Optimal Query Selection Using Multi-Armed Bandits

Aziz Koçanaoğulları, Yeganeh M. Marghi, Murat Akçakaya, and Deniz Erdoğmuş

***Abstract*—Query selection for latent variable estimation is conventionally performed by opting for observations with low noise or optimizing information-theoretic objectives related to reducing the level of estimated uncertainty based on the current best estimate. In these approaches, typically, the system makes a decision by leveraging the current available information about the state. However, trusting the current best estimate results in poor query selection when truth is far from the current estimate, and this negatively impacts the speed and accuracy of the latent variable estimation procedure. We introduce a novel sequential adaptive action value function for query selection using the multi-armed bandit framework, which allows us to find a tractable solution. For this adaptive-sequential query selection method, we analytically show: 1) performance improvement in the query selection for a dynamical system; and 2) the conditions where the model outperforms competitors. We also present favorable empirical assessments of the performance for this method, compared to alternative methods, both using Monte Carlo simulations and human-in-the-loop experiments with a brain–computer interface typing system, where the language model provides the prior information.**

***Index Terms*—Subset selection, query optimization, misleading prior, multi-armed bandit framework.**

I. Introduction

RECURSIVE state estimation contributes a key role in signal processing and system identification. The recursive paradigm is often used to extract information from model parameters or the states of a dynamic system in real time, given noisy observations. On the other hand, Bayesian methods are valuable decision making approaches, since they take into account a variety of prior knowledge about the system due to the experience and previous observations (history). In stochastic dynamic systems, to estimate the state variables, maximum a posteriori (MAP) inference is commonly used. To estimate the state with a high confidence (usually pre-defined), the system probes the environment through multiple recursions of sequences of queries, which reduces the rate of state estimation convergence. Therefore, the queries need to be designed specifically to optimize both the speed and the accuracy of the state estimation. Query selection/optimization in the recursive state estimation is often performed by greedy selections using: (i) expected posterior maximization [1]; (ii) Fisher information-based approaches [2], [3], and (iii) information theory-based approaches such as entropy minimization or maximum mutual information (MMI) [4]–[7]. It is shown that all these approaches for optimum sequence design through query selection lead to the selection of *N-best* queries with respect to the current belief [8], [9].

In estimation problems, the system may have access to an additional knowledge called prior information on the state of the system which can improve the estimation process. However, imprecise prior information may lead to incorrect posterior beliefs given the same set of observations. Accordingly, choosing the N-best queries by trusting the current belief does not always offer the best query optimization. In dynamic systems, the prior information about the environment can be adversarial due to the transition noise, observation noise, change of environment distributions and being outdated; hence resulting in longer decision cycles or the wrong state estimation [10], [11]. Many applications involving state estimation, system identification or sequential decision making using prior information encounter these challenges: recommender systems [12], communication networks [13], [14], radar systems [15] and clinical studies [16], [17]. The common problem in all of these applications is that the misleading information can extremely impact the final decision or estimation. Another category of methods to overcome the misleading information, is variance based methods [15], [18], [19] that can be extended using Fisher information [20] to either explore or exploit. The main drawback of these methods is that they only commit to either explore or exploit for the query selection, which leads to the same solution provided by the *N-best* method [8], [9].

In this letter, we propose an information theoretic query selection to discard the ambiguity (exploitation) in the state estimation while also measuring the credibility of the prior information (exploration). The proposed objective function is a linear combination of exploration and exploitation. Moreover, we reformulate the query selection as a multi-armed bandit (MAB) problem. We denote this framework as MAB based on State-Measurement-MI and State-Posterior-Momentum for RBSE. The MAB framework is a well-studied approach to formulate the learning process under available observations

Manuscript received August 10, 2018; revised October 13, 2018; accepted October 18, 2018. Date of publication October 26, 2018; date of current version November 7, 2018. The work of A. Koçanaoğulları, Y. M. Marghi, and D. Erdoğmuş was supported in part by the National Science Foundation (NSF) under Grant IIS-1149570 and Grant CNS-1544895, in part by the National Institute on Disability, Independent Living, and Rehabilitation Research under Grant 90RE5017-02-01, and in part by the National Institutes of Health under Grant R01DC009834. The work of M. Akçakaya was supported in part by the NSF under Grant IIS-1717654 and in part by the Air Force Office of Scientific Research through the Dynamic Data Driven Applications Systems Program under Grant FA9550-16-1-0386. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mario Huemer. *(Aziz Koçanaoğulları and Yeganeh M. Marghi contributed equally to this work.) (Corresponding author: Aziz Koçanaoğulları.)*

A. Koçanaoğulları, Y. M. Marghi, and D. Erdoğmuş are with Northeastern University, Boston, MA 02115 USA (e-mail: akocanaogullari@ece.neu.edu; ymmarghi@ece.neu.edu; erdogmus@ece.neu.edu).

M. Akçakaya is with the University of Pittsburgh, Pittsburgh, PA 15260 USA (e-mail: akcakaya@pitt.edu).

This letter has supplementary downloadable material available at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2878066

[21]–[24]. MAB has been proposed for decision making, predictive entropy search for sequential action selection and estimation applications [25]–[28], [29], [30]. Such applications consider sequential selection based on MAB, in which the approaches can be considered to repeatedly make choices among elements of a finite set of state elements [25], [28]. Such a formulation enables us to analytically demonstrate that the proposed policy for query selection containing exploration and exploitation performs at least as good as the methods that only rely on the exploitation of the current belief. MAB framework also enables the formulation of subset query selection optimization as a tractable problem especially through sequential selection with theoretical guarantees [31], [21].

The novel contributions of this letter can be summarized in: (i) introducing a new action-value function for query selection using changes in the posterior to encourage exploration in the MAB setting, (ii) providing a short-term policy evaluation to demonstrate that the proposed method has theoretical guarantees under certain assumptions, and (iii) evaluating the proposed method in an actual human-in-the-loop typing scheme employing a language-model-assisted Electroencephalogram (EEG)-based BCI typing system called RSVP Keyboard. Because of space limitations, we present the proofs of analytical propositions in the supplementary material. The system code is under revision and the current version can be accessed at https://github.com/BciPy [32].

II. Preliminaries

In the framework of the state estimation problem, We refer σ as the (unknown) state which is an element of a finite set $\mathcal{A}$. The system (learner) proceeds with the estimation through a sequential decision making process containing *sequences* indexed by s of multiple trials indexed by i. We denote a list of variables with $\{:\}$, for example $\Phi_{0:s}$ represents a sequence of variables from 0 to s. The query and evidence sets at sequence s are denoted by $\Phi_s \triangleq \{\phi_s^1, \ldots, \phi_s^K\}$ and $\varepsilon_s \triangleq \{\varepsilon_s^1, \ldots, \varepsilon_s^K\}$, respectively. Here, $K \in \mathbb{N}$ denotes number of trials in a sequence. We use the query class definition presented in [8] and assume each σ has a corresponding query defined with the class conditional representation. Therefore, without loss of generality, it is assumed that all of the observations are noisy and comes from two unimodal probability distributions conditioned on state and query tuples. Assuming all trials are independent and the current observation is only function of the current query and independent of the task history, $\mathcal{H}_s \triangleq \{\varepsilon_{1:s}, \Phi_{1:s}\}$, the posterior probability at time s can be expressed as:

$$p(\sigma|\varepsilon_s, \Phi_s, \mathcal{H}_{s-1}) = p(\sigma|\mathcal{H}_0) \prod_{j=1}^{s} \prod_{i=1}^{t_i} \frac{p(\varepsilon_j^i|\sigma, \phi_j^i)}{p(\varepsilon_j^i|\phi_j^i)}$$

where $p(\sigma|\mathcal{H}_0)$ is a prior information. Using maximum a posteriori (MAP) estimation [16] the learner attempts to estimate σ with re-occuring evidence collection. Based on the collected evidence if a decision is not possible, the system decides on a subset of queries for the upcoming sequence to improve its confidence. Accordingly, the query selection process is formulated by the following optimization:

$$\Phi_s = \arg\max_{\Phi} q_s(\varepsilon_{1:s-1}, \Phi_{1:s-1}, \Phi) \tag{1}$$

where q_s denotes the objective term, which we call it action-value function. Following querying, evidence ε_s is observed and accordingly the posterior is updated. In next section we propose an action value function that balances exploration and exploitation.

III. Method and Analysis

By imposing the MAB settings to the context of the state estimation problem, each query can be represented as an arm of a MAB. This reformulation allows us to solve the subset selection problem through a greedy approach by optimizing the action value function for each arm with theoretical guarantees [31]. Therefore, selection of arms (queries in each sequence) with highest action value one by one allows us to form the query in a computationally efficient way. In MAB formulation, for the design of upcoming sequence, we assume that multiple arms are pulled according to the state posterior probability that depends on the task history. The goal is then to define an objective that specifies the subset of queries (arms) to be picked at each step. Independency between trails allows us to perform optimization per trial at each sequence. Therefore, set optimization in (1) reduces to single query selection.

Conventionally, query selection is achieved through MMI [5], [6], which is equivalent to entropy minimization when the evidence ε_s corresponding to the sequence being designed is not observed; and hence, σ is independent from ϕ_s. Query selection using mutual information can be written as the following policy:

$$\begin{aligned} \phi_s^i &= \arg\max_{\phi} I(\sigma, \varepsilon_s^i|\phi, \mathcal{H}_{s-1}) \\ &= \arg\max_{\phi} -H(\sigma|\varepsilon_s^i, \phi, \mathcal{H}_{s-1}) \end{aligned} \tag{2}$$

In this letter, we consider three different action-value functions for the MAB formulation: (i) mutual information objective (2) (ii) history-based objective (iii) combination of (i) and (ii).

We introduce a new term called *Momentum* that is function of posterior changes across sequences, such that:

$$\begin{aligned} m(\phi|\mathcal{H}_j) = \mathbf{E}_{p(\sigma|\mathcal{H}_{j-1})} \big[&\log p(\sigma|\varepsilon_j^i, \phi, \mathcal{H}_{j-1}) \\ &- \log p(\sigma|\mathcal{H}_{j-1}) \big] \mathbb{1}_{\phi}(\sigma) \end{aligned} \tag{3}$$

where $\mathbb{1}_{\phi}(\sigma)$ denotes the indicator function which equates 1 if $\phi = \sigma$. Since $m(\phi|\mathcal{H}_j)$ is the summation of probability displacement multiplied by the probability mass along axes of the state space, we call it Momentum. Additionally, $m(\phi|\mathcal{H}_0) = 0$, $\forall \phi$, in words, without collecting any evidence we can not infer the trend of a particular state in estimation. Accordingly, for the history-based approach, the objective is defined as the average of Momentum as follows.

$$M(\phi|\mathcal{H}_{s-1}) = \frac{1}{s-1} \sum_{j=1}^{s-1} m(\phi|\mathcal{H}_j)\, \mathbb{1}_{\Phi_j}(\phi) \tag{4}$$

For this approach, the query selection policy becomes:

$$\phi_s^i = \arg\max_{\phi} M(\phi|\mathcal{H}_{s-1}) \tag{5}$$

We present a new action-value function for the query selection based on combination of mutual information and history-based objectives in (4) to balance exploration and exploitation. Accordingly, the action-value function and policy

can be defined as:

$$q_s^i(\phi) = I(\sigma, \varepsilon_s^i|\phi, \mathcal{H}_{s-1}) + \lambda M(\phi|\mathcal{H}_{s-1}), \quad \lambda \geq 0 \tag{6}$$

$$\begin{aligned}\phi_s^i &= \arg\max_{\phi} q_s(\phi) \\ &= \arg\max_{\phi} -H(\sigma|\varepsilon_s^i, \phi, \mathcal{H}_{s-1}) + \lambda M(\phi|\mathcal{H}_{s-1})\end{aligned} \tag{7}$$

where, λ is a tuning parameter that balances MMI and Momentum-based policies. The objective function (7) can be written as (8) by replacing the entropy term as shown in our previous work [8].

$$\begin{aligned}\phi_s^i = \arg\max_{\phi} &\mathbf{E}_{p(\sigma|\mathcal{H}_{s-1})}\mathbf{E}_{p(\varepsilon_s^i|\sigma,\phi)}\left[\log p(\varepsilon_s^i|\sigma,\phi)\right. \\ &\left. - \log p(\varepsilon_s^i|\phi)\right] + \frac{\lambda}{s-1}\sum_{j=1}^{s-1} m(\phi|\mathcal{H}_j)\mathbb{1}_{\Phi_j}(\phi)\end{aligned} \tag{8}$$

For the same given task history, we show that when policy in (8) is used in MAB formulation, the target state has higher probability to be chosen to appear in the query subset compared to other policies in (2) and (5). Here we propose the analysis of the correctness of the statement. To save space we use $I(a, \varepsilon_s^i|\phi = a, \mathcal{H}_{s-1}) = I_s(a)$ and $M(\phi = b|\mathcal{H}_{s-1}) = M_s(a)$ notations for the following lemmas.

Lemma 1: Given $a, b \in \mathcal{A}$ where $a \neq b$ and $\lambda \geq 0$, if $\exists\, \mathcal{H}_{s-1}$ s.t. $p(a|\mathcal{H}_{s-1}) < p(b|\mathcal{H}_{s-1})$, then

$$p\left(I_s(a) + \lambda M_s(a) > I_s(b) + \lambda M_s(b)\right) \geq p\left(I_s(a) > I_s(b)\right)$$

Lemma 1 shows that the probability of a (assuming to be the target state) having a higher action-value than b is larger when policy (8) is used instead of (2). Although the probability of a given, the task history is lower than the probability of b given the task history. This means that even if a is less likely than b according to the prior information and observations, using the proposed policy, a has more chance to appear in the query subset compared to policy in (2).

Lemma 2: Given $a, b \in \mathcal{A}$ where $a \neq b$ and $\lambda \geq 0$, if $\exists\, \mathcal{H}_{s-1}$ s.t. $p(a|\mathcal{H}_{s-1}) > p(b|\mathcal{H}_{s-1})$, then

$$p(I_s(a) + \lambda M_s(a) > I_s(b) + \lambda M_s(b)) \geq p\left(M_s(a) > M_s(b)\right)$$

Lemma 2 represents the case where the prior knowledge is supporting a rather than being adversarial. It shows that compared to using (4) as the action-value function, when a (unknown state) has higher probability given the task history compared to b, using action value function (6) has higher probability to choose a. These two lemmas show that the proposed query selection policy provides a balance between (2) and (5), and accordingly between the adversarial and supporting prior information. This balance is achieved through the selection of λ. Detailed proofs of Lemmas are provided in the Supplementary Materials.

The query λ should be updated dynamically such that the emphasis on mutual information component of the proposed function is increased with the number of sequences; i.e., the λ value should be decreased as the number of sequences is increasing [33]. We introduce a theorem that defines upper and lower bounds for the λ value to satisfy to guarantee that including the target state in the selected subset has higher probability when the proposed policy in (8) is used in MAB formulation compared to other policies (2) and (5).

Theorem 1: Let $\sigma \in \mathcal{A}$ be the target state and $|\mathcal{A}|$ represent the size of the finite set $\mathcal{A}$. *Consider three query selection policies as follows*:

$$\phi_{i,s}^{\pi_1} = \arg\max_{\phi}\ I(\sigma, \varepsilon_s^i|\phi, \mathcal{H}_{s-1})$$

$$\phi_{i,s}^{\pi_2} = \arg\max_{\phi}\ M(\phi|\mathcal{H}_{s-1})$$

$$\phi_{i,s}^{\pi_3} = \arg\max_{\phi}\ I(\sigma, \varepsilon_s^i|\phi, \mathcal{H}_{s-1}) + \lambda M(\phi|\mathcal{H}_{s-1}).$$

If $p(\phi^{\pi_3} = \sigma) \geq p(\phi^{\pi_i} = \sigma)$ for $i = 1, 2$, then $\exists\lambda$ that satisfies

$$\begin{aligned}&\frac{2(s-1)|\mathcal{A}|\left(d^2_{(p_s^{\phi^{\pi_1}},U)} - d^2_{(p_s^{\phi^{\pi_3}},U)}\right)}{\sum_{j=1}^{s-1}\left[d^2_{p_{i,j}^{\phi^{\pi_3}}} - d^2_{p_{i,j}^{\phi^{\pi_1}}}\right]\mathbb{1}_{\phi}(\sigma)} \leq \lambda \\ &\leq \frac{(s-1)\left(d^2_{(p_s^{\pi_3},U)} - d^2_{(p_s^{\pi_2},U)}\right)}{2\sum_{j=1}^{s-1}\left[\log\left(\frac{p(\sigma|\varepsilon_j^i, \phi^{\pi_2}, \mathcal{H}_{j-1})}{p(\sigma|\varepsilon_j^i, \phi^{\pi_3}, \mathcal{H}_{j-1})}\right)\right]}\end{aligned} \tag{9}$$

where

$$d^2_{(p_{i,s}^{\pi},U)} = \sum_{a\in\mathcal{A}}\left(p(a|\varepsilon_s^i, \phi^{\pi}, \mathcal{H}_{s-1}) - \frac{1}{|\mathcal{A}|}\right)^2$$

$$d^2_{p_{i,s}^{\pi}} = \left|p(\sigma|\mathcal{H}_{s-1}) - p(\sigma|\varepsilon_s^i, \phi^{\pi}, \mathcal{H}_{s-1})\right|^2.$$

This theorem shows the existence of the parameter λ makes the joint objective optimal. The proof of this theorem, which uses Pinsker's Theorem [34] and the results of Lemmas 1 and 2 can be found in the Supplementary Materials.

IV. Results

To assess the performance of the proposed query selection method, an language-model-assisted EEG-based BCI typing system called RSVP Keyboard [35] has been used as a test framework. Ten healthy participants (six females and four males), 20–35 years old were recruited under IRB-130107 protocol approved by Northeastern University. A DSI-24 Wearable Sensing EEG Headset was used for data acquisition, at a sampling rate of 300 Hz with active dry electrodes. EEG signals were acquired from 20 sensors according to International 10-20 System locations: Fp1, Fp2, Fz, F3, F4, F7, F8, Cz, C3, C4, T3, T4, T5, T6, P3, P4, O1, O2, A1 and A2. All participants were asked to perform two sessions including *calibration* and *Copy Phrase*. During calibration, the users were asked to attend to predefined target symbols within randomly ordered sequences to enable the system learn the class conditional EEG evidence distributions. Here, calibration session contains 100 sequences; each sequence includes five trials (letters); and one trial in each sequence is the target symbol which is displayed on the screen prior to each sequence. The time interval between trials is 200 ms. Optimal parameters for both target and non-target class distributions were learned using the calibration data which are used in Copy Phrase task. In Copy Phrase, participants were tasked to write a missing word in a pre-defined phrase using the system (a total of 6 words with various difficulties based on LM were typed). In addition to Copy Phrase, we

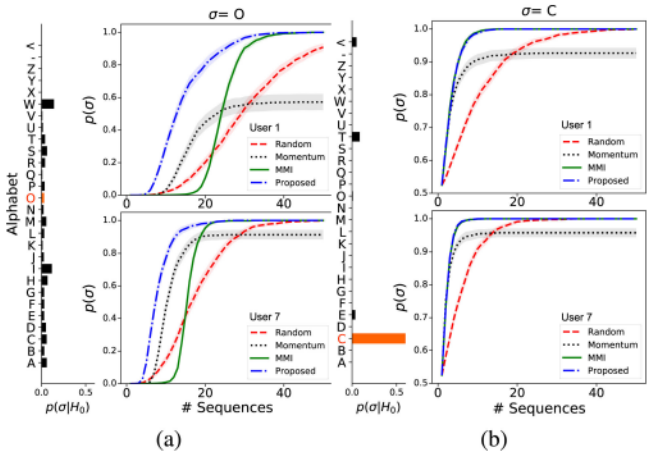

Fig. 1. Probability of the letter completion for 500 Monte-Carlo simulations for typing three target symbols in phrase '*OCCURRED*'. Intended symbols contain: (a)'*O*' and (b)'*C*' in the target phrase. Simulation results are reported for two users with different calibration performances. User 1 with AUC = 0.67 has lower performance than user 7 with AUC = 0.82. Bar plots show the LM prior probability over all typing symbols.

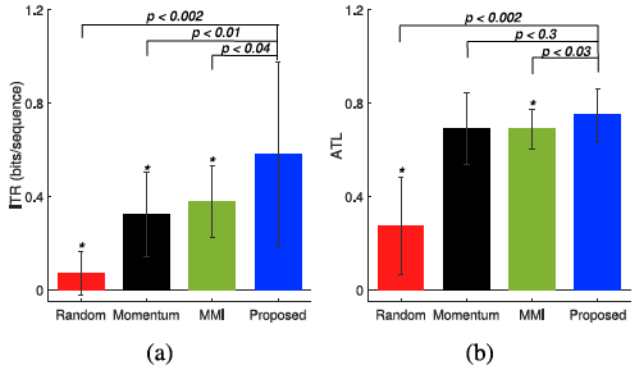

Fig. 2. Average of information transfer rate for four query selection methods. All of the results belongs to 10 users attending the copy phrase task in RSVP Keyboard experiment. p corresponds to the Wilcoxon signed-rank test.

also use the calibration data from each participant for BCI performance simulation. We present both simulation and real-time experiment results.

To evaluate the empirical performance of the proposed query selection we first utilized simulation. For our simulations we used conditional evidence distributions which are learned in the calibration session. More details about the simulation framework can be found in [35] study. During simulations, through Monte Carlo simulations samples from these conditional distributions were drawn to type '*O*' and '*C*' for each simulation in the phrase '*IT_OCCURRED_RANDOMLY*'. These letters are chosen because they have different difficulties to be typed based on the language model (prior information). The number of Monte-Carlo simulations is chosen to be 500. Fig. 1 shows the simulation results. Fig. 1 presents the typing performance for two users with different calibration performance which is quantified by the area under the receiver operating characteristics curve (AUC); $\text{AUC}_{u_1} = 0.82$ and $\text{AUC}_{u_7} = 0.67$. The bar plot next to each plot shows the prior information provided by the language model (LM) at the beginning of a decision cycle. For instance, it can be seen that the LM probability for '*O*' is very low and it is not quite likely to start a word with this letter. Accordingly, MMI method highly influenced by the LM prior, needs more sequences to estimate the target letter. In the early sequences of the decision process, the Momentum-based approach on average is faster than MMI to pick the intended letter for the query subset. Although, due to noise in the EEG evidence and miss-classification of observations, Momentum gets close to zero and does not pick the intended state for the query subset. Overall, the proposed method outperforms the other methods. However, when there is a likely letter like '*C*', MMI and the proposed method perform similarly. By comparing the simulation results of user 1 (lower AUC) with user 7 (higher AUC), it can be seen that all of the query selections are faster for the user 7 with higher calibration performance. It can also be observed that for user 1 even when there is more overlap between class conditional distributions (because of low AUC), the proposed method can estimate the target letter quite fast. On the other hand, it is more difficult for Momentum-based approach to capture the target letter. Moreover, MMI also requires more number of sequences.

As described above, the participants also attended Copy Phrase sessions after the calibration. Each participant attended four Copy Phrase tasks with different query selection methods. The order of the tasks were randomly assigned for each participant to avoid the learning impact.

Fig. 2 shows the average performance of all the query selection methods for all users including a statistical test results for Copy Phrase sessions. We reported the query selection performance in terms of two measures: accuracy in typing a letter correctly (ATL) that is the total number of correctly typed letters divided by the total number of typed letters and the information transfer rate (ITR) [36]. ITR summarizes the accuracy and speed into a single metric and it is commonly used to measure BCI performance. These results show that the proposed method outperforms the other methods both in terms of speed and accuracy. We used the Wilcoxon signed-rank test as a non-parametric statistical hypothesis test to perform a paired-comparison between the proposed method and the other query selection methods. The proposed method significantly enhanced the ITR compared to the other methods. Our statistical analysis also shows that the proposed method significantly improved the ATL compared to MMI and random query selection.

V. Conclusion

Being motivated by the MAB framework, we proposed a tractable solution to the subset query optimization for recursive Bayesian state estimation to enhance the estimation speed and accuracy. More specifically, a new action-value function was introduced for query selection, which uses a linear combination of mutual information and a momentum term which is a function of logarithmic changes of the posterior probability across sequences. We have also presented a bound for the action-value tuning parameter, which guarantees that the proposed query selection policy outperforms the others. An BCI typing system has been used as a test framework to assess the performance of the proposed method. Our results for both simulation and the human-in-the-loop experiment showed that the proposed method outperforms the alternative methods as shown by analytical results.

REFERENCES

[1] A. Wilson, A. Fern, and P. Tadepalli, "A Bayesian approach for policy learning from trajectory preference queries," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1133–1141.

[2] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 417–424.

[3] S. P. Chepuri and G. Leus, "Sparsity-promoting sensor selection for non-linear measurement models," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 684–698, Feb. 2015.

[4] D. Golovin, A. Krause, and D. Ray, "Near-optimal Bayesian active learning with noisy observations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 766–774.

[5] M. Higger *et al.*, "Recursive Bayesian coding for BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 704–714, Jun. 2017.

[6] B. Jedynak, P. I. Frazier, and R. Sznitman, "Twenty questions with noise: Bayes optimal policies for entropy loss," *J. Appl. Probab.*, vol. 49, no. 1, pp. 114–136, 2012.

[7] M. Moghadamfalahi, M. Akcakaya, H. Nezamfar, J. Sourati, and D. Erdogmus, "An active RBSE framework to generate optimal stimulus sequences in a BCI for spelling," *IEEE Trans. Signal Process.*, vol. 65, no. 20, pp. 5381–5392, Oct. 2017.

[8] A. Koçanaoğulları, M. Akçakaya, and D. Erdogmus, "On analysis of active querying for recursive state estimation," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 743–747, Jun. 2018.

[9] T. Tsiligkaridis, B. M. Sadler, and A. O. Hero, "Collaborative 20 questions for target localization," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2233–2252, Apr. 2014.

[10] S. Ungarala, E. Dolence, and K. Li, "Constrained extended kalman filter for nonlinear state estimation," *IFAC Proc. Vol.*, vol. 40, no. 5, pp. 63–68, 2007.

[11] R. Schneider and C. Georgakis, "How to not make the extended Kalman filter fail," *Ind. Eng. Chem. Res.*, vol. 52, no. 9, pp. 3354–3362, 2013.

[12] M. Quadrana, P. Cremonesi, and D. Jannach, "Sequence-aware recommender systems," 2018, arXiv:1802.08452.

[13] S. Haykin, K. Huber, and Z. Chen, "Bayesian sequential state estimation for MIMO wireless communications," *Proc. IEEE*, vol. 92, no. 3, pp. 439–454, Mar. 2004.

[14] T. Zhao and A. Nehorai, "Distributed sequential Bayesian estimation of a diffusive source in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1511–1524, Apr. 2007.

[15] P. Woodward, *Probability and Information Theory, With Applications to Radar* (International Series of Monographs on Electronics and Instrumentation), vol. 3. New York, NY, USA: Pergamon, 2014.

[16] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[17] H. G. Schmidt, M. L. De Volder, W. S. De Grave, J. H. Moust, and V. L. Patel, "Explanatory models in the processing of science text: The role of prior knowledge activation through small-group discussion," *J. Educ. Psychol.*, vol. 81, no. 4, pp. 610–619, 1989.

[18] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han, "A Bayesian approach to discovering truth from conflicting sources for data integration," *Proc. VLDB Endowment*, vol. 5, no. 6, pp. 550–561, 2012.

[19] B. Zhao and J. Han, "A probabilistic model for estimating real-valued truth from conflicting sources," *Proc. 10th Int. Workshop Qual. Databases*, 2012.

[20] G. Lidoris, D. Wollherr, and M. Buss, "Bayesian state estimation and behavior selection for autonomous robotic exploration in dynamic environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2008, pp. 1299–1306.

[21] Y. Baram, R. E. Yaniv, and K. Luz, "Online choice of active learning algorithms," *J. Mach. Learn. Res.*, vol. 5, no. Mar. pp. 255–291, 2004.

[22] T. Kocák, G. Neu, M. Valko, and R. Munos, "Efficient learning by implicit exploration in bandit problems with side observations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 613–621.

[23] H.-M. Chu and H.-T. Lin, "Can active learning experience be transferred?" in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 841–846.

[24] T. Lin, J. Li, and W. Chen, "Stochastic online greedy learning with semi-bandit feedbacks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 352–360.

[25] A. Lazaric *et al.*, "Sequential transfer in multi-armed bandit with finite set of models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2220–2228.

[26] D. E. Acuña and P. Schrater, "Structure learning in human sequential decision-making," *PLoS Comput. Biol.*, vol. 6, 2010, Art. no. e1001003.

[27] C. D. Rosin, "Multi-armed bandits with episode context," *Ann. Math. Artif. Intell.*, vol. 61, no. 3, pp. 203–230, 2011.

[28] R. McInerney, S. Roberts, and I. Rezek, "Sequential Bayesian decision making for multi-armed bandit," in *Proc. 5th Workshop Multi-Agent Sequential Decis. Making Uncertain Domains*, Toronto, ON, Canada, 2010, p. 38.

[29] E. Wang, H. Kurniawati, and D. P. Kroese, "CEMAB: A cross-entropy-based method for large-scale multi-armed bandits," in *Proc. Australas. Conf. Artif. Life Comput. Intell.*, 2017, pp. 353–365.

[30] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani, "Predictive entropy search for efficient global optimization of black-box functions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 918–926.

[31] V. F. Farias and R. Madan, "The irrevocable multiarmed bandit problem," *Oper. Res.*, vol. 59, no. 2, pp. 383–399, 2011.

[32] T. Memmott *et al.*, "BciPy: A python framework for brain-computer interface research," in *Proc. 7th Int. BCI Meeting*, Asilomar, CA, USA, 2018, pp. 183–184.

[33] J. A. Bilmes, "Dynamic Bayesian multinets," in *Proc. 16th Conf. Uncertainty Artif. Intell.*, 2000, pp. 38–45.

[34] A. A. Fedotov, P. Harremoës, and F. Topsoe, "Refinements of Pinsker's inequality," *IEEE Trans. Inf. Theory*, vol. 49, no. 6, pp. 1491–1498, Jun. 2003.

[35] U. Orhan *et al.*, "Probabilistic simulation framework for EEG-based BCI design," *Brain-Comput. Interfaces*, vol. 3, no. 4, pp. 171–185, 2016.

[36] B. Obermaier, C. Neuper, C. Guger, and G. Pfurtscheller, "Information transfer rate in a five-classes brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 9, no. 3, pp. 283–288, Sep. 2001.