Approximation Algorithms for Budgeted Learning Problems *

Sudipto Guha† Kamesh Munagala‡

In this talk we investigate general techniques for designing simple approximately optimal policies for the problems similar to the optimal design of experiments under a budget constraint. As a simplest example, consider first the following "coins" problem: We are given n coins and told that the probability p of obtaining head on tossing coin i is distributed according to a known prior distribution $\mathcal{R}_i$ over $[0, 1]$. We can toss any coin any number of times – but we are limited to 100 tosses. If we wish to maximize the probability of heads, how should we proceed? What happens if the coin has 3 or more outcomes?

The above is a classic problem in budgeted learning and can be posed in the bandit framework. We are given a bandit with n arms. The reward when arm i is played follows some fixed distribution R_i. This distribution is over K non-negative values $\{a_1, a_2, \ldots, a_K\}$. In the above coins example, K was equal to 2. We assume that $0 = a_1 \leq a_2 \ldots, a_K = 1$ by suitable scaling. The arms are assumed to be independent, so that the distributions for any i are independent of the corresponding distributions for an $i' \neq i$. The distributions R_i are *not known* to the player, but the player knows the *prior distribution* $\mathcal{R}_i$ from which the corresponding R_i is drawn. Whenever the arm i is played, the player gets an outcome that depends on the true fixed distribution R_i. In case of the coin toss ($K = 2$), the R_i will be the probability vector $(1 - p_i, p_i)$ corresponding to coin i and we were told of the distribution of p_i. The player can estimate R_i by playing the arm i and observing the reward. There is a cost c_i for playing arm i and a cost budget C is given. The arms are played in some adaptively determined order depending on the outcome of the previous plays, as long as the total cost of playing does not exceed C. When the budget C is exhausted, each arm has been played a certain number of times, we choose an arm which maximizes the posterior expected reward based on the outcomes. The **goal** is to devise an adaptive budget-constrained exploration strategy whose expected exploitation gain is maximized.

Computing the optimal solution to the above problem is NP-Hard even when at most one play is allowed per arm [8, 1], and even with unit cost ($c_i = 1$) for any play [6]. Several heuristics have been proposed for the budgeted multi-armed bandit problem [8, 9]. Extensive empirical comparison of these heuristics is presented in [8]. They show that unlike the celebrated stochastic infinite-horizon discounted reward multi-armed bandit problem [5, 10], the optimal policy in the budgeted case *is not an index policy*. Obtaining policies with non-trivial bounds were posed as an open problem in [8].

Our Results. Based on our work in [7], we present simple policies for the budgeted multi-armed bandit problem which are guaranteed to achieve 1/4 of the optimal exploitation gain. These represent the first polynomial time algorithms which produce policies with *any non-trivial* guarantees on the quality of the policy generated. Our policy design is via a linear programming formulation and stochastic packing [3, 4, 2]. Our framework also extends naturally to solve Lagrangean variants where there is no explicit budget constraint, and the optimization objective is the *difference* between the exploitation gain and the total cost spent in exploration. The policies obtained through the linear program are natural and simple, and serve as the basis for heuristics which empirically have near-optimal performance. Furthermore, though our policies have natural interpretations based on retirement rewards, they are *not* index policies. This raises several interesting issues including the applicability of our techiques to designing polynomial-time computable near-optimal policies for more general Markov Decision Processes, and the possibility that the performance of several well known index heuristics can be analyzed.

*Learning Algorithms: Oral Presentation

†Department of Computer and Information Sciences, University of Pennsylvania. Email: sudipto@cis.upenn.edu. Research supported in part by an Alfred P. Sloan Research Fellowship and by an NSF Award CCF-0430376.

‡Department of Computer Science, Duke University, Durham NC 27708-0129. Email: kamesh@cs.duke.edu. Research supported in part by NSF CNS-0540347. **Will Attend**

References

[1] V. Conitzer and T. Sandholm. Definition and complexity of some basic metareasoning problems. In *IJCAI*, pages 1099–1106, 2003.

[2] B. Dean. *Approximation Algorithms for Stochastic Scheduling Problems*. PhD thesis, MIT, 2005.

[3] B. C. Dean, M. X. Goemans, and J. Vondrak. Approximating the stochastic knapsack problem: The benefit of adaptivity. In *FOCS '04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 208–217, 2004.

[4] B. C. Dean, M. X. Goemans, and J. Vondrák. Adaptivity and approximation for stochastic packing problems. In *Proc.* 16^{th} *ACM-SIAM Symp. on Discrete algorithms*, pages 395–404, 2005.

[5] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. *Progress in statistics (European Meeting of Statisticians)*, 1972.

[6] A. Goel, S. Guha, and K. Munagala. Asking the right questions: Model-driven optimization using probes. In *Proc. ACM Symp. on Principles of Database Systems*, 2006.

[7] S. Guha and K. Munagala. Approximation algorithms for budgeted learning problems. *Manuscript, under review. Email for copy.*, 2006.

[8] O. Madani, D. J. Lizotte, and R. Greiner. Active model selection. In *UAI '04: Proc. 20th Conf. on Uncertainty in Artificial Intelligence*, pages 357–365, 2004.

[9] J. Schneider and A. Moore. Active learning in discrete input spaces. *The 34th Interface Symposium, Montreal, Quebec*, 2002.

[10] J. N. Tsitsiklis. A short proof of the Gittins index theorem. *Annals of Applied Probability*, 4(1):194–199, 1994.