MIT Libraries | DSpace@MIT

MIT Open Access Articles

Performance of bandit methods in acoustic relay positioning

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Cheung, Mei Yi, Joshua Leighton, Urbashi Mitra, Hanumant Singh, and Franz S. Hover. "Performance of Bandit Methods in Acoustic Relay Positioning." 2014 American Control Conference (June 2014).

As Published: http://dx.doi.org/10.1109/ACC.2014.6859385

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: http://hdl.handle.net/1721.1/98404

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike

Massachusetts Institute of Technology

Performance of Bandit Methods in Acoustic Relay Positioning

Mei Yi Cheung, Joshua Leighton, Urbashi Mitra, Hanumant Singh and Franz S. Hover

Abstract— **We consider the problem of maximizing underwater acoustic data transmission, by adaptively positioning a mobile relay. This is a classic exploration vs. exploitation scenario well-described by a multi-armed bandit formulation, which in its canonical form is optimally solved by the Gittins index rule. For an ocean vehicle traveling between distant waypoints, however, switching costs are significant, and the MAB with switching costs has no optimal index policy. To address this we have developed a strong adaptation of the Gittins index rule that employs limited-horizon enumeration. We describe autonomous shallow-water field experiments conducted in the Charles River (Boston, MA) with unmanned vehicles and acoustic modems, and compare the performance of different algorithms. Our switching-costs-aware MAB heuristic offers both superior real-time performance in decision-making and efficient learning of the unknown field.**

I. Introduction

Acoustic underwater communication has a wide range of ocean applications, including data collection from sensor networks, and point-to-point communication and control of untethered mobile robots. Although research in modern channel estimation and coding schemes has greatly improved throughput and reliability (e.g. [1], [2]), performance of the acoustic channel remains critically dependent on local environmental conditions [3]. In shallow water, multipath interference from the surface and bottom inhibits successful packet decoding. In addition, in harbors and other man-made environments, structures also contribute to multipath, and ambient noise can be a problem. Ray- and beam-tracing algorithms are routinely applied to predict performance, but they may be computationally expensive even in a two-dimensional setting, and typically require measuring or modeling water properties [4], [5]. Such detailed modeling may not be practical in real-time operations. While power, frequency, and coding schemes can be heuristically tuned on some modems [6], acoustic communication remains notoriously unreliable due to poorly understood spatial variability.

In this paper we describe a tractable and near-optimal procedure for adaptive placement of an acoustic relay, with the goal of improving cumulative data transmission from source to destination through the relay. This is a relevant and specific problem in underwater networking, representative of the decisions that have to be made whenever multiple agents communicate wirelessly. Despite the apparent simplicity of the single-relay problem, adaptive placement of acoustic modems has not been systematically studied until recently [7]. Following the discussion of Akyildiz *et al*. [8], Detweiler *et al*. [9], observed extreme variability of packet success rate with no apparent pattern, in 10m-deep water with custom equipment. Data from our field experiments in the Charles River (Boston MA), using Woods Hole Oceanographic Institution (WHOI) MicroModems in similar water depths, show these properties as well as pronounced day-to-day variations. Thus, the first major element of our approach is that we assume *no prior knowledge* about the dependence of channel performance on relay location – except the usual spreading law suggesting that the relay be situated somewhere between the source and destination nodes.

Gaining knowledge about the acoustic field, perhaps best achieved through a systematic survey, has to be balanced against increasing throughput, best achieved by spending time at promising sites. If the channel process is stationary, this balancing problem can be formulated as a canonical multi-armed bandit (MAB), with an optimal solution in the form of Gittins' indices [10]. The MAB algorithm is valuable for our application because it does not require prior modeling or heuristic tuning, although several mild assumptions on the reward processes are needed. More broadly, the MAB structure is useful for any poorly known environment with position-dependent fading. In previous work, we showed that field implementation of a decision rule based on Gittins indices improved cumulative data transmission by 14% and 19% for two trials over a simple touring strategy [7]. However, 60% of the total mission time during these experiments was spent in transit between locations.

This transit fraction is important in practice because the performance of an acoustic link is usually degraded by propulsion and self-noise [11]. Further, the physical support of a relay modem, e.g., hanging from a pole on a ship, may require removal for transit. These considerations strongly motivate augmenting the MAB with switching costs (MABSC). No optimal index policy solution exists for the MABSC [12], and most research on this topic has focused on deriving properties of the optimal policy [13], developing special cases [14], and bounding approximations to the optimal policy [15]. For a recent survey, see Jun [16]. The MABSC question has also been studied via a semi-Markov multi-armed restless bandit, addressed by marginal productivity indices (MPI) [17] and a linear programming relaxation (LP) [18], based on work by Bertsimas and Niño-Mora [19]. These include switching costs as a natural extension of the restless bandit [20], in which processes are non-

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. Urbashi Mitra is with the University of Southern California, Los Angeles CA 90089, ubli at usc.edu. Hanumant Singh is with the Woods Hole Oceanographic Institution, 266 Woods Hole Road, Woods Hole MA 02543, hsingh at whoi.edu

stationary. However, both the MPI[1] and LP treatments of the restless bandit trade the advantages of an exact and general problem statement against a lookahead horizon limited to one step. The alternative — enumeration — incurs exponential computing cost, which is clearly undesirable.

In our approach, we exploit a small state space that allows for a moderately deep enumeration, but at the same time seek methods by which the computational load can be reduced. In particular, applying a key result of Asawa and Teneketzis [13] allows us to leverage the Gittins index policy for substantially reducing the frequency of enumeration. This paper builds on our prior MABSC work for acoustic relaying [21], [22]. The major contribution here is in exploring more fully the performance of our heuristic MABSC solution on both synthetic data and on new field data. These confirm robustness of the approach, and show that an enumeration horizon of three to five is suitable for this problem scale.

The paper is arranged as follows. In Section II we formulate the MABSC, and its application to adaptive acoustic relay positioning. In Section III, we describe field experiments with fully autonomous MABSC decision-making, and a large, new "hybrid" dataset specifically developed to allow a fair comparison of different algorithms. Channel statistics explicitly support the stationarity assumption. In Section IV, we assess the MABSC algorithm and compare the performance of MAB, MABSC, and ϵ-greedy methods, which are popular and practical competitors [23], [24].

II. Problem Formulation

The stochastic multi-armed bandit problem considers dynamic allocation of a single resource amongst several competing reward processes, in order to maximize expected total reward. We describe 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 the state is a real, non-negative random variable. The multi-armed bandit process is a collection of N independent one-armed bandit machines, indexed by i. We denote the number of times machine i has been operated by n_i, and its state by $x_i(t)$, 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 $\bar{x}(t)$, containing $\{x_1(t) \cdots x_N(t)\}$. 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.

[1]For a stationary process with switching costs, the MPI is equivalent to Asawa & Teneketzis's switching index [17]

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:

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

Gittins and Jones [10] showed that the optimal policy is to play the machine with the largest expected reward per unit discounted time, maximized over all stopping times $\tau > 1$: $i_{t+1} = \underset{i}{\operatorname{argmax}}(\nu_i(x_i(t)))$, where[2]

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

Index ν_i is a function only of $x_i(t)$, allowing the MAB to be decomposed into N independent stopping time problems. Various algorithms to calculate the Gittins index have been reported, recently by Sonin [25] and Niño-Mora [26].

B. A Heuristic Adaptation for Switching Costs

The MAB formulation assumes instantaneous and costless switching between reward processes; an assumption ill-suited to the physical relay application. We define constant 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 $c(i, j) = \lfloor t_v(i, j)/t_r(i, j) \rfloor$ — approximately 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 cost models relevant to the application, and later we investigate several. The optimal solution to the MABSC is one that maximizes:

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

where we define $i_{-1} = i_0$.

As noted previously, switching costs do not admit an index-type optimal policy [12]. For this problem, we describe a solution of the *priority-index policy* form, where separate "continuation" and "decision" indices are used [17]. At every decision epoch, the continuation index is computed to decide if the current arm is to be played again. If it is not continued, the decision index is then computed to decide which arm to switch to. The continuation index ν_i is taken to be the Gittins index previously defined. If the 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 showed that it is optimal to continue

[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.

playing an arm up to its stopping time τ, only making a decision to switch when the stopping time is achieved (A&T Thm. 3.1) [13]. This occurs when the Gittins index of the current arm falls below any value it has previously reached, thereby defining the efficient continuation rule:

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

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 policies include choosing to return to the current location after switching away. The value of being in a final state $\hat{x}_i$ is accounted for with an updated Gittins index ν_i for that policy, and location-based switching costs are included simply as:

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

where

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

The adapted decision rule for MABSC is then,

$$i_{t+1} = \underset{i_1}{\text{argmax}}(\eta_\pi(\bar{x}(t))), \quad (8)$$

where 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 & Teneketzis.

III. Field Experiments and Datasets

A. The Relay Positioning Scenario and Bernoulli Transmission Model

We consider a one-way, two-link acoustic transmission in the Charles River Basin (Boston, MA). A source modem at the MIT Sailing Pavilion broadcasts a data message, which is repeated by a mobile relay on a robotic 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; direct source-to-destination through-transmissions are possible but do not impact the relay's behavior. For learning and decision-making by the mobile acoustic relay within the MAB framework, we discretize the physical space into $N = 9$ potential relay locations and define each location as an independent arm of the bandit. The candidate locations were chosen in a grid pattern centered on the line between the source and destination nodes, where no prior knowledge of the channel was assumed (Fig. 1). 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 two-hop transmission made by the relay is naturally described by a Bernoulli trial:

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

A computational method for calculating indices for a Bernoulli reward process is described in Gittins [27]. The state vector comprises only n_i, the number of plays, and s_i, the number of successes: the index is thus $\nu_i(n_i, s_i)$, which can be stored in a lookup table. We note that programmable modem parameters such as packet encoding scheme could be included combinatorially as additional machines [28]; we have fixed these for simplicity.

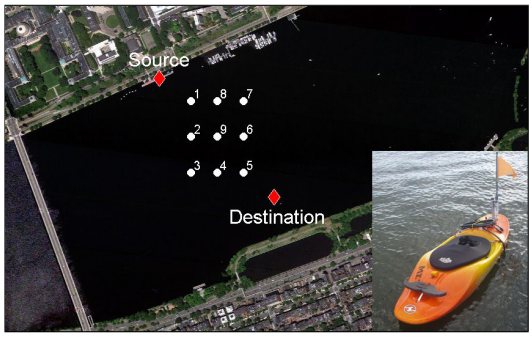

Fig. 1. Charles River Basin with autonomous surface vehicle inset. Relay locations are shown in white. Source and destination are shown in red.

B. Equipment and Operations

For all tests, Site 1 was designated as the relay's starting location. Each adaptive algorithm (MAB, heuristic MABSC, ϵ-greedy, and ϵ-decreasing) was initialized with the assumption of 100% packet success probability at every site[3]. The look-ahead horizon for policy enumeration was constrained by a maximum computation time of fifteen seconds[4], the time required for one two-hop transmission.

All field experiments were conducted with custom autonomous surface vehicles towing acoustic modem transducers at a fixed depth for underwater communications, with full benefits of GPS and WiFi connectivity for controlled experiments. The vehicles travel at 1.5m/s and maintain a station-keeping circle approximately 10m in diameter on location. We use WHOI Micro-Modems [6], 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"). Data was sent at PSK Rate 1, with a fixed message size of 192 bytes, We describe two major field experiments conducted with the MAB framework[5].

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

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

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

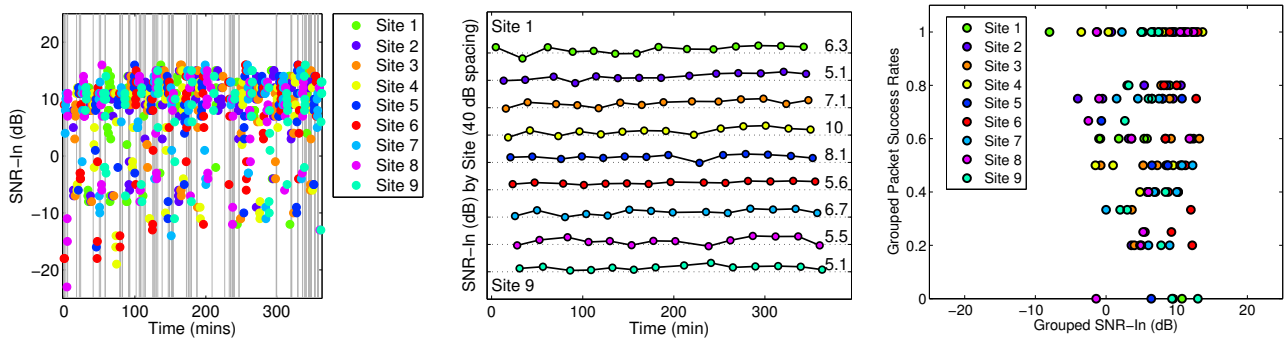

Fig. 2. Per-transmission SNR-In values over time (left), with lost packets shown in grey. Grouped SNR-In values over time (center), and Grouped Packet Success Rates against SNR-In (right). Data is for source to relay transmission only. Each averaging group consists of five transmissions.

C. Autonomous MABSC Dataset

We implemented the MABSC policy as a fully autonomous decision-maker for the mobile relay [22]. For acknowledgment of packet receipt at the destination back to the relay, we utilized the Micro-Modem frequency-shift-keying (FSK) Mini-packet, a 13-bit message with robust performance. Our tests have consistently shown Mini-packet loss rates of less than 5%. A touring survey consisting of a single circuit with ten transmissions at each site was performed immediately before the autonomous MABSC experiment to provide a comparison. The total mission time for each case was 53 minutes, with 90 observations for the survey and 100 for the MABSC.

D. "Hybrid" Tour Dataset

Fairly assessing the performance of competing algorithms online in acoustic positioning is difficult. Relays running competing algorithms would have to share the physical space and channel, experimental run times are on the order of hours, and weather conditions change daily. To address this, we constructed a "hybrid" scenario that systematically gathers data over a tour among the nine sites. We can then sample from this dataset with different algorithms, accounting for travel times by assuming constant speed. The mobile relay transmitted five times at each location per visit, and acknowledgments were communicated over WiFi for simplicity. A total of 1199 transmissions from source to relay and 1082 transmissions from relay to destination were sent, of which 754 packets were successful. The total experiment time was approximately 6.5 hours.

IV. Results

A. Acoustic Environment

Despite high SNR-In values, multi-path interference in the shallow Charles River makes packet decoding difficult and performance unpredictable. Altimetry data reveals highly irregular bottom topography [7]. As shown in Fig. 2 based on the hybrid tour data, the spread of SNR values is wide and there is no clear spatial or temporal structure. Remarkably, as seen from Fig. 2 (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. We note that this lack of correlation implies the modem is not operating with its optimal settings. Packet success rates at different sites show a narrow spread, with a maximum difference of 0.18: Site 4 was the best performing with a mean of 0.77 and Site 6 was the worst with a mean of 0.59. In comparison, a similar dataset collected on a different day showed a much higher spread in means, to a maximum of 0.43 [21]. As no clear trend in SNR-In values is discernible temporally or spatially, we assume the Bernoulli transmission processes are stationary over the time scale of this experiment.

B. Autonomous MABSC

Fig. 3 shows the cumulative performance (i.e. cumulative successful transmissions) of the autonomous MABSC and touring survey methods.

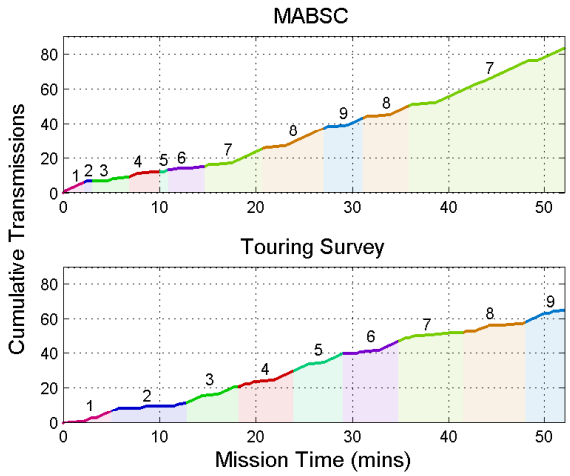

Fig. 3. Cumulative performance (successful transmissions) by mission time in minutes. The current location of the relay is color-coded by site and the site number is shown above the plot segment.

The top subplot demonstrates the behavior of the MABSC to quickly identify Sites 2 and 5 as poorly performing, and to allocate more time revisiting high-performing sites 7 and 8. Fig. 4 (left) shows the cumulative performance of the MABSC directly in comparison with the tour. During the first 30 minutes of the mission, performance is similar as the MABSC learns the field. For the remainder of the mission, however, the MABSC settles to a high-performing site, and achieves a final reward rate 77% higher, and an average reward rate 28% higher than the touring survey.

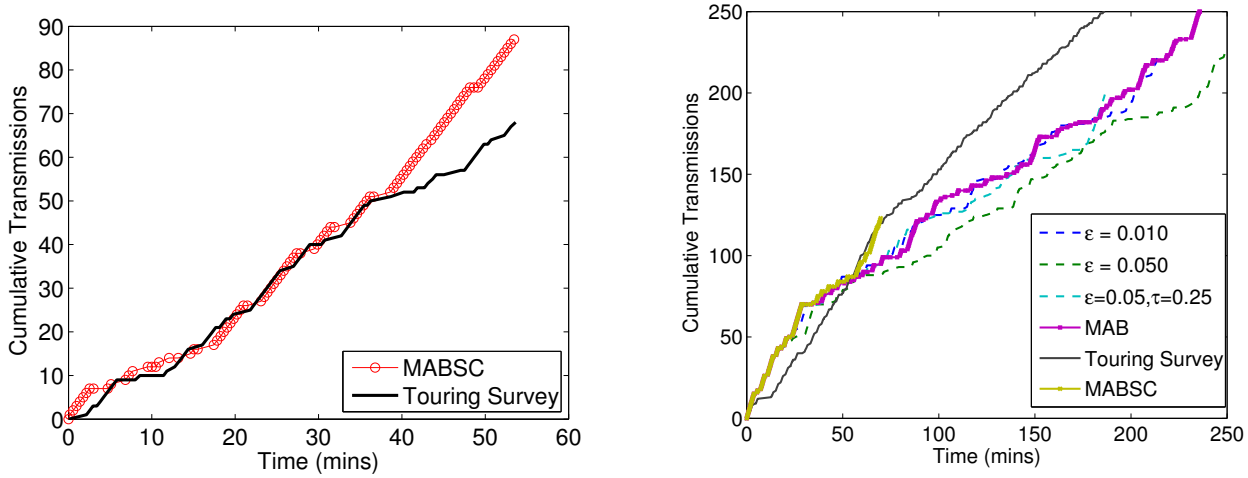

Fig. 4. Autonomy dataset cumulative performance (left) [22] and hybrid cumulative performance of MAB, MABSC and tuned ϵ-greedy and ϵ-decreasing algorithms (right) by mission time.

C. MABSC on Hybrid Tour Data

We now compare performance of the MAB, heuristic MABSC, ϵ-greedy, ϵ-decreasing algorithms, and the touring survey. Only the MABSC explicitly takes switching costs into account.

An ϵ-greedy algorithm plays the best arm $(1-\epsilon)$ of the time and switches to a random arm ϵ of the time. ϵ-decreasing is a variation on ϵ-greedy where the value of ϵ decreases in time. Kuleshov and Precup [24] have shown numerical simulation results suggesting these simple stochastic methods often outperform theoretically attractive approaches such as UCB (upper confidence bound), although the Gittins index rule was not included in the study. We tuned ϵ-dependent algorithms with $\epsilon = \{1\text{E-}3, 5\text{E-}3, 1\text{E-}2, 5\text{E-}2, 1\text{E-}1, 5\text{E-}1\}$ and $\tau = 0.1, 0.15, 0.2, 0.25, 0.3$; only the subset with the best performance is presented here for the sake of brevity.

Transmission data was sampled from the tour dataset in chronological order, terminating when unavailable data was requested. Since the number of physical transmissions available for each site is a constant, algorithms that spend more time at fewer sites exhaust the data more quickly.

In Fig. 4 (right), the touring survey initially performed worse than others, but then was exceptional up to about seventy-five minutes. The direct MAB formulation is competitive in real-time and its performance is significantly improved by the adaptation to switching costs. The MABSC algorithm i) found that the current node had the highest Gittins index 46% of the time (i.e. it is the best possible node), ii) made use of Asawa's theorem to continue without further computation 37% of the time, and iii) performed enumeration for 17% of decisions. At a computation horizon of six stages, enumeration takes 1.95% of total mission time, as compared to 7.8% without using the switching index and 11.3% if enumerating at every decision.

This assessment with field data augments the analysis developed in [21]. Taken together, we observe that in a switching cost scenario, decisions made by the MAB do not perform consistently better than a well-tuned greedy algorithm; the MABSC heuristic, however, is consistently the best.

D. MABSC with Synthetic Data

Using synthetic data allows us to further investigate the asymptotic performance of these algorithms with different probability spreads, enumeration horizons and switching cost models. Packet success probabilities were randomly assigned to the nine sites in two structures: a narrow range between 0.7 and 0.8, and a wide range between 0.1 and 0.9. We average the results of 100 trials, each trial consisting of 450 observations for each algorithm. Table I shows the average percentage improvement in reward rate (data transmitted per unit time) over a touring survey, for the MAB and MABSC with several look-ahead horizons, and for ϵ strategies. The touring survey takes five samples at each location, for a total of ten circuits. We use rate of reward for comparison as it factors in the time cost of switching. As in Fig. 4, only the tuned result for ϵ and τ is presented.

TABLE I
PERCENTAGE REWARD RATE IMPROVEMENT OVER TOURING SURVEY

Horizon	Narrow Range	Wide Range	Time per Enumeration (s)
ϵ-g	41.71	156.74	-
ϵ-d	54.14	174.96	-
MAB	60.97	183.95	-
MABSC, m=1	63.46	183.82	1.3e-04
MABSC, m=2	65.07	187.11	1.8e-03
MABSC, m=3	65.97	190.13	1.7e-02
MABSC, m=5	67.23	190.43	1.3e-00
MABSC, m=7	67.34	190.40	1.0e+02

The MAB improves significantly over simple ϵ strategies, and the MABSC provides further gains with apparently diminishing returns. The computation time required for a single decision enumeration is reported where applicable[6].

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

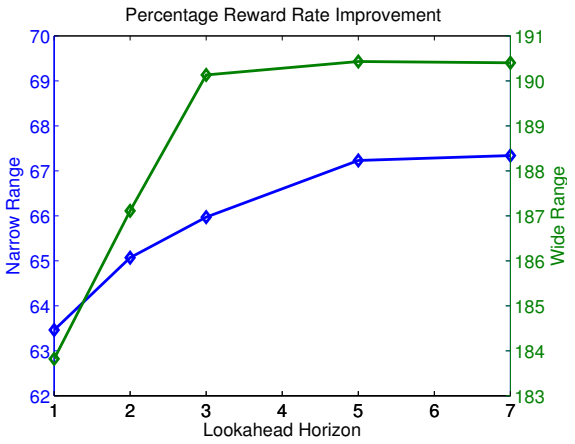

Fig. 5. Percentage improvement in reward rate over a touring survey for narrow and wide ranges of probabilities.

Fig. 5 shows the key result that longer enumeration horizons are very likely to be better than those which would be provided by the one-step lookahead MPI or LP restless bandit solutions, while a horizon of three to five represents a good tradeoff between computation time and gains.

We also consider three modifications of the previously described switching cost model based on travel time, and present results in Table II. A model with constant cost of one (i.e. transit cost is equal to one successful transmission irrespective of location) captures the effect of edge-independent switching costs. A normalized model scales nominal (edge-dependent) transit costs to be on a par with the reward, so that the average transit cost is equal to one successful transmission. Finally, an inflated model scales transit costs up ten times larger than the nominal value. Improvement was comparable across all cost models considered, confirming that the MABSC performs consistently well with a range of switching cost models and probability spreads.

TABLE II
PERCENTAGE REWARD RATE IMPROVEMENT OVER TOURING SURVEY FOR DIFFERENT SWITCHING COST MODELS, HORIZON THREE

Switching Cost	Narrow Range	Wide Range
Travel Time	65.97	190.13
Constant	64.78	188.80
Normalized	64.02	187.05
Inflated	68.35	188.37

E. Information

Finally, we recall that the MAB formalism optimally trades exploration and exploitation. The heuristic MABSC should similarly achieve successful exploitation through effective exploration. To assess this, we estimate each algorithm's information gain by computing the sum-of-squared error over all relay sites, where the error is taken between the algorithm's current estimate of a site's Bernoulli parameter, and the best possible estimate obtained from the entire dataset for that site. Fig. 6 shows the evolution of this error metric for each algorithm, using both the autonomous and hybrid datasets. As expected, the MAB and MABSC achieve a good field model in short time; they easily surpass learning from a tour and are competitive with the greedy methods. Interestingly, for our prior hybrid dataset [21], we found *no* greedy algorithms close to the learning rate of the MAB or MABSC, and additionally that the MAB was much better than the MABSC. As with the throughput performance discussed in Section IV-C of the present paper, these different outcomes across multiple experiments highlight the intrinsic challenges of underwater acoustic communication.

V. CONCLUSION

The multi-armed bandit is a powerful framework for addressing the poorly-known coupling between channel properties and physical location, as observed in acoustic relay positioning. We augmented the canonical MAB to include switching costs based on proven properties of the optimal solution, and have demonstrated in the field and using synthetic data that the MABSC scheme can achieve more throughput than can a tuned greedy method. Concerning partial enumeration as a practical approach, there is strong evidence in our immediate application for near-optimality with a horizon of three to five; this is entirely tractable from a computational perspective. These attributes should make the MABSC useful to practitioners seeking enhanced network performance underwater and elsewhere.

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] 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.

[7] 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.

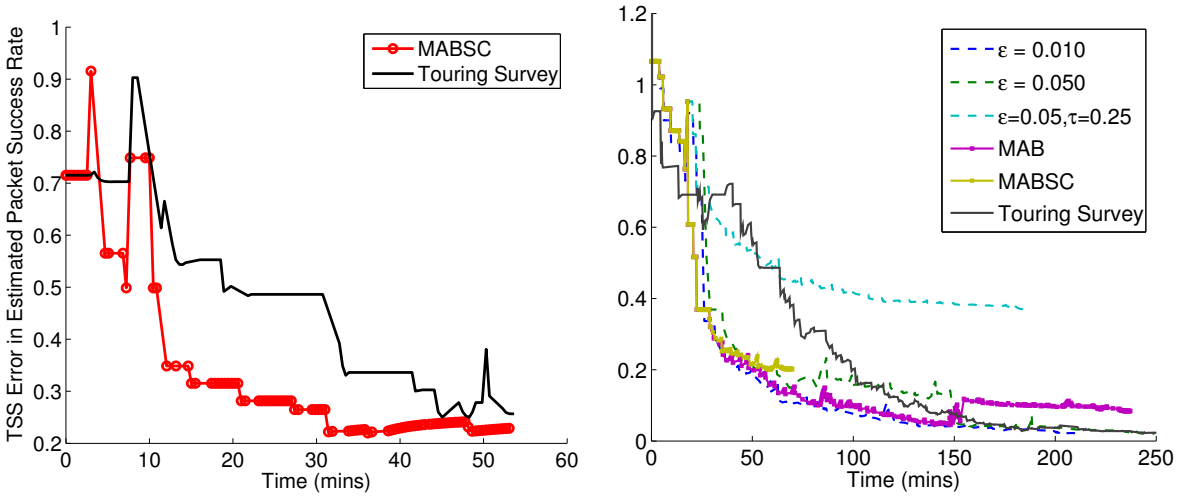

Fig. 6. Autonomous [22] and hybrid datasets total sum of squared differences (TSS) over time, summed over all sites.

[8] 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.

[9] 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.

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

[11] 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.

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

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

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

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

[21] 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.

[22] M. Cheung, J. Leighton, and F. Hover. A multi-armed bandit with switching costs for autonomous acoustic relay positioning. *Proc. International Symposium on Unmanned Untethered Submersible Technology (UUST)*, 2013.

[23] J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. pages 437–448. Springer, 2005.

[24] V. Kuleshov and D. Precup. Algorithms for the multi-armed bandit problem. *Journal of Machine Learning*, 1:1—48, 2010.

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

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

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

[28] S. Shankar and M. Chitre. Tuning an underwater communication link. In *OCEANS-Bergen, 2013 MTS/IEEE*, pages 1–9. IEEE, 2013.