Mean Estimation from One-Bit Measurements

Alon Kipnis[*] and John C. Duchi[*†]
[*]Stanford University, Department of Statistics
[†]Stanford University, Department of Electrical Engineering.

Abstract

We consider the problem of estimating the mean of a symmetric log-concave distribution under the following constraint: only a single bit per sample from this distribution is available to the estimator. We study the mean squared error risk in this estimation as a function of the number of samples, and hence the number of bits, from this distribution. Under an adaptive setting in which each bit is a function of the current sample and previously recorded bits, we show that the optimal relative efficiency compared to the sample mean is the efficiency of the median. We also consider a distributed setting where each bit is only a function of a single sample. We show that the optimal efficiency of the adaptive setting can be attained by splitting the sample into two parts and allowing only a single adaptation. Otherwise, we show that no one sample estimation procedure can attain the optimal efficiency uniformly over all points in the parameter space. Our results indicate that estimating the mean from one-bit measurements is equivalent to estimating the sample median from these measurements. In the adaptive case, this estimate can be done with vanishing error for any point in the parameter space. In the distributed case, this estimate can be done with vanishing error only for a finite number of possible values for the unknown mean.

I. Introduction

Estimating parameters from data collected and processed by multiple units may be limited due to communication constraints between these units. For example, this scenario arises in sensor arrays where information is collected at multiple physical locations and transmitted to a central estimation unit. In these situations, the ability to estimate a particular parameter from the data is dictated not only by the quality of observations and their number but also by the available bandwidth for communicating between the sensors and the central estimator. The question that we ask is to what extent a parametric estimation task is affected by this constraint on communication, and what are the fundamental performance limits in estimating a parameter subject to such restriction.

This paper answers this question in a particular setting: the estimation of the mean θ of a symmetric log-concave distribution with finite variance, under the constraint that only a single bit can be communicated on each sample from this distribution. As it turns out, the ability to share information before committing on each one-bit message dramatically affects the performance in estimating θ. We, therefore, distinguish among the three settings illustrated in Figure 1

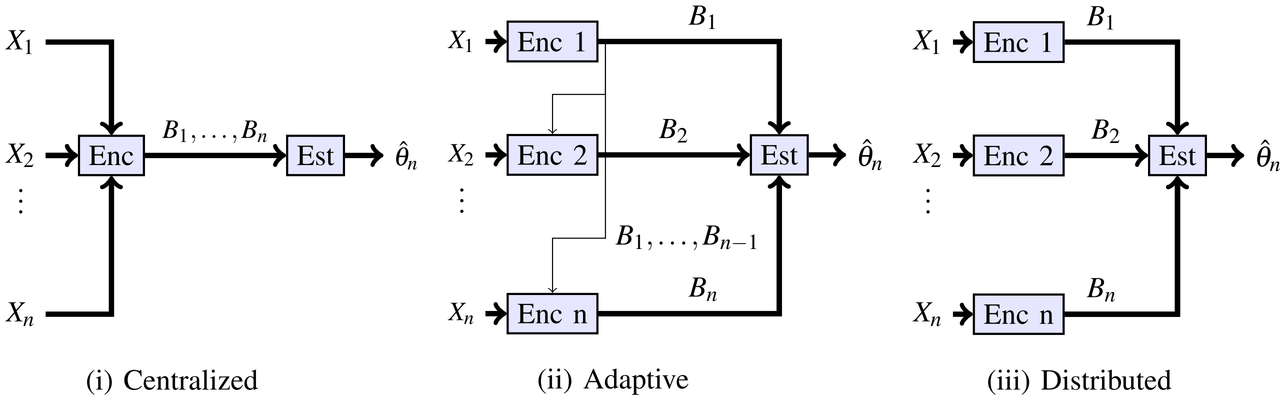

Fig. 1. Three encoding settings: (i) Centralized – an encoder sends n bits after observing n samples. (ii) Adaptive (sequential) – each encoder sends a bit B_n depending on its private sample X_n and previous bits $B_1,\dots,B_{n-1}$. (iii) Distributed – each encoder sends a single bit B_i only based on its private sample X_i.

(i) *Centralized* encoding: all n encoders confer and produce a single message consists of n bits.
(ii) *Adaptive* or *sequential* encoding: The nth encoder observes the nth sample and the $n-1$ previous bits.
(iii) *Distributed* encoding: The nth message is only a function of the nth sample.

Evidently, as far as information sharing is concerned, settings (iii) is a more restrictive version of (ii) which is more restrictive than (i). Below are three application examples for each of the settings (i)-(iii) above, respectively:

- **Signal acquisition:** A quantity is measured n times at different instances. The results are averaged in order to reduce measurement noise and the averaged result is then stored or communicated using n bits.
- **Analog-to-digital conversion:** A sigma-delta modulator (SDM) converts an analog signal into a sequence of bits by sampling the signal at a very high rate and then using one-bit threshold detector combined with a feedback loop to update an accumulated error state [1]. Therefore, the MSE in tracking an analog signal using an SDM falls under our setting (ii) when we assume that the signal at the input to the modulator is a constant (direct current) corrupted by, say, thermal noise [2]. Since the sampling rates in SDM are usually many times more than the bandwidth of its input, analyzing SDM under a constant input provides meaningful lower bound even for non-constant signals.
- **Privacy:** A business entity is interested in estimating the average income of its clients. In order to keep this information as confidential as possible, each client independently provides an answer to a yes/no question related to its income.

We measure the performance in estimating θ by the mean squared error (MSE) risk. For an asymptotically normal estimator with a finite MSE risk R_n, we are interested in particular in the limit

$$\lim_{n\to\infty} \frac{\sigma^2}{nR_n},$$

describing the *asymptotic relative efficiency* (ARE) of the estimator compared to asymptotically normal estimators whose variances decreases as $\sigma^2/n + o(n^{-1})$. Estimators of this form include the empirical mean of the samples, and, under some conditions, the optimal Bayes estimator.

In addition to the examples above, the excess risk in estimating a fixed parameter due to a one-bit per measurement constraint is useful in bounding from below the excess risk in estimating a signal from its noisy measurements. Namely, the excess MSE or ARE we derive serves as the most optimistic estimate for the risk in estimating under the one-bit per measurement constraint a signal changing over time or space. Such estimation settings are considered in [3], [4], [5], [6], [7].

In setting (i), the estimator can evaluate the optimal mean estimator (e.g., the sample mean in the Gaussian case) and then quantize it using n bits. Since the accuracy in describing the empirical mean decreases exponentially in n, the error due to quantization is negligible compared to the MSE in estimating the mean. Therefore, the ARE in this setting is 1. Namely, asymptotically, there is no loss in performance due to the communication constraint under centralized encoding. In this paper we show that a similar result does not hold in setting (ii): the ARE of any adaptive estimation scheme is at most the ARE of the sample median. Specifically, when the samples are drawn from the normal distribution, this ARE equals $2/\pi$, showing that the one-bit constraint increases the effective sample size in estimating θ by at least $\pi/2$ compared to estimating it without the bit constraint. We also show that this lower bound on the ARE is tight by providing an estimator that attains it. In fact, we show that only one adaptiveness, as illustrated in Figure 3 is enough to achieve the optimal ARE. Clearly, the penalty on the sample size in setting (iii) is at least as large as that in setting (ii). We show, however, that unlike in setting (ii), there is no distributed estimation scheme that is uniformly optimal in the sense that it attains the ARE of the sample median for all θ in the parameter space. This result is obtained by establishing local asymptotic normality of regularly enough encoding procedures for setting (iii), resulting in an exact characterization of the ARE of such procedures. To summarize our contributions, we establish that the optimal ARE in settings (i) is 1, the optimal ARE in setting (ii) is ARE of the sample median, and no uniformly optimal procedure exists in setting (iii).

It is important to note that although the ARE in setting (i) is one, this scheme already poses a non-trivial challenge for the design and analysis of an optimal encoding and estimation procedures. Indeed, the task of representing an unknown random quantity using n bits is equivalent to designing a 2^n levels scalar quantizer [8]. However, the optimal design of this quantizer depends on the distribution of its input, which is the goal of our estimation problem and hence its exact value is unknown. As a result, a non-trivial exploration-exploitation trade-off arises even in this case. Therefore, while uncertainty due to quantization decreases exponentially in the number of bits n, hence the ARE is 1, an exact expression for the MSE in this setting is still difficult to derive. The situation is even more involved in the adaptive encoding setting (ii): an encoding and estimation strategy that is optimal for $n-1$ adaptive one-bit messages of a sample of size $n-1$ may not lead to a globally optimal strategy upon the recipient of the nth sample. Conversely, any one-step optimal strategy, in the sense that it finds the best one-bit message as a function of the current sample and the previous $n-1$ messages, is not guaranteed to be globally optimal. Therefore, while we characterize the optimal one-bit message given the previous messages, this characterization does not necessarily lead to an upper bound on the ARE. Instead, our result on the maximal ARE is obtained by bounding the Fisher information of any n adaptive messages and using an appropriate information inequality.

Related Works

When the variance σ^2 is negligible compared the to desired accuracy, the task of finding θ using one-bit queries in the adaptive setting (ii) is solved by a bisection style method over the parameter space. Therefore, the general case of non-zero variance is reminiscent of the noisy binary search problem with a possibly infinite number of unreliable tests [9], [10]. However, since we assume a continuous parameter space, a more closely related problem is that of one-bit analog-to-digital conversion of a constant input corrupted by Gaussian noise. Using an SDM, Wong and Gray [2] showed that the output of the modulator converges to the true constant input almost surely, so that an SDM provides a consistent estimator for setting (ii). The rate of this convergence, however, was not analyzed and cannot be derived from the results of [2]. In particular, our results for setting (ii) imply that the asymptotic rate of convergence of the MSE in SDM to a constant input under an additive white Gaussian noise is at most $\sigma^2\pi/2$ over the number of feedback iterations. Baraniuk et. al [3] also considered adaptive one-bit measurements in the context of analog-to-digital conversion, although without noise at the input. By establishing the lower bound of $\sigma^2\pi/2n$ on the MSE, we show that the main results of [3], an exponential MSE decaying rate, does not hold in the noisy setting. Stated otherwise, the MSE in the setting of [3] may decay exponentially up to the noise level, after which it decays at most as $\sigma^2\pi/2n$.

One-bit measurements in the distributed setting (iii) was considered in [11], [12], [13], [14], [15], but without optimizing the encoders and their detection rules. Consequently, the performance derived in these works are not optimal. The work of [16] addresses the counterpart of our setting (iii) in the case of hypothesis testing, although the results there cannot be extended to parametric estimation. When the parameter space Θ is finite, Tsitsiklist [17] showed that when the cardinality of Θ is at most M and the probability of error criterion is used, then no more than $M(M-1)/2$ different detection rules are necessary in order to attain probability of error decreasing exponentially with the optimal exponent. Furthermore, in a version of this problem for the adaptive setting [18], it was shown that, with specific two-stage feedback, there is no gain in feedback compared to the fully distributed setting. Our results imply that the ARE in the distributed setting with threshold detection rules is strictly larger than that in the adaptive setting for almost all points in Θ, suggesting that the case where the cardinality of Θ is finite is different from the case where Θ is an open set.

As we explain in detail in Section III, the remote multiterminal source coding problem, also known as the CEO problem [19], [20], [21], [22], leads to lower bounds on the MSE in setting (iii). For the case of a Gaussian distribution, this lower bound bounds the ARE to be at most $3/4$. Thus, while this bound on the ARE provides no new information compared to the upper bound of $2/\pi$ we derive for setting (ii), it shows that the distributed nature of the problem is not a limiting factor in achieving MSE close to optimum even under the one bit per sample constraint.

Finally, we note that minimax estimation under limited communication with an arbitrary number of bits per node was considered in [23], [24]. The specialization of the results in [23], [24] to our settings (ii) and (iii) leads to looser lower bounds then the ones we derive in this paper. Looser bounds can also be obtained from works considering general inference and distributed estimation problems under data compression constraint [25], [26], [27], [28], [29]. In particular, the subsequent work of [29] uses an approach similar to ours to derive lower bound on the error in various parametric estimation problems from quantized measurements.

The remainder of this paper is organized as follows. In Section II we describe the problem and useful notation. In Section III we provide two simple bounds on the efficiency and MSE. Our main results for the adaptive and distributed cases are given in Sections IV and V, respectively. In Section VI we provide concluding remarks.

II. Problem Formulation

Let $f(x)$ be a symmetric and log-concave density function with a finite second moment σ^2. For $\theta \in \Theta$, denote by P_X the probability distribution with density $f(x-\theta)$. Therefore, P_X is an absolutely continuous log-concave distribution with mean θ and variance σ^2. Symmetry and log-concavity of $f(x)$ imply that P_X is strongly unimodal (has a unique local maxima) with its mode at $x=\theta$ [30]. We further assume that the *parameter space* Θ is a closed interval of the real line.

In some situations it is useful to assume that θ is drawn once from the prior distribution π on Θ. In this case we assume that π is an absolutely continuous distribution, and denote its density by $\pi(\theta)$, i.e., $\pi(d\theta)=\pi(\theta)d\theta$.

The random variables $X_1,\ldots,X_n$ represent n independent samples from P_X. We are interested in estimating θ from a set of n binary messages $B_1,\ldots,B_n$, obtained from $X_1,\ldots,X_n$ under three possible scenarios illustrated in Figure 1:

(i) Centralized $B_i(X_1,\ldots,X_n)$, $i=1,\ldots,n$.
(ii) Adaptive $B_i(X_i,B_1,\ldots,B_{i-1})$, $i=2,\ldots,n$.
(iii) Distributed $B_i(X_i)$, $i=1,\ldots,n$.

The performance of an estimator $\hat{\theta}_n \triangleq \hat{\theta}_n(B^n)$ in any of these cases is measured according to the mean squared error (MSE) risk:

$$R_n \triangleq \mathbb{E}\left(\hat{\theta}_n-\theta\right)^2, \tag{1}$$

where the expectation is taken with respect to the distribution of $X_1,\ldots,X_n$ and, whenever available, a prior distribution $\pi(\theta)$ over Θ. The main problems we consider in this paper are the minimal value of (1), as a function of n and $f(x)$, under different choices of the encoding functions in cases (i), (ii), and (iii).

We give particular attention to the ARE of estimators with respect to an asymptotically normal efficient estimator that is not subject to the bit constraint. Specifically, let $\{a_n, n\in\mathbb{N}\}$ be a sequence such that

$$\sqrt{a_n}\left(\hat{\theta}_n-\theta\right) \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

Then the ARE of $\hat{\theta}_n$ with respect to an unconstrained efficient estimator for θ is defined as [31, Def. 6.6.6]

$$\mathsf{ARE}(\hat{\theta}_n) \triangleq \lim_{n\to\infty}\frac{a_n}{n}.$$

Note that in the special case where there exists $V\in\mathbb{R}$ such that

$$a_n\mathbb{E}\left(\hat{\theta}_n-\theta\right)^2 = V+o(1),$$

with $o(1)\to 0$ as $n\to\infty$, then the ARE of $\hat{\theta}_n$ is finite and equals σ^2/V.

In addition to the notation above, we also denote by $F(x)$ the cumulative distribution function of X_i, and define

$$\eta(x) \triangleq \frac{f^2(x)}{F(x)(1-F(x))} = \frac{f(x)f(-x)}{F(x)F(-x)}, \tag{2}$$

where the last equality is due to symmetry of $f(x)$. We note that

$$\eta(x) = h(x)h(-x) = f(x)\left(h(x)+h(-x)\right), \tag{3}$$

where

$$h(x) \triangleq \frac{f(x)}{1-F(x)} = \frac{f(x)}{F(-x)}$$

is the *hazard* function (a.k.a. *failure rate* or *force of mortality*), which is a monotone increasing function since $f(x)$ is log-concave [32]. For $f(x)$ the normal density, it is shown in [33] and [34] that $\eta(x)$ is a strictly decreasing function of $|x|$, as illustrated in Fig. 2. In this paper we only consider the normal distribution and other log-concave symmetric distributions for which this property of $\eta(x)$ holds. Specifically, we require the following:

Assumptions 1: $\eta(x)$ has a unique maximum (at the origin) and is a non-increasing function of $|x|$.
Under this assumption we have,

$$4f^2(x) \leq \eta(x) \leq \eta(0),$$

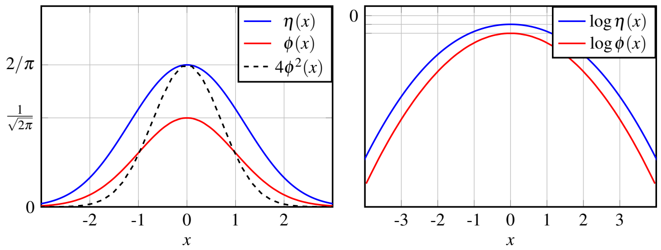

Fig. 2. The function $\eta(x) = f^2(x)/F(x)F(-x)$ for $f(x) = \phi(x)$ the standard normal density.

where $\eta(0) = 4f^2(0)$ is the asymptotic variance of the sample median. Combined with log-concavity of $f(x)$, Assumption 1 implies that $\eta(x)$ vanishes as $|x| \to \infty$.

Assumption 1 is satisfied, for example, by the generalized normal distributions with a shape parameter between 1 and 2 (including normal and Laplace distributions). Symmetric log-concave distributions that do not satisfy Assumption 1 include the uniform distribution and the generalized normal distribution with shape parameter greater than 2.

III. Consistent Estimation and Off-the-shelf Bounds

We next settle such skepticism by deriving lower and upper bounds to the relative efficiency under setting (iii).

In this section, we derive lower and upper bounds on the relative efficiency under setting (iii) using known results. These bounds establish the following facts:

I. A consistent estimator with an asymptotically normal distribution always exists in setting (iii), and hence in setting (ii).
II. For the normal distribution, the ARE in setting (iii) is at most $3/4$. Namely, under setting (iii), all estimators are strictly inferior compared to the sample mean.

A. *Consistent Estimation*

Fix $\theta_0 \in \mathbb{R}$ and define the ith message by

$$B_i = \mathbf{1}_{X_i > \theta_0},$$

where $\mathbf{1}_A$ is the indicator of the event A. We have

$$P_n \triangleq \frac{1}{n}\sum_{i=1}^{n} B_i \overset{a.s.}{\to} F(\theta - \theta_0),$$

so that

$$\hat{\theta}_n = \theta_0 + F^{-1}(P_n) \tag{4}$$

is a consistent estimator for θ in the distributed setting of Figure 1-(iii), where we note that $F(x)$ is invertible over the support of $f(x)$ which is a connected set due to log-concavity. Furthermore, the variance of P_n is $F(\theta-\theta_0)F(\theta_0-\theta)$, and hence the delta method implies that $\hat{\theta}_n$ is asymptotically normal with variance

$$\frac{F(\theta-\theta_0)F(\theta_0-\theta)}{f^2(\theta-\theta_0)} = \frac{1}{\eta(\theta-\theta_0)}. \tag{5}$$

In particular, the ARE of $\hat{\theta}_n$ equals $\eta(\theta-\theta_0)\sigma^2$. In other words, for a prescribed accuracy, $\hat{\theta}_n$ of (4) estimates θ with sample size that is $\eta(\theta-\theta_0)\sigma^2$ times the samples size required for the sample mean.

Assumption 1 implies that for all n large enough the ARE of $\hat{\theta}_n$ of (4) is never greater than $\eta(0)$, which is the ARE of the sample median. This ARE is attained only when $\theta_0 = \theta$, whereas θ is apriori unknown. Since $\eta(x)$ vanishes as $|x| \to \infty$, the ARE of $\hat{\theta}_n$ may be very small when θ is away from θ_0. As an example, when $f(x)$ is a normal density, the ARE of $\hat{\theta}_n$ is less than ≈ 0.14 when θ_0 is 2 standard deviations from θ. Therefore, the estimator $\hat{\theta}_n$ has little practical value unless the radius of Θ is small compared to the standard deviation.

It is suggested that we can attain lower variance in estimation by updating the threshold value θ_0 in (4) after observing one or a batch of the single bit messages. Schemes with such adaptation fall within the adaptive setting of Figure 1-(ii) which we consider in Section IV.

B. Multiterminal Source Coding

The CEO setting considers the estimation of a sequence $\theta_1, \theta_2 \ldots$, where a noisy version of each θ_i is available at n terminals. At each terminal i, an encoder observes the k noisy samples

$$X_{i,j} = \theta_j + Z_{i,j}, \qquad j = 1,\ldots,k, \qquad i = 1,\ldots,n,$$

and transmits $R_i k$ bits to a central estimator [19]. The central estimator produces an estimates $\hat{\theta}_1,\ldots,\hat{\theta}_k$ with the goal of minimizing the average MSE:

$$MSE_{\mathsf{CEO}} = \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}\left[\left(\theta_j - \hat{\theta}_j\right)^2\right].$$

Note that any distributed encoding scheme using one-bit per sample can be replicated k times and thus leads to a legitimate encoding and estimation scheme for the CEO problem with $R_1 = \ldots = R_n = 1$. It follows that, assuming that θ is drawn once from the prior $\pi(d\theta)$, our mean estimation problem from one-bit samples under distributed encoding corresponds to the CEO setting with $k = 1$ realization of θ observed under noise at n different locations, and communicated at each location using an encoder sending a single bit. Consequently, a lower bound on the MSE in estimating θ in the distributed encoding setting is given by the minimal MSE in the CEO setting as $k \to \infty$. Note that the difference between the CEO setting and ours lays in the privilege of each of the encoders to describe k realizations of θ using k bits with MSE averaged over these realizations, rather than a single realization using a single bit in ours.

When the prior on θ and the noise corrupting it at each location are Gaussian, the optimal encoding scheme and its asymptotic risk as $k \to \infty$ were fully characterized in [22]. Furthermore, the work of [35] provided an expression for MSE attained in the CEO setting under Gaussian priors. Adapting to our setting, this expression provides the following proposition:

Proposition 1: Assume that $\Theta = \mathbb{R}$ and $\pi(\theta) = \mathcal{N}(0, \sigma_\theta^2)$. Then any estimator $\hat{\theta}_n$ of θ in the distributed setting satisfies

$$n\mathbb{E}\left(\theta - \theta_n\right)^2 \geq \frac{4}{3}\sigma^2 + O(n^{-1}), \tag{6}$$

where the expectation is with respect to θ and X^n.

Proof: See Appendix A. □

From the formulation of the CEO problem, it follows that the difference between the MSE lower bound (6) and the actual MSE in the distributed setting (case (iii)) is attributed to the ability to perform coding over blocks. Namely, each encoder in the CEO setting may encode an arbitrary number of k independent realizations of θ using k bits, versus only one realization with one bit in ours. In other words, it is the ability to exploit the geometry of a high-dimensional product probability space that distinguishes between the CEO problem with one bit per encoder on average and the mean estimation problem from one-bit measurements in the distributed setting.

IV. Adaptive Estimation

The first main results of this paper, as described in Theorem 2 below, states that the ARE of any adaptive estimator cannot be larger than $\eta(0)\sigma^2$, which is the ARE of the median of the sample $X_1,\ldots,X_n$. We also provide a particular adaptive estimation scheme that attains this maximal efficiency.

A. Limited efficiency in the adaptive setting

Our first result asserts that the ARE of any adaptive encoding and estimation scheme is bounded from above by $\eta(0)\sigma^2$.

Theorem 2 (maximal relative effeciency): Let $\hat{\theta}_n$ be any estimator of θ in the adaptive setting of Figure 1-(ii). Assumes that $\pi(\theta)$ converges to zero at the endpoints of the interval Θ. Then

$$n\mathbb{E}\left[(\theta-\hat{\theta}_n)^2\right] \geq \frac{n}{4f^2(0)n+I_0},$$

where

$$I_0 = \mathbb{E}\left(\frac{d}{d\theta}\log\pi(\theta)\right)^2$$

is the Fisher information with respect to a location model in θ.

Sketch of Proof: The main idea in the proof is to bound from above the Fisher information of any set of n one-bit messages with respect to θ. Once this bound is achieved, the result follows by using the van-Trees inequality which bounds from below the risk of any estimator of θ by the inverse of the expected value of the aforementioned Fisher information plus I_0. The details are in Appendix C.

Theorem 2 implies that any estimator $\hat{\theta}_n$ from any adaptive encoding scheme satisfies

$$n\mathbb{E}\left[(\theta-\theta_n)^2\right] \leq \frac{1}{4f^2(0)} + O(n^{-1}),$$

and

$$\mathsf{ARE}(\hat{\theta}_n) \leq 4f^2(0)\sigma^2 = \eta(0)\sigma^2.$$

Next, we present an adaptive encoding and estimation scheme that attains the maximal ARE of $\eta(0)\sigma^2$.

B. Asymptotically optimal estimator

Let $\{\gamma_n, n\in\mathbb{N}\}$ be a strictly positive sequence. Consider the following estimator $\hat{\theta}_n$ for θ:

$$\theta_n = \theta_{n-1} + \gamma_n B_n, \quad n = 1,2,\ldots, \tag{7}$$

where

$$B_n = B_n(X_n, \theta_{n-1}) = \operatorname{sgn}(X_n - \theta_{n-1}). \tag{8}$$

Define the nth step estimation as

$$\bar{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i. \tag{9}$$

We have the following results:

Theorem 3: Consider the sequence $\left\{\bar{\theta}_n, n\in\mathbb{N}\right\}$ defined by (9).

(i) Assume that $\{\gamma_n, n\in\mathbb{N}\}$ satisfies

$$\begin{cases} \frac{\gamma_n-\gamma_{n+1}}{\gamma_n} = o(\gamma_n), \\ \sum_{n=1}^{\infty}\frac{\gamma_n^{(1+\lambda)/2}}{\sqrt{n}} < \infty, \quad \text{for some } 0<\lambda\leq 1 \end{cases} \tag{10}$$

(e.g., $\gamma_n = n^{-\beta}$ for $\beta\in(0,1)$). Then

$$\sqrt{n}\left(\bar{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\eta(0)}\right).$$

(ii) Assume in addition that $f(x)$ is continuously differentiable with finite Fisher information for location

$$I_\theta = \int_{\mathbb{R}}\left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx < \infty.$$

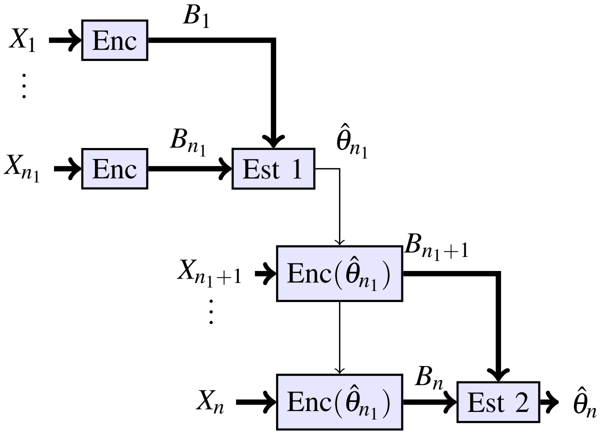

Fig. 3. Distributed encoding with a single interaction: The estimation obtained from the first n_1 bits in a distributed manner is to obtain another $n-n_1$ bits in a distributed manner.

For $\theta \in \Theta$, $h \in \mathbb{R}$ and n large enough such that $\theta + h/\sqrt{n} \in \Theta$, let $P^n_{\theta,h,n}$ be the measure with density $f(x-\theta-h/\sqrt{n})$. Then

$$\sqrt{n}\left(\bar{\theta}_n - \theta\right) \overset{d}{\to} \mathcal{N}\left(h, \frac{1}{\eta(0)}\right), \tag{11}$$

under $P^n_{\theta,h,n}$.

(iii) Assume that, in addition to (10), $\{\gamma_n, n \in \mathbb{N}\}$ satisfies

$$\begin{cases} \gamma_n = o(n^{-2/3}), \\ \sum_{n=1}^{\infty} \gamma_n = \infty. \end{cases} \tag{12}$$

(e.g., $\gamma_n = n^{-\beta}$ with $\beta \in (2/3,1)$). Then

$$\lim_{n\to\infty} n\mathbb{E}\left[\left(\bar{\theta}_n - \theta\right)^2\right] = \frac{1}{\eta(0)}.$$

Proof: See Appendix D.

Theorem 3 implies that the estimator $\hat{\theta}_n$, defined by (9) and (7), attains the maximal ARE as established by Theorem 2. Furthermore, it is locally minimax in the sense that it attains the minimal MSE for all local alternatives of θ. The update step (7) can be seen as a gradient descent step for the function $x \to |x|$ at the point $x = X_n - \theta_{n-1}$. Consequently, the procedure above is known as averaged stochastic gradient descent for minimizing $x \to |x|$ given the data $X_1,\ldots,X_n$. The minimal value of this optimization is the sample median, and Theorem 3 provides conditions for the sequence of gradient steps so that the algorithm converges to this minimum.

In the encoding and estimating procedure (7) and (9), each one-bit message B_n depends on its private sample as well as the current gradient descent estimate θ_{n-1}. In this sense, each encoder in this algorithm interacts with previous one by using the current estimate. As we explain next, it is possible to obtain the optimal efficiency of $\eta(0)\sigma^2$ with only one round of interactions among the encoders.

C. Optimality using Single Threshold Adaptation

In Section III we considered an estimator that is based on messages of the form

$$B_i = \mathbf{1}_{X_i > \theta_0}, \qquad i = 1,\ldots,n,$$

and showed that it is asymptotically normal with variance $1/\eta(\theta-\theta_0)$. We now show that a similar encoding leads to an asymptotically normal estimator with the minimal variance $1/\eta(0)$, provided we allow at least once to update the threshold value θ_0. In this procedure we separate the sample into two disjoint sets: $X_1,\ldots,X_{n_1}$ and

$X_{n_1+1},\dots,X_n$ for some $n_1 < n$. We first use the estimator (4) to obtain an estimate $\hat{\theta}_{n_1}$ based on $B_1,\dots,B_{n_1}$, and then use $\hat{\theta}_{n_1}$ as the new threshold value to obtain messages $B_{n_1+1},\dots,B_n$. Figure 3 illustrates a diagram of this procedure. The specific encoding and estimation scheme, as well as its asymptotic performance, are given by the following theorem:

Theorem 4: For $i = 1,\dots,n$ set

$$B_i = \begin{cases} \mathbf{1}_{X_i \ge \theta_0} & i = 1,\dots,n_1, \\ \mathbf{1}_{X_i \ge \hat{\theta}_{n_1}} & i = n_1+1,\dots,n, \end{cases}$$

where $n = n_1 + n_2$ and

$$\hat{\theta}_{n_1} \triangleq \theta_0 + F^{-1}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} B_i\right).$$

Let

$$\hat{\theta}_{n_2} = \hat{\theta}_{n_1} + F^{-1}\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n_2} B_i\right)$$

and assume that $n_1(n) \to \infty$ and $n_1(n)/n \to 0$. Then:

$$\sqrt{n}\left(\hat{\theta}_{n_2} - \theta\right)^2 \xrightarrow{d} \mathcal{N}\left(0, 1/\eta(0)\right).$$

Proof: For $t \in \mathbb{R}$, set

$$p_{n_2}(t) \triangleq \frac{1}{n_2}\sum_{i=n_1+1}^{n} \mathbf{1}_{X_i \ge t}.$$

From the central limit theorem

$$\begin{aligned} &\sqrt{n}\left(p_{n_2}(t) - F(\theta - t)\right) = \sqrt{\frac{n}{n_2}}\sqrt{n_2}\left(p_{n_2}(t) - F(\theta - t)\right) \\ &\xrightarrow{d} \mathcal{N}\left(0, V(t)\right), \end{aligned}$$

where

$$V(t) \triangleq F(\theta - t)F(t - \theta).$$

Applying the delta method to $p_{n_2}(t)$ with $g(x) = F^{-1}(x)$ we obtain

$$\begin{aligned} &\sqrt{n}\left(t + F^{-1}(\hat{p}_{n_2}(t)) - \theta\right) \\ &= \sqrt{n}\left(g(\hat{p}_{n_2}(t)) - g\left(F(\theta - t)\right)\right) \\ &\xrightarrow{d} \mathcal{N}\left(0, 1/\eta(\theta - t)\right). \end{aligned}$$

By the law of large numbers we also have

$$p_{n_1} \triangleq \frac{1}{n_1}\sum_{i=1}^{n_1} B_i \xrightarrow{a.s.} F(\theta - \theta_0),$$

so that $\hat{\theta}_{n_1}$ converges almost surely to θ as n goes to infinity, and thus

$$\eta(\hat{\theta}_{n_1} - \theta) \xrightarrow{a.s.} \eta(0).$$

By Slutsky's theorem we get

$$\begin{aligned} &\sqrt{n}\left(\hat{\theta}_{n_2} - \theta\right) \\ &= \sqrt{n}\left(\hat{\theta}_{n_1} + F^{-1}(\hat{p}_{n_2}(\hat{\theta}_{n_1})) - \theta\right) \\ &\xrightarrow{d} \mathcal{N}\left(0, 1/\eta(0)\right). \end{aligned}$$

□

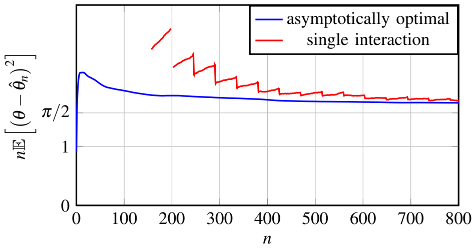

Fig. 4. Normalized empirical risk versus number of samples n for 10,000 Monte Carlo trials with $f(x)$ the standard normal density. In each trial, θ is chosen uniformly over the interval $(-1.64, 1.64)$. The single interaction strategy uses $n_1 = \lfloor \sqrt{n} \rfloor$ samples for its first stage.

V. Distributed Estimation

We now consider the distributed encoding setting in Figure 1-(iii). In this setting each one-bit message B_i is only a function of its private sample X_i, and hence B_i characterized by its *detection region*, defined as

$$A_i = \{x \in \mathbb{R} : B_i(x) = 1\}.$$

Consequently, B_i is of the form

$$B_i = \begin{cases} 1 & X_i \in A_i, \\ -1 & X_i \notin A_i, \end{cases} \quad i \in \mathbb{N},$$

where the detection region A_i is a Borel set that is independent of $X_1, \ldots, X_n$.

As a first step, we provide conditions under which the messages $B_1, B_2, \ldots$ define a local asymptotic normal family.

Theorem 5: For $n \in \mathbb{N}$ and $A_n \subset \mathbb{R}$, define

$$L_n(A_1, \ldots, A_n; \theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \frac{\left(\frac{d}{d\theta}\mathbb{P}(X_i \in A_i)\right)^2}{\mathbb{P}(X_i \in A_i)\left(1 - \mathbb{P}(X_i \in A_i)\right)}. \tag{13}$$

Consider the following conditions:

(i) The pdf $f(x)$ of $X_n - \theta$ is a log-concave, differentiable and symmetric density function such that $\eta(x)$ is unimodal.

(ii) A_n is a finite union of disjoint intervals.

(iii) The limit

$$\kappa(\theta) \triangleq \lim_{n\to\infty} L_n(A_1, \ldots, A_n; \theta) \tag{14}$$

exists.

For $i = 1, \ldots, n$ set

$$B_n = \begin{cases} 1 & X_n \in A_n, \\ -1 & X_n \notin A_n. \end{cases}$$

For any θ, $f(x)$ and a sequence of sets $A_1, A_2, \ldots$ such that (i)-(iii) hold, and any $h \in \mathbb{R}$, we have

$$\log \frac{\mathbb{P}_{\theta+h/\sqrt{n}}(B_1, \ldots, B_n)}{\mathbb{P}_\theta(B_1, \ldots, B_n)} \xrightarrow{d} \mathcal{N}\left(-\frac{1}{2}h^2\kappa(\theta), h^2\kappa(\theta)\right).$$

Proof: See Appendix F. □

Theorem 5 provides conditions under which $B_1,\ldots,B_n$ defines a LAN family with a precision parameter given by the limit in (14). Among these conditions, (iii) is arguably the strongest and hardest to verify. As we show in Section V-B below, this condition is satisfied, for example, when $A_1,\ldots,A_n$ are half lines whose starting points are drawn independently from some probability measure on $\mathbb{R}$. Similar ideas imply that condition (iii) holds whenever we choose the intervals consisting each A_i according to some pre-specified distribution.

An important conclusion of Theorem 5 follows from the local asymptotic minimax property of estimators in LAN models (e.g. [36]):

Corollary 6: Let $\hat{\theta}_n$ be an estimator of $\theta \in \Theta$ from $B_1,\ldots,B_n$ with detection regions $A_1,\ldots,A_n$ such that conditions (i)-(iii) of Theorem 5 hold. Then for any bounded, symmetric, and quasi-convex function L,

$$\liminf_{c\to\infty}\liminf_{n\to\infty} \sup_{\tau:\,|\theta-\tau|\leq\frac{c}{\sqrt{n}}} \mathbb{E}\left[L\left(\sqrt{n}(\hat{\theta}_n-\tau)\right)\right] \\ \geq \mathbb{E}\left[L(Z/\sqrt{\kappa(\theta)})\right],$$

where $Z \sim \mathcal{N}(0,1)$. In particular, for $L(x)=x^2$,

$$\liminf_{c\to\infty}\liminf_{n\to\infty} \sup_{\tau:\,|\theta-\tau|\leq\frac{c}{\sqrt{n}}} n\mathbb{E}\left(\hat{\theta}_n-\tau\right)^2 \geq 1/\kappa(\theta).$$

Corollary 6 says that when the messages define a LAN model, no estimator can attain MSE smaller than $1/\kappa(\theta)n+O(1/n)$ where $\kappa(\theta)$ is the precision parameter of the model at θ. This fact poses the upper bound of $\kappa(\theta)\sigma^2$ for the ARE of estimators in such models.

Next, we show that under LAN no estimator can attain the optimal ARE of $\eta(0)\sigma^2$ uniformly for all $\theta\in\Theta$.

A. Non-existence of a Uniformly Optimal Strategy

We now show that under LAN models, the optimal minimal risk $1/\eta(0)$ can only be attained at a finite number of points within Θ. This fact implies in particular that, unlike in the adaptive setting, no distributed estimation scheme has ARE of $\eta(0)\sigma^2$ for all $\theta\in\Theta$.

Theorem 7: Under conditions (i)-(iii) in Theorem 5, assume that each A_i is a union of at most K intervals. The number of points $\theta\in\Theta$ satisfying $\kappa(\theta)=\eta(0)$ is at most $2K$.

Proof: See Appendix **??** □

We next consider the case where each detection region is a half-open interval, i.e., the ith message is obtained by comparing X_i against a single threshold. As we explain next, the existence of a density for the sequence of thresholds is enough to establish local asymptotic normality and leads to a closed form expression for the precision parameter and the ARE.

B. Threshold Detection

Assume now that each B_i is of the form

$$B_i = \operatorname{sgn}(t_i - X_i) = \begin{cases} 1 & X_i < t_i, \\ -1 & X_i > t_i, \end{cases} \tag{15}$$

where $t_i \in \mathbb{R}$ is the *threshold* of the ith encoder. In other words, the detection region of B_i is $A_i = (t_i,\infty)$ and $\mathbb{P}(X_i \in A_i) = F\left(B_i(t_i-\theta)\right)$. It follows that

$$L_n(A_1,\ldots,A_n;\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(f(t_i-\theta)\right)^2}{F\left(t_i-\theta\right)F\left(\theta-t_i\right)} = \frac{1}{n}\sum_{i=1}^{n}\eta\left(t_i-\theta\right). \tag{16}$$

A natural condition for the existence of the limit (16) as $n \to \infty$ is that the empirical distribution of the threshold values converges to a probability measure. Specifically, for an interval $I \subset \mathbb{R}$ define

$$\lambda_n(I) = \frac{\text{card}\left(I \cap \{t_1, t_2, \ldots\}\right)}{n}.$$

Theorem 5 implies:

Corollary 8: Let $\{t_n\}_{n=1}^{\infty}$ be a sequence of threshold values such that λ_n converges (weakly) to a probability measure $\lambda(dt)$ on $\mathbb{R}$. Then $\{B_i = \text{sgn}(X_i - t_i)\}_{i=1}^{n}$ is a LAN family with precision parameter

$$\kappa(\theta) = \int_{\mathbb{R}} \eta(t-\theta)\lambda(dt).$$

The condition that λ_n converges to a probability measure is satisfies, for example, whenever the $t_1, \ldots, t_n$s are drawn independently from a probability distribution $\lambda(dt)$ on $\mathbb{R}$.

Due to local asymptotic normality of $\{B_n\}_{n=1}^{\infty}$, the maximum likelihood estimator (ML) of θ from $B_1, \ldots, B_n$, denoted here by $\hat{\theta}_n^{ML}$, is local asymptotic minimax in the sense that

$$\sqrt{n}\left(\hat{\theta}_n^{ML} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, 1/\kappa(\theta)\right).$$

It follows that when the density of the threshold values converges to a probability measure, the ARE of the ML estimator is $\kappa(\theta)\sigma^2$, and this ARE is maximal with respect to all local alternative estimators for θ. We note that $\hat{\theta}_n^{ML}$ is given by the root of

$$\sum_{i=1}^{n} B_i \frac{f(t_i - \theta)}{F(B_i(t_i - \theta))}, \tag{17}$$

which is the derivative of the log-likelihood function. This root is unique since the log-likelihood function is concave. Furthermore, for any $n \in \mathbb{R}$, we have that $\hat{\theta}_n^{ML} \in [t_{(1)}, t_{(n)}]$ where $t_{(i)}$ denotes the ith element of $\{t_1, t_2 \ldots\}$. Therefore, if $\{t_1, t_2 \ldots\}$ is bounded (for example $\{t_1, t_2 \ldots\} \subset \Theta$), then

$$\lim_{n\to\infty} n\left(\hat{\theta}_n^{ML} - \theta\right) = 1/\kappa(\theta),$$

so that the ML estimator attains the local asymptotic MSE of Corollary 6.

Since $\eta(x)$ attains its maximum at the origin, we conclude that

$$\kappa(\theta) \le \sup_{t\in\mathbb{R}} \eta\left(t-\theta\right) = \eta(0).$$

This upper bound on $\kappa(\theta)$ implies that the ARE of any distributed estimator based on a sequence of threshold detectors does not exceed $\eta(0)\sigma^2$, a fact that agrees with the lower bound under adaptive estimation derived in Theorem 2. This upper bound on $\kappa(\theta)$ is attained only when λ is the mass distribution at θ. Since θ is apriori unknown, we conclude that estimation in the distributed setting using threshold detection is strictly sub-optimal compared to the adaptive setting. In other words, the ability to choose the threshold values in an adaptive manner based on previous messages strictly improves relative efficiency compared to a non-adaptive threshold selection.

We conclude this section by considering the density of the threshold values that maximizes the ARE $\kappa(\theta)$ under the worst choice of $\theta \in \Theta$.

C. Minimax Threshold Density

The distribution $\lambda(dt)$ that maximizes $\kappa(\theta)$, and thus minimizes $1/\kappa(\theta)$, over the worst choice of θ in $\Theta = [-T, T]$ is given as the solution to the following optimization problem:

$$\begin{aligned} &\text{maximize} \quad \inf_{\theta\in[-T,T]} \int \eta(t-\theta)\lambda(dt) \\ &\text{subject to} \quad \lambda(dt) \ge 0, \quad \int \lambda(dt) \le 1. \end{aligned} \tag{18}$$

The objective function in (18) is concave in $\lambda(dt)$ and hence this problem can be solved using a convex program. We denote by $\kappa^{\star}(T)$ the maximal value of (18) and by $\lambda^{\star}(dt)$ the density that achieves this maximum.

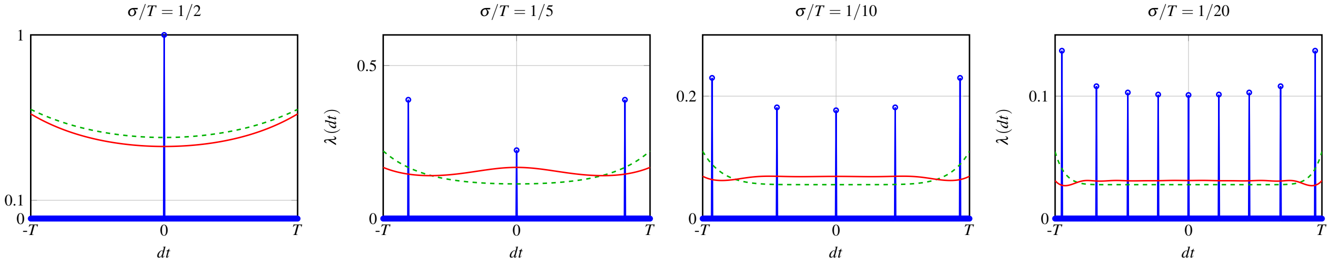

Fig. 5. Optimal threshold density $\lambda^{\star}(dt)$ (blue) that maximizes the ARE for $f(x) = \mathcal{N}(\theta, \sigma^2)$ and $\theta \in \Theta = [-T, T]$. The continuous curve (red) represents the reciprocal of the asymptotic risk for at a fixed $\theta \in \Theta$ under the optimal density, so the minimax risk is the inverse of its minimal value. The dashed curve (green) is the reciprocal of the asymptotic risk for a fixed θ under a uniform distribution of threshold values over Θ, hence its minimal value is the inverse of (19).

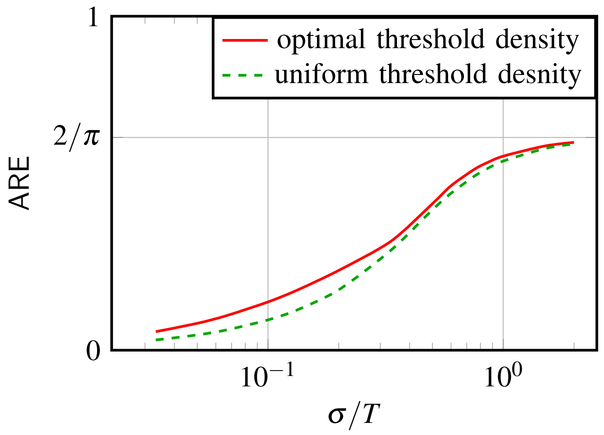

Fig. 6. Minimax ARE versus σ/T for $f(x) = \mathcal{N}(\theta, \sigma^2)$ and $\theta \in \Theta = [-T, T]$. The dashed curve (green) is the ARE under a uniform threshold density over Θ given by $K_{\mathsf{unif}}\sigma^2$, where κ_{unif} is given by (19).

Figure 5 illustrates an approximating to $\lambda^{\star}(dt)$ obtained by solving a discretized version of (18) for the case when $f(x)$ is the normal density with variance σ^2. The minimal asymptotic risk $\kappa^{\star}(\theta)$ obtained this way is illustrated in Fig. 6 as a function of the support size T. Also illustrated in these Figures is κ_{unif} which is the precision parameter corresponding to threshold values uniformly distribution over $\Theta = [-T, T]$, namely

$$\begin{aligned} \kappa_{\mathsf{unif}} &\triangleq \min_{\theta \in [-T,T]} \frac{1}{2T} \int_{-T}^{T} \eta\left(t - \theta\right) dt \\ &= \frac{1}{2T} \int_{-T}^{T} \eta\left(t \pm T\right) dt = \frac{1}{2T} \int_{0}^{2T} \eta(t) dt. \end{aligned} \tag{19}$$

From Corollary 8, we conclude that the ARE under a uniform distribution is $\kappa_{\mathsf{unif}}\sigma^2$.

We consider now the problem of minimizing the asymptotic Bayes risk $R_{\pi,\lambda} \triangleq \mathbb{E}K^{-1}(\theta)$ over all probability measures $\lambda(dt)$ with support in $\mathbb{R}$. This optimization problem can be written as follows:

$$\begin{aligned} \text{minimize} \quad & R_{\pi,\lambda} = \int \frac{\pi(d\theta)}{\int \eta\left(t - \theta\right) \lambda(dt)}. \\ \text{subject to} \quad & \lambda(dt) \geq 0, \quad \int \lambda(dt) = 1. \end{aligned} \tag{20}$$

We denote by $R_{\pi}^{\star}$ the minimal value of the objective function in (20). Since the function $x \to 1/x$ is convex for positive values, (20) defines a convex optimization problem in λ whose solution depends on the prior π. The solution to this problem is approximated by considering λ and π over a discrete set. On Figure 7 we illustrate

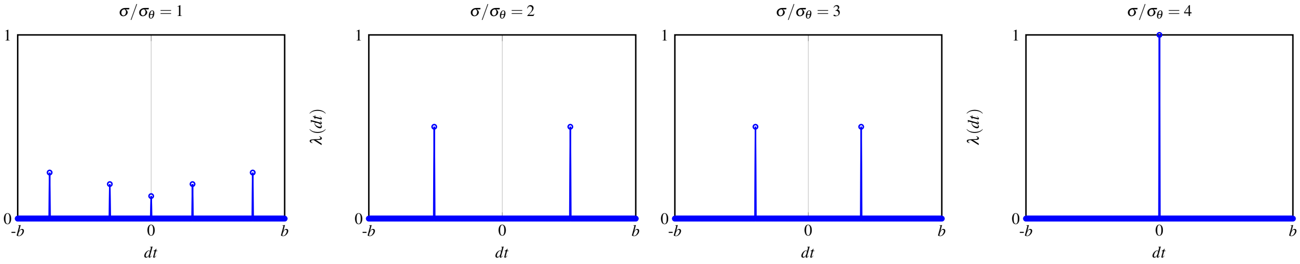

Fig. 7. Optimal threshold density $\lambda^\star(dt)$ that minimizes the asymptotic Bayes risk (20) for a uniform prior with $\sigma/\sigma_\theta = 1,2,3,4$, where $\sigma_\theta^2 = b^2/3$ is the variance of the prior.

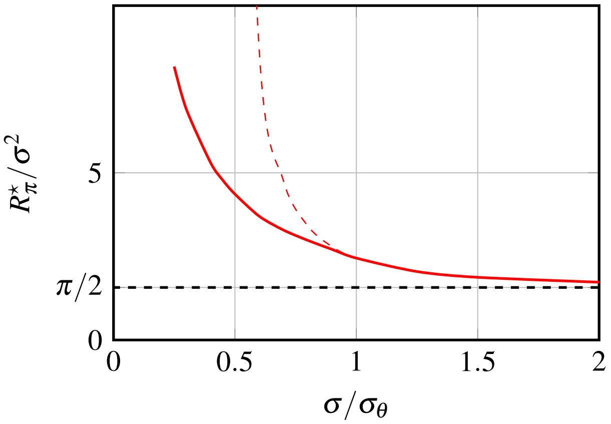

Fig. 8. Asymptotic Bayes risk $R^\star_\pi$ in estimating the mean of a normal distribution ($P_X = \mathcal{N}(\theta,\sigma^2)$under an optimal threshold distribution $\lambda^\star$ for π the uniform distribution over $\Theta = [-0.5, 0.5]$. The distribution $\lambda^\star$ is the minimizer of (20). It is illustrated for various cases in Fig. 7. The dashed curve represents the upper bound (21).

the solution to (20) for the case of where $f(x)$ is the normal distributon ($P_X = \mathcal{N}(\theta,\sigma^2)$) and the prior on θ is uniform over $\Theta = [-b,b]$. Figure 8 illustrates the corresponding Bayes risk.

As can be seen in Fig. 7 for the case $P_X = \mathcal{N}(\theta,\sigma^2)$ and a uniform π, when the radius of Θ is small compared to σ, the optimal distribution $\lambda^\star$ is a mass distribution. In this case, the ML estimator reduces to the estimator $\hat{\theta}_n$ of (4). As the following proposition shows, the Bayes risk for this choice of λ is maximal, and thus provides an upper bound on the Bayes risk under any λ. (20).

Proposition 9: For any prior $\pi(d\theta)$ and θ_0 in the support of $\eta(x)$ we have

$$R^\star_\pi \leq \int \frac{\pi(d\theta)}{\eta\left(\theta_0 - \theta\right)}, \tag{21}$$

Furthermore, assuming that $\theta_0 = \mathbb{E}\theta$ and that π has a finite third moment σ_θ^3, we have:

$$R^\star_\pi \leq \frac{1}{4f^2(0)} + \left(\frac{1}{4f^2(0)}\frac{-f''(0)}{f(0)} - 1\right)\sigma_\theta^2 + O(\sigma_\theta^3). \tag{22}$$

Proof: The function $x \to 1/x$ is convex for positive values, hence Jensen's inequality implies

$$\left(\int \eta\left(t-\theta\right)\lambda(dt)\right)^{-1} \leq \int \frac{\lambda(dt)}{\eta\left(t-\theta\right)}.$$

Therefore, the expected value of $\kappa^{-1}(\theta)$ satisfies

$$\mathbb{E}\frac{1}{\kappa(\theta)} \leq \int\int \frac{\pi(d\theta)\lambda(dt)}{\eta\left(t-\theta\right)}. \tag{23}$$

The bound (21) is obtained by taking λ to be a mass distribution at any θ_0 in the support of $\eta(x)$. Finally, (22) is obtained by expanding $1/\eta(x)$ to a third order Taylor series around zero and taking its expectation with respect to π at $x = \theta_0 - \theta$. □

We note that the function $1/\eta(x)$ is quasi-convex and symmetric around zero, so taking $\theta_0 = \mathbb{E}\theta$ minimizes the RHS of (21) among all θ_0 in the support of $\eta(x)$.

The bound (21) is not trivial as long as the integral in the RHS of (21) is finite, i.e., whenever the tail of $\pi(\theta)$ vanishes fast enough compared to $\eta(x)$. The expansion (22) implies that this bound becomes tight whenever the support of the optimal distribution is a mass distribution at $\mathbb{E}\theta$, in which case the expected value of $\kappa^{-1}(\theta)$ approaches $1/\eta(0) = 1/4f^2(0)$.

Example 1: In the normal case $P_X = \mathcal{N}(\theta, \sigma^2)$, the bound in (22) implies

$$\frac{R^\star_\pi}{\sigma^2} \leq \frac{\pi}{2} + \left(\frac{\pi}{2} - 1\right)\left(\frac{\sigma_\theta}{\sigma}\right)^2 + O\left(\frac{\sigma_\theta}{\sigma}\right)^3.$$

It follows that the ARE approaches its maximal value of $2/\pi$ whenever σ_θ/σ is small. The exact value of (21) in this case, as well as the Bayes ARE with the optimal threshold density $\lambda^\star$ for a uniform π, are illustrated in Fig. 8.

VI. Conclusions

We considered the MSE risk in estimating the mean of a symmetric and log-concave distribution from a sequence of bits, where each bit is obtained by encoding a single sample from this distribution. In an adaptive encoding setting, we showed that no estimator can attain asymptotic relative efficiency (ARE) larger than that of the median of the samples. We also showed that this bound is tight by presenting two adaptive encoding and estimation procedures that are as efficient as the median.

In the distributed setting we provided conditions for local asymptotic normality of the encoded samples, which implies asymptotic minimax bound on both the risk and ARE. We conclude that under such conditions, the optimal estimation performance derived for the adaptive case can only be attained over a finite number of points, i.e., no scheme is uniformly optimal in the distributed setting. We further considered the special case of messages obtained by comparing against a prescribed sequence of thresholds. We characterized the performance of the optimal estimator from such messages using the density of these thresholds, and consider the threshold density that minimizes the minimax risk.

APPENDIX

A. Proof of Proposition 1

First, we note that when $f(x) = \mathcal{N}(0,\sigma^2)$ and $\pi(\theta) = \mathcal{N}(0,\sigma_\theta^2)$, our distributed setting reduces to the Gaussian CEO problem with $R_1 = \ldots = R_n = 1$ and $k = 1$. Therefore, any MSE attained in our distributed setting is also attained by the CEO and thus the optimal CEO strategy (with arbitrary k) provides a lower bound to the MSE in our setting.

Consider the minimal MSE $D^\star$ in the Gaussian CEO setting with L observers and under a total sum-rate $R_\Sigma = R_1 + \ldots + R_L$ from [35, Eq. 10]:

$$R_\Sigma = \frac{1}{2}\log^+ \left[\frac{\sigma_\theta^2}{D^\star} \left(\frac{D^\star L}{D^\star L - \sigma^2 + D^\star \sigma^2/\sigma_\theta^2} \right)^L \right]. \tag{24}$$

For the special case of $R_\Sigma = n$ and $L = n$, we get

$$n = \frac{1}{2}\log_2 \left[\frac{\sigma_\theta^2}{D^\star} \left(\frac{D^\star n}{D^\star n - \sigma^2 + D^\star \sigma^2/\sigma_\theta^2} \right)^n \right]. \tag{25}$$

$D^\star$ satisfying (25) describing the MSE under an optimal allocation of the sum-rate $R_\Sigma = n$ among the n encoders. Therefore, this $D^\star$ provides a lower bound to the CEO MSE with $R_1 = \ldots, R_n = 1$ and hence a lower bound to the minimal MSE in estimating θ in the distributed setting. By considering $D^\star$ in (25) as $n \to \infty$, we see that

$$D^\star = \frac{4\sigma^2}{3n + 4\sigma^2/\sigma_\theta^2} + o(n^{-1}) = \frac{4\sigma^2}{3n} + o(n^{-1}).$$

We note that although the lower bound (6) was derived assuming the optimal allocation of n bits per observation among the encoders, this bound cannot be tightened by considering the CEO MSE while enforcing the condition $R_1 = \ldots = R_n = 1$. Indeed, an upper bound for the CEO MSE under the condition $R_1 = \ldots = R_n = 1$ follows from [37], and leads to

$$D^\star \leq \left(\frac{1}{\sigma_\theta^2} + \frac{3n}{4\sigma^2 + \sigma_\theta^2} \right)^{-1} = \frac{4\sigma^2}{3n} + \frac{\sigma_\theta^2}{3n} + O(n^{-2}),$$

which is equivalent to (6) when σ_θ is small.

B. Proof of Lemma 16

Denote

$$\delta_n \triangleq \delta_n(x_1,\ldots,x_n) \triangleq \sum_{k=1}^{n} (-1)^{k+1} f(x_k),$$

$$\Delta_n \triangleq \Delta_n(x_1,\ldots,x_n) \triangleq \sum_{k=1}^{n} (-1)^{k+1} F(x_k),$$

so

$$\eta(x) = \frac{(\delta_1(x))^2}{\Delta_1(x)(1-\Delta_1(x))} = \frac{(f(x))^2}{F(x)\,(1-F(x))^2}.$$

We use induction on $n \in \mathbb{N}$ to show that the LHS of (51) is bounded from above by $\max_i \eta(x_i)$. The case $n = 1$ is trivial. Assume that

$$\frac{(\delta_n)^2}{\Delta_n\,(1-\Delta_n)} \leq \max_i \eta(x_i) \tag{26}$$

for all integers up to $n = N-1$ and consider the case $n = N$. The maximal value of the LHS of (26) is attained for the same $(x_1,\ldots,x_N) \in \mathbb{R}^N$ that attains the maximal value of

$$g(x_1,\ldots,x_N) \triangleq 2\log\delta_N - \log\Delta_N - \log(1-\Delta_N),$$

The derivative of $g(x_1,\ldots,x_N)$ with respect to x_k is given by

$$\frac{\partial g}{\partial x_k} = \frac{2(-1)^{k+1}f'(x_k)}{\delta_N} - \frac{(-1)^{k+1}f(x_k)}{\Delta_N} + \frac{(-1)^{k+1}f(x_k)}{1-\Delta_N},$$

and we conclude that the gradient of g vanishes if and only if

$$\frac{f'(x_k)}{f(x_k)} = \frac{\delta_N}{2}\left(\frac{1}{\Delta_N} - \frac{1}{1-\Delta_N}\right), \quad k = 1,\ldots,N. \tag{27}$$

Since $f(x)$ is log-concave, symmetric, and differentiable, we may write $f(x) = e^{c(x)}$ where $c(x)$ is concave, symmetric, and differentiable. We have

$$\frac{f'(x)}{f(x)} = c'(x), \quad x \in \mathbb{R},$$

which is anti-symmetric, non-negative for $x < 0$, non-positive for $x > 0$, and non-increasing since $c(x)$ is concave. Therefore, if $c'(x_i) = c'(x_{i+1})$ for some $i = 1,\ldots,N-1$, then either (1) $x_i = x_{i+1}$ or (2) $c'(x)$ is the zero function. Since (2) violates the assumption that $f(x)$ is a density function, we conclude that $c'(x)$ is an injection. As a result, (27) is satisfied if and only if $x_1 = \ldots = x_N$. For odd N and $x_1 = \ldots = x_N$, the LHS of (26) equals $\eta(x_1) = \max_i \eta(x_i)$ hence the statement holds. For even N and any constant d, the limit of the LHS of (26) as $(x_1,\ldots,x_N) \to (d,\ldots,d)$ exists and equals zero. Therefore, the maximum of the LHS of (26) is not attained at the line $x_1 = \ldots = x_N$). We now consider the possibility that the LHS of (26) is maximized at the borders. That is, as one or more of the coordinates of $(x_1,\ldots,x_N)$ approaches $\pm\infty$, or $\pm\varepsilon$. As we assumed $x_1 \geq \ldots \geq x_N$, if $x_i = x_{i+1}$ for some i than their contribution to (26) is zero and thus this case reduces to the case $n = N-2$. A similar reduction holds if $x_N, x_{N-1} \to -\infty$, $x_1, x_2 \to \infty$, or x_i, x_{i+1} for some i. It is left to consider the cases:

(1) $x_N \to -\infty$.
(2) $x_1 \to \infty$.

Under case (1) we have

$$\lim_{x_N \to -\infty} \frac{\delta_N^2}{\Delta_N(1-\Delta_N)} = \frac{\left(\sum_{k=1}^{N-1}(-1)^{k+1}f(x_k)\right)^2}{\left(\sum_{k=1}^{N-1}(-1)^{k+1}F(x_k)\right)\left(1-\sum_{k=1}^{N-1}(-1)^{k+1}F(x_k)\right)},$$

which is smaller than $\max_i \eta_i(x_i)$ by the induction hypothesis. Under case (2) we have

$$\begin{aligned}
&\lim_{x_1\to\infty} \frac{\delta_N}{\Delta_N(1-\Delta_N)} \\
=&\frac{\left(\sum_{k=2}^{N}(-1)^{k+1}f(x_k)\right)}{\left(1+\sum_{k=2}^{N}(-1)^{k+1}F(x_k)\right)\left(1-1-\sum_{k=2}^{N}(-1)^{k+1}F(x_k)\right)} \\
=& \frac{\left(-\sum_{m=1}^{N}(-1)^{m+1}f(x'_m)\right)^2}{\left(1-\sum_{m=1}^{N-1}(-1)^{m+1}F(x'_m)\right)\left(\sum_{m=1}^{N-1}(-1)^{m+1}F(x'_m)\right)},
\end{aligned}$$

where $x'_m = x_{m+1}$. The last expression is also smaller than $\max_i \eta_i(x_i)$ by the induction hypothesis. □

C. Proof of Theorem 2

We first prove the following lemma:

Lemma 10: Let X be a random variable with a symmetric, log-concave, and continuously differentiable density function $f(x)$ such that $\eta(x)$ is unimodal. For a Borel measurable A set,

$$M(X) = \begin{cases} 1, & X \in A, \\ -1, & X \notin A. \end{cases}$$

Then the Fisher information of M with respect to θ is bounded from above by $\eta(0)$.

Proof of Lemma 10: When $f(x)$ is the normal density function, this lemma follows from [29, Thm. 3]. The proof below is based on a different techique than in [29], and is valid for any log-concave symmetric density satisfying Assumption 1.

The Fisher information of M with respect to θ is given by

$$\begin{aligned} I_\theta &= \mathbb{E}\left[\left(\frac{d}{d\theta} \log P(M|\theta) \right)^2 |\theta \right] \\ &= \frac{\left(\frac{d}{d\theta} P(M=1|\theta)\right)^2}{P(M=1|\theta)} + \frac{\left(\frac{d}{d\theta} P(M=-1|\theta)\right)^2}{P(M=-1|\theta)} \\ &= \frac{\left(\frac{d}{d\theta} \int_A f(x-\theta)\,dx\right)^2}{P(M=1|\theta)} + \frac{\left(\frac{d}{d\theta} \int_A f(x-\theta)\,dx\right)^2}{P(M=-1|\theta)} \\ &\overset{(a)}{=} \frac{\left(-\int_A f'(x-\theta)\,dx\right)^2}{P(M=1|\theta)} + \frac{\left(-\int_A f'(x-\theta)\,dx\right)^2}{P(M=-1|\theta)} \\ &= \frac{\left(\int_A f'(x-\theta)\,dx\right)^2}{P(M=1|\theta)\,(1-P(M=1|\theta))}, \\ &= \frac{\left(\int_A f'(x-\theta)\,dx\right)\left(\int_A f'(x-\theta)\,dx\right)}{\left(\int_A f(x-\theta)\,dx\right)\left(1-\int_A f(x-\theta)\,dx\right)}, \end{aligned} \tag{28}$$

where differentiation under the integral sign in (a) is possible since $f(x)$ is differentiable with continuous derivative $f'(x)$. Regularity of the Lebesgue measure implies that for any $\varepsilon > 0$, there exists a finite number k of disjoint open intervals $I_1, \ldots I_k$ such that

$$\int_{A \setminus \cup_{j=1}^k I_j} dx < \varepsilon,$$

which implies that for any $\varepsilon' > 0$, the set A in (28) can be replaced by a finite union of disjoint intervals without increasing I_θ by more than ε'. It is therefore enough to proceed in the proof assuming that A is of the form

$$A = \cup_{j=1}^k (a_j, b_j),$$

with $\infty \le a_1 \le \ldots a_k$, $b_1 \le b_k \le \infty$ and $a_j \le b_j$ for $j = 1, \ldots, k$. Under this assumption we have

$$\begin{aligned} \mathbb{P}(B_n = 1|\theta) &= \sum_{j=1}^k \mathbb{P}(X_n \in (a_j, b_j)) \\ &= \sum_{j=1}^k \left(F(b_j - \theta) - F(a_j - \theta)\right), \end{aligned}$$

so (28) can be rewritten as

$$\begin{aligned} &= \frac{\left(\sum_{j=1}^k f(a_j-\theta) - f(b_j-\theta)\right)^2}{\left(\sum_{j=1}^k F(b_j-\theta) - F(a_j-\theta)\right)} \\ &\times \frac{1}{1-\left(\sum_{j=1}^k F(b_j-\theta) - F(a_j-\theta)\right)} \end{aligned} \tag{29}$$

It follows from Lemma 16 that for any $\theta \in \mathbb{R}$ and any choice of the intervals endpoints, (29) is smaller than $4f^2(0)$. □

We now finish the proof of Theorem 2. In order to bound from above the Fisher information of any set of n one-bit messages with respect to θ, we first note that without loss of generality, each message B_i can is of the form

$$B_i = \begin{cases} X_i \in A_i & 1, \\ X_i \notin A_i & -1, \end{cases} \tag{30}$$

where $A_i \subset \mathbb{R}$ is a Borel measurable set. Consider the conditional distribution $P(B^n|\theta)$ of B^n given θ. We have

$$P\left(B^n|\theta\right) = \prod_{i=1}^{n} P\left(B_i|\theta, B^{i-1}\right), \tag{31}$$

where $P\left(B_i = 1|\theta, B^{i-1}\right) = \mathbb{P}\left(X_i \in A_i\right)$, so that the Fisher information of B^n with respect to θ is given by

$$I_\theta(B^n) = \sum_{i=1}^{n} I_\theta(B_i|B^{i-1}), \tag{32}$$

where $I_\theta(B_i|B^{i-1})$ is the Fisher information of the distribution of B_i given B^{i-1}. From Lemma 10 it follows that $I_\theta(B_i|B^{i-1}) \le 4f^2(0)$. The Van Trees inequality [38], [39] now implies

$$\begin{aligned}\mathbb{E}\left(\theta_n - \theta\right)^2 &\ge \frac{1}{\mathbb{E}I_\theta(B^n) + I_0} \\ &= \frac{1}{\sum_{i=1}^{n} I_\theta(B_i|B^{i-1}) + I_0} \\ &\ge \frac{1}{4f^2(0)n + I_0}.\end{aligned}$$

□

D. Proof of Theorem 3

The algorithm given in (7) and (9) is a special case of a more general class of estimation procedures given in [40] and [41].

Proof of (i): Consider the following simplified version of [40, Thm. 4]:

Theorem 11: [40, Thm. 4] Let

$$X_i = \theta + Z_i, \quad i = 1, \ldots, n,$$

where the Z_is are i.i.d. with zero means and finite variances. Define

$$\begin{aligned}\theta_i &= \theta_{i-1} + \gamma_i \varphi(X_i - \theta_{i-1}), \\ \bar{\theta}_n &= \frac{1}{n}\sum_{i=0}^{n-1} \theta_i,\end{aligned} \tag{33}$$

where in addition, assume the following:

(i) There exits K_1 such that $|\varphi(x)| \le K_1(1+|x|)$ for all $x \in \mathbb{R}$.

(ii) The sequence $\{\gamma_i\}_{i=1}^{\infty}$ satisfies conditions (10).

(iii) The function $\psi(x) \triangleq \mathbb{E}\left[\varphi(x+Z_1)\right]$ is differentiable at zero with $\psi'(0) > 0$, and satisfies $\psi(0) = 0$ and $x\psi(x) > 0$ for all $x \neq 0$. Moreover, assume that there exists K_2 and $0 < \lambda \le 1$ such that

$$\left|\psi(x) - \psi'(0)x\right| \le K_2|x|^{1+\lambda}. \tag{34}$$

(iv) The function $\chi(x) \triangleq \mathbb{E}\left[\varphi^2(x+Z_1)\right]$ is continuous at zero.

Then $\bar{\theta}_n \to \theta$ almost surely and $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to $\mathcal{N}(0,V)$, where

$$V = \frac{\chi(0)}{\psi'^2(0)}.$$

Using the notation above, we set $\varphi(x) = \text{sgn}(x)$ and $Z_i = X_i - \theta$. We have that $\chi(x) = \mathbb{E}\text{sgn}^2(x+Z_1) = 1$, so $\chi(0) = 1$. In addition,

$$\begin{aligned}\psi(x) = \mathbb{E}\left[\text{sgn}(x+Z_1)\right] &= \int_{-\infty}^{\infty} \text{sgn}(x+z) f(z) dz \\ &= \int_{-x}^{\infty} f(z)dz - \int_{-\infty}^{-x} f(z)dz.\end{aligned}$$

Using the symmetry of $f(x)$ around zero, it follows that $\psi'(x) = 2f(x)$ and thus $\psi'(0) = 2f(0)$. It is now easy to verify that the rest of the conditions in Theorem 11 are fulfilled for any $\lambda > 0$. Since

$$\frac{\chi(0)}{\psi'^2(0)} = \frac{1}{4f^2(0)} = \frac{1}{\eta(0)},$$

it follows from Theorem 11 that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, 1/\eta(0)\right).$$

Proof of (ii): The proof requires the following refinement of Theorem 11.

Theorem 12: Let $\Delta_i = \theta_i - \theta$ and $\bar{\Delta}_i = \frac{1}{n}\sum_{i=1}^{n} \bar{\Delta}_i$. Assume that, in addition to Assumptions (i)-(iv) of Theorem 11, there exists K_1 and $\lambda > 0$ such that

$$\mathbb{E}\left[|\varphi(Z_1) - \varphi(x + Z_1)|\right] \leq K_1 |x|^{1+\lambda}. \tag{35}$$

Then:

(i)

$$\sqrt{n}\bar{\Delta}_n = -\frac{1}{\sqrt{n}}\frac{1}{\psi'(0)}\sum_{i=1}^{n-1} \varphi(Z_i) + o_{p.n}(1). \tag{36}$$

where $o_{p,n}(1)$ converges in probability to 0 as $n \to \infty$.

(ii) If Z_1 has continuously differentiable density $f(x)$ with finite location Fisher information

$$I_\theta = \int_{\mathbb{R}} \left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx,$$

then

$$\sqrt{n}\left(\bar{\Delta}_n\right) \xrightarrow{d} \mathcal{N}\left(\frac{1}{\psi'(0)}\int_{\mathbb{R}} \varphi(x)f'(x)dx, \frac{\chi(0)}{\psi'^2(0)}\right)$$

under the local alternative $Z_1, \ldots, Z_n \sim P^n_{h/\sqrt{n}}$ with density $f^n(x - h/\sqrt{n})$.

The proof of Theorem 12 is given in Subsection E below.

In our setting, the condition (35) holds since

$$\begin{aligned} &\operatorname{sgn}(Z_1) - \operatorname{sgn}(x + Z_1) \\ &= \begin{cases} 2 & Z_1 > 0, x + Z_1 < 0, \\ -2 & Z_1 < 0, x + Z_1 > 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Consequently,

$$\mathbb{E}\left[|\varphi(Z_1) - \varphi(x + Z_1)|\right] \leq \mathbb{P}\left(|Z_1| < x\right) \leq f(0)|x|,$$

In addition, by anti symmetry of $f'(x)$ around $x = 0$,

$$\int_{\mathbb{R}} \varphi(x)f'(x)dx = \int_{\mathbb{R}} \operatorname{sgn}(x)f'(x)dx = 2\int_0^\infty f'(x)dx = 2f(0).$$

Theorem 12, applied to the setting of Theorem 3, implies (11).

Proof of (iii): Consider the following result from [41]:

Theorem 13: [41, Thm. 2] Let

$$\begin{cases} U_n = U_{n-1} - \gamma_n \varphi(Y_n), & Y_n = g'(U_{n-1}) + Z_n \\ \bar{U}_n = \frac{1}{n}\sum_{i=1}^{n} U_n, & n = 1, 2, \ldots. \end{cases} \tag{37}$$

Assume that the function $g(x)$ is twice differentiable with a strictly positive and uniformly bounded second derivative. In particular, $g(x)$ is convex with a unique minimizer $x^\star \in \mathbb{R}$. Moreover, assume that the noises Z_n are uncorrelated and identically distributed with a distribution with a density for which the Fisher information exits. Let $\psi(x)$ and $\chi(x)$ be defined as in Theorem 11-(iii) and satisfy the conditions there. Assume in addition that $\chi(0) > 0$, condition (34) with $\lambda = 1$, and there exits K_3 such that

$$\mathbb{E}\left[|\varphi(x+Z_1)|^4\right] \leq K_3(1+|x|^4).$$

Finally, assume that the sequence $\{\gamma_n\}$ satisfies conditions (10) and (12). Then

$$V_n \triangleq \mathbb{E}\left[\left(\bar{U}_n - x^\star\right)^2\right] = n^{-1}\frac{\chi(0)}{(\psi'(0))^2(g''(x^\star))^2} + o(n^{-1}).$$

We now use Theorem 13 with $g(x) = 0.5(x-\theta)^2$, $\varphi(x) = \text{sgn}(x)$, $Z_n = \theta - X_n$. From (37) we have

$$\begin{aligned} U_n &= U_{n-1} + \gamma_n \text{sgn}(\theta - U_{n-1} - Z_n) \\ &= U_{n-1} + \gamma_n \text{sgn}(X_n - U_{n-1}), \end{aligned}$$

so the estimator $\bar{U}_n$ equals to the one defined by (9) and (7). Note that

$$\mathbb{E}\left[|\varphi(x+Z_1)|^4\right] = 1 \leq K_3(1+|x|^4)$$

for any $K_3 \geq 1$, the Fisher information of Z_1 is σ^2, $\chi(x) = 1 > 0$, and that the conditions in Theorem 13 on $\psi(x)$ and $\chi(x)$ were verified to hold in the first part of the proof. In particular, $\psi'(0) = (2f(0))^{-2}$. Since $f(x)$ satisfies the conditions above with $x^\star = \theta$ and $g''(x) = 1$. Theorem 13 implies that for any $\theta \in \mathbb{R}$,

$$nV_n = n\mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] = \frac{1}{4f^2(0)} + o(1).$$

E. *Proof of Theorem 12*

Proof of (i): The proof of part (i) of Theorem 12 requires the following two additional results from [40]:

Lemma 14: [40, Lem. 2] Consider the process $\{\Delta_i^1\}_{i=0}^{\infty}$ defined by

$$\Delta_i^1 = \Delta_{i-1}^1 - \gamma_i(A\Delta_{i-1} + \xi_i), \quad i = 1, 2 \ldots.$$

Assume that $A > 0$ and condition (ii) of Theorem 11 hold. Then

$$\sqrt{n}\bar{\Delta}_n^1 = \frac{1}{\sqrt{n}}\sum_{i=0}^{n-1}\Delta_i^1 = \frac{\alpha_n \Delta_0^1}{\sqrt{n}\gamma_0} + \frac{1}{\sqrt{n}A}\sum_{i=1}^{n-1}\xi_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1} w_i^n \xi_i, \tag{38}$$

where α_n, and w_i^n are real numbers such that $|\alpha_n| \leq K$ and $|w_i^n| \leq K$ for some $K < \infty$, and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n-1}|w_i^n| = 0.$$

Lemma 15: Under the conditions of Theorem 11,

$$\sum_{i=1}^{\infty}\frac{|\Delta_i|^{1+\lambda}}{\sqrt{i}} < \infty$$

almost surely.

Lemma 15 follows from the proof of Theorem 2 in [40].

We separate the proof of part (i) into two steps.

Step I: The expansion (36) holds for the process $\bar{\Delta}_i^1$ defined as follows:

$$\Delta_i^1 = \Delta_{i-1}^1 - \gamma_i \psi'(0)\Delta_{i-1}^1 - \gamma_i \varphi(Z_i), \qquad \delta_0^1 = \Delta_0 \tag{39}$$

$$\bar{\Delta}_i^1 = \frac{1}{n}\sum_{i=0}^{n-1} \Delta_i^1. \tag{40}$$

In order to show this claim, use Lemma 14 with $A = \psi'(0)$ and $\xi_i = -\varphi(Z_i)$. The first expression on the RHS on (38) goes to zero in variance. In addition,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1} w_i^n \xi_i\right)^2\right] &= \frac{1}{n}\sum_{i=1}^{n}(w_i^n)^2\mathbb{E}\left[\xi_i^2\right] + \frac{1}{n}\sum_{i\neq j}^{n} w_i^n w_j^n \mathbb{E}\left[\xi_i\xi_j\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}(w_i^n)^2\mathbb{E}\left[\varphi(Z_i)^2\right] = \chi(0)\frac{1}{n}\sum_{i=1}^{n}(w_i^n)^2 \to 0.\end{aligned}$$

We obtain

$$\sqrt{n}\bar{\Delta}_n^1 = -\frac{1}{\sqrt{n}}\frac{1}{\psi'(0)}\sum_{i=1}^{n-1}\varphi(Z_i) + o_{p.n}(1), \tag{41}$$

Step II: $\bar{\Delta}_n$ and $\bar{\Delta}_n^1$ are asymptotically equivalent.

From (33) and (39), the difference $\delta_i = \Delta_i - \Delta_i^1$ satisfies the recursion

$$\delta_i = \delta_{i-1} - \gamma_i\psi'(0)\delta_{i-1} + \gamma_i\left(\psi'(0)\Delta_{i-1} + \varphi(Z_i) - \varphi(\Delta_{i-1}+Z_i)\right),$$

where $\delta_0 = 0$. Use Lemma 14 with $\xi_i = \psi'(0)\Delta_{i-1} + \varphi(Z_i) - \varphi(\Delta_{i-1}+Z_i)$ to obtain

$$\sqrt{n}\bar{\delta}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}\left(\frac{1}{\psi'(0)} + w_i^n\right)\xi_i \tag{42}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}\left(\frac{1}{\psi'(0)} + w_i^n\right)\left(\psi'(0)\Delta_{i-1} - \psi(\Delta_{i-1})\right) \tag{43}$$

$$+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}\left(\frac{1}{\psi'(0)} + w_i^n\right)\left(\psi(\Delta_{i-1}) + \varphi(Z_i) - \varphi(\Delta_{i-1}+Z_i)\right) \tag{44}$$

For the term (43) and using (34), there exists K_1 and K_2 such that

$$\begin{aligned}(43) &\leq K_1\sum_{i=1}^{\infty}\frac{1}{\sqrt{i}}\left|\left(\psi'(0)\Delta_{i-1} - \psi(\Delta_{i-1})\right)\right| \\ &\leq K_2\sum_{i=1}^{\infty}\frac{|\Delta_i|^{1+\lambda}}{\sqrt{i}}.\end{aligned}$$

Lemma 15 shows that

$$\sum_{i=1}^{\infty}\frac{|\Delta_i|^{1+\lambda}}{\sqrt{i}} < \infty, \tag{45}$$

hence the Kronecker lemma implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}\left(\frac{1}{\psi'(0)} + w_i^n\right)\left(\psi'(0)\Delta_{i-1} - \psi(\Delta_{i-1})\right) \to 0.$$

For the term (44), set

$$\varepsilon_i \triangleq \psi(\Delta_{i-1}) + \varphi(Z_i) - \varphi(\Delta_{i-1}+Z_i).$$

For $a > 0$ and $n \in \mathbb{N}$, define the event

$$A_{n,a} \triangleq \left\{\sum_{i=1}^{n-1}\frac{|\varepsilon_i|}{\sqrt{i}} \geq a\right\}.$$

By Markov's inequality, we have

$$\mathbb{E}\left[\mathbf{1}_{A_{n,a}} \mid \Delta_0, \ldots, \Delta_{n-1}\right] \leq \frac{1}{a} \sum_{i=1}^{n-1} \frac{\mathbb{E}\left[|\varepsilon_i| \Delta_{i-1} |\right]}{\sqrt{i}}. \tag{46}$$

Using (34) and (35), there exists K' and $\lambda' > 0$ such that

$$\begin{aligned}&\mathbb{E}\left[|\varepsilon_i||\Delta_{i-1}\right] \leq |\psi(\Delta_{i-1})| + |\varphi(Z_i) - \varphi(\Delta_{i-1} + Z_i)|\\ &\leq K'|\Delta_{i-1}|^{1+\lambda}.\end{aligned}$$

Plugging this bound in (46) and using Lemma 15,we obtain

$$\mathbb{P}(A_{n,a}) = \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{A_{n,a}} \mid \Delta_0, \ldots, \Delta_{n-1}\right]\right] \leq \frac{K''}{a}$$

for some constant K''. It follows that for any ε, we may choose a large enough such that

$$\sup_n \mathbb{P}(A_{n,a}) < \varepsilon.$$

This implies that for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{\infty} \frac{|\varepsilon_i|}{\sqrt{i}} < \infty\right) \geq 1 - \varepsilon,$$

and hence

$$\sum_{i=1}^{\infty} \frac{|\varepsilon_i|}{\sqrt{i}} < \infty$$

almost surely. The Kronecker lemma now implies

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} |\varepsilon_i| \to 0,$$

hence the term (44) satisfies

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} \left(\frac{1}{\psi'(0)} + w_i^n\right) \varepsilon_i \leq \frac{K'''}{\sqrt{n}} \sum_{i=1}^{n-1} |\varepsilon_i| \to 0.$$

This conclude the proof of part (i).

Part (ii): Use (36) to write

$$\sqrt{n}\bar{\Delta}_n = G_n + o_{p,n}(1),$$

where

$$G_n \triangleq -\frac{1}{\sqrt{n}} \frac{1}{\psi'(0)} \sum_{i=1}^{n} \varphi(Z_i).$$

From [36, Exm. 7.8], the location model $f(x-\theta)$ with continuously differentiable $f(x)$ is differentiable in quadratic mean with the score function $-f'(x-\theta)/f(x-\theta)$. This fact implies the following expansion [36, Thm. 7.2]:

$$\log \frac{\mathbb{P}^n_{h/\sqrt{n}}}{\mathbb{P}^n_0}(Z_1, \ldots, Z_n) = \log \prod_{i=1}^{n} \frac{f(Z_i - h/\sqrt{n})}{f(X_i - \theta)} = hJ_n - \frac{1}{2} h^2 I_\theta + o_{p,n}(1), \tag{47}$$

where

$$J_n \triangleq -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{f'(Z_i)}{f(Z_i)}.$$

We have

$$\begin{aligned}\mathbb{E}\left[G_n J_n\right] &= \frac{1}{\psi'(0)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\varphi(Z_i) \frac{f'(Z_i)}{f(Z_i)}\right]\\ &= \frac{1}{\psi'(0)} \mathbb{E}\left[\varphi(Z_1) \frac{f'(Z_1)}{f(Z_1)}\right] = \frac{1}{\psi'(0)} \int_{\mathbb{R}} \varphi(x) f'(x) dx,\end{aligned}$$

Since both $\sqrt{n}G_n$ and $\sqrt{n}J_n$ are the sum of n i.i.d. random variables with zero mean and finite variance, we obtain, from the central limit theorem and Slutsky's theorem, that

$$\left(\sqrt{n}\bar{\Delta}_n, \log\frac{\mathbb{P}^n_{h/\sqrt{n}}}{\mathbb{P}^n_0}\right) \xrightarrow{d} \mathcal{N}\left(\left(0, -\frac{h^2}{2}I_\theta\right), \begin{pmatrix} \frac{\chi(0)}{\psi'^2(0)} & \frac{-h}{\psi'(0)}\int_{\mathbb{R}}\varphi(x)f'(x)dx \\ \frac{-h}{\psi'(0)}\int_{\mathbb{R}}\varphi(x)f'(x)dx & h^2 I_\theta \end{pmatrix}\right)$$

Le Cam's third lemma [36, Exm. 6.7] implies that under $\mathbb{P}^n_{h/\sqrt{n}}$,

$$\sqrt{n}\bar{\Delta}_n \xrightarrow{d} \mathcal{N}\left(\frac{-h}{\psi'(0)}\int_{\mathbb{R}}\varphi(x)f'(x)dx, \frac{\chi(0)}{\psi'^2(0)}\right).$$

□

F. Proof of Theorem 5

The log probability mass distribution of $B^n = (B_1, \ldots, B_n)$ is given by

$$\log\mathbb{P}_\theta(b^n) = \sum_{i=1}^{n}\left(\frac{b_i+1}{2}\log\mathbb{P}(X_i\in A_i) + \frac{1-b_i}{2}\log\mathbb{P}(X_i\in A_i)\right), \quad b^n\in\{-1,1\}^n.$$

Consequently,

$$\log\frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(b^n)}{\mathbb{P}_\theta(b^n)} = \sum_{i=1}^{n}\frac{b_i+1}{2}\log\frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i\in A_i)}{\mathbb{P}_\theta(X_i\in A_i)} + \sum_{i=1}^{n}\frac{1-b_i}{2}\log\frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i\notin A_i)}{\mathbb{P}_\theta(X_i\notin A_i)}. \tag{48}$$

For each $i = 1, \ldots, n$, write

$$A_i = \bigcup_{k=1}^{K_i}(t_{i,k}, t_{i,k+1}),$$

where $t_{i,1} < \ldots < t_{i,K_i}$ and, with a slight abuse of notation, $t_{i,1}$ and t_{i,K_i} may also be $-\infty$ or $+\infty$, respectively. Thus

$$\mathbb{P}_\theta(X_i\in A_i) = \sum_{k=1}^{K_i}(-1)^k F(x_{i,k}-\theta).$$

In particular, since f is differentible, $\mathbb{P}_\theta(X_i\in A_i)$ is twice differentiable, and we may write

$$\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i\in A_i) = \mathbb{P}_\theta(X_i\in A_i) + \frac{d}{d\theta}\mathbb{P}_\theta(X_i\in A_i)\frac{h}{\sqrt{n}} + o(h),$$

and thus

$$\begin{aligned}\log\frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i\in A_i)}{\mathbb{P}_\theta(X_i\in A_i)} &= \log\left(1 + \frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i\in A_i)}{\mathbb{P}_\theta(X_i\in A_i)}\frac{h}{\sqrt{n}} + o(h)\right) \\ &= \frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i\in A_i)}{\mathbb{P}_\theta(X_i\in A_i)}\frac{h}{\sqrt{n}} - \frac{h}{2n}\left(\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i\in A_i)}{\mathbb{P}_\theta(X_i\in A_i)}\right)^2 + o(h^2).\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\log\frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i\notin A_i)}{\mathbb{P}_\theta(X_i\notin A_i)} \\ &= \frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i\notin A_i)}{\mathbb{P}_\theta(X_i\notin A_i)}\frac{h}{\sqrt{n}} - \frac{h}{2n}\left(\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i\notin A_i)}{\mathbb{P}_\theta(X_i\notin A_i)}\right)^2 + o(h^2).\end{aligned}$$

From (48) we obtain

$$\log \frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(b^n)}{\mathbb{P}_\theta(b^n)} = \frac{h}{\sqrt{n}} \sum_{i=1}^n \left(\frac{b_i+1}{2} \frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)} + \frac{1-b_i}{2} \frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \notin A_i)}{\mathbb{P}_\theta(X_i \notin A_i)} \right)$$

$$- \frac{h^2}{2n} \sum_{i=1}^n \left(\frac{b_i+1}{2} \left(\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)} \right)^2 + \frac{1-b_i}{2} \left(\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \notin A_i)}{\mathbb{P}_\theta(X_i \notin A_i)} \right)^2 \right) + o(h^2)$$

Noting that

$$\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \notin A_i)}{\mathbb{P}_\theta(X_i \notin A_i)} = \frac{-\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{1-\mathbb{P}_\theta(X_i \in A_i)},$$

the proof is completed by proving the following two claims:

I. For $i = 1, \ldots, n$ denote

$$U_i = \frac{B_i+1}{2} \frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)} + \frac{1-B_i}{2} \frac{-\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{1-\mathbb{P}_\theta(X_i \in A_i)}.$$

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n U_i \xrightarrow{d} \mathcal{N}(0, \kappa(\theta)).$$

II. For $i = 1, \ldots, n$ denote

$$V_i = \frac{B_i-1}{2} \left(\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)} \right)^2 + \frac{1-B_i}{2} \left(\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{1-\mathbb{P}_\theta(X_i \in A_i)} \right)^2.$$

Then

$$\frac{1}{n} \sum_{i=1}^n V_i \xrightarrow{a.s.} \kappa(\theta).$$

Proof of Claim I: First note that

$$\mathbb{E}[U_i] = \frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)} \mathbb{P}(B_i = 1) + \frac{-\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{1-\mathbb{P}_\theta(X_i \in A_i)} \mathbb{P}(B_i = -1) = 0.$$

In addition,

$$\mathbb{E}U_i^2 = \left(\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)} \right)^2 \mathbb{P}(B_i = 1) + \left(\frac{-\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{1-\mathbb{P}_\theta(X_i \in A_i)} \right)^2 \mathbb{P}(B_i = -1)$$

$$= \frac{\left(\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right)^2}{\mathbb{P}_\theta(X_i \in A_i)} + \frac{\left(\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right)^2}{1-\mathbb{P}_\theta(X_i \in A_i)}$$

$$= \frac{\left(\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right)^2}{\mathbb{P}_\theta(X_i \in A_i)\left(1-\mathbb{P}_\theta(X_i \in A_i)\right)}$$

Therefore

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}U_i^2 = L_n(A_1, \ldots, A_n) \xrightarrow{a.s.} \kappa(\theta)$$

for any $\theta \in \Theta$ such that the limit above exists. We now verify that the sequence $\{U_i, i = 1, 2, \ldots\}$ satisfies Lyaponov's condition for his version of the central limit time: for any $\delta > 0$ we have that

$$\mathbb{E}\left|U_i\right|^{2+\delta} = \frac{\left|\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right|^{2+\delta}}{\left(\mathbb{P}_\theta(X_i \in A_i)\right)^{1+\delta}} + \frac{\left|\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right|^{2+\delta}}{\left(1-\mathbb{P}_\theta(X_i \in A_i)\right)^{1+\delta}}$$

and

$$\frac{\sum_{i=1}^{n}\mathbb{E}\left|U_i\right|^{2+\delta}}{\left(\sum_{i=1}^{n}\mathbb{E}U_i^2\right)^{\delta}} = \frac{\frac{1}{n^{1+\delta}}\sum_{i=1}^{n}\mathbb{E}|U_i|^{2+\delta}}{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}U_i^2\right)^{\delta}}. \tag{49}$$

Next, we claim that there exits $\delta > 0$ and $K > 0$, that are independent of n, such that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|U_i|^{2+\delta} < M \tag{50}$$

for all n large enough. To see this, note that

$$\begin{aligned}
\mathbb{E}|U_i|^{2+\delta} &= \frac{\left|\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right|^{2+\delta}}{(\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}} + \frac{\left|\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right|^{2+\delta}}{(1-\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}} \\
&= \frac{\left|\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right|^{2+\delta}}{(\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}(1-\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}}\left((1-\mathbb{P}_\theta(X_i \in A_i))^{1+\delta} + (\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}\right) \\
&\leq \frac{\left|\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right|^{2+\delta}}{(\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}(1-\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}},
\end{aligned}$$

where the last transition is because

$$\left((1-\mathbb{P}_\theta(X_i \in A_i))^{1+\delta} + (\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}\right) \leq 1.$$

We now use the fact that each A_i is a finite union of interval and consider the following lemma, proof of which is given in Appendix B.

Lemma 16: Let $f(x)$ be a log-concave, symmetric, and differentiable PDF such that $\eta(x)$ is unimodal. There exists $\delta > 0$ such that for any $x_1 \geq \ldots \geq x_n \in \mathbb{R}$,

$$\frac{\left|\sum_{k=1}^{n}(-1)^{k+1}f(x_k)\right|^{2+\delta}}{\left(\sum_{k=1}^{n}(-1)^{k+1}F(x_k)\right)^{1+\delta}\left(1-\sum_{k=1}^{n}(-1)^{k+1}F(x_k)\right)^{1+\delta}} \leq 2^{\delta} f^{2+\delta}(0). \tag{51}$$

Lemma 16 implies

$$\begin{aligned}
&\leq \frac{\left|\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right|^{2+\delta}}{(\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}(1-\mathbb{P}_\theta(X_i \in A_i))^{1+\delta}} = \frac{\left|\sum_{k=1}^{K_i}(-1)^k f(x_{i,k}-\theta)\right|^{2+\delta}}{\left(\sum_{k=1}^{K_i}(-1)^k F(x_{i,k}-\theta)\right)^{1+\delta}\left(1-\sum_{k=1}^{K_i}(-1)^k F(x_{i,k}-\theta)\right)^{1+\delta}} \\
&\leq 2^{\delta} f^{2+\delta}(0).
\end{aligned}$$

In particular, we conclude that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|U_i|^{2+\delta} \leq 2^{\delta} f^{2+\delta}(0),$$

and thus for any $\delta > 0$ the numerator of (49), as well as the entire expression, goes to zero. From Lyaponov's central limit theorem we conclude that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i \xrightarrow{d} \mathcal{N}(0, \kappa(\theta)).$$

Proof of Claim II: We have:

$$\begin{aligned}
\mathbb{E}V_i &= \frac{\left(\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right)^2}{\mathbb{P}_\theta(X_i \in A_i)} + \frac{\left(\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right)^2}{1-\mathbb{P}_\theta(X_i \in A_i)} \\
&= \frac{\left(\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\right)^2}{\mathbb{P}_\theta(X_i \in A_i)\left(1-\mathbb{P}_\theta(X_i \in A_i)\right)}
\end{aligned}$$

We conclude that:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}V_i = L_n(A_1, \ldots, A_n) \to \kappa(\theta) \tag{52}$$

Since the V_is are independent of each other, Kolmogorov's law of large numbers implies

$$\frac{1}{n}\sum_{i=1}^{n} V_i \xrightarrow{a.s.} \kappa(\theta)$$

for any $\theta \in \Theta$ for which the limit (52) exists.

□

Proof of Lemma 16: Denote

$$\delta_n \triangleq \delta_n(x_1,\ldots,x_n) \triangleq \sum_{k=1}^{n}(-1)^{k+1} f(x_k),$$

$$\Delta_n \triangleq \Delta_n(x_1,\ldots,x_n) \triangleq \sum_{k=1}^{n}(-1)^{k+1} F(x_k),$$

and

$$\eta_\delta(x) \triangleq \frac{(\delta_1(x))^{2+\delta}}{(\Delta_1(x))^{1+\delta}(1-\Delta_1(x))^{1+\delta}} = \frac{(f(x))^{2+\delta}}{(F(x))^{1+\delta}(1-F(x))^{1+\delta}}.$$

The proof is by induction on n. For the case $n=1$ the LHS of (51) equals $\eta_\delta(x)$. Note that $\eta_\delta(0) = 2^\delta f^{2+\delta}(0)$, so it is enough to prove that $\eta_\delta(0)$ attains it maximum at $x=0$. We have $\eta_\delta(x) = \eta^{1+\delta}(x)/f^\delta(x)$, and thus

$$\log \eta_\delta'(x) = (1+\delta)\frac{\eta'(x)}{\eta(x)} - \delta \frac{f'(x)}{f(x)}.$$

By Assumption 1 both terms above are negative for $x>0$, so that $\log \eta_\delta'(x) \le 0$ if and only if

$$(1+\delta)\left|\frac{\eta'(x)}{\eta(x)}\right| \ge \delta \left|\frac{f'(x)}{f(x)}\right|, \qquad x>0. \tag{53}$$

Also by Assumption 1, for $x>0$ we have

$$0 > (\log \eta(x))' = \frac{h'(x)}{h(x)} - \frac{h'(-x)}{h(-x)},$$

so that (53) is satisfies for $\delta > 0$ small enough.

Next, assume that

$$\frac{(\delta_n)^{2+\delta}}{(\Delta_n)^{1+\delta}(1-\Delta_n)^{1+\delta}} \le \eta_\delta(0) = 2^\delta f^{2+\delta}(0). \tag{54}$$

for all integers up to $n = N-1$ and consider the case $n=N$. The maximal value of the LHS of (54) is attained for the same $(x_1,\ldots,x_N) \in \mathbb{R}^N$ that attains the maximal value of

$$g(x_1,\ldots,x_N) \triangleq (2+\delta)\log\delta_N - (1+\delta)\log\Delta_N - (1+\delta)\log(1-\Delta_N),$$

The derivative of $g(x_1,\ldots,x_N)$ with respect to x_k is given by

$$\frac{\partial g}{\partial x_k} = \frac{(2+\delta)(-1)^{k+1} f'(x_k)}{\delta_N} - \frac{(1+\delta)(-1)^{k+1} f(x_k)}{\Delta_N} + \frac{(1+\delta)(-1)^{k+1} f(x_k)}{1-\Delta_N}.$$

We conclude that the gradient of g vanishes if and only if

$$\frac{f'(x_k)}{f(x_k)} = \frac{\delta_N(1+\delta)}{2+\delta}\left(\frac{1}{\Delta_N} - \frac{1}{1-\Delta_N}\right), \quad k=1,\ldots,N. \tag{55}$$

From the same reason as in the proof of Lemma 16, (55) is satisfied if and only if $x_1 = \ldots = x_N$. For odd N and $x_1 = \ldots = x_N$, the LHS of (54) equals $\eta_\delta(x_1)$ which was shown to be smaller than $\eta_\delta(\varepsilon)$. For even N and any constant d, the limit of the LHS of (54) as $(x_1,\ldots,x_N) \to (d,\ldots,d)$ exists and equals zero. Therefore, the maximum of the LHS of (54) is not attained at the line $x_1 = \ldots = x_N$). We now consider the possibility that the LHS of (54) is maximized at the borders. That is, as one or more of the coordinates of $(x_1,\ldots,x_N)$ approaches $\pm\infty$, or $\pm\varepsilon$. As we assumed $x_1 \ge \ldots \ge x_N$, if $x_i = x_{i+1}$ for some i than their contribution to (54) is zero and thus this case reduces to the case $n = N-2$. A similar reduction holds if $x_N, x_{N-1} \to -\infty$, $x_1, x_2 \to \infty$, $x_i, x_{i+1} = -\varepsilon$ for some i, or $x_i, x_{i+1} = \varepsilon$ for some i. It is therefore enough to consider the cases:

(1) $x_N \to -\infty$.

(2) $x_1 \to \infty$.

Assume first $x_N \to -\infty$. Then

$$\frac{(\delta_N)^{2+\delta}}{(\Delta_N)^{1+\delta}(1-\Delta_N)^{1+\delta}} = \frac{\left(\sum_{k=1}^{N-1}(-1)^{k+1}f(x_k)\right)^{2+\delta}}{\left(\sum_{k=1}^{N-1}(-1)^{k+1}F(x_k)\right)^{1+\delta}\left(1-\sum_{k=1}^{N-1}(-1)^{k+1}F(x_k)\right)^{1+\delta}},$$

which is smaller than $\eta_\delta(0)$ by the induction hypothesis. Assume now that $x_1 \to \infty$. Then

$$\begin{aligned}&\frac{(\delta_N)^{2+\delta}}{(\Delta_N)^{1+\delta}(1-\Delta_N)^{1+\delta}}\\ =&\frac{\left(\sum_{k=2}^{N}(-1)^{k+1}f(x_k)\right)^{2+\delta}}{\left(1+\sum_{k=2}^{N}(-1)^{k+1}F(x_k)\right)^{1+\delta}\left(1-1-\sum_{k=2}^{N}(-1)^{k+1}F(x_k)\right)^{1+\delta}}\\ &=\frac{\left(-\sum_{m=1}^{N}(-1)^{m+1}f(x'_m)\right)^{2+\delta}}{\left(1-\sum_{m=1}^{N-1}(-1)^{m+1}F(x'_m)\right)^{1+\delta}\left(\sum_{m=1}^{N-1}(-1)^{m+1}F(x'_m)\right)^{1+\delta}},\end{aligned}$$

where $x'_m = x_{m+1}$ for $m = 1,\ldots,N-1$. The last expression is smaller than $\eta_\delta(0)$ by the induction hypothesis. □

G. *Proof of Theorem 7*

Let Ξ be the set of points $\theta \in \Theta$ for which $\kappa(\theta) = \eta(0)$. Since $B_1, B_2, \ldots$ satisfy the conditions in Theorem 5, for $\theta \in \Xi$ if and only if

$$\lim_{n\to\infty} L_n(A_1,\ldots,A_n;\theta) = \eta(0). \tag{56}$$

By assumption, we have $B_i^{-1} = A_i$ where A_i can be expressed as

$$A_i = \cup_{i=1}^{K}(a_{i,k}, b_{i,k}),$$

where $a_{i,1} \le b_{i,1} \le \ldots \le a_{i,K}, b_{i,K}$, and $a_{i,1}$ and $b_{i,K}$ may take the values $-\infty$ and ∞, respectively. Denote

$$\mathscr{B}_i = \cup_{k=1}^{K}\{a_{i,k}, b_{i,k}\}.$$

For any θ and $\varepsilon > 0$, denote

$$S_n(\theta,\varepsilon) \triangleq \{i \le n : (\theta-\varepsilon, \theta+\varepsilon) \cap \mathscr{B}_i \neq \emptyset\}$$

In words, S_n contains all integers smaller than n in which an ε-ball around θ contains an endpoint of one of the intervals consisting A_i. We now claim that if $\theta \in \Xi$ then $\mathrm{card}(S_n(\theta,\varepsilon))/n \to 1$. Indeed, for such θ we have

$$\begin{aligned}&L_n(A_1,\ldots,A_n;\theta)\\ &=\frac{1}{n}\sum_{i\in S_n(\varepsilon,\theta)}\frac{\left(\sum_{k=1}^{K}f(\theta-b_{i,k})-f(\theta-a_{i,k})\right)^2}{\sum_{k=1}^{K}\left(F(\theta-b_{i,k})-F(\theta-a_{i,k})\right)\left(1-\sum_{k=1}^{K}\left(F(\theta-b_{i,k})-F(\theta-a_{i,k})\right)\right)}\\ &+\frac{1}{n}\sum_{i\notin S_n(\varepsilon,\theta)}\frac{\left(\sum_{k=1}^{K}f(b_{i,k}-\theta)-f(a_{i,k}-\theta)\right)^2}{\sum_{k=1}^{K}\left(F(\theta-b_{i,k})-F(\theta-a_{i,k})\right)\left(1-\sum_{k=1}^{K}\left(F(\theta-b_{i,k})-F(\theta-a_{i,k})\right)\right)}\\ &\overset{(a)}{\le}\frac{\mathrm{card}\left(S_n(\theta,\varepsilon)\right)}{n}\eta(0)+\frac{n-\mathrm{card}\left(S_n(\theta,\varepsilon)\right)}{n}\eta(\varepsilon)\end{aligned} \tag{57}$$

where (a) follows from Lemma 16 and the fact that for $i \in S_n(\theta,\varepsilon)$,

$$\max\left\{\max_k \eta(b_{i,k}-\theta), \max_k \eta(a_{i,k}-\theta)\right\} \le \eta(\varepsilon) < \eta(0).$$

Unless $\mathrm{card}\left(S_n(\theta,\varepsilon)\right)/n \to 1$, (57), and thus $L_n(A_1,\ldots,A_n;\theta)$, are bounded from above by a constant that is smaller then $\eta(0)$ in contradiction to the fact that $\theta \in \Xi$.

For $k \in \mathbb{N}$, assume by contradiction that there exists $N \geq 2K+1$ distinct elements $\theta_1,\ldots,\theta_N \in \Xi$. Since each A_i consists of at most K intervals, we have that

$$\mathrm{card}(\cup_{i=1}^{n}\mathscr{B}_i) \leq 2nK. \tag{58}$$

Fix $\varepsilon > 0$ such that

$$\varepsilon < \frac{1}{2}\min_{i\neq j}|\theta_i - \theta_j|.$$

Since for each $\theta \in \Theta$ we have $S_n(\theta,\varepsilon) \to 1$, there exists n large enough such that

$$\mathrm{card}\left(S_n(\theta_i,\varepsilon)\right) \geq n\left(1 - \frac{1}{2N}\right)$$

for all $i = 1,\ldots,N$. However, $S_n(\theta_1,\varepsilon),\ldots S_n(\theta_N,\varepsilon)$ are disjoint, so the cardinallity of their union is at least $n\left(1-\frac{1}{2N}\right)N$ which is grater than $2nK + n/2$ in contradiction to (58).

References

[1] J. Candy, "A use of limit cycle oscillations to obtain robust analog-to-digital converters," *IEEE Transactions on Communications*, vol. 22, no. 3, pp. 298–305, Mar 1974.

[2] P. W. Wong and R. M. Gray, "Sigma-delta modulation with i.i.d. Gaussian inputs," *IEEE Transactions on Information Theory*, vol. 36, no. 4, pp. 784–798, Jul 1990.

[3] R. G. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters, "Exponential decay of reconstruction error from binary measurements of sparse signals," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3368–3385, 2017.

[4] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.

[5] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1275–1297, 2013.

[6] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive mimo systems," *IEEE Trans. Signal Process*, vol. 65, no. 15, pp. 4075–4089, 2017.

[7] J. Choi, J. Mo, and R. W. Heath, "Near maximum-likelihood detector and channel estimator for uplink multiuser massive mimo systems with one-bit adcs," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 2005–2018, 2016.

[8] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, Oct 1998.

[9] F. Cicalese, D. Mundici, and U. Vaccaro, "Least adaptive optimal search with unreliable tests," *Theoretical Computer Science*, vol. 270, no. 1, pp. 877–893, 2002.

[10] R. M. Karp and R. Kleinberg, "Noisy binary search and its applications," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '07. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 881–890.

[11] W. Shi, T. W. Sun, and R. D. Wesel, "Quasi-convexity and optimal binary fusion for distributed detection with identical sensors in generalized Gaussian noise," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 446–450, Jan 2001.

[12] P. Venkitasubramaniam, L. Tong, and A. Swami, "Quantization for maximin are in distributed estimation," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3596–3605, July 2007.

[13] A. Vempaty, H. He, B. Chen, and P. K. Varshney, "On quantizer design for distributed bayesian estimation in sensor networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 20, pp. 5359–5369, Oct 2014.

[14] H. Chen and P. K. Varshney, "Performance limit for distributed estimation systems with identical one-bit quantizers," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 466–471, 2010.

[15] ——, "Performance limit for distributed estimation systems with identical one-bit quantizers," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 466–471, Jan 2010.

[16] M. Longo, T. D. Lookabaugh, and R. M. Gray, "Quantization for decentralized hypothesis testing under communication constraints," *IEEE Transactions on Information Theory*, vol. 36, no. 2, pp. 241–255, Mar 1990.

[17] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 1, no. 2, pp. 167–182, 1988.

[18] W. P. Tay and J. N. Tsitsiklis, "The value of feedback for decentralized detection in large sensor networks," in *International Symposium on Wireless and Pervasive Computing*, Feb 2011, pp. 1–6.

[19] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem [multiterminal source coding]," *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 887–902, 1996.

[20] H. Viswanathan and T. Berger, "The quadratic Gaussian CEO problem," *IEEE Transactions on Information Theory*, vol. 43, no. 5, pp. 1549–1559, 1997.

[21] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1057–1070, 1998.

[22] V. Prabhakaran, D. Tse, and K. Ramachandran, "Rate region of the quadratic Gaussian CEO problem," in *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on.* IEEE, 2004, p. 119.

[23] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, “Information-theoretic lower bounds for distributed statistical estimation with communication constraints,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2328–2336.

[24] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang, “Optimality guarantees for distributed statistical estimation,” *arXiv preprint arXiv:1405.0782*, 2014.

[25] Y. Han, A. Özgür, and T. Weissman, “Geometric lower bounds for distributed parameter estimation under communication constraints,” *CoRR*, vol. abs/1802.08417, 2018. [Online]. Available: http://arxiv.org/abs/1802.08417

[26] Z. Zhang and T. Berger, “Estimation via compressed information,” *IEEE Transactions on Information Theory*, vol. 34, no. 2, pp. 198–211, 1988.

[27] Y. Han, P. Mukherjee, A. Ozgur, and T. Weissman, “Distributed statistical estimation of high-dimensional and nonparametric distributions,” in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 506–510.

[28] A. Xu and M. Raginsky, “Information-theoretic lower bounds on bayes risk in decentralized estimation,” *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1580–1600, 2017.

[29] L. Barnes, Y. Han, and A. Ozgur, “A geometric characterization of fisher information from quantized samples with applications to distributed statistical estimation,” in *2018 56st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2018.

[30] I. A. Ibragimov, “On the composition of unimodal distributions,” *Theory of Probability & Its Applications*, vol. 1, no. 2, pp. 255–260, 1956.

[31] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.

[32] M. Bagnoli and T. Bergstrom, “Log-concave probability and its applications,” *Economic theory*, vol. 26, no. 2, pp. 445–469, 2005.

[33] M. R. Sampford, “Some inequalities on mill’s ratio and related functions,” *The Annals of Mathematical Statistics*, vol. 24, no. 1, pp. 130–132, 1953.

[34] J. Hammersley, “On estimating restricted parameters,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 12, no. 2, pp. 192–240, 1950.

[35] J. Chen, X. Zhang, T. Berger, and S. Wicker, “An upper bound on the sum-rate distortion function and its corresponding rate allocation schemes for the CEO problem,” *Selected Areas in Communications, IEEE Journal on*, vol. 22, no. 6, pp. 977–987, Aug 2004.

[36] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[37] A. Kipnis, S. Rini, and A. J. Goldsmith, “Compress and estimate in multiterminal source coding,” 2017, unpublished. [Online]. Available: https://arxiv.org/abs/1602.02201

[38] H. L. Van Trees, *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[39] R. D. Gill and B. Y. Levit, “Applications of the van Trees inequality: a Bayesian Cramér-Rao bound,” *Bernoulli*, pp. 59–79, 1995.

[40] B. T. Polyak and A. B. Juditsky, “Acceleration of stochastic approximation by averaging,” *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[41] B. T. Polyak, “New stochastic approximation type procedures,” *Automat. i Telemekh*, vol. 7, no. 98-107, p. 2, 1990.