A SAMPLE-EFFICIENT SCHEME FOR DATA-DRIVEN DISTRIBUTED RESOURCE ALLOCATION IN NETWORKED ESTIMATION

Marcos M. Vasconcelos and Urbashi Mitra

Ming Hsieh Department of Electrical Engineering, University of Southern California, USA

ABSTRACT

Remote estimation over communication channels of limited capacity is an area of research with applications spanning many economically relevant areas, including cyber-physical systems and the Internet of Things. One popular choice of communication/scheduling policies used in remote estimation is the class of event-triggered policies. Typically, an event-triggering threshold is optimized, assuming complete knowledge of the system's underlying probabilistic model. However, this information is seldom available in real-world applications. This paper addresses the learning of an optimal threshold policy based on data samples collected at the sensor. Leveraging symmetry, quasi-convexity, and the method of Kernel density estimation, we propose a data-driven algorithm, which is guaranteed to converge to a globally optimal solution. Moreover, empirical evidence suggests that our algorithm is more sample-efficient than traditional learning approaches based on empirical risk minimization.

***Index Terms*—** Remote estimation, threshold policies, collision channel, machine learning, optimization

1. INTRODUCTION

Remote estimation systems constitute a broad class of problems with applications in many important technological fields such as cyber-physical systems and the Internet of Things [1, 2]. For the most part, results either prove the optimality of threshold policies for a particular problem [3, 4], or assume that the sensors use threshold policies and optimize performance of the system over the threshold [5]. In either case, the underlying probabilistic model is assumed to be fully available to the system designer, which is often not the case in practice. In this paper, we study a system where each sensor learns the optimal threshold from independent and identically distributed (IID) data samples drawn from an unknown symmetric probability density function (PDF). To the best of our knowledge, our approach is the first to establish a connection between a remote estimation problem and the area of statistical learning theory [6]. In particular, we provide a sample-efficient algorithm which learns the optimal threshold with guaranteed convergence.

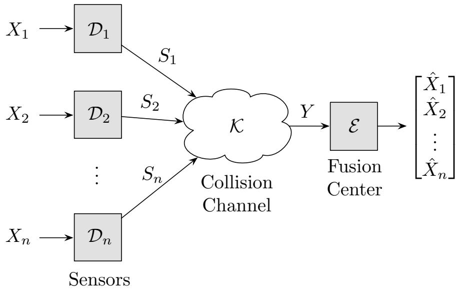

Fig. 1: System diagram for remote estimation over the collision channel.

2. PROBLEM FORMULATION

Consider a system where multiple sensors make measurements and communicate them to a remote fusion center. We will focus on the case where sensors make IID observations and decides whether to transmit them to the fusion center or not. The communication medium is modeled by a collision channel: If two or more nodes transmit in the same time slot, then the packets interfere with each other and do not get delivered at the receiver [1, 7].

There are $n \geq 2$ sensors, and at each time slot the i-th sensor makes a measurement $X_i = x_i$ and decides whether to communicate it or not to the fusion center using an event-triggered policy characterized by a single threshold δ, as follows:

$$S_i = \begin{cases} (i, x_i) & \text{if } |x_i - \mu| \geq \delta \\ \mathbf{S} & \text{otherwise.} \end{cases} \tag{1}$$

We make three assumptions: 1. all the measurements are identically distributed:

$$X_i \sim f_X, \;\; i \in \{1, \cdots, n\}; \tag{2}$$

2. the PDF is symmetric around the mean:

$$f_X(x - \mu) = f_X(-x + \mu); \tag{3}$$

This work was supported in part by the following agencies: ONR under grant N00014-15-1-2550, NSF under grants CNS-1213128, CCF-1718560, CCF-1410009, CPS-1446901 and AFOSR under grant FA9550-12-1-0215.

3. the PDF is supported on the real line:

$$f_X(x) > 0, \quad x \in \mathbb{R}. \tag{4}$$

The channel operates as follows: if two or more sensors transmit at the same time, $Y = \mathbf{C}$, (the "collision" symbol). If only the i-th sensor transmits, $Y = (i, x_i)$. If no sensors transmit $Y = \mathbf{S}$, (the "silence" symbol). Under the symmetry assumption of the PDF and the transmission policy, the optimal receiver is of the following form [5, 8]:

$$\hat{X}_i = \begin{cases} x_i & \text{if } Y = (i, x_i) \\ \mu & \text{otherwise.} \end{cases} \tag{5}$$

The goal is to choose a threshold δ such that the following normalized mean squared error is minimized

$$\mathcal{J}(\delta) = \frac{1}{n}\mathbf{E}\Big[\sum_{i=1}^{n}(X_i - \hat{X}_i)^2\Big]. \tag{6}$$

Without loss of generality (and due to space constraints), we will consider in this paper only the case in which $\mu = 0$ and $n = 2$. The general case will be considered in a forthcoming journal publication.

Proposition 1 *Under event-triggered policies characterized by a threshold δ, the cost function*[1] *is given by:*

$$\mathcal{J}(\delta) := \mathbf{E}[X^2] - \mathbf{E}[X^2\mathbf{1}(|X| \geq \delta)]\mathbf{E}[\mathbf{1}(|X| < \delta)]. \tag{7}$$

In [5], it was shown that the objective function in Eq. (7) satisfies the following important property.

Proposition 2 *For any symmetric and continuous PDF supported on the real line, the objective function $\mathcal{J}(\delta)$ in Eq.* (7) *is strictly quasi-convex.*

Proposition 2 is very attractive because it guarantees the existence of a single global optimal threshold, and that there are simple numerical algorithms to compute it (e.g. stochastic gradient descent).

Obviously, any algorithm used to compute such globally optimal threshold must rely on the full knowledge of f_X. Instead, we are interested in the case where the PDF is unknown, and each sensor needs to learn its threshold using a finite data set of IID samples drawn from the PDF f_X available to it, $\mathcal{D}_i := \{x_{i,k}\}_{k=1}^{M_i}$. The threshold design is decentralized, i.e., the data is not exchanged among sensors. This is due to both practicality and privacy constraints. The purpose of this paper is to provide a sample efficient data-driven algorithm that can be used to learn the optimal threshold $\delta^\star$ without exchanging data among sensors.

3. MAIN RESULTS

One data-driven approach to computing an estimate of the optimal threshold is to use the data set $\mathcal{D}_i$ to construct an approximate $\tilde{\mathcal{J}}_{\mathcal{D}_i}(\delta)$, and optimize it with respect to δ, i.e.,

$$\delta^\star_{\mathcal{D}_i} := \arg\min \tilde{\mathcal{J}}_{\mathcal{D}_i}(\delta). \tag{8}$$

The most natural approach is to replace the expectations in the cost function by their empirical means, this is called sample average approximation (SAA) [9] or empirical risk minimization (ERM) [10]. However, the cost function in Eq. (7) involves the product of expectations, which requires a slightly modified SAA in order to guarantee that the approximate objective function is an unbiased estimate of the expected cost. The following is an unbiased estimate of the objective function based on empirical means [11]:

$$\tilde{\mathcal{J}}^{\mathrm{SAA}}_{\mathcal{D}_i}(\delta) := \frac{1}{M_i}\sum_{k=1}^{M_i} x_{i,k}^2 - \frac{2}{M_i}\sum_{k=1}^{M_i/2} x_{i,k}^2\mathbf{1}(|x_{i,k}| \geq \delta)] \cdot \frac{2}{M_i}\sum_{j=M_i/2+1}^{M_i}\mathbf{1}(|x_{i,k}| < \delta). \tag{9}$$

Remark 1 *Notice that the SAA in Eq.* (9) *is inefficient because the data set must be partitioned into disjoint sets to approximate each of the expectations in the product term. Moreover, Fig. 2 shows that Eq.* (9) *is non-smooth, and in some cases, may not be quasi-convex, which is the feature that enables efficient numerical algorithms.*

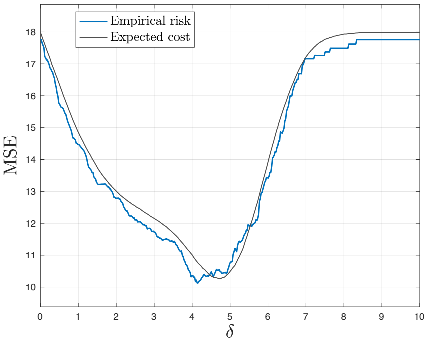

Fig. 2: Objective function and its sample average approximation using 500 data samples from the unknown density $f_X := \frac{1}{3}\Big(\mathcal{N}(-5,1) + \mathcal{N}(0,2) + \mathcal{N}(5,1)\Big)$.

This approach works well if the data set is sufficiently large, but fails in the finite data regime for two reasons: the first is the lack of smoothness, which prevents the use of any first order methods such as stochastic gradient descent. The second is that the function also ceases to be quasi-concave, which leads to multiple local minima. Therefore, the global

[1] The function $\mathbf{1}(\mathfrak{S})$ denotes the indicator function of the statement $\mathfrak{S}$.

maximum must be found by exhaustive search over the entire real line, which is not numerically efficient. We propose an approach which is data-driven but retains the smoothness and quasi-convexity of the objective function for data sets of any size such that a first-order method that is guaranteed to converge to the globally optimal solution can be efficiently implemented.

Theorem 1 *Let $\mathcal{D} = \{x_k\}_{k=1}^M$ denote a the data set of M IID samples drawn from symmetric PDF f_X. Let $\tilde{f}_{\mathcal{D}}(x)$ denote a PDF that approximates f_X constructed from $\mathcal{D}$. Let $\tilde{\mathcal{J}}_{\mathcal{D}}(\delta)$ be defined as:*

$$\tilde{\mathcal{J}}_{\mathcal{D}}(\delta) := \int_{\mathbb{R}} x^2 \tilde{f}_D(x)dx - \Big[\int_{\mathbb{R}} x^2 \mathbf{1}(|x| \geq \delta)\tilde{f}_D(x)dx\Big] \times \Big[\int_{\mathbb{R}} \mathbf{1}(|x| < \delta)\tilde{f}_D(x)dx\Big] \quad (10)$$

and $\delta^\star_{\mathcal{D}}$ where

$$\delta^\star_{\mathcal{D}} := \arg\min \tilde{\mathcal{J}}_{\mathcal{D}}(\delta). \quad (11)$$

Then,

$$\delta^\star_{\mathcal{D}} \overset{\text{w.p.1}}{\longrightarrow} \delta^\star, \quad M \to \infty, \quad (12)$$

where $\delta^\star$ is the global minimizer of Eq. (7).

Proof. Due to space constraints, the proof can be found in □

The approach we propose is to approximate the PDF f_X from data using a non-parametric method called *Kernel denstity estimation* (KDE) [12]. The main difference between KDE and empirical risk minimization is that while the empirical risk approach is a nonsmooth unbiased estimate of the objective function, KDE yields a smooth *biased* estimate of the PDF, which in turns implies in a biased estimate of the objective function, which allows for first-order optimization methods. additionally, if we use the knowledge of the PDF's symmetry, KDE preserves the quasi-concavity of the objective function for any realization of the data set.

The KDE approximation for a symmetric PDF is given by the following expression [13]: Let $\mathcal{D} = \{x_k\}_{k=1}^M$, where $x_k \sim f_X$, with f_X symmetric. Then,

$$\tilde{f}_{\mathcal{D}}(x) := \frac{1}{2Mh_M}\sum_{k=1}^M \Big[\mathcal{K}\Big(\frac{x - x_k}{h_M}\Big) + \mathcal{K}\Big(\frac{x + x_k}{h_M}\Big)\Big], \quad (13)$$

where $\mathcal{K}$ is the Gaussian kernel given by $\mathcal{K}(x) := \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{x^2}{2}\big)$, and the parameter h_M is the so-called bandwidth parameter and must be appropriately chosen. Typically, h_M is chosen according to the so-called *Normal reference rule* [12] as follows:

$$h_M := 1.06 \cdot M^{-1/5} \cdot \min\Big\{\sqrt{\frac{\sum_{k=1}^n x_k^2}{n-1}}, \frac{Q_{\mathcal{D}}}{1.34}\Big\}, \quad (14)$$

where $Q_{\mathcal{D}}$ is the data's interquartile range, which is defined as the difference between the 75th and 25th percentiles. Figure 3 shows the KDE based approximation based on a data set of $M = 500$ samples from a symmetric Gaussian mixture PDF. Given an approximation of the density computed using KDE, we can compute the approximate optimal threshold from data using the bissection method used to compute the unique solution of the first order optimality condition:

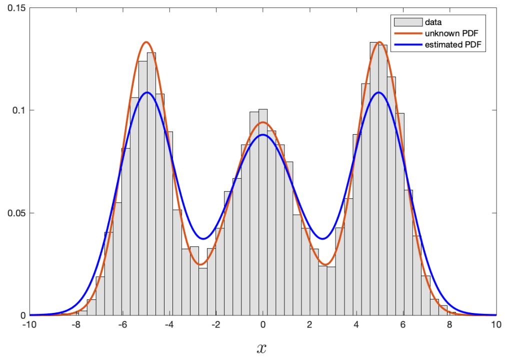

Fig. 3: Illustrative KDE approximation based on $M = 500$ samples from a symmetric Gaussian mixture PDF $f_X := \frac{1}{3}\Big(\mathcal{N}(-5,1) + \mathcal{N}(0,2) + \mathcal{N}(5,1)\Big)$.

$$\nabla \tilde{\mathcal{J}}_{\mathcal{D}}(\delta) = 0. \quad (15)$$

Theorem 2 *The global minimizer $\delta^\star_{\mathcal{D}}$ is the unique solution of $\tilde{\mathcal{F}}_{\mathcal{D}}(\delta) = 0$, where*

$$\begin{aligned}\tilde{\mathcal{F}}_{\mathcal{D}}(\delta) := {} & h_M^2 + \frac{1}{M}\sum_{k=1}^M x_k^2 \\ & - \frac{1}{M}\sum_{k=1}^n \frac{1}{\sqrt{2\pi}}\Big\{h_M \cdot (x_k - \delta)\cdot \exp\Big(-\frac{1}{2}\Big(\frac{x_k+\delta}{h_M}\Big)^2\Big) \\ & - h_M \cdot (x_k + \delta)\cdot \exp\Big(-\frac{1}{2}\Big(\frac{x_k-\delta}{h_M}\Big)^2\Big) \\ & + \sqrt{\frac{\pi}{2}}\cdot(h_M^2 + x_k^2)\cdot\Big[\mathrm{erf}\Big(\frac{x_k+\delta}{h_M\sqrt{2}}\Big) - \mathrm{erf}\Big(\frac{x_k-\delta}{h_M\sqrt{2}}\Big)\Big]\Big\} \\ & - \delta^2 \frac{1}{2M}\sum_{k=1}^M \Big[\mathrm{erf}\Big(\frac{x_k+\delta}{h_M\sqrt{2}}\Big) - \mathrm{erf}\Big(\frac{x_k-\delta}{h_M\sqrt{2}}\Big)\Big]. \quad (16)\end{aligned}$$

Proof. Due to space constraints, the proof can be found in □

To find the unique solution of $\tilde{\mathcal{F}}_{\mathcal{D}}$ can be computed using the following algorithm:

- **Initialization:** Set $\underline{\delta}^{(0)} := 0$, and $\bar{\delta}^{(0)} := L$, where L is any constant such that $\tilde{\mathcal{F}}_{\mathcal{D}}(L) < 0$
- **At the k-th iteration:**

1. Set $\delta^{(k+1)} := 0.5(\underline{\delta}^{(k)} + \bar{\delta}^{(k)})$
2. If $\tilde{\mathcal{F}}_n(\delta^{(k+1)}) = 0$, set $\delta^\star := \delta^{(k+1)}$
3. If $\tilde{\mathcal{F}}_n(\delta^{(k+1)}) > 0$, set $\underline{\delta}^{(k+1)} := \delta^{(k+1)}$
4. If $\tilde{\mathcal{F}}_n(\delta^{(k+1)}) < 0$, set $\bar{\delta}^{(k+1)} := \delta^{(k+1)}$

- **Stopping criterion:** $\bar{\delta}^{(k+1)} - \underline{\delta}^{(k+1)} < \epsilon$

This algorithm converges to the optimal solution. Setting the desired accuracy $\epsilon > 0$ in the beginning of the algorithm implies that we achieve a solution within ϵ of the optimal one in $\mathcal{O}(\log(1/\epsilon))$ iterations. Figure 4 illustrates the convergence of the approximate function $\tilde{\mathcal{F}}_{\mathcal{D}}(\delta)$ to the true function $\mathcal{F}(\delta) := \nabla\mathcal{F}(\delta)/2f_X(\delta)$ as $M \to \infty$.

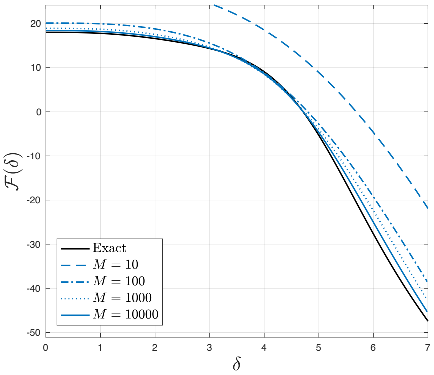

Fig. 4: Convergence of the approximate functions $\tilde{\mathcal{F}}_{\mathcal{D}}$ to the true $\mathcal{F}$ as the data set size M increases.

A real world implementation of this remote estimation system would be as follows: Each sensor would have a finite memory buffer M_i, at time k each sensor has a threshold estimate $\delta_i(k-1)$ and acquires a new data sample $x_i(k)$. Then the sensor updates its data set by dropping the oldest sample and adding the new one, i.e., $\mathcal{D}_i(k) = \{x_i(\ell)\}_{\ell=k-M_i+1}^{k}$. Based on this new data set, the sensor updates its threshold by computing: $\delta_i(k) = \arg\min \tilde{\mathcal{J}}_{\mathcal{D}_i(k)}(\delta)$. Finally, for a system with two sensors, we need 1 extra round of communication used for the sensors to agree on a single threshold. Let $i, j \in \{1, 2\}$ and $i \neq j$: $\delta_i(k) \leftarrow \frac{1}{2}\Big[\delta_i(k) + \delta_j(k)\Big]$. In a more general setup with a network of $n > 2$ sensors, a few rounds of local communications implementing a consensus/gossip algorithm would be required [8].

4. NUMERICAL RESULTS

We compare our algorithm based on KDE with the approach based on empirical risk minimization. Simulating the remote estimation system with sensors of data buffers of size M for 1000 sample paths with data samples drawn from $f_X = \frac{1}{3}\big(\mathcal{N}(-5,1) + \mathcal{N}(0,2) + \mathcal{N}(5,1)\big)$, we observe in Fig. 5 that as M increases the threshold converges to the true optimal solution in both cases. In fact, for $M \geq 100$, the average threshold (Fig. 5) and associated average performance (Fig. 6), are essentially the same. However, the shaded regions, which display the empirical variance of the simulations, show that the ERM scheme is much noisier than the scheme based on KDE. This stems from the fact that the ERM scheme minimizes a non-smooth, non-convex function as opposed to our algorithm which minimizes a smooth, quasi-convex function.

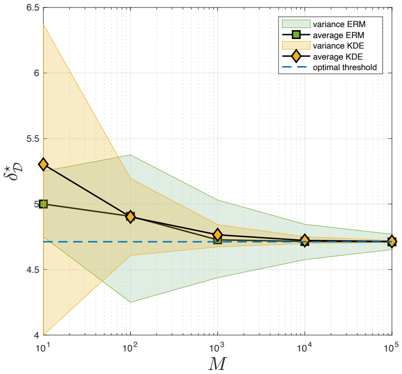

Fig. 5: Approximate threshold computed from data sets of increasing size M for the ERM and KDE approaches.

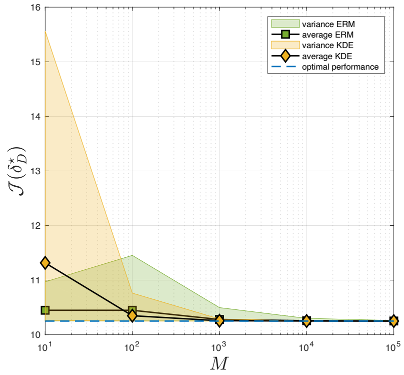

Fig. 6: True performance of thresholds computed from data sets of increasing size M for the ERM and KDE approaches.

Comparing the performance of both schemes in Figs. 5 and 6, we see that to achieve the same level of confidence on the learned threshold we need approximately 10 times more data samples in the ERM scheme. Therefore, we can state with confidence that our algorithm based on KDE is much more sample efficient than ERM.

5. CONCLUSIONS AND FUTURE WORK

This paper contains an algorithm based on KDE that can be used to learn an optimal threshold from data in remote estimation systems characterized by an unknown symmetric PDF.

One promising future research direction is to come up with online threshold learning schemes. Another future research topic is to optimize the bandwidth (h_M) selection method with respect to the performance of the remote estimation system. Finally, we would like to generalize our results to asymmetric unknown densities.

6. REFERENCES

[1] M. M. Vasconcelos and N. C. Martins, “Optimal estimation over the collision channel,” *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 321–336, 2017.

[2] S. Wu, K. Ding, P. Cheng, and L. Shi, “Optimal scheduling of multiple sensors over lossy and bandwidth limited channels,” *IEEE Transactions on Control of Network Systems*, vol. 7, no. 3, pp. 1188–1200, 2020.

[3] G. M. Lipsa and N. C. Martins, “Remote state estimation with communication costs for first-order LTI systems,” *IEEE Transactions on Automatic Control*, vol. 56, no. 9, pp. 2013–2025, 2011.

[4] J. Yun, C. Joo, and A. Eryilmaz, “Optimal real-time monitoring of an information source under communication costs,” in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 4767–4772.

[5] X. Zhang, M. M. Vasconcelos, W. Cui, and U. Mitra, “An optimal symmetric threshold strategy for remote estimation over the collision channel,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 9195–9199.

[6] V. Vapnik, *The nature of statistical learning theory*, Springer science & business media, 2013.

[7] X. Chen, X. Liao, and S. S. Bidokhti, “Remote estimation in decentralized random access channels,” *arXiv preprint arXiv:2007.03652*, 2020.

[8] X. Zhang, M. M. Vasconcelos, W. Cui, and U. Mitra, “Distributed remote estimation over the collision channel with and without local communication,” *arXiv preprint arXiv:2005.11438*, 2020.

[9] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello, “The sample average approximation method for stochastic discrete optimization,” *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2002.

[10] S. Mei, Y. Bai, and A. Montanari, “The landscape of empirical risk for nonconvex losses,” *The Annals of Statistics*, vol. 46, no. 6A, pp. 2747–2774, 2018.

[11] A. Lee, S. Tiberi, and G. Zanella, “Unbiased approximations of products of expectations,” *Biometrika*, vol. 106, no. 3, pp. 708–715, 2019.

[12] L. Wasserman, *All of nonparametric statistics*, Springer Science & Business Media, 2006.

[13] C. H. Kraft, Y. Lepage, and C. Van Eeden, “Estimation of a symmetric density function,” *Communications in Statistics - Theory and Methods*, vol. 14, no. 2, pp. 273–288, 1985.