ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414772

ACTIVE ESTIMATION FROM MULTIMODAL DATA

Arpan Mukherjee and Ali Tajer

Rensselaer Polytechnic Institute

Pin-Yu Chen and Payel Das

IBM Thomas J. Watson Research Center

ABSTRACT

The paper considers the problem of estimating a covariate parameter shared by multiple statistical models. Under the objective of estimating the parameter with target reliability with the fewest number of samples from these models, a fundamental question is how to glean samples from the statistical models. This question is especially important when the models are not equally descriptive or informative about the parameter, each being the most informative only for a specific regime of the parameter. This paper provides 1) an active sampling framework that specifies how the samples should be collected from different models over time in a data-adaptive fashion; 2) a stopping criterion specifying when the collected data is informative enough to form a reliable estimate for the covariate parameter; and 3) a terminal estimation rule. These rules, collectively, are shown to admit certain optimality guarantees. Numerical evaluations are provided to compare the performance with relevant existing approaches.

Index Terms— Active sampling, sequential estimation.

1. INTRODUCTION

Consider the canonical estimation problem in which a covariate parameter $\theta \in \Theta$ is embedded in two data streams with probability measures $\mathbb{P}_1(\cdot \mid \theta)$ and $\mathbb{P}_2(\cdot \mid \theta)$. Assume that we have the freedom to collect data from any of these two models in order to estimate θ. A fundamental question pertains to which of the two data streams should we sample from in order to form a sufficiently reliable estimate of θ with the fewest number of samples possible. If one can determine, *a priori*, that one of the two distributions is expected to be more informative for the entire range of θ, then the decision is trivial: always sample from that distribution. In general, however, it is often the case that each model is a more reliable model for a particular regime of θ. For instance, consider

$$\mathbb{P}_1(\cdot \mid \theta) \sim \mathcal{N}(0, \theta) \quad \text{and} \quad \mathbb{P}_2(\cdot \mid \theta) \sim \mathcal{N}(0, 1-\theta)\,, \quad (1)$$

for $\theta \in (0,1)$. It can be readily verified that when $\theta \in (0, 1/2)$, model $\mathbb{P}_1$ is a more reliable source of estimating θ, and when $\theta \in (1/2, 1)$, model $\mathbb{P}_2$ is more reliable. In such circumstances, pre-selecting and exclusively sampling from one of the two sequences is strongly sub-optimal, and one needs to design a mechanism that can dynamically switch between the two streams. In this paper we consider a general setting in which we have K models and design a sequential and data-adaptive mechanism for actively selecting the information sources over time such that an overall notion of optimality is ensured.

Designing such a mechanism involves co-designing the sampling and estimation processes. A closely related approach to designing such coupled sampling and decision-making process is *controlled (active) sensing*, originally developed by Chernoff for binary composite hypothesis testing through incorporating a controlled information gathering process that dynamically decides about taking one of a finite number of possible actions at each time [1]. Under the assumption of uniformly distinguishable hypotheses and having independent control actions, Chernoff's rule decides in favor of the action with the best *immediate* return according to proper information measures and achieves optimal performance in the asymptote of a diminishing rate of erroneous decisions. Chernoff's rule, specifically, at each time, identifies the most likely decision based on the collected data and takes the action that reinforces the decision. Extensions of the Chernoff's rule to various settings are studied in [2–5] including the more recent studies that are relevant to the scope of this paper in [6–16].

Unlike the rather extensive literature on active sampling design mechanisms for detection and classification problems, those for parameter estimation is far less investigated. The existing studies relevant to the scope of this paper include [17–20]. Specifically, [17] assigns a fixed sampling cost for collecting each measurement. It models the trade-off between collecting more measurements to improve the estimation fidelity, and stopping the sampling process to increase the agility using a cost function that linearly aggregates the costs associated with sampling and estimation. An asymptotically optimal sequential procedure is then prescribed for minimizing the unified cost function. This work was later generalized to the setting of multiple controls in [18], where each control depends on a control-specific parameter which was assumed to be different for every control. This framework was further extended to the case of models sharing a

This research was supported in part by RPI-IBM Artificial Intelligence Research Center, and the U. S. National Science Foundation under the grants CAREER ECCS-155448 and ECCS-1933107.

common unknown parameter in [19]. The models investigated in [17–19] optimize a cost function (linear combination of delay and estimation costs) that cannot provide an explicit guarantee on the estimation quality. To address this with an explicit emphasis on the estimation cost, a more natural formulation has been considered in [20], where the aim is to reduce the average sample complexity, such that the estimation cost falls below a pre-prescribed threshold. However, the setting in [20] considers only one data stream, in which the notion of actively sampling the information sources is not relevant. Our contribution is a generalization of [20] for multiple information streams, with the emphasis on designing an active sampling procedure.

2. ACTIVE SAMPLING FOR ESTIMATION

2.1. Data Model

Consider K information sources $\{S_1, \ldots, S_K\}$. Source S_i, for $i \in [K] \triangleq \{1, \cdots, K\}$ generates a time series consisting of independent and identically distributed (i.i.d.) random variables denoted by $\mathcal{X}^i \triangleq \{X_t^i \ : \ t \in \mathbb{N}\}$. The samples in $\mathcal{X}^i$ are generated according to a statistical model with the probability density function (pdf) $f_i(\cdot \mid \theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^m$ is an unknown parameter and Θ is a compact set.

In this paper we consider the canonical Bayesian parameter estimation problem, in which the objective is to form a reliable estimate for θ by collecting samples from $\{\mathcal{X}^1, \ldots, \mathcal{X}^K\}$. We denote the prior pdf of θ by π. To capture the fidelity of an estimate U, we adopt the quadratic estimation cost function denoted by $\ell(\theta, U) = \|U - \theta\|_2^2$.

2.2. Active Sampling

We consider a fully sequential data-acquisition mechanism, according to which we select one source at-a-time to collect a sample from. The objective is to identify an optimal sequence of source selections, such that with the fewest number of samples, on an average, we can form a sufficiently accurate estimate for θ. It is imperative to note that different sources have potentially distinct statistical models, rendering different estimation qualities. In general, we will not have a source whose estimation quality will be dominantly stronger than all other sources for the entire range of θ. If that happens, e.g., source S_1 offers the most reliable model for estimating θ for all possible values of θ, then a trivial sampling decision is to collect all the samples from S_1. Otherwise, an effective sampling strategy should be able to explore different sources in order to converge to and sample from the sources that are deemed the stronger sources.

Based on this pivotal premise, the key question is what order of sampling results in forming a reliable estimate with the fewest number of samples. To formalize this, we consider the following general sampling model that dynamically selects the sources over time. Samples are collected sequentially, such that at any time t and based on the information accumulated up to that time, the sampling procedure takes one of the following actions:

- A_1) *Exploration:* Due to lack of sufficient confidence, forming an estimate is deferred, and one more sample is taken.
- A_2) *Stopping:* Data collection-acquisition is terminated, indicating that there is sufficient data to form a reliable estimate for θ.
- A_3) *Estimation:* After stopping, an estimate is formed.

This process can be expressed uniquely by the data-adaptive rule for selecting sources over time, a stopping rule, and a final detection decision rule. To formalize the exploration process, define $\psi \ : \ \mathbb{N} \to [K]$, where $\psi(t)$ returns the index of the source to be selected at time t. We also denote the sample collected from sensor $S_{\psi(t)}$ by Y_t. Accordingly, we define

$$Y^t \triangleq \big(Y_1, \cdots, Y_t\big) \,, \ \text{and} \ \psi^t \triangleq \big(\psi(1), \cdots, \psi(t)\big) \,. \quad (2)$$

The information accumulated generates a σ-algebra denoted by $\{\mathcal{F}_t \ : \ t \in \mathbb{N}\}$ where $\mathcal{F}_t \triangleq \sigma(Y^t, \psi^t)$. We define $N \in \mathbb{N}$ as the stopping time of the sampling process that is $\mathcal{F}_t$-measurable. Finally, we define $\hat{\theta}_t$ as the estimate of θ at time t. Based on these decision rules, we define $\Delta \triangleq \big(N, \psi^N, \hat{\theta}_N\big)$ to specify the sampling strategy and the decisions involved.

2.3. Formulation

The two key figures of merit involved are the average sampling complexity N and the quality of the estimate. There exists an inherent tension between these two, as improving one penalizes the other one. In this subsection, we provide a formulation that explicitly captures this dichotomy and resolves the tension between them in a natural way. To proceed, we note that it can be readily verified that the posterior distribution of θ at time t is given by

$$\pi_t(\theta) \ = \ \frac{\pi(\theta) \prod_{i=1}^{t} f_{\psi(i)}(x_i \mid \theta)}{\displaystyle\int_{v \in \Theta} \pi(v) \prod_{i=1}^{t} f_{\psi(i)}(x_i \mid v) \, \mathrm{d}v} \,. \quad (3)$$

Based on this, for any given t, we define the *average posterior cost* function as

$$\mathsf{C}(\hat{\theta}_t \mid \mathcal{F}_t) \ \triangleq \ \mathbb{E}_t\big[\ell(\hat{\theta}_t, \theta) \mid \mathcal{F}_t\big] \,, \quad (4)$$

where $\mathbb{E}_t$ denotes expectations with respect to π_t. Based on these, we aim to find a sampling strategy Δ that minimizes N,

such that the estimation cost does not exceed a pre-specified threshold $\beta > 0$, i.e.,

$$\mathcal{P}(\beta) \triangleq \left\{ \begin{array}{ll} \min\limits_{\Delta} & \mathbb{E}[N] \\ \text{s.t.} & \mathsf{C}(\hat{\theta}_N \mid \mathcal{F}_N) \leq \beta \end{array} \right. . \tag{5}$$

In order to compare the performance of any strategy Δ, we consider a Bayesian counterpart of $\mathcal{P}(\beta)$. For this purpose the delay and estimation costs are integrated into a unified cost function denoted by

$$J(\Delta \mid \mathcal{F}_N) \triangleq \mathbb{E}[N] + c_\beta \cdot \mathsf{C}(\hat{\theta}_N \mid \mathcal{F}_N) . \tag{6}$$

It can be readily verified that for any β, there exists the constant c_β such that a solution to (5) can be found by equivalently solving $\min_\Delta J(\Delta \mid \mathcal{F}_N)$. By leveraging this relationship, we use the notion of *weak*-asymptotic pointwise optimality (w-APO), which was first introduced in [17], for assessing the relative efficiency of any solution Δ.

Definition 2.0. *A sequential procedure Δ is called* w*-APO if for any other $\bar{\Delta}$, and for any $\epsilon > 0$ we have*

$$\mathbb{P}\left\{ \frac{J(\Delta \mid \mathcal{F}_N)}{J(\bar{\Delta} \mid \mathcal{F}_{\bar{N}})} \leq 1 + \epsilon \right\} \to 1, \quad \text{as} \quad c_\beta \to \infty . \tag{7}$$

3. COST-AWARE ACTIVE ESTIMATION

In this section, we formalize a *cost-aware active estimation* (CAE) framework for active sampling and estimation. To specify different decision rules, we define $\lambda_i(x \mid \theta) \triangleq \log f_i(x \mid \theta)$ as the log-likelihood associated with the distribution of source S_i for $i \in [K]$. We assume that $\mathbb{E}_i[|\lambda_i(x \mid \theta)|] < +\infty$ for all $\theta \in \Theta$. Furthermore, we assume that $\lambda_i(x \mid \theta)$ is twice differentiable everywhere in θ such that $\frac{\partial^2}{\partial \theta^2}\lambda_i(\mathbf{x}^n;\theta)$ exists and is bounded. Accordingly, we denote the Fisher information (FI) associated with source S_i by

$$\mathcal{I}_i(\theta) \triangleq -\mathbb{E}_i\left[\frac{\partial^2}{\partial \theta^2} \lambda_i(x;\theta) \right] . \tag{8}$$

Finally, we assume that the likelihood functions under two sufficiently distinguishable parameters θ and ϕ are also distinguishable, that is,

$$\mathbb{E}_\theta\big[\sup\{\lambda_i(x;\theta) - \lambda_i(x;\phi) \, : \, |\theta - \phi| > \epsilon\} \big] < 0 . \tag{9}$$

3.1. Decision Rules

Estimator. We start by fixing the stopping time N and the sampling sequence ψ^N. We first specify an estimator that minimizes the average posterior estimation cost function $\mathsf{C}(\hat{\theta}_N \mid \mathcal{F}_N)$ for any given N and ψ^N. Assuming that $\mathbb{E}[N] < +\infty$ we can write

$$\mathsf{C}(\hat{\theta}_N \mid \mathcal{F}_N) = \mathbb{E}_t\left[\sum_{t=0}^{\infty} \ell(\hat{\theta}_N, \theta) \mathbb{1}_{\{N=t\}} \mid \mathcal{F}_t \right] \tag{10}$$

$$= \sum_{t=0}^{\infty} \mathbb{E}_t\left[\ell(\hat{\theta}_N, \theta) \mid \mathcal{F}_t \right] \mathbb{1}_{\{N=t\}} \tag{11}$$

$$\geq \sum_{t=0}^{\infty} \inf_{\hat{\theta}_N} \mathbb{E}_t\left[\ell(\hat{\theta}_N, \theta) \mid \mathcal{F}_t \right] \mathbb{1}_{\{N=t\}} . \tag{12}$$

Note that the indicator function $\mathbb{1}_{\{N=t\}}$ can be factored out since it is $\mathcal{F}_t$-measurable. The optimum Bayes estimator satisfies

$$\hat{\theta}_t = \arg\inf_{\hat{\theta}_t} \mathbb{E}_t\left[\ell(\hat{\theta}_t, \theta) \mid \mathcal{F}_t \right] , \tag{13}$$

and under the mean-square error (MSE) criterion it is well-known that the optimum Bayesian estimator is

$$\nu_t \triangleq \mathbb{E}_t\big[\theta \mid \mathcal{F}_t\big] . \tag{14}$$

Hence, by denoting the conditional average cost by $\mathsf{C}_t \triangleq \mathbb{E}_t\left[\ell(\nu_t, \theta) \mid \mathcal{F}_t\right]$, from (10)-(12) we have

$$\mathsf{C}(\hat{\theta}_N \mid \mathcal{F}_N) \geq \sum_{t=0}^{\infty} \mathsf{C}_t \mathbb{1}_{\{N=t\}} = \mathsf{C}_N . \tag{15}$$

Hence, for any arbitrary stopping time N, if we apply the optimum Bayes estimator at the stopping time, the conditional expected cost $\mathsf{C}(\hat{\theta}_N \mid \mathcal{F}_N)$ matches the lower bound C_N. We adopt the Bayes estimator in (14) as our estimator.

Active Sampling Rule. The source selection and sampling rule follows Chernoff's principle for sequential design of experiments for binary composite hypothesis testing through incorporating a controlled information gathering process that dynamically decides about taking one of a finite number of possible actions at each time [1]. Under the assumption of uniformly distinguishable hypotheses and having independent control actions, Chernoff's rule decides in favor of the action with the best immediate return according to proper information measures and achieves optimal performance in the asymptote of a diminishing rate of erroneous decisions. Chernoff's rule, specifically, at each time, identifies the most likely decision on the collected data and takes the action that reinforces the decision.

In the context of the estimation problem considered in this paper, at each time-step $t \in \mathbb{N}$, we wish to identify the *most informative* source that is expected to reduce the estimation costs by the largest margin. As a relevant measure for comparing the informativeness of different sources, we adopt the FI measure. More specifically, at any given instant $t \in \mathbb{N}$, based on ν_{t-1} as the the most updated estimate of θ given $\mathcal{F}_{t-1}$, we select the source with the highest FI measure, i.e., we select source

$$\psi(t) = \arg\max_{i \in [K]} \mathcal{I}_i(\nu_{t-1}) , \tag{16}$$

and any tie is resolved by tossing a fair coin.

Stopping Rule. Finally, given the estimators and the sampling rules, we specify the stopping rule. This rule is directly driven by the decision quality constraint specified in the formulation of problem $\mathcal{P}(\beta)$ in (5). Specifically, based on $\mathcal{P}(\beta)$, we are interested in minimizing the number of samples such that the average posterior estimation cost falls below the target level. Based on this, we set the stopping time as the first time that the cost $\mathsf{C}(\hat{\theta}_N \mid \mathcal{F}_N)$ falls below β, i.e.,

$$N \triangleq \inf \left\{ n \in \mathbb{N} : \mathsf{C}(\hat{\theta}_N \mid \mathcal{F}_N) \leq \beta \right\} . \quad (17)$$

3.2. Performance Guarantee

Next, we show that the combination of the decision rules specified in the previous subsection, collectively, admit w-APO optimality, formalized in the following theorem.

Theorem 3.1. *The combination of the estimator in* (14)*, active sampling rule in* (16)*, and the stopping rule in* (17)*, is* w*-APO.*

It is noteworthy that although this theorem is showing asymptotic optimality, in certain regimes the specific rules admit stronger optimality properties. For instance, in the special case of $K = 1$, i.e., only one information source, our framework and proposed algorithm reduces to that of [20], in which case stronger optimality properties are established.

4. NUMERICAL EXPERIMENT

A simple setup consisting of two sensors is considered. Each sensor generates i.i.d. random measurements according to exponential and Erlang-2 models:

$$\text{Sensor 1}: \quad X \mid \theta \sim \exp(x \mid \theta) , \quad (18)$$

$$\text{Sensor 2}: \quad X \mid \theta \sim \text{Erlang}(x; 2 \mid \theta) . \quad (19)$$

The prior distribution of θ is also exponential with parameter a. Subsequently, the posterior distribution of θ at time t is Gamma$(\theta, 2t - \gamma_t + 1, y_t)$, where we have defined

$$y_t \triangleq a + \sum_{i=1}^{t} x_i , \quad \text{and} \quad \gamma_t \triangleq \sum_{i=1}^{t} \mathbb{1}_{\{\psi(i)=1\}} . \quad (20)$$

γ_t counts the number of times that the sampling action selects sensor S_1. The MMSE estimate and the associated average posterior cost function are given by

$$\hat{\theta}_t = \frac{2n - \gamma_t + 1}{y_t} \quad \text{and} \quad \mathsf{C}(\hat{\theta}_t \mid \mathcal{F}_t) = \frac{2t - \gamma_t + 1}{y_t^2} . \quad (21)$$

Figure 1 shows the variations of the average sample complexity $\mathbb{E}[N]$ versus the estimation cost constraint β for the choices of $\theta = 0.5$ and $a = 2$. This figure compares the performance of three approaches to sensor selection:

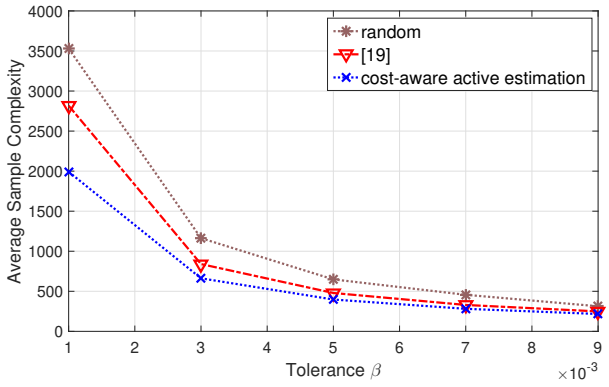

Fig. 1. Average sample complexity versus varying levels of prescribed tolerance β

1. The baseline model is a random selection of the sensors, according to which at each time one of the sensors is selected with probability $\frac{1}{2}$ for sampling.
2. The performance of the approach of [19]. This approach involves a tuning parameter c the controls the tradeoff between average delay and estimation performance and does not have an explicit performance guarantee on the estimation quality (a counterpart of β in our model). In order to facilitate comparisons, for any given β, we determine what choices of c yields an estimation cost β for [19], and use that to generate the $\mathbb{E}[N]$ versus β curve.
3. Finally, we plot the curve of the active estimation approach of this paper.

As the comparisons show, our proposed approach yields a (considerable) performance improvement for the smaller values of β, which is the more important regime accounting for high-accuracy estimation. As β increases, expectedly, the gap among the three methods diminishes.

5. CONCLUSION

In this paper, we have investigated an active estimation framework for estimating a covariate parameter shared by multiple statistical models. This framework provides a co-design for actively sampling the models over time and estimating the covariate parameter. The main observation is that when collecting the samples sequentially over time, a combination of forming a maximum likelihood estimate of the covariate parameter and subsequently selecting the model that has the largest Fisher information measure associated with the estimate renders an asymptotically optimal rule. The stopping rule of this framework consist of comparing the posterior average estimation cost with a pre-specified threshold that controls the estimation fidelity.

6. REFERENCES

[1] H. Chernoff, "Sequential design of experiments," *The Annals of Mathematical Statistics*, vol. 30, no. 3, pp. 755–770, Sep. 1959.

[2] S. Bessler, *Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments: Part I Theory*, Ph.D. thesis, Department of Statistics, Stanford University, March 1960.

[3] A. E. Albert, "The sequential design of experiments for infinitely many states of nature," *The Annals of Mathematical Statistics*, vol. 32, no. 3, pp. 774–799, Sep. 1961.

[4] G. E. P. Box and W. J. Hill, "Discrimination among mechanistic models," *Technometrics*, vol. 9, no. 1, pp. 57–71, Feb. 1967.

[5] W. J. Blot and D. A. Meeter, "Sequential experimental design procedures," *Journal of the American Statistical Association*, vol. 68, no. 343, pp. 586–593, Sep. 1973.

[6] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli, "Controlled sensing for multihypothesis testing," *IEEE Transactions on Automatic Control*, vol. 58, no. 10, pp. 2451–2464, May 2013.

[7] S. Nitinawarat and V. V. Veeravalli, "Controlled sensing for sequential multihypothesis testing with controlled Markovian observations and non-uniform control cost," *Sequential Analysis*, vol. 34, no. 1, pp. 1–24, Feb. 2015.

[8] K. Cohen and Q. Zhao, "Active hypothesis testing for anomaly detection," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1432–1450, Mar. 2015.

[9] N. K. Vaidhiyan and R. Sundaresan, "Active search with a cost for switching actions," *arXiv:1505.02358v1*, May 2015.

[10] J. Heydari, A. Tajer, and H. V. Poor, "Active sampling for the quickest detection of Markov networks," *arXiv 1711.04268*, 2020.

[11] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," *Annals of Statistics*, vol. 41, no. 6, pp. 2703–2738, Dec. 2013.

[12] J. Wang, "Bayes-optimal sequential multi-hypothesis testing in exponential families," *arXiv:1506.08915v1*, Jun. 2015.

[13] K. Cohen, Q. Zhao, and A. Swami, "Optimal index policies for anomaly localization in resource-constrained cyber systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4224–4236, Aug. 2014.

[14] K. Cohen and Q. Zhao, "Asymptotically optimal anomaly detection via sequential testing," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2929–2941, Jun. 2015.

[15] J. Heydari and A. Tajer, "Quickest search and learning over correlated sequences: Theory and application," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 638–651, February 2019.

[16] J. Heydari, A. Tajer, and H. V. Poor, "Quickest Linear Search over Correlated Sequences," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5786–5808, October 2016.

[17] P. J. Bickel and J. A. Yahav, "Asymptotically pointwise optimal procedures in sequential analysis," in *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, Calif., 1967, pp. 401–413, University of California Press.

[18] V. J. Yohai, "Asymptotically optimal Bayes sequential design of experiments for estimation," *Annals of Statistics*, vol. 1, no. 5, pp. 822–837, September 1973.

[19] G. Atia and S. Aeron, "Asymptotic optimality results for controlled sequential estimation," in *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, October 2013.

[20] G. V. Moustakides and T. Yaacoub and Y. Mei, "Sequential estimation based on conditional cost," in *Proc. IEEE International Symposium on Information Theory*, Aachen, Germany, July 2017.