2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

# Autonomous Mobile Acoustic Relay Positioning as a Multi-Armed Bandit with Switching Costs

Mei Yi Cheung, Joshua Leighton, Franz S. Hover

*Abstract*— **Underwater acoustic communication channels display highly variable and stochastic performance, especially in multipath-limited shallow-water and harbor environments. A mobile acoustic node can, however, learn the channel's properties as it moves about. Maximizing the cumulative data transmission through adaptive node positioning is a clean exploitation vs. exploration scenario because learning about poorly characterized locations must be balanced against exploiting known ones. While this problem is well described with the stochastic multi-armed bandit formalism, the classical assumption of costless switching is untenable in the field, where slow-moving vehicles often cover large distances. We present a heuristic adaptation to the MAB Gittins index rule with limited policy enumeration to account for switching costs, and describe field experiments conducted in the Charles River (Boston MA). The field data establish that the MAB and its switching cost extension are tractable in this application, and that performance is consistently superior to that of $\epsilon$-greedy policies.**

## I. Introduction

Acoustic communications is the main practical means of underwater wireless data transmission. It has wide-ranging ocean applications, such as data collection from underwater sensor networks for monitoring ocean processes, remote control of untethered mobile robots and point-to-point communication. In recent years, research on underwater acoustic communications systems has led to significant improvements in performance, through modern channel estimation, coding and error correction schemes (e.g. [1], [2]). However, channel performance remains strongly dependent on local environment properties such as temperature, salinity, wind and waves, and may vary on several time scales [3]. In shallow water, surface and bottom conditions as well as man-made structures cause multipath interference that affects performance in a spatially complex manner. Sophisticated ray and beam-tracing algorithms can be used to predict these effects but may be computationally expensive even in a two-dimensional setting and require measuring or modeling of environmental properties [4], [5].

It is challenging to predict the performance of the acoustic channel, especially in relation to the position of the nodes. Thus, in previous work [6], we posed the problem of improving cumulative data transmitted between fixed source and destination nodes by adaptively positioning a mobile relay. The relay learns about local channel performance when it transmits and must trade off exploring poorly characterized sites with exploiting known ones for throughput. Formulating this problem as a multi-armed bandit (MAB) allowed us to use an elegant and optimal decision rule in the form of Gittins indices [7] for a discretized space of potential relay locations. Using autonomous vehicles in the field, we showed that the Gittins index rule improved cumulative data transmission by 14% and 19% over a simple touring strategy [6]. However, the experiments also showed that 60% of the total mission time was spent in transit between locations. For an underwater vehicle traveling between distant waypoints, generally slow vehicle speeds mean that choosing to switch location may take much longer than choosing to sample again at the present location, a tradeoff not accounted for by the MAB algorithm. Additionally, in practice, the performance of the acoustic link may be adversely affected during transit by factors like increased vehicle propulsion noise [8].

In this paper, we address switching costs explicitly from an algorithmic point of view and describe an adaptation of the Gittins index rule with limited-horizon lookahead policy enumeration. Introducing switching costs into the canonical MAB framework removes the property of arm-independence, and Banks and Sundarum proved that no optimal index policy solution exists for the MABSC [9]. Subsequent research has focused on deriving general properties of the optimal policy [10], deriving explicit optimal policies for special cases [11], [12] and bounding approximations to the optimal policy [13]. The problem has also been reformulated as a semi-Markov multi-armed restless bandit, for which marginal productivity indices are a near-optimal solution[1] [14]. For a recent survey, see Jun [15]. Applying the main result of Asawa and Teneketzis [10] allows us to adapt the Gittins index policy for switching costs with limited-horizon policy enumeration by reducing the frequency of expensive computations.

$\epsilon$-greedy algorithms, which combine myopic greedy behavior with heuristic random sampling, are well-known alternatives to the MAB [16]. An $\epsilon$-greedy algorithm plays the best arm $(1-\epsilon)$ of the time and switches to a random arm $\epsilon$ of the time. $\epsilon$-decreasing is a variation where the value of $\epsilon$ decreases with time constant $\tau$. These algorithms are simple to implement and must be tuned heuristically for good performance. Drawing samples from a large survey dataset collected in the field, we apply the MAB, MABSC and $\epsilon$-greedy decision policies to the same data and compare the performance of each of these decision rules. We show that the real-time performance of the MABSC heuristic is competitive with that of $\epsilon$-greedy algorithms, while at the same time gathering more information on the field.

M. Cheung, J. Leighton. and F. Hover are with the Department of Mechanical Engineering, MIT, Cambridge, MA 02139, USA {mc2922, jleight,hover} at mit.edu

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

Section II of this paper describes the MAB formulation for the adaptive positioning problem, a heuristic for the inclusion of switching costs and asymptotic policy bounds. In Section III, we describe a hybrid field experiment conducted with autonomous surface vehicles, and in Section IV we compare the MAB algorithm with and without a switching cost heuristic to well-known alternatives.

## II. Problem Formulation

### A. Adaptive Relay Positioning as a Multi-Armed Bandit

The multi-armed bandit framework considers the problem of dynamically allocating a resource between $N$ competing processes or "arms" so as to maximize the total expected reward. Each arm is described by a sequence of states $x(1), \cdots, x(n)$, where $x(n)$ is a random variable representing the state of the arm after it has been operated $n$ times. In the multi-armed process, we denote the number of times each arm $i$ has been operated by $n_i$, and denote its state by $x_i(t)$, where $t$ is the current global decision epoch:

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

In general, the reward returned by each state $R(x_i(t))$ is a real non-negative random variable. We denote the state of the multi-arm process at a given time by $\bar{x}$, which is the vector $(x_1 \cdots x_N)$. At each decision epoch, the process allocates the available resource to a single arm, reaping its associated state-dependent reward while the states of all other arms remain frozen.

For the relay positioning problem, we define each potential relay location as an arm. The relay "plays" the arm by relaying on location, each acoustic transmission being described by a Bernoulli trial:

$$B_i(t) = \begin{cases} 1 & \text{if transmission success} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and the reward process of each arm is Bernoulli.

A dynamic allocation policy $\pi$ is optimal if it defines at each decision epoch $t$ an arm for allocation $i_t$, such that the expected value of the total expected reward $V_\pi$ is maximized. For a discount factor $0 < \beta < 1$, and an infinite horizon, the total expected reward is:

$$V_\pi(\bar{x}) = E\left[\sum_{t=0}^{\infty} \beta^t R(x_{i_t}(t)) \mid \bar{x}(0) = \bar{x}\right]. \tag{3}$$

Gittins and Jones showed that the optimal solution is an index policy, with the value of playing an arm at any time represented by an index $\nu_i$ that is a function only of that machine's current state $x_i(t)$ [7]. The Gittins index can be understood as the expected discounted reward per unit time maximized over all stopping times $\tau > 1$, conditioned on the process state at global time $t$:

$$\nu_i(x_i(t)) = \underset{\tau>1}{max} \frac{E\left[\sum_{k=0}^{\tau-1} \beta^k R(x_i(k)) \mid x_i(0) = x_i(t)\right]}{E\left[\sum_{k=0}^{\tau-1} \beta^k \mid x_i(0) = x_i(t)\right]}. \tag{4}$$

The optimal rule at each decision epoch is to play the machine with the largest current Gittins index, defining the policy $\pi$ as:

$$i_t = \underset{i}{\operatorname{argmax}}(\nu_i(x_i(t))). \tag{5}$$

A computational method for calculating the Gittins indices for a Bernoulli reward process is described in Gittins [17]. The infinite horizon is approximated with a large finite horizon and backwards induction is used to solve for indices.

### B. Policy Enumeration for Switching Costs

We define constant costs $c(i, j)$ to achieve a practical representation of the attractiveness of taking one more sample before investing in a lengthy transit. If $t_v(i, j)$ is the time taken to travel from $i$ to $j$, and $t_t$ is the time taken to relay one transmission, then $c(i, j) = t_v(i, j)/t_t$ — the number of transmissions the relay could have made on location if it had chosen to sample instead of travelling, or the approximate potential value given up by choosing to transit. These costs are the elements of a constant symmetric matrix whose diagonals are 0. In the presence of switching costs, the index policy was shown to be suboptimal [9]. The optimal solution to the MABSC is one that maximizes:

$$V_\pi(\bar{x}) = E\Bigg\{ \sum_{t=0}^{\infty} \beta^t \big[R(x_{i_t}(t)) - c(x_{i_t}(t), x_{i_t}(t+1))\big] \mid \bar{x}(0) = \bar{x} \Bigg\}. \tag{6}$$

The assumption of arm-independence necessary to decompose the $N$-dimensional problem into $N$ one-dimensional optimization problems no longer holds, since the reward returned by a process no longer depends solely on the number of times $n$ an arm has been operated. For this problem, we describe a solution of the *priority-index policy* form, where separate "continuation" and "decision" indices are used [18]. The continuation index $\nu_i$ is the Gittins index previously defined, and determines if an arm previously played should be continued. Asawa and Teneketzis showed that it is optimal to continue playing an arm up to its stopping time $\tau$, and only making a decision to switch when the stopping time is achieved (A&T Thm. 2.1) [10]. The stopping time is achieved when the Gittins index of the current arm falls below any value it has previously reached, thereby defining the continuation rule:

$$\text{if } \min_{k<t} \nu_{i_k}(x_{i_k}(k)) \leq \nu_i(x_i(t)), \text{ set } i_{t+1} = i_t. \tag{7}$$

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. The decision index is used to determine which arm to switch to when the stopping time is achieved. We calculate a "policy" decision index by maximizing an $m$-horizon look-ahead enumeration of the expected reward rate over all possible policies $\pi$, where $\pi$ is any possible sequence of plays $i_1, ..., i_m \; \forall i \in 1, ..., N$.

Location-based switching costs are simple to include in this formulation, and the value of being in the final state is accounted for with an updated Gittins index $\nu_i$:

$$\eta_\pi(\bar{x}(t)) = \frac{e(\bar{x}(t))}{E\left[\sum_{k=0}^{m} \beta^k \mid \bar{x}(0) = \bar{x}(t)\right]} + \nu_i(\hat{x}_{i_{t+m}}(t+m)), \quad (8)$$

where

$$e(\bar{x}(t)) = E\Big\{ \sum_{k=0}^{m} \beta^k \big[R(x_{i_k}(k)) - c(x_{i_k}(k), x_{i_{k+1}}(k+1))\big] \mid \bar{x}(0) = \bar{x}(t) \Big\}. \quad (9)$$

The adapted decision rule for MABSC is then

$$i_{t+1} = \underset{i_1}{\operatorname{argmax}}(\eta_\pi(\bar{x}(t))). \quad (10)$$

If $m = 0$, canonical Gittins indices are returned. If $m = 1$, switching indices described in [10] is returned. In general, the number of possible policies scales exponentially with the number of arms and the length of the lookahead horizon. 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.

### C. Asymptotic Efficiency

Asymptotically efficient suboptimal decision rules are guaranteed to converge to the optimal arm as the number of observations increases, and perform asymptotically as well as optimal policies. We follow the construction of Agrawal [13], who extended [19] to the case with switching costs, to determine asymptotic efficiency for the MABSC adaptation. We assume each reward distribution can be parametrized by an unknown parameter $\theta$ and denote the optimal mean by $\mu^*$. An allocation rule's "regret" is defined to be the loss in reward due to suboptimal action:

$$R_t(\bar{\theta}) = R'_t(\bar{\theta}) + SW \quad (11)$$

where $R'_t(\bar{\theta})$ is the sampling regret, $SW$ is the switching regret and $\bar{\theta} = \{\theta_1, ..., \theta_N\}$. For $t \leq \infty$, let $T_t(i)$ is the number of times arm $i$ is sampled,

$$T_t(i) = \sum_{k=1}^{t} 1\{i_k = i\}. \quad (12)$$

Similarly, let $S_t(i,j)$ is the number of times a switch $i$ to $j$ was made. Then, the sampling and switching regret are:

$$R'_t(\bar{\theta}) = \sum_{i:\mu(\theta_i)<\mu^*} (\mu^* - \mu(\theta_i)) E[T_t(i)] \quad (13)$$

$$SW = \sum_{i=1}^{N} \sum_{j=1}^{N} c(i,j) E[S_t(i,j)], \quad (14)$$

For the lower bound to hold, the decision rule's regret must not increase sharply, i.e. the rule must be uniformly good. A rule is uniformly good if it satisfies the following condition for every parameter and cost configuration $\bar{\theta}$, $c(i,j)$,

$$R_t(\bar{\theta}, c) = o(t^a) \ \forall\, a > 0 \quad (15)$$

Assuming no arm is impossible to switch away from, this rule is satisfied by the adapted MABSC decision rule as suboptimal arms with decreasing Gittins indices are not continued. For the class of uniformly good allocation rules, Lai and Robbins [19] proved the following lower bound:

$$\liminf_{t\to\infty} R_t(\bar{\theta}) \geq \left[\sum_{i:\mu(\theta_i)<\mu^*} \frac{(\mu^* - \mu(\theta_i))}{I(\theta_i, \theta^*)}\right] \log \text{ t} \quad (16)$$

where $I(\theta_i, \theta^*)$ is the Kulback-Leibler number. Agrawal showed directly from Eqn. 16 that any asymptotically efficient decision rule takes about $O(\log \text{ t})$ samples from any inferior process up to time $t$ and proved that a block allocation scheme where the expected number of switches is controlled to $o(\log \text{ t})$ a priori is asymptotically efficient [13]. Similarly, the adapted MABSC decision rule is a block allocation rule where the use of Gittins continuation indices guarantees asymptotic continuation of the optimal process and decreased sampling of inferior processes. Given that every arm can be switched to and away from, policy enumeration by the decision process converges to the optimal process and the number of switches decreases with time. In comparison, the $\epsilon$-greedy algorithm does not achieve the lower bound as it must switch away from the optimal arm $\epsilon\%$ of the time, while the $\epsilon$-decreasing algorithm does [20].

## III. Experiment Description

We consider a one-way, two-link network, originating from a source modem at the MIT Sailing Pavilion on the Charles River Basin, and repeated by the mobile relay to a station-keeping vehicle 580m across the river. A transmission is considered successful if both links are successful. All field experiments were conducted with custom autonomous surface vehicles (Fig. 1) towing acoustic modem transducers at fixed depth to simulate underwater communications, with the benefit of GPS (noise covariance on the order of $10m^2$) and WiFi connectivity for controlled experiments. The vehicles travel at 1.5 m/s and maintain a station-keeping circle ten meters in diameter on location.

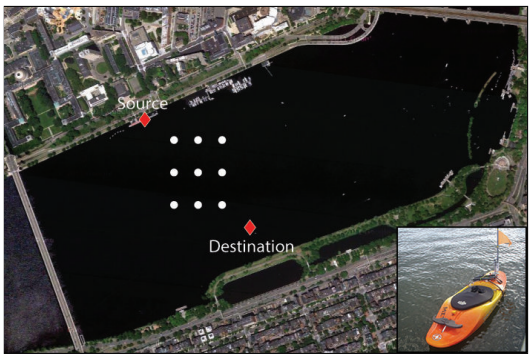


Fig. 1. Autonomous surface vehicle operating off the MIT Sailing Pavilion.

We use Woods Hole Oceanographic Institution (WHOI) Micro-Modems [21], 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"). Micro-Modem transmissions were fixed at PSK Rate 2, with a message size of 192 bytes and an average two-hop transmission time of fifteen seconds. We note that although programmable modem parameters such as packet encoding schemes can be included combinatorially as additional machines, we have fixed these for simplicity. No prior knowledge of the acoustic channel was assumed beyond the usual spreading law to choose nine candidate relay locations spaced 100m apart in a grid pattern centered on the line between the source and destination nodes (Fig. 1). The time taken to switch ranges from 1.1 to 3.2 minutes. In practice, the choice of potential relay locations will be influenced by mission constraints. The depth of the Charles River Basin ranges from two to twelve meters.

In the field it is difficult to compare the performance of several competing algorithms as multiple relays would share the same physical space and channel, resulting in transmissions experiencing acoustic interference or extended wait times. Conducting experiments on different days is also undesirable as changing weather and surface conditions make it difficult to objectively evaluate the improvement in performance due to action by the algorithms. Thus, we construct a hybrid experiment; first, by collecting a large dataset of transmissions on a single experimental day. A touring survey taking five transmissions at every location was conducted for several hours. Then, each decision algorithm was applied to the same dataset, i.e. transmission results were sampled from the dataset for the appropriate time and location and used to update the algorithm's information state. The shallow-water acoustic environment is in general difficult to model and using field data allows us to capture complex spatially-dependent behavior. The hybrid dataset contained 835 detected transmissions from source to relay and 636 detected transmissions from relay to destination, with 493 of these being successfully decoded relayed transmissions.

## IV. Experiment Results

### A. Acoustic Channel Statistics

SNR-In values reported for all acoustic transmissions during the data collection mission are presented in Fig. 2. We note that the data demonstrates no clear spatial structure and does not noticeably distinguish between locations of different performance, with Site 7 as a possible exception. The highest and lowest performing locations are situated surprisingly close by. Altimetry data in the area visited by the relay shows irregular bottom topography and a shallower shelf to the northeast where Site 7 is situated [6]. Though not visible, a deeper channel is also present towards the south (Boston) bank where the destination node is situated. Fig. 3 shows SNR-In values by mission time and is color-coded by location, with lost packets shown as dotted lines. Despite high SNR-In values, multi-path interference in the shallow-water environment makes packet decoding challenging.

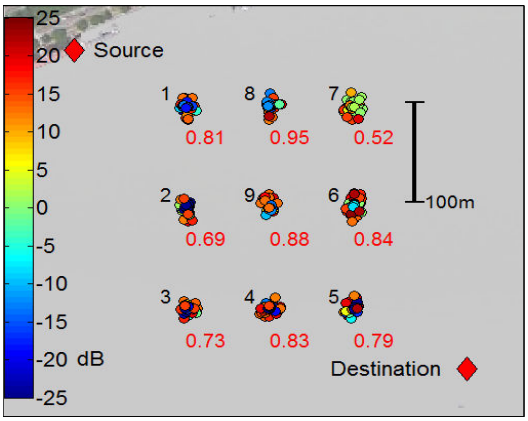


Fig. 2. SNR-In of transmissions at each of nine potential relay locations in the Charles River Basin. Site number is shown in black and final packet success rates estimates over the entire mission are shown in red.

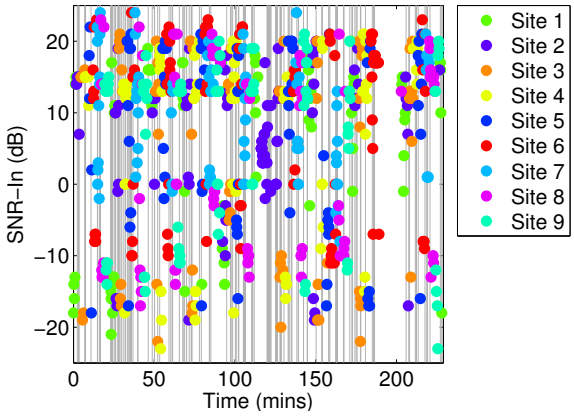


Fig. 3. SNR-In values over time, color-coded by site, with lost packets shown as grey lines. Sites are visited in the same order each tour.

SNR-In values for the five transmissions taken each tour were averaged and Fig. 4 (left) shows the progression in each site for the time-averaged values of SNR-In. There is no clear trend in these values temporally and thus we assume the Bernoulli transmission processes to be acceptably stationary over the time scale of the experiment. Remarkably, as illustrated in Fig. 4 (right), 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.

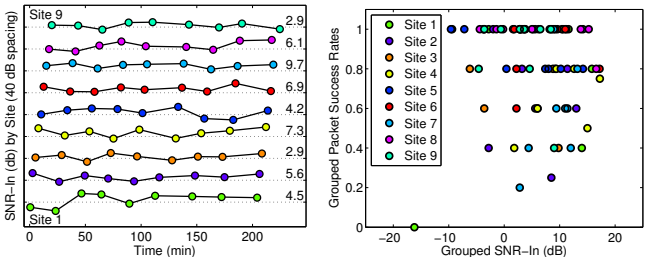


Fig. 4. HybridSetA Grouped SNR-In values over time (left), with overall average noted at right, and Grouped Packet Sucess Rates against SNR-In (right). Data is for source to relay transmission only. Each averaging group consists of five transmissions on location.

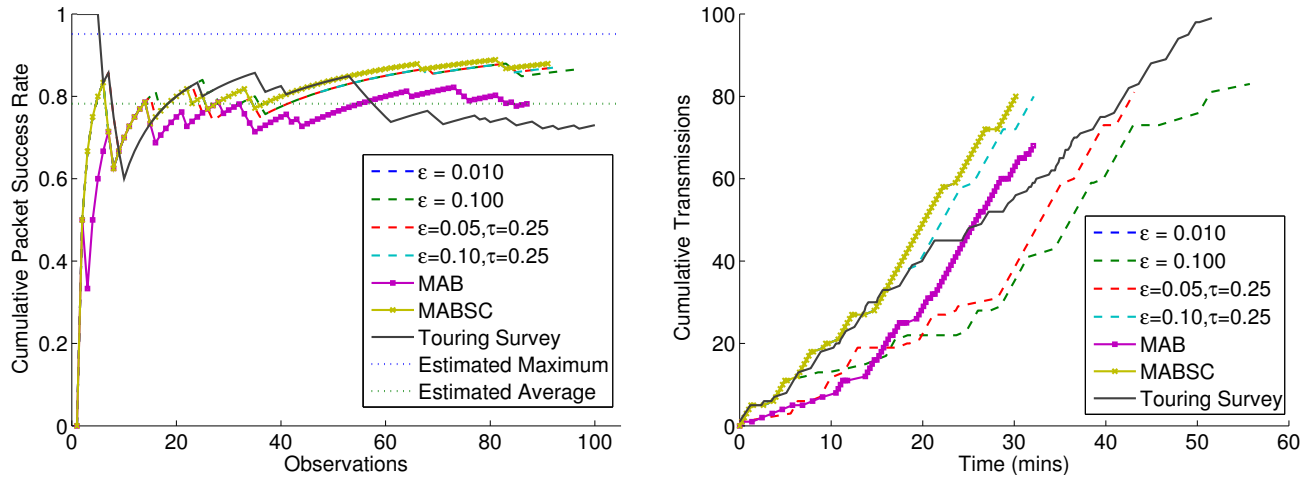


Fig. 5. Cumulative performance of MAB, MABSC and tuned $\epsilon$-greedy and $\epsilon$-decreasing algorithms by observations (left), where 1 indicates 100% success rate, and cumulative successful transmissions by calculated mission times (right).

### B. Algorithm Comparison

We compare the performance of the Bernoulli Gittins index solution (MAB), the switching cost adaptation (MABSC), $\epsilon$-greedy, $\epsilon$-decreasing algorithms and the initial subset of the touring survey. Each algorithm was initialized with an estimate of one at each site.[2] and Site 1 was designated as the starting site. Initializing $\epsilon$-greedy algorithms with a tour did not demonstrate consistently improved algorithm performance, and in the interest of comparison we fix the same start conditions for each algorithm. Indices were initialized to one where appropriate. There was no restriction on the expected number of switches for algorithms other than algorithmically for the adapted MABSC. The look-ahead horizon for policy enumeration was constrained by a maximum computation time of 15 seconds or the time taken to sample once, and a horizon of five took an average of one second.[3] $\epsilon$-dependent algorithms were tuned with $\epsilon$ and $\tau$ of differing orders of magnitude and only a high performing subset is presented here for the sake of clarity. We evaluate each algorithm's performance in terms of the average packet success rate achieved, which is computed from the cumulative number of successful transmissions drawn from the experimental dataset. Transmissions were drawn from the dataset in chronological order, terminating when unavailable data was requested. Since the number of transmissions at each location is limited by the total mission time of the touring survey conducted in the field, fewer observations are generated for greedier algorithms sampling at one location more often.

Fig. 5 shows the cumulative performance level by observations for the algorithms considered, where the one on the y-axis corresponds to 100% cumulative packet success rate. The estimated final and average success rates were computed from the entire data set. From the final estimated means of Fig. 2, we note that the highest performing Site 8 performed significantly better than the next nearest competitor. Thus, the canonical MAB formulation experienced decreased cumulative performance in the beginning from learning about lower performing sites and returns a cumulative packet success rate close to the estimated average. In comparison, MABSC improves the cumulative performance and is closely competitive with tuned $\epsilon$-greedy and $\epsilon$-decreasing algorithms.

Although their performances are comparable on a per observation basis, we demonstrate the impact of switching (travel) times on the cumulative performance for each algorithm (Fig. 5). Transit times were determined by assuming the vehicle was traveling at 1.5 m/s and the time to relay was taken as fifteen seconds. This model closely matches what we observed during field tests. $\epsilon$-greedy with the greatest value of $\epsilon$ demonstrates slow overall rate of transmission as expected, while decreasing values of $\epsilon$ demonstrate higher rates. The MAB formulation is shown to be competitive in real time and its performance is significantly improved by the adaptation to switching costs. The MAB algorithm was found to switch for 12% of decisions, while the MABSC algorithm switched for 8.8%.

Further, the information gained by each algorithm is also of use for practical planning purposes, and we expect the multi-armed bandit to perform efficient exploration, weighted towards characterizing high-performing sites with confidence. We consider the value of information obtained by each algorithm by calculating the total sum of squared differences for each packet success rate estimation from the mean estimated using the whole dataset (Fig. 6).

The MAB steadily improves the estimate at each site, while $\epsilon$-greedy and $\epsilon$-decreasing algorithms improve their estimate only with small probability. While the MABSC does not gain the same amount of information as the canonical bandit, it is able to systematically improve the error while reaping the benefit of improved performance in real-time. The MAB and MABSC's ability to gain significantly more

[2]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.

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

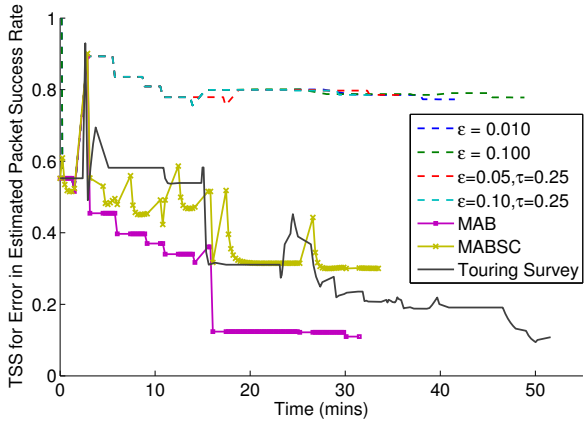


Fig. 6. Total sum of squared differences for each estimated packet success rate, as compared to the success rate estimated with the entire dataset.

information in the same time, number of observations and without compromising overall cumulative data transmission makes these algorithms more practically desirable, so that short missions may have the maximum impact possible.

## V. Conclusion

Adaptive relay positioning using mobile acoustic nodes addresses the relationship between the performance of the acoustic channel and the node's physical location, a relationship that is in practice difficult to model and predict. The multi-armed bandit formulation well-describes the exploration vs. exploitation problem defined by maximizing cumulative performance in an unknown environment, and the Gittins index policy for Bernoulli reward processes provides an optimal, elegant solution without the need for costly parameter tuning. However, in the presence of switching costs, the Gittins indices solution is suboptimal, and data transmission rates are slower as the vehicle spends more time in transit between waypoints. We have described an adaption for switching costs that uses Gittins indices as continuation indices and a limited horizon lookahead policy enumeration to calculate decision indices, with asymptotically efficient performance. The use of separate continuation and decision indices allow us to leverage a proven property of the optimal policy (A&T Thm. 2.1) to reduce computation costs and increase the lookahead horizon.

Algorithm comparison with field data shows that MABSC consistently improves on the performance of MAB and provides comparable real time performance to myopic greedy policies with tuned heuristic parameters. Further, the MAB and MABSC also show significantly improved error estimates compared with greedy algorithms. These algorithmic properties imply that there is little reason to rely on simple surveys or greedy strategies for either throughput or information-gathering goals.

## 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. *Journal of the Acoustical Society of America*, 118(1):263–278, 2005.

[3] B. Tomasi, J. Preisig, G. B. Deane, and M. Zorzi. A study on the wide-sense stationarity of the underwater acoustic channel for non-coherent communication systems. *11th European Wireless Conference - Sustainable Wireless Technologies*, pages 1–6, Apr. 2011.

[4] J. B. Bowlin, J. L. Spiesberger, T. F. Duda, and L. E. Freitag. Ocean acoustical ray-tracing : Software Ray. 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*, pages 1–9, May 2010.

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

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

[8] 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.

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

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

[11] D. Bergemann and J. Välimäki. Stationary multi-choice bandit problems. *Journal of Economic Dynamics and Control*, 25(10):1585–1594, 2001.

[12] 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.

[13] 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.

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

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

[16] J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. In *Proc. 16th European Conference on Machine Learning (ECML)*, pages 437–448. Springer, 2005.

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

[18] J. Niño Mora. Computing an index policy for bandits with switching penalties. In *Proc. 2nd International Conference on Performance evaluation methodologies and tools*, ValueTools, pages 76:1–76:10. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2007.

[19] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4 – 22, 1985.

[20] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2-3):235–256, 2002.

[21] 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.