A MULTI-ARMED BANDIT WITH SWITCHING COSTS FOR AUTONOMOUS ACOUSTIC RELAY POSITIONING

Mei Yi Cheung, Joshua Leighton and Franz S. Hover
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge MA 02139
{mc2922, jleight, hover}@mit.edu

***Abstract*—We address the goal of maximizing underwater acoustic data transmission in an unknown environment by adaptively positioning an autonomous mobile relay so as to learn and exploit spatial variations in channel performance. This is a classic exploration vs. exploitation problem, which can be elegantly formulated as a multi-armed bandit (MAB) with an optimal solution in the form of Gittins indices. However, for an autonomous ocean vehicle traveling between distant waypoints, switching costs in the form of travel times are significant. In the presence of these switching costs, the MAB has no optimal index policy. Thus, we adapt the Gittins index decision rule with limited-lookahead policy enumeration and use proven properties of the optimal policy to reduce computation costs. We describe shallow-water field experiments conducted in the Charles River (Boston, MA) with autonomous surface vehicles and acoustic modems, and show that the switching-costs-aware algorithm offers superior real-time performance in decision-making and efficient learning of the unknown field.**

I. INTRODUCTION

Acoustic underwater communications is critical to many ocean applications, such as underwater sensor networks and point-to-point communications with autonomous underwater vehicles. "Acomms" has seen great progress in recent years with the development of modern channel estimation and coding schemes (e.g. [1], [2]), however, the performance of the acoustic channel remains strongly dependent on local environmental properties – i.e. temperature, salinity and bathymetry [3]. In shallow-water environments, surface and bottom conditions are sources of complex multi-path interference. Similarly, in harbors and other man-made environments, multi-path from structures and ambient noise present challenges to successful packet decoding. Ray- and beam-tracing algorithms can be used to predict performance, but they may be computationally expensive even in a two-dimensional setting, and typically require measuring or modeling the water properties [4], [5]. Detailed measuring and modeling of these properties may not be practical or desirable in complex, shallow-water environments, or short missions in unknown surroundings. These factors result in the notorious unreliability and complex spatial variation of acoustic channel performance. With improved acomms performance, autonomous underwater vehicles and teams of them could play an enhanced role in the ocean community.

Portions of this work appeared in IEEE International Conference on Robotics and Automation, May 2013

To this end, we formulate a procedure for autonomous, adaptive placement of a mobile modem that functions as an acoustic relay between fixed source and destination nodes, with the overall goal of improving cumulative data transmitted from the source to the destination. A mobile relay can learn about the performance of a physical acoustic link as it transmits on location. Positioning a mobile acoustic relay is a classic exploration vs. exploitation scenario, since the relay cannot completely sacrifice throughput as it searches for the best location. We seek an algorithm by which the relay explores the physical space in real-time to find locations with good acomms performance without compromising the mission-level goal of cumulative data transmission and given only statistical information collected on location. Despite the apparent simplicity of this acoustic relaying problem, adaptive placement of acoustic modems has not been systematically studied until recently [6]. Following the discussion of Akyildiz *et al.* [7], Detweiler *et al.* [8], observed extreme variability and no clear pattern of packet success rate, in 10m-deep water with custom equipment. Data from our field experiments in the Charles River (Boston, MA), using Woods Hole Oceanographic Institution (WHOI) Micro-Modems in similar water depths, show the same properties as well as pronounced day-to-day variations. Thus, we assume the relay has no prior knowledge about the dependence of the acoustic channel's performance on its location, except the usual spreading law suggesting the relay should be situated between the source and destination nodes.

The stochastic multi-armed bandit problem (MAB) is a formulation of this type of resource allocation problem with a rich history in optimization. Scheduling problems naturally lend themselves to a bandit-based optimal solution and the MAB formulation has been used in many applications, such as beam scheduling for array tracking systems [9], radio channel allocation [10], and website ranking [11]. Its canonical formulation admits an elegant and precise solution in the form of Gittins indices [12], [13], which define a decision rule by which the relay can choose an optimal location at every decision epoch from a set of points in discretized space. The MAB algorithm is a practical solution which does not require prior modeling or heuristic tuning. More broadly, the MAB formulation we present may also be applied to any poorly known environment with position-dependent fading and interference, such as ad-hoc wireless networks [14], [15].

In practice, the performance of the acoustic link while transiting is often adversely affected by propulsion and self-noise [16]. For a slow-moving ocean vehicle traveling between distant waypoints, the time taken to change location is significant compared to transmission times. These considerations motivate augmenting the MAB to account for switching costs (MABSC). However, no optimal index policy solution exists for the MABSC [17], and most research on this topic has focused on deriving general properties of the optimal policy [18], deriving explicit optimal policies for special cases [19], and bounding approximations to the optimal policy [20]. For a recent survey, see Jun [21]. The problem has also been reformulated as a semi-Markov multi-armed restless bandit, addressed by marginal productivity indices (MPI) [22] and a linear programming relaxation (LP) [23], based on work by Bertsimas and Niño-Mora [24]. These include switching costs as a natural extension of the restless bandit [25], in which processes are non-stationary. However, both the MPI[1] and LP treatments of the restless bandit trade the advantage of an exact and general problem statement for a lookahead horizon limited to one switch/step, while the alternative, enumeration of decision paths, incurs an exponential computational cost. Here, we exploit a small state space that allows for a much deeper enumeration, but also seek methods by which this load can be reduced. In particular, applying the key result of Asawa and Teneketzis [18] allows us to adapt the Gittins index policy while greatly reducing the computation cost of decision-making required.

In Section II we formulate the MABSC, and its application to the problem of adaptive acoustic relay positioning. In Section III, we describe field experiments conducted with mobile surface vehicles towing underwater acoustic modems under fully autonomous control, with the MAB and our MABSC adaptation. In Section IV, we demonstrate the attractiveness of the MABSC algorithm using experimental data. For an extensive comparison of the MAB and MABSC with competing algorithms, see [26].

II. Problem Formulation

The stochastic multi-armed bandit problem considers dynamic allocation of a single resource amongst several competing processes, in order to maximize total expected reward. We formulate the stationary multi-armed bandit in generality and then develop the switching cost heuristic.

A. The Canonical MAB Problem

The one-armed bandit is defined as a sequence of process states $x(1), \cdots, x(n)$, where $x(n)$ is a random variable representing the state of the machine after it has been operated n times. In general, the reward $R(x(n))$ from each state is a real, non-negative random variable. The multi-armed bandit process is a collection of N independent one-armed bandit machines, here indexed by i. n_i denotes the number of times machine i has been operated, and its state is $x_i(t)$, where t is the current global decision epoch:

$$t = \sum_{i=1}^{N} n_i. \tag{1}$$

We denote the state of the multi-arm process as a whole by the vector $\bar{x}(t)$, which is $\{x_1(t) \cdots x_N(t)\}$. At each decision epoch, the process samples a single machine, updating the state and reaping the associated reward, while the states of all other machines remain frozen.

The classical MAB problem has as its optimal solution a dynamic allocation policy π that defines at each decision epoch the machine for allocation i_t, such that the expected value of the total reward V_π is maximized. For the discount factor $0 < \beta < 1$ and an infinite horizon, this reward is:

$$V_\pi(\bar{x}) = E\bigg[\sum_{k=0}^{\infty} \beta^k R(x_{i_k}(k)) \mid \bar{x}(0) = \bar{x} \bigg]. \tag{2}$$

Gittins and Jones [12] showed that the optimal policy is to play the machine with the largest expected discounted reward rate, or the Gittins index, maximized over all stopping times $\tau > 1$:
$i_{t+1} = \underset{i}{\operatorname{argmax}}(\nu_i(x_i(t)))$, where[2]

$$\nu_i(x_i(t)) = \max_{\tau>1} \frac{E\bigg[\sum_{k=0}^{\tau-1} \beta^k R(x_i(k)) \mid x_i(0) = x_i(t) \bigg]}{E\bigg[\sum_{k=0}^{\tau-1} \beta^k \mid x_i(0) = x_i(t) \bigg]}. \tag{3}$$

Crucially, the index ν_i is a function only of $x_i(t)$, allowing the MAB to be decomposed into N independent stopping time problems. Efficient algorithms to calculate the Gittins index have been reported [27], Niño-Mora [28].

B. Enumeration as an Adaptation for Switching Costs

We define constant switching costs $c(i, j)$ to reflect the undesirability of switching from machine i to machine j; in the context of relay positioning, the cost is that of time spent in transit. If $t_v(i, j)$ is the time taken to travel from i to j, and $t_r(i, j)$ is the time taken to relay, we can set $c(i, j) = \lfloor t_v(i, j)/t_r(i, j) \rfloor$ — the number of transmissions the relay could have made on location if it had chosen to sample instead of traveling. This is only one of many applicable cost models. The optimal solution to the MABSC is one that maximizes:

$$V_\pi(\bar{x}) = E\bigg\{ \sum_{k=0}^{\infty} \beta^t \big[R(x_{i_k}(k)) - c(i_k, i_{k-1})\big] | \bar{x}(0) = \bar{x} \bigg\} \tag{4}$$

where we define $i_{-1} = i_0$. As noted previously, the MABSC does not admit an optimal index policy because the reward returned by a process no longer depends solely on the number of times n_i an arm has been operated and the problem cannot be decomposed [17]. For this problem, we describe a solution of the *priority-index policy* form, where separate "continuation" and "decision" indices are used [22]. This scheme separates the decision process into two modes. At every decision epoch, the continuation index is computed and used to decide if the current arm is continued. If it is time to switch, the decision index is computed to decide which arm to switch to. The continuation index ν_i is taken to be the Gittins index. If the

[1]For a stationary process with switching costs, the MPI is equivalent to Asawa & Teneketzis's switching index [22]

[2]This standard notation directly shows the form of expected discounted reward over discounted time, although in our formulation we assume β to be constant and independent of state.

current arm has the highest Gittins index of the field, it can be continued without further decision. However, even if it is not the current maximum, Asawa and Teneketzis (A&T Thm. 2.1) showed that it is optimal to continue playing an arm up to its stopping time τ, only making a decision to switch when the stopping time is achieved: when the Gittins index of the current arm falls below any value it has previously reached [18]. This defines the continuation rule:

$$\text{if } \min_{k<t} \nu_{i_k}(x_{i_k}(k)) \leq \nu_i(x_i(t)), \text{ set } i_{t+1} = i_t. \tag{5}$$

When the stopping time is achieved, i.e., the above condition does not hold, the decision index determines which arm to switch to. The continuation rule can only increase the number of times an arm is played, and reduces the required computation frequency of the decision index, thus admitting an accurate and flexible solution for a problem of this scope. We calculate the decision index by maximizing an m-horizon look-ahead enumeration of the expected reward rate over all possible policies π, where π is any possible sequence of plays $i_1, ..., i_m \, \forall i \in 1, ..., N$. We do not enumerate the action of remaining in the current location, although future actions can include choosing to return to the current location after switching away. The value of being in the final state $\hat{x}_i$ is accounted for with an estimated Gittins index ν_i for that policy. Location-based switching costs are simple to include in this formulation:

$$\eta_\pi(\bar{x}(t)) = \frac{e(\bar{x}(t))}{E\Big[\sum_{k=0}^{m} \beta^k \mid \bar{x}(0) = \bar{x}(t)\Big]} + \nu_i(\hat{x}_{i_{t+m}}(t+m)), \tag{6}$$

where

$$e(\bar{x}(t)) = E\Big\{ \sum_{k=0}^{m-1} \beta^k \big[R(x_{i_k}(k)) - c(x_{i_k}(k), x(k+1))\big] \mid x_{i_k}(0) = x_{i_t}(t) \Big\}. \tag{7}$$

The adapted decision rule for MABSC is then

$$i_{t+1} = \underset{i_1}{\operatorname{argmax}}(\eta_\pi(\bar{x}(t))). \tag{8}$$

If $m = 0$, this rule is identical to the MAB Gittins index rule. If $m = 1$, this rule is identical to the switching index defined by Asawa. Since enumeration is computation-intensive, we apply A&T Thm. 2.1 to reduce the number of required decision index computations. Thus, longer horizons can be enumerated, allowing the algorithm to capture the benefits of efficient routing where a more myopic policy would not.

III. Field Experiments and Datasets

We describe autonomous field experiments conducted with the MAB and MABSC decision policies[3]; Section IV assesses and interprets the results more fully.

[3] These experimental datasets are provided online for download at *http://web.mit.edu/hovergroup/resources.html*

A. Bernoulli Gittins Index Formulation

For learning and decision-making by the mobile relay within the MAB and MABSC framework, we discretize the physical space into N potential relay locations and define each location as an independent arm of the bandit. The agent plays an arm by relaying through that location, updating its state information on the arm, and then deciding which location to play next. Each relayed transmission on location is naturally described by a Bernoulli trial defined as:

$$X_i = \begin{cases} 1 & \text{if transmission success;} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

A computational method for calculating indices for a Bernoulli reward process is described in Gittins [29]. The state vector in this case comprises simply n_i, the number of plays on this location, and s_i, the number of successes at this location: the index is thus $\nu_i(n_i, s_i)$. We use the Bernoulli Gittins index as the continuation index for the MABSC since the Bernoulli reward process naturally describes the series of individual transmissions made on location by the relay.

B. Normal Gittins Index Formulation

For the MAB autonomous experiment, we heuristically account for switching costs by designating five transmissions as one observation, so that the time spent on location is at least as long as the shortest transit time away. We define the reward $\bar{\theta}_i$ of each machine as the estimated mean of the Bernoulli random variable $X_i(t)$ for those five transmissions. The estimate of the mean and variance of $X_i(t)$ can be recomputed with each new sample as a function of n_i, following [30]:

$$\bar{\theta}_i(t) = \begin{cases} \frac{n_i - 1}{n_i}\bar{\theta}_i(t-1) + \frac{1}{n_i}X_i(t) & \text{if } i \text{ was played} \\ \bar{\theta}_i(t-1) & \text{otherwise.} \end{cases} \tag{10}$$

$$\hat{\sigma}_i^2(t) = \begin{cases} \frac{n_i - 2}{n_i - 1}\hat{\sigma}_i(t-1) + \frac{1}{n_i}(X_i(t) - \bar{\theta}_i(t-1))^2 & \text{if } i \text{ was played and } n_i \geq 2 \\ \hat{\sigma}_i^2(t-1) & \text{if } i \text{ was not played.} \end{cases} \tag{11}$$

Thus, $\bar{\theta}_i$ is the best estimate of the probability of packet success at location i — a practical measure of the acoustic channel's performance that can be updated with incoming samples. Its standard deviation $\bar{\sigma}_i$ is given by:

$$\bar{\sigma}_i^2 = \frac{1}{n_i}\hat{\sigma}_i^2, \tag{12}$$

which determines the weighted benefits of exploration as defined by the index calculation. By the Central Limit Theorem, assuming the Bernoulli trials are independent, the reward distribution approximates a normal distribution as the number of observations gets large. Gittins [12] showed that the index for this reward process is a function of the mean (expected reward) and its standard deviation (uncertainty of estimate). This is expressed as:

$$\nu(\bar{\theta}_i, n, \bar{\sigma}_i) = \bar{\theta}_i + \bar{\sigma}_i \nu(0, n, 1). \tag{13}$$

$\nu(0,n,1)$ are Gittins indices for the normal distribution with mean 0, variance 1, and have been previously tabulated [29]. Indices are stored in a lookup table and accessed in real time by the relay, which updates as new information becomes available.

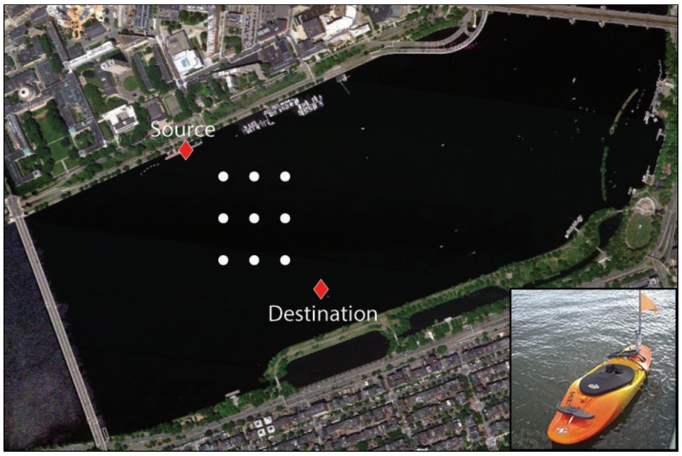

Fig. 1. Charles River Basin (Boston, MA) with Autonomous Surface Vehicle *Nostromo* inset. Relay locations are shown in white. Source and destination locations are shown in red.

We consider a one-way, two-link acoustic transmission in the Charles River Basin. A source modem at the MIT Sailing Pavilion broadcasts a data message, which is repeated by an acoustic modem towed by a robotic surface vehicle. The destination node is a second robotic vehicle station-keeping 580m across the river from the source. A transmission is considered successful if both hops succeed, i.e., the relay decodes the source packet, and the destination decodes the relay packet. Source transmissions may reach the destination directly; this through-transmission success rate reflects the performance of the acoustic channel with no relay. A two-hop transmission takes a total of fifteen seconds in our experiments. In setting up the adaptive positioning experiments, we assumed no prior knowledge of the acoustic channel beyond the usual spreading law. Nine candidate relay locations were chosen in a grid pattern centered on the line between the source and destination nodes (Fig. 1). In practice, such a choice would be influenced by mission constraints. For all experiments, Site 1 was the starting location.

All field experiments are conducted with custom autonomous surface vehicles (Fig. 1) towing acoustic modem transducers at a fixed depth to simulate underwater communications, with the benefits of GPS and WiFi connectivity for controlled experiments. The vehicles travel at 1.5m/s and maintain a station-keeping circle ten meters in diameter on location. We use WHOI Micro-Modems [31], an established and commercially available technology for underwater acoustic data transmission, and report SNR values from before the equalizer on the receiving modem ("SNR-In"). We note that although programmable modem parameters such as packet encoding scheme can be included combinatorially as additional machines, we have fixed these for simplicity.

C. Autonomous MAB

For the MAB algorithm, a sample was designated as five transmissions so that the time required to sample once was comparable to the time of transit to another location. All data transmissions were sent at phase-shift keying (PSK) Rate 2, with a fixed message size of 192 bytes. Estimates of the packet success mean and its variance were initialized with a touring survey consisting of two samples at each location. This initialization was required as at least two observations are required to calculate a Normal Gittins index. Subsequently, the relay executed the algorithm, selecting the location with the highest index and breaking ties by favoring shorter travel distances. Two experimental trials were conducted on the same day. Acknowledgements of transmission receipts were communicated from the destination to the relay over Wi-Fi for simplicity.

D. Autonomous MABSC

We employed the Bernoulli reward scheme with each sample consisting of a single two-hop transmission sent at PSK Rate 1. The MABSC algorithm was initialized with the assumption of 100% packet success probability at each site[4]. The look-ahead horizon for policy enumeration was constrained to a maximum computation time of fifteen seconds. For a look-ahead horizon of five, the average computation time is on the order of one second[5]. For an underwater vehicle, learning the result of the relayed transmission (success or failure) from the destination robot must be done with acoustic acknowledgments. To simulate this with our surface vehicles, we utilized the Micro-Modem frequency-shift-keying (FSK) Mini-packet, a 13-bit message with robust performance. Our experiments have consistently shown packet loss rates of less than 5% for the FSK Mini-packet, thus it is substantially more reliable than the PSK 192-byte packets used for data transmission. If the acknowledgment is lost, the two-hop transmission is considered a failure by the relay. A touring survey consisting of a single circuit with ten transmissions at each site was performed before the experiment to provide a comparison measure, and the MABSC algorithm was run for the same mission time (55 minutes). Subsequently, the relay robot executed the MABSC algorithm autonomously.

IV. Results

Fig. 2 shows SNR-In values reported for all acoustic transmissions during the MAB field experiment, Trial 2. The spread of values is -wide (25 to 25dB) and there is no clear spatial structure to the distribution. The closest locations for each respective link (Sites 1 and 5) do not have discernibly higher SNR-In values. Despite relatively high SNR-In values, the complex, shallow-water environment makes packet decoding difficult and introduces the spatial variation that the multi-armed bandit exploits. Altimetry data reveals highly irregular bottom topography in the area visited by the relay, especially a very shallow (1-3m) shelf to the northeast and a deeper trench towards the south (Boston) bank where the destination node is situated [6].

[4]Practically, the choice of initialization represents an acceptable performance threshold. Unexplored sites may never be chosen if a previous site maintains performance above or equal to the threshold. Here we have prioritized exploration of all possible locations.

[5]Computed with Matlab R2012b on Windows 7 (64bit), Intel i5-3450, 16GB of RAM

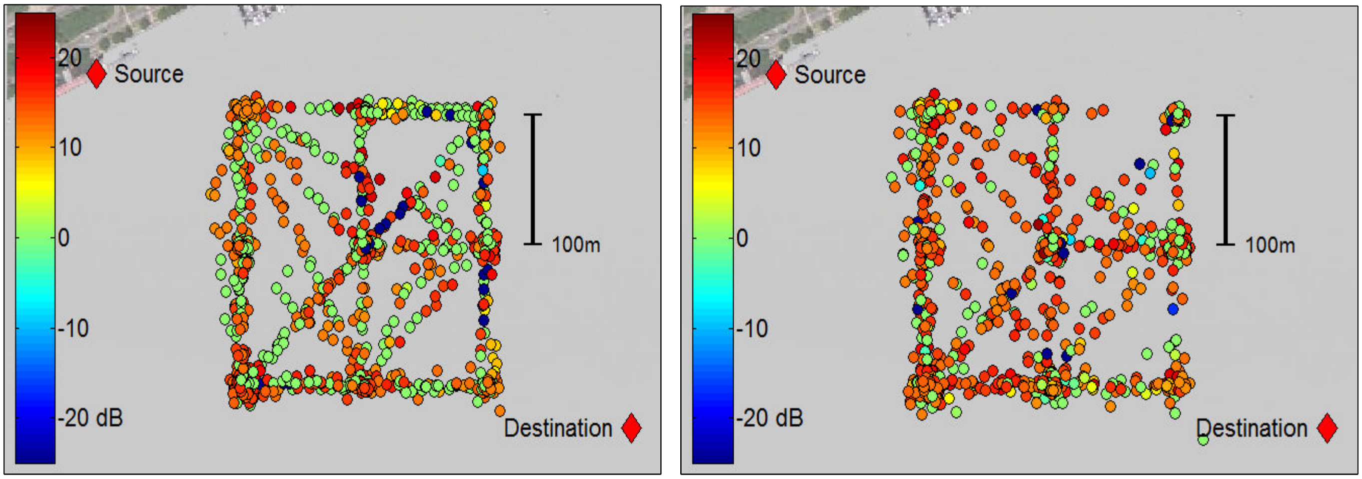

Fig. 2. SNR-In for data transmissions from MAB field experiment, Trial 2. Direction of transmission is shore to relay (left) and relay to destination (right).

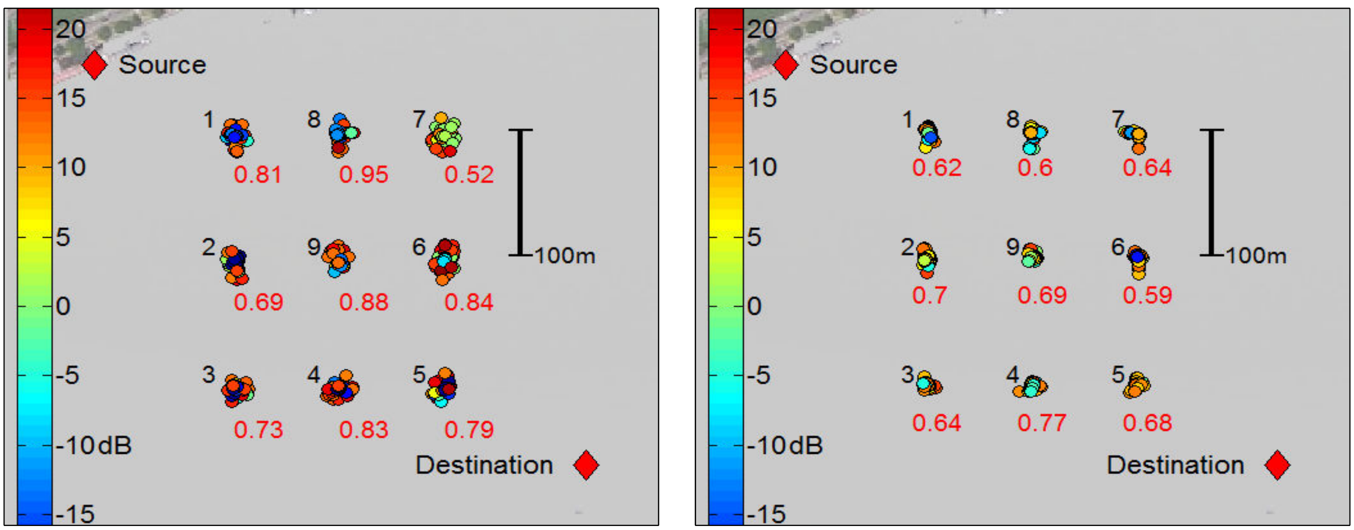

Fig. 3. SNR-In for data transmissions from TourA (left) and B (right). Direction of transmission is source to relay. Site number is shown in black and final packet success rates estimates over the whole mission is shown in red.

Fig. 3 shows SNR-In values and final packet success rates from two touring surveys (TourA and B) of the nine relay way-points conducted on different days. For these touring surveys, the mobile relay transmitted five times at each location per visit, at PSK Rate 2 with a fixed message size of 192 bytes, and acknowledgments of transmission receipts were communicated over Wi-Fi for simplicity and shorter cycle times. There is significant spatial variation in final packet success rates by location. TourA showed a spread in means of 43%, with Site 8 outperforming the next nearest by 7%. In comparison, TourB shows a narrower spread of 18% and Site 4 was the highest performing and Site 8 exhibiting poor performance. Variation from day to day motivates the need for a fast and adaptive learning algorithm without the need for re-tuning heuristic parameters. The MAB and MABSC formulations adaptively explore these space-varying properties of the acoustic channel without sacrificing overall data transmission or requiring prior knowledge of channel characteristics.

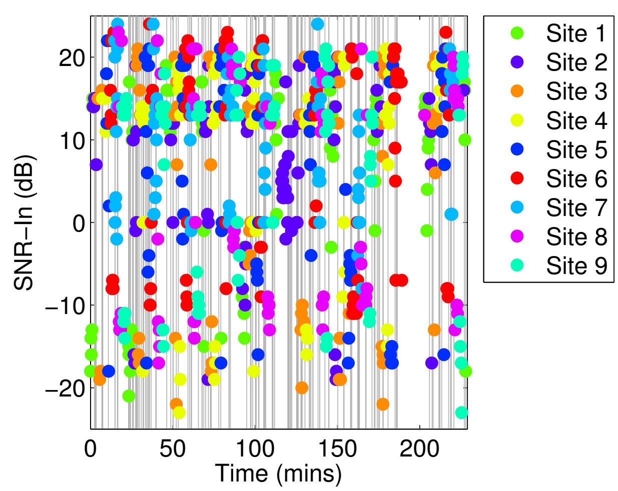

Fig. 4. SNR-In over time for TourA. Grey lines represent lost packets. Sites are visited in the same order and direction of transmission is source to relay.

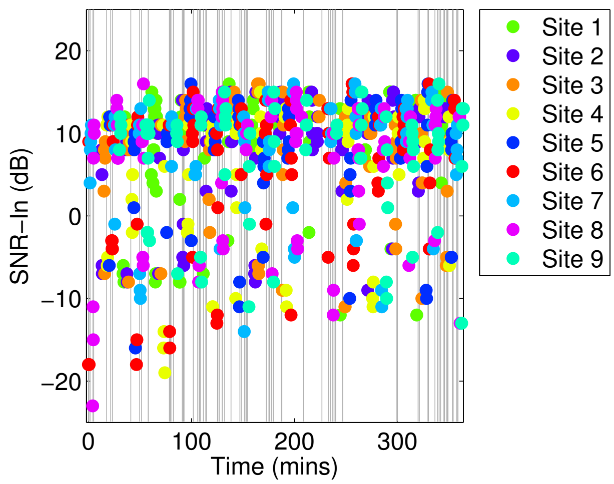

Fig. 5. SNR-In over time for TourB. Grey lines represent lost packets. Sites are visited in the same order and direction of transmission is source to relay.

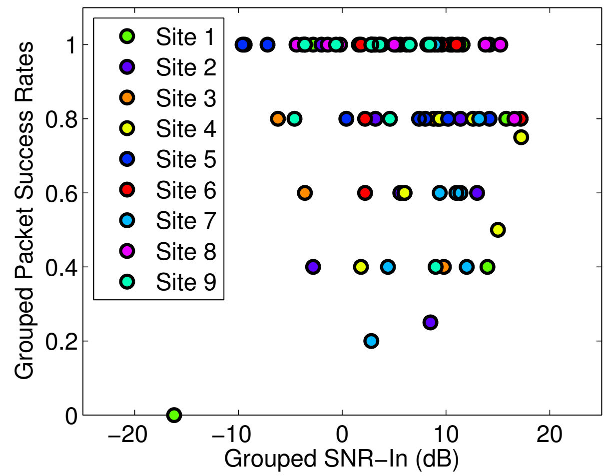

Fig. 7. TourA Grouped Packet Sucess Rates against SNR-In (right). Data is for source to relay transmission only. Each averaging group consists of five transmissions on location.

Figs. 4 and 5 show SNR-In values from the touring surveys. No clear trend can be identified in SNR-In temporally or spatially from these systematic surveys. Fig. 6 shows Grouped SNR-In time series by site for the touring survey datasets. As no clear trend in SNR-In values is discernible temporally or spatially, we are able to assume the Bernoulli transmission processes are stationary over the time scale of this experiment. Remarkably, as illustrated in Fig. 7, there is essentially no correlation of SNR-In with the corresponding grouped packet success rates of those transmissions, with high variation in SNR-In even for 100% success.

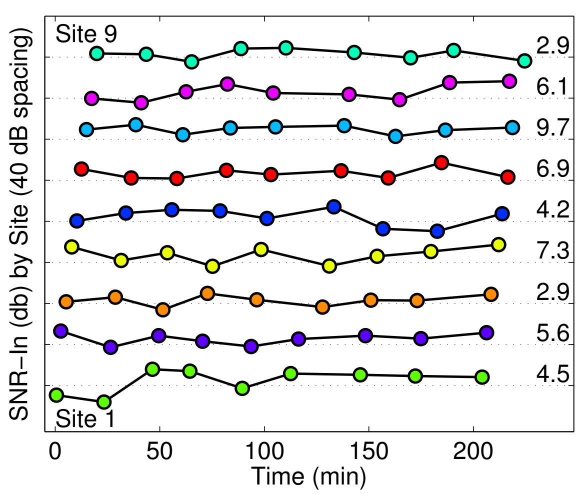

Fig. 6. TourA Grouped SNR-In values over time, with overall average noted at right. Data is for source to relay transmission only. Each averaging group consists of five transmissions on location.

A. Autonomous MAB

Fig. 8 presents the evolution by observation of the algorithm's three key parameters for Trial 2: the Gittins index, the estimated mean, and the variance of the estimate of each machine. The initial choice of exploration for Location 7 with low mean but high variance is clear, as well as eventual settling into the high-performing Location 3. As the number of observations increases, the Normal Gittins index falls away exponentially, and the algorithm asymptotically favors one or several points with comparatively higher performance. Evaluation of the algorithm's performance is considered in terms of cumulative bytes of data successfully transmitted via the relay (Fig. 9). For comparison with a method that does not take into account location-dependent acoustic performance, we extrapolate the expected reward from the initial touring survey and present the difference in reward obtained, in Fig 10. During the exploration regime following the initialization period, MAB decision-making performs similarly or worse than the touring approach as the vehicle continues to visit locations with poor performance but higher standard deviations. For Trial 1, the cumulative data transmitted eventually exceeds that of the touring extrapolation by 14% and the final data transmission rate is improved by 33%, while for Trial 2, the cumulative data transmitted and the rate of transmission increased by 19.6% and 37% respectively. In both trials, the MAB behavior was effective at improving cumulative data transmission as compared to the initial survey. However, approximately 60% of mission time was spent in transit, highlighting the significance of switching costs.

B. Autonomous MABSC

We compare the performance of the autonomous MABSC decision algorithm to the touring survey conducted on the same day and consisting of ten transmissions at each site over one circuit. The two algorithms were limited to the same mission duration. Figs. 11 and 12 show the performance of the MABSC in comparison with the tour. During the first thirty minutes of the mission, performance was similar as the MABSC explored the unknown field. For the remainder of the mission, the MABSC settled to a high-performing site, and achieved a final reward rate 77.2% higher, and an average reward rate 28.4% higher than the touring survey. Since the MABSC is able to discard poorly-performing locations while accounting for the cost of switching locations, it was able to carry out more transmissions (Fig. 11).

We have shown that the MABSC decision policy efficiently maximizes data transmission and outperforms a simple touring

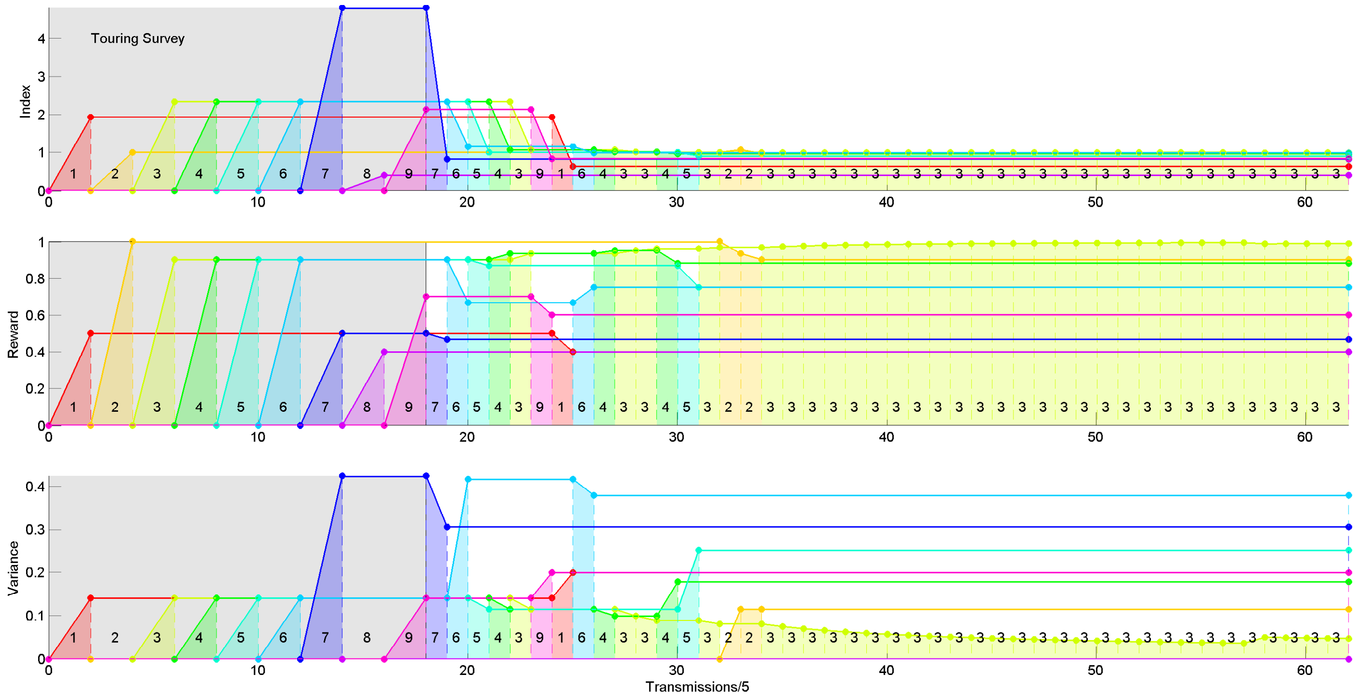

Fig. 8. Evolution of the Gittins Index, the mean, and the variance of the estimate for Trial 2.

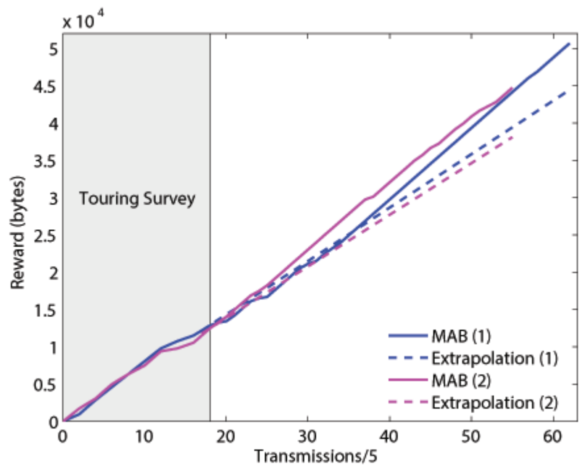

Fig. 9. Cumulative bytes transmitted for Trials 1 and 2 and extrapolations of inital touring survey.

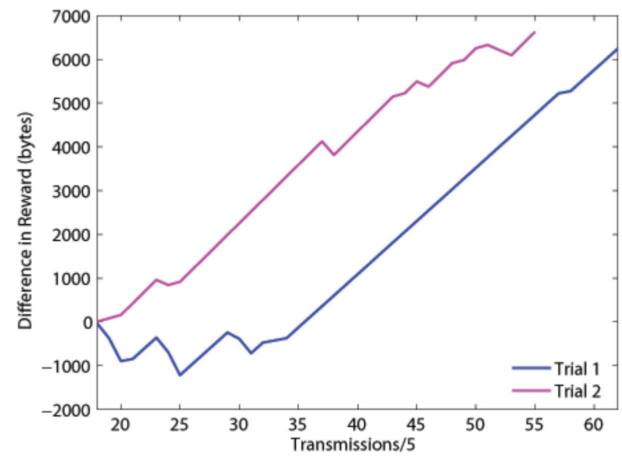

Fig. 10. Improvement in cumulative bytes transmitted with respect to extrapolations of inital touring survey.

survey. Finally, we recall that the multi-armed bandit formulation optimally trades-off exploration and exploitation, and we expect the MABSC to gain information about the field effectively. Thus, we assess the performance of the touring survey and the MABSC policy in exploration. We value an algorithm's information gain by computing the total sum of squares error for each algorithm summed over all relay sites. The squared error is taken between the algorithm's current estimate of the site's performance, and the best possible estimate obtained from both datasets. Fig. 13 shows the evolution of this error by observation for the touring survey and the MABSC. It is clear the MABSC gathers information about the performance field significantly faster than a simple touring survey, and does so without sacrificing data transmission.

V. Conclusion

We have shown that a multi-armed bandit formulation for autonomous, adaptive acoustic relay positioning addresses the direct but complex coupling between channel properties and an acoustic relay's physical location. In particular, the MAB algorithm optimally trades off real-time exploration of the field with exploitation of high-performing sites. We implemented the MAB as part of a fully autonomous system, with mobile surface robots and acoustic modems in shallow water, and evaluated the decision policy's performance in the field. These experiments showed the significance of switching costs, and we augmented the canonical MAB to include switching costs based on proven properties of the optimal solution. As demonstrated by autonomous field experiments, the MABSC is an ef-

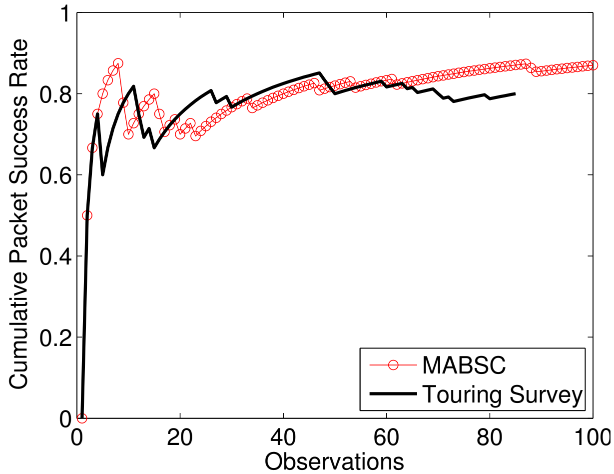

Fig. 11. Cumulative packet success rate by observation, where unity indicates 100% success rate.

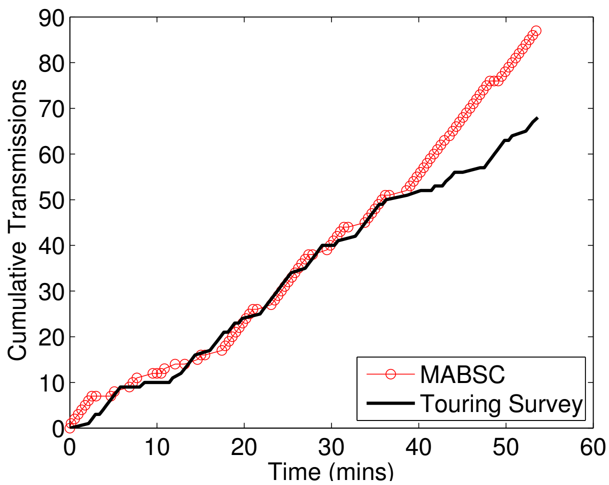

Fig. 12. Cumulative data transmitted by mission time.

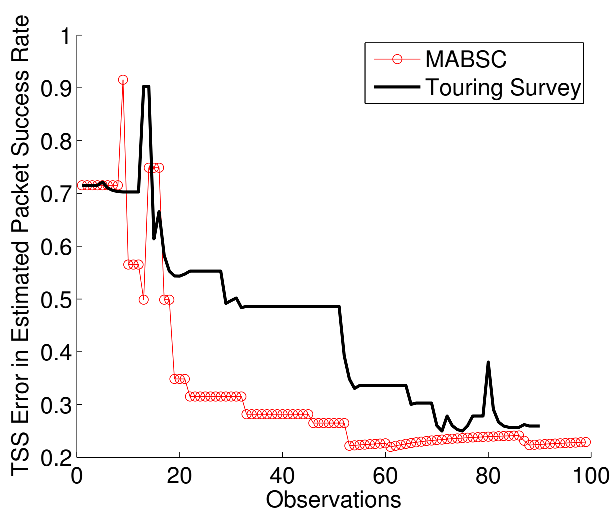

Fig. 13. Total sum of squares error by observation.

ficient and practical procedure for identifying high-performing relay sites and maximizing cumulative data transmission in an unknown and complex shallow-water environment.

Acknowledgements

This work is supported by the Office of Naval Research, Grant N00014-09-1-0700, the National Science Foundation, Contract CNS-1212597, and Finmeccanica. We thank Toby Schneider and Mike Benjamin at MIT; Keenan Ball and Sandipa Singh at WHOI; and MIT Sailing Master Franny Charles.

References

[1] M. Chitre, S. Shahabudeen, and M. Stojanovic. Underwater acoustic communications and networking: Recent advances and future challenges. *Marine Technology Society Journal*, 42(1):103–116, 2008.

[2] J. C. Preisig. Performance analysis of adaptive equalization for coherent acoustic communications in the time-varying ocean environment. *The Journal of the Acoustical Society of America*, 118(1):263–278, 2005.

[3] M. Stojanovic and J. Preisig. Underwater acoustic communication channels: Propagation models and statistical characterization. *Communications Magazine, IEEE*, 47(1):84–89, Jan. 2009.

[4] J. B. Bowlin, J. L. Spiesberger, T. F. Duda, and L. E. Freitag. Ocean acoustical ray-tracing : Software Ray. Technical report, WHOAS at MBLWHOI Library, 1992.

[5] B. Tomasi, G. Zappa, K. McCoy, P. Casari, and M. Zorzi. Experimental study of the space-time properties of acoustic channels for underwater communications. In *Proc. IEEE OCEANS, Sydney*, pages 1–9, 2010.

[6] M. Cheung, J. Leighton, and F. Hover. Multi-armed bandit formulation for autonomous mobile acoustic relay adaptive positioning. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

[7] I. F. Akyildiz, D. Pompili, and T. Melodia. State-of-the-art in protocol research for underwater acoustic sensor networks. In *Proc. 1st ACM International Workshop on Underwater Networks*, pages 7–16, 2006.

[8] C. Detweiler, M. Doniec, I. Vasilescu, E. Basha, and D. Rus. Autonomous depth adjustment for underwater sensor networks. In *Proc. 5th ACM International Workshop on Underwater Networks*, pages 12:1–12:4, 2010.

[9] V. Krishnamurthy and R. J. Evans. Hidden markov model multiarm bandits: a methodology for beam scheduling in multitarget tracking. *IEEE Transactions on Signal Processing*, 49(12):2893–2908, Dec. 2001.

[10] Y. Gai, B. Krishnamachari, and R. Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *Proc. IEEE Symposium on New Frontiers in Dynamic Spectrum*, pages 1–9, Apr. 2010.

[11] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proc. 25th International Conference on Machine Learning (ICML)*, pages 784–791, 2008.

[12] J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

[13] P. Whittle. Multi-armed Bandits and the Gittins Index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):143–149, 1980.

[14] M. Grossglauser and D. Tse. Mobility increases the capacity of ad-hoc wireless networks. In *Proc. 20th Joint Conference of the IEEE Computer and Communications Societies (INFOCOM).*, volume 3, pages 1360–1369, 2001.

[15] A. So and B. Liang. Effect of relaying on capacity improvement in wireless local area networks. In *Proc. IEEE Wireless Communications and Networking Conference*, volume 3, pages 1539–1544, 2005.

[16] J. D. Holmes, W. M. Carey, J. F. Lynch, A. E. Newhall, and A. Kukulya. An autonomous underwater vehicle towed array for ocean acoustic measurements and inversions. In *Proc. IEEE OCEANS-Europe*, volume 2, pages 1058–1061, 2005.

[17] J. S. Banks and R. K. Sundaram. Switching costs and the Gittins Index. *Econometrica*, 62(3):687–694, 1994.

[18] M. Asawa and D. Teneketzis. Multi-armed bandits with switching penalties. *IEEE Transactions on Automatic Control*, 41(3):328–348, 1996.

[19] F. Dusonchet and M.O. Hongler. Optimal hysteresis for a class of deterministic deteriorating two-armed bandit problem with switching costs. *Automatica*, 39(11):1947–1955, 2003.

[20] R. Agrawal, M.V. Hedge, and D. Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10):899–906, 1988.

[21] T. Jun. A survey on the bandit problem with switching costs. *De Economist*, 152(4):513–541, 2004.

[22] J. Niño-Mora. A faster index algorithm and a computational study for bandits with switching costs. *INFORMS J. on Computing*, 20(2):255–269, 2008.

[23] J. Le Ny and E. Feron. Restless bandits with switching costs: Linear programming relaxations, performance bounds and limited lookahead policies. In *Proc. IEEE American Control Conference*, pages 6–12, 2006.

[24] D. Bertsimas and J. Niño Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

[25] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, pages 287–298, 1988.

[26] M. Cheung, J. Leighton, and F. Hover. Autonomous mobile acoustic relay positioning as a multi-armed bandit with switching costs. In *Proc. IEEE International Conference on Intelligent Robots and Systems (IROS), to appear*, 2013.

[27] I. M. Sonin. A generalized Gittins index for a Markov chain and its recursive calculation. *Statistics and Probability Letters*, 78(12):1526–1533, 2008.

[28] J. Niño-Mora. A (2/3)n3 Fast-Pivoting Algorithm for the Gittins Index and Optimal Stopping of a Markov Chain. *INFORMS J. on Computing*, 19(4):596–606, Oct. 2007.

[29] J. C. Gittins, R. Weber, and K. D. Glazebrook. *Multi-armed bandit allocation indices*, volume 25. Wiley Online Library, 1989.

[30] W. B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. Wiley-Blackwell, 2007.

[31] L. Freitag, M. Grund, S. Singh, J. Partan, P. Koski, and K. Ball. The WHOI micro-modem: an acoustic communications and navigation system for multiple platforms. In *Proc. MTS/IEEE OCEANS*, volume 2, pages 1086–1092, Sept. 2005.