha(x) = \tilde{\pi}_j^\alpha(x) p_{ji}^\alpha(x)$, because both sides equal

$$\frac{f_i^\alpha(x) f_j^\alpha(x) \mathbb{I}\{j \in \mathcal{N}(i)\}}{\sum_\ell (f_\ell^\alpha(x) \sum_{k\in\mathcal{N}(\ell)} f_k^\alpha(x))},$$

where $\mathbb{I}\{j \in \mathcal{N}(i)\} = \mathbb{I}\{i \in \mathcal{N}(j)\}$. So $\pi^\alpha(x) = \tilde{\pi}^\alpha(x)$.

We apply the 'o.d.e. approach' to our problem. Thus let $\varphi_i^\alpha(x) := f_i^\alpha(x) \sum_{j\in\mathcal{N}(i)} f_j^\alpha(x)/x_i$ and consider the o.d.e.

$$\dot{x}_i(t) = \frac{x_i(t)\varphi_i^\alpha(x(t))}{\sum_k x_k(t)\varphi_k^\alpha(x(t))} - x_i(t). \quad (11)$$

Note that every equilibrium of (11) satisfies the fixed point equation

$$\pi(i) = h_i(\pi) := \frac{f_i^\alpha(\pi) \sum_{j\in\mathcal{N}(i)} f_j^\alpha(\pi)}{\sum_k f_k^\alpha(\pi) \sum_{\ell\in\mathcal{N}(k)} f_\ell^\alpha(\pi)} \quad \forall i. \quad (12)$$

Set $h(\cdot) := [h_1(\cdot, \cdots, h_m(\cdot)]$. By irreducibility, every such π must be in $\text{int}(\mathcal{S}_m)$.

Lemma 3.1: The o.d.e. (11) has the same trajectories and the same asymptotic behavior as the o.d.e.

$$\dot{z}_i(t) = z_i(t) \left(\varphi_i^\alpha(z(t)) - \sum_j z_j(t)\varphi_j^\alpha(z(t)) \right), \quad (13)$$

i.e., $z(t) = x(\tau(t))$ for some $t \in [0, \infty) \mapsto \tau(t) \in [0, \infty)$ which is strictly increasing and satisfies $t \uparrow \infty \iff \tau(t) \uparrow \infty$.

Proof: Since the r.h.s. of (13) is locally Lipschitz in the interior of $\mathcal{S}_m$, (13) has a unique solution when $z(0) \in \text{int}(\mathcal{S}_m)$. We obtained (13) from (11) by multiplying the r.h.s. of (11) by the positive scalar valued bounded function $q(t) := \sum_k x_k(t)\varphi_k^\alpha(x(t))$, which is bounded away from zero uniformly in t. This amounts to a pure time scaling $t \mapsto \tau(t)$ where $\tau(\cdot)$ is specified by the well-posed differential equation $\dot{\tau}(t) = q(\tau(t))$. Then $z(t) := x(\tau(t))$. (The same device was used in [7], p. 368.) Also, for suitable $\infty > c_2 > c_1 > 0$, $c_1 t \leq \tau(t) \leq c_2 t$. In particular, $\tau(t) \uparrow \infty$ as $t \uparrow \infty$, so the entire trajectory is covered. The claim follows. ∎

The dynamics (13) is a special case of *replicator dynamics* [28] (as is equation (3), [7], p. 368, in a similar context). Note also that an equilibrium z^* of (13) must satisfy

$$z_i^* > 0 \implies \varphi_i^\alpha(z^*) = \sum_j z_j^* \varphi_j^\alpha(z^*). \quad (14)$$

In particular, $\varphi_i^\alpha(z^*) \equiv$ a constant for $i \in$ the support of z^*.

Let $A := [[a_{ij}]]_{i,j\in\mathcal{V}}$ be the (symmetric) adjacency matrix of $\mathcal{G}$. Then for $x = [x_1, \cdots, x_m] \in \mathcal{S}_m$,

$$\varphi_i^\alpha(x) = \frac{\partial}{\partial x_i} \Psi^\alpha(x) \text{ for } \Psi^\alpha(x) := \frac{1}{2\alpha} \sum_{i,j} a_{ij} f_i^\alpha(x) f_j^\alpha(x).$$

Thus (13) corresponds to the replicator dynamics for a potential game with potential $-\Psi^\alpha$ [28]. In what follows, by

local maximum of a function we mean a point in its domain where a local maximum is attained and not the function value there. We make the following assumption which is generically true (i.e., true for almost all parameter values, see, e.g., [25], Chapter 2).

(A2) *The equilibrium points of (11) (i.e., the fixed points of (12)) are isolated and hyperbolic, i.e., the Jacobian matrix of h at these points does not have eigenvalues on the imaginary axis. Also, their stable and unstable manifolds, which exist by hyperbolicity, intersect transversally if they do.*[2]

In view of the preceding discussion, this amounts to the requirement that the Hessian of Ψ^α be nonsingular at its critical points in int$(\mathcal{S}_m)$.

Theorem 3.2: For each $\alpha > 0$, the local maxima of $\Psi^\alpha : \mathcal{S}_m \mapsto \mathbb{R}$ are stable equilibria of (11) and the iterates of (1) converge to the set thereof, a.s.

Proof: Since (11) and (13) are obtained from each other by a time scaling $t \mapsto \tau(t)$ that satisfies $\tau(t) = \Theta(t)$, it suffices to consider only (13). We have

$$\begin{aligned} &\frac{d}{dt}\Psi^\alpha(z(t)) \\ &= \sum_i z_i(t)\left(\varphi_i^\alpha(z(t)) - \sum_j z_j(t)\varphi_j^\alpha(z_j(t))\right)^2 \\ &\geq 0. \end{aligned} \tag{15}$$

Thus $-\Psi^\alpha$ serves as a Lyapunov function for (13), implying that it converges to the set of critical and Kuhn-Tucker points of Ψ^α. The local maxima will then correspond to stable equilibria. We next argue that the iterates converge to some local maximum a.s. By Corollary 8, p. 74, [12], for stochastic approximation with Markov noise, combined with the first bullet of section 2.2, p. 16, and Corollary 4, p. 18, [12] (both of which which work with Markov noise for exactly identical reasons) and (A2), the iterates converge a.s. to a single, possibly sample path dependent, critical or Kuhn-Tucker point of Ψ^α. That it must be a stable equilibrium, i.e., a local maximum, follows by a variant of the theory developed in section 4.3, pp. 40-47, [12]. This argument is very technical and is sketched in Appendix II. ∎

The next lemma is similar to Theorem 6.3 of [7], see also Theorem 5.1 of [3], reproduced as Chapter 10 of [2]. We sketch a brief proof for the sake of completeness.

Lemma 3.3: The probability of convergence of $\{x(n)\}$ in (1) to any local maximum of Ψ^α in $\mathcal{S}_m$ is strictly positive.

Proof: Let x^* be a local maximum and O its domain of attraction for (11). Since the graph is irreducible and the probability of next choice being j is strictly positive $\forall\ j \in \mathcal{N}(i)$ when the current choice is i, it follows that the probability of $\{x(n)\}$ reaching O from any initial condition in finitely many steps is strictly positive. Once in O, the probability of convergence to x^* is strictly positive by Theorem III.4 of [23], implying the claim. ∎

[2]This makes it a special case of a 'Morse-Smale system'.

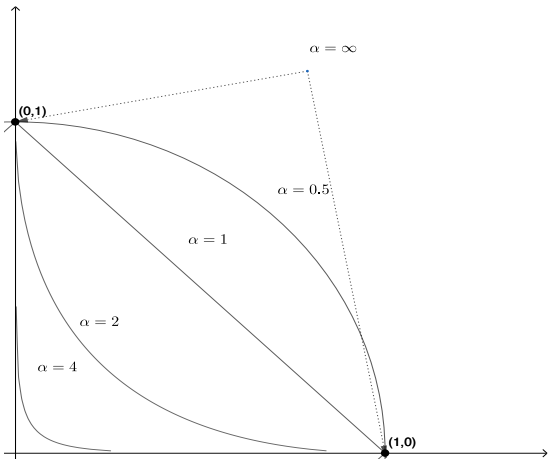


Fig. 1: An illustration of the collapse of sets B^α to B^∞.

We have

$$\Psi^\alpha(\pi) = \frac{1}{2\alpha}\sum_{i,j}(\mu_i\mu_j)^\alpha a_{ij}\pi_i^\alpha\pi_j^\alpha.$$

Corollary 3.4: The local maxima of Ψ^α are of the form $\pi(i) = z(i)^{\frac{1}{\alpha}}$ where z is a local maximum of the quadratic form in $\{x_i\}$ given by $\sum_{i,j} x_i x_j (\mu_i\mu_j)^\alpha a_{ij}$, over the set

$$B^\alpha := \{y : y(i) \geq 0\ \forall i,\ \sum_i y(i)^{\frac{1}{\alpha}} = 1\}.$$

IV. 'ANNEALED' DYNAMICS

In this section, we consider the 'annealed' dynamics. That is, taking a cue from simulated annealing [22], we consider the asymptotics as $\alpha \uparrow \infty$, corresponding to the 'temperature' $T := 1/\alpha \downarrow 0$, slowly with time. A behavioral interpretation is that the agents exhibit a herd behavior, weighing in public opinion more and more with time. We first analyze the optimization problem described in Corollary 3.4 as $\alpha \uparrow \infty$. The set of limit points of B^α as $\alpha \uparrow \infty$ is given by (see Fig. 1) $B^\infty := \cap_{\alpha>0}\overline{(\cup_{\alpha'>\alpha}B^{\alpha'})} \supset B^* := \{e_i, 1 \leq i \leq m\}$, where $e_i, 1 \leq i \leq m$, are the unit coordinate vectors. Let

$$D := \{i \in \mathcal{V} : \mu_i = \max_j \mu_j\} \tag{16}$$

and $\Pi^\alpha := \{\pi \in \mathcal{S}_m : \pi$ is a local maximum of $\Psi^\alpha\}, \alpha > 0$.

Lemma 4.1: If $\alpha_n \uparrow \infty$ and $\pi_n \in \Pi^{\alpha_n}$, then $\pi_n \to B^*$.

Proof: We are concerned here only about the relative sizes (i.e., ratios) of the summands in the definition of Ψ^α. So we may assume that $\max_i \mu_i = 1$ and drop the factor $\frac{1}{2\alpha}$ in the definition of Ψ^α. This simplifies the analysis while not affecting the location of local maxima and the relative magnitudes of the function values there. Let $S^* := \{i : \mu_i = 1\}$. Then $\sum_{i,j}(\mu_i\mu_j)^\alpha a_{ij}x(i)x(j) \overset{\alpha\uparrow\infty}{\to} 0$ uniformly outside any relatively open neighborhood of B^* in $\mathcal{S}_m$. Hence

$$\begin{aligned} &\max_{x\in\mathcal{S}_m}\sum_{i,j}(\mu_i\mu_j)^\alpha a_{ij}x(i)x(j) \\ &\overset{\alpha\uparrow\infty}{\to} \max_{x\in B^\infty}\sum_{i,j}(\mu_i\mu_j)^\alpha a_{ij}x(i)x(j) = 1, \end{aligned}$$

which is attained at some $e_i, i \in S^*$. The claim follows. ∎

Recall from (12) that π^α is a (not necessarily unique) solution to the fixed point equation

$$\pi^\alpha(i) \;:=\; \frac{f_i^\alpha(\pi^\alpha)\sum_{k\in\mathcal{N}(i)} f_k^\alpha(\pi^\alpha)}{\sum_\ell (f_\ell^\alpha(\pi^\alpha)\sum_{k\in\mathcal{N}(\ell)} f_k^\alpha(\pi^\alpha))}. \tag{17}$$

Decrease $T := 1/\alpha$ slowly according to the iteration

$$T(n+1) = (1-b(n))T(n),\ n \geq 0, \tag{18}$$

where $1 > b(n) \downarrow 0$ are stepsizes satisfying

$$\sum_n b(n) = \infty,\ nb(n) \stackrel{n\uparrow\infty}{\to} 0,\ b(n) = o(c(n)). \tag{19}$$

The second condition implies $\sum_n b(n)^2 < \infty$. Assume that $x(0) \in \text{int}(\mathcal{S}_m)$. This is not a restriction, since $x(n) \in \text{int}(\mathcal{S}_m)$ from some n on when all possible choices have been made at least once and the above requirement can be ensured simply by counting time from then on. Our main result is the following, reminiscent of '*stochastically stable*' equilibria of [33].

Theorem 4.2: $\sum_{i\in D} x_i(n) \to 1$ a.s.

Proof: The second and third conditions in (19) render the pair (1), (18) a two time scale stochastic approximation with (1) run on a fast time scale and (18) run on a slower time scale. In fact the situation is simpler than the general two time scale schemes because the latter does not depend on the former, the dependence is unidirectional. We shall use the results of [31]. In [31], stochastic recursive *inclusions* involving set-valued maps on both time scales are considered. In (1), (18), we have instead single valued Lipschitz maps for which assumptions A1-A8 of [31] are easily verified. Our slow iteration (18) has a unique limit 0, whence A10 of [31] is trivially satisfied. This leaves the verification of assumption A9 of [31]. Consider (1) for fixed $\alpha = 1/T, \varepsilon(n) \equiv 0$, and define:

$$D_0^T := \{\pi : \pi \text{ satisfies the fixed point equation (17)}\}.$$

Let $D^T :=$ the closed convex hull of D_0^T. Using the fact that $T(n)$ update on a slower time scale and hence are 'quasi-static' for the faster time scale of $x(n)$ (cf. the 'two time scale' methodology of [12], secion 6.1), we first 'freeze' the slow components $T(n) \approx T$ and analyze the fast iterate (1). By the theory of stochastic approximation with Markov noise (see [12], Chapter 6), it tracks the o.d.e. (11), a time-scaled version of (13) as observed earlier. Thus it converges to D^T by Theorem 3.2. We next show that as $T = T(n) \downarrow 0$ and $\tilde{\pi}_n \in D^{T(n)}$ $n \geq 1$, $\tilde{\pi}_n \to$ the set D defined in (16). Consider a subsequence $\tilde{T}(n) \downarrow 0$ such that

$$\tilde{\pi}_n := \pi^\alpha\Big|_{\alpha=1/\tilde{T}(n)} \to \pi^*$$

for some $\pi^* \in \mathcal{S}_m$ with support S^*. Rewrite (17) as

$$\begin{aligned}\tilde{\pi}_n(i) &= \frac{\sum_{j\in\mathcal{N}(i)}[\mu_i\mu_j\tilde{\pi}_n(i)\tilde{\pi}_n(j)]^{1/\tilde{T}(n)}}{\sum_{i'}\sum_{j\in\mathcal{N}(i')}[\mu_{i'}\mu_j\tilde{\pi}_n(i')\tilde{\pi}_n(j)]^{1/\tilde{T}(n)}} \\ &= \frac{\frac{\sum_{j\in\mathcal{N}(i)}[(\mu_i\mu_j\tilde{\pi}_n(i)\tilde{\pi}_n(j)]^{1/\tilde{T}(n)}}{\max_{k,l\in\mathcal{N}(k)}[\mu_k\mu_l\tilde{\pi}_n(k)\tilde{\pi}_n(l)]^{1/\tilde{T}(n)}}}{\sum_{i'}\frac{\sum_{j\in\mathcal{N}(i')}[\mu_{i'}\mu_j\tilde{\pi}_n(i')\tilde{\pi}_n(j)]^{1/\tilde{T}(n)}}{\max_{k,l\in\mathcal{N}(k)}[\mu_k\mu_l\tilde{\in}_n(k)\tilde{\pi}_n(l)]^{1/\tilde{T}(n)}}}.\end{aligned}$$

As $\tilde{T}(n) \downarrow 0$, this concentrates on the set of $(i,j) \in \mathcal{E}$ for which

$$\begin{aligned}&\mu_i\pi^*(i)\sum_{j\in\mathcal{N}(i)\cap S^*}\mu_j\pi^*(j) \\ &\quad= \max_k\left(\mu_k\pi^*(k)\sum_{\ell\in\mathcal{N}(k)\cap S^*}\mu_\ell\pi^*(\ell)\right).\end{aligned}$$

Combined with Lemma 4.1, this implies that the measure will concentrate on the i such that

$$\mu(i)^2 = \max_j \mu(j)^2,$$

i.e., on D. Setting $D^{1/T} = D$ when $T = 0$, this verifies A9 of [31] for our purposes[3]. Then Theorem 4, p. 1435, [31], holds. We note that in the notation of this theorem, $\mathcal{Y} = \{0\}$ and $\lambda(y) = D^{1/y}$, whence the claim follows. ■

V. The unconstrainted case

In this section we consider the case without graphical constraints, i.e., when the graph $\mathcal{G}$ is fully connected, where we can say more. The case without graphical constraints can be viewed as a special case with $\mathcal{G} =$ the complete graph, i.e., $a_{ij} = 1\ \forall\ i,j$. Then $\Psi^\alpha(x) = \left(\sum_i f_i^\alpha(x)\right)^2$, which is convex for $\alpha \geq 1$, where the absence of graphical constraints does allow us to make stronger statements. Unfortunately this does not buy us stronger results for the $\alpha \uparrow \infty$ asymptotics. However, the story is different for a fixed $\alpha \in (0,1)$, where we indeed can say much more than in the graphically constrained case. Specifically, we get desired convergence guarantees even for a fixed α in this range, and make an analogy with Ant Colony Optimization [1], [13].

For $\alpha \in (0,1)$, since the expression being squared is non-negative, we can equivalently consider the problem of maximizing $\psi^\alpha(x) := \sum_i f_i^\alpha(x)$, which is *strictly concave*. Hence it has a unique maximum on $\mathcal{S}_m$ to which our scheme will converge even without annealing. In fact, in this case, the stationary solution can be specified explicitly using the Lagrange multiplier technique as:

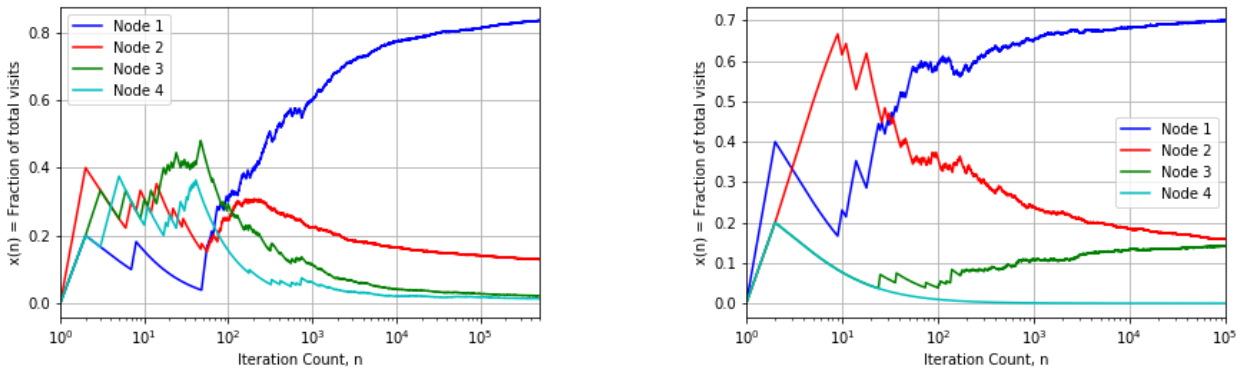
$$x_i(\infty) = \frac{\mu_i^{\alpha/(1-\alpha)}}{\sum_{k=1}^m \mu_k^{\alpha/(1-\alpha)}}\ . \tag{20}$$

From (20), as $\alpha \to 1$, the frequencies $x_i(\infty)$ start to concentrate on D defined in (16). As seen in the simulation section, in practice one does not need to take α very close to one. If $\alpha = 1$, the replicator dynamics has the well studied linear payoffs and converges to a solution with only one nonzero component by standard arguments.

Now consider the case of $\alpha > 1$ with $\varepsilon(n) \equiv$ a constant $\varepsilon > 0$. Note that in the unconstrained case, given x, the transition probability matrix $[[p_{ij}^{\alpha,\varepsilon}(x)]]$ is a stationary probability matrix with the identical rows $\pi^{\alpha,\varepsilon}(x)$ given by

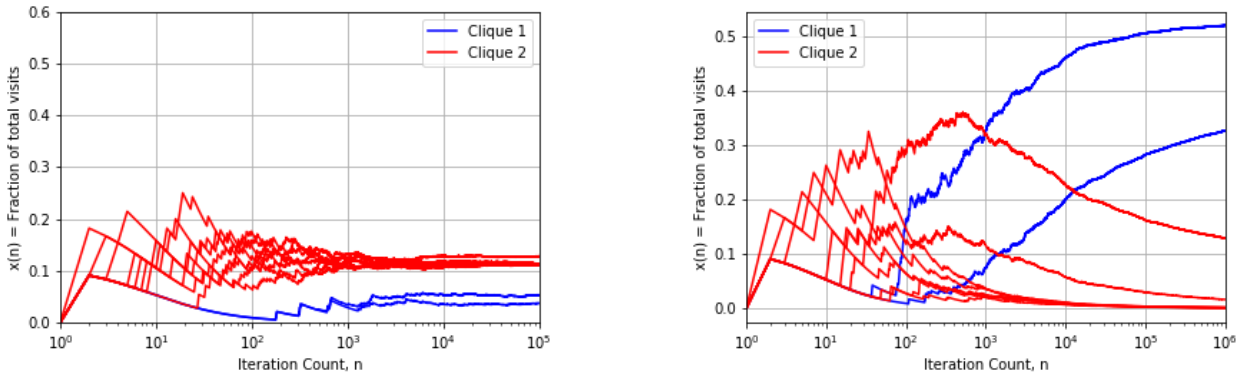
$$\pi_i^{\alpha,\varepsilon}(x) := (1-\varepsilon)\frac{f_i^\alpha(x)}{\sum_k f_k^\alpha(x)} + \varepsilon\frac{1}{m}.$$

[3]It is also clear that the limiting measure will be uniform on D.



(a) Linear Topology (b) Star Topology

Fig. 2: Fraction of total Visits, $x(n)$ Vs. Iteration Count for Linear and Star Topology.



(a) Initialize in clique-2, α fixed. (b) Initialize in clique-2, $\alpha \to \infty$.

Fig. 3: Fraction of total Visits, $x(n)$ Vs. Iteration Count for the two clique experiment

Hence its stationary distribution coincides with its (identical) rows. By Corollary 8, p. 74, [12], the sequence $\{x(n)\}$ tracks the o.d.e.

$$\dot{x}_i(t) = \pi^{\alpha,\varepsilon}(x(t)) - x_i(t), \tag{21}$$

i.e.,

$$\dot{x}_i(t) = (1-\varepsilon)\frac{f_i^\alpha(x(t))}{\sum_k f_k^\alpha(x(t))} + \varepsilon\frac{1}{m} - x_i(t).$$

The stationarity condition for the above o.d.e. gives

$$(1-\varepsilon)\frac{f_i^\alpha(x)}{\sum_k f_k^\alpha(x)} + \varepsilon\frac{1}{m} - x_i = 0 \quad \forall i. \tag{22}$$

If $\varepsilon \to 1$, then by standard continuity arguments, $x \to$ the set of solutions to (22) corresponding to $\varepsilon = 1$. This is a singleton consisting of the uniform distribution $x_i = \frac{1}{m}\ \forall i$. The map

$$(x,\varepsilon) \mapsto F(x,\varepsilon) := (1-\varepsilon)(\sum_k f_k^\alpha(x))^{-1}[f_1^\alpha(x), \cdots, f_m^\alpha] + \frac{\varepsilon}{m}I - x$$

has a nonsingular Jacobian matrix $-I$ w.r.t. x in $\text{int}(\mathcal{S}_m)$ at $\varepsilon = 1$. Hence by the implicit function theorem, the fixed point x^ε of (22) is an analytic function in a small neighborhood of the uniform distribution [6], i.e.,

$$x_i(\varepsilon) = \frac{1}{m} + (1-\varepsilon)x_i^{(1)} + \dots .$$

Substituting this expansion in the stationarity condition (22) and equating terms with the same powers of $1-\varepsilon$ yields

$$x_i^{(1)} = \frac{\mu_i^\alpha}{\sum_{k=1}^m \mu_k^\alpha} - \frac{1}{m}.$$

This implies that the states with indices in the set D will obtain a larger fraction of visits in comparison with the other states. This is reminiscent of the Ant Colony Optimization algorithm of [1], [13] where the initial randomness itself builds up the bias in favor of the optimum, to which the scheme converges *with high probability*. A very fine analysis of the $\alpha > 1$ case for a related model appears in [10].

The payoff functions $\{\varphi_i^\alpha(\cdot)\}$ in (13) are of the form $\varphi_i^\alpha(z) = g_i(z_i)h(z)$ for $h(\cdot) : \mathcal{S}_m \mapsto (0,\infty)$ and $g_i : [0,1] \mapsto \mathbb{R}^+$, where the latter are monotone increasing. As shown in Lemma 4, p. 14, [13], corners of $\mathcal{S}_m$, i.e., $\{e_i\}$, are stable equilibria for (13) and the only ones to be so. Moreover, the domain of attraction of e_i is $\{z \in \mathcal{S}_m : z_i > z_j,\ j \neq i\}$. In view of the foregoing, this makes it clear how the bias for the optimum builds up starting from a uniform prior.

VI. Simulation Experiments

In this section we empirically demonstrate our theoretical results on a star and linear graph topology (with $m = 4$, see Fig. 2 and 3). For the linear topology, $\mu = (2, \frac{1}{4}, \frac{1}{2}, 1)$, designed so as to demonstrate the hill descending capabilities (i.e. jump out of the local maximum at node 4) of the algorithm. The noise $\zeta_i(\cdot)$ is assumed to be $N(0, 0.1)$. The random exploration parameter is set as $\varepsilon(n) := \frac{1}{\log(n+1)}$. As can be seen in Fig. 2, $x_1(n)$ (the fraction of visits to the node with the highest μ) converges to 1 as $n \uparrow \infty$. We remark here that the cooling schedule $\{\alpha(n)\}$ is the most important (and sensitive) parameter of the algorithm. A too fast or constant cooling schedule may tend to make the algorithm get stuck in the local maximum at node 4. The cooling schedule we used was $\alpha(n+1) = \alpha(n)\left(1 - \frac{1}{n \log n}\right)^{-1}$. For initial few iterations, we keep $\alpha = 10^{-2}$ fixed to promote exploration. For the star topology, $\mu = (1, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. The cooling schedule was the same as before. Here, the central node, i.e. the node connected to all other nodes, is node 4. For comparison purposes, we have also tried $\mu = (2, \frac{1}{4}, \frac{1}{2}, 1)$ with the fixed $\alpha = 0.85 < 1$ in the complete graph setting. The dynamics always converges to the stationary solution $(0.98, 0.000, 0.000, 0.019)$. This demonstrates our conclusion from Section V that in the unconstrained case for the values of $\alpha < 1$ even not so close to one, a very significant portion of the mass is concentrated on the optimal node.

Our next numerical experiment is aimed at highlighting the importance of annealing for convergence of $x(n)$ to D. We consider a graph composed of two cliques connected through a single edge. The number of nodes for clique-1 is 2 and those for clique-2 is 8. We set the noise $\zeta_i = 0$ for all i for this experiment. The results have been plotted in Fig. 3. We set $\mu_i = 1$ for $i \in$ clique-1 and $\mu_i = 0.5$ for $i \in$ clique-2. Some points to note are:

- If we initialize the walk in clique-2 and *do not* increase $\alpha \to \infty$, then the relative frequencies converge to non-zero values for nodes in clique-2. (In Fig. 3(a), we have set $T = 0.1$ $(\alpha = 10)$.)
- If we initialize the walk in clique-2 and *do* increase $\alpha \to \infty$, then the chain moves to clique-1 and stays there.

With linear topology, we make an important comparison with the multiarmed bandit literature. With nodes labeled $\{1, 2, 3, 4\}$, the $\alpha \uparrow \infty$ limit corresponds to the transition probabilities

$$p(1|1), p(1|2), p(4|3), p(4|4) = 1, \;\; p(i|j) = 0 \text{ otherwise.}$$

That is, the chain moves deterministically to the neighbor (including itself) with the highest reward. It has two communicating classes $\{1, 2\}$ and $\{3, 4\}$. For $\epsilon \in (0, 1)$, the ϵ-greedy policy has a stationary distribution that is seen to concentrate equally on $1, 4$ as $\epsilon \downarrow 0$ by the symmetry of the problem. In particular, it is a suboptimal distribution. A simple two time scale argument applied to (1) then shows that $x(n)$ converges this suboptimal distribution. In contrast, if we consider the corresponding fully connected graph with the same reward structure, the purely greedy policy given by the $\alpha \uparrow \infty$ limit has $p(1|i) = 1 \;\forall i$ and the stationary distribution is seen to concentrate on the optimal node 1. In the fully connected case the $\varepsilon(n)$-greedy policy with $\varepsilon(n) = \frac{1}{n}$ converges to the optimal, as shown in Theorem 3 of [4]. Thus, a standard bandit algorithm can fail in the graph-constrained framework.

In Fig. 4, we provide a comparison of the proposed algorithm with Simulated Annealing. We briefly describe the details of the modified version of SA we use here. The SA algorithm consists of a discrete time inhomogeneous Markov chain, whose transition mechanism $P(n) := [[p_{xy}(n)]]_{x,y \in \mathcal{V}}$ for temperature T_n can be formally written as:

$$p_{x,y}(n) = \begin{cases} 0, & \text{if } y \notin \mathcal{N}(x) \\ \frac{1}{|\mathcal{N}(x)|} \exp\left\{\frac{-\left(\hat{\mu}_x(n) - \hat{\mu}_y(n)\right)^+}{T_n}\right\}, & \text{otherwise} \end{cases}$$

and

$$p_{x,x}(n) = 1 - \sum_{i \in \mathcal{N}(x)} p_{x,i}(n),$$

where $(x)^+ := \max(0, x)$ and $\hat{\mu}_x(n)$ is the empirical mean estimate at time n of object x. To keep the comparison to our algorithm fair we update the empirical mean in the same manner as (4).

Judging from Fig. 4, our algorithm achieves a better medium and long run performance in terms of relative frequency of the optimal reward for both linear and star topology. The time step for SA is kept equal to $\frac{\gamma}{\log(1+k)}$, where $\gamma = 0.1$ is selected empirically to give the best performance.

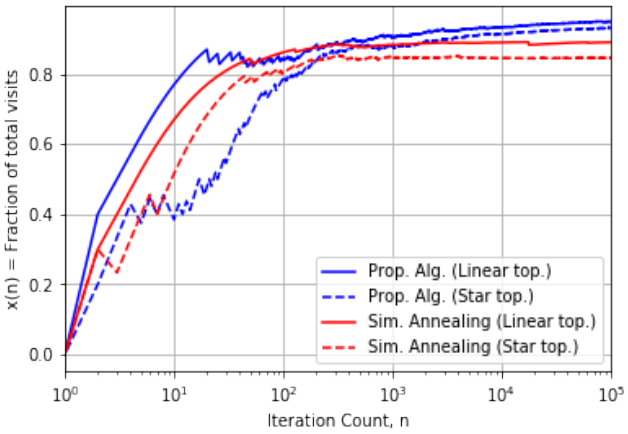


Fig. 4: An empirical comparison of SA with the proposed algorithm for star and linear topology. The reward vectors are kept the same as the previous experiments.

Appendix I

Proof of Lemma 2.1 : This follows from the strong law of large numbers if

$$S_i(n) \uparrow \infty, \tag{23}$$

and our convergence analysis applies. But (23) follows from the fact $\sum_n \varepsilon(n) = \infty$, because by the conditional Borel-Cantelli lemma (Lemma 17, p. 49, of [12]),

$$\sum_n \mathbb{I}\{\xi(n+1) = i\} = \infty \iff \sum_n P(\xi(n+1) = i | \mathcal{F}_n) = \infty$$

a.s. Now $\chi(\xi(n))$ assigns mass $\frac{1}{|\mathcal{N}(i)|} \geq \frac{1}{m}$ to i when $\xi(n) \in \mathcal{N}(i)$ and 0 otherwise. Hence

$$\sum_n P(\xi(n+1) = i|\mathcal{F}_n) \geq \frac{1}{m} \sum_{j \in \mathcal{N}(i)} \sum_n \varepsilon(n) I\{\xi(n) = j\},$$

By the conditional Borel-Cantelli lemma,

$$\frac{1}{m} \sum_{j \in \mathcal{N}(i)} \sum_n \varepsilon(n) I\{\xi(n) = j\} = \infty$$
$$\iff \frac{1}{m} \sum_{j \in \mathcal{N}(i)} \sum_n \varepsilon(n) P(\xi(n) = j|\mathcal{F}_{n-1}) = \infty.$$

Using a similar bound for $P(\xi(n) = j|\mathcal{F}_{n-1})$ yields

$$\frac{1}{m} \sum_{j \in \mathcal{N}(i)} \sum_n \varepsilon(n) P(\xi(n) = j|\mathcal{F}_{n-1})$$
$$\geq \frac{1}{m^2} \sum_{k \in \mathcal{N}(j)} \sum_{j \in \mathcal{N}(i)} \sum_{n \geq 1} \varepsilon(n)^2 I\{\xi(n-1) = k\},$$

and so on, so combining all these inequalities and using (6),

$$\frac{1}{m^m} \sum_{n \geq m} \varepsilon(n)^m = \infty \Longrightarrow \sum_n \mathbb{I}\{\xi(n+1) = i\} = \infty$$

a.s. Thus (23) holds.

Appendix II

Here we sketch the proof of the 'avoidance of unstable equilibria a.s.' (also known as 'avoidance of traps') result invoked in the proof of Theorem 3.2. This is based on the results of section 4.3, [12], pp. 44-51, originally from [11]. These in turn depend on the estimates of section 4.1, pp. 31-41 of [12]. We sketch the main steps, referring the reader to the above for details common to both and highlight only the differences between the present set-up and that of section 4.3, [12]. For later reference, we use (An)*, $n \geq 1$, to denote the assumptions of *ibid.* and simply (An) to refer to our own.

The proof of *ibid.* is broadly in two parts. The bulk of the work is for the first part, which is to show that the iterates will keep getting pushed away from the stable manifolds of unstable equilibria sufficiently often, a.s. This is an argument based on the conditional Borel-Cantelli lemma. In [12], this argument relies on showing that the aggregated martingale noise over an interval approaches a non-degenerate gaussian distribution under suitable scaling, by the central limit theorem for martingale arrays. This is ensured by assumption (A6)*. The topological assumption (A5)* then ensures that there is enough probability of the iterates getting pushed away adequately and often enough that they move away from the manifold, to the domain of attraction of stable equilibria. The second part then says that it will converge to a stable equilibrium almost surely. This uses a concentration result from section 4.1 of [12], which quantifies the probability of convergence to a stable equilibrium given that the current iterate is in its domain of equilibrium. For us, the second part simply amounts to replacing the latter result by its counterpart for Markov noise from [23]. The first part is what takes the most effort. While (A5)* can be ensured by imposing a reasonable assumption, (A6)* turns out to be more elusive, precisely because of graph constraints that imply motion only to neighboring nodes. Thus, the natural counterpart of (A6)* that would require the conditional covariance of $\xi(n+1)$ given $\mathcal{F}_n$ to be non-singular is simply false. Luckily, we need such non-singularity to hold in an average sense. Bulk of our work below will be towards establishing this. The condition (A7)* is simply replaced by its suitable counterpart here, so it is not a major issue.

It should also be added that the assumptions and proof of [11] followed here are among many such for 'avoidance of traps' results, see [15], [27], to name some others. Thus it seems eminently possible to adapt these to give alternative sets of assumptions and corresponding proofs for Markov noise.

We begin by discussing the key assumptions (A5)*-(A8)* in section 4.3, [12], that are specific to the results therein. Assumptions (A1)*-(A4)* of *ibid.* are generic assumptions for stochastic approximation that are already covered here. Let $m_i = |\mathcal{N}(i)|$. Define the $\{\mathcal{F}_n\}$-martingale difference sequence

$$\begin{aligned} M_i(n+1) &= I\{\xi(n+1) = i\} - (1 - \varepsilon(n)) p^{\alpha}_{\xi(n)i}(x(n)) \\ &\quad - \varepsilon(n) I\{i \in \mathcal{N}(\xi(n))\}/m_i. \end{aligned} \tag{24}$$

Let $a(n) := \frac{1}{n+1}, n \geq 0$. Then (1) can be written as

$$\begin{aligned} x_i(n+1) &= x_i(n) + a(n) \Bigg[(1 - \varepsilon(n)) p^{\alpha}_{x_i(n)j}(x(k)) + \\ &\quad \frac{\varepsilon(n)}{m_i} \Bigg] + a(n) M_i(n+1), \ 1 \leq i \leq m. \end{aligned} \tag{25}$$

Let W denote the complement of the union of the domains of attraction of stable equilibria, i.e., the local maxima of Ψ. One important implication of (A2) is the following. Define the truncated open cone

$$C_\kappa := \left\{ x \in \mathcal{S}_m : 1 < x_1 < 2, \ \left| \sum_{i=2}^m x_i^2 \right|^{1/2} < \kappa x_1 \right\}$$

for some $\kappa > 0$. For any orthogonal matrix O, $x \in \mathbb{R}^d$ and $a > 0$, we let OD, $x + D$ and aD denote respectively, the rotation of D by O, translation of D by x, and scaling of D by a. Then (A2) implies:

(A2') There exists $\kappa > 0$ such that for any $x \in \mathcal{S}_m$ and sufficiently small $a > 0$, there exists an orthogonal matrix $O_{a,x}$ such that $B(x, a, \kappa) := x + aO_{x,a}C_\kappa$ satisfies: any $y \in B(x, a, \kappa)$ is at least distance a away from W.

This means in particular that for any sufficiently small $a > 0$, we can plant a version of the truncated cone scaled down by a near x by means of suitable translation and rotation, in such a manner that it lies entirely in W. This ensures that any point in $\mathbb{R}^m$ cannot have points in the complement of W arbitrarily close to it in all directions. This replaces (A5)*. Next we consider (A6)*. This is not appropriate for the 'Markov noise' framework here, hence will have to be modified. We modify it by replacing $Q(x)$ there by $Q^n_i(x), 1 \leq i \leq m$, where $Q^n_{\xi(n)}(x(n))$ is the conditional covariance matrix of the

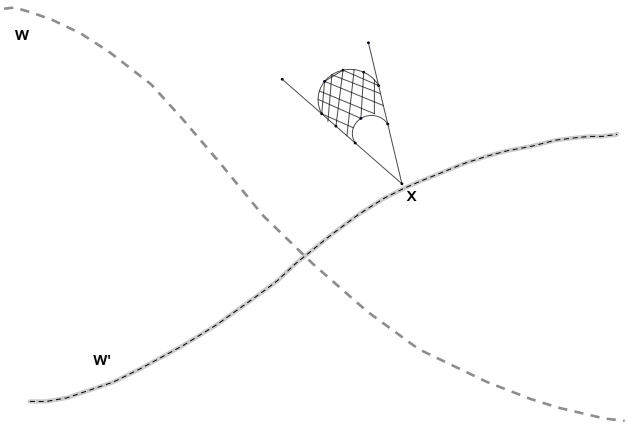


Fig. 5: An illustration of Assumption 2'.

random vector $[I\{\xi(n+1)=1\},\cdots,I\{\xi(n+1)=m\}]$ conditioned on $\xi(n), x(n)$, which is the same as 'conditioned on $\mathcal{F}_n$' by virtue of conditional independence. Then $Q_i^n(x)$ has a $m_i \times m_i$ diagonal block $\bar{Q}_i^n(x)$, corresponding to rows and columns indexed by elements of $\mathcal{N}(i)$. Note also that $I\{\xi(n+1)=j\}$, $j \in \mathcal{N}(i)$, conditioned on $\xi(n), x(n)$, are conditionally Bernoulli random variables, albeit correlated. The remaining rows and columns of $Q_i^n(x)$ are zero. Thus $Q_i^n(x)$ is singular for each i, n, and the obvious counterpart of (A6)*, which would require the least eigenvalue of the $Q_i^n(x)$'s to be bounded away from zero, is not tenable. However, a closer scrutiny of the arguments of section 4.3, [12], specifically the last part of the proof of Lemma 16 there, shows that the actual requirement is weaker. We exploit this fact below.

An additional complication is that the smallest eigenvalue of the diagonal submatrices $\bar{Q}_i^n(x)$ is also zero because of the fact that $\sum_{j\in\mathcal{N}(i)} I\{\xi(n+1)=j\}=1$ when $\xi(n)=i$ introduces degeneracy: the vector $\mathbf{1} := [1,\cdots,1]^T$ is always an eigenvector corresponding to eigenvalue 0. However, our dynamics is confined to the probability simplex, a compact manifold with boundary, to which $\mathbf{1}$ is orthogonal. Thus we need to consider only the linear transformations

$$y \in \mathcal{R}^{m_i} \mapsto$$
$$\mathcal{S}^i := \{z \in \mathcal{R}^{m_i} : z_j \geq 0, 1 \leq j \leq m_i, \sum_{j=1}^{m_i} z_j = 1\}.$$

We show later that the least eigenvalue $\lambda_n(i)$ of $\bar{Q}_i^n(x)\Big|_{S^i}$ satisfies

$$\lambda_n(i) \geq \frac{\varepsilon(n)}{m}, \tag{26}$$

which in turn implies that

$$\bar{Q}_i(x)\Big|_{S^i} \geq \frac{\varepsilon(n)}{m} J_i\Big|_{S^i}, \tag{27}$$

where $J_i :=$ the diagonal matrix with diagonal elements $=1$ for rows and columns corresponding to $\mathcal{N}(i)$ and $=0$ otherwise. The inequality in (27) is w.r.t. the usual partial order for positive semidefinite matrices. Denote by I the m-dimensional identity matrix and by $D\pi^\alpha$ the Jacobian matrix of π^α. Also define

$$\varphi(n) = \left(\sum_{k=n}^{s(n)} a(k)^2 \varepsilon(k)\right)^{\frac{1}{2}}.$$

where $s(n) := \min\{k \geq n : \sum_{\ell=1}^{k} a(k) \geq T\}$ for a prescribed $T > 0$. Then as in p. 48, [12], we have,

$$\begin{aligned}
&\frac{1}{\varphi(n)^2} \sum_{j=s(n)}^{s(n)+i-1} a(j)^2 \times \\
&\quad \left(\prod_{k=j+1}^{s(n)+i-1} (I + a(k)(D\pi^\alpha(x(n)) - I))\right) \\
&\quad \times Q_{\xi(n)}(x(n)) \times \\
&\quad \left(\prod_{k=j+1}^{s(n)+i-1} (I + a(k)(D\pi^\alpha(x(n)) - I))\right)^T \\
&\quad \geq \frac{1}{m\varphi(n)^2} \times \\
&\sum_{j=s(n)}^{s(n)+i-1} a(j)^2 \left(\prod_{k=j+1}^{s(n)+i-1} (I + a(k)(D\pi^\alpha(x(n)) - I))\right) \\
&\quad \times \varepsilon(k) J_{\xi(n)} \left(\prod_{k=j+1}^{s(n)+i-1} (I + a(k) D\pi^\alpha(x(n)) - I))\right)^T.
\end{aligned} \tag{28}$$

Define the random probability vector $\nu(n) = [\nu_1(n),\cdots,\nu_s(n)]$ by

$$\nu_i(n) := \frac{\sum_{k=n}^{s(n)} a(k)^2 \varepsilon(k) I\{\xi(k)=i\}}{\sum_{k=n}^{s(n)} a(k)^2 \varepsilon(k)}$$

for $i \in S$. Then an argument analogous to that of Lemma 6, pp. 73-74, [12], shows that a.s., every limit point π^* of $\{\nu(n)\}$ is some stationary distribution π^α for $\{\xi(n)\}$. In particular, it has full support by virtue of (17). By dropping to a further subsequence if necessary, consider a limit point of the r.h.s. of (28). This will be of the form $\frac{1}{m}\int_0^t \Phi(T,s)(\sum_i \pi^*(i) J_i)\Phi(T,s)^T ds$ for some $t \geq 0$, where $\Phi(\cdot,\cdot)$ is the fundamental matrix for the linearization of the o.d.e. (11) restricted to $\mathcal{S}_M$. This is clearly positive definite when restricted to $\mathcal{S}_M$ (because $\sum_i \pi^*(i) J_i$ is). The argument leading to Corollary 18 in [12], pp. 49, then goes through as before.

(A7)* is used in section 4.3, [12], on p. 50 alone. One key step in its application there is the use of the estimate of trapping probability (i.e., the probability of convergence to a stable equilibrium conditioned on the iterates being in its domain of attraction), from Theorem 8, pp. 37, [12]. This is used to conclude the proof in section 4.3 of [12]. That estimate cannot be used here because we are dealing with Markov noise. However, we can use the (stronger) concentration result from Theorem III.4, [23] to conclude our desired result in a completely analogous manner. That said, we still need to verify, as in p. 50 of [12], that

$$\sum_{k\geq n}\frac{1}{(k+1)^2}=o(\varphi(n))=o\left(\sqrt{\sum_{k=n}^{s(n)}\left(\frac{\varepsilon(k)}{k+1}\right)^2}\right). \quad (29)$$

The l.h.s. is $\Theta\left(\frac{1}{n}\right)$. The r.h.s. is $\Theta\left(\sqrt{\frac{T\varepsilon(s(n))^2}{s(n)}}\right)=\Theta\left(\frac{\varepsilon(n)}{\sqrt{n}}\right)$ because $s(n)=\Theta\left(ne^T\right)$. Thus (29) amounts to $\frac{1/n}{\varepsilon(n)/\sqrt{n}}\to 0$, i.e., $\varepsilon(n)=\omega\left(\frac{1}{\sqrt{n}}\right)$. This is the second condition in (7).

(A8)* can be seen to hold in the interior of $\mathcal{S}_m$, which is our state space of interest, because it follows from (17) that the equilibria will be in the interior of $\mathcal{S}_m$.

We have ignored the errors due to time variation of $\hat{\mu}_n, T(n)$ because they do not affect the analysis. Both get multiplied by $a(n)$ and are therefore $o(a(n))$ in the 'drift' (i.e., the driving vector field) of the algorithm and contribute only an asymptotically negligible error. (See again the second bullet on p. 17 of [12] which applies to stochastic approximation with Markov noise as well.) The factor $a(n)\varepsilon(n)$ on the other hand multiplies the noise and therefore is what matters for 'avoidance of traps'.

Derivation of (26):

For $\xi_n=i$,

$$\begin{aligned}p_n(j) &:= (1-\varepsilon(n))p_{ij}^{\alpha}(x(n))I\{j\in\mathcal{N}(i)\}\\ &\quad+\varepsilon(n)I\{j\in\mathcal{N}(i)\}/m_i. \quad (30)\end{aligned}$$

Then $p_n(j)\geq\frac{\varepsilon(n)}{m_i}\ \forall j\in\mathcal{N}(i)$. Fix n. Let $p=[p(1),\cdots,p(m_i)]$ be a probability vector in $S_0^i:=$ the simplex of probability vectors in $\mathcal{R}^{m_i}$ with each component $\geq\frac{\varepsilon(n)}{m_i}$ (in particular, $p_n(\cdot)\in S_0^i$). Let $y=[y_1,\cdots,y_{m_i}]^T\in\mathcal{R}^{m_i}$ satisfy $\|y\|_2=1$ and $y\perp\mathbf{1}$ (i.e., $\sum_i y_i=0$). Then

$$y^T\bar{Q}_i(x(n))y\geq\min_{p\in S_0^i}\left(\sum_{j\in\mathcal{N}(i)}p(j)y_j^2-\left(\sum_j p(j)y_j\right)^2\right).$$

The function of $p(\cdot)$ in parentheses on the right is concave in $p(\cdot)$ for a fixed x and will achieve its minimum at some corner of S_0^i, say (without loss of generality) at

$$p:=\left[1-\frac{(m_i-1)\varepsilon(n)}{m_i},\frac{\varepsilon(n)}{m_i},\cdots,\frac{\varepsilon(n)}{m_i}\right]. \quad (31)$$

Then

$$\begin{aligned}y^T\bar{Q}_i(x(n))y &\geq (1-\varepsilon(n))y_1^2+\frac{\varepsilon(n)}{m_i}\sum_i y_j^2-\\ &\quad\left((1-\varepsilon(n))y_1+\frac{\varepsilon(n)}{m_i}\sum_i y_i\right)^2\\ &= ((1-\varepsilon(n))-(1-\varepsilon(n))^2)y_1^2+\frac{\varepsilon(n)}{m_i}\\ &\geq \frac{\varepsilon(n)}{m_i},\end{aligned}$$

where we use the identities $\sum_i y_i=0$, $\sum_i y_i^2=1$. This completes the proof.

Appendix III

In this appendix, we provide an example of $\{c(n)\}$ in (5). Let $c(n)=\frac{1}{1+(n+1)\log(n+1)}$ in (5). Then we have

$$\begin{aligned}\varepsilon(n) &= \prod_{k=1}^{n}\Big(1-\frac{1}{1+(k+1)\log(k+1)}\Big)\varepsilon(0)\\ &< \exp\Big(-\sum_{k=1}^{n}\frac{1}{1+(k+1)\log(k+1)}\Big)\varepsilon(0)\\ &< \exp\Big(-\log\log n\Big)\upsilon\varepsilon(0)\\ &= \frac{\upsilon\varepsilon(0)}{\log n}\end{aligned}$$

for some $\upsilon>0$. Thus $\varepsilon(n)=O\left(\frac{1}{\log n}\right)$. Next we show that $\varepsilon(n)=\Omega\left(\frac{1}{\log n}\right)$. For this we use the fact for $x\in(0,1)$,

$$\log\left(\frac{1}{1-x}\right)\leq\frac{x}{1-x}\Longrightarrow 1-x\geq e^{-\frac{x}{1-x}}.$$

Letting $\varepsilon(0)=1$ without loss of generality,

$$\begin{aligned}\varepsilon(n) &= \prod_{k=1}^{n}\left(1-\frac{1}{1+(k+1)\log(k+1)}\right)\\ &\geq \prod_{k=1}^{n}e^{-\frac{p_k}{1-p_k}}\ \text{for}\ p_k:=\frac{1}{1+(k+1)\log(k+1)}\\ &= e^{-\sum_{k=1}^{n}\frac{p_k}{1-p_k}}.\end{aligned}$$

As $p\downarrow 0$, $\frac{p}{1-p}=p(1+o(1))$. Thus

$$\varepsilon(n)\geq e^{-\sum_{k=1}^{n}p_k(1+o(1))}.$$

But

$$\begin{aligned}\sum_{k=1}^{n}p_k &\leq p_1+\int_0^n\frac{1}{1+(1+y)\log(1+y)}dy\\ &\leq \log\log(n+1)+\log C'\end{aligned}$$

for suitable $C'>0$. Hence for suitable $C>0$,

$$\begin{aligned}\varepsilon(n) &\geq Ce^{-(1+o(1))\sum_{k=1}^{n}p_k}\\ &\geq Ce^{-(1+\epsilon(n))(\log\log(n+1))}\\ &\quad\text{where}\ \epsilon(n)\overset{n\uparrow\infty}{\to}0,\\ &= \frac{C}{(\log(n+2))^{1+\epsilon(n)}}.\end{aligned}$$

That is, $\varepsilon(n)=\Theta((\log n)^{-1})$.

Using the above, it is easy to verify that $\{c(n)\}$ satisfies the stipulated conditions.

REFERENCES

[1] H. B. Ammar, K. Tuyls and M. Kaisers, M. "Evolutionary dynamics of ant colony optimization", *Proc. German Conference on Multiagent System Technologies*, Springer, Berlin-Heidelberg, 2012, 40-52.

[2] W. B. Arthur, *Increasing Returns and Path Dependence in the Economy*, The University of Michigan Press, Ann Arbor, MI, 1994.

[3] W. B. Arthur, Y. M. Ermoliev and Y. M. Kaniovski, "Strong laws for a class of path-dependent stochastic processes with applications", in *Proc. Intl. Conf. on Stochastic optimization, Kiev 1984* (V. Arkin, A. Shiryayev and R. Wets, eds.), Springer Verlag, Berlin-Heidelberg, 1986, 287-300.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem", *Machine learning*, 47(2-3), 2002, 235-256.

[5] K. Avrachenkov and V. S. Borkar, "Metastability in Stochastic Replicator Dynamics", *Dynamic Games and Applications*, 9(2), 2019, 366-390.

[6] K. E. Avrachenkov, J. A. Filar and P. G. Howlett, *Analytic Perturbation Theory and Its Applications*, SIAM, 2013.

[7] M. Benaim, "Vertex-reinforced random walks and a conjecture of Pemantle", *Annals of Probability* 25(1), 1997, 361-392.

[8] M. Benaim and P. Tarres, "Dynamics of vertex-reinforced random walks", *Annals of Probability* 39(6), 2011, 2178-2223.

[9] M. Benaim and O. Raimond, "A class of self-interacting processes with applications to games and reinforced random walks", *SIAM Journal on Control and Optimization*, 48(7), 2010, 4707-4730.

[10] M. Benaim, O. Raimond and B. Schapira, "Strongly Vertex-Reinforced-Random-Walk on the complete graph", *ALEA Lat. Am. J. Probab. Math. Stat.*, 10(2), 2013, 767-782.

[11] V. S. Borkar, "Avoidance of traps in stochastic approximation", *Systems and Control Letters* 50(1), 2003, 1-9.

[12] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems View*, Hindustan Publishing Agency, New Delhi, and Cambridge University Press, Cambridge, UK, 2008.

[13] V. S. Borkar and D. Das, "A novel ACO algorithm for optimization via reinforcement and initial bias", *Swarm Intelligence*, 3(1), 2009, 3-34.

[14] C. S. Bouttier and I. Gavra, "Convergence rate of a simulated annealing algorithm with noisy observations", *The Journal of Machine Learning Research*, 20(1), 2019, 127-171.

[15] O. Brandiere and M. Duflo, M., "Les algorithmes stochastiques contournent - ils les pieges?" *Annales de l'IHP Probabilités et Statistiques* 32(3), 1996, 395-427.

[16] O. Catoni, "Rough large deviation estimates for simulated annealing: Application to exponential schedules", *Annals of Probability* 20(3), 1992, 1109-1146.

[17] C. P. Chamley, *Rational Herds: Econmic Models of Social Learning*, Cambridge University Press, Cambridge, UK, 2004.

[18] J. Cohen, A. Heliou and P. Mertikopoulos, "Learning with bandit feedback in potential games", *Advances in Neural Information Processing Systems* 30, 2017, 6369-6378.

[19] C. Dempsey, "Join the crowdsourced effort to search for the missing Malaysian Airlines flight", March 10, 2014, https://www.geographyrealm.com/join-crowdsourced-effort-search-missing-malaysian-airlines-flight/

[20] S. B. Gelfand and S. K. Mitter, "Simulated annealing with noisy or imprecise energy measurements", *Journal of Optimization Theory and Applications* 62(1), 1989, 49-62.

[21] W. J. Gutjahr and G.C Pflug "Simulated annealing for noisy cost functions", *Journal of Global Optimization*, 8(1), 1996, 1-13.

[22] B. Hajek, "Cooling schedules for optimal annealing", *Mathematics of Operations Research* 13(2), 1988, 311-329.

[23] P. Karmakar and S. Bhatnagar, "Dynamics of stochastic approximation with iterate-dependent Markov noise under verifiable conditions in compact state space with the stability of iterates not ensured", *IEEE Trans. on Automatic Control*, 2022 (to appear, available online).

[24] T. Lattimore and C. Szepesvári, *Bandit algorithms*, Cambridge University Press, Cambridge, UK.

[25] Y. Matsumoto, *An Introduction to Morse Theory*, Trans. of Mathematical Monographs No. 208, American Math. Society, Providence, RI.

[26] E. Pariser, *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[27] R. Pemantle, "Nonconvergence to unstable points in urn models and stochastic approximations", *The Annals of Probability* 18(2), 1990, 698-712.

[28] W. H. Sandholm, *Population Games and Evolutionary Dynamics*, MIT Press, Cambridge, Mass., 2010.

[29] A. Sankararaman, A. Ganesh and S. Shakkottai, "Social learning in multi agent multi armed bandits", *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3(3), 2019, 1-35.

[30] V. Shah, J. Blanchet and R. Johari, "Bandit learning with positive externalities", *Advances in Neural Information Processing Systems* 31, 2018, 4918-4928.

[31] V. Yaji and S. Bhatnagar, "Stochastic recursive inclusions in two timescales with non-additive iterate dependent Markov noise", *Math. Op. Research* 45(4), 2020, 1405-1444.

[32] H. P. Young, "The evolution of conventions", *Econometrica* 61(1), 1993, 57-84.

[33] H. P. Young, *Individual Strategy and Social Structure*, Princeton University Press, Princeton, NJ, 1998.