Mean Estimation from Single-Bit Measurements

***Abstract*—We consider the problem of estimating the mean of a symmetric log-concave distribution under the constraint that 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 number of bits, from this distribution. Under an adaptive scheme 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 standard mean estimator without bit limitation, is the efficiency of the median. For the distributed scheme we consider the setting where each bit is obtained by comparing against a prescribed threshold value. We show that the maximum likelihood estimator in this case is asymptotically local minimax, and its asymptotic relative efficiency is finite although strictly larger than that of the sample median.**

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 P_X with a 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 single-bit message dramatically affects the performance in estimating θ. We therefore distinguish among three settings:

(i) *Centralized* encoding: all n encoders confer and produce a single n bit message.

(ii) *Adaptive* or *sequential* encoding: the ith encoder observes the ith sample and the $i-1$ previous messages.

(iii) *Distributed* encoding: the ith message is only a function of the ith 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 example of each of settings (i)-(iii), respectively:

(i) **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.

(ii) **analog-to-digital conversion via sigma-delta modulation (SDM):** in SDM 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 is 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.

(iii) **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 answers a single yes/no questions about its individual 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 $\sigma^2/n + o(n^{-1})$ where σ^2. Estimator of this form includes the empirical mean of $X_1, \ldots, X_n$, and under some conditions, optimal Bayes estimations. We note that the performance under estimating a signal that may vary with each noisy measurement is always harder that our setting where the parameter (the mean) remains

fixed. Therefore, the excess MSE risk due to the one-bit per measurement constraint we derive in this paper can serve as the most optimistic estimate for the degradation in performance in estimating a signal that vary in time.

In setting (i), the estimator can evaluate the empirical mean of the samples 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 results does not hold even in setting and (ii): the ARE of any adaptive estimation scheme is at least that of the sample median. Specifically, when P_X is normal, this ARE equals $2/\pi \approx 0.637$, showing that the one-bit constraint increases the effective sample size in estimating θ by at least $2/\pi \approx 1.57$ compared to estimating it without the bit constraint. We also show that this lower bound on the ARE and is tight by providing an estimator that attains it. Clearly, the minimal penalty on the MSE in setting (ii) also holds under setting (iii), although the question whether this MSE is achievable (or otherwise, what is the minimal MSE) remains open. Instead, in setting (iii) we restrict ourselves to estimators from messages obtained by comparison against a prescribed value (that may be different for each sample). We show that the maximum likelihood (ML) estimator for θ from these messages is asymptotically local minimax, and its asymptotic variance is strictly greater than the variance of the median. Thus, at least when limited to threshold detection, the ability to adapt the threshold allows for a more efficient estimation.

Even though 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 [2]. 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. Note that the only missing parameter in our setting is the mean, which, under setting (i), is known to the encoder with uncertainty interval proportional to $\sigma/\sqrt{n}$. Therefore, while it is clear that uncertainty due to quantization decreases exponentially in the number of bits n leading to ARE 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 detector given the previous messages, this characterization cannot be use to derive a lower bound on the ARE. Instead, our result on the median efficiency being the minimal ARE is obtained by bounding the Fisher information of any n adaptive messages and using the van Trees version of the information inequality [3].

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 easily solved by a bisection style method over the parameter space. Therefore, the general case of non-zero variance is a reminiscent of the noisy binary search problem with possibly infinite number of unreliable tests [4], [5]. 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 θ corrupted by a Gaussian noise. Using a SDM, Wong and Gray [1] showed that the output of the modulator converges to the true constant input almost surely, so that a 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 Gaussian nosie is at most $\sigma^2\pi/2$ over the number of feedback iterations. Baraniuk et. al [6] also considered adaptive one-bit measurements in the context of analog-to-digital conversion, although without noise at the input. Their result of an exponential MSE decaying rate clearly does not hold in the noisy setting (stated otherwise, the MSE can decay exponentially up to the noise level).

One-bit measurements in the distributed setting (iii) was considered in [7], [8], [9], [10], [11], but without optimizing the encoders or their detection rule. The work of [12] addresses the counterpart of our setting (iii) in the case of hypothesis testing, although the results there cannot be extended to parametric estimation.

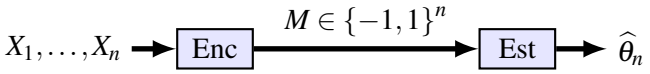

Fig. 1: Centralized encoding using one bit per sample on average.

More generally, when the parameter space Θ is finite, the characterization of the optimal detection rules was considered by Tsitsiklist in [13]. It was shown there that if the cardinality of Θ is at most M and the probability of error criterion is used, than 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 sestting [14], it was shown that, with a specific two-stages feedback, there is no gain in feedback compared to the fully distributed setting. Our results implies that the 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 than when Θ is an open set.

As we explain in details in Section III, the remote multiterminal source coding problem, also known as the CEO problem [15], [16], [17], [18], leads to a 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 one-bit quantization of each sample.

Finally, we note that our settings (ii) and (iii) can be obtained as special cases of [19] that consider adaptive and distributed estimation protocols for m machines, each has access to n/m independent samples. The main result of [19] 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). Other related works include statistical inference under multiterminal data compression [20], [21], and one-bit quantization constraints in MIMO detection in wireless communication [22].

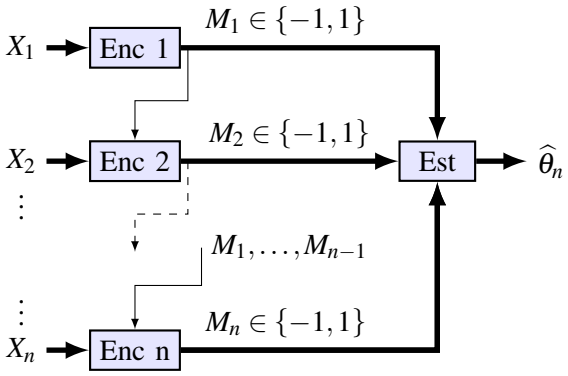

Fig. 2: Adaptive single-bit encoding: the ith encoder delivers a single bit message which is a function of its private sample X_i and the previous messages $M_1,\ldots,M_{i-1}$.

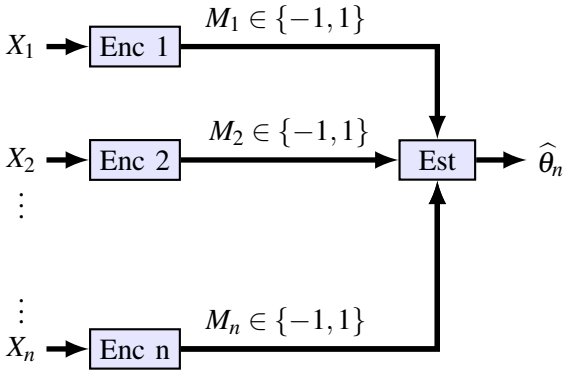

Fig. 3: Distributed single-bit encoding: the single-bit message produced by each encoder is only a function of its private sample X_i.

II. Problem Formulation

Let Θ be a closed interval of the real line. Let $f(x)$ be a symmetric log-concave density function with a finite second moment σ^2. We denote by P_X the probability distribution with density $f(x-\theta)$, where $\theta \in \Theta$. Therefore, P_X is an absolutely continuous log-concave distribution with mean θ, variance σ^2. Moreover, symmetry and log-concavity of $f(x)$ implies that P_X has a single mode at $x = \theta$.

In some cases we assume that θ is drawn once from the prior distribution π on the parameter space Θ. In this cases we assume that π is 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 the parameter θ from a set of n binary messages $M^n =$

$(M_1,\ldots,M_n)$, obtained from $X^n=(X_1,\ldots,X_n)$ under three possible scenarios:

(i) Centralized $M^n(X^n)$ (Fig. 1).
(ii) Adaptive $M_i(X_i,M^{i-1})$, $i=1,\ldots,n$ (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 the prior distribution $\pi(\theta)$. The main problem we consider is the minimal value of (1) as a function of n, under various choices of the encoding functions in cases (i), (ii), and (iii).

Particular interest is given to the ARE of estimators with respect to an asymptotically normal efficient estimator without the one-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 [23, 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

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

then the ARE of $\hat{\theta}_n$ is finite and equals σ^2/V.

III. Preliminary Results

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, one may suspect that estimation in these cases is trivial as in the centralized setting (i), where it is easy to attain ARE 1. In this section we settle this skepticism by showing that (1) a consistent estimator with an asymptotically normal distribution always exists in setting (iii), and (2) the ARE in setting (iii) with $P_X=\mathcal{N}(\theta,\sigma^2)$ is bounded from below by $3/4$.

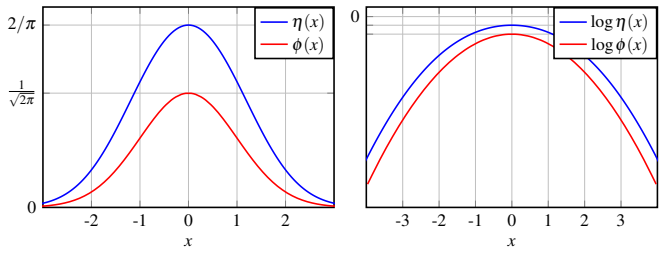

Fig. 4: The function $\eta(x)=f^2(x)/F(x)F(-x)$ (blue) for $f(x)=\phi(x)$ the standard normal density (red).

A. Consistent Estimation

Denote by M_i is the indicator of the event $X_i>\theta_0$ for some arbitrary θ_0. Note that

$$\frac{1}{n}\sum_{i=1}^n M_i\xrightarrow{a.s.}F(\theta),$$

where $F(x)$ is the cumulative distribution of $X_1-\theta$ that is assumed to be known. Therefore,

$$\hat{\theta}_n=F^{-1}\left(\frac{1}{n}\sum_{i=1}^n M_i\right) \tag{2}$$

is a consistent estimator for θ under setting (iii). Moreover, it can be verified, as we do in Section V, that $\hat{\theta}_n$ is asymptotically normal with variance that is the inverse of

$$\eta(\theta-\theta_0)\triangleq\frac{f^2(\theta-\theta_0)}{F(\theta-\theta_0)\left(1-F(\theta-\theta_0)\right)}. \tag{3}$$

In particular, the ARE of $\hat{\theta}_n$ is finite and equals $\sigma^2/\eta(\theta-\theta_0)$. For $f(x)$ the normal density, it is shown in [24] and [25] that the function $\eta(x)$ is a strictly decreasing function of $|x|$. The following proposition, generalizes this property of $\eta(x)$ to any log-concave distribution.

Proposition 1: Let $f(x)$ be log-concave, symmetric and differntiable probability density function. Then

(i) The function

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

is symmetric and strictly decreasing for any $x>0$ in the support of $f(x)$. In particular, $\eta(x)\le\eta(0)=4f^2(0)$.
(ii) $\eta(x)$ goes to 0 as $|x|$ goes to infinity.

Proof: See Appendix **??**. □

At $\theta = \theta_0$ the variance of $\hat{\theta}_n$ of (2) equals $1/(4f(0)^2)$ which is the variance of the median. However, since $\eta(x)$ vanishes as $|x| \to \infty$ (see illustration in Fig. 4 for the normal density), the variance of $\hat{\theta}_n$ may be very large when θ is away from θ_0. In this case, the estimator $\hat{\theta}_n$ has little practical value unless Θ is very small compared to σ^2.

In Section IV we present an asymptotically normal estimator that attains the variance of the median $1/\eta(0)$ regardless of the radius of Θ or the prior $\pi(\theta)$. We also show in Section IV that no estimator can attain variance that is smaller than $1/\eta(0)$, so that the ARE of any estimator is at least $\sigma^2\eta(0) = (2f(0)\sigma)^2$.

B. Lower bound from the CEO

The CEO setting includes n encoders, each has access to a noisy version of a random source sequence [15]. The ith encoder observes k noisy source symbols and transmit $R_i k$ bits to a central estimator.

Assuming that θ is drawn once from the prior $\pi(d\theta)$, our mean estimation problem from one-bit samples under distributed encoding (setting (iii) in the Introduction) corresponds to the Gaussian CEO setting with $k = 1$ source realization: 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. As a result, a lower bound on the MSE distortion in estimating θ in the distributed encoding setting is given by the MSE in the optimal source coding scheme for the CEO with: n terminals of codes rates $R_1 = \ldots = R_n = 1$, a Gaussian observation noise at each terminal of variance σ^2, and an arbitrary number of k independent draws of θ. Note that the difference between the CEO and ours lays in the privilege of each of the CEO encoders to describe k realizations of θ using k bits with MSE averaged over these realization, whereas our setting only allows $k = 1$.

By using an expression for the minimal MSE in the Gaussian CEO as the number of terminals goes to infinity, we conclude the following:

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

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

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

Proof: We consider the expression [26, Eq. 10] that gives the minimal distortion $D^\star$ in the CEO with L observers and under a total sum-rate $R_\Sigma = R_1 + \ldots + R_L$:

$$R_\Sigma = \frac{1}{2}\log^+\left[\frac{\sigma_\theta^2}{D^\star}\left(\frac{D^\star L}{D^\star L - \sigma^2 + D^\star\sigma^2/\sigma_\theta^2}\right)^L\right]. \tag{5}$$

Assuming $R_\Sigma = n$ and $L = n$, we get

$$n = \frac{1}{2}\log_2\left[\frac{\sigma_\theta^2}{D^\star}\left(\frac{D^\star n}{D^\star n - \sigma^2 + D^\star\sigma^2/\sigma_\theta^2}\right)^n\right]. \tag{6}$$

The value of $D^\star$ that satisfies the equation above describes the MSE under an optimal allocation of the sum-rate $R_\Sigma = n$ among the n encoders. Therefore, $D^\star$ 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 encoding setting. By considering $D^\star$ in (6) 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 (4) 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 [27], 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 (4) when σ_θ goes to zero.

From the formulation of the CEO problem, it follows that the difference between the MSE lower bound (4) 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 3 below, states that the ARE of any adaptive estimator cannot be larger than $(2f(0)\sigma)^2$, which the

ARE of the median of 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 message M_i that minimizes the MSE given X_i and the previous M^{i-1} messages is chosen. Numerical simulations for the case of $f(x)$ the normal density function shows that the ARE of the estimator described by this scheme is $(2f(0)\sigma)^2$.

A. Maximal efficiency of adaptive one-bit schemes

Our first results asserts that the ARE of any adaptive estimation scheme is bounded from above by $(2f(0)\sigma)^2$, as follows from the following theorem:

Theorem 3 (minimal relative effeciency): Let $\hat{\theta}_n$ be any estimator of θ in the adaptive setting of Fig. 2 and assumes that $\pi(\theta)$ converges to zero at the endpoints of the interval Θ. Then

$$\mathbb{E}\left[(\theta-\theta_n)^2\right] \geq \frac{\pi\sigma^2}{2n+\pi\sigma^2 I_0} = \frac{\sigma^2}{4f^2(0)n} + o(n^{-1}),$$

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 single-bit messages with respect to θ. Once this bound is achieved, the result follows by using the van-Trees inequality [28, Thm. 2.13],[3] 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 given in the Appendix.

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

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:

$$\theta_n = \theta_{n-1} + \gamma_n \mathrm{sgn}(X_n - \theta_{n-1}), \quad n=1,2,\ldots, \tag{7}$$

and set the nth step estimation as

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i. \tag{8}$$

For the estimator defined by (7) and (8) we have the following results:

Theorem 4: Consider the sequence $\left\{\hat{\theta}_n, n\in\mathbb{N}\right\}$ defined by (8).

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

$$\begin{cases} \frac{\gamma_n-\gamma_{n+1}}{\gamma_n} = o(\gamma_n), \\ \sum_{n=1}^{\infty}\frac{\gamma_n^{(1+\lambda)/2}}{\sqrt{n}} < \infty, \quad \text{for some } 0<\lambda\leq 1 \end{cases} \tag{9}$$

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

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

(ii) Assume that in addition to (9), $\{\gamma_n, n\in\mathbb{N}\}$ satisfies

$$\begin{cases} \gamma_n = o(n^{-2/3}), \\ \sum_{n=1}^{\infty}\gamma_n = \infty. \end{cases} \tag{10}$$

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

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

Proof: The asymptotic behavior of (8) is obtained from [29, Thm. 4] and [30, Thm. 2]. The details are in the Appendix.

Theorem 4 implies that the estimator $\hat{\theta}_n$, defined by (8) and (7), attains the maximal ARE as established by Theorem 3.

Note that θ_0 is not explicitly defined in equation (8). A reasonable initialization for θ_0 is $\theta_0 = \mathbb{E}[\theta]$, although Theorem 4 implies that the asymptotic behavior of the estimator is indifferent to this initialization. Thus, the optimal efficiency is attained regardless of the prior distribution on θ or the radius of the parameter space Θ. Nevertheless, the bound in Theorem 3 suggests that the non-asymptotic MSE can be reduced significantly whenever the location information I_0 is large. In what follows, we consider a greedy estimation procedure that uses the optimal decision rule given the prior distribution and the previous messages. In particular, this procedure exploit the prior information on θ provided by its prior.

C. 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 word, 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

$$M^\star = \mathrm{sgn}(X - \tau), \tag{11}$$

where τ satisfies the equation

$$\tau = \frac{m^-(\tau) + m^+(\tau)}{2}, \tag{12}$$

with

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

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

$$\mathbb{E}\left(\theta - \hat{\theta}(M)\right)^2 \geq \mathbb{E}\left(\theta - \mathbb{E}[\theta | M^\star]\right)^2, \tag{13}$$

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

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

$$2x = \frac{\int_x^{\infty} u f(u) du}{\int_x^{\infty} f(u) du} + \frac{\int_{-\infty}^{x} u f(u) du}{\int_{-\infty}^{x} f(u) du} \tag{14}$$

has a unique solution.

Lemma 7: Let U be an absolutely continuous random variable with pdf $P(du)$. Then the one-bit message $M^\star \in \{-1, 1\}$ that minimizes

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

is given by

$$M^\star = \mathrm{sgn}(U - \tau),$$

where τ is the unique solution to

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

□

Remark 1: The value of the optimal threshold as given in Theorem 5 is different than the one given in [31, Eq. 5] for apparently the same problem. Indeed, it seems like Equation 4 there is erroneous.

By applying at each step the optimal one-step decision rule from Theorem 5, 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

$$\begin{aligned} P_n(t) &= P(\theta = t | M^n) \\ &= \frac{P\left(\theta = t | M^{n-1}\right) P(M_n | \theta = t, M^{n-1})}{P(M_n | M^{n-1})} \\ &= \alpha_n P_{n-1}(t) F\left(M_n(t - \tau_{n-1})\right), \end{aligned} \tag{15}$$

  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

$$\theta_n = \mathbb{E}[\theta | M^n] = \int_{-\infty}^{\infty} t P_n(t) dt. \tag{16}$$

  (3) Obtain τ_n from equation (12) with the updated prior $P_n(t)$. Note that $F(x)$ is log-concave so the updated prior $P_n(t)$ remains log-concave and thus a unique solution to (12) is guaranteed by Lemma 6.
  (4) Update the $(n+1)$th message as

$$M_{n+1} = \mathrm{sgn}(X_{n+1} - \tau_n) \tag{17}$$

Since equation (12) has no analytic solution in general, it is hard to derive the asymptotic behavior of the estimator defined by (16) and (17). We conjecture, however, that it attains the asymptotic relative efficiency of $4f^2(0)$ under the same conditions on $f(x)$ in the problem formulation. In Fig. 5 we demonstrate this convergence 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 (7) and (8), as well as the MSE achieved by the sample mean for the same sample realization.

V. Distributed Estimation from Threshold Detectors

We now consider the distributed encoding setting described in Fig. 3, in which each single-bit message is independent of the other messages. Instead of searching for the optimal choice of the messages, we only consider messages that are obtained each sample against a

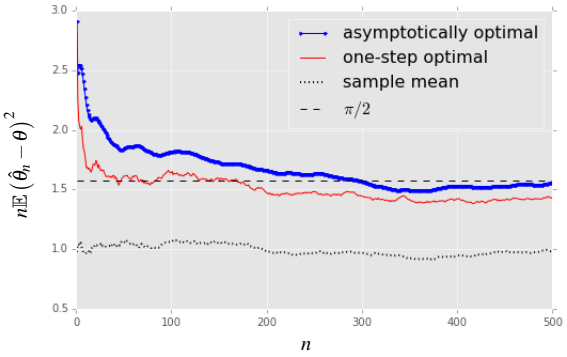

Fig. 5: Normalized empirical risk $n\left(\hat{\theta}_n-\theta\right)^2$ versus number of samples n for 500 Monte Carlo trials. In each trial, θ is chosen uniformly in the interval $(-3,3)$.

prescribed threshold. Namely, each message is of the form

$$M_i=\operatorname{sgn}(t_i-X_i)=\begin{cases}1 & X_i<t_i,\\ -1 & X_i>t_i,\end{cases} \tag{18}$$

where $t_i\in\mathbb{R}$ is the *threshold* of the ith detector applied by the encoder on it sample X_i. The *density* λ_n of the set of threshold values $\mathscr{T}_n\triangleq\{t_1,\ldots,t_n\}$ is defined as their empirical distribution, i.e., for an interval $I\subset\mathbb{R}$,

$$\lambda_n(I)=\frac{\operatorname{card}(\mathscr{T}_n\cap I)}{n}.$$

We further assume that λ_n converges (weakly) to a probability measure $\lambda(dt)$ on $\mathbb{R}$. For example, this convergence occurs whenever $t_1,\ldots,t_n$ are drawn independently from the probability distribution $\lambda(dt)$.

A. Local Asymptotic Minimax Estimation

In order to estimate θ from the messages M^n, we consider the maximum likelihood (ML) estimator. The log-likelihood function of θ is given by

$$l(M^n|\theta)=\sum_{i=1}^n\log F\left(M_i\left(t_i-\theta\right)\right).$$

Since $F(x)$ is a log concave function, the log-likelihood function has a unique maximizer $\hat{\theta}_n$ which is the ML estimator. Specifically, the ML estimator $\hat{\theta}_{ML}$ is the unique root of

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

Next, we show that under the assumptions above the sequence of messages M^n defines a local asymptotic normal (LAN) family of probability distributions. As a result, we conclude that the ML estimator in estimating θ from M^n satisfies a local asymptotic minimax property with respect to the *precision* parameter of this LAN family [32].

Theorem 8: Let $f(x)$ be a log-concave, symmetric, and twice differentiable probability density function. Consider the sequence M^n of threshold detectors

$$M_i=\operatorname{sgn}(X_i-t_i),\quad i=1,\ldots,n,$$

with threshold density converges to a probability measure λ. The the following two statements hold:

(i) Any estimator θ_n of θ which is a function of M^n satisfies

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

where

$$K(\theta)\triangleq\int\eta\left(t-\theta\right)\lambda(dt),$$

and

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

(ii) The asymptotic distribution of the ML estimator $\hat{\theta}_n$ is given by

$$\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{D}\mathscr{N}\left(0,1/K(\theta)\right).$$

Remark 2: The condition that $\lambda_n(dt)$ converges weakly to $\lambda(dt)$ can be relaxed to the condition that for any $\theta\in\Theta$,

$$\frac{1}{n}\sum_{i=1}^n\eta\left(t_i-\theta\right)$$

converges to $K(\theta)$.

Proof: See Appendix. □

It follows from Theorem 8 that the asymptotic MSE of the ML estimator is $1/(nK(\theta))$, and that this MSE is asymptotically minimal with respect to all local alternative estimators for θ. In particular, we conclude that the ARE of any estimator based on threshold detection is $K(\theta)\sigma^2$.

It follows from Proposition 1 that $\eta(x)$ attains its maximum at the origin. Since the thresholds density λ integrates to 1, it follows that

$$K(\theta)\le\sup_{t\in\mathbb{R}}\eta\left(t-\theta\right)=\eta(0)=4f^2(0).$$

This upper bound on $K(\theta)$ implies that the ARE of any distributed estimator is not smaller than $(2f(0)\sigma)^2$, a

fact that agrees with the lower bound under adaptive estimation derived in Theorem 3. Furthermore, since this upper bound on $K(\theta)$ is attained whenever the density λ is the mass distribution at θ which is unknown apriori, Theorem 8 implies that estimation in the distributed setting using threshold detection is strictly sub-optimal compared to the adaptive setting. In other words, the ability to adaptively choose the threshold value strictly improves the relative efficiency compared to a non-adaptive threshold selection.

In what follow, we consider the asymptotic threshold density $\lambda(dt)$ that minimize the asymptotic MSE under the worst choice of $\theta \in \Theta$, or, what is the same, the expected value of $\mathbb{E}[1/K(\theta)]$ under the least favorable prior $\pi(\theta)$.

B. Minimax Threshold Density

The distribution $\lambda(dt)$ that minimizes the asymptotic MSE $K(\theta)$ over the worst choice of θ in $\Theta = [-b,b]$ is given as the solution to the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & \min_{\theta\in[-b,b]} \int \eta(t-\theta)\lambda(dt) \\ \text{subject to} \quad & \lambda(dt) \geq 0, \quad \int \lambda(dt) \leq 1. \end{aligned} \tag{19}$$

Denote the maximal value of the objective in (19) by $K^\star(b)$. Since the objective function in (19) is concave in λ, this problem can be solved using a convex program. By discretizing the interval $[-b,b]$ using N_θ values $\theta_1,\ldots,\theta_{N_\theta}$ and the real line using N_λ values $\lambda_1,\ldots,\lambda_{N_\lambda}$, the discrete version of (19) is the following linear program (LP) in the variables $K \in \mathbb{R}$ and $\lambda \in \mathbb{R}^{N_\lambda}$:

$$\begin{aligned} \text{maximize} \quad & K \\ \text{subject to} \quad & K \leq \mathbf{H}\lambda \\ & \lambda \geq 0, \quad \mathbf{1}^T\lambda \leq 1, \end{aligned} \tag{20}$$

where $\mathbf{H}_{i,j} = \eta(t_i - \theta_j)$, $i = 1,\ldots,N_\lambda$, $j = 1,\ldots,N_\theta$.

Remark 3: The number of variables in (20) is $N_\lambda + 1$ and number of constraints is $1 + N_\lambda + N_\theta$. Since an LP has an optimal solution at which the number of constraints for which equality holds is no smaller than the number of variables [33], there exists an optimal λ with support over no more than N_θ points. Therefore, in approximating the solution of (19) using (20), it is enough to take $N_\lambda = N_\theta$.

Figure 6 illustrates $\lambda^\star$ obtained as the solution to (20) and the function $\mathbf{H}\lambda^\star$ for the case of $f(x)$ the standard normal density. The minimal asymptotic ML risk $K^\star$ is

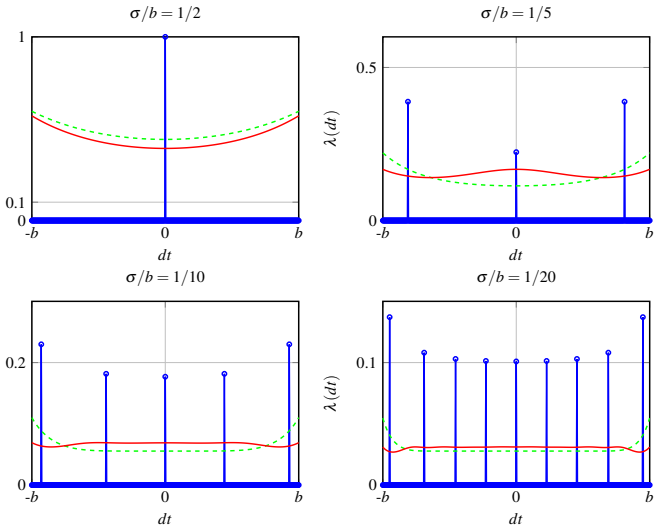

Fig. 6: The optimal density $\lambda^\star$ (blue) that minimizes the maximum asymptotic ML risk for $f(x)$ a standard normal density $\Theta = [-b,b]$, and various values of σ. The continuous curve (red) represents the asymptotic risk for a fixed $\theta \in [-b,b]$ under the optimal density, so its maximal value is minimax risk. The dashed curve is the asymptotic risk a fixed $\theta \in [-b,b]$ under a uniform distribution over $[-b,b]$, so its maximal value is given by (21).

illustrated in Fig. 7 as a function of the support size b. Also illustrated in these Figures is the MSE obtained when the threshold values are uniformly distribution over $\Theta = [-b,b]$, i.e., $\lambda(dt) = dt/(2b)$. In this case we have

$$\begin{aligned} K_{\mathsf{unif}} &\triangleq \min_{\theta\in[-b,b]} \frac{1}{2b}\int_{-b}^{b} \eta(t-\theta)\,dt \\ &= \frac{1}{2b}\int_{-b}^{b} \eta(t\pm b)\,dt = \frac{1}{2b}\int_{0}^{2b} \eta(t)dt = \frac{1}{4b}\int_{-2b}^{2b} \eta(t)dt, \end{aligned} \tag{21}$$

so the ARE under a uniform distribution equals $\sigma^2 K_{\mathsf{unif}}$.

C. Asymptotic Bayes Risk

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:

$$\begin{aligned} \text{minimize} \quad & R_{\pi,\lambda} = \int \frac{\pi(d\theta)}{\int \eta(t-\theta)\lambda(dt)}. \\ \text{subject to} \quad & \lambda(dt) \geq 0, \quad \int \lambda(dt) = 1. \end{aligned} \tag{22}$$

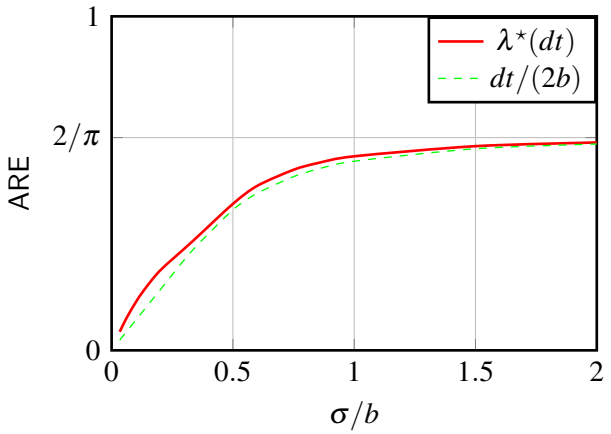

Fig. 7: Minimax ARE versus σ/b for $P_X = \mathcal{N}(\theta, \sigma^2)$. The dashed curve represents the minimal ARE under a uniform threshold density given by $\sigma^2/K_{\mathsf{unif}}$, where K_{unif} is given by (21).

.

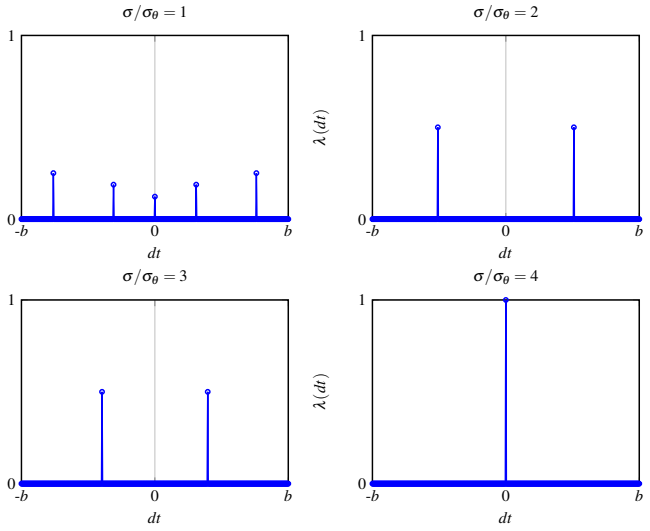

Fig. 8: The threshold density $\lambda(dt)$ that minimizes the asymptotic Bayes risk (22) for a uniform prior with $\sigma/\sigma_\theta = 1,2,3,4$, where $\sigma_\theta^2 = b^2/3$ is the variance of the prior.

We denote by $R_\pi^\star$ the minimal value of the objective function in (22). Since the function $x \to 1/x$ is convex for positive values, (22) 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 in a similar way that (20) is obtained from (19).

For the case of a normal distribution ($P_X = \mathcal{N}(\theta, \sigma^2)$)

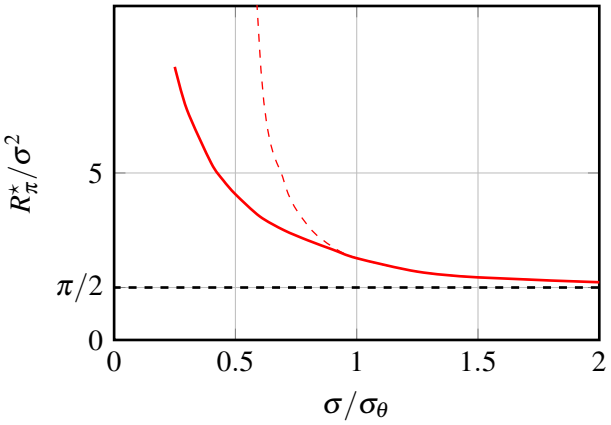

Fig. 9: Asymptotic Bayes risk $R_\pi^\star$ in estimating the mean of a normal distribution ($P_X = \mathcal{N}(\theta, \sigma^2)$under an optimal threshold distribution $\lambda^\star$ for π the uniform distribution over $\Theta = [-0.5, 0.5]$. The distribution $\lambda^\star$ is the minimizer of (22). It is illustrated for various cases in Fig. 8. The dashed curve represents the upper bound (23).

and a uniform prior ($\pi(d\theta) = dt/(2b)$), the optimal asymptotic threshold density obtained as the solution to (22) is illustrated in Fig. 8, whereas Fig. 9 illustrates the corresponding Bayes risk.

As can be seen in Fig. 8 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 (2). 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. (22).

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

$$R_\pi^\star \le \int \frac{\pi(d\theta)}{\eta(\theta_0 - \theta)}, \tag{23}$$

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

$$R_\pi^\star \le \frac{1}{4f^2(0)} + \left(\frac{1}{4f^2(0)}\frac{-f''(0)}{f(0)} - 1\right)\sigma_\theta^2 + O(\sigma_\theta^3). \tag{24}$$

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

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

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

$$\mathbb{E}\frac{1}{K(\theta)} \leq \int\int \frac{\pi(d\theta)\lambda(dt)}{\eta(t-\theta)}. \tag{25}$$

The bound (23) is obtained by taking λ to be a mass distribution at any θ_0 in the support of $\eta(x)$. Finally, (24) 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 (23) among all θ_0 in the support of $\eta(x)$.

The bound (23) is not trivial as long as the integral in the RHS of (23) is finite, i.e., whenever the tail of $\pi(\theta)$ vanishes fast enough compared to $\eta(x)$. The expansion (24) 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 $K^{-1}(\theta)$ approaches $1/\eta(0) = 1/4f^2(0)$.

Example 1: Consider the case $P_X = \mathcal{N}(\theta, \sigma^2)$. The bound in (24) 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 (23) in this case, as well as the Bayes ARE with the optimal threshold density $\lambda^\star$ for a uniform π, are illustrated in Fig. 9.

VI. Conclusions

We considered the MSE risk in estimating the mean of a normal distribution from a sequence of bits obtained by encoding each sample from the distribution using a single bit. Under an adaptive encoding of the samples, we showed that the no estimator can attain asymptotic MSE or relative efficiency larger than that of the median of the samples. We also showed that this lower bound is tight by presenting an adaptive estimation procedure from this bit sequence that is as efficient as the median. We also consider a one-step optimal scheme that minimizes the MSE given any set of previously obtained bits. Numerical simulations shows that the one-step optimal scheme attains the optimal relative efficiency for the normal distribution.

In the distributed setting, we considered the estimation from threshold detection under the assumption that the empirical distribution of the threshold values converges to a probability distribution. For this setting, we showed that the maximum likelihood estimator is local asymptotically minimax, and that the relative efficiency of this estimator is strictly smaller than that of the sample median.

Appendix

This appendix contains the proofs of Proposition 1, Theorems 3, 4, 5, and 8.

A. Proof of Proposition 1

(i) Since $f(x)$ is log-concave, so does $F(x)$. There exists $h(x)$ that is non-decreasing, concave, twice differntiable, such that $F(x) = e^{h(x)}$. In addition, $f(x) = h'(x)F(x)$ so $h'(x) \geq 0$, and $h'(x)$ is decreasing since $h''(x) \leq 0$ because $h(x)$ is concave. In fact, $h'(x) = f(x)/F(x)$ is non-constant for any x in the support of $f(x)$, so $h'(x)$ is strictly decreasing in this support. It easily verified that

$$\eta(x) = h'(x)h'(-x).$$

It follows that $\eta(x)$ is symmetric. We now show that $\eta(x)$ is decreasing for $x > 0$. Let $0 < x_1 < x_2$. Consider

$$\begin{aligned}\eta(x_2) - \eta(x_1) &= h'(x_2)h'(-x_2) - h'(x_1)h'(-x_1)\\ &= h'(x_2)h'(-x_1)\left(\frac{h'(-x_2)}{h'(-x_1)} - \frac{h'(x_1)}{h'(x_2)}\right).\end{aligned} \tag{26}$$

Now, because $h'(x)$ is decreasing, we have that $h'(x_1) \geq h'(x_2)$ and $h'(-x_2) \geq h'(-x_1)$. If follows that the brackets in the RHS of (26) are negative, whereas $h'(x_2)h'(-x_1)$ is non-negative since $h'(x)$ is non-negative. Hence $\eta(x_2) \leq \eta(x_1)$. It follows that $\eta(x)$ is a monotone decreasing function for $x > 0$. It is left to show that it is strictly decreasing for $x > 0$ in the support of $f(x)$. Since $h'(x)$ is strictly decreasing, $\eta(x)$ may be constant only for $x > 0$ for which $h'(x)/h'(-x)$ is a constant. But for all such xs, $h'(x)$ is strictly increasing, $h'(-x)$ is strictly decreasing, and thus $h'(x)/h'(-x)$ is strictly increasing.

(ii) Because $F(x)$ is differentiable, for any $a \in \mathbb{R}$ there exists $a \leq c \leq x$ such that $F(x) = F(a) + f(c)(x-a)$. For $a = 0$, and using the fact that $f(x)$ is non-increasing for $x > 0$ we have

$$\frac{f(x)}{F(x)} = \frac{f(x)}{0.5 + f(c)x} \leq \frac{f(x)}{f(c)}\frac{1}{x} \leq \frac{1}{x}.$$

From symmetry of $f(x)$ we get that $f(-x)/F(-x) \leq 1/x$ for $x > 0$. Finally, noting that $h(x) \to 0$ as $x \to \infty$, we conclude that for $x > 0$,

$$\eta(x) = h'(-x)h'(x) \leq \frac{h(x)}{x} \leq \frac{1}{x}.$$

Proof of Theorem 3

We first prove the following two lemmas:

Lemma 10: Let $f(x)$ be a log-concave, symmetric, and differentiable pdf. Denote by $F(x)$ the corresponding CDF. Then for any $x_1 \geq \ldots \geq x_n \in \mathbb{R}$, we have

$$\frac{\left(\sum_{k=1}^{n}(-1)^{k+1}f(x_k)\right)^2}{\left(\sum_{k=1}^{n}(-1)^{k+1}F(x_k)\right)\left(1-\sum_{k=1}^{n}(-1)^{k+1}F(x_k)\right)} \leq 4f^2(0). \tag{27}$$

Lemma 11: Let X be a random variable with a symmetric, log-concave, and continuously differentiable density function $f(x)$. Assume that for a Borel measurable A,

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

Then the Fisher information of M with respect to θ is bounded from above by $4f^2(0)$.

Proof of Lemma 10: We now use induction on $n \in \mathbb{N}$ to show that the LHS of (27) is bounded from above by $\eta(0)$.

The case $n = 1$ it follows from Proposition 1. Assume now that (27) holds for all integers up to some $n = N-1$ and consider the case $n = N$. The maximal value of the LHS of (27) is attained for the same $(x_1,\ldots,x_N) \in \mathbb{R}^N$ that attains the maximal value of

$$g(x_1,\ldots,x_N) \triangleq 2\log\left(\sum_{k=1}^{N}(-1)^{k+1}f(x_k)\right) - \log\left(\sum_{k=1}^{N}(-1)^{k+1}F(x_k)\right) - \log\left(1-\sum_{k=1}^{N}(-1)^{k+1}F(x_k)\right)$$
$$= 2\log\delta_N - \log\Delta_N - \log(1-\Delta_N),$$

where we denoted $\delta_N \triangleq \sum_{k=1}^{N}(-1)^{k+1}f(x_k)$ and $\Delta_N = \sum_{k=1}^{N}(-1)^{k+1}F(x_k)$. The derivative of $g(x_1,\ldots,x_N)$ with respect to x_k is given by

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

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

$$\frac{f'(x_k)}{f(x_k)} = \frac{\delta_N}{2}\left(\frac{1}{\Delta_N} - \frac{1}{1-\Delta_N}\right), \quad k = 1,\ldots,N. \tag{28}$$

From the case $n = 1$ and for x in the support of $f(x)$, we have

$$\frac{f'(x)}{f(x)} = h''(x) + \left(h'(x)\right)^2 < 0.$$

As a result, $\frac{f'(x)}{f(x)}$ is one to one and (28) implies $x_1 = \ldots = x_N$. If N is odd then for $x_1 = \ldots = x_N$ we have that the LHS of (27) equals $\eta(x_1)$ which was shown to be smaller than $\eta(0)$. If N is even, then for any constant c the limit of the LHS of (27) as $(x_1,\ldots,x_N) \to (c,\ldots,c)$ exists and equals zero. Therefore, the maximum of the LHS of (27) is not attained at the line $x_1 = \ldots = x_N$). We now consider the possibility that the LHS of (27) is maximized at the borders, as one or more of the coordinates of $(x_1,\ldots,x_N)$ approaches plus or minus infinity. For simplicity we only consider the cases where x_N goes to minus infinity or x_1 goes to plus infinity (the general case where the first m coordinates goes to infinity or the last m to minus infinity is obtained using similar arguments). Assume first $x_N \to -\infty$. Then the LHS of (27) equals

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

which is smaller than $\eta(0)$ by the induction hypothesis. Assume now that $x_1 \to \infty$. Then the LHS of (27) equals

$$\frac{\left(\sum_{k=2}^{N}(-1)^{k+1}f(x_k)\right)^2}{\left(1+\sum_{k=2}^{N}(-1)^{k+1}F(x_k)\right)\left(1-1-\sum_{k=2}^{N}(-1)^{k+1}F(x_k)\right)} = \frac{\left(-\sum_{m=1}^{N}(-1)^{m+1}f(x'_m)\right)^2}{\left(1-\sum_{m=1}^{N-1}(-1)^{m+1}F(x'_m)\right)\left(\sum_{m=1}^{N-1}(-1)^{m+1}F(x'_m)\right)},$$

where $x'_m = x_{m+1}$. The last expression is also smaller than $\eta(0)$ by the induction hypothesis. □

Proof of Lemma 11: The Fisher information of M with respect to θ is given by

$$\begin{aligned} I_\theta &= \mathbb{E}\left[\left(\frac{d}{d\theta}\log P(M|\theta)\right)^2|\theta\right] \\ &= \frac{\left(\frac{d}{d\theta}P(M=1|\theta)\right)^2}{P(M=1|\theta)} + \frac{\left(\frac{d}{d\theta}P(M=-1|\theta)\right)^2}{P(M=-1|\theta)} \\ &= \frac{\left(\frac{d}{d\theta}\int_A f(x-\theta)\,dx\right)^2}{P(M=1|\theta)} + \frac{\left(\frac{d}{d\theta}\int_A f(x-\theta)\,dx\right)^2}{P(M=-1|\theta)} \\ &\overset{(a)}{=} \frac{\left(-\int_A f'(x-\theta)\,dx\right)^2}{P(M=1|\theta)} + \frac{\left(-\int_A f'(x-\theta)\,dx\right)^2}{P(M=-1|\theta)} \\ &= \frac{\left(\int_A f'(x-\theta)\,dx\right)^2}{P(M=1|\theta)\left(1-P(M=1|\theta)\right)}, \\ &= \frac{\left(\int_A f'(x-\theta)\,dx\right)\left(\int_A f'(x-\theta)\,dx\right)}{\left(\int_A f(x-\theta)\,dx\right)\left(1-\int_A f(x-\theta)\,dx\right)}, \end{aligned} \tag{29}$$

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 (29) 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 (29) can be rewritten as

$$\begin{aligned} &= \frac{\left(\sum_{j=1}^{k} f\left(a_j - \theta\right) - f\left(b_j - \theta\right)\right)^2}{\left(\sum_{j=1}^{k} F\left(b_j - \theta\right) - F\left(a_j - \theta\right)\right)} \\ &\times \frac{1}{1 - \left(\sum_{j=1}^{k} F\left(b_j - \theta\right) - F\left(a_j - \theta\right)\right)} \end{aligned} \tag{30}$$

It follows from Lemma 10 that for any $\theta \in \mathbb{R}$ and any choice of the intervals endpoints, (30) is smaller than $4f^2(0)$. □

We now consider the proof of Theorem 3. In order to bound from above the Fisher information of any set of n single-bit messages with respect to θ, we first note that without loss of generality, each message M_i can is of the form

$$M_i = \begin{cases} X_i \in A_i & 1, \\ X_i \notin A_i & -1, \end{cases} \tag{31}$$

where $A_i \subset \mathbb{R}$ is a Borel measurable set. Indeed, any measurable function $M(X_i) \in \{-1, 1\}$ can be written in the form (31) with $A_i = M^{-1}(1)$. Consider the conditional distribution $P(M^n|\theta)$ of M^n given θ. We have

$$P\left(M^n|\theta\right) = \prod_{i=1}^{n} P\left(M_i|\theta, M^{i-1}\right), \tag{32}$$

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

$$I_\theta(M^n) = \sum_{i=1}^{n} I_\theta(M_i|M^{i-1}), \tag{33}$$

where $I_\theta(M_i|M^{i-1})$ is the Fisher information of the distribution of M_i given M^{i-1}. From Lemma 11 it follows that $I_\theta(M_i|M^{i-1}) \leq 4f^2(0)$. The Van Trees inequality [34], [3] 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 4

The algorithm given in (7) and (8) is a special case of a more general class of estimation procedures given in [29] and [30]. Specifically, (i) in Theorem 4 follows from the following simplified version of [29, Thm. 4]:

Theorem 12: [29, 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 $|\varphi(x)| \leq K_1(1 + |x|)$ for all $x \in \mathbb{R}$.

(ii) The sequence $\{\gamma_i\}_{i=1}^{\infty}$ satisfies conditions (9).

(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 $0 < \lambda \leq 1$ such that

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

(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 Thm. 12 are fulfilled for any $\lambda > 0$. Since

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

Thm. 4-(i) follows from Thm. 12.

In order to prove Thm. 4-(ii), we consider:

Theorem 13: [30, Thm. 2] Let

$$\begin{cases} U_n = U_{n-1} - \gamma_n \varphi(Y_n), & Y_n = g'(U_{n-1}) + Z_n \\ \bar{U}_n = \frac{1}{n}\sum_{i=1}^{n} U_n, & n = 1,2,\ldots. \end{cases} \tag{35}$$

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 Thm. 12-(iii) and satisfies the conditions there. Assume in addition that $\chi(0) > 0$, condition (34) with $\lambda = 1$, and there exits K_3 such that

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

Finally, assume that the sequence $\{\gamma_n\}$ satisfies conditions (9) and (10). 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 Thm. 13 with $g(x) = 0.5(x-\theta)^2$, $\varphi(x) = -\mathrm{sgn}(-x)$, $Z_n = \theta - X_n$ and $U_n = \theta_n$. From (35) 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 (8) and (7). Note that

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

for any $K_3 \ge 1$, the Fisher information of Z_1 is σ^2, $\chi(x) = 1 > 0$, and that the conditions in Thm. 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 $g''(x) = 1$. Thm. 13 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 single-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 [35] 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 (14) 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

$$1-\frac{f(\tau)\int_{-\infty}^{\tau}f(x)(\tau-x)dx}{\left(\int_{-\infty}^{\tau}f(x)dx\right)^2}. \tag{36}$$

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 (36) 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

We will prove that $P_\theta(m_i)=\mathbb{P}(M_i=m_i)$, $\imath=1,\dots,n$ is a local asymptotic normal (LAN) family with *precision* parameter $K(\theta)$. The statements in the theorem then follows from the local asymptotic minimax theorem of LAN families [32].

The probability mass distribution of M^n is given by

$$P_\theta(m^n)=\prod_{i=1}^{n}F\left(m_i(t_i-\theta)\right),\quad m^n\in\{-1,1\}^n.$$

Consider the log-likelihood ratio under a sequence of local alternatives $\theta'=\theta+h/\sqrt{n}$ for some $h\in\mathbb{R}$:

$$\log\frac{P_{\theta+\frac{h}{\sqrt{n}}}(m^n)}{P_\theta(m^n)}=\sum_{i=1}^{n}\log\left(F\left(m_i(t_i-\theta-h/\sqrt{n})\right)\right)-\sum_{i=1}^{n}\log\left(F\left(m_i(t_i\right.\right. \tag{37}$$

Since $f(x)$ is differentiable, $\log F(x)$ is twice differentiable and we may write

$$\log F(x+t)=\log F(x)+\frac{f(x)}{F(x)}h-\frac{t^2}{2}\left(\frac{f'(x)}{F(x)}-\frac{f^2(x)}{F^2(x)}\right)+o(t^2).$$

Therefore, (37) can be written as

$$-h\sum_{i=1}^{n}\frac{m_i}{\sqrt{n}}\frac{f\left(m_i(t_i-\theta)\right)}{F\left(m_i(t_i-\theta)\right)}-\frac{h^2}{2n}\sum_{i=1}^{n}\left(\frac{f'\left(m_i(t_i-\theta)\right)}{F\left(m_i(t_i-\theta)\right)}-\frac{f^2\left(m_i(t_i-\theta)\right)}{F^2\left(m_i(t_i-\theta)\right)}\right.$$

The proof is completed by proving the following two lemmas:

Lemma 14: For $i=1,\dots,n$ denote

$$U_i=-M_i\frac{f\left(M_i(t_i-\theta)\right)}{F\left(M_i(t_i-\theta)\right)}.$$

Then

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

Lemma 15: For $i=1,\dots,n$ denote

$$V_i=-\left[\frac{f'\left(M_i(t_i-\theta)\right)}{F\left(M_i(t_i-\theta)\right)}-\frac{f^2\left(M_i(t_i-\theta)\right)}{F^2\left(M_i(t_i-\theta)\right)}\right].$$

Then

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

Proof of Lemma 14: First note that

$$\begin{aligned}\mathbb{E}[U_i]&=\mathbb{E}\left[\mathbb{E}[U_i|M_i]\right]\\&=-\frac{f(t_i-\theta)}{F(t_i-\theta)}\mathbb{P}(M_i=1)+\frac{f(\theta-t_i)}{F(\theta-t_i)}\mathbb{P}(M_i=-1)\\&=-f(t_i-\theta)+f(\theta-t_i)=0.\end{aligned}$$

In addition,

$$\mathbb{E}U_i^2=\frac{f^2\left(t_i-\theta\right)}{F\left(t_i-\theta\right)\left(1-F\left(t_i-\theta\right)\right)},$$

and therefore

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}U_i^2\to K(\theta).$$

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

$$\mathbb{E}\left|U_i\right|^{2+\delta} = \left(\frac{f^{2+\delta}\left(t_i-\theta\right)}{F^{2+\delta}\left(t_i-\theta\right)} + \frac{f^{2+\delta}\left(t_i-\theta\right)}{1-F^{2+\delta}\left(t_i-\theta\right)}\right),$$

and

$$\frac{\sum_{i=1}^n \mathbb{E}\left|U_i\right|^{2+\delta}}{\left(\sum_{i=1}^n \mathbb{E}U_i^2\right)^{\delta}} = \frac{\frac{1}{n^{1+\delta}}\sum_{i=1}^n \mathbb{E}U_i^{2+\delta}}{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}U_i^2\right)^{\delta}}. \tag{38}$$

The functions $\log F(x)$ and $\log(1-F(x)) = \log F(-x)$ are differentiable, so that $f(x)/F(x)$, $f(x)/(1-F(x))$, and hence $\left(f(x)/F(x)\right)^{2+\delta}$ and $\left(f(x)/(1-F(x))\right)^{2+\delta}$ are integrable. As n goes to infinity, we have

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left|U_i\right|^{2+\delta} \to \int \left(\frac{f^{2+\delta}\left(t-\theta\right)}{F^{2+\delta}\left(t-\theta\right)} + \frac{f^{2+\delta}\left(t-\theta\right)}{1-F^{2+\delta}\left(t-\theta\right)}\right)\lambda(dt),$$

from which we conclude that the numerator in (38) goes to zero. Since the denumerator in (38) goes to $K^{\delta}(\theta)$, the entire expression goes to zero and hence Lyaponov's condition is satisfied. From Lyaponov's central limit theorem we conclude that

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

□

Proof of Lemma 15: We have:

$$\begin{aligned}\mathbb{E}V_i &= \left[\frac{f^2\left(t_i-\theta\right)}{F\left(t_i-\theta\right)} - f'\left(t_i-\theta\right)\right] + \left[\frac{f^2\left(\theta-t_i\right)}{F\left(\theta-t_i\right)} - f'\left(\theta-t_i\right)\right] \\ &= \frac{f^2(\theta-t_i)}{F(\theta-t_i)F(t_i-\theta)} = \eta(\theta-t_i).\end{aligned}$$

where above we used the facts $f'(-x) = -f'(x)$ and $F(x)+F(-x) = 1$. We conclude that:

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}V_i = \frac{1}{n}\sum_{i=1}^n \frac{f^2\left(t_i-\theta\right)}{F\left(t_i-\theta\right)\left(1-F\left(t_i-\theta\right)\right)} \to K(\theta).$$

Finally, it follows from Kolmogorov's law of large numbers (e.g. [36, Thm. 10.2.3]) that

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

□

REFERENCES

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

[2] R. Gray and D. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct 1998.

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

[4] 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.

[5] 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.

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

[7] 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 Trans. Inf. Theory*, vol. 47, no. 1, pp. 446–450, Jan 2001.

[8] 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.

[9] 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.

[10] 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.

[11] ——, "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.

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

[13] 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.

[14] 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.

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

[16] H. Viswanathan and T. Berger, "The quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1549–1559, 1997.

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

[18] 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.
[19] 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.
[20] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, 1987.
[21] Z. Zhang and T. Berger, "Estimation via compressed information," *IEEE Trans. Inf. Theory*, vol. 34, no. 2, pp. 198–211, 1988.
[22] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Trans. Commun.*, vol. 57, no. 12, 2009.
[23] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.
[24] 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.
[25] J. Hammersley, "On estimating restricted parameters," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 12, no. 2, pp. 192–240, 1950.
[26] 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.
[27] 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
[28] A. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. Springer New York, 2008.
[29] 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.
[30] B. T. Polyak, "New stochastic approximation type procedures," *Automat. i Telemekh*, vol. 7, no. 98-107, p. 2, 1990.
[31] P. Venkitasubramaniam, G. Mergen, L. Tong, and A. Swami, "Quantization for distributed estimation in large scale sensor networks," in *2005 3rd International Conference on Intelligent Sensing and Information Processing*, Dec 2005, pp. 121–127.
[32] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.
[33] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.
[34] H. L. Van Trees, *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004.
[35] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar 1982.
[36] P. K. Sen and J. M. Singer, *Large sample methods in statistics: an introduction with applications*. CRC Press, 1994, vol. 25.
[37] T. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2300–2324, Oct 1998.
[38] I. Csiszár, "The method of types [information theory]," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

The estimation problem we consider can be seen as the quantization, compression or encoding of the samples, and the decoding of the parameter θ from the encoded version of the sample. Therefore, this problem is closely related to various setting in coding and information theory. The goal of this section is to illustrates this connection and review related results that are relevant in our case.

B. Statistical Inference under Communication Constraints

The works [20], [21], [37] consider various problems of statistical inference under multiterminal lossy compression. In the setting of [20], [21], [37], each terminal I observes n samples from the distribution and is allotted nR_i bits to communicate its estimate.The main focus of these works is the difference between inference with communication constraint and the unconstrained vanilla statistical estimation setting, as the number of samples n goes to infinity subject to a total finite rate constraint. Our case (i) can be seen as a special case of this setting where with a single terminal and $R_1 = 1$. However, as explained in [37, Sec. III], in this setting the unconstrained inference performance is always attained when all samples are taken from the same distribution. Indeed, in an i.i.d setting the *type* of the sample [38] is a sufficient statistics for any estimation task, and the latter can be described using a number of codewords polynomial in n regardless of the distribution of the samples. For this reason, attention is given in these works to inference problems involving multiple distributions observed at different locations, hence the name *multiterminal*.

C. Sigma-Delta Encoding

and covariance function [1, Eq. 25]

$$R(k) = \mathbb{E} M_{n+k} M_n = \theta^2 + O\left(\frac{1}{\sigma\sqrt{k}}\right).$$

This is not enough to deduce convergence. Need to go over the paper and check if rate of convergence can be deduced. Also do simulations BEFORE that.

What is the error in estimating the mean of stationary ergodic process ? Since the SDM is a special case of the sequential scheme, we conclude that the MSE of any SDM with a noisy DC signal is bounded from below by $n^{-1}\sigma^2\pi/2$.

Source coding

With a full access to the sample as in setting (i), the problem of encoding and estimating θ is reduced to the MSE attained by a scalar quantizer adjusted to the sufficient statistics of the sample. Setting (ii) includes as a special case the sigma-delta modulation (SDM) analog-to-digital conversion scheme with a

constant input θ corrupted by Gaussian noise Z_i, as was considered in [1]. While it was shown there that the output of the modulator converges to the true constant input, the rate of this convergence was not analyzed and cannot be derived directly from the results of [1]. As a corollary from the results in this paper we conclude that the rate of convergence of a SDM to a constant input signal is at most $\sigma^2\pi/2$ over the number of feedback iterations. Finally, the remote multiterminal source coding setting of [15] corresponds to the case of n rate-constrained encoders, each observing a noisy version of an information source. The difference between this setting and ours is that in ours the parameter of interest is not an information source. By assuming a prior distribution on this parameter, the CEO provides a lower bound on the estimation error in the fully distributed setting (iii). This lower bound can be attained if we were to consider the average error in multiple independent realizations of our problem rather than a single realization as we do here.