

Research Statement

Varsha Dani

December, 2007

My research so far has focussed on the area of Online Optimization which was the subject of my Ph.D. thesis. I have also done some related (and unrelated) work in Machine Learning, Game Theory and Economics.

In the next few years, I hope to branch out into more areas within Machine Learning, which seems to be generating a lot of exciting new problems. In particular, I am interested in learning more about reinforcement learning and also applications of Machine Learning to Finance. Another direction I would like to go is in applying more sophisticated concentration-of-measure inequalities, such as Talagrand's inequality, in the context of Machine Learning, an approach which has proved very fruitful in my earlier work on Online Optimization. I would also generally like to broaden my scope and work on new and interesting problems.

I am also actively extending my previous work on Online Optimization, to deal with more complicated stochastic models and more kinds of loss functions, as discussed in the next section.

Online Optimization

The multi-armed bandit problem is a foundational model for sequential decision-making with missing information. First introduced by Robbins [14] in the context of the sequential design of statistical experiments such as clinical drug trials, it has subsequently been applied in such diverse areas as adaptive routing in networks, reinforcement learning and boosting, prediction markets, and financial portfolio rebalancing.

In the traditional multi-armed bandit problem, over a sequence of T rounds, a decision maker must choose a decision from a set of K options with unknown outcomes. Simultaneously, the environment assigns a cost to each outcome in an unknown way. After each such round, the algorithm gets the cost of its chosen decision as feedback which may help it make better decisions in future rounds. The canonical measure of performance is the "regret," defined as the difference between the cost incurred by the decision maker and the cost of the single best decision which could have been made with full knowledge of the environment. This model has been rather well-studied over the years and is now essentially completely solved (see Auer et al [3]). The best regret possible is \sqrt{KT} , up to a sublogarithmic factor.

Often we are confronted with decision sets D of very large size (possibly even infinite); if the environment can choose arbitrary cost functions, then, as noted, the above regret bound is the best that can be achieved. However in many settings, the large decision sets have some structure, and can be (isometrically) embedded into a low-dimensional Euclidean space so that the cost functions chosen by the environment are linear functions of the decisions. A case in point is the online network routing problem, also known as the "Drive to Work" problem. Here, on every round a path must be chosen from the set of all paths between two fixed nodes in the network. Although this decision set may be exponential in the size of the network, since the cost (or latency) of a path is the sum of the latencies on each of its edges, its inherent dimension is just the number of edges in the network.

In this setting, one hopes for a better regret bound than the $\sqrt{|D|T}$ bound guaranteed by the results of Auer et al. [3], one that depends nicely on the inherent dimensionality of the problem, rather than the size of the decision set, which may be exponential in the dimension. A simple approach to this, used by Awerbuch and Kleinberg [4] and McMahan and Blum [13], separates the timeline into exploration rounds and exploitation rounds. During the exploration rounds, their algorithms play random decisions from a basis for the decision space and use the observed costs to construct estimators for the true costs. During the exploitation rounds, these estimators are used to select what seems to be the best decision. On these rounds the feedback is discarded. Using this approach and optimizing the fraction of exploration rounds, they were able to prove expected regret bounds with a polynomial dependence on the dimension, n , at the expense of the dependence on the time horizon T ($O(\text{poly}(n)T^{2/3})$ in [4], $O(\text{poly}(n)T^{3/4})$ in [13] against a more powerful

model for the environment). In joint work with Tom Hayes, [10] we improved the regret bounds of McMahan and Blum [13] to $O(nT^{2/3})$ and showed that this was the best that could be achieved by any separated-timeline approach. In subsequent work with Tom Hayes and Sham Kakade [8] we used estimation techniques inspired by linear regression to construct an algorithm whose expected regret is at most $O(n^{3/2}\sqrt{T})$. Here again we use an exploration vs. exploitation tradeoff. However, unlike the aforementioned separated-timeline approach, here the feedback is never discarded, leading to the improved bound. In more recent work, with Bartlett *et al.* [5], we combined the above technique with a “dynamic variance compensation” approach used by Auer *et al.* [3], thereby obtaining high-probability regret bounds of $O(n^{3/2}\sqrt{T})$. The best known lower bound for this problem is $\Omega(n\sqrt{T})$.

The original 1952 paper of Robbins [14] introduced the multi-armed bandit problem for independent identically distributed costs drawn from a fixed but unknown underlying distribution. In this case it is reasonable to seek *deterministic* algorithms achieving low regret. To solve this problem, Lai and Robbins [12] proposed the following elegant idea (described here for the 2-arm bandit case): Most of the time, choose the apparently better decision, based on past observations, choosing the apparently worse decision just often enough to be very confident that its true mean is in fact worse. To this end, they proposed the idea of keeping track of an “upper confidence bound” for each population. Combining this idea with the power of Chernoff’s bounds leads to the following very clean formulation (see, e.g., Agrawal [1]). For each population, keep track of the empirical mean of observations, as well as the number of times it has been sampled, k . By Chernoff’s bound, the empirical mean plus $\sqrt{2\ln(1/\delta)/k}$ is an upper bound for the true mean with probability at least $1 - \delta$. Now simply, in each round, choose the population with the highest upper confidence bound. This algorithm has the desirable property of being *self-correcting*: each time the truly worse population is selected, we expect its upper confidence bound to decrease, thereby reducing our tendency to select it again. This algorithm achieves a regret of $O(\log T)$.

The above algorithm extends to any finite decision set with essentially no modifications. However the bounds obtained scale linearly with the size of the decision set, making them impractical for stochastic online linear optimization, when the cardinality of the decision set is large compared to its inherent dimension n . Auer [2] gave an algorithm achieving $O^*(\text{poly}(n)\sqrt{T})$ regret for finite decision sets. At the core of his algorithm is a natural generalization of the UCB algorithm described above, which, at every step, chooses the decision for which a certain upper confidence bound is maximized.

One may note that there is a substantial difference ($\log T$ vs. \sqrt{T}) between the bound obtained by Auer and those of his predecessors. One reason for this difference is a simple reversal of quantifiers. Lai and Robbins and Agrawal studied asymptotic regret in a model where the unknown distribution was fixed and the game was run for an arbitrarily long time. By contrast, in Auer’s work, the time horizon is fixed first, and the distribution of costs may depend on it. In this setting, there are lower bounds showing that $\Omega(\sqrt{T})$ regret can be forced.

Nevertheless, there really is a fundamental difference between the stochastic K -armed bandit problem and the stochastic bandit linear optimization problem. In joint work with Tom Hayes and Sham Kakade [9] we analyze a simpler version of Auer’s algorithm and show that it gets $O^*(n\sqrt{T})$ regret with high probability, in the model where the time horizon is selected first. Moreover, we show that this is optimal up to poly-logarithmic factors, that is, we present a lower bound that shows that we have the correct dependence on n as well. We then show that if the distribution is fixed first, and independently of T then under certain special circumstances, and certainly when the decision set is a finite set or a fixed polytope, we get similar asymptotic behaviour to that of Lai and Robbins and Agrawal, *i.e.*, we get $O(n^2 \log^3 T)$ regret. On the other hand, we show that this is not possible in general: there are decision sets for which $\Omega(n\sqrt{T})$ can be forced, even if the distribution is set independently of T .

I am actively working on a number of new directions within Online Optimization. For instance, I would like to extend our result for online linear optimization for i.i.d. random variables to handle more complicated stochastic models, such as exchangeable random variables, or hidden Markov models. In these settings, because the environment is somewhat more restricted in setting its cost functions, we may hope to prove improved bounds with stronger notions of regret. I would also like to extend our work to handle more complex types of loss functions, such as convex functions.

Machine Learning Reductions

The goal in supervised learning is to use a “training set” of labelled examples to produce a classifier which can (correctly) predict labels of previously unseen (and unlabelled) examples. In *binary classification*, the classifier must distinguish between two kinds of data (*e.g.*, is this a cancerous growth or not? is this mushroom edible or not? is this an image of a face or not?) Other examples of supervised learning problems include

- multiclass classification: correctly distinguish between multiple kinds of example, *e.g.*, which digit is represented by this handwritten character,
- importance-weighted classification: some examples are more important than others,
- cost-sensitive classification: some types of mistake are more costly than others, *e.g.* stopping at a green light is less costly than running a red light, and
- regression: fitting a numerical function to the data.

Considerable effort has been spent on each of these tasks, but especially on binary classification, which is ostensibly the easiest. But is this really true!?

If one can learn the binary classifications, “is this an A or not,” “is this a B or not,” and “is this a C or not,” then these classifiers can be combined to solve the 4-way classification problem “which letter is this, A,B,C or none of the above?”

A *reduction* between supervised learning tasks is a procedure allowing an algorithm for one of the tasks to be converted into a learning algorithm for the other task. Although there were already a number of reductions known in the literature, a coherent theory was lacking. In joint work with Beygelzimer, Hayes, Langford and Zadrozny [6], we introduced a formal notion of error limiting reductions, and constructed such a reduction from cost sensitive classification to binary classification. We also analyzed the error rates of several existing reductions under our new model.

The Wisdom of Crowds

In recent years information markets have attracted attention for their purported ability to predict such future events as election results, sporting events and terrorist activity. Several studies have suggested that such markets may be more accurate than traditional polls. But so far economics and game theory have been able to give no rigorous analysis of the extent to which markets efficiently aggregate information.

In joint work, with Omid Madani, David Pennock and Sumit Sanghai and Brian Galebach [11], using data from an on-line gaming site called ProbabilitySports, we did several experiments to compare the predictive power of several online and offline prediction algorithms, including a simulated information market. Our results suggest that it is hard to do much better than simple averaging algorithms on this data set. (This probably says more about sports fans than it does about learning algorithms and information markets in general.)

This work was done while I was a summer intern at Yahoo! Research Labs.

Fair Allocation

In the generalized cake-cutting problem, a cake is to be divided among n players, according to some notion of fair division. Two such notions are equitable division, where each player gets at least a $1/n$ fraction of the cake according to his or her own valuation (which is an arbitrary measure on the cake), and envy-free division, where no player prefers the piece allocated to someone else to their own piece. This may be different from equitable division since the players may have different valuations.

In the discrete version, instead of a cake, a collection of indivisible goods is to be shared by n players, who may have different valuations for them. Real-life examples include the redistribution of property in a divorce settlement or an inheritance, or negotiating the terms of a treaty or business contract. In such cases, the indivisibility of the goods can make it impossible to achieve perfectly equitable allocations (indeed, when

there is just one good, it can only be assigned to one player.) In joint work with Ivona Bezakova [7], we studied a notion of fairness called max-min fairness, in which the goal is to maximize the value to the player getting the smallest value, and subject to that, maximize the value to the player getting the second smallest value, and so on. Although we showed that the problem of finding an optimal such allocation is NP-hard, we were able to efficiently compute an approximately optimal allocation for the algorithmic problem where all the valuations are known.

References

- [1] R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078, 1995.
- [2] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, 2003.
- [3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.
- [4] B. Awerbuch and R. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the 36th ACM Symposium on Theory of Computing (STOC)*, 2004.
- [5] P. Bartlett, V. Dani, T. P. Hayes, S. M. Kakade, A. Rakhlin, and A. Tewari. High probability regret bounds for online linear optimization (working title). *Manuscript*, 2007.
- [6] Alina Beygelzimer, Varsha Dani, Thomas P. Hayes, John Langford, and Bianca Zadrozny. Error limiting reductions between classification tasks. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 49–56, New York, NY, USA, 2005. ACM.
- [7] Ivona Bezakova and Varsha Dani. Allocating indivisible goods. Technical Report 20, University of Chicago, Department of Computer Science Technical Report TR-2004-10, 2004.
- [8] V. Dani, T. P. Hayes, and S. M. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*. 2008.
- [9] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. *In submission to STOC 2008*, 2008.
- [10] Varsha Dani and Thomas P. Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.
- [11] Varsha Dani, Omid Madani, David M. Pennock, Sumit K. Sanghai, and Brian Galebach. An empirical comparison of algorithms for aggregating expert predictions. In *UAI*. AUAI Press, 2006.
- [12] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4, 1985.
- [13] H.B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, 2004.
- [14] H. Robbins. Some aspects of the sequential design of experiments. In *Bulletin of the American Mathematical Society*, volume 55, 1952.