SIAM J. CONTROL OPTIM.
Vol. 53, No. 2, pp. 1056–1081

ONLINE LEARNING OF NASH EQUILIBRIA IN CONGESTION GAMES[*]

WALID KRICHENE[†], BENJAMIN DRIGHÈS[‡], AND ALEXANDRE M. BAYEN[§]

Abstract. We study the repeated, nonatomic congestion game, in which multiple populations of players share resources and make, at each iteration, a decentralized decision on which resources to utilize. We investigate the following question: given a model of how individual players update their strategies, does the resulting dynamics of strategy profiles converge to the set of Nash equilibria of the one-shot game? We consider in particular a model in which players update their strategies using algorithms with sublinear discounted regret. We show that the resulting sequence of strategy profiles converges to the set of Nash equilibria in the sense of Cesàro means. However, convergence of the actual sequence is not guaranteed in general. We show that it can be guaranteed for a class of algorithms with a sublinear discounted regret and which satisfy an additional condition. We call such algorithms AREP (approximate replicator) algorithms, as they can be interpreted as a discrete-time approximation of the replicator equation, which models the continuous-time evolution of population strategies, and which is known to converge for the class of congestion games.

Key words. online learning, population dynamics, regret minimization, congestion games, nash equilibria

AMS subject classifications. 68W27, 91A13, 62L20, 90C25

DOI. 10.1137/140980685

1. Introduction.

Congestion games are noncooperative games that model the interaction of players who share resources. Each player makes a decision on which resources to utilize. The individual decisions of players result in a resource allocation at the population scale. Resources which are highly utilized become congested, and the corresponding players incur higher losses. For example, in routing games—a subclass of congestion games—the resources are edges in a network, and each player needs to travel from a given source vertex to a given destination vertex on the graph. Each player chooses a path, and the joint decision of all players determines the congestion on each edge. The more a given edge is utilized, the more congested it is, creating delays for those players using that edge.

The one-shot congestion game has been studied extensively, and a comprehensive introduction is given, for example, in [25]. In particular, congestion games are shown to be convex potential games, thus their Nash equilibria can be expressed as the solution to a convex optimization problem. Characterizing the Nash equilibria of the congestion game gives useful insights, such as the loss of efficiency due to the selfishness of players. One popular measure of inefficiency is the price of anarchy, introduced by Koutsoupias and Papadimitriou in [20] and studied in the case of routing games by Roughgarden and Tardos in [26]. While characterizing Nash equilibria of the one-shot game gives

[*] Received by the editors August 5, 2014; accepted for publication (in revised form) February 9, 2015; published electronically April 21, 2015. This work was supported in part by FORCES (Foundations of Resilient Cyber-Physical Systems), which receives support from the National Science Foundation (NSF awards CNS-1238959, CNS-1238962, CNS-1239054, CNS-1239166).
http://www.siam.org/journals/sicon/53-2/98068.html

[†] Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA 94720 (walid@eecs.berkeley.edu).

[‡] Ecole Polytechnique, Palaiseau 91128, France (benjamin.drighes@polytechnique.edu).

[§] Department of Electrical Engineering and Computer Sciences and Department of Civil and Environmental Engineering, UC Berkeley, Berkeley, CA 94720 (bayen@berkeley.edu).

many insights, it does not model how players *arrive to the equilibrium*. Studying the game in a repeated setting can help answer this question. Additionally, most realistic scenarios do not correspond to a one-shot setting but rather to a repeated setting in which players make decisions in an online fashion, observe outcomes, and may update their strategies given the previous outcomes. This motivates the study of the game and the population dynamics in an online learning framework.

Arguably, a good model for learning should be distributed and should not have extensive information requirements. In particular, one should not expect the players to have an accurate model of congestion of the different resources. Players should be able to learn simply by observing the outcomes of their previous actions and those of other players. No-regret learning is of particular interest here, as many regret-minimizing algorithms are easy to implement by individual players and only require the player losses to be revealed; see, for example, [10] and the references therein. The Hedge algorithm (also known as the multiplicative weights algorithm [1], or the exponentiated gradient method [18]) is a famous example of regret-minimizing algorithms. It was introduced to the machine learning community by Freund and Schapire in [14], a generalization of the weighted majority algorithm of Littlestone and Warmuth [21]. The Hedge algorithm will be central in our discussion, as it will motivate the study of the continuous-time replicator equation, and will eventually be shown to converge for congestion games.

No-regret learning and its resulting population dynamics have been studied in the context of routing games, a special case of congestion games [6, 5, 19]. For example, in [5], Blum, Even-Dar, and Ligett show that the sequence of strategy profiles converges to the set of ϵ-approximate Nash equilibria on a $(1-\epsilon)$-fraction of days. They also give explicit convergence rates which depend on the maximum slopes of the congestion functions. In [19], Kleinberg, Piliouras, and Tardos study the problem of online learning in atomic congestion games with finitely many players. Although the setting is different (we study nonatomic congestion games, which involve populations of infinitely many players), the problems are closely related. In particular, the authors in [19] make a connection between the discrete-time Hedge algorithm and the continuous-time replicator dynamics. We build on this connection, and previous results by Fischer and Vöcking [11] on convergence of the replicator dynamics, to prove stronger convergence results for a class of discrete-time dynamics, which includes, in particular, the Hedge algorithm.

Continuous-time dynamics have also been studied for several classes of population games and for congestion games in particular; see, for example, [29]. In [27], Sandholm studies convergence for the class of potential games. He shows that dynamics which satisfy a positive correlation condition with respect to the potential function of the game converge to the set of stationary points of the vector field (usually, a superset of Nash equilibria). In [16], Hofbauer and Sandholm study the convergence of EPT dynamics for the class of stable games. In [12], Fox and Shamma extend these convergence results to passive evolutionary dynamics and give a dynamical systems interpretation. While our discussion is mainly concerned with discrete-time dynamics, properties of continuous-time evolutionary dynamics will be used in our analysis, in particular convergence of solutions of the replicator ODE.

We will consider a model in which the losses are discounted over time, using a vanishing sequence of discount factors $(\gamma_\tau)_{\tau\in\mathbb{N}}$. This defines a discounted regret, and we will focus our attention on online learning algorithms with sublinear discounted regret. The sequence of discount factors will have several interpretations beyond its economic motivation. For example, we will observe that the Hedge algorithm has sublinear discounted regret if we use the sequence $(\gamma_\tau)_\tau$ as learning rates.

After defining the model and giving preliminary results in sections 2 and 3, we show in section 4 that when players use online learning algorithms with sublinear discounted regret, the sequence of strategy profiles converges to the set of Nash equilibria in the Cesàro sense. In order to obtain strong convergence, we first motivate the study of the replicator dynamics. Indeed, it can be viewed as a continuous-time limit of the Hedge algorithm with decreasing learning rates. In section 5, we recall the convergence result of the replicator dynamics. By discretizing the replicator equation (using the same discount sequence $(\gamma_\tau)_{\tau\in\mathbb{N}}$ as discretization time steps) we obtain a multiplicative-weights update rule with sublinear discounted regret, which we call the REP (replicator) algorithm. Finally, in section 6, we define a class of online learning algorithms we call the AREP (approximate replicator) algorithms, which can be expressed as a discrete REP algorithm with perturbations that satisfy a condition given in Definition 6.8. Using results from the theory of stochastic approximation, we show that strong convergence is guaranteed for AREP algorithms with sublinear discounted regret. We finally observe that both the REP algorithm and the Hedge algorithm belong to this class, which proves convergence for these two algorithms in particular.

2. The congestion game model. In the congestion game, a finite set $\mathcal{R}$ of resources is shared by a set $\mathcal{X}$ of players. The set of players is endowed with a structure of measure space, $(\mathcal{X}, \mathcal{M}, m)$, where $\mathcal{M}$ is a σ-algebra of measurable subsets, and m is a finite Lebesgue measure. The measure is nonatomic in the sense that single-player sets are null-sets for m. The player set is partitioned into K populations, $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$. For all k, the total mass of population $\mathcal{X}_k$ is assumed to be finite and nonzero. Each player $x \in \mathcal{X}_k$ has a task to perform, characterized by a collection of bundles $\mathcal{P}_k \subset \mathcal{P}$, where $\mathcal{P}$ is the power set of $\mathcal{R}$. The task can be accomplished by choosing any bundle of resources $p \in \mathcal{P}_k$. The action set of any player in $\mathcal{X}_k$ is then simply $\mathcal{P}_k$.

The joint actions of all players can be represented by an action profile $a : \mathcal{X} \to \mathcal{P}$ such that for all $x \in \mathcal{X}_k$, $a(x) \in \mathcal{P}_k$ is the bundle of resources chosen by player x. The function $x \mapsto a(x)$ is assumed to be $\mathcal{M}$-measurable ($\mathcal{P}$ is equipped with the counting measure). The action profile a determines the bundle loads and resource loads, defined as follows: for all $k \in \{1, \dots, K\}$ and $p \in \mathcal{P}_k$, the load of bundle p under population $\mathcal{X}_k$ is the total mass of players in $\mathcal{X}_k$ who chose that bundle

$$f_p^k(a) = \int_{x\in\mathcal{X}_k} 1_{(a(x)=p)} dm(x). \tag{2.1}$$

For any $r \in \mathcal{R}$, the resource load is defined to be the total mass of players utilizing r

$$\phi_r(a) = \sum_{k=1}^{K} \sum_{p\in\mathcal{P}_k : r\in p} f_p^k(a). \tag{2.2}$$

The resource loads determine the losses of all players: the loss associated to a resource r is given by $c_r(\phi_r(a))$, where the congestion functions c_r are assumed to satisfy the following.

Assumption 2.1. The congestion functions c_r are nonnegative, nondecreasing, Lipschitz-continuous functions.

The total loss of a player x such that $a(x) = p$ is $\sum_{r\in p} c_r(\phi_r(a))$. The congestion model is given by the tuple $(K, (\mathcal{X}_k)_{1\le k\le K}, \mathcal{R}, (\mathcal{P}_k)_{1\le k\le K}, (c_r)_{r\in\mathcal{R}})$. The congestion game is determined by the action set and the loss function for every player: for all $x \in \mathcal{X}_k$, the action set of x is $\mathcal{P}_k$, and the loss function of x, given the action profile a, is $\sum_{r\in a(x)} c_r(\phi_r(a))$.

2.1. A macroscopic view. The action profile a specifies the bundle of each player x. A more concise description of the joint action of players is given by the bundle distribution: the proportion of players choosing bundle p in population $\mathcal{X}_k$ is denoted by $\mu_p^k(a) = f_p^k(a)/m(\mathcal{X}_k)$, which defines a bundle distribution for population $\mathcal{X}_k$, $\mu^k(a) = (\mu_p^k(a))_{p\in\mathcal{P}_k} \in \Delta^{\mathcal{P}_k}$, and a bundle distribution across populations, given by the product distribution $\mu(a) = (\mu^1(a), \dots, \mu^K(a)) \in \Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_K}$. We say that the action profile a induces the distribution $\mu(a)$. Here $\Delta^{\mathcal{P}_k}$ denotes the simplex of distributions over $\mathcal{P}_k$, that is, $\Delta^{\mathcal{P}_k} = \{\mu \in \mathbb{R}_+^{\mathcal{P}_k} : \sum_{p\in\mathcal{P}_k} \mu_p = 1\}$.

The product of simplexes $\Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_K}$ will be denoted by Δ. This macroscopic representation of the joint actions of players will be useful in our analysis. We will also view the resource loads as linear functions of the product distribution $\mu(a)$. Indeed, we have from (2.2) and the definition of $\mu_p^k(a)$

$$\phi_r(a) = \sum_{k=1}^{K} m(\mathcal{X}_k) \sum_{p\in\mathcal{P}_k : r\in p} \mu_p^k(a) = \sum_{k=1}^{K} m(\mathcal{X}_k)(M^k\mu^k(a))_r,$$

where for all k, $M^k \in \mathbb{R}^{\mathcal{R}\times\mathcal{P}_k}$ is an incidence matrix defined as follows: for all $r \in \mathcal{R}$ and all $p \in \mathcal{P}_k$,

$$M_{r,p}^k = \begin{cases} 1 & \text{if } r \in p, \\ 0 & \text{otherwise.} \end{cases}$$

We write in vector form $\phi(a) = \sum_{k=1}^K m(\mathcal{X}_k)M^k\mu^k(a)$, and by defining the scaled incidence matrix $\bar{M} = \big(\; m(\mathcal{X}_1)M^1| \dots |m(\mathcal{X}_K)M^K \; \big)$, we have $\phi(a) = \bar{M}\mu(a)$

By abuse of notation, the dependence on the action profile a will be omitted, so we will write μ instead of $\mu(a)$ and ϕ instead of $\phi(a)$. Finally, we define the loss function of a bundle $p \in \mathcal{P}_k$ to be

$$\ell_p^k(\mu) = \sum_{r\in p} c_r(\phi_r) = \sum_{r\in p} c_r((\bar{M}\mu)_r) = M^\top c(\bar{M}\mu), \tag{2.3}$$

where M is the incidence matrix $M = \big(\; M^1| \quad \dots \quad |M^K \; \big)$ and $c(\phi)$ is the vector $(c_r(\phi_r))_{r\in\mathcal{R}}$. We denote by $\ell^k(\mu)$ the vector of losses $(\ell_p^k(\mu))_{p\in\mathcal{P}_k}$ and by $\ell(\mu)$ the K-tuple $\ell(\mu) = (\ell^1(\mu), \dots, \ell^K(\mu))$.

2.2. Nash equilibria of the congestion game. We can now define and characterize the Nash equilibria of the congestion game, also called Wardrop equilibria, in reference to [28].

Definition 2.2 (Nash equilibrium). *A product distribution μ is a Nash equilibrium of the congestion game if for all k, and all $p \in \mathcal{P}_k$ such that $\mu_p^k > 0$, $\ell_{p'}^k(\mu) \geq \ell_p^k(\mu)$ for all $p' \in \mathcal{P}_k$. The set of Nash equilibria will be denoted by $\mathcal{N}$.*

In finite player games, a Nash equilibrium is defined to be an action profile a such that no player has an incentive to unilaterally deviate [23], that is, no player can strictly decrease her loss by unilaterally changing her action. We show that this condition (referred to as the Nash condition) holds for *almost all players* whenever μ is a Nash equilibrium in the sense of Definition 2.2.

Proposition 2.3. *A distribution μ is a Nash equilibrium if and only if for any joint action a which induces the distribution μ, almost all players have no incentive to unilaterally deviate from a.*

Proof. First, we observe that, given an action profile a, when a single player x changes her strategy, this does not affect the distribution μ. This follows from the definition of the distribution, $\mu^k_p = \frac{1}{m(\mathcal{X}_k)} \int_{\mathcal{X}_k} 1_{(a(x)=p)} dm(x)$. Changing the action profile a on a null-set $\{x\}$ does not affect the integral.

Now, assume that almost all players have no incentive to unilaterally deviate. That is, for all k, for almost all $x \in \mathcal{X}_k$,

$$\forall p' \in \mathcal{P}_k,\ \ell^k_{p'}(\mu') \geq \ell^k_{a(x)}(\mu), \tag{2.4}$$

where μ' is the distribution obtained when x unilaterally changes her bundle from $a(x)$ to p'. By the previous observation, $\mu' = \mu$. As a consequence, condition (2.4) becomes for almost all x, and for all p', $\ell^k_{p'}(\mu) \geq \ell^k_{a(x)}(\mu)$. Therefore, integrating over the set $\{x \in \mathcal{X}_k : a(x) = p\}$, we have for all k, $\ell^k_{p'}(\mu)\mu^k_p \geq \ell^k_p(\mu)\mu^k_p$ for all p', which implies that μ is a Nash equilibrium in the sense of Definition 2.2. Conversely, if a is an action profile, inducing distribution μ, such that the Nash condition does not hold for a set of players with positive measure, then there exists k_0 and a subset $X \subset \mathcal{X}_{k_0}$ with $m(X) > 0$, such that every player in X can strictly decrease her loss by changing her action. Let $X_p = \{x \in X : a(x) = p\}$; then X is the disjoint union $X = \cup_{p \in \mathcal{P}_k} X_p$, and there exists p_0 such that $m(X_{p_0}) > 0$. Therefore

$$\mu^{k_0}_{p_0} = \frac{m\left(\{x \in \mathcal{X}_{k_0} : a(x) = p_0\}\right)}{m(\mathcal{X}_{k_0})} \geq \frac{m(X_{p_0})}{m(\mathcal{X}_{k_0})} > 0.$$

Let $x \in X_{p_0}$. Since x can strictly decrease her loss by unilaterally changing her action, there exists p_1 such that $\ell^{k_0}_{p_1}(\mu) < \ell^{k_0}_{a(x)}(\mu) = \ell^{k_0}_{p_0}(\mu)$. But since $\mu^{k_0}_{p_0} > 0$, μ is not a Nash equilibrium. □

Definition 2.2 also implies that, for a population $\mathcal{X}_k$, all bundles with nonzero mass have equal losses, and bundles with zero mass have greater losses. Therefore almost all players incur the same loss.

2.3. Mixed strategies. The Nash equilibria we have described so far are *pure strategy* equilibria, since each player x deterministically plays a single action $a(x)$. We now extend the model to allow mixed strategies. That is, the action of a player x is a random variable $A(x)$ with distribution $\pi(x)$ and with realization $a(x)$.

We show that when players use mixed strategies, provided they randomize independently, the resulting Nash equilibria are, in fact, the same as those given in Definition 2.2. The key observation is that under independent randomization, the resulting bundle distributions μ^k are random variables with zero variance, and thus they are essentially deterministic.

To formalize the probabilistic setting, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A mixed strategy profile is a function $A : \mathcal{X} \to \Omega \to \mathcal{P}$, such that for all k and all $x \in \mathcal{X}_k$, $A(x)$ is a $\mathcal{P}_k$-valued random variable, such that the mapping $(x, \omega) \mapsto A(x)(\omega)$ is $\mathcal{M} \times \mathcal{F}$-measurable. For all $x \in \mathcal{X}_k$ and $p \in \mathcal{P}_k$, let $\pi^k_p(x) = \mathbb{P}[A(x) = p]$. Similarly to the deterministic case, the mixed strategy profile A determines the bundle distributions μ^k, which are, in this case, random variables, as we recall that $\mu^k_p = \frac{1}{m(\mathcal{X}_k)} \int_{\mathcal{X}_k} 1_{(A(x)=p)} dm(x)$.

Nevertheless, assuming players randomize independently, the bundle distribution is almost surely equal to its expectation, as stated in the following proposition. The assumption of independent randomization is a reasonable one, since players are noncooperative.

Proposition 2.4. *Under independent randomization,*

$$\forall k, \text{ almost surely, } \mu^k = \mathbb{E}[\mu^k] = \frac{1}{m(\mathcal{X}_k)} \int_{\mathcal{X}_k} \pi^k(x) dm(x).$$

Proof. Fix k and let $p \in \mathcal{P}_k$. Since $(x, \omega) \mapsto 1_{(A(x)=p)}(\omega)$ is a nonnegative bounded $\mathcal{M} \times \mathcal{F}$-measurable function, we can apply Tonelli's theorem and write

$$\begin{aligned}\mathbb{E}\left[\mu_p^k\right] &= \mathbb{E}\left[\frac{1}{m(\mathcal{X}_k)} \int_{\mathcal{X}_k} 1_{(A(x)=p)} dm(x)\right] = \frac{1}{m(\mathcal{X}_k)} \int_{\mathcal{X}_k} \mathbb{E}\left[1_{(A(x)=p)}\right] dm(x) \\ &= \frac{1}{m(\mathcal{X}_k)} \int_{\mathcal{X}_k} \pi_p^k(x) dm(x).\end{aligned}$$

Similarly,

$$\begin{aligned}m(\mathcal{X}_k)^2 \operatorname{var}\left[\mu_p^k\right] &= \mathbb{E}\left(\int_{\mathcal{X}_k} 1_{(A(x)=p)} dm(x)\right)^2 - \left(\int_{\mathcal{X}_k} \pi_p^k(x) dm(x)\right)^2 \\ &= \int_{\mathcal{X}_k} \int_{\mathcal{X}_k} \mathbb{E}\, 1_{(A(x)=p; A(x')=p)} dm(x) dm(x') - \int_{\mathcal{X}_k} \int_{\mathcal{X}_k} \pi_p^k(x) \pi_p^k(x') dm(x) dm(x') \\ &= \int_{\mathcal{X}_k \times \mathcal{X}_k} \left(\mathbb{P}[A(x) = p; A(x') = p] - \pi_p^k(x) \pi_p^k(x')\right) d(m \times m)(x, x').\end{aligned}$$

Then observing that the diagonal $D = \{(x, x) \colon x \in \mathcal{X}_k\}$ is an $(m \times m)$-nullset (this follows, for example, from Proposition 251T in [13]), we can restrict the integral to the set $\mathcal{X}_k \times \mathcal{X}_k \setminus D$, on which $\mathbb{P}[A(x) = p; A(x') = p] = \pi_p^k(x)\pi_p^k(x')$, by the independent randomization assumption. This proves that $\operatorname{var}\left[\mu_p^k\right] = 0$. Therefore $\mu_p^k = \mathbb{E}[\mu_p^k]$ almost surely. □

2.4. The Rosenthal potential function. We now discuss how one can formulate the set of Nash equilibria as the solution of a convex optimization problem. Consider the function

$$V(\mu) = \sum_{r \in \mathcal{R}} \int_0^{(\bar{M}\mu)_r} c_r(u) du, \tag{2.5}$$

defined on the product of simplexes $\Delta^{\mathcal{P}_1} \times \cdots \times \Delta^{\mathcal{P}_K}$, which will be denoted by Δ. V is called the Rosenthal potential function and was introduced in [24] for the congestion game with finitely many players and later generalized to the infinite-players case. It can be viewed as the composition of the function $\bar{V} : \phi \in \mathbb{R}_+^{\mathcal{R}} \mapsto \sum_{r \in \mathcal{R}} \int_0^{\phi_r} c_r(u) du$ and the linear function $\mu \mapsto \bar{M}\mu$. Since for all r, c_r is, by assumption, nonnegative, $\bar{V}$ is differentiable and nonnegative and $\nabla \bar{V}(\phi) = (c_r(\phi_r))_{r \in \mathcal{R}}$. And since c_r are nondecreasing, $\bar{V}$ is convex. (One way to see this is by Taylor's theorem: for all ϕ^0, ϕ, t such that $\phi^0 \in \mathbb{R}_+^{\mathcal{R}}$ and $\phi^0 + t\phi \in \mathbb{R}_+^{\mathcal{R}}$, there exists t' between 0 and t such that $\bar{V}(\phi^0 + t\phi) = \bar{V}(\phi^0) + t\left\langle \nabla \bar{V}(\phi^0 + t'\phi), \phi \right\rangle \geq \bar{V}(\phi^0) + t\left\langle \nabla \bar{V}(\phi^0), \phi \right\rangle$; thus $\bar{V}$ satisfies the first-order convexity condition. See, for example, [7, Section 3.1].) Therefore V is convex as the composition of a convex and a linear function.

A simple application of the chain rule gives $\nabla V(\mu) = \bar{M}^\top c(\bar{M}\mu)$. If we denote by $\nabla_{\mu^k} V(\mu)$ the vector of partial derivatives with respect to μ_p^k, $p \in \mathcal{P}_k$, we have $\nabla_{\mu^k} V(\mu) = m(\mathcal{X}_k) {M^k}^\top c(\bar{M}\mu) = m(\mathcal{X}_k) \ell^k(\mu)$. Thus,

$$\forall k, \ \forall p \in \mathcal{P}_k, \quad \frac{\partial V}{\partial \mu_p^k}(\mu) = m(\mathcal{X}_k) \ell_p^k(\mu), \tag{2.6}$$

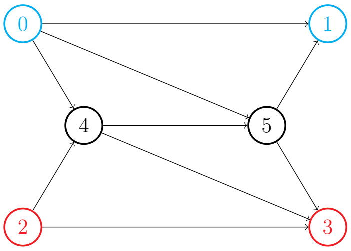

FIG. 1. *Routing game with two populations of players.*

and V is a potential function for the congestion game, as defined in [27], for example. Next, we show the relationship between the set of Nash equilibria and the potential function V.

THEOREM 2.5 (Rosenthal [24]). *$\mathcal{N}$ is the set of minimizers of V on the product of simplexes Δ. It is a nonempty convex compact set. We will denote by $V_{\mathcal{N}}$ the value of V on $\mathcal{N}$.*

Since the set of Nash equilibria can be expressed as the solution to a convex optimization problem, it can be computed in polynomial time in the size of the problem. Beyond computing Nash equilibria, we seek to model how players arrive at the set $\mathcal{N}$. This is discussed in section 3. But first, we define routing games, a special case of congestion games.

2.5. Example: Routing games.

A routing game is a congestion game with an underlying graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V}$ and edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. In this case, the resource set is equal to the edge set, $\mathcal{R} = \mathcal{E}$. Routing games are used to model congestion on transportation or communication networks. Each population $\mathcal{X}_k$ is characterized by a common source vertex $s_k \in \mathcal{V}$ and a common destination vertex $t_k \in \mathcal{V}$. In a transportation setting, players represent drivers traveling from s_k to t_k; in a communication setting, players send packets from s_k to t_k. The action set $\mathcal{P}_k$ is a set of paths connecting s_k to t_k. In other words, each player chooses a path connecting his or her source and destination vertices. The bundle load f^k_p is then called the flow on path p. The resource load ϕ_r is called the total edge flow. Finally, the congestion functions $\phi_r \mapsto c_r(\phi_r)$ determine the delay (or latency) incurred by each player.

We will use the routing game given in Figure 1 as an example to illustrate the convergence result of section 6. In this example, two populations of players share the network; the first population sends packets from v_0 to v_1, and the second population from v_2 to v_3. The paths (bundles) available to each population are given by $\mathcal{P}_1 = \{(v_0, v_1), (v_0, v_4, v_5, v_1), (v_0, v_5, v_1)\}$, $\mathcal{P}_2 = \{(v_2, v_3), (v_2, v_4, v_5, v_3), (v_2, v_4, v_3)\}$.

3. Online learning in congestion games.

We now describe the online learning framework for the congestion game and present the Hedge algorithm in particular.

3.1. The online learning framework.

Suppose that the game is played repeatedly for infinitely many iterations, indexed by $\tau \in \mathbb{N}$. During iteration τ, each player chooses a bundle simultaneously. The decision of all players can be represented, as defined above, by an action profile $a^{(\tau)} : \mathcal{X} \to \mathcal{P}$. This induces, at the level of each population $\mathcal{X}_k$, a bundle distribution $\mu^{k^{(\tau)}}$. These, in turn, determine the resource loads and the bundle losses $\ell^k_p(\mu^{(\tau)})$. The losses for bundles $p \in \mathcal{P}_k$ are revealed to all players in population $\mathcal{X}_k$, which marks the end of iteration τ. Players can then use this information to update their strategies before the start of the next iteration.

A note on the information assumptions. Here, we assume that at the end of the iteration, a player observes the losses of all bundles $p \in \mathcal{P}_k$. Instead, one could assume that a player can only observe the losses she incurs. This is often called the multiarmed-bandit setting, in reference to armed-bandit slot machines, in which a gambler can choose, at each iteration, one machine to play and is only revealed the loss of that machine. Making this restriction requires players to use additional exploration of bundles. A comprehensive presentation of online learning algorithms in the multiarmed-bandit setting, both stochastic and deterministic, can be found, for example, in [2]. Regret bounds are also given in [10, Section 6.7, pp. 156–159], and [9, 8], as well as [3, 15] for the online shortest path problem. We choose to use the full feedback assumption to simplify our discussion, leaving the multiarmed-bandit setting as a possible extension. We believe this is a reasonable model in many games, since bundle losses could be announced publicly. In the special case of routing games, this can be achieved by having a central authority measure and announce the delays. This is particularly true in transportation networks, in which many agencies and online services measure delays and make this information publicly available. Assuming the full vector of bundle losses is revealed does not mean, however, that players have access to the individual resource loads $\phi_r^{(\tau)}$, or to the congestion functions $c_r(\cdot)$, which is consistent with our initial argument that, in a realistic model, players should only rely on the observed value of the bundle losses.

Each player $x \in \mathcal{X}_k$ is assumed to draw her bundle from a randomized strategy $\pi^{(\tau)}(x) \in \Delta^{\mathcal{P}_k}$ (the deterministic case is a special case in which $\pi^{(\tau)}(x)$ is a vertex on the simplex, i.e., a pure strategy). As discussed in section 2.3, players randomize independently. At the end of iteration τ, player x updates her strategy using an *update rule* or *online learning algorithm*, as defined below.

Definition 3.1 (online learning algorithm for the congestion game). *An online learning algorithm (or update rule) for the congestion game, applied by a player $x \in \mathcal{X}_k$, is a sequence of functions $\left({}^xU^{(\tau)}\right)_{\tau\in\mathbb{N}}$, fixed a priori, that is, before the start of the game, such that for each τ,*

$$\begin{aligned} {}^xU^{(\tau)} : &\left(\mathbb{R}^{\mathcal{P}_k}\right)^\tau \times \Delta^{\mathcal{P}_k} \to \Delta^{\mathcal{P}_k} \\ &\Big((\ell^k(\mu^{(t)}))_{t\le\tau}, \pi^{(\tau)}(x)\Big) \mapsto \pi^{(\tau+1)}(x) \end{aligned}$$

is a function which maps, given the history of bundle losses $(\ell^k(\mu^{(t)}))_{t\le\tau}$, the strategy on the current day $\pi^{(\tau)}(x)$ to the strategy on the next day $\pi^{(\tau+1)}(x)$. The online learning framework is summarized in Algorithm 1.

Algorithm 1. Online learning framework for the congestion game.

```
For every player $x \in \mathcal{P}_k$, an initial mixed strategy $\pi^{(0)}(x) \in \Delta^{\mathcal{P}_k}$ and an online learning algorithm $({}^xU^{(\tau)})_{\tau\in\mathbb{N}}$
for each iteration $\tau \in \mathbb{N}$ do
    Every player $x$ independently draws a bundle according to her strategy $\pi^{(\tau)}(x)$, i.e., $A^{(\tau)}(x) \sim \pi^{(\tau)}(x)$.
    The vector of bundle losses $\ell^k(\mu^{(\tau)})$ is revealed to all players in $\mathcal{P}_k$. Each player incurs the loss of the bundle she chose.
    Players update their mixed strategies: $\pi^{(\tau+1)}(x) = {}^xU^{(\tau)}((\ell_p^k(\mu^{(t)}))_{t\le\tau}, \pi^{(\tau)}(x))$.
end for
```

We will focus our attention on algorithms which have vanishing upper bounds on the average discounted regret, defined in the next section.

3.2. Discounted regret.

Since the game is played for infinitely many iterations, we assume that the losses of players are discounted over time. This is a common technique in infinite-horizon optimal control, for example, and can be motivated from an economic perspective by considering that losses are devalued over time. We also give an interpretation of discounting in terms of learning rates, as discussed in section 3.4.

Let $(\gamma_\tau)_{\tau\in\mathbb{N}}$ denote the sequence of discount factors. We make the following assumption.

Assumption 3.2. The sequence of discount factors $(\gamma_\tau)_{\tau\in\mathbb{N}}$ is assumed to be positive decreasing with $\lim_{\tau\to\infty}\gamma_\tau = 0$ and $\lim_{T\to\infty}\sum_{\tau=0}^T \gamma_\tau = \infty$.

On iteration τ, a player $x \in \mathcal{X}_k$ who draws an action $A^{(\tau)}(x) \sim \pi^{(\tau)}(x)$ incurs a discounted loss given by $\gamma_\tau \ell^k_{A^{(\tau)}(x)}(\mu^{(\tau)})$, where $\mu^{(\tau)}$ is the distribution induced by the profile $A^{(\tau)}$. The cumulative discounted loss for player x, up to iteration T, is then defined to be

$$L^{(T)}(x) = \sum_{\tau=0}^{T} \gamma_\tau \ell^k_{A^{(\tau)}(x)}(\mu^{(\tau)}). \tag{3.1}$$

We observe that this is a random variable, since the action $A^{(\tau)}(x)$ of player x is random, drawn from a distribution $\pi^{(\tau)}(x)$. The expectation of the cumulative discounted loss is then $\mathbb{E}[L^{(T)}(x)] = \sum_{\tau=0}^T \gamma_\tau\, \mathbb{E}[\ell^k_{A^{(\tau)}(x)}(\mu^{(\tau)})] = \sum_{\tau=0}^T \gamma_\tau \left\langle \pi^{(\tau)}(x), \ell^k(\mu^{(\tau)})\right\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product on $\mathbb{R}^{\mathcal{P}_k}$. Similarly, we define the cumulative discounted loss for a fixed bundle $p \in \mathcal{P}_k$,

$$\mathscr{L}_p^{k\,(T)} = \sum_{\tau=0}^{T} \gamma_\tau \ell^k_p(\mu^{(\tau)}). \tag{3.2}$$

We can now define the discounted regret.

Definition 3.3 (discounted regret). *Let $x \in \mathcal{X}_k$, and consider an online learning algorithm for the congestion game, given by the sequence of functions $\left({}^xU^{(\tau)}\right)_{\tau\in\mathbb{N}}$. Let $(\mu^{(\tau)})_{\tau\in\mathbb{N}}$ be the sequence of distributions, determined by the mixed strategy profile of all players. Then the discounted regret up to iteration T, for player x, under algorithm U, is the random variable*

$$R^{(T)}(x) = L^{(T)}(x) - \min_{p\in\mathcal{P}_k} \mathscr{L}_p^{k\,(T)}. \tag{3.3}$$

The algorithm U is said to have sublinear discounted regret if for any sequence of distributions $(\mu^{(\tau)})_{\tau\in\mathbb{N}}$, and any initial strategy $\pi^{(0)}$,

$$\frac{1}{\sum_{\tau=0}^T \gamma_\tau}\left[R^{(T)}(x)\right]^+ \to 0 \text{ almost surely as } T\to\infty. \tag{3.4}$$

If we have convergence in the L^1-norm, $\frac{1}{\sum_{\tau=0}^T\gamma_\tau}\left[\mathbb{E}\left[R^{(T)}(x)\right]\right]^+ \to 0$, we say that the algorithm has sublinear discounted regret in expectation.

We observe that in the definition of the regret, one can replace the minimum over the set $\mathcal{P}_k$ by a minimum over the simplex $\Delta^{\mathcal{P}_k}$, $\min_{p\in\mathcal{P}_k} L_p^{(T)} = \min_{\pi\in\Delta^{\mathcal{P}_k}} \left\langle \pi, L^{(T)}\right\rangle$, since the minimizers of a bounded linear function lie on the set of extremal points of the feasible set. Therefore, the discounted regret compares the performance of the online learning algorithm to the *best stationary strategy in hindsight*. Indeed, $\left\langle \pi, L^{(T)} \right\rangle$ is the cumulative discounted loss of a stationary strategy π, and minimizing this expression over $\pi \in \Delta^{\mathcal{P}_k}$ yields the best stationary strategy in hindsight: one cannot know a priori which strategy will minimize the expression until all losses up to T are revealed. If the algorithm has sublinear regret, its average performance is, asymptotically, as good as the performance of any stationary strategy, regardless of the sequence of distributions $(\mu^{(\tau)})_{\tau\in\mathbb{N}}$.

A note on monotonicity of the discount factors. A similar definition of discounted regret is used, for example, by Cesa-Bianchi and Lugosi in section 3.2 of [10]. However, in their definition, the sequence of discount factors is *increasing*. This can be motivated by the following argument: present observations may provide better information than past, stale observations. While this argument is accurate in many applications, it does not serve our purpose of convergence of population strategies. In our discussion, the standing assumption is that discount factors are *decreasing*.

Finally, we observe that the cumulative discounted loss and regret are bounded, uniformly in x, since the congestion functions are continuous on a compact set.

PROPOSITION 3.4. *There exists $\rho \geq 0$ such that for all k, all $p \in \mathcal{P}_k$, and all $\mu \in \Delta$, $\ell^k_p(\mu) \in [0, \rho]$; and for all $x \in \mathcal{X}_k$, $\frac{1}{\sum_{\tau=0}^T \gamma_\tau} L^{(T)}(x) \in [0, \rho]$ and $\frac{1}{\sum_{\tau=0}^T \gamma_\tau} \left[R^{(T)}(x)\right]^+ \in [0, \rho]$.*

3.3. Populationwide regret. We have defined the discounted regret $R^{(T)}(x)$ for a single player x. In order to analyze the population dynamics, we define a populationwide cumulative discounted loss $L^{k^{(T)}}$ and discounted regret $R^{k^{(T)}}$ as follows:

$$L^{k^{(T)}} = \frac{1}{m(\mathcal{X}_k)} \int_{\mathcal{X}_k} L^{(T)}(x) dm(x), \tag{3.5}$$

$$R^{k^{(T)}} = \frac{1}{m(\mathcal{X}_k)} \int_{\mathcal{X}_k} R^{(T)}(x) dm(x) = L^{k^{(T)}} - \min_{p\in\mathcal{P}_k} \mathscr{L}^{k^{(T)}}_p. \tag{3.6}$$

Since $L^{(T)}(x)$ is random for all x, $L^{k^{(T)}}$ is also a random variable. However, it is, in fact, almost surely equal to its expectation. Indeed, recalling that $\mu^{k^{(\tau)}}_p$ is the proportion of players who chose bundle p at iteration τ (also a random variable), we can write

$$L^{k^{(T)}} = \sum_{\tau=0}^T \gamma_\tau \frac{1}{m(\mathcal{X}_k)} \sum_{p\in\mathcal{P}_k} \int_{\{x\in\mathcal{X}_k : A^{(\tau)}(x)=p\}} \ell^k_p(\mu^{(\tau)}) dm(x) = \sum_{\tau=0}^T \gamma_\tau \sum_{p\in\mathcal{P}_k} \mu^{k^{(\tau)}}_p \ell^k_p(\mu^{(\tau)}),$$

thus assuming players randomize independently, $\mu^{(\tau)}$ is almost surely deterministic by Proposition 2.4, and so is $L^{k^{(T)}}$. The same holds for $R^{k^{(T)}}$.

PROPOSITION 3.5. *If almost every player $x \in \mathcal{X}_k$ applies an online learning algorithm with sublinear regret in expectation, then the populationwide regret is also sublinear.*

Proof. By the previous observation, we have, almost surely,

$$R^{k^{(T)}} = \mathbb{E}\left[R^{k^{(T)}}\right] = \frac{1}{m(\mathcal{X}_k)} \int_{\mathcal{X}_k} \mathbb{E}\left[R^{(T)}(x)\right] dm(x),$$

where the second equality follows from Tonelli's theorem. Taking the positive part and using Jensen's inequality, we have

$$\frac{1}{\sum_{\tau=0}^{T}\gamma_\tau}\left[R^{k^{(T)}}\right]^+ \le \frac{1}{m(\mathcal{X}_k)}\int_{\mathcal{X}_k}\frac{1}{\sum_{\tau=0}^{T}\gamma_\tau}\left[\mathbb{E}\left[R^{(T)}(x)\right]\right]^+ dm(x).$$

By assumption, $\frac{1}{\sum_{\tau=0}^{T}\gamma_\tau}\left[\mathbb{E}\left[R^{(T)}(x)\right]\right]^+$ converges to 0 for all x, and by Proposition 3.4, it is bounded uniformly in x. Thus the result follows by applying the dominated convergence theorem. □

3.4. Hedge algorithm with vanishing learning rates. We now present one particular online learning algorithm with sublinear regret. Consider a congestion game, and let ρ be an upper bound on the losses. The existence of such an upper bound was established in Proposition 3.4.

DEFINITION 3.6 (Hedge algorithm). *The Hedge algorithm, applied by player $x \in \mathcal{X}_k$, with initial distribution $\pi^{(0)} \in \Delta^{\mathcal{P}_k}$ and learning rates $(\eta_\tau)_{\tau\in\mathbb{N}}$, is an online learning algorithm $({}^xU^{(\tau)})_{\tau\in\mathbb{N}}$ such that the τth update function is given by*

$$(3.7)\quad {}^xU^{(\tau)}((\ell^k(\mu^{(t)}))_{t\le\tau},\pi^{(\tau)}(x)) = \pi^{(\tau+1)}(x) \propto \left(\pi_p^{(\tau)}(x)\exp\left(-\eta_\tau\frac{\ell_p^k(\mu^{(\tau)})}{\rho}\right)\right)_{p\in\mathcal{P}_k}$$

Intuitively, the Hedge algorithm updates the distribution by computing, at each iteration, a set of bundle weights, then normalizing the vector of weights. The weight of a bundle p is obtained by multiplying the probability at the previous iteration, $\pi_p^{(\tau)}$, by a term which is exponentially decreasing in the bundle loss $\ell_p^k(\mu^{(\tau)})$; thus the higher the loss of bundle p at iteration τ, the lower the probability of selecting p at the next iteration. The parameter η_τ can be interpreted as a learning rate, as the Hedge update rule (3.7) is the solution to the following optimization problem:

$$(3.8)\qquad \pi^{(\tau+1)} \in \arg\min_{\pi\in\Delta^{\mathcal{P}_k}} \left\langle \pi, \frac{\ell^k(\mu^{(\tau)})}{\rho}\right\rangle + \frac{1}{\eta_\tau}D_{\mathrm{KL}}(\pi\|\pi^{(\tau)}),$$

where $D_{\mathrm{KL}}(\pi\|\nu) = \sum_{p\in\mathcal{P}_k}\pi_p\log\frac{\pi_p}{\nu_p}$ is the Kullback–Leibler divergence of distribution π with respect to ν (see, for example, [18]).

The objective function in (3.8) is the sum of an instantaneous loss term $\langle\pi, \frac{\ell^k(\mu^{(\tau)})}{\rho}\rangle$ and a regularization term $\frac{1}{\eta_\tau}D_{KL}(\pi\|\pi^{(\tau)})$ which penalizes deviations from the previous distribution $\pi^{(\tau)}$, with a regularization coefficient $\frac{1}{\eta_\tau}$. The greedy problem (with no regularization term) would yield a pure strategy which concentrates all the mass on the bundle which had minimal loss on the previous iteration. With the regularization term, the player "hedges her bet" by penalizing too much deviation from the previous distribution. The coefficient η_τ determines the relative importance of the two terms in the objective function. In particular, as $\eta_\tau \to 0$, the solution to the problem (3.8) converges to $\pi^{(\tau)}$ since the regularization term dominates the instantaneous loss term. In other words, as η_τ converges to 0, the player stops learning from new observations, which justifies calling η_τ a *learning rate*.

Remark 3.7. The sequence of distributions given by the Hedge algorithm also satisfies

$$\pi^{(\tau+1)} \propto \left(\pi_p^{(0)} \exp\left(-\sum_{t=0}^{\tau} \eta_t \frac{\ell_p^k(\mu^{(t)})}{\rho}\right)\right)_{p\in\mathcal{P}_k}. \tag{3.9}$$

This follows from the update equation (3.7) and a simple induction on τ. In particular, when $\eta_\tau = \gamma_\tau$, the term $\sum_{t=0}^{\tau} \eta_t \ell_p^k(\mu^{(t)})$ coincides with the cumulative discounted loss $\mathscr{L}_p^{k\,(\tau)}$ defined in (3.2). This motivates using the discount factors γ_τ as learning rates. We discuss this in the next proposition.

PROPOSITION 3.8. *Consider a congestion game with a sequence of discount factors* $(\gamma_\tau)_{\tau\in\mathbb{N}}$ *satisfying Assumption* 3.2. *Then the Hedge algorithm with learning rates* (γ_τ) *satisfies the following regret bound: for any sequence of distributions* $(\mu^{(\tau)})_\tau$ *and any initial strategy* $\pi^{(0)}$,

$$\mathbb{E}[R^{(T)}(x)] \le -\rho \log \pi_{\min}^{(0)} + \frac{\rho}{8}\sum_{\tau=0}^{T} \gamma_\tau^2,$$

where $\pi_{\min}^{(0)} = \min_{p\in\mathcal{P}_k} \pi_p^{(0)}$.

Proof. Given an initial strategy $\pi^{(0)}$, define $\xi \colon u \in \mathbb{R}^{\mathcal{P}_k} \mapsto \log(\sum_{p\in\mathcal{P}_k} \pi_p^{(0)} \exp(-\frac{u_p}{\rho}))$. Recalling the expression of the cumulative bundle loss $\mathscr{L}_p^{k\,(\tau)} = \sum_{t=0}^{\tau} \gamma_t \ell_p^k(\mu^{(t)})$, we have for all $\tau \ge 0$

$$\begin{aligned}
\xi(\mathscr{L}^{k\,(\tau+1)}) - \xi(\mathscr{L}^{k\,(\tau)}) &= \log\left(\sum_{p\in\mathcal{P}_k} \frac{\pi_p^{(0)} \exp\left(-\frac{\mathscr{L}_p^{k\,(\tau)}}{\rho}\right)}{\sum_{p'\in\mathcal{P}_k} \exp\left(-\frac{\mathscr{L}_{p'}^{k\,(\tau)}}{\rho}\right)} \exp\left(-\gamma_{\tau+1}\frac{\ell_p^k(\mu^{(\tau+1)})}{\rho}\right)\right)\\
&= \log\left(\sum_{p\in\mathcal{P}_k} \pi_p^{(\tau+1)} \exp\left(-\gamma_{\tau+1}\frac{\ell_p^k(\mu^{(\tau+1)})}{\rho}\right)\right)\\
&\le -\gamma_{\tau+1}\sum_{p\in\mathcal{P}_k} \pi_p^{(\tau+1)} \frac{\ell_p^k(\mu^{(\tau+1)})}{\rho} + \frac{\gamma_{\tau+1}^2}{8}.
\end{aligned}$$

The last inequality follows from Hoeffding's lemma, since $0 \le \frac{\ell_p^k(\mu^{(\tau)})}{\rho} \le 1$. Summing over $\tau \in \{-1, \ldots, T-1\}$, we have for all p

$$\xi(\mathscr{L}^{k\,(T)}) - \xi(\mathscr{L}^{k\,(-1)}) \le -\sum_{\tau=0}^{T} \gamma_\tau \sum_{p\in\mathcal{P}^k} \pi_p^{(\tau)} \frac{\ell_p^k(\mu^{(\tau)})}{\rho} + \frac{1}{8}\sum_{\tau=0}^{T} \gamma_\tau^2,$$

where $\xi(\mathscr{L}^{(-1)}) = \xi(0) = 0$. By monotonicity of the log function, we have for all $p_0 \in \mathcal{P}_k$, $\log(\pi_{p_0}^{(0)} \exp(-\frac{\mathscr{L}_{p_0}^{k\ (T)}}{\rho})) \le \xi(\mathscr{L}^{k\,(T)})$; thus

$$-\frac{\mathscr{L}_{p_0}^{k\ (\tau)}}{\rho} + \log \pi_{p_0}^{(0)} \le \xi(\mathscr{L}^{k\,(T)}) \le -\sum_{\tau=0}^{T} \gamma_\tau \sum_{p\in\mathcal{P}^k} \pi_p^{(\tau)} \frac{\ell_p^k(\mu^{(\tau)})}{\rho} + \frac{1}{8}\sum_{\tau=0}^{T} \gamma_\tau^2.$$

Rearranging, we have for all $p \in \mathcal{P}_k$

$$\sum_{\tau=0}^{T} \gamma_\tau \sum_{p \in \mathcal{P}_k} \pi_p^{(\tau)} \ell_p^k(\mu^{(\tau)}) - \mathscr{L}_{p_0}^{k\,(T)} \leq -\frac{\rho}{8} \log \pi_{p_0}^{(0)} + \rho \sum_{\tau=0}^{T} \gamma_\tau^2,$$

and we obtain the desired inequality by maximizing both sides over $p_0 \in \mathcal{P}_k$. □

The previous proposition provides an upper bound on the expected regret of the Hedge algorithm, of the form

$$\frac{\mathbb{E}\left[R^{(T)}(x)\right]}{\sum_{\tau \leq T} \gamma_\tau} \leq -\rho \pi_{\min}^{(0)} \frac{1}{\sum_{\tau \leq T} \gamma_\tau} + \frac{\rho}{8} \frac{\sum_{\tau \leq T} \gamma_\tau^2}{\sum_{\tau \leq T} \gamma_\tau}.$$

Given Assumption 3.2 on the discount factors, we have $\lim_{T \to \infty} \frac{\sum_{\tau \leq T} \gamma_\tau^2}{\sum_{\tau \leq T} \gamma_\tau} = 0$, which proves that the discounted regret is sub-linear. This also provides a bound on the convergence rate. For example, if $\gamma_\tau \sim \frac{1}{\tau}$, then the upper bound is equivalent to $\frac{c}{\log T}$, which converges to zero as $T \to \infty$, albeit slowly. A better bound can be obtained for sequences of discount factors which are not square-summable, for example, taking $\gamma_\tau \sim \frac{1}{\sqrt{\tau}}$, the upper bound is equivalent to $\frac{c \log T}{\sqrt{T}}$.

We now have one example of an online learning algorithm with sublinear discounted regret. Furthermore, we have an interpretation of the sequence (γ_τ) as learning rates, which provides additional intuition on Assumption 3.2 on (γ_τ): decreasing the learning rates will help the system converge.

In the next section, we start our analysis of the population dynamics when all players apply a learning algorithm with sublinear discounted regret.

4. Convergence in the Cesàro sense.

As discussed in Proposition 3.5, if almost every player applies an algorithm with sublinear discounted regret in expectation, then the populationwide discounted regret is sublinear (almost surely). We now show that whenever the population has sublinear discounted regret, the sequence of distributions $(\mu^{(\tau)})_\tau$ converges in the sense of Cesàro. That is, $\sum_{\tau \leq T} \gamma_\tau \mu^{(\tau)} / \sum_{\tau \leq T} \gamma_\tau$ converges to the set of Nash equilibria. We also show that we have convergence of a dense subsequence. First, we give some definitions.

Definition 4.1 (convergence in the sense of Cesàro). *Fix a sequence of positive weights $(\gamma_\tau)_{\tau \in \mathbb{N}}$. A sequence $(u^{(\tau)})_{\tau \in \mathbb{N}}$ of elements of a normed vector space $(F, \|\cdot\|)$ converges to $u \in F$ in the sense of Cesàro means with respect to $(\gamma_\tau)_\tau$ if $\lim_{T \to \infty} \frac{\sum_{\tau \in \mathbb{N}: \tau \leq T} \gamma_\tau u^{(\tau)}}{\sum_{\tau \in \mathbb{N}: \tau \leq T} \gamma_\tau} = u$. We write $u^{(\tau)} \xrightarrow{(\gamma_\tau)} u$.*

The Stolz–Cesàro theorem states that if $(u^{(\tau)})_\tau$ converges to u, then it converges in the sense of Cesàro means with respect to any nonsummable sequence $(\gamma_\tau)_\tau$; see, for example, [22]. The converse is not true in general. However, if a sequence converges *absolutely* in the sense of Cesàro means, i.e., $\|u^{(\tau)} - u\| \xrightarrow{(\gamma_\tau)} 0$, then a dense subsequence of $(u^{(\tau)})_\tau$ converges to u. To show this, we first show that absolute Cesàro convergence implies statistical convergence, as defined below.

Definition 4.2 (statistical convergence). *Fix a sequence of positive weights $(\gamma_\tau)_\tau$. A sequence $(u^{(\tau)})_{\tau \in \mathbb{N}}$ of elements of a normed vector space $(F, \|\cdot\|)$ converges to $u \in F$ statistically with respect to (γ_τ) if for all $\epsilon > 0$, the set of indexes $\mathcal{I}_\epsilon = \{\tau \in \mathbb{N} \colon \|u^{(\tau)} - u\| \geq \epsilon\}$ has zero density with respect to (γ_τ). The density of a subset of integers $\mathcal{I} \subset \mathbb{N}$, with respect to the sequence of positive weights (γ_τ), is defined to be the limit, if it exists, $\lim_{T \to \infty} \frac{\sum_{\tau \in \mathcal{I}: \tau \leq T} \gamma_\tau}{\sum_{\tau \in \mathbb{N}: \tau \leq T} \gamma_\tau}$.*

LEMMA 4.3. *If $(u^{(\tau)})_\tau$ converges to u absolutely in the sense of Cesàro means with respect to (γ_τ), then it converges to u statistically with respect to (γ_τ).*

Proof. Let $\epsilon > 0$. We have for all $T \in \mathbb{N}$,

$$0 \le \frac{\sum_{\tau\in\mathcal{I}_\epsilon:\ \tau\le T}\gamma_\tau\epsilon}{\sum_{\tau\in\mathbb{N}:\ \tau\le T}\gamma_\tau} \le \frac{\sum_{\tau\in\mathbb{N}:\tau\le T}\gamma_\tau\|u^{(\tau)}-u\|}{\sum_{\tau\in\mathbb{N}:\tau\le T}\gamma_\tau},$$

which converges to 0 since $(u^{(\tau)})_\tau$ converges to u absolutely in the sense of Cesàro means. Therefore $\mathcal{I}_\epsilon$ has zero density for all ϵ. ▯

We can now show convergence of a dense subsequence.

PROPOSITION 4.4. *If $(u^{(\tau)})_{\tau\in\mathbb{N}}$ converges to u absolutely in the sense of Cesàro means with respect to (γ_τ), then there exists a subset of indexes $\mathcal{T}\subset\mathbb{N}$ of density one, such that the subsequence $(u^{(\tau)})_{\tau\in\mathcal{T}}$ converges to u.*

Proof. By Lemma 4.3, for all $\epsilon > 0$, the set $\mathcal{I}_\epsilon = \{\tau \in \mathbb{N}\colon \|u^{(\tau)} - u\| \ge \epsilon\}$ has zero density. We will construct a set $\mathcal{I} \subset \mathbb{N}$ of zero density, such that the subsequence $(u_\tau)_{\tau\in\mathbb{N}\setminus\mathcal{I}}$ converges. For all $k \in \mathbb{N}^*$, let $p_k(T) = \sum_{\tau\in\mathcal{I}_{\frac{1}{k}}:\ \tau\le T}\gamma_\tau$. Since $\frac{p_k(T)}{\sum_{\tau\in\mathbb{N}:\ \tau\le T}\gamma_\tau}$ converges to 0 as $T\to\infty$, there exists $T_k > 0$ such that for all $T \ge T_k$, $\frac{p_k(T)}{\sum_{\tau\in\mathbb{N}:\ \tau\le T}\gamma_\tau} \le \frac{1}{k}$. Without loss of generality, we can assume that $(T_k)_{k\in\mathbb{N}^*}$ is increasing. Now, let $\mathcal{I} = \bigcup_{k\in\mathbb{N}^*}(\mathcal{I}_{\frac{1}{k}}\cap\{T_k,\dots,T_{k+1}-1\})$. Then we have for all $k\in\mathbb{N}^*$, $\mathcal{I}\cap\{0,\dots,T_{k+1}-1\} = \left(\cup_{j=1}^k\mathcal{I}_{\frac{1}{j}}\right)\cap\{0,\dots,T_{k+1}-1\}$. But since $\mathcal{I}_1\subset\mathcal{I}_{\frac{1}{2}}\subset\cdots\subset\mathcal{I}_{\frac{1}{k}}$, we have $\mathcal{I}\cap\{0,\dots,T_{k+1}-1\}\subset\mathcal{I}_{\frac{1}{k}}\cap\{0,\dots,T_{k+1}-1\}$; thus for all T such that $T_k \le T < T_{k+1}$, we have

$$\frac{\sum_{\tau\in\mathcal{I}:\ \tau\le T}\gamma_\tau}{\sum_{\tau\in\mathbb{N}:\ \tau\le T}\gamma_\tau} \le \frac{\sum_{\tau\in\mathcal{I}_{\frac{1}{k}}:\ \tau\le T}\gamma_\tau}{\sum_{\tau\in\mathbb{N}:\ \tau\le T}\gamma_\tau} = \frac{p_k(T)}{\sum_{\tau\in\mathbb{N}:\ \tau\le T}\gamma_\tau} \le \frac{1}{k},$$

which proves that $\mathcal{I}$ has zero density.

Let $\mathcal{T} = \mathbb{N}\setminus\mathcal{I}$. We have that $\mathcal{T}$ has density one, and it remains to prove that the subsequence $(u^{(\tau)})_{\tau\in\mathcal{T}}$ converges to u. Since $\mathcal{T}$ has density one, it has infinitely many elements, and for all k, there exists $S_k\in\mathcal{T}$ such that $S_k\ge T_k$. For all $\tau\in\mathcal{T}$ with $\tau\ge S_k$, there exists $k'\ge k$ such that $T_{k'}\le\tau<T_{k'+1}$. Since $\tau\notin\mathcal{I}$ and $T_{k'}\le\tau<T_{k'+1}$, we must have $\tau\notin\mathcal{I}_{\frac{1}{k'}}$; therefore $\|u^{(\tau)}-u\|<\frac{1}{k'}\le\frac{1}{k}$. This proves that $(u^{(\tau)})_{\tau\in\mathcal{T}}$ converges to u. ▯

We now present the main result of this section, which concerns the convergence of a subsequence of population distributions $(\mu^{(\tau)})$ to the set $\mathcal{N}$ of Nash equilibria. We say that $(\mu^{(\tau)})$ converges to $\mathcal{N}$ if $d(\mu^{(\tau)},\mathcal{N})\to 0$, where $d(\mu,\mathcal{N}) = \inf_{\nu\in\mathcal{N}}\|\mu-\nu\|$.

THEOREM 4.5. *Consider a congestion game with discount factors $(\gamma_\tau)_\tau$ satisfying Assumption 3.2. Assume that for all $k\in\{1,\dots,K\}$, population k has sublinear discounted regret. Then the sequence of distributions $(\mu^{(\tau)})_\tau$ converges to the set of Nash equilibria in the sense of Cesàro means with respect to (γ_τ). Furthermore, there exists a dense subsequence $(\mu_\tau)_{\tau\in\mathcal{T}}$ which converges to $\mathcal{N}$.*

Proof. First, we observe the following fact.

LEMMA 4.6. *A sequence $(\nu^{(\tau)})$ in Δ converges to $\mathcal{N}$ only if $(V(\nu^{(\tau)}))$ converges to $V_\mathcal{N}$, the value of V on $\mathcal{N}$.*

Indeed, suppose by contradiction that $V(\nu^{(\tau)})\to V_\mathcal{N}$ but $\nu^{(\tau)}\not\to\mathcal{N}$. Then there would exist $\epsilon>0$ and a subsequence $(\nu^{(\tau)})_{\tau\in\mathcal{T}}$, $\mathcal{T}\subset\mathbb{N}$, such that $d(\nu^{(\tau)},\mathcal{N})\ge\epsilon$ for all $\tau\in\mathcal{T}$. Since Δ is compact, we can extract a further subsequence $(\nu^{(\tau)})_{\tau\in\mathcal{T}'}$, which converges to some $\nu \notin \mathcal{N}$. But by continuity of V, $(V(\nu^{(\tau)}))_{\tau \in \mathcal{T}'}$ converges to $V(\nu) > V_{\mathcal{N}}$, a contradiction.

Consider the potential function V defined in (2.5). By convexity of V and the expression (2.6) of its gradient, we have for all τ and for all $\mu \in \Delta$,

$$V(\mu^{(\tau)}) - V(\mu) \leq \left\langle \nabla V(\mu^{(\tau)}), \mu^{(\tau)} - \mu \right\rangle = \sum_{k=1}^{K} m(\mathcal{X}_k) \left\langle \ell^k(\mu^{(\tau)}), \mu_p^{k^{(\tau)}} - \mu_p^k \right\rangle,$$

then taking the weighted sum up to iteration T,

$$\begin{aligned}\sum_{\tau=0}^{T} \gamma_\tau (V(\mu^{(\tau)}) - V(\mu)) &\leq \sum_{k=1}^{K} m(\mathcal{X}_k) \left[\sum_{\tau=0}^{T} \gamma_\tau \left\langle \mu^{k^{(\tau)}}, \ell^k(\mu^{(\tau)}) \right\rangle - \left\langle \mu^k, \sum_{\tau=0}^{T} \gamma_\tau \ell^k(\mu^{(\tau)}) \right\rangle \right] \\ &= \sum_{k=1}^{K} m(\mathcal{X}_k) \left[L^{k^{(T)}} - \left\langle \mu^k, \mathscr{L}^{k^{(T)}} \right\rangle \right] \leq \sum_{k=1}^{K} m(\mathcal{X}_k) R^{k^{(T)}},\end{aligned}$$

where for the last inequality, we use the fact that $\left\langle \mu^k, \mathscr{L}^{k^{(T)}} \right\rangle \geq \min_{p \in \mathcal{P}_k} \mathscr{L}_p^{k^{(T)}}$. In particular, when μ is a Nash equilibrium, by Theorem 2.5, $V(\mu) = \min_{\mu \in \Delta} V(\mu) = V_{\mathcal{N}}$, and thus

$$\frac{\sum_{\tau=0}^{T} \gamma_\tau |V(\mu^{(\tau)}) - V_{\mathcal{N}}|}{\sum_{\tau=0}^{T} \gamma_\tau} \leq \sum_{k=1}^{K} m(\mathcal{X}_k) \frac{R^{k^{(T)}}}{\sum_{\tau=0}^{T} \gamma_\tau}.$$

Since the populationwide regret $R^{k^{(T)}}$ is assumed to be sublinear for all k, we have $|V(\mu^{(\tau)}) - V_{\mathcal{N}}| \xrightarrow{(\gamma_\tau)} 0$. By Proposition 4.4, there exists $\mathcal{T} \subset \mathbb{N}$ of density one, such that $(V(\mu^{(\tau)}))_{\tau \in \mathcal{T}}$ converges to $V_{\mathcal{N}}$. And it follows that $(\mu^{(\tau)})_{\tau \in \mathcal{T}}$ converges to $\mathcal{N}$. This proves the second part of the theorem. To prove the first part, we observe that, by convexity of V,

$$V_{\mathcal{N}} \leq V\left(\frac{\sum_{\tau=0}^{T} \gamma_\tau \mu^{(\tau)}}{\sum_{\tau=0}^{T} \gamma_\tau} \right) \leq \frac{\sum_{\tau=0}^{T} \gamma_\tau V(\mu^{(\tau)})}{\sum_{\tau=0}^{T} \gamma_\tau} = V_{\mathcal{N}} + \frac{\sum_{\tau=0}^{T} \gamma_\tau (V(\mu^{(\tau)}) - V_{\mathcal{N}})}{\sum_{\tau=0}^{T} \gamma_\tau},$$

and the upper bound converges to $V_{\mathcal{N}}$. Therefore $\left(\frac{\sum_{\tau \leq T} \gamma_\tau \mu^{(\tau)}}{\sum_{\tau \leq T} \gamma_\tau} \right)_T$ converges to $\mathcal{N}$. □

To conclude this section, we observe that the Cesàro convergence result of Theorem 4.5 can be generalized to any game with a convex potential function.

5. Continuous-time dynamics.

We now turn to the harder question of convergence of $(\mu^{(\tau)})_\tau$: we seek to derive sufficient conditions under which the sequence $(\mu^{(\tau)})$ converges to $\mathcal{N}$. In this section, we study a continuous-time limit of the update equation given by the Hedge algorithm. The resulting ODE, known as the replicator equation, will be useful in proving strong convergence results in the next section.

5.1. The replicator dynamics.

To motivate the study of the replicator dynamics from an online learning point of view, we first derive the continuous-time replicator dynamics as a limit of the discrete Hedge dynamics, as discussed below. Assume that in each population $\mathcal{X}_k$, all players start from the same initial distribution $\pi^{k^{(0)}} \in \Delta^{\mathcal{P}_k}$, and apply the Hedge algorithm with learning rates (γ_τ). As a result, the sequence of distributions $(\mu^{k^{(\tau)}})$ satisfies the Hedge update rule (3.7). Now suppose the existence of an underlying continuous time $t \in \mathbb{R}_+$ and write $\boldsymbol{\mu}(t)$ the distribution at time t. Suppose that the updates occur at discrete times T_τ, $\tau \in \mathbb{N}$, such that the time steps are given by a decreasing, vanishing sequence ϵ_τ. That is, $T_{\tau+1} - T_\tau = \epsilon_\tau$. Then we have for all k and all $p \in \mathcal{P}_k$, using Landau notation,

$$\begin{aligned}\boldsymbol{\mu}_p^k(T_{\tau+1}) = \mu_p^{k^{(\tau+1)}} &= \mu_p^{k^{(\tau)}} \frac{e^{-\gamma_\tau \frac{\ell_p^k(\mu^{(\tau)})}{\rho}}}{\sum_{p' \in \mathcal{P}_k} \mu_{p'}^{k^{(\tau)}} e^{-\gamma_\tau \frac{\ell_{p'}^k(\mu^{(\tau)})}{\rho}}} \\ &= \mu_p^{k^{(\tau)}} \frac{1 - \gamma_\tau \frac{\ell_p^k(\mu^{(\tau)})}{\rho} + o(\gamma_\tau)}{1 - \gamma_\tau \sum_{p' \in \mathcal{P}_k} \mu_{p'}^{k^{(\tau)}} \frac{\ell_{p'}^k(\mu^{(\tau)})}{\rho} + o(\gamma_\tau)} \\ &= \boldsymbol{\mu}_p^k(T_\tau) \left[1 + \gamma_\tau \frac{\bar{\ell}^k(\mu^{(\tau)}) - \ell_p^k(\mu^{(\tau)})}{\rho}\right] + o(\gamma_\tau).\end{aligned}$$

Thus,

$$\frac{\boldsymbol{\mu}_p^k(T_{\tau+1}) - \boldsymbol{\mu}_p^k(T_\tau)}{T_{\tau+1} - T_\tau} \frac{\epsilon_\tau}{\gamma_\tau} = \boldsymbol{\mu}_p^k(T_\tau) \frac{\bar{\ell}^k(\mu(\tau)) - \ell_p^k(\mu(\tau))}{\rho} + o(1).$$

In particular, if we take the discretization time steps ϵ_τ to be equal to the sequence of learning rate γ_τ, the expression simplifies, and taking the limit as $\gamma_\tau \to 0$, we obtain the following ODE system:

$$\begin{cases} \boldsymbol{\mu}(0) \in \mathring{\Delta} \\ \forall k,\ \forall p \in \mathcal{P}_k, \frac{d\boldsymbol{\mu}_p^k(t)}{dt} = \boldsymbol{\mu}_p^k(t) \frac{\bar{\ell}^k(\boldsymbol{\mu}(t)) - \ell_p^k(\boldsymbol{\mu}(t))}{\rho}, \end{cases} \tag{5.1}$$

where $\mathring{\Delta} = \{\mu \in \Delta \colon \forall k,\ \forall p \in \mathcal{P}_k, \mu_p^k > 0\}$ is the relative interior of Δ. Here, we require that the initial distribution have positive weights on all bundles for the following reason: whenever $\boldsymbol{\mu}_p^k(0) = 0$, any solution trajectory will have $\boldsymbol{\mu}_p^k(t) \equiv 0$. It is impossible for such trajectories to converge to the set of Nash equilibria $\mathcal{N}$ if the support of equilibria in $\mathcal{N}$ contains p. In other words, the replicator dynamics cannot expand the support of the initial distribution; therefore we require that the initial distribution be supported everywhere.

Equation (5.1) defines a vector field $F : \Delta \to \mathcal{H}$, where $\mathcal{H}$ is the product $\mathcal{H} = \mathcal{H}^{\mathcal{P}_1} \times \cdots \times \mathcal{H}^{\mathcal{P}_K}$ and $\mathcal{H}^{\mathcal{P}_k} = \{v \in \mathbb{R}^{\mathcal{P}_k} : \sum_{p \in \mathcal{P}} v_p = 0\}$ is the linear hyperplane parallel to the simplex $\Delta^{\mathcal{P}_k}$. Indeed, we have for all $\mu \in \Delta$ and for all k, $\sum_{p \in \mathcal{P}_k} F_p^k(\mu) = \sum_{p \in \mathcal{P}_k} \ell_p^k(\mu)\mu_p^k - \bar{\ell}^k(\mu) \sum_{p \in \mathcal{P}_k} \mu_p^k = 0$.

The following proposition ensures that the solutions remain in the relative interior and are defined on all times.

PROPOSITION 5.1. *The ODE* (5.1) *has a unique solution* $\mu(t)$ *which remains in* $\mathring{\Delta}$ *and is defined on* $\mathbb{R}_+$.

Proof. First, since the congestion functions c_r are assumed to be Lipschitz continuous, so is the vector field F. We thus have existence and uniqueness of a solution by the Cauchy–Lipschitz theorem.

To show that the solution remains in the relative interior of Δ, we observe that for all k, $\frac{d}{dt} \sum_{p \in \mathcal{P}_k} \boldsymbol{\mu}_p^k(t) = \sum_{p \in \mathcal{P}_k} F_p^k(\boldsymbol{\mu}(t)) = 0$ by the previous observation. Therefore,

$\sum_{p\in\mathcal{P}_k} \boldsymbol{\mu}_p^k(t)$ is constant and equal to 1. To show that $\boldsymbol{\mu}_p^k(t) > 0$ for all t in the solution domain, assume by contradiction that there exists $t_0 > 0$ and $p_0 \in \mathcal{P}_k$ such that $\boldsymbol{\mu}_{p_0}^k(t_0) = 0$. Since the solution trajectories are continuous, we can assume, without loss of generality, that t_0 is the infimum of all such times (thus for all $t < t_0$, $\boldsymbol{\mu}_{p_0}(t) > 0$). Now consider the new system given by

$$\begin{aligned}\dot{\tilde{\boldsymbol{\mu}}}_p &= \frac{1}{\rho}(\bar{\ell}(\tilde{\boldsymbol{\mu}}) - \ell_p(\tilde{\boldsymbol{\mu}}))\tilde{\boldsymbol{\mu}}_p \quad && \forall p \neq p_0,\\ \tilde{\boldsymbol{\mu}}_p(t_0) &= \boldsymbol{\mu}_p(t_0) && \forall p \neq p_0,\end{aligned}$$

and $\tilde{\boldsymbol{\mu}}_{p_0}(t)$ is identically equal to 0. Any solution of the new system, defined on $(t_0 - \delta, t_0]$, is also a solution of (5.1). Since $\boldsymbol{\mu}(t_0) = \tilde{\boldsymbol{\mu}}(t_0)$, we have $\boldsymbol{\mu} \equiv \tilde{\boldsymbol{\mu}}$ by uniqueness of the solution. This leads to a contradiction since by assumption, for all $t < t_0$, $\boldsymbol{\mu}_p(t) > 0$ but $\tilde{\boldsymbol{\mu}}_p(t) = 0$.

This proves that $\boldsymbol{\mu}$ remains in $\mathring{\Delta}$. Furthermore, since Δ is compact, we have by Theorem 2.4 in [17] that the solution is defined on $\mathbb{R}_+$ (otherwise it would eventually leave any compact set). ▯

5.2. Stationary points of the replicator dynamics.

We first give a characterization of stationary points of the replicator dynamics applied to the congestion game.

Proposition 5.2. *A product distribution μ is a stationary point for the replicator dynamics* (5.1) *if and only if the bundle losses $\ell_p^k(\mu)$ are equal on the support of μ^k.*

This follows immediately from (5.1). We observe in particular that all Nash equilibria are stationary points, but a stationary point may not be a Nash equilibrium in general: one may have a stationary point μ such that $\mu_p^k = 0$ but $\ell_p^k(\mu)$ is strictly lower than losses of bundles in the support, which violates the condition in Definition 2.2 of a Nash equilibrium.

A stationary point μ with support $\mathcal{P}'_1 \times \cdots \times \mathcal{P}'_K$ can be viewed as a Nash equilibrium of a modified congestion game, in which the bundle set of each population $\mathcal{X}_k$ is restricted to $\mathcal{P}'_k$. For this reason, stationary points have been called *restricted Nash equilibria* by Fischer and Vöcking in [11]. We will denote the set of stationary points by $\mathcal{RN}$, in reference to the aforementioned paper.

Remark 5.3. By the previous observation, a stationary point with support $\mathcal{P}'_1 \times \cdots \times \mathcal{P}'_K$ is a minimizer of the potential function V on the product $\Delta^{\mathcal{P}'_1} \times \cdots \times \Delta^{\mathcal{P}'_K}$. As the number of support sets is finite, the set of potential values of stationary points $V(\mathcal{RN})$ is also finite.

5.3. Convergence of the replicator dynamics.

In [11], Fischer and Vöcking prove, using a Lyapunov argument, that all solution trajectories of the replicator system asymptotically approach the set of stationary points $\mathcal{RN}$. Unfortunately, this result only guarantees convergence to a superset of Nash equilibria. However, this will be useful in the next section.

Proposition 5.4 (Fischer and Vöcking [11]). *Every solution of the system* (5.1) *converges to the set of stationary points $\mathcal{RN}$.*

5.4. A discrete-time replicator equation: The REP update rule.

Inspired by the continuous-time replicator dynamics, we propose a discrete-time multiplicative update rule by discretizing the ODE (5.1). The resulting algorithm has many desirable properties, such as sublinear discounted regret and simplicity of implementation. We call it the REP algorithm in reference to the replicator ODE.

The vector field F can be written in the following form: for all k, $F^k(\boldsymbol{\mu}) = G^k(\boldsymbol{\mu}, \ell(\boldsymbol{\mu}))$, where for all p,

$$G_p^k(\boldsymbol{\mu}, \ell) = \boldsymbol{\mu}_p^k \frac{\langle \boldsymbol{\mu}^k, \ell^k \rangle - \ell_p^k}{\rho}.$$

This motivates the following update rule for a player $x \in \mathcal{X}_k$ with distribution $\pi^{(\tau)}(x)$:

$$\pi^{(\tau+1)}(x) = \pi^{(\tau)}(x) + \eta_\tau G^k(\pi^{(\tau)}(x), \ell(\mu^{(\tau)})).$$

Definition 5.5 (discrete replicator algorithm). *The REP algorithm, applied by player $x \in \mathcal{X}_k$, with initial distribution $\pi^{(0)} \in \Delta^{\mathcal{P}_k}$ and learning rates $(\eta_\tau)_{\tau\in\mathbb{N}}$ with $\eta_\tau \leq 1$, is an online learning algorithm $({}^xU^{(\tau)})_{\tau\in\mathbb{N}}$ such that the τth update function is given by ${}^xU^{(\tau)}((\ell^k(\mu^{(t)}))_{t\leq\tau}, \pi^{(\tau)}) = \pi^{(\tau+1)}$, such that*

$$\pi_p^{(\tau+1)} - \pi_p^{(\tau)} = \eta_\tau \pi_p^{(\tau)} \frac{\langle \pi^{(\tau)}, \ell^k(\mu^{(\tau)}) \rangle - \ell_p^k(\mu^{(\tau)})}{\rho}. \tag{5.2}$$

Here, $\langle \pi^{(\tau)}, \ell^k(\mu^{(\tau)}) \rangle - \ell_p^k(\mu^{(\tau)})$ is the expected instantaneous regret of the player, with respect to bundle p. Thus the REP update can also be expressed in terms of the previous distribution and the expected instantaneous regret.

Under the REP update, the sequence of strategy profiles $\pi^{(\tau)}$ remains in the product of simplexes Δ, provided $\eta_\tau \leq 1$ for all τ. Indeed, for all $\tau \in \mathbb{N}$, we have $\sum_{p\in\mathcal{P}_k} \pi_p^{(\tau+1)} = \sum_{p\in\mathcal{P}_k} \pi_p^{(\tau)} + \frac{\eta_\tau}{\rho}[\bar{\ell}^k(\mu^{(\tau)}) - \sum_{p\in\mathcal{P}_k} \mu_p^{(\tau)} \ell_p^k(\mu^{(\tau)})] = \sum_{p\in\mathcal{P}_k} \pi_p^{(\tau)}$ and $1 + \eta_\tau \frac{\bar{\ell}^k(\mu^{(\tau)}) - \ell_p^k(\mu^{(\tau)})}{\rho} \geq 1 - \eta_\tau \geq 0$ if $\eta_\tau \leq 1$, which guarantees that $\pi^{(\tau)}$ remains in Δ.

We now show that the REP update rule with learning rates (γ_τ) has sublinear discounted regret. First, we prove the following lemma for general online learning problems with signed losses.

Lemma 5.6. *Consider a discounted online learning problem, with sequence of discount factors (γ_τ), with $\gamma_\tau \leq \frac{1}{2}$ for all τ. Let $\mathcal{P}_k$ be the finite decision set, and assume that the losses are signed and bounded, $m_p^{(\tau)} \in [-1, 1]$ for all τ and $p \in \mathcal{P}$. Then the multiplicative-weights algorithm defined by the update rule*

$$\pi^{(\tau+1)} \propto \left(\pi_p^{(\tau)}(1 - \gamma_\tau m_p^{(\tau)})\right)_{p\in\mathcal{P}_k} \tag{5.3}$$

has the following regret bound: for all T and all $p \in \mathcal{P}_k$,

$$\sum_{0\leq\tau\leq T} \gamma_\tau \left\langle m^{(\tau)}, \pi^{(\tau)} \right\rangle \leq -\log \pi_{\min}^{(0)} + \sum_{0\leq\tau\leq T} \gamma_\tau m_p^{(\tau)} + \sum_{0\leq\tau\leq T} \gamma_\tau^2 |m_p^{(\tau)}|,$$

where $\pi_{\min}^{(0)} = \min_{p\in\mathcal{P}_k} \pi_p^{(0)}$.

Proof. We extend the proof of Theorem 2.1 in [1] to the discounted case. By a simple induction, we have for all T, $\pi^{(T)}$ is proportional to the vector $w^{(T)}$ defined by

$$w_p^{(T)} = \pi_p^{(0)} \prod_{0\leq\tau<T} (1 - \gamma_\tau m_p^{(\tau)}).$$

Define the function $\xi^{(T)} = \sum_p w_p^{(T)}$. Then $\pi_p^{(T)} = \frac{w_p^{(T)}}{\xi^{(T)}}$, and we have for all T

$$\begin{aligned}\xi^{(T+1)} &= \sum_p w_p^{(T+1)} = \sum_p w_p^{(T)}(1 - \gamma_T m_p^{(T)}) = \xi^{(T)} - \gamma_T \sum_p m_p^{(T)} \pi_p^{(T)} \xi^{(T)} \\ &= \xi^{(T)} \left(1 - \gamma_T \left\langle m^{(T)}, \pi^{(T)} \right\rangle\right) \leq \xi^{(T)} e^{-\gamma_T \left\langle m^{(T)}, \pi^{(T)} \right\rangle}.\end{aligned}$$

Thus, by induction on T, $\xi^{(T+1)} \le \exp(-\sum_{0\le\tau\le T}\gamma_\tau \left\langle m^{(\tau)}, \pi^{(\tau)}\right\rangle)$. We also have for all p, $\xi^{(T+1)} \ge w_p^{(T+1)} \ge \pi_{\min}^{(0)} \prod_{0\le\tau\le T}(1-\gamma_t m_p^{(\tau)})$. Combining the bounds on $\xi^{(\tau)}$ and taking logarithms, we have

$$\sum_{0\le\tau\le T} \gamma_\tau \left\langle m^{(\tau)}, \pi^{(\tau)}\right\rangle \le -\log \pi_{\min}^{(0)} - \sum_{0\le\tau\le T} \log(1-\gamma_\tau m_p^{(\tau)}).$$

To obtain the desired bound, it suffices to observe that for all $m \in [-1,1]$ and $\gamma \in [0,\frac{1}{2}]$, $-\log(1-\gamma m) \le \gamma m + \gamma^2 |m|$. □

PROPOSITION 5.7. *If the sequence of discounts (γ_τ) satisfies Assumption* 3.2 *and is bounded by $\frac{1}{2}$, then the REP algorithm with learning rates γ_τ has sublinear discounted regret.*

Proof. Let

$$r_p^{(\tau)} = \left\langle \pi^{(\tau)}, \ell^k(\mu^{(\tau)})\right\rangle - \ell_p^k(\mu^{(\tau)}) \in [-\rho, \rho]$$

be the *instantaneous regret* of the player. Then the REP update can be viewed as a multiplicative-weights algorithm with update rule (5.3), in which the vector of signed losses is given by $m_p^{(\tau)} = -\frac{r_p{}^{(\tau)}}{\rho} \in [-1,1]$, and discount factors (γ_τ). Observing that $\left\langle r^{(\tau)}, \pi^{(\tau)}\right\rangle = 0$, we have by Lemma 5.6, for all $p \in \mathcal{P}_k$,

$$\frac{1}{\rho}\sum_{0\le\tau\le T} \gamma_\tau r_p^{(\tau)} \le -\log \pi_{\min}^{(0)} + \sum_{0\le\tau\le T} \gamma_\tau^2.$$

Rearranging and taking the maximum over $p \in \mathcal{P}_k$, we obtain the bound on the discounted regret,

$$R^{(T)}(x) \le -\rho \log \pi_{\min}^{(0)} + \rho \sum_{0\le\tau\le T} \gamma_\tau^2,$$

which shows $\limsup_{T\to\infty} \frac{1}{\sum_{\tau\le T}\gamma_\tau} R^{(T)}(x) \le 0$. □

Interestingly, the REP update can also be obtained as the solution to a regularized version of the greedy update $\min_{\pi\in\Delta^{\mathcal{P}_k}} \langle \pi, \frac{\ell^k(\mu^{(\tau)})}{\rho}\rangle$, similarly to the Hedge update, but with a different regularization function.

PROPOSITION 5.8. *The REP update rule is solution to the following problem:*

$$\{\pi^{(\tau+1)}\} = \arg\min_{\pi\in\Delta} \left\langle \pi, \frac{\ell^k(\mu^{(\tau)})}{\rho}\right\rangle + \frac{1}{\eta_\tau} R(\pi \| \pi^{(\tau)}),$$

where $R(\pi\|\nu) = \frac{1}{2}\sum_{p\in\mathcal{P}_k} \pi_p \left(\frac{\pi_p}{\nu_p} - 1\right)^2$.

Proof. Define the partial Lagrangian function

$$\mathcal{L}(\pi;\lambda) = \sum_{p\in\mathcal{P}} \pi_p \frac{\ell^k(\mu^{(\tau)})}{\rho} + \frac{1}{2\gamma_\tau}\sum_{p\in\mathcal{P}_k} \pi_p^{(\tau)} \left(\frac{\pi_p}{\pi_p^{(\tau)}} - 1\right)^2 - \lambda\left(\sum_{p\in\mathcal{P}_k} \pi_p - 1\right),$$

where $\lambda \in \mathbb{R}$ is the dual variable for the constraint $\sum_{p\in\mathcal{P}_k} \pi_p = 1$. Its gradient is

$$\frac{\partial}{\partial \pi_p}\mathcal{L}(\pi;\lambda) = \frac{\ell_p^k(\mu^{(\tau)})}{\rho} + \frac{1}{\gamma_\tau}\left(\frac{\pi_p}{\pi_p^{(\tau)}} - 1\right) - \lambda \;\forall p \in \mathcal{P}_k,$$

$$\frac{\partial}{\partial \lambda}\mathcal{L}(\pi;\lambda) = -\sum_{p\in\mathcal{P}_k} \pi_p + 1,$$

and $(\pi^\star, \lambda^\star)$ are primal-dual optimal if and only if

$$\frac{\pi_p^\star}{\pi_p^{(\tau)}} = 1 + \gamma_\tau \left(\lambda - \frac{\ell_p^k(\mu^{(\tau)})}{\rho}\right) \text{ and } \sum_{p \in \mathcal{P}_k} \pi_p^\star = 1.$$

Multiplying by $\pi_p^{(\tau)}$ and taking the sum over $p \in \mathcal{P}_k$, we have $1 = 1 + \gamma_\tau \lambda^\star - \gamma_\tau \langle \pi^{(\tau)}, \frac{\ell^k(\mu^{(\tau)})}{\rho} \rangle$, i.e. $\lambda^\star = \langle \pi^{(\tau)}, \frac{\ell^k(\mu^{(\tau)})}{\rho} \rangle$; thus the solution $\pi^\star$ satisfies the REP update rule (5.2). ▯

6. Strong convergence of discounted no-regret learning.

In this section, we give sufficient conditions which guarantee convergence of the sequence of population strategies. The idea is to show that, under these conditions, the discrete process $(\mu^{(\tau)})_{\tau \in \mathbb{N}}$ approaches, in a certain sense, the trajectories of the continuous-time replicator dynamics. Then one can show, using a Lyapunov function, that any limit point of the discrete process must lie in the set of stationary points $\mathcal{RN}$. With an additional argument, we show that, in fact, limit points lie in the set $\mathcal{N}$ of Nash equilibria.

We start by reviewing results from the theory of stochastic approximation, which we use in the proof of Theorem 6.10.

6.1. Results from the theory of stochastic approximation.

We summarize results from [4] due to Benaïm. Let $\mathcal{D} \subset \mathbb{R}^n$, and consider a dynamical system given by the ODE

$$\dot{\mu} = F(\mu), \tag{6.1}$$

where $F : \mathcal{D} \to \mathbb{R}^n$ is a continuous globally integrable vector field, with unique integral curves which remain in $\mathcal{D}$. Let Φ be the associated flow function such that $t \mapsto \Phi_t(\mu^{(0)})$ is the solution trajectory of (6.1) with initial condition $\mu(0) = \mu^{(0)}$.

6.1.1. Discrete-time approximation.

We now define what it means for a discrete process to approach the trajectories of the system (6.1).

Let $(\mu^{(\tau)})_\tau$ be a discrete-time process with values in $\mathcal{D}$. $(\mu^{(\tau)})_\tau$ is said to be a discrete-time approximation of the dynamical system (6.1) if there exists a sequence $(\gamma_\tau)_{\tau \in \mathbb{N}}$ of nonnegative real numbers such that $\sum_{\tau \in \mathbb{N}} \gamma_\tau = \infty$ and $\lim_{\tau \to \infty} \gamma_\tau = 0$, and a sequence of deterministic or random perturbations $U^{(\tau)} \in \mathbb{R}^n$ such that for all τ,

$$\mu^{(\tau+1)} - \mu^{(\tau)} = \gamma_\tau \left(F(\mu^{(\tau)}) + U^{(\tau+1)}\right). \tag{6.2}$$

Given such a discrete-time approximation, we can define the affine interpolated process of $(\mu^{(\tau)})$: let $T_\tau = \sum_{t=0}^{\tau} \gamma_t$ as in section 5.1.

DEFINITION 6.1 (affine interpolated process). *The continuous time affine interpolated process of the discrete process $(\mu^{(\tau)})_{\tau \in \mathbb{N}}$ is the function $M : \mathbb{R}_+ \to \mathbb{R}^n$ defined as*

$$M(T_\tau + s) = \mu^{(\tau)} + s \frac{\mu^{(\tau+1)} - \mu^{(\tau)}}{\gamma_\tau} \quad \forall \tau \in \mathbb{N} \textit{ and } \forall s \in [0, \gamma_\tau).$$

The next proposition gives sufficient conditions for an affine interpolated process to be an asymptotic pseudotrajectory (APT).

PROPOSITION 6.2 (Proposition 4.1 in [4]). *Let M be the affine interpolated process of the discrete-time approximation $(\mu^{(\tau)})$, and assume the following:*

1. *For all $T > 0$,*

$$\lim_{\tau_1 \to \infty} \max_{\tau_2 \colon \sum_{\tau=\tau_1}^{\tau_2} \gamma_\tau < T} \left\| \sum_{\tau=\tau_1}^{\tau_2} \gamma_\tau U^{(\tau+1)} \right\| = 0. \tag{6.3}$$

2. $\sup_{\tau \in \mathbb{N}} \|\mu^{(\tau)}\| < \infty$.

Then M is an APT of the flow Φ induced by the vector field F.

Furthermore, we have the following sufficient condition for property (6.3) to hold.

Proposition 6.3. *Let $(\mu^{(\tau)})_{\tau \in \mathbb{N}}$ be a discrete time approximation of the system (6.1). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_\tau)_{\tau \in \mathbb{N}}$ a filtration of $\mathcal{F}$. Suppose that the perturbations satisfy the Robbins–Monro conditions: for all $\tau \in \mathbb{N}$,*

(i) *$U^{(\tau)}$ is measurable with respect to $\mathcal{F}_\tau$,*

(ii) *$\mathbb{E}[U^{(\tau+1)} | \mathcal{F}_\tau] = 0$.*

Furthermore, suppose that there exists $q \geq 2$ such that

$$\sup_{\tau \in \mathbb{N}} \mathbb{E}[\|U^{(\tau)}\|^q] < \infty \qquad \text{and} \qquad \sum_{\tau \in \mathbb{N}} \gamma_\tau^{1+q/2} < \infty.$$

Then, condition 1 of Proposition 6.2 holds with probability one.

6.1.2. Chain transitivity.

We next give an important property of limit points of bounded APTs, given in Theorem 6.6.

Definition 6.4 (pseudoorbit and chain transitivity). *A (δ, T)-pseudoorbit from $a \in \mathcal{D}$ to $b \in \mathcal{D}$ is a finite sequence of partial trajectories. It is given by a sequence of points $(t_i, y_i), i \in \{0, \dots, k-1\}$ (with $t_i \geq T$ for all i) and the corresponding sequence of partial trajectories*

$$\{\Phi_t(y_i) \colon 0 \leq t \leq t_i\};\ i = 0, \dots, k-1,$$

such that $d(y_0, a) < \delta$, $d(\Phi_{t_i}(y_i), y_{i+1}) < \delta$ for all i, and $y_k = b$.

The conditions are illustrated in Figure 2. We write $\Phi \colon a \hookrightarrow_{\delta,T} b$ if there exists a (δ, T)-pesudoorbit from a to b. We write $a \hookrightarrow b$ if $a \hookrightarrow_{\delta,T} b$ for all $\delta, T > 0$. The flow Φ is said to be chain transitive if $a \hookrightarrow b$ for all $a, b \in \mathcal{D}$.

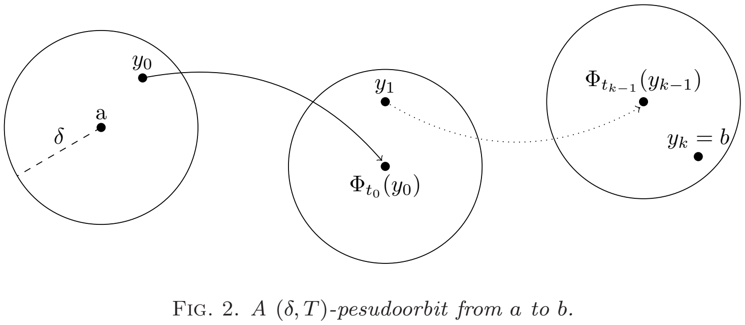

Fig. 2. *A (δ, T)-pesudoorbit from a to b.*

In the remainder of this section, let $\Gamma \subset \mathcal{D}$ be a compact invariant set for Φ, that is, $\Phi_t(\Gamma) \subseteq \Gamma$ for all $t \in \mathbb{R}^+$.

Definition 6.5 (internally chain transitive set). *The compact invariant set Γ is internally chain transitive if the restriction of Φ to Γ is chain transitive.*

Theorem 6.6 (Theorem 5.7 in [4]). *Let X be a bounded APT of (6.1). Then the limit set $L(X) = \bigcap_{t \geq 0} \overline{\{X(s) \colon s \geq t\}}$ is internally chain transitive.*

Finally, we give the following property of internally chain transitive sets.

PROPOSITION 6.7 (Proposition 6.4 in [4]). *Let $\Gamma \subset \mathcal{D}$ be a compact invariant set and suppose that there exists a Lyapunov function $V : \mathcal{D} \to \mathbb{R}$ for Γ (that is, V is continuous and $\frac{d}{dt}V(x(t)) = \langle \nabla V(x(t)), F(x(t))\rangle < 0$ for all $x \notin \Gamma$) such that $V(\Gamma)$ has empty interior. Then every internally chain transitive set L is contained in Γ and V is constant on L.*

6.2. The AREP class. Now, we are ready to define a class of online learning algorithms which we call AREP. An AREP online algorithm can be viewed as a perturbed version of the replicator algorithm.

DEFINITION 6.8 (AREP algorithm). *An online learning algorithm, applied by player $x \in \mathcal{X}_k$, with output sequence $(\pi^{(\tau)})_{\tau\in\mathbb{N}}$, is said to be an AREP algorithm if its update equation can be written as*

$$\pi_p^{(\tau+1)} - \pi_p^{(\tau)} = \gamma_\tau \left(\pi_p^{(\tau)} \frac{\langle \pi^{(\tau)}, \ell^k(\mu^{(\tau)})\rangle - \ell_p^k(\mu^{(\tau)})}{\rho} + U_p^{(\tau)} \right), \tag{6.4}$$

where $(U^{(\tau)})_{\tau\in\mathbb{N}}$ is a bounded sequence of stochastic perturbations with values in $\mathbb{R}^{\mathcal{P}_k}$, and which satisfies condition (6.3).

In particular, the REP algorithm given in Definition 5.5 is an AREP algorithm with zero perturbations. It turns out that the Hedge algorithm also belongs to the AREP class.

PROPOSITION 6.9. *The Hedge algorithm with learning rates $(\gamma_\tau)_\tau$ satisfying Assumption* 3.2 *is an AREP algorithm.*

Proof. Let $(\pi^{(\tau)})_{\tau\in\mathbb{N}}$ be the sequence of strategies, and let $(\mu^{(\tau)})_\tau$ be any sequence of population distributions. By definition of the Hedge algorithm, we have

$$\pi_p^{(\tau+1)} = \pi_p^{(\tau)} \exp\left(-\gamma_\tau \frac{\ell_p^k(\mu^{(\tau)})}{\rho}\right) \Bigg/ \sum_{p'\in\mathcal{P}_k} \pi_{p'}^{(\tau)} \exp\left(-\gamma_\tau \frac{\ell_{p'}^k(\mu^{(\tau)})}{\rho}\right),$$

which we can write in the form of (6.4) with perturbation terms

$$U_p^{(\tau+1)} = \frac{\pi_p^{(\tau)}}{\gamma_\tau}\left[\exp\left(-\gamma_\tau \frac{\ell_p^k(\mu^{(\tau)}) - \tilde{\ell}^{k(\tau)}}{\rho}\right) + \gamma_\tau \frac{\ell_p^k(\mu^{(\tau)}) - \tilde{\ell}^{k(\tau)}}{\rho} - 1\right] + \pi_p^{(\tau)} \frac{\tilde{\ell}^{k(\tau)} - \bar{\ell}^{k(\tau)}}{\rho},$$

where

$$\begin{aligned} \tilde{\ell}^{k(\tau)} &= -\frac{\rho}{\gamma_\tau} \log \sum_{p'\in\mathcal{P}_k} \pi_{p'}^{(\tau)} \exp\left(-\gamma_\tau \frac{\ell_{p'}^k(\mu^{(\tau)})}{\rho}\right), \\ \bar{\ell}^{k(\tau)} &= \left\langle \pi(\tau), \ell^k(\mu(\tau))\right\rangle. \end{aligned}$$

Letting $\theta(x) = e^x - x - 1$, we have for all $p \in \mathcal{P}_k$

$$U_p^{(\tau+1)} = \frac{\pi_p^{(\tau)}}{\gamma_\tau} \theta\left(-\gamma_\tau \frac{\ell_p^k(\mu^{(\tau)}) - \tilde{\ell}^{k(\tau)}}{\rho}\right) + \frac{\pi_p^{(\tau)}}{\rho}(\tilde{\ell}^{k(\tau)} - \bar{\ell}^{k(\tau)}).$$

The first term is a $O(\gamma_\tau)$ as $\theta(x) \sim_0 x^2/2$. To bound the second term, we have by concavity of the logarithm

$$\tilde{\ell}^{k(\tau)} = -\frac{\rho}{\gamma_\tau} \log \sum_{p'\in\mathcal{P}_k} \pi_{p'}^{(\tau)} \exp\left(-\gamma_\tau \frac{\ell_{p'}(\mu^{(\tau)})}{\rho}\right) \le \sum_{p'\in\mathcal{P}_k} \pi_{p'}^{(\tau)} \ell_{p'}^k(\mu^{(\tau)}) = \bar{\ell}^{k(\tau)}.$$

And by Hoeffding's lemma,

$$\log \sum_{p' \in \mathcal{P}_k} \pi_{p'} \exp\left(-\gamma_\tau \frac{\ell_{p'}(\mu^{(\tau)})}{\rho}\right) \leq -\gamma_\tau \sum_{p' \in \mathcal{P}_k} \pi_{p'}^{(\tau)} \frac{\ell_{p'}(\mu^{(\tau)})}{\rho} + \frac{\gamma_\tau^2}{8}.$$

Rearranging, we have $0 \leq \bar{\ell}^k(\tau) - \tilde{\ell}^k(\tau) \leq \frac{\rho\gamma_\tau}{8}$, therefore $U_p(\tau+1) = O(\gamma_\tau)$, and

$$\left\| \sum_{\tau=\tau_1}^{\tau_2} \gamma_\tau U(\tau+1) \right\| = O\left(\sum_{\tau=\tau_1}^{\tau_2} \gamma_t^2\right).$$

Finally, since $\gamma_\tau \downarrow 0$, for any fixed T, $\max_{\tau_2 : \sum_{\tau=\tau_1}^{\tau_2} \gamma_\tau \leq T} \sum_{\tau_1}^{\tau_2} \gamma_\tau^2$ converges to 0 as $\tau_1 \to \infty$; therefore condition (6.3) is verified. □

6.3. Convergence of AREP algorithms with sublinear discounted regret.

We now give the main convergence result.

THEOREM 6.10. *Suppose that the population strategies $(\mu^{(\tau)})_\tau$ obey an AREP update rule with sublinear discounted regret. Then $(\mu^{(\tau)})$ converges to the set of Nash equilibria $\mathcal{N}$.*

Proof. By assumption, we have

$$\mu_p^{(\tau+1)} - \mu_p^{(\tau)} = \gamma_\tau \left(G_p^k\left(\mu^{(\tau)}, \ell(\mu^{(\tau)})\right) + U_p^{(\tau+1)}\right) = \gamma_\tau\left(F_p^k(\mu^{(\tau)}) + U_p^{(\tau+1)}\right),$$

where, by definition of the AREP class, the perturbations $U^{(\tau)}$ satisfy condition 1 of Proposition 6.2. Condition 2 is also satisfied since the sequence $(\mu^{(\tau)})_\tau$ lies in the compact set Δ. Thus by Proposition 6.2, the affine interpolated process M of $(\mu^{(\tau)})_\tau$ is an APT of the continuous-time replicator system $\dot{\boldsymbol{\mu}} = F(\boldsymbol{\mu})$. Thus by Theorem 6.6, the limit set $L(M)$ is internally chain transitive.

Consider the set of restricted Nash equilibria $\mathcal{RN}$. This set is invariant ($\mathcal{RN}$ is the set of stationary points of the vector field) and compact ($\mathcal{RN}$ is the finite union of compact sets by Remark 5.3). The Rosenthal potential function V is a Lyapunov function for $\mathcal{RN}$ (see the proof of Theorem 4.5), and $V(\mathcal{RN})$ has an empty interior since it is a finite set by Remark 5.3. Therefore we can apply Proposition 6.7 to conclude that the set of limit points $L(M)$ is contained in $\mathcal{RN}$ and V is constant over $L(M)$. Let v be this constant value.

Next, we show that the sequence of potentials $V(\mu^{(\tau)})$ converges. Let $\hat{v}$ be a limit point of $V(\mu^{(\tau)})$. Then by Lemma 4.6, $\hat{v} = V(\hat{\mu})$, where $\hat{\mu}$ is a limit point of $(\mu^{(\tau)})$. In particular, $\hat{\mu} \in L(M)$, thus $\hat{v} = V(\hat{\mu}) = v$. This shows that the bounded sequence $(V(\mu^{(\tau)}))$ has a unique limit point v; therefore it converges to v, and it remains to show that $v = V_{\mathcal{N}}$ to conclude (by Lemma 4.6).

To show that $v = V_{\mathcal{N}}$, we first observe that since $V(\mu^{(\tau)}) \to v$, we also have $V(\mu^{(\tau)}) \xrightarrow{(\gamma_\tau)} v$. But the population dynamics is also assumed to have sublinear discounted regret; thus by Theorem 4.5, $V(\mu^{(\tau)}) \xrightarrow{(\gamma_\tau)} V_{\mathcal{N}}$. By uniqueness of the limit, we must have $v = V_{\mathcal{N}}$. □

Note that Theorem 6.10 assumes that the AREP update rule is applied to the population dynamics $(\mu^{(\tau)})$, not to individual strategies $\pi^{(\tau)}(x)$. One sufficient condition for $\mu^{(\tau)}$ to satisfy an AREP update is that for each k, all players in $\mathcal{X}_k$ start from a common initial distribution $\pi^{k^{(0)}} = \mu^{k^{(0)}}$, and apply the same update rule. This guarantees that for all τ and for all x, $\mu^{(\tau)} = \pi^{(\tau)}(x)$.

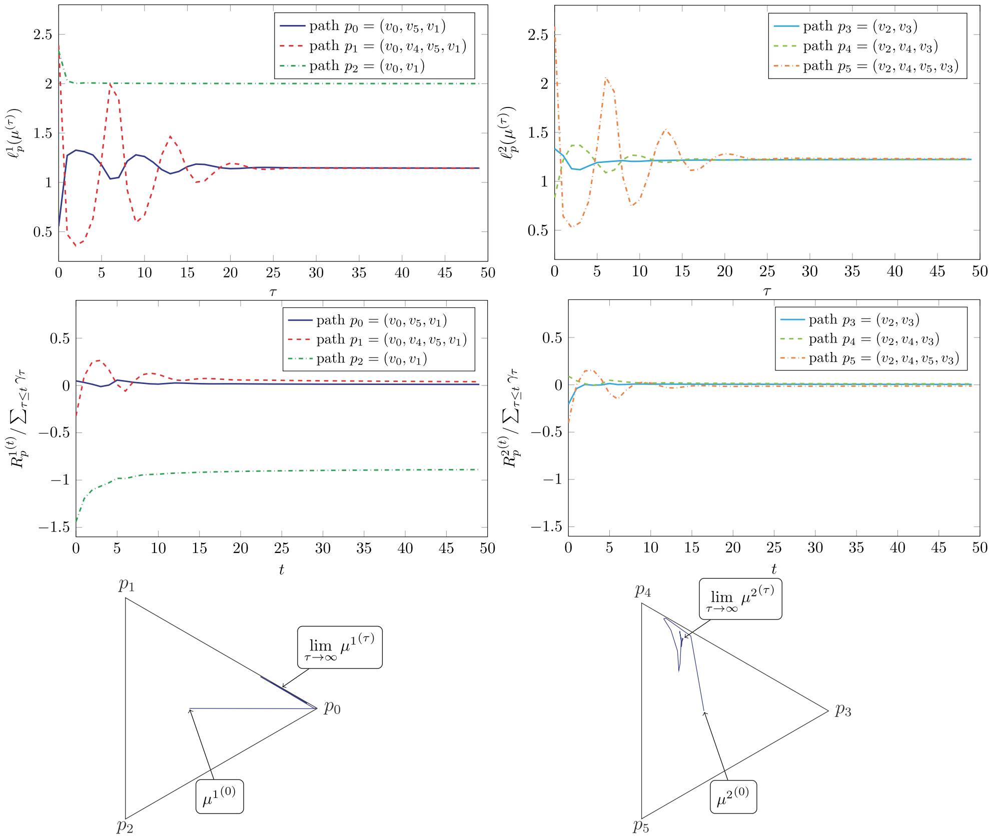

FIG. 3. *Simulation of the population dynamics under the discounted Hedge algorithm, initialized at the uniform distribution. The trajectories of the population strategies* $\mu^{k(\tau)}$ *are given in the* 2*-simplex for each population (bottom). The path losses* $\ell^k_p(\mu^{(\tau)})$ *for both populations (top) converge to a common value on the support on the Nash equilibrium. The sequences of discounted regrets (middle) confirm that the population regret is sublinear, i.e.,* $\limsup_{t\to\infty} \frac{R^{k(t)}}{\sum_{\tau\le t}\gamma_\tau} \le 0$.

6.4. Convergence of the REP and Hedge algorithms. We apply Theorem 6.10 to show convergence of the REP and Hedge algorithms.

COROLLARY 6.11. *If* $(\mu^{(\tau)})$ *obeys the REP update rule with learning rates* γ_τ *satisfying Assumption* 3.2 *and such that* $\gamma_\tau \le \frac{1}{2}$*, then* $\mu^{(\tau)} \to \mathcal{N}$.

Proof. The REP update rule is a discounted no-regret algorithm by Proposition 5.7, and it is an AREP algorithm with zero perturbations, so we can apply Theorem 6.10. ▯

COROLLARY 6.12. *If* $(\mu^{(\tau)})$ *obeys the discounted Hedge update rule with learning rates* γ_τ *satisfying Assumption* 3.2*, then* $\mu^{(\tau)} \to \mathcal{N}$.

Proof. By Propositions 3.8 and 6.9, the discounted Hedge algorithm with rates γ_τ is an AREP algorithm with sublinear discounted regret, and we can apply Theorem 6.10. ▯

We illustrate this convergence result with a routing game on the example network introduced in section 2.5. We simulate the population dynamics under the discounted Hedge algorithm with a harmonic sequence of learning rates, $\gamma_\tau = \frac{20}{10+\tau}$. The results are shown in Figure 3.

7. Conclusion.

We studied the convergence of online learning dynamics in the nonatomic congestion game. We showed that dynamics with sublinear discounted population regret guarantee the convergence of $(\bar{\mu}^{(\tau)})$, the sequence of Cesàro means of population strategies. To obtain convergence of the actual sequence of strategies $(\mu^{(\tau)})$, we introduced the AREP class of approximate replicator dynamics, inspired by the replicator ODE. We showed that whenever the population strategies obey an AREP dynamics and have sublinear discounted regret, the sequence converges. These results assume that the sequence of discount factors (γ_τ) is identical for all players. One question is whether this assumption can be relaxed, so that different players can use different learning rates.

Acknowledgments. We thank the anonymous reviewers and the editor for many insightful comments and suggestions.

REFERENCES

[1] S. Arora, E. Hazan, and S. Kale, *The multiplicative weights update method: A meta-algorithm and applications*, Theory Comput., 8 (2012), pp. 121–164.

[2] J.-Y. Audibert and S. Bubeck, *Minimax policies for adversarial and stochastic bandits*, in Proceedings of COLT, 2009.

[3] B. Awerbuch and R. D. Kleinberg, *Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches*, in Proceedings of the 36th Annual ACM Symposium on Theory of Computing, STOC '04, New York, 2004, pp. 45–53.

[4] M. Benaïm, *Dynamics of stochastic approximation algorithms*, in Séminaire de probabilités XXXIII, Lecture Notes in Math. 1709, Springer, New York, 1999, pp. 1–68.

[5] A. Blum, E. Even-Dar, and K. Ligett, *Routing without regret: On convergence to Nash equilibria of regret-minimizing algorithms in routing games*, in Proceedings of the 25th Annual ACM Symposium on Principles of Distributed Computing, New York, 2006, pp. 45–52.

[6] A. Blum and Y. Mansour, *Learning, regret minimization, and equilibria*, in Algorithmic Game Theory, Cambridge University Press, Cambridge, UK, 2007, pp. 79–101.

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2010.

[8] S. Bubeck and N. Cesa-Bianchi, *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*, Found. Trends Machine Learning, 5 (2012), pp. 1–122.

[9] S. Bubeck, V. Perchet, and Philippe Rigollet, *Bounded Regret in Stochastic Multi-armed Bandits*, CoRR abs/1302.1611, 2013.

[10] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, Cambridge, UK, 2006.

[11] S. Fischer and B. Vöcking, *On the evolution of selfish routing*, in Algorithms: Proceedings of ESA 2004, Lecture Notes in Comput. Sci. 3221, Springer, New York, 2004, pp. 323–334.

[12] M. J. Fox and J. S. Shamma, *Population games, stable games, and passivity*, Games, 4 (2013), pp. 561–583.

[13] D. H. Fremlin, *Measure Theory*, vol. 4, Torres Fremlin, Colchester, UK, 2000.

[14] Y. Freund and R. E. Schapire, *Adaptive game playing using multiplicative weights*, Games Econom. Behav., 29 (1999), pp. 79–103.

[15] A. György, T. Linder, G. Lugosi, and G. Ottucsák, *The on-line shortest path problem under partial monitoring*, J. Mach. Learn. Res., 8 (2007), pp. 2369–2403.

[16] J. Hofbauer and W. H. Sandholm, *Stable games and their dynamics*, J. Econom. Theory, 144 (2009), pp. 1665–1693.

[17] H. K. Khalil, *Nonlinear Systems*, Macmillan, New York, 1992.

[18] J. Kivinen and M. K. Warmuth, *Exponentiated gradient versus gradient descent for linear predictors*, Inform. Comput., 132 (1997), pp. 1–63.

[19] R. Kleinberg, G. Piliouras, and E. Tardos, *Multiplicative updates outperform generic no-regret learning in congestion games: Extended abstract*, in Proceedings of the 41st Annual ACM Symposium on Theory of Computing, 2009, pp. 533–542.

[20] E. Koutsoupias and C. Papadimitriou, *Worst-case equilibria*, in Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science, 1999, pp. 404–413.

[21] N. Littlestone and M. K. Warmuth, *The weighted majority algorithm*, in 30th Annual Symposium on Foundations of Computer Science, IEEE, 1989, pp. 256–261.

[22] M. Muresan, *A Concrete Approach to Classical Analysis*, Springer, New York, 2009.

[23] J. Nash, *Non-cooperative games*, Ann. Math., 54 (1951), pp. 286–295.

[24] R. W. Rosenthal, *A class of games possessing pure-strategy Nash equilibria*, Internat. J. Game Theory, 2 (1973), pp. 65–67.

[25] T. Roughgarden, *Routing games*, in Algorithmic Game Theory, Cambridge University Press, Cambridge, UK, 2007, pp. 461–486.

[26] T. Roughgarden and É. Tardos, *How bad is selfish routing?*, J. ACM, 49 (2002), pp. 236–259.

[27] W. H. Sandholm, *Potential games with continuous player sets*, J. Econom. Theory, 97 (2001), pp. 81–108.

[28] J. G. Wardrop, *Some theoretical aspects of road traffic research*, in ICE Proceedings: Engineering Divisions, vol. 1, Thomas Telford, London, 1952, pp. 325–362.

[29] J. W. Weibull, *Evolutionary Game Theory*, MIT Press, Cambridge, MA, 1997.