Playing Congestion Games with Bandit Feedbacks (Extended Abstract)

Po-An Chen*
National Chiao Tung University
1001 University Road
Hsinchu, Taiwan
poanchen@nctu.edu.tw

Chi-Jen Lu
Academia Sinica
128 Academia Road, Sec. 2
Nankang, Taipei 115, Taiwan
cjlu@iis.sinica.edu.tw

ABSTRACT

Almost all convergence results from each player adopting specific "no-regret" learning algorithms such as multiplicative updates or the more general mirror-descent algorithms in repeated games are only known in the more generous information model, in which each player is assumed to have access to the costs of all possible choices, even the unchosen ones, at each time step. This assumption in general may seem too strong, while a more realistic one is captured by the bandit model, in which each player at each time step is restricted to know only the cost of her currently chosen path, but not any of the unchosen ones. Can convergence still be achieved in such a more challenging bandit model? We answer this question positively. While existing bandit algorithms do not seem to work here, we develop a new family of bandit algorithms based on the mirror-descent algorithm with such a guarantee in atomic congestion games.

Categories and Subject Descriptors

Theory of computation [**Theory and algorithms for application domains**]: Algorithmic game theory and mechanism design—*Convergence and learning in games*

Keywords

Mirror-descent algorithm; No-regret dynamics; Convergence

1. INTRODUCTION

Recurrent strategic scenarios are usually modeled as repeated games in game theory, and "no-regret" algorithms from the area of online learning [3] are good candidates to model learning players. It is known that in congestion games, when each player adopts specific "no-regret" learning algorithms such as multiplicative updates or the more general mirror-descent algorithms, their joint strategy profile will quickly approach an approximate Nash equilibrium. These results, however, are all based on somewhat generous information models. For instance of congestion games, edge cost functions are assumed common knowledge in the full-information model of [7]. More stringent information model was considered for the load-balancing games in [6] and later for the congestion games in [4], in which the edge cost functions are not common knowledge anymore, but still the cost values of all paths at each step are assumed available through "bulletin board" posting. With such global information, players can get a grasp of the costs corresponding to played and even unplayed strategies, which allows players to update their strategies better and makes convergence of the whole system potentially easier. However, such a strong assumption on the information availability may not always be realistic and may limit the applicability of these results.

The bandit model in online learning on the other hand considers a probably more realistic and prevalent setting, in which at each time step each player only knows the cost of her (or his) own currently played strategy, but not any costs of unplayed strategies. In the area of online learning, many bandit algorithms with no-regret guarantee have been developed, including for example those based on the multiplicative updates algorithm for the experts problem [1] and those based on the gradient-descent algorithm for online linear or convex optimization [5]. However, not much is known in the area of game theory for playing repeated games in the bandit model. Although similar convergence results of average plays can be established immediately for bandit algorithms with no-regret guarantee, we are not aware of any previous result establishing convergence of actual plays in the bandit setting. This motivates us to ask the question: are there natural classes of bandit algorithms which selfish players individually have incentive to adopt and the whole system will quickly converge to an approximate Nash equilibrium that is also beneficial to the whole in terms of natural social cost measures?

We answer this question affirmatively. For the class of atomic congestion games, we propose a family of bandit algorithms based on the mirror-descent algorithms [2, 4], and we show that when each player individually adopts such a bandit algorithm, their joint strategy profile quickly approaches an approximate Nash equilibrium. In addition, as in [4], one can show that the approached approximate equilibria also have good quality in terms of some measures of social cost, including the average individual cost and maximum individual cost.

2. PRELIMINARIES

In this paper, we consider the following *atomic congestion game*, described by $(N, E, (\mathcal{S}_i)_{i\in N}, (c_e)_{e\in E})$: N is the set of

*Supported in part by NSC 102-2221-E-009-061-MY2.

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

players, E is the set of edges (resources), $\mathcal{S}_i \subseteq 2^E$ is the collection of allowed paths (subsets of resources) for player i, and c_e is the cost function of edge e, which is a nondecreasing function of the amount of flow on it. Let us assume that there are n players, each player has at most d allowed paths, each path has length at most m, each allowed path of a player intersects at most k allowed paths (including that path itself) of that player, and each player has a flow of amount $1/n$ to route. The strategy of each player i is to send her entire flow on a single path, chosen randomly according to some distribution over her allowed paths, which can be represented by a $|\mathcal{S}_i|$-dimensional vector $\pi_i = (\pi_{i,s})_{s\in\mathcal{S}_i}$, where $\pi_{i,s} \in [0,1]$ is the probability of choosing path s. It turns out to be more convenient for us to represent each player's strategy π_i by an equivalent form $x_i = (1/n)\pi_i$, where $1/n$ is the amount of flow each player has. That is, for every $i \in N$ and $s \in \mathcal{S}_i$, $x_{i,s} = (1/n)\pi_{i,s} \in [0, 1/n]$ and $\sum_{s\in\mathcal{S}_i} x_{i,s} = 1/n$. Let $\mathcal{K}_i$ denote the feasible set of all such $x_i \in [0,1/n]^{|\mathcal{S}_i|}$ for player i, and let $\mathcal{K} = \mathcal{K}_1 \times \cdots \times \mathcal{K}_n$, which is the feasible set of all such joint strategy profiles $x = (x_1, \ldots, x_n)$ of the n players.

As in [4], we consider the following potential function:

$$\Phi(x) = \sum_{e\in E} \int_0^{\ell_e(x)} c_e(y)dy, \qquad (1)$$

for $x \in \mathcal{K}$, where $\ell_e(x) = \sum_{j\in N} \sum_{r\in\mathcal{S}_j : e\in r} x^t_{j,r}$. According to [4], Φ is a convex function. As in [6, 4], we assume that the cost functions satisfy the property that there exist constants a, b such that for any $e \in E$,

$$ay \le c_e(y) \le by \text{ and } c''_e(y) \le b, \text{ for any } y \in [0,1]. \qquad (2)$$

Then the potential function defined with respect to such cost functions is smooth in the following sense.

Definition 1. *A function Φ over $\mathcal{K}$ is called (α,β,λ)-smooth if for any $x \in \mathcal{K}$, $\Phi(x) \le \alpha, \|\nabla\Phi(x)\|_\infty \le \beta$, and $\nabla^2\Phi(x) \preceq \lambda I$.*

Proposition 2. *The function Φ defined in (1) with cost functions satisfying condition (2) is (α,β,λ)-smooth, for $\alpha = bm/2, \beta = bm$, and $\lambda = bmk$.*

3. OUR RESULTS

In this more challenging bandit model, each player i does not know the whole vector $\nabla_i\Phi(x^t)$. Instead, the only information player i has after choosing a path s_i is $c_{s_i}(X^t) = \sum_{e\in s_i} c_e(\ell_e(X^t))$, where X^t is the choice vector of players at step t. In order to follow the framework of [4], each player i needs to have a way to approximate the vector $\nabla_i\Phi(x^t)$, her portion of the gradient vector $\nabla\Phi(x^t)$. Our basic idea is for each player to divide the time steps into episodes, each consisting of a consecutive number of steps, and to do the following in each episode. During episode τ, each player i plays some fixed strategy x_i^τ for all the steps (instead of playing different strategies in different steps), collects statistics to obtain an estimate $\hat{g}_i^\tau$ for $\nabla_i\Phi(x^\tau)$, and at the end of the episode uses $\hat{g}_i^\tau$ to update her strategy for the next episode. Two keys are: how to come up with the estimate $\hat{g}_i^\tau$ and how to update the next strategy.

By setting the number of steps in each episode τ properly, we have the following.

Lemma 3. *With high probability, each player i in each episode τ can have $\|\hat{g}_i^\tau - \nabla_i\Phi(x^\tau)\|_\infty \le 4bm/n$.*

Having obtained a good estimate $\hat{g}_i^\tau$, as guaranteed by Lemma 3, we then follow [4] and consider the following update rule for each player i's strategy of the next episode:

$$x_i^{\tau+1} = \arg\min_{z_i\in\bar{\mathcal{K}}_i} \left\{ \eta_i \langle \hat{g}_i^\tau, z_i\rangle + \mathcal{B}^{R_i}(z_i, x_i^\tau) \right\} \qquad (3)$$

Here, $\eta_i > 0$ is some learning rate, R_i is some regularization function, and $\mathcal{B}^{R_i}(\cdot,\cdot)$ is the Bregman divergence with respect to R_i.

We analyze the behavior of the system of players adopting the algorithm described in the previous section. Our main result is the following, which shows that the system indeed quickly converges, in the sense that the value of the potential function $\Phi(x^\tau)$ quickly comes near the minimum value $\Phi(q)$, where $q = \arg\min_{z\in\mathcal{K}} \Phi(z)$, and then stays close afterwards.

Theorem 4. *Consider any atomic congestion game of n players, with a potential function Φ which is (α,β,λ)-smooth, and let $q = \arg\min_{z\in\mathcal{K}} \Phi(z)$. Suppose each player i updates her strategy according to the rule in (3), with $\eta_i \le 1/\lambda$, using $\hat{g}_i^\tau$ with the guarantee that $\|\hat{g}_i^\tau - \nabla_i\Phi(x^\tau)\|_\infty \le \epsilon$. Consider any η such that $\eta \le \eta_i$ for any i and $\theta = \sqrt{\eta\Gamma\epsilon n} \le 1$ such that $\Gamma \cdot \mathcal{B}^{R_i}(x_i, y_i) \le \|x_i - y_i\|_2^2$, for any i and any feasible x_i, y_i, and let $\delta = 6\epsilon + \theta\beta d\Lambda$ for some parameter Λ, which implies that each path s is chosen with probability $\pi_{i,s} \ge \Lambda$. Then for $\tau_0 = \alpha/\delta$ and $\triangle = 3\delta/\theta$, it holds that for any $\tau \ge \tau_0$,*

$$\Phi(x^\tau) \le \Phi(q) + \triangle.$$

Future Work

One of the immediate future work could be extending such framework to other partial-information models by other suitable gradient estimation methods. A different line of future work would be to consider appropriate bandit scenarios for market equilibrium problems and to see if generalized mirror-descents with approximate gradients works there.

REFERENCES

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1), 2003.

[2] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

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

[4] P.-A. Chen and C.-J. Lu. Generalized mirror descents in congestion games with splittable flows. In *Proc. AAMAS*, 2014.

[5] A. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proc. SODA*, 2005.

[6] R. Kleinberg, G. Piliouras, and E. Tardos. Load balancing without regret in the bulletin board model. In *Proc. PODC*, 2009.

[7] R. Kleinberg, G. Piliouras, and E. Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proc. STOC*, 2009.