Mean Estimation from One-Bit Measurements

John C. Duchi*† and Alon Kipnis*
*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 (MSE) 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 the previously observed bits, we show that the optimal relative efficiency compared to the sample mean is the efficiency of the median. For example, in estimating the mean of a normal distribution, a constraint of one bit per sample incurs a penalty of $\pi/2$ in sample size compared to the unconstrained case. We also consider a distributed setting where each one-bit message is only a function of a single sample. We derive lower bounds on the MSE in this setting, and show that the optimal efficiency can only be attained at a finite number of points in the parameter space. Finally, we analyze a distributed setting where the bits are obtained by comparing each sample against a prescribed threshold. Consequently, we consider the threshold density that minimizes the maximal MSE. 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 three settings:

(i) *Centralized* encoding (Fig. 1): all n encoders confer and produce a single n bit message.
(ii) *Adaptive* or *sequential* encoding (Fig. 2): the nth encoder observes the nth sample and the $n-1$ previous messages (bits).
(iii) *Distributed* encoding (Fig. 3): the nth message is only a function of the nth observation.

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 settings (i)-(iii) above, respectively:

- **Signal acquisition:** a quantity is measured n times at different instances, and the results are averaged in order to reduce measurement noise. The averaged result is then stored using one of n states.
- **Analog-to-digital conversion:** in sigma-delta modulation, an analog signal is converted into a sequence of bits by sampling it at a very high rate and then using one-bit threshold detector combined with a feedback loop to update an accumulated error state. Therefore, the MSE in tracking an analog signal using a 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 [1]. 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.

- **Differential 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. We are interested in particular in the *asymptotic relative efficiency* (ARE) of estimators in the constrained setting 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 a signal that varies in time or space under the one bit per measurement constraint. Such situations are considered in [2], [3], [4], [5], [6].

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 $\pi/2$, showing that the one-bit constraint increases the effective sample size in estimating θ by at least $\pi/2 \approx 1.57$ 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. Clearly, the penalty on the sample size in setting (iii) is at least as large as that in setting (ii). Unlike in setting (ii), we show that 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.

We note that although the ARE in setting (i) is 1, this scheme already poses a non-trivial challenge for the design and analysis of an optimal encoding and estimation procedures. Indeed, the standard technique to encode an unknown random quantity using n bits is equivalent to the design of a scalar quantizer [7]. 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 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.

In addition to encoding and estimation schemes that lead to optimal results, we also consider two additional "natural" schemes. Specifically, in setting (ii) we consider the one-bit optimal scheme, i.e., the case of a greedy encoder that given the nth sample and the previous $n-1$ bits, provides a bit that minimizes the n-step MSE. In setting (iii) we also consider the case where the messages are obtained by comparing each sample against a prescribed threshold. This threshold may be different across samples and is assumed deterministic (independent of the data).

Related Works

As the variance σ^2 goes to zero, 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 [8], [9]. However, since we

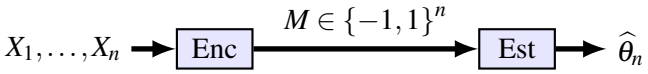

Fig. 1. Centralized one-bit encoding: encoder sends n bits after observing n samples.

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 [1] 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 [1]. 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 [2] 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 [2], an exponential MSE decaying rate, does not hold in the noisy setting. Stated otherwise, the MSE in the setting of [2] 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 [10], [11], [12], [13], [14], but without optimizing the encoders and their detection rules. Consequently, the performance derived in these works are not optimal. The work of [15] 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 [16] 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 [17], 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 they ARE in the distributed setting with threshold detection rules is strictly larger than that in the adaptive setting, suggesting that the case of a finite Θ is very 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 [18], [19], [20], [21], 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 optimal even under the one bit per sample constraint.

Finally, we note that our settings (ii) and (iii) can be obtained as special cases of [22] that consider adaptive and distributed estimation protocols for m machines, each has access to n/m independent samples. The main result of [22] are bounds on the estimation error as a function of the number of bits R each machine uses for communication. The specialization of their result to our setting, by taking $m = n$ and $R = 1$, leads to looser lower bounds then we derive here for cases (ii) and (iii). Looser bounds can also be obtained from [23], [24], [**?**] that considered inference and distributed estimation under data compression constraint [25], [23], [24], [26].

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 with its mode at $x = \theta$ [27]. 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 $M_1,\ldots,M_n$, obtained from $X_1,\ldots,X_n$ under three possible scenarios:

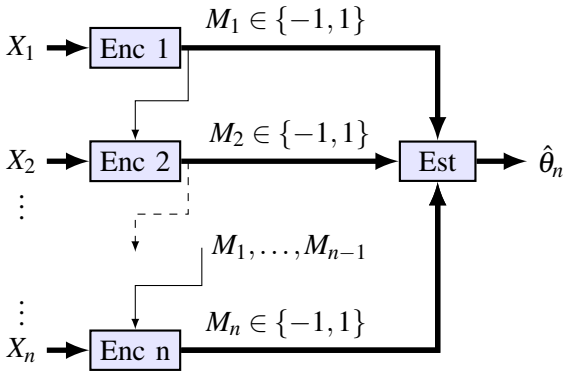

Fig. 2. Adaptive one-bit encoding: the nth encoder delivers a single bit message which is a function of its private sample X_n and the previous messages $M_1,\ldots,M_{n-1}$.

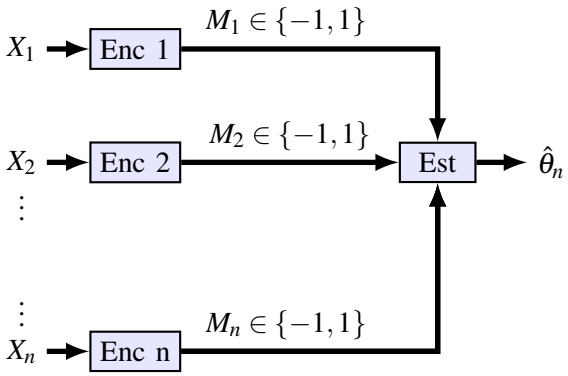

Fig. 3. Distributed one-bit encoding: the one-bit message produced by each encoder is only a function of its private sample X_i.

(i) Centralized $M_i(X_1,\ldots,X_n)$, $i=1,\ldots,n$ (Fig. 1).
(ii) Adaptive $M_i(X_i,M_1,\ldots,M_{i-1})$, $i=2,\ldots,n$ (Fig. 2).
(iii) Distributed $M_i(X_i)$, $i=1,\ldots,n$ (Fig. 3).

The performance of an estimator $\hat{\theta}_n \triangleq \hat{\theta}_n(M^n)$ 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^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 [28, 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.

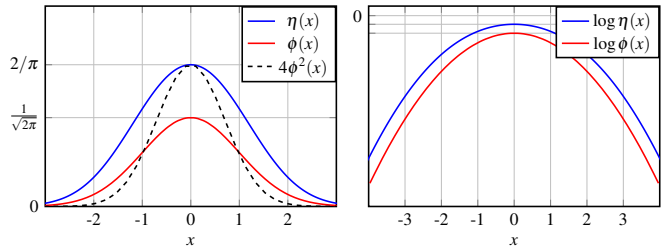

Fig. 4. The function $\eta(x) = f^2(x)/F(x)F(-x)$ for $f(x) = \phi(x)$ the standard normal density.

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 [29]. For $f(x)$ the normal density, it is shown in [30] and [31] that $\eta(x)$ is a strictly decreasing function of $|x|$, as illustrated in Fig. 4. 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: $\eta(x)$ is strongly unimodal.

Under this assumption we have,

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

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. Consisten Estimation and Of-the-shelf Bounds

On a first impression, it may not be clear whether consistent estimation of the mean is even possible in the adaptive and centralized settings. On the other hand, it may seem as if estimation in these cases is trivial as in the centralized setting (i). We next settle such skepticism by deriving lower and upper bounds to the relative efficiency under setting (iii). We show that:

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

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

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

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

so that

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

is a consistent estimator for θ in the distributed setting of Fig. 3, where we note that $F(x)$ is invertible over the support of $f(x)$ which is a connected set. 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

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

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 $\theta_0 = \theta$, although θ is unknown apriori. 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 ≈ 0.07 when θ_0 is 2.33 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 lower variance can be obtained by an estimator of the form (4) if, after observing a batch of the single bit messages, one can update the threshold value θ_0. Such a scheme falls within the adaptive setting of Fig. 2 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 [18].

Assuming that θ is drawn once from the prior $\pi(d\theta)$, our mean estimation problem from one-bit samples under distributed encoding in Fig. 3 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. As a result, 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.

An optimal scheme for the CEO and its corresponding MSE is known only for the case where the prior on θ, as well as the noise corrupting it at each location, are Gaussian. Namely, the Gaussian CEO is obtained from our setting when the ith encoder uses $R_i = 1$ bits to transmit a message that is a function of $X_i = \theta + \sigma Z_i$, where Z_i is standard normal. Consequently, a lower bound on the MSE in our distributed setting is obtained by considering the minimal MSE in the Gaussian CEO as the number of encoders n goes to infinity. This leads to 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

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

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

Proof: We consider the minimal distortion $D^\star$ in the Gaussian CEO setting with L observers and under a total sum-rate $R_\Sigma = R_1 + \ldots + R_L$ from [32, Eq. 10]:

$$R_\Sigma = \frac{1}{2}\log^+\left[\frac{\sigma_\theta^2}{D^\star}\left(\frac{D^\star L}{D^\star L - \sigma^2 + D^\star\sigma^2/\sigma_\theta^2}\right)^L\right]. \tag{7}$$

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

$$n = \frac{1}{2}\log_2\left[\frac{\sigma_\theta^2}{D^\star}\left(\frac{D^\star n}{D^\star n - \sigma^2 + D^\star\sigma^2/\sigma_\theta^2}\right)^n\right]. \tag{8}$$

$D^\star$ satisfying (8) 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 distortion with $R_1 = \ldots, R_n = 1$ and hence a lower bound to the minimal MSE in estimating θ in the distributed setting of Fig. 3. By considering $D^\star$ in (8) 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 distortion while enforcing the condition $R_1 = \ldots = R_n = 1$. Indeed, an upper bound for the CEO distortion under the condition $R_1 = \ldots = R_n = 1$ follows from [33], and leads to

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

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

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 exclusively attributed to the ability to perform coding over blocks. Namely, each CEO encoder 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 and the mean estimation problem from one-bit measurements in the distributed setting of Fig. 3.

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 $X_1, \ldots, X_n$. Next, we provide a particular adaptive estimation scheme that attains this maximal efficiency. Finally, in Theorem 5, we provide an adaptive estimation scheme that is one-step optimal in the sense that at each step i, the chosen message M_i minimizes the MSE given X_i and the previous $i-1$ messages.

A. Maximal efficiency in 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 Fig. 2. Assumes that $\pi(\theta)$ converges to zero at the endpoints of the interval Θ. Then

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

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 [34, Thm. 2.13],[35] which bounds from below the MSE of any estimator of θ by the inverse of the expected value of the aforementioned Fisher information plus I_0. The details are in the Appendix.

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:

$$\theta_n = \theta_{n-1} + \gamma_n M_n, \quad n=1,2,\ldots, \tag{9}$$

where

$$M_n = M_n(X_n, \theta_{n-1}) = \mathrm{sgn}(X_n - \theta_{n-1}). \tag{10}$$

Define the nth step estimation as

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i. \tag{11}$$

We have the following results:

Theorem 3: Consider the sequence $\left\{\hat{\theta}_n, n\in\mathbb{N}\right\}$ defined by (11).

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

$$\begin{cases} \frac{\gamma_n-\gamma_{n+1}}{\gamma_n} = o(\gamma_n), \\ \sum_{n=1}^{\infty}\frac{\gamma_n^{(1+\lambda)/2}}{\sqrt{n}} < \infty, \quad \text{for some } 0<\lambda\leq 1 \end{cases} \tag{12}$$

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

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

(ii) Assume that in addition to (12), $\{\gamma_n, n\in\mathbb{N}\}$ satisfies

$$\begin{cases} \gamma_n = o(n^{-2/3}), \\ \sum_{n=1}^{\infty}\gamma_n = \infty. \end{cases} \tag{13}$$

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

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

Proof: The asymptotic behavior of (11) follows from [36, Thm. 4] and [37, Thm. 2]. The details are in the Appendix.

Theorem 3 implies that the estimator $\hat{\theta}_n$, defined by (11) and (9), attains the maximal ARE as established by Theorem 2. The update step (9) 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 (9) and (11), each one-bit message M_n is a function of the current gradient descent estimate θ_{n-1} and its private sample. As we explain next, it is possible to obtain the optimal efficiency of $\eta(0)\sigma^2$ using only a single "access" to such state.

C. Optimality using a Single Interaction

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

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

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 may update once the threshold value θ_0. In this procedure we separate the sample into two disjoint sets: $X_1, \ldots, X_{n_1}$ and $X_{n_1+1}, \ldots, X_n$ for some $n_1 < n$. We first use the estimator (4) to obtain an estimate $\hat{\theta}_{n_1}$ based on $M_1, \ldots, M_{n_1}$, and then use $\hat{\theta}_{n_1}$ as the new threshold value to obtain messages $M_{n_1+1}, \ldots, M_n$. The specific encoding and estimation scheme, as well as its asymptotic performance, are given by the following theorem:

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

$$M_i = \begin{cases} \mathbf{1}_{X_i \geq \theta_0} & i = 1, \ldots, n_1, \\ \mathbf{1}_{X_i \geq \hat{\theta}_{n_1}} & i = n_1 + 1, \ldots, n, \end{cases}$$

where $n = n_1 + n_2$ and

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

Let

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

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

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

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

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

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

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

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

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

□

To conclude the adaptive estimation setting, we now consider estimation from one-bit messages obtained via an encoding scheme that is one-step optimal.

D. One-step optimal estimation

We now consider an estimation scheme that posses the property of *one-step optimality*: at each step n, the nth encoder designs the detection region $M_n^{-1}(1)$ such that the MSE given M^n is minimal. In other words, this scheme designs the messages in a greedy manner, such that the MSE at step n is minimal given the current state of the estimation described by M^{n-1}.

The following theorem determine the message, i.e., the decision rule $(M^{n-1},X_n)\to M_n$, that minimizes the next step MSE:

Theorem 5 (optimal one-step estimation): Let $\pi(\theta)$ be an absolutely continuous log-concave probability distribution. Given a sample X from a log-concave distribution with mean θ, define

$$M^\star=\operatorname{sgn}(X-\tau), \tag{14}$$

where τ satisfies the equation

$$\tau=\frac{m^-(\tau)+m^+(\tau)}{2}, \tag{15}$$

with

$$\begin{aligned}m^-(\tau)&=\frac{\int_{-\infty}^{\tau}\theta\pi(d\theta)}{\int_{-\infty}^{\tau}\pi(d\theta)},\\m^+(\tau)&=\frac{\int_{\tau}^{\infty}\theta\pi(d\theta)}{\int_{\tau}^{\infty}\pi(d\theta)}.\end{aligned}$$

For any estimator $\hat{\theta}$ which is a function of $M(X)\in\{-1,1\}$,

$$\mathbb{E}\left(\theta-\hat{\theta}(M)\right)^2\geq\mathbb{E}\left(\theta-\mathbb{E}[\theta|M^\star]\right)^2. \tag{16}$$

Proof: The proof is the result of the following two lemmas, proofs of which can be found in the Appendix:

Lemma 6: Let $f(x)$ be a log-concave density function. Then the equation

$$2x=\frac{\int_x^{\infty}uf(u)du}{\int_x^{\infty}f(u)du}+\frac{\int_{-\infty}^{x}uf(u)du}{\int_{-\infty}^{x}f(u)du} \tag{17}$$

has a unique solution.

Lemma 7: Let U be an absolutely continuous random variable with PDF $P(du)$. Set

$$M^\star(u)=\operatorname{sgn}(u-\tau),$$

where τ is the unique solution to

$$2\tau=\frac{\int_{\tau}^{\infty}uP(du)}{\int_{\tau}^{\infty}P(du)}+\frac{\int_{-\infty}^{\tau}uP(du)}{\int_{-\infty}^{\tau}P(du)}.$$

Then for any measurable $M:\mathbb{R}\to\{-1,1\}$,

$$\int\left(u-\mathbb{E}[U|M^\star(u)]\right)^2P(du)\leq\int\left(u-\mathbb{E}[U|M(u)]\right)^2P(du).$$

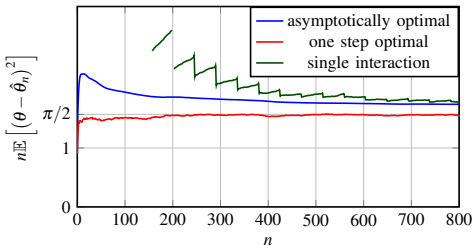

Fig. 5. Normalized empirical risk versus number of samples n for 10,000 Monte Carlo trials with $f(x)$ the standard normal density. In each trial, θ is chosen uniformly over the interval $(-1.64, 1.64)$. The single interaction strategy uses $n_1 = \lfloor \sqrt{n} \rfloor$ samples for its first stage.

□

Remark 1: The optimal threshold given in Theorem 5 is different than the one given in [38, Eq. 5] for apparently the same problem. Indeed, it seems like Equation 4 in [38] is erroneous.

By applying the optimal one-step decision rule from Theorem 5 at each step, we arrive at the following adaptive encoding and estimation scheme:

- Initialization: set $P_0(t) = \pi(t)$ and $\tau_0 = \mathbb{E}\theta$.
- For $n \geq 1$:

(1) Update the prior as

$$\begin{aligned} P_n(t) &= P(\theta = t | M^n) \\ &= \frac{P\left(\theta = t | M^{n-1}\right) P(M_n | \theta = t, M^{n-1})}{P(M_n | M^{n-1})} \\ &= \alpha_n P_{n-1}(t) F\left(M_n(t - \tau_{n-1})\right), \end{aligned} \tag{18}$$

where α_n is given by

$$\alpha_n = \left(\int_{\mathbb{R}} P_{n-1}(t) F\left(M_n(t-\tau_{n-1})\right) dt \right)^{-1}.$$

(2) The nth estimate for θ is the conditional expectation of θ given M^n, namely

$$\theta_n = \mathbb{E}\left[\theta | M^n\right] = \int_{-\infty}^{\infty} t P_n(t) dt. \tag{19}$$

(3) Obtain τ_n from equation (15) with the updated prior $P_n(t)$. Note that if $P_{n-1}(t)$ and $F(x)$ are log-concave, so does $P_n(t)$. Therefore, $P_n(t)$ is log-concave by induction and a unique solution to (15) is guaranteed by Lemma 6.

(4) Update the $(n+1)$th message as

$$M_{n+1} = \operatorname{sgn}(X_{n+1} - \tau_n) \tag{20}$$

The normalized MSE of the estimator defined by (19) and (20) is illustrated in Fig. 5 in the case where $f(x)$ is the standard normal density. Also shown in Fig. 5 are the normalized MSE of the asymptotically optimal estimator defined by (9) and (11), as well as the MSE achieved by the sample mean for the same sample realization.

V. Distributed Estimation

We now consider the distributed encoding setting described in Fig. 3 (Setting (iii) in the Introduction). In this setting each one-bit message M_i is only a function of its private sample X_i, and hence M_i characterized by its *detection region*, defined as

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

Consequently, M_i is of the form

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

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 $M_1,M_2,\ldots$ define a local asymptotic normal family.

Theorem 8: For $n\in\mathbb{N}$ and $A_n\subset\mathbb{R}$, define

$$L_n(A_1,\ldots,A_n;\theta)\triangleq\frac{1}{n}\sum_{i=1}^{n}\frac{\left(\frac{d}{d\theta}\mathbb{P}(X_i\in A_i)\right)^2}{\mathbb{P}(X_i\in A_i)\left(1-\mathbb{P}(X_i\in A_i)\right)}. \tag{21}$$

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

$$\kappa(\theta)\triangleq\lim_{n\to\infty}L_n(A_1,\ldots,A_n;\theta) \tag{22}$$

exists.

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

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

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

$$\log\frac{\mathbb{P}_{\theta+h/\sqrt{n}}(M_1,\ldots,M_n)}{\mathbb{P}_{\theta}(M_1,\ldots,M_n)}\xrightarrow{D}\mathcal{N}\left(-\frac{1}{2}h^2\kappa(\theta),h^2\kappa(\theta)\right).$$

Theorem 8 provides conditions under which $M_1,\ldots,M_n$ defines a LAN family with a precision parameter given by the limit in (22). An important conclusion of this theorem follows from the local asymptotic minimax property of estimators in LAN models (e.g. [39]):

Corollary 9: Let $\hat{\theta}_n$ be an estimator of $\theta\in\Theta$ from $M_1,\ldots,M_n$ with detection regions $A_1,\ldots,A_n$. Assume that as n goes to infinity, conditions (i)-(iii) of Theorem 8 hold. Then for any bounded, symmetric, and quasi-convex function L,

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

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

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

Corollary 9 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 10: Under conditions (i)-(iii) in Theorem 8, 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 M_i is of the form

$$M_i = \operatorname{sgn}(t_i - X_i) = \begin{cases} 1 & X_i < t_i, \\ -1 & X_i > t_i, \end{cases} \tag{23}$$

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

$$L_n(A_1, \ldots, A_n; \theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(f(t_i - \theta)\right)^2}{F\left(t_i - \theta\right) F\left(\theta - t_i\right)} = \frac{1}{n} \sum_{i=1}^{n} \eta(t_i - \theta). \tag{24}$$

A natural condition for the existence of the limit (24) 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

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

Theorem 8 implies:

Corollary 11: 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 $\{M_i = \operatorname{sgn}(X_i - t_i)\}_{i=1}^{n}$ is a LAN family with precision parameter

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

Due to local asymptotic normality of $\{M_n\}_{n=1}^{\infty}$, the maximum likelihood estimator (ML) of θ from $M_1, \ldots, M_n$, 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

$$\sum_{i=1}^{n} M_i \frac{f\left(t_i - \theta\right)}{F\left(M_i(t_i - \theta)\right)}, \tag{25}$$

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 9.

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

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

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

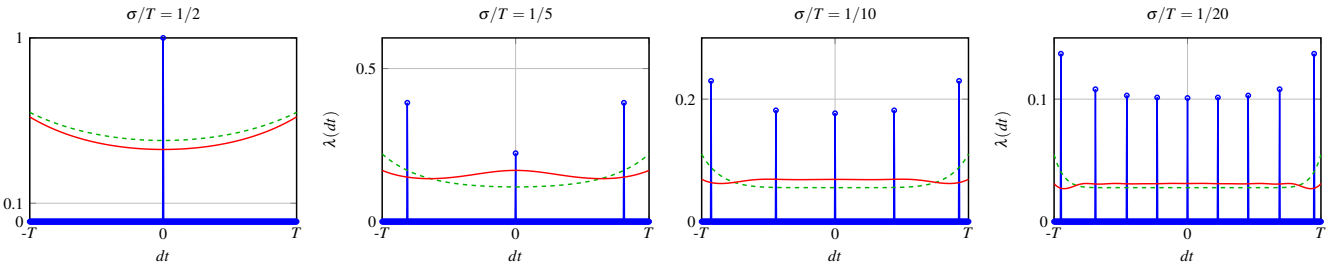

Fig. 6. Optimal threshold density $\lambda^\star(dt)$ (blue) that maximizes the ARE for $f(x) = \mathcal{N}(\theta, \sigma^2)$ and $\theta \in \Theta = [-T, T]$. The continuous curve (red) represents the reciprocal of the asymptotic risk for at a fixed $\theta \in \Theta$ under the optimal density, so 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 (27).

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:

$$\begin{aligned} \text{maximize} \quad & \inf_{\theta \in [-T,T]} \int \eta(t-\theta)\lambda(dt) \\ \text{subject to} \quad & \lambda(dt) \geq 0, \quad \int \lambda(dt) \leq 1. \end{aligned} \tag{26}$$

The objective function in (26) 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 (26) and by $\lambda^\star(dt)$ the density that achieves this maximum.

Figure 6 illustrates an approximating to $\lambda^\star(dt)$ obtained by solving a discretized version of (26) 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. 7 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

$$\begin{aligned} \kappa_{\mathsf{unif}} &\triangleq \min_{\theta \in [-T,T]} \frac{1}{2T} \int_{-T}^{T} \eta\left(t-\theta\right) dt \\ &= \frac{1}{2T} \int_{-T}^{T} \eta\left(t \pm T\right) dt = \frac{1}{2T} \int_{0}^{2T} \eta(t) dt. \end{aligned} \tag{27}$$

From Corollary 11, we conclude that the ARE under a uniform distribution is $\kappa_{\mathsf{unif}}\sigma^2$.

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. We also characterized the one-step optimal scheme in this adaptive setting, i.e., the scheme that minimizes the risk given any set of previously obtained bits.

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

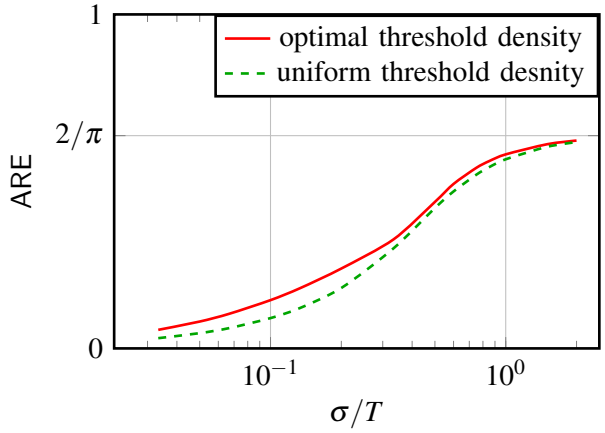

Fig. 7. Minimax ARE versus σ/T for $f(x) = \mathcal{N}(\theta, \sigma^2)$ and $\theta \in \Theta = [-T, T]$. The dashed curve (green) is the ARE under a uniform threshold density over Θ given by $K_{\mathsf{unif}}\sigma^2$, where κ_{unif} is given by (27).

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
PROOFS

Lemma 12: Let $f(x)$ be a log-concave, symmetric, and differentiable density such that $\eta(x)$ is strongly unimodal. For any $x_1 > \ldots > x_n \in \mathbb{R}$,

$$\frac{\left(\sum_{k=1}^{n}(-1)^{k+1}f(x_k)\right)^2}{\left(\sum_{k=1}^{n}(-1)^{k+1}F(x_k)\right)\left(1-\sum_{k=1}^{n}(-1)^{k+1}F(x_k)\right)} \leq \max_i \eta(x_i). \tag{28}$$

In particular, if $|x_i| > \varepsilon > 0$ for all $i = 1,\ldots,n$, then

$$\frac{\left(\sum_{k=1}^{n}(-1)^{k+1}f(x_k)\right)^2}{\left(\sum_{k=1}^{n}(-1)^{k+1}F(x_k)\right)\left(1-\sum_{k=1}^{n}(-1)^{k+1}F(x_k)\right)} \leq \eta(\varepsilon) < 0.$$

Proof of Lemma 12: 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

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

We use induction on $n \in \mathbb{N}$ to show that the LHS of (28) is bounded from above by $\max_i \eta(x_i)$. The case $n = 1$ is trivial. Assume that

$$\frac{(\delta_n)^2}{\Delta_n\left(1-\Delta_n\right)} \leq \max_i \eta(x_i) \tag{29}$$

for all integers up to $n = N-1$ and consider the case $n = N$. The maximal value of the LHS of (29) is attained for the same $(x_1,\ldots,x_N) \in \mathbb{R}^N$ that attains the maximal value of

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

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

$$\frac{f'(x_k)}{f(x_k)} = \frac{\delta_N}{2}\left(\frac{1}{\Delta_N} - \frac{1}{1-\Delta_N}\right), \quad k = 1,\ldots,N. \tag{30}$$

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, (30) is satisfied if and only if $x_1 = \ldots = x_N$. For odd N and $x_1 = \ldots = x_N$, the LHS of (29) 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 (29) as $(x_1,\ldots,x_N) \to (d,\ldots,d)$ exists and equals zero. Therefore, the maximum of the LHS of (29) is not attained at the line $x_1 = \ldots = x_N$). We now consider the possibility that the LHS of (29) 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 (29) 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. □

Proof of Theorem 2

We first prove the following lemma:

Lemma 13: 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 13: The Fisher information of M with respect to θ is given by

$$\begin{aligned}
I_\theta &= \mathbb{E}\left[\left(\frac{d}{d\theta}\log P(M|\theta)\right)^2 |\theta\right] \\
&= \frac{\left(\frac{d}{d\theta}P(M=1|\theta)\right)^2}{P(M=1|\theta)} + \frac{\left(\frac{d}{d\theta}P(M=-1|\theta)\right)^2}{P(M=-1|\theta)} \\
&= \frac{\left(\frac{d}{d\theta}\int_A f(x-\theta)\,dx\right)^2}{P(M=1|\theta)} + \frac{\left(\frac{d}{d\theta}\int_A f(x-\theta)\,dx\right)^2}{P(M=-1|\theta)} \\
&\overset{(a)}{=} \frac{\left(-\int_A f'(x-\theta)\,dx\right)^2}{P(M=1|\theta)} + \frac{\left(-\int_A f'(x-\theta)\,dx\right)^2}{P(M=-1|\theta)} \\
&= \frac{\left(\int_A f'(x-\theta)\,dx\right)^2}{P(M=1|\theta)\left(1-P(M=1|\theta)\right)}, \\
&= \frac{\left(\int_A f'(x-\theta)\,dx\right)\left(\int_A f'(x-\theta)\,dx\right)}{\left(\int_A f(x-\theta)\,dx\right)\left(1-\int_A f(x-\theta)\,dx\right)},
\end{aligned} \tag{31}$$

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 (31) 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 $\infty \leq a_1 \leq \ldots a_k$, $b_1 \leq b_k \leq \infty$ and $a_j \leq b_j$ for $j = 1, \ldots, k$. Under this assumption we have

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

so (31) can be rewritten as

$$\begin{aligned}&= \frac{\left(\sum_{j=1}^{k} f\left(a_j - \theta\right) - f\left(b_j - \theta\right)\right)^2}{\left(\sum_{j=1}^{k} F\left(b_j - \theta\right) - F\left(a_j - \theta\right)\right)} \\ &\times \frac{1}{1 - \left(\sum_{j=1}^{k} F\left(b_j - \theta\right) - F\left(a_j - \theta\right)\right)}\end{aligned} \tag{32}$$

It follows from Lemma 12 that for any $\theta \in \mathbb{R}$ and any choice of the intervals endpoints, (32) 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 M_i can is of the form

$$M_i = \begin{cases} X_i \in A_i & 1, \\ X_i \notin A_i & -1, \end{cases} \tag{33}$$

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

$$P\left(M^n|\theta\right) = \prod_{i=1}^{n} P\left(M_i|\theta, M^{i-1}\right), \tag{34}$$

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

$$I_\theta(M^n) = \sum_{i=1}^{n} I_\theta(M_i|M^{i-1}), \tag{35}$$

where $I_\theta(M_i|M^{i-1})$ is the Fisher information of the distribution of M_i given M^{i-1}. From Lemma 13 it follows that $I_\theta(M_i|M^{i-1}) \leq 4f^2(0)$. The Van Trees inequality [40], [35] now implies

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

□

Proof of Theorem 3

The algorithm given in (9) and (11) is a special case of a more general class of estimation procedures given in [36] and [37]. Specifically, (i) in Theorem 3 follows from the following simplified version of [36, Thm. 4]:

Theorem 14: [36, 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

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

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 (12).

(iii) The function $\psi(x) \triangleq \mathbb{E}\varphi(x+Z_1)$ 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

$$\left|\psi(x) - \psi'(0)x\right| \le K_2|x|^{1+\lambda}. \tag{36}$$

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

Then $\hat{\theta}_n \to \theta$ 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 $\varphi(x) = \mathrm{sgn}(x)$ and $Z_i = X_i - \theta$. We have that $\chi(x) = \mathbb{E}\mathrm{sgn}^2(x+Z_1) = 1$, so $\chi(0) = 1$. In addition,

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

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 14 are fulfilled for any $\lambda > 0$. Since

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

Theorem 3-(i) follows from Theorem 14.

In order to prove part (ii) of Theorem 3 we use the following result from [37]:

Theorem 15: [37, 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 for which the Fisher information exits. Let $\psi(x)$ and $\chi(x)$ be defined as in Theorem 14-(iii) and satisfies the conditions there. Assume in addition that $\chi(0) > 0$, condition (36) with $\lambda = 1$, and there exits K_3 such that

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

Finally, assume that the sequence $\{\gamma_n\}$ satisfies conditions (12) and (13). Then

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

We now use Theorem 15 with $g(x) = 0.5(x-\theta)^2$, $\varphi(x) = -\mathrm{sgn}(-x)$, $Z_n = \theta - X_n$ and $U_n = \theta_n$. From (37) we have

$$\begin{aligned}\theta_n &= \theta_{n-1} + \gamma_n\mathrm{sgn}(\theta - \theta_{n-1} - Z_n) \\ &= \theta_{n-1} + \gamma_n\mathrm{sgn}(X_n - \theta_{n-1}),\end{aligned}$$

so the estimator $\hat{\theta}_n$ defined by $\hat{\theta}_n$ equals to the one defined by (11) and (9). Note that

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

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 15 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 15 implies

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

□

Proof of Theorem 5

In this subsection we prove Lemmas 7 and 6 which together implies Theorem 5.

Proof of Lemma 7: Any one-bit message $M(u) \in \{0,1\}$ is characterized by two decision region $A_1 = M^{-1}(1)$ and $A_{-1} = M^{-1}(-1)$, so that $\mathbb{E}[U|M(U)]$ assumes only two values: $\mu_1 = \mathbb{E}[U|M(U) = 1]$ and $\mu_{-1} = \mathbb{E}[U|M(U) = -1]$. We claim that a necessary condition for $M(u)$ to be optimal is that the sets A_1 and A_{-1} are the Voronoi sets on $\mathbb{R}$ corresponding to the points μ_1 and μ_{-1}, respectively, modulo a set of measure $P(du)$ zero. Indeed, assume by contradiction that for such an optimal partition there exists a set $B \subset A_1$ with $\mathbb{P}(U \in B) > 0$ such that $(b-\mu_1)^2 > (b-\mu_{-1})^2$. The expected square error in this partition satisfies:

$$\begin{aligned}
&\int_{\mathbb{R}} (u - \mathbb{E}[U|M(u)])^2 P(du) = \\
&\int_{A_1} (u-\mu_1)^2 P(du) + \int_{A_{-1}} (u-\mu_{-1})^2 P(du) \\
&= \int_{A_1 \setminus B} (u-\mu_1)^2 P(du) + \int_B (u-\mu_1)^2 P(du) \\
&\quad + \int_{A_{-1}} (u-\mu_{-1})^2 P(du) \\
&> \int_{A_1 \setminus B} (u-\mu_1)^2 P(du) + \int_B (u-\mu_2)^2 P(du) \\
&\quad + \int_{A_{-1}} (u-\mu_{-1})^2 P(du),
\end{aligned}$$

so the partition $A_1' = A_1 \setminus B$, $A_{-1}' = A_{-1} \cup B$ attains lower error variance which contradicts the optimality assumption and proves our claim. It is evident that Voronoi partition of the real line corresponding to μ_1 and μ_{-1} is of the form $A_{-1} = (-\infty, \tau)$, $A_1 = (\tau, \infty)$ where the point τ is of equal distance from μ_1 and μ_{-1}, namely $\tau = \frac{\mu_1 + \mu_{-1}}{2}$. From these two conditions (which are a special case of the conditions derived in [41] for two quantization regions) we conclude that τ must satisfy the equation

$$2\tau = \frac{\int_\tau^\infty u P(du)}{\int_\tau^\infty P(du)} + \frac{\int_{-\infty}^\tau u P(du)}{\int_{-\infty}^\tau P(du)}.$$

□

Proof of Lemma 6: Any solution to (17) is a solution to $h^+(x) = h^-(x)$ where

$$h^+(x) = \frac{\int_x^\infty u f(u) du}{\int_x^\infty f(u) du} - x$$

and

$$h^-(x) = x - \frac{\int_{-\infty}^x u f(u) du}{\int_{-\infty}^x f(u) du}.$$

We now prove that $h^+(x)$ is monotonically decreasing while $h^-(x)$ is increasing, so they meet at most at one point. The derivative of $h^-(x)$ is given by

$$1 - \frac{f(\tau) \int_{-\infty}^\tau f(x)(\tau - x) dx}{\left(\int_{-\infty}^\tau f(x) dx\right)^2}. \tag{38}$$

Denote $F(x) = \int_{-\infty}^{x} f(u)du$. Using integration by parts in the numerator and from the fact that $\lim_{\tau\to-\infty} \tau \int_{-\infty}^{\tau} f(x)dx = 0$, the last expression can be written as

$$1 - \frac{f(\tau)\int_{-\infty}^{\tau} F(x)dx}{(F(\tau))^2}.$$

Log-concavity of $f(x)$ implies log-concavity of $F(x)$, so that we can write $F(x) = e^{g(x)}$ for some concave and differentiable function $g(x)$. Moreover, we have $f(x) = g'(x)e^{g(x)}$ where, by concavity of $g(x)$, the derivative $g'(x)$ of $g(x)$ is non-increasing. With these notation we have

$$\begin{aligned}\frac{f(\tau)\int_{-\infty}^{\tau} F(x)dx}{(F(\tau))^2} &= \frac{g'(\tau)e^{g(\tau)}\int_{-\infty}^{\tau} e^{g(x)}dx}{e^{2g(\tau)}} \\ &= e^{-g(\tau)}\int_{-\infty}^{\tau} g'(\tau)e^{g(x)}dx \\ &\le e^{-g(\tau)}\int_{-\infty}^{\tau} g'(x)e^{g(x)}dx \\ &= e^{-g(\tau)}F(\tau) = 1.\end{aligned}$$

(where the second from the last step follows since $g'(x) \le g'(\tau)$ for any $x \le \tau$). If follows that (38) is non-negative and thus $h^-(x)$ is monotonically increasing. Since

$$h^+(-x) = x - \frac{\int_{-\infty}^{x} uf(-u)du}{\int_{-\infty}^{x} f(-u)du},$$

the fact that $h^+(x)$ is monotonically decreasing follows from similar arguments. Moreover, since the derivatives of $h^+(x)$ and $h^-(x)$ never vanish at the same time over any open interval, their difference cannot be constant over any interval. Finally, since

$$\lim_{x\to-\infty} h^+(x) = \lim_{x\to\infty} h^-(x)$$

and since non of these functions are constant, monotonicity of $h^+(x)$ and $h^-(x)$ implies that they must meet at some $x \in \mathbb{R}$. □

Proof of Theorem 8

The log probability mass distribution of $M^n = (M_1, \ldots, M_n)$ is given by

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

Consequently,

$$\log \frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(m^n)}{\mathbb{P}_\theta(m^n)} = \sum_{i=1}^{n} \frac{m_i + 1}{2} \log \frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)} + \sum_{i=1}^{n} \frac{1 - m_i}{2} \log \frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i \notin A_i)}{\mathbb{P}_\theta(X_i \notin A_i)}. \tag{39}$$

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

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

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

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

Similarly, we have

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

From (39) we obtain

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

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

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

Then

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

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

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

Then

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

Proof of Claim I: First note that

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

In addition,

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

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

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

and

$$\frac{\sum_{i=1}^n\mathbb{E}\left|U_i\right|^{2+\delta}}{\left(\sum_{i=1}^n\mathbb{E}U_i^2\right)^\delta} = \frac{\frac{1}{n^{1+\delta}}\sum_{i=1}^n\mathbb{E}|U_i|^{2+\delta}}{\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}U_i^2\right)^\delta}. \tag{40}$$

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

$$\frac{1}{n}\sum_{i=1}^n\mathbb{E}|U_i|^{2+\delta}<M \tag{41}$$

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

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

where the last transition is because

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

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 the appendix:

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 $x_1\ge\ldots\ge x_n\in\mathbb{R}$,

$$\frac{\left|\sum_{k=1}^n(-1)^{k+1}f(x_k)\right|^{2+\delta}}{\left(\sum_{k=1}^n(-1)^{k+1}F(x_k)\right)^{1+\delta}\left(1-\sum_{k=1}^n(-1)^{k+1}F(x_k)\right)^{1+\delta}}\le 2^\delta f^{2+\delta}(0). \tag{42}$$

Lemma 16 implies

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

In particular, we conclude that

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

and thus for any $\delta > 0$ the numerator of (40), as well as the entire expression, goes to zero. From Lyaponov's central limit theorem we conclude that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i \xrightarrow{D} \mathscr{N}\left(0, \kappa(\theta)\right).$$

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:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}V_i = L_n(A_1,\ldots,A_n) \to \kappa(\theta) \tag{43}$$

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 (43) 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

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

The proof is by induction on n. For the case $n=1$ the LHS of (42) 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

$$(1+\delta)\left|\frac{\eta'(x)}{\eta(x)}\right| \ge \delta\left|\frac{f'(x)}{f(x)}\right|, \qquad x>0. \tag{44}$$

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 (44) is satisfies for $\delta > 0$ small enough.

Next, assume that

$$\frac{(\delta_n)^{2+\delta}}{(\Delta_n)^{1+\delta}\left(1-\Delta_n\right)^{1+\delta}} \le \eta_\delta(0) = 2^\delta f^{2+\delta}(0). \tag{45}$$

for all integers up to $n = N-1$ and consider the case $n = N$. The maximal value of the LHS of (45) 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

$$\frac{f'(x_k)}{f(x_k)} = \frac{\delta_N(1+\delta)}{2+\delta}\left(\frac{1}{\Delta_N} - \frac{1}{1-\Delta_N}\right), \quad k = 1,\ldots,N. \tag{46}$$

From the same reason as in the proof of Lemma 12, (46) is satisfied if and only if $x_1 = \ldots = x_N$. For odd N and $x_1 = \ldots = x_N$, the LHS of (45) 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 (45) as $(x_1,\ldots,x_N) \to (d,\ldots,d)$ exists and equals zero. Therefore, the maximum of the LHS of (45) is not attained at the line $x_1 = \ldots = x_N$). We now consider the possibility that the LHS of (45) 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 (45) 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. □

Proof of Theorem 10

Let Ξ be the set of points $\theta \in \Theta$ for which $\kappa(\theta) = \eta(0)$. Since $M_1, M_2, \ldots$ satisfy the conditions in Theorem 8, for $\theta \in \Xi$ if and only if

$$\lim_{n\to\infty} L_n(A_1,\ldots,A_n;\theta) = \eta(0). \tag{47}$$

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

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

where $a_{i,1} \leq b_{i,1} \leq \ldots \leq a_{i,K}, b_{i,K}$, 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 $\text{card}(S_n(\theta,\varepsilon))/n \to 1$. Indeed, for such θ we have

$$\begin{aligned}
&L_n(A_1,\ldots,A_n;\theta) \\
&= \frac{1}{n} \sum_{i \in S_n(\varepsilon,\theta)} \frac{\left(\sum_{k=1}^K f(\theta-b_{i,k}) - f(\theta-a_{i,k})\right)^2}{\sum_{k=1}^K \left(F(\theta-b_{i,k}) - F(\theta-a_{i,k})\right)\left(1-\sum_{k=1}^K \left(F(\theta-b_{i,k}) - F(\theta-a_{i,k})\right)\right)} \\
&+ \frac{1}{n} \sum_{i \notin S_n(\varepsilon,\theta)} \frac{\left(\sum_{k=1}^K f(b_{i,k}-\theta) - f(a_{i,k}-\theta)\right)^2}{\sum_{k=1}^K \left(F(\theta-b_{i,k}) - F(\theta-a_{i,k})\right)\left(1-\sum_{k=1}^K \left(F(\theta-b_{i,k}) - F(\theta-a_{i,k})\right)\right)} \\
&\overset{(a)}{\le} \frac{\text{card}\left(S_n(\theta,\varepsilon)\right)}{n}\eta(0) + \frac{n-\text{card}\left(S_n(\theta,\varepsilon)\right)}{n}\eta(\varepsilon)
\end{aligned} \tag{48}$$

where (a) follows from Lemma 12 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 $\text{card}\left(S_n(\theta,\varepsilon)\right)/n \to 1$, (48), 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 $N \ge 2K+1$ distinct elements $\theta_1,\ldots,\theta_N \in \Xi$. Since each A_i consists of at most K intervals, we have that

$$\text{card}(\cup_{i=1}^n \mathscr{B}_i) \le 2nK. \tag{49}$$

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

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

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 (49).