Mean Estimation from Single-bit Measurements

Abstract

We consider the problem of estimating the mean of a normal distribution under the constraint that only a single bit per sample from this distribution is available to the estimator. We study the mean squared error risk in this estimation as a function of the number of samples, and hence number of bits, from this distribution. We consider an adaptive scheme in which each bit is a function of the current sample and the previously observed bits, as well as a fully distributed setting in which each bit is only a function of the current sample. For the adaptive scheme, we show that the optimal relative efficiency, compared to the standard mean estimator without bit limitation, is $\pi/2$. Namely, the number of samples required to attain the mean to a prescribed accuracy is only $\pi/2$ times larger due to the bit constraint. For the distributed scheme we consider the setting where each bit is obtained by comparing the current sample to a prescribed value. For this setting, we show that the expected asymptotic efficiency converges to $\pi/2$ provided that the variance of the distribution is high compared to the apriori uncertainty about the mean. Our results indicate that, rather surprisingly, parametric estimation from the coarsest form of quantization without significant penalty in the number of measurements is possible.

One-bit with constant input is important, since estimating when the parameter varies with time is always harder. Our result implies than the error in any estimation with one-bit per sample is at least $pi/2$ times the MSE with an unlimited number of bits per sample.

Check connection to the μ-sum problem (Kim-ElGamal Ch. 21) that (maybe ?) gives a lower bound, and can be reduced to the CEO.

I. Background

Processing information on multiple physical locations may impose communication constraints on the amount of information that can be shared among different units of the system. In this situation, the ability to estimate a particular parameter is dictated not only by quality of observations but also by the constraint on the communication rate of the sensor, which may be severely restricted due to power, bandwidth, or size of data. The question that we ask is to what extent a parametric estimation task is affected by such communication constraints, and what are the fundamental performance of estimation subject to these restrictions.

In this paper we provide a precise answer to this question in a limited setting: the estimation of the mean θ of a normal distribution with a known variance σ^2, under the constraint that only a single bit can be communicated to the estimator on each sample from this distribution. As it turns out, the ability to share information before communicating it dramatically affects the performance in recovering θ. We distinguish in particular among three different settings:

(i) *Centralized* encoding: all the n encoders confer and produce a single n bit message.
(ii) *Adaptive* or *sequential* encoding: the ith encoder observes the ith sample and the $i-1$ previously encoded observations.
(iii) *Distributed* encoding: the output of the ith encoder is only a function of the ith observation.

As far as information sharing is concerned, it is evident that settings (iii) is a more restrictive version of (ii) which is more restrictive than (i). We are concerned with the mean squared error (MSE) risk in estimating θ as the number of observations n goes to infinity. In particular, we are interested in the *relative efficiency* of estimators in settings (i)-(iii), defined as the ratio between their MSE to σ^2/n. The latter is the MSE attained by the sample mean, which is the minimal risk estimator when no constraints on the communication are imposed.

It is well known that the relative efficiency in setting (i) is 1 [1], [2], [3], i.e., asymptotically, there is no loss in performance due to the communication constraint in this setting. Indeed, under setting (i) the sample mean can be described with an error exponentially decaying in n, leading to MSE of $\sigma^2/n+o(n)$. In this work we show that a similar result does not hold even in setting (ii): the relative efficiency of any adaptive estimator is at least $\pi/2$. Namely, a single bit per sample communication constraint in the adaptive setting incurs a minimal penalty of at least $\pi/2 \approx 1.57$ compared to an unconstrained estimator, or to the optimal estimator in setting (i). In addition to this negative statement, we provide an estimator which asymptotically attains this minimal relative efficiency, i.e., we show that this lower bound is tight. Moreover, we show that even in the distributed setting (iii), a relative efficiency of $\pi/2$ is possible as long as σ^2 is large compared to the variance of the prior distribution on θ. These results are quite surprising since they show that, even under the coarsest form of quantization, an estimator based on $\approx 1.59n$ measurements achieves accuracy in MSE or confidence interval comparable to an estimator that uses n samples with infinite bit-precision.

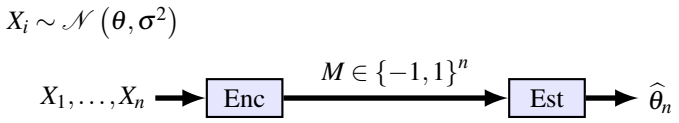

Fig. 1: Centralized encoding using one bit per sample on average.

Related Works

Shamir [**?**] showed that in hide-and-seek hypothesis testing problem, bitrate constraint may lead to a significant reduction in detection performance compared to the unrestricted case. This result is then applied to online learning, stochastic optimization and sparse PCA. Shamir's results shows that the penalty is in the dimension (d/R instead of $lod(d)$). In our setting the dimension is 1.

Baranuik et. al. [**?**] showed that adaptive one-bit measurements leads to exponential decay of MSE, but they did not consider noise!!

Statistical inference problems under communication constraints were considered in information theory in [4], [3], [1]. These works consider a multi-terminal source coding setting in which the ith terminal observes n samples from the distribution and is allotted nR_i bits to communicate its estimate. The main focus of these works is the difference between inference with communication constraint and the unconstrained vanilla statistical estimation setting, as the number of samples n goes to infinity subject to a total finite rate constraint. Our setting (i) can be seen as a special case of this setting where with a single terminal and $R_1 = 1$.However, as explained in [1, Sec. III], a single terminal in this setting always lead to the unconstrained inference performance. Indeed, when all samples are taken from the same distribution, the *type* of the sample [5] is a sufficient statistics for any inference task, and the latter can be described using a number of codewords polynomial in n regardless of the distribution of the samples. For this reason, attention is given in these works to inference problems involving multiple distributions observed at different locations and hence the results there do not apply to our setting. In fact, with a full access to the sample as in setting (i), the problem of encoding and estimating θ is reduced to the MSE attained by a scalar quantizer adjusted to the sufficient statistics of the sample. Finally, we note that setting (ii) includes as a special case the sigma-delta modulation (SDM) analog-to-digital conversion scheme with a constant input θ corrupted by Gaussian noise Z_i, as was considered in [6]. While it was shown there that the output of the modulator converges to the true constant input, the rate of this convergence was not analyzed and cannot be derived directly from the results of [6]. In fact, a corollary from the results in this paper we conclude that the rate of convergence of a SDM to a constant input signal is at most $\sigma^2\pi/2$ over the number of feedback iterations. Finally, the multiterminal source coding setting denoted as the CEO problem [7] can be seen as a generalization of our setting (iii), where multiple draws of θ and rounds of n bit communication is permitted. In particular, the minimal CEO distortion with 1 bit per terminal provides a lower bound on the MSE risk.

More related to our setting is the work [2] that considers the minimax MSE risk in estimating a vector of parameters from m machines, where each machine has access to n/m samples a limited number of bits to describe these samples. The main results of this work are lower bounds on the estimation error as a function of the total number of bits can be sent by each machine. In the terminology of [2], our settings (ii) and (iii) fall under the category of *interactive*, and *independent* communication protocols, respectively. However, unlike in [2], we focus on achievable schemes and tight bounds in n for the asymptotic MSE risk, rather than its scaling rate.

The rest of this paper is organized as follows: the main problem is defined in Section II. Our main results are presented and discussed in Sections V and VI. Concluding remarks are given in Section VII.

II. Problem Formulation

Let X_i, $i = 1, \ldots, n$, be n independent samples from the normal distribution $P(X) = \mathcal{N}(\theta, \sigma^2)$ with mean θ and variance σ^2. We also assume that the mean θ is drawn once from a prior distribution $\pi(\theta)$ on the parameter space Θ, which is a closed interval of the real line. We moreover assume that $\pi(\theta)$ is absolutely continuous with respect to the Lebesgue measure with a bounded second moment. We denote the variance of θ with respect to $\pi(\theta)$ by σ_θ^2. The problem we consider is the estimation of the parameter θ under the following constraints on the communication between the samples $X_1, \ldots, X_n$ and a centralized estimator:

(i) The estimator at time n is only a function of the n messages $M^n = (M_1, \ldots, M_n)$ (Fig. 1).
(ii) For each $i = 1, \ldots, n$, the ith message M_i is a function of the sample X_i and the $i-1$ previous messages M^{i-1} (Fig. 2).
(iii) The ith message M_i takes only two possible values, say 1 and -1 (Fig. 3).

In other words, the only information on the sample $X^n(X_1, \ldots, X_n)$ available to the estimator is the messages M^n. In addition, each message M_i is a function from the real line into the set $\{-1, 1\}$ which is measurable with respect to the sigma algebra

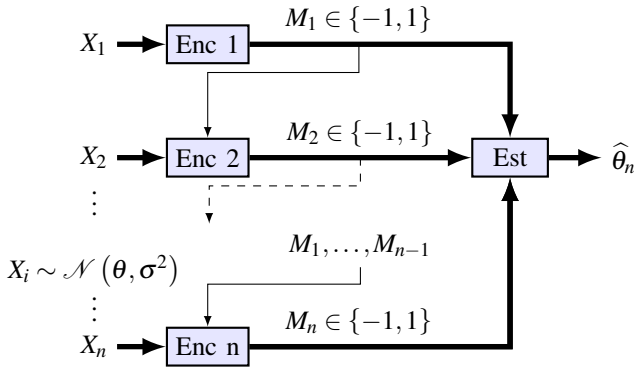

Fig. 2: Adaptive single-bit encoding: the ith encoder delivers a single bit message which is a function of its private sample X_i and the previous messages $M_1,\ldots,M_{i-1}$.

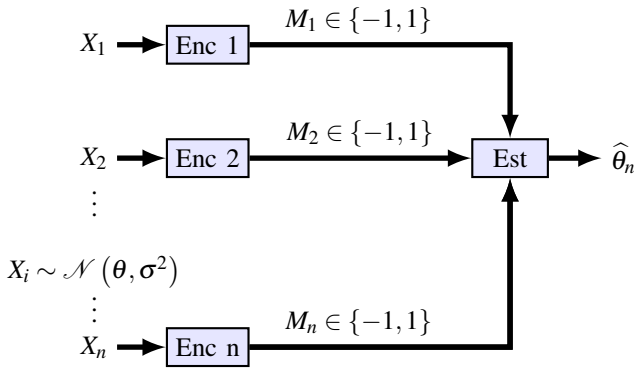

Fig. 3: Distributed single-bit encoding: the single-bit message produced by each encoder is only a function of its private sample X_i.

generated by M^{i-1} and X_i. The estimator, upon observing M^n, produces an estimate $\widehat{\theta}_n(M^n)$ of θ. A system describing the above scheme is illustrated in Fig. 2.

The performance of this estimation is measured by the mean squared error (MSE) risk function:

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

where the expectation is taken with respect to the distribution of X^n and the prior distribution $\pi(\theta)$.

The main problem we consider is the minimal value that can be attain in (1) as a function of n. Note that this minimization is the combination of the following two processes: (1) selecting the ith message M_i based on past messages and current observation X_i, and (2) estimating θ given messages M^n. We are interested in particular in the asymptotic relative efficiency of estimators $\widehat{\theta}_n$ for θ compared to the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Since the MSE attained by the latter is σ^2/n, this relative efficiency is defined as

$$\lim_{n\to\infty} \frac{n}{\sigma^2} R_n. \tag{2}$$

In addition to the notations defined above, we denote by $\phi(x)$ the standard normal density and by $\Phi(x)$ the standard normal cumulative distribution function. Prime denote derivative with respect to θ.

III. Relation to Remote Multiterminal Source Coding

In this section we draw a connection between our general problem of mean estimation from one-bit samples, to the remote multiterminal source coding problem [7], also known as the CEO problem. The setting of the CEO includes n encoders, each has access to a noisy version of a random source sequence. The ith encoder observes k noisy source symbols and transmit $R_i k$ bits to a central estimator.

Assuming that θ is drawn randomly once from a prior $\pi(\theta)$, mean estimation from one-bit under distributed encoding (setting (iii) in the Introduction) corresponds to the CEO setting with block length $k=1$, where the ith encoder observes

$$X_i = \theta + \sigma Z_i, \tag{3}$$

with $Z_1,\ldots,Z_n$ i.i.d. standard normal, and uses $R_i = 1$ bits to transmit its observation. As a result, the quadratic distortion in the optimal source coding scheme for the CEO with n terminals at rates $R_1 = \ldots = R_n = 1$ and Gaussian observation noise at each estimate of variance σ^2, provides a lower bound on the MSE distortion in estimating θ in the distributed encoding setting.

A closed-form expression for the MSE distortion in the CEO is known only for the case where the source sequence is sampled independently from the normal distribution [8]. By using the characterization of this minimal distortion as the number of terminals goes to infinity, we conclude the following:

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

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

Proof: We consider the expression [9, Eq. 10] that provides the minimal distortion $D^\star$ in the CEO with L observers and under a total sum-rate $R_\Sigma = R_1 + \ldots + R_L$:

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

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

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

The value of $D^\star$ that satisfies the equation above describes the MSE under an optimal allocation of the sum-rate $R_\Sigma = n$ among the n encoders. Therefore, $D^\star$ provides a lower bound to the CEO distortion with $R_1 = \ldots, R_n = 1$ and hence a lower bound to the minimal MSE in estimating θ in the distributed encoding setting. By considering $D^\star$ in (6) as $n \to \infty$, we conclude that

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

□

Since the lower bound (4) was derived assuming optimal allocation of the n bits among the encoders, it may seem as if a tighter bound can be obtained in our case by considering the CEO distortion with $R_1 = \ldots, R_n = 1$. However, an upper bound for the CEO distortion under this condition follows from [10, Prop. 5.2], and leads to

$$D_{CEO} \leq \left(\frac{1}{\sigma_\theta^2} + \frac{3n}{4\sigma^2 + \sigma_\theta^2}\right)^{-1} = \frac{4\sigma^2}{3n} + \frac{\sigma_\theta^2}{3n} + o(n^{-1}),$$

which is equivalent to (4) when σ_θ is small.

We conclude therefore that the difference between the MSE lower bound (4) and the actual MSE in the distributed encoding setting, is attributed exclusively to the ability to perform coding over blocks. Namely, the ability to consider $k > 1$ independent realizations of θ versus only one in ours. In other words, it is the ability to exploit the geometry of a high-dimensional product space, rather than the distributed nature of the problem, that distinguishes between the CEO distortion and the mean estimation from one-bit samples.

In the next section we show that the ARE in adaptive encoding setting does not exceeds $\pi/2$, and thus provides a tighter lower bound for the distributed encoding setting than $4/3$ of (4).

IV. Relation to Decentralized Detection

Once the decision rule of each sensor or encoder is fixed, the optimal estimation of θ is determined using the maximum a posteriori probability rule. Hence, the distributed setting is reduced to finding the optimal decision rule of each encoder. When the parameter space Θ is finite, the characterization of the optimal decision rules was considered by Tsitsiklist in [11]. It was shown there that if the cardinality of Θ is at most M and the probability of error criterion is used, than no more than $M(M-1)/2$ different decision rules are necessary in order to attain probability of error decreasing exponentially with the optimal exponent. That is, some beyond $M(M-1)/2$ some decision rules can be repeated. A version of this problem for the adaptive scenario was also considered in [12], where it was shown that with a specific two-stages feedback, asymptotically there is no gain in feedback compared to the fully distributed setting.

Wei et. al. [13] considered the case where a single threshold is applied to all quantizers, and showed how to chose the optimal threshold. ARE and minimax ARE of estimation from distributed quantized measurements that is based on Fisher information was considered in [14], [15], [16], [17]. However, the error criterion considered in these works is the Fisher information rather than a risk or ARE criterion. Moreover, these works provide optimality conditions only under the assumption the the same quantizer is used at each encoder.

V. Adaptive Estimation

The first main results of this paper, as described in Thm. 2 below, states that the ARE of any adaptive estimator cannot be lower than $\pi/2$. Next, we provide a particular adaptive estimation scheme and show in Thm. 3 that its efficiency is $\pi/2$. Finally, in Thm. 4, we provide an adaptive estimation scheme that is one-step optimal in the sense that at each step i, the message M_i that minimizes the MSE given X_i and the previous M^{i-1} messages is chosen. While it is not clear whether the efficiency of this last scheme is $\pi/2$, numerical simulations suggests that the MSE of this scheme times n also converges to $\pi/2$.

A. A lower bound on adaptive one-bit schemes

Our first results asserts that the ARE (2) of any adaptive estimation scheme is bounded from below by $\pi/2$, as follows from the following theorem:

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

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

where

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

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

Sketch of Proof: The main idea in the proof is to bound from above the Fisher information of any set of n single-bit messages with respect to θ. Once this bound is achieved, the result follows by using the van-Trees inequality [18, Thm. 2.13],[19] which bounds from below the MSE of any estimator of θ by the inverse of the expected value of the aforementioned Fisher information plus I_0. The details are given in the Appendix.

Next, we present an adaptive estimation scheme which attains ARE of $\pi/2$.

B. Asymptotically optimal estimator

Consider the following estimator $\widehat{\theta}_n$ for θ: set

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

where $\{\gamma_n\}_{n=1}^{\infty}$ is any strictly positive sequence satisfying

(i) $\frac{\gamma_n-\gamma_{n+1}}{\gamma_n} = o(\gamma_n)$

(ii) $\sum_{n=1}^{\infty}\frac{\gamma_n^{(1+\lambda)/2}}{\sqrt{n}} < \infty$ for some $0<\lambda\leq 1$.

(e.g. $\gamma_n = n^{-\beta}$ for $\beta\in(0,1)$). The nth step estimation is defined by

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

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

Theorem 3: The sequence $\widehat{\theta}_n$ of (8) satisfies

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \pi\sigma^2/2\right).$$

Proof: The asymptotic behavior of (8) is a special case of [20, Thm. 4]. The details are provided in the Appendix.

In other words, Thm. 3 implies that the estimator $\widehat{\theta}_n$, defined by (8) and (7), attains the minimal asymptotic efficiency as established by Thm. 2.

Note that θ_0 is not explicitly defined in equation (8). While a reasonable initialization is $\theta_0 = \mathbb{E}[\theta]$, Thm. 3 implies that the asymptotic behavior of the estimator is indifferent to this initialization. Thus, the optimal efficiency is attained regardless of the prior distribution on θ or the size of the parameter space Θ. Nevertheless, the bound in Thm. 2 suggests that the non-asymptotic estimation error can be greatly reduced whenever the location information I_0 is large. In contrast, the one-step optimal scheme presented in the following subsection exploits the prior information on θ provided by $\pi(\theta)$.

C. One-step optimal estimation

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

The following theorem determine the structure of the message that minimizes the next step MSE:

Theorem 4 (optimal one-step estimation): Let $\pi(\theta)$ be an absolutely continuous log-concave probability distribution. Given a sample X from the distribution $\mathcal{N}(\theta,\sigma^2)$, define

$$M = \operatorname{sgn}(X-\tau), \tag{9}$$

where τ satisfies the equation

$$\tau = \frac{m^-(\tau)+m^+(\tau)}{2}, \tag{10}$$

with

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

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

$$\mathbb{E}\left(\theta-\widehat{\theta}(M')\right)^2 \geq \mathbb{E}\left(\theta-\mathbb{E}[\theta|M]\right)^2, \tag{11}$$

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

Lemma 5: Let $f(x)$ be a log-concave probability density function. Then the equation

$$2x = \frac{\int_x^{\infty}uf(u)du}{\int_x^{\infty}f(u)du} + \frac{\int_{-\infty}^{x}uf(u)du}{\int_{-\infty}^{x}f(u)du} \tag{12}$$

has a unique solution.

Lemma 6: Let U be an absolutely continuous random variable with pdf $P(du)$. Then the one-bit message $M^\star\in\{-1,1\}$ that minimizes

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

is given by

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

where τ is the unique solution to

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

□

Remark 1: The value of the optimal threshold as given in Theorem 4 is different than the one given in [21, Eq. 5] for apparently the same problem. Indeed, it seems like Equation 4 there is erroneous.

Thm. 4 suggests the following adaptive encoding and estimation scheme:

- Initialization: set $P_0(t) = \pi(\theta)$.
- For $n \geq 1$:

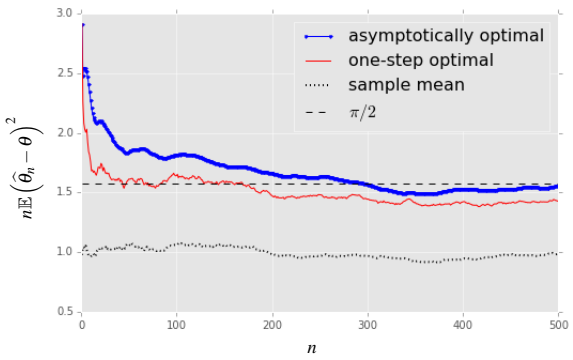

Fig. 4: Normalized empirical risk $n\left(\widehat{\theta}_n-\theta\right)^2$ versus number of samples n for 500 Monte Carlo trials. In each trial, θ is chosen uniformly in the interval $(-3,3)$.

1) Update the prior as

$$P_n(t) = P(\theta = t|M^n) \tag{13}$$
$$= \frac{P\left(\theta = t|M^{n-1}\right)P(M_n|\theta = t, M^{n-1})}{P(M_n|M^{n-1})}$$
$$= \alpha_n P_{n-1}(t)\Phi\left(M_n\frac{t-\tau_{n-1}}{\sigma}\right), \tag{14}$$

where α_n is a normalization coefficient that equals to

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

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

$$\theta_n = \mathbb{E}[\theta|M^n] = \int_{-\infty}^{\infty} tP_n(t)dt. \tag{15}$$

3) Solve equation (10) with the updated prior $P_n(t)$ instead of $P(d\theta)$. Note that since the standard normal cdf $\Phi(x)$ is log-concave, the updated prior $P_n(t)$ remains log-concave and thus a unique solution to (10) is guaranteed by Lem. 5.
4) Update the $(n+1)$th message as

$$M_{n+1} = \text{sgn}(X_{n+1}-\tau_n) \tag{16}$$

Since equation (10) has no analytic solution in general, it is hard to derive the asymptotic behavior of the estimator defined by (15) and (16). We conjecture, however, that it attains the asymptotic relative efficiency of $\sigma^2\pi/2$, as can be observed from the numerical simulation illustrated in Fig. 4. Also shown in Fig. 4 are the normalized MSE of the asymptotically optimal estimator defined by (7) and (8), as well as the MSE achieved by the sample mean for the same sample realization.

VI. Distributed Estimation

We now consider the case where a single bit message is available on each sample from the distribution, and this message is independent of the other messages. Moreover, we assume that each message is characterized by a threshold value, and that "1" is sent by the ith encoder whenever x_i is above this value.

A. Threshold Detection

We assume that each message is of the form

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

where $t_i \in \mathbb{R}$ is the *threshold* of the ith detector applied by the encoder on it sample X_i. In order to characterize the asymptotic behavior of estimation from these sequence of messages, we consider the *density* of n threshold values defined as:

$$\lambda_n([a,b]) = \frac{1}{n}\left|\mathscr{T}\cap[a,b]\right|,$$

where $\mathscr{T} = \{t_n, n \in \mathbb{N}\}$ is the collection of all threshold values. We further assume that λ_n converges (weakly) to a probability measure $\lambda(t)$ on $\mathbb{R}$.

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

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

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

$$\sum_{i=1}^{n} M_i \frac{\phi\left(\frac{t_i-\theta}{\sigma}\right)}{\Phi\left(M_i i^{\frac{t_i-\theta}{\sigma}}\right)}$$

Next, we show that under the assumptions above the sequence of messages M^n defines a local asymptotic normal (LAN) family of probability distributions. Therefore, the MSE in estimating θ from the messages M^n satisfies a local asymptotic minimax property with respect to the *precision* parameter of this LAN family [22]. Moreover, we conclude the ML estimator is locally asymptotic minimax.

Theorem 7: Consider the sequence of threshold detectors M^n with threshold density converges to a probability measure λ. The the following two statements hold:

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

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

where

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

and

$$\eta(x) = \frac{\phi^2(x)}{\Phi(x)(1-\Phi(x))}.$$

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

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} \mathscr{N}\left(0, \sigma^2/K(\theta)\right).$$

Remark 2: The condition that λ_n converges weakly to λ can be relaxed to the condition

$$\frac{1}{n}\sum_{i=1}^{n} \eta\left(\frac{t_i-\theta}{\sigma}\right) \xrightarrow{n\to\infty} \int \eta\left(\frac{t-\theta}{\sigma}\right)\lambda(dt)$$

(pointwise for any $\theta \in \Theta$).

Sketch of Proof: The proof shows that the probability distribution of the vector of messages M^n define a local asymptotic normal (LAN) family of probability distributions with precision parameter $K(\theta)/\sigma^2$. The statements in the theorem then follows from the local asymptotic minimax theorem of LAN families [22]. The details are given in the Appendix.

It follows from Thm. 7 that the asymptotic risk of the ML estimator is $\sigma^2/K(\theta)$, and that this risk is asymptotically minimal over all local alternative estimators for θ. Moreover, the relative efficiency of the ML in the threshold detection scheme equals $1/K(\theta)$. Since the thresholds density $\lambda(t)$ integrates to 1, and from the bound on the function $\eta(x)$ obtained in the proof of Lemma 10, it follows that

$$K(\theta) \le \sup_{t\in\mathbb{R}} \eta\left(\frac{t-\theta}{\sigma}\right)\int \lambda(dt) \le \frac{2}{\pi}.$$

This upper bound on $K(\theta)$ implies that the relative efficiency of any distributed estimator is at least $\pi/2$, a fact that agrees with the lower bound under adaptive estimation derived in Thm. 2. Unfortunately, this upper bound on $K(\theta)$ is attained whenever the density λ is the mass distribution at θ, which is unknown apriori. Therefore, in the section below we consider the asymptotic threshold density function λ that minimize the expected asymptotic risk $\mathbb{E}\left[\sigma^2/K(\theta)\right]$.

B. Worst case θ

Next, we consider the threshold density λ that minimizes the asymptotic risk over the worst choice of θ in the parameter space Θ, namely

$$\sup_{\lambda} \inf_{\theta\in\Theta} K(\theta) = \int \eta\left(\frac{t-\theta}{\sigma}\right)\lambda(dt), \tag{17}$$

where the supremum is over all probability measure λ on $\mathbb{R}$. By setting $\Theta = (-b\sigma, b\sigma)$ for some $b>0$, the supremum in (17) is independent of σ and we can write

$$K^\star(b) = \sup_{\lambda} \inf_{\tau\in(-b,b)} \int \eta\,(t-\tau)\,\lambda(dt). \tag{18}$$

In order to evaluate $K^\star(b)$, we consider the following variational optimization problem:

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

Since the objective function in (19) is concave in λ, this problem can be solved using a convex program. In fact, by discretizing the interval $(-b,b)$ using N_τ values $\tau_1,\ldots,\tau_{N_\tau}$ and the real line using N_λ values $\lambda_1,\ldots,\lambda_{N_\lambda}$, the discrete version of (19) is the following linear program (LP) in the variables $K\in\mathbb{R}$ and $\lambda\in\mathbb{R}^{N_\lambda}$:

$$\begin{aligned} \text{maximize} \quad & K \\ \text{subject to} \quad & K \leq \mathbf{H}\lambda \end{aligned} \tag{20}$$

$$\lambda \geq 0, \quad \mathbf{1}^T\lambda \leq 1, \tag{21}$$

where $H_{i,j} = \eta(t_i-\tau_j)$.

Remark 3: The number of variables in problem (20) is $N_\lambda+1$ and number of constraints is $1+N_\lambda+N_\tau$. Since an LP has an optimal solution at which the number of constraints for which equality holds is no smaller than the number of variables [**?**], there exists an optimal λ with support over no more than N_τ points.

The solution $\lambda^\star$ to (20) and the function $\mathbf{H}\lambda^\star$ are illustrated in Fig. 5, for various values of the support size of the parameters space b. The minimal asymptotic ML risk as a function of b is illustrated in Fig. 6.

C. Uniform Threshold Distribution

We now consider the case where the threshold distribution λ is uniform over the range $(-a,a)$ for some $a>0$, namely $\lambda(dt) = dt/(2b\sigma)$. The asymptotic ML risk in this case is σ^2 times the inverse of

$$K_{\mathsf{unif}} = \inf_{\theta\in(-b\sigma,b\sigma)} \frac{1}{2b\sigma}\int_{-b\sigma}^{b\sigma} \eta\left(\frac{t-\theta}{\sigma}\right)dt = \frac{1}{2b}\int_{-b}^{b}\eta(t-b)dt = \frac{1}{4b}\int_{-b}^{b}\eta(t)dt. \tag{22}$$

The value of $K_{\mathsf{unif}}/(2b)$ is illustrated in Fig. 6.

D. Asymptotic Bayes Risk

Consider problem of minimizing the expected asymptotic risk $\mathbb{E}[\sigma^2/K(\theta)]$ over all probability measures $\lambda(dt)$. This optimization problem can be written as follows:

$$\begin{aligned} \text{minimize} \quad & \sigma^2\int\frac{\pi(d\theta)}{K(\theta)} = \sigma^2\int\frac{\pi(d\theta)}{\int\eta\left(\frac{t-\theta}{\sigma}\right)\lambda(dt)}. \\ \text{subject to} \quad & \lambda(dt)\geq 0, \quad \int\lambda(dt)=1. \end{aligned} \tag{23}$$

Since the function $x\to 1/x$ is convex for positive values, (23) defines a convex optimization problem in λ whose solution depends on the the prior $\pi(\theta)$. The optimal asymptotic threshold density obtained as the solution to (23) for a normal prior is illustrated in Fig. 7, whereas Fig. 8 illustrates the corresponding expected risk. The expected asymptotic risk for the uniform distribution is illustrated in Fig. 9.

As can be seen from Fig. 7, for a normal prior with a small variance compared to the variance of the samples σ^2, the optimal density λ is a mass distribution. As it turns out from the following proposition, the choice of λ to be a mass distribution leads to a general upper bound on the expected asymptotic risk, and not only on the minimal value of (23).

Proposition 8: Let $\theta_0 = \mathbb{E}\theta$. Then

$$\mathbb{E}\frac{\sigma^2}{K(\theta)} \leq \sigma^2\int\frac{\pi(d\theta)}{\eta\left(\frac{\theta_0-\theta}{\sigma}\right)}, \tag{24}$$

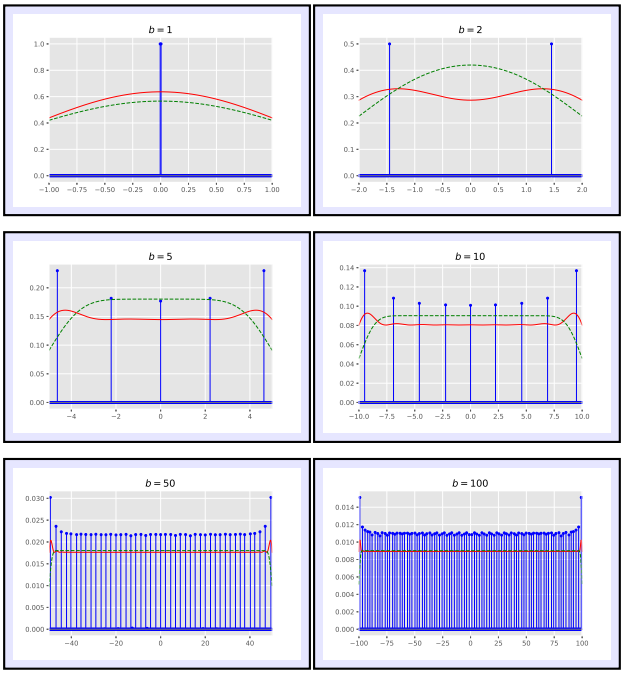

Fig. 5: Support of the optimal density $\lambda^\star$ that minimizes the asymptotic ML risk for the worst choice of $\theta \in (-\sigma b, \sigma b)$ where the parameter space is the interval $(-b,b)$, for $b = 1,2,5,10,50,100$. The continuous curve represents the inverse of the asymptotic risk under the optimal density and the dashed curves corresponds the inverse of the asymptotic risk under a uniform distribution over $(-a,a)$ with $a = b+2$.

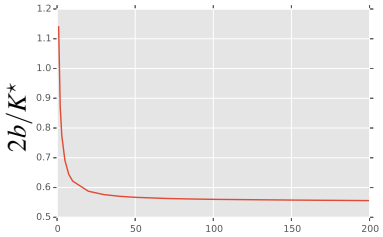

Fig. 6: Asymptotic risk of the ML estimator under an optimal choice of the threshold density λ versus the support of the parameter space $\Theta = (-b,b)$.

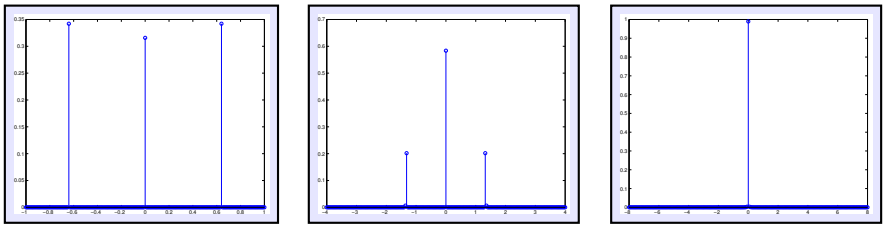

Fig. 7: The optimal asymptotic threshold density λ that minimizes the expected asymptotic ML risk (23) for a Gaussian prior for $\sigma/\sigma_\theta = 0.25, 1, 2$ (left to right), where σ_θ^2 is the variance of the prior.

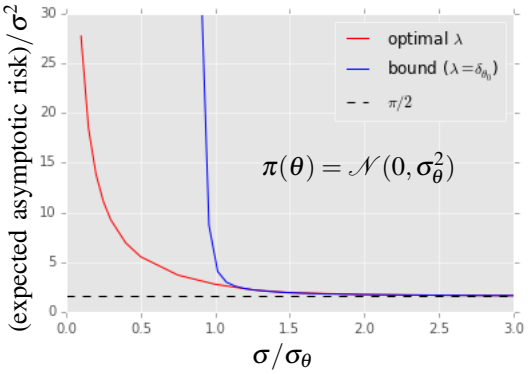

Fig. 8: Expected asymptotic risk $\mathbb{E}(1/K(\theta))$ in estimation from threshold detectors versus σ/σ_θ for a Gaussian prior with variance σ_θ^2. The optimal asymptotic threshold density λ (red) is obtained by solving (23). The bound (blue) is obtained using (24).

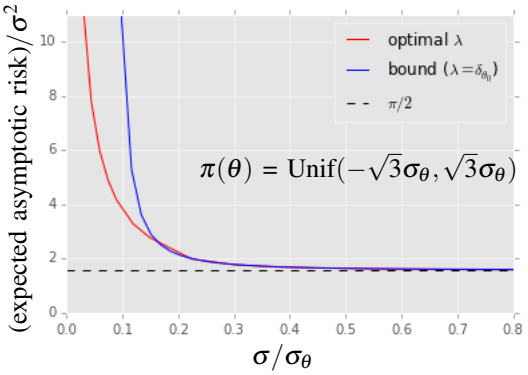

Fig. 9: Expected asymptotic risk $\mathbb{E}(1/K(\theta))$ in estimation from threshold detectors versus σ/σ_θ for a uniform prior with variance σ_θ^2. The optimal asymptotic threshold density λ (red) is obtained by solving (23). The bound (blue) is obtained using (24).

Proof: Since the function $x \to 1/x$ is convex for positive values, Jensen's inequality applied to the RHS of (23) leads to

$$\frac{\sigma^2}{K(\theta)} \leq \sigma^2 \int \frac{\lambda(dt)}{\eta\left(\frac{t-\theta}{\sigma}\right)}.$$

Therefore, the expected value of $\sigma^2/K(\theta)$ satisfies

$$\mathbb{E}\frac{\sigma^2}{K(\theta)} \leq \sigma^2 \int\int \frac{\pi(d\theta)\lambda(dt)}{\eta\left(\frac{t-\theta}{\sigma}\right)}. \tag{25}$$

The bound (24) follows by using $\lambda = \delta_{\theta_0}$. □

We note that since the function $1/\eta(x)$ is quasi-convex and symmetric, taking $\lambda = \lambda_{\theta_0}$ minimizes the RHS of (25) and leading to the tightest bound among the family bounds obtained after using Jensen's inequality.

The bound (24) is not trivial as long as the tail of the prior $\pi(\theta)$ is small enough such that the integral in the RHS of (24) is finite. In general, this bound is tight whenever the support of the optimal distribution is a mass distribution at the origin. The following proposition implies that this bound is always tight as σ/σ_θ goes to infinity.

Proposition 9: For any prior $\pi(\theta)$ for which

$$\int \frac{\pi(d\theta)}{\eta(\theta)} < \infty,$$

we have

$$\mathbb{E}[1/K(\theta)] = \frac{\pi}{2} + \left(\frac{\pi}{2} - 1\right)\left(\frac{\sigma_\theta}{\sigma}\right)^2 + o\left(\frac{1}{\sigma}\right)^3. \tag{26}$$

as $\sigma_\theta/\sigma \to 0$.

Proof: It is enough to prove that (26) holds for the upper bound (24). First note that the condition in the proposition implies that all the moments of $\pi(\theta)$ exists, and in particular its second and third. The results follows by expanding the function $1/\eta(x)$ in a power series around zero as

$$1/\eta(x) = \frac{\pi}{2} + \left(\frac{\pi}{2} - 1\right)x^2 + o(x^3).$$

□

E. Discussion

It follows from Prop. 9 that the asymptotic expected risk converges to $\pi/2$ when the ratio between the variance of the distribution of X_i to the variance of the prior tends to infinity. Therefore, we conclude that the minimal risk in the adaptive setting can be attained even in the distributed setting, as long as this ratio is large enough. Prop. 9 also bounds the convergence rate of the expected risk to its optimal value, where we note that higher order terms in (26) can be obtained by considering a higher order power series expansion of $1/\eta(x)$.

When the ratio σ^2/σ_θ^2 is large, the bound (24) is tight and the probability measure λ that minimizes the expected risk converges to a mass distribution at the expected value of θ according to its prior $\pi(\theta)$. Namely, all the encoders in this case report "1" whenever X_i is larger than this expected value and "-1" otherwise. Intuitively, an accurate estimation of θ in this case is possible only if a sufficient mix of "1"s and "-1"s is obtained from the sample. When σ_θ^2 is high, so is the probability of having a realization of θ away from its mean θ_0. In this case, a threshold detector located at θ_0 may have a strong bias toward one of "1" or "-1", unless the variance of the samples σ^2 is also relatively large. Indeed, as illustrated in Figs 7 and 8, when σ^2 is small compared to σ_θ^2, the bound in Prop. 8 is not tight and the optimal asymptotic threshold density λ is supported by multiple points. As illustrated in Figs 8 and Figs 9, when the ratio σ/σ_θ is small the expected asymptotic risk, and therefore the relative efficiency of any asymptotic local minimax optimal estimator, is higher than $\pi/2$ and cannot be uniformly bounded in this ratio regardless of the threshold distribution λ. For example, if the prior information on θ indicates that it lays in the interval $[-1,1]$ with a uniform probability and the variance of the observations is $\sigma^2 = 0.02$, then the asymptotic expected relative efficiency is ≈ 8.25. This number goes down to ≈ 3.02 for $\sigma^2 = 0.2$.

To summarize, we conclude that when the prior information on θ provides a reasonable localization of its value compared to the variance of the measurements, then the relative efficiency (2) of the ML estimator is $\pi/2$. On the other hand, the relative efficiency of any local asymptotic minimax estimator can be very high if the variance of the measurements is very small compared to the apriori uncertainty in θ, even under the optimal threshold distribution λ.

VII. Conclusions

We considered the relative efficiency in estimating the mean of a normal distribution from a single-bit encoding of each sample from this distribution. For the adaptive setting, we have shown that this minimal relative efficiency is $\pi/2$, namely there is a penalty factor of at least $\pi/2$ on the asymptotic mean square error risk in estimating the mean compared to an estimator that has full access to the sample. We also showed that this lower bound is tight by presenting an adaptive estimation procedure that attains it. In addition, we characterized the single-bit message that minimizes the next step MSE, and described an estimation procedure that is based on a sequence of one-step optimal messages. We leave open the questions whether this estimator attains the minimal efficiency and whether it leads to an estimation scheme which is globally optimal in the adaptive setting.

For the distributed setting, we considered the estimation from threshold detection under the assumption that density of the threshold values converges to a probability distribution. For this setting, we characterized the risk attained by the maximum likelihood estimator, and showed that this estimator is local asymptotic minimax. In addition, we showed that when the variance of the underlying normal distribution is high compared to the size of the variance of the prior on the unknown mean, the asymptotic relative efficiency in this setting converges to $\pi/2$. Namely, under this conditions, there is no loss in efficiency compared to the adaptive case. Nevertheless, this relative efficiency can be very high whenever the variance of the underlying distribution is very small compared to the apriori uncertainty in θ. We leave open the question whether there exists a distributed one-bit estimation scheme that achieves relative efficiency close to $\pi/2$ even when the apriori uncertainty in θ is high compared to the variance of the underlying normal distribution.

APPENDIX

In this appendix we provide detailed proofs of our main results as described in Section V.

Proof of Theorem 2

We first prove the following two lemmas:

Lemma 10: For any $x_1 \geq \ldots \geq x_n \in \mathbb{R}$, we have

$$\frac{\left(\sum_{k=1}^{n}(-1)^{k+1}\phi(x_k)\right)^2}{\left(\sum_{k=1}^{n}(-1)^{k+1}\Phi(x_k)\right)\left(1-\sum_{k=1}^{n}(-1)^{k+1}\Phi(x_k)\right)} \leq \frac{2}{\pi}. \tag{27}$$

Lemma 11: Let $X \sim \mathcal{N}(\theta,\sigma^2)$ and assume that

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

Then the Fisher information of M with respect to θ is bounded from above by $2/(\pi\sigma^2)$.

Proof of Lemma 10: We use induction on $n \in \mathbb{N}$. For the base case $n = 1$ we have

$$\eta(x) \triangleq \frac{\phi^2(x)}{\Phi(x)\left(1-\Phi(x)\right)}. \tag{28}$$

Taking the logarithm of (28) and differentiating, we conclude that any point x that maximizes (28) satisfies

$$x = \phi(x)\left(\Phi(x) - \frac{1}{2}\right).$$

However, since $x > \phi(x)\left(\Phi(x)-\frac{1}{2}\right)$ for all $x > 0$, the only point that satisfies the last condition is $x = 0$. At this point (28) equals $2/\pi$.

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

$$\begin{aligned} g(x_1,\ldots,x_N) &\triangleq 2\log\left(\sum_{k=1}^{N}(-1)^{k+1}\phi(x_k)\right) - \log\left(\sum_{k=1}^{N}(-1)^{k+1}\Phi(x_k)\right) - \log\left(1-\sum_{k=1}^{N}(-1)^{k+1}\Phi(x_k)\right) \\ &= 2\log\delta_N - \log\Delta_N - \log\left(1-\Delta_N\right), \end{aligned}$$

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

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

Using the fact that $\phi'(x) = -x\phi(x)$, we conclude that the gradient of g vanishes only if

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

In particular, the condition above implies $x_1 = \ldots = x_N$. If N is odd then for $x_1 = \ldots = x_N$ we have that (27) equals $\eta(x_1)$, which was previously shown to be smaller than $\pi/2$. If N is even, then for any constant c the limit of (27) exits as $(x_1,\ldots,x_N) \to (c,\ldots,c)$ and equals zero. Therefore, the maximum of (27) is not attained at this line. We now consider the possibility that (27) is maximized at the borders, as one or more of the coordinates of $(x_1,\ldots,x_N)$ approaches plus or minus infinity. For simplicity we only consider the cases where x_N goes to minus infinity or x_1 goes to plus infinity (the general case where the first m coordinates goes to infinity or the last m to minus infinity is obtained using similar arguments). Assume first $x_N \to -\infty$. Then (27) equals

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

which is smaller than $2/\pi$ by the induction hypothesis. Assume now that $x_1 \to \infty$. Then (27) equals

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

where $x'_m = x_{m+1}$. The last expression is also smaller than $2/\pi$ by the induction hypothesis. This proves Lemma 10.

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

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

where differentiation under the integral sign in (a) is possible since $\phi(x)$ is differentiable with absolutely integrable derivative $\phi'(x) = -x\phi(x)$. 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

$$\int_{A\setminus\cup_{j=1}^k I_j} dx < \varepsilon\sigma^2,$$

which implies that for any $\varepsilon' > 0$, the set A in (29) can be replaced by a finite union of disjoint intervals without increasing I_θ by more than ε'. It is therefore enough to proceed in the proof assuming that A is of the form

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

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

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

so (29) can be rewritten as

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

The proof of Lemma 11 is completed since it follows from 10 that for any $\theta \in \mathbb{R}$ and any choice of the intervals endpoints, (30) is smaller than $2/(\sigma^2\pi)$.

We now consider the proof of Thm. 2. In order to bound from above the Fisher information of any set of n single-bit messages with respect to θ, we first note that without loss of generality, each message M_i can be written in the form

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

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

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

where $P\left(M_i = 1|\theta, M^{i-1}\right) = \mathbb{P}(X_i \in A_i)$. We now prove the following Lemma:

Going back to (32), it follows that the Fisher information of M^n with respect to θ is given by

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

where $I_\theta(M_i|M^{i-1})$ is the Fisher information of the distribution of M_i given M^{i-1}. From Lemma 11 it follows that $I_\theta(M_i|M^{i-1}) \leq 2/(\pi\sigma^2)$. The Van Trees inequality [23], [19] now implies

$$\begin{aligned}\mathbb{E}\,(\theta_n - \theta)^2 &\geq \frac{1}{\mathbb{E}I_\theta(M^n) + I_0} \\ &= \frac{1}{\sum_{i=1}^{n} I_\theta(M_i|M^{i-1}) + I_0} \\ &\geq \frac{1}{2n/(\pi\sigma^2) + I_0}.\end{aligned}$$

□

Proof of Theorem 3

The algorithm given in (7) and (8) is a special case of a more general class of estimation procedures given in [20]. In particular, Theorem 3 is follows directly from the following simplified version of [20, Thm. 4]:

Theorem 12: [20, Thm. 4] Let

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

Define

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

where the Z_is are $\mathcal{N}(0,\sigma^2)$ and independent of each other, the sequence $\{\gamma_i\}_{i=1}^{\infty}$ satisfies conditions (i) and (ii) in Theorem 3, and $|\varphi(x)| \leq K_1(1+x)$ for some K_1. Define $\psi(x) = \mathbb{E}\varphi(x+Z_1)$, $\chi(x) = \mathbb{E}\varphi^2(x+Z_1)$ and assume that $\psi(0) = 0$, $x\psi(x) > 0$ for all $x \neq 0$, $\chi(x)$ is continuous at zero, and that $\psi(x)$ is differentiable at zero with $\psi'(0) > 0$. Moreover, assume that there exists K_2 and $0 < \lambda \leq 1$ such that

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

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

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

Using the notation in Theorem 12, we set $\varphi(x) = \operatorname{sgn}(x)$ and $Z_i = X_i - \theta$. We have:

$$\chi(x) = \mathbb{E}\varphi^2(x+Z_1),$$

so $\chi(0) = 1$. In addition,

$$\begin{aligned}\psi(x) = \mathbb{E}\operatorname{sgn}(x+Z_1) &= \int_{-\infty}^{\infty} \operatorname{sgn}(x+z)\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{z^2}{\sigma^2}} dz \\ &= \int_{-x}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{z^2}{\sigma^2}} dz - \int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{z^2}{\sigma^2}} dz.\end{aligned}$$

This leads to

$$\psi'(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{\sigma^2}} dz + \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{z^2}{\sigma^2}} dz,$$

so $\psi'(0) = \frac{2}{\sqrt{2\pi\sigma}}$. It is now easy to verify that the rest of the conditions in Theorem 12 are fulfilled for any $\lambda > 0$. Since

$$\frac{\chi(0)}{\psi'^2(0)} = \frac{\pi\sigma^2}{2},$$

Theorem 3 follows from Theorem 12. □

Proof of Theorem 4

In this subsection we prove lemmas 6 and 5 which lead to Theorem 4.

Proof of Lemma 6: Since any single-bit message $M(u) \in \{0,1\}$ is characterized by two decision region $A_1 = M^{-1}(1)$ and $A_{-1} = M^{-1}(-1)$, it follows that $\mathbb{E}[U|M(U)]$ assumes only two values: $\mu_1 = \mathbb{E}[U|M(U)=1]$ and $\mu_{-1} = \mathbb{E}[U|M(U)=-1]$. We claim that a necessary condition for $M(u)$ to be optimal is that the sets A_1 and A_{-1} are, modulo a set of measure $P(du)$ zero, the Voronoi sets on $\mathbb{R}$ corresponding to the points μ_1 and μ_{-1}, respectively. Indeed, assume by contradiction that for such an optimal partition there exists a set $B \subset A_1$ with $\mathbb{P}(U \in B) > 0$ such that $(b-\mu_1)^2 > (b-\mu_{-1})^2$. The expected square error in this partition satisfies:

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

so clearly, the partition $A_1' = A_1 \setminus B$, $A_{-1}' = A_{-1} \cup B$ attains lower error variance which contradicts the optimality assumption and proves our claim. It is evident that Voronoi partition of the real line corresponding to μ_1 and μ_{-1} is of the form $A_{-1} = (-\infty, \tau)$, $A_1 = (\tau, \infty)$ where the point τ is of equal distance from μ_1 and μ_{-1}, namely $\tau = \frac{\mu_1 + \mu_{-1}}{2}$. From these two conditions (which are a special case of the conditions derived in [24] for two quantization regions) we conclude that τ must satisfy the equation

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

□

Proof of Lemma 5: Any solution to (12) is a solution to $h^+(x) = h^-(x)$ where

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

and

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

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

$$1 - \frac{f(\tau) \int_{-\infty}^\tau f(x)(\tau - x) dx}{\left(\int_{-\infty}^\tau f(x) dx\right)^2}. \tag{34}$$

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

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

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

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

(where the second from the last step follows since $g'(x) \le g'(\tau)$ for any $x \le \tau$). If follows that (34) is non-negative and thus $h^-(x)$ is monotonically increasing. Since

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

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

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

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

Proof of Theorem 7

We will prove that the distribution of M^n defines a local asymptotic normal (LAN) family of probability distributions with *precision* parameter $K(\theta)/\sigma^2$. The statements in the theorem then follows from the local asymptotic minimax theorem of LAN families [22].

The probability mass distribution of M^n is given by

$$P_\theta(m^n) = \prod_{i=1}^n \Phi\left(m_i \frac{t_i-\theta}{\sigma}\right).$$

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

$$\log \frac{P_{\theta+\frac{h}{\sqrt{n}}}(M^n)}{P_\theta(M^n)} = \sum_{i=1}^n \log\left(\Phi\left(\frac{M_i}{\sigma}(t_i-\theta-h/\sqrt{n})\right)\right) - \sum_{i=1}^n \log\left(\Phi\left(M_i\frac{t_i-\theta}{\sigma}\right)\right). \tag{35}$$

Using the Taylor expansion of $\log\Phi(x)$, (35) can be written as

$$-h\sum_{i=1}^n \frac{M_i}{\sqrt{n}\sigma}\frac{\phi\left(M_i\frac{t_i-\theta}{\sigma}\right)}{\Phi\left(M_i\frac{t_i-\theta}{\sigma}\right)} - \frac{h^2}{2\sigma^2 n}\sum_{i=1}^n\left(\frac{\phi'\left(M_i\frac{t_i-\theta}{\sigma}\right)}{\Phi\left(M_i\frac{t_i-\theta}{\sigma}\right)} - \frac{\phi^2\left(M_i\frac{t_i-\theta}{\sigma}\right)}{\Phi^2\left(M_i\frac{t_i-\theta}{\sigma}\right)}\right) + o(1)$$

The proof is completed by proving the following two lemmas:

Lemma 13: For $i = 1,\ldots,n$ denote

$$U_i = -M_i\frac{\phi\left(M_i\frac{t_i-\theta}{\sigma}\right)}{\Phi\left(M_i\frac{t_i-\theta}{\sigma}\right)}.$$

Then

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

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

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

Then

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

Proof of Lemma 13: We have that $\mathbb{E}U_i = 0$. In addition,

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

and therefore

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

We now verify that the sequence $\{U_i, i = 1,2,\ldots\}$ satisfies Lyaponov's condition: for any $\delta > 0$ we have that

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

and

$$\frac{\sum_{i=1}^{n}\mathbb{E}\left|U_i\right|^{2+\delta}}{\left(\sqrt{\sum_{i=1}^{n}\mathbb{E}U_i^2}\right)^{2+\delta}}=\frac{\frac{1}{n^{1+\delta}}\sum_{i=1}^{n}\mathbb{E}U_i^{2+\delta}}{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}U_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}U_i^2\right)^{\delta}}. \tag{36}$$

As n goes to infinity, we have

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

so the numerator in (36) goes to zero. Since the denumerator in (36) goes to $(K(\theta))^{1+\delta}$), the entire expression goes to zero and hence Lyaponov's condition is satisfied. From Lyaponov's central limit theorem we conclude that

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

□

Proof of Lemma 14: Since $\phi'(x)=-x\phi(x)$, we have that

$$\mathbb{E}V_i=\frac{\phi^2\left(\frac{t_i-\theta}{\sigma}\right)}{\Phi\left(\frac{t_i-\theta}{\sigma}\right)\left(1-\Phi\left(\frac{t_i-\theta}{\sigma}\right)\right)},$$

and thus

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

It follows from Kolmogorov's law of large numbers (e.g. [25, Thm. 10.2.3]) that

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

□

REFERENCES

[1] T. Han and S.-I. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2300–2324, Oct 1998.

[2] 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.

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

[4] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, 1987.

[5] I. Csiszár, "The method of types [information theory]," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

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

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

[8] 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.

[9] 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.

[10] 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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] ——, "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.

[18] A. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. Springer New York, 2008.

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

[20] 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.
[21] P. Venkitasubramaniam, G. Mergen, L. Tong, and A. Swami, "Quantization for distributed estimation in large scale sensor networks," in *2005 3rd International Conference on Intelligent Sensing and Information Processing*, Dec 2005, pp. 121–127.
[22] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.
[23] H. L. Van Trees, *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004.
[24] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar 1982.
[25] P. K. Sen and J. M. Singer, *Large sample methods in statistics: an introduction with applications*. CRC Press, 1994, vol. 25.
[26] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inf. Theory*, vol. 9, no. 3, pp. 136–143, 1963.
[27] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback–i: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. 12, no. 2, pp. 172–182, Apr 1966.
[28] M. Varasteh, O. Simeone, and D. Gündüz, "Joint source-channel coding with one-bit ADC front end," *CoRR*, vol. abs/1604.06578, 2016. [Online]. Available: http://arxiv.org/abs/1604.06578

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

A. Statistical Inference under Communication Constraints

The works [4], [3], [1] consider various problems of statistical inference under multiterminal lossy compression. In the setting of [4], [3], [1], each terminal I observes n samples from the distribution and is allotted nR_i bits to communicate its estimate.The main focus of these works is the difference between inference with communication constraint and the unconstrained vanilla statistical estimation setting, as the number of samples n goes to infinity subject to a total finite rate constraint. Our case (i) can be seen as a special case of this setting where with a single terminal and $R_1 = 1$. However, as explained in [1, Sec. III], in this setting the unconstrained inference performance is always attained when all samples are taken from the same distribution. Indeed, in an i.i.d setting the *type* of the sample [5] is a sufficient statistics for any estimation task, and the latter can be described using a number of codewords polynomial in n regardless of the distribution of the samples. For this reason, attention is given in these works to inference problems involving multiple distributions observed at different locations, hence the name *multiterminal*.

B. Sigma-Delta Encoding

and covariance function [6, Eq. 25]

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

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

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

Source coding

With a full access to the sample as in setting (i), the problem of encoding and estimating θ is reduced to the MSE attained by a scalar quantizer adjusted to the sufficient statistics of the sample. Setting (ii) includes as a special case the sigma-delta modulation (SDM) analog-to-digital conversion scheme with a constant input θ corrupted by Gaussian noise Z_i, as was considered in [6]. While it was shown there that the output of the modulator converges to the true constant input, the rate of this convergence was not analyzed and cannot be derived directly from the results of [6]. As a corollary from the results in this paper we conclude that the rate of convergence of a SDM to a constant input signal is at most $\sigma^2\pi/2$ over the number of feedback iterations. Finally, the remote multiterminal source coding setting of [7] corresponds to the case of n rate-constrained encoders, each observing a noisy version of an information source. The difference between this setting and ours is that in ours the parameter of interest is not an information source. By assuming a prior distribution on this parameter, the CEO provides a lower bound on the estimation error in the fully distributed setting (iii). This lower bound can be attained if we were to consider the average error in multiple independent realizations of our problem rather than a single realization as we do here.

Relation to channel coding

The problem we consider is the estimation of the parameter θ by observing

$$X_i = \theta + Z_i, \quad i = 1, \ldots, n, \tag{37}$$

where the Z_is are independent, centered normal with variance σ^2. The problem of estimating θ from the X_is can be seen as the attempt to transmit a real number θ over an additive white Gaussian noise (AWGN) channel. In particular, setting (ii) can be seen as the feedback version of this transmission, which is closely related to [26], [27]. Using the joint source and channel coding theorem, the number of channel uses required to attain MSE of δ in estimating θ is given by

$$n \geq \frac{R_\theta(\delta)}{\frac{1}{2}\log\left(1+\sigma_\theta^2/\sigma^2\right)}, \tag{38}$$

where $R_\theta(\delta)$ is the rate-distortion function of the prior $\pi(\theta)$, and the variance of this distribution is σ_θ^2. From (38) we obtain the following lower bound on the asymptotic minimal MSE:

$$\delta \geq D_\theta\left(\frac{n}{2}\log\left(1+\sigma_\theta^2/\sigma^2\right)\right), \tag{39}$$

where $D_\theta(R)$ is the inverse of $R_\theta(D)$. On the other hand, even if $\sigma^2 \ll \sigma_\theta^2$ than distortion of at least $D_\theta(n)$ is expected. So that the lower bound is the maximum between (39) and $D_\theta(n)$. For example, when $\pi(\theta)$ is Gaussian, we have

$$\mathbb{E}\left(\theta-\widehat{\theta}\right)^2 \geq \max\left\{\sigma_\theta^2\left(1+\frac{\sigma_\theta^2}{\sigma^2}\right)^{-n}, \sigma_\theta^2 2^{-2n}\right\}.$$

Unfortunately, the bound in (39) is strict and cannot be attained unless in trivial cases. The reason is our setting θ is drawn once from a prior distribution, so the communication in (37) is limited to a repetition code. Moreover, in setting (ii) we have the additional constraint that the message is a causal function of the previous messages, where in case (iii) it is completely memoryless. Works in information theory which consider such setting includes AWGN channel with 1-bit ADC at the receiver [28].