Mean Estimation from One-Bit Measurements

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

Abstract

We consider the problem of estimating the mean of a symmetric log-concave distribution under constraint that only a single bit per sample from this distribution is available to the estimator. We study the mean squared error as a function of the sample size (and hence number of bits). We consider three settings: first, a centralized setting, where an encoder may release n bits given a sample of size n, and for which there is no asymptotic penalty for quantization; second, an adaptive setting in which each bit is a function of the current observation and previously recorded bits, where we show that the optimal relative efficiency compared to the sample mean is precisely the efficiency of the median; lastly, we show that in a distributed setting where each bit is only a function of a local sample, no estimator can achieve optimal efficiency uniformly over the parameter space. We additionally complement our results in the adaptive setting by showing that *one* round of adaptivity is sufficient to achieve optimal mean-square error.

I. Introduction

We consider estimation of parameters from data collected by multiple units under communication constraints between the units. Such scenarios arise in sensor arrays, where sensor motes collect information, which they transmit to a central estimation unit [1], [2]. More generally, communication is substantially more expensive than computation in modern computing infrastructure [3]. It is thus of interest to understand the extent to which communication constraints induce fundamental accuracy and efficiency limits in parametric estimation problems.

We answer this question in a sylized version of this problem: the estimation of the mean θ of a symmetric log-concave distribution under the constraint that only a single bit can be communicated about each observation from this distribution. Different information sharing schemes strongly affect the performance of estimators for θ; we illustrate the three main settings we consider in Figure 1.

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

The distributed setting (iii) is the most restrictive; as it turns out, (ii) is slightly more restrictive than the fully centralized setting (i), and in our setting, a variant of the adaptive setting (ii) in which there is only *one* round of

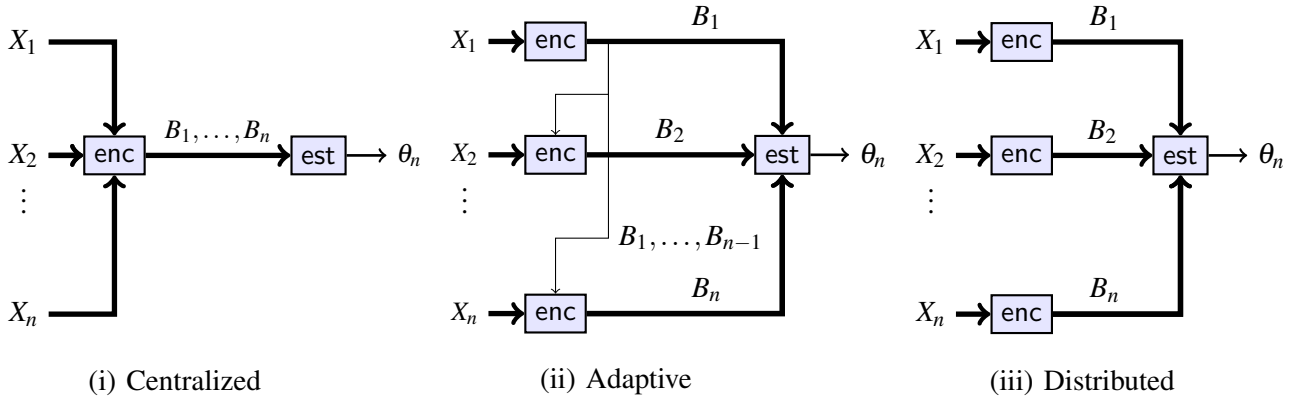

Fig. 1. Three encoding settings: (i) Centralized – an encoder sends n bits after observing n samples. (ii) Adaptive (sequential) – the ith encoder sends the bit B_i depending on its private sample X_i and previous bits $B_1,\ldots,B_{i-1}$. (iii) Distributed – each encoder send the bit B_i based on its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

adaptivity—as we make formal later—is enough to achieve the same efficiency as the fully sequential setting (ii). Each setting has natural applications:

- **Signal acquisition (i):** A quantity is measured n times at different instances. The results are averaged in order to reduce measurement noise and the averaged result is then stored or communicated using n bits.
- **Analog-to-digital conversion (ii):** A sigma-delta modulator (SDM) converts an analog signal into a sequence of bits by sampling the signal at a very high rate and then using one-bit threshold detector combined with a feedback loop to update an accumulated error state [4]. Therefore, the expected error in tracking an analog signal using an SDM falls under our setting (ii) when we assume that the signal at the input to the modulator is a constant (direct current) corrupted by, say, thermal noise [5]. Since the sampling rates in SDM are usually many times more than the bandwidth of its input, analyzing SDM under a constant input provides meaningful lower bound even for non-constant signals.
- **Privacy (ii)–(iii):** 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 [6].

Let us provide an informal description of our results and setting. For an estimator θ_n with finite quadratic risk (mean squared error (MSE)) $R_n = \mathbb{E}_\theta[(\theta_n - \theta)^2]$, we study the limit

$$\limsup_{n\to\infty} nR_n. \tag{1}$$

By comparing this quantity to achievable rates of convergence without communication constraints, we can evaluate the efficiency losses—asymptotic relative efficiency—of the estimator to appropriately optimal (unconstrained) estimators. (We shall be more formal in the sequel.) By lower bounding the quantity (1), we also provide limits on estimation of single-bit-per-measurement constrained signals in more general settings [7], [8], [9], [10], [11].

In setting (i), the estimator can evaluate any optimal estimator of location (e.g., the sample mean if the data is Gaussian), then quantize it using n bits. As the accuracy in describing the empirical mean decreases exponentially in n, the quantization error is negligible compared to the statistical error in mean estimation [12]. That is, centralized encoding induces no asymptotic efficiency loss. The story is different in settings (ii) and (iii). Precisely, we show that in the adaptive setting (ii), the optimal efficiency of a one-bit scheme is (asypmtotically) precisely that of the sample median, and that this efficiency is achievable. As a concrete example, when X_i are i.i.d. Gaussian, we necessarily lose a factor of $\pi/2 \approx 1.57$ in the asymptotic risk; the one-bit constraint decreases the effective sample size by a factor of $\pi/2$ compared to estimating it without the bit constraint. It turns out that, in the settings we consider, only a *single round* of adaptivity (see Fig. 3 for an illustration) is sufficient to achieve optimal convergence rates. In distinction from setting (ii), in setting (iii) when the messages must be independent, there is no distributed estimation scheme that achieves the efficiency of the sample median uniformly over θ. We establish this result via Le Cam's local asyptotic normality theory, allowing us to provide exact characterizations of the asymptotic efficiency of suitably regular encoding schemes.

Our asymptotic setting is important in that it allows us to elide difficulties present in finite sample settings. For example, in setting (i), developing an optimal quantizer at finite n requires choosing a 2^n level scalar quantizer, which is non-trivial [13]. In interactive and sequential settings (e.g. (ii)), the situation is more challenging, as it is unclear whether any type of compositionality applies, in that an $n-1$-step optimal estimator may be only vaguely related to the n-step optimal estimator. Thus, to provide our lower bounds, we rely on stronger information-based inequalities, including the Van Trees inequality [14] and Le Cam's local asymptotic normality theory [15], [16], [17].

Related Work

The many challenges of estimation under communication constraints have given rise to a large literature investigating different aspects of constrained estimation. While our setting—in which we observe a single bit per signal X_i—is restrictive, it inspires substantial work. Perhaps the most related is that of Wong and Gray [5], who study one-bit analog-to-digital conversion of a constant input corrupted by Gaussian noise using a Sigma-Delta Modulator (SDM). They show almost sure convergence, but provide no rate (and no rates follow from their analysis); in contrast, we provide an optimal procedure and matching lower bound achieving risk $\frac{\pi}{2}\sigma^2$ in the limit (1) when

$X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$. A growing literature on one-bit measurements in high-dimensional problems [7], [18], [19] shows how to reconstruct sparse signals, where Baraniuk et al. [7] show that in noiseless settings, exponential decay in MSE is possible; our results make precise the penalty for noise under one-bit sensing, showing that the error can decay (under Gaussian noise) at best as $\frac{\pi}{2}\frac{\sigma^2}{n}$.

In fully distributed settings (iii), the challenges are different, and there is also a substantial literature with one-bit (quantized) measurements [20], [21], [22], [23], [24]. We complement these results by providing precise lower bounds and optimality results; previous performance bounds are suboptimal. Work on the remote multiterminal source coding problem, or CEO problem [25], [26], [27], [28], provides lower bounds on the MSE in setting (iii); because of the somewhat distinct setting, these bounds are looser than ours (which have optimal constants). In settings more similar to our statistical estimation scenario—such as estimation of parameters in a multi-dimensional linear model—a line of work provides lower bounds on statistical estimation [29], [30], [31], [32], [33], [34], [35], [36]. These results are finite sample and apply more broadly than ours, but as a consequence, they have unusable constants, while our stylized model allows precise identification of exact constants. Work subsequent to the initial draft of this paper [37] uses an approach similar to ours—bounding quantized Fisher information—to derive lower bounds on the error in parametric estimation problems from quantized measurements in non-adaptive settings.

Testing (and discrete estimation) problems also enjoy a robust literature, though as a consequence of our results to come, the results for testing, i.e., when the parameter space Θ is finite, are quite different from those for estimation, as it is possible to construct optimal decision (testing) rules in a completely distributed fashion. In this context, Longo et al. [38] propose procedures for distributed testing based on optimizing a Bhattacharyya distance. Tsitsiklis [39] shows 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 with optimal exponent. Moreover, in a distributed setting, feedback is unnecessary for optimal testing/detection [40], in strong distinction to the estimation case we consider.

The remainder of this paper is organized as follows. In Section II we describe the problem, notation, and our basic assumptions. 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 and Notation

Let $f : \mathbb{R} \to \mathbb{R}_+$ be a symmetric and log-concave probability density, which necessarily has finite second moment σ^2, and let $\Theta \subset \mathbb{R}$ be closed and convex. For $\theta \in \mathbb{R}$, let P_θ be the probability distribution with density $f(x-\theta)$, so that θ indexes the location family $\{P_\theta\}_{\theta\in\Theta}$. The log-concavity and symmetry $f(x)$ imply that P_θ has a unique mean and median at θ [41]. We observe a sample $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P_\theta$, where θ is unknown, and wish to estimate θ given only binary messages $B_1, \ldots, B_n \in \{0,1\}$ about each X_i. We study this under three distinct computational scenarios, which we illustrate in Figure 1:

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

We also consider a hybrid of the fully distributed setting (where the bits B_i are independent) and the adaptive setting (where each bit B_i may depend on the previous bits) to a *one-step* adaptive setting, where the quantization scheme may be modified to depend on one fixed function of the previous information.

(ii') One-step adaptive, where for some function g and a (fixed) t, if $i \le t$ then $B_i = B_i(X_i)$ while if $i > t$, then $B_i = B_i(X_i, g(B_1, \ldots, B_t))$.

We measure the performance of an estimator $\theta_n \triangleq \theta_n(B_1^n)$ by one of a few notions. In the simplest case, we assume a prior π on θ (which may be a point mass) and consider the quadratic risk

$$R_n = R_n(\pi) \triangleq \int \mathbb{E}_\theta \left(\theta_n - \theta\right)^2 d\pi(\theta), \tag{2}$$

where the expectation is taken with respect to the distribution of $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P_\theta$. The main problems we consider in this paper are the minimal value of the risk (2) as a function of the sample size n and the density f, under different choices of the encoding functions in cases (i)–(iii). The quadratic risk (2) may be infinite in some cases;

we defer discussion of this case to later sections, as it is technically demanding and detracts from the presentation here.

Now, let $\sigma_f^2 \triangleq \mathbb{E}[\frac{f'(X)^2}{f(X)^2}]$ be the Fisher information for the location in the family $\{P_\theta\}$, which is finite when f is log-concave and symmetric. We give particular attention to the asymptotic relative efficiency (ARE) of estimators with respect to asymptotically normal efficient estimators achieving the information bound [17]. In this case, if $\{m(n), n \in \mathbb{N}\}$ is a sequence such that

$$\sqrt{m(n)}(\theta_n - \theta) \overset{d}{\rightsquigarrow} \mathcal{N}(0, \sigma_f^2),$$

then the ARE of the estimator is [42, Def. 6.6.6]

$$\mathsf{ARE}(\theta_n) \triangleq \liminf_{n\to\infty} \frac{m(n)}{n}. \tag{3}$$

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

$$m(n)R_n = m(n)\mathbb{E}_\theta\,(\theta_n - \theta)^2 = V + o(1),$$

the ARE of θ_n is σ_f^2/V, so that θ_n requires a sample V/σ_f^2-times larger than that of an efficient estimator for comparable accuracy to the (information) efficient estimator.

Notation and basic assumptions

To describe our results and make them formal, we require some additional notation and one main assumption, which restricts the class of distributions we consider. We use the typical notation that $F(x) = \int_{-\infty}^{x} f(t)dt$ is the cumulative distribution function of the X_i, and we let

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

be the *hazard* function (or the *failure rate* or *force of mortality*), which is monotone increasing as f is log-concave [43]. Given the centrality of the median to our efficiency bounds, it is unsurprising that the quantity

$$\eta(x) \triangleq \frac{f^2(x)}{F(x)(1-F(x))} \overset{(\star)}{=} \frac{f(x)f(-x)}{F(x)F(-x)} \tag{4}$$

appears throughout our development. (Here equality $(\star)$ is immediate by the symmetry of f.) For $p \in (0,1)$ and $x = F^{-1}(p)$,

$$\frac{1}{\eta(x)} = \frac{1}{\eta(F^{-1}(p))} = \frac{p(1-p)}{f(F^{-1}(p))^2} \tag{5}$$

is of course the familiar asymptotic variance of the pth quantile of the sample $X_1, \ldots, X^n$ (cf. [17], Ch. 21).

For f the normal density, classical results [44], [45] show that that $\eta(x)$ is a strictly decreasing function of $|x|$, as we illustrate in Fig. 2. We consider log-concave symmetric distributions sharing this property. Specifically, we require the following.

Assumption A1: The origin $x = 0$ uniquely maximizes $\eta(x)$, and $\eta(x)$ is non-increasing in $|x|$.

Under this assumption,

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

where $\eta(0) = 4f^2(0)$ is the asymptotic variance of the sample median (Eq. (5) at $p = 1/2$). Combined with log-concavity of $f(x)$, Assumption A1 implies that $\eta(x)$ vanishes as $|x| \to \infty$. Several distributions satisfy Assumption A1, including the generalized normal distributions with a shape parameter between 1 and 2 (including the normal and Laplace distributions). Symmetric log-concave distributions that failing Assumption A1 include the uniform distribution and the generalized normal distribution with shape parameter greater than 2. (See Appendix A for a brief discussion on the uniform distribution, which illustrates why some restriction of the class of distributions is necessary to develop our results; in brief, even one-step adaptive estimators with a single bit can achieve rates faster than the paramateric $\sqrt{n}$ convergence for location estimation in the uniform distribution.)

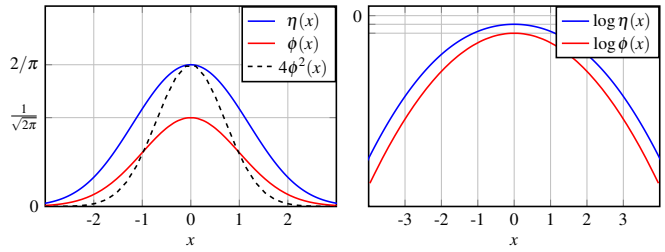

Fig. 2. The function $\eta(x) = f^2(x)/F(x)F(-x)$ for $f(x) = \phi(x)$ the standard normal density.

III. Consistent Estimation and Off-the-shelf Bounds

We begin our technical treatment by deriving a few bounds on the efficiency of estimators in setting (iii). These bounds establish the following facts:

1. A consistent estimator with an asymptotically normal distribution always exists in setting (iii), and hence in the adaptive settings (ii) and (ii').
2. For the normal distribution, the asymptotic relative efficiency (3) in the distributed setting (iii) is at most $3/4$. No estimator can be as efficient as the sample mean.

A. Consistent Estimation

The simplest estimator is simply to invert a quantile. Indeed, fix $\theta_0 \in \mathbb{R}$ and define the ith message by

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

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

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

so that

$$\theta_n = \theta_0 - F^{-1}(P_n) \tag{6}$$

is a consistent estimator for θ in the distributed setting of Figure 1-(iii), where we note that F is invertible over the support of f. As the variance of P_n is $F(\theta_0-\theta)(1-F(\theta_0-\theta))$, a delta method calculation [17, Ch. 23] implies that θ_n is asymptotically normal with variance

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

In the Gaussian case where the $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$, the ARE of θ_n is $\eta(\theta_0-\theta)\sigma^2$.

Assumption A1 implies that the optimal asymptotic variance for an estimator of the form (6) is $1/\eta(0)$, the asymptotic of the sample median. Unfortunately, as θ is (by definition) *a priori* unknown and $\eta(x)$ monotonically decreases in $|x|$, this naive estimator θ_n may be very inefficient when θ is far from the initial guess θ_0. As an example, when f is a the normal density, the ARE of θ_n is less than 0.15 when $|\theta_0 - \theta| \geq 2\sigma$, and more broadly, $\mathsf{ARE}(\theta_n)$ asymptotes to $|\theta_0|\exp(-\theta_0^2/2)/\sqrt{2\pi}$ as $|\theta_0-\theta|$ gets large. Yet that $\theta_0 = \theta$ minimizes this asymptotic variance, and η is continuous, is suggestive: if we can use a suitably good initial estimate θ_n^{init} for θ, it is possible that a one-step adaptive estimator (recall (ii')) may be asymptotically strong, as we see in Section IV.

B. Multiterminal Source Coding

A related problem is the CEO problem, which 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 [25]. The central estimator produces an estimates $\theta_1, \ldots, \theta_k$ with the goal of minimizing the quadratic risk:

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

Note that any distributed encoding scheme using one-bit per sample can be replicated k times and thus leads to a legitimate encoding and estimation scheme for the CEO problem with $r_1 = \ldots = r_n = 1$. It follows that, assuming that θ is drawn once from the prior $\pi(d\theta)$, our mean estimation problem from one-bit samples under distributed encoding corresponds to the CEO setting with $k = 1$ realization of θ observed under noise at n different locations, and communicated at each location using an encoder sending a single bit. Consequently, a lower bound on the MSE in estimating θ in the distributed encoding setting is given by the minimal MSE in the CEO setting as $k \to \infty$. Note that the difference between the CEO setting and ours lays in the privilege of each of the encoders to describe k realizations of θ using k bits with MSE averaged over these realizations, rather than a single realization using a single bit in ours.

When the prior on θ and the noise corrupting it at each location are Gaussian, Prabhakaran et al. [28] characterize the optimal encoding and its asymptotic risk as $k \to \infty$. Chen et al. [46] also provide an expression for the quadratic risk in the CEO setting under Gaussian priors. Adapting to our setting, this expression provides the following proposition:

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

$$n\mathbb{E}\left[(\theta - \theta_n)^2\right] \geq \frac{4}{3}\sigma^2 + O(n^{-1}), \tag{7}$$

where the expectation is with respect to θ and X^n.
See Appendix B for a proof.

As we shall see, this bound is loose: the difference between the MSE lower bound (7) and the actual MSE in the distributed setting (case (iii)) occurs because in the CEO setting, each encoder may encode an arbitrary number of k independent realizations of θ using k bits; in our situation, $k = 1$. That blocking allows more efficient encoding and exploiting the high-dimensional geometry of the product probability space in the CEO problem is perhaps unsurprising, and our goal in the sequel will be to characterize the performance degradation one bit encoding engenders.

IV. Adaptive Estimation

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

A. Limited efficiency in the adaptive setting

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

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

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

where

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

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

Proof: See Appendix C.

Theorem 2 implies that any estimator θ_n from any adaptive encoding scheme satisfies

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

and

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

It is possible to extend Theorem 2 to any other loss functions using a more subtle version of the Van Tress inequality provided in [47]. See also [48].

Next, we provide 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 θ_n for θ:

$$\theta_n = \theta_{n-1} + \gamma_n B_n, \quad n=1,2,\ldots, \tag{8}$$

where

$$B_n = B_n(X_n,\theta_{n-1}) = \operatorname{sgn}(X_n-\theta_{n-1}). \tag{9}$$

Define the nth step estimation as

$$\bar{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i. \tag{10}$$

We have the following results:

Theorem 3: Consider the sequence $\{\bar{\theta}_n\}_{n\in\mathbb{N}}$ defined by (10).

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

$$\begin{cases} \frac{\gamma_n-\gamma_{n+1}}{\gamma_n} = o(\gamma_n), \\ \sum_{n=1}^{\infty}\frac{\gamma_n^{(1+\lambda)/2}}{\sqrt{n}} < \infty, \quad \text{for some } 0<\lambda\leq 1 \end{cases} \tag{11}$$

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

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

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

$$\sigma_f^2 \triangleq \mathbb{E}\left[\frac{f'(X)}{f(X)}\right]$$

for the location in the family $\{P_\theta\}$. Then for any bounded, symmetric, and quasi-convex function L,

$$\liminf_{c\to\infty}\liminf_{n\to\infty}\sup_{\tau:|\theta-\tau|\leq\frac{c}{\sqrt{n}}}\mathbb{E}\left[L\left(\sqrt{n}(\bar{\theta}_n-\tau)\right)\right] = \mathbb{E}\left[L(Z/\sqrt{\eta(0)})\right], \tag{12}$$

where $Z\sim\mathcal{N}(0,1)$.

(iii) Assume that, in addition to (11), $\{\gamma_n\}_{n\in\mathbb{N}}$ 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 $\beta\in(2/3,1)$). Then

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

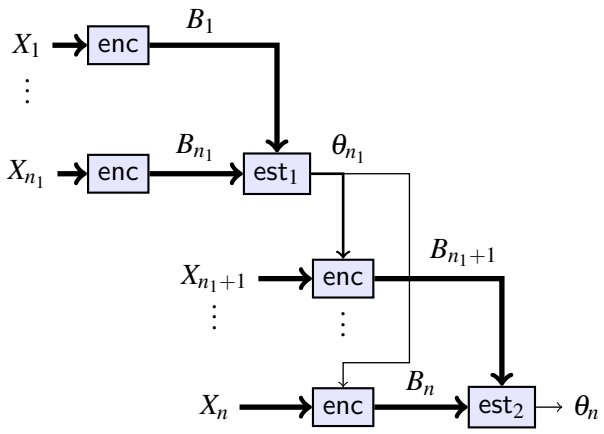

Fig. 3. Distributed encoding with a single interaction: The estimation obtained from the first n_1 bits in a distributed manner is to obtain another $n-n_1$ bits in a distributed manner.

Proof: See Appendix E.

Theorem 3 implies that the estimator θ_n defined by (10) and (8) attains the maximal ARE, as established by Theorem 2. The update step (8) 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 particular, parts (i) and (ii) of Theorem 3 provides conditions under which $\bar{\theta}$ is asymptotically normal and local asymptotically minimax, in the sense that it attains the local asymptotic minimax bound [49]. Part (iii) provides an additional condition under which $\bar{\theta}$ also converges in its second moment.

In the encoding and estimating procedure (8) and (10), each one-bit message B_n depends on its private sample as well as the current gradient descent estimate θ_{n-1}. In this sense, each encoder in this algorithm interacts with previous one by using the current estimate. As we explain next, it is possible to obtain the optimal efficiency of $\eta(0)\sigma^2$ with only one round of interactions among the encoders.

C. Maximal Efficiency using One Round of Threshold Adaptation

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

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

and deduced 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 attaining the lower variance bound $1/\eta(0)$, provided we allow to update the threshold value θ_0 based on previously observed bits at least once. 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 (6) to obtain an estimate θ_{n_1} based on $B_1, \ldots, B_{n_1}$, and then use θ_{n_1} as the new threshold value to obtain messages $B_{n_1+1}, \ldots, B_n$. Figure 3 illustrates a diagram of this procedure. The specific encoding and estimation scheme, as well as its asymptotic performance, are given by the following theorem:

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

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

where $n = n_1 + n_2$ and

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

Let

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

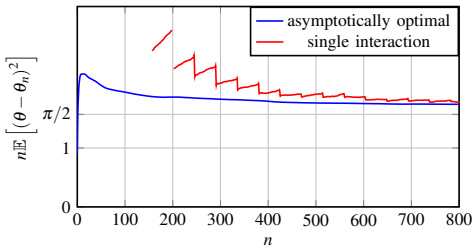

Fig. 4. Normalized empirical risk versus number of samples n for 10,000 Monte Carlo trials with $f(x)$ the standard normal density. In each trial, θ is chosen uniformly over the interval $(-1.64, 1.64)$. The single interaction strategy uses $n_1 = \lfloor \sqrt{n} \rfloor$ samples for its first stage.

and assume that, as $n \to \infty$, $n_1(n) \to \infty$ and $n_2(n)/n \to 1$. Then:

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

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

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

Since $n_2(n)/n \to 1$, if follows from the Berry Esseen theorem that there exists C, independent of n and t, such that

$$\sup_{t \in \mathbb{R}} \left| \sqrt{n} \frac{p_{n_2}(t) - F(\theta - t)}{F(\theta - t)F(t - \theta)} - \Phi(t) \right| \leq C/\sqrt{n}.$$

where $Z \sim \mathcal{N}(0,1)$. It follows that

$$\sqrt{n} \frac{p_{n_2}(Y_n) - F(0)}{F^2(0)} \xrightarrow{d} \mathcal{N}(0,1) \tag{14}$$

for any sequence of random variables $\{Y_n\}$ converging in probability to θ. In particular, since

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

by the law of large numbers, (14) holds for $Y_n = \theta_{n_1}$. Finally, we use the delta method with $g(x) = F^{-1}(x)$ to conclude

$$\begin{aligned}
&\sqrt{n}\,(\theta_{n_2} - \theta) = \sqrt{n}\left(\theta_{n_1} + F^{-1}(p_{n_2}(\theta_{n_1})) - \theta\right) \\
&= \sqrt{n}\left(g(p_{n_2}(\theta_{n_1})) - g\left(F(\theta - \theta_{n_1})\right)\right) \xrightarrow{d} \mathcal{N}\left(0, F^2(0)/f^2(0)\right).
\end{aligned}$$

□

V. Distributed Estimation

In this section, we now consider the distributed encoding setting in Figure 1-(iii) where each one-bit message B_i is only a function of its private sample X_i. In this case, the ith encoder is fully characterized ~~its *detection* region~~, defined as

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

Consequently, B_i is of the form

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

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

As a first step, the following theorem provides conditions under which the messages $B_1, B_2, \ldots$ define a local asymptotic normal family. The proof of this theorem is in Appendix G.

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

$$L_n(A_1,\ldots,A_n;\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \frac{\left(\frac{d}{d\theta}\mathbb{P}(X_i \in A_i)\right)^2}{\mathbb{P}(X_i \in A_i)\left(1-\mathbb{P}(X_i \in A_i)\right)}. \tag{15}$$

Consider the following conditions:

(i) The probability density function $f(x)$ of $X_n - \theta$ is log-concave, differentiable, and symmetric. Furthermore, there exists $\delta > 0$ such that the function $\eta^{1+\delta}(x)/f^{\delta}(x)$ is non-increasing in $|x|$.

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

(iii) The limit

$$\kappa(\theta) \triangleq \lim_{n\to\infty} L_n(A_1,\ldots,A_n;\theta) \tag{16}$$

exists.

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

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

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

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

Theorem 5 provides conditions under which $B_1,\ldots,B_n$ defines a LAN family with a precision parameter given by the limit in (16). Condition (i) is satisfied, for example, by the normal distribution with $\delta \le 2$. Condition (iii) is arguably the strongest and hardest to verify. As we show in Section V-B below, this condition is satisfied, for example, when $A_1,\ldots,A_n$ are half lines whose starting points are drawn independently from some probability measure on $\mathbb{R}$. Similar ideas imply that condition (iii) holds whenever we choose the intervals consisting each A_i according to some pre-specified distribution.

An important conclusion of Theorem 5 follows from the local asymptotic minimax property of estimators in LAN models [49]):

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

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

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

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

Corollary 6 says that when the sequence of one-bit messages defines a LAN model with respect to θ, no estimator can attain MSE smaller than $1/\kappa(\theta)n + O(1/n)$ where $\kappa(\theta)$ is the precision parameter of the model. Consequently, the ARE in such case is at most $\kappa(\theta)\sigma^2$. Furthermore, as we show next, no estimator in this setting can attain the optimal ARE of $\eta(0)\sigma^2$ uniformly over Θ.

A. Non-existence of a Uniformly Optimal Strategy

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

Theorem 7: Under conditions (i)-(iii) in Theorem 5, assume that each A_i is the 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 H. □

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

B. Threshold Detection

Assume now that each B_i is of the form

$$B_i = \operatorname{sgn}(t_i - X_i) = \begin{cases} 1 & X_i < t_i, \\ -1 & X_i > t_i, \end{cases} \tag{17}$$

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

$$L_n(A_1, \ldots, A_n; \theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{(f(t_i - \theta))^2}{F(t_i - \theta) F(\theta - t_i)} = \frac{1}{n} \sum_{i=1}^{n} \eta(t_i - \theta). \tag{18}$$

A natural condition for the existence of the limit (18) 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

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

Theorem 5 implies:

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

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

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

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

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

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 θ_n^{ML} is given by the root of

$$\sum_{i=1}^{n} B_i \frac{f(t_i - \theta)}{F(B_i(t_i - \theta))}, \tag{19}$$

This root is unique since the log-likelihood function is concave. Furthermore, for any $n \in \mathbb{R}$, we have that $\theta_n^{ML} \in [t_{(1)}, t_{(n)}]$ 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

$$\lim_{n \to \infty} n\mathbb{E}\left[\left(\theta_n^{ML} - \theta\right)^2\right] = 1/\kappa(\theta),$$

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

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

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

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

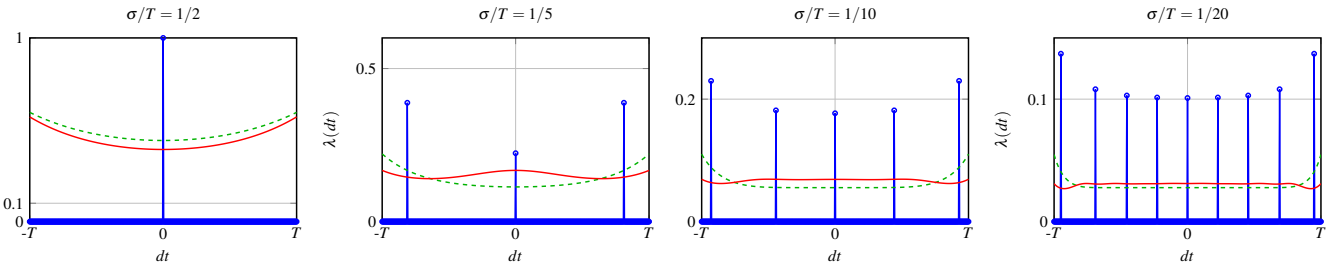

Fig. 5. Optimal threshold density $\lambda^\star(dt)$ (blue) that maximizes the ARE for $f(x) = \mathcal{N}(\theta, \sigma^2)$ and $\theta \in [-T, T]$. The continuous curve (red) represents the reciprocal of the asymptotic risk at a fixed $\theta \in [-T, T]$ under the optimal density, i.e., the minimax risk is the inverse of the minimal value of this curve. The dashed curve (green) is the reciprocal of the asymptotic risk at a fixed $\theta \in [-T, T]$ when the threshold values are uniformly distributed over $[-T, T]$, hence its minimal value is the inverse of (21).

unknown, we conclude that estimation in the distributed setting using threshold detection is strictly sub-optimal compared to the adaptive setting. In other words, the ability to choose the threshold values in an adaptive manner based on previous messages necessarily 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 Θ is given as the solution to the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & \inf_{\theta \in \Theta} \int \eta(t-\theta)\lambda(dt) \\ \text{subject to} \quad & \lambda(dt) \geq 0, \quad \int \lambda(dt) \leq 1. \end{aligned} \tag{20}$$

The objective function in (20) is concave in $\lambda(dt)$ and hence this problem can be solved using a convex program. We denote by $\kappa^\star$ the maximal value of (20) and by $\lambda^\star(dt)$ the density that achieves this maximum. Theorem 5 and Corollary 8 imply that for any $\theta \in \Theta$, the ARE of any local asymptotic minimax estimator (such as the ML estimator) from $\{B_i, i \in N\}$ of the form (17) with $t_i \sim \lambda^\star$ i.i.d. is at least $\kappa^\star$.

Figure 5 illustrates an approximating to $\lambda^\star(dt)$ obtained by solving a discretized version of (20) for the case when $f(x)$ is the normal density with variance σ^2 and $\Theta = [-T, T]$. The minimax asymptotic precision parameter $\kappa^\star$ obtained this way is illustrated in Fig. 6 as a function of half 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

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

Corollary 8 implies that the ARE under threshold values uniformly distributed is $\kappa_{\text{unif}}\sigma^2$.

When a prior π on Θ is provided, one may be interested in the threshold density λ minimizing the Bayes risk $R_n(\pi)$. The resulting minimization problem is

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

which is convex in λ since $x \to 1/x$, $x > 0$, is convex. Figure 7 illustrates the solution to (22) for the case of where $f(x)$ is the normal distribution and π is the uniform over $\Theta = [-T, T]$.

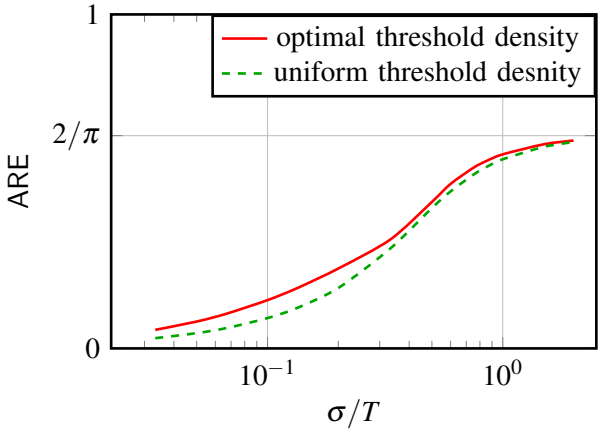

Fig. 6. Minimax ARE versus σ/T for $f(x) = \mathcal{N}(\theta, \sigma^2)$ and $\theta \in \Theta = [-T, T]$. The dashed curve (green) is the ARE under a uniform threshold density over Θ given by $K_{\mathsf{unif}}\sigma^2$, where κ_{unif} is given by (21).

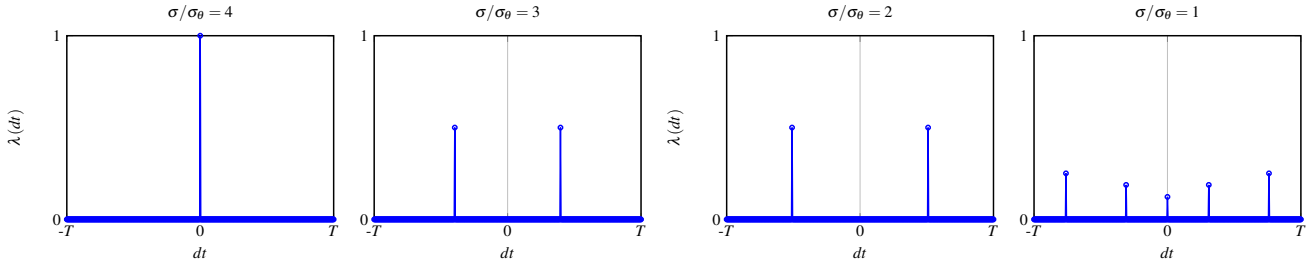

Fig. 7. Optimal threshold density $\lambda^\star(dt)$ that minimizes the asymptotic Bayes risk (22) for a uniform prior with $\sigma/\sigma_\theta = 1, 2, 3, 4$, where $\sigma_\theta^2 = T^2/3$ is the variance of the prior.

VI. Conclusions

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

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

Discussion? Future: high dim, memory constraints, adaptive (many-bit). Connection to constrained minimax. Quantized NNs?

Acknowledgments

WRITE THIS SECTION The work of A. Kipnis was supported in part by funding from the NSF under Grant No. DMS-1418362 and DMS-1407813.

Also... limitations of bespoke tools? Mis-specification bound. Quantized gradient communication could be useful (and generally optimal?). Strong similar to privacy. (And maybe 1-step-type ests.)

APPENDIX

A. *Fast convergence of uniform estimators under bit constraints*

Here we consider the uniform distribution as our location family, demonstrating that in the adaptive setting (ii) or even the one-step adaptive setting (ii'), constrained estimators can attain rates faster than the $1/\sqrt{n}$ rates regular estimands allow. Indeed, define $c(x) = -\log 2$ for $x \in [-1,1]$ and $c(x) = -\infty$ for $x \notin [-1,1]$. Then $f(x) = e^{-c(x)}$ is log-concave and symmetric, and we may consider the location family with densities $f(x-\theta)$. For notational simplicity, we assume we have a sample of size $2n$. We provide a proof sketch that there is a one-step adaptive estimator θ_n such that

$$\sup_{|\theta| \le \log n} P_\theta \left(|\theta_n - \theta| \ge \frac{8 \log n}{n^{3/4}} \right) \le \frac{2}{n^2}. \tag{23}$$

for all large n, and so (by the Borel-Cantelli lemmas), for any $\theta \in \mathbb{R}$ we have $P_\theta(|\theta_n - \theta| \le \sqrt{2\log n}/n^{3/4}$ eventually$) = 1$. This is of course faster than the $1/\sqrt{n}$ rates we prove throughout.

To prove inequality (23), we proceed in two steps, both quite similar. First, we define an initial estimator θ_n^{init}. Let $\varepsilon > 0$, which we will determine presently, though we will take $n\varepsilon \to \infty$ as $n \to \infty$, so that we may assume w.l.o.g. that $\theta \in [-n\varepsilon/2, n\varepsilon/2]$. Take the interval $[-n\varepsilon, n\varepsilon]$, and construct m thresholds at intervals of size $2n\varepsilon/m$; let the jth such threshold be

$$t_j \triangleq -n\varepsilon + \frac{2n(j-1)\varepsilon}{m}$$

Then we "assign" observations to each pair of thresholds, so that threshold j corresponds to observations $I_j \triangleq \{\frac{n(j-1)}{m} + 1, \ldots, \frac{nj}{m}\}$, of which there are n/m. For each index $i \in I_j$, we set

$$B_i = \begin{cases} 1 & \text{if } X_i \ge t_j \\ 0 & \text{otherwise.} \end{cases}$$

Then we simply set θ_n^{init} to be the maximal threshold for which $B_i = 0$ for all observations X_i corresponding to that threshold.

Let us now consider the probability that θ_n^{init} is substantially wrong. For notational simplicity, let $U_i = (1 + X_i - \theta)/2$, so that the U_i are uniform on $[0,1]$. First, note that we always have $\theta_n^{\text{init}} \ge \theta - \frac{2n\varepsilon}{m}$, because no observations will be below the appropriate threshold. Let $j^\star$ be the smallest index j for which $\theta \le t_{j^\star}$, and consider the index sets $I_{j^\star}$, $I_{j^\star+1}$, and so on. The event $\theta_n^{\text{init}} \ge t_{j^\star} + \frac{2n\varepsilon}{m}$ may occur only if for each of the n/m observations in the set $I_{j^\star+1}$, we have $U_i \ge \frac{n\varepsilon}{m}$. Thus,

$$P_\theta \left(\theta_n^{\text{init}} \ge t_{j^\star} + \frac{2n\varepsilon}{m} \right) \le \left(1 - \frac{n\varepsilon}{m}\right)^{\frac{n}{m}} \le \exp\left(-\frac{n^2\varepsilon}{m^2}\right).$$

Setting the number of bins $m = \sqrt{n}$, the resolution $\varepsilon = 2\log n/n$, we obtain $P_\theta(\theta_n^{\text{init}} \ge t_{j^\star} + 4\log n/\sqrt{n}) \le n^{-2}$. Thus we have

$$\sup_{|\theta| \le \log n} P_\theta \left(|\theta_n^{\text{init}} - \theta| \ge \frac{8 \log n}{\sqrt{n}} \right) \le \frac{1}{n^2}. \tag{24}$$

The second stage estimator follows roughly the same strategy, except that the resolution of the bins is tighter. In particular, let us assume that $|\theta_n^{\text{init}} - \theta| \le \frac{8\log n}{\sqrt{n}}$, which happens eventually by inequality (24). (We will assume this tacitly for the remainder of the argument.) Consider the interval $\Theta_n \triangleq \theta_n^{\text{init}} + [-\frac{16\log n}{\sqrt{n}}, \frac{16\log n}{\sqrt{n}}]$ centered at θ_n^{init}; we know that the interval includes $[\theta - \frac{8\log n}{\sqrt{n}}, \theta + \frac{8\log n}{\sqrt{n}}]$. Without loss of generality we assume $\theta_n^{\text{init}} = 0$. Following precisely the same discretization strategy as that for θ_n^{init}, we divide Θ_n into m equal intervals, with thresholds $t_j = -\frac{16\log n}{\sqrt{n}} + \frac{32(j-1)\log n}{m\sqrt{n}}$; let $\varepsilon_n = \frac{32\log n}{m\sqrt{n}}$ be the width of these intervals. Then following exactly the same reasoning as above, we assign indices $I_j = \{\frac{n(j-1)}{m} + 1, \ldots, \frac{nj}{m}\}$ and for $i \in I_j$, set $B_i = 1$ if $X_i \ge t_j$. We define θ_n to be the maximal threshold t_j for which $B_i = 0$ for all observations $X_i \in I_j$. Then following precisely the reasoning above, we have (on the event that $|\theta_n^{\text{init}} - \theta| \le \frac{8\log n}{\sqrt{n}}$)

$$P_\theta(|\theta_n - \theta| \ge 2\varepsilon_n) \le (1 - \varepsilon_n)^{\frac{n}{m}} \le \exp\left(-\frac{n\varepsilon_n}{m}\right) = \exp\left(-\frac{32\sqrt{n}\log n}{m^2}\right).$$

Set $m = 4n^{1/4}$ to obtain the claimed result (23).

B. Proof of Proposition 1

Denote by $D^\star$ the optimal MSE in the Gaussian CEO with L observers and under a total sum-rate $r = r_1 + \ldots + r_L$. An expression for $D^\star$ as a function of r is give as [46, Eq. 10]:

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

For the special case where $r = n$ and $L = n$, we have

$$n = \frac{1}{2}\log_2\left[\frac{\sigma_\theta^2}{D^\star}\left(\frac{D^\star n}{D^\star n - \sigma^2 + D^\star\sigma^2/\sigma_\theta^2}\right)^n\right]. \tag{26}$$

Consider the distributed encoding setting (iii) in the case where $f(x) = \mathcal{N}(0,\sigma^2)$ and the prior on Θ is $\pi = \mathcal{N}(0,\sigma_\theta^2)$. The Gaussian CEO problem of [26] with a unit bitrate $r_1 = \ldots = r_n = 1$ at each terminal and blocklength $k = 1$ reduces to our distributed setting (iii). Since $D^\star$ satisfying (26) describes the MSE in the CEO setting under an optimal allocation of the sum-rate $r = n$ among n encoders, it provides a lower bound to the minimal MSE in estimating θ in the distributed setting. By considering the limit $n \to \infty$ in (26), we see that

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

This implies Proposition 1.

C. Proof of Theorem 2

Consider the following two Lemmas:

Lemma 9: Let $f(x)$ be log-concave, symmetric, and differentiable density function such that Assumption 1 holds. For any $x_1 \geq \ldots \geq x_n \in \mathbb{R}$,

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

Lemma 10: Let X be a random variable with a symmetric, log-concave, and continuously differentiable density function $f(x)$ such that Assumption 1 holds. For a Borel measurable A set,

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

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

Lemma 9 is obtained as the special case $\delta = 0$ of Lemma 11. We now prove Lemma 10.

Proof of Lemma 10: When $f(x)$ is the normal density function, this lemma follows from [37, Thm. 3]. The proof below is based on a different techique than in [37], and is valid for any log-concave symmetric density satisfying Assumption A1.

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

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

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

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

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

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

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

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

so (28) can be rewritten as

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

It follows from Lemma 9 that for any $\theta \in \mathbb{R}$ and any choice of the intervals endpoints, (29) is smaller than

$$\max_{t \in \{a_j, b_j, j=1,\ldots,k\}} 4f^2(t) \le 4f^2(0),$$

where the last transition is due to Assumption 1. □

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

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

where $A_i \subset \mathbb{R}$ is a Borel measurable set. Consider the conditional distribution $P(B_1,\ldots,B_n|\theta)$ of $(B_1,\ldots,B_n)$ given θ. We have

$$P(B_1,\ldots,B_n|\theta) = \prod_{i=1}^{n} P(B_i|\theta,B_1,\ldots,B_{i-1}), \tag{31}$$

where $P(B_i = 1|\theta,B_1,\ldots,B_{i-1}) = \mathbb{P}(X_i \in A_i)$, so that the Fisher information of $B_1,\ldots,B_n$ with respect to θ is given by

$$I_\theta(B_1,\ldots,B_N) = \sum_{i=1}^{n} I_\theta(B_i|B_1,\ldots,B_{i-1}), \tag{32}$$

where $I_\theta(B_i|B_{i-1},\ldots,B_1)$ is the Fisher information of the distribution of B_i given $B_1,\ldots,B_{i-1}$. From Lemma 10 it follows that $I_\theta(B_i|B_{i-1},\ldots,B_1) \le 4f^2(0)$. The Van Trees inequality [51], [52] now implies

$$\begin{aligned}\mathbb{E}\left[(\theta_n-\theta)^2\right] &\ge \frac{1}{\mathbb{E}\left[I_\theta(B_1,\ldots,B_n)\right]+I_0}\\ &= \frac{1}{\sum_{i=1}^{n} I_\theta(B_i|B^{i-1})+I_0}\\ &\ge \frac{1}{4f^2(0)n+I_0}.\end{aligned}$$

□

D. Isoperimetric Lemma

The following lemma is used in the proof Theorems 2, 5, and 7.

Lemma 11: Let $f(x)$ be log-concave, symmetric, and differentiable density function. Let $\delta \ge 0$. Assume that the function

$$\eta_\delta(x) \triangleq \eta^{1+\delta}(x)/f^\delta(x) = \frac{(f(x))^{2+\delta}}{(F(x)(1-F(x)))^{1+\delta}}$$

is non-increasing in $|x|$. For any $x_1 \ge \ldots \ge x_n \in \mathbb{R}^n$,

$$\frac{\left|\sum_{i=1}^{n}(-1)^{i+1}f(x_i)\right|^{2+\delta}}{\left|\sum_{i=1}^{n}(-1)^{i+1}F(x_i)\right|^{1+\delta}\left|1-\sum_{k=1}^{n}(-1)^{i+1}F(x_i)\right|^{1+\delta}} \le \max_i \eta_\delta(x_i). \tag{33}$$

In particular,

$$\frac{\left|\sum_{i=1}^{n}(-1)^{i+1}f(x_i)\right|^{2+\delta}}{\left|\sum_{i=1}^{n}(-1)^{i+1}F(x_i)\right|^{1+\delta}\left|1-\sum_{i=1}^{n}(-1)^{i+1}F(x_i)\right|^{1+\delta}} \le \eta_\delta(0) = 4^{1+\delta}f^{2+\delta}(0).$$

Proof of Lemma 11: Denote

$$\delta_n(x_1,\ldots,x_n) \triangleq \sum_{i=1}^{n} s_i f(x_i),$$

$$\Delta_n(x_1,\ldots,x_n) \triangleq \sum_{i=1}^{n} s_i F(x_i),$$

where $s_i \triangleq (-1)^{i+1}$. We use induction on $n \in \mathbb{N}$ to show that

$$\frac{|\delta_n(x_1,\ldots,x_n)|^{2+\delta}}{|\Delta_n(x_1,\ldots,x_n)\,(1-\Delta_n(x_1,\ldots,x_n))|^{1+\delta}} \le \max_i \eta_\delta(x_i). \tag{34}$$

Since

$$\eta_\delta(x) = \frac{|\delta_1(x)|^{2+\delta}}{|\Delta_1(x)(1-\Delta_1(x))|^{1+\delta}},$$

The case $n=1$ is trivial. Assume that (34) holds for all integers up to $n=N$ and for any $x_1 \geq \ldots \geq x_N$. Consider the case $n=N+1$. Let i^* be the index such that x_{i^*} has minimal absolute value among $x_1,\ldots,x_N$. The assumption on $\eta_\delta(x)$ implies that

$$\eta_\delta(x_{i^*}) = \max_i \eta_\delta(x_i).$$

Since the LHS of (33) is invariant to a sign flip of all $x_1,\ldots,x_{N+1}$, we may assume that x_{i^*} is positive without loss of generality. Set $x^* = x_{i^*}$ and let $k = i^* - 1$. In what follows, variables with subscript of non-positive index are ignored in summations and in lists of arguments to functions. Consider the function

$$g(y_1,\ldots,y_N) \triangleq g(y_1,\ldots,y_N|x^*,k) \tag{35}$$

$$\triangleq \frac{|\delta_{N+1}(y_1,\ldots,y_k,x^*,y_{k+1}\ldots,y_N))|^{2+\delta}}{|\Delta_{N+1}(y_1,\ldots,y_k,x_{i^*},y_{k+1}\ldots,y_N)(1-\Delta_{N+1}(y_1,\ldots,y_k,x^*,y_{k+1}\ldots,y_N)|^{1+\delta}}. \tag{36}$$

The LHS of (34) is obtained by taking $y_i = x_{k_i}$ where k_i is the ith element in $\{1,\ldots,N+1\}\setminus\{i^*\}$. It is therefore enough to prove that

$$\max_{(y_1,\ldots,y_N)\in A_N} g(y_1,\ldots,y_N) \leq \eta_\delta(x^*),$$

where

$$A_N(x^*,k) \triangleq \left\{(y_1,\ldots,y_N)\in\mathbb{R}^N : y_1 \geq y_k \geq x^* \geq -x^* \geq y_{k+1}\ldots \geq y_N\right\}.$$

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 with derivative

$$c'(x) \triangleq \frac{f'(x)}{f(x)}$$

that is non-increasing. We first prove the lemma under the assumption that $c'(x)$ is an injection, or, equivalently, that $c(x)$ is strictly decreasing for all $x\in\mathbb{R}$.

The maximal value of $g(y_1,\ldots,y_N)$ is attained for the same $(y_1,\ldots,y_N)\in A_N(x^*,k)$ that maximizes

$$\log(g)(y_1,\ldots,y_N) = (2+\delta)\log(\delta_N) - (1+\delta)\log(\Delta_N) - (1+\delta)\log(1-\Delta_N),$$

where in the last display and henceforth we suppress the arguments $y_1,\ldots,y_k,x^*,y_{k+1},\ldots,y_N$ of the functions δ_N and Δ_N. Within the interior of $A_N(x^*,k)$, all three expressions in (36) within an absolute value are positive. It follows that partial derivative of $\log(g)(y_1,\ldots,y_N)$ with respect to y_i within the interior of $A_N(x^*,k)$ is given by

$$\frac{\partial\log(g)}{\partial y_i} = \frac{(2+\delta)s_i f'(x_i)}{\delta_N} - \frac{(1+\delta)s_i f(x_i)}{\Delta_N} + \frac{(1+\delta)s_i f(y_i)}{1-\Delta_N}.$$

We conclude that the gradient of $\log(g)$ vanishes if and only if

$$c'(y_i) = \frac{f'(y_i)}{f(y_i)} = \frac{1+\delta}{2+\delta}\frac{\delta_N}{2}\left(\frac{1}{\Delta_N} - \frac{1}{1-\Delta_N}\right), \quad i=1,\ldots,N. \tag{37}$$

Since we assumed that $c'(x)$ is an injection, (37) is satisfied if and only if $y_1=\ldots=y_N$. In this case, $g(y_1,\ldots,y_N) = \eta_\delta(x_{i^*})$ if N is even. If N is odd and $y_1=\ldots=y_N>x^*$, then

$$g(y_1,\ldots,y_N) = \frac{|f(y_1)-f(x_{i^*})|^{2+\delta}}{|F(y_1)-F(x_{i^*})|^{1+\delta}\,|1-(F(y_1)-F(x_{i^*}))|^{1+\delta}}$$

which is bounded from from above by $\eta_\delta(x_{i^*})$ by the induction hypothesis. The case where N is odd and $-x^*\leq y_1 = \ldots = y_N$ is similar. We now consider the possibility that the maximum of $g(y_1,\ldots,y_N)$ is attained at the boundaries of $A_N(x^*,k)$. At boundary points for which $y_i=y_{i+1}$ for some i, the contribution of y_i and y_{i+1} to $g(y_1,\ldots,y_N)$ is zero and the induction assumption for $n=N-1$ implies that

$$g(y_1,\ldots,y_N) \leq \eta_\delta(x^*)$$

The remaining boundary points of $A_N(x^*,k)$ are covered by the following cases:

(1) $y_N \to -\infty$.
(2) $y_1 \to \infty$.

(3) $y_k = x_{i^*}$.
(4) $y_{k+1} = -x_{i^*}$.

For case (1),

$$g(y_1,\ldots,y_N) \to \frac{\left|\sum_{i=1}^{k} s_i f(y_i) + s_{i^*} f(x_{i^*}) - \sum_{i=k+1}^{N-1} s_i f(y_i)\right|^{2+\delta}}{\left|\sum_{i=1}^{k} s_i F(y_i) + s_{i^*} F(x_{i^*}) - \sum_{i=k+1}^{N-1} s_i F(y_i)\right|^{1+\delta} \left|1 - \left(\sum_{i=1}^{k} s_i F(y_i) + s_{i^*} F(x_{i^*}) - \sum_{i=k+1}^{N-1} s_i F(y_i)\right)\right|^{1+\delta}},$$

which is smaller than $\eta_\delta(x_{i^*})$ by the induction hypothesis. Similarly, under case (2),

$$g(y_1,\ldots,y_N) \to \frac{\left|\sum_{i=2}^{k} s_i f(y_i) + s_{i^*} f(x_{i^*}) - \sum_{i=k+1}^{N} s_i f(y_i)\right|^{2+\delta}}{\left|1 + \sum_{i=2}^{k} s_i F(y_i) + s_{i^*} F(x_{i^*}) - \sum_{i=k+1}^{N} s_i F(y_i)\right|^{1+\delta} \left|-\left(\sum_{i=2}^{k} s_i F(y_i) + s_{i^*} F(x_{i^*}) - \sum_{i=k+1}^{N} s_i F(y_i)\right)\right|^{1+\delta}},$$

$$= \frac{\left|-\sum_{i=2}^{k} s_i f(y_i) - s_{i^*} f(x_{i^*}) + \sum_{i=k+1}^{N} s_i f(y_i)\right|^{2+\delta}}{\left|1 - \left(-\sum_{i=2}^{k} s_i F(y_i) - s_{i^*} F(x_{i^*}) + \sum_{i=k+1}^{N} s_i F(y_i)\right)\right|^{1+\delta} \left|-\sum_{i=2}^{k} s_i F(y_i) - s_{i^*} F(x_{i^*}) + \sum_{i=k+1}^{N} s_i F(y_i)\right|^{1+\delta}}$$

which is smaller than $\eta_\delta(x_{i^*})$ by the induction hypothesis. Under case (3), the terms in δ_N and Δ_N corresponding to y_k and x_{i^*} cancel each other. As a result, $g(y_1,\ldots,y_N)$ reduces to an expression with $n = N-1$ variables hence this case is handled by the induction hypothesis. Finally, under case (4), set

$$d \triangleq s_k F(-x^*) + s_{i^*} F(x^*) = s_{i^*}\left(1 - 2F(-x^*)\right),$$

$$\sigma \triangleq \sum_{i=1}^{k-1} s_i f(y_i) - \sum_{i=k+1}^{N} s_i f(y_i).$$

and

$$\Sigma \triangleq \sum_{i=1}^{k-1} s_i F(y_i) - \sum_{i=k+1}^{N} s_i F(y_i).$$

We have

$$g(y_1,\ldots,y_N) =$$

$$= \frac{\left|\sum_{i=1}^{k-1} s_i f(y_i) - \sum_{i=k+1}^{N} s_i f(y_i)\right|^{2+\delta}}{\left|\sum_{i=1}^{k-1} s_i F(y_i) + d(x^*) - \sum_{i=k+1}^{N} s_i F(y_i)\right|^{1+\delta} \left|1 - \sum_{i=1}^{k-1} s_i F(y_i) - d(x^*) + \sum_{i=k+1}^{N} s_i F(y_i)\right|^{1+\delta}},$$

$$= \frac{|\sigma|^{2+\delta}}{|\Sigma + d|^{1+\delta}\,|1 - \Sigma - d|^{1+\delta}} = \frac{|\sigma|^{2+\delta}}{|\Sigma|^{1+\delta}\,|1-\Sigma|^{1+\delta}} \left|\frac{\Sigma(1-\Sigma)}{\Sigma(1-\Sigma) + d(1-2\Sigma) - d^2}\right|^{1+\delta}.$$

By the induction hypothesis,

$$\frac{|\sigma|^{2+\delta}}{|\Sigma|^{1+\delta}\,|1-\Sigma|^{1+\delta}} \le \eta_\delta(x^*),$$

hence it is left to show that

$$\frac{\Sigma(1-\Sigma)}{\Sigma(1-\Sigma) + d(1-2\Sigma) - d^2} \le 1.$$

Whenever $d > 0$,

$$\frac{\Sigma(1-\Sigma) + d(1-2\Sigma) - d^2}{\Sigma(1-\Sigma)} \ge 1 \Leftrightarrow 1 - 2\Sigma \ge d,$$

while for $d < 0$,

$$\frac{\Sigma(1-\Sigma) + d(1-2\Sigma) - d^2}{\Sigma(1-\Sigma)} \ge 1 \Leftrightarrow 1 - 2\Sigma \le d.$$

Therefore, it is enough to show that $\Sigma \le F(-x^*)$ if $s_{i^*} = 1$ and $\Sigma \ge F(-x^*)$ if $s_{i^*} = -1$. Indeed, if $s_{i^*} = 1$, then $s_{k+1} = -1$ and monotonicity of $F(x)$ implies that

$$\Sigma + d \le F(y_1) - F(y_k) + F(x^*) - F(-x^*) + F(y_{k+2}) - F(y_N),$$

and hence

$$\Sigma \leq 1 - F(x^*) = F(-x^*).$$

Similarly, if $s_{i^*} = -1$ then

$$1 - \Sigma \leq 1 - F(-x^*).$$

This conclude the proof in the case where $c'(x)$ is an injection.

In the case where $c'(x)$ is not necessarily strictly decreasing, we approximate $c(x)$ using another concave symmetric function whose derivative is always negative except, perhaps, at the origin. For $\alpha > 0$ consider the function $f_\alpha(x) = \kappa(\alpha) e^{\mathrm{sgn}(x)|c(x)|^{1+\alpha}}$, where $\kappa(\alpha)$ is chosen such that $f_\alpha(x)$ is a probability density function. Then $c_\alpha(x)$ is concave, symmetric, and differentiable with

$$c'_\alpha(x) \triangleq \frac{f'_\alpha(x)}{f_\alpha(x)} = (1+\alpha)|c(x)|^\alpha c'(x).$$

Now $c'_\alpha(x)$ is non-increasing since it is the derivative of a concave function. Furthermore, since $c(x)$ is non-constant on any interval and $c'(x)$ is non-increasing, $c'_\alpha(x)$ is non-constant on any interval hence an injection. It follows from the first part of the proof that, for any $\alpha > 0$,

$$\frac{(\delta_{n,\alpha})^2}{\Delta_{n,\alpha}(1-\Delta_{n,\alpha})} \leq \max_i \eta_\alpha(x_i), \tag{38}$$

where

$$\delta_{n,\alpha} \triangleq \sum_{k=1}^{n} (-1)^{k+1} f_\alpha(x_k),$$

$$\Delta_{n,\alpha} \triangleq \sum_{k=1}^{n} (-1)^{k+1} F_\alpha(x_k),$$

and

$$\eta_{\delta,\alpha}(x) \triangleq \frac{(f_\alpha(x))^{2+\delta}}{(F_\alpha(x)(1-F(x)))^{1+\delta}}.$$

The proof is completed by noting that

$$\lim_{\alpha\to 0} \frac{(\delta_{n,\alpha})^{2+\delta}}{(\Delta_{n,\alpha}(1-\Delta_{n,\alpha}))^{1+\delta}} = \frac{(\delta_n)^{2+\delta}}{(\Delta_n(1-\Delta_n))^{1+\delta}},$$

and, since the maximum is over a finite set,

$$\lim_{\alpha\to 0} \max_i \eta_{\delta,\alpha}(x_i) = \max_i \eta_\delta(x_i).$$

□

E. Proof of Theorem 3

The estimation algorithm of (8) and (10) is a special case of a more general class of estimation procedures given in [53] and [54]. The proof of Theorem 3 relies on various results from these works.

Proof of (i): Consider the following simplified version of [53, Thm. 4]:

Theorem 12: [53, 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

$$\theta_i = \theta_{i-1} + \gamma_i \varphi(X_i - \theta_{i-1}), \qquad \bar{\theta}_n = \frac{1}{n}\sum_{i=0}^{n-1} \theta_i, \tag{39}$$

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 (11).

(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

$$\left|\psi(x) - \psi'(0)x\right| \le K_2 |x|^{1+\lambda}. \tag{40}$$

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

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

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

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

$$\psi(x) = \mathbb{E}[\operatorname{sgn}(x+Z_1)] = \int_{-\infty}^{\infty} \operatorname{sgn}(x+z) f(z) dz$$
$$= \int_{-x}^{\infty} f(z)dz - \int_{-\infty}^{-x} f(z)dz.$$

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

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

it follows from Theorem 12 that

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

Proof of (ii): We first show that the estimator $\bar{\theta}_n$ is regular in the following sense: For $\theta \in \Theta$, $h \in \mathbb{R}$ and n large enough such that $\theta + h/\sqrt{n} \in \Theta$, let $\mathbb{P}_{\theta,n}$ be a product probability measure on $\mathbb{R}^n$ with density $f(x - \theta - h/\sqrt{n})$ in each of its n coordinates. Then

$$\sqrt{n}\left(\bar{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(h, \frac{1}{\eta(0)}\right), \tag{41}$$

under $\mathbb{P}_{\theta,n}$. In order to show (41) we use the following refinement of Theorem 12, proof of which is given in Subsection F below.

Theorem 13: Set $\Delta_i = \theta_i - \theta$ and $\bar{\Delta}_i = \frac{1}{n}\sum_{i=1}^{n} \Delta_i$. Assume that, in addition to Assumptions (i)-(iv) of Theorem 12, there exists K_1 and $\lambda > 0$ such that

$$\mathbb{E}\left[|\varphi(Z_1) - \varphi(x+Z_1)|\right] \le K_1 |x|^{\lambda}. \tag{42}$$

Then:

(i)

$$\sqrt{n}\bar{\Delta}_n = -\frac{1}{\sqrt{n}} \frac{1}{\psi'(0)} \sum_{i=1}^{n-1} \varphi(Z_i) + o_{p.n}(1). \tag{43}$$

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

(ii) If Z_1 has continuously differentiable density $f(x)$ with finite Fisher information for location σ_f^2, then for any converging sequence $h_n \to h$,

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

under the local alternative $Z_1, \ldots, Z_n \sim \mathbb{P}^n_{h_n/\sqrt{n}}$ with density $\prod_{i=1}^n f(x_i - h_n/\sqrt{n})$.

In our setting, we have

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

It follows that

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

and hence condition (42) is fulfilled. In addition, by anti-symmetry of $f'(x)$ around $x = 0$,

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

Theorem 13, applied to the setting of Theorem 3, implies (41).

Under the assumption that $f(x)$ is continuously differentiable with a finite Fisher information for location, the model $\{Z_n + \theta\}_{n\in\mathbb{N}}$ is differentiable in quadratic mean [49, Exm. 7.8] and hence local asymptotically normal (LAN) in the sense that

$$\log\left(\frac{\mathbb{P}_{\theta,n}(X^n)}{\mathbb{P}_\theta(X^n)}\right) = h\eta^{-1/2}(0)Z - \frac{1}{2}h^2\eta^{-1}(0) + o_{p,n}(1),$$

where $Z \sim \mathcal{N}(0,1)$ and $o_{p,n}(1) \to 0$ as $n \to \infty$ under $\mathbb{P}_\theta$. The proof is concluded since any regular estimator in LAN model satisfies (12) [55].

Proof of (iii): Consider the following result from [54]:

Theorem 14: [54, Thm. 2] Let

$$\begin{cases} U_n = U_{n-1} - \gamma_n \varphi(Y_n), & Y_n = g'(U_{n-1}) + Z_n \\ \bar{U}_n = \frac{1}{n}\sum_{i=1}^n U_n, & n = 1, 2, \ldots. \end{cases} \tag{44}$$

Assume that the function $g(x)$ is twice differentiable with a strictly positive and uniformly bounded second derivative. In particular, $g(x)$ is convex with a unique minimizer $x^\star \in \mathbb{R}$. Moreover, assume that the noises Z_n are uncorrelated and identically distributed with a distribution with a density for which the Fisher information exits. Let $\psi(x)$ and $\chi(x)$ be defined as in Theorem 12-(iii) and satisfy the conditions there. Assume in addition that $\chi(0) > 0$, condition (40) 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 (11) 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 14 with $g(x) = 0.5(x-\theta)^2$, $\varphi(x) = \operatorname{sgn}(x)$, $Z_n = \theta - X_n$. From (44) we have

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

so the estimator $\bar{U}_n$ equals to the one defined by (10) and (8). 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 14 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 14 implies that for any $\theta \in \mathbb{R}$,

$$V_n = \mathbb{E}\left[(\theta_n - \theta)^2\right] = \frac{1}{4nf^2(0)} + o(n^{-1}).$$

F. Proof of Theorem 13

Proof of (i): The proof of part (i) of Theorem 13 requires the following two additional results from [53]:

Lemma 15: [53, Lem. 2] Consider the process $\{\Delta_i^1\}_{i=0}^\infty$ defined by

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

Assume that $A > 0$ and condition (ii) of Theorem 12 holds. Then

$$\sqrt{n}\bar{\Delta}_n^1 = \frac{1}{\sqrt{n}}\sum_{i=0}^{n-1}\Delta_i^1 = \frac{\alpha_n\Delta_0^1}{\sqrt{n}\gamma_0} + \frac{1}{\sqrt{n}A}\sum_{i=1}^{n-1}\xi_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}w_i^n\xi_i, \tag{45}$$

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

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

Lemma 16: Under the conditions of Theorem 12,

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

almost surely.

Lemma 16 follows from the proof of Theorem 2 in [53].

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

Step I: The expansion (43) holds for the process $\{\bar{\Delta}_i^1\}_{i=1}^\infty$ defined as follows:

$$\Delta_i^1 = \Delta_{i-1}^1 - \gamma_i\psi'(0)\Delta_{i-1}^1 - \gamma_i\varphi(Z_i), \qquad \delta_0^1 = \Delta_0 \tag{46}$$

$$\bar{\Delta}_i^1 = \frac{1}{n}\sum_{i=0}^{n-1}\Delta_i^1. \tag{47}$$

In order to prove this claim, use Lemma 15 with $A = \psi'(0)$ and $\xi_i = -\varphi(Z_i)$. The first expression on the RHS of (45) goes to zero in variance. In addition,

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

We obtain

$$\sqrt{n}\bar{\Delta}_n^1 = -\frac{1}{\sqrt{n}}\frac{1}{\psi'(0)}\sum_{i=1}^{n-1}\varphi(Z_i) + o_{p.n}(1), \tag{48}$$

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

From (39) and (46), the difference $\delta_i = \Delta_i - \Delta_i^1$ satisfies the recursion

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

where $\delta_0 = 0$. Use Lemma 15 with $\xi_i = \psi'(0)\Delta_{i-1} + \varphi(Z_i) - \varphi(\Delta_{i-1} + Z_i)$ to obtain

$$\sqrt{n}\bar{\delta}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} \left(\frac{1}{\psi'(0)} + w_i^n\right) \xi_i \tag{49}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} \left(\frac{1}{\psi'(0)} + w_i^n\right) \left(\psi'(0)\Delta_{i-1} - \psi(\Delta_{i-1})\right) \tag{50}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} \left(\frac{1}{\psi'(0)} + w_i^n\right) \left(\psi(\Delta_{i-1}) + \varphi(Z_i) - \varphi(\Delta_{i-1} + Z_i)\right) \tag{51}$$

For the term (50) and using (40), there exists K_1 and K_2 such that

$$(50) \le K_1 \sum_{i=1}^{\infty} \frac{1}{\sqrt{i}} \left|\left(\psi'(0)\Delta_{i-1} - \psi(\Delta_{i-1})\right)\right|$$

$$\le K_2 \sum_{i=1}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{\sqrt{i}}.$$

by Lemma 16 shows that

$$\sum_{i=1}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{\sqrt{i}} < \infty, \tag{52}$$

hence the Kronecker lemma implies

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

For the term (51), set

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

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

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

By Markov's inequality, we have

$$\mathbb{E}\left[\mathbf{1}_{A_{n,a}} \mid \Delta_0, \ldots, \Delta_{n-1}\right] \le \frac{1}{a} \sum_{i=1}^{n-1} \frac{\mathbb{E}\left[|\varepsilon_i| \Delta_{i-1}|\right]}{\sqrt{i}}. \tag{53}$$

Using (40) and (42), there exists K' and $\lambda' > 0$ such that

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

Plugging this bound in (53) and using Lemma 16, we obtain

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

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

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

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

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

and hence

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

almost surely. The Kronecker lemma now implies

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

hence the term (51) satisfies

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

This conclude the proof of part (i).

Part (ii): Use (43) to write

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

where

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

From [49, Exm. 7.8], the location model $f(x-\theta)$ with continuously differentiable $f(x)$ is differentiable in quadratic mean. This fact implies the following expansion [49, Thm. 7.2]:

$$\log \frac{\mathbb{P}^n_{h_n/\sqrt{n}}}{\mathbb{P}^n_0}(Z_1, \ldots, Z_n) = \log \prod_{i=1}^{n} \frac{f(Z_i - h_n/\sqrt{n})}{f(X_i - \theta)} = hJ_n - \frac{1}{2} h^2 I_\theta + o_{p,n}(1), \tag{54}$$

for any converging sequence $h_n \to h$, where

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

We have

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

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

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

Le Cam's third lemma [49, Exm. 6.7] implies that under $\mathbb{P}^n_{h/\sqrt{n}}$,

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

□

G. *Proof of Theorem 5*

The log-probability mass distribution of $(B_1,\dots,B_n)$ is given by

$$\log \mathbb{P}_\theta(b_1,\dots,b_n) = \sum_{i=1}^n \left(\frac{b_i+1}{2} \log \mathbb{P}(X_i \in A_i) + \frac{1-b_i}{2} \log \mathbb{P}(X_i \notin A_i) \right), \quad b_i \in \{-1,1\}, \quad i=1,\dots,n.$$

Consequently,

$$\log \frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(b_1,\dots,b_n)}{\mathbb{P}_\theta(b_1,\dots,b_n)} = \sum_{i=1}^n \frac{b_i+1}{2} \log \frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)} + \sum_{i=1}^n \frac{1-b_i}{2} \log \frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i \notin A_i)}{\mathbb{P}_\theta(X_i \notin A_i)}. \tag{55}$$

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

$$A_i = \bigcup_{k=1}^{K_i} \left(t_{i,2k-1}, t_{i,2k}\right),$$

where $t_{i,1} < \dots < t_{i,K_i}$ and, with a slight abuse of notation, $t_{i,1}$ and t_{i,K_i} may take the values $-\infty$ or $+\infty$, respectively. Thus

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

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

$$\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(X_i \in A_i) = \mathbb{P}_\theta(X_i \in A_i) + \frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)\frac{h}{\sqrt{n}} + o(h^2),$$

and thus

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

Similarly, we have

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

From (55) we obtain

$$\log \frac{\mathbb{P}_{\theta+\frac{h}{\sqrt{n}}}(b_1,\dots,b_n)}{\mathbb{P}_\theta(b_1,\dots,b_n)} = \frac{h}{\sqrt{n}} \sum_{i=1}^n \left(\frac{b_i+1}{2}\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)} + \frac{1-b_i}{2}\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \notin A_i)}{\mathbb{P}_\theta(X_i \notin A_i)} \right)$$
$$- \frac{h^2}{2n}\sum_{i=1}^n \left(\frac{b_i+1}{2}\left(\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \in A_i)}{\mathbb{P}_\theta(X_i \in A_i)}\right)^2 + \frac{1-b_i}{2}\left(\frac{\frac{d}{d\theta}\mathbb{P}_\theta(X_i \notin A_i)}{\mathbb{P}_\theta(X_i \notin A_i)}\right)^2 \right) + o(h^2)$$

Noting that

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

the proof is completed by proving the following two claims:

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

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

Then

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

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

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

Then

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

Proof of Claim I: First note that

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

In addition,

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

Therefore

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

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

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

and

$$\frac{\sum_{i=1}^n \mathbb{E}\left[|U_i|^{2+\delta}\right]}{\left(\sum_{i=1}^n \mathbb{E}\left[U_i^2\right]\right)^{\delta}} = \frac{\frac{1}{n^{1+\delta}}\sum_{i=1}^n \mathbb{E}\left[|U_i|^{2+\delta}\right]}{\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[U_i^2\right]\right)^{\delta}}. \tag{56}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[|U_i|^{2+\delta}\right] < K(\delta) \tag{57}$$

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

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

We now use the fact that each A_i is a finite union of intervals. Under the assumption that there exists $\delta > 0$ such that $\eta^{1+\delta}(x)/f^{\delta}(x)$ is uniquely maximized by the origin and is non-increasing in $|x|$, Lemma 11 implies

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

$$\leq 4^{1+\delta}(f(0))^{2+\delta}.$$

It follows that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[|U_i|^{2+\delta}\right] \leq 4^{1+\delta}(f(0))^{2+\delta},$$

and thus the numerator of (56), as well as the entire expression, goes to zero. From Lyaponov's central limit theorem we obtain

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

Proof of Claim II: We have:

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

We conclude that:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[V_i] = L_n(A_1,\ldots,A_n) \to \kappa(\theta) \tag{58}$$

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 (58) exists. □

H. Proof of Theorem 7

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

$$\lim_{n\to\infty} L_n(A_1,\ldots,A_n;\theta) = \eta(0). \tag{59}$$

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

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

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

$$C_i = \cup_{k=1}^{K}\{a_{i,k}, b_{i,k}\}.$$

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

$$S_n(\theta,\varepsilon) \triangleq \{i \leq n : (\theta-\varepsilon, \theta+\varepsilon) \cap C_i \neq \emptyset\}$$

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

$$\begin{aligned}
&L_n(A_1,\ldots,A_n;\theta) \\
&= \frac{1}{n}\sum_{i\in S_n(\varepsilon,\theta)} \frac{\left(\sum_{k=1}^K f(\theta-b_{i,k}) - f(\theta-a_{i,k})\right)^2}{\sum_{k=1}^K \left(F(\theta-b_{i,k})-F(\theta-a_{i,k})\right)\left(1-\sum_{k=1}^K \left(F(\theta-b_{i,k})-F(\theta-a_{i,k})\right)\right)} \\
&+ \frac{1}{n}\sum_{i\notin S_n(\varepsilon,\theta)} \frac{\left(\sum_{k=1}^K f(b_{i,k}-\theta) - f(a_{i,k}-\theta)\right)^2}{\sum_{k=1}^K \left(F(\theta-b_{i,k})-F(\theta-a_{i,k})\right)\left(1-\sum_{k=1}^K \left(F(\theta-b_{i,k})-F(\theta-a_{i,k})\right)\right)} \\
&\overset{(a)}{\leq} \frac{\text{card}\left(S_n(\theta,\varepsilon)\right)}{n}\eta(0) + \frac{n-\text{card}\left(S_n(\theta,\varepsilon)\right)}{n}\eta(\varepsilon)
\end{aligned} \tag{60}$$

where (a) follows from Lemma 11 with $\delta = 0$, and the fact that for $i \in S_n(\theta,\varepsilon)$,

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

Unless $\text{card}\left(S_n(\theta,\varepsilon)\right)/n \to 1$, we get that (60), hence $L_n(A_1,\ldots,A_n;\theta)$, are bounded from above by a constant that is smaller then $\eta(0)$ in contradiction to the fact that $\theta \in \Xi$.

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

$$\text{card}(\cup_{i=1}^n \mathscr{B}_i) \leq 2nK. \tag{61}$$

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

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

for all $i = 1,\ldots,N$. However, $S_n(\theta_1,\varepsilon),\ldots S_n(\theta_N,\varepsilon)$ are disjoint, so the cardinallity of their union is at least $n\left(1-\frac{1}{2N}\right)N$ ~~which is grater than~~ $2nK + n/2$ in contradiction to (61).

References

[1] V. Lesser, C. Ortiz, and M. Tambe, Eds., *Distributed Sensor Networks: A Multiagent Perspective*. Kluwer Academic Publishers, 2003, vol. 9.

[2] D. Li, K. Wong, Y. Hu, and A. Sayeed, "Detection, classification and tracking of targets in distributed sensor networks," in *IEEE Signal Processing Magazine*, 2002, pp. 17–29.

[3] S. Fuller and L. Millett, *The Future of Computing Performance: Game Over or Next Level?* National Academies Press, 2011.

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

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

[6] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax optimal procedures for locally private estimation (with discussion)," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 182–215, 2018.

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

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

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

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

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

[12] T. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2300–2324, Oct 1998.

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

[14] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. Springer, 2009.

[15] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[16] L. Le Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[17] A. W. van der Vaart, *Asymptotic Statistics*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[18] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "One-bit matrix completion," *Information and Inference*, p. to appear, 2015.

[19] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 482–494, 2013.

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

[21] 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.

[22] 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.

[23] 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.

[24] ——, "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.

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

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

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

[28] 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.

[29] 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.

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

[31] A. Garg, T. Ma, and H. L. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *Advances in Neural Information Processing Systems 27*, 2014.

[32] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," in *Proceedings of the Forty-Eighth Annual ACM Symposium on the Theory of Computing*, 2016. [Online]. Available: https://arxiv.org/abs/1506.07216

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

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

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

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

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

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

[39] 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.

[40] 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.

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

[42] E. L. Lehmann and G. Casella, *Theory of Point Estimation, Second Edition*. Springer, 1998.

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

[44] 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.

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

[46] 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.

[47] S. Y. Efroimovich, "Information contained in a sequence of observations," *Problems in Information Transmission*, vol. 15, pp. 24–39, 1980.

[48] E. Aras, K. Lee, A. Pananjady, and T. A. Courtade, "A family of bayesian cramér-rao bounds, and consequences for log-concave priors," *CoRR*, vol. abs/1902.08582, 2019. [Online]. Available: http://arxiv.org/abs/1902.08582
[49] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.
[50] 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
[51] H. L. Van Trees, *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004.
[52] R. D. Gill and B. Y. Levit, "Applications of the van Trees inequality: a Bayesian Cramér-Rao bound," *Bernoulli*, pp. 59–79, 1995.
[53] 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.
[54] B. T. Polyak, "New stochastic approximation type procedures," *Automat. i Telemekh*, vol. 7, no. 98-107, p. 2, 1990.
[55] R. Beran, "The role of Hájek's convolution theorem in statistical theory," *Kybernetika*, vol. 31, no. 3, pp. 221–237, 1995.