REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
02-08-2016	Final Report	29-Jun-2012 - 28-Jun-2016

4. TITLE AND SUBTITLE
Final Report: Stochastic Online Learning in Dynamic Networks under Unknown Models

5a. CONTRACT NUMBER
W911NF-12-1-0271

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Qing Zhao

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Davis
Sponsored Programs
1850 Research Park Drive, Suite 300
Davis, CA 95618 -6153

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
62018-NS.24

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This research aims to develop fundamental theories and practical algorithms for distributed, robust, and real-time learning in dynamic tactical networks. The overall objective is to significantly move the frontiers of knowledge in stochastic learning in the classic multi-armed bandit by systematically relaxing traditionally adopted restrictive assumptions.

15. SUBJECT TERMS
Online learning, multi-armed bandit, dynamic networks

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Qing Zhao
UU	UU	UU	UU		19b. TELEPHONE NUMBER 530-752-7390

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

Report Title

Final Report: Stochastic Online Learning in Dynamic Networks under Unknown Models

ABSTRACT

This research aims to develop fundamental theories and practical algorithms for distributed, robust, and real-time learning in dynamic tactical networks. The overall objective is to significantly move the frontiers of knowledge in stochastic learning in the classic multi-armed bandit by systematically relaxing traditionally adopted restrictive assumptions.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
08/30/2013	1.00	Haoyang Liu, Keqin Liu, Qing Zhao. Learning in A Changing World: Restless Multi-Armed Bandit with Unknown Dynamics, IEEE TRANSACTIONS ON Information Theory, (03 2013): 1902. doi:
08/30/2013	7.00	Sattar Vakali, Keqin Liu, Qing Zhao. Deterministic Sequencing of Exploration and Exploitation for Multi-Armed Bandit Problems, IEEE Journal of Selected Topics in Signal Processing, (10 2013): 0. doi:
TOTAL:	**2**	

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received		Paper
TOTAL:		

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper
08/01/2016 19.00	. Mean Variance and Value at Risk in Multi-Armed Bandit Problems, The 53nd Annual Allerton Conference on Communication, Control, and Computing. 01-OCT-15, Monticello, IL. : ,
08/01/2016 20.00	. Oligopoly Dynamic Pricing: A Repeated Game with Incomplete Information, IEEE International Conference on Acoustics, Speech, and Signal Processing. 20-MAR-16, Shanghai, China. : ,
08/02/2016 21.00	. Online Learning and Optimization of Markov Jump Linear Models, IEEE International Conference on Acoustics, Speech, and Signal Processing. 20-MAR-16, Shanghai, China. : ,
08/02/2016 22.00	. Online Learning and Pricing for Demand Response in Smart Distribution Networks, IEEE Statistical Signal Processing Workshop. 26-JUN-16, Spain. : ,
TOTAL:	**4**

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received		Paper
08/01/2016	4.00	Pouya Tehrani, Qing Zhao. Distributed Online Learning of the Shortest Path under Unknown Random Edge Weights, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 26-MAY-13, Vancouver, Canada. : ,
08/01/2016	5.00	Pouya Tehrani, Qing Zhao. Stochastic Online Learning under Unknown Time-Varying Models, IEEE Asilomar Conference on Signals, Systems, and Computers. 04-NOV-12, Pacific Grove, CA. : ,
08/01/2016	12.00	Sattar Vakili, Qing Zhao. Distributed Node-Weighted Connected Dominating Set Problems, The 47th IEEE Asilomar Conference on Signals, Systems, and Computers. 03-NOV-13, Pacific Grove, CA. : ,
08/01/2016	13.00	Sattar Vakili, Qing Zhao. Achieving Complete Learning in Multi-Armed Bandit Problems, The 47th IEEE Asilomar Conference on Signals, Systems, and Computers. 03-NOV-13, Pacific Grove, CA. : ,
08/01/2016	11.00	Pouya Tehrani, Qing Zhao, Tara Javidi. Opportunistic Routing under Unknown Stochastic Models, IEEE Workshops on Computational Advances in Multi-Channel Sensor Array Processing (CAMSAP). 15-DEC-13, Saint Martin, France. : ,
08/01/2016	10.00	Yixuan Zhai, Qing Zhao. Online Learning for Network Optimization under Unknown Models, IEEE Global Conference on Signal and Information Processing (GlobalSIP). 03-DEC-13, Austin, Texas. : ,
08/01/2016	14.00	Sattar Vakili, Qing Zhao. Risk-Averse Online Learning under Mean-Variance Measures, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 19-APR-15, Brisbane, Australia. : ,
08/01/2016	15.00	Yixuan Zhai, Qing Zhao. Competitive Dynamic Pricing under Demand Uncertainty, The 48th IEEE Asilomar Conference on Signals, Systems, and Computers. 02-NOV-14, Pacific Grove, CA. : ,
08/01/2016	16.00	Sattar Vakili, Qing Zhao, Yuan Zhou. Time-Varying Stochastic Multi-Armed Bandit, The 48th IEEE Asilomar Conference on Signals, Systems, and Computers. 02-NOV-14, Pacific Grove, CA. : ,
TOTAL:	**9**	

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
08/30/2013	3.00	Keqin Liu, Sattar Vakili, Qing Zhao. Stochastic Online Learning for Network Optimization under Random Unknown Weights, IEEE TRANSACTIONS ON Signal Processing (05 2013)
TOTAL:		**1**

Number of Manuscripts:

Books

Received		Book
TOTAL:		

Received		Book Chapter
TOTAL:		

Patents Submitted

Patents Awarded

Awards

In 2013, the PI was selected for elevation to IEEE Fellow for "contributions to learning and decision theory in dynamic systems and cognitive networking."

In 2014, the PI received the Outstanding Mid-Career Faculty Research Award from the UC Davis College of Engineering.

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Sattar Vakili	0.80	
Pouya Tehrani	0.15	
Yixuan Zhai	0.25	
Yuan Zhou	0.10	
FTE Equivalent:	**1.30**	
Total Number:	**4**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Qing Zhao	0.20	
FTE Equivalent:	**0.20**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME
Dianqi Han
Total Number: 1

Names of personnel receiving PHDs

NAME
Pouya Tehrani
Yixuan Zhai
Total Number: 2

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Please see Attachment.

Technology Transfer

During this ARO project, the PI completed a two-year 6.2 project with CERDEC and ARL on ``Distributed Learning for Quality of Information'' which ended on June 30,
2014. This two-year project focused on extending the learning techniques based on multi-armed bandit theory developed under the ARO project to support distributed battlefield environments. A major objective of this 6.2 project was to transition the distributed learning algorithms developed under the ARO project to the CERDEC-STCD CNEDAT (Cognitive Network Design and Analysis Toolset) currently under development. Based on CERDEC's interest in online routing and establishing virtual network backbone via connected dominating sets (CDS) in unknown communication environments, we extended learning techniques developed under the ARO project to this more complex problem. We developed a discrete event simulation platform for the developed distributed learning algorithms for connected dominating set and shortest path routing. The working source code and executables were transitioned to CERDEC-STCD and ARL.

Scientific Progress

This research aims to develop fundamental theories and practical algorithms for distributed, robust, and real-time learning in dynamic tactical networks. The overall objective is to significantly move the frontiers of knowledge on stochastic learning beyond the limitations and boundaries of the classic formulations. The online learning theories and algorithms under development will have broad military applications, include cognitive and opportunistic networking, dynamic resource allocation in unknown environments and under unpredictable demands, and optimal activation and decision making under unknown models and incomplete observations.

The technical approach rests on a stochastic online learning framework based on the multi-armed bandit (MAB) theory which is a fundamental mathematical tool for optimal sequential decision making and learning in uncertain environments. However, three assumptions in the classic MAB formulation significantly limit its applications: that different actions yields independent outcomes, that the observations are independent and identically distributed over time, and that there is only a single decision maker. Furthermore, even under these restrictive assumptions, existing learning algorithms can only handle cases where the random observations (i.e., the reward or outcomes of a given action) obey a bounded-support distribution or a few light-tailed distributions. Integrating theories and techniques in large deviations principle, distributed statistical inference, decision theory, and stochastic control, we systematically address these limitations and application barriers of the classic MAB theory and develop theories and algorithms capable of

- learning under heavy-tailed reward distributions;
- learning from dependencies across actions;
- learning under restless and time-varying network models;
- learning in a decentralized setting with multiple distributed decision makers;
- learning from side information;
- learning under risk constraints.

Below we summarize the scientific progress and major accomplishments of this project.

Learning under Heavy-Tailed Reward Distributions:

The existing online learning algorithms developed under the classic MAB formulation can only handle finite-support reward distributions and several specific infinite-support light-tailed distributions such as Gaussian, exponential, and Poisson (when the distribution type is known). This significantly limits the applications of these learning algorithms. For instance, in the application of route selection, the cost (i.e., negative reward) of each route may reflect the latency which can have a general, potentially heavy-tailed distribution.

In [1], we developed a general approach based on a Deterministic Sequencing of Exploration and Exploitation (DSEE) for constructing MAB learning algorithms. We

showed that DSEE achieves the optimal logarithmic order for both light-tailed and heavy-tailed distributions without knowing the specific distribution type (other than whether it is light-tailed or heavy-tailed). We have thus successfully provided a complete solution to the problem of learning under general reward distributions.

More importantly, different from the classic online learning policies proposed for MAB, the DSEE approach has a clearly defined tunable parameter---the cardinality of the exploration sequence---which can be adjusted according to the ``hardness'' (in terms of learning) of the reward distributions and observation models. It is thus more easily extendable to handle a wide range of variations of MAB.

Learning from Action Dependencies for Network Optimization:

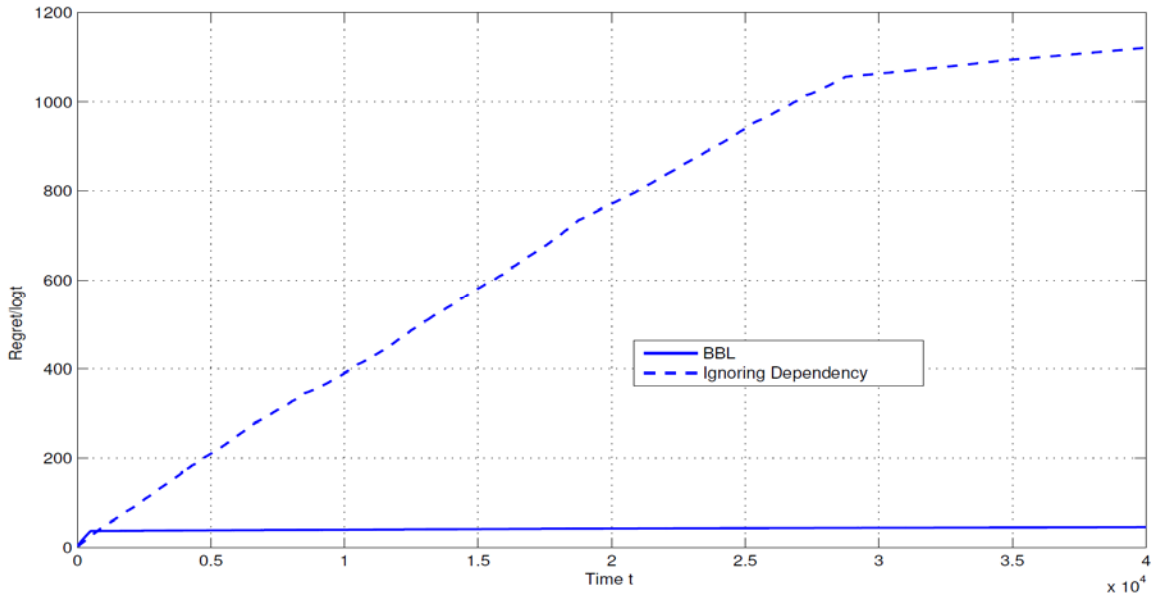
In the classical MAB formulation, arms (i.e., different actions) are independent. Reward observations from one arm do not provide information about the quality of other arms. As a result, regret grows linearly with the number of arms and logarithmically with time. However, many network optimization problems (such as optimal activation for online detection and shortest-path adaptive routing and minimum spanning under unknown and time-varying edge weights) lead to MAB with a large number of arms (e.g., the number of paths) dependent through a small number of unknowns (e.g., the number of edges). While the dependency across arms can be ignored in learning and existing learning algorithms directly apply, such a naive approach yields poor performance in terms of regret, storage, and computation.

In [2], we investigated the problem of online learning for network optimization under unknown random weights. Such network optimization problems include finding the shortest path, the minimum spanning tree, and the minimum dominating set as examples. For such combinatorial network optimization problems, the number of choices often grows exponentially with the number of unknowns (e.g., the number of paths can grow exponentially with the number of edges). A direct application of existing MAB learning algorithms will yield a regret that grows exponentially with the network size. In [2], we showed that by exploiting arm dependencies, we can achieve a regret that grows polynomially (rather than exponentially) with the network size while maintaining its optimal logarithmic order with time. Our approach is based on dimension reduction of the learning space by constructing a set of properly chosen basis functions of the underlying network. Simulation results demonstrate that our basis-based learning algorithm achieves orders of magnitude improvement in both the steady state performance and the learning convergence rate.

In [3,4], we further extended the proposed basis-based learning approach to distributed routing where there is no central control unit and each node only knows its neighbors but not the entire network topology. Without the knowledge of the network topology, it is no longer possible to construct basis of the network. The key is to develop online learning strategies at each individual node. Specifically, through local information exchange with its neighbors, each node decides which neighbor to route the current packet to, aiming at minimizing the expected total cost of the resulting path by learning from past

observations of its outgoing edges. We proposed a distributed learning algorithm that preserves the logarithmic regret order with time and the polynomial regret order with the network size as in the centralized case.

In [5], we tackled distributed learning of *Minimum Connected Dominating Set* (MCDS) in ad hoc networks. Applications of the MCDS problem includes finding a virtual backbone (also referred to as a spine) in wireless ad-hoc networks. Such a virtual backbone is important in routing, broadcasting and connectivity management (topological control) in wireless ad-hoc networks for efficient resource utilization. We proposed a fully distributed learning algorithm for NW-MCDS. The proposed distributed algorithm achieves the same optimal logarithmic regret order with time as the centralized algorithm. It also offers the optimal linear regret order with the network size. Shown in the figure below is the normalized regret over time for learning the minimum connected dominating set for constructing a virtual backbone in an ad hoc network. We observe from the figure that comparing with the approach that ignores dependencies among CDS's, our learning algorithm BBL offers orders of magnitude improvement in both the steady-state performance and the convergence rate to the logarithmic regret order.



Learning in a Changing World:

The classical MAB formulation assumes an i.i.d. (identically and independently distributed) or a rested Markov reward model which applies only to systems without memory or systems whose dynamics only result from the player's action. In many practical applications such as cognitive radio networks, scheduling in queueing and communication systems, and target tracking, the system has memory and continues to evolve even when it is not engaged by the player. For example, channels continue to evolve even when they are not sensed; queues continue to grow due to new arrivals even when they are not served; targets continue to move even when they are not monitored. For such applications, the classic policies no longer apply.

In [6], we developed adaptive learning algorithms capable of handling a restless reward model where the reward state of each arm continues to evolve (according to unknown stochastic processes) no matter the arm is activated or not. We constructed an online learning algorithm with an interleaving exploration and exploitation epoch structure that achieves a regret with logarithmic order. We further extended the problem to a decentralized setting where multiple distributed players share the arms without information exchange. Under both an exogenous restless model and an endogenous restless model, we showed that a decentralized extension of the proposed learning algorithm preserves the logarithmic regret order as in the centralized setting. In [7,8,9], we studied the problem under unknown Markov jump affine models. We developed an online learning algorithm based on simultaneous perturbations stochastic approximation that achieves the best regret order. We also showed that this learning algorithm converges to the optimal control input almost surely as well as in the mean square sense.

In [10,11], we considered multi-armed bandit with time varying reward distributions, which, has received little attention in the literature. Our objective was to characterize the impact of model variations on learning and address the fundamental question of when it is possible to track unknown model variations through learning. In particular, we established in [10] sufficient conditions on the rate of model variations under which learning can or cannot improve the regret order. In [11], we considered a general time-varying MAB model in which the unknown reward distribution of each arm can change arbitrarily at any time. We established that the regret of the stochastic MAB with arbitrary reward model variations is lower bounded by the order of the squared root of T. We further showed that this lower bound is achieved by an online learning algorithm, thus demonstrating the tightness of the lower bound and the order optimality of the algorithm. We also considered a time variation model where the number of changes experienced by the reward distributions is constrained by a constant that is a function of the horizon length T. Referred to as the piece-wise stationary model, this time variation model allows reward distributions to change at arbitrary time instants, but the total number of change points is no more than $H(T)-1$. In other words, the reward distribution sequence consists of up to $H(T)$ stationary segments. Under this time variation model, we showed that a regret order of the squared root of $TH(T)\log T$ is achievable, which is sublinear in T provided that $H(T)$ is of a lower order than $T\log T$.

Distributed Learning with Multiple Decision Makers and Incomplete Observations:

The classical MAB formulation considers only a single player, which, in a network setting, can only handle centralized configurations where all players act collectively as a single entity by exchanging observations and making decisions jointly. In many applications, however, information exchange among players and joint decision making can be costly or even infeasible. We often face a decentralized problem in which multiple distributed players learn from their local observations and make decisions independently. While other players' observations and actions are unobservable, players' actions affect each other: conflicts occur when multiple players choose the same arm at the same time and conflicting players can only share the reward offered by the arm, not necessarily with

conservation (for instance, when multiple users access the same channel at the same time, no one may succeed).

In [1], we tackled a decentralized MAB problem in which multiple distributed players learn from their local observations and make decisions independently. We further assumed that reward observations are incomplete, in particular, collisions among players are unobservable. In other words, a player does not know whether it is involved in a collision, or equivalently, whether the received reward reflects the true state of the arm. Collisions thus not only result in immediate reward loss, but also corrupt the observations that a player relies on for learning the arm rank. Such decentralized learning problems arise in communication networks where multiple distributed users share the access to a common set of channels, each with unknown communication quality. If multiple users access the same channel at the same time, no one transmits successfully or only one captures the channel through certain signaling schemes such as carrier sensing. Another application is multi-agent systems in which M agents search or collect targets in N locations. When multiple agents choose the same location, they share the reward in an unknown way that may depend on which player comes first or the number of colliding agents.

The deterministic separation of exploration and exploitation in DSEE, however, can ensure that collisions are contained within the exploitation sequence. Learning in the exploration sequence is thus carried out using only reliable observations. In particular, we showed that under the DSEE approach, the system regret, defined as the total reward loss with respect to the ideal scenario of known reward models and centralized scheduling among players, grows at the same orders as the regret in the single-player MAB under the same conditions on the reward distributions. These results hinge on the extendability of DSEE to targeting at arms with arbitrary ranks (not necessarily the best arm) and the sufficiency in learning the arm rank solely through the observations from the exploration sequence.

In [12,13], we also studied distribution learning in a game theoretic framework for the application of multi-seller dynamic pricing with unknown demand models. We formulated the problem as an infinitely repeated game with incomplete information and developed a dynamic pricing strategy referred to as Competitive and Cooperative Demand Learning (CCDL). The strategy was shown to be a subgame perfect Nash equilibrium and Pareto efficient. We further showed that the proposed competitive pricing strategy achieves a bounded regret, where regret is defined as the total expected loss in profit with respect to the ideal scenario of a known demand model. This demonstrates the learning and cooperation efficiency of the proposed strategy.

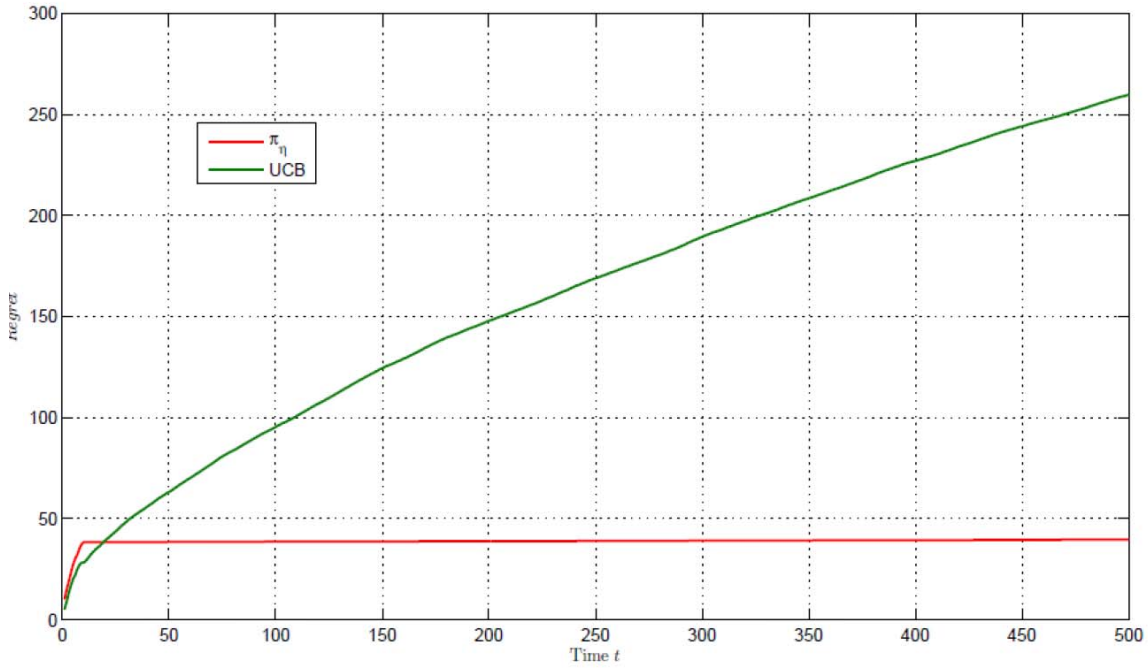
Learning from Side Information:

Under the classic multi-armed bandit formulation, a fundamental limit of any online learning algorithm is that regret at least grows logarithmically with time (as shown in Lai and Robbins's seminar work in 1985). This implies that the cost of learning must grow unboundedly with time, and for any online learning algorithm, mistakes in selecting a

suboptimal arm occur infinitely often, even though with reduced frequency. In other words, the learning algorithm will never converge to the best arm, but rather needs to continue learning by selecting suboptimal arms to refine its learning results.
One fundamental question we have pursued in the second year of this project is whether any side information on the reward model can lead to bounded regret, and if yes, what the minimum side information is for achieving bounded regret. Note that if a learning algorithm achieves a bounded regret over an infinite time horizon, the total expected number of mistakes in selecting a suboptimal arm by this learning algorithm is bounded. In other words, the algorithm will converge to the optimal arm with probability 1. This case is thus referred to as complete learning.

In [14], we were able to provide a positive answer to the above question: we showed that if a single value between the expected reward of the best arm and the expected reward of the second best arm is known, complete learning can be achieved. Furthermore, we constructed a simple practical learning algorithm that achieves a bounded regret under this minimum side information.

In the figure below, we compare the regret (as a function of time) of the optimal learning algorithm that does not use any side information (the green curve) and the regret of our algorithm (the red curve) that exploits the minimal side information. We can see that for the former algorithm, regret grows unboundedly over time, while our algorithm quickly converges to the best arm as indicated by the regret remaining constant after a very short period of time. Orders of magnitude of improvement in learning efficiency is thus achieved with minimal side information.



Learning under Risk Constraints:

In the classic MAB formulation, the performance measure is on the *expected* return of an online learning algorithm. There is no control over risk or deviation from the average performance. In many applications, however, the decision maker may be more interested in reducing the uncertainty (i.e., risk) in the outcome, rather than achieving the highest ensemble average. In [15,16,17], we addressed the issue of risk management in online learning and formulated and studied *risk-averse* MAB problems under commonly adopted risk measures: mean variance and value at risk. Introduced by Nobel Prize laureate Markowits in 1952, mean-variance of a random variable is given by a linear combination of the mean and the variance of the random variable, where the weight coefficient in the linear combination can be understood as the Lagrangian multiplier in a constrained optimization problem of maximizing the expected return subject to a risk constraint on the variance or minimizing the risk subject to a constraint on the minimum expected return. The second risk measure---value at risk---controls the probability of unacceptable deviation from the mean. Equivalently, the objective is to maximize a high-probability minimum return.

Risk-averse MAB has received little attention in the literature. There are a couple of scattered studies, mostly heuristic and empirical. In particular, the fundamental limit on learning efficiency under risk measures is unknown. In [15,16,17], we obtained the first set of analytical results on risk-averse learning, in which we established the fundamental performance limit and developed risk-averse learning algorithms that achieve the performance limit.

References:

[1]. S. Vakili, K. Liu, and Q. Zhao, "Deterministic Sequencing of Exploration and Exploitation for Multi-Armed Bandit Problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759 – 767, October, 2013.

[2]. Y. Zhai, Q. Zhao, "Online Learning for Network Optimization under Unknown Models," in *Proc. of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, December, 2013.

[3]. P. Tehrani, Q. Zhao, "Distributed Online Learning of the Shortest Path under Unknown Random Edge Weights," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May, 2013.

[4.] P. Tehrani, Q. Zhao, T. Javidi, "Opportunistic Routing under Unknown Stochastic Models," in *Proc. of IEEE Workshops on Computational Advances in Multi-Channel Sensor Array Processing (CAMSAP)*, December, 2013.

[5.] S. Vakili, Q. Zhao, "Distributed Node-Weighted Connected Dominating Set Problems," in *Proc. of the 47th IEEE Asilomar Conference on Signals, Systems, and Computers*, November, 2013.

[6]. H. Liu, K. Liu, and Q. Zhao, "Learning in A Changing World: Restless Multi-Armed Bandit with Unknown Dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902-1916, March 2013.

[7.] S. Baltaoglu, L. Tong, Q. Zhao, "Online Learning and Optimization of Markov Jump Linear Models," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March, 2016.

[8.] S. Baltaoglu, L. Tong, Q. Zhao, "Online Learning and Pricing for Demand Response in Smart Distribution Networks," in *Proc. of IEEE Statistical Signal Processing Workshop (SSP)*, June, 2016.

[9.] S. Baltaoglu, L. Tong, Q. Zhao, "Online Learning and Optimization of Markov Jump Affine Models," submitted to *IEEE Transactions on Information Theory.*

[10]. P. Tehrani and Q. Zhao, "Stochastic Online Learning under Unknown Time-Varying Models," in *Proc. of the 46th IEEE Asilomar Conference on Signals, Systems, and Computers*, November, 2012.

[11.] S. Vakili, Q. Zhao, Y. Zhou, "Time-Varying Stochastic Multi-Armed Bandit," in *Proc. of the 48th IEEE Asilomar Conference on Signals, Systems, and Computers*, November, 2014.

[12.] Y. Zhai, Q. Zhao, "Competitive Dynamic Pricing under Demand Uncertainty," in *Proc. of the 48th IEEE Asilomar Conference on Signals, Systems, and Computers*, November, 2014.

[13.] Y. Zhai, Q. Zhao, "Oligopoly Dynamic Pricing: A Repeated Game with Incomplete Information," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March, 2016.

[14.] S. Vakili, Q. Zhao, "Achieving Complete Learning in Multi-Armed Bandit Problems," in *Proc. of the 47th IEEE Asilomar Conference on Signals, Systems, and Computers*, November, 2013.

[15.] S. Vakili and Q. Zhao, "Risk-Averse Online Learning under Mean-Variance Measures," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April, 2015.

[16.] Vakili, Q. Zhao, "Mean Variance and Value at Risk in Multi-Armed Bandit Problems," in *Proc. of the 53nd Annual Allerton Conference on Communication, Control, and Computing*, October, 2015.

[17.] S. Vakili, Q. Zhao, "Risk-Averse Multi-Armed Bandit Problems under Mean-Variance Measure," to appear in *IEEE Journal of Selected Topics in Signal Processing*.