Monotone multi-armed bandit allocations

Aleksandrs Slivkins SLIVKINS@MICROSOFT.COM
Microsoft Research Silicon Valley, Mountain View, CA 94043, USA

We present a novel angle for multi-armed bandits (henceforth abbreviated MAB) which follows from the recent work on *MAB mechanisms* (Babaioff et al., 2009; Devanur and Kakade, 2009; Babaioff et al., 2010). The new problem is, essentially, about designing MAB algorithms under an additional constraint motivated by their application to MAB mechanisms.

This note is self-contained, although some familiarity with MAB is assumed; we refer the reader to Cesa-Bianchi and Lugosi (2006) for more background.

1. Problem formulation

We start with a slightly non-standard formalism for MAB.

Definition 1 (MAB) *In each round, an algorithm selects among k alternatives (*arms*), and collects a reward. The rewards are fixed in advance, but not revealed to the algorithm. Specifically, one fixes a table whose (i,t)-th entry is the reward of arm i in round t; such table is called a* realization *(of the rewards). The realization is generated by a random process (*generator*). The algorithm only knows that the generator belongs to some set (of generators) called the* MAB domain.[1]

Two natural special cases are *stochastic rewards*: for each arm i, the reward of this arm in every given round is an independent sample from the same distribution $\mathcal{D}_i$, and *adversarial rewards*: the realization is chosen by an adversary.

Definition 2 (MAB allocation rule) *An* MAB allocation rule *is an MAB algorithm which initially inputs a vector of* bids $b_i \in [0,1]$ *for each arm i; the rewards received from arm i (*raw rewards*) are then scaled by a factor of b_i. An MAB allocation rule is called* monotone *(resp.,* ex-post monotone*) if for each agent arm i, increasing b_i and keeping all other bids fixed cannot decrease the raw reward from arm i, in expectation over the realization (resp., for every realization).*

Now we are ready to state the problem:

Problem 1 *For a given MAB domain, design an MAB allocation rule so as to maximize the total reward subject to the constraint of (ex-post) monotonicity.*

[1] The notation "MAB domain" is not standard; we adopt it here for the ease of presentation.

It is worth emphasizing that Problem 1 is well-defined for each of the numerous MAB domains studied in the literature: stochastic or adversarial, with or without priors on rewards, with or without auxiliary or contextual information, and so forth. For each MAB domain there is a corresponding version of Problem 1. The main qualitative issue is whether and by how much the additional constraint of (ex-post) monotonicity impacts performance.

2. Brief motivation

The motivating example is the following idealized model for selling ad slots to potential advertisers. Several advertisers, each with a single ad, are competing for a single ad slot that is displayed to multiple users over time. Each advertiser derives value only if her ad is clicked by a user. Thus, each arm corresponds to an advertiser, the bid corresponds to the value per click, and the raw reward in each round is 1 if clicked and 0 otherwise.

To motivate the (ex-post) monotonicity property we need to extend the above model to a simple auction, called *MAB auction*: first all advertisers submit their bids, then the ad slots are allocated using an MAB allocation rule, and finally the ad platform assigns how much each agent needs to pay. Crucially, the value per click (v_i) is a *private information*: it is known to the advertiser, but not to the ad platform. Thus, an advertiser may *lie* about it ($b_i \neq v_i$) if she thinks it may benefit her. This setting have been defined independently and concurrently in Devanur and Kakade (2009) and Babaioff et al. (2009), and further studied in Babaioff et al. (2010); see Appendix A for some background and motivation.

In an MAB auction, the monotonicity property of the MAB allocation rule is necessary and sufficient (with appropriate payments) to incentivize the advertisers to submit truthful bids. Monotonicity corresponds to incentives in expectation over realizations, whereas a stronger and more desirable ex-post monotonicity corresponds to incentives for every given realization. The details are deferred to Appendix A; these details are not necessary for understanding Problem 1.

It is worth noting that the *social welfare* of an MAB auction – a standard performance benchmark for auctions – coincides with the total reward of the corresponding allocation rule. Social welfare is defined as the total utility: the sum of utilities of all advertisers and the utility (profit) of the ad platform (so payments cancel out).

3. Current status

The problem has been resolved for stochastic rewards in the strongest possible sense: there exists an ex-post monotone MAB allocation rule whose regret is essentially optimal among all MAB allocation rules (Babaioff et al., 2010). Moreover, an MAB allocation rule based on (a version of) a well-known MAB algorithm `UCB1` (Auer et al., 2002a) is proved to be monotone; this result extends to a more general family of MAB allocation rules. However, `UCB1`-based MAB allocation rule is *not* ex-post monotone as is; making it ex-post monotone seems to require significant modifications.

For all other MAB domains the problem is open. One particularly appealing target is adversarial rewards.[2] Babaioff et al. (2009) exhibit an ex-post monotone MAB allocation rule (based on the MAB algorithm in Awerbuch and Kleinberg (2008)) which separates

[2] Note that for adversarial rewards, the notions of ex-post monotonicity and monotonicity coincide.

exploration and exploitation and has regret $\tilde{O}(T^{2/3})$. This is in contrast to optimal MAB algorithms such as EXP3 (Auer et al., 2002b) that achieve regret $\tilde{O}(\sqrt{T})$. We conjecture that EXP3-based MAB allocation rule is *not* ex-post monotone. It is not clear whether an MAB allocation rule with regret $\tilde{O}(\sqrt{T})$ can be ex-post monotone.

Acknowledgments

We thank Robert Kleinberg for his comments and suggestions.

References

Aaron Archer and Éva Tardos. Truthful mechanisms for one-parameter agents. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 482–491, 2001.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a. Preliminary version in *15th ICML*, 1998.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b. Preliminary version in *36th IEEE FOCS*, 1995.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, February 2008. Preliminary version appeared in *36th ACM STOC*, 2004.

Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *10th ACM Conf. on Electronic Commerce (EC)*, pages 79–88, 2009. Full version available at http://arxiv.org/abs/0812.2291.

Moshe Babaioff, Robert Kleinberg, and Aleksandrs Slivkins. Truthful mechanisms with implicit payment computation. In *11th ACM Conf. on Electronic Commerce (EC)*, pages 43–52, 2010. Best Paper Award. Full version available at http://arxiv.org/abs/1004.3630.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

Nikhil Devanur and Sham M. Kakade. The price of truthfulness for pay-per-click auctions. In *10th ACM Conf. on Electronic Commerce (EC)*, pages 99–106, 2009.

Roger B. Myerson. Optimal Auction Design. *Mathematics of Operations Research*, 6:58–73, 1981.

N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani (eds.). *Algorithmic Game Theory.* Cambridge University Press., 2007.

Tim Roughgarden. An algorithmic game theory primer. IFIP International Conference on Theoretical Computer Science (TCS). An invited survey., 2008.

Balasubramanian Sivan and Christopher A. Wilkens. Single-call mechanisms, 2010. Available at http://arxiv.org/abs/1011.6134.

Appendix A. MAB auctions: motivation and background

Sponsored search and MAB auctions. Sponsored search on the Web is a billions dollar market in which ad slots are sold to potential advertisers. The ad slots are usually sold via auctions, with numerous auctions running every second. In a given auction, multiple advertisers compete for a well-defined selection of ad slots, such as ad slots that appear next to search results on a given keyword. The predominant model, called *pay-per-click* (*PPC*) auctions, is that an advertiser pays only if her ad is clicked by a user. The underlying assumption here is that an advertiser derives value only if her ad is clicked, and the PPC model shields advertisers from the risk that the allocation of ads to ad slots may be suboptimal for a given advertiser. A typical PPC auction proceeds as follows: advertisers submit their preferences, then the ad platform allocates ad slots between the advertisers, records the clicks, and then the payments are assigned based on the submitted preferences, the allocation, and the observed clicks.

In general, optimizing performance in a PPC auction requires knowing or estimating the rates at which the ads are clicked (*click-through rates*, or *CTRs*). Most prior work assumes that these rates are known or estimated externally (see Babaioff et al. (2009) for a more thorough discussion).

A recent line of work (Devanur and Kakade, 2009; Babaioff et al., 2009, 2010) considers the problem of designing truthful PPC auctions when the process of learning the CTRs is explicitly treated as a part of the game. The motivation is that the self-interested advertisers would take this process into account, as it influences their utility. These papers are mainly interested in the interplay of the two key issues: the *strategic* issue (the issue of incentives) and the issue of learning CTRs from the clicks that are observed over time. They define and study a clean model, called *MAB auctions*, which abstracts these two issues. This model (which we described in Section 2) can be seen as a natural *strategic* version of MAB.

Incentives and monotonicity. As we alluded in Section 2, the issue of incentives is crucial. The goal is to design mechanisms that are *incentive-compatible* in the following sense: every agent maximizes her expected utility by bidding truthfully, for any bids of the others. This is a standard notion in the literature on Mechanism Design.[3] A stronger notion, *ex-post incentive-compatibility*, is the same for each realization of the clicks (rather than in expectation). Thus, we design MAB auctions so as to maximize performance subject to the constraint of (ex-post) incentive-compatibility. Two standard performance measures in the literature are social welfare and total profit.

A well-known general result in Mechanism Design (Myerson, 1981; Archer and Tardos, 2001) implies that an MAB allocation rule can be "extended" to an (ex-post) incentive-compatible MAB mechanism if and only if the MAB allocation rule is (ex-post) monotone. However, naively computing the payments in the "if" direction of this result requires information about the realization that is not revealed during a given run of the mechanism. Devanur and Kakade (2009) and Babaioff et al. (2009) show that incomputability of pay-

[3] Mechanism Design is a branch of Economics that is concerned with, essentially, the design of auctions. (A *mechanism* is a slightly more general technical term.) A Computer Science take on Mechanism Design (termed *Algorithmic Mechanism Design*) is a design of algorithms whose inputs are provided by self-interested agents, which brings about additional incentive constraints. For background see the book Nisan et al. (2007) and a survey Roughgarden (2008).

ments is a real obstacle. They show that ex-post truthfulness implies severe limitations on the structure of a deterministic MAB allocation rule: essentially, it must separate exploration and exploitation. For the natural example of stochastic clicks it leads to very suboptimal performance (compared to arbitrary MAB allocation rules), both in terms of social welfare (Babaioff et al., 2009) and in terms of profit (Devanur and Kakade, 2009).

Babaioff et al. (2010) overcome this obstacle for *randomized* MAB mechanisms: they provide a general procedure to transform any (ex-post) monotone MAB mechanism to a randomized (ex-post) incentive-compatible MAB mechanism with a very minor loss in performance.[4] This result motivates Problem 1.

[4] The downside of this transformation is a relatively high variance in payments, see Babaioff et al. (2010) for further discussion. However, a very recent result of Sivan and Wilkens (2010) shows that this amount of variance is essentially optimal for any such general transformation.